

Mva_Project1.R

thindprateek

2020-10-01

```
##### Assignment 3 EDA, Data Cleaning and Tests #####
```

```
#Getting working directory  
getwd()
```

```
## [1] "/Users/thindprateek/Desktop/Multivariate Analysis"
```

```
#Setting directory to load dataset  
setwd("/Users/thindprateek/Desktop/Multivariate Analysis")
```

```
#Reading the data into a dataframe  
df <- read.csv(file = 'US_Accidents_June20.csv')
```

```
# Printing first few columns of dataset for inference  
head(df)
```

```
##   ID  Source TMC Severity      Start_Time      End_Time Start_Lat  
## 1 A-1 MapQuest 201      3 2016-02-08 05:46:00 2016-02-08 11:00:00 39.86515  
## 2 A-2 MapQuest 201      2 2016-02-08 06:07:59 2016-02-08 06:37:59 39.92806  
## 3 A-3 MapQuest 201      2 2016-02-08 06:49:27 2016-02-08 07:19:27 39.06315  
## 4 A-4 MapQuest 201      3 2016-02-08 07:23:34 2016-02-08 07:53:34 39.74775  
## 5 A-5 MapQuest 201      2 2016-02-08 07:39:07 2016-02-08 08:09:07 39.62778  
## 6 A-6 MapQuest 201      3 2016-02-08 07:44:26 2016-02-08 08:14:26 40.10059  
##   Start_Lng End_Lat End_Lng Distance.mi.  
## 1 -84.05872    NA     NA      0.01  
## 2 -82.83118    NA     NA      0.01  
## 3 -84.03261    NA     NA      0.01  
## 4 -84.20558    NA     NA      0.01  
## 5 -84.18835    NA     NA      0.01  
## 6 -82.92519    NA     NA      0.01  
##                                         Description  
## 1 Right lane blocked due to accident on I-70 Eastbound at Exit 41 OH-235 State Route 4.  
## 2                               Accident on Brice Rd at Tussing Rd. Expect delays.  
## 3           Accident on OH-32 State Route 32 Westbound at Dela Palma Rd. Expect delays.  
## 4           Accident on I-75 Southbound at Exits 52 52B US-35. Expect delays.  
## 5           Accident on McEwen Rd at OH-725 Miamisburg Centerville Rd. Expect delays.  
## 6           Accident on I-270 Outerbelt Northbound near Exit 29 OH-3 State St. Expect delays.  
##   Number          Street Side      City      County State  
## 1     NA            I-70 E     R      Dayton Montgomery OH  
## 2    2584          Brice Rd    L Reynoldsburg Franklin OH  
## 3     NA        State Route 32    R Williamsburg Clermont OH
```

```

## 4      NA          I-75 S     R      Dayton Montgomery OH
## 5      NA Miamisburg Centerville Rd   R      Dayton Montgomery OH
## 6      NA          Westerville Rd   R  Westerville Franklin OH
## Zipcode Country  Timezone Airport_Code Weather_Timestamp Temperature.F.
## 1      45424      US US/Eastern    KFFO 2016-02-08 05:58:00  36.9
## 2      43068-3402  US US/Eastern    KCMH 2016-02-08 05:51:00  37.9
## 3      45176      US US/Eastern    KI69 2016-02-08 06:56:00  36.0
## 4      45417      US US/Eastern    KDAY 2016-02-08 07:38:00  35.1
## 5      45459      US US/Eastern    KMGY 2016-02-08 07:53:00  36.0
## 6      43081      US US/Eastern    KCMH 2016-02-08 07:51:00  37.9
## Wind_Chill.F. Humidity... Pressure.in. Visibility.mi. Wind_Direction
## 1      NA          91      29.68      10      Calm
## 2      NA          100     29.65      10      Calm
## 3      33.3        100     29.67      10      SW
## 4      31.0        96      29.64       9      SW
## 5      33.3        89      29.65       6      SW
## 6      35.5        97      29.63       7      SSW
## Wind_Speed.mph. Precipitation.in. Weather_Condition Amenity Bump Crossing
## 1      NA          0.02     Light Rain False False False
## 2      NA          0.00     Light Rain False False False
## 3      3.5         NA      Overcast False False False
## 4      4.6         NA      Mostly Cloudy False False False
## 5      3.5         NA      Mostly Cloudy False False False
## 6      3.5         0.03     Light Rain False False False
## Give_Way Junction No_Exit Railway Roundabout Station Stop Traffic_Calming
## 1      False     False False False False False False False
## 2      False     False False False False False False False
## 3      False     False False False False False False False
## 4      False     False False False False False False False
## 5      False     False False False False False False False
## 6      False     False False False False False False False
## Traffic_Signal Turning_Loop Sunrise_Sunset Civil_Twilight Nautical_Twilight
## 1      False     False Night  Night  Night  Night
## 2      False     False Night  Night  Night  Night
## 3      True      False Night  Night  Night  Day
## 4      False     False Night  Night  Day    Day
## 5      True      False Day   Day   Day   Day
## 6      False     False Day   Day   Day   Day
## Astronomical_Twilight
## 1      Night
## 2      Day
## 3      Day
## 4      Day
## 5      Day
## 6      Day

```

```
## DATA EXPLORATION AND REDUCTION ##
```

```
## Setting random seed to shuffle data before splitting
set.seed(23)
```

```
#Checking number of rows
rows<-sample(nrow(df))
```

```

#Shuffling the data
mva<-df[rows,]

#Taking the required number of instances from the shuffled data to reduce any biases
mva<-mva[500000:1000000,]

```

```

#Checking the structure of the dataset
str(mva)

```

```

## 'data.frame':      500001 obs. of  49 variables:
##   $ ID                  : chr  "A-210675" "A-1806002" "A-1932273" "A-11427" ...
##   $ Source              : chr  "MapQuest" "MapQuest" "MapQuest" "MapQuest" ...
##   $ TMC                 : num  241 201 201 201 NA 201 245 201 NA 201 ...
##   $ Severity             : int  3 2 2 2 2 2 3 2 2 3 ...
##   $ Start_Time           : chr  "2016-10-05 18:56:21" "2018-06-04 07:56:44" "2018-04-04 16:38:49" "2018-04-04 16:38:49" ...
##   $ End_Time              : chr  "2016-10-05 19:41:07" "2018-06-04 08:26:29" "2018-04-04 17:23:33" "2018-04-04 17:23:33" ...
##   $ Start_Lat             : num  40.8 40.3 33.7 37.6 38.6 ...
##   $ Start_Lng             : num  -74.3 -75.7 -117.9 -122.1 -121.6 ...
##   $ End_Lat               : num  NA NA NA NA 38.6 ...
##   $ End_Lng               : num  NA NA NA NA -122 ...
##   $ Distance.mi.          : num  0 0 0 0.01 0 0.01 0 0 0 0 ...
##   $ Description            : chr  "Left lane blocked due to accident on I-280 Westbound at Exits 4A 4B" ...
##   $ Number                : num  NA 1535 NA NA NA ...
##   $ Street                 : chr  "I-280 E" "Sell Rd" "Corona del Mar Fwy N" "W Jackson St" ...
##   $ Side                   : chr  "R" "L" "R" "R" ...
##   $ City                   : chr  "Roseland" "Pottstown" "Costa Mesa" "Hayward" ...
##   $ County                 : chr  "Essex" "Montgomery" "Orange" "Alameda" ...
##   $ State                  : chr  "NJ" "PA" "CA" "CA" ...
##   $ Zipcode                : chr  "07068" "19464" "92626" "94544" ...
##   $ Country                : chr  "US" "US" "US" "US" ...
##   $ Timezone               : chr  "US/Eastern" "US/Eastern" "US/Pacific" "US/Pacific" ...
##   $ Airport_Code            : chr  "KCDW" "KPTW" "KSNA" "KHWD" ...
##   $ Weather_Timestamp       : chr  "2016-10-05 18:53:00" "2018-06-04 07:54:00" "2018-04-04 16:53:00" "2018-04-04 16:53:00" ...
##   $ Temperature.F.          : num  60.1 54 62.1 54 65 ...
##   $ Wind_Chill.F.           : num  NA NA NA NA 65 NA NA NA 55 NA ...
##   $ Humidity...              : num  67 93 72 59 15 87 96 54 38 12 ...
##   $ Pressure.in.            : num  30.3 29.8 30 30.1 29.9 ...
##   $ Visibility.mi.          : num  10 8 9 7 10 10 0.8 10 10 10 ...
##   $ Wind_Direction           : chr  "Calm" "NE" "SW" "NE" ...
##   $ Wind_Speed.mph.         : num  NA 4.6 8.1 9.2 17 9.2 NA NA 10 NA ...
##   $ Precipitation.in.        : num  NA NA NA NA 0 0 0.11 0 0 NA ...
##   $ Weather_Condition        : chr  "Clear" "Scattered Clouds" "Partly Cloudy" "Partly Cloudy" ...
##   $ Amenity                 : chr  "False" "False" "False" "False" ...
##   $ Bump                     : chr  "False" "False" "False" "False" ...
##   $ Crossing                 : chr  "False" "False" "False" "False" ...
##   $ Give_Way                 : chr  "False" "False" "False" "False" ...
##   $ Junction                 : chr  "False" "False" "False" "False" ...
##   $ No_Exit                  : chr  "False" "False" "False" "False" ...
##   $ Railway                  : chr  "False" "False" "False" "False" ...
##   $ Roundabout                : chr  "False" "False" "False" "False" ...
##   $ Station                   : chr  "False" "False" "False" "False" ...
##   $ Stop                      : chr  "False" "False" "False" "False" ...
##   $ Traffic_Calming           : chr  "False" "False" "False" "False" ...

```

```

## $ Traffic_Signal      : chr  "False" "False" "False" "False" ...
## $ Turning_Loop         : chr  "False" "False" "False" "False" ...
## $ Sunrise_Sunset        : chr  "Night" "Day" "Day" "Day" ...
## $ Civil_Twilight        : chr  "Day" "Day" "Day" "Day" ...
## $ Nautical_Twilight     : chr  "Day" "Day" "Day" "Day" ...
## $ Astronomical_Twilight: chr  "Day" "Day" "Day" "Day" ...

# Checking the number of rows and columns in the current uncleaned dataset
ncol(mva)

## [1] 49

nrow(mva)

## [1] 500001

# Printing all the column names to find and filter the relevant and irrelevant attributes
names<-names(mva)
names

## [1] "ID"                  "Source"              "TMC"
## [4] "Severity"            "Start_Time"          "End_Time"
## [7] "Start_Lat"            "Start_Lng"            "End_Lat"
## [10] "End_Lng"              "Distance.mi."        "Description"
## [13] "Number"               "Street"               "Side"
## [16] "City"                 "County"               "State"
## [19] "Zipcode"              "Country"              "Timezone"
## [22] "Airport_Code"         "Weather_Timestamp"   "Temperature.F."
## [25] "Wind_Chill.F."        "Humidity..."          "Pressure.in."
## [28] "Visibility.mi."       "Wind_Direction"      "Wind_Speed.mph."
## [31] "Precipitation.in."    "Weather_Condition"   "Amenity"
## [34] "Bump"                 "Crossing"             "Give_Way"
## [37] "Junction"             "No_Exit"              "Railway"
## [40] "Roundabout"            "Station"              "Stop"
## [43] "Traffic_Calming"       "Traffic_Signal"       "Turning_Loop"
## [46] "Sunrise_Sunset"         "Civil_Twilight"       "Nautical_Twilight"
## [49] "Astronomical_Twilight"

## DATA CLEANING ##

#Dropping the surplus attributes which do not contribute to the analysis
mva <- mva[-c(1:3,7:10,13,14,19,21:23,33,47:49)]

#Checking for any null values in the present dataset
is.na(mva[,])

#Checking which rows have all the values filled and complete
complete.cases(mva)

#Making a new dataframe with only the rows that have complete information and all values filled
Mva<-na.omit(mva)

```

```

#Verifying for missing values in the new dataframe
#complete.cases(Mva)

#Checking the number of rows and columns in the new CLEANED dataframe
ncol(Mva)

## [1] 32

nrow(Mva)

## [1] 182849

## EXPLORATORY DESCRIPTIVE STATISTICS ##

#Getting some basic initial statistical values for insights into the dataset
summary(Mva)

##      Severity      Start_Time      End_Time      Distance.mi.
##  Min.   :1.00  Length:182849  Length:182849  Min.   : 0.0000
##  1st Qu.:2.00  Class  :character  Class  :character  1st Qu.: 0.0000
##  Median :2.00  Mode   :character  Mode   :character  Median : 0.0000
##  Mean   :2.28                               Mean   : 0.2903
##  3rd Qu.:3.00                               3rd Qu.: 0.0000
##  Max.   :4.00                               Max.   :79.4400
##      Description      Side      City      County
##  Length:182849  Length:182849  Length:182849  Length:182849
##  Class  :character  Class  :character  Class  :character  Class  :character
##  Mode   :character  Mode   :character  Mode   :character  Mode   :character
##
##      State      Country      Temperature.F.      Wind_Chill.F.
##  Length:182849  Length:182849  Min.   :-24.00  Min.   :-50.10
##  Class  :character  Class  :character  1st Qu.: 48.00  1st Qu.: 46.00
##  Mode   :character  Mode   :character  Median  : 63.00  Median  : 63.00
##                               Mean   : 61.03  Mean   : 59.57
##                               3rd Qu.: 75.00  3rd Qu.: 75.00
##                               Max.   :115.00  Max.   :115.00
##      Humidity...      Pressure.in.      Visibility.mi.      Wind_Direction
##  Min.   : 1.00  Min.   :20.19  Min.   : 0.000  Length:182849
##  1st Qu.: 49.00 1st Qu.:29.13  1st Qu.: 10.000  Class  :character
##  Median : 69.00  Median :29.62  Median : 10.000  Mode   :character
##  Mean   : 65.53  Mean   :29.30  Mean   :  8.856
##  3rd Qu.: 85.00 3rd Qu.:29.92  3rd Qu.: 10.000
##  Max.   :100.00  Max.   :31.08  Max.   :100.000
##      Wind_Speed.mph.      Precipitation.in.      Weather_Condition      Bump
##  Min.   : 0.000  Min.   :0.0000000  Length:182849  Length:182849
##  1st Qu.: 3.000  1st Qu.:0.0000000  Class  :character  Class  :character
##  Median : 7.000  Median :0.0000000  Mode   :character  Mode   :character
##  Mean   : 7.568  Mean   :0.007555
##  3rd Qu.:10.000 3rd Qu.:0.0000000
##  Max.   :142.000 Max.   :10.020000

```

```

##      Crossing          Give_Way        Junction        No_Exit
## Length:182849    Length:182849    Length:182849    Length:182849
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##      Railway          Roundabout       Station         Stop
## Length:182849    Length:182849    Length:182849    Length:182849
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##      Traffic_Calming  Traffic_Signal   Turning_Loop   Sunrise_Sunset
## Length:182849    Length:182849    Length:182849    Length:182849
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##
```

```

#Checking for datatypes of each of the attributes in the dataset
sapply(Mva,class)

```

```

##      Severity        Start_Time       End_Time     Distance.mi.
## "integer"      "character"      "character"    "numeric"
## Description        Side           City          County
## "character"      "character"      "character"    "character"
## State            Country        Temperature.F. Wind_Chill.F.
## "character"      "character"      "numeric"      "numeric"
## Humidity...       Pressure.in.  Visibility.mi. Wind_Direction
## "numeric"        "numeric"       "numeric"      "character"
## Wind_Speed.mph. Precipitation.in. Weather_Condition Bump
## "numeric"        "numeric"       "character"    "character"
## Crossing          Give_Way        Junction        No_Exit
## "character"      "character"      "character"    "character"
## Railway          Roundabout       Station         Stop
## "character"      "character"      "character"    "character"
##      Traffic_Calming  Traffic_Signal   Turning_Loop   Sunrise_Sunset
## "character"      "character"      "character"    "character"

```

```

# Creating new dataframe with only the numerical attributes to perform statistical functions
num<-Mva[,c(1,4,11:15,17,18)]

```

```

# Finding the correlation between all numerical variables
corr<-cor(num,method=c("pearson","kendall","spearman"))
corr

```

```

##      Severity Distance.mi. Temperature.F. Wind_Chill.F.
## Severity          1.000000000  0.18320436   -0.02780004  -0.03350551
## Distance.mi.      0.183204361  1.000000000  -0.02185492  -0.02466257
## Temperature.F.   -0.027800040 -0.02185492   1.000000000  0.99375385

```

```

## Wind_Chill.F.      -0.033505511  -0.02466257      0.99375385  1.000000000
## Humidity...        0.052701523   0.02176980     -0.43135406  -0.41426578
## Pressure.in.       -0.003922202  -0.03934736      0.03946663  0.04070670
## Visibility.mi.    -0.025895193  -0.01784896      0.31676910  0.32580072
## Wind_Speed.mph.   0.052353286   0.02074942     -0.01090524  -0.06539002
## Precipitation.in. 0.019554182   0.00344414     -0.03308594  -0.03359791
##                               Humidity...  Pressure.in.  Visibility.mi.  Wind_Speed.mph.
## Severity            0.05270152  -0.003922202  -0.02589519  0.05235329
## Distance.mi.        0.02176980  -0.039347359  -0.01784896  0.02074942
## Temperature.F.     -0.43135406  0.039466627  0.31676910  -0.01090524
## Wind_Chill.F.      -0.41426578  0.040706699  0.32580072  -0.06539002
## Humidity...          1.00000000  0.198936163  -0.43067894  -0.14819125
## Pressure.in.        0.19893616  1.000000000  -0.08965456  -0.04925739
## Visibility.mi.     -0.43067894  -0.089654562  1.00000000  -0.02254997
## Wind_Speed.mph.    -0.14819125  -0.049257386  -0.02254997  1.00000000
## Precipitation.in.  0.10380281  0.017843267  -0.14656459  0.03132604
##                               Precipitation.in.
## Severity            0.01955418
## Distance.mi.        0.00344414
## Temperature.F.     -0.03308594
## Wind_Chill.F.      -0.03359791
## Humidity...          0.10380281
## Pressure.in.        0.01784327
## Visibility.mi.     -0.14656459
## Wind_Speed.mph.    0.03132604
## Precipitation.in.  1.00000000

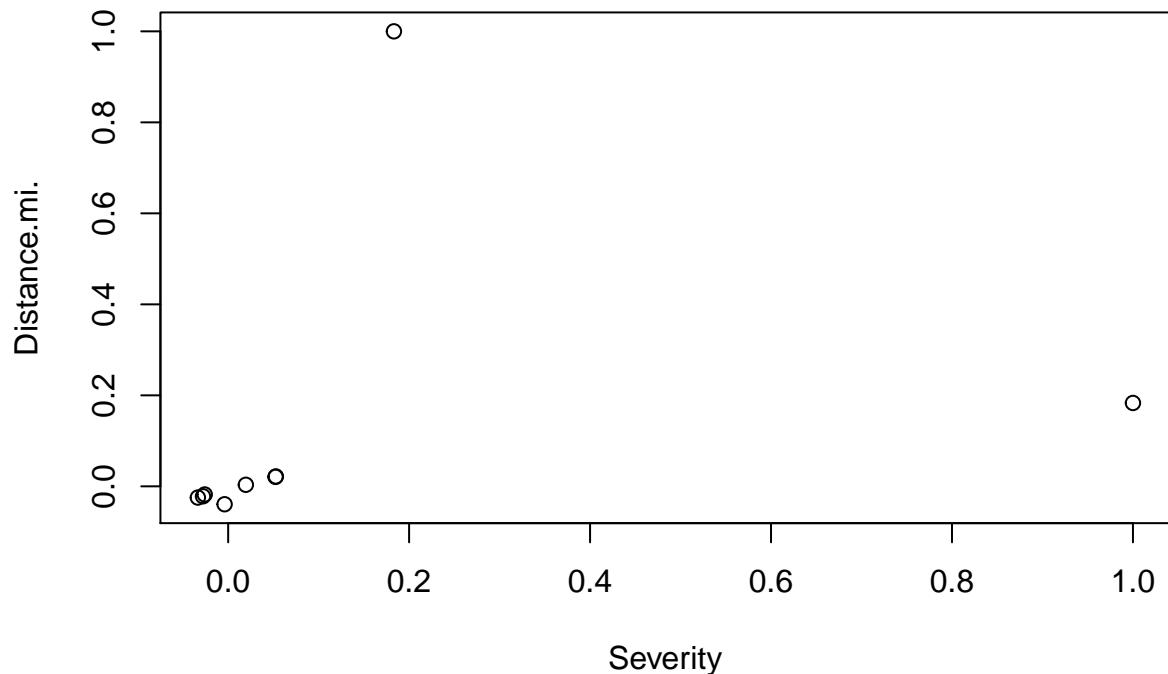
```

```
# OR
```

```
#Plotting the correlation graph
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
plot(corr)
```

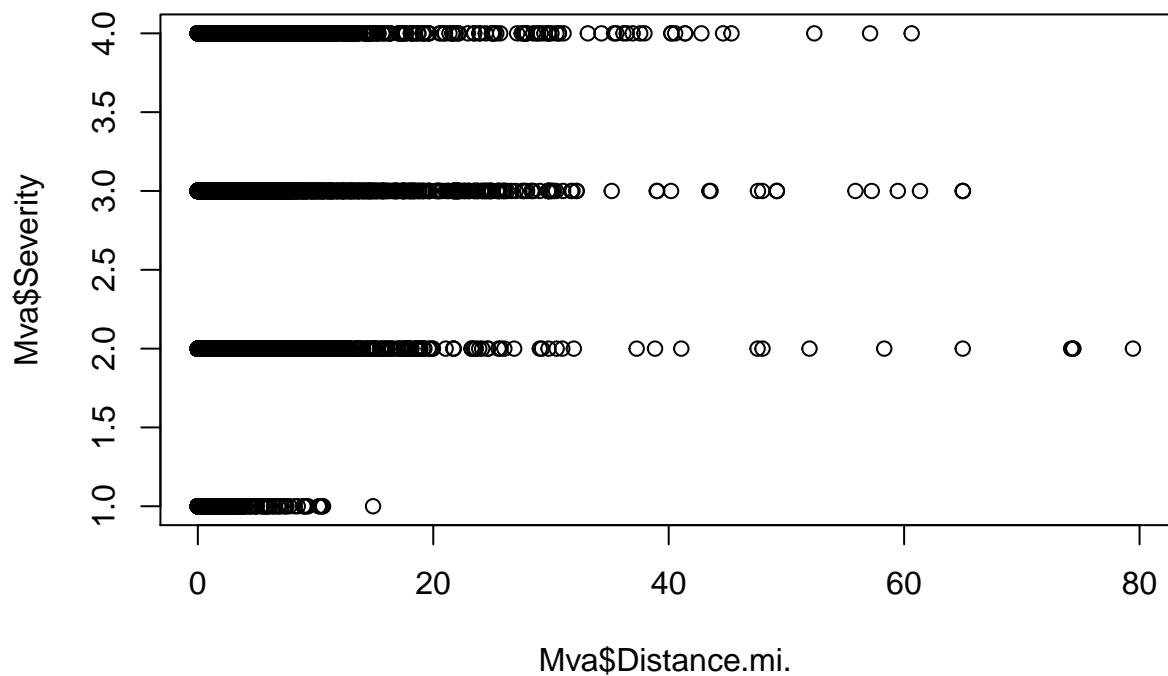


```
#We observe that there are strong correlations between Temp and visibility, wind chill and visibility,
#Humidity and pressure etc, strong neg corrs are observed between Humidity and wind chill for instance
# A higher pos value between 0 and 1 is positive corr and neg value between 0 and -1 is a neg corr
```

```
# DATA EXPLORATION AND VISUALIZATION ##
```

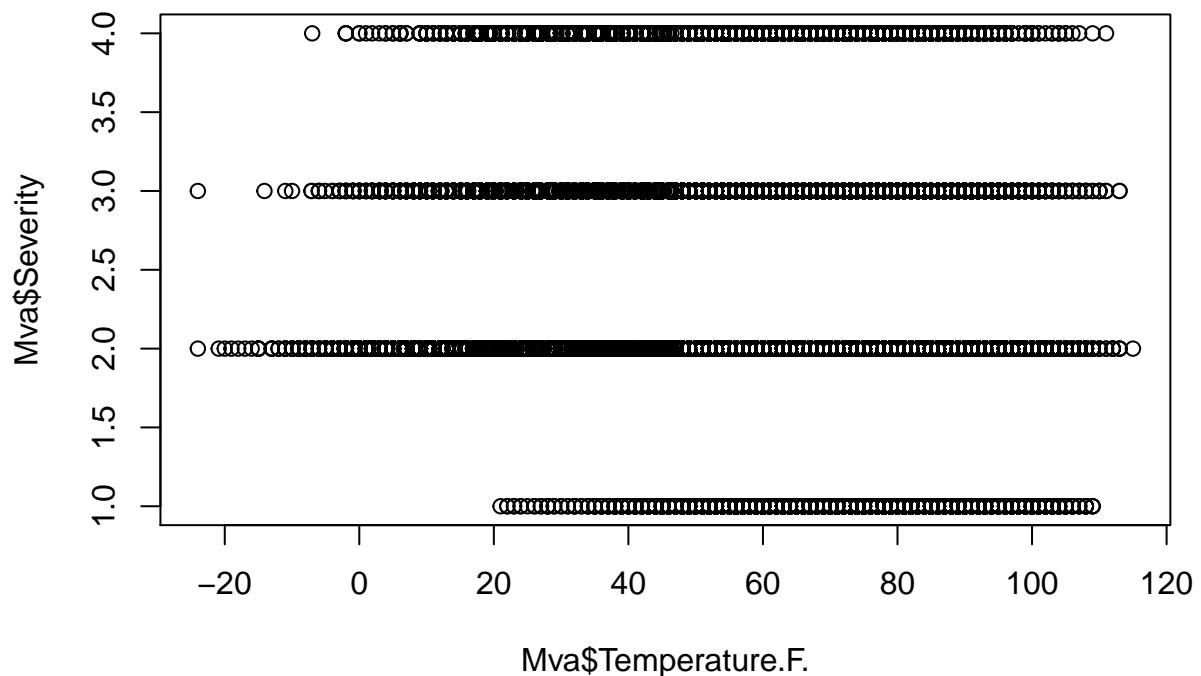
```
#Ques 1. How are the numerical weather variables related to the KPI Severity of accident ??
```

```
## Plotting Scatter plots between each variable and the KPI(Severity)
plot(Mva$Distance.mi., Mva$Severity)
```

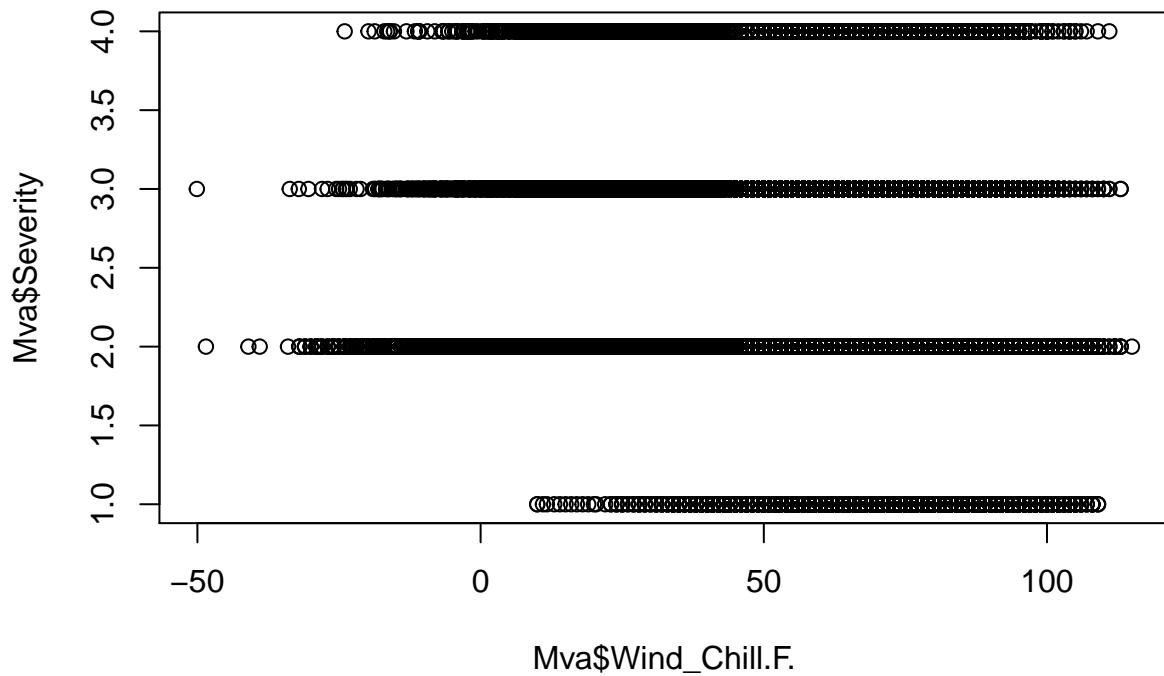


```
#Some outliers are found in the distance column for distances larger than 20 mtrs, most categories of severity have been removed
```

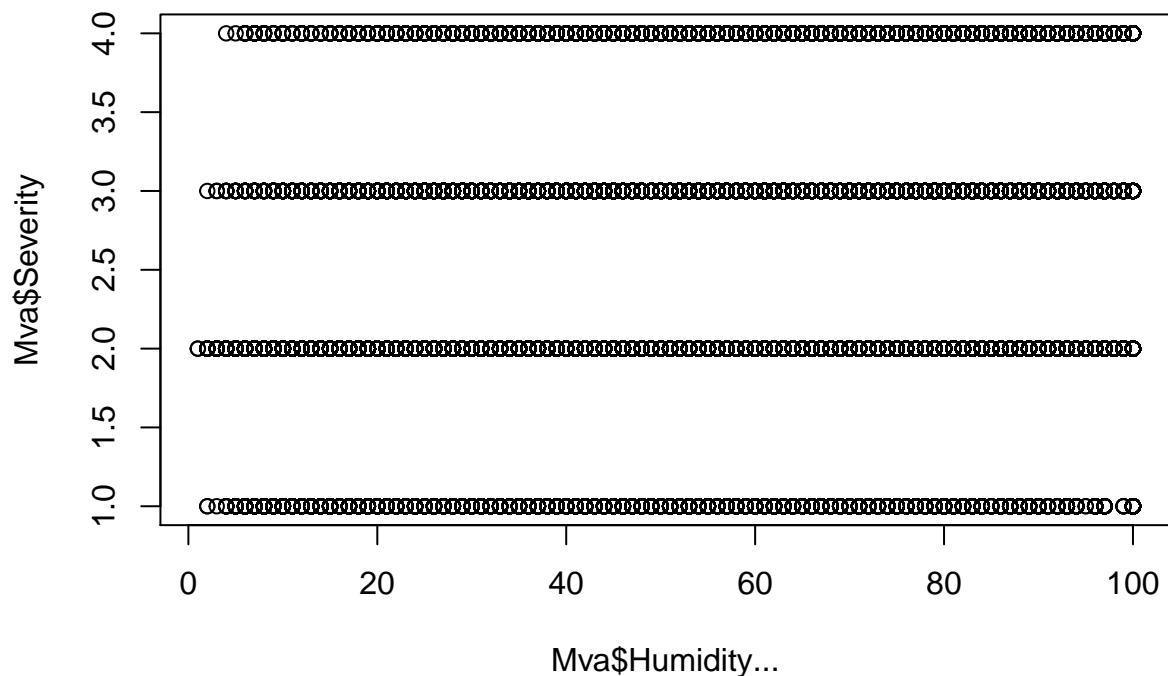
```
plot(Mva$Temperature.F.,Mva$Severity)
```



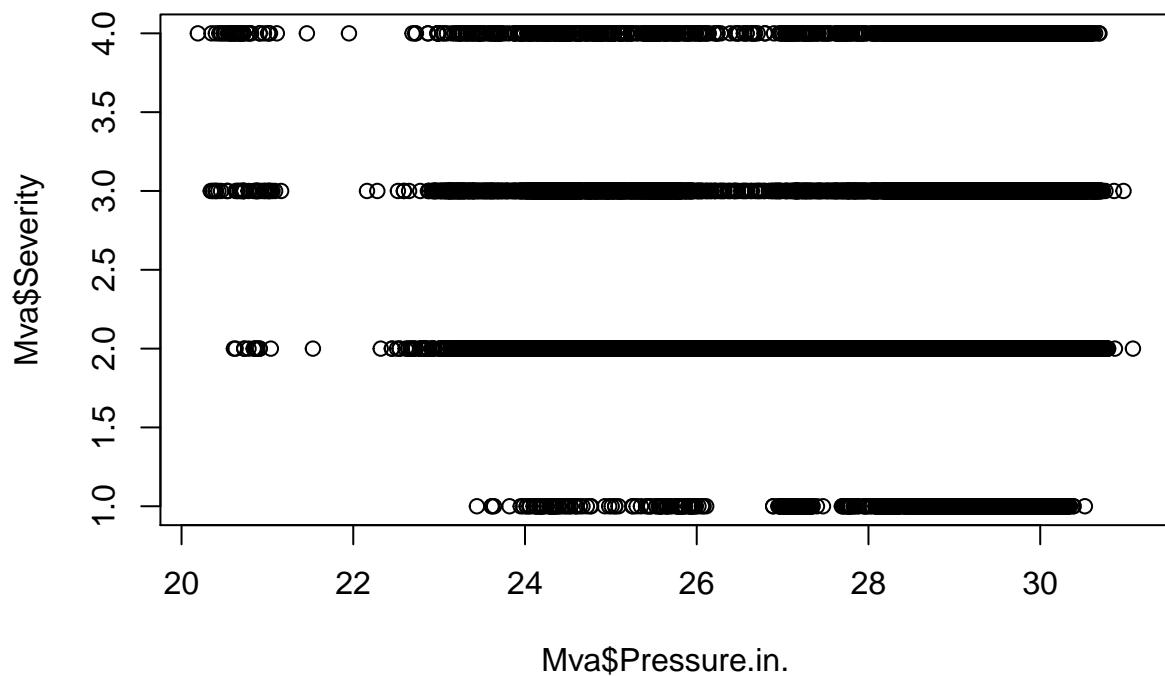
```
#From the apparent density of the data points on the severity vs temp plot, for slight rise in temp the  
plot(Mva$Wind_Chill.F.,Mva$Severity)
```



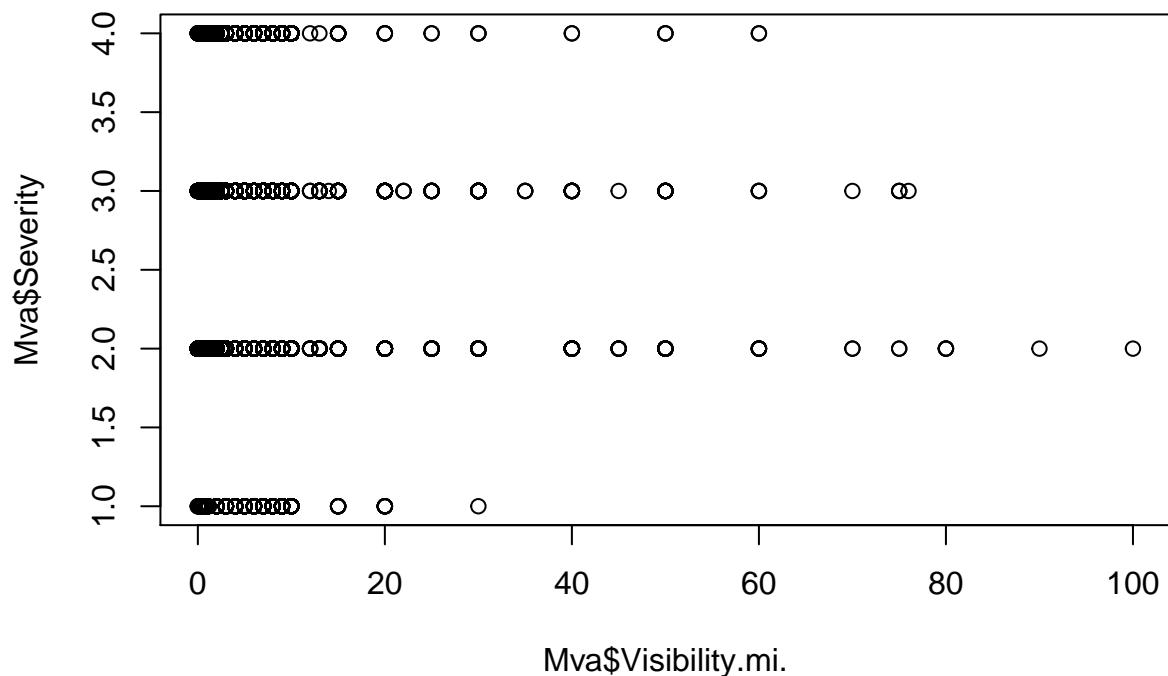
```
#A similar effect is observed , as wind chills temp rises severity is increasing , also no category 1 a  
plot(Mva$Humidity...,Mva$Severity)
```



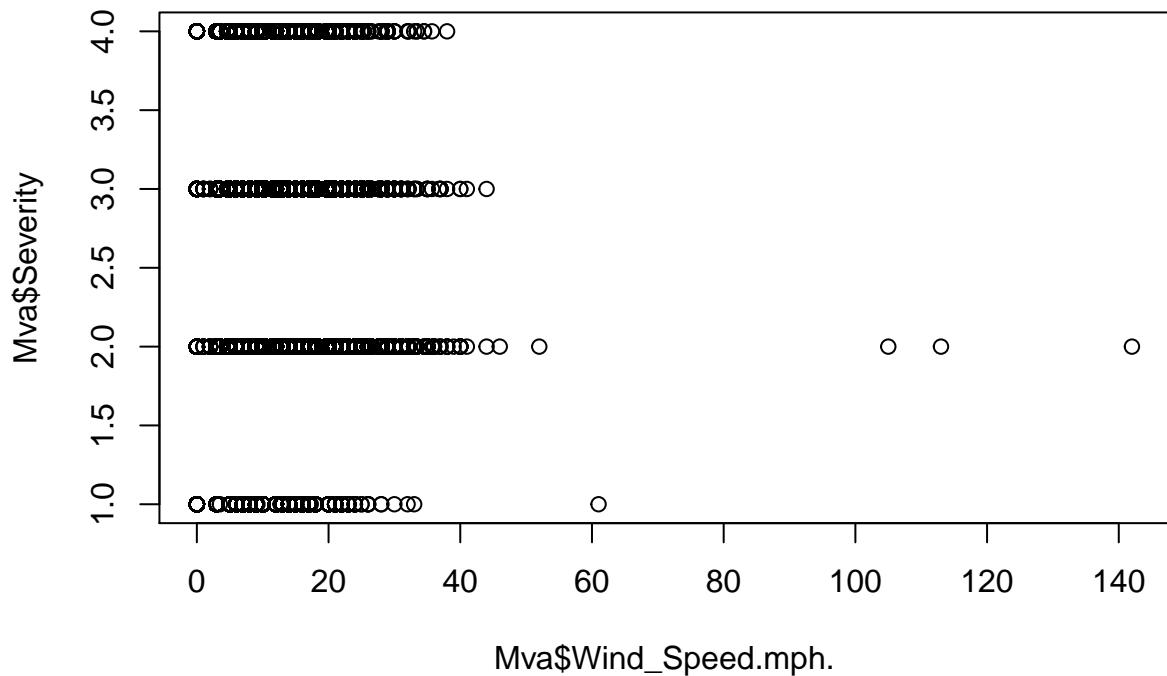
```
#Humidity seems to almost have no co relation with severity here  
plot(Mva$Pressure.in.,Mva$Severity)
```



```
#No cat 1 accidents for pressures less than 23 inches, for other categories , the data seem pretty skewed
plot(Mva$Visibility.mi.,Mva$Severity)
```



```
# Almost all categories of accidents have taken place when Visibility was 0, with few outliers in each
plot(Mva$Wind_Speed.mph., Mva$Severity)
```



```

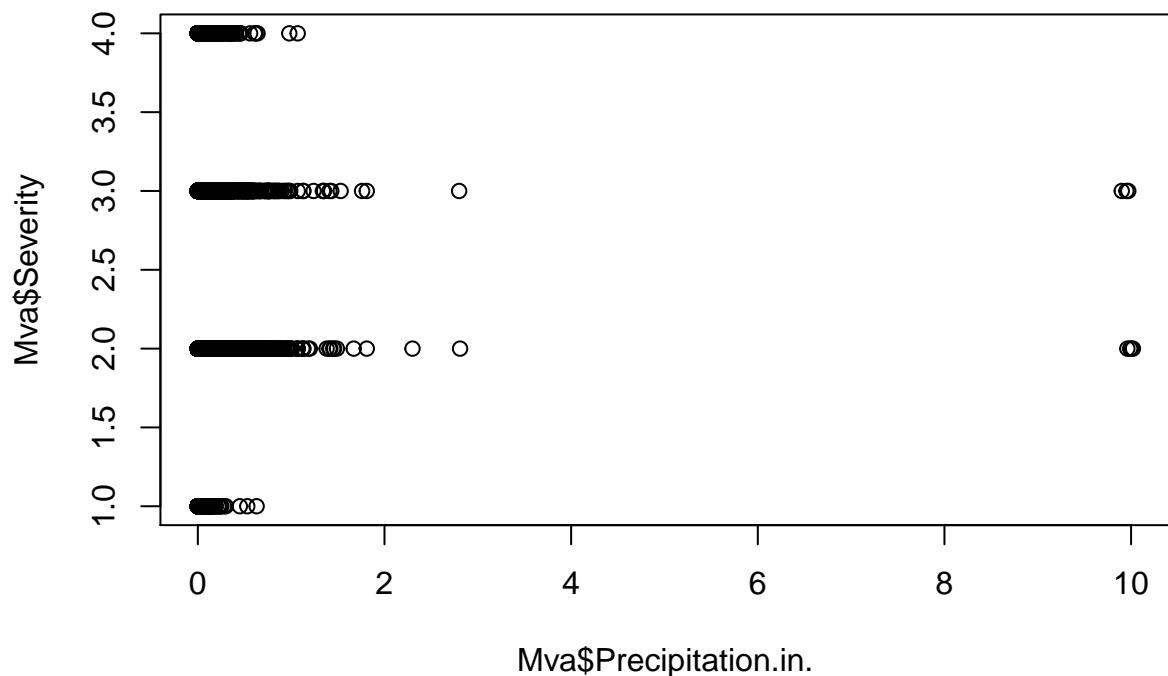
plot(Mva$Precipitation.in.,Mva$Severity)

#Scatterplots here show how the variables are distributed and their frequencies in each category of severity

#Ques 2. How are the attributes distributed, what values or ranges of each numerical attribute is more frequent

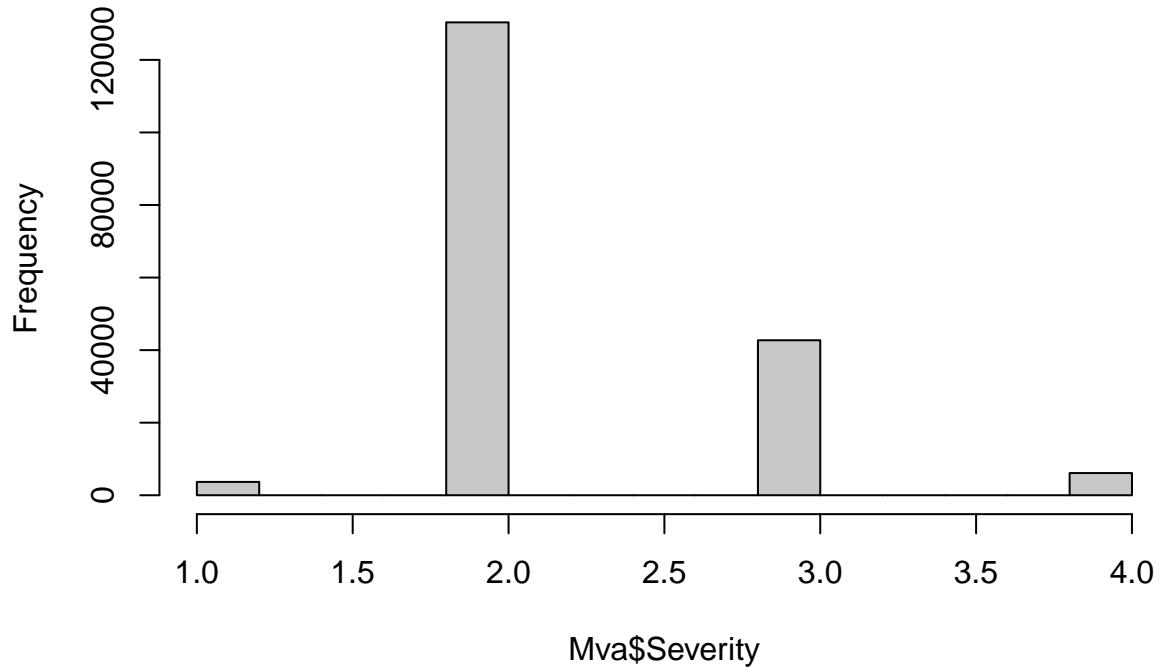
#Plotting histograms
library(ggplot2)

```



```
#Which was the category that most accidents fell under??  
hist(Mva$Severity)
```

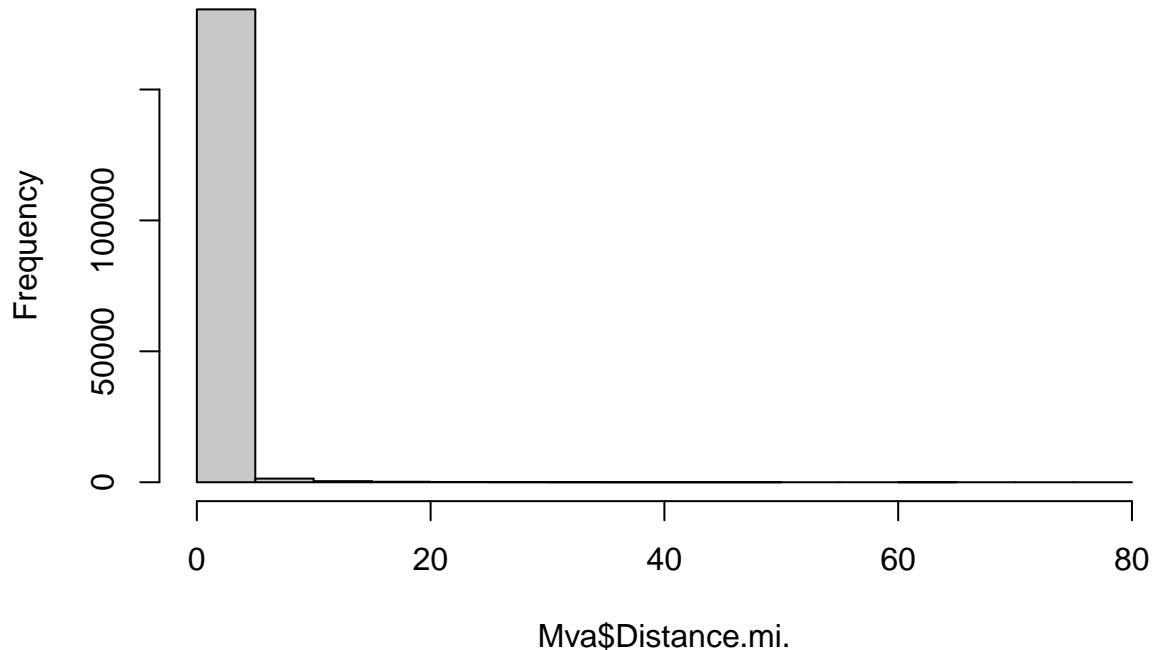
Histogram of Mva\$Severity



```
# Category 2 severity was most prevalent
```

```
#What was the distance distribution over which the accident took place??
hist(Mva$Distance.mi.)
```

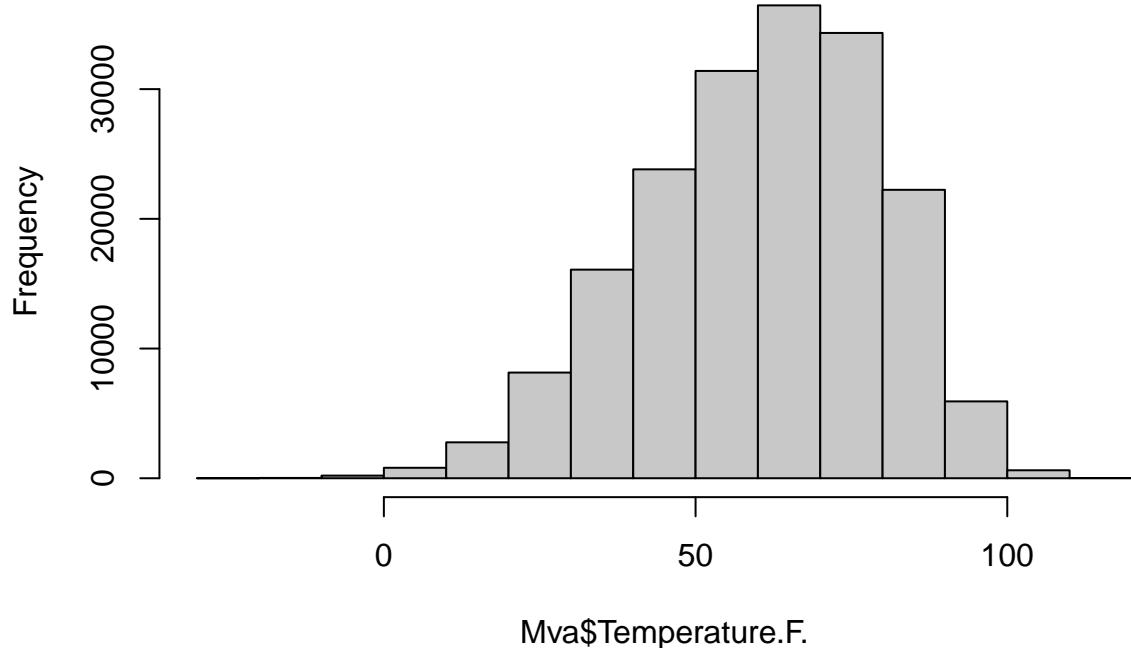
Histogram of Mva\$Distance.mi.



```
#Mostly between 0 to 20 meters
```

```
# What temp contributed most to the accidents??
hist(Mva$Temperature.F.)
```

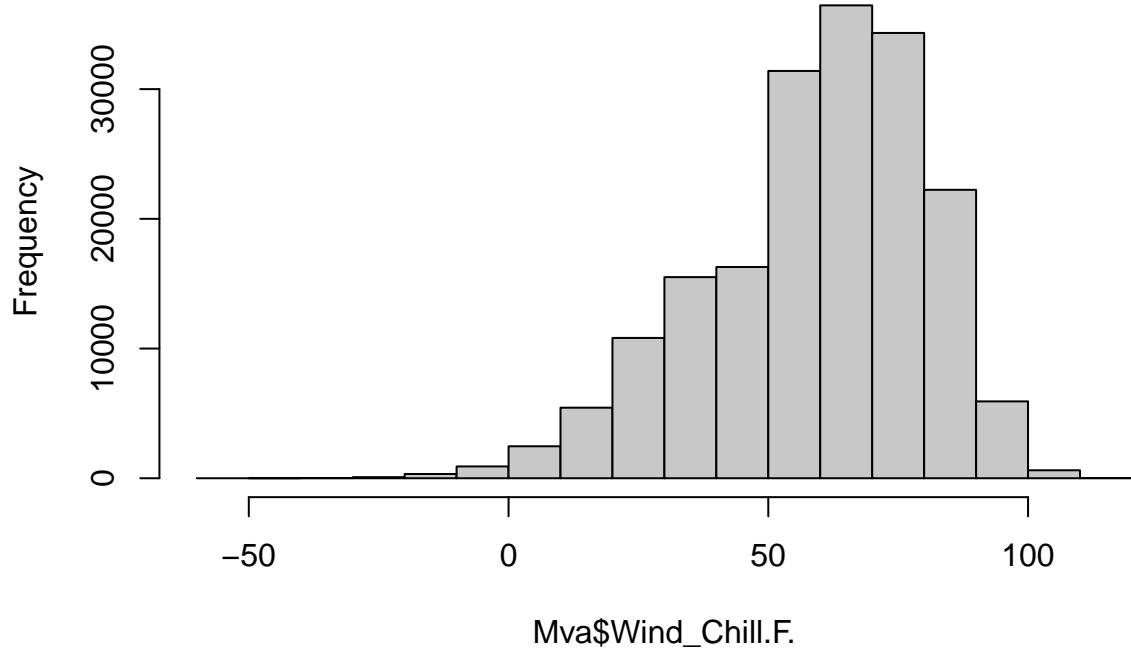
Histogram of Mva\$Temperature.F.



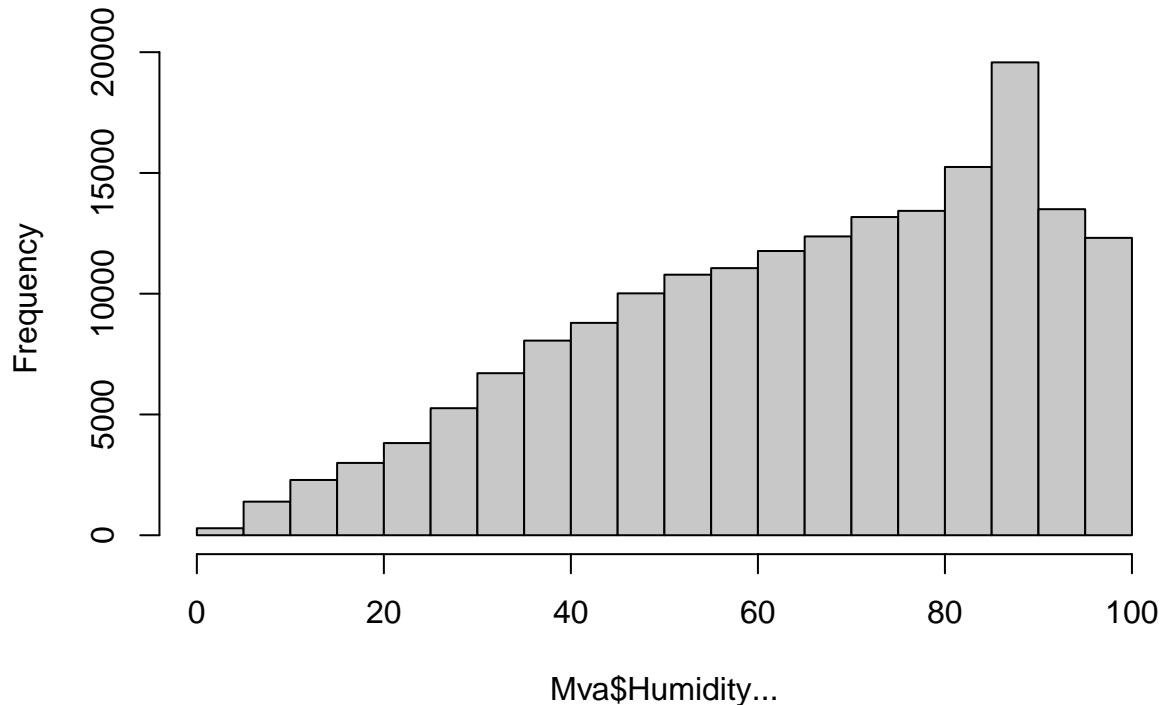
```
# Most accidents took place when the temp was between 60-80 deg F
```

```
# How did wind chill contribute to the accident??
hist(Mva$Wind_Chill.F.)
```

Histogram of Mva\$Wind_Chill.F.



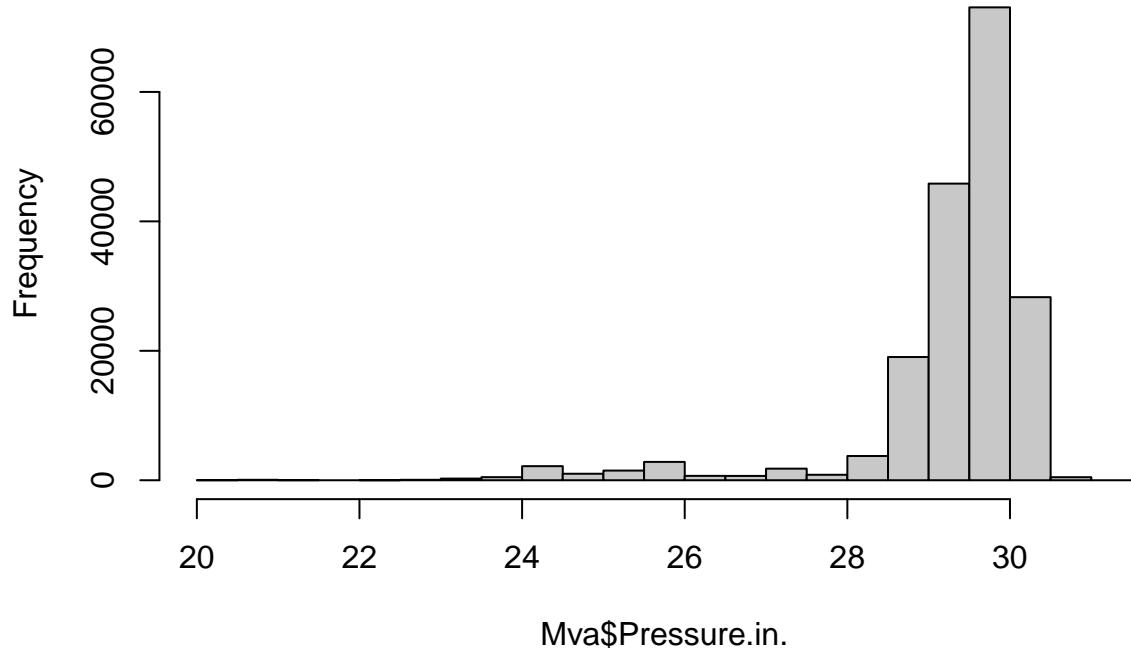
Histogram of Mva\$Humidity...



#Shows a strong positive trend , as the humidity increased so did the number of accidents

```
#Pressure contribution??  
hist(Mva$Pressure.in.)
```

Histogram of Mva\$Pressure.in.

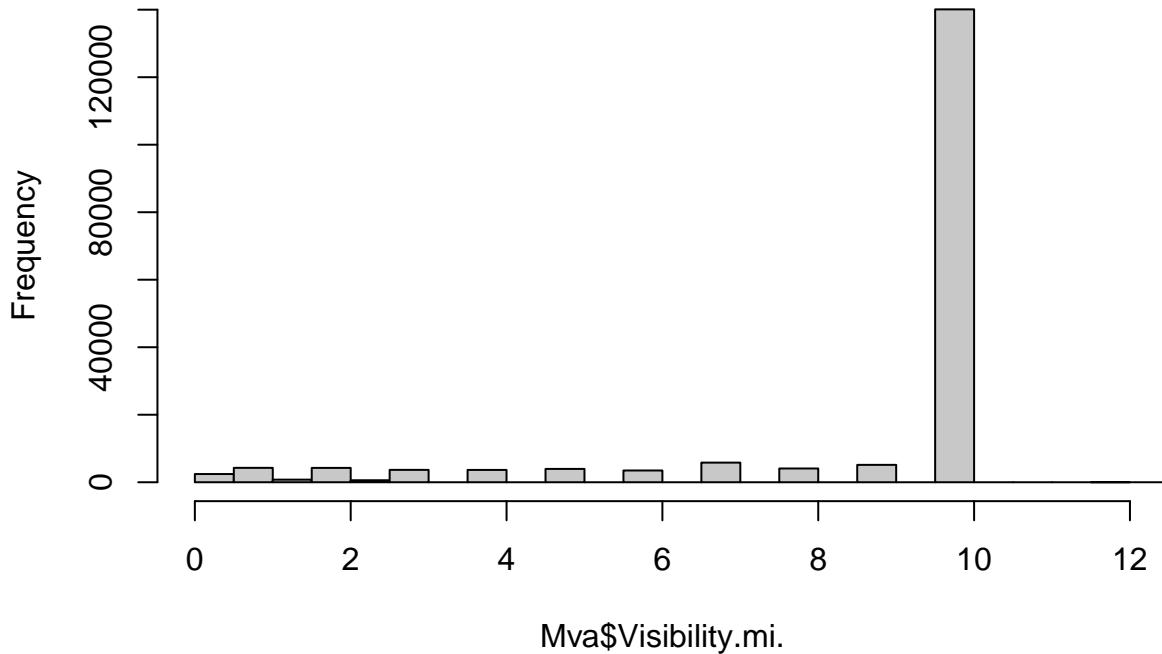


```
#Again we see a positive relation, as the higher values of pressures show higher numbers of accidents
```

```
#Visibility vs no of accidents??
```

```
hist(Mva$Visibility.mi., xlim=c(0,12), breaks=200)
```

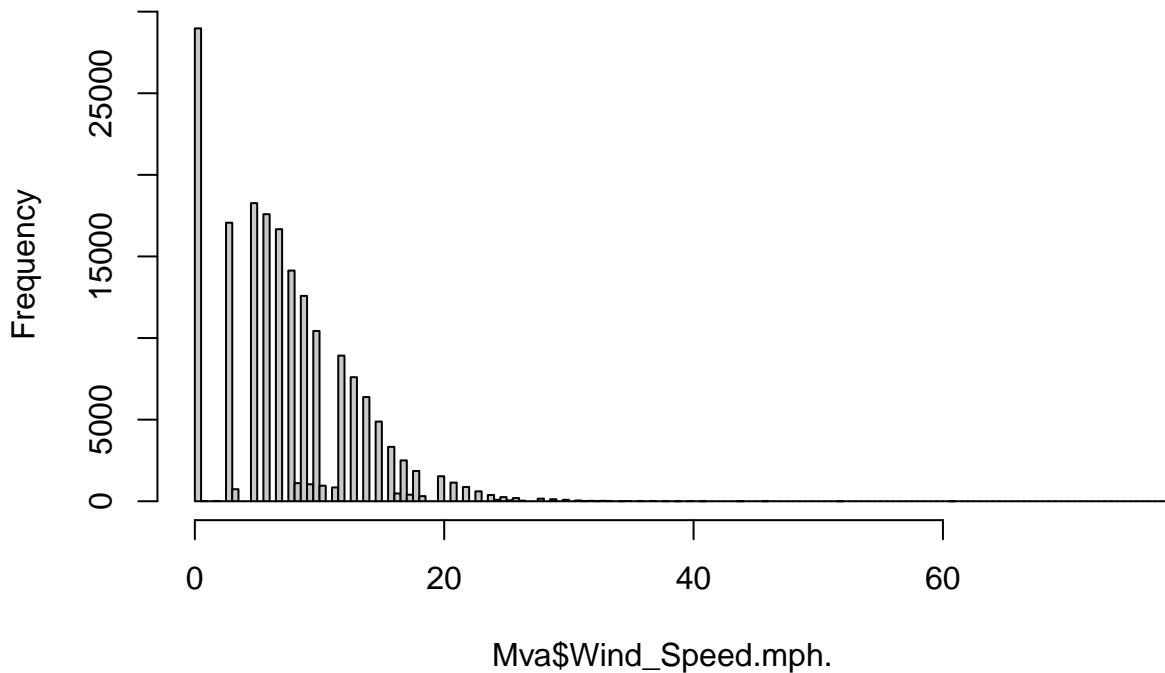
Histogram of Mva\$Visibility.mi.



```
# A low visibility is the most prevalent for accident and thus has much more accidents associated with
#AS visibility increased accidents decreased

#Was Wind speed a factor in accident event??
hist(Mva$Wind_Speed.mph., xlim=c(0,75), breaks=500)
```

Histogram of Mva\$Wind_Speed.mph.

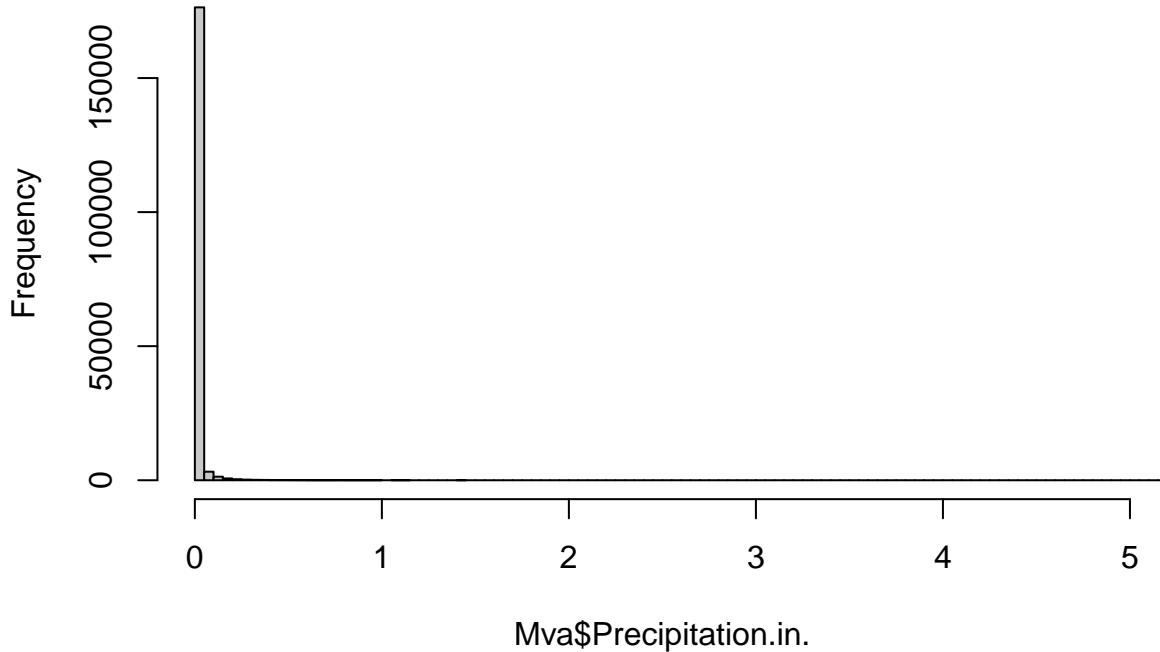


```
#Wind chills seem to be negatively related to accidents, maybe when high wind chills were present , less
```

```
#Effect of precipitation?
```

```
hist(Mva$Precipitation.in.,breaks=200,xlim=c(0,5))
```

Histogram of Mva\$Precipitation.in.



```
#Precipitation shows no significant relation to no of accidents
```

```
# Histograms here show the frequencies of each attribute within a certain range,
```

```
#Finding covariance and mahalanobis, Euclidean distance
```

```
cov(num)
```

```
##          Severity Distance.mi. Temperature.F. Wind_Chill.F.
## Severity      0.308812204  0.1638066450   -0.29085445  -0.39486552
## Distance.mi.    0.163806645  2.5887908367   -0.66203474  -0.84153587
## Temperature.F. -0.290854452 -0.6620347391   354.45910223  396.77804128
## Wind_Chill.F.  -0.394865518 -0.8415358735   396.77804128  449.75031319
## Humidity...     0.674941335  0.8072323952  -187.15980637 -202.46988132
## Pressure.in.   -0.002500267 -0.0726228340    0.85235831  0.99028626
## Visibility.mi. -0.041726178 -0.0832728841   17.29291115  20.03456907
## Wind_Speed.mph. 0.158426351  0.1817986534   -1.11803257  -7.55150430
## Precipitation.in. 0.000814851  0.0004155474   -0.04671086  -0.05343051
##          Humidity... Pressure.in. Visibility.mi. Wind_Speed.mph.
## Severity       0.6749413 -0.002500267  -0.04172618   0.15842635
## Distance.mi.    0.8072324 -0.072622834  -0.08327288   0.18179865
## Temperature.F. -187.1598064  0.852358309   17.29291115  -1.11803257
## Wind_Chill.F.  -202.4698813  0.990286255   20.03456907  -7.55150430
## Humidity...      531.1182985  5.259185569  -28.78004693  -18.59749464
## Pressure.in.     5.2591856  1.315885952  -0.29821106  -0.30769237
## Visibility.mi.  -28.7800469 -0.298211062   8.40784040  -0.35606101
## Wind_Speed.mph. -18.5974946 -0.307692373  -0.35606101  29.65326254
```

```

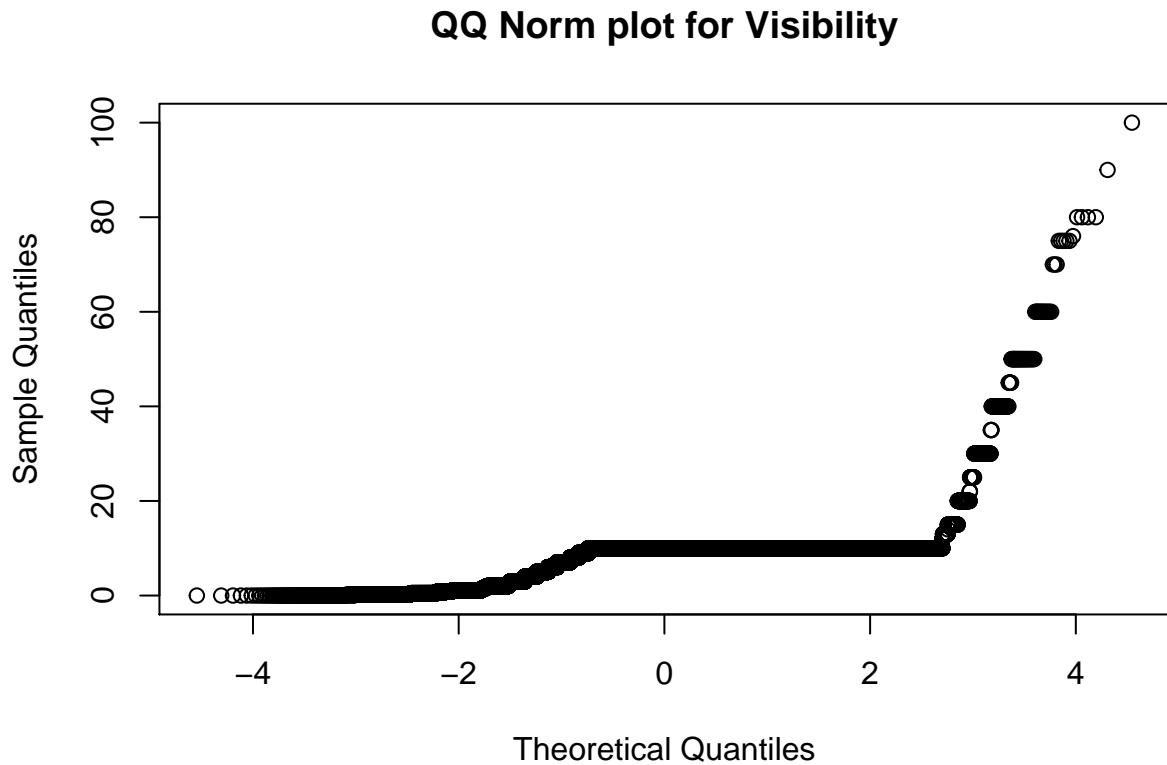
## Precipitation.in.    0.1793891  0.001534881   -0.03186857      0.01279184
##                               Precipitation.in.
## Severity             0.0008148510
## Distance.mi.        0.0004155474
## Temperature.F.     -0.0467108622
## Wind_Chill.F.       -0.0534305107
## Humidity...          0.1793891075
## Pressure.in.         0.0015348814
## Visibility.mi.      -0.0318685722
## Wind_Speed.mph.     0.0127918428
## Precipitation.in.    0.0056231887

#Again here we find strong positive relationships between Temp and wind chill,Temp and visibility etc
#Strong neg relation between Humidity and visibility
# For the KPI that is severity, Humidity, wind speed,Distance were the attributes that showed slightly pos
#dist(num,method = "euclidean")
#mahalanobis(x=Mva$Distance.mi.,data.y=Mva$Temperature.F.)

## FINDING NORMALITY OF DATA ATTRIBUTES (NUMERICAL) ##

#Finding qqnorm normality in dataset, similarly will be plotted for all numerical attributes,(finding n
qqnorm(num[,c("Visibility.mi.")],main="QQ Norm plot for Visibility")

```

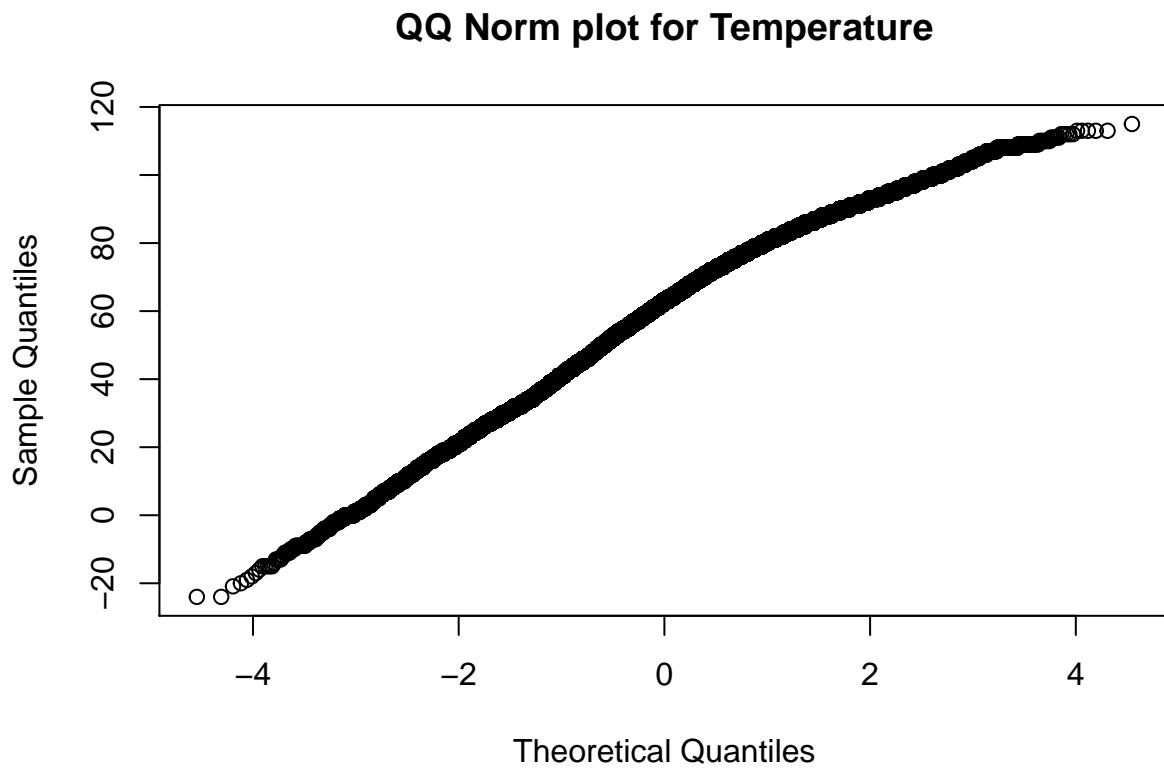


```

#Using the scale function to standardize the column values
Mva$Visibility.mi.<-scale(Mva$Visibility.mi.)
#Above plot for visibility shows that the data is not normal, and therefore would be an issue while
#Statistical testing and hence the visibility attribute needs to be normalized

qqnorm(num[,c("Temperature.F.")],main="QQ Norm plot for Temperature")

```



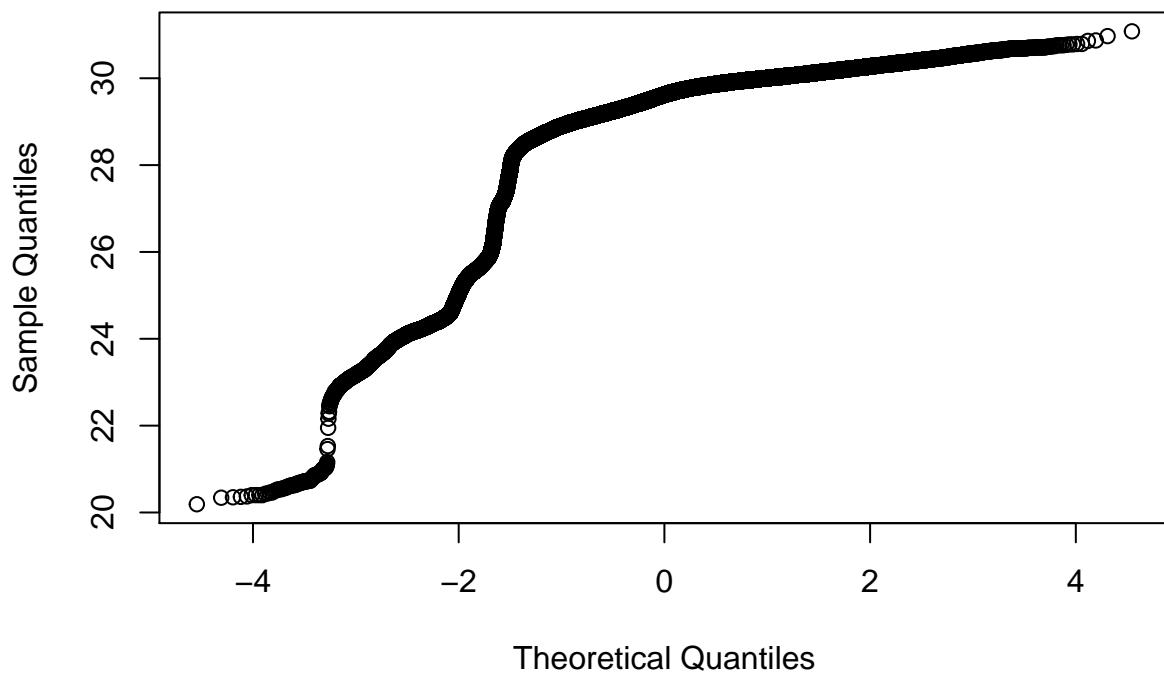
```

#The qq plot for temperature shows that the data is almost perfectly normal

qqnorm(num[,c("Pressure.in.")],main="QQ Norm plot for Pressure")

```

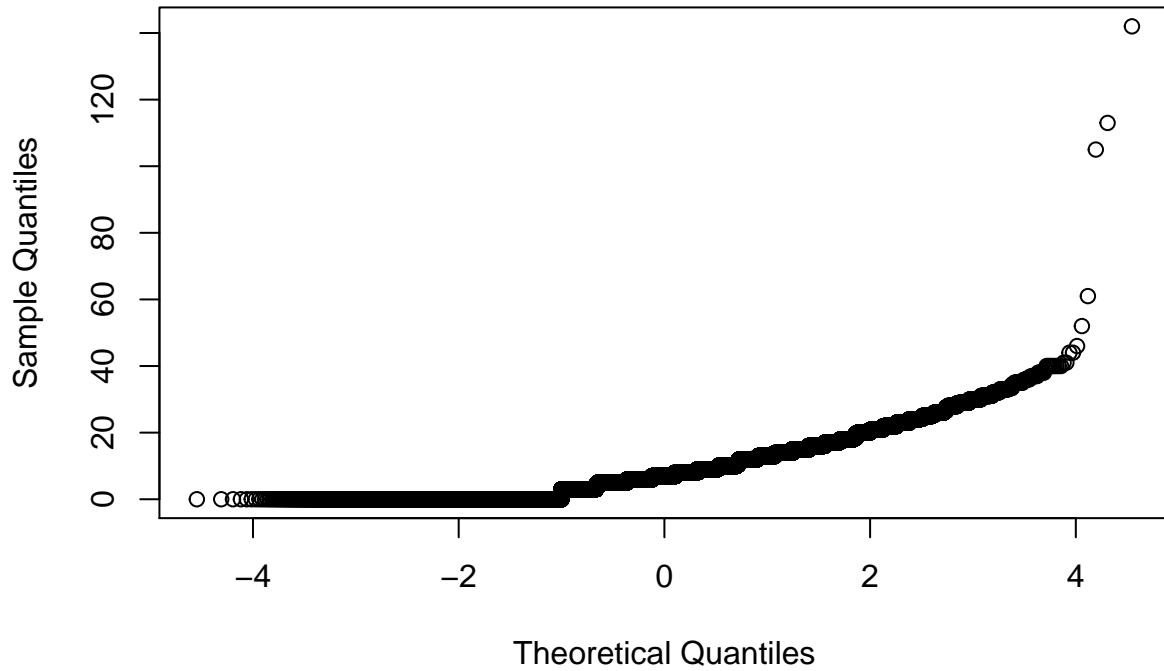
QQ Norm plot for Pressure



```
#The above plot also has a skewed distribution and needs to be normalized
```

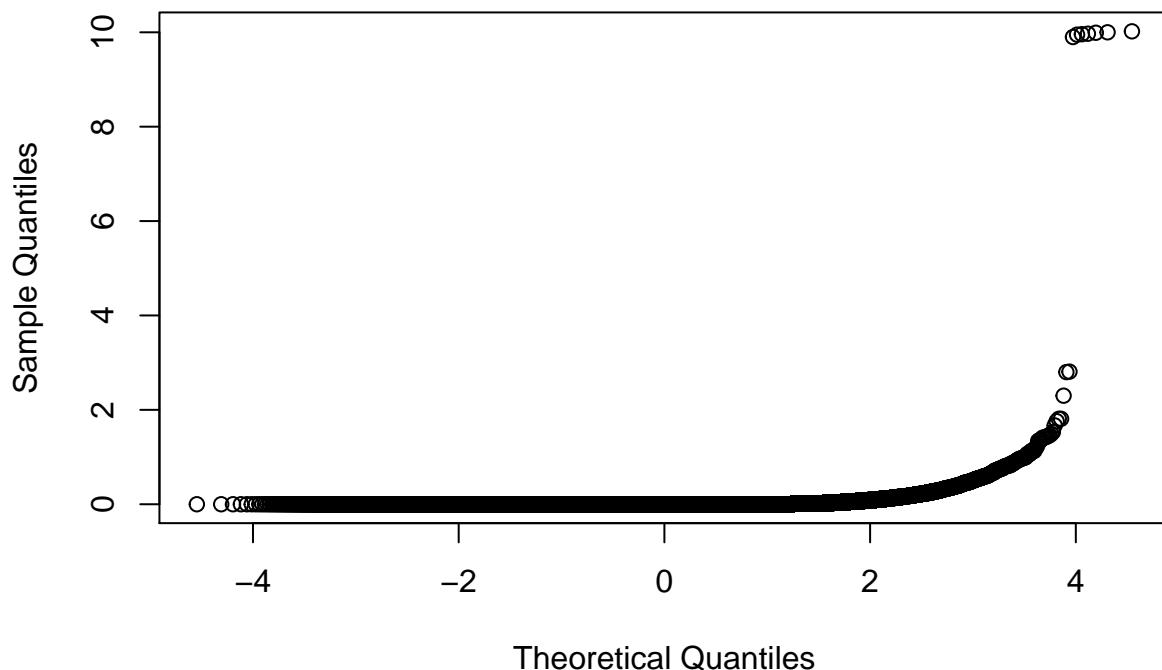
```
qqnorm(num[,c("Wind_Speed.mph.")],main="QQ Norm plot for Wind speed")
```

QQ Norm plot for Wind speed



```
#above plot shows slight skewness in the wind speed normality with presence of outliers that must be el  
qqnorm(num[,c("Precipitation.in.")],main="QQ Norm plot for Precipitation")
```

QQ Norm plot for Precipitation



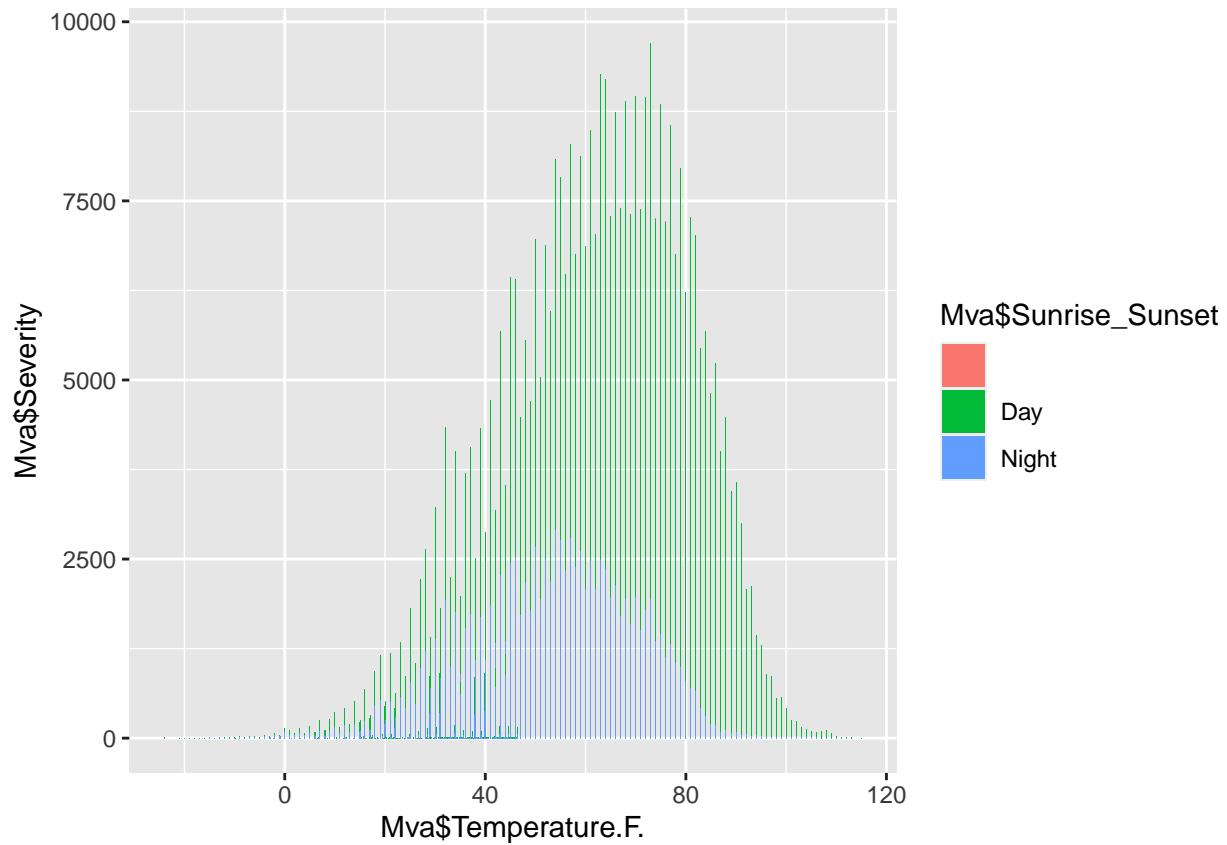
```
#Scaling the data
Mva$Precipitation.in.<-scale(Mva$Precipitation.in.)
#qqnorm(num[,c("Precipitation.in.")],main="QQ Norm plot for Precipitation")
#Above plot shows presence of outliers and slight skewness

## CATEGORICAL VARIABLE EXPLORATION ##

library(ggplot2)

#Some examples of observations noted

#1. Sunrise_Sunset : whether accidents happened more at night or day
ggplot(Mva,aes(fill=Mva$Sunrise_Sunset,y=Mva$Severity,x=Mva$Temperature.F.))+
  geom_bar(stat='identity')
```

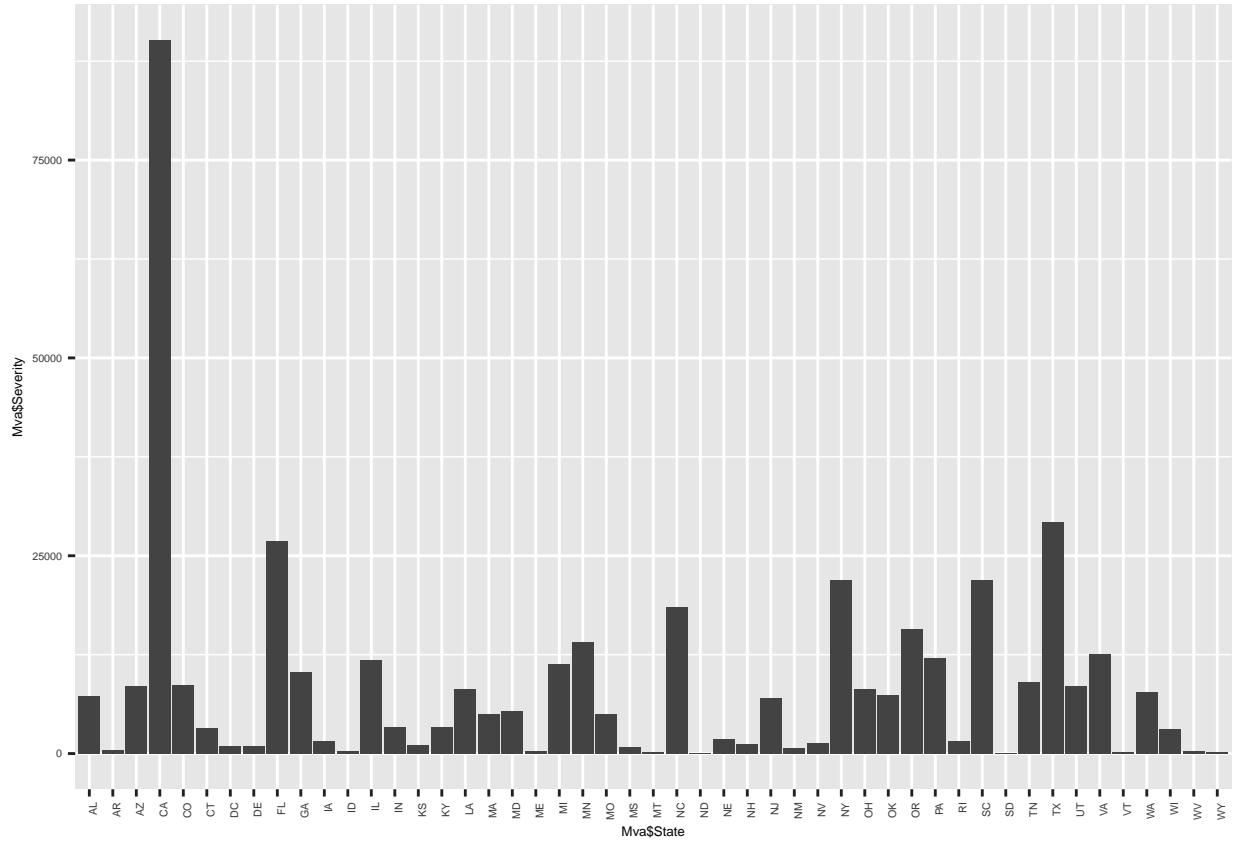


#Above graph shows that for the normally distributed temperature data, a high majority of the road accidents happened during the day
#Also, the peak was around 60 degrees F

#2. State :exploring which states had the most accidents and at what severity

```
ggplot(Mva, aes(x=Mva$State,y=Mva$Severity)) +
  geom_histogram(stat='identity',binwidth=50) +
  theme(text = element_text(size=5),
        axis.text.x = element_text(angle=90, hjust=1))
```

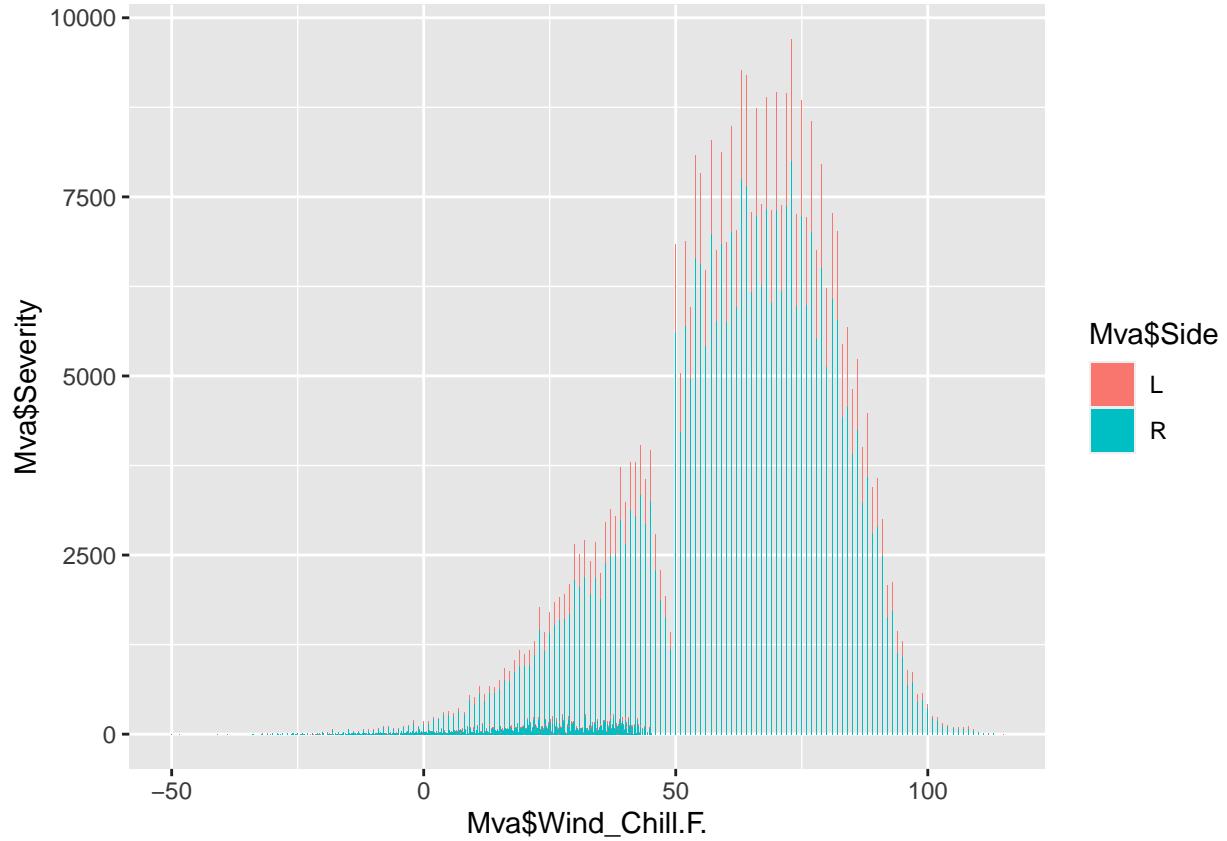
Warning: Ignoring unknown parameters: binwidth, bins, pad



#Here we observe that the most number of accidents took place in the state of California, Texas and Florida.

#Ques 3. Side: To find which side of the road was more prone to the event of accident.

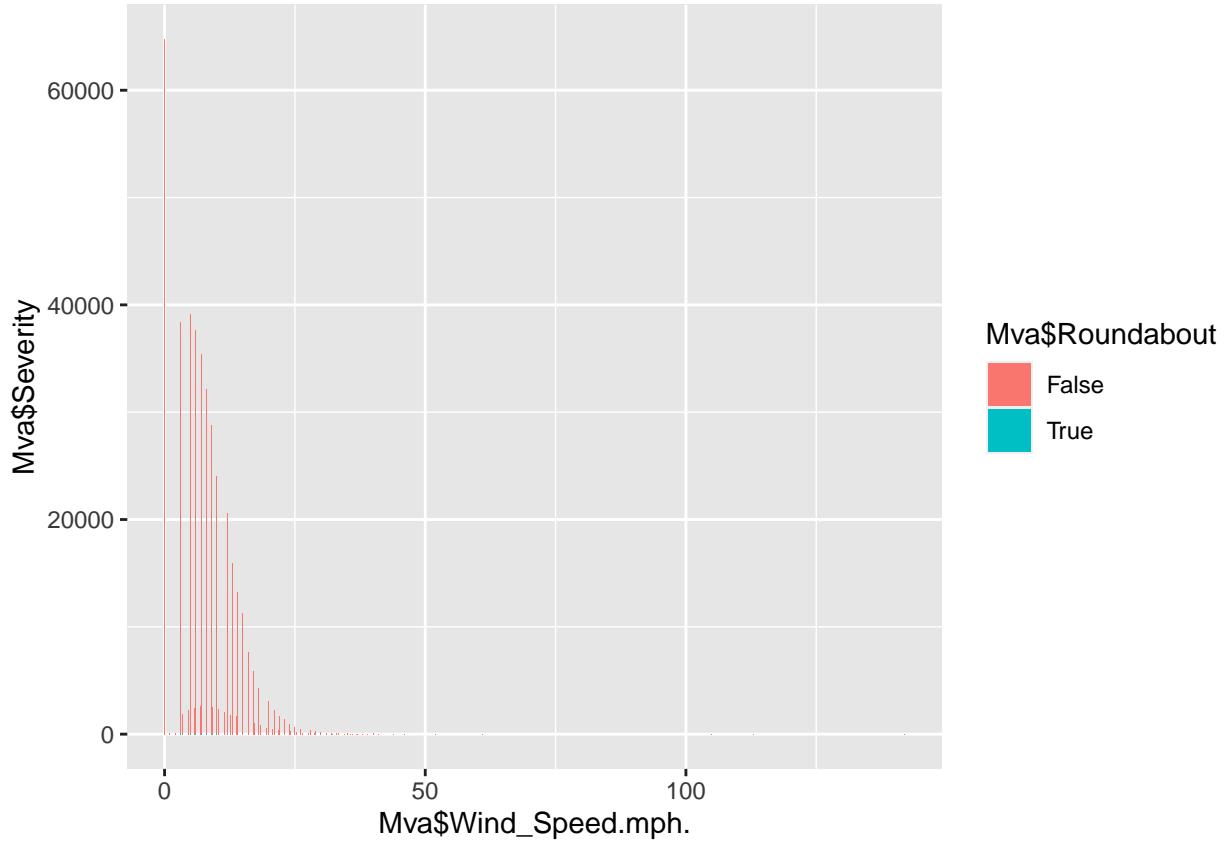
```
ggplot(Mva, aes(fill=Mva$Side, y=Mva$Severity, x=Mva$Wind_Chill.F.)) +
  geom_bar(position="stack", stat="identity")
```



```
#We observe that most accidents took place on the right side of the road.
```

```
#Ques 4. To fond whether the roundabout presence or absence was significant for the accident.
ggplot(Mva, aes(fill=Mva$Roundabout, y=Mva$Severity, x=Mva$Wind_Speed.mph.)) +
  geom_histogram(position="stack", stat="identity", xlim=c(0,1))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad, xlim
```



```
#Here we observe that mostly the roundabout was absent when the accident took place.
```

```
## PERFORMING HYPOTHESIS TESTS ## will be expanded more after further processing and normalizing of the
```

```
#t tests here determine whether the null hypothesis that the mean of the two samples is equal will be r
# Performing t-tests on various combinations of variables
```

```
t.test(Mva$Humidity...~Mva$Station)
```

```
##
## Welch Two Sample t-test
##
## data: Mva$Humidity... by Mva$Station
## t = 2.3901, df = 3848.4, p-value = 0.01689
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.155066 1.570797
## sample estimates:
## mean in group False mean in group True
##           65.54438          64.68145
```

```
# The p value comes out to be really small therefore the null hypothesis that the mean of both samples
```

```
t.test(Mva$Severity~Mva$Side,data=Mva)
```

```
##  
## Welch Two Sample t-test  
##  
## data: Mva$Severity by Mva$Side  
## t = -66.356, df = 59222, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.2013200 -0.1897682  
## sample estimates:  
## mean in group L mean in group R  
## 2.121470 2.317015
```

#Here again it is rejected, since p-value is extremely small