NCDC Storm Events Database Analysis

MGMT 582: BUSINESS INTELLIGENCE/ANALYTICS – FALL 2016

Jack Welch

Synopsis

I am an MBA student at Bridgewater State University, enrolled in MGMT 582 – Business Intelligence/Analytics. This is a study in data mining using public data sources published at Data.gov with a popular data mining software application called R Studio. While completing my course at BSU in the Fall of 2016, I supplemented my studies with an online course offered from Coursera.org, entitled Reproducible Research from John Hopkins University. My studies in reproducible research demonstrated the importance of documentation in the data mining process as well as the advantages of sharing your research for the benefit of yourself and others. An instrumental part of the conclusion of this capstone project at BSU will include the publication of my findings at GitHub.com in a reproducible format using the R Markdown language and the Knitr Package which are important built-in features of the R Studio graphical user interface or GUI.

After browsing the vast amount of data that has been published at Data.gov, I elected to work with the *NCDC Storm Events Database* published and maintained by the US National Oceanic and Atmospheric Administration (NOAA).[1] Storm event data is so important as storms can be a source of tremendous impact on property and on our health here in the United States and across the globe. With a rise in global temperatures, experts are predicting that there will a tremendous increase in the frequency of occurrence and the magnitude of weather events. The purpose of this study will be to see if we can visualize a trend in the historical data that supports these claims.

Data.gov & Data Source

Data.gov is the United States Government repository for its data resources which have been opened for access to the general public. According to Wikipedia, President Obama and his administration, opened this website in March of 2009 for the purpose encouraging data enthusiasts and professionals all over the world to explore and to study this data for our mutual benefit. Data is a powerful tool today which allows us to make more intelligent predictions about our future based on historical patterns within existing behavior. Our government recognizes the power of putting this data into the hands of the public and to then encourage that same public to affect change based on our study of this past. Today, Data.gov maintains nearly 200K data sets categorized as Agriculture, Climate, Consumer, Education, Energy, Finance, Health, Science and more.

The National Weather Service has collected storm event data from 1950 to present. NOAA maintains this data on the following page at Data.gov - NCDC Storm Events Database. This data contains statistics related to personal injuries, property damage, casualties and more. This dataset is organized to record the different types of storms, the date of occurrence, and the magnitude of their impact on our environment. The data is stored in CSV files on an FTP server at the following URL - NOAA FTP site. You will notice too that complete documentation is available at this FTP site that describes the data format which has been used to prepare these data export files from 1950 to present. The analysis presented

herein will maintain complete integrity with this original data source as we will utilize a script which will get this data directly from its source. Any manipulation of this data necessary to "clean" it for the purposes of our analysis will be completely documented and automated with our R programming scripts. We will address this important step in further detail within the Data Analysis section of this report. Further, one of the primary focuses of this study is to use R Studio to conduct our data mining techniques. By this, I mean that R Studio is an effective tool which will allow us to get to know this data set even where documentation of its format is unavailable. Some of the first few steps of our data mining study will be to explore the overall scope of this data set including the column names, number of observations, and patterns which exist within the dataset.

R Studio

R Studio is an extremely popular choice for a graphical user interface and development platform capable of conducting data mining and advanced analytical techniques on a variety of data sources. R Studio is open source, distributed for FREE, and installs quite easily on popular personal computers. R Studio supports a host of default and custom loadable R Packages like ggplot, dplyr, R Markdown, Knitr and more. An R package is a distributed software solution or library which simplifies the act of writing custom R programs. An R package is generally a collection of specialized functions and algorithms which provides the heart of a custom program application or analytical technique.

GitHub.com & Git Bash

GitHub.com is a cloud-based repository for open-source software storage, distribution, and development. GitHub not only offers a place to host software source-code for convenient distribution to others, but GitHub encourages collaborative development with others with full-featured software version control called Git. Software developers can "branch" or "fork" an existing public repository, download the software to his/her personal computer, modify the software for general improvement, and then publish these changes back to the cloud for adoption and approval by the software owner. Another software tool called GitBash allows software developers the ability to transfer software to and from a GitHub repository. WikiPedia reports that, as of April 2016, GitHub reports more than 14 million users and more than 35 million repositories. Now that is a sign of world-wide collaboration and development.[5]

Reproducible Research

This data analysis project will make use of a host of R packages yet I want to highlight the importance of the use of a dynamic report generating tool like Knitr. Knitr is an R package, deployed within R Studio, which allows a data analyst to write and to execute R programming steps inside a descriptive report. This technique allows a data analyst to document and to explain the research techniques from data source to visualization. Effective use of Knitr allows a data mining project to be reproducible. Reproducibility adds credibility to an analytical technique and most importantly it speeds further exploration and study of our vast data resources. Data is coming available to us all in high volumes, rapid speed, and every increasing size. Effective management of a complex data analysis project counts on the documentation and repeatability of each and every step of the process.

Data Analysis

Get Data from DATA.gov

My first step in the preparation of my analysis was the writing of a script to get the data at Data.gov. After following instructions provided at Data.gov, I found that the raw data sits in zipped files located on an FTP server at <ftp://ftp.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/>. I prepared a script that automatically gets this data from its FTP location, unzips each file, and then appends each file to a larger master file that will ultimately be held in memory for further analysis within R Studio. The development of this script was an iterative process, as I needed to be cautious of memory requirements.

With my first pass at the collection of data from Data.gov, I collected 10 years of data back to 2006. I found that my personal computer could handle this task easily, and the resulting stormdata.csv file was approximately 200MB in size. Since my computer has nearly 8GB of RAM, I decided to go ahead and collect the full 65 years of storm data. After pulling 65 years of data, I observed that my final stormdata.csv file was only 400MB in size. This appears to be a very manageable data set that can be conveniently explored with R Studio.

The only "cleaning" or modification of this data that we implemented was the removal the episode and event narrative columns contained within the original data sets. The columns were removed to reduce the size of the assembled file as this process removed lengthy text fields from the data sets. Our focus within this analysis will be to look for trends in storm magnitude and destruction. So the important fields within each weather event observation will be the date, storm type, magnitude, and destruction information.

Initial Table Exploration and Observation

After successfully assembling 65 years of data from the NCDC Storm Events Database, we now have the ability to hold in memory a dataset with storm event data from calendar year 1951 through the summer of 2016. With the R dim() function, we can quickly observe that the dataset contains 1,401,817 observations of 50 different variables.

```
# Check the dimensionality
dim(stormdata)
## [1] 1401817      50
```

This simply means that our table contains 1,401,817 rows and 50 columns. If we now want to observe the contents of these columns, we can make use of the names() function:

```
# Output the names of the columns
names(stormdata)
##  [1] "X"                    "BEGIN_YEARMONTH"     "BEGIN_DAY"
##  [4] "BEGIN_TIME"           "END_YEARMONTH"       "END_DAY"
```

```
##  [7] "END_TIME"           "EPISODE_ID"         "EVENT_ID"
## [10] "STATE"              "STATE_FIPS"         "YEAR"
## [13] "MONTH_NAME"         "EVENT_TYPE"         "CZ_TYPE"
## [16] "CZ_FIPS"            "CZ_NAME"            "WFO"
## [19] "BEGIN_DATE_TIME"    "CZ_TIMEZONE"        "END_DATE_TIME"
## [22] "INJURIES_DIRECT"    "INJURIES_INDIRECT"  "DEATHS_DIRECT"
## [25] "DEATHS_INDIRECT"    "DAMAGE_PROPERTY"    "DAMAGE_CROPS"
## [28] "SOURCE"             "MAGNITUDE"          "MAGNITUDE_TYPE"
## [31] "FLOOD_CAUSE"        "CATEGORY"           "TOR_F_SCALE"
## [34] "TOR_LENGTH"         "TOR_WIDTH"          "TOR_OTHER_WFO"
## [37] "TOR_OTHER_CZ_STATE" "TOR_OTHER_CZ_FIPS"  "TOR_OTHER_CZ_NAME"
## [40] "BEGIN_RANGE"        "BEGIN_AZIMUTH"      "BEGIN_LOCATION"
## [43] "END_RANGE"          "END_AZIMUTH"        "END_LOCATION"
## [46] "BEGIN_LAT"          "BEGIN_LON"          "END_LAT"
## [49] "END_LON"            "DATA_SOURCE"
```

From this step above, we can quickly observe that we have the fields available in order to complete a trend analysis showing storm magnitude, frequency, and damage from weather events over time. We will certainly be focused on the fields highlighted in yellow but before we charge ahead, it is important to better understand this data set.

It is important to check and make sure that we have the proper format for the fields where we need them. By this, I mean we need text where we expect it and numbers where we need them. We simply can not sum up text fields and sometimes data transformation or further "cleaning" is necessary at this stage.

We will make use of the str() function in order to explore the data structure of the dataframe being held in memory:

```
# List the structure of the file
str(stormdata)
```

```
## 'data.frame':    1401817 obs. of  50 variables:
##  $ X                 : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ BEGIN_YEARMONTH   : int  195109 195106 195103 195105 195107 195105 1951
03 195105 195106 195107 ...
##  $ BEGIN_DAY         : int  9 17 28 9 15 8 30 11 27 21 ...
##  $ BEGIN_TIME        : int  915 2200 510 1830 1620 1800 1500 1330 2204 110
0 ...
```

```
##  $ END_YEARMONTH     : int  195109 195106 195103 195105 195107 195105 1951
03 195105 195106 195107 ...

##  $ END_DAY           : int  9 17 28 9 15 8 30 11 27 21 ...

##  $ END_TIME          : int  915 2200 510 1830 1620 1800 1500 1330 2204 110
0 ...

##  $ EPISODE_ID        : int  NA NA NA NA NA NA NA NA NA NA ...

##  $ EVENT_ID          : int  10047282 10028729 10120421 10099717 10099742 1
0028691 10104933 10104934 10104935 10104936 ...

##  $ STATE             : Factor w/ 69 levels "","ALABAMA","ALASKA",..: 40 26
61 52 52 26 54 54 54 54 ...

##  $ STATE_FIPS        : int  28 20 48 40 40 20 42 42 42 42 ...

##  $ YEAR              : int  1951 1951 1951 1951 1951 1951 1951 1951 1951 1
951 ...

##  $ MONTH_NAME        : Factor w/ 12 levels "April","August",..: 12 7 8 9 6
9 8 9 7 6 ...

##  $ EVENT_TYPE        : Factor w/ 74 levels "Astronomical Low Tide",..: 63
63 63 63 63 63 63 63 63 63 ...

##  $ CZ_TYPE           : Factor w/ 3 levels "C","M","Z": 1 1 1 1 1 1 1 1 1 1
...

##  $ CZ_FIPS           : int  87 63 225 33 73 203 1 111 5 15 ...

##  $ CZ_NAME           : Factor w/ 5135 levels "","5NM E OF FAIRPORT MI TO R
OCK ISLAND PASSAGE",..: 2507 1673 1939 1011 2195 4978 17 4003 116 356 ...

##  $ WFO               : Factor w/ 542 levels "","$AC","$AG",..: NA NA NA NA
NA NA NA NA NA NA ...

##  $ BEGIN_DATE_TIME   : Factor w/ 754787 levels "01-APR-00 00:00:00",..: 22
3320 409904 687802 219403 357636 193110 733184 271822 660342 505338 ...

##  $ CZ_TIMEZONE       : Factor w/ 26 levels "AKST-9","AST",..: 7 7 7 7 7 7
7 7 7 7 ...

##  $ END_DATE_TIME     : Factor w/ 746369 levels "01-APR-00 00:05:00",..: 21
9760 404183 679406 215886 352362 190068 724347 267679 652124 498484 ...

##  $ INJURIES_DIRECT   : int  0 0 0 0 0 0 1 0 0 ...

##  $ INJURIES_INDIRECT : int  0 0 0 0 0 0 0 0 0 ...

##  $ DEATHS_DIRECT     : int  0 0 0 0 0 0 0 0 0 ...

##  $ DEATHS_INDIRECT   : int  0 0 0 0 0 0 0 0 0 ...

##  $ DAMAGE_PROPERTY   : Factor w/ 2618 levels "",".01K",".01M",..: 206 206
924 924 47 47 924 1168 924 47 ...

##  $ DAMAGE_CROPS      : Factor w/ 1108 levels "",".01K",".01M",..: 59 59 59
59 59 59 59 59 59 59 ...

##  $ SOURCE            : Factor w/ 73 levels "","911 Call Center",..: NA NA
NA NA NA NA NA NA NA NA ...
```

```
##  $ MAGNITUDE          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MAGNITUDE_TYPE     : Factor w/ 7 levels "","E","EG","ES",..: NA NA NA NA
NA NA NA NA NA NA ...
##  $ FLOOD_CAUSE        : Factor w/ 8 levels "","Dam / Levee Break",..: NA NA
NA NA NA NA NA NA NA NA ...
##  $ CATEGORY           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TOR_F_SCALE        : Factor w/ 14 levels "","EF0","EF1",..: 11 9 10 10 1
1 10 10 11 11 9 ...
##  $ TOR_LENGTH         : num  0.1 0.7 0.5 0 0 0 0.1 8 19.7 0.1 ...
##  $ TOR_WIDTH          : num  100 33 17 33 100 33 20 33 33 33 ...
##  $ TOR_OTHER_WFO      : Factor w/ 90 levels "","ABQ","ABR",..: NA NA NA NA
NA NA NA NA NA NA ...
##  $ TOR_OTHER_CZ_STATE: Factor w/ 40 levels "","AL","AR","CA",..: NA NA NA
NA NA NA NA NA NA NA ...
##  $ TOR_OTHER_CZ_FIPS : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TOR_OTHER_CZ_NAME : Factor w/ 697 levels "","ADAIR","ADAMS",..: NA NA N
A NA NA NA NA NA NA NA ...
##  $ BEGIN_RANGE        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BEGIN_AZIMUTH      : Factor w/ 31 levels "","E","Eas","EE",..: NA NA NA
NA NA NA NA NA NA NA ...
##  $ BEGIN_LOCATION     : Factor w/ 58057 levels "","- 1 N Albion",..: NA NA
NA NA NA NA NA NA NA NA ...
##  $ END_RANGE          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ END_AZIMUTH        : Factor w/ 24 levels "","E","ENE","ENESE",..: NA NA
NA NA NA NA NA NA NA NA ...
##  $ END_LOCATION       : Factor w/ 47039 levels "","- .5 NNW",..: NA NA NA N
A NA NA NA NA NA NA ...
##  $ BEGIN_LAT          : num  33.5 39.1 31.4 34.5 34.8 ...
##  $ BEGIN_LON          : num  -88.4 -100.2 -95.6 -98.3 -94.8 ...
##  $ END_LAT            : num  NA NA NA NA NA ...
##  $ END_LON            : num  NA NA NA NA NA ...
##  $ DATA_SOURCE        : Factor w/ 4 levels "CSV","PDC","PDS",..: 4 4 4 4 4
4 4 4 4 4 ...
```

This is such an important step and one that will hurt your data analytics efforts if it is missed. From this step above, we can observe that most of the variables are in a format which we will find useful. We have integer types where we need them and descriptive text fields where they are necessary as well. We do however have an issue with the DAMAGE_PROPERTY and DAMAGE_CROPS fields. These fields

are in a type called 'factor' which by observation is a combination of numbers and text. This data format does not allow us to perform any type of direct mathematical computation and these fields will require further exploration and transformation. In the field of data analytics, this process is often called "cleaning".

Data Cleansing

Per the documentation contained on the FTP site, the DAMAGE_PROPERTY and DAMAGE_CROP fields contain numerical digits followed by the letter "K" for thousands of dollars, the letter "M" denoting millions of dollars, and the letter "B" for billions. If we are to use these data fields, we should take steps within our analysis to create a new numerical field on which we can complete computational techniques. For the purposes of this study, we will not need this data field so we will opt to ignore this condition now.

Trend Analysis

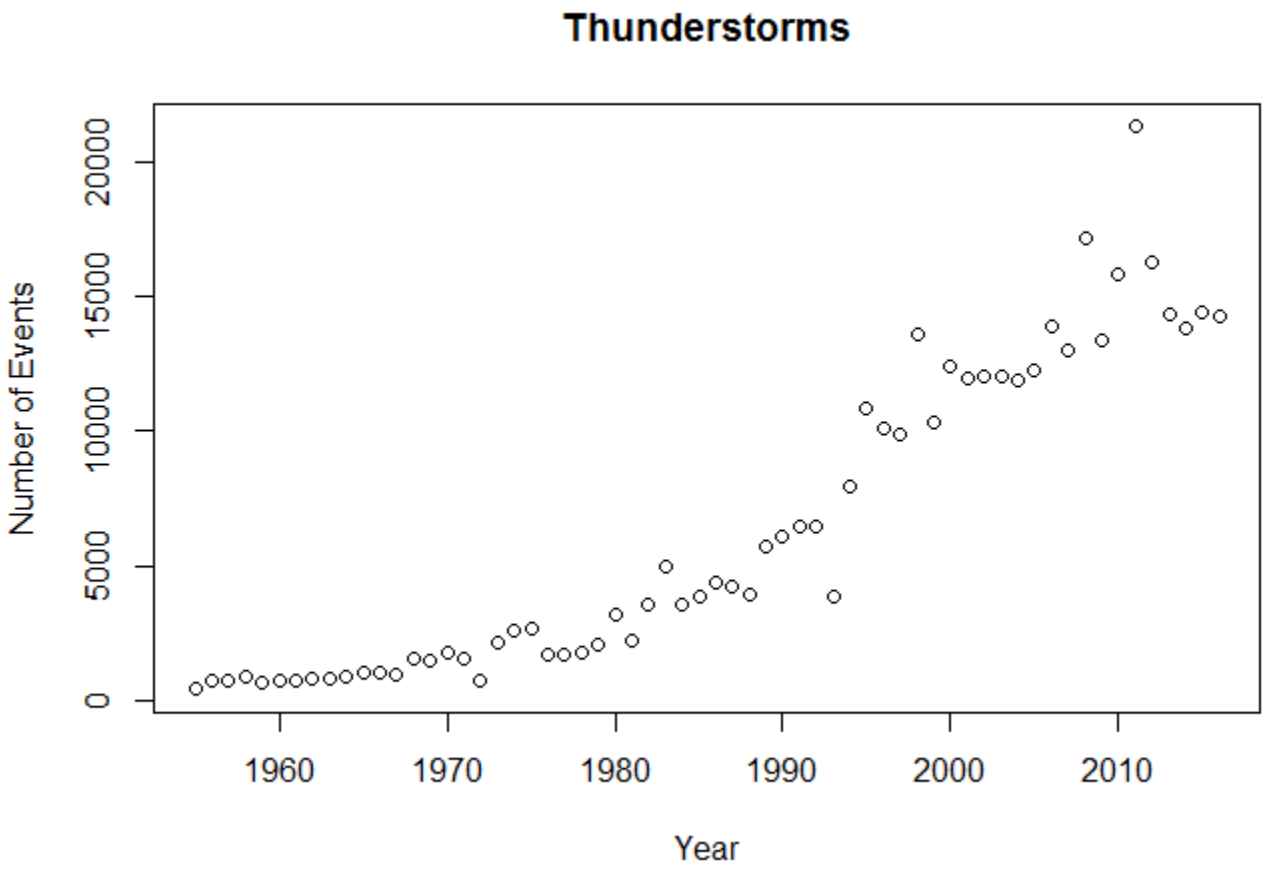
It is now desirable to complete a trend analysis whereby we can visualize the frequency of events by year. We want to be able to see if there is an upward trend in the number of occurrences of these major weather events. I will complete my trend analysis on five of the top ten most frequently occurring types of events in the US.

To complete my analysis, I need to perform a sub-setting technique from the original data frame whereby I extract the data that is subject to analysis. I want to perform, with the use of R Studio, the equivalent of an SQL query formatted as follows:

```
SELECT YEAR, COUNT(*)
        FROM stormdata
        WHERE EVENT_TYPE='Thunderstorm Wind'
        GROUP BY YEAR
        ORDER BY YEAR
```

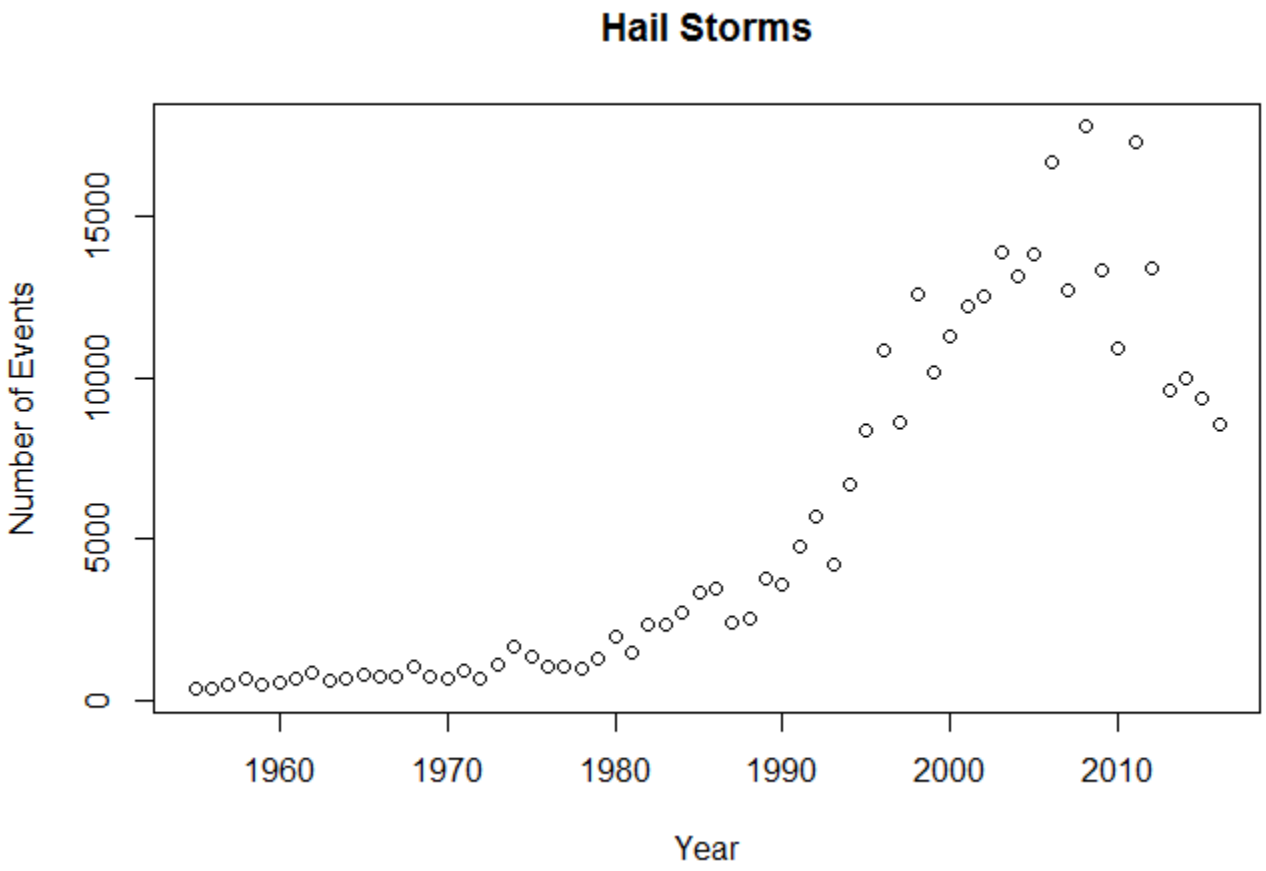
Thunderstorm Events

```
# Subset the stormdata data frame
thunderstorms <- stormdata[stormdata$EVENT_TYPE=="Thunderstorm
Wind",c('YEAR', 'EVENT_TYPE', 'INJURIES_DIRECT', 'INJURIES_INDIRECT',
'DEATHS_DIRECT', 'DEATHS_INDIRECT', 'DAMAGE_PROPERTY', 'DAMAGE_CROPS',
'MAGNITUDE', 'MAGNITUDE_TYPE')] %>% group_by(YEAR) %>% summarise(number =
n())

# Create Simple Plot
plot(thunderstorms$YEAR, thunderstorms$number, type="p",
main="Thunderstorms", xlab="Year", ylab="Number of Events")
```



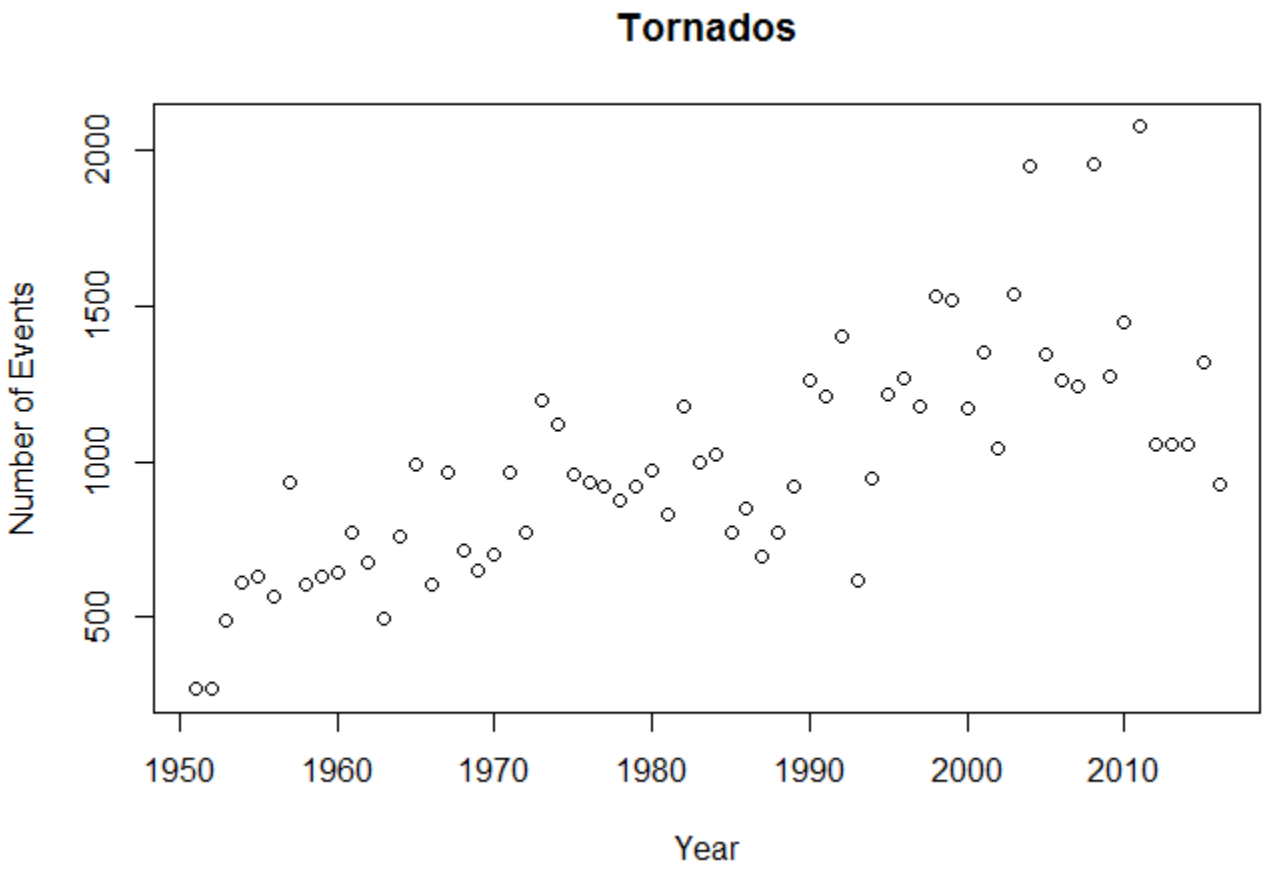
Hail Storm Events

```
# Subset the stormdata data frame
hailstorms <- stormdata[stormdata$EVENT_TYPE=="Hail",c('YEAR', 'EVENT_TYPE',
'INJURIES_DIRECT', 'INJURIES_INDIRECT', 'DEATHS_DIRECT', 'DEATHS_INDIRECT',
'DAMAGE_PROPERTY', 'DAMAGE_CROPS', 'MAGNITUDE', 'MAGNITUDE_TYPE')] %>%
group_by(YEAR) %>% summarise(number = n())

# Create Simple Plot
plot(hailstorms$YEAR, hailstorms$number, type="p", main="Hail Storms",
xlab="Year", ylab="Number of Events")
```



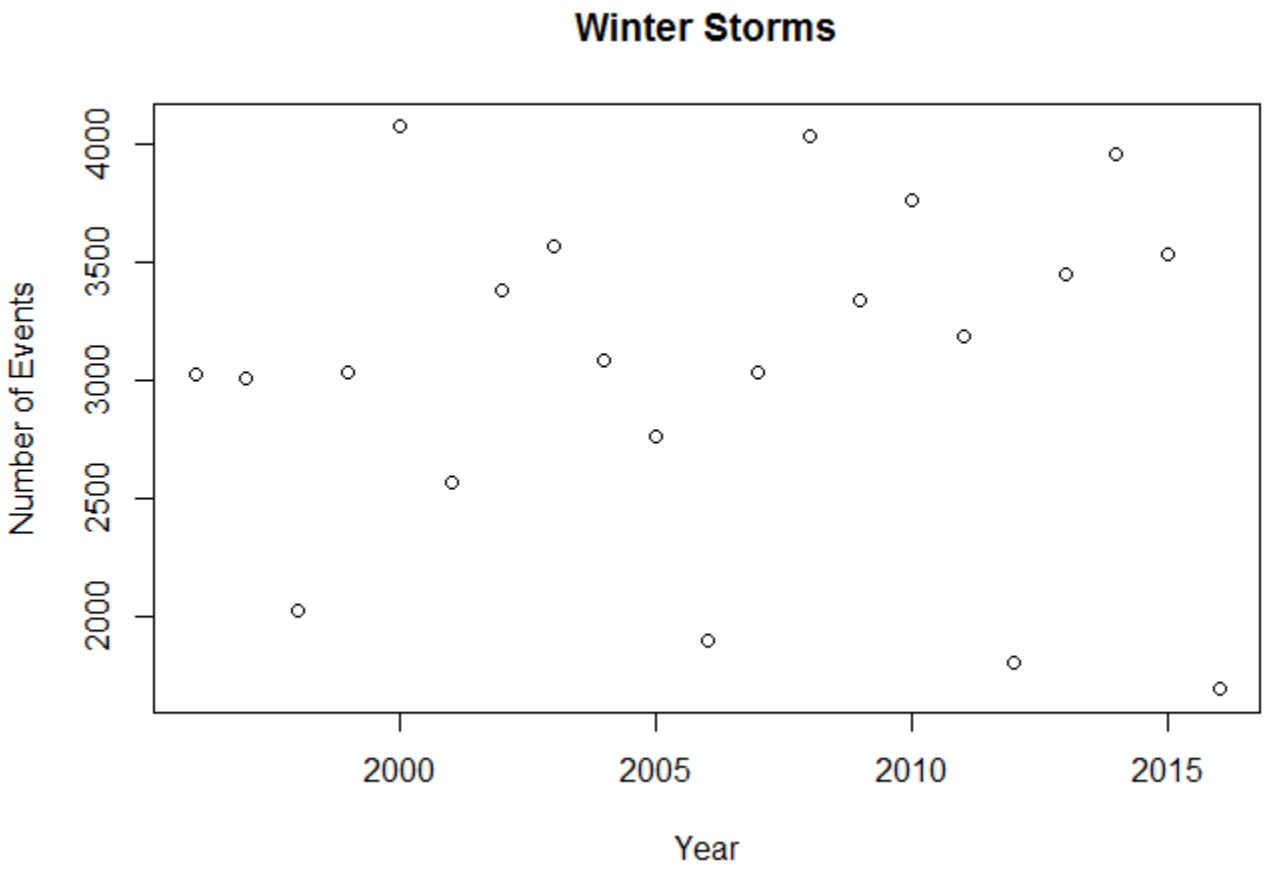
Tornado Events

```
# Subset the stormdata data frame
tornados <- stormdata[stormdata$EVENT_TYPE=="Tornado",c('YEAR', 'EVENT_TYPE',
'INJURIES_DIRECT', 'INJURIES_INDIRECT', 'DEATHS_DIRECT', 'DEATHS_INDIRECT',
'DAMAGE_PROPERTY', 'DAMAGE_CROPS', 'MAGNITUDE', 'MAGNITUDE_TYPE')] %>%
group_by(YEAR) %>% summarise(number = n())

# Create Simple Plot
plot(tornados$YEAR, tornados$number, type="p", main="Tornados", xlab="Year",
ylab="Number of Events")
```



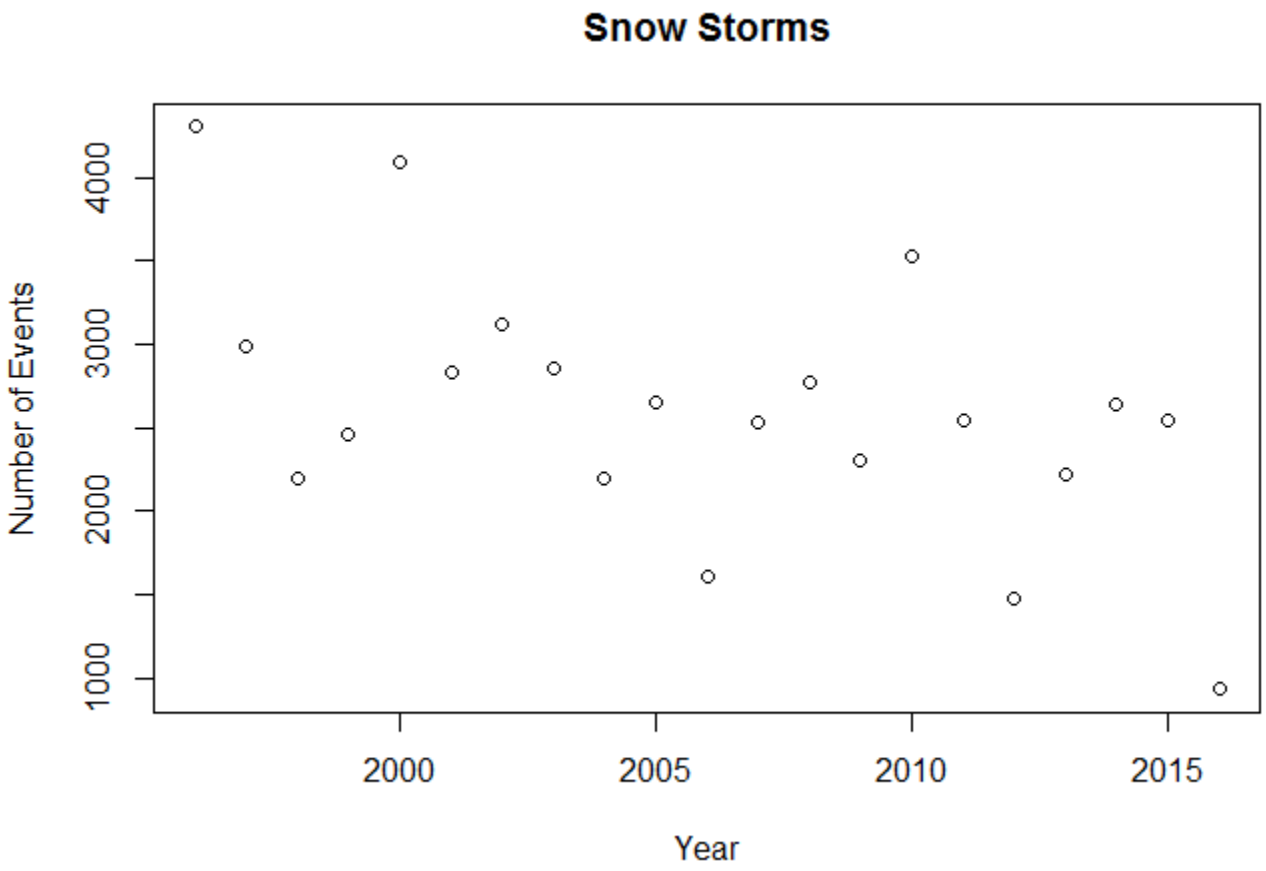
Winter Storm Events

```
# Subset the stormdata data frame
winterstorms <- stormdata[stormdata$EVENT_TYPE=="Winter Storm",c('YEAR',
'EVENT_TYPE', 'INJURIES_DIRECT', 'INJURIES_INDIRECT', 'DEATHS_DIRECT',
'DEATHS_INDIRECT', 'DAMAGE_PROPERTY', 'DAMAGE_CROPS', 'MAGNITUDE',
'MAGNITUDE_TYPE')] %>% group_by(YEAR) %>% summarise(number = n())

# Create Simple Plot
plot(winterstorms$YEAR, winterstorms$number, type="p", main="Winter Storms",
xlab="Year", ylab="Number of Events")
```



Snow Storm Events

```
# Subset the stormdata data frame
snowstorms <- stormdata[stormdata$EVENT_TYPE=="Heavy Snow",c('YEAR',
'EVENT_TYPE', 'INJURIES_DIRECT', 'INJURIES_INDIRECT', 'DEATHS_DIRECT',
'DEATHS_INDIRECT', 'DAMAGE_PROPERTY', 'DAMAGE_CROPS', 'MAGNITUDE',
'MAGNITUDE_TYPE')] %>% group_by(YEAR) %>% summarise(number = n())

# Create Simple Plot
plot(snowstorms$YEAR, snowstorms$number, type="p", main="Snow Storms",
xlab="Year", ylab="Number of Events")
```


Snow Storms
Number of Events
1000
2000
3000
4000
2000
2005
2010
2015
Year

Conclusion

This data analysis project has certainly exposed me to many tools and resources which are available to complete a data mining project using important historical data sets. We have gathered a total of 65 years of weather data for this exercise from the United States government via Data.gov, and we have explored the content of this data with the aid of R Studio. We set out to see if we could create a visualization of this data that demonstrates claims that global warming could be contributing to an increased occurrence of weather events here within the United States. The historical data certainly shows a significant increase in the occurrence of all weather events (with the only exception where there is limited historical data). In fact, the records show that the increased frequency of storms is shockingly true at what appears to be logarithmic rates. What we don't know from this data is whether the sharp upward trend reflects our increased ability to monitor and to gather this data or if there is an upward rate in actual occurrence of these weather events.

This exercise was not intended to break any ground by creating any new awareness of what has been studied many times before this one. The main purpose of my personal study was to gather an increased awareness of how to complete a data mining study with some rather interesting data.

References

1. September 1, 2016. *NCDC Storm Events Database*. Data.gov. Obtained from https://catalog.data.gov/dataset/ncdc-storm-events-database on December 3, 2016.
2. *Reproducible Research*. Coursera.org. Obtained from https://www.coursera.org/learn/reproducible-research/home/welcome on December 3, 2016.
3. *Data Exploration*. RDataMining.com. Obtained from http://www.rdatamining.com/examples/exploration on December 10, 2016.
4. *Data.gov*. Wikipedia.org. Obtained from https://en.wikipedia.org/wiki/Data.gov on December 10, 2016.
5. *GitHub*. Wikipedia.org. Obtained from https://en.wikipedia.org/wiki/GitHub on December 11, 2016.