STAT 3032 - Assignment2

Saksham Goel

January 26, 2018

A Simple Data Analysis

This assignment focus on using R and RStudio specifically to create a .Rmd file using RStudio. In this assignment I will also study the hurricanes dataset, consisting of 287 hurricanes that had made landfall in the US between 1851 and 2013.

Step 1 - Importing the Dataset

To import the dataset in R I use a simple "read.csv" command which read a csv file and then I just store all the data stored in a variable called hurricanes as follows:

```
hurricanes = read.csv("hurricanes.csv")
```

Step 2 - Printing the Dataset Summary

Once I have imported the dataset I can use a simple summary() function to get a summary of all the data from the dataset. This function provides a brief summary containing mean, median, mode, range etc for all the individual columns or variables in the dataset. The summary command and the summary is as follows:

```
summary(hurricanes)
```

```
##       Year          Month          HighestSS     Pressure
##  Min.   :1851   Min.   : 6.000   Min.   :1   Min.   : 892.0
##  1st Qu.:1888   1st Qu.: 8.000   1st Qu.:1   1st Qu.: 954.0
##  Median :1926   Median : 9.000   Median :2   Median : 967.0
##  Mean   :1928   Mean   : 8.557   Mean   :2   Mean   : 965.8
##  3rd Qu.:1964   3rd Qu.: 9.000   3rd Qu.:3   3rd Qu.: 980.0
##  Max.   :2012   Max.   :11.000   Max.   :5   Max.   :1003.0
##                                              NA's   :1
##     MaxWind                 Name           AL                CT
##  Min.   : 65.00   -----       :149   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 70.00               :  4   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 87.50   "Galveston":  3   Median :0.0000   Median :0.00000
##  Mean   : 87.99   Bob         :  3   Mean   :0.1498   Mean   :0.06272
##  3rd Qu.:100.00   Cindy       :  3   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :160.00   Bonnie      :  2   Max.   :3.0000   Max.   :3.00000
##  NA's   :43       (Other)     :123
##        DE                FL                GA                LA
##  Min.   :0.000000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.006969   Mean   :0.8328   Mean   :0.1254   Mean   :0.4007
##  3rd Qu.:0.000000   3rd Qu.:1.5000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.000000   Max.   :5.0000   Max.   :4.0000   Max.   :5.0000
##
##        MA                MD                ME                 MS
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000  
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.0000  
##  Mean   :0.05923   Mean   :0.01394   Mean   :0.02439   Mean   :0.1568  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000  
##  Max.   :3.00000   Max.   :2.00000   Max.   :2.00000   Max.   :5.0000  
##                                                                        
##        NC               NH                NJ                NY         
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000  
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000  
##  Mean   :0.3136   Mean   :0.01045   Mean   :0.01394   Mean   :0.08362  
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000  
##  Max.   :4.0000   Max.   :2.00000   Max.   :1.00000   Max.   :3.00000  
##                                                                        
##        PA                  RI                SC               TX        
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.000000   Median :0.00000   Median :0.0000   Median :0.0000  
##  Mean   :0.003484   Mean   :0.06272   Mean   :0.1742   Mean   :0.4286  
##  3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000  
##  Max.   :1.000000   Max.   :3.00000   Max.   :4.0000   Max.   :4.0000  
##                                                                        
##        VA         
##  Min.   :0.00000  
##  1st Qu.:0.00000  
##  Median :0.00000  
##  Mean   :0.04878  
##  3rd Qu.:0.00000  
##  Max.   :2.00000  
## 
```

Step 3 - Printing the First 6 rows of Dataset

Printing the actual data is easy and can be accomplished as given in the piece of code below. Here we have specified that we just want the data for the first 6 rows.

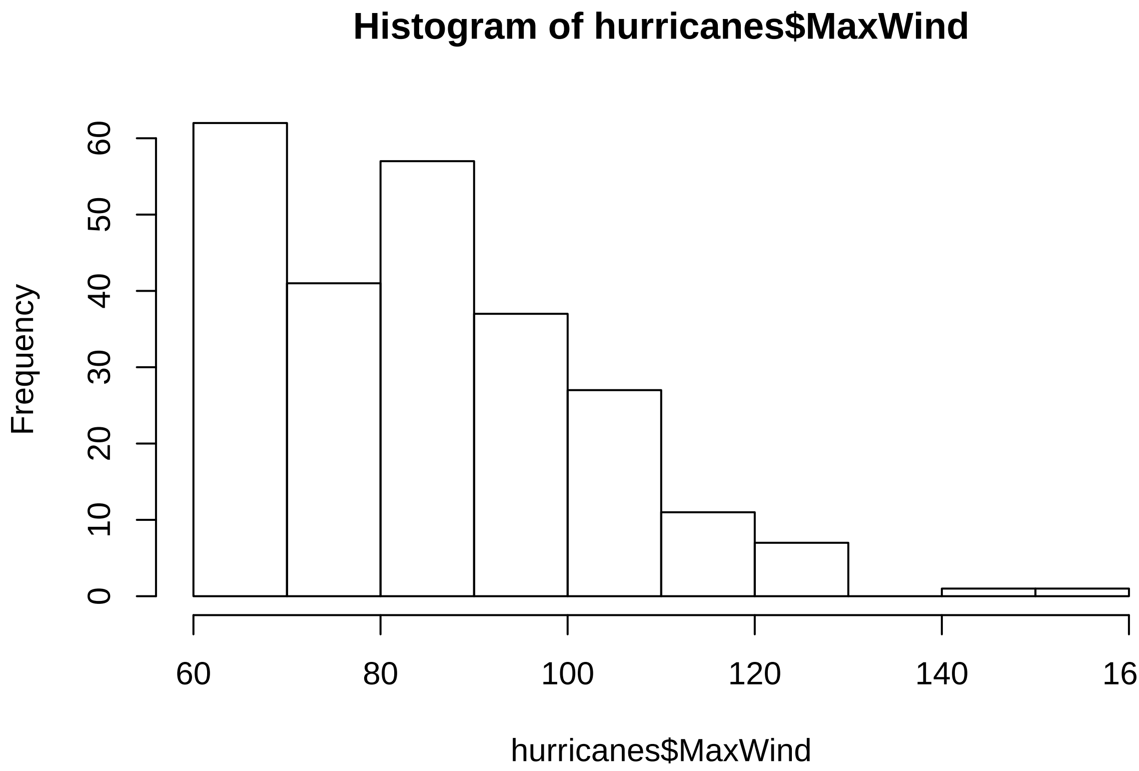
```
hurricanes[1:6,]
```

```
##   Year Month HighestSS Pressure MaxWind                   Name AL CT DE FL
## 1 1851     6         1      974      80                  -----  0  0  0  0
## 2 1851     8         3      955     100 "Great Middle Florida"  0  0  0  3
## 3 1852     8         3      961     100         "Great Mobile"  3  0  0  2
## 4 1852     9         1      982      70                  -----  0  0  0  1
## 5 1852    10         2      965      90       "Middle Florida"  0  0  0  2
## 6 1853    10         1      965      70                  -----  0  0  0  0
##   GA LA MA MD ME MS NC NH NJ NY PA RI SC TX VA
## 1  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0
## 2  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 3  0  2  0  0  0  3  0  0  0  0  0  0  0  0  0
## 4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 5  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 6  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

Step 4 - Making a Histogram

To print out a histogram containing just data regarding one column can be achived by the code as given below:

```
hist(hurricanes$MaxWind)
```



Step 5 - Analyzing the Histogram

Analyzing the Histogram: In the above histogram we can see that frequency of the first bin corresponding to 60-70 is highest. The frequency of the highest bin seems to be more than 60 and also the graph seems to follow a general trend of decreasing fequency (linear decrease) with increase in the max wind speed. One shocking thing is that the frequency of the max_wind in range 70-80 is less than that of 80-90 which does not follow the usual pattern. I suppose that this can be attributed to the fact regarding how the data was collected and also that the data is a sample and thus can have some aberrant behaviours.