Road Traffic Accidents Analysis

Ayman A. Tuffaha

4/22/2022

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Introduction

The dataset used at this assignment includes Road Traffic Accidents records in US from the year of 2016 to the year of 2021. Datasets can be found here:
https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?select=US_Accidents_Dec21_updated.csv

For more information and description about the dataset, Please visit the below URL:
https://smoosavi.org/datasets/us_accidents

Shows structure of the data frame:

```
## 'data.frame':    2845342 obs. of  47 variables:
##  $ ID                   : chr  "A-1" "A-2" "A-3" "A-4" ...
##  $ Severity             : int  3 2 2 2 3 2 2 2 2 2 ...
##  $ Start_Time           : chr  "2016-02-08 00:37:08" "2016-02-08
05:56:20" "2016-02-08 06:15:39" "2016-02-08 06:51:45" ...
##  $ End_Time             : chr  "2016-02-08 06:37:08" "2016-02-08
11:56:20" "2016-02-08 12:15:39" "2016-02-08 12:51:45" ...
##  $ Start_Lat            : num  40.1 39.9 39.1 41.1 39.2 ...
##  $ Start_Lng            : num  -83.1 -84.1 -84.5 -81.5 -84.5 ...
##  $ End_Lat              : num  40.1 39.9 39.1 41.1 39.2 ...
##  $ End_Lng              : num  -83 -84 -84.5 -81.5 -84.5 ...
##  $ Distance.mi.         : num  3.23 0.747 0.055 0.123 0.5 ...
##  $ Description          : chr  "Between Sawmill Rd/Exit 20 and OH-
315/Olentangy Riv Rd/Exit 22 - Accident." "At OH-4/OH-235/Exit 41 -
Accident." "At I-71/US-50/Exit 1 - Accident." "At Dart Ave/Exit 21 -
Accident." ...
```

```
##  $ Number               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Street               : chr  "Outerbelt E" "I-70 E" "I-75 S" "I-77
N" ...
##  $ Side                 : chr  "R" "R" "R" "R" ...
##  $ City                 : chr  "Dublin" "Dayton" "Cincinnati"
"Akron" ...
##  $ County               : chr  "Franklin" "Montgomery" "Hamilton"
"Summit" ...
##  $ State                : chr  "OH" "OH" "OH" "OH" ...
##  $ Zipcode              : chr  "43017" "45424" "45203" "44311" ...
##  $ Country              : chr  "US" "US" "US" "US" ...
##  $ Timezone             : chr  "US/Eastern" "US/Eastern"
"US/Eastern" "US/Eastern" ...
##  $ Airport_Code         : chr  "KOSU" "KFFO" "KLUK" "KAKR" ...
##  $ Weather_Timestamp    : chr  "2016-02-08 00:53:00" "2016-02-08
05:58:00" "2016-02-08 05:53:00" "2016-02-08 06:54:00" ...
##  $ Temperature.F.       : num  42.1 36.9 36 39 37 35.6 33.8 33.1 39
32 ...
##  $ Wind_Chill.F.        : num  36.1 NA NA NA 29.8 29.2 NA 30 31.8
28.7 ...
##  $ Humidity...          : num  58 91 97 55 93 100 100 92 70 100 ...
##  $ Pressure.in.         : num  29.8 29.7 29.7 29.6 29.7 ...
##  $ Visibility.mi.       : num  10 10 10 10 10 10 3 0.5 10 0.5 ...
##  $ Wind_Direction       : chr  "SW" "Calm" "Calm" "Calm" ...
##  $ Wind_Speed.mph.      : num  10.4 NA NA NA 10.4 8.1 2.3 3.5 11.5
3.5 ...
##  $ Precipitation.in.    : num  0 0.02 0.02 NA 0.01 NA NA 0.08 NA
0.05 ...
##  $ Weather_Condition    : chr  "Light Rain" "Light Rain" "Overcast"
"Overcast" ...
##  $ 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" "True" "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" "Night" "Night" "Night" ...
##  $ Civil_Twilight       : chr  "Night" "Night" "Night" "Night" ...
##  $ Nautical_Twilight    : chr  "Night" "Night" "Night" "Day" ...
##  $ Astronomical_Twilight: chr  "Night" "Night" "Day" "Day" ...

## [1] 47
```

```
## [1] 2845342

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

Data Exploration:

Shows descriptive statistics of each category:

```
## Rows: 2,845,342
## Columns: 47
## $ ID                    <chr> "A-1", "A-2", "A-3", "A-4", "A-5", "A-
6", "A-7",~
## $ Severity              <int> 3, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2,
2, 3, 3, ~
## $ Start_Time            <chr> "2016-02-08 00:37:08", "2016-02-08
05:56:20", "2~
## $ End_Time              <chr> "2016-02-08 06:37:08", "2016-02-08
11:56:20", "2~
## $ Start_Lat             <dbl> 40.10891, 39.86542, 39.10266,
41.06213, 39.17239~
## $ Start_Lng             <dbl> -83.09286, -84.06280, -84.52468, -
81.53784, -84.~
## $ End_Lat               <dbl> 40.11206, 39.86501, 39.10209,
41.06217, 39.17048~
## $ End_Lng               <dbl> -83.03187, -84.04873, -84.52396, -
81.53547, -84.~
## $ Distance.mi.          <dbl> 3.230, 0.747, 0.055, 0.123, 0.500,
1.427, 0.227,~
## $ Description           <chr> "Between Sawmill Rd/Exit 20 and OH-
```

```
315/Olentangy~
## $ Number                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, ~
## $ Street                <chr> "Outerbelt E", "I-70 E", "I-75 S", "I-
77 N", "I-~
## $ Side                  <chr> "R", "R", "R", "R", "R", "R", "R",
"R", "R", "R"~
## $ City                  <chr> "Dublin", "Dayton", "Cincinnati",
"Akron", "Cinc~
## $ County                <chr> "Franklin", "Montgomery", "Hamilton",
"Summit", ~
## $ State                 <chr> "OH", "OH", "OH", "OH", "OH", "OH",
"OH", "OH", ~
## $ Zipcode               <chr> "43017", "45424", "45203", "44311",
"45217", "45~
## $ Country               <chr> "US", "US", "US", "US", "US", "US",
"US", "US", ~
## $ Timezone              <chr> "US/Eastern", "US/Eastern",
"US/Eastern", "US/Ea~
## $ Airport_Code          <chr> "KOSU", "KFFO", "KLUK", "KAKR",
"KLUK", "KI69", ~
## $ Weather_Timestamp     <chr> "2016-02-08 00:53:00", "2016-02-08
05:58:00", "2~
## $ Temperature.F.        <dbl> 42.1, 36.9, 36.0, 39.0, 37.0, 35.6,
33.8, 33.1, ~
## $ Wind_Chill.F.         <dbl> 36.1, NA, NA, NA, 29.8, 29.2, NA,
30.0, 31.8, 28~
## $ Humidity...           <dbl> 58, 91, 97, 55, 93, 100, 100, 92, 70,
100, 100, ~
## $ Pressure.in.          <dbl> 29.76, 29.68, 29.70, 29.65, 29.69,
29.66, 29.63,~
## $ Visibility.mi.        <dbl> 10.0, 10.0, 10.0, 10.0, 10.0, 10.0,
3.0, 0.5, 10~
## $ Wind_Direction        <chr> "SW", "Calm", "Calm", "Calm", "WSW",
"WSW", "SW"~
## $ Wind_Speed.mph.       <dbl> 10.4, NA, NA, NA, 10.4, 8.1, 2.3, 3.5,
11.5, 3.5~
## $ Precipitation.in.     <dbl> 0.00, 0.02, 0.02, NA, 0.01, NA, NA,
0.08, NA, 0.~
## $ Weather_Condition     <chr> "Light Rain", "Light Rain",
"Overcast", "Overcas~
## $ Amenity               <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Bump                  <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Crossing              <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Give_Way              <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Junction              <chr> "False", "False", "True", "False",
```

```
"False", "Fal~
## $ No_Exit               <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Railway               <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Roundabout            <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Station               <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Stop                  <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Traffic_Calming       <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Traffic_Signal        <chr> "False", "False", "False", "False",
"False", "Tr~
## $ Turning_Loop          <chr> "False", "False", "False", "False",
"False", "Fa~
## $ Sunrise_Sunset        <chr> "Night", "Night", "Night", "Night",
"Day", "Day"~
## $ Civil_Twilight        <chr> "Night", "Night", "Night", "Night",
"Day", "Day"~
## $ Nautical_Twilight     <chr> "Night", "Night", "Night", "Day",
"Day", "Day", ~
## $ Astronomical_Twilight <chr> "Night", "Night", "Day", "Day", "Day",
"Day", "D~

## [1] "data.frame"

## [1] 2845342      47

##       ID                Severity      Start_Time          End_Time
##  Length:2845342     Min.   :1.000   Length:2845342
Length:2845342
##  Class :character   1st Qu.:2.000   Class :character   Class
:character
##  Mode  :character   Median :2.000   Mode  :character   Mode
:character
##                     Mean   :2.138
##                     3rd Qu.:2.000
##                     Max.   :4.000
##
##    Start_Lat       Start_Lng          End_Lat         End_Lng
##  Min.   :24.57   Min.   :-124.55   Min.   :24.57   Min.   :-124.55
##  1st Qu.:33.45   1st Qu.:-118.03   1st Qu.:33.45   1st Qu.:-118.03
##  Median :36.10   Median : -92.42   Median :36.10   Median : -92.42
##  Mean   :36.25   Mean   : -97.11   Mean   :36.25   Mean   : -97.11
##  3rd Qu.:40.16   3rd Qu.: -80.37   3rd Qu.:40.16   3rd Qu.: -80.37
##  Max.   :49.00   Max.   : -67.11   Max.   :49.08   Max.   : -67.11
##
##   Distance.mi.      Description            Number           Street
```

```
##  Min.   :  0.0000   Length:2845342     Min.   :      0
Length:2845342
##  1st Qu.:  0.0520   Class :character   1st Qu.:   1270   Class
:character
##  Median :  0.2440   Mode  :character   Median :   4007   Mode
:character
##  Mean   :  0.7027                      Mean   :   8089
##  3rd Qu.:  0.7640                      3rd Qu.:   9567
##  Max.   :155.1860                      Max.   :9999997
##                                        NA's   :1743911
##      Side               City              County              State
##  Length:2845342     Length:2845342     Length:2845342
Length:2845342
##  Class :character   Class :character   Class :character   Class
:character
##  Mode  :character   Mode  :character   Mode  :character   Mode
:character
##
##
##
##
##    Zipcode            Country            Timezone
Airport_Code
##  Length:2845342     Length:2845342     Length:2845342
Length:2845342
##  Class :character   Class :character   Class :character   Class
:character
##  Mode  :character   Mode  :character   Mode  :character   Mode
:character
##
##
##
##
##  Weather_Timestamp  Temperature.F.   Wind_Chill.F.     Humidity...
##  Length:2845342     Min.   :-89.00   Min.   :-89.0    Min.   :  1.00
##  Class :character   1st Qu.: 50.00   1st Qu.: 46.0    1st Qu.: 48.00
##  Mode  :character   Median : 64.00   Median : 63.0    Median : 67.00
##                     Mean   : 61.79   Mean   : 59.7    Mean   : 64.37
##                     3rd Qu.: 76.00   3rd Qu.: 76.0    3rd Qu.: 83.00
##                     Max.   :196.00   Max.   :196.0    Max.   :100.00
##                     NA's   :69274    NA's   :469643   NA's   :73092
##   Pressure.in.   Visibility.mi.  Wind_Direction     Wind_Speed.mph.
##  Min.   : 0.00   Min.   :  0.0   Length:2845342     Min.   :   0.0
##  1st Qu.:29.31   1st Qu.: 10.0   Class :character   1st Qu.:   3.5
##  Median :29.82   Median : 10.0   Mode  :character   Median :   7.0
##  Mean   :29.47   Mean   :  9.1                      Mean   :   7.4
##  3rd Qu.:30.01   3rd Qu.: 10.0                      3rd Qu.:  10.0
##  Max.   :58.90   Max.   :140.0                      Max.   :1087.0
##  NA's   :59200   NA's   :70546                      NA's   :157944
##  Precipitation.in. Weather_Condition    Amenity             Bump
```

```
##  Min.   : 0        Length:2845342     Length:2845342
Length:2845342
##  1st Qu.: 0        Class :character   Class :character   Class
:character
##  Median : 0        Mode  :character   Mode  :character   Mode
:character
##  Mean   : 0
##  3rd Qu.: 0
##  Max.   :24
##  NA's   :549458
##    Crossing           Give_Way           Junction           No_Exit
##  Length:2845342     Length:2845342     Length:2845342
Length:2845342
##  Class :character   Class :character   Class :character   Class
:character
##  Mode  :character   Mode  :character   Mode  :character   Mode
:character
##
##
##
##
##    Railway            Roundabout          Station               Stop
##  Length:2845342     Length:2845342     Length:2845342
Length:2845342
##  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:2845342     Length:2845342     Length:2845342
Length:2845342
##  Class :character   Class :character   Class :character   Class
:character
##  Mode  :character   Mode  :character   Mode  :character   Mode
:character
##
##
##
##
##  Civil_Twilight     Nautical_Twilight  Astronomical_Twilight
##  Length:2845342     Length:2845342     Length:2845342
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##
```

Number of Accidents By State:

```
[1] 49

 [1] "OH" "IN" "KY" "WV" "MI" "PA" "CA" "NV" "MN" "TX" "MO" "CO" "OK"
"LA" "KS" "WI" "IA" "MS" "NE" "ND"

[21] "WY" "SD" "MT" "NM" "AR" "IL" "NJ" "GA" "FL" "NY" "CT" "RI" "SC"
"NC" "MD" "MA" "TN" "VA" "DE" "DC"

[41] "ME" "AL" "NH" "VT" "AZ" "UT" "ID" "OR" "WA"
```

```
  x   freq

1  AL  19322

2  AR  10935

3  AZ  56504

4  CA 795868

5  CO  25340

6  CT  29762

7  DC   9133

8  DE   4842

9  FL 401388

10 GA  40086

11 IA   9607

12 ID   8544

13 IL  47105

14 IN  20850

15 KS   9033

16 KY   6638

17 LA  47232

18 MA   6392

19 MD  65085
```

```
20 ME   2193
21 MI  43843
22 MN  97185
23 MO  29633
24 MS   5320
25 MT  15964
26 NC  91362
27 ND   2258
28 NE   3320
29 NH   3866
30 NJ  52902
31 NM   2370
32 NV   6197
33 NY 108049
34 OH  24409
35 OK   8806
36 OR 126341
37 PA  99975
38 RI   4451
39 SC  89216
40 SD    201
41 TN  52613
42 TX 149037
43 UT  49193
44 VA 113535
45 VT    365
46 WA  32554
47 WI   7896
```

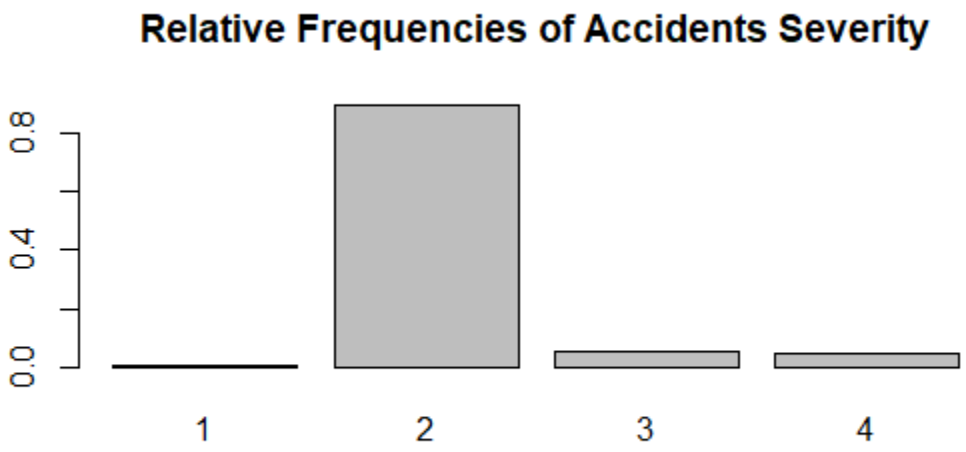
```
48 WV   7632

49 WY    990
```

Analyse and visualize data:

1) Analyse Data: Severity of accidents Analysis

We noticed the most number of accidents Severity is severity 2

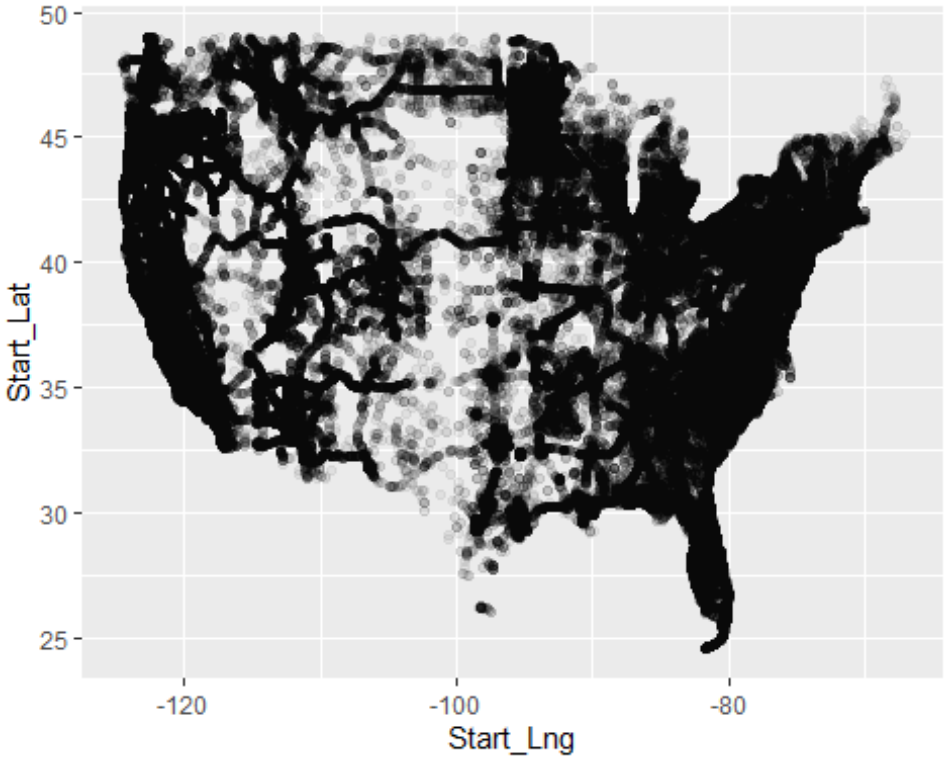


2) Analyse Data: Longitude and Latitude Analysis

Distribution of accidents across US map, We found that more accidents occur in urban areas across the east, west, and middle areas of the USA

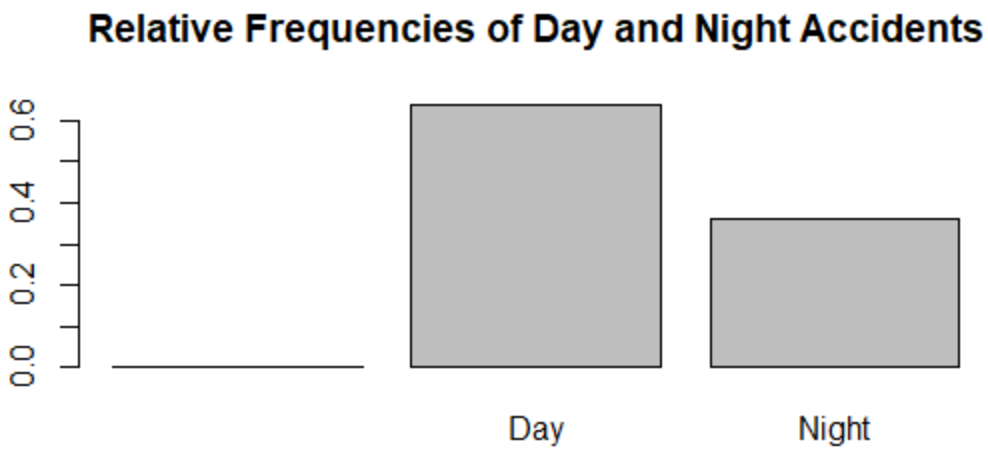
```
##               Number     Precipitation.in.         Wind_Chill.F.
##              1743911                549458                469643
##       Wind_Speed.mph.          Humidity...       Visibility.mi.
##               157944                 73092                 70546
##        Temperature.F.         Pressure.in.                    ID
##                69274                 59200                     0
##              Severity           Start_Time              End_Time
##                     0                    0                     0
##             Start_Lat           Start_Lng               End_Lat
##                     0                    0                     0
##               End_Lng         Distance.mi.           Description
##                     0                    0                     0
##                Street                 Side                  City
##                     0                    0                     0
##                County                State               Zipcode
##                     0                    0                     0
##               Country             Timezone          Airport_Code
```

```
##                    0                    0                    0
##     Weather_Timestamp        Wind_Direction     Weather_Condition
##                    0                    0                    0
##              Amenity                 Bump             Crossing
##                    0                    0                    0
##             Give_Way             Junction              No_Exit
##                    0                    0                    0
##              Railway           Roundabout              Station
##                    0                    0                    0
##                 Stop      Traffic_Calming      Traffic_Signal
##                    0                    0                    0
##         Turning_Loop       Sunrise_Sunset      Civil_Twilight
##                    0                    0                    0
##     Nautical_Twilight Astronomical_Twilight
##                    0                    0
```



3) Analyse Data: Day and Night Accidents Analysis

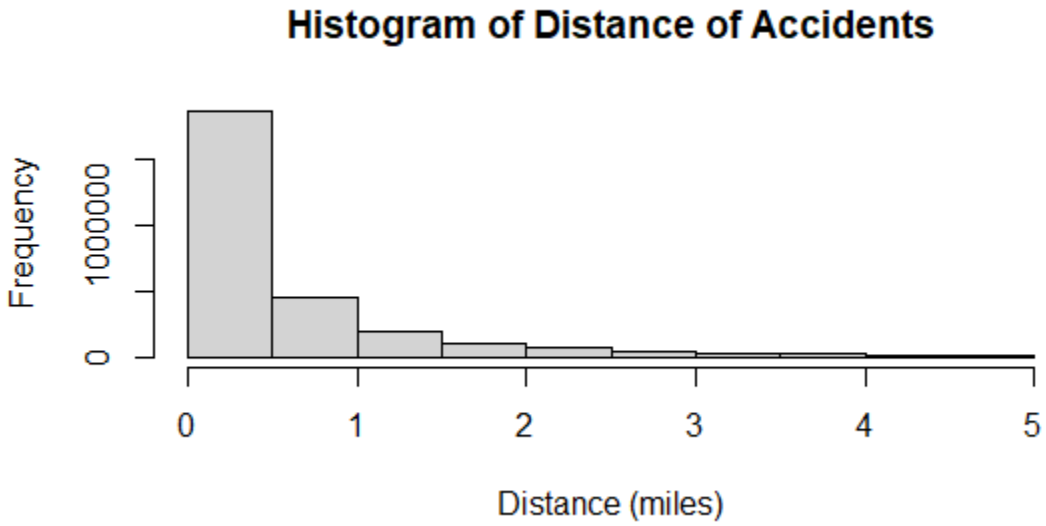
This frequency table is for the sunrise_sunset variable. The results shows the majority of accidents occurs at the Day. The explanation of this results because most of the people outdoor at day and most of Ruch hour / crowdies happened at day.



4) Analyse Data: Distance of Accidents Analysis

This histogram analysis is for the Distance.mi variable. Which means the amount of road space affected by the accidents.

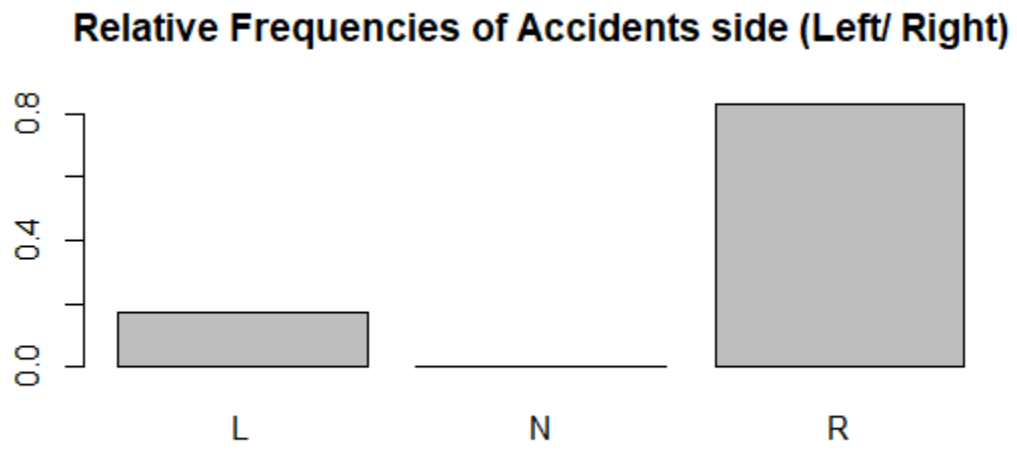
The results shows the majority of accidents not affect that much of the road space and the results was less than 1 mile.



5) Analyse Data: Side of Accidents Analysis (Left/ Right Side)

This frequency table is for the Side variable. That means which side of the road the accidents were reported is it Left/ Right Side.

The results shows the majority of accidents located at Right side of the road.



Create a Linear Regression Model of a Data:

```
## Call:
## lm(formula = as.numeric(Severity) ~ State + Pressure.in. +
##     Humidity... + Temperature.F. + Wind_Speed.mph., data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4152 -0.3693 -0.3333  0.6177  1.7735
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.061e+00  1.196e-01  17.238  < 2e-16 ***
## State            -1.463e-05  2.531e-05  -0.578 0.563279
## Pressure.in.     1.139e-02  3.963e-03   2.874 0.004061 **
## Humidity...      4.680e-04  1.246e-04   3.755 0.000174 ***
## Temperature.F.  -5.820e-04  1.625e-04  -3.582 0.000341 ***
## Wind_Speed.mph.  1.562e-03  5.418e-04   2.884 0.003933 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5367 on 34995 degrees of freedom
## Multiple R-squared:  0.004742,   Adjusted R-squared:  0.004572
## F-statistic: 27.79 on 6 and 34995 DF,  p-value: < 2.2e-16
```

```
##     (Intercept)      State       Pressure.in.     Humidity...
##       7.8526958       0.9999854       1.0114533       1.0004681
##  Temperature.F. Wind_Speed.mph.
##       0.9994181       1.0015636
```

Summary:

1- The Dataset is huge and I found difficulties in the training the dataset.

2- I believe there are missed some important variables/ Features like Driver gender, Age, Years of license issuance which was not exist at the dataset

3- The Precipitation variable contains a lot of NA values in the dataset.

4- The most dangerous turn in the US is turning right.