Data Science Tools

R • Excel • KNIME • OpenOffice

C. Greco

Data Science Tools

LICENSE, DISCLAIMER OF LIABILITY, AND LIMITED WARRANTY

By purchasing or using this book (the "Work"), you agree that this license grants permission to use the contents contained herein, but does not give you the right of ownership to any of the textual content in the book or ownership to any of the information or products contained in it. *This license does not permit uploading of the Work onto the Internet or on a network (of any kind) without the written consent of the Publisher.* Duplication or dissemination of any text, code, simulations, images, etc. contained herein is limited to and subject to licensing terms for the respective products, and permission must be obtained from the Publisher or the owner of the content, etc., in order to reproduce or network any portion of the textual material (in any media) that is contained in the Work.

Mercury Learning and Information ("MLI" or "the Publisher") and anyone involved in the creation, writing, or production of the companion disc, accompanying algorithms, code, or computer programs ("the software"), and any accompanying Web site or software of the Work, cannot and do not warrant the performance or results that might be obtained by using the contents of the Work. The author, developers, and the Publisher have used their best efforts to insure the accuracy and functionality of the textual material and/or programs contained in this package; we, however, make no warranty of any kind, express or implied, regarding the performance of these contents or programs. The Work is sold "as is" without warranty (except for defective materials used in manufacturing the book or due to faulty workmanship).

The author, developers, and the publisher of any accompanying content, and anyone involved in the composition, production, and manufacturing of this work will not be liable for damages of any kind arising out of the use of (or the inability to use) the algorithms, source code, computer programs, or textual material contained in this publication. This includes, but is not limited to, loss of revenue or profit, or other incidental, physical, or consequential damages arising out of the use of this Work.

The sole remedy in the event of a claim of any kind is expressly limited to replacement of the book, and only at the discretion of the Publisher. The use of "implied warranty" and certain "exclusions" vary from state to state, and might not apply to the purchaser of this product.

Data Science Tools

R, Excel, KNIME, & OpenOffice

Christopher Greco

Mercury Learning and Information
Dulles, Virginia
Boston, Massachusetts
New Delhi

Publisher: David Pallai
Mercury Learning and Information
22841 Quicksilver Drive
Dulles, VA 20166
info@merclearning.com
www.merclearning.com
(800) 232-0223

C. Greco. *Data Science Tools: R, Excel, KNIME, & OpenOffice.*
ISBN: 978-1-68392-583-5

Library of Congress Control Number: 2020937123

202122321 Printed on acid-free paper in the United States of America

CONTENTS

PREFACE

Data Science is all the rage. There is a great probability that every book you read, every Web site that you visit, every advertisement that you receive, is a result of data science and, with it, data analytics. What used to be "statistics" is now referenced as data analytics or data science. The concepts behind data science are myriad and complex, but the underlying concept is that very basic statistical concepts are vital to understanding data. This book really has a two-fold purpose. The first is to review briefly some of the concepts that the reader may have encountered while taking a course (or courses) in statistics, while the second is to demonstrate how to use tools to visualize those statistical concepts.

There are several caveats that must accompany this book. The first one is that the tools are of a certain version, which will be described below. This means that there will undoubtedly be future versions of these tools that might perform differently on your computer. I want to be very clear that this performance does not mean that these tools will perform better. Three of these are free and open source tools, and, as such, perform as well as the group of developers dictate they will in their most current versions. In most instances, the tool will be enhanced in the newer version, but there might be a different "buttonology" that will be associated with newer functions. You will see the word "buttonology" throughout this book in the form of the mechanics of the tool itself. I am not here to teach the reader statistics or the different concepts that compose the topics of this book. I am here to show you how the free and open source tools are applied to these concepts.

Now it is time to get to the very heart of the text, the tools of data science. There will be four tools that will encompass the content of this book. Three are open source tools (FOSS or Free and Open Source), with one being COS (Common Off the Shelf) software, but all four will require some instruction in their use. These are not always intuitive or self-explanatory,

so there will be many screen pages for each mechanical function. I feel that visual familiarization trumps narrative, so you will not see a lot of writing, mostly descriptions and step-by-step mechanics. A few of you may be wondering how to practice these skills, and for those readers there is a final chapter that has several scenarios that allow the reader to apply what they have learned from these tools.

The organization of this book will be on the statistical concept, not the tool, which means that each chapter will encompass an explanation of the statistical concept, and then how to apply each tool to that concept. By using this presentation method, readers can go to the prescribed concept and use the tool most comfortably applied. Each section will be labeled accordingly, so they will both be in the table of contents and the index. This makes it simpler for individuals to see their choice of tools and the concepts they have to apply to those tools.

C. Greco
April 2020

ACKNOWLEDGMENTS

When I have done these in the past, I always mentioned my wife, children, and grandchildren, which to me was not just necessary but mandatory, because they are the ones that impact me every day. Thanks to my brothers and sisters, who always set the bar high enough for excellence, but not so high that I would injure myself getting over it. You all always provided me with the motivation to do better. Now, I have to add a few people that have helped me get this book into print and in the electronic media. The first and foremost is Jim Walsh of Mercury Learning, who took a risk having me write a book on free and open source applications. I truly believe in this book, and he trusted me to put my best foot forward, but in addition he made suggestions along the way that helped me to be a better writer and contributor to the bigger publishing picture. I truly appreciate all your help, Jim.

The other editors and writers at Mercury Learning are like looking at a Science, Technology, Engineering, and Math (STEM) Hall of Fame. I am truly honored and privileged to even have a book title with this noble group. Thanks for all the guidance.

Finally, my father, who told me in no uncertain words that I should never try to study "hard sciences" but stick with the "soft sciences," since I really stunk at math. Thanks, Dad, for giving me that incentive to pursue statistics and data analysis. I owe it all to you.

NOTES ON PERMISSIONS

- Microsoft Corporation screenshots fall under the guidelines seen here: *https://www.microsoft.com/en-us/legal/intellectualproperty/permissions/default.aspx.*
- OpenOffice screenshots fall under the guidelines seen here: *https://www.openoffice.org/license.html.*
- R / RStudio screenshots are permitted through the RStudio license and permission *https://rstudio.com/about/software-license-descriptions/.*
- R Foundation: *http://www.r-project.org.*
- Rattle screenshots are used with permission and also cited in:
  Graham Williams. (2011). Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Use R! New York, NY: Springer.
- KNIME screenshots are permitted through KNIME licensing and permission: *https://www.knime.com/downloads/full-license.*

CHAPTER 1

FIRST STEPS

1.1 INTRODUCTION TO DATA TOOLS

People have different motivations for pursuing what interests them. Ask someone about a car and they might say that they hate sedans, or love SUVs, or would never get anything other than an electric car, or maybe not get a car at all! People have different preferences and this does not change with data science (statistical) tools. Some people love Excel, to the point where they will use nothing other than that software for anything from keeping a budget to analyzing data. There are many reasons for maintaining dedication, but the main reason from my experience is familiarization with the object. A person who has only driven a stick shift loves the clutch, while those that have never driven a stick will not be as prone to prefer one with a manual gear shifter.

What reasons are there for preferring one software application to another? From my experience, there are three main points:

1. The software is easy to use
2. The software is available from anywhere
3. The software is updated regularly

Normally it could be put that software is inexpensive, but with the age of subscriptions software licenses are no longer perpetual, so a monthly payment is all that is necessary to ensure that the reader has access to the software as long as the subscription is current. Let's explore each point and elaborate.

1.1.1 The Software Is Easy to Use

If an analyst can select a few buttons and—voilà—the result appears, it is much easier than the "p" word. What is the "p" word? Programming! If an analyst has to do programming, it makes it difficult to get the result. Of course, analysts do not realize that once something is programmed, it is easier to apply that programming, but that is for another book at another time. The main point to get here is that Graphic User Interface (GUI) software seems to be preferred to programming software. The COS software is well known and also known to be easy to use. Some of the FOSS software will require more preparation.

1.1.2 The Software Is Available from Anywhere

In this age of cloud computing, being able to access software seems trivial. After speaking with colleagues, they like the fact that they can perform and save their work online so they will not lose it. They also like the fact that updates are transparent and performed while they are using the tool. Finally, they like the fact that they do not have to worry about installing the software and using their memory or disk space.

1.1.3 The Software Is Updated Regularly

The previous section covers this, so we will not elaborate. However, it is important to note that the tools that will be covered in this book are updated regularly. Unfortunately, the analyst will have to be the one to opt-in to the updates.

1.1.4 Summary

Now that we have covered why analysts prefer certain tools, a description of the ones covered in this book will be given in table form to simplify the presentation and (as stated previously) minimize the written word.

Software	Ease (1=Easy, 5=Hard)	Available	Updated
Excel	1	24/7	Company
R(RStudio / Rattle)	3	24/7	Analyst
KNIME	4	24/7	Analyst
OpenOffice	2	24/7	Analyst

1.2 WHY DATA ANALYSIS (DATA SCIENCE) AT ALL?

The world today is a compendium of data. Data exist in everything we do, whether it is buying groceries or researching to buy a house. There are so many free applets and applications that are available to us that we have a hard time saying no to any of these. As one reference put it, and this author has generalized, if what you are downloading is free, then *you* are the product (Poundstone, 2019). This is poignant, because free and open source (FOSS) is something that is commonly accessible and available to all of us. However, why do we need data science to analyze all of this information? In my knowledge, there are a number of reasons why data science exists. First, it exists to corral the trillions of bytes of information that is gathered by companies and government agencies to determine everything from the cost of milk to the amount of carbon emissions in the air. Forty years ago, most data were collected, retrieved, and filed using paper. Personal computers were a dream, and data science was called *archiving* or something similar. Moving toward electronic media, databases turned mounds of paper into kilo-, mega-, giga-, and even petabytes. But with that amount of data, analysis turned from pencil and paper into personal computers, or any computer. Analysts started to realize that dynamic software was the means to getting data analysis into a more usable form.

Data science grew out of this data analytic effort and uses conventional statistical methods coupled with the power of computing in order to make data science readily available to all private and public entities. With the power to analyze marketing, technical, and personnel data, companies now have the ability to calculate the probability of their product succeeding, or their revenue growing the next year. With the growth of data science comes the many tools that make data analytics a possibility.

1.3 WHERE TO GET DATA

Now that we have an introduction to the "why" of data science, the next subject is "where." Where do you get data to use with data science tools? The answer to that question, especially now, is that data is available on many web sites for analysis (Williams, 2011). Some of these web sites include:

1. *www.data.gov*, which contains pages of data from different government agencies. If you want to know about climate data, or census, or disease control, this is the place to go.

2. *www.kaggle.com*, which not only contains data, but has contests with existing data that anyone can join. One dataset contains the various data collected from the Titanic, including how many died or survived and all the demographics for analysis and correlation.
3. Just about any federal government agency. If you do not want to go to a general web site, then go to *www.cdc.gov*, *www.census.gov*, *www.noaa.gov*, or any separate government web site for data pertaining to things like Social Security (*www.ssa.gov*) or even intelligence (*www.nsa.gov*) for some historical data.

Now that you have the "whys" and "wheres" associated with data science and tools, you now move on to the next step—actually using the tools with real data. Besides, you have no doubt had enough of this stage setting.

The data for this book was retrieved at the site, *https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/*, which has the tornado tracking data for the United States from 1951 until 2018. The government agency NOAA stands for the National Oceanic and Atmospheric Agency. The recommendation is to download these files (as many as you like) and use them separately for the examples in the book. This book will focus on the 1951 tornado tracking to make it relatively straightforward. Once you download the data, then the next step is to import the data into your favorite statistical tool.

CHAPTER 2

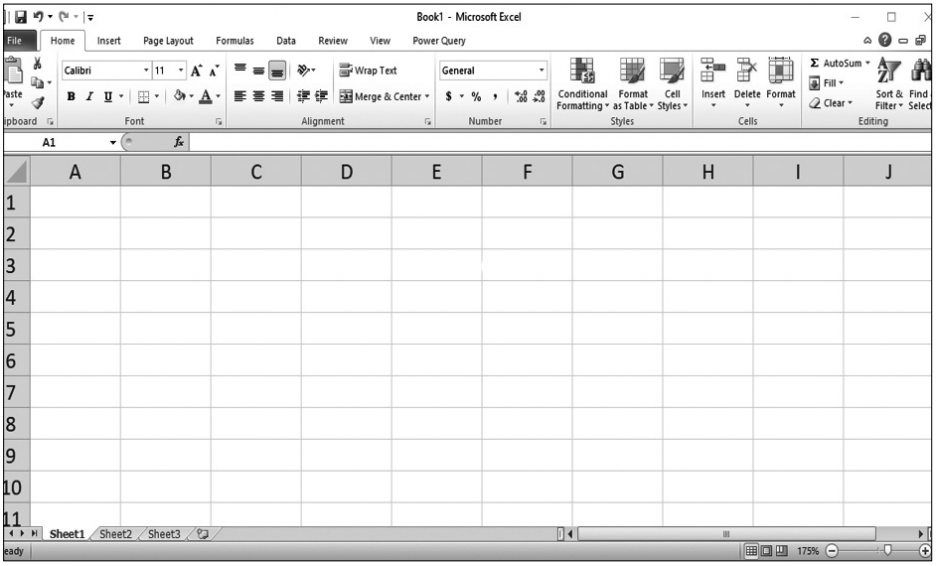
IMPORTING DATA

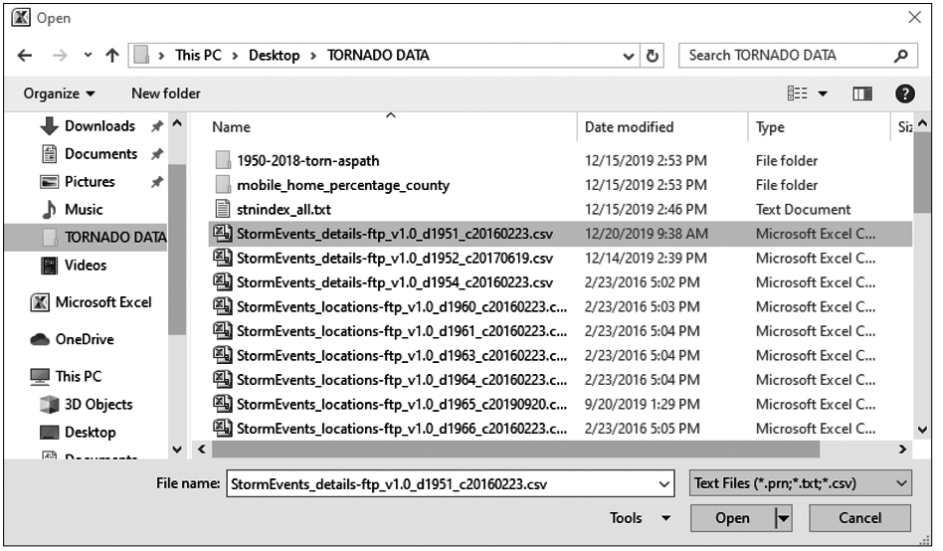
The first step to analyzing data is to import the data into the appropriate tool. This first section will show how to import data using each of the tools—Excel, R, KNIME, and OpenOffice. Since most analysts are familiar with Excel, Excel will be the first one addressed and then OpenOffice, since it is very close to Excel in functionality, for a good introduction to importing data.

2.1 EXCEL

The version for this text will be Microsoft Excel 2016, because that is the version that appears in many federal government agencies. As of the writing of this book, Excel 2019 is available but not used in public service at this point.

Importing data into Excel could not be easier. The file that has been downloaded is a Comma Separated Value (CSV) file, so to import the file into Excel, go to the file location and double-click on the file. The file will appear in Excel if the computer defaults to all spreadsheets going into Excel. If not, open Excel and choose "File" and "Open" to go to the file location and open the file. The following screens illustrate the operation.





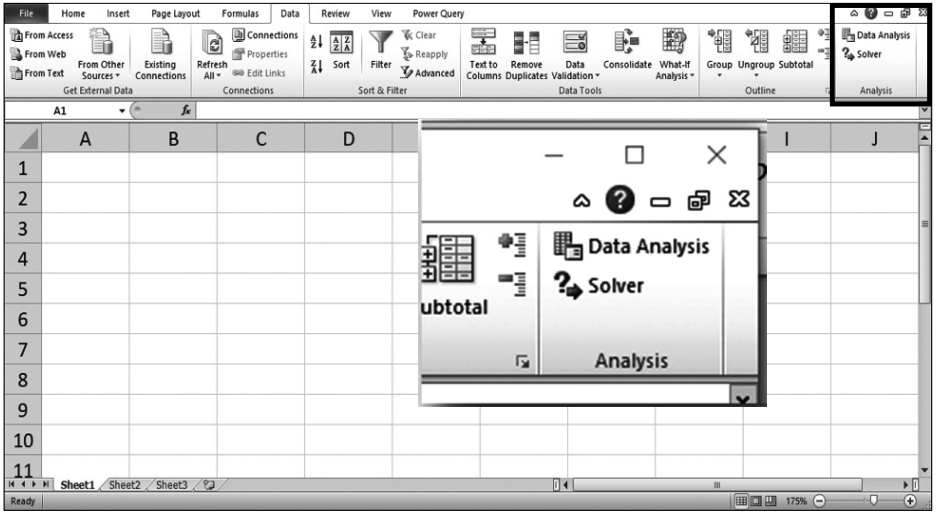
One caveat at this point with Excel. When opening a file, the default extension for Excel is the worksheet extension or "xlsx." If the worksheet is a CSV, then that default has to be changed, as demonstrated in the preceding process. Once the extension is changed, click "OPEN" and the spreadsheet

will appear in Excel. If the purpose is to stay as a CSV, then save it as such when you complete the work on the spreadsheet. Otherwise, save it as an "XLSX" file so that all the functionality of Excel remains with the spreadsheet as the analysis continues.

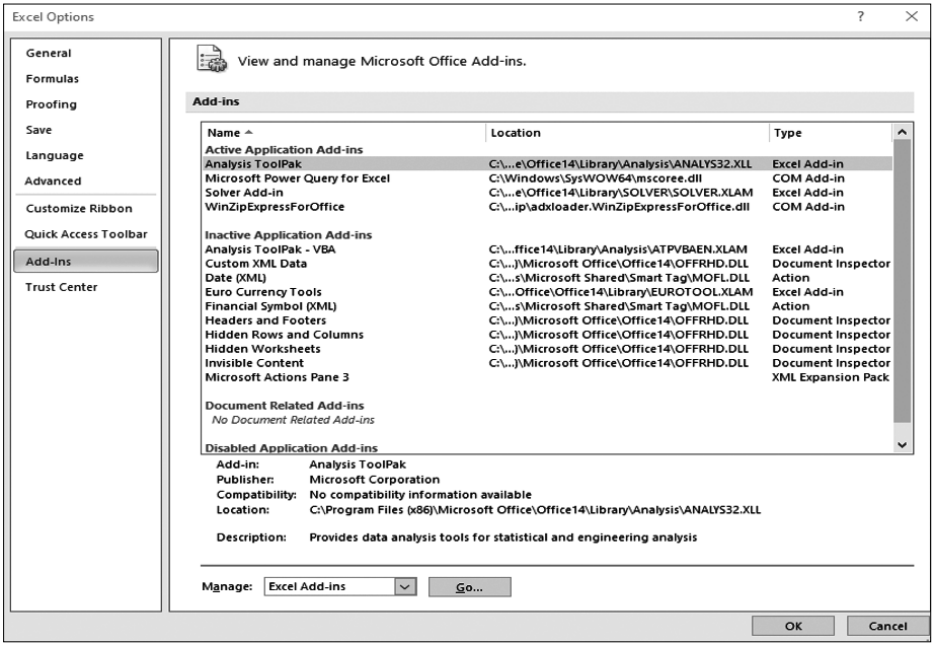
This is probably the easiest import for any of the applications presented because of the intuitive nature of Excel.

2.1.1 Excel Analysis ToolPak

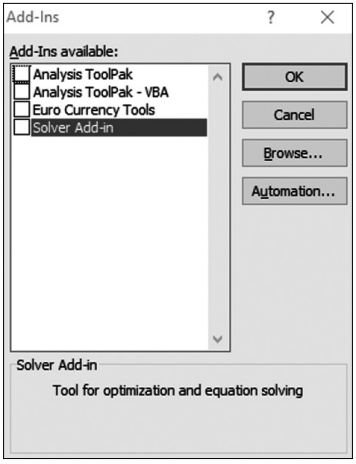
From this point forward, for any statistical analysis with Excel, we will be using the Analysis ToolPak, which will need to be installed as an add-on through Excel. If the Analysis ToolPak is already installed, it will show in the "Data" tab of Excel as shown here.



If the Analysis ToolPak is not showing in the Data toolbar, the analyst can add it simply by going to the "File" tab and choosing "Options" at the bottom of the left column. A screen will appear showing all the possibilities in the left column. The analyst chooses "Add-Ins" and the screen below will appear, showing all the add-ins that are available or not available. Take a second and look at the add-ins that are available as part of the Excel installation. There are a number of them, and they are very useful in data analytics. Take time to explore these add-ins to see how they can enhance your analysis, but in the meantime, finish installing the Analysis ToolPak add-in to complete this analysis.



When selecting Options, the next screen will reveal a number of choices in the left-hand side column. Choose "Add-Ins" and there will be a list of possible add-ins for Excel. Choose "Analysis ToolPak," which will at this point be in "Inactive Application Add-Ins," and go down to the bottom of the screen where it says "Manage:" to ensure that "Excel Add-Ins" is in the text box. Click on the "Go…" button and the following screen will appear.



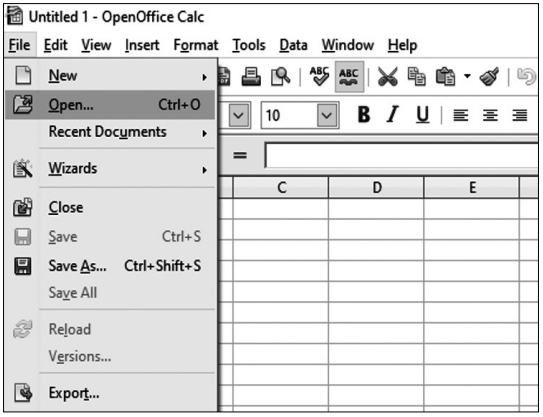
Click in the checkbox next to "Analysis ToolPak" in order to activate the add-in, and it will appear in the Excel toolbar. If it does not, try to close out of Excel and try the process again. It should work at that point. If it does not work after repeated attempts and the computer is a government computer, there may be a firewall in place that will prevent the use of this add-in. If the system administrator cannot provide the computer with access, there is a description at the end of this book that will demonstrate the buttonology to substitute for the Analysis ToolPak.

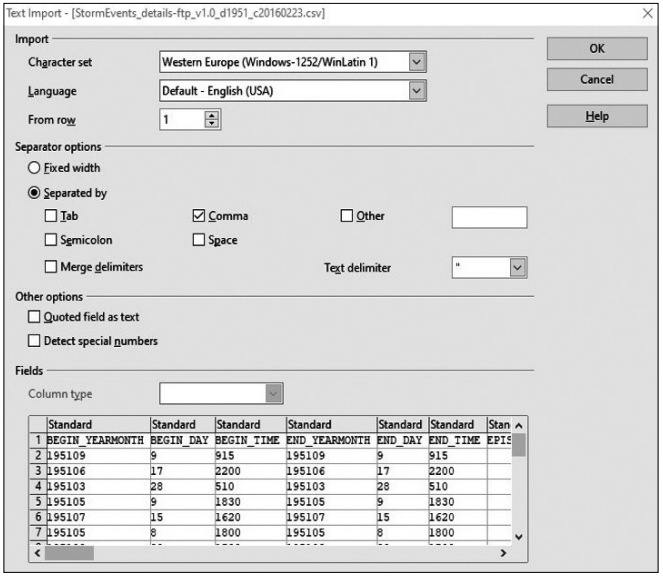
2.2 OPENOFFICE

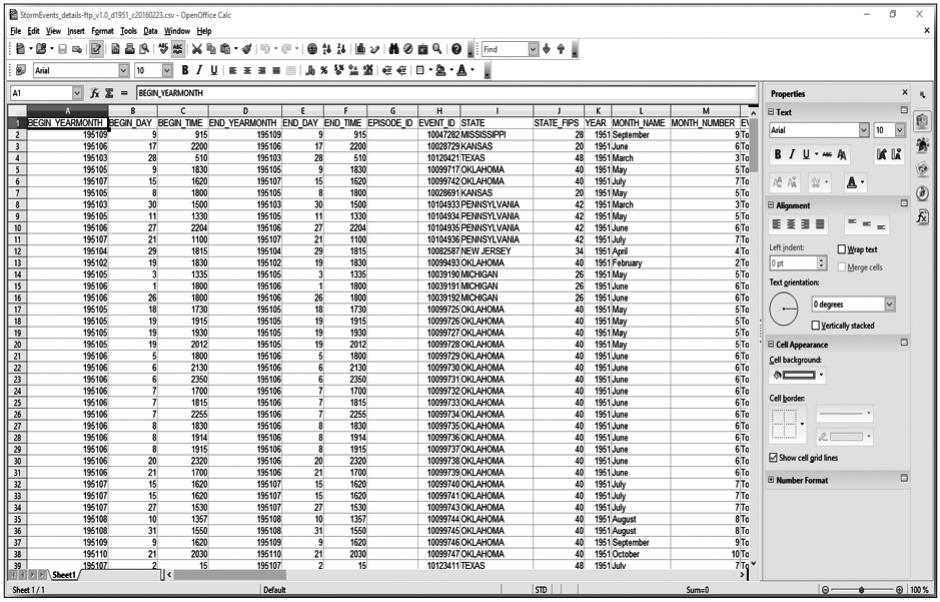
The first step to using OpenOffice is to download the software from the OpenOffice website (*www.openoffice.org*), which is relatively straightforward. The current version of the software is 4.1.7, which will be the version that we will be using in this book. When you install OpenOffice you do not have to install all the different functionalities, and in this instance you just need the spreadsheet program, so when you open the splash screen you will see the following:

At this point, select Spreadsheet and this screen will appear, which will look very much like Excel. In fact, having used Excel between 1998 and 2000, it will look very much like those versions. What this means is that the functionality is not exactly the same, but it will be everything you need for the statistics concepts in this book.

The first task will be to import data retrieved from the Internet. In this case it will be the data from a site that tracks tornados occurring in the United States from 1950–2018. This data will be imported by using the same technique as in Excel—through the "open" command in the File Menu as depicted here:







BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EPISODE_ID	EVENT_ID	STATE	STATE_FIPS	YEAR	MONTH_NAME	MONTH_NUMBER	EV
195109	9	915	195109	9	915		10047282	MISSISSIPPI	28	1951	September	9	To
195106	17	2200	195106	17	2200		10028729	KANSAS	20	1951	June	6	To
195103	28	510	195103	28	510		10120421	TEXAS	48	1951	March	3	To
195105	9	1830	195105	9	1830		10099717	OKLAHOMA	40	1951	May	5	To
195107	15	1620	195107	15	1620		10099742	OKLAHOMA	40	1951	July	7	To
195105	8	1800	195105	8	1800		10028691	KANSAS	20	1951	May	5	To
195103	30	1500	195103	30	1500		10104933	PENNSYLVANIA	42	1951	March	3	To
195105	11	1330	195105	11	1330		10104934	PENNSYLVANIA	42	1951	May	5	To
195106	27	2204	195106	27	2204		10104935	PENNSYLVANIA	42	1951	June	6	To
195107	21	1100	195107	21	1100		10104936	PENNSYLVANIA	42	1951	July	7	To
195104	29	1815	195104	29	1815		10082587	NEW JERSEY	34	1951	April	4	To
195102	19	1830	195102	19	1830		10099493	OKLAHOMA	40	1951	February	2	To
195105	3	1335	195105	3	1335		10039190	MICHIGAN	26	1951	May	5	To
195106	1	1800	195106	1	1800		10039191	MICHIGAN	26	1951	June	6	To
195106	26	1800	195106	26	1800		10039192	MICHIGAN	26	1951	June	6	To
195105	18	1730	195105	18	1730		10099725	OKLAHOMA	40	1951	May	5	To
195105	19	1915	195105	19	1915		10099726	OKLAHOMA	40	1951	May	5	To
195105	19	1930	195105	19	1930		10099727	OKLAHOMA	40	1951	May	5	To
195105	19	2012	195105	19	2012		10099728	OKLAHOMA	40	1951	May	5	To
195106	5	1800	195106	5	1800		10099729	OKLAHOMA	40	1951	June	6	To
195106	6	2130	195106	6	2130		10099730	OKLAHOMA	40	1951	June	6	To
195106	6	2350	195106	6	2350		10099731	OKLAHOMA	40	1951	June	6	To
195106	7	1700	195106	7	1700		10099732	OKLAHOMA	40	1951	June	6	To
195106	7	1815	195106	7	1815		10099733	OKLAHOMA	40	1951	June	6	To
195106	7	2255	195106	7	2255		10099734	OKLAHOMA	40	1951	June	6	To
195106	8	1830	195106	8	1830		10099735	OKLAHOMA	40	1951	June	6	To
195106	8	1914	195106	8	1914		10099736	OKLAHOMA	40	1951	June	6	To
195106	8	1915	195106	8	1915		10099737	OKLAHOMA	40	1951	June	6	To
195106	20	2320	195106	20	2320		10099738	OKLAHOMA	40	1951	June	6	To
195106	21	1700	195106	21	1700		10099739	OKLAHOMA	40	1951	June	6	To
195107	15	1620	195107	15	1620		10099740	OKLAHOMA	40	1951	July	7	To
195107	15	1620	195107	15	1620		10099741	OKLAHOMA	40	1951	July	7	To
195107	27	1530	195107	27	1530		10099743	OKLAHOMA	40	1951	July	7	To
195108	10	1357	195108	10	1357		10099744	OKLAHOMA	40	1951	August	8	To
195108	31	1550	195108	31	1550		10099745	OKLAHOMA	40	1951	August	8	To
195109	9	1620	195109	9	1620		10099746	OKLAHOMA	40	1951	September	9	To
195110	21	2030	195110	21	2030		10099747	OKLAHOMA	40	1951	October	10	To
195107	2	15	195107	2	15		10123411	TEXAS	48	1951	July	7	To

Now comes the cleaning and transforming of the data in preparation for analysis. However, in order to make this file available to other tools, it might be advantageous to save it as an Excel file, or even a text file. For those that like Comma Separated Value (CSV) files, most of the data that is found on many data sites seem to default to CSV files, so leaving this file in the CSV extension would be fine.

2.3 IMPORT INTO R AND RATTLE

Importing data into the R statistical application is relatively easy if the reader would download both the R and the RStudio applications. R can be found in the Comprehensive R Archive Network (CRAN) site for the R application (*https://cran.r-project.org/*), while RStudio can be found at *https://rstudio.com/products/rstudio/*. Both will need to be installed in order to make R less program-centric and a little more graphic user interface (GUI). For the purpose of this book, R will refer to version 3.6.2 and RStudio to version 1.2.5019. This will afford some standardization to the different screens and functions, but we have found that functionality may differ but has never decreased with later versions. For instance, "GGobi" is one function that does not seem to work with recent Rattle versions, but we have also found that

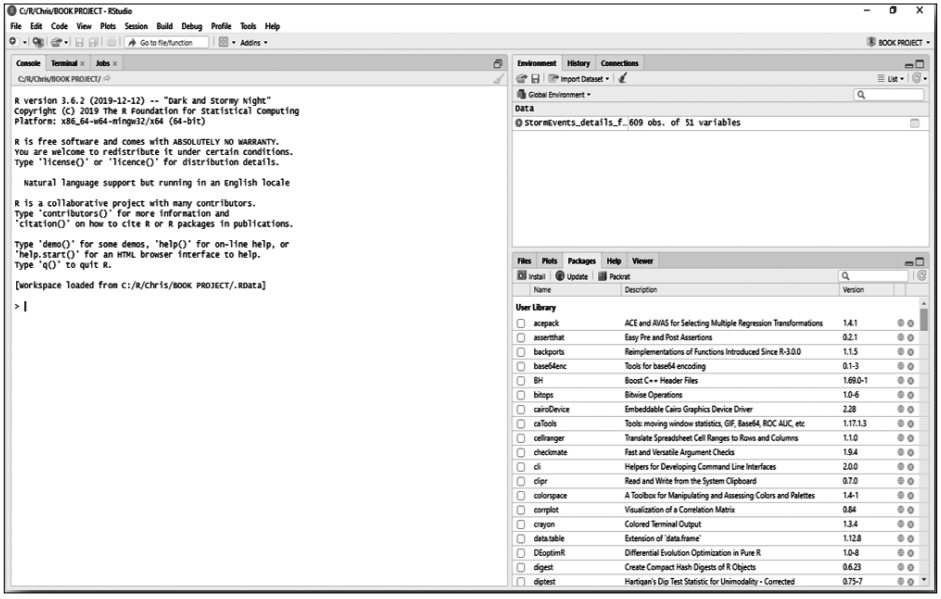
"GGRaptr" works just as well, so GGobi has been replaced, and there is some work to do on the analysts' part to get to that conclusion. In making these references, there is an important point that anyone using R must understand. GGRaptr and GGobi are part of literally thousands of "packages" that are available to work with R. These packages reside on the CRAN network or linked networks that are part of this open source effort. The book will show you how to install these packages and make them available to your analysis. These packages are so robust and dynamic that some of them are specifically made for some of the statistical tests that are in this book. However, as the analyst will find with R, not everything is set out like a buffet; some of the items have to be cooked.

2.4 IMPORT INTO RSTUDIO

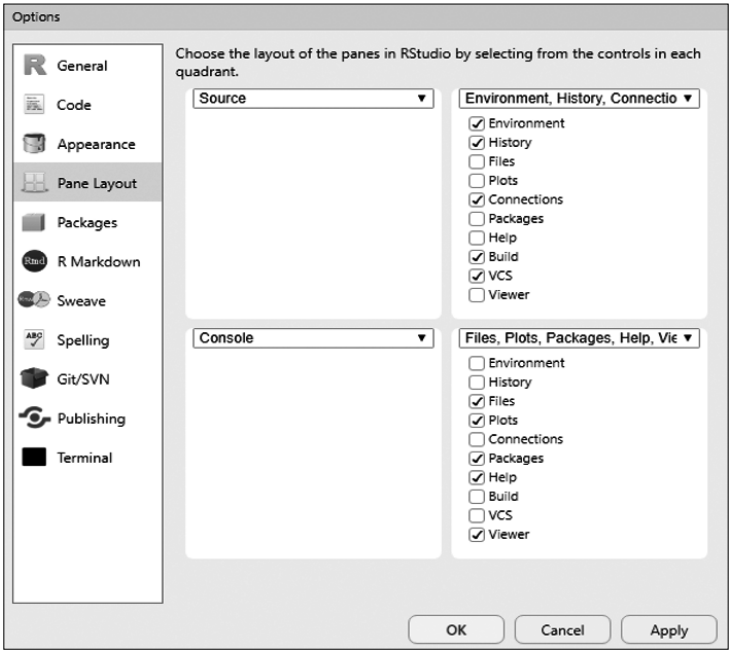
Once RStudio is installed and opened for the first time, this default work environment screen will appear. There are several things that are important to know before making any import attempts. First, did installing RStudio go into the "documents" folder or the "C" drive? This may make a difference in how RStudio responds to some commands and "packages." In order to eliminate any possible problems with R or RStudio, it might be advisable to start the application as an administrator if it is a Windows Operating System. In this way, the application will automatically have access to files that reside on protected folders and files.

When downloading an open source product, please ensure that there is active antivirus software on your machine. Additionally, scan the executable that has been downloaded *before* activating the product. Finally, if the plan is to do the analysis online, ensure there is an active Virtual Private Network (VPN) purchased and active on the machine. There are many VPNs available online, so pick one and use it. This will prevent any possible active intrusion that could happen while working with the open source application. People will avoid open source for these reasons, but understand that some expensive statistical applications have had some security problems, so just be prepared and that will prevent any possible mishaps with these software products.

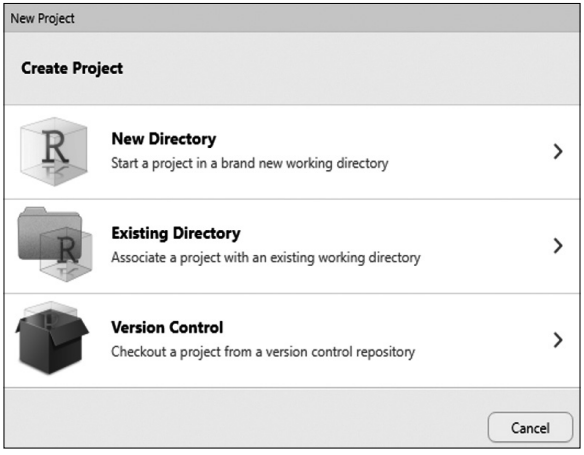
Now let's move on with the import for R and RStudio. When the installation of R and RStudio is done, the first time RStudio is opened, the screen will appear as the following:



Each of these areas on the screen represents a "pane." Customizing these panes is done by clicking in the "View" in the top toolbar. Let's explain each pane separately. The one on the left is the "Console" pane where programming is performed. Although this book is not centered on programming, there are times when the analyst must enter certain commands to perform a task. This pane is where it will happen. This left-hand side pane acts as two when a file is imported. At the moment the file is imported, another pane will appear called the "Source" pane, which will reveal the dataset in its entirety. More on this after the import. The two right-hand side panes show the history of the commands that are entered (top) and the different packages that are installed (bottom). There are tabs at each of these panes which apply to each pane's function. What is great about RStudio (and there are plenty of great features about RStudio) is that if you click on the "View" and select "Pane Layout," you will see the following screen, which can help you decide where you want each of the panes during development. You can choose exactly where you want each part of the development scheme.



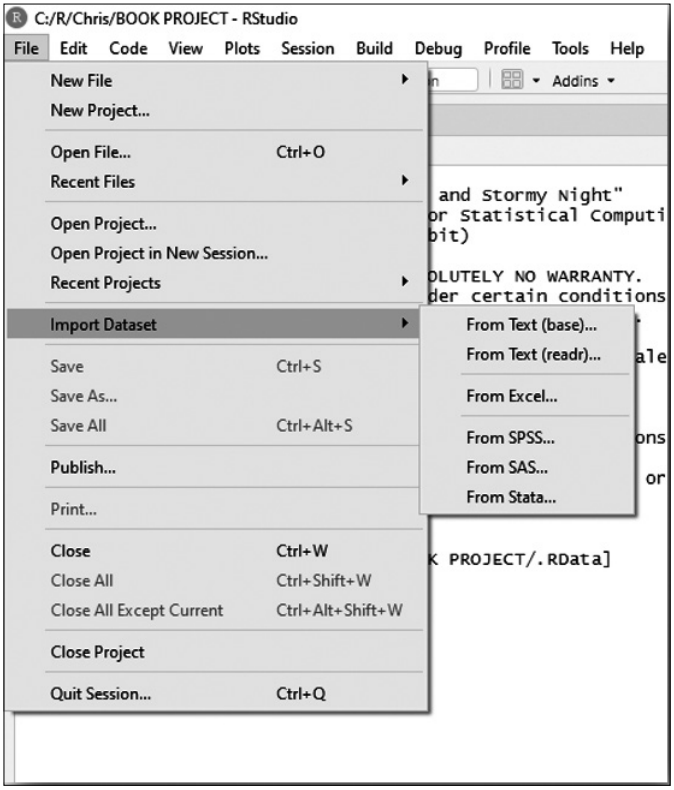
One caveat at this point, but this caveat is optional. While using RStudio, you can set where you want your project to be stored. From experience, some analysts do not save their project or even make a project, but instead rely on RStudio to do so automatically. RStudio will save files to the main R directory, but you can save them to a more specific folder which will hold your project material. The method to open a new project and save that project is to select "New Project" from the File menu, and you will get this screen.



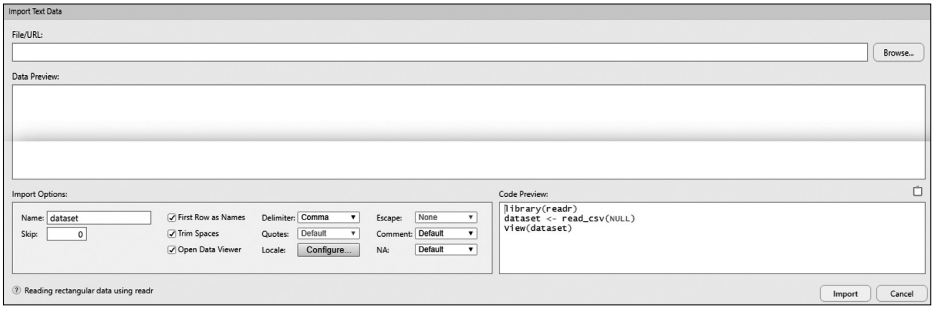
The choices are self-explanatory, so we will let you explore where you want to place your project files. Once you do that, RStudio will open in that project. If you want it to open another project, you guessed it, you use the "Open" selection in the File menu.

Those that have used R before might prefer the "basic" R screen without the assistance of RStudio, which is appreciated. RStudio will show the programming that is incorporated into the different mouse clicks, which will be shown later. First, importing the data is the next step to get RStudio (and Rattle) working.

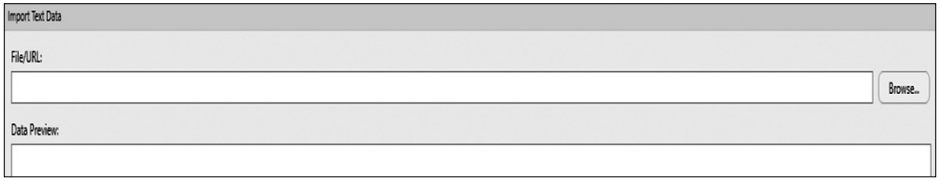
To import data into RStudio, you select "import dataset" from the File Menu. This is shown in the following screen grab. Ensure that "From Text (readr)" is selected to include the CSV files that are being imported.



When the choice is selected, the following screen will appear with plenty of blank text boxes. Reviewing these separately will help to make sense of those text boxes.

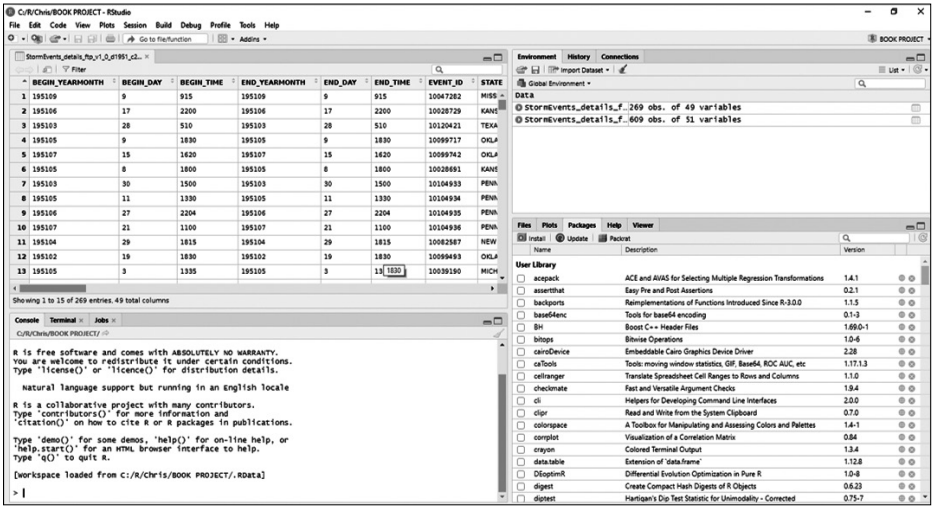


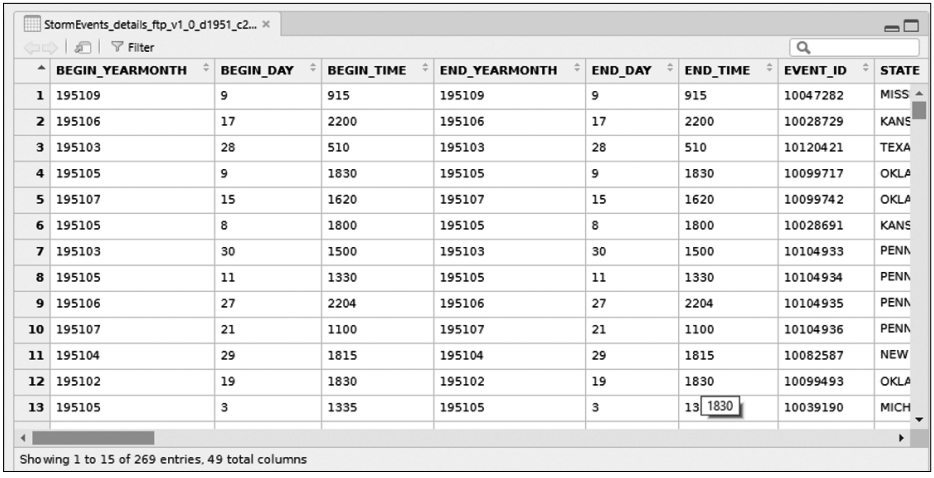
There are many components to this screen, but the main one is the top text box where the file name is placed to retrieve it either from the Internet or your computer. For the purposes of this book, the focus will be on already downloaded files that exist on the computer. The same file used in previous examples, which is the 1951 Tornado Tracking, will be used here also.



Once the file is inserted into the "File/URL" box, usually through using the "Browse…" button, then the file will appear as a preview in the large open text box. An avid R analyst may wish to know the background programming, and that is in the bottom right text box. If one has R, one can cut and paste the code and get the same results, except that the file will be saved in your R file rather than the RStudio area (most of the time they are the same, given that the analyst installs both R and RStudio in the same folder).

Once the file is imported into RStudio, the analyst will see the file in the File Pane, which in this case is in the top left-hand side of the screen, shown as follows with the main screen first and the file pane second.



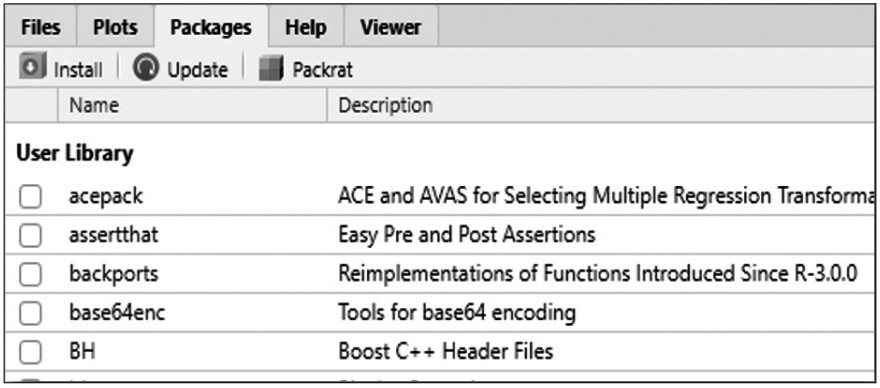


There is a caveat here that is vital when using RStudio. When a file is imported into RStudio, it becomes a "tibble." This is a term that means the dataset is of a particular type, and as such will need certain R packages in order to expeditiously analyze the data. No worries, since the tibble is also analyzed using conventional R tools, which can be used through RStudio.

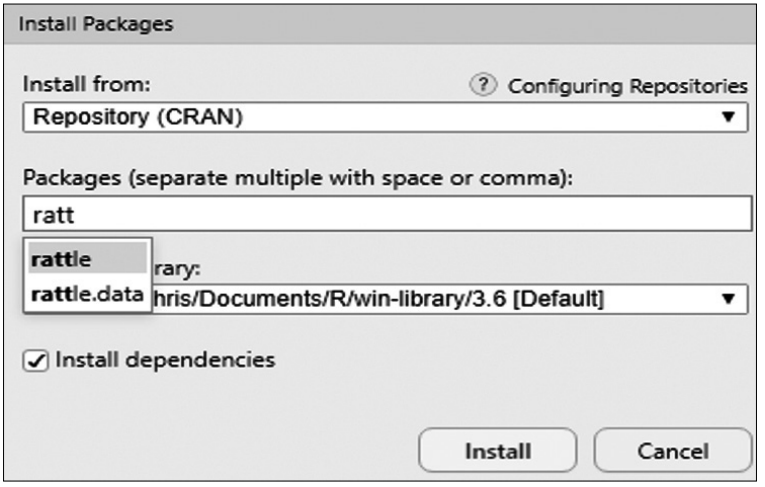
2.5 RATTLE IMPORT

R has a particularly robust package called Rattle that is so useful that it must be separated from R while describing importing (or any other function for that matter). Installing Rattle begins with the RStudio pane called "*Files*, *Plots*, *Packages*, and *Help*" (the lower right-hand side pane). As depicted in the following screen, this contains a number of packages that are already installed in the R, and subsequently RStudio, application. When first installing R and RStudio, the number of packages will be limited to those that are included in that installation. The other packages are installed either separately or come as a joining of other packages in order to activate the main package being installed. This all sounds confusing, so describing the process for installing Rattle should clear this up rapidly.

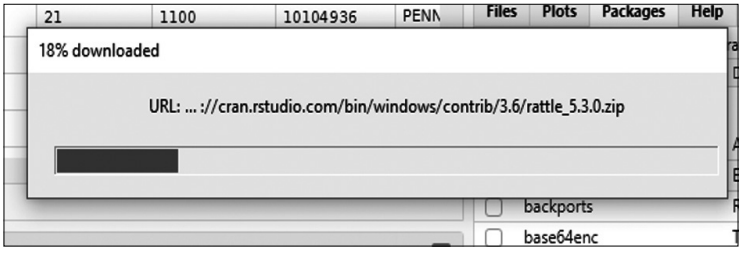
The first step when installing a package is to ensure the "Packages" tab is selected as shown in the following. Remember that this pane is located at the bottom right of the RStudio work environment. Notice the "install" button at the top left-hand side of the screen. This is the one we will be using to install the packages.



When choosing "install" the following popup will appear, showing a CRAN server where the package is stored (and can be downloaded and installed) along with a blank text box for the package. *Caveat*: the computer must be connected to the Internet or this part will fail. Start typing Rattle into the blank text box and, without finishing the word, "rattle" will appear. Notice that "Install dependencies" is checked. This is important since many packages have sub-packages that are independent, but to which this package has links in order to function. Leave this in its default mode for now.

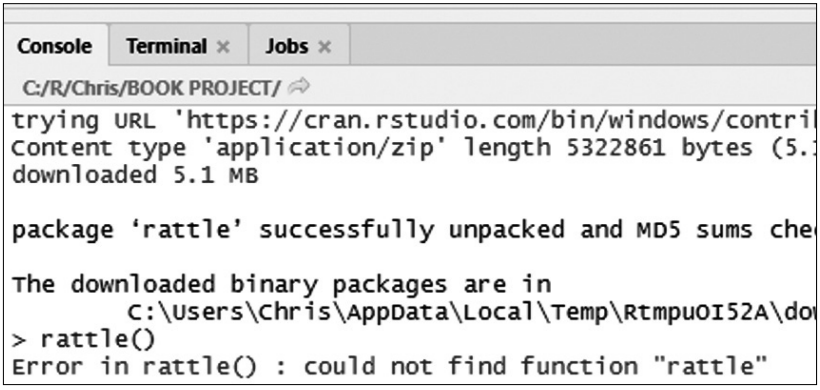


Click on the "Install" button and there will be a flurry of activity on the bottom left pane of RStudio. This is good because that means that RStudio found the server where the package resides and is downloading and installing the package.




```
Console  Terminal ×  Jobs ×
C:/R/Chris/BOOK PROJECT/
downloaded 5.1 MB

package 'rattle' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Chris\AppData\Local\Temp\RtmpuOI52A\downloaded_packages
> install.packages("rattle")
Installing package into 'C:/Users/Chris/Documents/R/win-library/3.6'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/rattle_5.3.0.zip'
Content type 'application/zip' length 5322861 bytes (5.1 MB)
downloaded 5.1 MB

package 'rattle' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Chris\AppData\Local\Temp\RtmpuOI52A\downloaded_packages
>
```

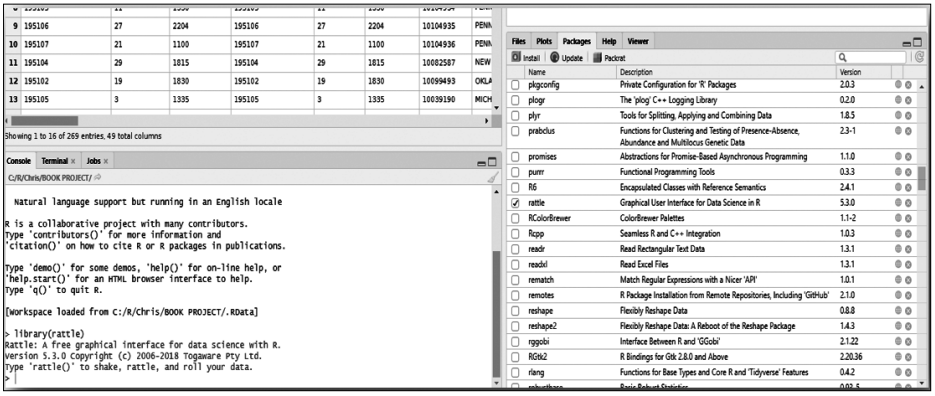
Now that Rattle has been installed, there is still one more step that must be accomplished, actually activating the package on R and RStudio. If the analyst types "Rattle" on the screen without loading the package, the message is clear.

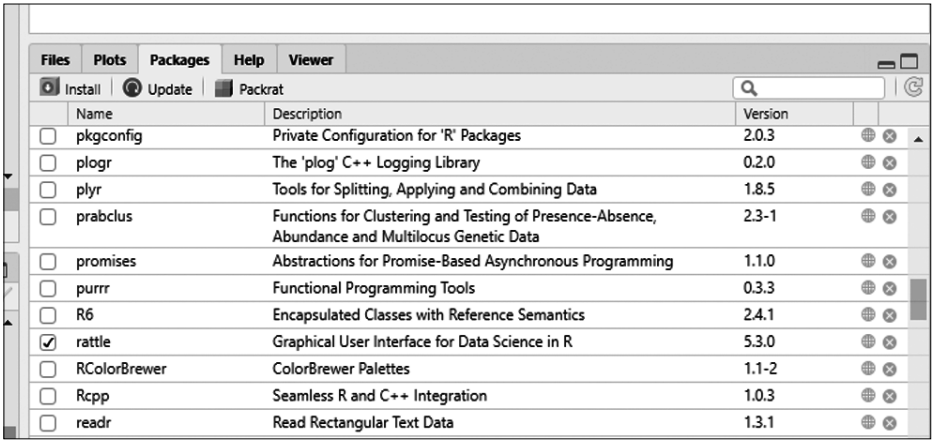


Rattle has to be loaded into R in order for it to be active. To do this is simple. One way is to type the following in R:

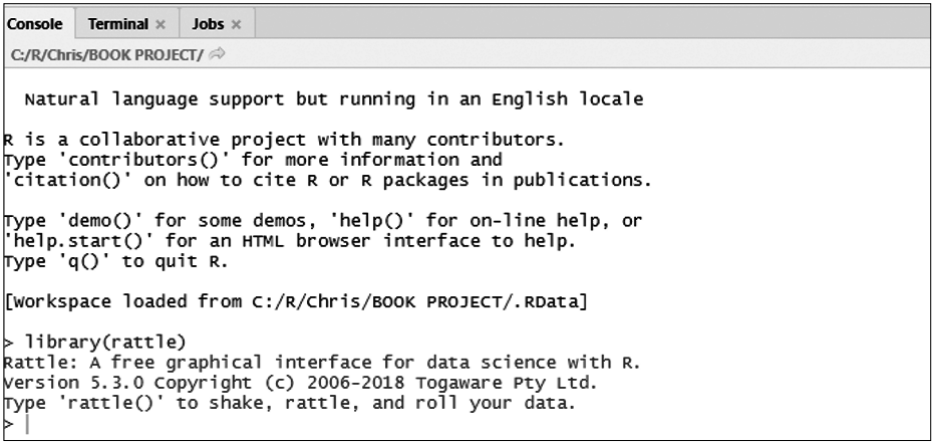
```
>library (rattle)
```

Another is to use the "packages" tab in the screen to the bottom right (in this book's configuration of the viewing pane) and check the checkbox next to Rattle (as shown in the following screen). Since RStudio is attached to R, the code will appear as if by magic in R.





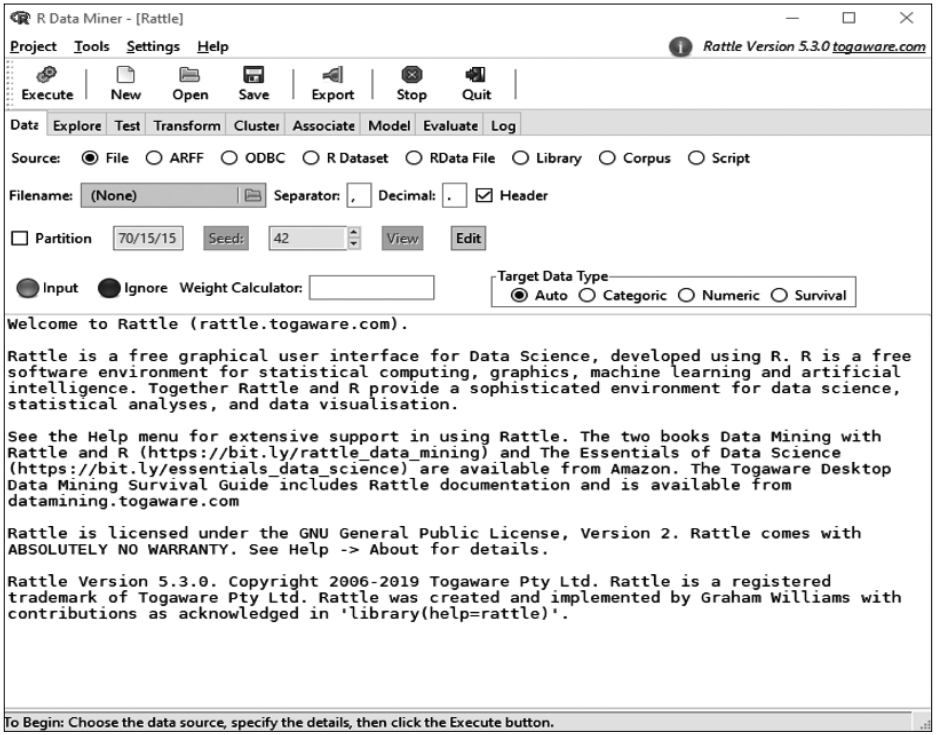
The previous screen shows the Packages tab with rattle checked. The moment an analyst performs this selection, the programming pane will come to life as follows and load the Rattle package, along with any dependencies that may come with the package that were not installed the first time. In some ways, R and RStudio anticipate what the analyst will require before they need it.



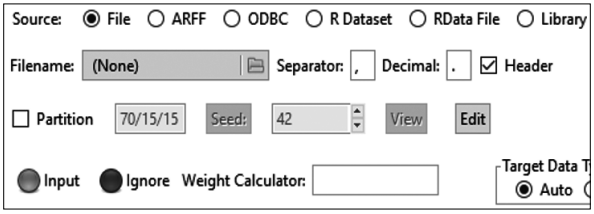
To activate Rattle, type the following in the programming pane:

```
>rattle()
```

At this point, the analyst has installed and loaded the package, so Rattle will show the first screen in a separate window that will be will appear as:

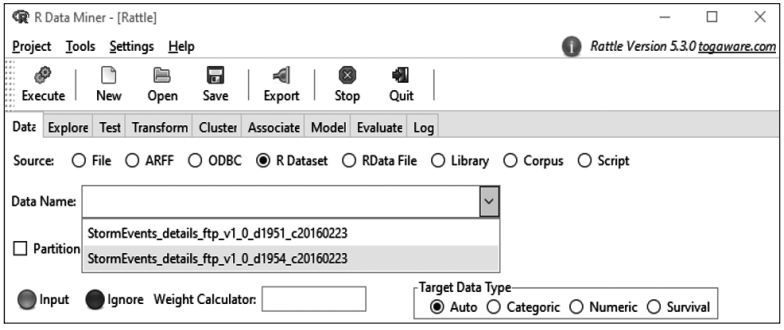


This screen is the home screen for Rattle and where the functionality of the tool is performed. The first step is to import the data into this tool. This is where R and Rattle are linked. Once the data is imported into R (or RStudio in this case), then it is made available to all other tools, in this case Rattle. In order to import the data into Rattle, use the "Filename" box in the main screen as follows:

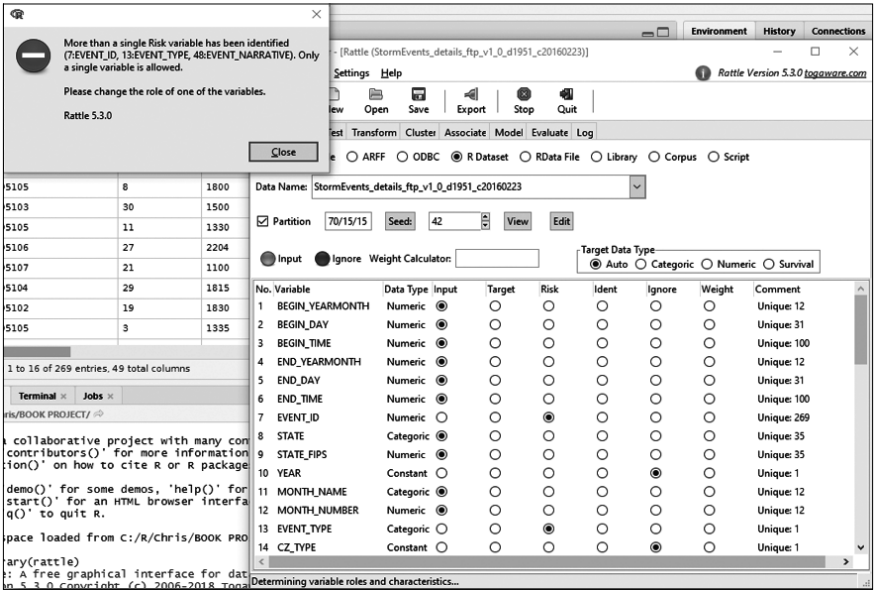


In this case, using the radio button choice of "File" compels you to reveal a filename in order to import the data. Use the same location of the data you did for OpenOffice and ensure that the "Separator" is a comma (since it is a CSV). Also, ensure that "Header" is checked, since this data does have a header.

However, since the data has already been loaded into R, the analyst can choose the "R Dataset" radio button as follows to reveal the dataset already in R. Choose the first file and click on "Execute" in the first iconic toolbar and the data will be imported to Rattle.



No matter how easy it seems, there are always some configuration changes to the dataset in order to make it more amenable to Rattle. In this case, once the data is imported (executed), a warning message appears as follows. This is easily fixed by just selecting one risk variable rather than having the numerous variables that Rattle picked.



Since choosing whether a variable is input, target, risk, ident, ignore, or weight is done with a touch of the mouse, it is easy to fix this by just picking one variable to be the risk. But before this is done, a little explanation is necessary to describe the different types of variables that the analyst will associate with each of these data types.

The following table gives a brief description of each of these data types taken from the CRAN. When the word "dynamic" is used, that means that these can be changed by the analyst any time a different model or evaluation is performed. The analyst simply changes the radio button choice and clicks on "Execute" again. The dataset is automatically changed. What is important about this last section is that any time the analyst wants to change the target or risk variables, a simple backtrack to the dataset and Execute will change the dataset variables. In some cases, as will be explained later in this text, some charts allow for interactive changes which will automatically change the target or risk factors in the dataset. The best is yet to come.

Type	Description
Input	The independent variables
Target	The dependent variables
Risk	Dynamic value that is used in risk charts
Ident	Have a unique value for each record
Ignore	Variables that are removed
Weight	Variables that are valued more than other variables in order to show importance

Now that the file is imported into RStudio and Rattle, the next tool used for import will be KNIME.

2.6 IMPORT INTO KNIME

KNIME is a data analysis tool developed in Europe and, in this author's experience, combines conventional statistical analysis with systems engineering process flow. The tool has "nodes" or modules that are self-contained analysis and transformation mini-tools to break down the dataset and analyze that dataset into the desired components. Before importing, the analyst must download the KNIME application.

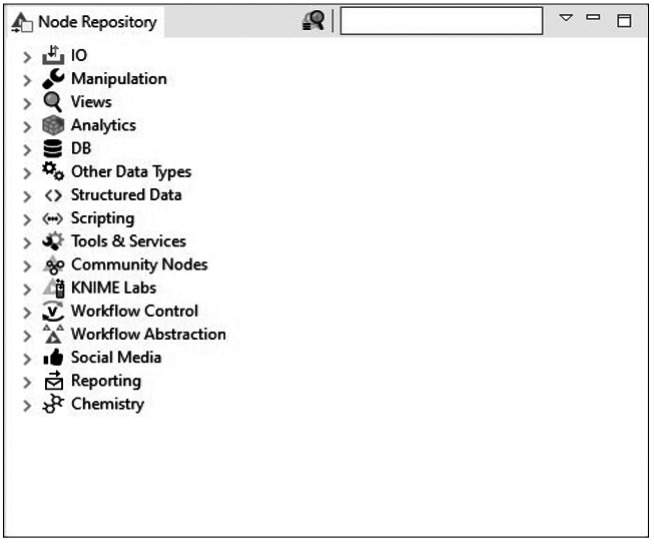
Remember that the same warning for this open source tool applies. Make sure the antivirus software is active and activated and that you scan the executable before activating the KNIME software. Caution is important here, but the tool will be worth all this effort.

After over twenty years of performing data analysis and teaching the same, this author has never seen a tool such as KNIME. This should not be construed as an unwavering endorsement of KNIME, but for data analytics, the tool provides a wide variation of "nodes" that, when placed in a flowchart-like process, can provide the analyst with a continuing stream of dataset transformation and analysis. We highly recommend that any analyst at least try this software to see if it fits your needs.

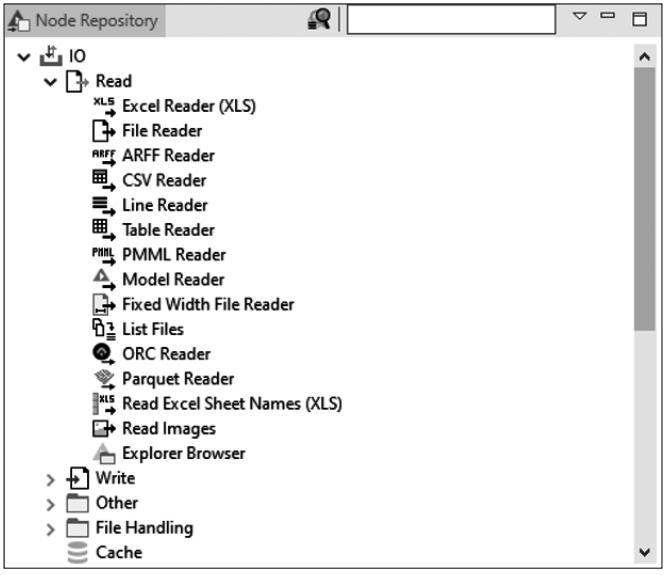
The installation is relatively easy. Go to the KNIME site (*www.knime.com*) and download the most current version of the application. For the purposes of this book, the version will be 4.1.0, but again this application has some very fundamental functions that will not change in the foreseeable future, so no worries on updates to this text for a while.

One caveat: This is a very "memory hungry" application, so it is advisable to have at least 8 GB RAM, and if the computer has 12–16 GB the application will work faster. Once you download the application and open it, this splash screen will appear. It may be a few seconds for this to appear, and it may not contain all the icons that are in this one, but later on this will be explained.

After opening the KNIME application, the following screen will appear with many different panes. For the moment, the "node repository" pane will be the main focus. The following screenshots show the first splash screen, the entire KNIME work area followed by the node repository pane. One recommendation is to explore the KNIME tool thoroughly, since there are many ways to combine these nodes or modules into one process flow, as will be demonstrated later with just one or two of these nodes.

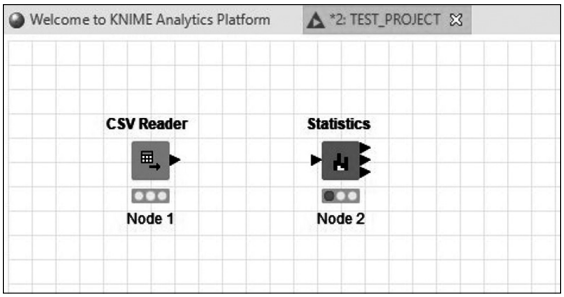


The "IO" portion of the node repository will be the one that will import the dataset for analysis. Since the original file was a CSV, that is the one that will be imported into KNIME. The following screens show the process for selecting and importing the data. Please pay special attention to the other options available for importing to see the plethora of choices from KNIME.

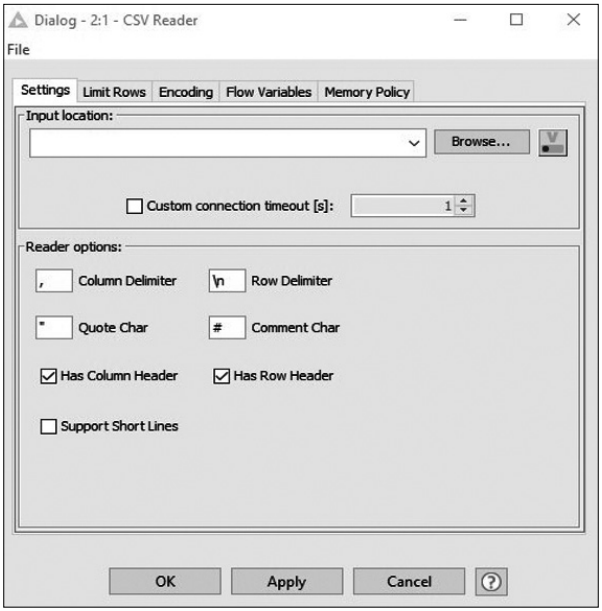


As you can see, the "CSV Reader" is one of the KNIME nodes. The node is active within the workspace by clicking and dragging the node into that

workspace. Once that is completed, the workspace will appear as this (with an added node for effect). The "CSV Reader" node (or node 1) has a yellow indicator, which means that the data is either not available or not "cleaned," which means there needs to be configuration with the data to turn the light green.



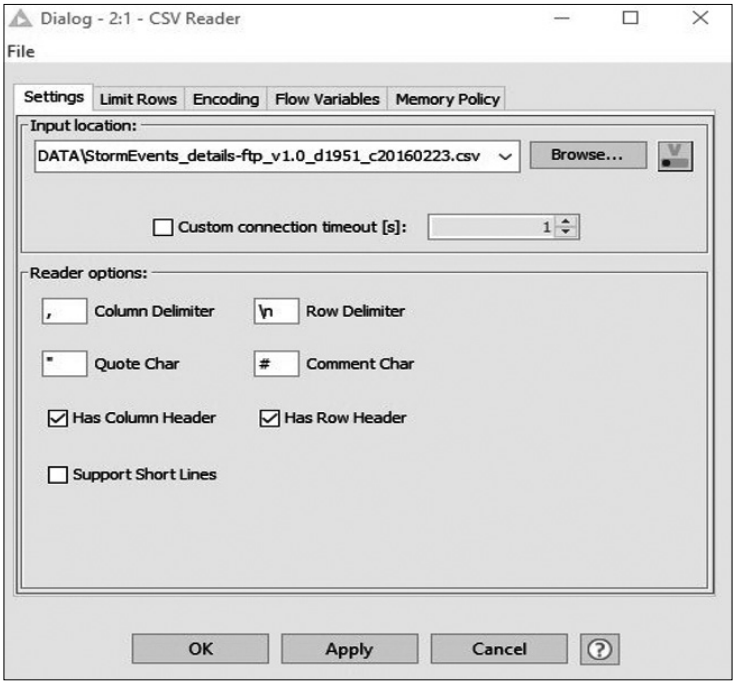
Just as a note, the analyst can also name the workspace as if opening a new project in R. This will be discussed later; currently, double-click on the "CSV Reader" node after placing it in the main workspace and this screen will appear. Notice that there are several default choices in the configuration, including those in "Reader options," and for now those are fine. What is essential is that "Has Column Header" and "Has Row Header" are checked and that "Column Delimiter" shows a comma.



At this point, the next step would be to use the "Browse…" button to search your computer for the file to be imported. Once that is done, the screen should appear similar to the image below.

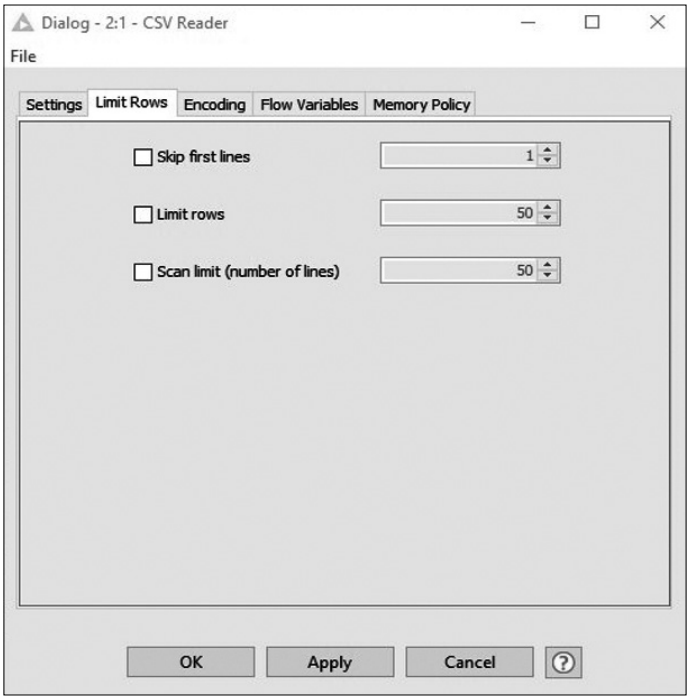
Once this is completed, click on the "Apply" or the "OK" button to import the data. At this point, the data is imported into KNIME. However, before clicking "Apply" explore the other tabs to see how they are part of the overall configuration of this dataset.

The first tab is the "Limit Rows" tab, shown as follows. This tab will assist the analyst to determine which rows to include in the dataset. This is one of the fundamental concepts of data science, which is to "understand the data" (part of CRISP-DM). If the analyst does not understand what requirements are associated with the data, it will be difficult to determine which data is useful. In this case, the analyst can skip the first line or lines as is determined by the data content, along with limiting the number of rows to scan.

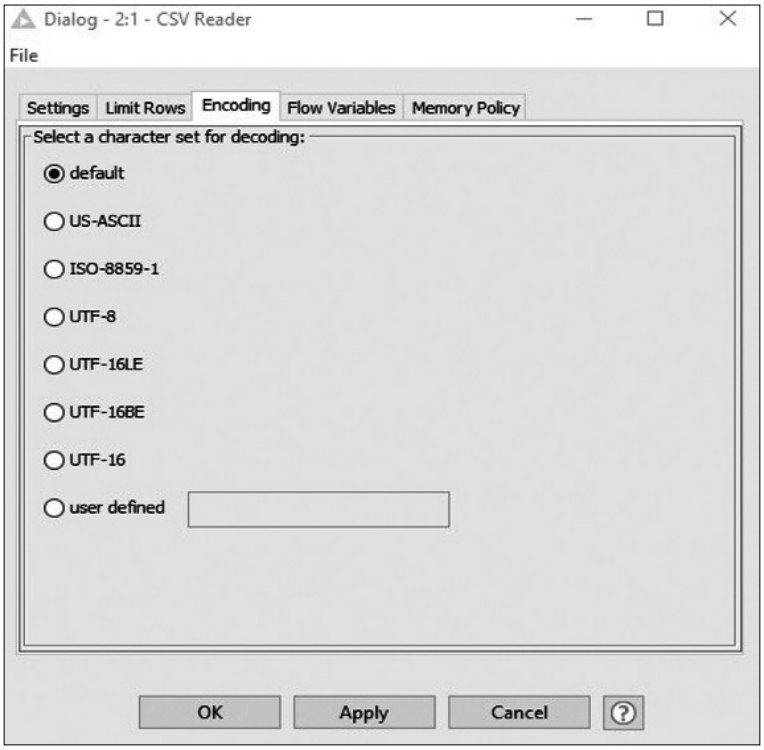


Why would an analyst do this? There are terms associated with only taking part of the data for analysis (such as "training data"), but suffice it is to say that performing analysis on a part of the data is much quicker than performing it on all the data, and the evaluation can take place later on a larger portion of the data. This tab helps to accomplish this function without much effort.

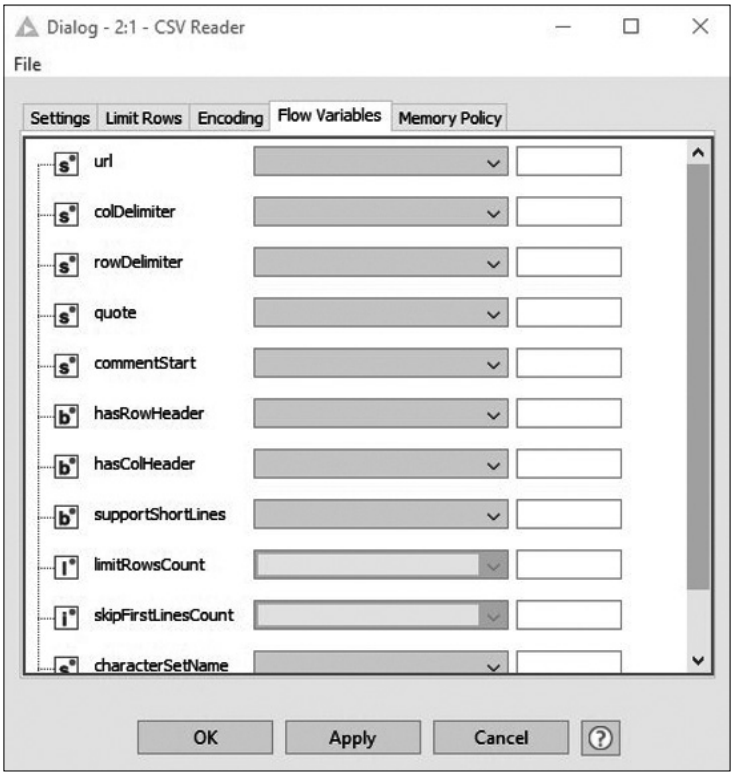
This type of configuration customization is just one feature that makes KNIME very flexible as a software product. Fortunately, because the tool is open source, there are many community sites and collaboration efforts that help to describe these screens and their function. The reason for the detail here is that there are times when the analyst needs immediate association between the function and the statistical concept. This is provided here.

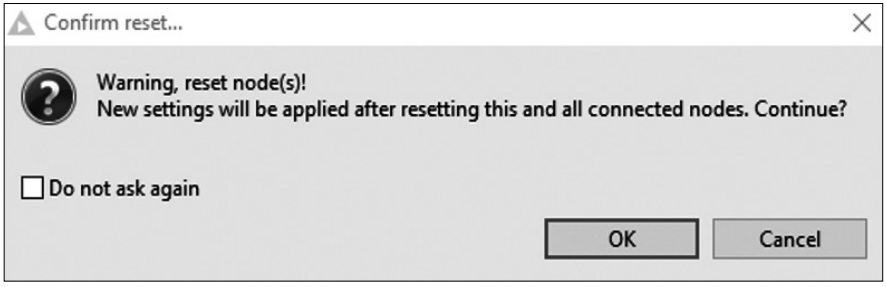


The next tab is the "Encoding" tab which is shown as follows. There are times when certain text encoding is important for programming or other analytical efforts. The analyst may never apply these settings, but it is important to know where they are in case encoding needs to be applied to the dataset. In most cases, the "Default" radio button is the one that applies, so there is no need at this point to change that choice. However, in cases where raw text is imported, some of the other choices may make the transition to KNIME both smoother and more useful to the analyst.



The next tab is one that exists throughout the KNIME application called "Flow Variables." Although this will not be used in these statistical applications, they can be used in future node flows. These are analyst-defined values that exist so that each node does not have to be set individually as they refer to the dataset. This book will not delve into these, but the KNIME websites give a more than adequate explanation to the use of these variables. What will be said about flow variables is that the analyst can set them at the text box that is next to each variable. This will "force" the variable across each of the subsequent nodes in the flow diagram. The analyst needs to remember that the flow variables, once set and applied, must be removed and the dataset refreshed each time the flow is executed. This is simply done by clicking on the "Apply" button after each reconfiguration. A message will appear that tells the analyst that the node has been changed, shown as follows.

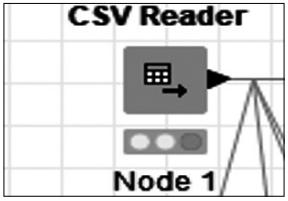


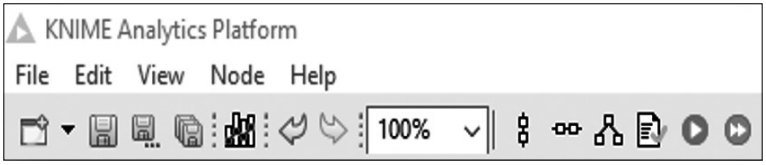


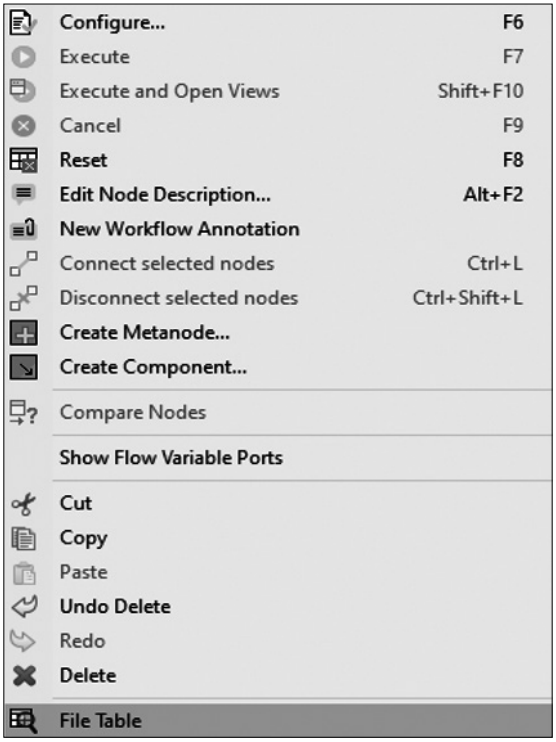
Once the "OK" has been clicked, the dataset will have a new configuration and the flow must be "Re-Executed." It is at this point that a review is necessary of the "Stoplight" approach to KNIME. If the reader would like more information than found in this book, the references to all the open source tools are in the Reference section of this book.

2.6.1 Stoplight Approach

Each node in KNIME is governed by looking at the node to see the status of that node. If the node has a "green" light, that node has been executed or activated within the flow. The following "CSV Reader" node has the green light, which shows that it has been executed. If anything is changed in the "Configuration" of this node, then the light will change to either "yellow" or have a "caution triangle" below the node. If the light is yellow, then right-clicking on the node and clicking on the "Execute" choice will execute the local node (the one that is the focus of the node at that time). To execute ALL the nodes attached to the one in focus, go to the main toolbar and click on the "double arrow" icon shown as follows to execute ALL the nodes attached to the selected node. If a node has a caution triangle, then that means that something is wrong with either a node that "feeds" the caution node, or something is wrong with the configuration. The best way to reduce or eliminate these caution nodes is to ensure the configuration on the feeding node is correct by testing the node through right-clicking on the feeding node and looking at the result of that node. An example of this follows.







In the previous screen, "File Table" is selected, since that is the result of the node. This will show the table that is the result of executing that node. Once confirmed, the node will show a correct configuration for the process flow. If incorrect, double-click the node to reveal the configuration screen, or right-click and select "Configure…" to enter the configuration screen.

CHAPTER 3

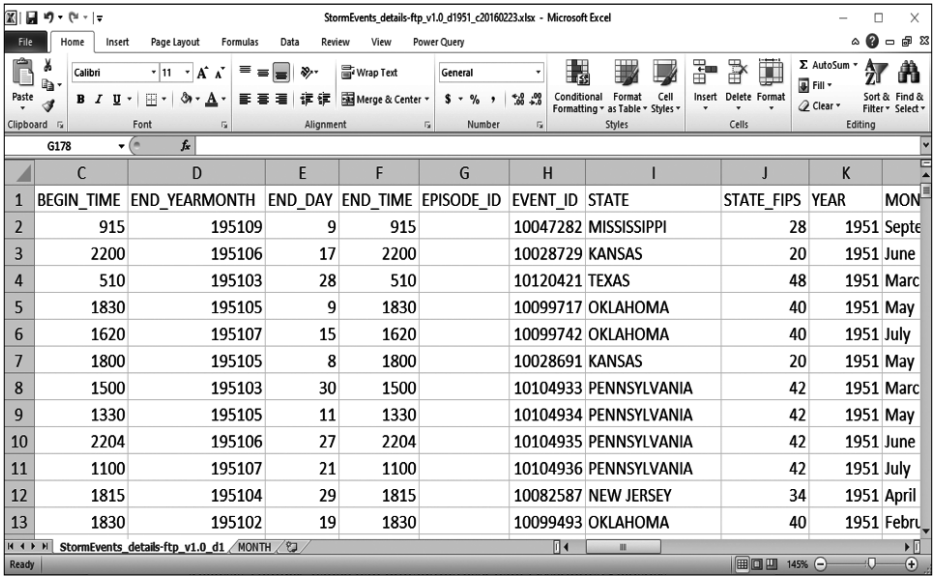
STATISTICAL TESTS

3.1 DESCRIPTIVE STATISTICS

The topic that is commonly introduced in statistics is *descriptive statistics*. Many students have already been exposed to many of these, including *mean*, *median*, *mode*, *variance*, and *standard deviation*. As promised, this book is not going to delve into the formulas for these or force the student to do them by hand. The main reason for stating them here is to apply each data science tool to show these descriptive statistics, hopefully with one function within the tool.

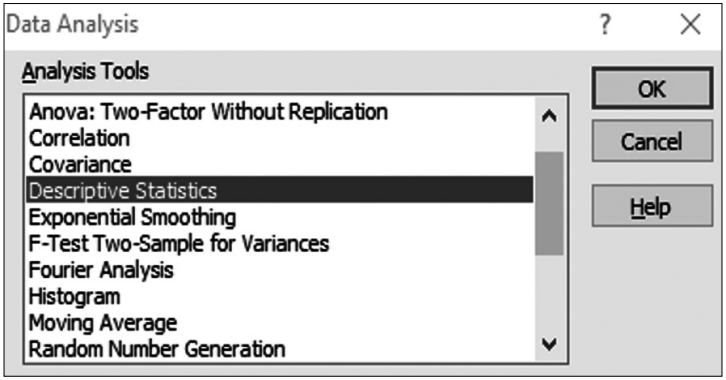
3.1.1 Excel

Excel, as discussed previously, has this magical function called the Analysis ToolPak, which will provide the analyst with descriptive statistics without each calculation being inputted into the application. The first step is to open the dataset and then open the Analysis ToolPak, as depicted in the following two screens.

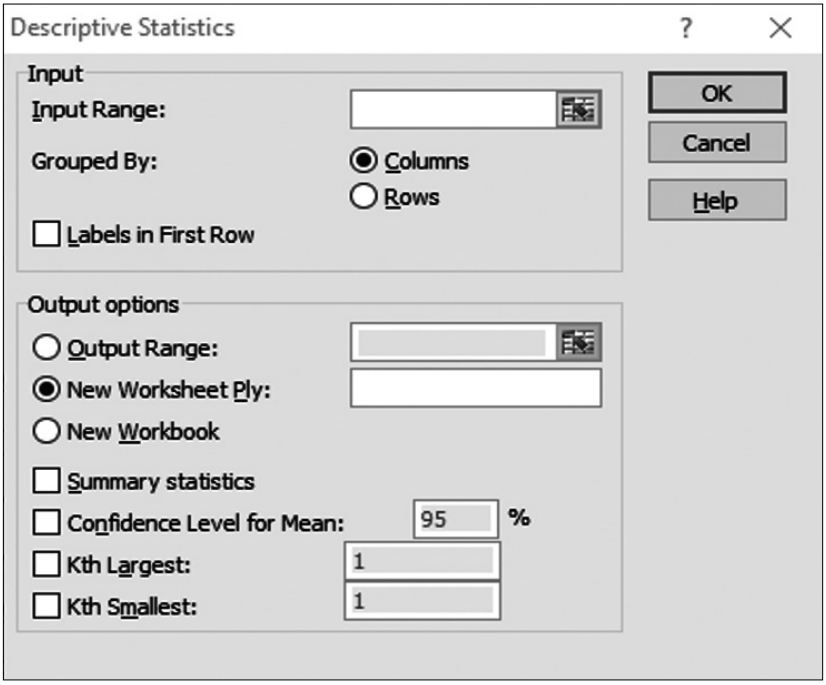


StormEvents_details-ftp_v1.0_d1951_c20160223.xlsx - Microsoft Excel

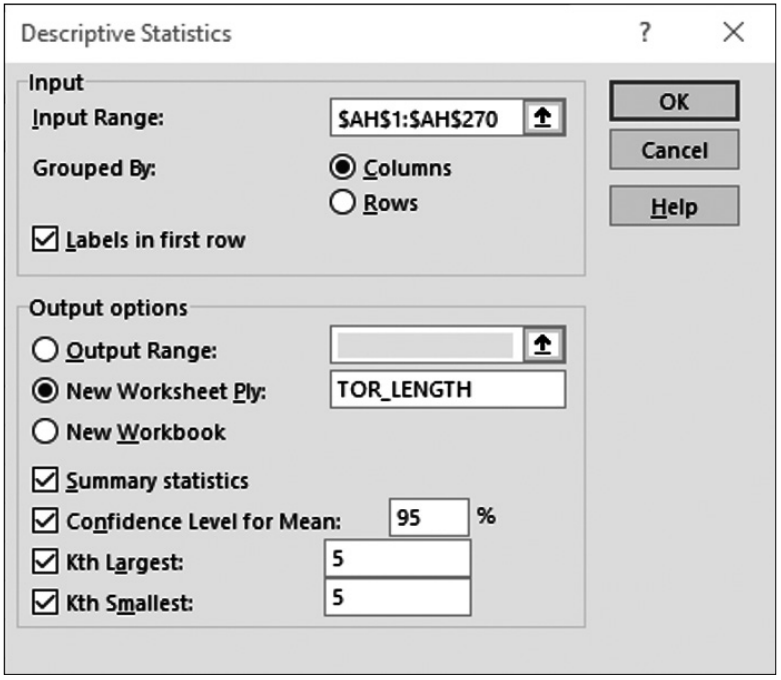
	C	D	E	F	G	H	I	J	K	
1	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EPISODE_ID	EVENT_ID	STATE	STATE_FIPS	YEAR	MON
2	915	195109	9	915		10047282	MISSISSIPPI	28	1951	Septe
3	2200	195106	17	2200		10028729	KANSAS	20	1951	June
4	510	195103	28	510		10120421	TEXAS	48	1951	Marc
5	1830	195105	9	1830		10099717	OKLAHOMA	40	1951	May
6	1620	195107	15	1620		10099742	OKLAHOMA	40	1951	July
7	1800	195105	8	1800		10028691	KANSAS	20	1951	May
8	1500	195103	30	1500		10104933	PENNSYLVANIA	42	1951	Marc
9	1330	195105	11	1330		10104934	PENNSYLVANIA	42	1951	May
10	2204	195106	27	2204		10104935	PENNSYLVANIA	42	1951	June
11	1100	195107	21	1100		10104936	PENNSYLVANIA	42	1951	July
12	1815	195104	29	1815		10082587	NEW JERSEY	34	1951	April
13	1830	195102	19	1830		10099493	OKLAHOMA	40	1951	Febru



When the analyst selects the "Data Analysis" icon at the far right of the previous screen (after selecting the "Data" tab), the pop-up screen appears with many options for using the data analysis functions. The one that is the focus on this chapter is the "Descriptive Statistics" choice (in blue). The analyst will select this option and click the OK button, and the next screen will appear. At once, the analyst will notice that there are many text blanks to fill, but this is not a problem as long as there is a dataset on which to apply this function.



The first blank to fill is the column or columns that need to be resolved to descriptive statistics. The analyst can do this either manually or by selecting the columns from the dataset. There is a caveat here, mainly pertaining to columns versus rows. If the dataset has column names and data going down the column and THIS is the data you need resolved, then ensure the "Group by:" choice has the "Columns" radio button selected. If "Rows" is selected, then the analysis will be done by row rather than column. Most datasets are configured or organized by column, so that is why the column radio button is already selected. Once that is completed, the next step is to fill out the screen as follows to get descriptive statistics for TOR_LENGTH, which is basically the length of the tornado from the first sighting to dissipation. Ensure that "Summary Statistics" is checked; otherwise, the results will not be what the analyst expects.



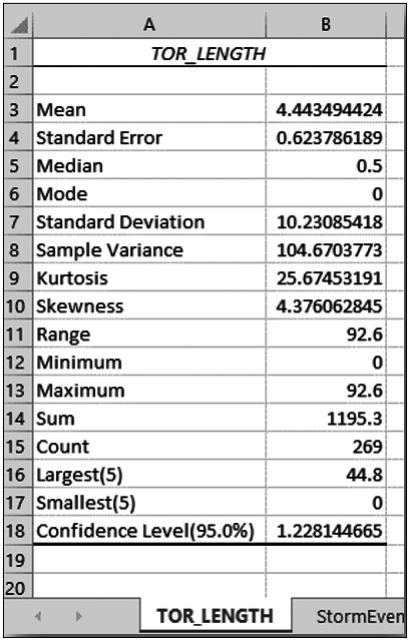
One word of caution is necessary at this point. The default selection under "Output options" is to designate an output range. If the analyst does this, the results will be in the same worksheet as the dataset. This could prove to crowd out the worksheet so that the analyst will have to scroll beyond the dataset cells to see the functional results. It is recommended to always use the "New Worksheet Ply:" option and name the worksheet something similar to the previous title. This will ensure that the dataset sheet will remain *just* the dataset, rather than adding unnecessary columns to that worksheet.

Another item that is necessary is the "Level in first row" checkbox, which is important. Normally, column headings (or labels) are important for naming each column. Not checking this box will tell Excel that there are data, not names, in the first row. That could prove hazardous should there be column headings in that row.

Notice that there is a "95%" in the text box next to "Confidence Level for Mean:" which indicates that, should this dataset be a sample of the larger dataset, this function would show you a range where there would be a 95% chance that it would contain the mean. This will be covered under "Confidence Intervals" in the next section, but suffice it to say that it will be important to check this box and set it for 95% (since this is the conventional confidence level for Statistics).

"Kth Largest" and "Kth Smallest" have been checked and marked with a "5" to demonstrate how to use this option. Basically, what this will show is the "5th Largest" and "5th Smallest" values in the dataset. This might be useful if the analyst wanted to find out how a certain value ranked among all the other values.

When the "OK" is clicked, the following will appear in an additional worksheet in your Excel workbook. The analyst will immediately notice that there are many terms that are recognizable and those that are not.



	A	B
1	*TOR_LENGTH*	
2		
3	Mean	4.443494424
4	Standard Error	0.623786189
5	Median	0.5
6	Mode	0
7	Standard Deviation	10.23085418
8	Sample Variance	104.6703773
9	Kurtosis	25.67453191
10	Skewness	4.376062845
11	Range	92.6
12	Minimum	0
13	Maximum	92.6
14	Sum	1195.3
15	Count	269
16	Largest(5)	44.8
17	Smallest(5)	0
18	Confidence Level(95.0%)	1.228144665
19		
20		

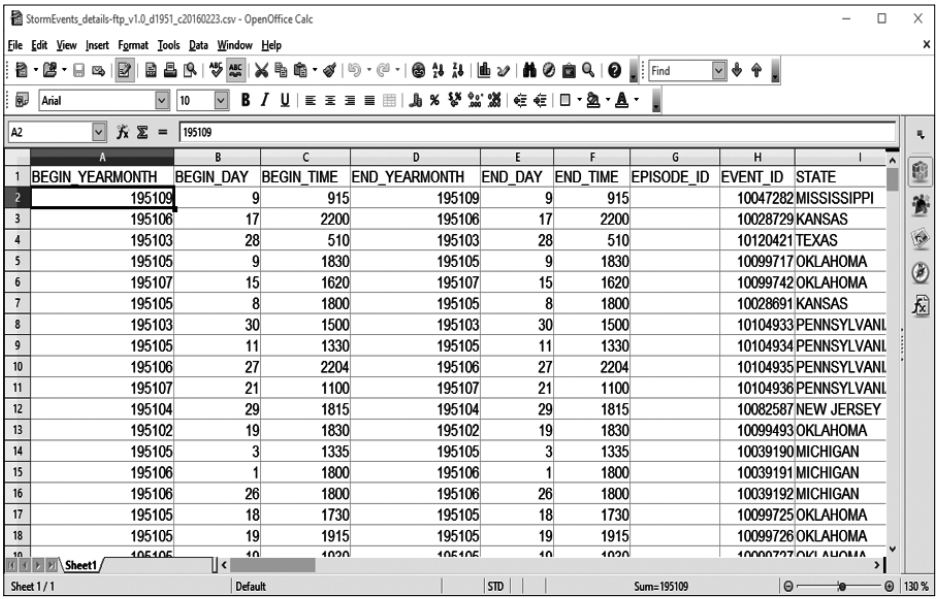
TOR_LENGTH StormEven

This result should provide the analyst with a description of the data, much like seeing a person should provide a description of that person. Again, this book is not a statistics primer, so it will not delve into the specifics of each of these titles. Most importantly, remember that this function will give you a good preview at a dataset variable even before any graphs or charts are produced.

3.1.2 OpenOffice

OpenOffice, although much like Excel, does not have an Analysis ToolPak that is available for that software. As such, it will take more effort to have the same result as Excel.

The first step is to open the OpenOffice Spreadsheet and open the dataset that was imported in the last step. The result should look like the following screen:



	A	B	C	D	E	F	G	H	I
1	BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EPISODE_ID	EVENT_ID	STATE
2	195109	9	915	195109	9	915		10047282	MISSISSIPPI
3	195106	17	2200	195106	17	2200		10028729	KANSAS
4	195103	28	510	195103	28	510		10120421	TEXAS
5	195105	9	1830	195105	9	1830		10099717	OKLAHOMA
6	195107	15	1620	195107	15	1620		10099742	OKLAHOMA
7	195105	8	1800	195105	8	1800		10028691	KANSAS
8	195103	30	1500	195103	30	1500		10104933	PENNSYLVANI
9	195105	11	1330	195105	11	1330		10104934	PENNSYLVANI
10	195106	27	2204	195106	27	2204		10104935	PENNSYLVANI
11	195107	21	1100	195107	21	1100		10104936	PENNSYLVANI
12	195104	29	1815	195104	29	1815		10082587	NEW JERSEY
13	195102	19	1830	195102	19	1830		10099493	OKLAHOMA
14	195105	3	1335	195105	3	1335		10039190	MICHIGAN
15	195106	1	1800	195106	1	1800		10039191	MICHIGAN
16	195106	26	1800	195106	26	1800		10039192	MICHIGAN
17	195105	18	1730	195105	18	1730		10099725	OKLAHOMA
18	195105	19	1915	195105	19	1915		10099726	OKLAHOMA

At this point, the descriptive statistics will include the following items:

1. Mean
2. Median
3. Mode
4. Standard Deviation
5. Kurtosis
6. Skew
7. Minimum
8. Maximum
9. Confidence Level for Means

All of the previous items were part of the descriptive statistics included in the Excel Analysis ToolPak. This is the case with OpenOffice. However, follow the formulas and it makes it repeatable.

The first formula will be for *mean* (or *average*). The formula will appear as this when placed in the formula bar of OpenOffice:

```
=AVERAGE(AH2:AH270)
```

This formula should be placed after the AH270 cell to prevent any circular calculations. The next formulas should be placed after (lower than) the AVERAGE calculation. One word of caution is necessary at this point. Ensure that you make AH2:AH270 an absolute reference (dollar signs before both AH2 and AH270 to look like this—$AH$2 and $AH$270). This will prevent the AVERAGE result from being included in the calculation below it and so on. The formulas should appear as the following:

```
=MEDIAN($AH$2:$AH$270)
=MODE($AH$2:$AH$270)
=STDEV($AH$2:$AH$270)
=KURT($AH$2:$AH$270)
=SKEW($AH$2:$AH$270)
=MIN($AH$2:$AH$270)
=MAX(AH2:AH270)
=CONFIDENCE(0.95;AH274;269)
```

The results of these calculations are as follows (to the left). Next to those results are the results from Excel (the screen to the right). This is important! How do the results compare? Are they substantially different, or relatively similar? This is now to verify your tool accuracy.

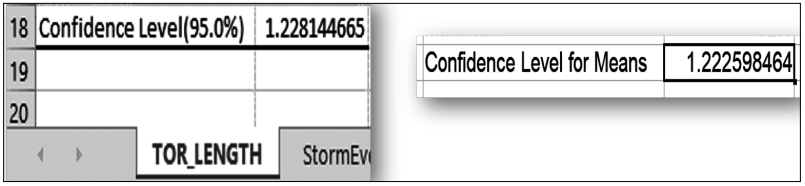
Mean	4.4434944238
Median	0.5
Mode	0
Standard Dev	10.2308541821
Kurtosis	25.6745319148
Skew	4.376062845
Minimum	0
Maximum	92.6
Confidence Level for Means	0.039115622

	A	B
1	*TOR_LENGTH*	
2		
3	Mean	4.443494424
4	Standard Error	0.623786189
5	Median	0.5
6	Mode	0
7	Standard Deviation	10.23085418
8	Sample Variance	104.6703773
9	Kurtosis	25.67453191
10	Skewness	4.376062845
11	Range	92.6
12	Minimum	0
13	Maximum	92.6
14	Sum	1195.3
15	Count	269
16	Largest(5)	44.8
17	Smallest(5)	0
18	Confidence Level(95.0%)	1.228144665
19		
20		

TOR_LENGTH StormEven

Most noticeable is the large discrepancy between the Confidence Levels of OpenOffice and Excel. After reviewing the formulas, the "alpha" that is desired for OpenOffice is the difference between "1" and the Confidence Level, which means that, for OpenOffice, the proper number is ".05" (or 1 – .95), NOT ".95" as originally submitted. After this formula change (illustrated as follows), the Confidence Level result is also below the formula. The discrepancy no longer exists. This is important to remember—not every formula is exactly the same between Excel and OpenOffice.

```
=CONFIDENCE(0.05;AH274;269)
```

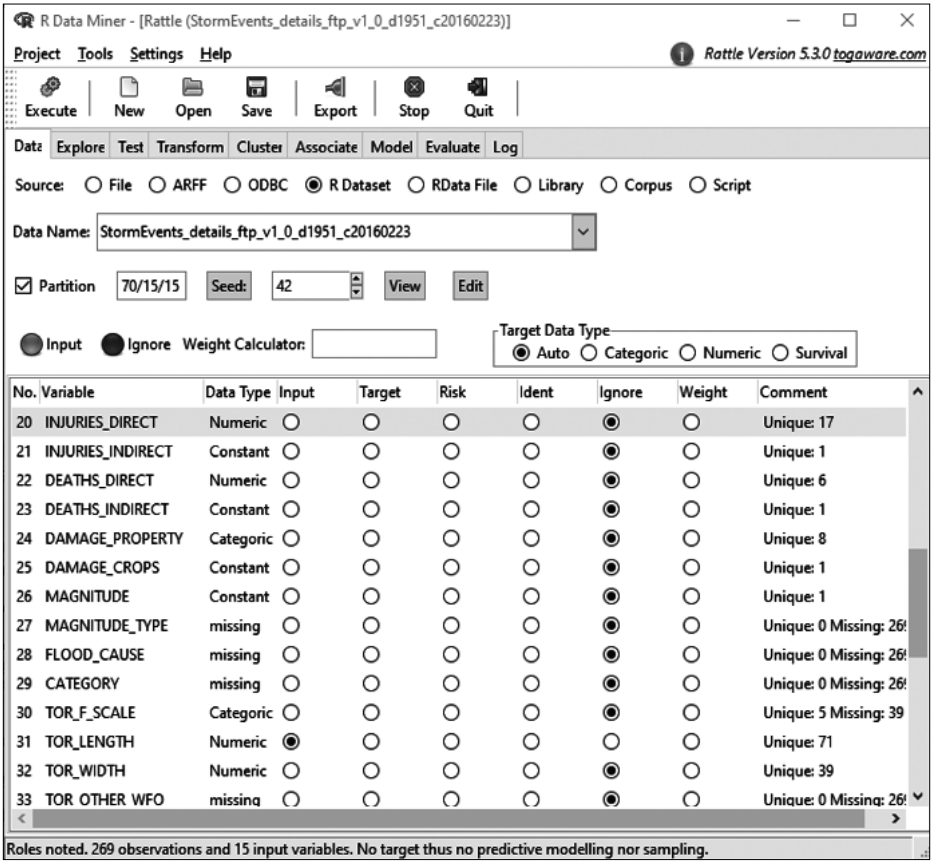


Another difference between Excel and OpenOffice is that any separation in the formula must be done with a "," in Excel and a ";" in OpenOffice. The analyst will receive an error message when the wrong symbol is used in OpenOffice. The warning is to relieve some angst for analysts that normally feel there is a problem with the software should an error appear. In this case, it is simple to change from a comma to a semicolon.

With Excel and OpenOffice formulating descriptive statistics, RStudio/Rattle and KNIME will be a little more challenging, but certainly not insurmountable. RStudio is first and then KNIME.

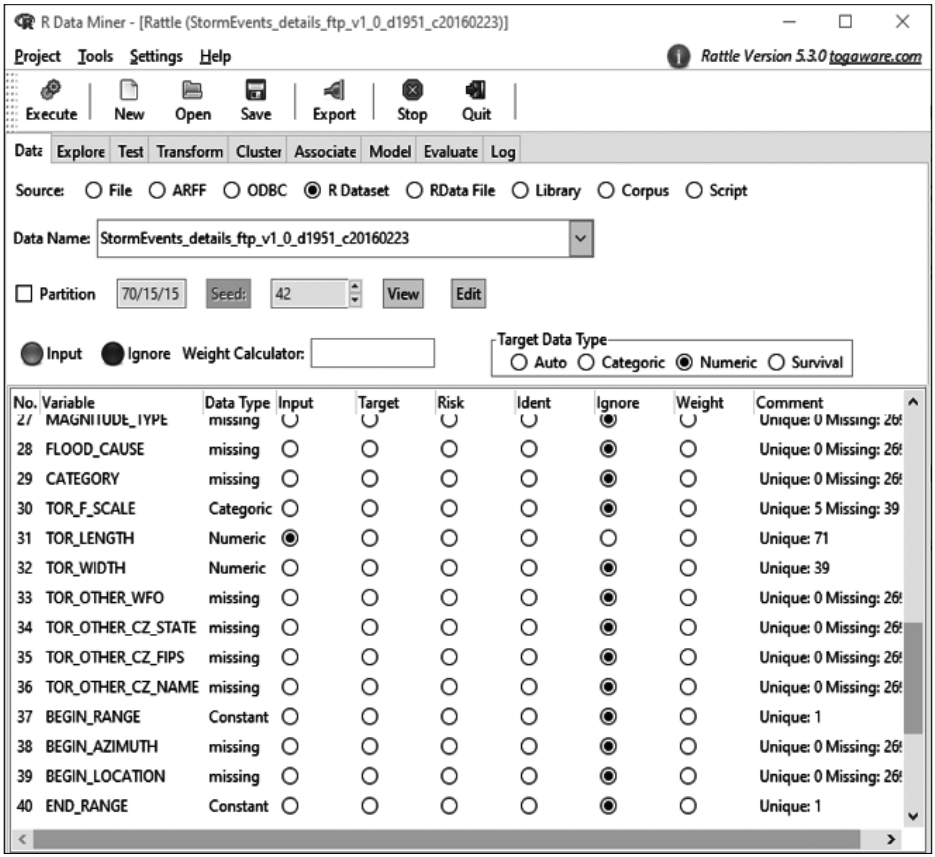
3.1.3 RStudio/Rattle

The first step is straightforward—open the RStudio application and Rattle package as was mentioned in the Importing Data section to receive to the following screen.

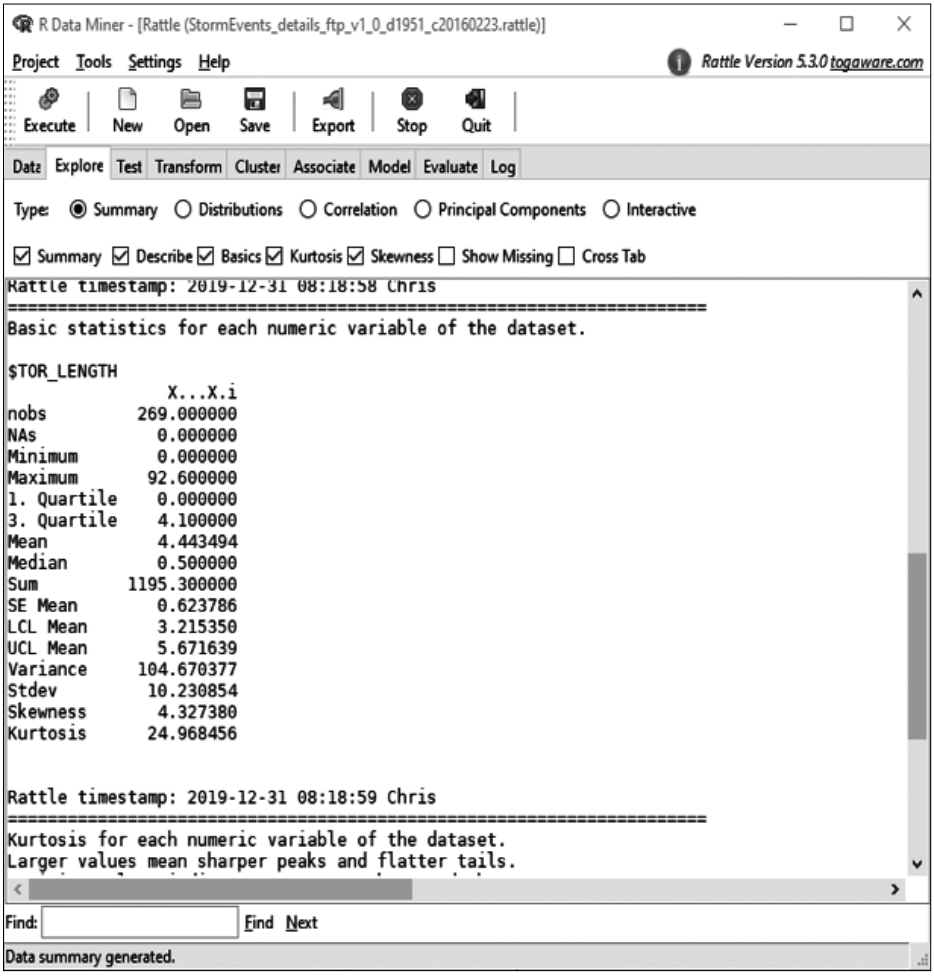


Notice "Roles noted" at the bottom of this screen. The analyst can pick any target for determining statistics. In this case "TOR_LENGTH" was chosen to be consistent with the same variable used in the previous sections/tools. Please notice the part of the screen marked "Partition," which is checked. What this means is that Rattle will automatically separate the data into percentages for training and validation; in this case 70% of the data will be used for training. This sampled dataset is *not* the entire dataset but is automatically randomly selected so that the analyst can test the different functions on a sample of the dataset rather than using the entire dataset. For a small dataset this is not necessary, but for larger datasets, this not only advantageous but necessary to save computing time. In this case the box is checked, but uncheck this

box (shown as follows) so that descriptive statistics are only performed against the entire dataset, to be consistent with other sections.



Once this screen is attained, the next step would be to ensure that "Execute" is clicked to lock in the dataset and the variables. Once that is completed, select the radio button next to the "Data" tab called the "Explore" tab. The following screen will appear:



Notice from the previous screen the radio buttons that are selected and the boxes that are checked. What this combination does is provide the analyst with the most information from clicking on the "Execute" icon. In this case it provides many different results similar to the ones that were produced by both Excel and OpenOffice. Going screen by screen will reveal these results in a more organized fashion.

The first few parts of the descriptive statistics screen in Rattle show a column with all the different summary statistics, which is nearly similar to the

Excel and OpenOffice screens. The Excel screen is placed beside the summary results screen of Rattle to show the similarities. Again, it would seem that the numbers match the original results from Excel, which verifies the algorithm in (so far) all of the different tools. This is a good thing! Consistency is the key for statistics, so having the same results shows consistency. The one result that is slightly different is the kurtosis, but this is probably because of rounding in the calculation and of no concern to the data analyst. A definition of each of the row names in the Rattle result is included in the following table. The aspects that are very interesting, and of special interest, are the "LCL" and "UCL," which are included in the Rattle result. This designates the "Lower Confidence Level" and "Upper Confidence Level," which are the same as the "95% Confidence Level" in Excel. Since this looks different, take the mean and add the "95% Confidence Level" from Excel, and then take the mean and subtract the "95% Confidence Level" from Excel, and you will get the UCL and LCL respectively. Rattle presents the same result in a different manner. Please do not let that throw you as an analyst. Different tools may present results differently, but that does not mean that they are inconsistent in result, just different in format.

```
Basic statistics for each numeric variable of the dataset.

$TOR_LENGTH
                X...X.i
nobs         269.000000
NAs            0.000000
Minimum        0.000000
Maximum       92.600000
1. Quartile    0.000000
3. Quartile    4.100000
Mean           4.443494
Median         0.500000
Sum         1195.300000
SE Mean        0.623786
LCL Mean       3.215350
UCL Mean       5.671639
Variance     104.670377
Stdev         10.230854
Skewness       4.327380
Kurtosis      24.968456
```

	A	B
1	*TOR_LENGTH*	
2		
3	Mean	4.443494424
4	Standard Error	0.623786189
5	Median	0.5
6	Mode	0
7	Standard Deviation	10.23085418
8	Sample Variance	104.6703773
9	Kurtosis	25.67453191
10	Skewness	4.376062845
11	Range	92.6
12	Minimum	0
13	Maximum	92.6
14	Sum	1195.3
15	Count	269
16	Largest(5)	44.8
17	Smallest(5)	0
18	Confidence Level(95.0%)	1.228144665
19		
20		

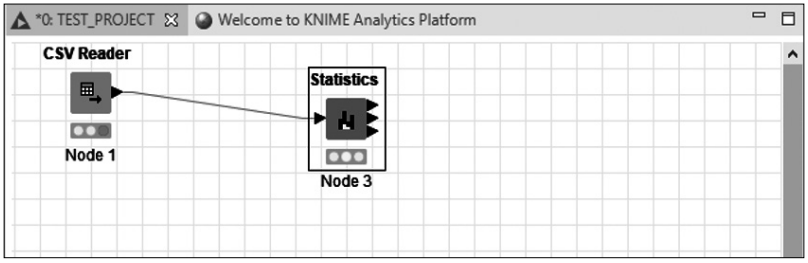
TOR_LENGTH StormEven

Row Name	Definition
Nobs	Number of Objects
NAs	Missing Data
Minimum	Minimum Value
Maximum	Maximum Value
1. Quartile	First Quartile (25th Percentile)
3. Quartile	Third Quartile (75th Percentile)
Mean	Arithmetic center of data
Median	Physical center of data
Sum	Additive values of all data
SE mean	Standard Error (standard deviation/ square root of objects)
LCL Mean	Lower Confidence Level of Mean
UCL Mean	Upper Confidence Level of Mean
Variance	Sum of Squares difference between each value and mean
Stdev	Standard Deviation (square root of Variance)
Skewness	Positive means right skew, negative means left skew, 0 means normal distribution (or close)
Kurtosis	The "peak" of the data (higher value means sharper peak)

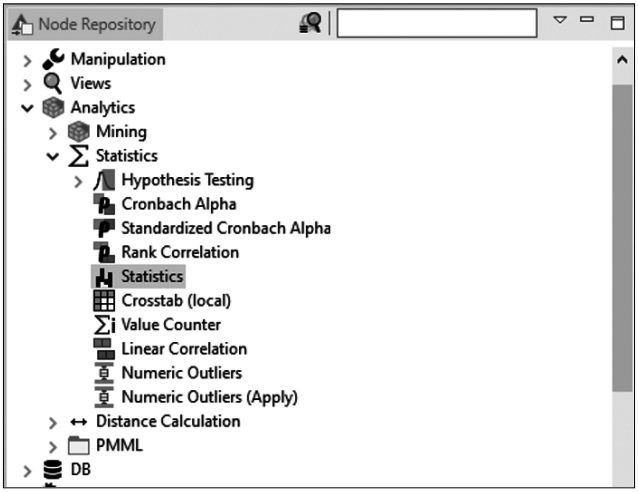
So far, a comparison of all the results seems to point to some very small discrepancies that can be caused by the type of formula or by rounding in that formula. Fortunately, the tools are consistent in their main figures, which says that using several tools to verify the results is something that needs further exploration, which will happen in this book.

3.1.4 KNIME

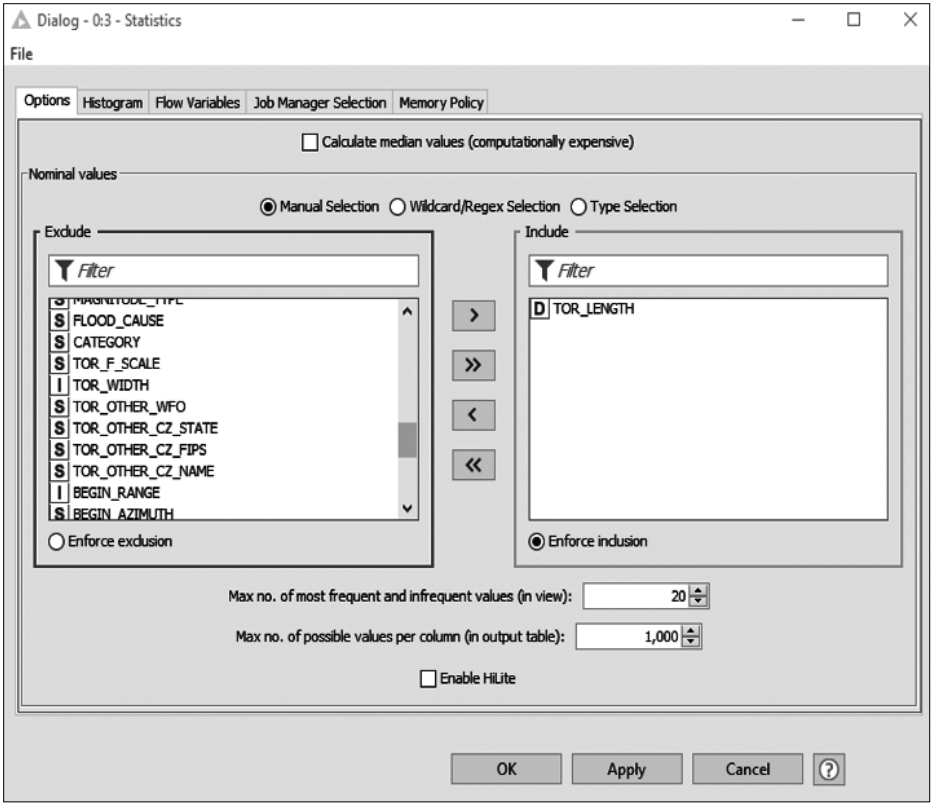
KNIME is modular in nature, so using the node that will give you descriptive statistics is the best way to perform this function as easily as possible. There is some data preparation that must be done prior to this step. The first step is to open KNIME to either a new project or one that you have already saved as shown.



As one can see, an analyst can add any of the nodes that are available within KNIME to the workspace. But which one to add? Where is information on that node? The answers are available right in the workspace of KNIME's application. Choose the "Statistics" node located within the "Statistics" part of the "Analytics" category of nodes. The location is shown as follows. After locating the node, left-click and hold, pulling the node onto the workspace. After you relocate the node, connect the nodes by clicking and holding onto the "black triangle" located on the right side of the "CSV Reader" node and connect it to the "Statistics" node. The workspace is now ready for performing the function of the node.



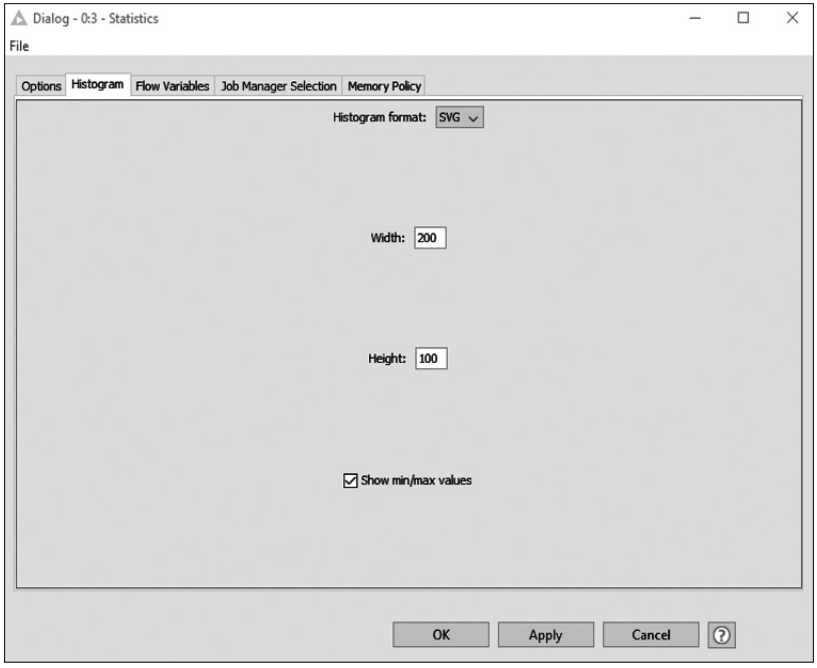
After placing and connecting the nodes, double-click on the "Statistics" node and this screen will appear. What this screen does is perform descriptive statistics on the variable or variables that the analyst chooses. In this case, only "TOR_LENGTH" will be chosen, to be consistent with the other tool functions. The way to clear the right side of the screen (where the variables the analyst wants evaluated exist) is to click on the double left arrow (<<) and then on the left side of the screen choose just TOR_LENGTH, moving it to the right side of the screen with a click of the single right arrow (>). Now the node is configured with the variable, but other preparation needs to be set before clicking on "Execute."



Notice the two text boxes below the selection screens. What they denote is the maximum number of different values and the maximum number of values in the variables that the analyst chose. If there are more than 20 different values, KNIME will ignore them. It might be beneficial to change this to a

higher number to account for a higher number of value changes in case there are unique values in the variable or column. In this case, a change to 100 should suffice for TOR_LENGTH. The second block is sufficient since the number of rows in this dataset was about 300, so the number placed in this block should more than cover this column.

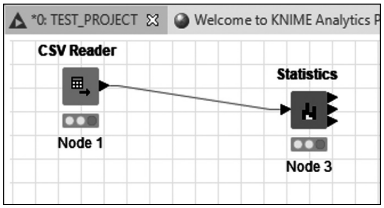
Once that is completed move on to the next tab, "Histogram," which will look like the following screen. There is not much to change here unless the analyst needs more pixel space or a larger image. The screen will default to SVG, but it can be changed to PNG if desired. Explore both to see which one will suit your presentation or article.



Once all the screens meet the analyst's needs, click OK or Apply and a message may appear stating that the node has been changed and asking if the analyst wants this change. Click OK again if this message appears and you want to change the node.

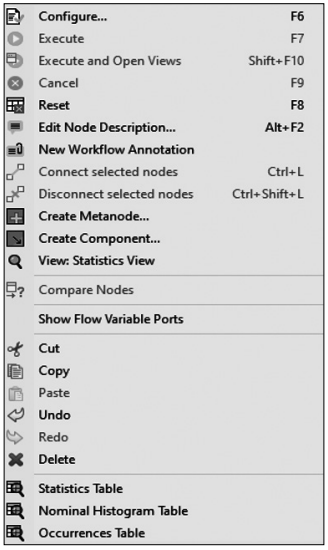
One note of caution is necessary here. On the first screen, the analyst will notice that the block labeled "Calculate Median Values" is not checked. This is by design, since calculating median values is computationally arduous, especially with many variables. Unfortunately, for the purposes of this demonstration, median values are part of the descriptive statistics that are necessary, so this block will have to be checked.

After the configuration is complete, the analyst should see the "CSV Reader" and "Statistics" nodes in a flow configuration connected with the output of CSV Reader to Statistics. By clicking on the "double green arrow" (illustrated as follows), the analyst will execute all the nodes and green lights should appear on all the nodes as in the screen following the double green arrow illustration.



If the analyst right-clicks the "Statistics" node, they will see a choice to view the summary table shown as follows. This will present the result to the analyst of the function that was just performed. Following are both this screen and the screen for the results.

Once the table appears, the analyst will notice that the table contains all the different variables. If the analyst scrolls to see TOR_LENGTH, it will show the same results as other tools.



File Hilite Navigation View

Table "default" - Rows: 25 | Spec - Columns: 16 | Properties | Flow Variables

Row ID	S Column	D Min	D Max	D Mean	D Std. deviation	D Variance	D Skewness	D Kurtosis	D Overall sum
TOR_LENGTH	TOR_LENGTH	0	92.6	4.443	10.231	104.67	4.376	25.675	1,195.3

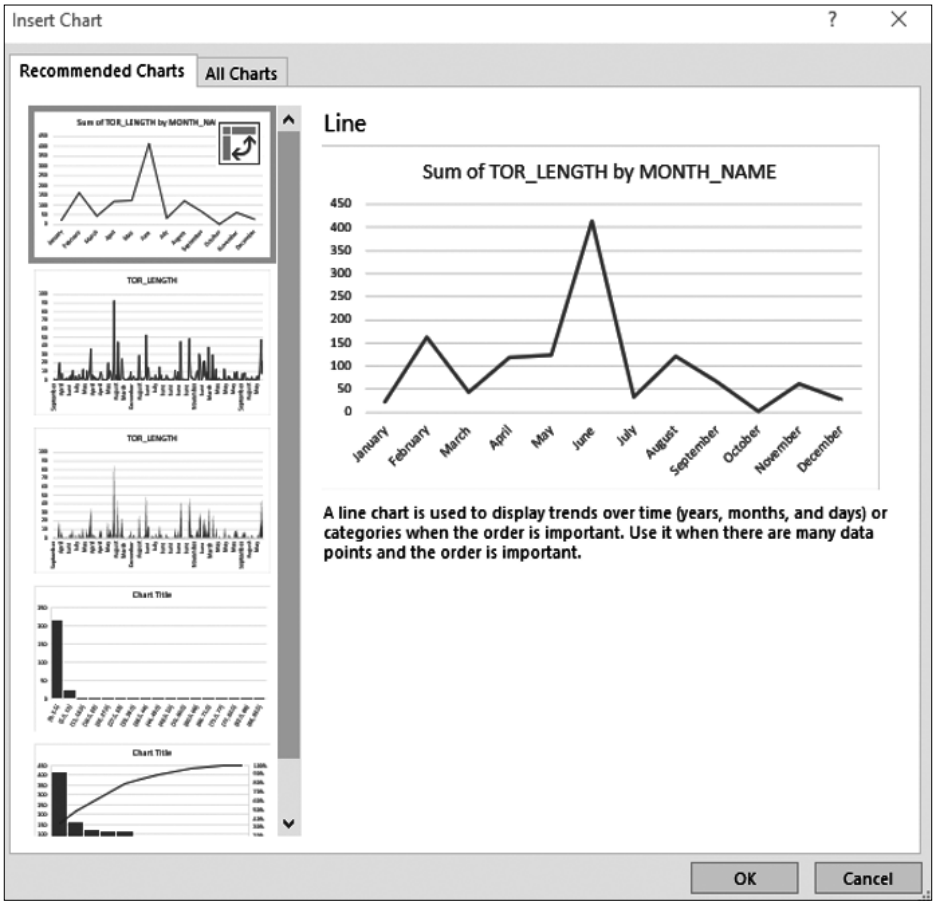
3.2 CUMULATIVE PROBABILITY CHARTS

Although covered in a very rudimentary way in most statistics classes, a *cumulative probability chart* is a very useful way of presenting data to show where the main issues arise. As an example, if a manager wanted to see which departments were taking the most paid time off (PTO) a year, by month, the manager might use this chart in order to determine which departments (and which months) seem to have the most inclination toward PTO for employees. It always surprised me that, when performing this statistical function, the main months for PTO were not December or January, but more toward the spring and autumn. This correlated with graduations (college and high school), along with football games (specifically away games). In all, this is useful for many industries, from banks to metal fabrication. This book will address each tool with the same dataset that has been used all along and will focus on two variables, TOR_LENGTH and MONTH_NAME, in order to see if the tornados occur the most during certain months and use the probability chart to show this data.

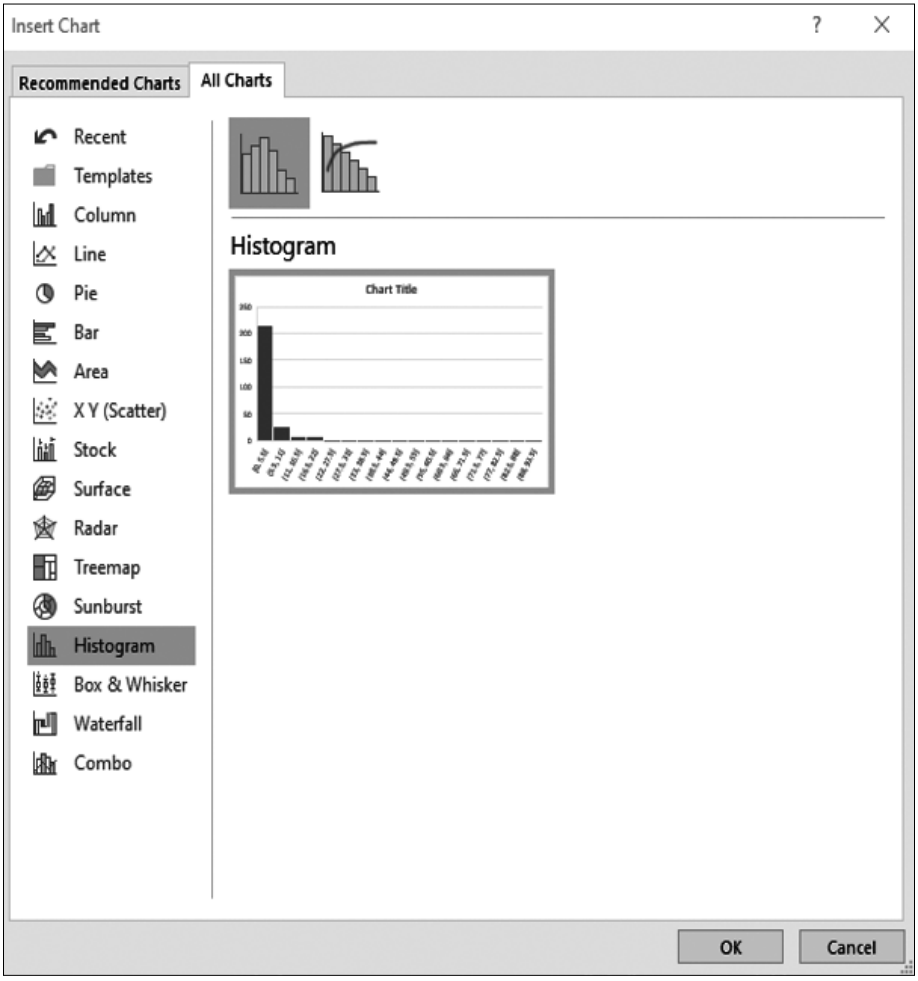
3.2.1 Excel

Excel has an already existing function prepared for cumulative probability charts, otherwise called *Pareto charts*. The analyst does not have to proceed through a pivot table or chart and can automatically make a chart from the dataset selection.

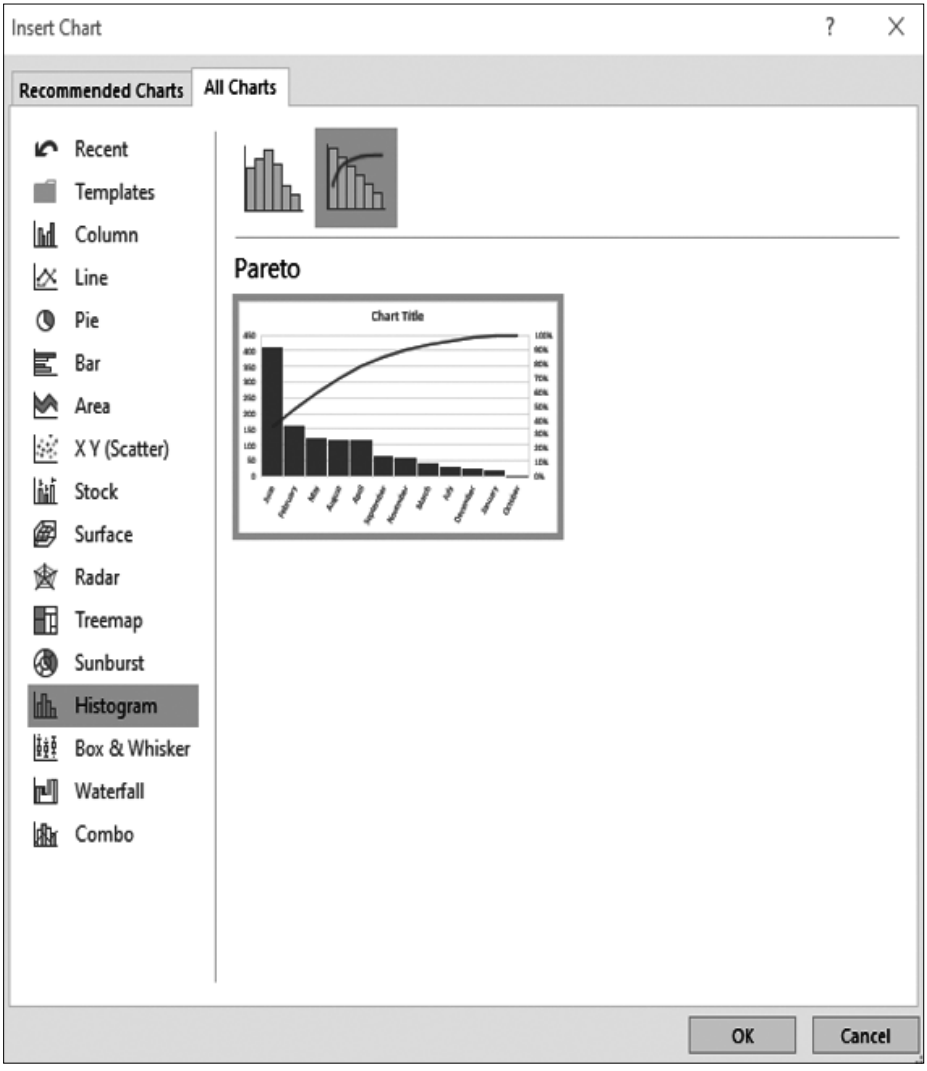
As stated before, the analyst should select the two columns that will be charted, namely TOR_LENGTH and MONTH_NAME. To do this, select the first column and hold the "CTRL" key down while selecting the second column. After this is completed, select "Insert" from the main toolbar and select "Recommended Charts." This will produce the following screen.



At this point, select the "All Charts" tab (to the right of the selected tab) and you will see the following screen.

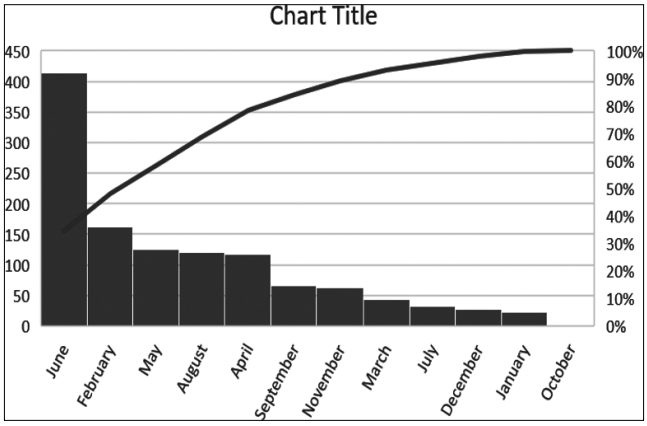


Once you select "Histogram" from the left menu, two sub-choices appear in the right screen at the upper-left column. The one to the left is a conventional histogram, and the one to the right is the cumulative probability plot. Choose the one to the right, and the finished chart will appear in preview as shown.



Click OK and the chart will appear, but what does it mean? The interpretation is that there are greater tornado lengths (basically longer tornados) in some months than in others. In order to do the conventional Pareto measurement, keep looking right until the 80% mark is achieved, and that will show the months that produce 80% of the longer tornados. The months would be June, February, May, August, and April. Remember that this only takes into consideration the year 1951.

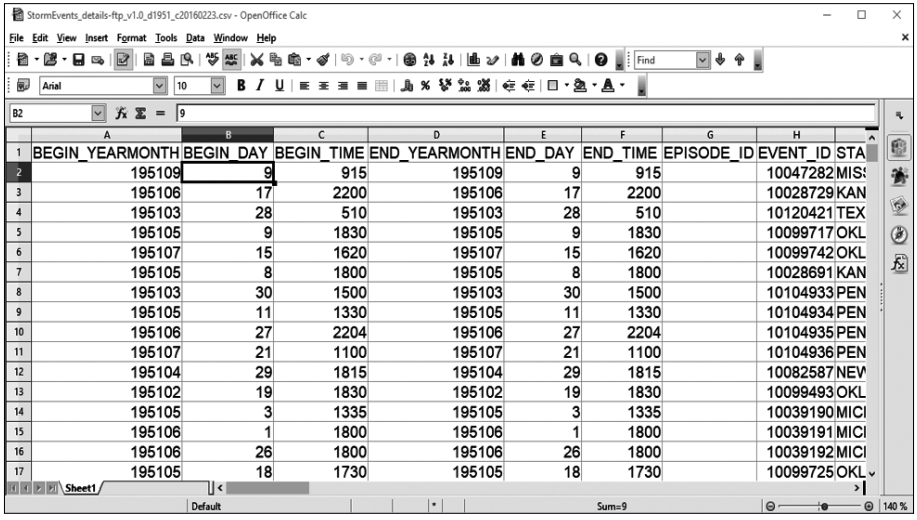
The analyst can now do a similar chart with states and see the states in the United States where 80% of the longer tornados occur. Try it and you will be shocked by the states that have the longer tornados. Not the ones you would expect! The magic of dataset analysis.



3.2.2 OpenOffice

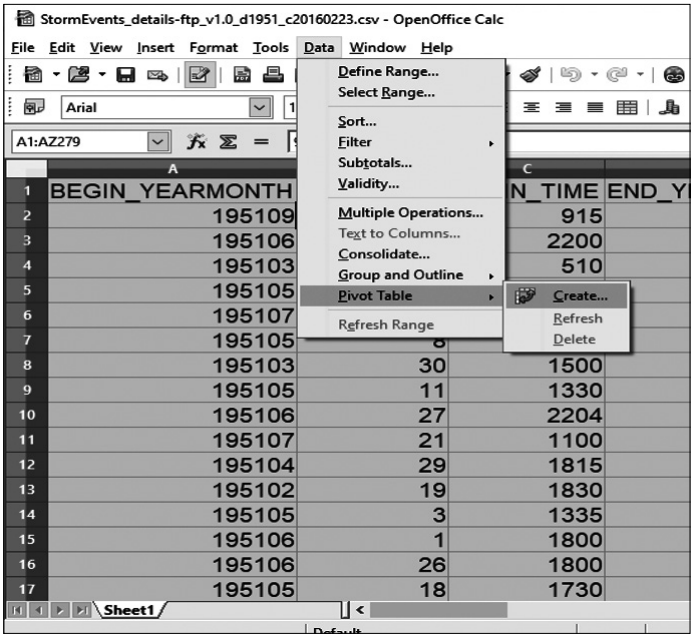
OpenOffice does not have the "magic" button that Excel possesses, but it does have the ability to produce a cumulative probability chart within its *pivot table* capability.

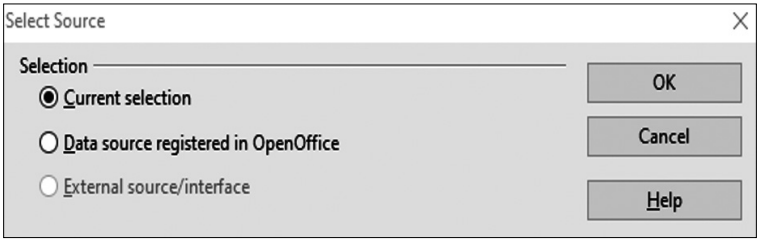
The first step is to open OpenOffice and ensure that the same dataset that was loaded into Excel is loaded in OpenOffice. The screen at this point should be the same as the following:



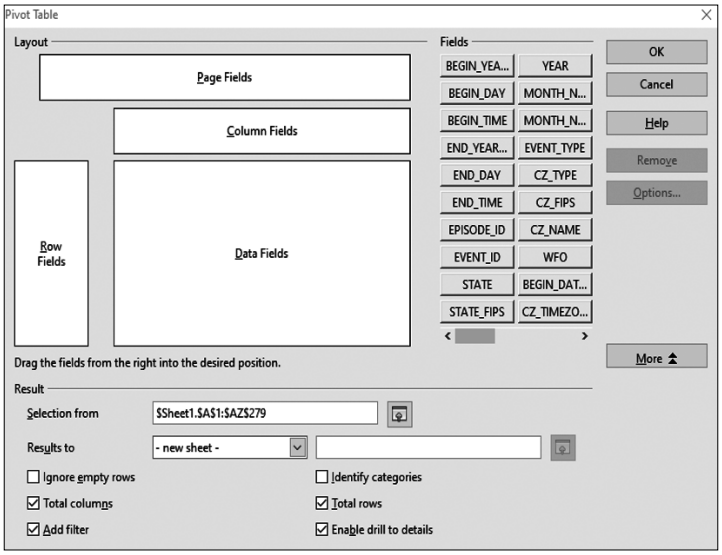
	A	B	C	D	E	F	G	H	
1	BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EPISODE_ID	EVENT_ID	STA
2	195109	9	915	195109	9	915		10047282	MIS
3	195106	17	2200	195106	17	2200		10028729	KAN
4	195103	28	510	195103	28	510		10120421	TEX
5	195105	9	1830	195105	9	1830		10099717	OKL
6	195107	15	1620	195107	15	1620		10099742	OKL
7	195105	8	1800	195105	8	1800		10028691	KAN
8	195103	30	1500	195103	30	1500		10104933	PEN
9	195105	11	1330	195105	11	1330		10104934	PEN
10	195106	27	2204	195106	27	2204		10104935	PEN
11	195107	21	1100	195107	21	1100		10104936	PEN
12	195104	29	1815	195104	29	1815		10082587	NEV
13	195102	19	1830	195102	19	1830		10099493	OKL
14	195105	3	1335	195105	3	1335		10039190	MIC
15	195106	1	1800	195106	1	1800		10039191	MIC
16	195106	26	1800	195106	26	1800		10039192	MIC
17	195105	18	1730	195105	18	1730		10099725	OKL

The next step to do is to insert a pivot chart (just as with Excel) in order to use the "group-by" approach to the data and enable the cumulative probability function. This is located in the "Data" area of the main toolbar and is shown as follows. Select "Pivot Table" and then "Create," at which time the next screen will appear.

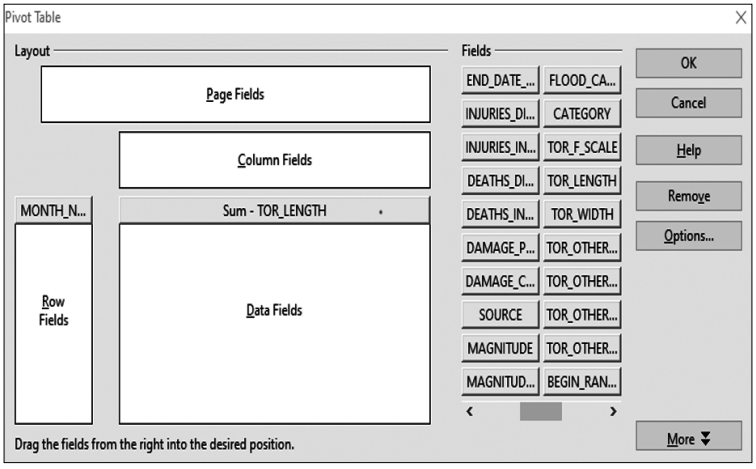




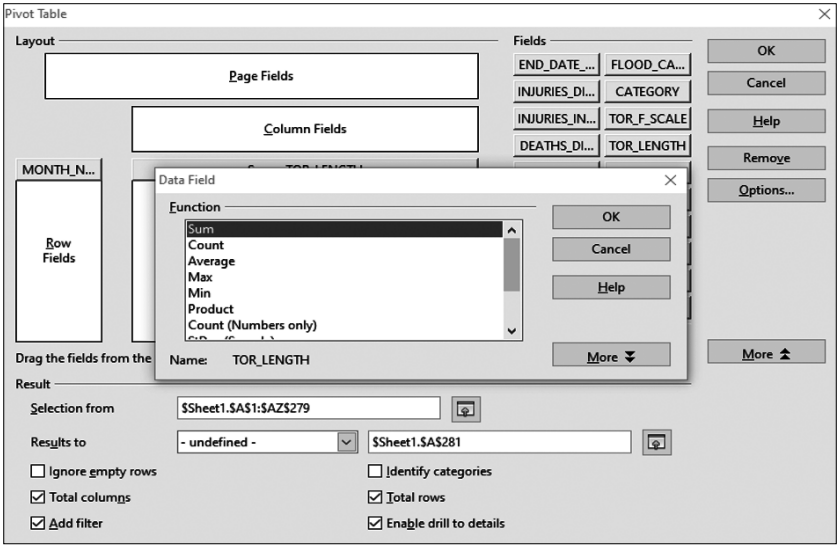
The "Current selection" radio button will be the default in this situation; click OK to display the table as follows.



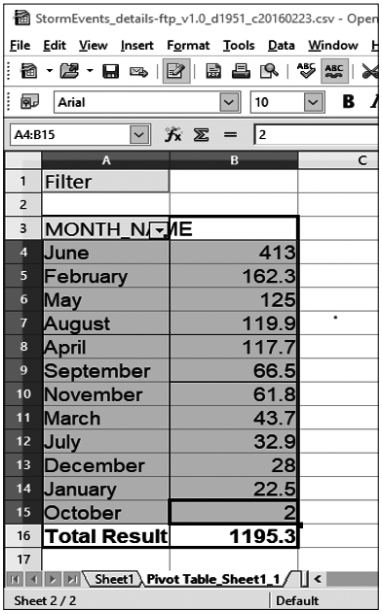
There is a lot of activity in this screen, but the first step is to ensure that the information below the "More" button is correct. Ensure that "Selection from" represents the data you want in the pivot chart and that "Results to" is to a new sheet. This way, the data will not be "shoehorned" into the same worksheet as the other data. Explore the "checkboxes" so that they match the analyst's configuration in order to get the most analysis for the function. After that configure the Pivot Table as follows for this example. The reasoning is that the requirement is to understand how tornado length is associated with the month of the year. The analyst will place "MONTH_NAME" into the Row Fields and "TOR_LENGTH" into the Data Fields as follows.



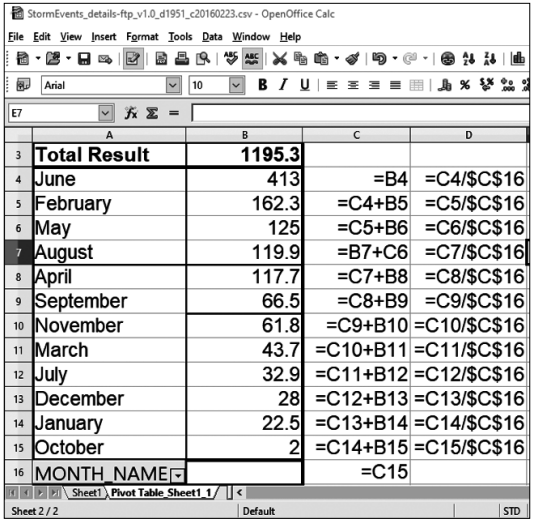
Notice that TOR_LENGTH has "Sum" to the left of the column header. The sum is appreciated, but the average tornado length is where the real requirement is located. In order to change "Sum" to "Average," left-click on the gray bar marked "Sum-TOR_LENGTH" and look to the right to see there is an "Options…" selection that is dark gray. Click on that alternative and the following screen will appear. Choose "Average" and click OK to get the label change on TOR_LENGTH. At this point, click OK and the data will then appear as a regular dataset with just month and tornado length.

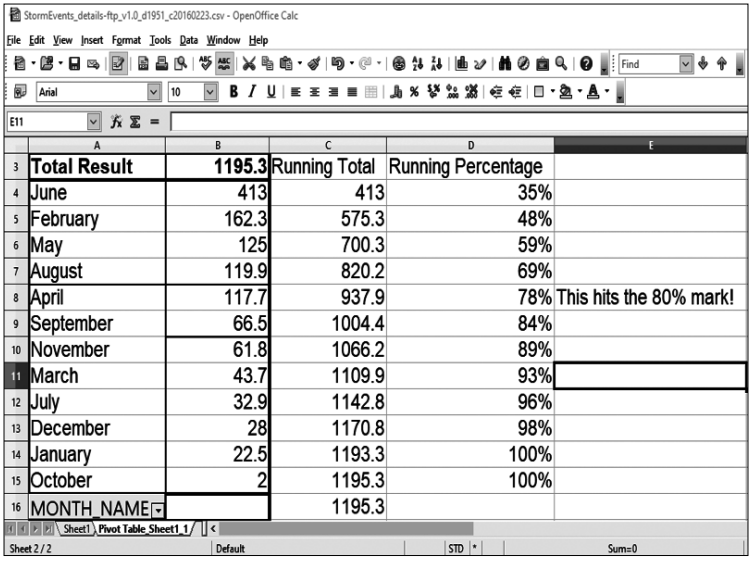


Now, the analyst must revert back to their knowledge of Excel prior to Excel having the ability to place the same variable in the rows for different purposes. OpenOffice does not allow this, but part of the work is already done. The analyst must sort the numerical data in descending order to show the same sequence of months as in Excel. That screen is as follows:



What has been accomplished thus far is just part of the requirement for the cumulative probability chart. Once this is done, add two more columns and perform a running total; the formulas are shown in the next screen for the running total and the running percentage. When this is finished, the next step will be to insert a chart to properly display the results.




StormEvents_details-ftp_v1.0_d1951_c20160223.csv - OpenOffice Calc

File Edit View Insert Format Tools Data Window Help

Arial 10

E11

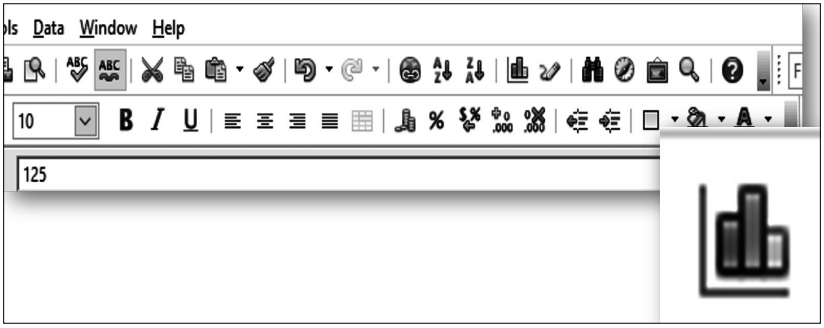
	A	B	C	D	E
3	**Total Result**	**1195.3**	Running Total	Running Percentage	
4	June	413	413	35%	
5	February	162.3	575.3	48%	
6	May	125	700.3	59%	
7	August	119.9	820.2	69%	
8	April	117.7	937.9	78%	This hits the 80% mark!
9	September	66.5	1004.4	84%	
10	November	61.8	1066.2	89%	
11	March	43.7	1109.9	93%	
12	July	32.9	1142.8	96%	
13	December	28	1170.8	98%	
14	January	22.5	1193.3	100%	
15	October	2	1195.3	100%	
16	MONTH_NAME		1195.3		

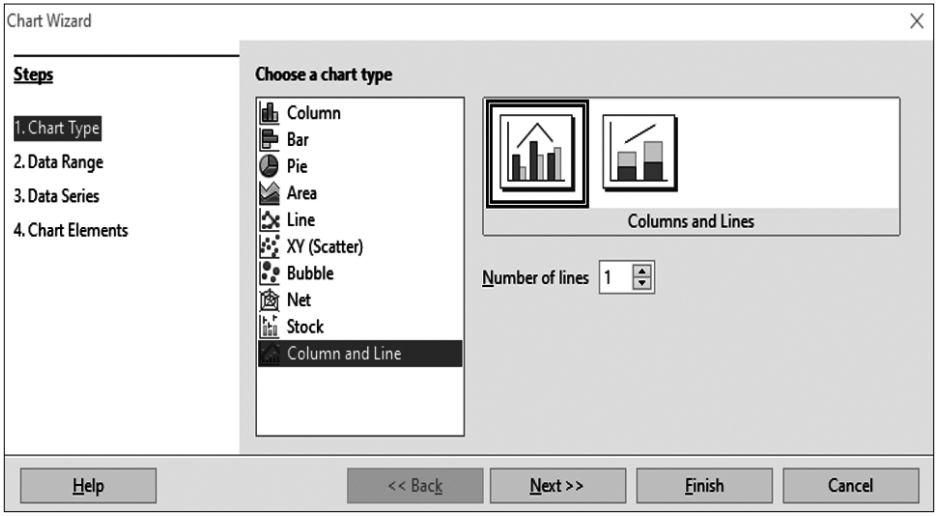
Sheet1 Pivot Table_Sheet1_1

Sheet 2 / 2 Default STD Sum=0

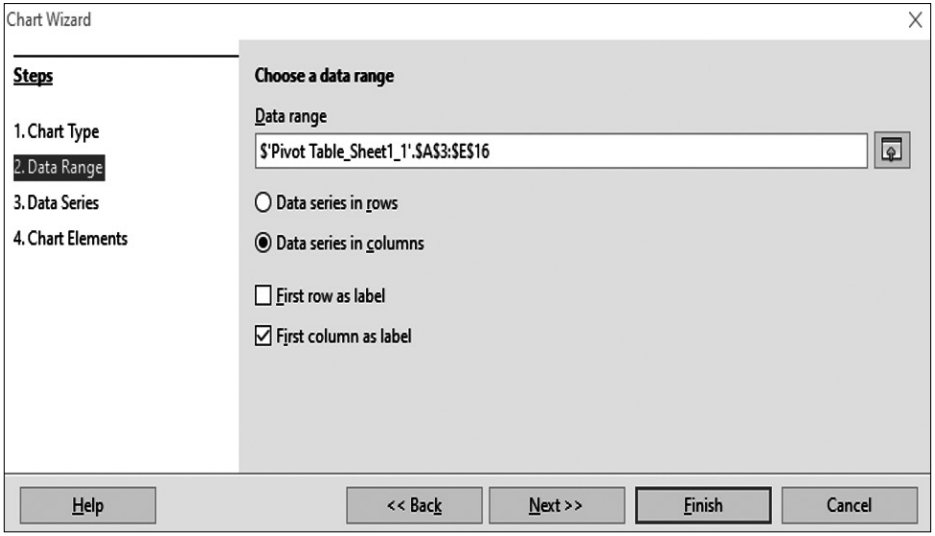
There are some configuration steps that must be accomplished before the analyst gets the same results as with Excel. Now, let's look closely at the necessary steps.

The first step is to use the "Insert" toolbar to insert a chart; in this case the normal bar chart is fine, but there is a combination bar line chart depicted as follows that works very well in this situation. Once this is selected, the next screen will appear to show the different chart choices.

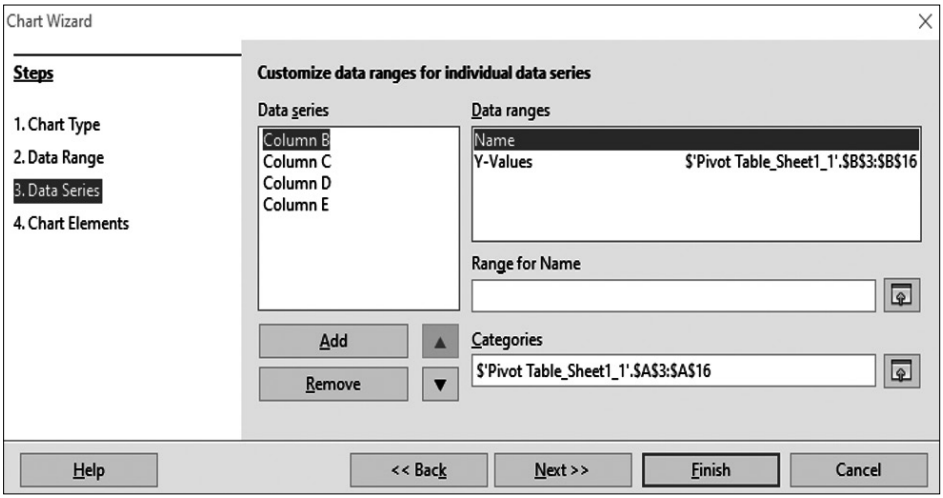




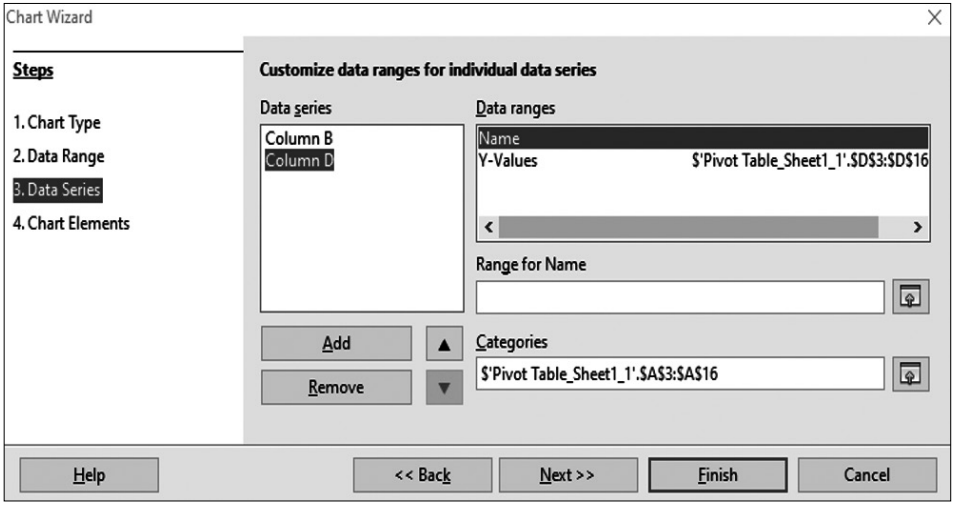
Once the Column and Line chart is selected, go to "2. Data Range" to see the range of data included in the chart.



It would appear the data is as desired, but there needs to be removal of some of the data to make the table "cleaner." In this case, move on to "3. Data Series" to view all the series that are on the chart.



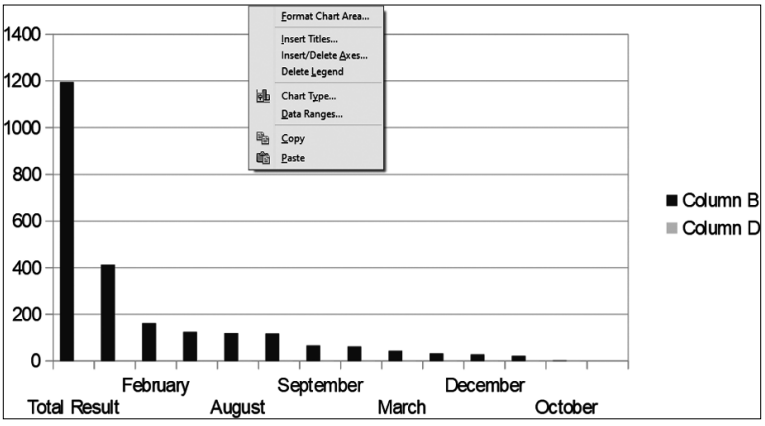
As the analyst reviews the chart and the table, it would be advisable to remove Column E, which does not contribute to the table, and Column C, which is just the running total, leaving Column B, which shows in descending order the tornado lengths by month, and Column D, which is the percentage of those lengths. Removing them is simple—select the column and click on the "Remove" button. The finished screen is as follows:

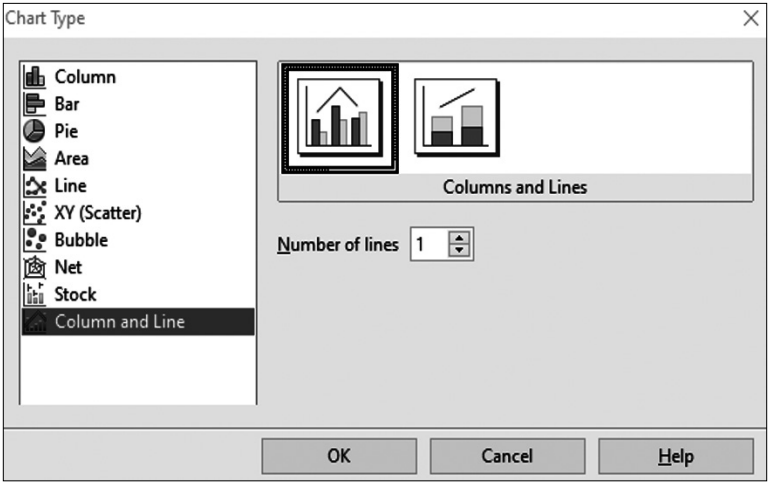


At this juncture, the analyst needs to turn their attention to the chart, which now shows what seems to be just one element. The reason for this is

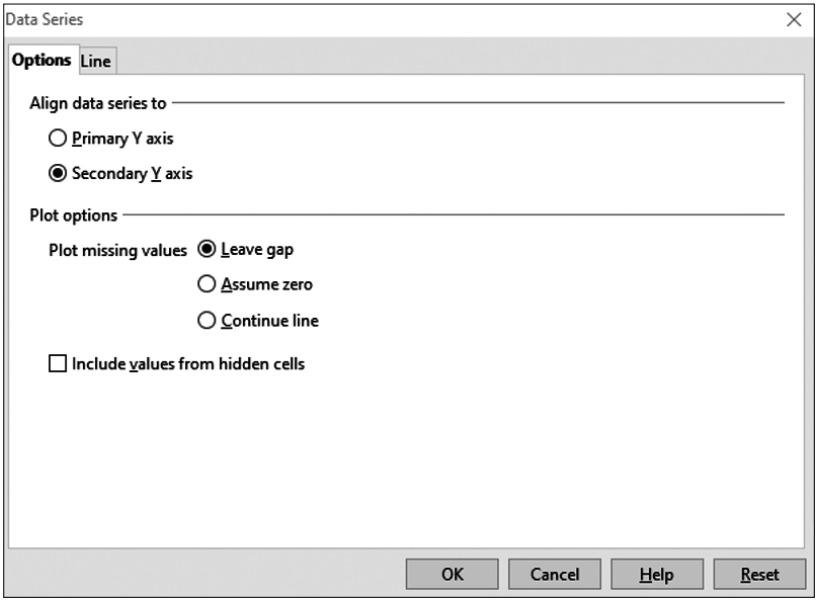
because the other element's highest value is 100 (100%) and does not appear within the range of the numbers. The analyst will need to make a secondary axis for this percentage, and that will enable the data to appear.

The next step will be to establish the secondary axis. This is done by first establishing a line for the percentage running total (Column D). Right-click on the chart and choose "Chart Type" as shown:

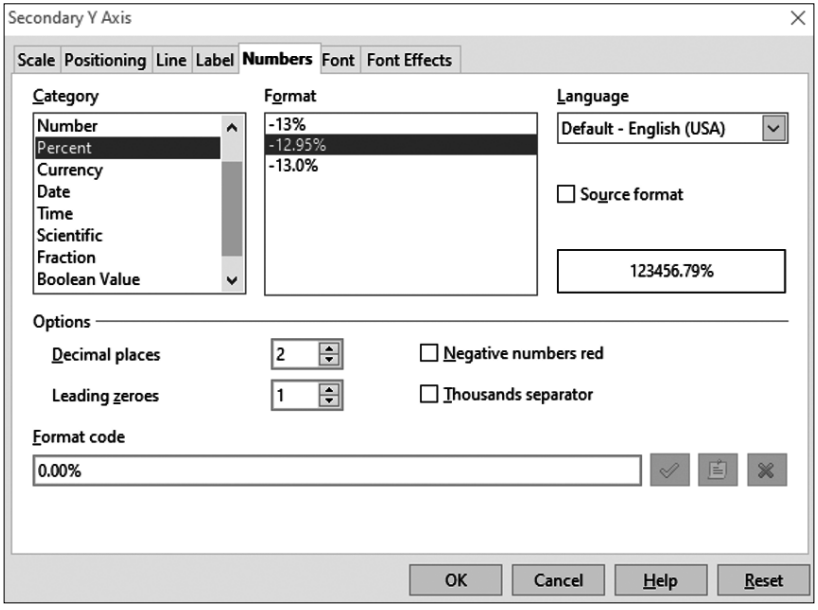




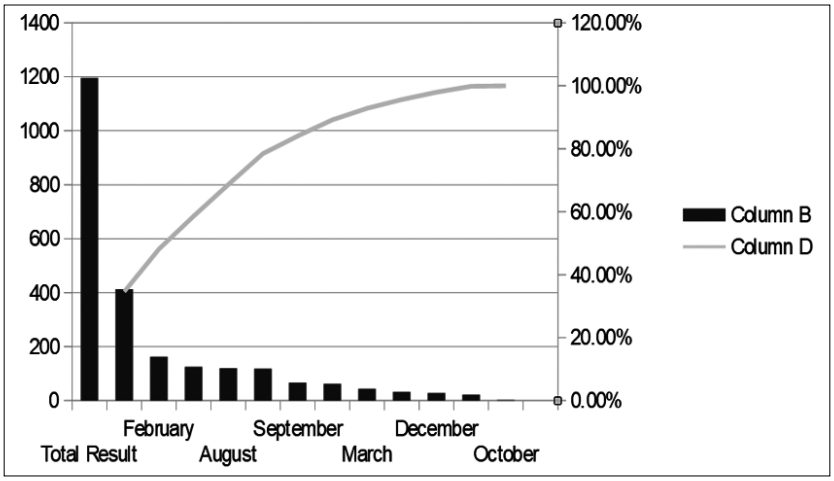
By performing this function, the analyst will then reveal a yellow line that goes across the x-axis. This line represents the percentage running total. The goal is to get this line to display on the same graph as the columns. Double-click on the yellow line and the following screen will appear. Ensure that you attribute this line to the "Secondary y-axis," since the goal is to display both the raw numbers and percentages on the same graph.



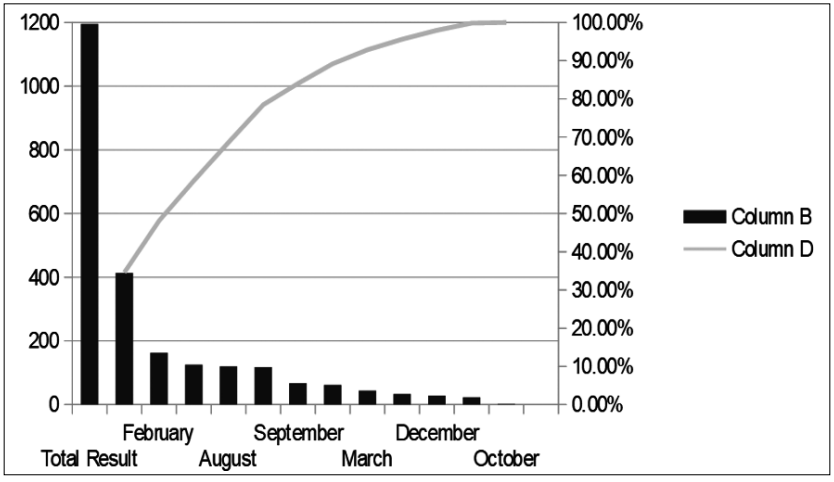
This will produce the following chart, which does not show the true percentage but more of probability between 0 and 1. To change this to percentage, double-click on the right-hand side numbers and this screen will appear.



Click on the tab marked "Numbers" and remove the check in the checkbox labeled "Source Format" and then choose "Percent" from the left-hand list. Click OK and the following chart will appear.



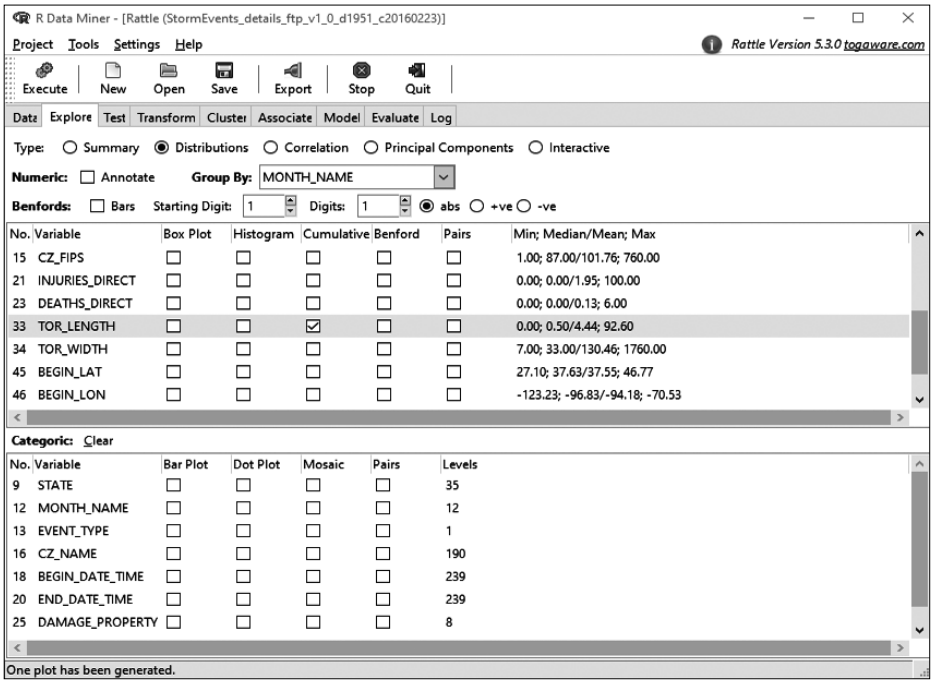
At this point, explore some of the other functions within the chart to reduce the numbers in order to eliminate the additional space at the top of the values so that the chart looks like the one that follows. A significant amount of work, but the same result. Once the analyst practices this function, it will become second nature.



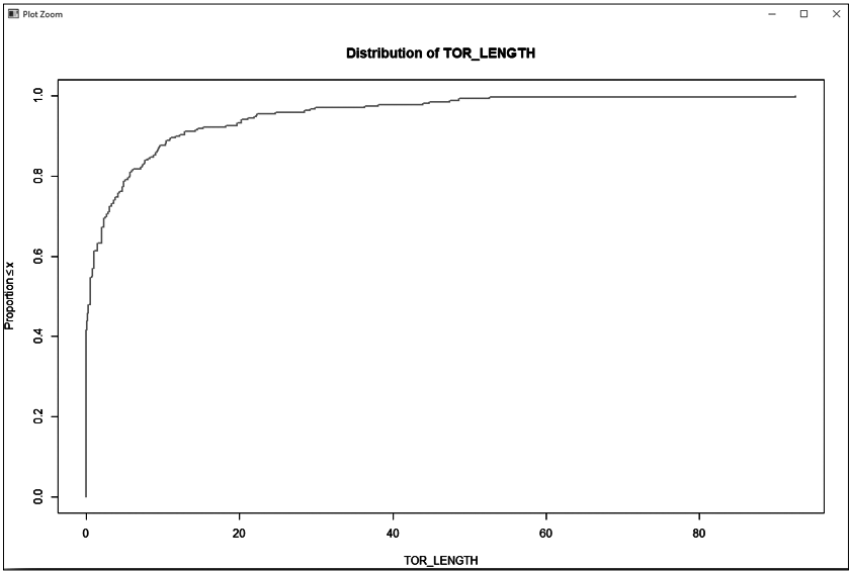
3.2.3 R/RStudio/Rattle

The process for producing graphs in Rattle is very simple. However, the process for producing a graph similar to the ones that have been covered is much more complicated. Since the process for producing conventional graphs is more straightforward, the subject shall take a slight turn off the main road for this tool in this convention.

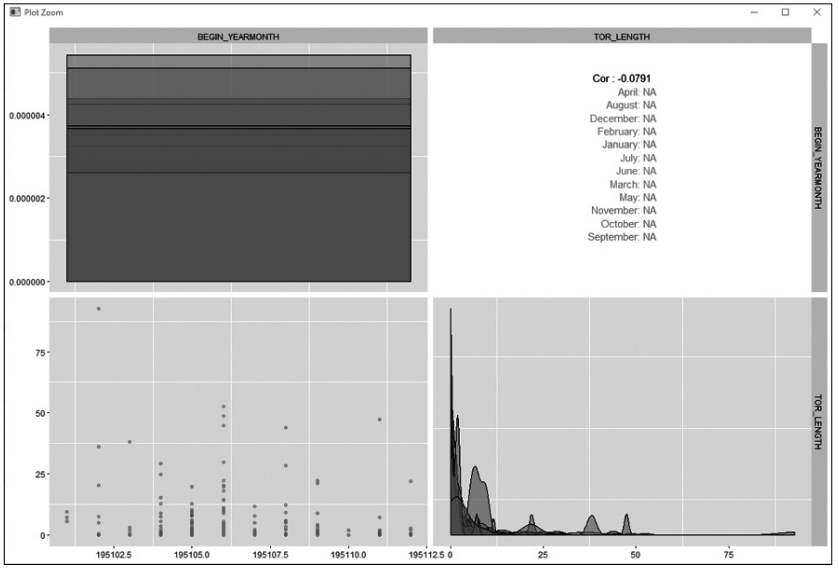
In Rattle, there is an "Explore" tab that provides many different functions for the analyst. One of those is "Distributions," which offers the analyst a wide array of data visualizations and, together, could do the same as the cumulative probability plot presented in previous sections. The first distribution is the "Cumulative" chart, which is simple enough to choose using the checkbox on the "Distributions" area, which is shown as follows.



On this screen, there is a "Group By:" drop-down box where the analyst can choose the variable by which the main categoric variable is grouped, the same as in a pivot table. In this instance, the numeric variable TOR_LENGTH is not paired with any other variable. In other words, the resulting graph pictured as follows does not have an association with the MONTH_NAME as in the previous sections.



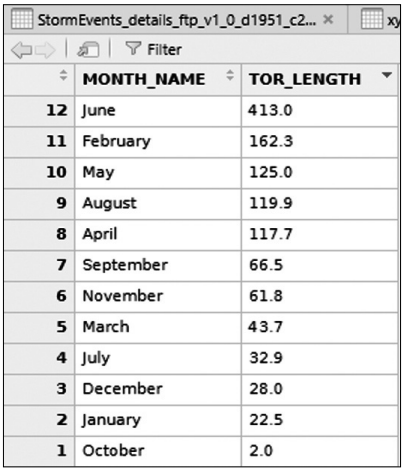
What the previous chart shows is that about 80% of the tornados have lengths that are under 15, but it does not relate length to MONTH_NAME. There is a relationship chart that does depict this, but not in the same way as in previous sections. This is the "Pairs" checkbox, where the analyst can select both TOR_LENGTH and MONTH_NAME as follows:



In order to have a true cumulative probability chart, a transformation of the data will have to be completed and programmatic plots of the data will have to take place to produce something similar to the previous results. In order to do this, the analyst will have to rely on RStudio in order to conduct this section rather than Rattle. We would urge the analyst to explore more Rattle distributions and similar functions to see it separately.

In the meantime, back to RStudio and the cumulative probability chart. The first step to do is to import the dataset into RStudio, which is covered in the section on Importing Data. The next step is to isolate the variables that we want to use for our cumulative probability chart, which is MONTH_NAME and TOR_LENGTH. RStudio provides a very nice Integrated Development Environment (IDE) that is on default at the bottom left-hand side of the screen.

The following screen shows some of the programming done to isolate the variables and make a cumulative raw sum of the tornado lengths. The resulting table is as follows, along with the programming.


StormEvents_details_ftp_v1_0_d1951_c2...

Filter

	MONTH_NAME	TOR_LENGTH
12	June	413.0
11	February	162.3
10	May	125.0
9	August	119.9
8	April	117.7
7	September	66.5
6	November	61.8
5	March	43.7
4	July	32.9
3	December	28.0
2	January	22.5
1	October	2.0

```
xyz<-xy%>%group_by(MONTH_NAME)%>%
summarize(TOR_LENGTH=sum(TOR_LENGTH))
```

What the previous programming achieves is grouping the data by MONTH_NAME and TOR_LENGTH, summing up the raw tornado lengths and grouping them by month, much like it was done in the Pivot Tables in the previous sections. Now comes the task of performing a cumulative percentage and then plotting that in RStudio. The results depend on adding a column

with the cumulative sums and then taking those sums and changing them to percentages. The screen and programming show the table necessary to make the cumulative probability chart.

Now comes the challenging part. Everything in the table is exactly the way it should be when making the chart in applications like Excel or OpenOffice. Unfortunately, R is not made for charts that have two y-axes as must be done in this particular instance. Therefore, it is time to do some more programming in order to properly display this chart. The challenge is doing it so that the result will mirror the results from other tools.

StormEvents_details_ftp_v1_0_d1951_c2... × | xyzadd × | xyz × | xyzsort ×

Filter

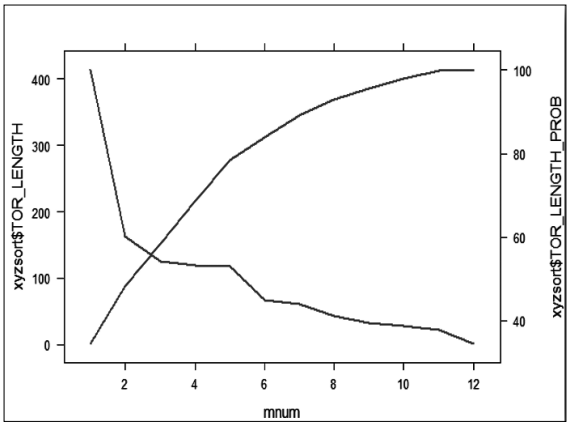
	MONTH_NAME	TOR_LENGTH	TOR_LENGTH_CUM	TOR_LENGTH_PROB
1	June	413.0	413.0	34.55200
2	February	162.3	575.3	48.13018
3	May	125.0	700.3	58.58780
4	August	119.9	820.2	68.61876
5	April	117.7	937.9	78.46566
6	September	66.5	1004.4	84.02911
7	November	61.8	1066.2	89.19936
8	March	43.7	1109.9	92.85535
9	July	32.9	1142.8	95.60780
10	December	28.0	1170.8	97.95031
11	January	22.5	1193.3	99.83268
12	October	2.0	1195.3	100.00000

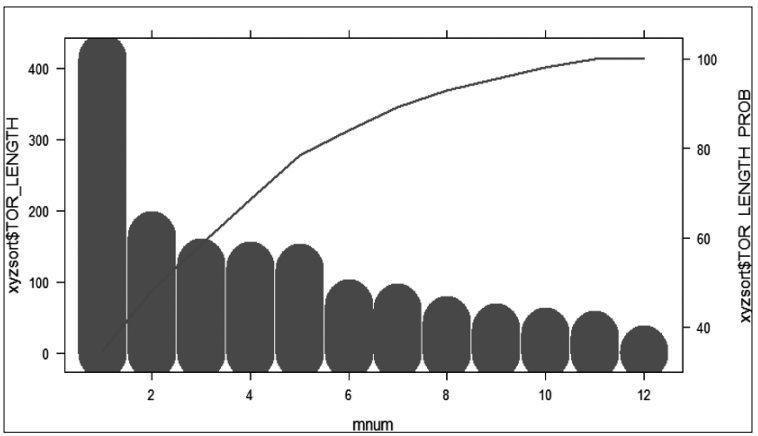
```
> xyzsort[,"TOR_LENGTH_CUM"]<-cumsum(xyzsort$TOR_LENGTH)
```

```
> xyzsort[,"TOR_LENGTH_PROB"]<-cumsum(xyzsort$TOR_
                    LENGTH/1195.3)
```

```
> xyzsort[,"TOR_LENGTH_PROB"]<-cumsum((xyzsort$TOR_
                  LENGTH/1195.3)*100)
```

```
> install.packages("latticeExtra")
```





The previous chart is pretty close to the previous cumulative probability charts, but there is room for some formatting. There is plenty of room to explore R, RStudio, and Rattle, and we will let the analyst continue to do the fine-tuning of this application. The programming to get this result is as follows, and the resources for these programming tips are located in the References section.

```
xy<-StormEvents_details_ftp_v1_0_d1951_c20160223_FIXED

COMMENT: Using "xy" as variable for imported dataset

xyz<-xy%>%group_by(MONTH_NAME)%>%summarize(length=sum
                        (TOR_LENGTH))
```

```
COMMENT: This acts like a Pivot Table, taking the tornado
  length (TOR_LENGTH) and grouping them by month (MONTH_
                          NAME)

      xyzsort<-arrange(xyz,desc(xyz$TOR_LENGTH))

  COMMENT: This sorts the tornado lengths in the dataset

  xyzsort[,"TOR_LENGTH_CUM"]<-cumsum(xyzsort$TOR_LENGTH)

  COMMENT: This allows the cumulative sum of the tornado
              lengths after the arranging

        mnum<-c(1,2,3,4,5,6,7,8,9,10,11,12)

    COMMENT: This provides for the row number that is
 associated with the MONTH_NAME (caution: This means that
   "1" = June, NOT January as the row sequence makes a
                       difference.)

        obj1<-xyplot(xyzsort$TOR_LENGTH ~ mnum,
              xyzsort,type="h",lwd=50)

   COMMENT: This requires the "LATTICEXTRA" package be
 installed and sets up the first series for the chart. The
                 "h" type mean histogram

obj2<-xyplot(xyzsort$TOR_LENGTH_PROB ~ mnum, xyzsort,type
              ="l",lwd=2,col="steelblue")

 COMMENT: This also requires the LATTICEXTRA package and
 sets up the second series for the chart. The "l" (lower
                  case L) type is line.

doubleYScale(obj1,obj2, add.ylab2=TRUE,use.style=FALSE)

  COMMENT: This provides the dual y-axis that you need
 for the chart and plots one as bar chart and the second
              (cumulative) as line chart
```

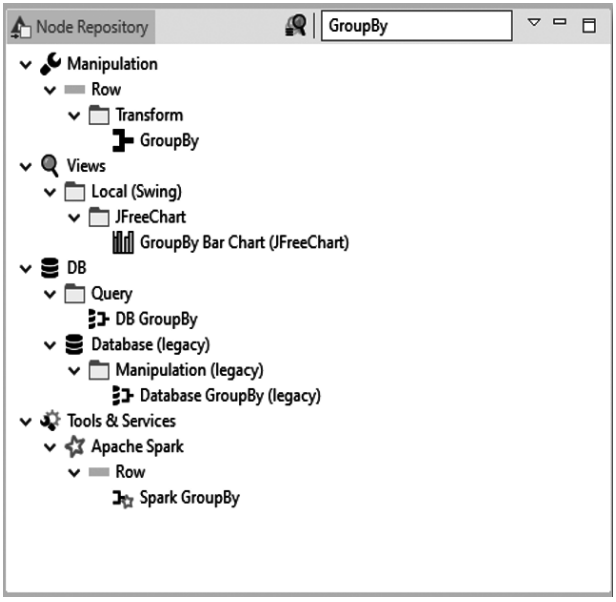
Yes, this is very complicated, but until R provides for an automated cumulative probability chart, this is just one of the many programming ways to perform this application. Unfortunately, programming is not just necessary in this instance, but required. Fortunately, there are numerous references for most of these types of applications. Please explore them using your favorite search tool.

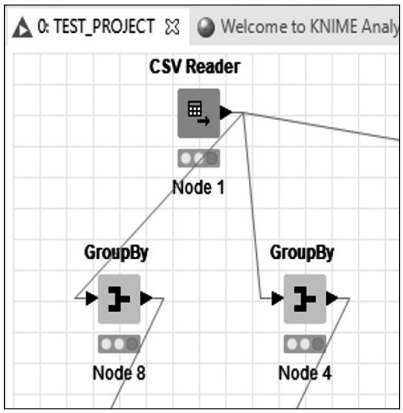
3.2.4 KNIME

The KNIME tool is also relatively complicated when it comes to cumulative probability charts, but there are nodes available to transition the dataset to the chart that, once done, can be adapted to other datasets of similar construction. Like a flow chart, the nodes form a step-by-step approach to the transformation of data into a chart. This process will proceed node by node for clarity.

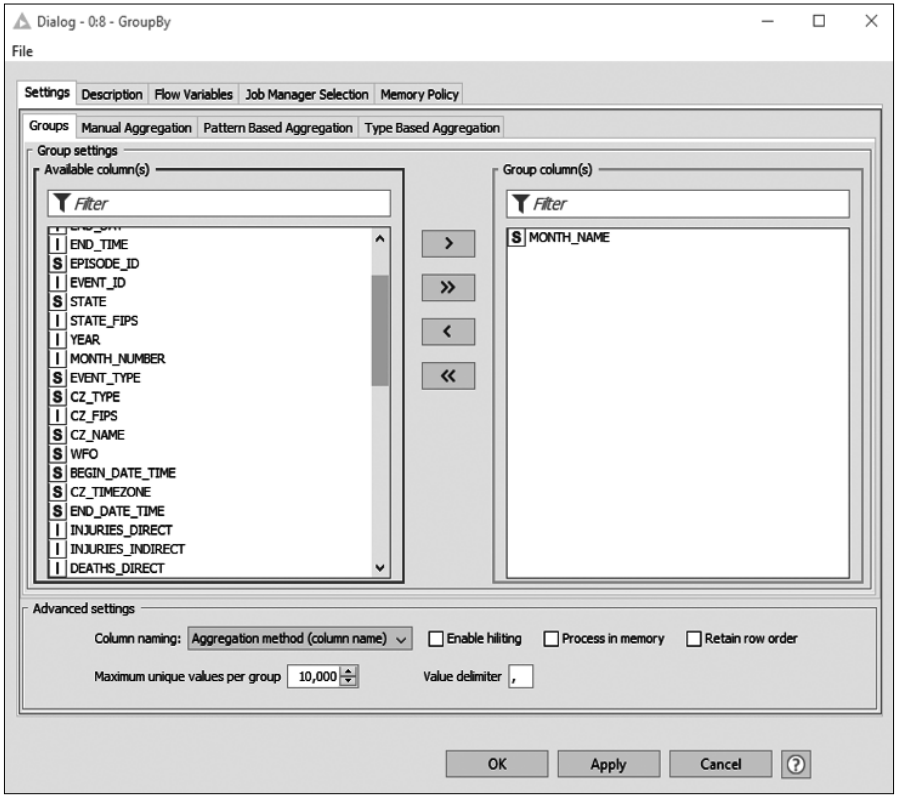
During the last section on KNIME, the statistics node was attached to the CSV Reader node in order to obtain descriptive statistics. In this case, there will be a number of nodes in order to get a table that is reflective of the ones in the past sections. Each of these nodes will transform the data into an end-product that the analyst can then export into Excel, OpenOffice, or R. In this case, export the finished table into OpenOffice as shown in the following and use the section on OpenOffice to make the CSV output into the cumulative probability chart. Sometimes it is just easier to export and use another tool and then try to complicate a chart result. In this case, exporting is easier.

The first node that is necessary to make the table is the "GroupBy" node, found by using the search block in the lower left-hand side menu grouping as shown. The analyst can also go to Manipulation -> Row -> Transform -> GroupBy in order to get to the node. Left-click on the node and drag it to the workspace and connect it to the CSV Reader node that the analyst had used previously. The screen should now appear as the following screen.





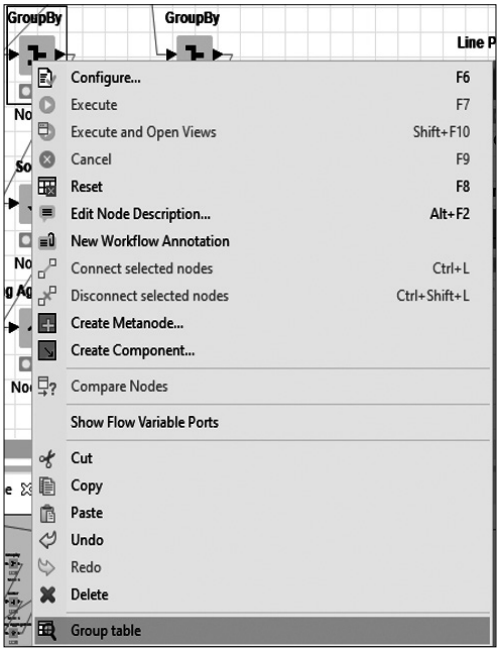
The screen shows two GroupBy nodes side by side connecting to the CSV Reader node. This is deliberate since the analyst will need both these nodes to incorporate into the final new table. The first node will be the one on the right (labeled Node 4). Double-click on this node to reveal the screen.



Use the single arrow (>) to move MONTH_NAME over to include it in Group Settings. What this does is to group the next column or columns by the Group name, which is MONTH_NAME. In essence you are doing the same as moving MONTH_NAME into the "Columns" space of the Pivot Table in Excel or OpenOffice. Once this is completed, please explore this screen to see some of the other options. None of these will be changed with the application, but in the future, exploring these options will present the analyst with many more derivations of this tool and node.

Next, move from the "Groups" tab to the "Manual Aggregation" tab to see the following screen. Moving TOR_LENGTH into the column as shown will now group TOR_LENGTH by MONTH_NAME. Ensure that the option "sum" is selected from the drop-down for this variable.

Click OK or Apply and OK to configure the node and remember to execute the node using either the single green or double green arrow. Once that is completed, right-click on Node 8 and select the last option at the bottom of the sub-menu for that node as depicted in the following screen.



Once the Group table option is selected, the resulting table will appear as in the following screen. This is a "preview" table and would need to be exported if the process was ended here. For now, this will serve as a placeholder to ensure the results are what the analyst expects.

Group table - 0:4 - GroupBy

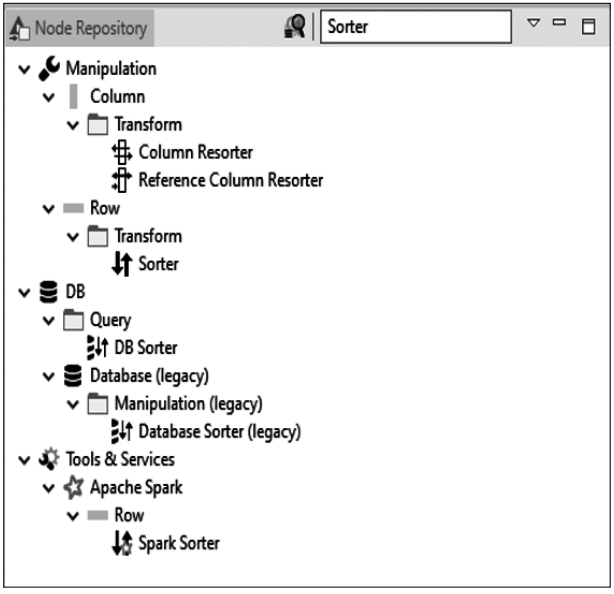
File Hilite Navigation View

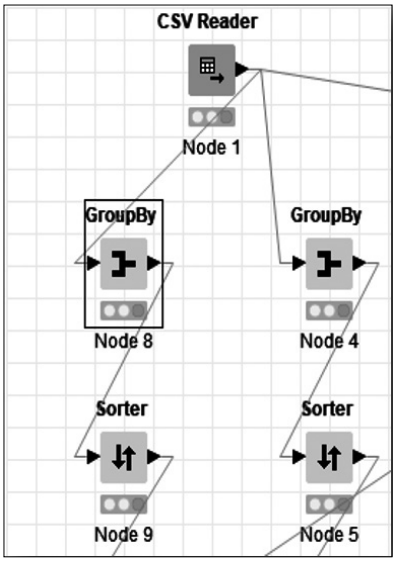
Table "default" - Rows: 12 | Spec - Columns: 2 | Properties

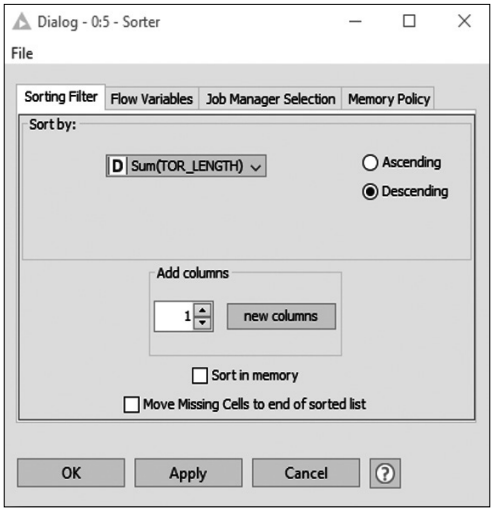
Row ID	S MONTH...	D Sum(T...
Row0	April	117.7
Row1	August	119.9
Row2	December	28
Row3	February	162.3
Row4	January	22.5
Row5	July	32.9
Row6	June	413
Row7	March	43.7
Row8	May	125
Row9	November	61.8
Row10	October	2
Row11	September	66.5

So far, so good. What this table shows is the sum of the tornado lengths per month, but notice that the months are in alphabetical order. The tornado lengths have to be sorted in descending order so that the bar graph portion of the cumulative probability chart is complete.

The next node to implement is the Sorter node, located here in the lower left sub-menu. Once that has been selected, dragged and dropped, and connected to the GroupBy node, double-click on the Sorter node to reveal a screen showing the different configuration options.

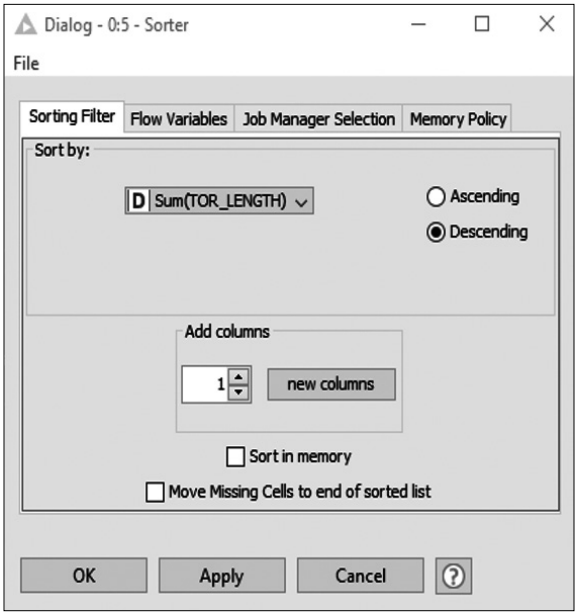






One note before continuing with this process. Please notice that the GroupBy and now the Sorter are located under the "Row" category. This may at first be confusing because the analyst is looking to sort the column, but KNIME places this Sorter in the Row category because each row is being sorted as part of the column. This may sound as if it is not intuitive, but the analyst must think of this throughout using KNIME. It is not wrong, just a different perspective.

Once the Sorter node is dragged, dropped, and connected, double-click the node to reveal this screen. Again, any time the analyst double-clicks the node, it is telling the application to configure that node. This is the easiest way to move into the configuration mode. Also, remember at this point that we are working on Node 5, the right-hand side of the duplicate nodes.



This configuration screen is straightforward. First, set "Sort by:" with "Sum(TOR_LENGTH)" from the down arrow and ensure that "Descending" is selected from the right side. Do not worry about the area marked "Add columns," since this would add columns that would be placed in the sorting hierarchy, much like choosing additional sorting columns in other applications. Do not worry about the rest of the choices; click Apply and OK or just OK to configure the node. At this point, the node will be "yellow" and will need to be executed to activate the flow. Click on the single green arrow while this node is selected, or the double green arrow to execute all the nodes. The resulting table is recovered by right-clicking the node and choosing Sorted Table. That screen is displayed as follows:

Sorted Table - 0:5 - Sorter

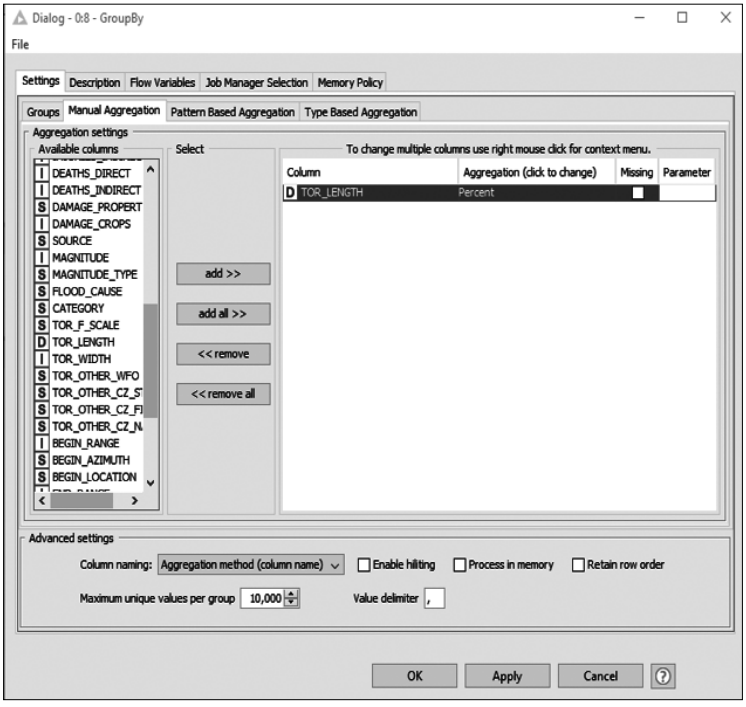
File Hilite Navigation View

Table "default" - Rows: 12 | Spec - Columns: 2 | Propert

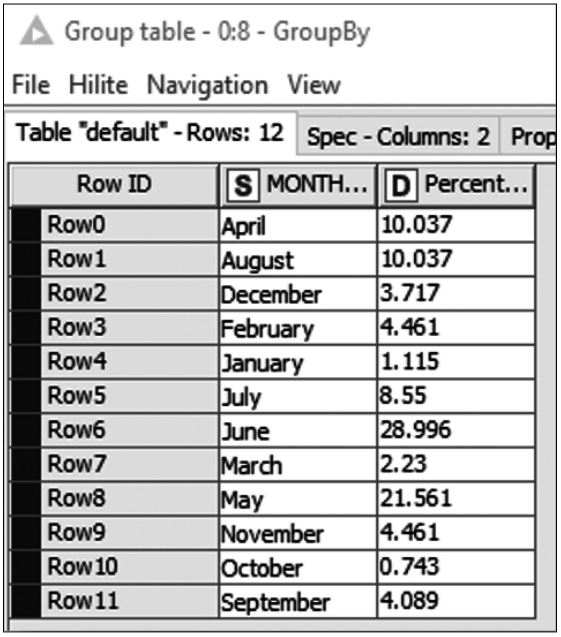
Row ID	S MONTH...	D Sum(T...
Row6	June	413
Row3	February	162.3
Row8	May	125
Row1	August	119.9
Row0	April	117.7
Row11	September	66.5
Row9	November	61.8
Row7	March	43.7
Row5	July	32.9
Row2	December	28
Row4	January	22.5
Row10	October	2

If the analyst refers back to other tools at this point, they will see that the numbers are the same. This is, as stated before, a good way to verify the analysis process. Now that the table is summed and sorted, the next step is to calculate the probability so that the table can be used to track both the tornado lengths and the probability of those lengths (percentage actually, but essentially the same) compared to the total lengths throughout the months.

Please double-click on Node 8 to get to the following screen, which is the configuration for the percent (or probability) of the tornado lengths.

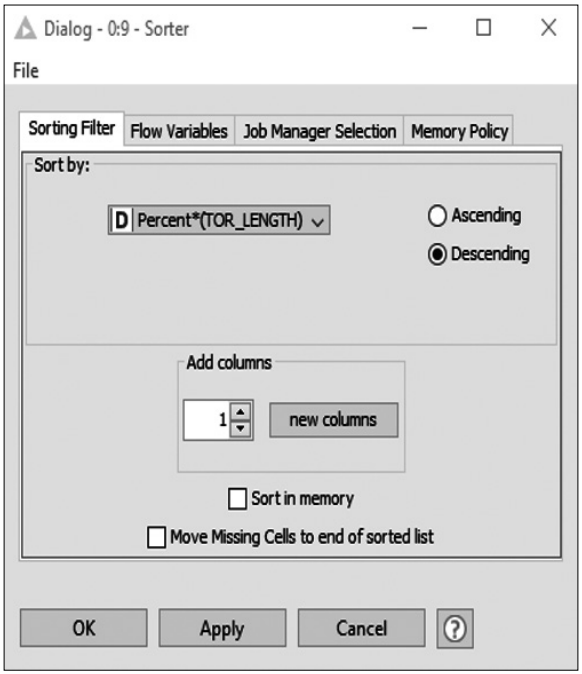


The analyst will notice that the "Aggregation" is now "Percent" rather than "Sum" so that the column will now have a percent. The "Groups" tab will still have the MONTH_NAME column chosen, since this is the first column that the analyst wants to see in the table. Once Node 8 is configured and executed, right-click and choose "Group Table" to see the table that resulted from that flow.



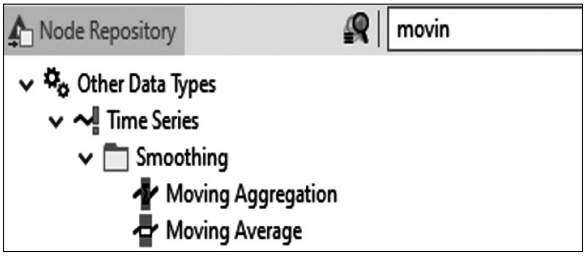
Row ID	S MONTH...	D Percent...
Row0	April	10.037
Row1	August	10.037
Row2	December	3.717
Row3	February	4.461
Row4	January	1.115
Row5	July	8.55
Row6	June	28.996
Row7	March	2.23
Row8	May	21.561
Row9	November	4.461
Row10	October	0.743
Row11	September	4.089

As the analyst can see, the percentages are not sorted and will need to be sorted so that the table will match the raw numbers that were sorted earlier. This will necessitate another "Sorter" node (Node 9 in this case), and the Sorter node will be about the same as Node 5, except instead of the sum, the analyst will be sorting by descending order the percentage of the tornado lengths.



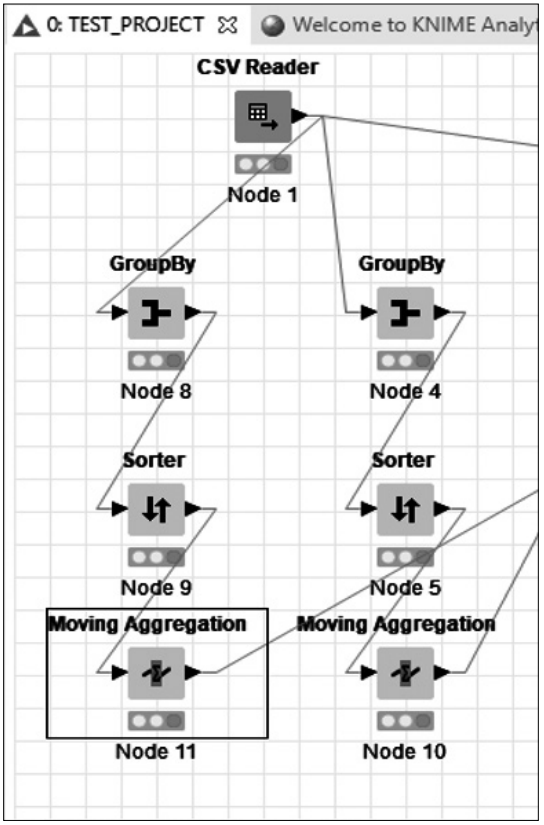
As the analyst can see, the "Sorting Filter" tab looks exactly the same in this node as in Node 5, so there is no real difference. As stated before, please explore the other tabs and the other options within those tabs, since there are undoubtedly some that will enhance the KNIME experience with data analysis.

Once the calculating and sorting is finished, it is time for the tornado lengths to be accumulated. This is done through a node called "Moving Aggregation," which is located at the left-hand side sub-menu as shown here:

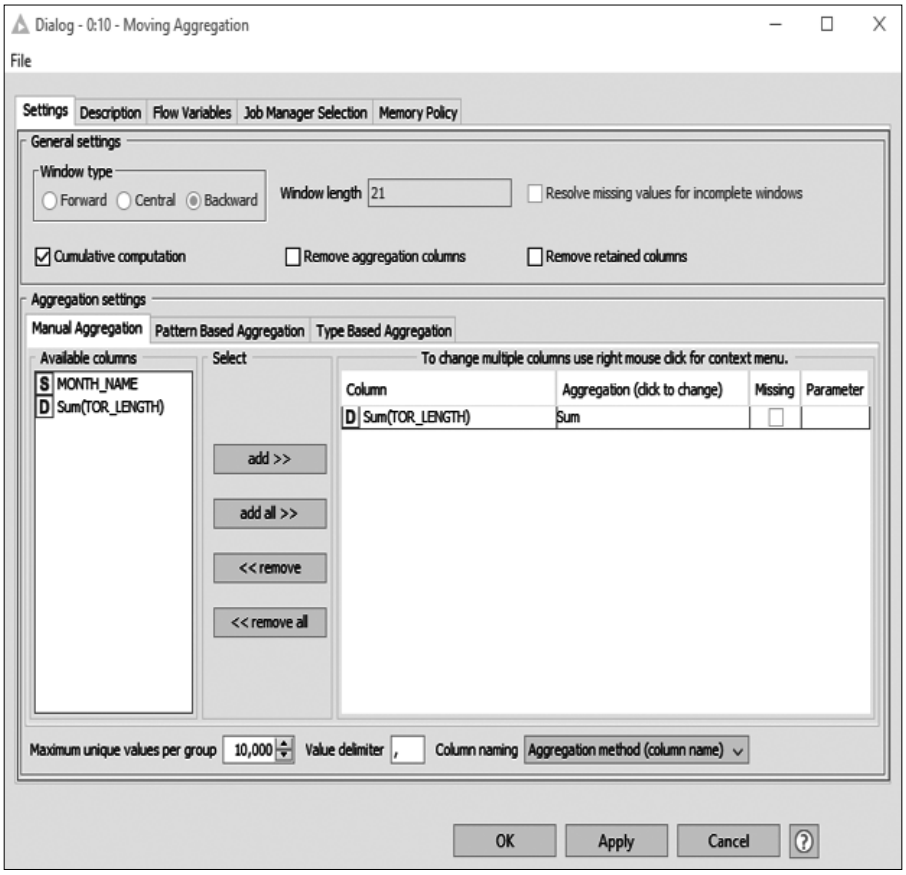


As the analyst can see, the name "movin" is in the search box, which means that the full node title is not necessary to get to the node that the analyst needs.

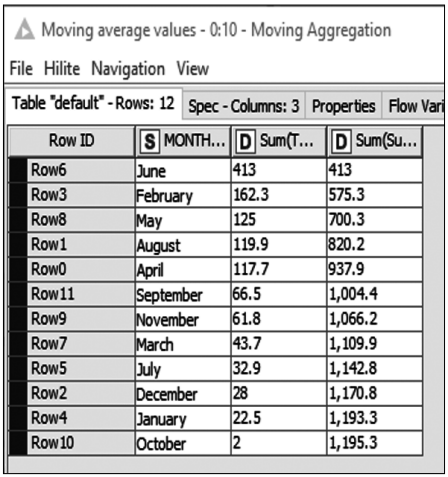
Once the Moving Aggregation node is dragged, dropped, and connected, the flow will look similar to the one that follows. Please note that the analyst can move nodes around the screen and place them in any configuration. This helps reading the workflow diagram for the analyst and those that may use the flow after the analyst is finished as well as the flow transitions to others as a template for further enhancement.



The first node that will be discussed is Node 10 (the bottom right-hand side of the previous screen). This node is going to provide a cumulative calculation of the sorted tornado length sums above it (Node 5). Double-clicking Node 10 will produce this screen. Note that there are several tabs to this screen.



Please take a close look at the screen. There are several configuration choices that are important in this step. First, ensure that the "Cumulative computation" box is selected, which will gray out the "Window length" choice. Also ensure that the correct column is added using the "add >>" button; in this case "Sum(TOR_LENGTH)" is the column the analyst wants to accumulate. Once the OK button is clicked, then the analyst must remember to execute the node in order to activate the flow process and complete the calculations. The table produced from this entire process is the following screen. Please notice the additional column produced by this process.

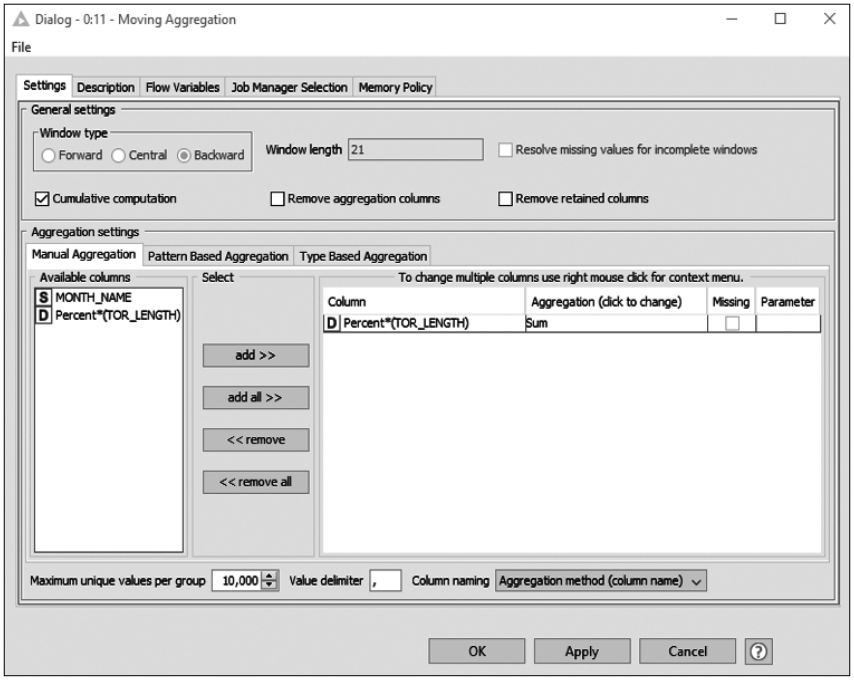

Moving average values - 0:10 - Moving Aggregation

File Hilite Navigation View

Table "default" - Rows: 12 | Spec - Columns: 3 | Properties | Flow Vari

Row ID	S MONTH...	D Sum(T...	D Sum(Su...
Row6	June	413	413
Row3	February	162.3	575.3
Row8	May	125	700.3
Row1	August	119.9	820.2
Row0	April	117.7	937.9
Row11	September	66.5	1,004.4
Row9	November	61.8	1,066.2
Row7	March	43.7	1,109.9
Row5	July	32.9	1,142.8
Row2	December	28	1,170.8
Row4	January	22.5	1,193.3
Row10	October	2	1,195.3

Now that the number accumulation is completed, the next step is to accumulate the percentages, which is done in the same way (with the same type nodes) as the preceding section. This time, focus on Node 11, which is illustrated as follows. The only difference is that the analyst is now concentrating on the percent column instead of the sum column, but the calculation required would still be "sum" as depicted here.


Dialog - 0:11 - Moving Aggregation

File

Settings | Description | Flow Variables | Job Manager Selection | Memory Policy

General settings

Window type: Forward / Central / Backward

Window length: 21

Resolve missing values for incomplete windows

Cumulative computation

Remove aggregation columns

Remove retained columns

Aggregation settings

Manual Aggregation | Pattern Based Aggregation | Type Based Aggregation

Available columns: S MONTH_NAME, D Percent*(TOR_LENGTH)

Select

add >> | add all >> | << remove | << remove all

To change multiple columns use right mouse click for context menu.

Column	Aggregation (click to change)	Missing	Parameter
D Percent*(TOR_LENGTH)	Sum		

Maximum unique values per group: 10,000 | Value delimiter: , | Column naming: Aggregation method (column name)

OK | Apply | Cancel

Once the node is configured and the OK button is clicked, please remember to click on Execute to activate the flow. The completion of this step will produce the following table, which is the accumulated percentages (the total should be 100).

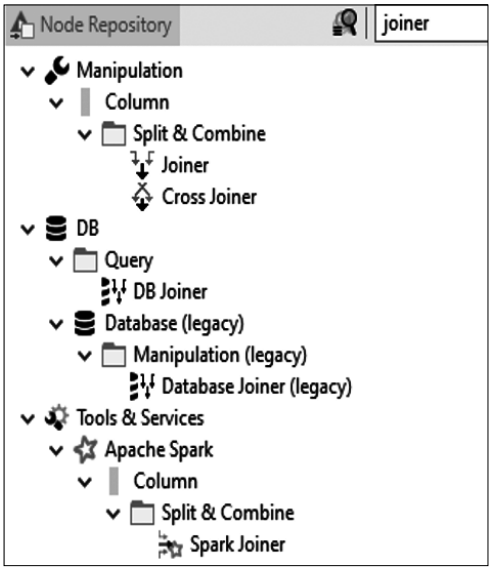
Moving average values - 0:11 - Moving Aggregation

File Hilite Navigation View

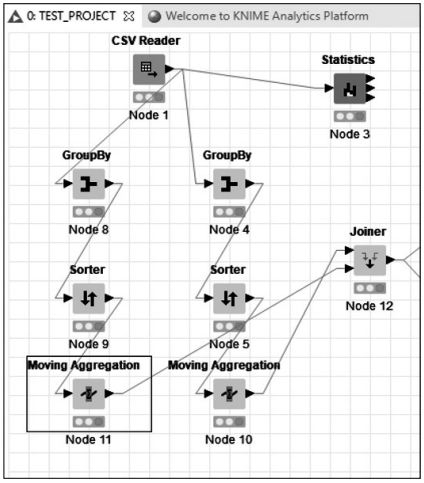
Table "default" - Rows: 12 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	S MONTH...	D Percent...	D Sum(Pe...
Row6	June	28.996	28.996
Row8	May	21.561	50.558
Row0	April	10.037	60.595
Row1	August	10.037	70.632
Row5	July	8.55	79.182
Row3	February	4.461	83.643
Row9	November	4.461	88.104
Row11	September	4.089	92.193
Row2	December	3.717	95.911
Row7	March	2.23	98.141
Row4	January	1.115	99.257
Row10	October	0.743	100

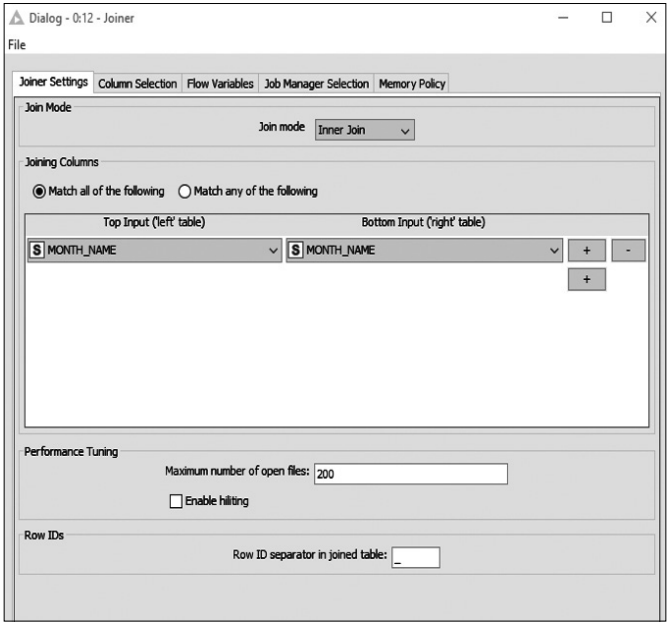
The resulting tables from both the numbers and percentages must now be joined into one table. There is a node for just about everything in KNIME, and joining is no different. Find the "Joiner" node in the sub-menu on the bottom left of the KNIME screen by typing the name in the search box.



In this case, the Joiner node is located as part of the "Column" category. The purpose of the Joiner is to join columns, which is exactly what the analyst wants in this case. Once the analyst drags and connects the Joiner node, the resulting process will look similar to this screen.



Please ensure that both nodes that are Moving Aggregation are connected to the Joiner node. There are two input connections to the Joiner node to make the joining of two columns possible. Double-clicking the Joiner node will reveal the configuration screen.



The first tab is "Joiner Settings," which helps the analyst decide what column to "key" the joining. In this case the "MONTH_NAME" column exists (and is the same) between the two sets of columns used, so that is the keying column. "Join Mode" has several options, but the default option is the one used this time. Remember that "Top Input" and "Bottom Input" should both have the same column to key if that is the purpose of the node. In order to keep the columns with the same calculations and in the correct sorted order, there is nothing else that the analyst needs to do at this point.

The resulting table is found by right-clicking the Joiner node and choosing the option at the bottom of that sub-menu called "Joined Table." The table appears as in the following screen.

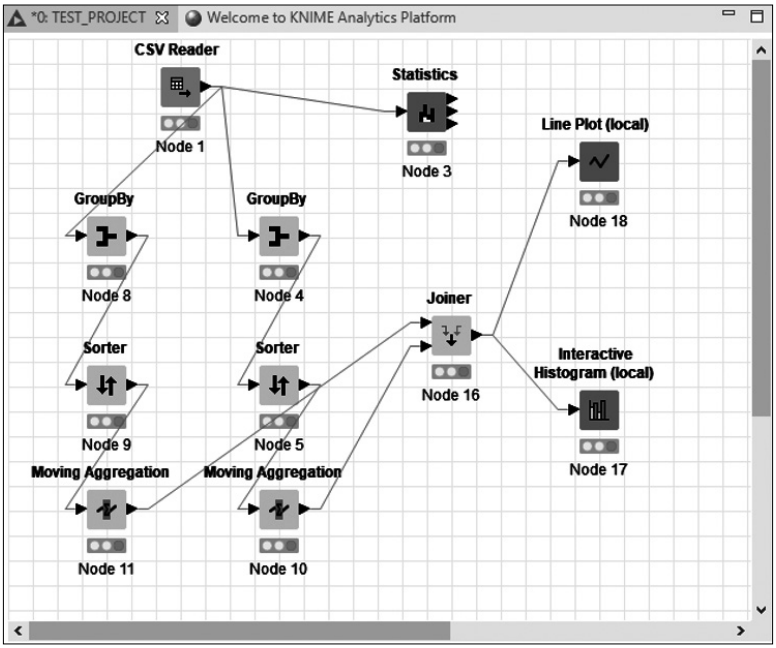
Joined table - 0:16 - Joiner

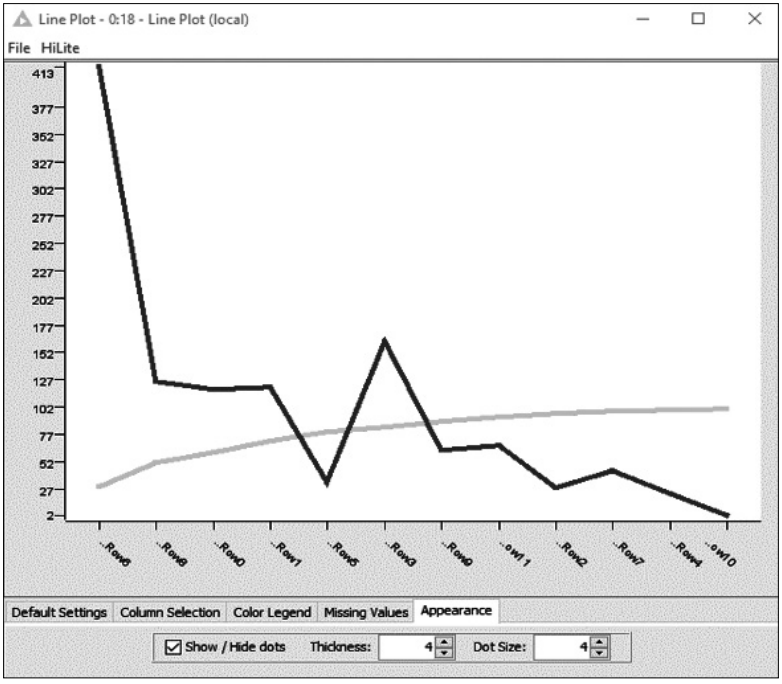
File Hilite Navigation View

Table "default" - Rows: 12 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	S MONTH...	D Percent...	D Sum(Pe...	D Sum(T...	D Sum(Su...
Row6_Row6	June	28.996	28.996	413	413
Row8_Row8	May	21.561	50.558	125	700.3
Row0_Row0	April	10.037	60.595	117.7	937.9
Row1_Row1	August	10.037	70.632	119.9	820.2
Row5_Row5	July	8.55	79.182	32.9	1,142.8
Row3_Row3	February	4.461	83.643	162.3	575.3
Row9_Row9	November	4.461	88.104	61.8	1,066.2
Row11_Row11	September	4.089	92.193	66.5	1,004.4
Row2_Row2	December	3.717	95.911	28	1,170.8
Row7_Row7	March	2.23	98.141	43.7	1,109.9
Row4_Row4	January	1.115	99.257	22.5	1,193.3
Row10_Row10	October	0.743	100	2	1,195.3

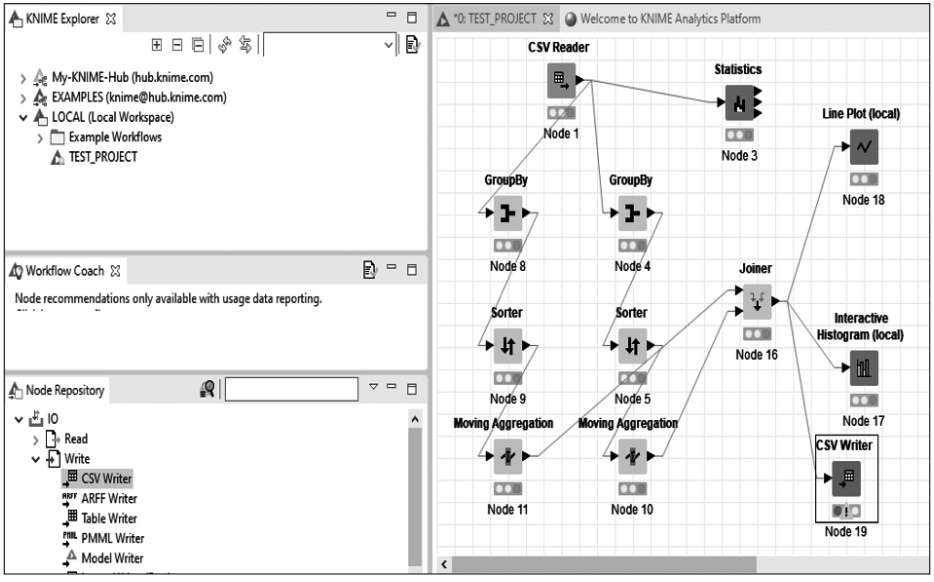
This is the table that would be completed with the charts. Unfortunately, KNIME does not have a cumulative probability chart, and the amount of programming necessary to produce this chart is beyond the scope of this book. KNIME does have the nodes necessary for line charts or histograms, which are located in the sub-menus. An example of the node is depicted as follows, with the result following that screen. This is about as close as you can get to a cumulative probability chart with the available KNIME nodes.


*0: TEST_PROJECT
Welcome to KNIME Analytics Platform
CSV Reader
Node 1
Statistics
Node 3
Line Plot (local)
Node 18
GroupBy
Node 8
GroupBy
Node 4
Joiner
Node 16
Sorter
Node 9
Sorter
Node 5
Interactive Histogram (local)
Node 17
Moving Aggregation
Node 11
Moving Aggregation
Node 10

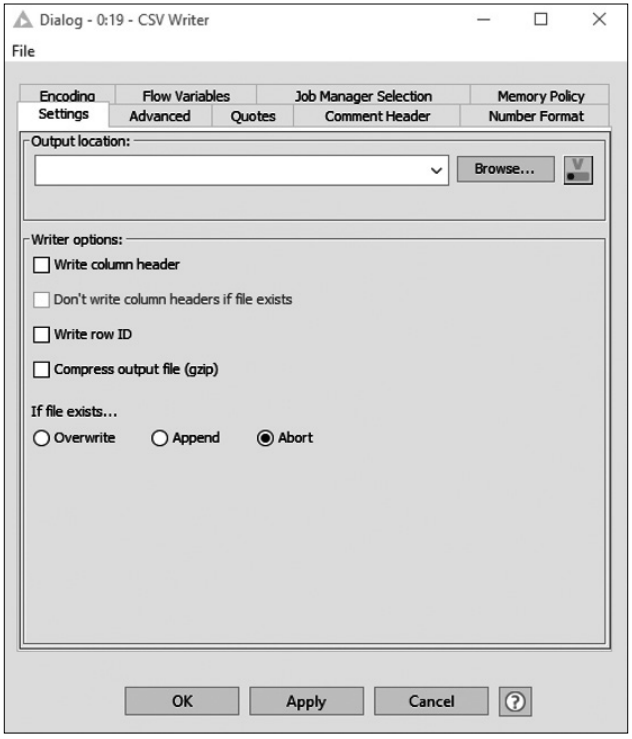

Line Plot - 0:18 - Line Plot (local)
File HiLite
413
377
352
327
302
277
252
227
202
177
152
127
102
77
52
27
2
Default Settings
Column Selection
Color Legend
Missing Values
Appearance
Show / Hide dots
Thickness:
4
Dot Size:
4

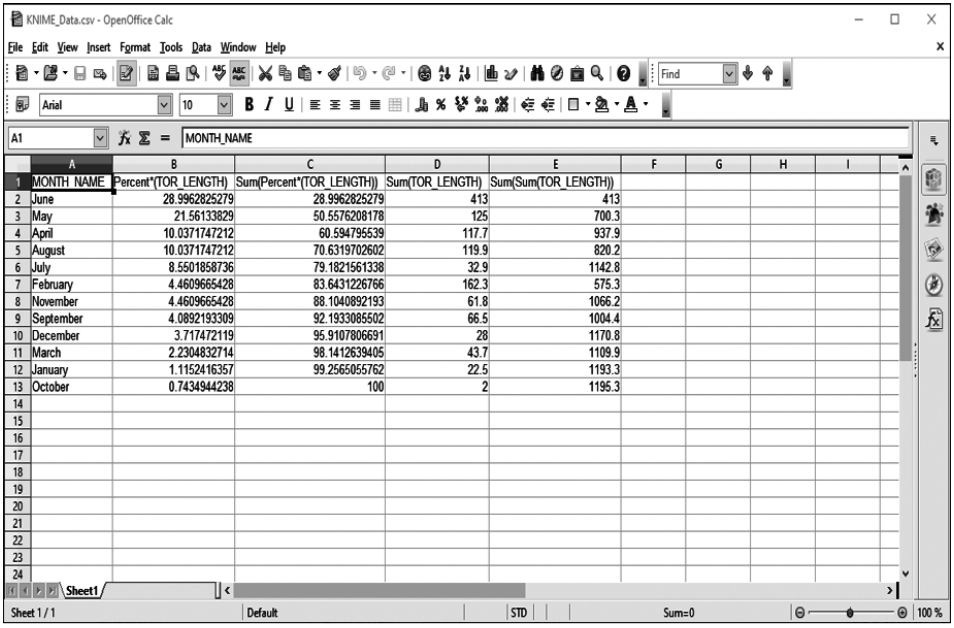
Fortunately, KNIME does have the capability to export (write) files to many of the applications that perform data analytics, some of which are in this book. The one that will suffice for this section is the CSV type file.

In order to export or write the file, there is a node called CSV Writer which, once dragged, placed, and connected, provides the ability to export the finished table to a CSV file. The workflow diagram follows along with the location of the CSV writer.



Now double-click on the CSV Writer node to reveal the configuration screen. In this screen, the main entry is the location and file name of the export file. In this case, the analyst can make this location anywhere from the local computer to a network server. Also, ensure that the "Write column header" box is checked. Otherwise, the columns will have no headers. Once that is done, the file can be opened by a tool that opens CSV files. In this case, this would be OpenOffice. Once that is done, the section on OpenOffice cumulative probability charts can be implemented.





	A	B	C	D	E
1	MONTH_NAME	Percent*(TOR_LENGTH)	Sum(Percent*(TOR_LENGTH))	Sum(TOR_LENGTH)	Sum(Sum(TOR_LENGTH))
2	June	28.9962825279	28.9962825279	413	413
3	May	21.56133829	50.5576208178	125	700.3
4	April	10.0371747212	60.594795539	117.7	937.9
5	August	10.0371747212	70.6319702602	119.9	820.2
6	July	8.5501858736	79.1821561338	32.9	1142.8
7	February	4.4609665428	83.6431226766	162.3	575.3
8	November	4.4609665428	88.1040892193	61.8	1066.2
9	September	4.0892193309	92.1933085502	66.5	1004.4
10	December	3.717472119	95.9107806691	28	1170.8
11	March	2.2304832714	98.1412639405	43.7	1109.9
12	January	1.1152416357	99.2565055762	22.5	1193.3
13	October	0.7434944238	100	2	1195.3
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					

Remember, if the chart is not available in the tool, export the file and use the available tool to make the chart. If a feature is not available in one tool, it will be available in the other (and probably easier too).

3.3 T-TEST (PARAMETRIC)

The t-test is something that compares data from a perspective of means. The test is very valuable when comparing items such as test grades, inventory, or even if the amount of a product placed in a bag meets the standard for that quantity (such as candy or nails). The background of the t-test is interesting, but that is best left to the statistics instructor, since these types of stories help to build a better understanding of why the concept was initiated. For those that are interested, it is best to use a search engine and type in "History of Student T-Test." There will be more than enough results to get a very good understanding of the concept. Suffice it to say that the t-test is used when the analyst has a sample of data and the population standard deviation is not known. This is true in many data analytics cases. Finding a population standard deviation is not always possible.

Some might question why the word "Parametric" was placed in parentheses next to the title of this section. The word parametric when associated with statistics means that the method is related to an algorithm as part of a table or normal distribution. There are non-parametric tests such as the Wilcoxon class of testing, but that is beyond the scope of this book. Suffice it to say that parametric tests are those that most analysts have used in the past, whether they be chi-square, t-tests, or Z-tests. Please explore this to become more familiar with the lexicon of data analytics.

3.3.1 Excel

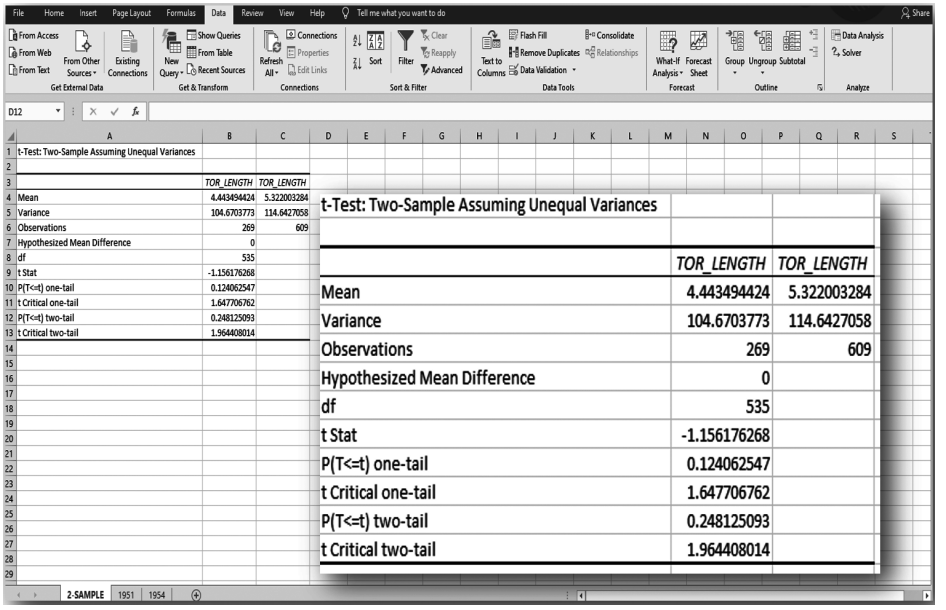
Excel, through the Analysis ToolPak, provides a perfect platform for the t-test. The process for performing this statistical test is relatively straightforward.

First, the analyst opens the Excel application and the dataset, in this case the same one used for other concepts and tools. The resulting screen is as follows, but understand that this worksheet contains two pieces of data. The first is from 1951 and the second is from 1954. What we are trying to see is if the average tornado length was greater in 1954 than in 1951, even though there was more tornado activity recorded in 1954. If this were a hypothesis, the null hypothesis would be that the average tornado length

of 1951 = average tornado length of 1954; while the alternative hypothesis would be that the average tornado length of 1951 was less than (<) the average tornado length of 1954.

One assumption that will be made during this section is that the variances between these two datasets are unequal. Using the F-Test that is provided in the Analysis ToolPak will prove this, which will be discussed in a subsequent section.

The first step to do is to combine the tornado datasets from 1954 and from 1951. These datasets are available from the site mentioned in the section on where to get data. Ensure that both of these datasets are in the same worksheet. The next thing that the analyst must do is to open the Analysis ToolPak, which is in the Data tab. After activating the ToolPak, choose "t-Test: Two-Sample Assuming Unequal Variances," and at this point fill in the two text boxes with the columns of "TOR_LENGTH" from 1951 (first) and then 1954 (the one below the first text box). Ensure that "Labels" is checked to account for column headers, and pick the location as being a new worksheet. Once this is completed and you click OK, you will receive the following result.



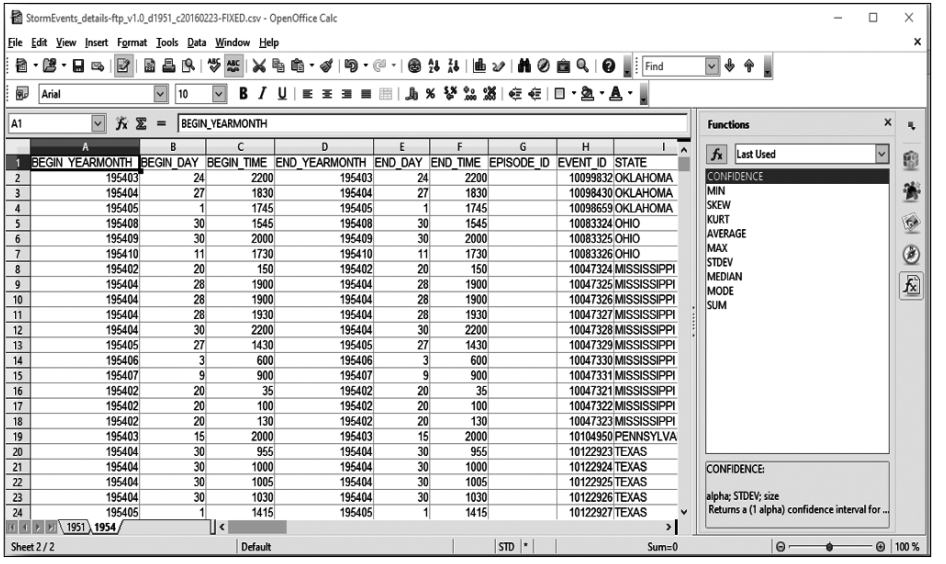
t-Test: Two-Sample Assuming Unequal Variances		
	TOR_LENGTH	*TOR_LENGTH*
Mean	4.443494424	5.322003284
Variance	104.6703773	114.6427058
Observations	269	609
Hypothesized Mean Difference	0	
df	535	
t Stat	-1.156176268	
P(T<=t) one-tail	0.124062547	
t Critical one-tail	1.647706762	
P(T<=t) two-tail	0.248125093	
t Critical two-tail	1.964408014	

At this point, it is important to understand what this result describes. Basically, the test was completed at a 95% confidence level, which means that if the "P(T<=t) one-tail" is .05 or less (<.05), then the null hypothesis is rejected and the two tornado length averages are the same, while if the "P(T<=t) one-tail" is greater than .05 (>.05), the null hypothesis is not rejected (in some statistical circles, the word is *accepted*, but there are many arguments about this word or *not rejected*). This means that, according to the t-test, the analyst has shown that the average tornado length is not different between 1951 and 1954 at the 95% confidence level. This again assumes unequal variances.

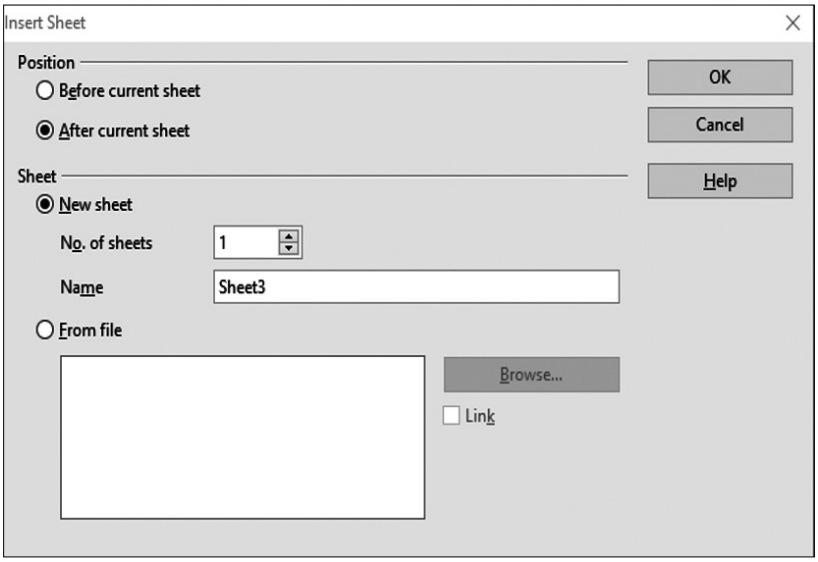
3.3.2 OpenOffice

While Excel provides a nice app for performing t-tests, OpenOffice relies on formulas. This next section will perform the same testing on the same datasets and get the same results.

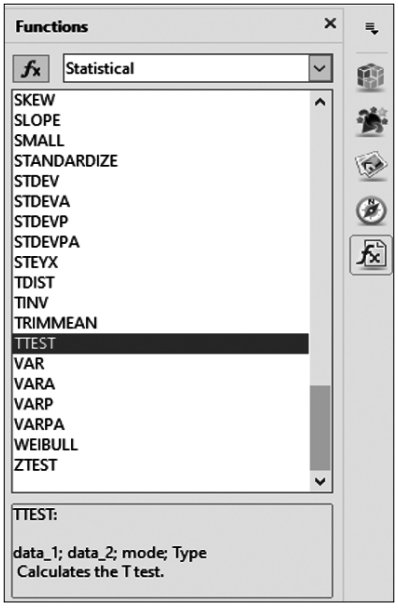
The first step, as with all data analysis, is to import the appropriate datasets. This is done with the import function of OpenOffice, which was covered in the section on importing data. The method for combining the worksheets into one workbook is the same as with Excel, resulting in the following screen. There is an alternative way of inserting worksheets, as shown in the screen following the OpenOffice workbook. In this case, the analyst is inserting a sheet from a file, which can also be done in Excel.



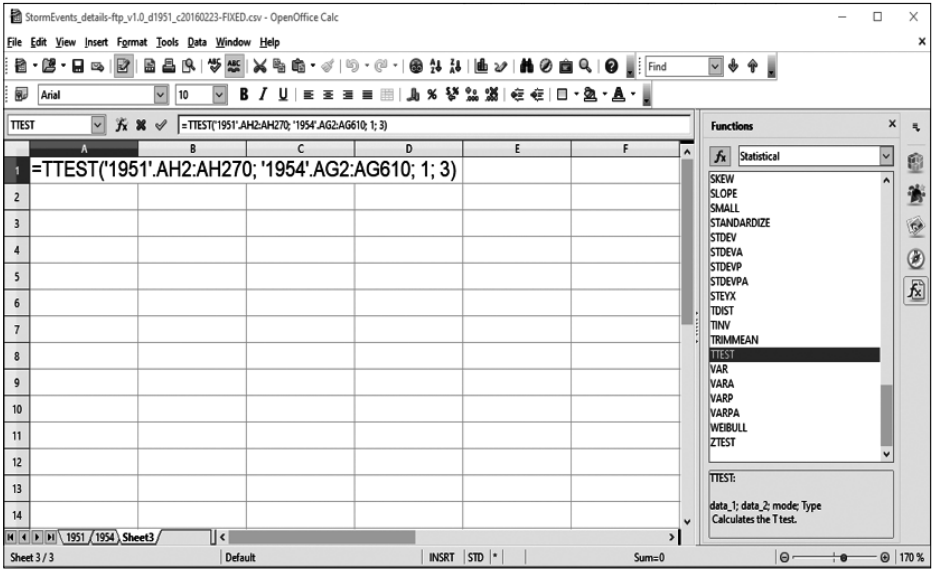
	A	B	C	D	E	F	G	H	I
1	BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EPISODE_ID	EVENT_ID	STATE
2	195403	24	2200	195403	24	2200		10099832	OKLAHOMA
3	195404	27	1830	195404	27	1830		10098430	OKLAHOMA
4	195405	1	1745	195405	1	1745		10098659	OKLAHOMA
5	195408	30	1545	195408	30	1545		10083324	OHIO
6	195409	30	2000	195409	30	2000		10083325	OHIO
7	195410	11	1730	195410	11	1730		10083326	OHIO
8	195402	20	150	195402	20	150		10047324	MISSISSIPPI
9	195404	28	1900	195404	28	1900		10047325	MISSISSIPPI
10	195404	28	1900	195404	28	1900		10047326	MISSISSIPPI
11	195404	28	1930	195404	28	1930		10047327	MISSISSIPPI
12	195404	30	2200	195404	30	2200		10047328	MISSISSIPPI
13	195405	27	1430	195405	27	1430		10047329	MISSISSIPPI
14	195406	3	600	195406	3	600		10047330	MISSISSIPPI
15	195407	9	900	195407	9	900		10047331	MISSISSIPPI
16	195402	20	35	195402	20	35		10047321	MISSISSIPPI
17	195402	20	100	195402	20	100		10047322	MISSISSIPPI
18	195402	20	130	195402	20	130		10047323	MISSISSIPPI
19	195403	15	2000	195403	15	2000		10104950	PENNSYLVA
20	195404	30	955	195404	30	955		10122923	TEXAS
21	195404	30	1000	195404	30	1000		10122924	TEXAS
22	195404	30	1005	195404	30	1005		10122925	TEXAS
23	195404	30	1030	195404	30	1030		10122926	TEXAS
24	195405	1	1415	195405	1	1415		10122927	TEXAS



At this point, there will be a need to use some formulas in order to get the t-test result. On the far right of the screen are five icons, the bottom of which is the formula wizard. Please activate that wizard, which will show another screen pane called "Functions."



Before TTEST is selected, it is recommended that a new worksheet is inserted and the cell A1 is selected. That will provide a place for the result of the TTEST to reside. The following screen shows the finished formula with a legend below it to show what some of the parameters within the formula provide.



Looking between the semicolons, which is the same as the comma for separations between parameters in Excel, the first two parameters are the cell contents of the 1951 and 1954 tornado lengths. The last two denote the mode and type of t-test. The "1" means that this is a one-tailed test. Yes, that means that "2" would mean a two-tailed test. The last parameter is a "3," which means that this is a two-sample test with unequal variances. The number "1" means a paired sample test, and the "2" means that it is a two-sample test with equal variances. A great site to get the lowdown on t-tests for OpenOffice is located here: *https://wiki.openoffice.org/wiki/Documentation/How_Tos/Calc:_TTEST_function*.

Once the formula is activated with an ENTER press, the following number will appear: - 0.1240626151. If the analyst would look back at the Excel output, this number is the p-value for the one-tailed test. It means the same here as it did in the preceding analysis. It appears the tornado lengths are considered equal under this statistical test.

3.3.3 R/RStudio/Rattle

The R tool, and Rattle specifically, are challenging when conducting the t-test, but with a little patience and a little programming, everything will be fine.

To start with Rattle, first open the package by the method discussed in the section referencing importing data using Rattle, but there will be a slight twist to this import. First, transform the data using R so that you have three columns—MONTH_NAME, TOR_LENGTH_1951, and TOR_LENGTH_1954. In this way, the t-test will be set like the previous sections in OpenOffice and Excel.

To establish three columns, first import the 1951 dataset, which should already be completed, and add the 1954 dataset, which is done the same way as importing the 1951 dataset. After that is completed, then the analyst will want to isolate and make a table with the three columns.

The screen showing both datasets in the RStudio source pane is shown as follows. Remember that the analyst wants to shorten these file names with letters and numbers. It will be much easier in the programming if those are shortened.

```
tor1951<-StormEvents_details_ftp_v1_0_d1951_c20160223_
                          FIXED

tor1954<-StormEvents_details_ftp_v1_0_d1954_c20160223
```

Once that is completed, then the analyst can make a t-test from the following commands, again assuming unequal variances and using a 95% confidence level. The following are the commands, followed by the resulting t-test. The results are exactly the same as in the previous sections, with an exception of the negative confidence interval, which will be explained in a later section.

```
t.test(tor1951$TOR_LENGTH,tor1954$TOR_LENGTH,alternative=

   "less",paired=FALSE,var.equal=FALSE,conf.level=0.95)

                Welch Two Sample t-test

data: tor1951$TOR_LENGTH and tor1954$TOR_LENGTH
t = -1.1562, df = 534.86, p-value = 0.1241
alternative hypothesis: true difference in means is less
than 0
95 percent confidence interval:
   -Inf 0.3734851
sample estimates:
mean of x mean of y
4.443494 5.322003
```

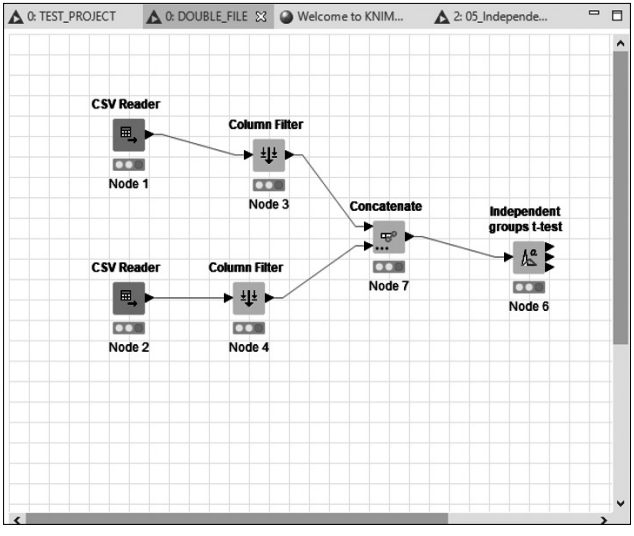
The previous result matches the p-value (.1241) from previous sections. The R application even writes out the alternative hypothesis for you, which is convenient. So far, all tools agree with each other, which proves valuable when convincing someone that the results have been verified.

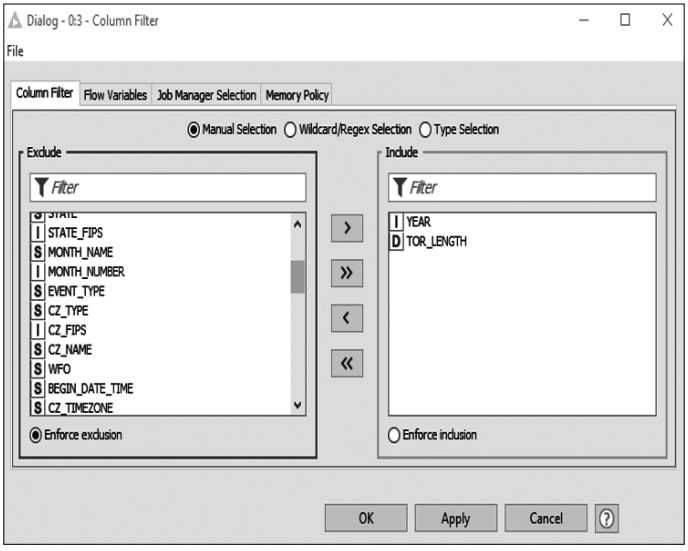
3.3.4 KNIME

The main advantage that KNIME has over other tools is that, if a node exists that performs the test function, then setting that node in the process flow allows for the transformation and testing of that dataset. There is a node for the t-test that exists in KNIME. However, the analyst is faced with combining two datasets so that the tornado lengths can be compared with the same accuracy as in previous sections.

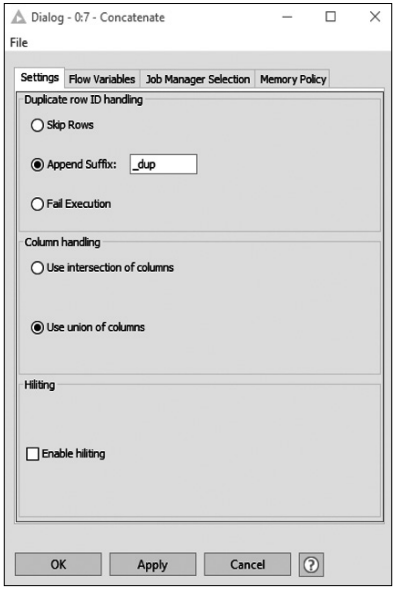
In KNIME, if there are two datasets used, then simply add another CSV Reader node at the beginning of the flow and place the second file into that node for further use. Once both files are imported via nodes, the challenge then comes to ensure the t-test is properly found, dragged, placed, and connected. The t-test node is located in the sub-menu on the left (use the search box and type "t-test" in that box). The placed and connected node for the t-test is shown on the following screen along with the necessary nodes to perform the t-test.

The first step is to isolate the columns that will be used in the test, in this case the 1951 and 1954 tornado lengths. This isolation will be done with the Column Filter nodes, which will be placed one against each CSV as shown. The screens for this are shown after the workflow.



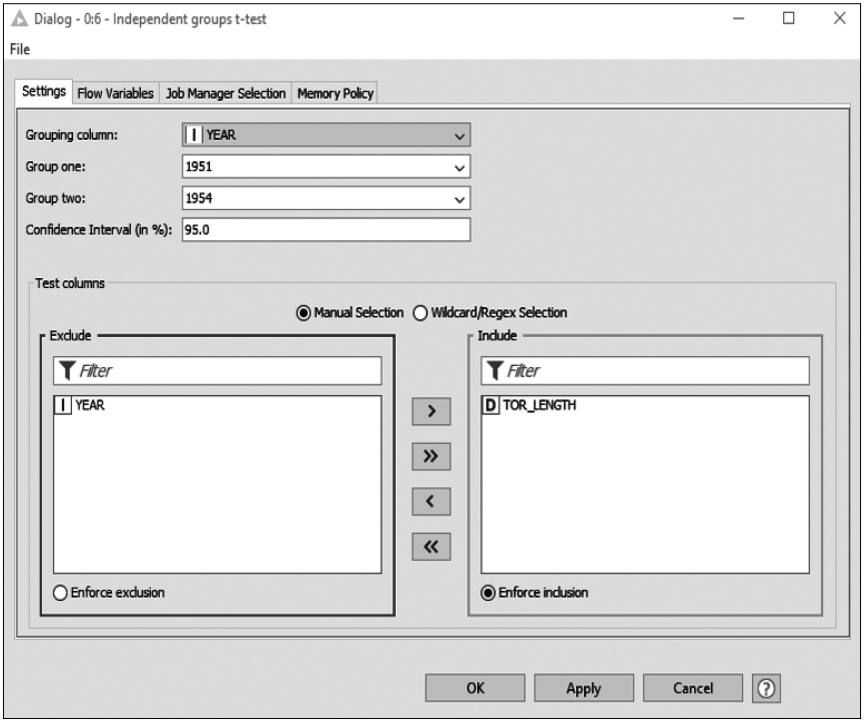


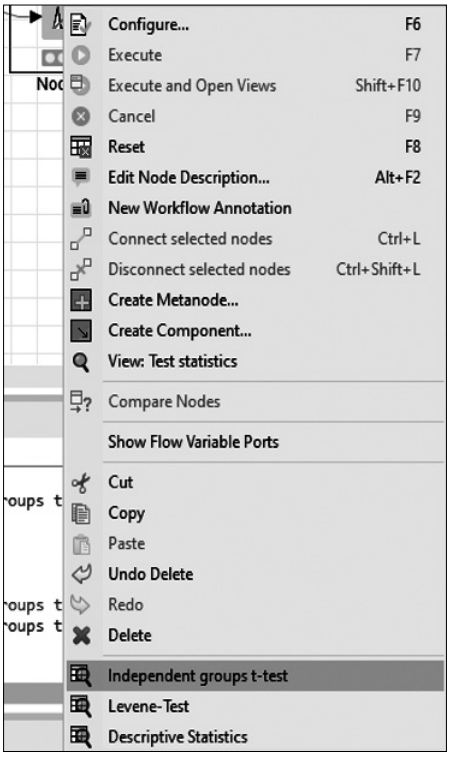
In both nodes for the Column Filter, the screen will look the same, but it is important to note that these will have two separate years, 1951 and 1954. That is why choosing the YEAR column will help differentiate the rows once the two columns are joined. The joining is done using the CONCATENATE node, which is shown as follows. In this case, the analyst will want a union of the rows, since that will add the rows from 1954 to the rows of 1951. This is vital, since the next step will use the different years.



The end product of this node is illustrated as follows. Note that there is now one column—TOR_LENGTH—with years 1951 until that column is exhausted, and then 1954. This will be an important distinction when the analyst will add the t-test node.

At this point, it is time to add the node that will actually perform the statistical test—the t-test node. In this case, the name of the node is the *Independent groups t-test*, which the analyst can find by typing that into the search box. The configuration screen for this node (once it is connected) is as follows. Note the different settings in the configuration box, since this is necessary to get the most accurate response. Also, remember to place the 1951 group in the first box, and the 1954 group in the second box. At this point, the hypothesis is the same as in the previous sections, that the null consists of both 1951 and 1954 having the same average tornado lengths, with the alternative being that 1951 has less of an average tornado length than 1954. The results for this node are viewed by right-clicking on the node and selecting the first option from the bottom of that child window as shown.





Once the analyst chooses "Independent groups t-test" from the menu, the following screen will appear. Please compare these results with the results from the other sections. There will be some slight differences between this one and the three others. However, the results are the same and the null hypothesis is not rejected, meaning that the there is no statistical significance to the point that 1951 is less than 1954 in tornado lengths.

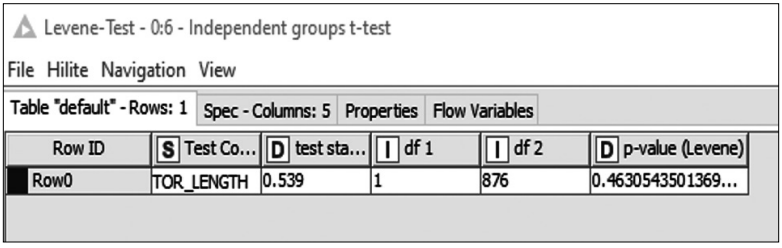
Independent groups t-test - 0:6 - Independent groups t-test

File Hilite Navigation View

Table "default" - Rows: 2 | Spec - Columns: 10 | Properties | Flow Variables

Row ID	S Test Co...	S Variance Assumption	D t	D df	D p-value (2-tailed)
Row0	TOR_LENGTH	Equal variances assumed	-1.136	876	0.2562790899191453
Row1	TOR_LENGTH	Equal variances not assumed	-1.156	534.858	0.2481252302253254

This is just one of three outputs from this node. The other two outputs include an "F-test" (Levene-test) which can give the analyst the probability that the two samples have equal or unequal variances. That screen is illustrated as follows. The result is that the test is greater than the alpha (.05 is smaller than the result), which would mean that the variances are unequal. Given that the overall number of 1954 tornados are triple that of 1951, this would seem logical. However, the test helps confirm the observation. Of course, remember that not all tools are created equal. If the analyst has some doubt as to the veracity of this result, refer to the other tools and conduct similar tests to ensure the accuracy and consistency of your answer.


Levene-Test - 0:6 - Independent groups t-test

File Hilite Navigation View

Table "default" - Rows: 1 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	S Test Co...	D test sta...	I df 1	I df 2	D p-value (Levene)
Row0	TOR_LENGTH	0.539	1	876	0.4630543501369...

CHAPTER 4

MORE STATISTICAL TESTS

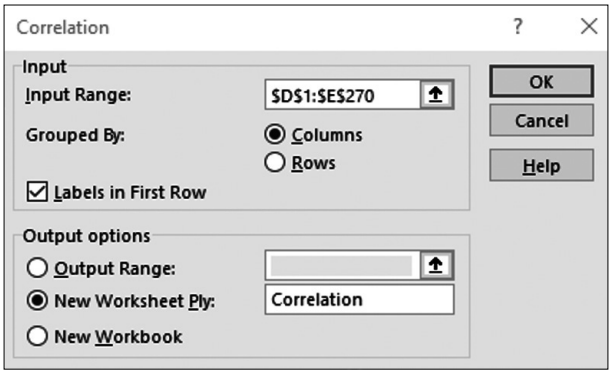
4.1 CORRELATION

Correlation is probably one of the most recognizable statistical concepts, in this author's opinion. Whenever someone hears correlation, they may think that one factor causes the other, but as many statisticians and data analysts will state—correlation does not mean causation. However, correlation is still a powerful concept that can be readily performed with these tools in a somewhat straightforward fashion. In this book the correlation will not be shown on a scatterplot (that may come later), but it will be shown with a matrix showing the variables and how they are associated with one another through a correlation number. This number is between 0 and 1, showing the relationship between these two variables. For instance, if there is a correlation of .90, that is considered to be a very high positive correlation. What that means is that, as one variable increases, the other variable increases. If the correlation is –.90, this is a very high negative correlation, which means as one variable increases, the other decreases. An example of a negative correlation would be the years on a car and its price. The correlation is something that exists in all of the tools and will be addressed one tool at a time, much like the other sections.

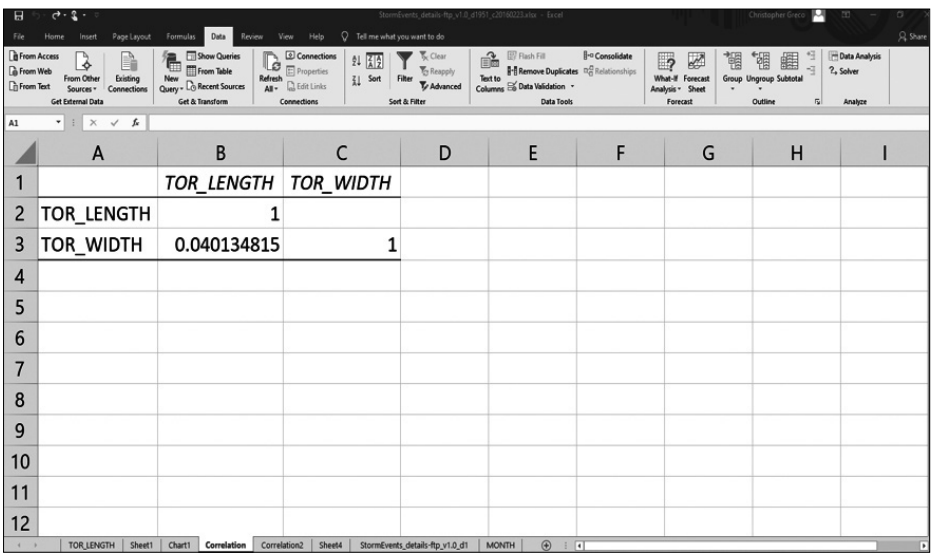
4.1.1 Excel

Correlation is very easy in Excel, especially when using the Analysis ToolPak. In this section, the analyst will make a test of correlation between the tornado length and the tornado width, or how much area the tornado occupied. To do this, the analyst will first use the same file that has been used in the previous sections, mainly the 1951 tornado survey. Once the file is imported (or opened

once saved), the analyst will then open the Analysis ToolPak in the same way as in other sections. The analyst will then pick the Correlation choice from the Analysis ToolPak as shown in the following screen. The screen is already completed showing the two columns that must be chosen, and the fact that they must be side by side. This may require moving some of the columns next to each other, but with Excel this is also somewhat simple.

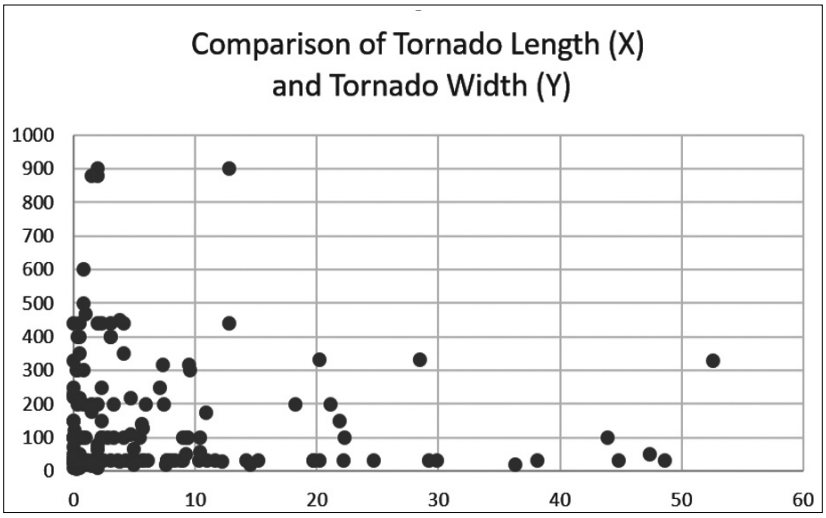


The correlation result for this combination is as follows, but what does it mean? The meaning of this result is that there is less than a .10 relationship between tornado length and tornado width (remember for 1951 only), so no matter the length of the tornado, predicting how wide it will be is almost impossible.



	A	B	C	D	E	F	G	H	I
1		*TOR_LENGTH*	*TOR_WIDTH*						
2	TOR_LENGTH	1							
3	TOR_WIDTH	0.040134815	1						
4									
5									
6									
7									
8									
9									
10									
11									
12									

If the analyst were to plot this on a scatterplot, again using Excel, the chart would look similar to this, which indicates no predictability between tornado length and tornado width.



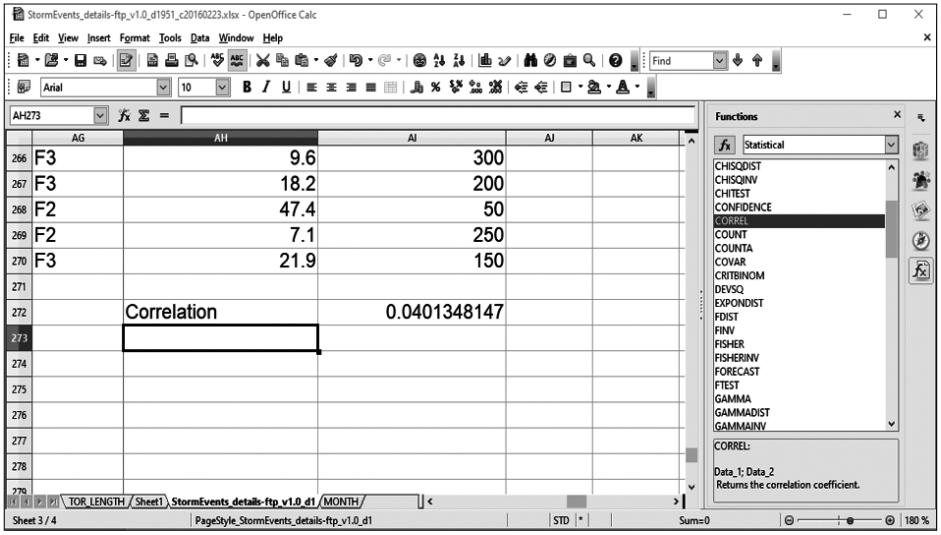
As long as the analyst has columns that are beside one another, a correlation can be conducted to do what is called a multiple correlation. Basically, it is done the same way as the correlation done previously, but just with more correlations in the matrix. There is some information on this featured in Chapter 7 of this book.

4.1.2 OpenOffice

OpenOffice is very similar to Excel in the correlation area, but instead of using the Analysis ToolPak, regular formulas are the conventional method for OpenOffice. In this case, the file will be the same with the same two variables. The big difference is that the first step will be to pick a blank cell to place the formula and then use the following formula for the correlation between tornado length and tornado width:

```
=CORREL(AH2:AH270;AI2:AI270)
```

When the analyst presses the ENTER button, the following screen appears. This shows the correlation between these two variables.

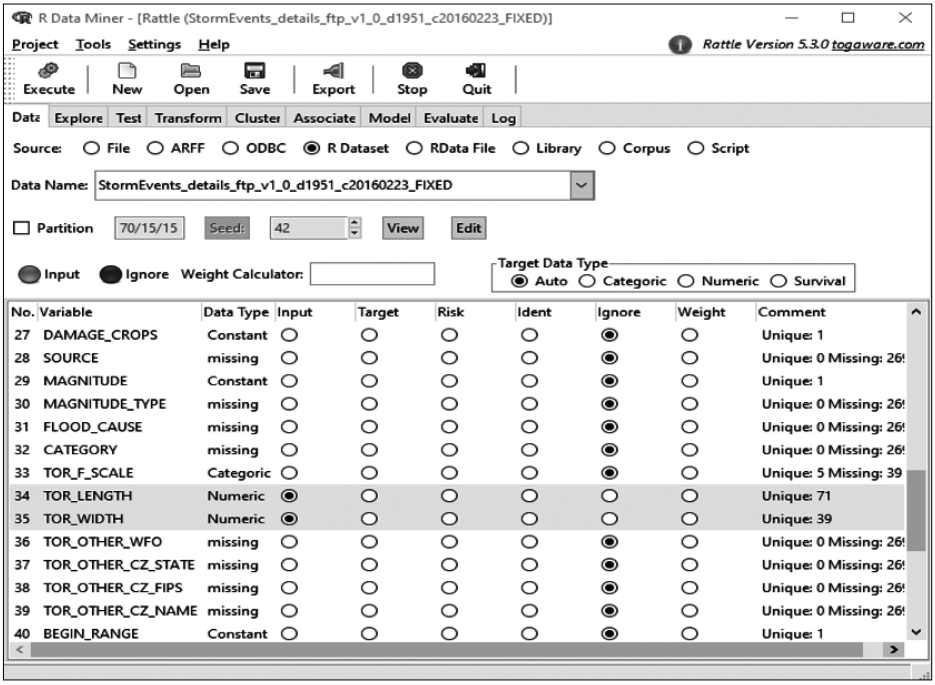


As the analyst can see, the correlation result matches the result from Excel. If a multiple correlation is necessary, then that will be covered in the next section on regression, since there is functionality within OpenOffice to do multiple regression and therefore multiple correlation.

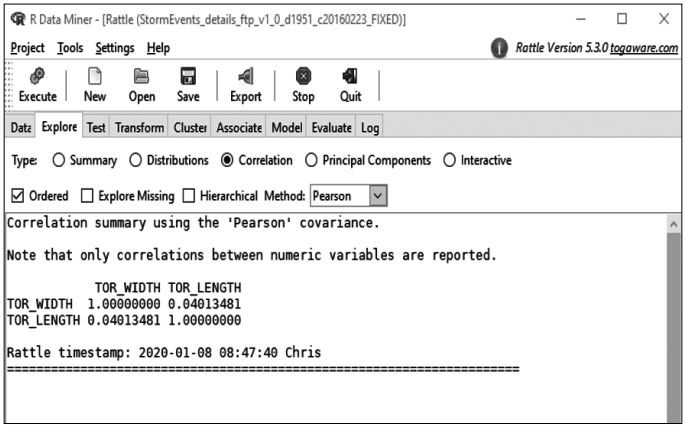
4.1.3 R/RStudio/Rattle

The R application is very versatile as it applies to conventional testing, and correlation is no exception. The process for performing the correlation testing is more intuitive than with many other tools. In this section, Rattle will be used to perform the correlation function, but discussion will also entail the use of RStudio in the programming functions behind the correlation process.

First, ensure the file with the 1951 tornado tracking is imported into the Rattle package that should be activated after opening the RStudio application. This is shown in the previous section in importing data. After the importing, ensure that you click on Execute so that the data is loaded, and if error messages appear (which will in this case), assign each variable to either "input" or "ignore." In this case, assign all the variables by TOR_LENGTH and TOR_WIDTH to the "ignore" radio buttons in order to limit the correlation. The analyst can choose to do correlations on all the variables, called multiple correlations, but this can be cumbersome and memory consuming. The finished screen should resemble the image that follows:



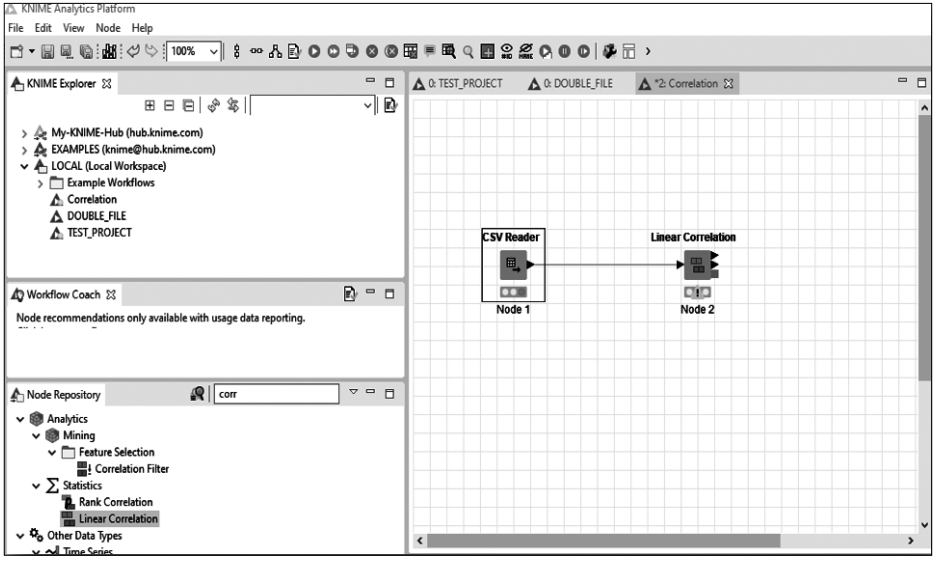
After clicking on the "Execute" icon in the toolbar, go to the "Explore" tab in order to use the correlation function. At this point, the analyst will follow the choices in this screen in order to receive the results that are displayed. If the results appear different from this screen, please ensure that the "Pearson" method is chosen under the drop-down box. If other methods are chosen, different results will appear, and somewhat drastically different, so be cautious and check the work.



As is shown, the correlation number (.04013) is the same as those that were presented in other sections. The ease with which this was done is pretty impressive. For single file functions, Rattle is a good option for data analysis using correlation.

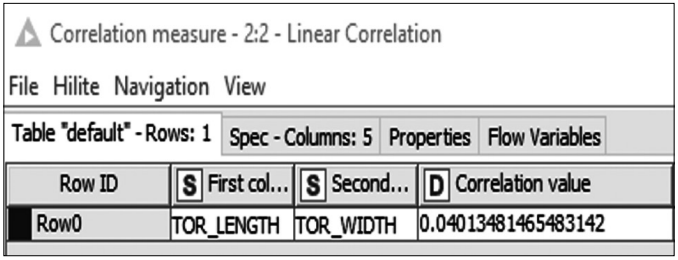
4.1.4 KNIME

The flexibility of KNIME is based on the node functionality, and the KNIME tool does have the ability to do a correlation with a node. To do this, open KNIME to the ongoing project, or even create a new project and start with importing the 1951 tornado tracking as was done in the import section. Once this is completed, add the correlation node to the CSV Reader node as depicted in the following screen.



The configuration of the Linear Correlation node consists of identifying the two variables to be tested, and in this case it will be the TOR_LENGTH and the TOR_WIDTH. These should be set in the configuration screen as shown in the following manner. Please ensure that the columns or variables the analyst identifies are correct, because this tool, like any other tool, will give the analyst the results that have been inputted, since it cannot predict what the analyst wanted, just what they chose. The analyst will want to right-click on the Correlation node and choose the option at the bottom of the sub-menu called Correlation Measure. This will reveal the following screen

and give the analyst an indication of the correlation between these two variables. Please remember that part of data analysis is not always picking the best correlation variables. There is a constant testing and reforming of hypotheses in order to analyze correctly. Please take that into consideration as this next screen is presented.



The analyst will immediately notice that the "Correlation value" is the same as the ones in previous sections. The main reason for using the same values and variables was to demonstrate to the analyst that the tool used for the function would not present different results. In this case, all tools presented the exact same result. The main cause of inconsistency could be a rounding issue or using a different formula, but in this case, all tools used the Pearson correlation method, and the formula for that method is very consistent across the many statistical texts. Some of these texts are included as references to this book.

4.2 REGRESSION

The one statistical concept that data analysts seem to understand, at least the ones that this author has taught, is regression. In fact, this author has seen regression applied to datasets that did not need this type of analysis. However, linear regression is somewhat important and needs to be addressed with respect to these tools. A quick review of the concept is necessary in order to set the stage for the subsequent demonstrations.

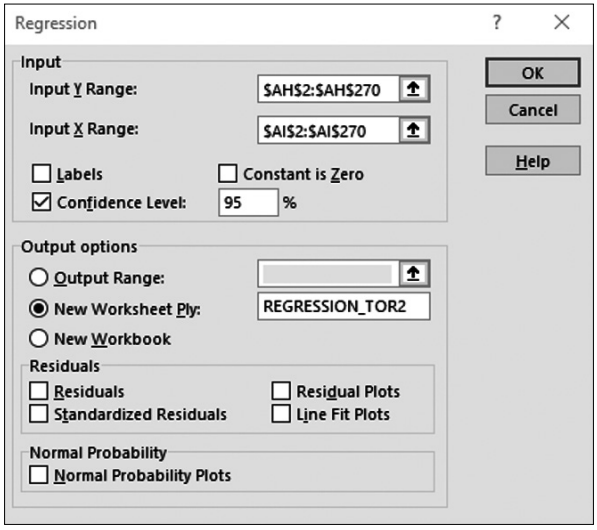
Linear regression is using a linear equation (sorry about that, it is necessary to relive the nightmare of high school math) to plot a possible prediction of future values based on past results. In essence, X-Y coordinate points are plotted using the two variables in the dataset that the analyst has chosen, and an equation is formulated from the plotting of those points. The equation that is formulated is a linear equation (conventionally) forming a line that tries to

split the data points where one half of the points are on one side of the line and half are on the other (approximately). In this section, the analyst will be applying tools to formulate this predictive equation. More explanation of the equation will be given after the first tool, in this case Excel.

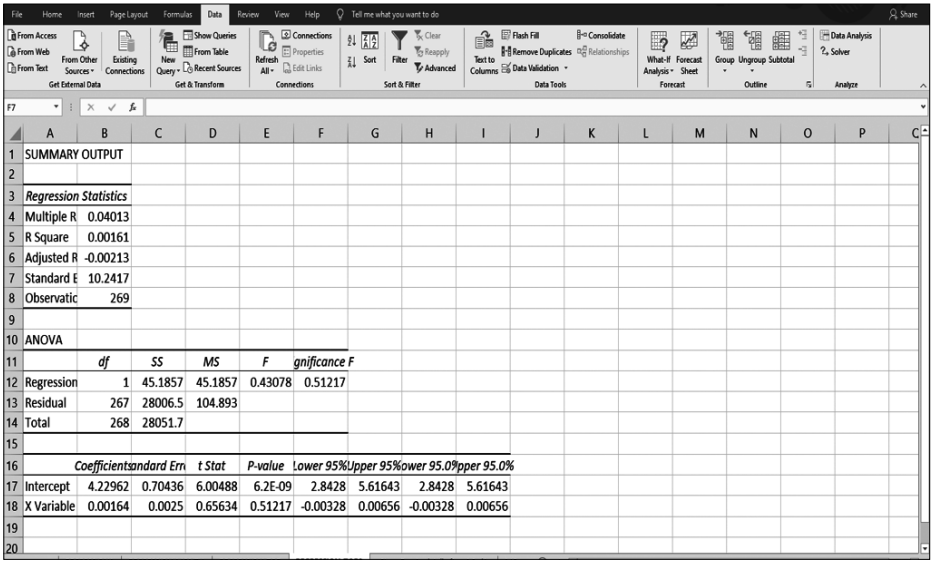
4.2.1 Excel

Excel, through the Analysis ToolPak, has a Regression function that does the job of presenting the Regression results in a quick fashion that can be used in presentations. The first step is to import the data of the 1951 tornado tracking so that the analyst can implement regression against two or more variables. In this instance, the analyst will be using regression against two variables, the tornado length (TOR_LENGTH) and the tornado width (TOR_WIDTH), for demonstration purposes.

After activating the Analysis ToolPak and choosing "Regression" from the menu, the analyst will select two variables for testing. The first will be the y-axis variable, which is called the "response" variable or "dependent" variable, and the second will be the x-axis variable, which is the "predicted" variable. In essence, at the completion of this task, what the analyst will place in the "x" will result in a "y." In this case, the analyst can place a width in the "x" and the result will be a tornado length. Again, this is for demonstration purposes and should not be construed as true predictive analysis for predicting tornado lengths. Remember that this is just one year of data.



The result of the regression analysis is as follows. Please notice all the different numbers on this screen. The ones that are important to the analyst immediately will be those that encompass the regression equation.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	C
1	SUMMARY OUTPUT																
2																	
3	*Regression Statistics*																
4	Multiple R	0.04013															
5	R Square	0.00161															
6	Adjusted R	-0.00213															
7	Standard E	10.2417															
8	Observatic	269															
9																	
10	ANOVA																
11		*df*	*SS*	*MS*	*F*	*gnificance F*											
12	Regression	1	45.1857	45.1857	0.43078	0.51217											
13	Residual	267	28006.5	104.893													
14	Total	268	28051.7														
15																	
16		*Coefficients*	*andard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*								
17	Intercept	4.22962	0.70436	6.00488	6.2E-09	2.8428	5.61643	2.8428	5.61643								
18	X Variable	0.00164	0.0025	0.65634	0.51217	-0.00328	0.00656	-0.00328	0.00656								
19																	
20																	

The numbers that are of interest to the analyst include the ones that are in the column "Coefficients" for both "Intercept" and "X Variable" (which in this case is TOR_LENGTH). The resulting regression equation will be y=0.0164x + 4.22962. This means that if the analyst wants to know what tornado width they should expect, the analyst places the tornado length for the "x" variable, multiplies it by .0164, and adds 4.33962 to find the approximate tornado width. Here is where regression can be misused. First, this is for one year only, taking into consideration 12 months, not all of which have tornados, or tornados of any length. Second, the correlation, as covered in the last section, is very slim—being .040—which means that there is less than .1 correlation, an extremely low correlation between day and tornado length. If the analyst wants to do a multiple regression, that can be done through this tool, but the columns for the data must be next to each other, so that demonstration will be discussed later. The main purpose for these variables is to ensure that the tools give a consistent result. The main point here is that these two variables are not a good combination for the purpose of regression, given the correlation and the lack of longitudinal data.

4.2.2 OpenOffice

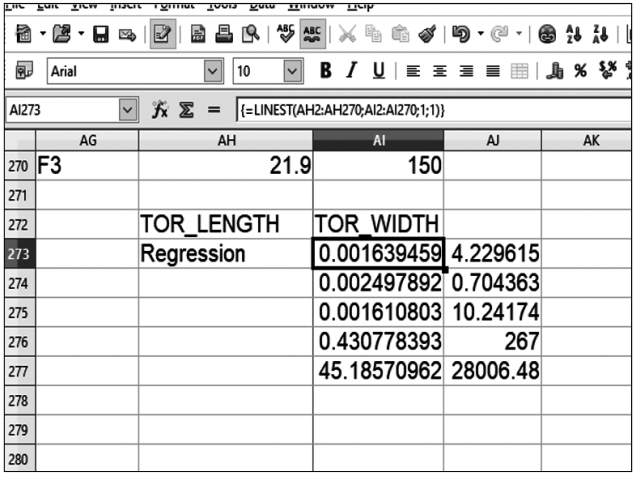
The OpenOffice regression function is comparable to the "pre-Analysis ToolPak" formula for Excel. The formula function is called "linest" and is called an "array formula." What this means is that the result of the formula is carried across several cells. To make a formula an array formula, before pressing the ENTER key, combine the CTRL-SHIFT-ENTER keys to transform the formula into an array formula. The formula will be enclosed in "curly" brackets ({}) rather than parentheses.

The first step to using the regression formula in OpenOffice is to open the file that has been used in the previous sections and ensure that the variables selected are TOR_LENGTH and TOR_WIDTH for this demonstration.

The next step is to place the formula for regression in a blank cell (much like an analyst does in Excel). Remember that a semicolon separates the parameters of the formula—not commas. The formula for regression will look similar to this (for the specific columns/variables mentioned previously).

```
=LINEST(AH2:AH270;AI2:AI270;1;1)
```

At first blush, this formula would look very much the LINEST formula in Excel, except for the semicolons. The difference between this and other formulas in OpenOffice is remembering the CTRL-SHIFT-ENTER in order to see the entire result of the regression analysis. The following screen shows what happens when you finish the formula in this fashion.



	AG	AH	AI	AJ	AK
270	F3	21.9	150		
271					
272		TOR_LENGTH	TOR_WIDTH		
273		Regression	0.001639459	4.229615	
274			0.002497892	0.704363	
275			0.001610803	10.24174	
276			0.430778393	267	
277			45.18570962	28006.48	
278					
279					
280					

What do all these numbers mean? How do they relate to the result in Excel? The following table will shed a little light on the previous cells in OpenOffice. The site to get the full translation is located here: *https://wiki.openoffice.org/wiki/Documentation/How_Tos/Calc:_LINEST_function*.

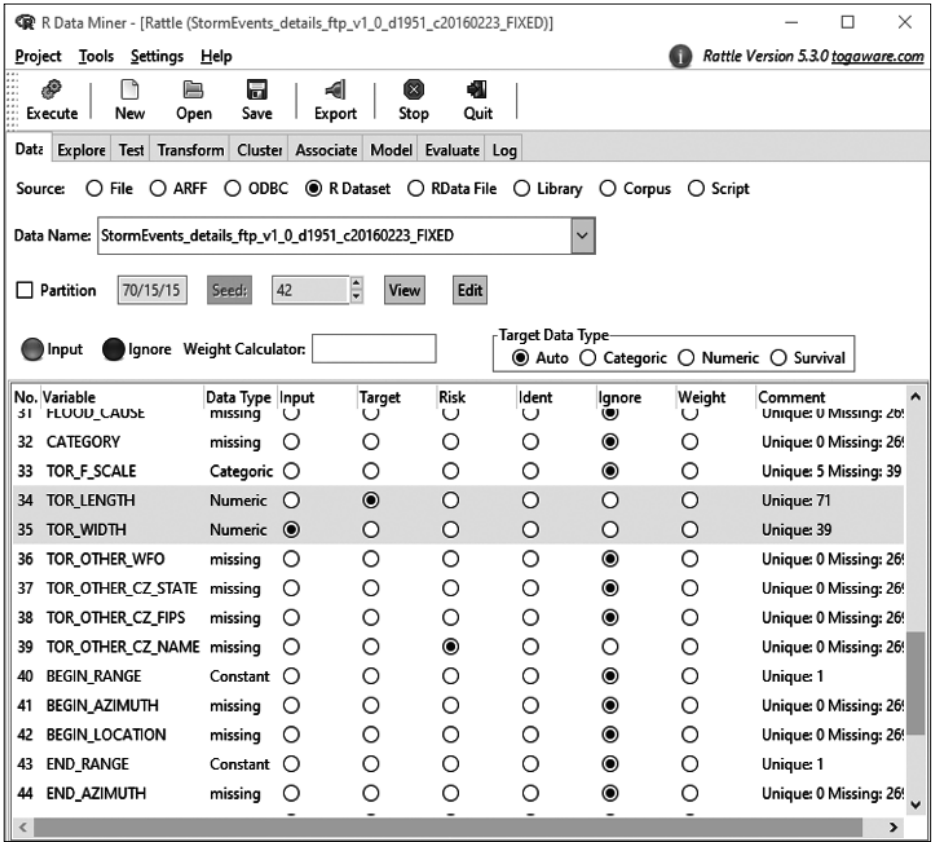
For this book, just a few of the cells in the table are included.

"x" value	"y" intercept
R^2 (explained in this section)	
"F" (Levene Test explained in this book)	Degrees of Freedom (beyond scope of this book)

What does this table tell us? The first cell on the left is the same as the "X Variable" from the Excel result (AI 273 in this case). The second cell ("Y" Variable) is the same as the "intercept" (AJ 273). The "r^2" is the square of the "r" correlation value and would reflect the value marked "R Square" in the Excel table. Those cells in the previous table are the most important at this juncture and the most valuable to the data analyst using the tool. Using the table cells explained, a linear equation can now be formulated, and the correlation can be calculated. Not as "pretty" as Excel, but it gives the same results.

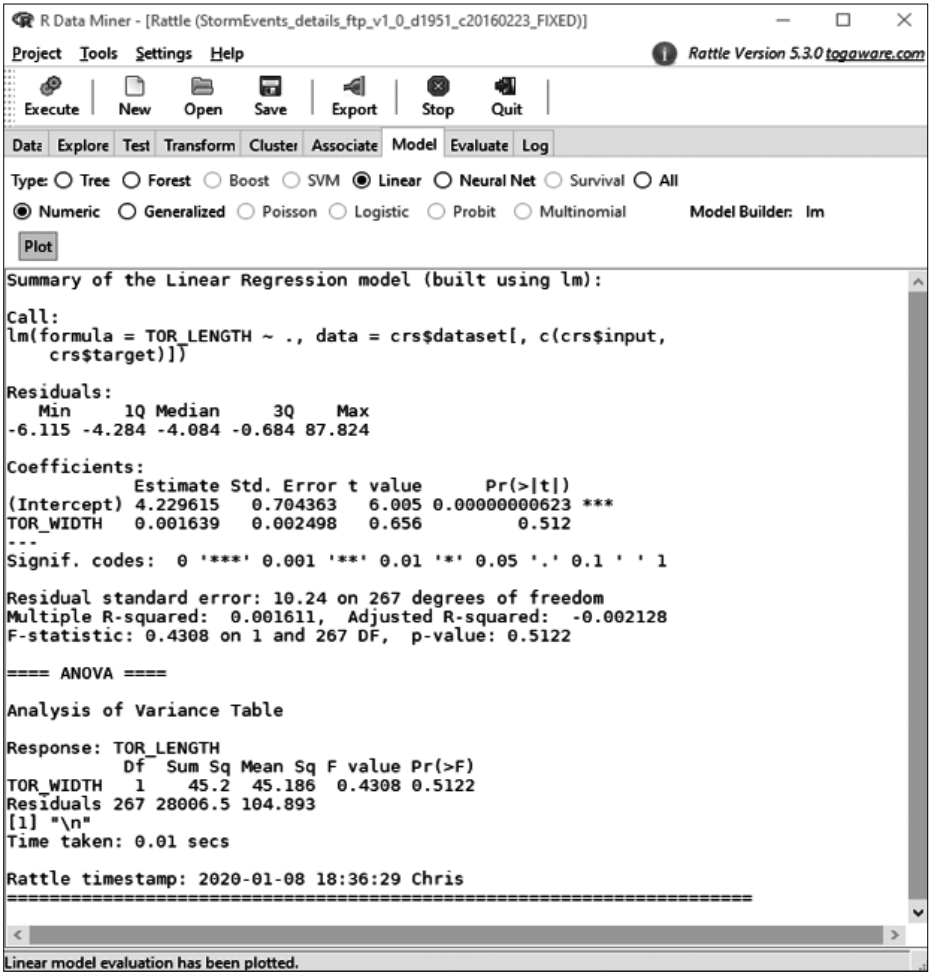
4.2.3 R/RStudio/Rattle

The Rattle package within R provides a perfectly adequate method of regression analysis. The steps for setting up the data to be analyzed in this fashion are the same as before—import the same data as has been used in the past sections and use TOR_LENGTH and TOR_WIDTH for the regression analysis. The Rattle screen for the data should appear as the following:



Please notice that all of the other variables have been placed in "Ignore" since the analyst will have no need of them at this point. TOR_LENGTH has been placed in the "Target" column, while TOR_WIDTH will be placed in the "Input" column. This is the same as placing TOR_LENGTH in the "y-axis" and TOR_WIDTH in the "X-axis." For the future use of multiple regression, the other variables may be reentered into the "Input" column to include them. If any changes are made to the dataset, ensure that the analyst clicks on the "Execute" icon. Otherwise, the table will be as it was before the changes. Execute saves the changes.

Once the data is set and the Execute icon is clicked, go to the "Model" tab and select the "Linear" type and "Numeric" below that choice. After that, click on the "Execute" icon and the following screen will appear. It is evident that the results here by Rattle match the results from the other tools.

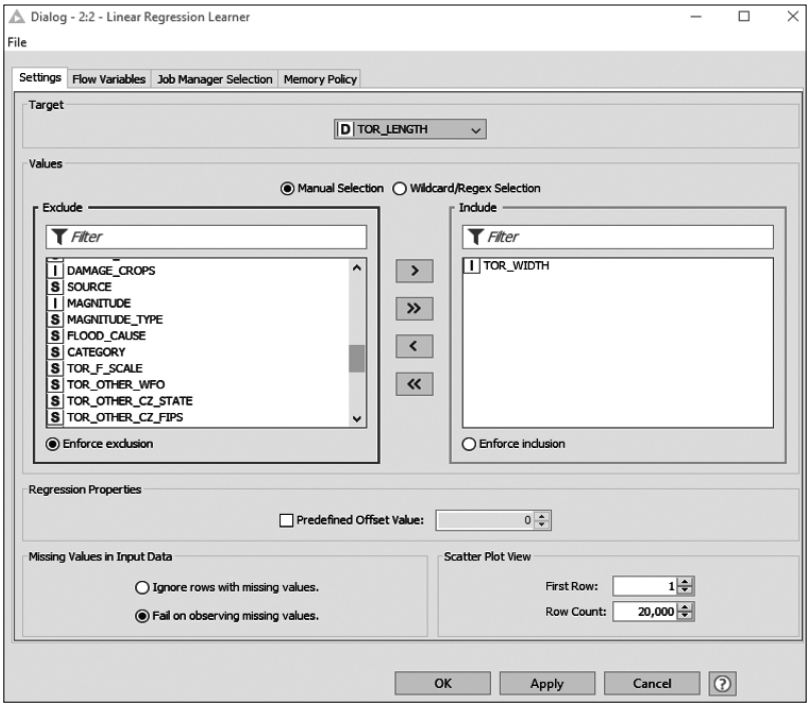


For now, it appears that every tool has agreed with the others concerning this concept, so this would be a great way to verify results from regression. With the function being so readily available and relatively simple to use, there is no reason why verification of results would not be undertaken in this situation.

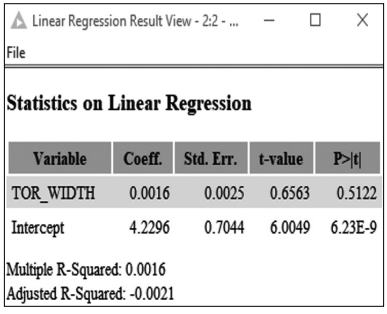
4.2.4 KNIME

KNIME is the final tool that will be addressed in this regression concept. As in the other cases, the KNIME application does have a node available for regression analysis. Once the data is imported through the CSV Reader, the analyst can connect the data to the regression node, called the Linear

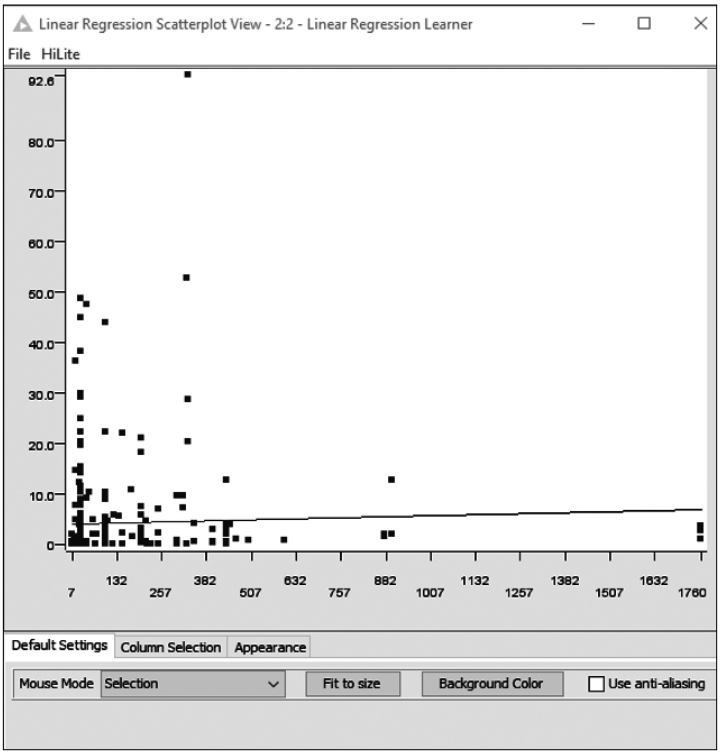
Regression Learner node, found by typing in the word "regression" in the search box. Once the node is dragged, placed, and connected, double-click on it to open it, revealing the following configuration screen.



Configure the screen exactly as it appears in the previous figure, using TOR_LENGTH as the target (the same as Rattle) and TOR_WIDTH as the column to include against this target. Click OK and execute the node (remember the green arrow). Once the node is executed, right-click on that node and choose "View: Linear Regression Result View." This will produce this window, which shows the same results as in the past sections.



If the analyst desires, they can also select the menu choice below the one producing this window to show the scatterplot result, which will match the scatterplot in Excel.



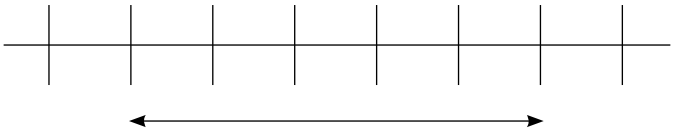
It is very important that the analyst understand that regression is a model and must be verified and tested through evaluation techniques. This is beyond the scope of this book, but the tools presented here have a wide array of testing functions for this purpose. As stated before, exploring each or all of these tools will enhance knowledge of both data analytics and statistics.

4.3 CONFIDENCE INTERVAL

The confidence interval has been making a comeback in the statistics arena. In his book, *Statistics Done Wrong*, Alex Reinhart explains that the confidence interval is a simple statistical method that has not been used as often as necessary, which has led to some interesting, if not inaccurate, results

(Reinhart, 2015). Since confidence intervals are a relatively simple method to gauge the effect of a value on another value, a brief explanation is necessary before heading into the tool use.

When a data analyst addresses a confidence interval, they are using the one-dimensional perspective of a confidence level. Although this may sound confusing, a simple illustration will probably help to clarify the concept. If an analyst were to remove the bell curve from the standard normal curve, the result would be as follows:



The center vertical line represents the mean, and the lines to the right and left represent the different standard deviations plus or minus that mean. The arrows represent the 95% confidence interval based on the 95% confidence level. In essence, what this means is that, given a sample of a population, the confidence interval will tell the analyst what the chances are that the mean is located between those intervals. In other words, at the 95% confidence interval, there is a 95% chance that the mean lies somewhere between the Upper Confidence Level (UCL) and the Lower Confidence Level (LCL).

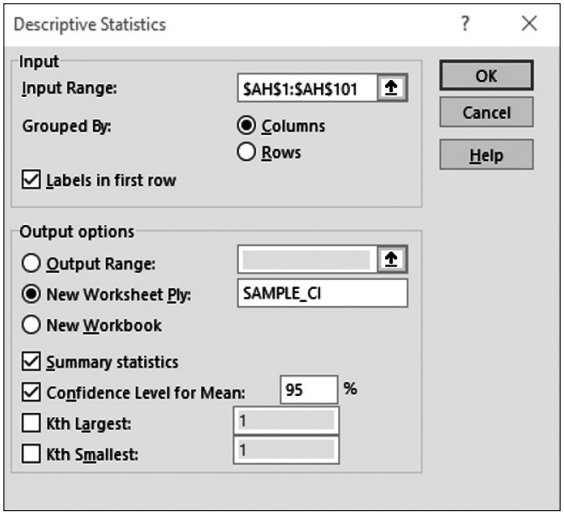
How does this help with data analytics? The answer again derives from Reinhart, who addresses confidence interval with an almost reverent tone in his book. Throughout the book, he gives a wide variety of examples, most from real studies, that provides explicit defense of using the confidence interval. The bottom line in confidence intervals is that it helps to provide verification of the results of studies. For instance, if a sample shows the average length of a person's employment is 8 years and, using a 95% confidence interval, the range is between 6 and 10, the analyst can state that with 95% certainty, that the average of a person's employment in the population is somewhere between 6 and 10 years. (Reinhart, 2015). The whole idea of the confidence interval is to provide the analyst with a reading on how well the study was conducted. Reinhart stated that if the interval is too wide, as in our example the result was 1–20, that would mean that there was not enough sampling done and that it is necessary to gather more samples in order to narrow the interval (Reinhart, 2015).

The reason for learning the confidence interval is to incorporate it into any study or analytical project the analyst has to perform, and to show that the actual method is simple, effective, and available within a wide variety of tools.

4.3.1 Excel

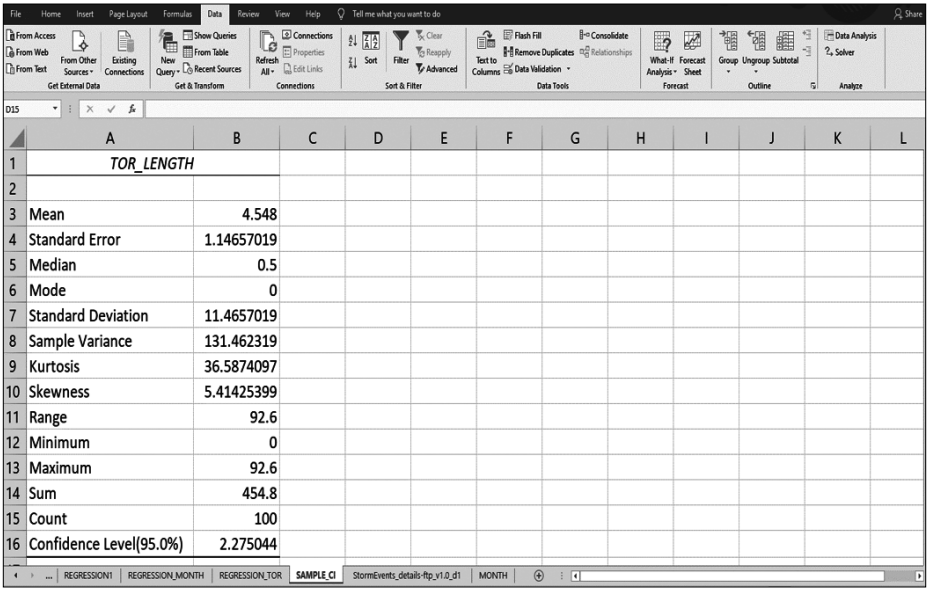
Excel, through the Analysis ToolPak, has the ability to present the confidence interval as part of the overall descriptive statistics portion of the ToolPak. This procedure starts with the import of the data, again using the 1951 tornado tracking data, and opening the Analysis ToolPak. Select Descriptive Statistics from the ToolPak menu and use TOR_LENGTH as the column to be used, but this time only choose the first 100 rows. This will be the sample that will be used to compare the confidence interval to the actual population mean at the end. Please realize that just picking the first 100 rows is not a true random sample, but that will be discussed in a later section. For this demonstration, this will suffice.

Once the Analysis ToolPak is opened and the column and rows are chosen, the screen will appear as the following screen:



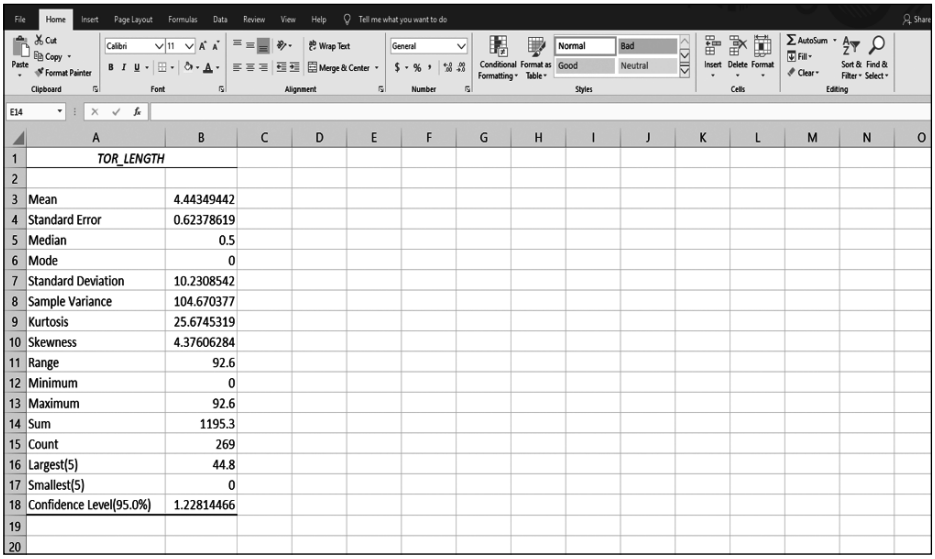
A word of caution concerning this screen. Please notice that "Labels in first row" is checked. If there are no labels in the selection, Excel will automatically take the first row and use it as a label, not reporting it in the data analysis. This will lead to problems with any analysis done on the data. Just a word of warning to those that use the default settings or configurations. Please check these setting before clicking OK. Also, please notice that Summary Statistics and Confidence Level for Means are both checked. If the analyst does not check the Summary Statistics block, that will be a problem, but

if they do not check the Confidence Level for Means block, the result that is desired in this situation will not appear. As stated before, please ensure that the configuration is the way that is desired before executing. The result of the confidence level is in the following screen. But what does it mean?



	A	B
1	*TOR_LENGTH*	
2		
3	Mean	4.548
4	Standard Error	1.14657019
5	Median	0.5
6	Mode	0
7	Standard Deviation	11.4657019
8	Sample Variance	131.462319
9	Kurtosis	36.5874097
10	Skewness	5.41425399
11	Range	92.6
12	Minimum	0
13	Maximum	92.6
14	Sum	454.8
15	Count	100
16	Confidence Level(95.0%)	2.275044

The Confidence Level (95.0%) row in the result states that the confidence level is 2.275044. What this means is that, with 95% confidence, the population mean is somewhere between ±2.275044, which means, taking into consideration that the sample mean is 4.548 according to the summary results, the population mean is somewhere between 2.272956 and 6.823044. To further show this result, take the entire TOR_LENGTH column into the mean for the summary statistics, using the ToolPak with the following result.

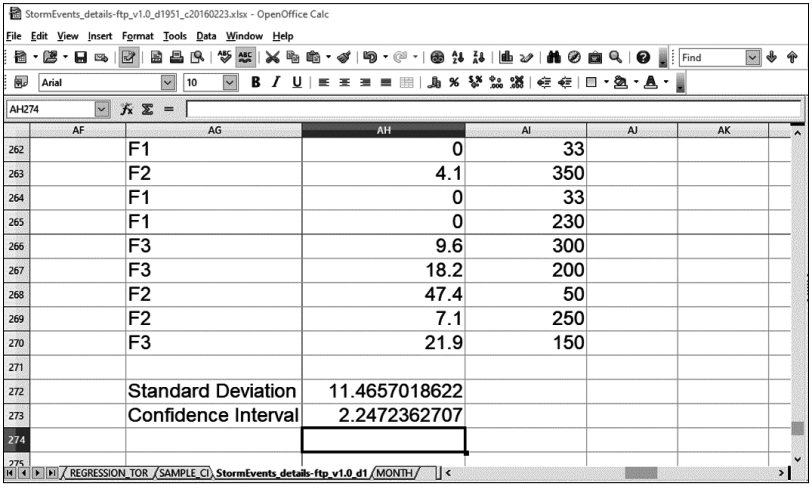


	A	B
1	*TOR_LENGTH*	
2		
3	Mean	4.44349442
4	Standard Error	0.62378619
5	Median	0.5
6	Mode	0
7	Standard Deviation	10.2308542
8	Sample Variance	104.670377
9	Kurtosis	25.6745319
10	Skewness	4.37606284
11	Range	92.6
12	Minimum	0
13	Maximum	92.6
14	Sum	1195.3
15	Count	269
16	Largest(5)	44.8
17	Smallest(5)	0
18	Confidence Level(95.0%)	1.22814466
19		
20		

The population mean is 4.4349442, which is between 2.27 and 6.82. If the confidence level is lower, the range will also be lower, so a confidence level of 80% will produce a range that will be narrower.

4.3.2 OpenOffice

The OpenOffice function of the confidence interval does not have the same convenience of the Analysis ToolPak of Excel, but it does the job. The first step is the same as other sections; import the 1951 tornado data and pick an empty cell in that worksheet to display the confidence interval as shown.



	AF	AG	AH	AI	AJ	AK
262		F1	0	33		
263		F2	4.1	350		
264		F1	0	33		
265		F1	0	230		
266		F3	9.6	300		
267		F3	18.2	200		
268		F2	47.4	50		
269		F2	7.1	250		
270		F3	21.9	150		
271						
272		Standard Deviation	11.4657018622			
273		Confidence Interval	2.2472362707			
274						
275						

In this example, the Confidence Interval results are placed at the bottom of the TOR_LENGTH column. In this way, selecting the column should be much easier. Remember that only the first 100 rows will be selected, but in this instance start at AH2, not AH1. OpenOffice does not have the same regard for headings, so ensure they are not included in the data pull. The formula for the confidence interval must include the standard deviation, which is STDEVA or the standard deviation of a sample. This formula should be placed in AH272.

```
=STDEVA(AH2:AH101)
```

Now what this does is provide the ability of the confidence interval to employ the standard deviation into the confidence interval formula, which should be placed in AH273, and should look like this:

```
=CONFIDENCE(0.05;AH272;100)
```

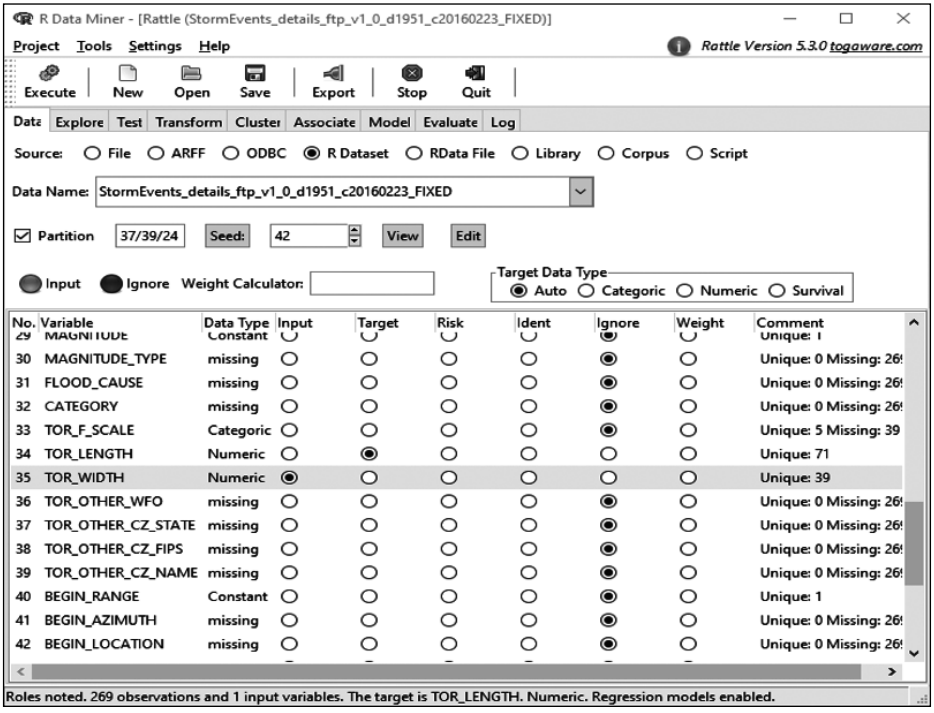
An explanation is necessary on the previous formula. The "0.05" in the beginning is what is called an "alpha value." When statistics refers to "alpha" it means the probability that there is a "false positive" from the results. This is also called a Type 1 error, but the real calculation to determine the alpha is taking 1 – Confidence Level. In this case, the confidence level is 95% or .95. If the analyst takes 1–.95 the result is .05. A long explanation, but it is one that is necessary with OpenOffice, since the tool uses alpha much more than confidence level. The result does not match exactly those of Excel, but again this could be a rounding situation or one where the formula for the confidence interval has a more exact calculation than the other. The difference is not drastic (.03 difference), so these results could be used to verify each other. The standard deviation is exactly the same, which gives much credence to the veracity and consistency of these tools.

One more point before leaving OpenOffice. Remember, the more sampling, the narrower the range of the confidence interval. This will hold true in all the other tools, and in any other statistical tool involved in data analysis. As with any study, the more sampling, the more accurate the results, as long as the sampling is done randomly.

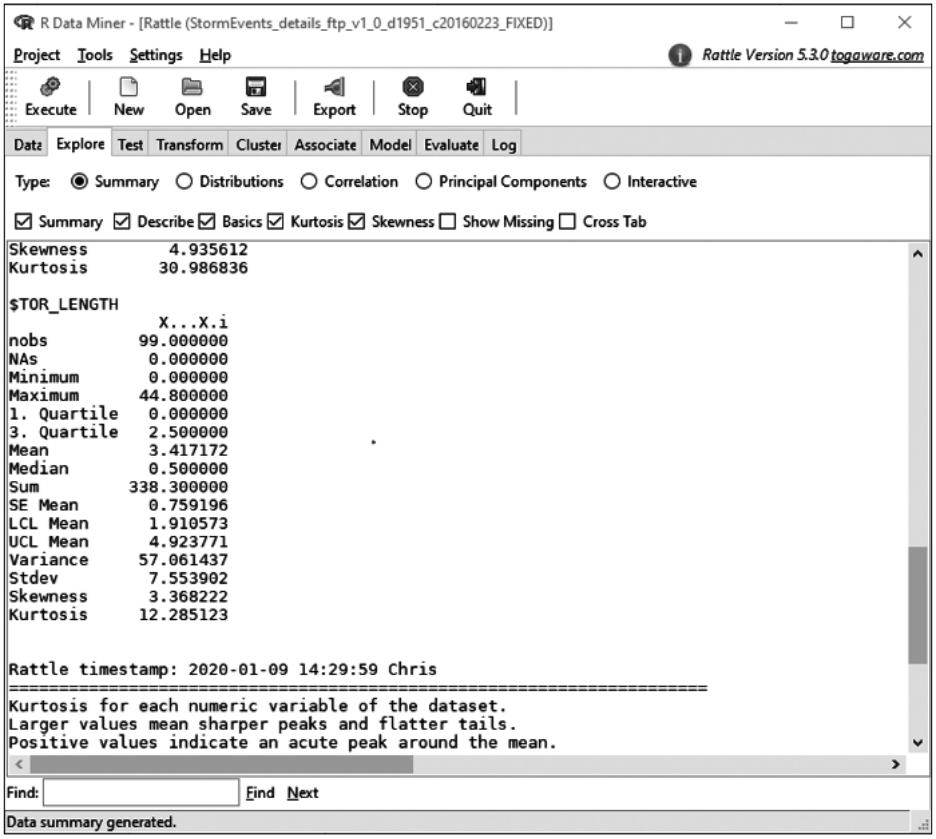
4.3.3 R/RStudio/Rattle

The use of Rattle for confidence interval is about the same as with the Analysis ToolPak in Excel. The first step is to import the data and have it prepared for

the method. In this case, the only active input will be TOR_LENGTH. Also notice that the "partition" block is checked. The notation "70/15/15" means that the data is split into 70% training, 15% validation, and 15% testing. What that means is that 70% is sampled, which would be about 140 rows. This is slightly more than what this test entails, so change the partition block to 37/39/24, which will make the training data at 37%, making the rows 99 and very close to the previous sections. The newly configured screen is as follows:

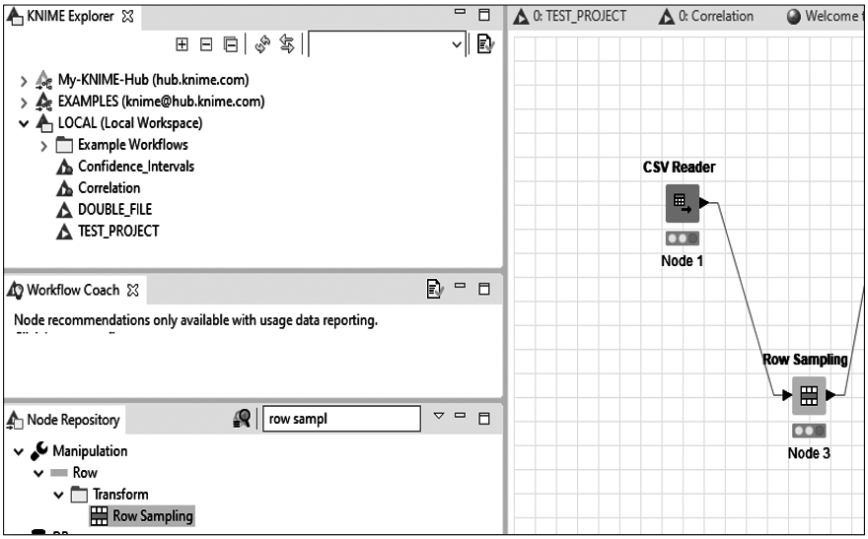


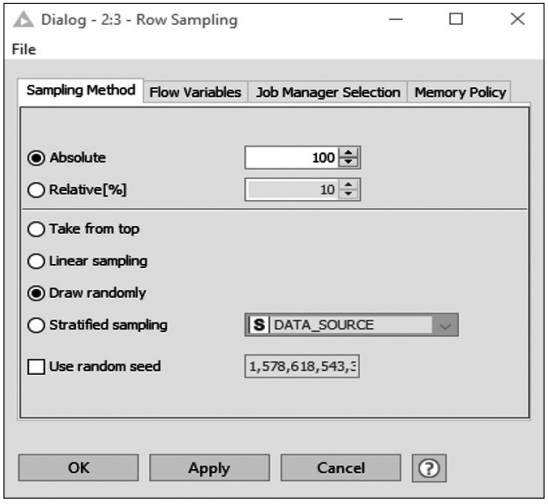
This is where Rattle goes the extra mile automatically. Notice that next to the partition block there is a "seed" button. What this does is randomly sample the dataset using the seed value as a beginning point. This means that this data will be randomly sampled, and that will be tested with the next screen for descriptive statistics, illustrated as follows. Ensure that the boxes are checked almost all the way across except for the last two.



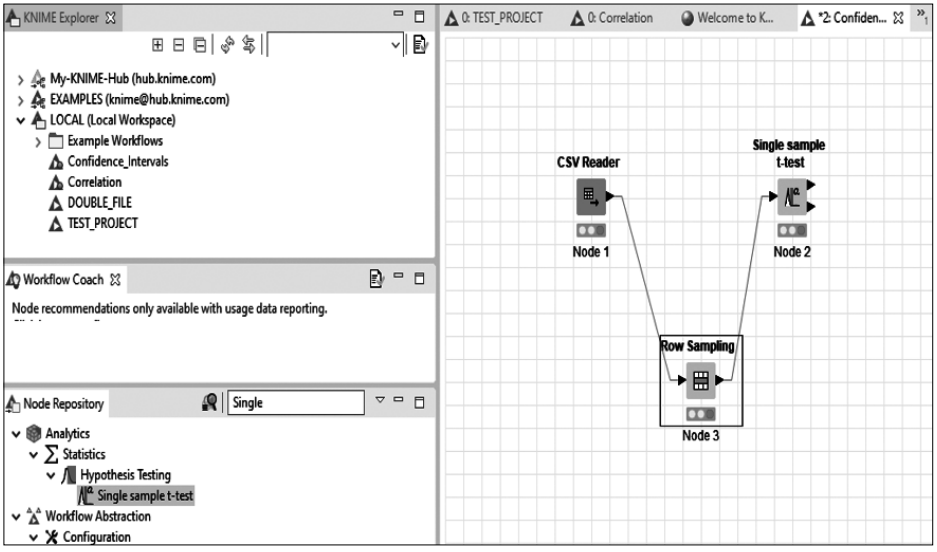
4.3.4 KNIME

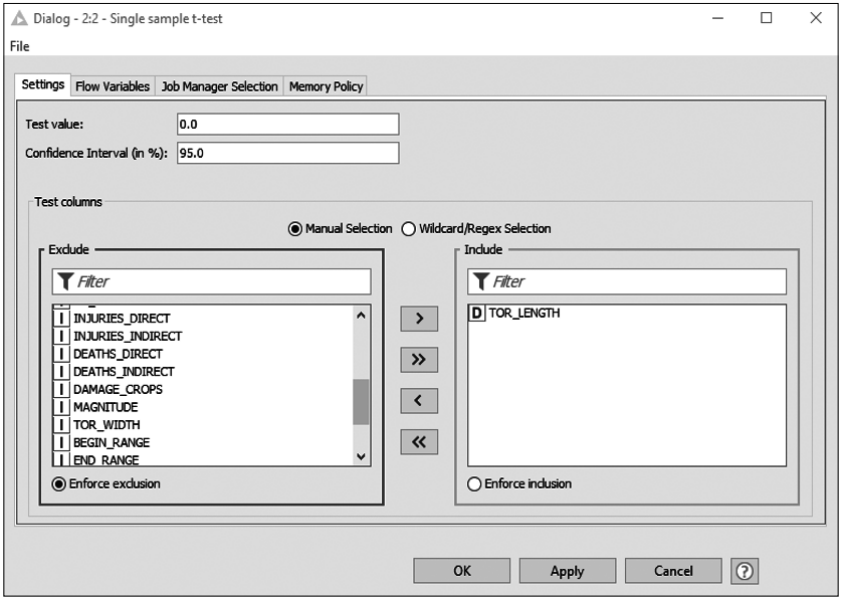
KNIME provides a node for confidence intervals, but it is part of another node, so it is important to use research in order to see how these nodes can be used in several types of methods. First, it is important to use the sampling of TOR_LENGTH as we did in the other sections. In this case, the node to include is called "Row Sampling." Once that is dragged, placed, and connected, the configuration should look like this for random sampling. However, please note the many combinations of sampling that are available with this node. The configuration to be used in this section is as follows:





As the analyst can see from the configuration screen, the current sample is for 100 values that are drawn randomly. KNIME provides one node for this sampling and, from experience, it is a good way to get a sampling with one step. The sampling is completed, but there still needs to be a node that provides the confidence interval, and that node comes from the Single Sample T-Test node. Once that node is dragged, placed, and connected, the following screens will show the final flow and the configuration screen for this node.





As the analyst can see from these screens, the configuration for the t-test node is straightforward. The Confidence Interval is 95%, which can be changed, and the variable is TOR_LENGTH. Notice that the "Test value" is 0. There is a reason for this in this instance. Under normal t-tests with one

sample, the null hypothesis would be based on the mean of the sample being equal to a value. The test value is that value, so it is good to know why that block is available within this node. This will not be used in this case. After executing all the nodes, the following result is available by right-clicking on the t-test node and choosing "View: Test Statistics." The analyst can see the CI (Lower Bound) and CI (Upper Bound). That means that the population mean is somewhere between these two values. There is a 95% chance that it is between these two values. Do not worry about the any of the other numbers in this row, since they relate to the t-test. However, as the analyst can see, the numbers are very close to the other sections, which shows that the results are at least somewhat consistent with the sample of 100.

Test statistics - 2:2 - Single sample t-test

File

Single Sample T-Test

Descriptive Statistics

	N	Missing Count	Mean	Standard Deviation	Standard Error Mean
TOR_LENGTH	100	0	4.716	11.3613	1.1361

Single Sample Test

Confidence Interval (CI) Probability: 95.0%

	Test Value	t	df	p-value (2-tailed)	Mean Difference	CI (Lower Bound)	CI (Upper Bound)
TOR_LENGTH	0.0	4.1509	99	7.02E-5	4.716	2.4617	6.9703

4.4 RANDOM SAMPLING

Through years of teaching statistics and data analytics, random sampling is a commonly misunderstood concept from students. Students often consider picking the first 10 or 20 values in a dataset random sampling when it is not random. Random is taking into consideration all values equally (Reinhart, 2015). This is what some researchers do not do by reason of convenience or necessity (Reinhart, 2015). However, in order to properly measure the center of a dataset or to accurately predict the effect of an event on another event in a population, an accurate sample is necessary. There are several methods to

perform this function, and all the tools mentioned have ways of random sampling. Some of these have been mentioned in the previous sections, but this is a refresher on those methods.

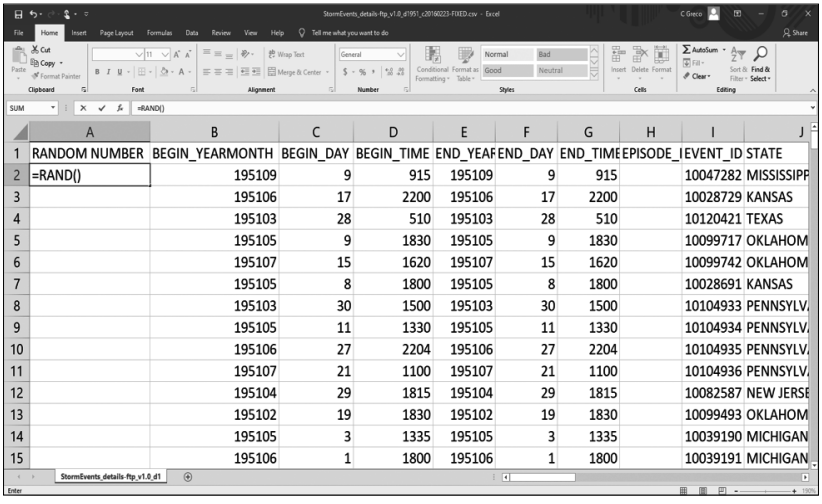
4.4.1 Excel

Excel has the Analysis ToolPak, which in turn has a function for sampling, but when the analyst does this, they will get duplicates in the process. A way of preventing duplicates is the assignation of a unique value to every event, thereby preventing duplicates. The first step in this process is to load the dataset that has been used in past sections—the 1951 tornado tracking. Once that has been completed, insert a column at the beginning of the data and name it "Random Numbers," since this will be the unique number that will be assigned to each row.

In the first row of that column, place the following formula:

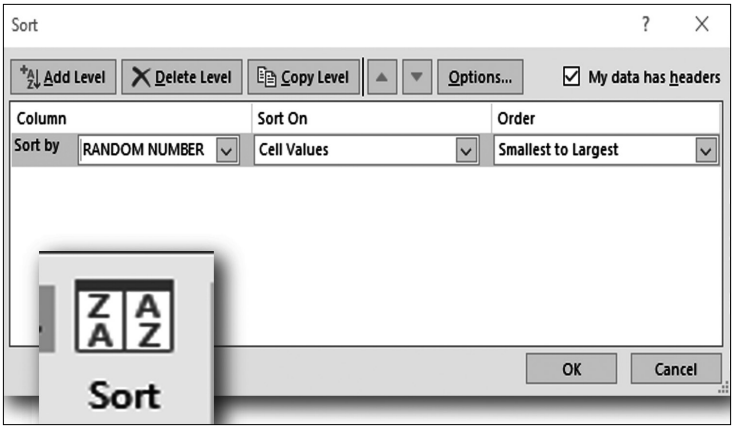
```
= RAND()
```

After that step, ensure that all of that column has the same formula by double-clicking the "fill handle" located at the bottom right-hand side of the formula cell (the "fill handle" will resemble a bold plus sign). By double-clicking on the fill handle, all the blank cells in the formula column will be filled with a random number. In fact, every time the analyst does any calculation or presses the ENTER key, the rows will reshuffle. This will make any sample of the data truly random. The finished dataset, before shuffling and after shuffling, is shown as follows:



	A	B	C	D	E	F	G	H	I	J
1	RANDOM NUMBER	BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEAR	END_DAY	END_TIME	EPISODE_I	EVENT_ID	STATE
2	=RAND()	195109	9	915	195109	9	915		10047282	MISSISSIPP
3		195106	17	2200	195106	17	2200		10028729	KANSAS
4		195103	28	510	195103	28	510		10120421	TEXAS
5		195105	9	1830	195105	9	1830		10099717	OKLAHOM
6		195107	15	1620	195107	15	1620		10099742	OKLAHOM
7		195105	8	1800	195105	8	1800		10028691	KANSAS
8		195103	30	1500	195103	30	1500		10104933	PENNSYLV
9		195105	11	1330	195105	11	1330		10104934	PENNSYLV
10		195106	27	2204	195106	27	2204		10104935	PENNSYLV
11		195107	21	1100	195107	21	1100		10104936	PENNSYLV
12		195104	29	1815	195104	29	1815		10082587	NEW JERSE
13		195102	19	1830	195102	19	1830		10099493	OKLAHOM
14		195105	3	1335	195105	3	1335		10039190	MICHIGAN
15		195106	1	1800	195106	1	1800		10039191	MICHIGAN

In order to sort to ensure that the random number covers all the columns, use the sort option in the Data tab of the screen, the one that resembles the following screen. By clicking on this icon in the toolbar, the following menu will appear. Use the down arrow to choose the column with the random number and click OK.

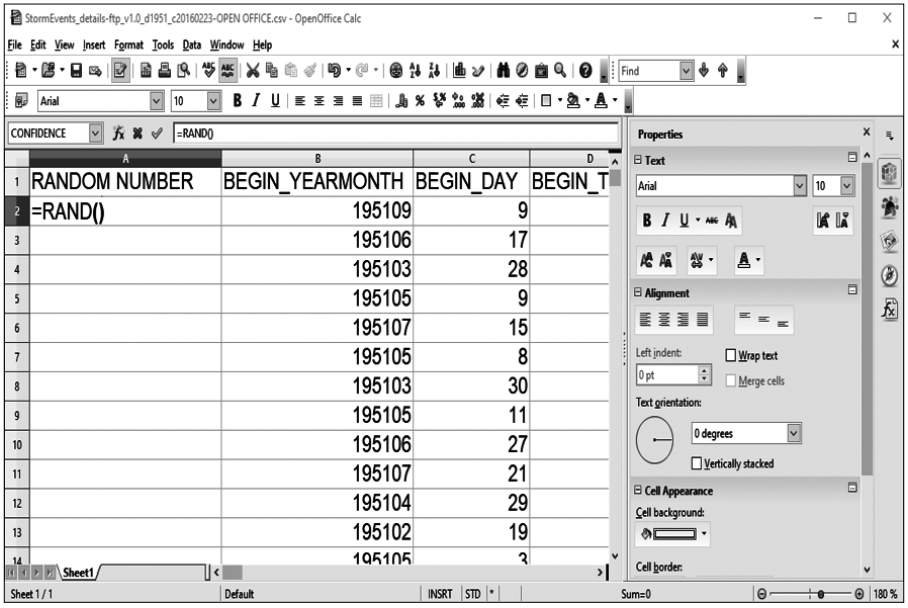


One word of caution at this point for this function. Please ensure the cell selected is the one in the RANDOM NUMBER column. Otherwise, the analyst will only be sorting based on the column selected. If the selected column is a month column, then the sorting will be on the month name—not on the random number. If the selected cell is the random number, the analyst will see the columns shuffled accordingly. This is a simple way of shuffling the deck without fear of the data being biased through systematic sampling without randomness.

4.4.2 OpenOffice

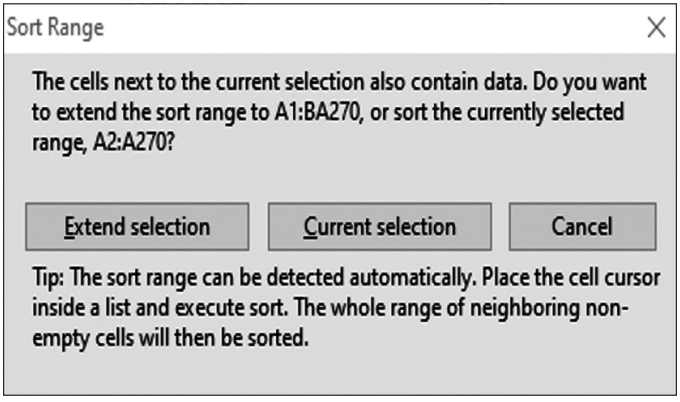
Think of OpenOffice as very similar to Excel, with some very slight differences in formulas. Because of this, random sampling in OpenOffice will be very similar to the process in Excel.

The first step is, of course, to import the dataset, after which the analyst performs the same functions as Excel by inserting a column and naming it RANDOM NUMBERS and placing the same formula as Excel in that first cell of that column. The screen should resemble this when that step is completed:

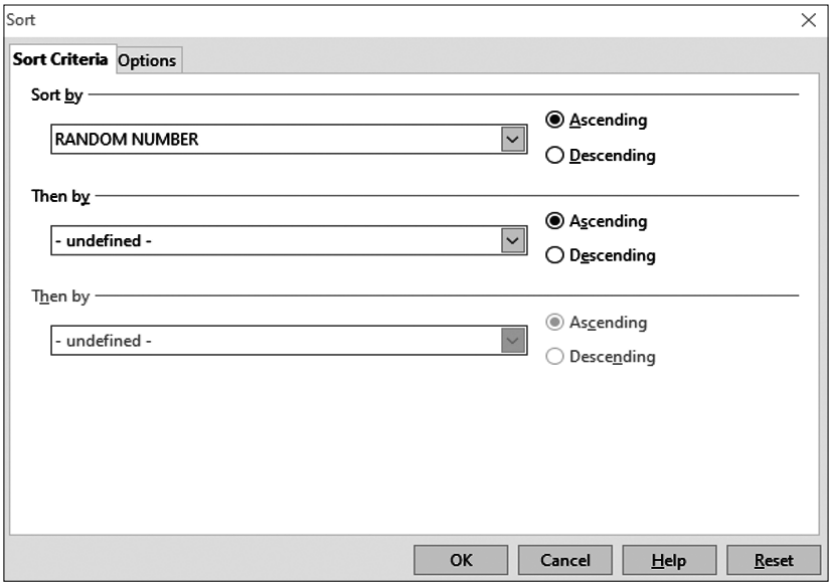


If this screen looks familiar, it should be, since the formatting is very similar to the prior section. Oh, and the double-clicking of the "fill handle" also works with this tool. This makes it much easier to copy the formulas down a column, including all rows. A bit of caution here is necessary. If any rows are blank or contain blank values next to the column that you are double-clicking, there may be a stop at the row, so it is beneficial for the analyst to check the column to ensure that all rows have a random number.

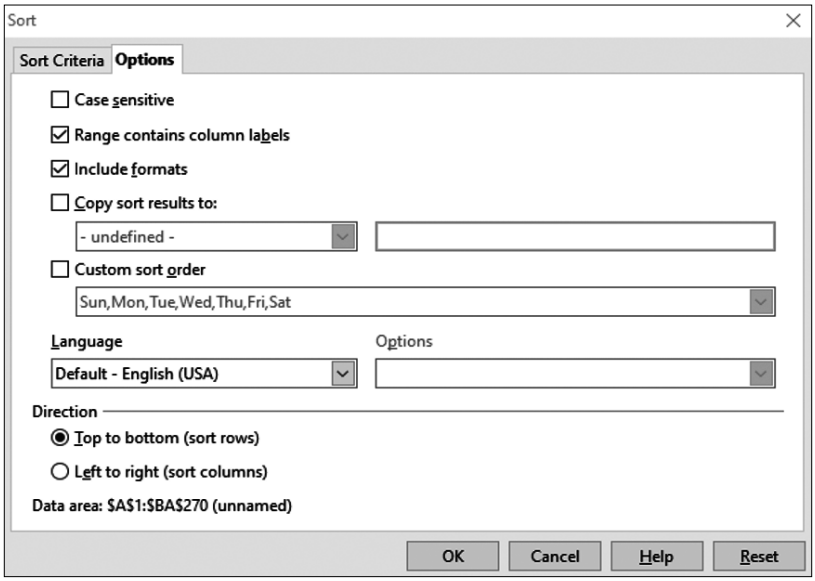
Once that step is completed, go to the Data tab of the toolbar and choose the Sort option. Once that is completed, the following screen will appear:



Choose the "Extend selection" button in order to have the entire dataset placed under the focus of the random number sort. In this way, the entire dataset (with rows intact, which is very important), will be sorted based on the random number column. Then simply continue sorting to shuffle the dataset. One recommendation on sorting the dataset is to not save the dataset until the analyst knows that all the rows are still intact.

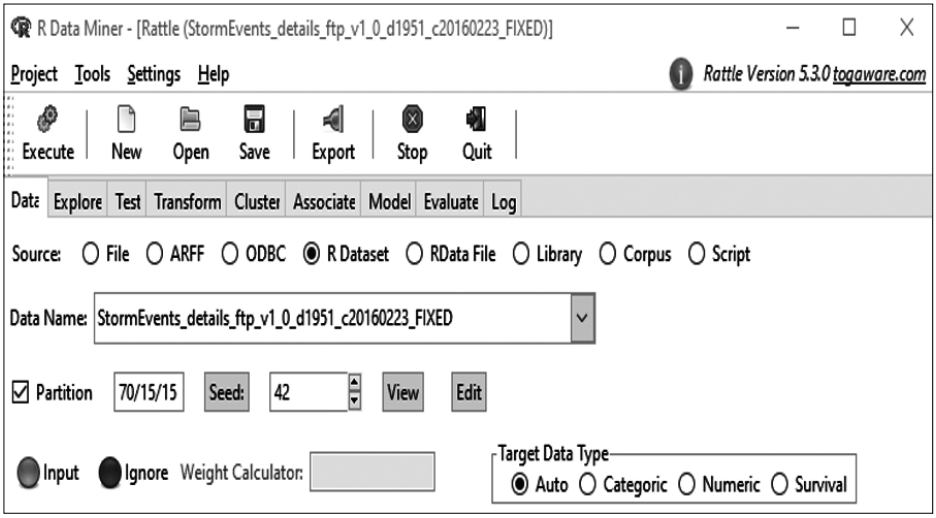


The following screen shows the "options" tab of the same menu as previously. Although the analyst may gloss over this tab, it is important to see the different options that are offered with this sort function to ensure that they match the analyst's intentions. A checkbox like "Range contains column labels" is important when sorting since, if the box is unchecked, the headers will be sorted with the data. This could be a messy result that will affect the different tests conducted by the analyst. In other words, be careful about these different screens and explore them whenever possible.

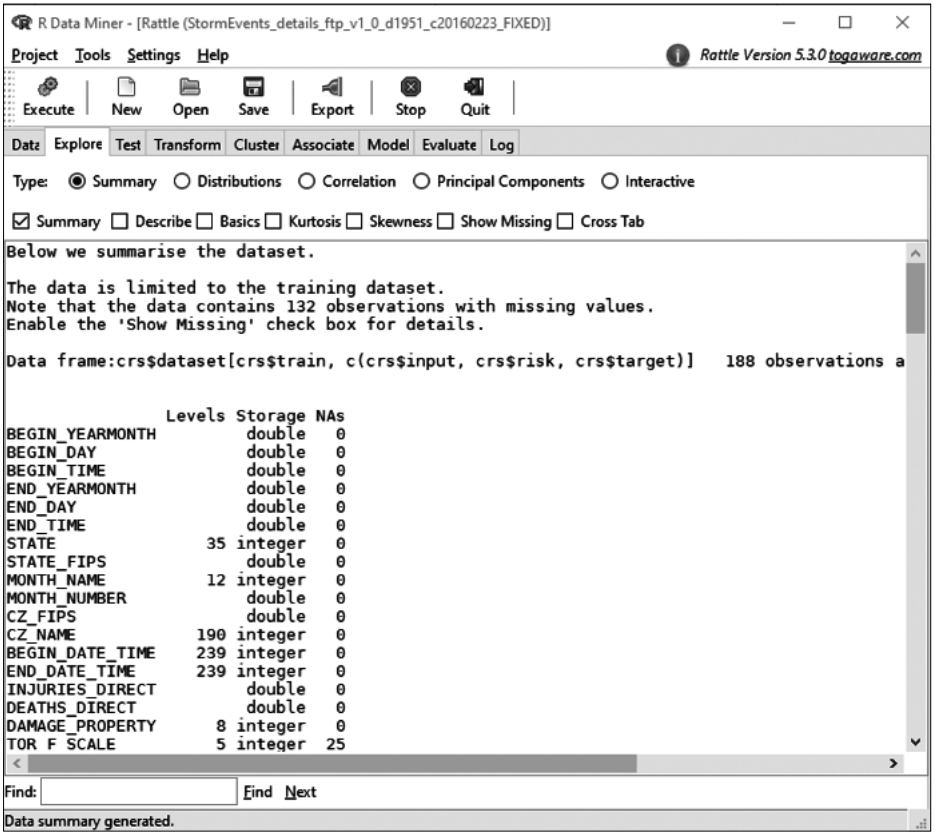


4.4.3 R/RStudio/Rattle

The sampling done with Rattle has been covered in past sections but will be refreshed with this specific topic. Once the Rattle package is activated within RStudio and the dataset is loaded, the next step will be to random sample the data. In this case, the entire dataset will be included in the sampling, whereas in the past section specific columns were identified in the Data tab. The Dat a tab should look like this if the analyst wants to sample 50% of the dataset (called the training dataset) and use that as a way of testing the different methods and functions. At this point, the analyst should ensure that there is only one "Risk" variable (the tool will warn you if there are more) and that the "Partition" values add up to 100 (the tool will warn you about this also). The following screen will sum it up for the sampling. One note here is important—the "seed" value is a value usually based on the computer clock, but by setting the same seed, the same random values are regenerated, which can be useful when comparing testing on the same dataset. However, the analyst can reset the seed by pressing the "seed" button. The default of 42 is fine to start this process.



How does the analyst know if the sampling actually occurred? Use a very simple function within Rattle (like summary statistics on the "Explore" tab), and the following screen shows that there are 188 observations, which is 50% of the entire dataset. The sampling worked, and there are the appropriate number of values in the testing. If the analyst has a specific number that they need to sample (that will be discussed in a supplement), then it is relatively simple to calculate that number. The total amount of rows is 279, so take the number that is needed to sample, say 140, and divide that number by the total number of rows. That percent is what you feed into the "partition" block of the Data tab in Rattle. In this case it would be 140/279, which would be approximately 50%. However, please remember that the sampling can be adjusted to whatever sampling is needed to either increase the "power" of the statistical test (which will be discussed later), or else any other factor that would help to increase the accuracy of the statistical method or test.

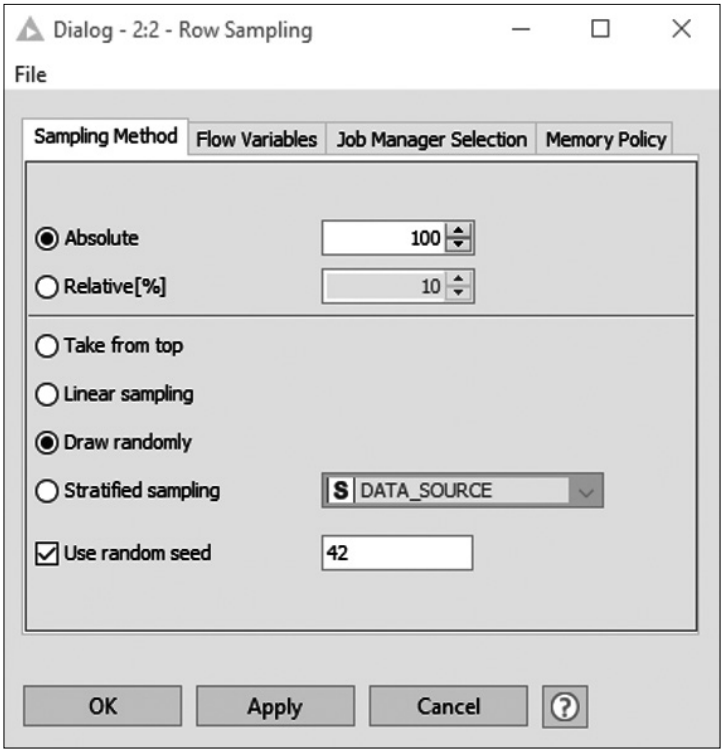


4.4.4 KNIME

KNIME has the ability to sample through, wait for it, a node for this purpose. There is one thing that needs to be explained about the nodes in KNIME. Unlike other tools, KNIME explicitly associates nodes with the rows in the dataset. This is important when using KNIME, because this is actually very accurate. When sampling is done on any dataset, if the column headers deal with variables, sampling is done with rows in mind. It is one of the areas in which KNIME is different from other tools, but that difference does not make it incorrect, just a different perspective with a different lexicon.

The first step with KNIME is dragging and placing the CSV Reader node with the imported dataset, and then dragging and placing the Row Sampling node, connecting it to the CSV Reader node. A quick reminder: each node has to be configured and executed before the process can be completed.

The next step is setting the sampling to what the analyst desires. This is done by double-clicking the Row Sampling node to reveal this screen.



Please notice that the analyst can select a percentage of the dataset to sample or an absolute number. Also notice the "Use random seed" block, which the analyst can set to the same seed that was set in Rattle. Along with that, if the Rattle partition is set to 50%, the analyst can set that percentage in this configuration screen to 50% also to see if the sampling produces a similar or different result. This is a great way to determine the different results using two different tools, both testing consistency and accuracy of the methods. In this case, the analyst sets the random seed to the same as with Rattle and the number to 188, which is the same as with Rattle. The executed node produces the following results:

TOR_LENGTH	TOR_LENGTH	0	92.6	4.874	10.944	119.769	4.446	25.933	916.3

The analyst can check this with the Rattle results to see their proximity in values. The bottom line is that the sampling methods in some of the tools are much easier, and little if any duplication is completed with random sampling in mind. Some tools have functions that are made for sampling, while others need a little more configuration. However, it is evident that sampling will continue with the data analyst for the future, since population analysis is somewhat arduous. Sampling is important and consistent if done randomly.

CHAPTER 5

STATISTICAL METHODS FOR SPECIFIC TOOLS

5.1 POWER

Power is something that an experienced undergraduate instructor in statistics would cover with some passing interest, but certainly not in any great detail. According to one reference, power is not only an option, but it should be a requirement (Reinhart, 2015). For a better understanding of power, a review of the types of errors is necessary. The focus of hypothesis testing in this book has been the Type 1 error (or false positive). By stating an "alpha" of .05, the analyst is conducting a Type 1 error, which means that there is only a 5% probability that there should be a false positive. A false positive means that the test reveals a result that may be incorrect, like a flu test result pronouncing someone with the flu that does not have it. In the power test, the Type 2 error is practiced, which is a false negative. Basically, what this means is that, if the power is 80% (or .8), there is an 80% chance of a test saying that someone does not have the flu that will in fact not have the flu. There is still a 20% chance of the false negative, or someone who was tested for flu who tested negative but really does have the flu. In the statistics world, 80% power is acceptable and conventional. The real challenge behind this is that there is a required number of events that must be sampled in order to produce this 80% power result. What is going to be demonstrated here is the process to get to that sampling result, and therefore a more accurate statistical result.

5.1.1 R/RStudio/Rattle

Excel does not have the ability to do power except by manually inputting a formula, and the same goes for OpenOffice. In order to make this as easy as possible for the analyst, this section will only focus on the tool in this text that can perform the power function straight from an existing function. This will be the R/RStudio/Rattle tool.

The first step to performing this procedure is to import the required datasets, which will be the 1951 and 1954 tornado tracking data, focusing on the TOR_LENGTH variables as in a previous section. Once this is accomplished, determine the function that will be needed; in this case it will be the "pwr" package, which can be installed like any other package in R or RStudio, as described in a previous section. Once this is completed, simply fill in the parameters of the formula with the values in order to get the missing value. For instance, if the analyst wants to know how many samples they need to have an 80% power (which is the same as having an "alpha" of .05), when they have two variables with means of 5 and 4, with a population standard deviation of 5, the analyst needs to find out how many samples they need in order to attain the 80% power. In R/RStudio, after installing the "pwr" package, the analyst needs to put the following formula into the RStudio workspace.

```
> pwr.norm.test(d=.2,sig.level=.05,power=.8,alternative=
                                                  "two.sided")

Mean power calculation for normal distribution with known
                                                      variance

          d = 0.2

          n = 196.2215

  sig.level = 0.05

      power = 0.8

alternative = two.sided
```

The reason for using "two.sided" is that the analyst does not care if one mean is less or greater than the other, just whether they are equal or not equal. The number of values needed to get an 80% power will be 197, since events are usually integers, so 196.2 is rounded up. What would happen if the analyst wanted to see if one mean was greater than the other? How many events would be needed then to get the 80% power? This is a simple change in the formula, so that the formula would now read as follows, but the alternative would be change to "greater" in order to properly address the alternative

hypothesis. The reader may remember that hypothesis testing was addressed in a previous section, and one aspect of hypothesis testing that remains consistent was that the null hypothesis is always about one value equaling the other value (such as mean1=mean2, etc.). The alternative hypothesis would be one of three: either "one value is less than the other value," "one value is greater than the other value," or "one value does not equal the other value." The "two. sided" means the third option, or that one mean does not equal the other mean. The "greater" option means that one mean is greater than the other mean. If the analyst changes the alternative to "greater," the result changes to the following:

```
> pwr.norm.test(d=.2,sig.level=.05,power=.8,alternative=
                                                    "greater")

Mean power calculation for normal distribution with known
                                                      variance

          d = 0.2
          n = 154.5639
  sig.level = 0.05
      power = 0.8
alternative = greater
```

As the analyst can see, the sample changes from 197 to 155. This means that, in order to get the 80% power, it would be less sampling effort if the alternative hypothesis is greater. At this point, the last option or "less" has not been chosen, but this is not possible with a "d" that is positive. The reason is that part of the calculation that goes into "d" is (mean1-mean2)/standard deviation. If the "d" is positive, then "less" is not an option because mean1-mean2 is positive. The analyst would have to change the "d" to a negative number in order to employ the "less" option. Spoiler alert: the value after doing this will be the same number as the "greater." The reason for this is because the analyst is testing a normal distribution (or what the analyst thinks might be a normal distribution).

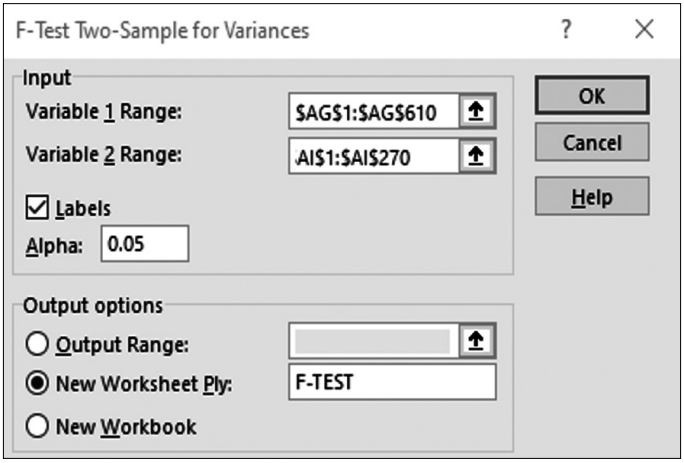
That is the R/RStudio answer to the power calculation. As one can see, it does take some effort on the part of the analyst, but it is still much simpler than performing the same function in any other tool. As such, this section will only address the R/RStudio tool for the power calculation. More information on the power tool and its importance is available by looking at the references located at the back of this book.

5.2 F-TEST

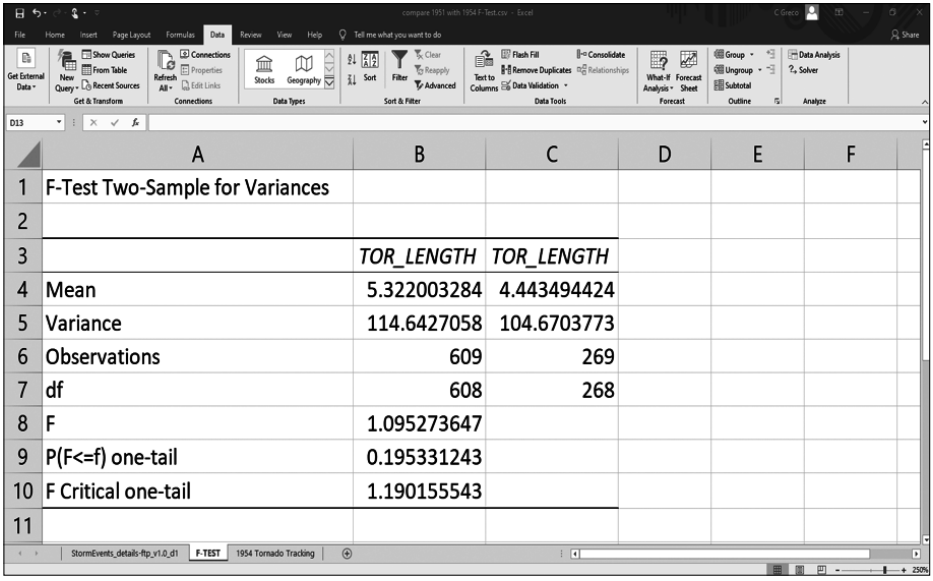
The F-Test is a way of testing whether the two variables being tested have equal or unequal variances. This is important whenever a two-sample t-test is being conducted, since the calculation to the T Statistic is different for equal or unequal variances. This test is also called the Levene Test, named after the author of an essay on this method (Levene, 1960), which detects with a conventional chance (usually 95%) whether the variances are equal between the different variables between datasets or within a dataset. Most of the tools have this function already available, but it is interesting that they do not seem to be used to check the variances prior to employing the t-test, which has slight variations in the formulas depending on whether there are equal variances or not. There is a fantastic website that can help the analyst with the Levene concept and explanation, along with formulas and tools (Technology, 2013). This is considered an out of the ordinary technique, because the analyst needs to employ this prior to conducting other tests.

5.2.1 Excel

Excel has the Analysis ToolPak, which can readily perform the Levene F-Test and actually has a selection for this in the ToolPak. The process for employing this is straightforward. The first step is to import the data, which in this case will be the 1951 and 1954 tornado tracking as was done for the t-test. After import, open the Analysis ToolPak and select F-Test Two Sample for Variances, which is just below Exponential Smoothing. At this point, please select the columns (two columns) of data to be compared, just as it appears on the following screen.



The analyst will notice that the "Labels" block is checked and that a new worksheet is being created to hold the results. For the purposes of this section, the 1951 and 1954 columns marked "TOR_LENGTH," which have been the staple of these demonstrations, is being used again for consistency. The analyst might question the reason for using any type of testing to see if the variances are different, since they are different if the analyst did a summary statistics of the datasets. However, because of the number of events in each sample, and the difference between them (279 vs over 600), making a judgement on the variances based on sight is not really adequate for statistical testing. By performing the Levene Test, the result tells the analyst what the chances are that the variances are unequal given the disparity in the sample numbers. This is important for the subsequent t-test. The result from the previous Excel Levene Test follows. What does it all mean?



	A	B	C	D	E	F
1	F-Test Two-Sample for Variances					
2						
3		*TOR_LENGTH*	*TOR_LENGTH*			
4	Mean	5.322003284	4.443494424			
5	Variance	114.6427058	104.6703773			
6	Observations	609	269			
7	df	608	268			
8	F	1.095273647				
9	P(F<=f) one-tail	0.195331243				
10	F Critical one-tail	1.190155543				
11						

The area on which the analyst will want to focus is the last three rows, which tell us if the null hypothesis (that both variances are equal) is correct. The "F" is 1.09 and the "F Critical one-tail" is 1.19. Since the F is less than the F Critical, the analyst will not reject the null hypothesis, which means that the two variances are equal. If the analyst wants to confirm, look at the "P(F<=f) one-tail," which shows a value of .193. This value is greater than the "alpha" that was set at the configuration screen, which was .05. If the p-value from the F-Test is greater than the alpha, then the null hypothesis is rejected, and

we can deduce that the two variances are equal. In this case, the p-value is greater than the alpha, pointing to a chance that the variances are equal. At this point, the analyst can then select the correct choice of t-test in order to run that procedure.

5.2.2 R/RStudio/Rattle

The Levene Test for Rattle is a relatively simple procedure. However, and this is important, Rattle only accommodates one dataset at a time. In order to do a combination of variables, the analyst will have to prepare the data prior to inserting it into Rattle, or else use the RStudio inherent programming feature. In this case, the RStudio programming feature is the choice, simply because it is just one or two lines of code.

The first step is the usual—import the data, which should already be accomplished. Then there is the necessary prerequisite step of ensuring that the proper package is installed. In this case, the var.test function is located within the STATS package, which is already installed in R and subsequently automatically installed in RStudio. The way that the analyst can find this will be explained in the supplemental information.

Once the proper package is installed and activated, it takes one line of code to ensure that both files are adequately compared. Just remember that both files have to be imported into RStudio in order for the comparison to happen.

The following lines of code and the results are included for review. Again, remember that these results may not be the same as with Excel. The main reason is that the underlying algorithm may be slightly different, but the results will be the same.

```
> tor1951<- StormEvents_details_ftp_v1_0_d1951_c20160223

> tor1954<- StormEvents_details_ftp_v1_0_d1954_c20160223

> var.test(tor1951$TOR_LENGTH,tor1954$TOR_LENGTH)

       F test to compare two variances

data: tor1951$TOR_LENGTH and tor1954$TOR_LENGTH

F = 0.91301, num df = 268, denom df = 608, p-value =
                                                    0.3907

alternative hypothesis: true ratio of variances is not
                                                  equal to 1
```

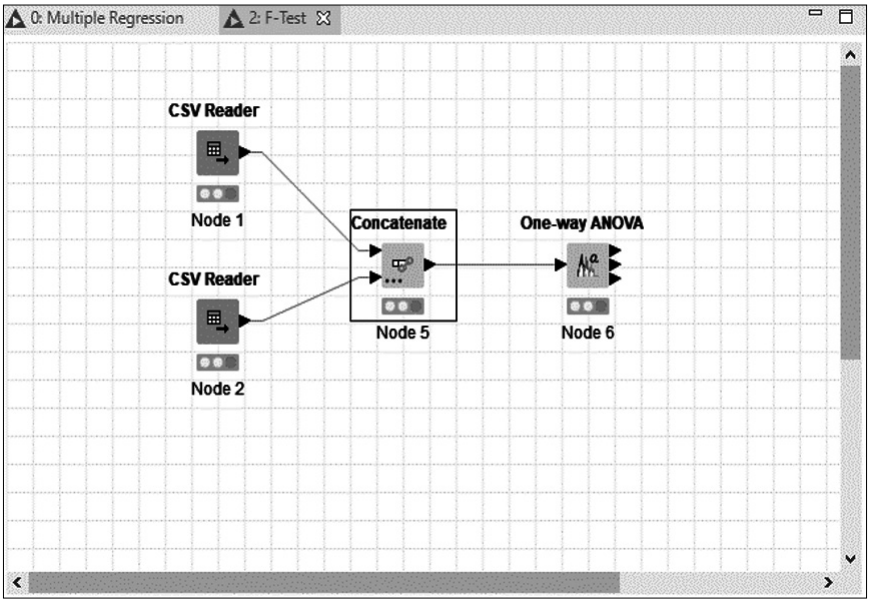
```
95 percent confidence interval:
0.7478819 1.1237249
sample estimates:
ratio of variances
      0.9130138
```

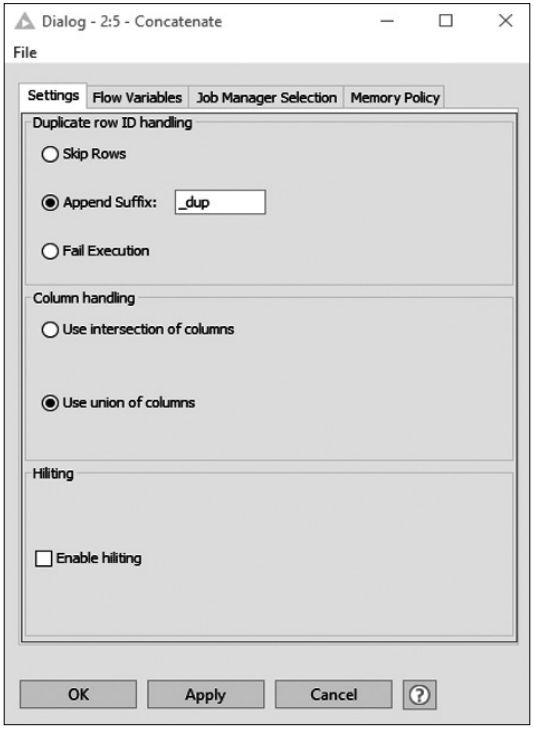
The analyst will notice that the RStudio package includes the alternative hypothesis, which is helpful. Basically, what this means is that if the p-value is less than the alpha (which, as has already been discussed, is .05), then the null hypothesis can be rejected. However, in this case the p-value is greater than the alpha value, so the null hypothesis is not rejected, which means that there is a statistical probability that the two variances are equal.

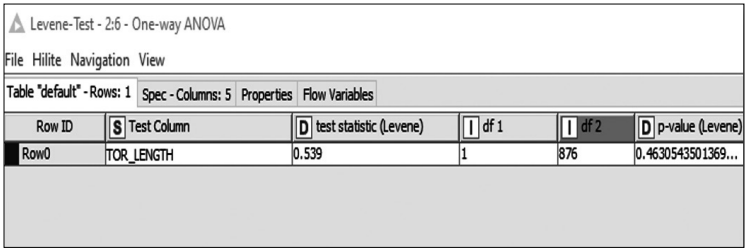
5.2.3 KNIME

KNIME has a node to perform the Levene F-Test (surprise!), but it is part of another node called the One-Way ANOVA, so doing a search on the Levene F-Test will not reveal the appropriate node. There is some preparation needed before the F-Test can be done.

The first step will be to import the two files (1951 and 1954 tornado tracking) via the *CSV Reader* node. There will be two *CSV Reader* nodes to accommodate the two files. Once that is completed, drag and connect the Concatenate node as shown in the following screen. Configure the node as shown after the workflow screen. There is something to remember about the results that are about to be revealed. They may be different from the other tools, but no worries. Again, the difference is usually because of the internal workings of the tools, and the results will be the same as to the hypothesis choice. In addition, if there is any difference between the original data and the data that is chosen for the analysis, there will be differences in the results. The one aspect of being consistent is to understand the data and ensure that all aspects of the data are the same for every test. As in experimentation, if the subjects are not the same in an aspect that is important to the test, the test will be biased.


0: Multiple Regression
2: F-Test
CSV Reader
Node 1
CSV Reader
Node 2
Concatenate
Node 5
One-way ANOVA
Node 6


Dialog - 2:5 - Concatenate
File
Settings
Flow Variables
Job Manager Selection
Memory Policy
Duplicate row ID handling
Skip Rows
Append Suffix:
_dup
Fail Execution
Column handling
Use intersection of columns
Use union of columns
Hiliting
Enable hiliting
OK
Apply
Cancel


Levene-Test - 2:6 - One-way ANOVA

File Hilite Navigation View

Table "default" - Rows: 1 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	S Test Column	D test statistic (Levene)	I df 1	I df 2	D p-value (Levene)
Row0	TOR_LENGTH	0.539	1	876	0.4630543501369...

It is important to emphasize that this result from KNIME leads to the same conclusion as the other tools. Since the p-value is .463 and the alpha is .05, the p-value is greater than the alpha, which will lead to the same result—that the two variances between the 1951 and 1954 tornado lengths have a very good probability of being equal.

5.3 MULTIPLE REGRESSION/CORRELATION

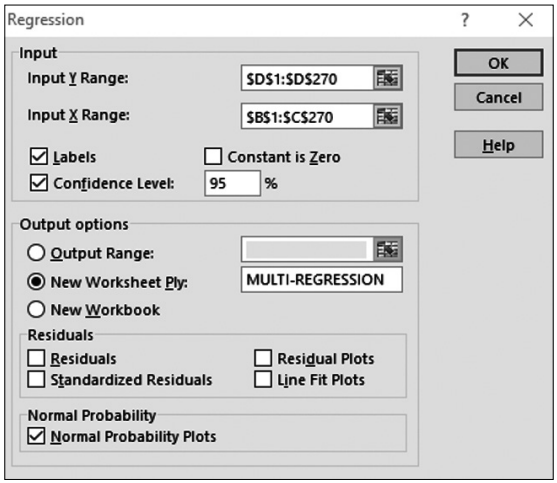
There are instances when analysis demands that several variables are tested for a relationship. In this case, some of the tools provide a direct method for performing this function. However, there is some caution when using multiple regression. According to one source, it is important to understand the consequences of multiple regression or correlation. One of these is called overfitting the data (Reinhart, 2015). In essence, in any dataset, using multiple correlation can usually result in at least one variable relating to another. The trick to this is to ensure that the analyst has the requirements prior to conducting this test, thereby reducing to eliminating this situation. There is an entire book on spurious correlations, which is the result of the analyst looking for a relationship instead of remaining unbiased. This book should be required reading for all future data analysts (Vigen, Tyler, *Spurious Correlations: Correlation Does not Equal Causation*, Hachette Books, New York, 2015.).

5.3.1 Excel

To perform either a multiple regression or correlation in Excel is simple given the Analysis ToolPak. Instead of selecting just one column for the "Y," select several columns. However, and this is important, all columns selected must be contiguous. There cannot be a column between those that the analyst wants to test. Therefore, the analyst will have to ensure that the data is properly formatted and cleaned prior to conducting this test. The procedure for performing

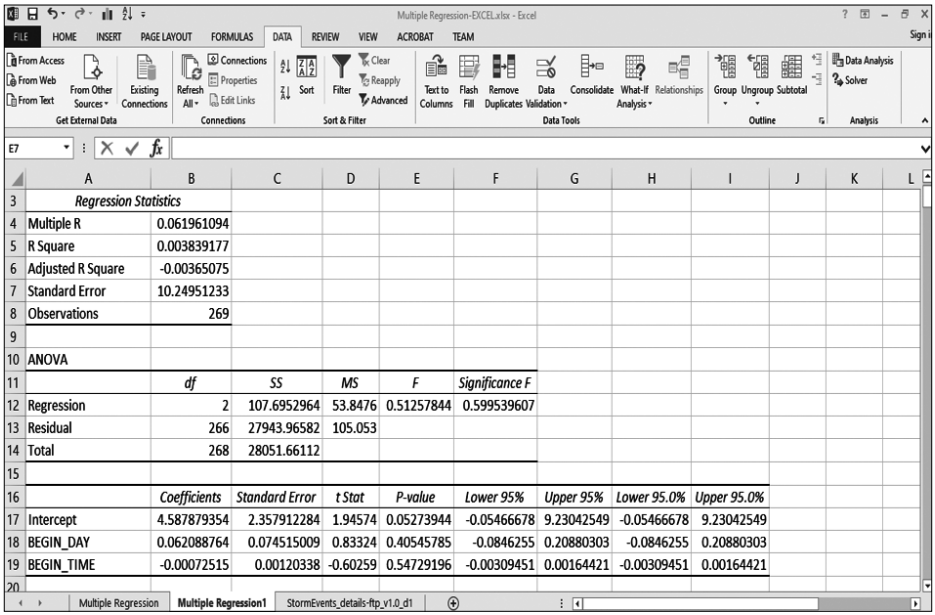
a multiple regression is to first consider the variables that the analyst will be regressing. In this case, it will be TOR_LENGTH or tornado length and BEGIN_DAY and BEGIN_TIME, which is the day and time when the tornado occurred, respectively. The analyst wants to know if they can predict the tornado length from the day and time the tornado began. In order to do this, the 1951 tornado file, which will be used in this case, has to be imported, and the Analysis ToolPak has to be loaded. There is one other item that needs to be considered. The columns being considered must be together, so the analyst will have to ensure that is completed prior to implementing the multiple regression. The next consideration is which variable to use as the dependent variable and the independent variable. The dependent variable is the "y" and the independent variable is the "x". This is important since it will determine which column to use in the function.

Once these steps are completed, the analyst can use the Analysis ToolPak and select "Regression." Once that is selected, the following screen will appear, and it's configured so that the "y-axis" or the dependent variable is the TOR_LENGTH and the "x-axis" or independent variables are the BEGIN_TIME and BEGIN_DAY. The result should be plugging in a time and day and getting an estimated tornado length based on those two variables. Please realize that this is approximate and needs to be validated through actual use. However, for the purposes of this text, this example will do nicely.



The result is the following screen. The equation shows the intercept (y) and the two x values (BEGIN_TIME and BEGIN_DAY). Although the relationship is tenuous, there is a workable formula resulting from this function. One more aspect of multiple regression is the correlation, which is very low

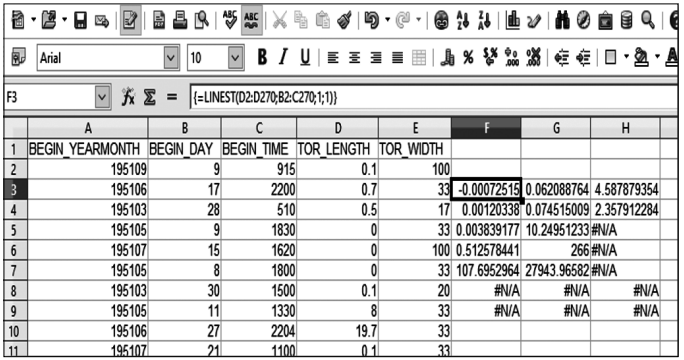
to the point of nonexistence. If the analyst is trying to predict tornado length from the two independent variables, it will produce a result, but the association of these two variables is tenuous with tornado length.



	A	B	C	D	E	F	G	H	I	J	K	L
3	*Regression Statistics*											
4	Multiple R	0.061961094										
5	R Square	0.003839177										
6	Adjusted R Square	-0.00365075										
7	Standard Error	10.24951233										
8	Observations	269										
9												
10	ANOVA											
11		*df*	*SS*	*MS*	*F*	*Significance F*						
12	Regression	2	107.6952964	53.8476	0.51257844	0.599539607						
13	Residual	266	27943.96582	105.053								
14	Total	268	28051.66112									
15												
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*			
17	Intercept	4.587879354	2.357912284	1.94574	0.05273944	-0.05466678	9.23042549	-0.05466678	9.23042549			
18	BEGIN_DAY	0.062088764	0.074515009	0.83324	0.40545785	-0.0846255	0.20880303	-0.0846255	0.20880303			
19	BEGIN_TIME	-0.00072515	0.00120338	-0.60259	0.54729196	-0.00309451	0.00164421	-0.00309451	0.00164421			
20												

5.3.2 OpenOffice

As with other functions, OpenOffice does not have the Analysis ToolPak to efficiently produce a result, but it can do multiple regression based on formulas. After importing the same data as used with Excel, the "linest" formula is used with contiguous columns, after which the formula is transformed into an array formula by using CTRL-SHIFT-ENTER, which produces the following result:



	A	B	C	D	E	F	G	H
1	BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	TOR_LENGTH	TOR_WIDTH			
2	195109	9	915	0.1	100			
3	195106	17	2200	0.7	33	-0.00072515	0.062088764	4.587879354
4	195103	28	510	0.5	17	0.00120338	0.074515009	2.357912284
5	195105	9	1830	0	33	0.003839177	10.24951233	#N/A
6	195107	15	1620	0	100	0.512578441	266	#N/A
7	195105	8	1800	0	33	107.6952964	27943.96582	#N/A
8	195103	30	1500	0.1	20	#N/A	#N/A	#N/A
9	195105	11	1330	8	33	#N/A	#N/A	#N/A
10	195106	27	2204	19.7	33			
11	195107	21	1100	0.1	33			

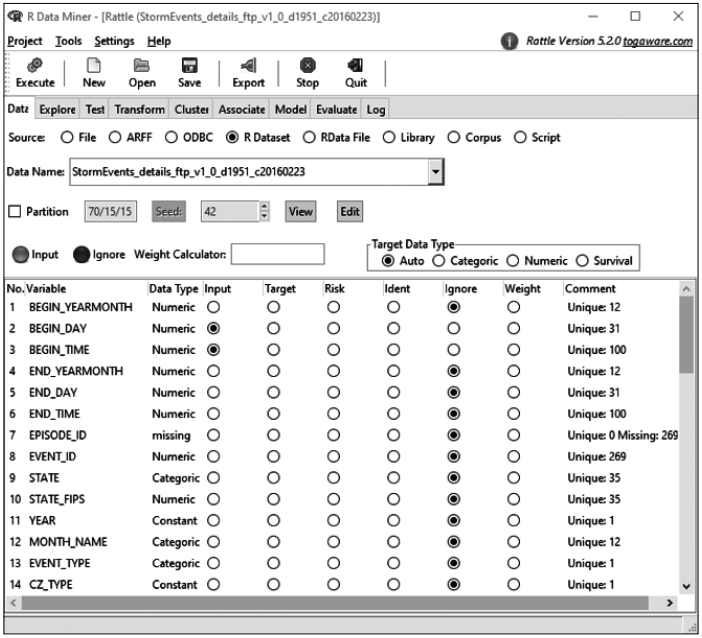
The way to read this result is to look at F3 (selected), and that is the same as the number in BEGIN_TIME in the Excel readout. In other words, it is one of the "x" values. The other "x" value, BEGIN_DAY, is located in G3, and the intercept is located at H3. Basically, what this means is the formula would read as:

$-0.00072515x+.062088764x_2+4.587879354$.

What this means is if an analyst wants to know what the tornado length would be on the 9th of a month at 0900, then the analyst would plug these numbers into "x" and "x_2" and add the intercept to get the tornado length. To reiterate, this is just an example and does not show a relationship between these factors. This is for demonstration only.

5.3.3 R/RStudio/Rattle

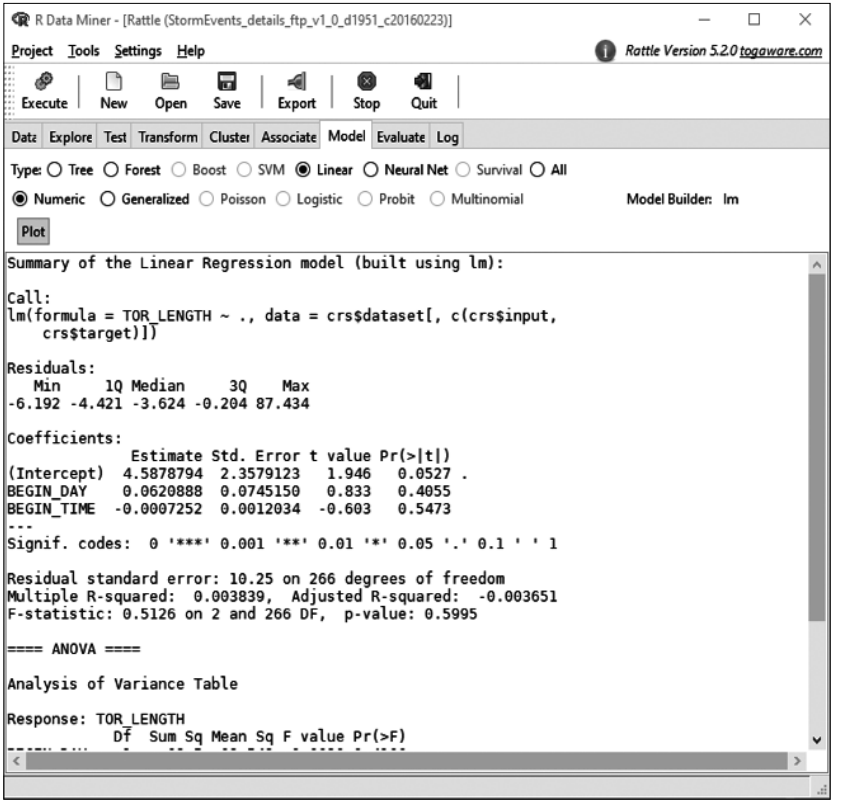
In multiple regression, Rattle is a good choice for this function because it is one available within the package. The configuration is the same as in other sections, the first step being to import and assign the particular variables the appropriate designator. As shown in the following screen, there has to be a "target" variable assigned, otherwise the regression function will not work. In this case, the target function will be the tornado length or TOR_LENGTH variable, since that is the one that will be the dependent variable. The other factors, the time and day, will be the independent variables, similar to the previous configuration in OpenOffice. After this is completed, click on the "Execute" icon and the following result will appear.



Once the data is configured, move to the "Model" tab in order to perform the regression. The screen is configured just like the one as follows and, once the "Execute" icon is pressed, the results will show as illustrated.

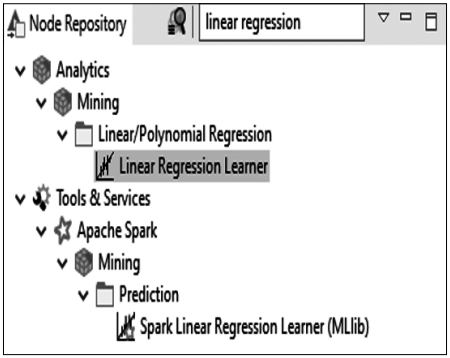
The entire screen can be somewhat overwhelming, but the results are very similar to the previous sections, in that the numbers across from BEGIN_TIME and BEGIN_DAY are the same as those provided by other tools, and reflect most closely the Excel readout.

One word of caution from configuring these screens. It was noted in previous Rattle sections that the analyst must pay attention to the data options to ensure that the dataset will include all the rows. Remember the "Partition" options? This is important, since performing any function in Rattle will produce different results with different settings in the Partition option. If the analyst is just using Rattle, there are preparation steps that are not just important, but vital in order to ensure consistent results.



5.3.4 KNIME

The multiple regression node located in KNIME is not immediately visible. The node is located at the location in this screen. Again, placing "regression" in the search block will identify the location of the node as follows:



The KNIME multiple regression is relatively straightforward. The Linear Regression Learner node is placed and connected to the CSV Reader node. By double clicking the Linear Regression Learner node, a screen will appear that will allow the user to choose which is the target column and which are the "independent" columns. This is done like other nodes of this type.

Once the node is configured, execute both nodes and after the user receives a "green light," then right click on the Linear Regression Learner node and choose the coefficients and statistics table to see the results. The user should use the same method of reading these results as described in previous sections.

It is important to remember that every node that the analyst needs is probably available in KNIME, but sometimes it takes some searching in order to get those nodes. One additional note is that there are plenty of community communications available if an analyst needs assistance in KNIME. In some cases, these communities include actual processes that are complete and available for download in order to test and see how the process works. Researching these nodes is practical and contributes to the learning process with these tools. Along with research is practice, which is extremely valuable.

5.4 BENFORD'S LAW

Benford's Law was developed to detect anomalies in numeric data, namely accounting inputs. The basic theory behind it is that numbers that are "normally distributed" reflect a descending curve from "1" to "9." By implementing Benford's Law, the analyst can detect if there are irregularities in numbers, which could lead to revealing fraudulent submissions. This is used by accountants and financial analysts to help curb fraud and accounting issues (Statistical Consultants Limited, 2011). The one tool that seemed to accomplish this with little effort was Rattle, since it has it as an option in the functions provided with the tool. One note of caution here is that Rattle (and R in general) works on packages, which means that there are times when there will be a package needed in order to complete a function within Rattle. When this happens, Rattle will tell the analyst that a package is needed and ask if the analyst wants the package installed. If the analyst picks the no option, the package will not be installed, and the process will end. The one point about analysts who depend on Rattle is that they trust the tool to download items and not install malware or other memory hungry files. In the years using Rattle, this has not happened to this author, but there is no guarantee on the 100% safety issue with this tool. However, the same can be said for more well-known tools that are trusted and used by companies and the federal government that have installed files that hackers used for malware. In every case, it would be wise for any data scientist to activate and continue any antivirus software that they have installed on their computer.

5.4.1 Rattle

The configuration of Benford's Law for Rattle is a little complicated, but it is still very robust compared to using formulas in some of the other tools. Rattle places Benford's Law in the "Explore" tab under "Distributions." When the analyst selects the Distributions radio button, the following screen is revealed and selecting Benford's Law is as easy as checking a box. However, there is some preparation that accompanies that check.

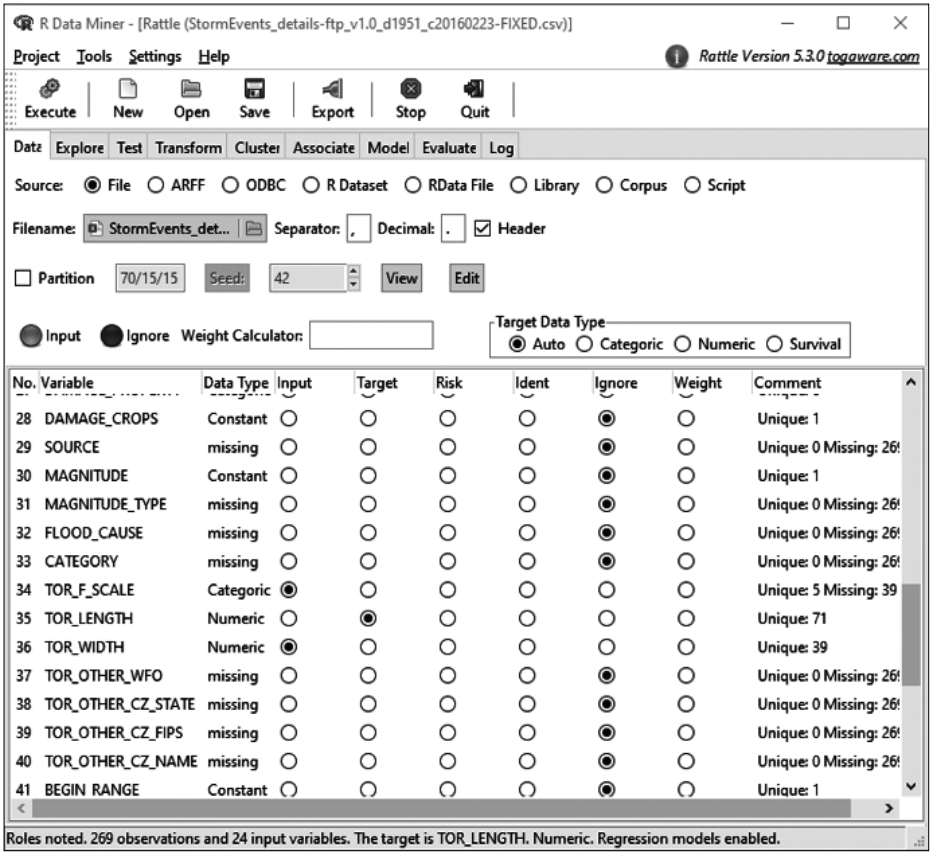
First, go back to the "Data" tab and ensure that TOR_LENGTH is selected as the "Target," since that is the variable that the analyst wants to use to see if it conforms to Benford's Law. Ensure that, once TOR_LENGTH is chosen, the analyst clicks on the "Execute" icon to activate that within the dataset. Also remember that there is no need to "ignore" all the other variables, since the target is the one that will be the primary variable considered under the

Benford's Law button. One reminder is that the "Explore" tab has a wide array of functions that are available for data analytics, so please attempt these different combinations to see if there is a function that will fit the analyst's need.

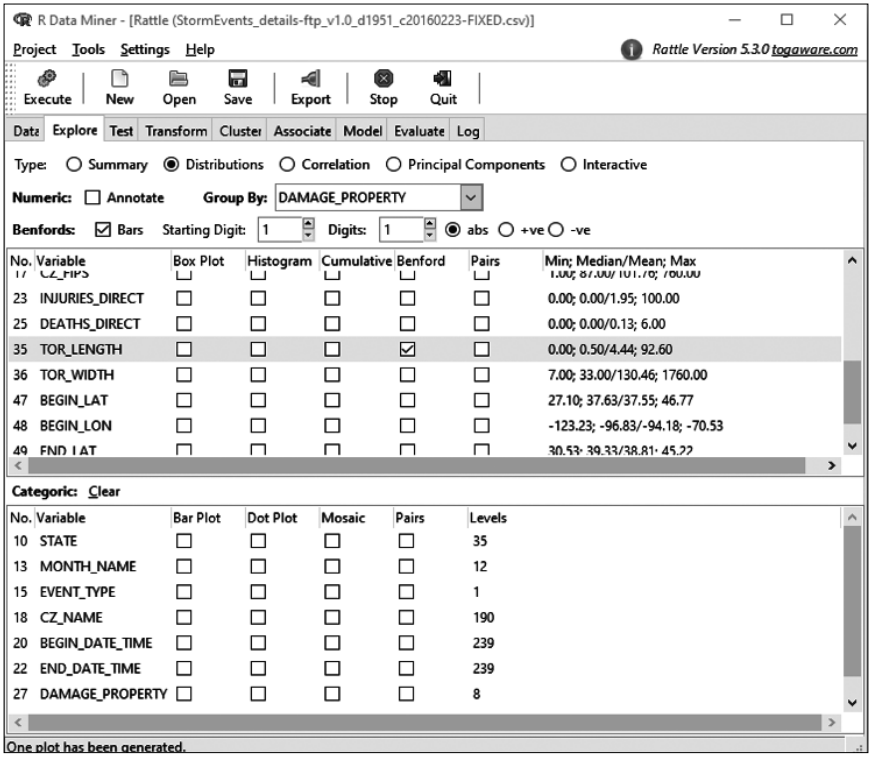
Once the data is appropriately selected as shown in the following, the next step will be to ensure that the configuration of the "Explore" tab is correct for the function to work. A word of warning here is to ensure that the dataset is appropriate for the function. When the analyst selects the data, there may be a temptation to use the "R Dataset" option within the "Data" tab. Although this would be fine for many of the functions, the Benford's Law option needs to be reading a "data frame," which is not the type of dataset resulting from the "R Dataset" choice. The dataset produced by that choice is called a "tibble," which is a type of dataset very flexible with many of the packages available within R and Rattle. However, it is not compatible with the Benford's Law function. Therefore, instead of using the "R Dataset" option, it is best to choose "File" as the source. In this way, by using the file directly from the computer, there is no transformation from R to make it into a tibble. The previous statement is just a suggestion, since there are commands that can change a tibble to a "data frame" right in R; but if programming is not preferable, then importing a regular computer file is the right option.

Once the data has been selected and imported, there is a need to make a variable a "target," and in this case TOR_LENGTH has been chosen. This will ensure that the appropriate function will recognize and focus on TOR_LENGTH as the factor to be considered. The following screen shows the appropriate choices. One note is that the "partition" checkbox is unchecked. In this case, all rows will be considered in this function, but as in the previous section on "training" datasets, the analyst can choose to determine what percentage is needed (sampling) to do the test and then validate it with another part of the entire dataset.

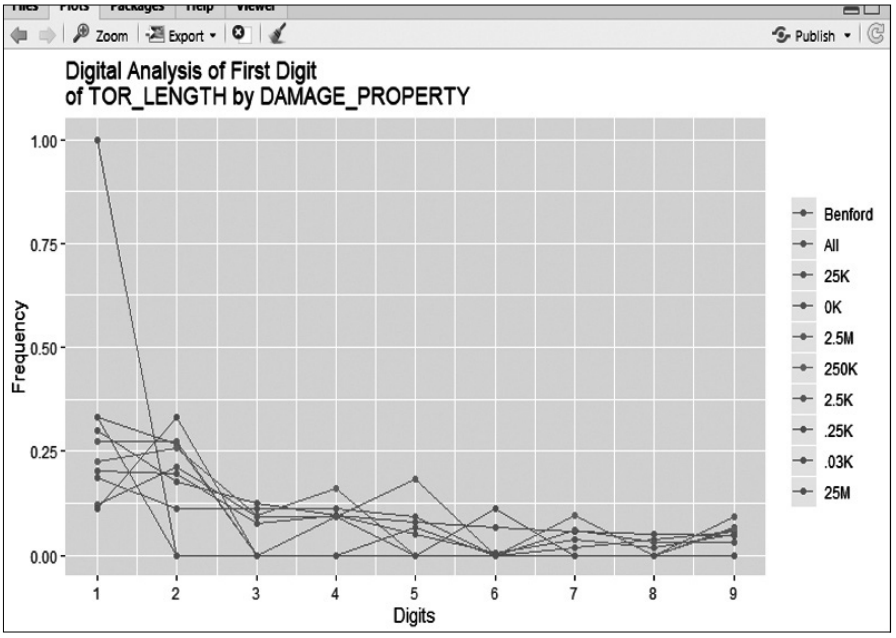
Once the dataset is imported and the data is configured, it is time to go to the "Explore" tab and process the choices necessary to activate the Benford's Law functionality within Rattle.



The first step would be to select "Distribution" from the options, and one window with two screens will appear. For now, the top one will be the focus of this section. Choose a variable for the Benford's Law option and then ensure that the "group by" choice is a variable that will show a valid association. In this case DAMAGE_PROPERTY was shown as the categorical variable. The choices should appear as follows:



Once the choices are made, click on the Execute icon and check the RStudio plot screen which is, as a default, located at the bottom right quarter of the screen. There the analyst should see the following screen which, as a warning, could seem very complicated. The main line on which the analyst should focus is the one marked "Benford," which shows the probability that the first digits will appear in normal data. For instance, looking at the Benford line (red), the "1" digit appears around .30 or 30% of the time. If the analyst is looking at the "red" dot that appears at the top of the graph, this is not the Benford line but one that is depicting the number of "1" digits that appear in TOR_LENGTH when addressing 25M or 25 million dollars of damage. However, with other DAMAGE_PROPERTY figures such as 2.5M or 25K, the line is somewhat close to the Benford line. This means that there is some similarity between those figures and the normality of the Benford Law.



However, even though the graph may not be discriminating, there is a program in R/RStudio that gives a judgement on whether Benford's Law is adhered to by the data. The programming line for this is shown as follows as it would appear in R:

```
> benford(tor1951$TOR_LENGTH,number.of.digits=1,sign=
                          "positive",discrete=FALSE,round=3)
```

From this line, the following result will appear. As the analyst can see, it judges the data as nonconforming to Benford's Law, but at the end qualifies that statement with stating that no real-world data will totally conform to Benford's Law. This is important, since reflection of real data to a theory is not a realistic outcome.

```
Benford object:

Data: tor1951$TOR_LENGTH
Number of observations used = 157
Number of obs. for second order = 69
First digits analysed = 1
```

```
Mantissa:

  Statistic Value
    Mean 0.476
     Var 0.089
 Ex.Kurtosis -1.098
  Skewness -0.112


The 5 largest deviations:

digits absolute.diff
1    5    13.57
2    6    9.51
3    1    9.26
4    2    8.35
5    7    2.10

Stats:

      Pearson's Chi-squared test

data: tor1951$TOR_LENGTH
X-squared = 28.47, df = 8, p-value = 0.0003927


      Mantissa Arc Test

data: tor1951$TOR_LENGTH
L2 = 0.0039588, df = 2, p-value = 0.5371

Mean Absolute Deviation (MAD): 0.03218442
MAD Conformity - Nigrini (2012): Nonconformity
Distortion Factor: -34.24091

Remember: Real data will never conform perfectly to
         Benford's Law. You should not focus on p-values!
```

This shows that this function can be done with relative ease from this tool, with no additional programming necessary. One additional comment is that, in order for this to work, the analyst may have to install the "benford.analysis"

package that is part of R but is not automatically installed with the base R or RStudio.

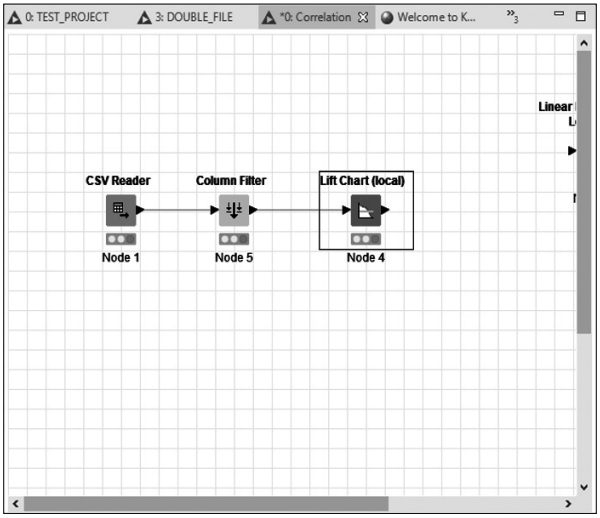
5.5 LIFT

Lift is a method of evaluating a predictive model. Many times, the analyst will conduct a model or test without evaluating the potential value of such a test. In this case, Lift can evaluate the predictive value before running the actual test. Think of the value of such a function. If the test is not valuable or not appropriate, why run the test? In the book *Data Science for Business* (found in the reference section) the authors present an example of someone going into a store and buying a combination of products. The lift will determine the feasibility of predicting whether someone buying a specific combination of those products, whether it be beer and eggs, or beer and potato chips. This is based on a probability of buying one product, and then the other product. The formula is included in the book, which this author would highly recommend every analyst read, especially if they are part of a large company that makes its revenue producing and selling products, specifically consumables (Provost, 2013). What this has done is show the feasibility of predicting the purchase of a combination of products. The best tool of the ones described in this text to perform the lift function is KNIME, since it has a node to perform this method.

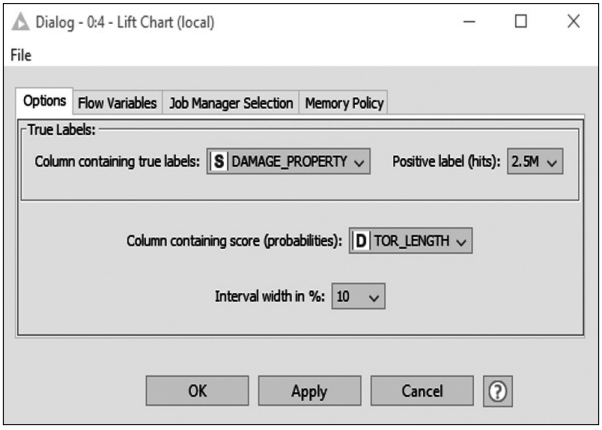
5.5.1 KNIME

As with many other functions, KNIME has a node for calculating lift, which the analyst can find through the search bar as shown in the following screen. Please remember that the analyst does not have to type the entire node name, since KNIME will search as the analyst is typing.

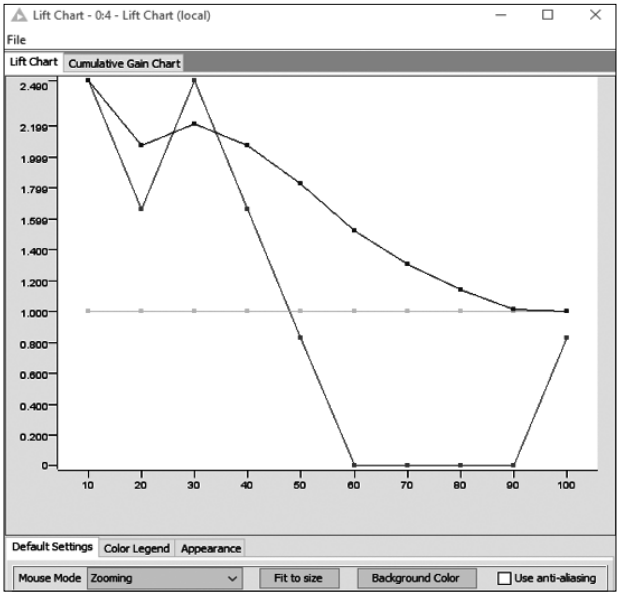
Once the data is imported through the *CSV Reader* and the Lift Chart is found, dragged, placed, and connected to the *CSV Reader*, the next step is to configure the Lift Chart node. The following screen is one configuration of this screen showing TOR_LENGTH with DAMAGE_PROPERTY as the predictor value. This means that the method is evaluating the potential of predicting tornado length from the damage done to the property by that tornado.

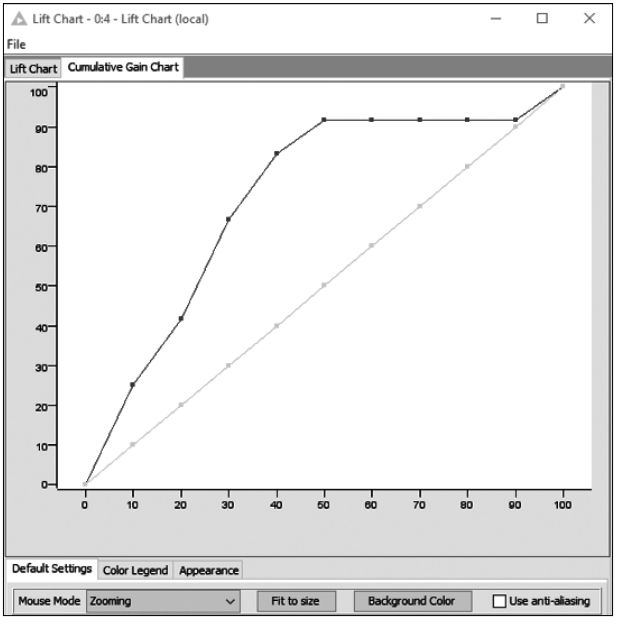


The following screen shows the configuration for the Lift Chart (local) node. As the analyst can see, DAMAGE_PROPERTY is going to be set against TOR_LENGTH to see if this will make a good predictive model. The user has to set "Positive label (hits)" specifically to the category of 2.5M (or 2.5 million dollars of damage) to see if it is worthwhile to have a predictive model against this figure. The analyst can use the down arrow to choose other damage amounts, but this one should be predictive against tornado lengths, showing an association between damage and lengths. The result, once executed, is below this screen. This node allows for both the lift chart and cumulative gain chart, both of which are useful to the analyst. The lift chart shows great distances between the measurement (red) and the baseline (green), which is a good indicator of prediction. Also, the cumulative gain chart shows the line rising above the baseline throughout the chart, which is also a good indicator of prediction.



The following screens are the results of the lift function, as stated previously. Please notice that there are some formatting options including legend colors and other specific options. Please explore these since it is always helpful to understand how these options may change the appearance of the chart and enhance the analyst's and the recipient's experience with this tool.





Once reviewing these charts, the analyst will know if it is worthwhile to process a prediction model against these two variables. One note of caution at this point. The analyst has only picked one category of damage. It might be worthwhile to check other damage amounts to see if there is any use to associate these two factors in a model.

Also, the analyst might notice that there is a node in between the CSV Reader node and the Lift Chart node. That node was placed in the process to limit the number of columns used for the Lift Chart. It is vital that only the factors being considered are available; otherwise, there is a slight chance that KNIME might try to include other factors into the mix, thinking that categorical characteristics are open. This has happened with some nodes, but the way to counter that is to only use those columns that apply, or stick to the nodes marked in parentheses (local), since those are the ones that seem to offer the basics but also seem the most stable. This is, of course, this author's opinion.

5.6 WORDCLOUD

Sometimes the analyst receives data that is all text and wonders how to transform the words into numbers for analysis. It is fortunate that the analyst no longer has to concern themselves with this transformation. Thanks to algorithms and research by other statisticians and analysts, there is now a function to take words and analyze those words for the most and least used word. Although this seems perfunctory, the function provides the analyst and the recipient of the analyzed data with a one-screen visual of the "corpus" or words in text. It is this that will be discussed next, specifically with the tools that allow this analysis to be completed with the least amount of arduous programming or functional steps.

5.6.1 R/RStudio

The R/RStudio combination allows for the quickest method to analyze words in text. In this case, the data will be a little different than in past sections. The data imported will be from the 1995 tornado tracking (details) data from the link described back in the first few sections of the book. Once the data is opened, delete all the columns except for the column marked "Event Narrative." The analyst will use this data as text to extract words that may present some patterns valuable to the analysis.

As a side note, there are some packages that are necessary in order for the wordcloud to function. Some are installed as part of the wordcloud package, but you may need to install the "tm" and "RColorBrewer" packages in order to produce a color visual. If an analyst wants to see what a wordcloud looks like, there are plenty of sites available to either view one or actually do small ones free of charge. Just place "wordcloud" in a search engine and there are plenty of examples available for viewing. There is also an example at the end of this section.

Once the narrative is isolated, copy all of the text and save the text as a ".txt" file in order to alleviate any extraneous characters that might be included as part of a word processing configuration. The ".txt" file is relatively simple and has very little extraneous characters to fog up the analysis.

Name the file "textanalysis.txt" in order to simplify the identification and import the file to R/RStudio. However, this time, to import the file as a text file will need some programming commands in order for the import to be functional. The following commands are taken from an analytics website that specializes in R functions (Sankhar, 2018).

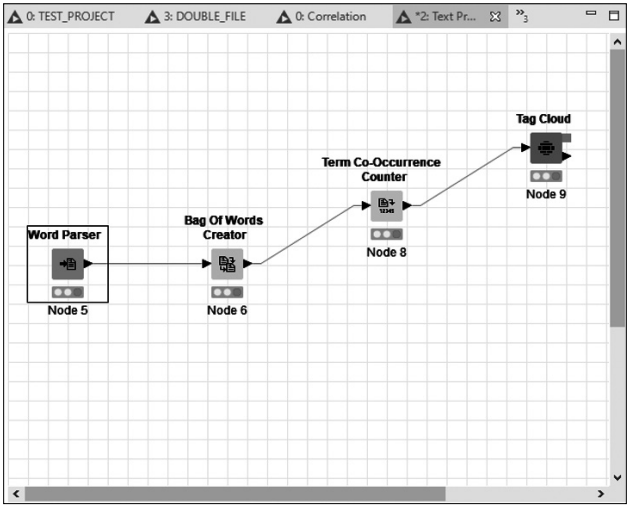
```
> library(tm)
> library(RColorBrewer)
> setwd("C:/")
> speech = "CORPUS/textanalysis.txt"
> library(wordcloud)
> speech_clean<-readLines(speech)
> wordcloud(speech_clean)
```

The previous programming is required in order for the wordcloud function to operate properly. The first two lines load R packages that help with the text cleaning and enhance the wordcloud package. The third line sets the working directory so that the entire file location (which can be a long line) can be shortened. The fourth line sets the variable "speech" to the text file. The fifth line opens the wordcloud package, while the sixth line uses the "readLines" function to read each line of the "speech" text and uses the variable "speech_clean" to store the result. The last line activates the wordcloud function on the final text. The result of the last line is illustrated as follows:

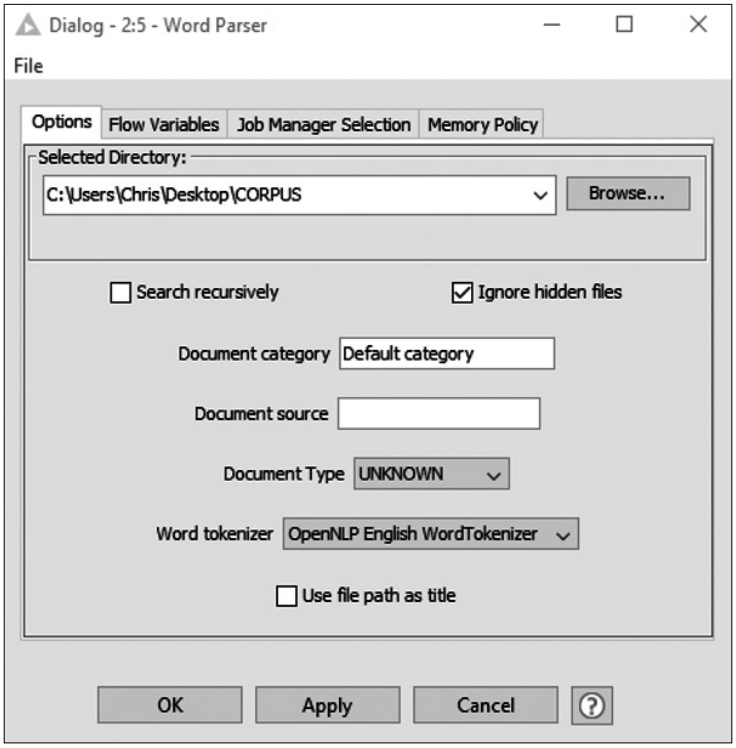
The previous wordcloud shows the words that appear the most in the narrative as the largest and those that appear less as smaller and smaller text. From this, it looks like "alabaster" appears more than "winston" or "scottsdale." What this may mean, since these are county names, is that one county may experience more severe storms than another. This can be done without any numerical analysis. This is not to say that this may answer all the questions the analyst may have, but it may put some light on an otherwise obscure dataset in another instance. Employing text analysis can help to reduce some confusion about the data.

5.6.2 KNIME

KNIME also has a wordcloud function in the node library called "tag cloud," which is illustrated in the workflow screen. Note the different nodes, since each one will be described as it appears left to right in the workflow.

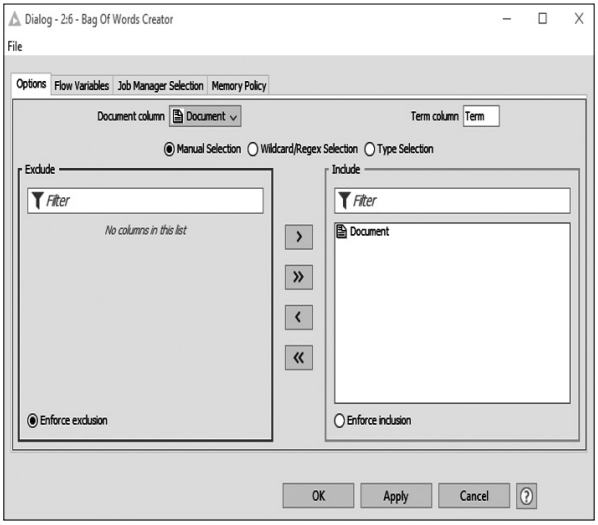


The first one will be the "Word Parser" node, which takes Microsoft Word documents and prepares them for text analysis. The configuration for this screen is as follows and shows the folder location to be analyzed. One warning is that this is not the file but the folder. The node will search the folder for Word files and use those for analysis.



The choice of "Document Type" is unknown, but there are several choices for this down arrow, including book and proceeding. It is up to the analyst which one they choose to analyze. The unknown choice fits well for this example. The "Word Tokenizer" is the default, and again the analyst can choose between a number of these types of parsing functions. It would benefit the analyst to try a number of these to see if they make a difference in the text mining. For this example, the default is chosen.

The next node in the workflow is the "Bag of Words Creator" node, which splits the text into words and uses numerical indicators to sum up how many times the word occurs in a sentence or group of sentences. To see this clearly, the configuration screen is as follows (first tab).



There is only one column included in the analysis called "Document," and the "Term" column is named "Term" by default. This is important since the next nodes will rely on the result of this node for further analysis. Once that is finished, the result of that node looks similar to this:

Documents output table - 2:6 - Bag Of Words Creator

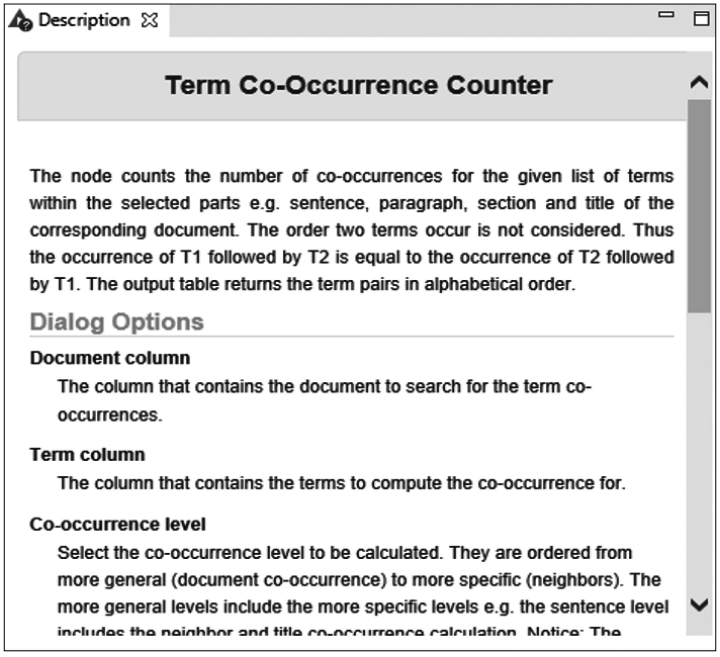
File Hilite Navigation View

Table "default" - Rows: 532 | Spec - Columns: 2 | Properties | Flow Variables

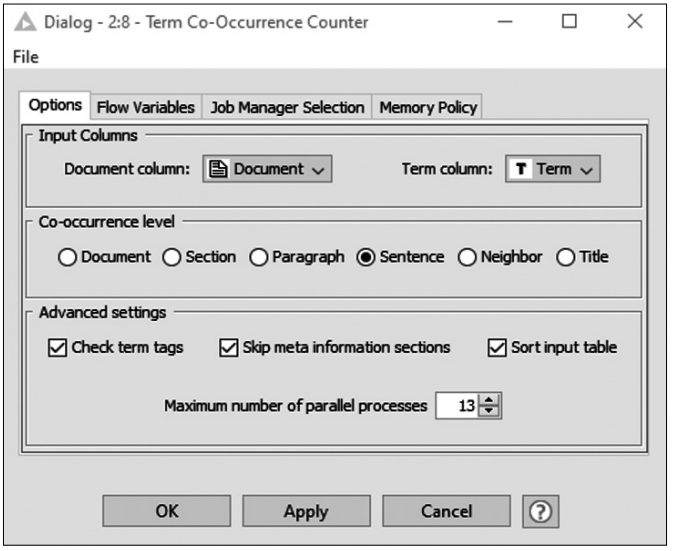
Row ID	Document	Term
Row0	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*[]
Row1	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Pinson[]
Row2	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	,03,1638CS...
Row3	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*,,1645CST[]
Row4	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	,[]
Row5	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	?[]
Row6	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	?,[]
Row7	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Hail[]
Row8	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	([]
Row9	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	1.75[]
Row10	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	)[]
Row11	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Three-quart...
Row12	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	inch[]
Row13	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	hail[]
Row14	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	was[]
Row15	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	reported[]
Row16	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	at[]
Row17	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Oxmoor[]
Row18	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Valley[]
Row19	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Golf[]
Row20	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Club[]
Row21	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	in[]
Row22	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Homewood[]
Row23	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	.[]
Row24	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	One[]
Row25	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	the[]
Row26	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Trace[]
Row27	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Crossings[]
Row28	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	area[]
Row29	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	of[]
Row30	"Pinson,03,1638CST- *,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Hoover[]

The node has split the sentences by each word and placed that word as a row. This would take many hours if done by hand, but this function makes it seem easy. The next node will take these words and count the number of times they occur with another one of the words in the text. This is used for some advanced analysis, such as how words are used in the combination of other words and such.

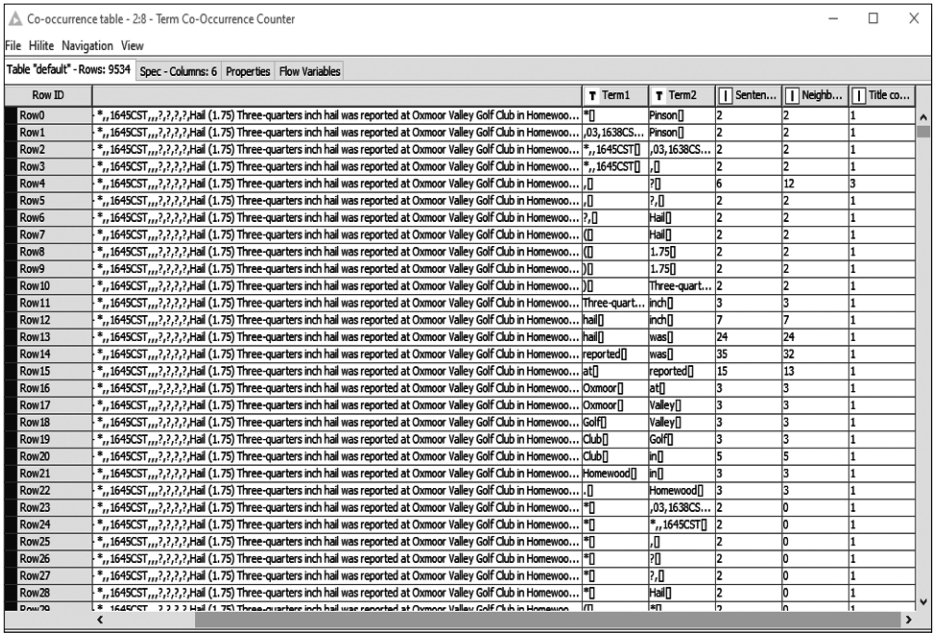
This node is called "Term Co-Occurrence Counter," and the description of this node, as with all nodes, appears when the analyst clicks on the node, as shown in the following. Usually these descriptions are enough for the analyst to know if the node will be useful in the workflow or something that might be useful in later processes. This is one part of KNIME that makes it very analyst friendly, in that every node is accompanied with a description.



The configuration screen for the node is depicted as follows. The analyst can choose a number of options for this node, and the ones chosen for this example work fine with the dataset.

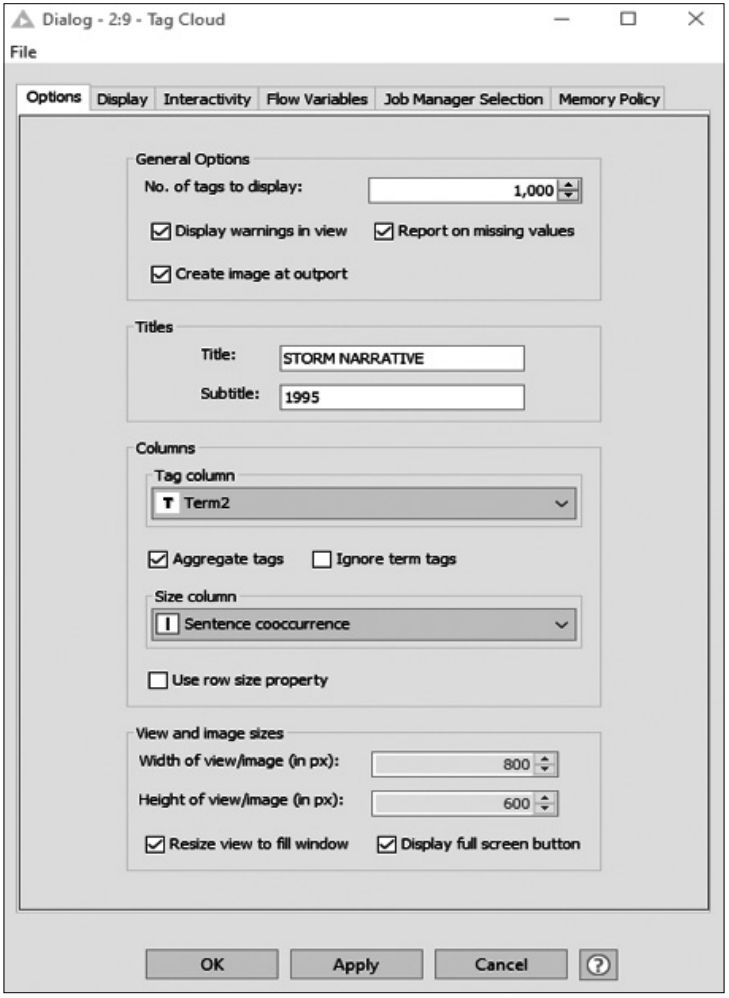


Notice that "Co-occurrence level" is set at "Sentence," but there are other choices that the analyze can pick, so please explore these nodes to see if there is a combination that provides the complete results that are needed for the analysis. The result from this node is the occurrence of the words with other words in sentences throughout the text. The table is shown as follows:

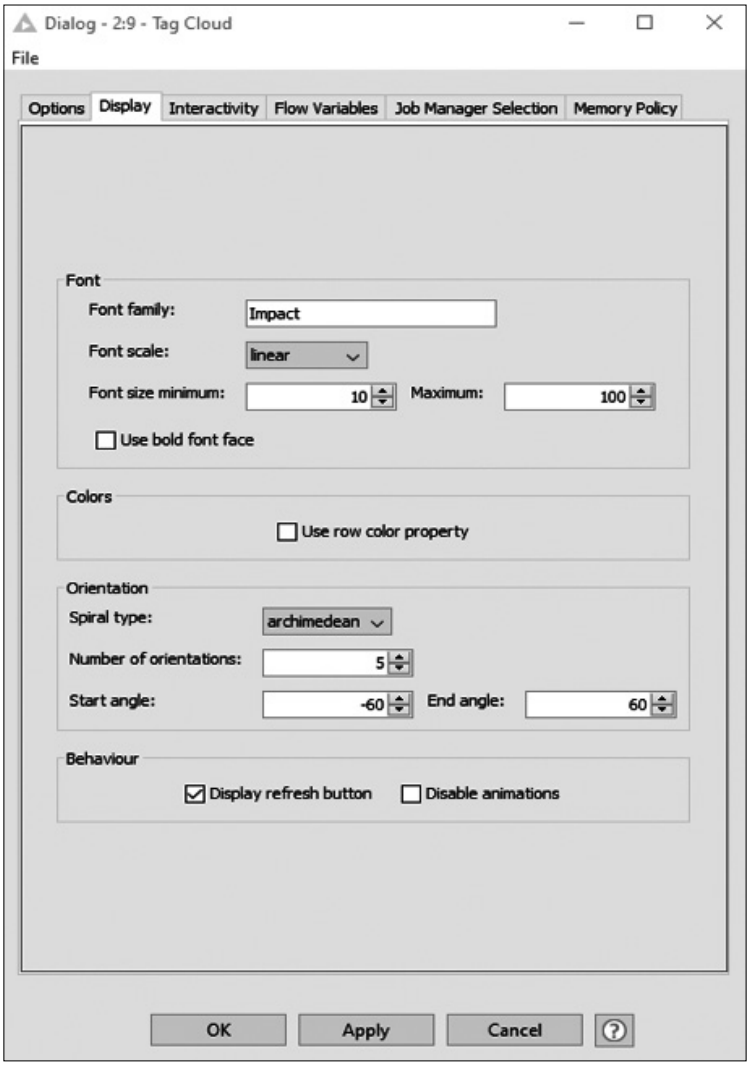


Row ID		Term1	Term2	Senten...	Neighb...	Title co...
Row0	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*[]	Pinson[]	2	2	1
Row1	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	,03,1638CS...	Pinson[]	2	2	1
Row2	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*,,1645CST[]	,03,1638CS...	2	2	1
Row3	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*,,1645CST[]	,[]	2	2	1
Row4	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	,[]	?[]	6	12	3
Row5	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	,[]	?,[]	2	2	1
Row6	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	?,[]	Hail[]	2	2	1
Row7	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	([]	Hail[]	2	2	1
Row8	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	([]	1.75[]	2	2	1
Row9	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	)[]	1.75[]	2	2	1
Row10	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	)[]	Three-quart...	2	2	1
Row11	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Three-quart...	inch[]	3	3	1
Row12	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	hail[]	inch[]	7	7	1
Row13	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	hail[]	was[]	24	24	1
Row14	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	reported[]	was[]	35	32	1
Row15	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	at[]	reported[]	15	13	1
Row16	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Oxmoor[]	at[]	3	3	1
Row17	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Oxmoor[]	Valley[]	3	3	1
Row18	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Golf[]	Valley[]	3	3	1
Row19	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Club[]	Golf[]	3	3	1
Row20	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Club[]	in[]	5	5	1
Row21	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	Homewood[]	in[]	3	3	1
Row22	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	.[]	Homewood[]	3	3	1
Row23	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*[]	,03,1638CS...	2	0	1
Row24	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*[]	*,,1645CST[]	2	0	1
Row25	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*[]	,[]	2	0	1
Row26	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*[]	?[]	2	0	1
Row27	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*[]	?,[]	2	0	1
Row28	*,,1645CST,,,?,?,?,?,Hail (1.75) Three-quarters inch hail was reported at Oxmoor Valley Golf Club in Homewoo...	*[]	Hail[]	2	0	1

Although interesting, the table is not that useful, but visually it would help the analyst determine those words that occur the most and those that occur the least. This visual is provided by the last node called the "Tag Cloud" which, according to the description, is the same code as provided on a site called "Wordle," which the analyst can see on the web page *www.wordle.net* and was developed by Jonathan Feinberg (as shown in credits on the web page). The Tag Cloud node configuration screen is shown, along with the configuration set for this example. Again, explore the different options for all these nodes, since they can produce some very interesting visual presentations for use in the data analysis.

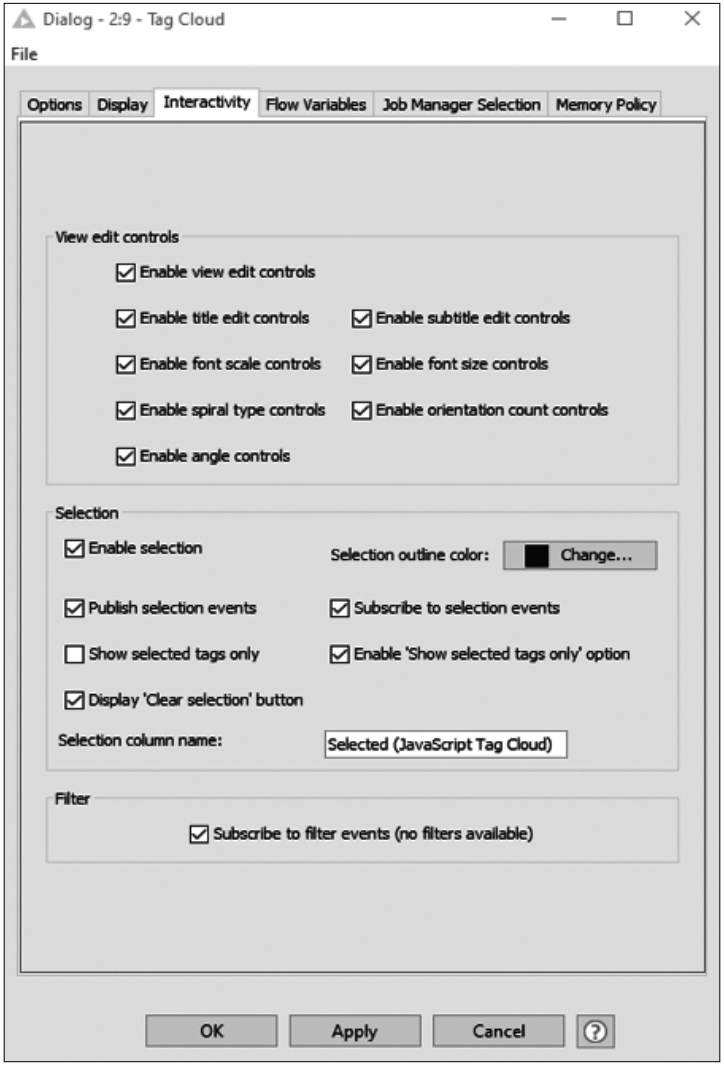


In this tab of the configuration screen, the analyst has set the title and subtitle of the visual, along with the tag column and size column. Size Column will determine the size of the word based on the occurrence or co-occurrence. The analyst could use other columns for this purpose and get a different result. For this example, these settings will perform the function. The second tab or "Display" is next with the following configuration.

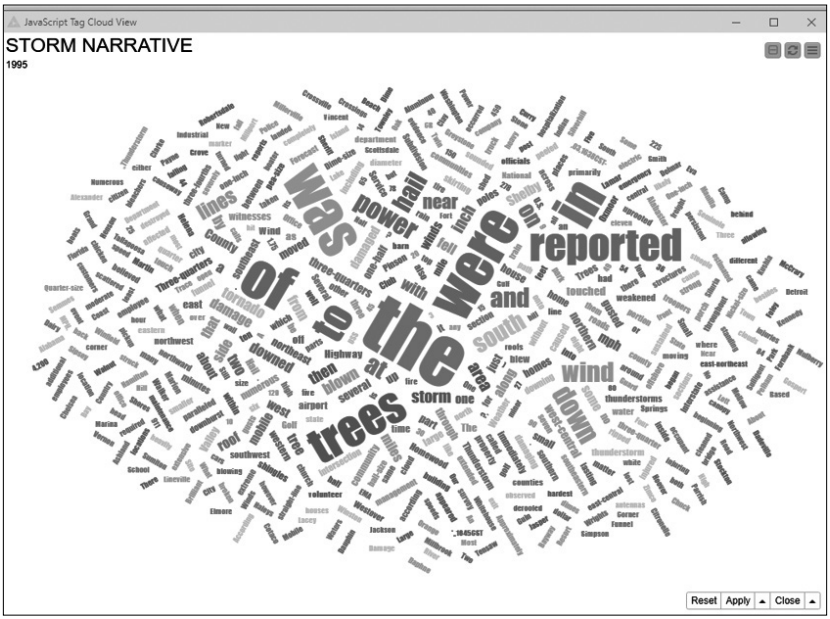


The one choice of "Font Scale" is based on growing the font linearly is this instance, but there are other scales available from the down arrow, and those might show some different visuals than the one that the analyst will see in the following.

The final tab of "Interactivity" provides the analyst with a way of manipulating the final visual, although these sometimes have a way of providing too many choices for the analyst, which "muddies" the analysis. However, these choices are here for the analyst to make should they so decide.



The final result, once the process is completed and connected (and executed), is the following visual. Notice how the words show a difference in size based on the occurrence of the word. As in the R version explained earlier, this type of analysis could point to some interesting trends and patterns. Although this example may not be the best to use, examples such as comments on surveys or text comments are great to use for this type of analysis.



These are just some of the tools available for text mining, and the more involved the analyst gets with this type of data, the more possibilities there are with analyzing actual text. If the analyst wants to test these functions, pick a speech from an online source and use that for these types of analytics. The possibilities are endless.

5.7 FILTERING

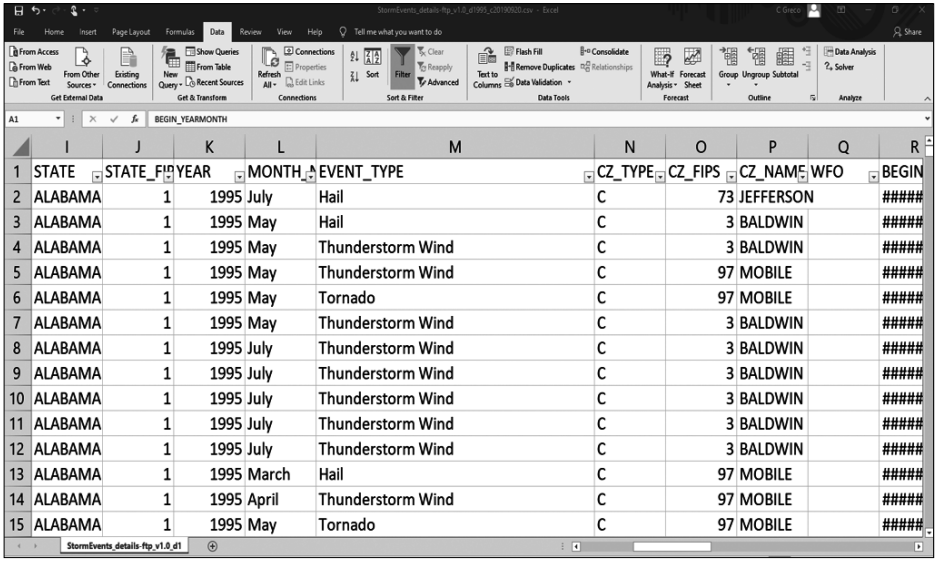
Probably the most fundamental task of any analyst is to clean the data so that only the data that is really pertinent is used in statistical testing. One of the primary ways of performing this function is by filtering the variables to ensure that only those variables that are needed are visible. This section will address that function and use the different tools to show how that can be accomplished.

5.7.1 Excel

In the situation with Excel, filtering is accomplished by two methods. The first method is choosing the "Filter" option from the "Data" tab, while the second is making the spreadsheet into a data table. Both of those methods will be demonstrated here.

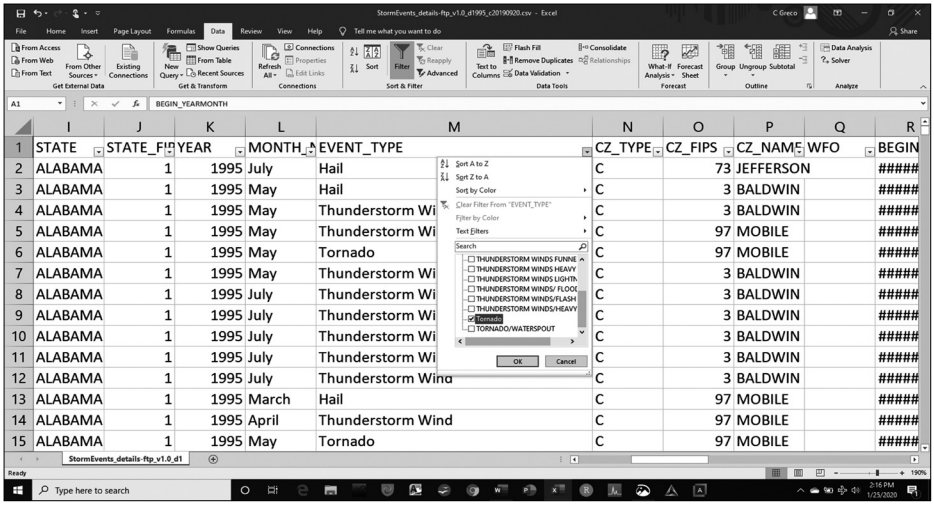
In the first method the analyst would first import the data; in this case the data will be the tornado data from 1995, since that contains more than just tornado data, and the analyst only wants tornado data. The filtering will eliminate all other data but tornado data.

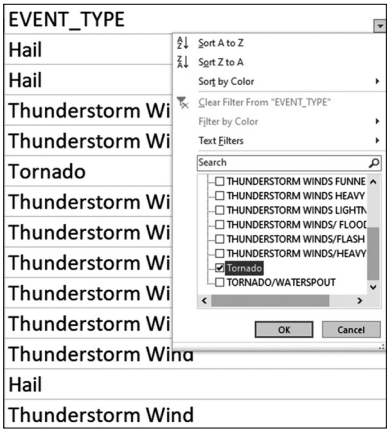
After importing the data, the analyst will go to the Data tab, and there the "Filter" choice (which looks like a funnel) exists. Click on the funnel and the filter down arrow will appear next to all variables (column headings in the data as shown).



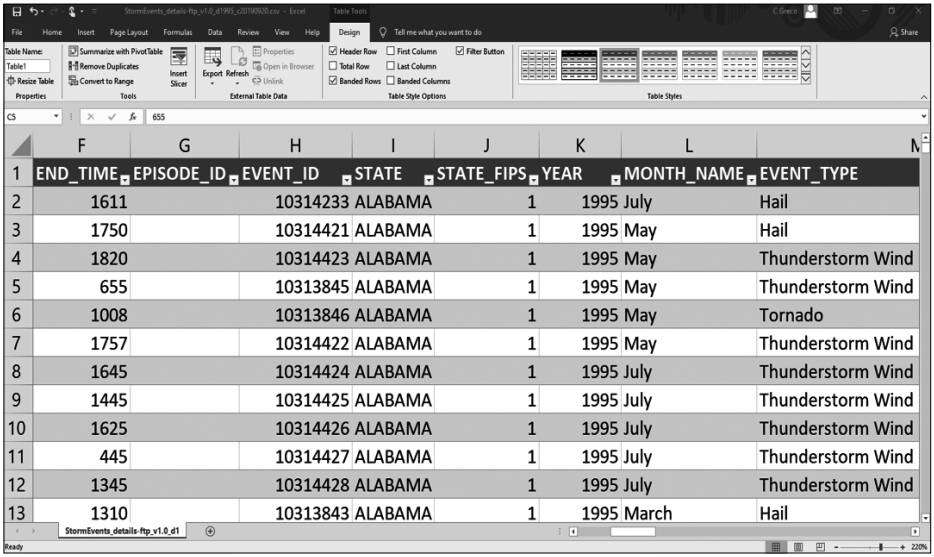
	I	J	K	L	M	N	O	P	Q	R
1	STATE	STATE_FIP	YEAR	MONTH_N	EVENT_TYPE	CZ_TYPE	CZ_FIPS	CZ_NAME	WFO	BEGIN
2	ALABAMA	1	1995	July	Hail	C	73	JEFFERSON		#####
3	ALABAMA	1	1995	May	Hail	C	3	BALDWIN		#####
4	ALABAMA	1	1995	May	Thunderstorm Wind	C	3	BALDWIN		#####
5	ALABAMA	1	1995	May	Thunderstorm Wind	C	97	MOBILE		#####
6	ALABAMA	1	1995	May	Tornado	C	97	MOBILE		#####
7	ALABAMA	1	1995	May	Thunderstorm Wind	C	3	BALDWIN		#####
8	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN		#####
9	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN		#####
10	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN		#####
11	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN		#####
12	ALABAMA	1	1995	July	Thunderstorm Wind	C	3	BALDWIN		#####
13	ALABAMA	1	1995	March	Hail	C	97	MOBILE		#####
14	ALABAMA	1	1995	April	Thunderstorm Wind	C	97	MOBILE		#####
15	ALABAMA	1	1995	May	Tornado	C	97	MOBILE		#####

Scroll until the "EVENT_TYPE" column appears as shown and use the down arrow to select just tornado from the different choices available. Once that is completed, only the tornado rows will appear. The analyst can then select the spreadsheet and copy it to another worksheet to work on just the tornado occurrences.





The second way to filter the columns is by changing the worksheet to a data table. The process for doing this is relatively straightforward. The first step is to import the data as before, but this time go to the Insert tab and choose "Table" in order to change the worksheet (or range) into a data table. The result of doing so is depicted in the following screen:

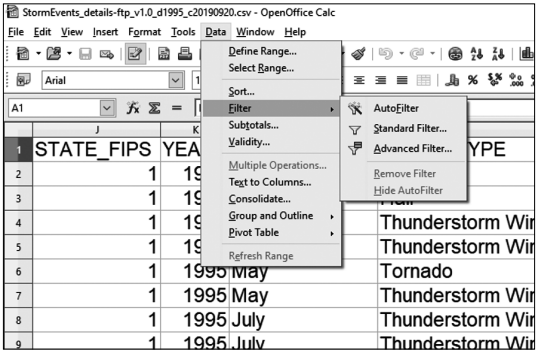


	F	G	H	I	J	K	L	M
1	END_TIME	EPISODE_ID	EVENT_ID	STATE	STATE_FIPS	YEAR	MONTH_NAME	EVENT_TYPE
2	1611		10314233	ALABAMA	1	1995	July	Hail
3	1750		10314421	ALABAMA	1	1995	May	Hail
4	1820		10314423	ALABAMA	1	1995	May	Thunderstorm Wind
5	655		10313845	ALABAMA	1	1995	May	Thunderstorm Wind
6	1008		10313846	ALABAMA	1	1995	May	Tornado
7	1757		10314422	ALABAMA	1	1995	May	Thunderstorm Wind
8	1645		10314424	ALABAMA	1	1995	July	Thunderstorm Wind
9	1445		10314425	ALABAMA	1	1995	July	Thunderstorm Wind
10	1625		10314426	ALABAMA	1	1995	July	Thunderstorm Wind
11	445		10314427	ALABAMA	1	1995	July	Thunderstorm Wind
12	1345		10314428	ALABAMA	1	1995	July	Thunderstorm Wind
13	1310		10313843	ALABAMA	1	1995	March	Hail

As the analyst can see, the data table comes equipped with filter down arrows already as part of the transformation, so the analyst can use these as in the previous paragraphs. There are many advantages to changing the worksheet to a data table, but they are beyond the scope of this book and are more than covered in the many Excel books that are available. This is included in this book only to use as a comparison to the other tools available here.

5.7.2 OpenOffice

OpenOffice has the same feel as older versions of Excel, so the natural place to start the filter process would be to go to the OpenOffice Spreadsheet, import the data, and go to the "Data" tab as in Excel. As shown in the following, it is a little different since under the filter option there are several choices.



The "AutoFilter" choice is fine for this example. As soon as that choice is selected, the same type of down arrows will appear next to the column headings and the analyst can choose how to filter the data. In this case, choosing "tornado" seems appropriate.

5.7.3 R/RStudio/Rattle

R has a package called "dplyr" which can be loaded and used within RStudio. This will filter the database so that only the columns needed will be viewed and can be loaded into Rattle as an R database. In this case, the analyst wants to limit the columns to only those rows that have "Tornado" in them, so that is the term used. However, in order to establish the new database as the filtered database, and to shortcut the long file name, the analyst decides to store the imported 1995 Severe Storm database into the TORNADO_1995 database. This provides for a much easier transition into the programming arena. The following commands will produce the result needed to continue with any further analysis.

```
> library(dplyr)
> TORNADO_1995<-StormEvents_details_ftp_v1_0_d1995_
                                                  c20190920
> TORNADO_1995<-filter(TORNADO_1995,EVENT_
                                              TYPE=="Tornado")
> View(TORNADO_1995)
```

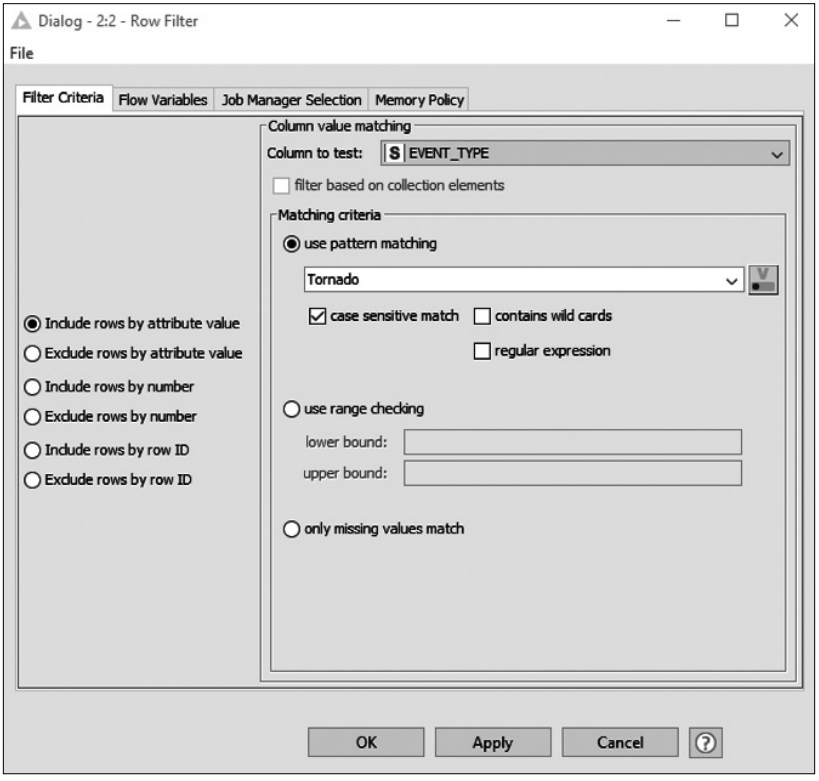
The last line (starting with "View") simply makes the data visible in the pane for viewing the data in a table format. This helps the analyst ensure that the filtering was done properly.

5.7.4 KNIME

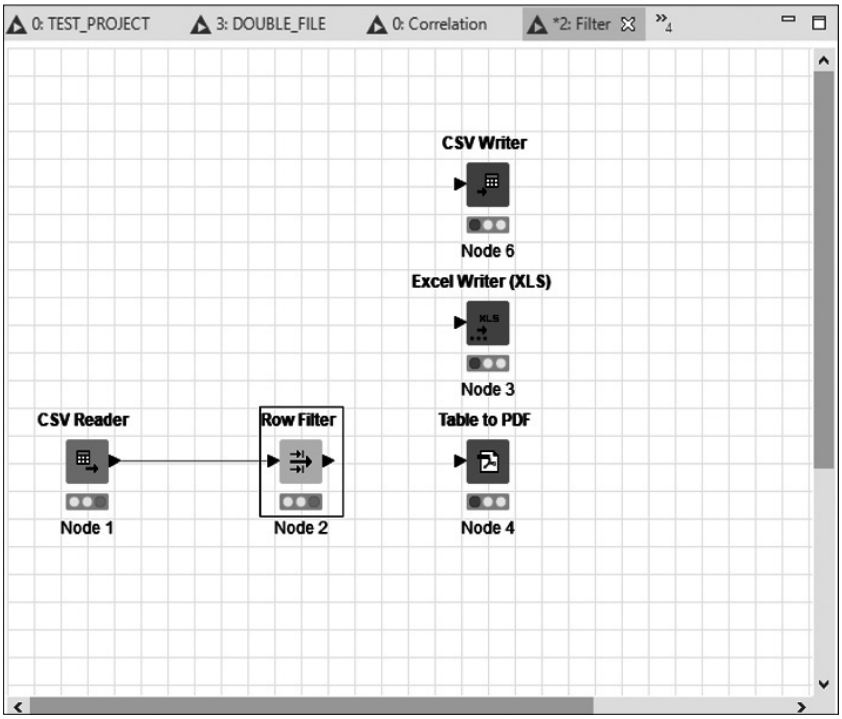
The KNIME tool can filter using a node for this purpose. The first step will be to import the 1995 tornado data into KNIME using the tried and true CSV Reader node and then filter the data using the "Row Filter" node. The configuration screen for the node is as follows, and you need to explore this configuration to best fit the needs of the analysis.

It is important to notice the different parts of this configuration screen. The column is selected at the top right and then "Matching criteria" is selected. In this case, the analyst only wants the tornado portion of the data, so that is chosen, but in addition the box for case sensitivity is checked to ensure that

the variable matches. Notice the "Include rows by attribute value" is selected, as it should be, since the analyst wants only those rows marked "Tornado." Please be careful whenever choosing any other option, because choosing the wrong one will eliminate all those rows the analyst wants to use!



The finished workflow screen for the KNIME filter workflow is as follows, and take special note of the added nodes. These nodes are there to show the different types of tools that KNIME can export, including two tools mentioned in this text. These nodes can be found in the "IO" category and can be a major enhancement to the data analysis, since the same data can be analyzed using different tools. These nodes can also be used to export a number of datasets in succession since, once the nodes are set in the workflow, the output can also be consistently determined.



The reason for including the "Table to PDF" option is that there are times when the finished analysis is best suited for a report, and there is nothing like converting the table to PDF to help include that reporting page in the most flexible document style. Besides, the PDF document can be imported into a number of tools and used in future data analytics, so the conversion to PDF only makes sense in the long run. Regardless of the reason, using the output nodes will help the analyst to be dynamic in their future analytics.

CHAPTER 6

SUMMARY

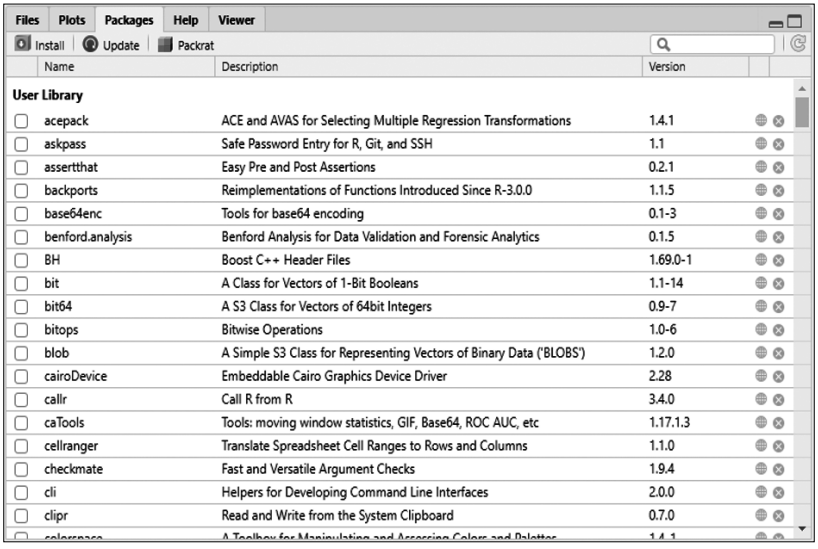
This text was based on a few different statistical concepts that exist currently within the data analyst "wheelhouse." If the data analyst is not familiar with any of the aforementioned concepts, please read the many statistical texts and references that are either at the end of this book or found in many online and brick and mortar bookstores. Start with some very basic texts and move to the more complex. Whatever the analyst does in the way of analyzing data, a good foundation of statistics is both necessary and productive. There is always more information on data analytics, data science, and statistics that is out there, so never let a possible learning possibility pass by. Also, as far as these tools are concerned, the very "topsoil" of functionality has been demonstrated with them. There is certainly more information available and more functionality possible with these applications. It is important that the analyst use these tools for the purpose of the analysis. Avoid using the tool to display a colorful graph or to visualize something that may not be valid. The very reputation of the analyst is at stake when taking loosely connected variables and attempting to connect them. That is not the purpose of analytic tools. The purpose of the tools is to provide the analyst the quickest method to calculate something that would take hours to do with manual methods like a calculator.

6.1 PACKAGES

There were several areas that were not discussed at the beginning of the text that need some clarification now. The first is "package" in relation to Rattle (and R in general). A package in R refers to a specific programmed function that acts as a "one-step" procedure to do certain tests and models. These are critical in making the analytical process as quick and efficient as possible, but there are some caveats that need to accompany these packages. First, the

package must be installed in order to be activated. Some are installed with the basic R installation, but there are many that are not. Rattle is actually a package that must be installed in order to work. Every time an analyst closes R (which will consequently close Rattle if it is open), the R base will convert back to not having the packages activated within R. The package will still be installed, but it will not be activated until the user does this through the programming window (which has been demonstrated), or by "checking the box" next to the package in the IDE right bottom pane of RStudio, depicted as follows. The packages that are shown are just a fraction of those available through the Comprehensive R Archive Network (or CRAN), from which any package can be found and installed. When the analyst installs R, they can choose which CRAN "mirror" (basically server) to load these packages from. In some cases, it is beneficial to click on the "install" button in the following screen and type in a function that needs to be activated. In most cases, there is a functional sub-program (package) that can do the "messy work" for the analyst. This is the power of R and Rattle—to help the analyst solve analytical problems without extraordinary programming expertise.

One more point about packages. These sub-programs may rely (depend) on other packages and may install these dependencies in order for the package to work properly. In many instances, this will be announced to the user and permission will be asked to install the package dependencies. If the user has any doubt as to the appropriateness of the package or the dependencies, they can say no to this request. However, that will mean that the package will not function properly.

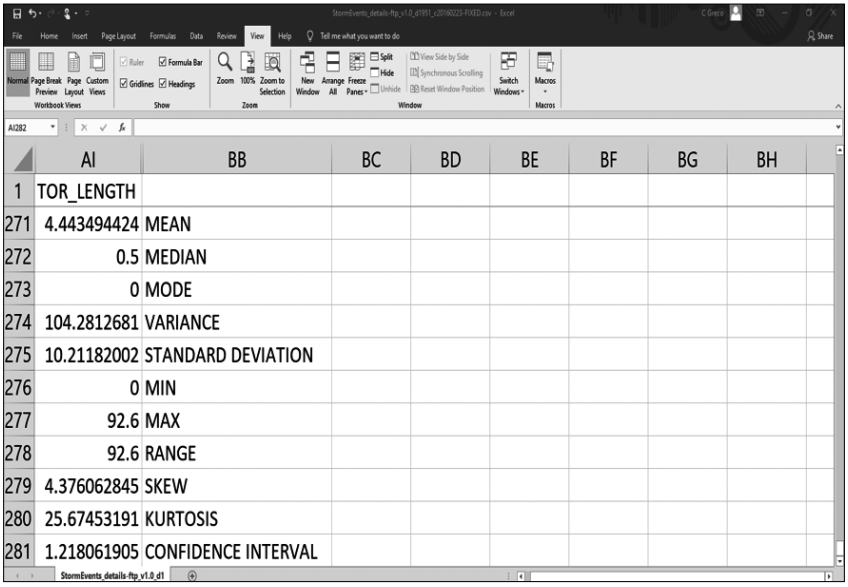


The nice aspect about RStudio is that, by checking the box next to the package, the R programming is automatically activated to place the package at the user's disposal. There is no other programming that the user has to accomplish, just make sure the box is checked. This is just one reason why downloading and installing RStudio is well worthwhile for anyone that wants to do data analytics with a FOSS application.

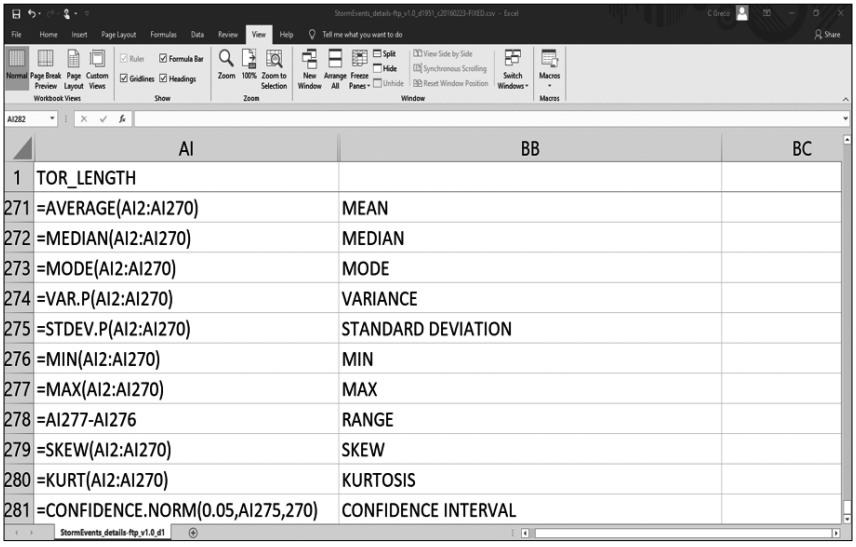
6.2 ANALYSIS TOOLPAK

The Analysis ToolPak in Excel may not be available for every user, specifically for those that may work in the federal government, since that is called an add-in and may be under additional agreements other than the base Microsoft Office. As a result, the add-in may not be available to those that need it. If this is the case, then an analyst can always do the "manual" approach to getting the same data as using the Analysis ToolPak. This is more complicated than using the convenience of the add-in, but with a little patience and persistence, the same results will appear.

In order to do this, the first thing is to import the data as done before, but this time using the bottom of the worksheet to list and use the different formulas to produce the descriptive summary as in the previous section. The following screen will show all the formulas necessary to provide the information for the tornado lengths (TOR_LENGTH) on the 1951 tornado data. Each of these formulas will be discussed and the results shown. One hint about showing formulas in Excel: if there is a need to show the formulas in the spreadsheet, go to the "Formulas" tab on the main toolbar and choose "Show Formulas." If there is a preference to use the keyboard shortcuts, then hold the CTRL key and press the "~" button, which is located just below the "Esc" key. This is a "toggle" key that the analyst can continually press to show the formulas or the results.



	AI	BB	BC	BD	BE	BF	BG	BH
1	TOR_LENGTH							
271	4.443494424	MEAN						
272	0.5	MEDIAN						
273	0	MODE						
274	104.2812681	VARIANCE						
275	10.21182002	STANDARD DEVIATION						
276	0	MIN						
277	92.6	MAX						
278	92.6	RANGE						
279	4.376062845	SKEW						
280	25.67453191	KURTOSIS						
281	1.218061905	CONFIDENCE INTERVAL						



	AI	BB	BC
1	TOR_LENGTH		
271	=AVERAGE(AI2:AI270)	MEAN	
272	=MEDIAN(AI2:AI270)	MEDIAN	
273	=MODE(AI2:AI270)	MODE	
274	=VAR.P(AI2:AI270)	VARIANCE	
275	=STDEV.P(AI2:AI270)	STANDARD DEVIATION	
276	=MIN(AI2:AI270)	MIN	
277	=MAX(AI2:AI270)	MAX	
278	=AI277-AI276	RANGE	
279	=SKEW(AI2:AI270)	SKEW	
280	=KURT(AI2:AI270)	KURTOSIS	
281	=CONFIDENCE.NORM(0.05,AI275,270)	CONFIDENCE INTERVAL	

Compare these results with the results from the section on Descriptive Statistics and there will be little if any difference. Also, the section showing these formulas in OpenOffice is different since OpenOffice uses ";" and Excel uses "," so remember these differences when moving between tools.

CHAPTER 7

SUPPLEMENTAL INFORMATION

This section will contain information that was neglected in the explanation of some of the other sections along with some exercises for the reader to use in order to better focus on the different concepts presented in the previous sections. The answers will include most of the tools, taking the best one for the problem and working toward others that might do the trick. Not all the tools will be included in all the answers separately, but each individual tool will be displayed in at least one of the answers. Please go out to the locations where data is accessible and use the data for exercises in order to just have some analytical fun. Otherwise, this book will end up on a shelf, never used but for a paperweight.

7.1 EXERCISE ONE – TORNADO AND THE STATES

The first exercise will explore using some of the tools to analyze which states seem to have a tornado more than other states. The analyst should never go into an analysis jumping to conclusions. There may be a hypothesis that the analyst want to make. This is not a conclusion, but really an assertion. For instance, the analyst might say that there were more tornados in Texas than in Connecticut in 2018. This is an assertion that can be verified by data and by basic analytics.

This section will focus on a particular statistical test and how to either reject or not reject the null hypothesis based on that test. The first step is to state the hypothesis as the null hypothesis and then make an alternative hypothesis. To make it clear, this does not have to be a formal process, but

using an informal hypothesis formulation helps the analysis to be more precise, since not stating one allows the analyst to "play the field" concerning the type of data variables to test and, by virtue of that, expand the relationships between these fields until one or more are related. This is a biased way of performing analysis and will result in possible spurious correlations or, worse, determining that a variable has a cause and effect relationship with another variable when in fact there is no relationship of that sort.

In this case, the assertion (or claim) is that there are more tornados in Texas than in Connecticut in 2018. The null hypothesis (not the claim in this case) would be that there are the same number of tornados in Texas as in Connecticut. There are analysts that would try to do a correlation or regression analysis in order to prove the assertion, but in this case a simple descriptive analysis is more than sufficient to make the case.

The first step will be to import the data to the tool and then perform descriptive statistics against that data. After performing that test, the analyst could show the relationship by a simple bar chart or similar visual. Remember that there are certain types of data that are more amenable to certain types of visual presentations. Discrete variables (those that are integers), which are not time dependent, are more adaptable to bar charts. Longitudinal studies, those that are based on succeeding years, are better suited to line charts. This is important since it is that type of association that will allow the analyst to make a very effective presentation without confusing the audience.

Find the answer to this exercise using any of the tools presented in this text along with the 2018 tornado dataset (ensure it says "details" on the file name) at the site mentioned in the first few sections of this text, specifically at *https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/*.

Remember that this file will have all the storm events, so the analyst will have to filter the tornados from the rest of the storm events to get the proper data to analyze. Filtering was addressed in an earlier section.

7.1.1 Answer to Exercise 7.1

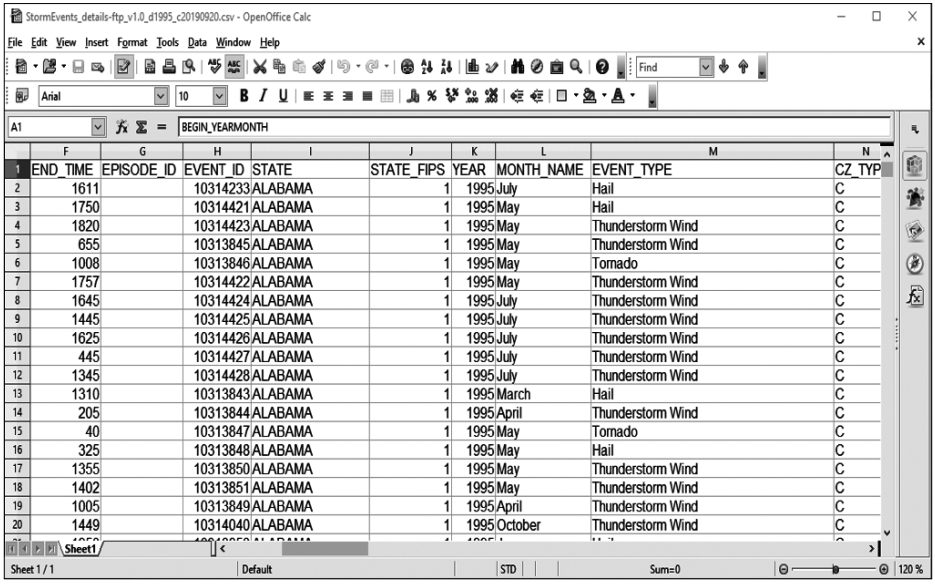
The answer to the preceding exercise will require data filtering to ensure that only tornado rows are included in the data. After that, a simple comparison of Texas and Connecticut for the count (or average) of tornados will suffice for the analysis. There is something that may be considered in this analysis. Factors such as population and land area might have some bearing on the actual data, specifically on the number of tornados. This would be standardized ("normalized") by using weights, which in this case is not considered, and is not within the scope of this book.

7.1.1.1 Answer According to OpenOffice

The answer to the question using OpenOffice would be very similar to Excel, except that OpenOffice does not have the Analysis ToolPak.

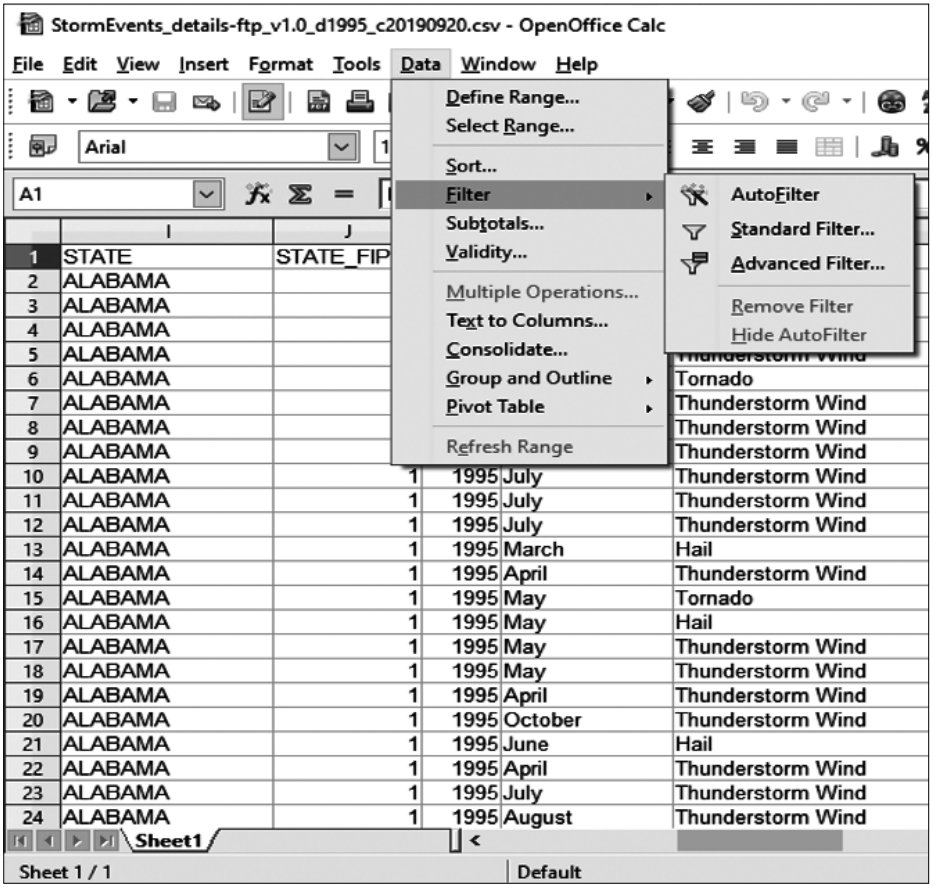
The steps to analyze the data, considering the hypothesis that there are fewer tornados in Connecticut than Texas, would be to import the data and then filter the data to only consider the tornado events. After that, present the data in a visual that would show the recipient the answer to the question (or whether or not to confirm the hypothesis).

The data in OpenOffice would look like the following screen, which shows the entire data, including all severe storms, which would need to be filtered. The next screen shows the filtered data; and the last screen shows a bar chart showing the number of tornados per state.



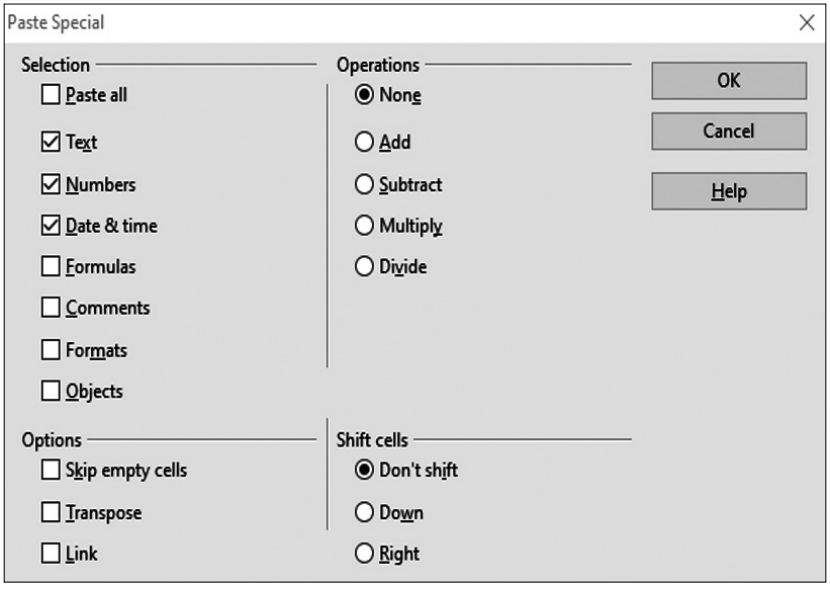
	F	G	H	I	J	K	L	M	N
1	END_TIME	EPISODE_ID	EVENT_ID	STATE	STATE_FIPS	YEAR	MONTH_NAME	EVENT_TYPE	CZ_TYP
2	1611		10314233	ALABAMA	1	1995	July	Hail	C
3	1750		10314421	ALABAMA	1	1995	May	Hail	C
4	1820		10314423	ALABAMA	1	1995	May	Thunderstorm Wind	C
5	655		10313845	ALABAMA	1	1995	May	Thunderstorm Wind	C
6	1008		10313846	ALABAMA	1	1995	May	Tornado	C
7	1757		10314422	ALABAMA	1	1995	May	Thunderstorm Wind	C
8	1645		10314424	ALABAMA	1	1995	July	Thunderstorm Wind	C
9	1445		10314425	ALABAMA	1	1995	July	Thunderstorm Wind	C
10	1625		10314426	ALABAMA	1	1995	July	Thunderstorm Wind	C
11	445		10314427	ALABAMA	1	1995	July	Thunderstorm Wind	C
12	1345		10314428	ALABAMA	1	1995	July	Thunderstorm Wind	C
13	1310		10313843	ALABAMA	1	1995	March	Hail	C
14	205		10313844	ALABAMA	1	1995	April	Thunderstorm Wind	C
15	40		10313847	ALABAMA	1	1995	May	Tornado	C
16	325		10313848	ALABAMA	1	1995	May	Hail	C
17	1355		10313850	ALABAMA	1	1995	May	Thunderstorm Wind	C
18	1402		10313851	ALABAMA	1	1995	May	Thunderstorm Wind	C
19	1005		10313849	ALABAMA	1	1995	April	Thunderstorm Wind	C
20	1449		10314040	ALABAMA	1	1995	October	Thunderstorm Wind	C

The next step is to filter the data so that just tornados are visible in the EVENT_TYPE column. This is done through the Data choice in the toolbar and choosing the Filter… option with the Auto Filter… sub-option as shown. What this will do is to place the funnel next to all the columns, and the analyst can then choose the variable desired.

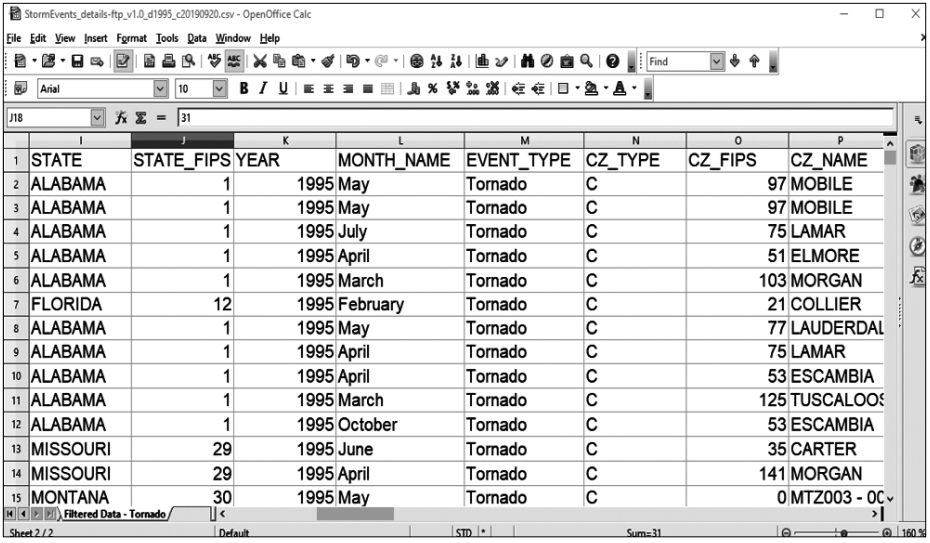


	I	J			
1	STATE	STATE_FIP			
2	ALABAMA				
3	ALABAMA				
4	ALABAMA				
5	ALABAMA				
6	ALABAMA				Tornado
7	ALABAMA				Thunderstorm Wind
8	ALABAMA				Thunderstorm Wind
9	ALABAMA				Thunderstorm Wind
10	ALABAMA	1	1995	July	Thunderstorm Wind
11	ALABAMA	1	1995	July	Thunderstorm Wind
12	ALABAMA	1	1995	July	Thunderstorm Wind
13	ALABAMA	1	1995	March	Hail
14	ALABAMA	1	1995	April	Thunderstorm Wind
15	ALABAMA	1	1995	May	Tornado
16	ALABAMA	1	1995	May	Hail
17	ALABAMA	1	1995	May	Thunderstorm Wind
18	ALABAMA	1	1995	May	Thunderstorm Wind
19	ALABAMA	1	1995	April	Thunderstorm Wind
20	ALABAMA	1	1995	October	Thunderstorm Wind
21	ALABAMA	1	1995	June	Hail
22	ALABAMA	1	1995	April	Thunderstorm Wind
23	ALABAMA	1	1995	July	Thunderstorm Wind
24	ALABAMA	1	1995	August	Thunderstorm Wind

The result of this filter is shown in the next screen. Please note that the tornado factor is now the only one showing. The next step will be to copy and paste the filtered data into another sheet. This is the same process as one done in Excel, so select all of the data and paste it into another sheet. Be careful with this step, however, since the desire is to paste as values and not just paste everything, since that will include all the unfiltered values, making the copied sheet contain all of the data, not just the tornado rows. In order to do this, right-click on the data to be copied and then select cell A1 in the blank sheet. Right-click in the blank sheet and there will be an option called Paste Special. When that is clicked, the following screen will appear:



If the checkbox is checked for Paste All, the copied sheet will be the same as the unfiltered sheet. If that is unchecked and the three are checked that are shown, the copied sheet will be for all intents and purposes a new sheet of just tornado events. That is the result that the analyst wants to achieve.



StormEvents_details-ftp_v1.0_d1995_c20190920.csv - OpenOffice Calc

File Edit View Insert Format Tools Data Window Help

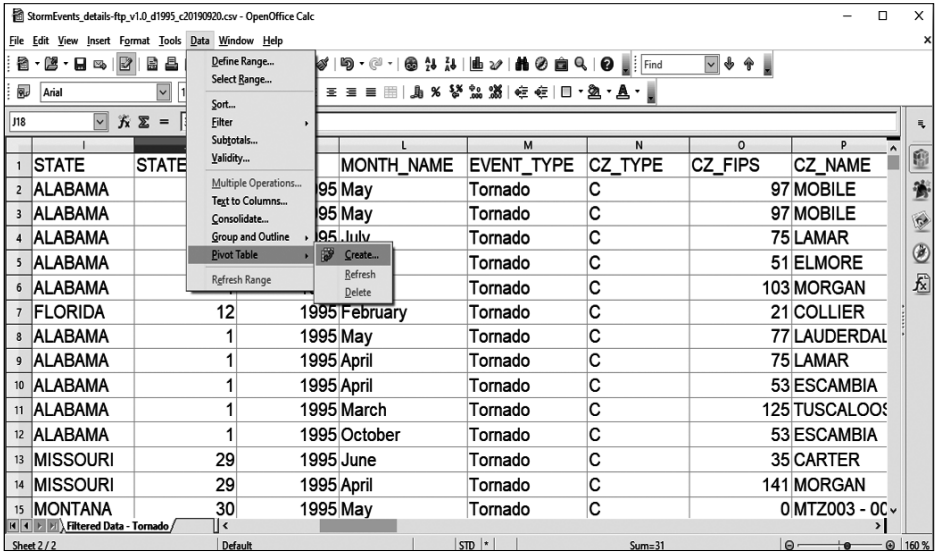
J18 | 31

	I	J	K	L	M	N	O	P
1	STATE	STATE_FIPS	YEAR	MONTH_NAME	EVENT_TYPE	CZ_TYPE	CZ_FIPS	CZ_NAME
2	ALABAMA	1	1995	May	Tornado	C	97	MOBILE
3	ALABAMA	1	1995	May	Tornado	C	97	MOBILE
4	ALABAMA	1	1995	July	Tornado	C	75	LAMAR
5	ALABAMA	1	1995	April	Tornado	C	51	ELMORE
6	ALABAMA	1	1995	March	Tornado	C	103	MORGAN
7	FLORIDA	12	1995	February	Tornado	C	21	COLLIER
8	ALABAMA	1	1995	May	Tornado	C	77	LAUDERDAL
9	ALABAMA	1	1995	April	Tornado	C	75	LAMAR
10	ALABAMA	1	1995	April	Tornado	C	53	ESCAMBIA
11	ALABAMA	1	1995	March	Tornado	C	125	TUSCALOOS
12	ALABAMA	1	1995	October	Tornado	C	53	ESCAMBIA
13	MISSOURI	29	1995	June	Tornado	C	35	CARTER
14	MISSOURI	29	1995	April	Tornado	C	141	MORGAN
15	MONTANA	30	1995	May	Tornado	C	0	MTZ003 - 0C

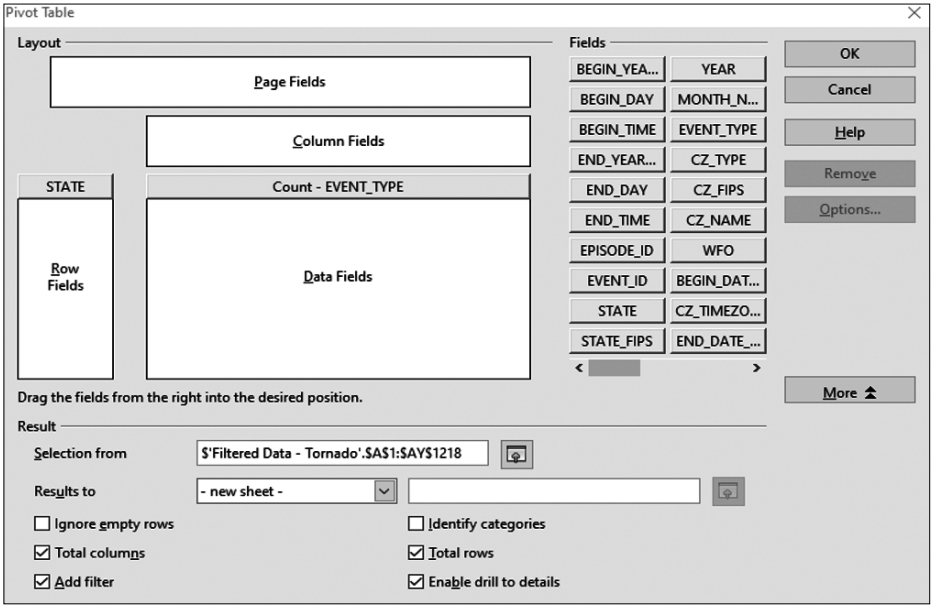
Filtered Data - Tornado

Sheet 2 / 2 Default STD Sum=31 160 %

From this result, the analyst can now make a pivot table with the STATE as the x-axis and the Tornado as the y-axis. The analyst will want to ensure that the two states being compared are Texas and Connecticut as shown in the following:

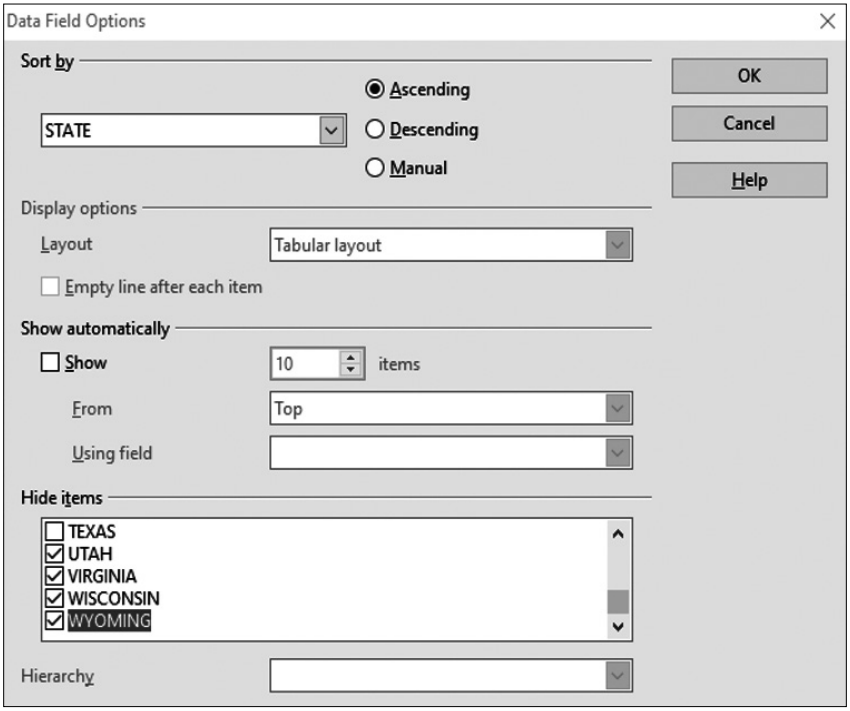


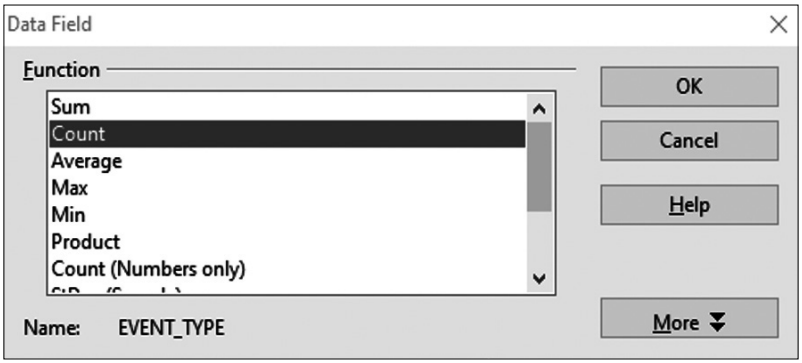
	STATE	STATE		MONTH_NAME	EVENT_TYPE	CZ_TYPE	CZ_FIPS	CZ_NAME
2	ALABAMA		95	May	Tornado	C	97	MOBILE
3	ALABAMA		95	May	Tornado	C	97	MOBILE
4	ALABAMA		95	July	Tornado	C	75	LAMAR
5	ALABAMA				Tornado	C	51	ELMORE
6	ALABAMA				Tornado	C	103	MORGAN
7	FLORIDA	12	1995	February	Tornado	C	21	COLLIER
8	ALABAMA	1	1995	May	Tornado	C	77	LAUDERDAL
9	ALABAMA	1	1995	April	Tornado	C	75	LAMAR
10	ALABAMA	1	1995	April	Tornado	C	53	ESCAMBIA
11	ALABAMA	1	1995	March	Tornado	C	125	TUSCALOOS
12	ALABAMA	1	1995	October	Tornado	C	53	ESCAMBIA
13	MISSOURI	29	1995	June	Tornado	C	35	CARTER
14	MISSOURI	29	1995	April	Tornado	C	141	MORGAN
15	MONTANA	30	1995	May	Tornado	C	0	MTZ003 - 0C

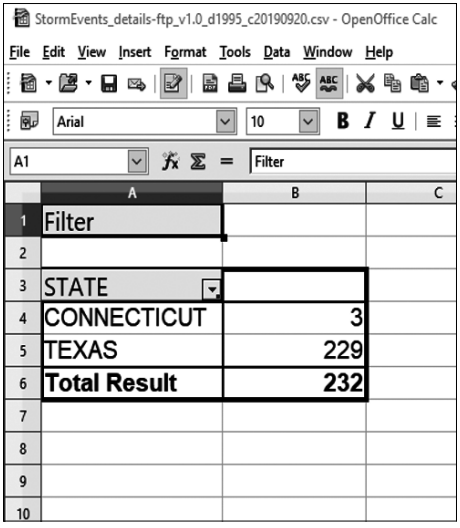


Please configure the Pivot Table screen as in the previous screen, but ensure that the pivot table is placed in another sheet. Otherwise, the pivot table will exist in the same sheet as the original data, which could cause issues if the analyst is not aware of this and selects an entire sheet, which could include the pivot table, polluting the data with additional numbers.

The result of the pivot table, with the STATE as the rows and the DATA as the count of tornados, is as follows, and it will be more than obvious that there are more tornados in Texas than in Connecticut. However, and this is important, the land area of Connecticut is much smaller than Texas and, should this be weighted, the numbers would be closer. That is beyond the scope of this book, but exploring these characteristics will only make the analyst a much more detailed user of the data.





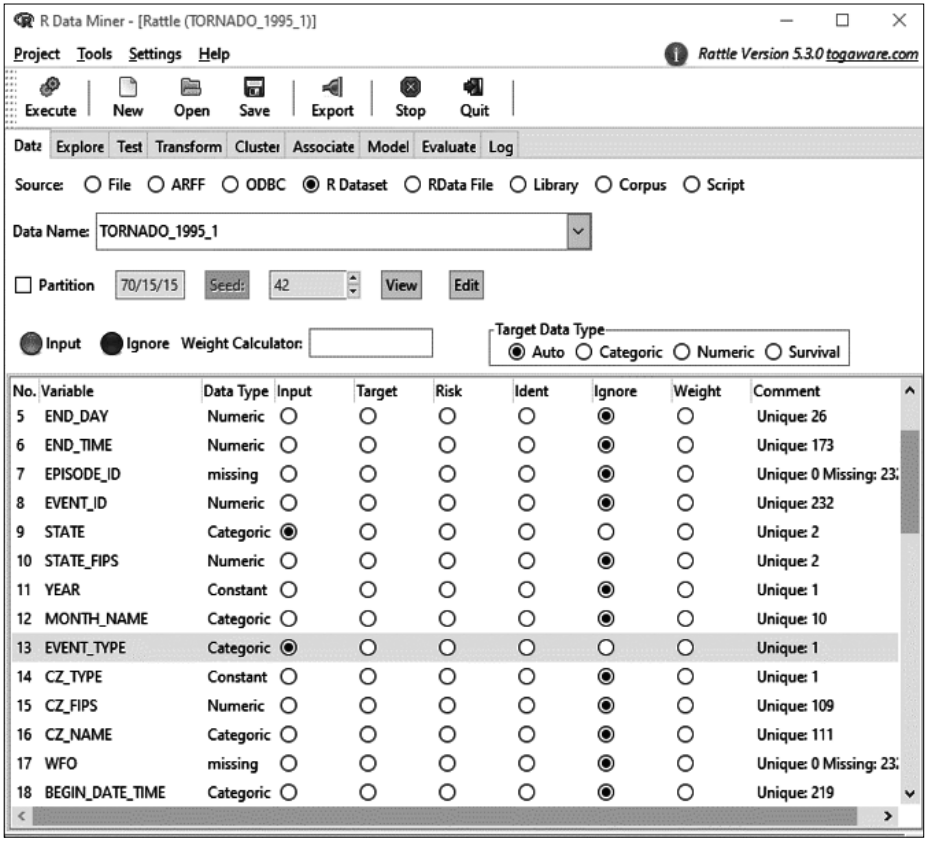


7.1.1.2 Answer According to Rattle

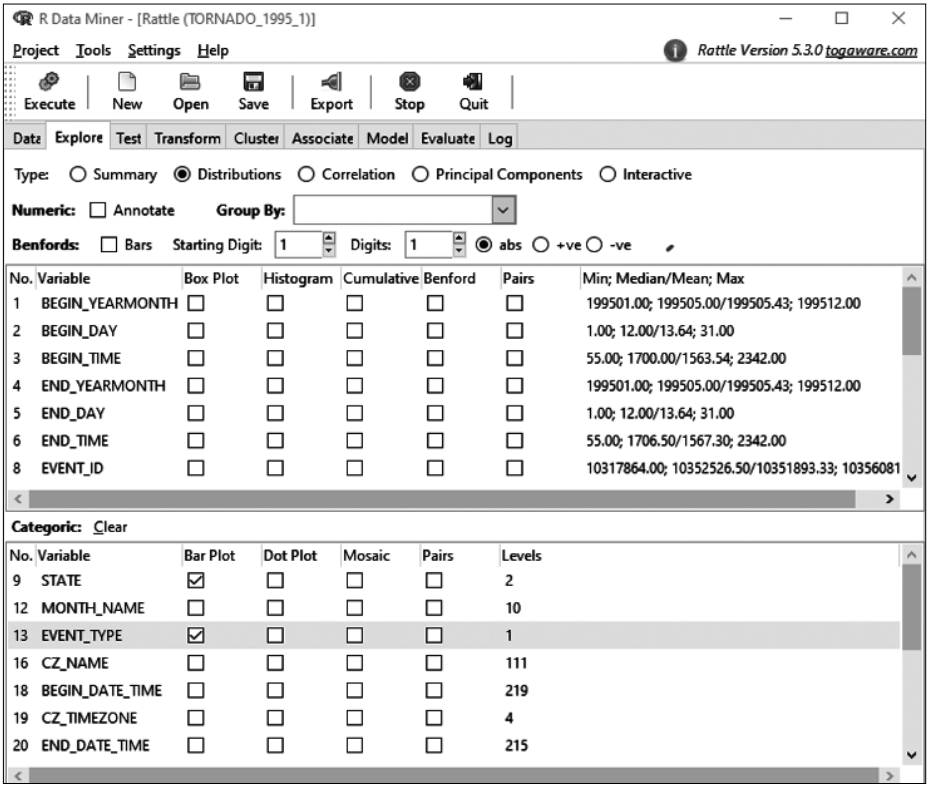
Forming a chart within Rattle is relatively straightforward. The most difficult part of this process is the filtering of the data, but that can be done within both R and Rattle. Filtering using R/RStudio was already covered in a prior section, so the file TORNADO_1995 already has just the tornado record and fields prepared. To import this into Rattle, use the data import function as described in a previous section. Before importing the data, it might be wise to eliminate all the STATES except for those needed. This is also done through the filter command with the command line shown here.

```
TORNADO_1995_1<- filter(TORNADO_1995,STATE=="TEXAS" |
                                     STATE=="CONNECTICUT")
```

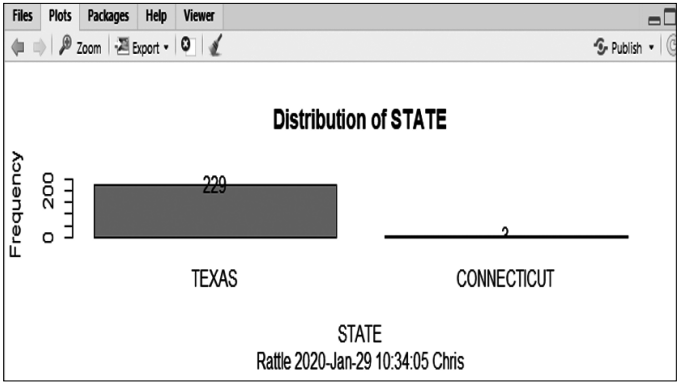
Notice that there is a pipe ("|") used for the "or" function, so this line states that I want the original dataset to be filtered to show only tornados that have occurred in Texas or Connecticut. After that is finished, then import the data into Rattle and ensure that the Execute icon is clicked. The user may have to use the radio button marked "ignore" to ensure that only those two fields of STATE and EVENT_TYPE are chosen. This is shown as follows:



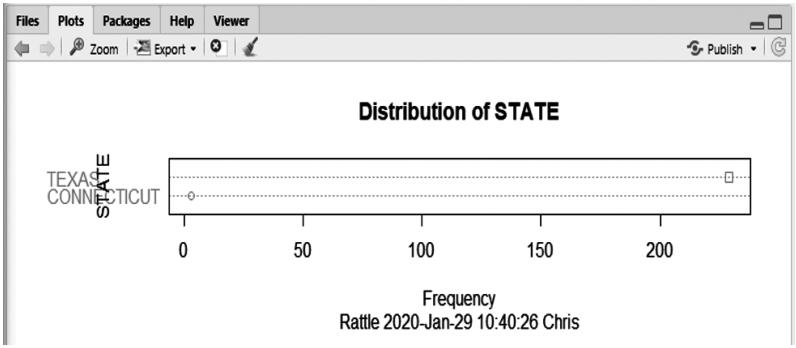
Once that is completed, move to the "Explore" tab and use the "Distributions" choice of the bar chart to display the two results. The screen to configure is as follows:



The result of clicking the Execute icon is shown as follows. The visual is self-explanatory and matches the results from the OpenOffice screen.



Just a few suggestions. First, the resulting visual will appear in a pane in RStudio, not in Rattle, so do not be expecting a result at that location. Second, the number for the Connecticut number of tornados is cut off; it should be 3, and there will be another graph following this that will show the total, which will prove that 3 appears here. If the user wants another way of displaying the result, use the "Dot Plot" function within the Distributions tab to produce this result.

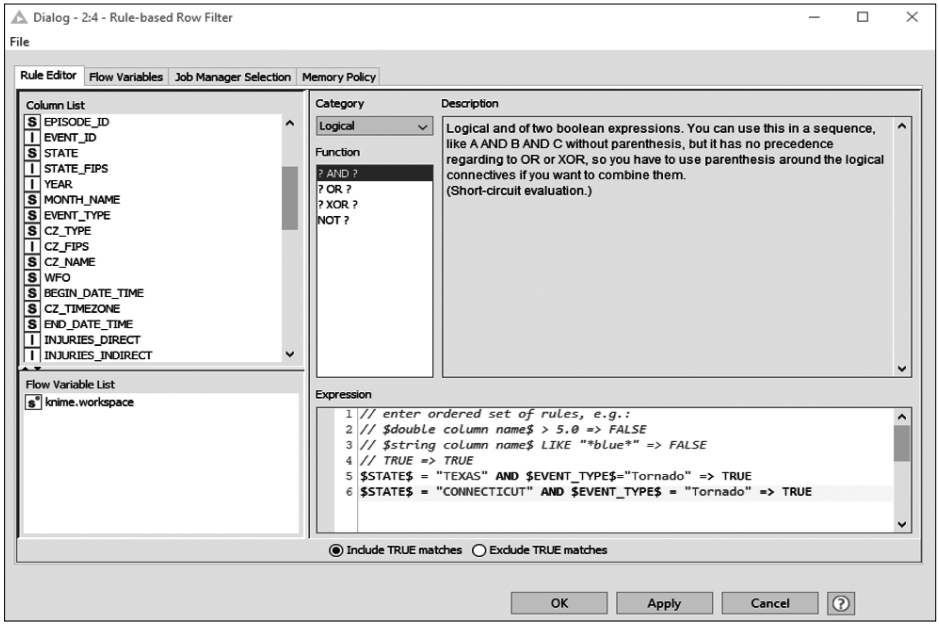


This shows the great disparity of tornados without the use of total numbers. This alone should prove that there are fewer tornados in Connecticut than Texas (nominally, in any case).

7.1.1.3 Answer According to KNIME

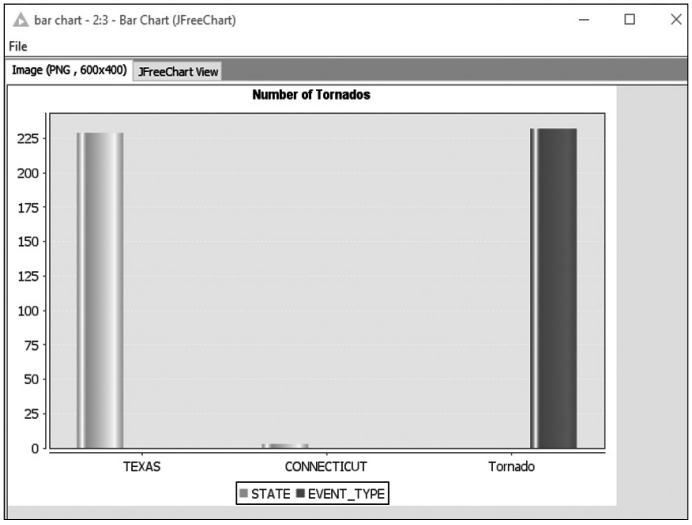
The KNIME tool has the ability to show the result of the analysis with very few nodes. The first node will be a CSV Reader which will read the entire dataset into the tool. The second node will be a Row Filter node to just show the STATE and EVENT_TYPE columns, and the third node will be to visualize the results.

There is an additional node that the user will need in order to make quick work of the filtering task. This node is called the Rule-Based Row Filter and will demand a little programming in order to make the filter work as a combination of STATE and EVENT_TYPE. The configuration for this node is illustrated as follows. The programming will be explained line by line.

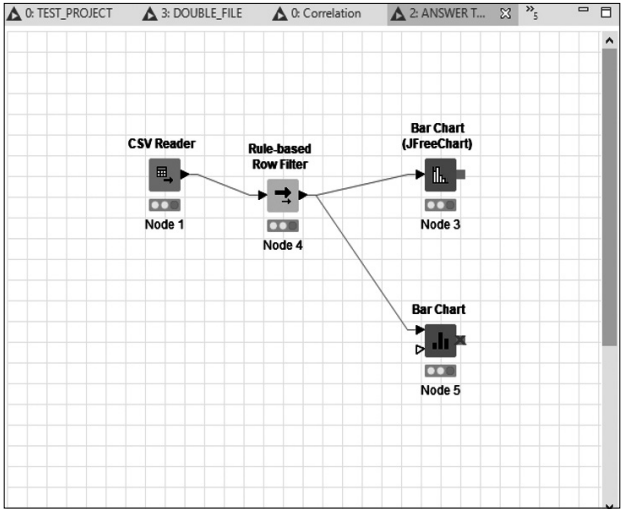


The first line (line #5 in the configuration screen) is basically setting the STATE rows to TEXAS and the EVENT_TYPE to Tornado. The second line (line #6 in the configuration screen) is setting the STATE rows to CONNECTICUT and the EVENT_TYPE to Tornado. By placing these lines one after the other, the user is placing an AND between them and the data. The resulting table will show only those tornados in Texas and Connecticut.

The last node in this process is a visualization node to help present the results. In this case, the Bar Chart node is perfectly acceptable. There are several to choose from, and two are included in this process. This is the result of the Bar Chart (JFree Chart) option. The picture is worth 1,000 words.



The third column in this chart shows the EVENT_TYPE as Tornado, which will count both Texas and Connecticut. Most of the tools show this as part of this function. The finished workflow is shown along with the named nodes. Remember that the row filter will demand just a little programming. However, also remember that the analyst can take any tornado tracking year and run it through this workflow to determine the difference in tornado rate between Texas and Connecticut. Once the workflow is set, no further configuration is required.



One last comment on KNIME. There are two Bar Chart nodes, one labeled as (JFree Chart) and the other plain. The plain Bar Chart is not as friendly as the JFree Chart, so it is recommended that the analyst use the JFree Chart if there is a choice. The plain Bar Chart is included to demonstrate that one node can feed two or more other nodes.

7.1.2 Pairing Exercise

One area of analysis that helps to determine changes from one event to another event is available through t-testing called pairing samples. In essence, what this entails is taking a sample of data that is paired one-for-one between one field and another field in order to determine if there really is a change between one and the other, taking into account chance. For instance, if the analyst wants to know if an individual can see better before and after cataract surgery, a pre- and post-testing method would help to determine if this is true using the paired sampling technique. In order to do this, the following exercise presents simulated data showing 100 students that have taken a test before and after a class. Each student is assigned a number (sequentially) that is the same for each test. Both tests have exactly the same questions, but the answers are randomized between the pre- and post-test in order to eliminate students just memorizing answers. The analyst must use the pairing sampling t-test in order to determine the following question: "Do students perform better in the post-test than the pre-test?" From this question, the following null hypothesis is generated:

H_0: Students' post-test scores and pre-test scores are not different

H_a: Students' scores are different between pre-test and post-test

This is called a "two-tailed" test, since it does not matter if the post-test scores are less than or greater than the pre-test scores. This is a simpler method of testing, and this will be the preferred option for this specific test.

The text for this exercise is located here, including the two columns necessary for the testing. The analyst can use the import instructions for each tool in order to get the data into the different analytical applications. Please note that there are commas included in order to make a comma separated value file that will help the import into the different tools. If the analyst wants to download the data, they may do so from the author's website at *www.grectech.com*. The analyst will navigate to the download page and will see the appropriate file listed under "Book Exercise 2."

Student, Pre-Testing, Post-Testing

1,	85,	82	39,	88,	82	77,	87,	87
2,	92,	92	40,	61,	90	78,	58,	82
3,	77,	82	41,	99,	99	79,	80,	82
4,	64,	79	42,	63,	78	80,	69,	73
5,	69,	70	43,	75,	91	81,	66,	94
6,	57,	89	44,	75,	93	82,	94,	89
7,	84,	79	45,	65,	88	83,	97,	95
8,	54,	70	46,	76,	99	84,	94,	90
9,	65,	97	47,	95,	75	85,	82,	94
10,	53,	88	48,	89,	88	86,	79,	87
11,	86,	80	49,	89,	77	87,	56,	74
12,	54,	82	50,	53,	78	88,	96,	94
13,	92,	90	51,	87,	97	89,	90,	85
14,	78,	99	52,	53,	79	90,	89,	72
15,	94,	89	53,	96,	89	91,	50,	98
16,	77,	95	54,	93,	83	92,	72,	80
17,	70,	79	55,	73,	71	93,	86,	88
18,	96,	98	56,	81,	93	94,	67,	99
19,	80,	86	57,	96,	98	95,	80,	78
20,	87,	81	58,	85,	81	96,	83,	75
21,	89,	99	59,	75,	92	97,	63,	84
22,	71,	81	60,	71,	94	98,	50,	90
23,	92,	98	61,	58,	91	99,	82,	97
24,	53,	97	62,	96,	72	100,	88,	76
25,	100,	99	63,	55,	77			
26,	54,	89	64,	90,	73			
27,	62,	94	65,	66,	74			
28,	65,	76	66,	87,	79			
29,	91,	92	67,	91,	90			
30,	76,	71	68,	85,	86			
31,	74,	84	69,	66,	95			
32,	95,	99	70,	80,	81			
33,	73,	83	71,	84,	89			
34,	50,	91	72,	55,	70			
35,	85,	81	73,	70,	88			
36,	73,	87	74,	59,	78			
37,	100,	91	75,	91,	83			
38,	52,	88	76,	95,	92			

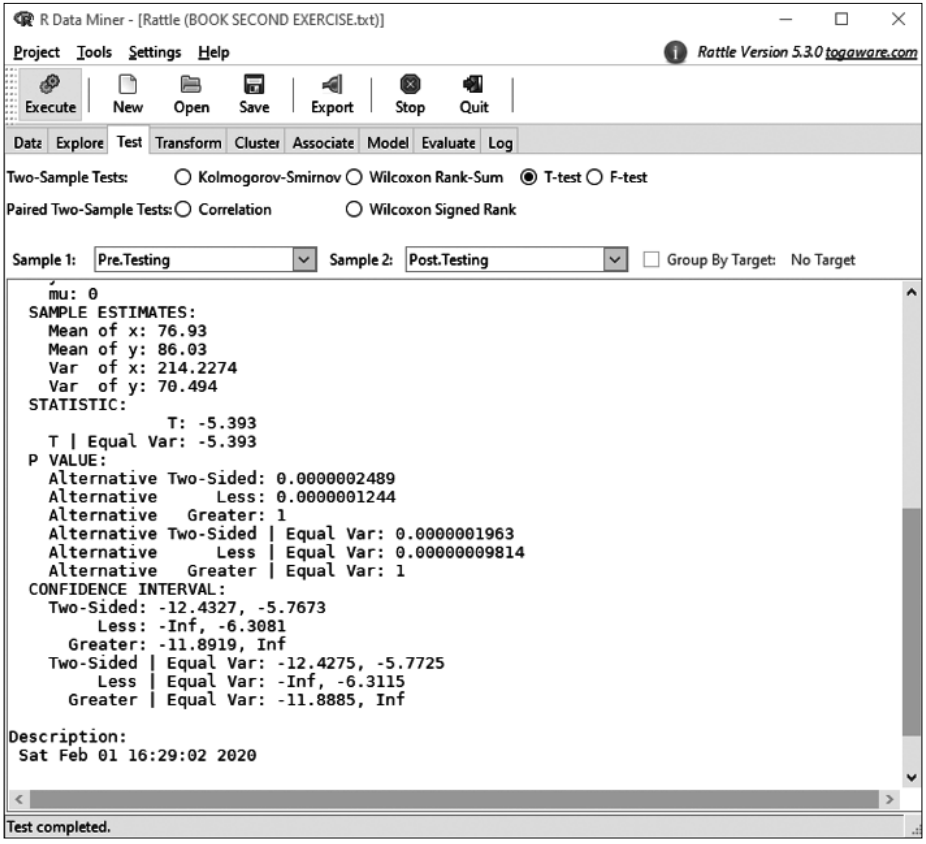
7.1.2.1 Answer to Exercise 2 – Rattle

The answer to this exercise can be solved within Rattle with just a little effort. The first step will be to import the data, which is a text file, into Rattle using the "DATA" tab and then move to "EXPLORE" in order to conduct the t-test. The finished configuration of the problem is in the next illustration. The explanation for this screen will take each of the sections and describe each for the analyst. Remember that the main reason for the test was to determine if there were in fact differences between the first and second test. Also remember that the analyst wanted a sample, which in this case was set with Rattle through the DATA tab, using the "partition" setting at 50/25/25, which means that 50% of the data would be sampled and tested.

The main area for focus in this result is the middle of the screen, where it shows the p-values for the various "tailed" tests. For the uninitiated in statistics, "tailed" tests refer to whether the analyst is testing if one sample's mean is less than or greater than the other sample's mean. In this case, it would be that the first sample (the pre-test) is less than the second sample (the post-test) if it is a right-tailed test, and a left-tailed test for the reverse (the first sample is greater than then second sample). What does this mean? It signifies that the alternative hypothesis asserts that, if the first sample mean does not equal the second sample mean, that the first sample is less than the second sample. In this case, the analyst wants to know if the post-test results are greater than the pre-test results. This would then assert that there was knowledge transfer and that the post-test shows that the students actually learned the material that they did not know during the pre-test (generally). In order to be more specific, each question between the students could be run through this testing to see if there is a significant difference between the pre-test and the post-test to see if the instructor did in fact increase the students' knowledge. This is important for the instructor, since it shows if the content helped or hindered the students. No instructor wants to see that students learn less in their classroom.

If the analyst wants to know the probability of the pre-test being less than the post-test, accounting for a chance event that the pre-test could register more than the post-test, then they would need to look at the "P-VALUE" section of the screen under "Alternative" and "Less" which gives this number: .0000001244. What this means is that there is basically a zero chance of the pre-test being less than the post-test with a 95% confidence level. It also means that, taken into consideration that chance is not a factor, that pre-test values are less overall then post-test values.

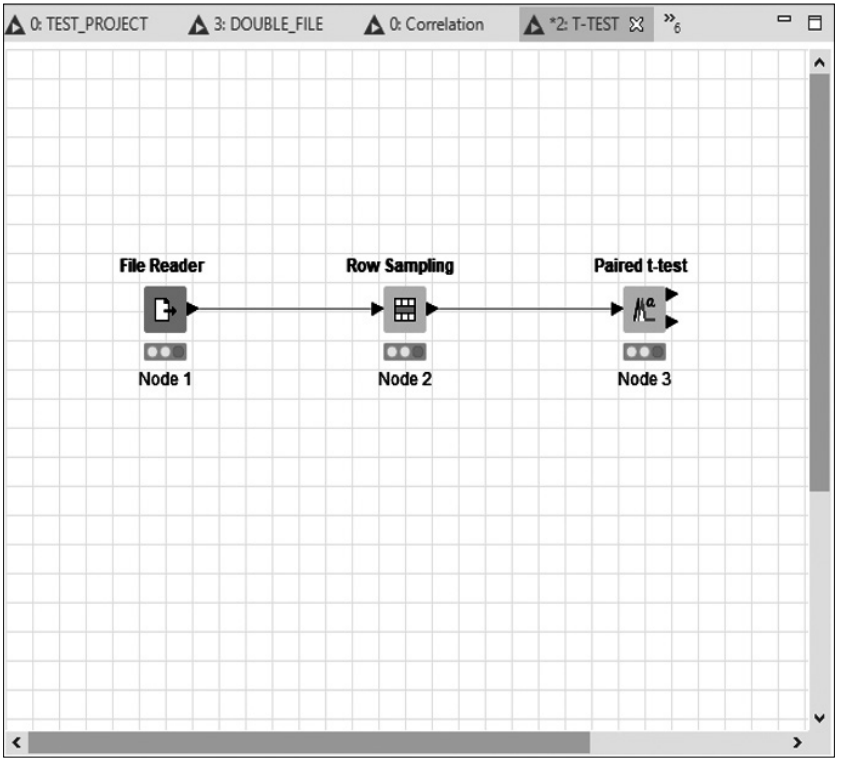
Another way of looking at this through this tool is the confidence interval, which is located on the screen at the "CONFIDENCE INTERVAL" area. If the analyst takes a look at the "Less:" they will see that the interval at 95% is –Infinity (–Inf) to –6.3081. If the analyst looks carefully, they will see that "0" is not included in that range, which means that the pre-test and post-test means are never 0, pointing to the fact that the pre-test is less than the post-test. This is one more factor in the overall statistical test and one, as stated earlier in this book, that is important in the overall analysis of data.



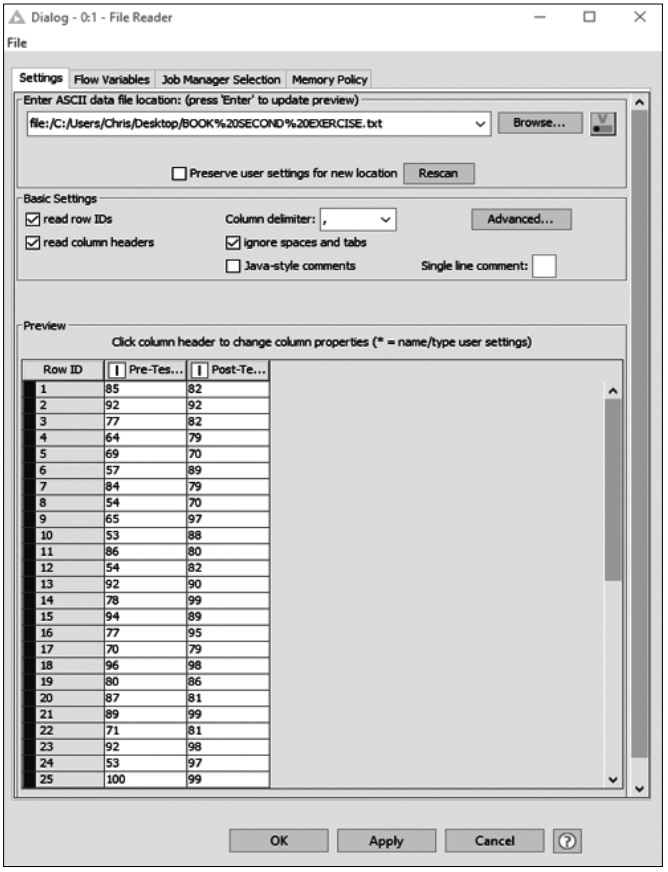
7.1.2.2 Answer to Exercise 2 – KNIME

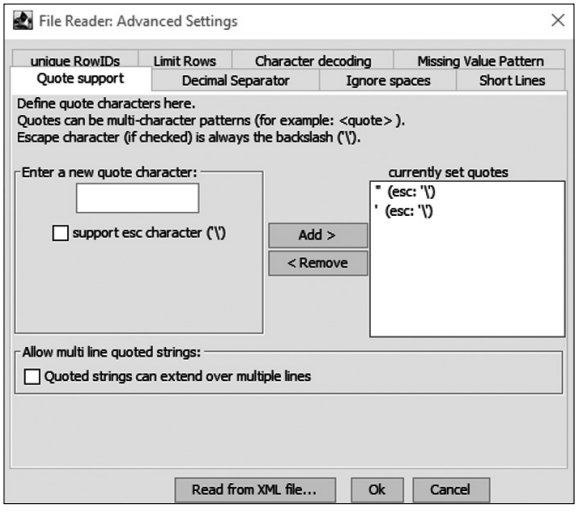
KNIME has a node that will assist with this issue, but first the analyst must import the data into the tool and work with the data. This is done the same way as in previous sections, but this time instead of using the CSV Reader

node, the analyst uses the File Reader node, which will read the text from the text file. The analyst could use the CSV Reader for this, but it is easier to use the File Reader for the possibility of the delimiter being something other than a comma. The workflow is shown in the following screen and will be explained node by node to ensure the analyst understands both the flow and the different nodes.

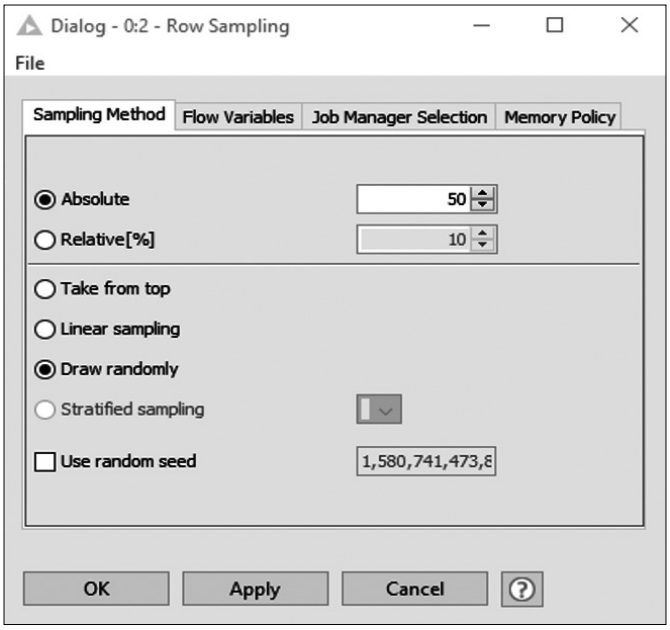


The first node is the File Reader node, which has the configuration screen as shown. Note that the checkboxes include those that would be similar to other nodes. There is also an "Advanced…" block that can be selected, and it has been included for the analyst to explore as necessary.


Dialog - 0:1 - File Reader
File
Settings
Flow Variables
Job Manager Selection
Memory Policy
Enter ASCII data file location: (press 'Enter' to update preview)
file:/C:/Users/Chris/Desktop/BOOK%20SECOND%20EXERCISE.txt
Browse...
Preserve user settings for new location
Rescan
Basic Settings
read row IDs
read column headers
Column delimiter: ,
ignore spaces and tabs
Java-style comments
Single line comment:
Advanced...
Preview
Click column header to change column properties (* = name/type user settings)
Row ID
Pre-Tes...
Post-Te...
1 85 82
2 92 92
3 77 82
4 64 79
5 69 70
6 57 89
7 84 79
8 54 70
9 65 97
10 53 88
11 86 80
12 54 82
13 92 90
14 78 99
15 94 89
16 77 95
17 70 79
18 96 98
19 80 86
20 87 81
21 89 99
22 71 81
23 92 98
24 53 97
25 100 99
OK
Apply
Cancel

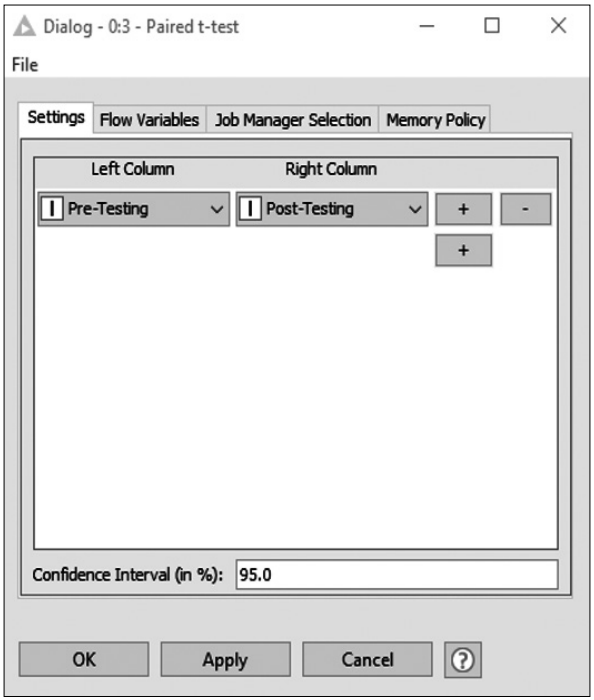

File Reader: Advanced Settings
unique RowIDs
Limit Rows
Character decoding
Missing Value Pattern
Quote support
Decimal Separator
Ignore spaces
Short Lines
Define quote characters here.
Quotes can be multi-character patterns (for example: <quote>).
Escape character (if checked) is always the backslash ('\').
Enter a new quote character:
support esc character ('\')
Add >
< Remove
currently set quotes
" (esc: '\')
' (esc: '\')
Allow multi line quoted strings:
Quoted strings can extend over multiple lines
Read from XML file...
Ok
Cancel

The second node is the Row Sampling node, which has been used before in the random sampling section. This node will sample the rows necessary to make the test valid. The "power" formula has not been used in this instance, so the original 50 rows have been sampled, which means the results for this test should not match the results from the previous answers (because they are randomly sampled rows). The configuration for this screen is as follows:



The final node is the Paired t-test node, which will perform the calculations necessary for the result. Remember to click on the Execute green arrow after every configuration change. There are no worries if you forget, since the application will give the analyst reminders that something has changed that will affect the workflow.

The result follows this screen, and the analyst should take note of the p-values that are attached along with the confidence intervals, as explained in the previous section. It is essential that the analyst consider the confidence intervals, since those are the types of test that are simple to perform and produce as effective a result as more complicated statistical tests. Remember that complicated does not mean correct.



Test statistics - 0:3 - Paired t-test

File

Paired T-Test

Paired Samples Statistics

	Column	N	Missing Count	Mean	Standard Deviation	Standard Error Mean
Pair 1	Pre-Testing	50	0	76.98	14.3506	2.0295
Pair 1	Post-Testing	50	0	84.54	8.505	1.2028

Paired Samples Test

Confidence Interval (CI) Probability: 95.0%

	Label	t	df	p-value (2-tailed)	Mean	Standard Deviation	Standard Error Mean	CI (Lower Bound)	CI (Upper Bound)
Pair 1	Pre-Testing - Post-Testing	-3.5053	49	0.001	-7.56	15.2505	2.1567	-11.8941	-3.2259

A quick look at these results shows that the p-value is .001 and the confidence interval (CI) ranges from –11 to –3. This means that "0" is not included in that range, denoting that the means are not the same. It also shows that the upper bound is –3, which means that the pre-testing is less than the post-testing, which is also extremely valuable for the analyst and can prompt some

additional testing. The analyst can make other tests that would help to specify any particular differences, but it does at least produce the result that there are in fact differences between the pre- and post-testing means.

REFERENCES

Levene, H. (1960). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (Ingram Olkin, Ed.). Redwood City, CA: Stanford University Press.

Poundstone, W. (2019). *The Doomsday Calculation: How an Equation That Predicts the Future is Transforming Everything We Know about Life and the Universe.* New York, NY: Hachette Book Group.

Provost, F., and Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.* Sebastopol: O'Reilly.

Reinhart, A. (2015). *Statistics Done Wrong: The Woefully Complete Guide.* San Francisco, CA: No Starch Press.

Sankhar, A. (2018, November 30). *How to Create a WordCloud in R*. Analytics Training. https://analyticstraining.com/how-to-create-a-word-cloud-in-r/.

Statistical Consultants Limited. (2011, May 14). *Benford's Law and Accounting Fraud Detection*. Statistical Consultants Limited. www.statisticalconsultants.co.nz/blog/benfords-law-and-accounting-fraud-detection.html

Technology, N. I. (2013, October 30). *Levene Test for Equality of Variances*. Engineering Statistics Handbook. https://www.itl.nist.gov/div898/handbook/index.htm.

Williams, G. (2011). *Data Mining with Rattle and R.* New York, NY: Springer Science+Business Media.

INDEX