Storms
Exploratory Analysis

Carson Brower, cbrower@bellarmine.edu

I. INTRODUCTION

This dataset is the NOAA Atlantic hurricane database best track data, https://www.nhc.noaa.gov/data/#hurdat. The data includes the positions and attributes of storms from 1975-2022. Storms from 1979 onward are measured every six hours during the lifetime of the storm. Storms in earlier years have some missing data. I chose this data because storms are interesting, and it had a good amount of quality variables and information to look at. I also chose this data because I thought it would be interesting to see the comparison between wind and pressure and other categories from the data.

II. DATA SET DESCRIPTION

This data set from R Studio called "storms" contains 19,537 obs. Of thirteen variables contained within the dataset. The data types include variables such as wind and pressure which are common storm variables with more listed in the table below. The data types include numerical data with nominal and interval areas. There was some missing data within the data set.

Table 1: Data Types

```
   name   year month   day  hour   lat  long status              category  wind pressure
   <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <fct>                  <dbl> <int>    <int>
 1 Amy    1975     6    27     0  27.5 -79   tropical depression       NA    25     1013
 2 Amy    1975     6    27     6  28.5 -79   tropical depression       NA    25     1013
 3 Amy    1975     6    27    12  29.5 -79   tropical depression       NA    25     1013
 4 Amy    1975     6    27    18  30.5 -79   tropical depression       NA    25     1013
 5 Amy    1975     6    28     0  31.5 -78.8 tropical depression       NA    25     1012
 6 Amy    1975     6    28     6  32.4 -78.7 tropical depression       NA    25     1012
 # i 2 more variables: tropicalstorm_force_diameter <int>, hurricane_force_diameter <int>
> str(data)
tibble [19,537 × 13] (S3: tbl_df/tbl/data.frame)
 $ name                          : chr [1:19537] "Amy" "Amy" "Amy" "Amy" ...
 $ year                          : num [1:19537] 1975 1975 1975 1975 1975 ...
 $ month                         : num [1:19537] 6 6 6 6 6 6 6 6 6 6 ...
 $ day                           : int [1:19537] 27 27 27 27 28 28 28 28 29 29 ...
 $ hour                          : num [1:19537] 0 6 12 18 0 6 12 18 0 6 ...
 $ lat                           : num [1:19537] 27.5 28.5 29.5 30.5 31.5 32.4 33.3 34 34.4 34 ...
 $ long                          : num [1:19537] -79 -79 -79 -79 -78.8 -78.7 -78 -77 -75.8 -74.8 ...
 $ status                        : Factor w/ 9 levels "disturbance",..: 7 7 7 7 7 7 7 7 8 8 ...
 $ category                      : num [1:19537] NA NA NA NA NA NA NA NA NA NA ...
 $ wind                          : int [1:19537] 25 25 25 25 25 25 25 30 35 40 ...
 $ pressure                      : int [1:19537] 1013 1013 1013 1013 1012 1012 1011 1006 1004 1002
...
 $ tropicalstorm_force_diameter: int [1:19537] NA NA NA NA NA NA NA NA NA NA ...
 $ hurricane_force_diameter    : int [1:19537] NA NA NA NA NA NA NA NA NA NA ...
```

Table 2: Missing Data

```
> sapply(data, function(x) sum(is.na(x)))
                          name                           year                          month
                             0                              0                              0
                           day                           hour                            lat
                             0                              0                              0
                          long                         status                       category
                             0                              0                          14734
                          wind                       pressure tropicalstorm_force_diameter
                             0                              0                           9512
    hurricane_force_diameter
                          9512
```

III. Data Set Summary Statistics

In this section of my findings from the data you will find two tables and a heating map. The first table represents statistical information regarding each categorical variable and the statistics based on the data set with information

like quartile range, mean, max, and min. The second table is a correlation table this provides correlations between each categorical variable within the dataset and how each variable correlate to the other. The heat map is in connection with the correlation table and explains based on colors where the darker the color the stronger the positive and negative correlation there is and the lighter the color the weaker the correlation is, based upon variable correlations.

Table 3: Summary Statistics for storms

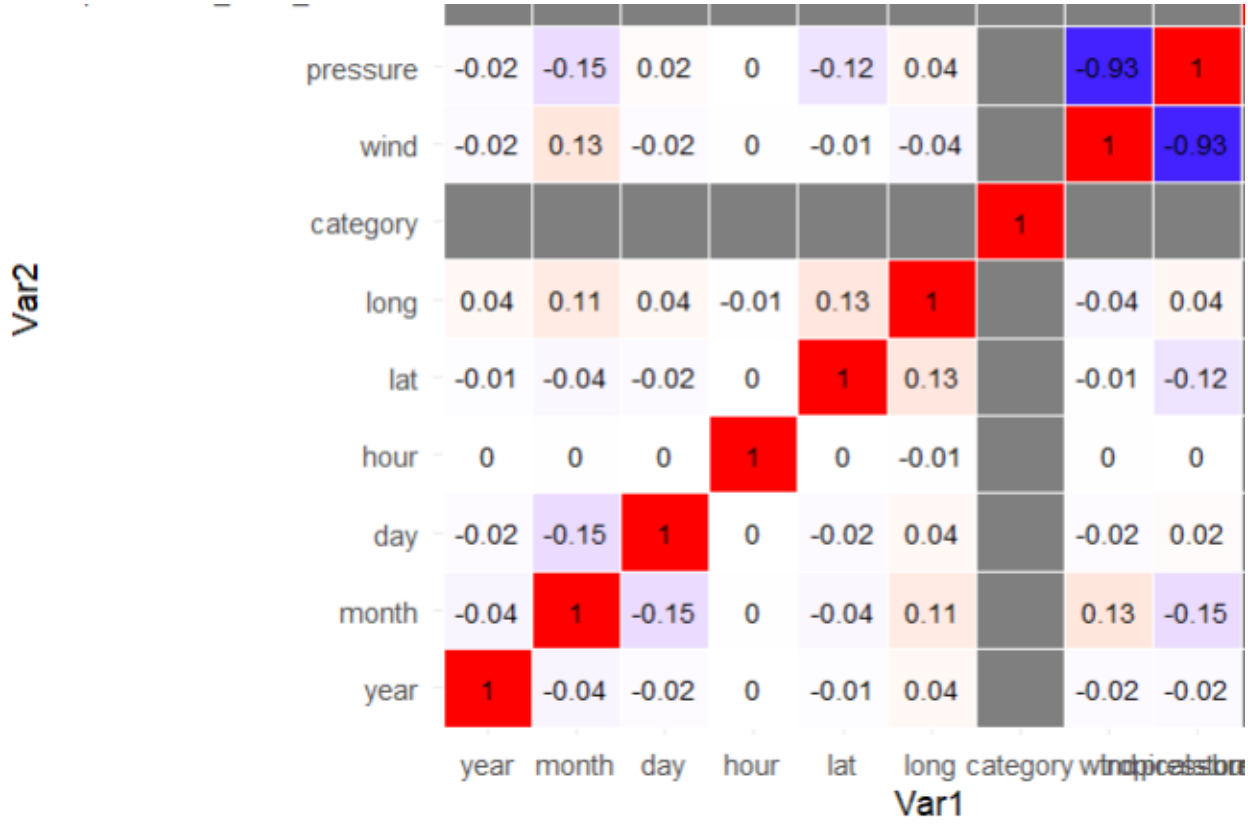
```
> summary(data)
     name                year          month             day              hour
 Length:19537       Min.   :1975   Min.   : 1.000   Min.   : 1.00   Min.   : 0.000
 Class :character   1st Qu.:1994   1st Qu.: 8.000   1st Qu.: 8.00   1st Qu.: 5.000
 Mode  :character   Median :2004   Median : 9.000   Median :16.00   Median :12.000
                    Mean   :2003   Mean   : 8.706   Mean   :15.73   Mean   : 9.101
                    3rd Qu.:2013   3rd Qu.: 9.000   3rd Qu.:24.00   3rd Qu.:18.000
                    Max.   :2022   Max.   :12.000   Max.   :31.00   Max.   :23.000

      lat             long                         status        category         wind
 Min.   : 7.00   Min.   :-136.90   tropical storm     :6830   Min.   :1.000   Min.   : 10.00
 1st Qu.:18.30   1st Qu.: -78.80   hurricane          :4803   1st Qu.:1.000   1st Qu.: 30.00
 Median :26.60   Median : -62.30   tropical depression:3569   Median :1.000   Median : 45.00
 Mean   :27.01   Mean   : -61.56   extratropical      :2151   Mean   :1.896   Mean   : 50.05
 3rd Qu.:33.80   3rd Qu.: -45.50   other low          :1453   3rd Qu.:3.000   3rd Qu.: 65.00
 Max.   :70.70   Max.   :  13.50   subtropical storm  : 298   Max.   :5.000   Max.   :165.00
                                   (Other)            : 433   NA's   :14734
    pressure      tropicalstorm_force_diameter hurricane_force_diameter
 Min.   : 882.0   Min.   :   0.0               Min.   :  0.00
 1st Qu.: 986.0   1st Qu.:   0.0               1st Qu.:  0.00
 Median :1000.0   Median : 110.0               Median :  0.00
 Mean   : 993.5   Mean   : 147.9               Mean   : 14.92
 3rd Qu.:1007.0   3rd Qu.: 220.0               3rd Qu.:  0.00
 Max.   :1024.0   Max.   :1440.0               Max.   :300.00
                  NA's   :9512                 NA's   :9512
```

Table 4: Correlation Table

	year	month	day	hour
year	1.000000e+00	-0.044164025	-0.0187583704	-1.047902e-05
month	-4.416402e-02	1.000000000	-0.1519341733	-4.814857e-03
day	-1.875837e-02	-0.151934173	1.0000000000	-2.435544e-04
hour	-1.047902e-05	-0.004814857	-0.0002435544	1.000000e+00
lat	-1.377481e-02	-0.035013211	-0.0218098000	-7.278459e-04
long	3.725233e-02	0.107107345	0.0382380232	-9.253437e-03
category	NA	NA	NA	NA
wind	-2.498871e-02	0.128475281	-0.0182713517	2.784341e-03
pressure	-1.901855e-02	-0.148516742	0.0184712706	6.135554e-04
tropicalstorm_force_diameter	NA	NA	NA	NA
hurricane_force_diameter	NA	NA	NA	NA

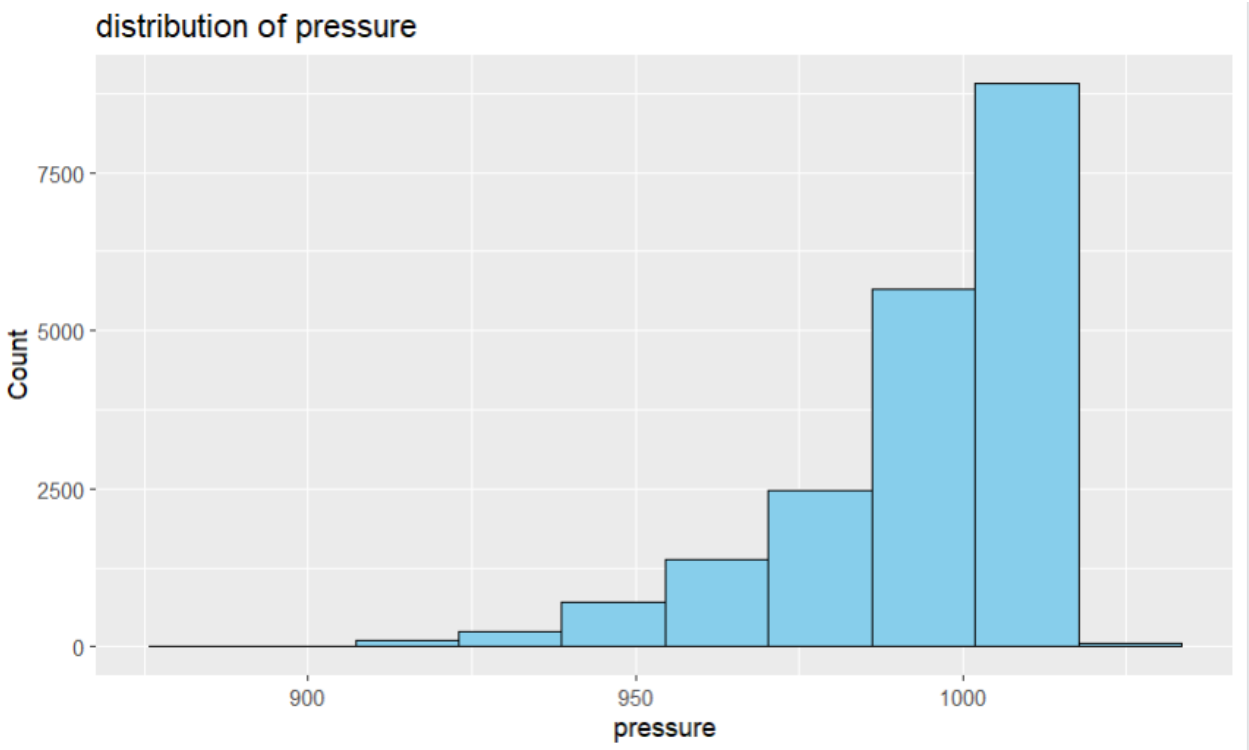
	lat	long	category	wind	pressure
year	-0.0137748059	0.037252331	NA	-0.024988710	-0.0190185469
month	-0.0350132111	0.107107345	NA	0.128475281	-0.1485167420
day	-0.0218098000	0.038238023	NA	-0.018271352	0.0184712706
hour	-0.0007278459	-0.009253437	NA	0.002784341	0.0006135554
lat	1.0000000000	0.128521819	NA	-0.005694302	-0.1161173404
long	0.1285218192	1.000000000	NA	-0.038128573	0.0418885537
category	NA	NA	1	NA	NA
wind	-0.0056943024	-0.038128573	NA	1.000000000	-0.9277700136
pressure	-0.1161173404	0.041888554	NA	-0.927770014	1.0000000000
tropicalstorm_force_diameter	NA	NA	NA	NA	NA
hurricane_force_diameter	NA	NA	NA	NA	NA

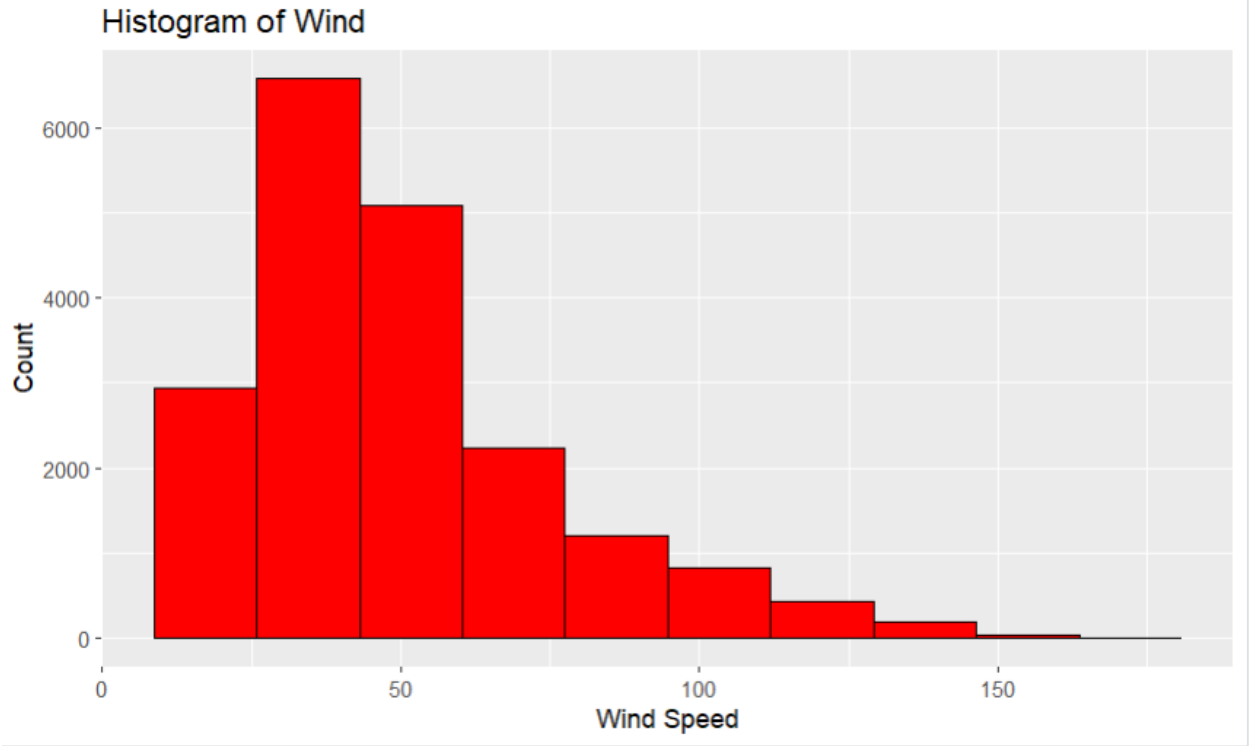
Table 5: Heatmap

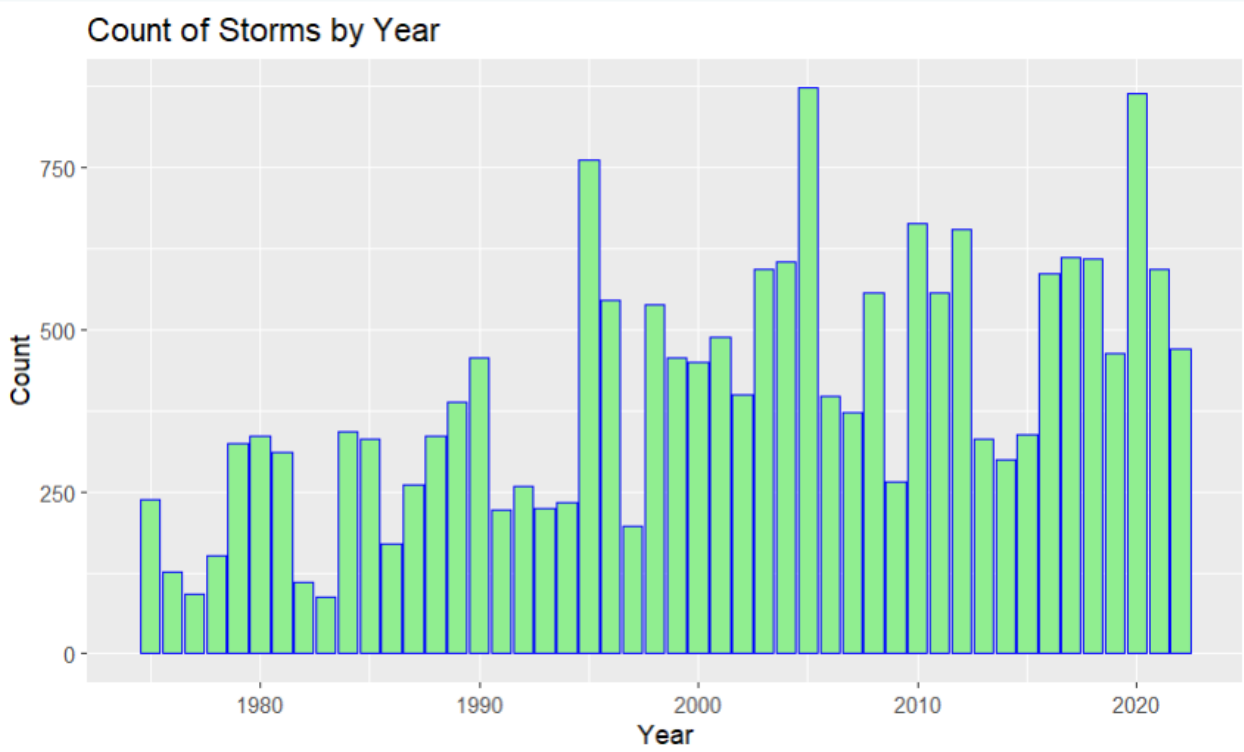


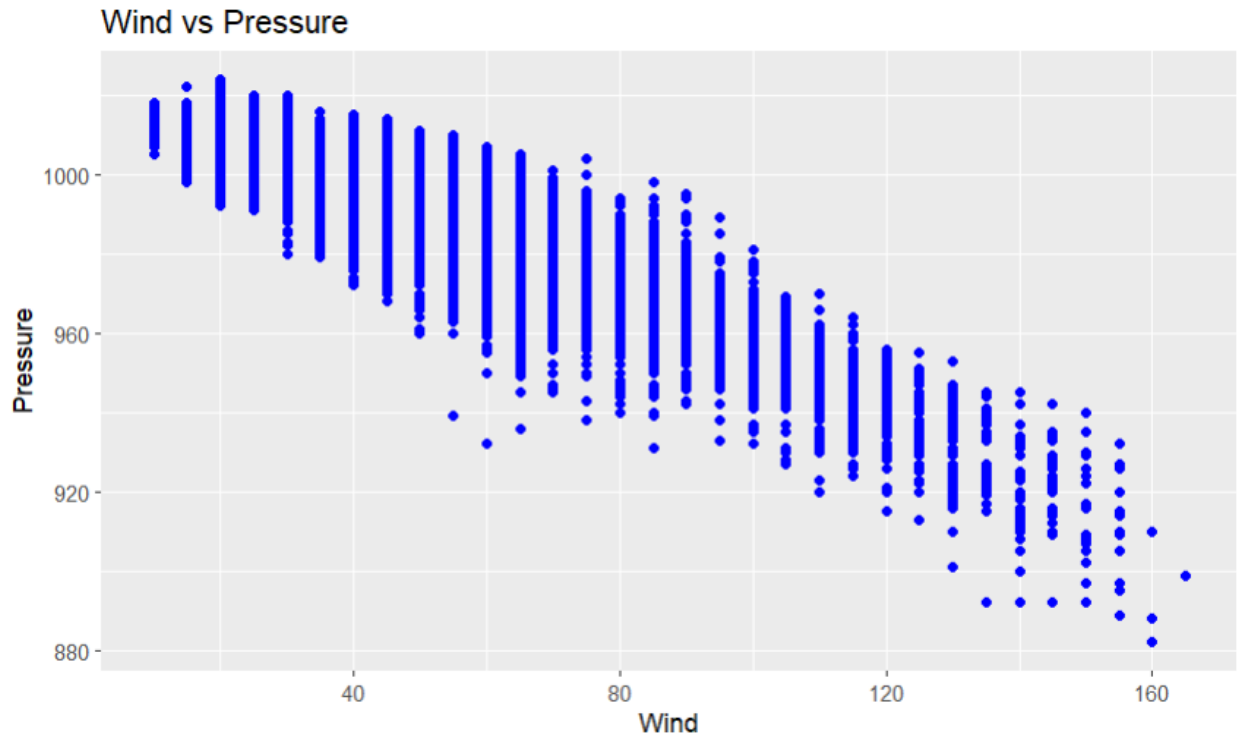
IV. DATA SET GRAPHICAL EXPLORATION

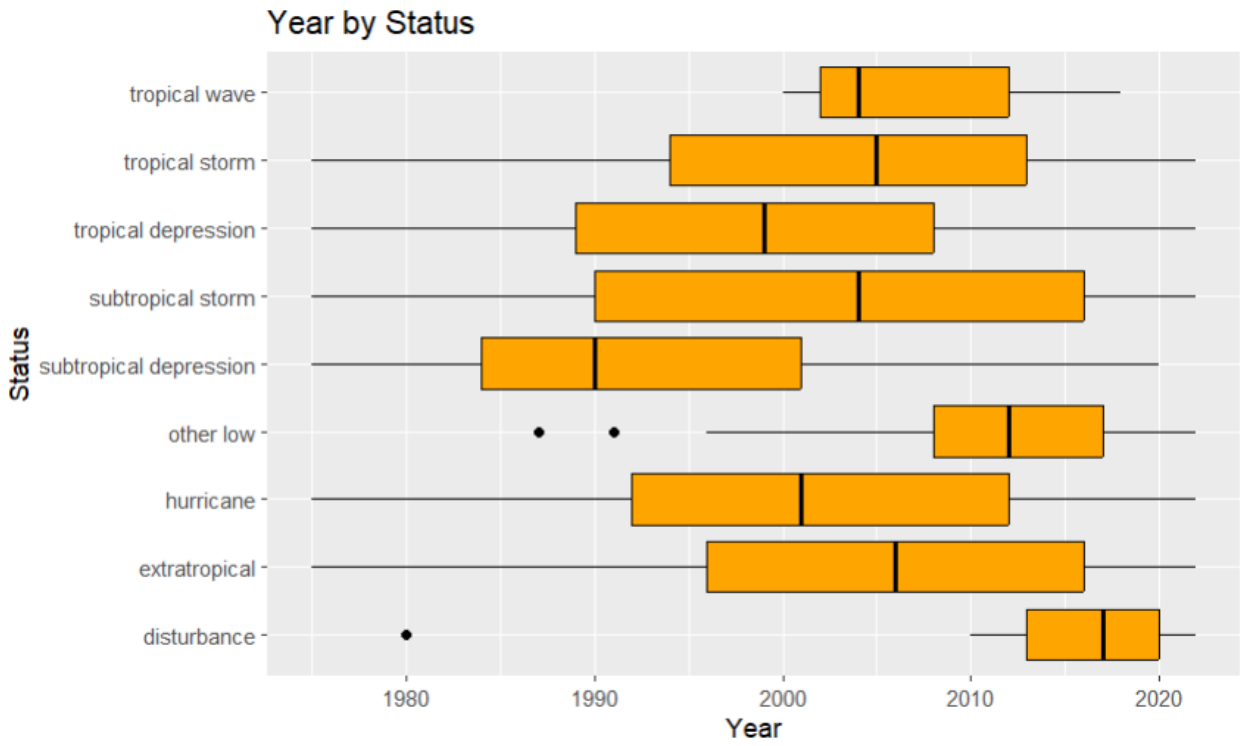
Within the data set graphical exploration section of the analysis, with the use of five graphs and charts to display our data and to answer the major question of what we were trying to find which was what variables directly correlated with each other. The first graph was a histogram displaying pressure distribution. The second graph was a histogram displaying distribution of wind. While the third graph was a bar graph comparing the number of storms per year. The fourth graph was a scatterplot comparing wind and pressure. The final graph was a boxplot displaying the storm status by year. These were the main graphs that represented the main representation of the data.


distribution of pressure
Count
7500
5000
2500
0
900
950
1000
pressure


Histogram of Wind
Count
6000
4000
2000
0
0
50
100
150
Wind Speed


Count of Storms by Year
Count
750
500
250
0
1980
1990
2000
2010
2020
Year


Wind vs Pressure
Pressure
1000
960
920
880
40
80
120
160
Wind



V. SUMMARY OF FINDINGS

After going through the dataset, finding statistical information, using multiple graphs, and discussing the data. Through this entire process the question I was looking at was what variables directly correlate with storms and an increase of larger storms? Through our exploration of the dataset, I concluded that there are certain categorical variables that are related and correlate to the question. Quite a few of the variables that are calculated from the data were pressure, wind, as well as the location of the storms. When looking at those variables from the results of the data what can be seen is the year to year the storms have become more powerful and more of them have been happening. Also, when pressure is lower from them there is higher wind which is well known in larger storms like hurricanes. The location is important as well based on whether it is closer to land which is smaller storms or out in the ocean which are larger storms. Overall, this data set was interesting to look at and gave and provided me with quality insight to variables connected with storms such as hurricanes and tropical storms.