TEAM 5 EDA - Accidents Dataset

Jayatha Chandra, Avanti Dorle, Lavina Talreja, Mansi Thanki

04/23/2023

Contents

Import necessary libraries

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.2.1     v stringr 1.5.0
## v tidyr   1.3.0     v forcats 0.5.2
## v purrr   1.0.1
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'purrr' was built under R version 4.1.2
```

```
## Warning: package 'stringr' was built under R version 4.1.2
```

```
## Warning: package 'forcats' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(RSQLite)
```

```
## Warning: package 'RSQLite' was built under R version 4.1.2
```

```
library(tidyr)
library(stringr)
library(modelr)
```

```
## Warning: package 'modelr' was built under R version 4.1.2
```

```
library(testthat)
```

```
## Warning: package 'testthat' was built under R version 4.1.2
```

```
##
## Attaching package: 'testthat'
##
## The following object is masked from 'package:purrr':
##
##     is_null
```

```
##
## The following object is masked from 'package:tidyr':
##
##     matches
##
## The following objects are masked from 'package:readr':
##
##     edition_get, local_edition
##
## The following object is masked from 'package:dplyr':
##
##     matches
```

```
library(assertive)
```

```
##
## Attaching package: 'assertive'
##
## The following objects are masked from 'package:testthat':
##
##     has_names, is_false, is_less_than, is_null, is_true
##
## The following objects are masked from 'package:purrr':
##
##     is_atomic, is_character, is_double, is_empty, is_formula,
##     is_function, is_integer, is_list, is_logical, is_null, is_vector
##
## The following object is masked from 'package:tibble':
##
##     has_rownames
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.1.2
```

```
## Loading required package: timechange
```

```
## Warning: package 'timechange' was built under R version 4.1.2
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

Import clean data that we generated from Python notebook after preprocessing.

```
df <- read.csv('clean_data.csv')
```

```
as.tibble(df)
```

```
## Warning: `as.tibble()` was deprecated in tibble 2.0.0.
## i Please use `as_tibble()` instead.
## i The signature and semantics have changed, see `?as_tibble`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## # A tibble: 2,845,342 x 46
##    ID    Severity Start_Time        End_Time Start_Lat Start_Lng End_Lat End_Lng
##    <chr>    <int> <chr>             <chr>        <dbl>     <dbl>   <dbl>   <dbl>
##  1 A-1          3 2016-02-08 00:37~ 2016-02~      40.1     -83.1    40.1   -83.0
##  2 A-2          2 2016-02-08 05:56~ 2016-02~      39.9     -84.1    39.9   -84.0
##  3 A-3          2 2016-02-08 06:15~ 2016-02~      39.1     -84.5    39.1   -84.5
##  4 A-4          2 2016-02-08 06:51~ 2016-02~      41.1     -81.5    41.1   -81.5
##  5 A-5          3 2016-02-08 07:53~ 2016-02~      39.2     -84.5    39.2   -84.5
##  6 A-6          2 2016-02-08 08:16~ 2016-02~      39.1     -84.0    39.1   -84.1
##  7 A-7          2 2016-02-08 08:15~ 2016-02~      39.8     -84.2    39.8   -84.2
##  8 A-8          2 2016-02-08 11:51~ 2016-02~      41.4     -81.8    41.4   -81.8
##  9 A-9          2 2016-02-08 14:19~ 2016-02~      40.7     -84.1    40.7   -84.1
## 10 A-10         2 2016-02-08 15:16~ 2016-02~      40.1     -83.0    40.1   -83.0
## # i 2,845,332 more rows
## # i 38 more variables: Distance.mi. <dbl>, Description <chr>, Street <chr>,
## #   Side <chr>, City <chr>, County <chr>, State <chr>, Zipcode <chr>,
## #   Country <chr>, Timezone <chr>, Airport_Code <chr>, Weather_Timestamp <chr>,
## #   Temperature.F. <dbl>, Wind_Chill.F. <dbl>, Humidity <dbl>, Pressure <dbl>,
## #   Visibility <dbl>, Wind_Direction <chr>, Wind_Speed <dbl>,
## #   Precipitation <dbl>, Weather_Condition <chr>, Amenity <chr>, ...
```

```
sum(is.na(df))
```

```
## [1] 0
```

```
length(unique(df$City))
```

```
## [1] 11681
```

Exploratory Data Analysis

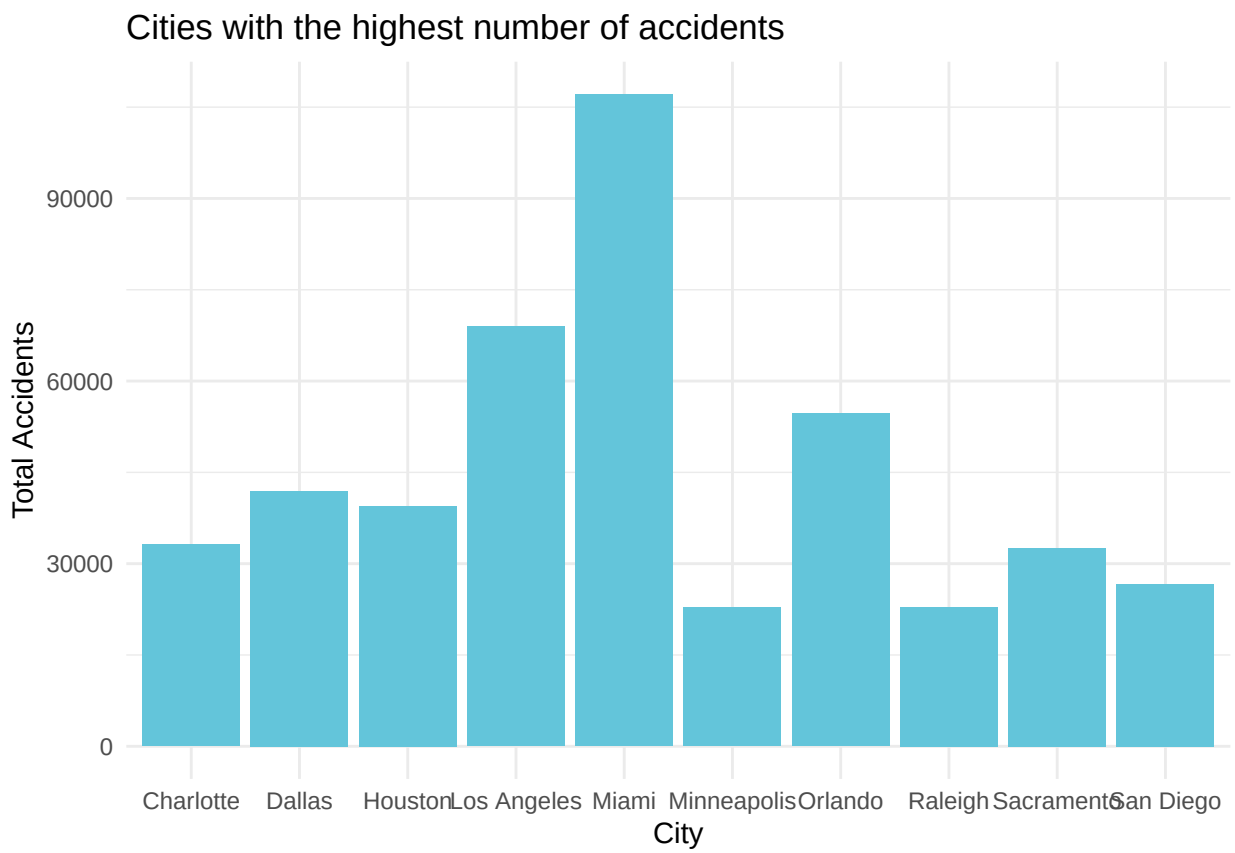
Location based analysis

Top 10 cities with highest number of accidents

```
top10_city <- df %>%
  group_by(City) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(10)

as.tibble(top10_city)
```

```
## # A tibble: 10 x 2
##    City         count
##    <chr>        <int>
##  1 Miami       107103
##  2 Los Angeles  68956
##  3 Orlando      54691
##  4 Dallas       41979
##  5 Houston      39448
##  6 Charlotte    33152
##  7 Sacramento   32559
##  8 San Diego    26627
##  9 Raleigh      22840
## 10 Minneapolis  22768
```

```
ggplot(top10_city, aes(x=City, y=count)) +
  geom_bar(stat = "identity", fill = "#63C5Da") + theme_minimal() +
  labs(title="Cities with the highest number of accidents",
       x="City", y="Total Accidents")
```



We can observe that the Miami has highest number of accidents whereas Minneapolis has the lowest number of accidents. This can be due to the population and number of people taking vacations in Miami.

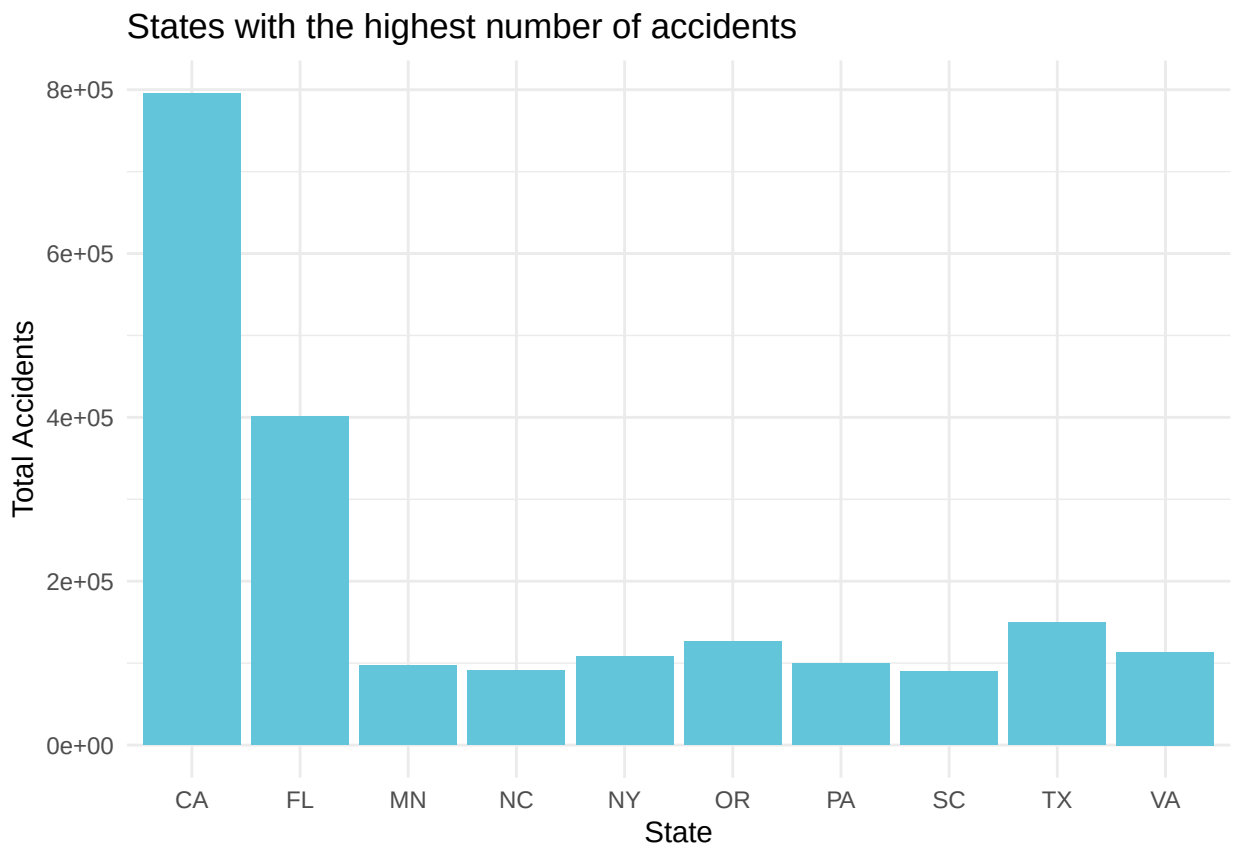
Top 10 states with highest number of accidents

```
top10_states <- df %>%
  group_by(State) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(10)

as.tibble(top10_states)
```

```
## # A tibble: 10 x 2
##    State  count
##    <chr>  <int>
##  1 CA    795868
##  2 FL    401388
##  3 TX    149037
##  4 OR    126341
##  5 VA    113535
##  6 NY    108049
##  7 PA     99975
##  8 MN     97185
```

```
##  9 NC     91362
## 10 SC     89216
```

```
ggplot(top10_states, aes(x=State, y=count)) +
  geom_bar(stat = "identity", fill = "#63C5Da") + theme_minimal() +
  labs(title="States with the highest number of accidents",
       x="State", y="Total Accidents")
```

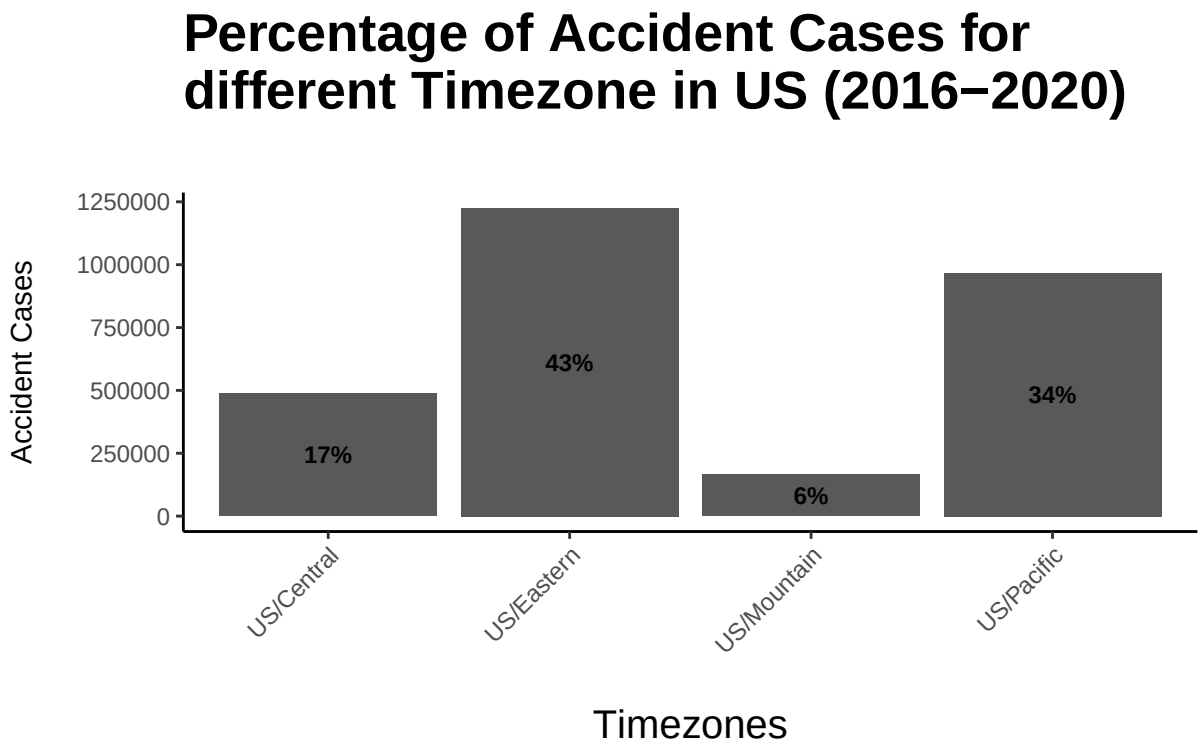


We can observe that the CA has highest number of accidents whereas SC has the lowest number of accidents. This can be due to the population and number of people taking vacations to CA.

Timezone based accidents

```
# Bar Plot showing number of accidents that occured in different time zones
library(ggplot2)
library(ggplotify)


timezone_df <- df %>%
  count(Timezone) %>%
  rename(Timezone = "Timezone", Cases = "n")


# create a bar plot with the specified dimensions and resolution
```

```r
fig <- ggplot(data = timezone_df, aes(x = Timezone, y = Cases)) +
  geom_bar(stat = 'identity') +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.border = element_blank(), axis.line = element_line(colour = 'black'),
        legend.position = 'bottom', legend.title = element_blank(),
        axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))

# add labels to the bars
total <- nrow(df)
fig + geom_text(aes(label = paste0(round(Cases/total * 100), '%')),
                position = position_stack(vjust = 0.5), size = 3, fontface = 'bold', color = 'black') +
  labs(title = '\nPercentage of Accident Cases for \ndifferent Timezone in US (2016-2020)\n',
       x = '\nTimezones\n', y = '\nAccident Cases\n') +
  theme(plot.title = element_text(size = 20, face = 'bold'),
        axis.title.x = element_text(size = 15))
```



Eastern time zone region in the US has the highest no. of road accident cases (43%) in past years. Mountain time zone region in the US has the lowest no. of road accident cases (6%) in past years.

Weather based analysis

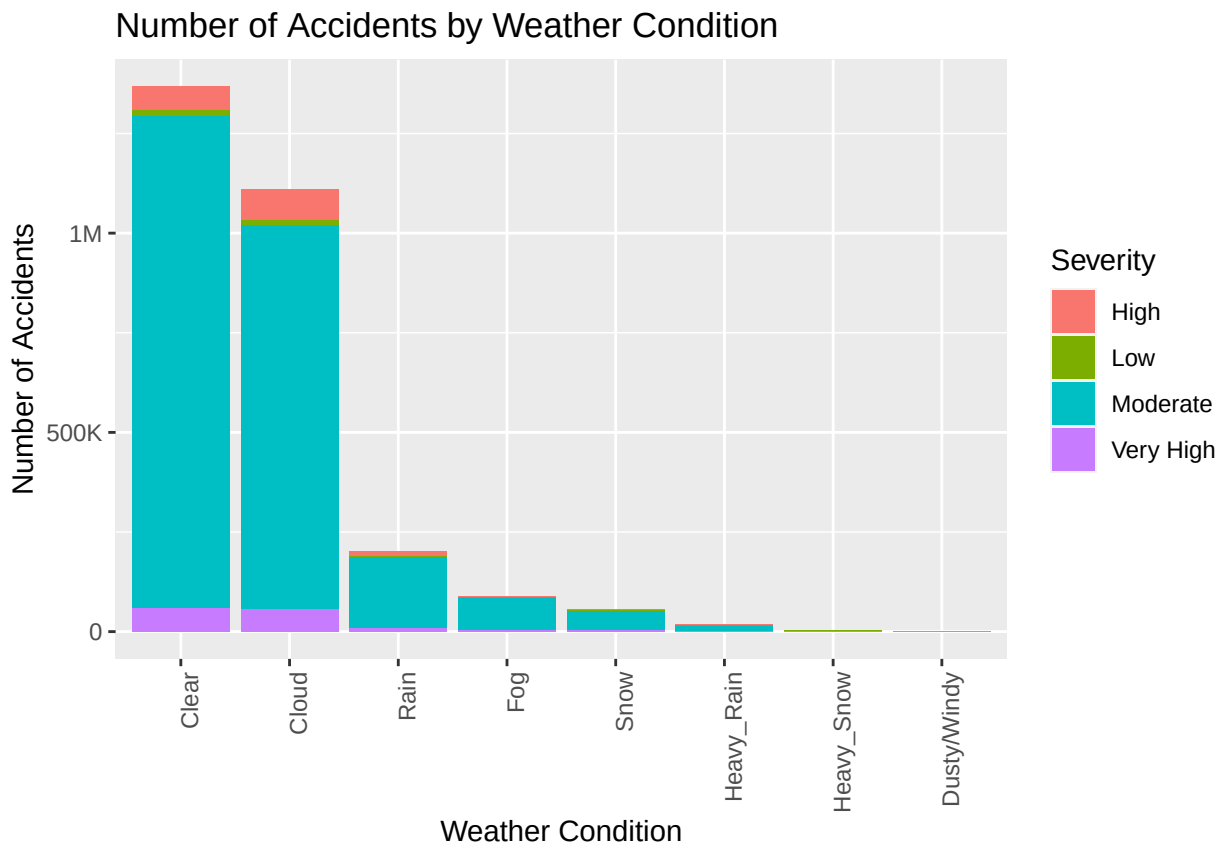
Severity of accidents based on Weather conditions

```
# Summarize the number of accidents by weather condition
accidents_weather <- df %>%
  mutate( Severity = case_when(
            `Severity` == 1 ~ "Low",
            `Severity` == 2 ~ "Moderate",
            `Severity` == 3 ~ "High",
            `Severity` == 4 ~ "Very High",
            TRUE ~ "Unknown"
         )) %>%
  drop_na() %>%
  group_by(Weather_Condition, Severity) %>%
  summarize(num_accidents = n()) %>%
  arrange(desc(num_accidents))
```

```
## `summarise()` has grouped output by 'Weather_Condition'. You can override using
## the `.groups` argument.
```

```
# Create a bar chart of the number of accidents by weather condition
ggplot(accidents_weather, aes(x = reorder(Weather_Condition, -num_accidents),
                              y = num_accidents, fill=Severity)) +
  geom_bar(stat = "identity") +
  xlab("Weather Condition") +
  ylab("Number of Accidents") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Number of Accidents by Weather Condition")+
  scale_y_continuous(labels = scales::label_number_si())
```

```
## Warning: `label_number_si()` was deprecated in scales 1.2.0.
## i Please use the `scale_cut` argument of `label_number()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



We can observe the clear weather condition mostly shows moderate level severity in most cases. The Cloudy weather does show some high and very high sever accidents.

Severity of accidents based on Precipitation, Wind speed, Temperature, Pressure

```
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.1.2
```

```
# plot 1: For Precipitation

plot1 <- df %>%
  mutate(PrecipitationRange = cut(Precipitation, breaks = seq(0, 1, 0.1),
                                  right = TRUE),
         Severity = case_when(
           `Severity` == 1 ~ "Low",
           `Severity` == 2 ~ "Moderate",
           `Severity` == 3 ~ "High",
           `Severity` == 4 ~ "Very High",
           TRUE ~ "Unknown"
         )) %>%
  drop_na() %>%
  group_by(PrecipitationRange, Severity) %>%
  summarise(AccidentCount = n()) %>%
```

```
  ggplot(aes(x = PrecipitationRange, y = AccidentCount, fill = Severity)) +
  labs(title = "Precipitation, Severity, and Accidents",
       x = "Precipitation Range (in inches)",
       y = "Number of Accidents") +
  geom_col() +
  scale_y_continuous(labels = scales::label_number_si()) +
  theme(axis.text.x = element_text(angle = 90))
```

```
## `summarise()` has grouped output by 'PrecipitationRange'. You can override
## using the `.groups` argument.
```

```
# plot 2: For Pressure
plot2 <- df %>%
  mutate(PressureRange = cut(Pressure, breaks = seq(0, 40, 2), right = TRUE),
         Severity = case_when(
           `Severity` == 1 ~ "Low",
           `Severity` == 2 ~ "Moderate",
           `Severity` == 3 ~ "High",
           `Severity` == 4 ~ "Very High",
           TRUE ~ "Unknown"
         )) %>%
  drop_na() %>%
  group_by(PressureRange, Severity) %>%
  summarise(AccidentCount = n()) %>%
  ggplot(aes(x = PressureRange, y = AccidentCount, fill = Severity)) +
  labs(title = "Pressure, Severity, and Accidents",
       x = "Pressure Range (in hPa)",
       y = "Number of Accidents") +
  geom_col() +
  scale_y_continuous(labels = scales::label_number_si()) +
  theme(axis.text.x = element_text(angle = 90))
```

```
## `summarise()` has grouped output by 'PressureRange'. You can override using the
## `.groups` argument.
```

```
# plot 3: For Windspeed

plot3 <- df_windspeed <- df %>%
  mutate(WindSpeedRange = cut(Wind_Speed, breaks = seq(0, 50, 2), right = FALSE),
   Severity = case_when(
           `Severity` == 1 ~ "Low",
           `Severity` == 2 ~ "Moderate",
           `Severity` == 3 ~ "High",
           `Severity` == 4 ~ "Very High",
           TRUE ~ "Unknown"
         )) %>%
  drop_na() %>%
  group_by(WindSpeedRange,Severity) %>%
  summarise(AccidentCount = n()) %>%
  ggplot(data = ., aes(x = WindSpeedRange, y = AccidentCount, fill=Severity)) +
  labs(title = "Wind Speed, Severity and Accidents",
       x = "Wind Speed Range (mph)",
```

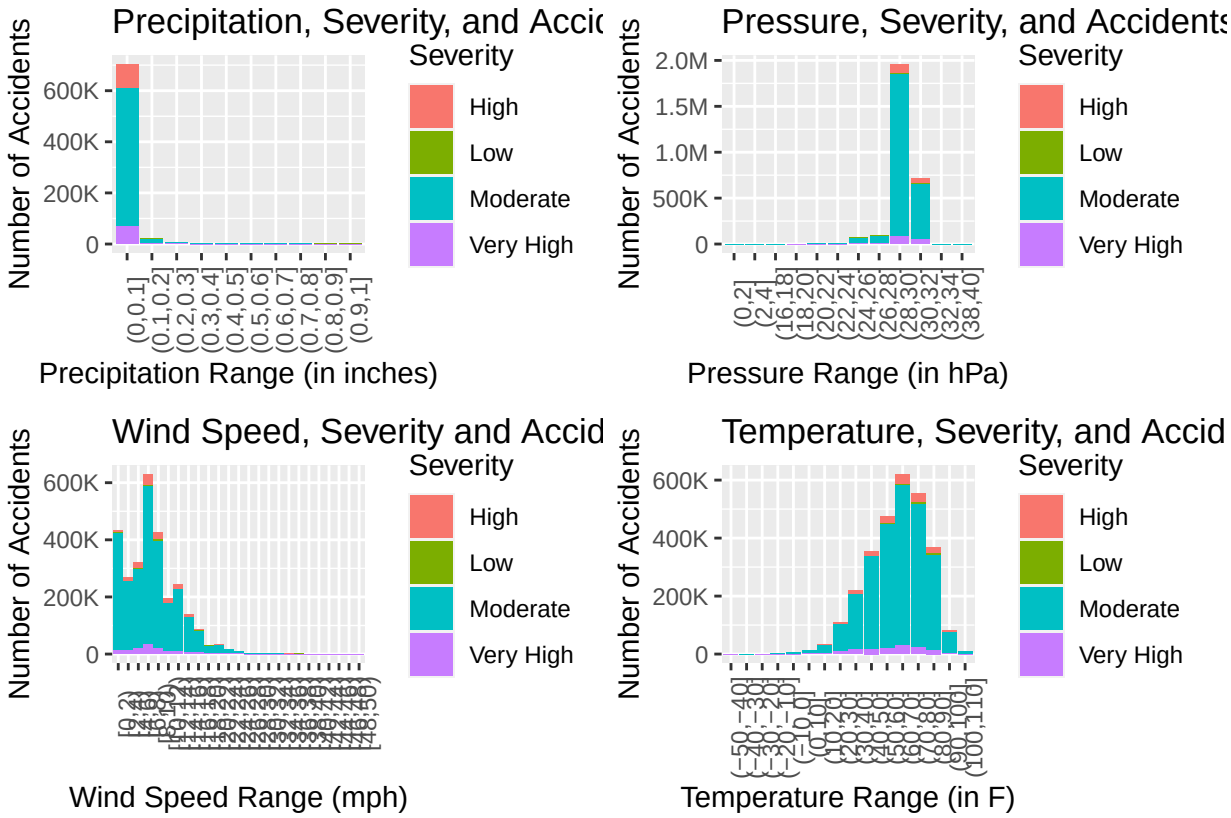
```
       y = "Number of Accidents") +
  geom_col() +
  scale_y_continuous(labels = scales::label_number_si())+
  theme(axis.text.x = element_text(angle = 90))
```

```
## `summarise()` has grouped output by 'WindSpeedRange'. You can override using
## the `.groups` argument.
```

```
# plot 4: For Temperature
plot4 <- df %>%
  mutate(TempRange = cut(Temperature.F., breaks = seq(-50, 110, 10),
                         right = TRUE),
         Severity = case_when(
           `Severity` == 1 ~ "Low",
           `Severity` == 2 ~ "Moderate",
           `Severity` == 3 ~ "High",
           `Severity` == 4 ~ "Very High",
           TRUE ~ "Unknown"
         )) %>%
  drop_na() %>%
  group_by(TempRange, Severity) %>%
  summarise(AccidentCount = n()) %>%
  ggplot(aes(x = TempRange, y = AccidentCount, fill = Severity)) +
  labs(title = "Temperature, Severity, and Accidents",
       x = "Temperature Range (in F)",
       y = "Number of Accidents") +
  geom_col() +
  scale_y_continuous(labels = scales::label_number_si()) +
  theme(axis.text.x = element_text(angle = 90))
```

```
## `summarise()` has grouped output by 'TempRange'. You can override using the
## `.groups` argument.
```

```
#combine plots
plot1 + plot2 + plot3 + plot4
```



We can observe there are more accidents when the precipitation is low, when the pressure is around 28-30 hPa, when the wind speed is around 4-8 mph and when the temperature is between 40 to 90 F. These are mostly normal conditions hence it shows that accidents occure in normal conditions.

```
library(patchwork)

# Plot 5
plot5 <- df %>%
  mutate(WindchillRange = cut(Wind_Chill.F., breaks = seq(-90,200, 20),
                              right = FALSE),
         Severity = case_when(
           `Severity` == 1 ~ "Low",
           `Severity` == 2 ~ "Moderate",
           `Severity` == 3 ~ "High",
           `Severity` == 4 ~ "Very High",
           TRUE ~ "Unknown"
         )) %>%
  drop_na() %>%
  group_by(WindchillRange, Severity) %>%
  summarise(AccidentCount = n()) %>%
  ggplot(aes(x = WindchillRange, y = AccidentCount, fill = Severity)) +
  labs(title = "Relationship between Windchill,
       Severity, and Number of Accidents",
       x = "Windchill Range (in F)",
       y = "Number of Accidents") +
  geom_col(position = "stack") +
```

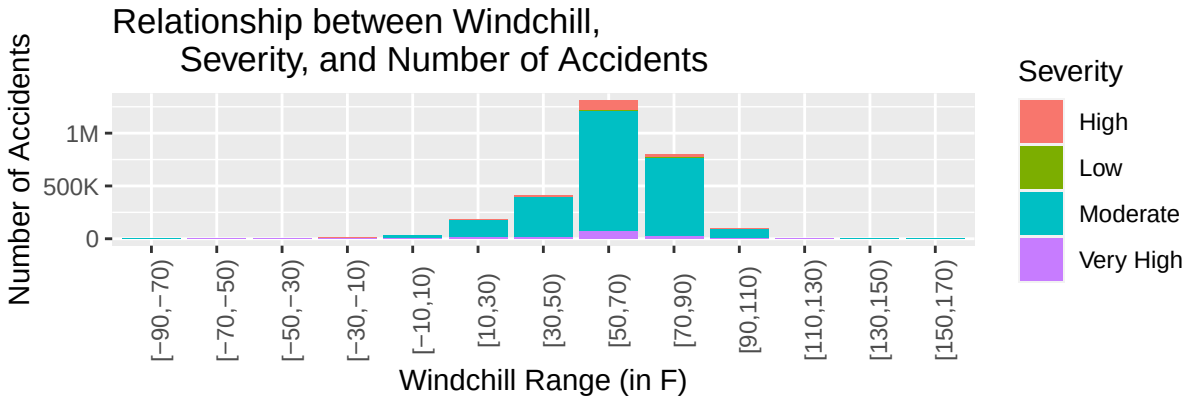
```
  scale_y_continuous(labels = scales::label_number_si())+
  theme(axis.text.x = element_text(angle = 90))
```

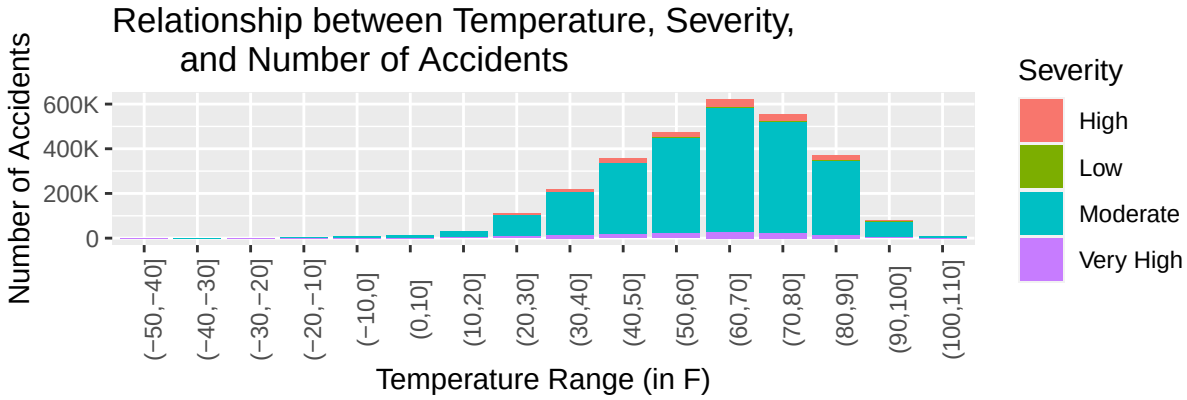
```
## `summarise()` has grouped output by 'WindchillRange'. You can override using
## the `.groups` argument.
```

```
plot6 <- df %>%
  mutate(TempRange = cut(Temperature.F., breaks = seq(-50, 110, 10),
                         right = TRUE),
         Severity = case_when(
           `Severity` == 1 ~ "Low",
           `Severity` == 2 ~ "Moderate",
           `Severity` == 3 ~ "High",
           `Severity` == 4 ~ "Very High",
           TRUE ~ "Unknown"
         )) %>%
  drop_na() %>%
  group_by(TempRange, Severity) %>%
  summarise(AccidentCount = n()) %>%
  ggplot(aes(x = TempRange, y = AccidentCount, fill = Severity)) +
  labs(title = "Relationship between Temperature, Severity,
       and Number of Accidents",
       x = "Temperature Range (in F)",
       y = "Number of Accidents") +
  geom_col() +
  scale_y_continuous(labels = scales::label_number_si()) +
  theme(axis.text.x = element_text(angle = 90))
```

```
## `summarise()` has grouped output by 'TempRange'. You can override using the
## `.groups` argument.
```

```
# combine plots
plot5 + plot6 + plot_layout(ncol = 1, nrow=2)
```





Time based analysis

```
library(lubridate)
unique(year(df$Start_Time))
```

```
## [1] 2016 2017 2021 2020 2018 2019
```

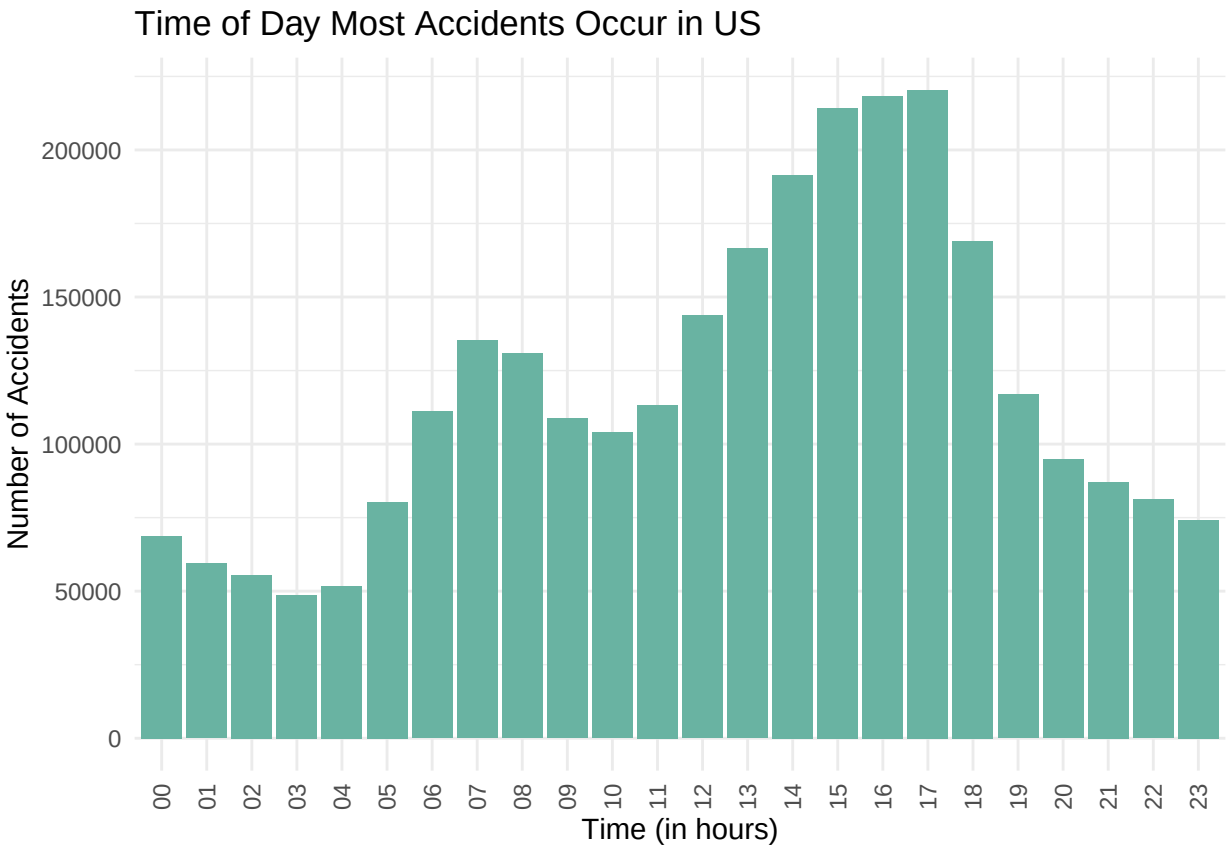
```
sum(is.na(df$Start_Time))
```

```
## [1] 0
```

Accidents based on time of the day

```
# What time of day most of the accidents occur?

library(ggplot2)
library(dplyr)

# Converting the Start_Time column to a datetime format
df$Start_Time <- as.POSIXct(df$Start_Time, format="%Y-%m-%d %H:%M:%S", tz="GMT")
```

```r
# Counting the number of accidents per hour of the day
hour_count <- df %>%
              mutate(hour = format(as.POSIXct(Start_Time), format = "%H")) %>%
              count(hour = hour) %>%
              arrange(hour)

ggplot(hour_count, aes(x=hour, y=n)) +
  geom_bar(stat="identity", fill="#69b3a2") +
  labs(title="Time of Day Most Accidents Occur in US", x="Time (in hours)",
       y="Number of Accidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

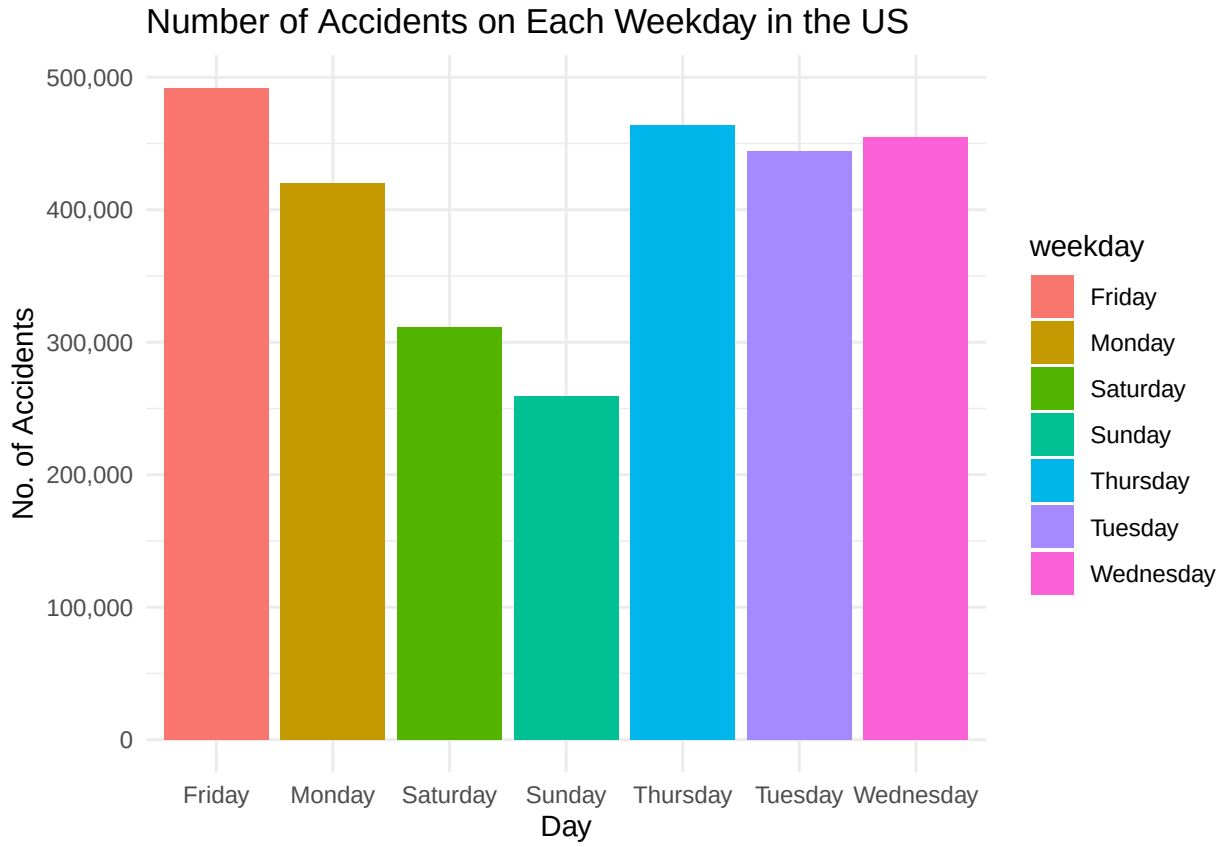


It can be observed that most of the accidents have occured during peak working hours such as 6 AM to 8 AM and more during 3 PM to 5 PM. This can be because of the traffic accumulated on the road.

Accidents based on day of the week

```r
# In which day of the week most of the accident occurs?

library(ggplot2)
library(dplyr)

# Counting the number of accidents per weekday
```

```
weekday_count <- df %>%
                 mutate(weekday = weekdays(as.Date(Start_Time))) %>%
                 count(weekday, sort = TRUE)

ggplot(weekday_count, aes(x = weekday, y = n, fill=weekday)) +
  geom_bar(stat = "identity") +
  labs(title = "Number of Accidents on Each Weekday in the US",
       x = "Day", y = "No. of Accidents") +
  scale_y_continuous(labels = scales::comma) +
 theme_minimal()
```

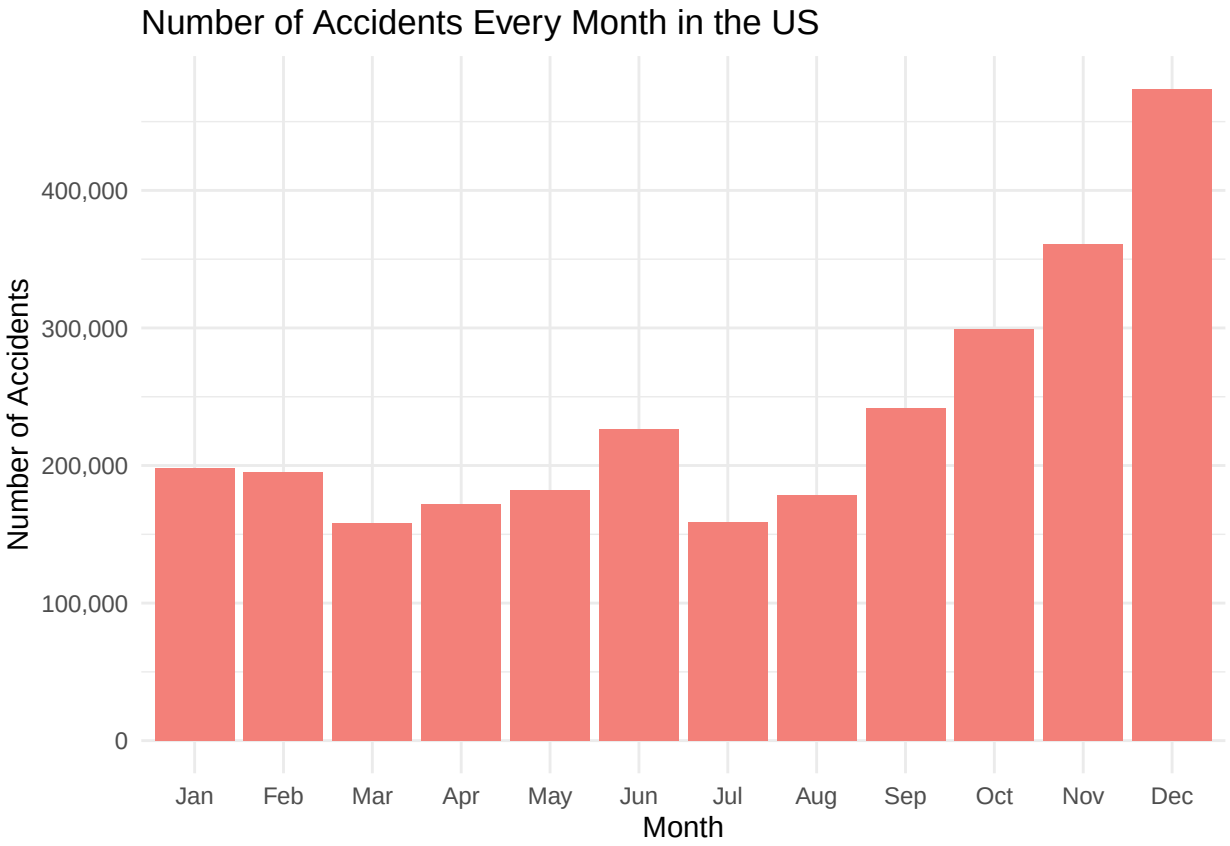


It is evident that most of the accidents occur during the Friday and other week days compared to weekends. This can be because of the lesser traffic during the weekends.

Accidents based on month of the year

```
# In which month most of the accident occurs?

library(ggplot2)
library(dplyr)
library(lubridate)

# Counting the number of accidents per month
month_count <- df %>%
```

```
                mutate(month = month(Start_Time, label = TRUE)) %>%
                count(month, sort = TRUE)

ggplot(month_count, aes(x = month, y = n)) +
  geom_bar(stat = "identity", fill = "#F38079") +
  labs(title = "Number of Accidents Every Month in the US",
       x = "Month", y = "Number of Accidents") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```

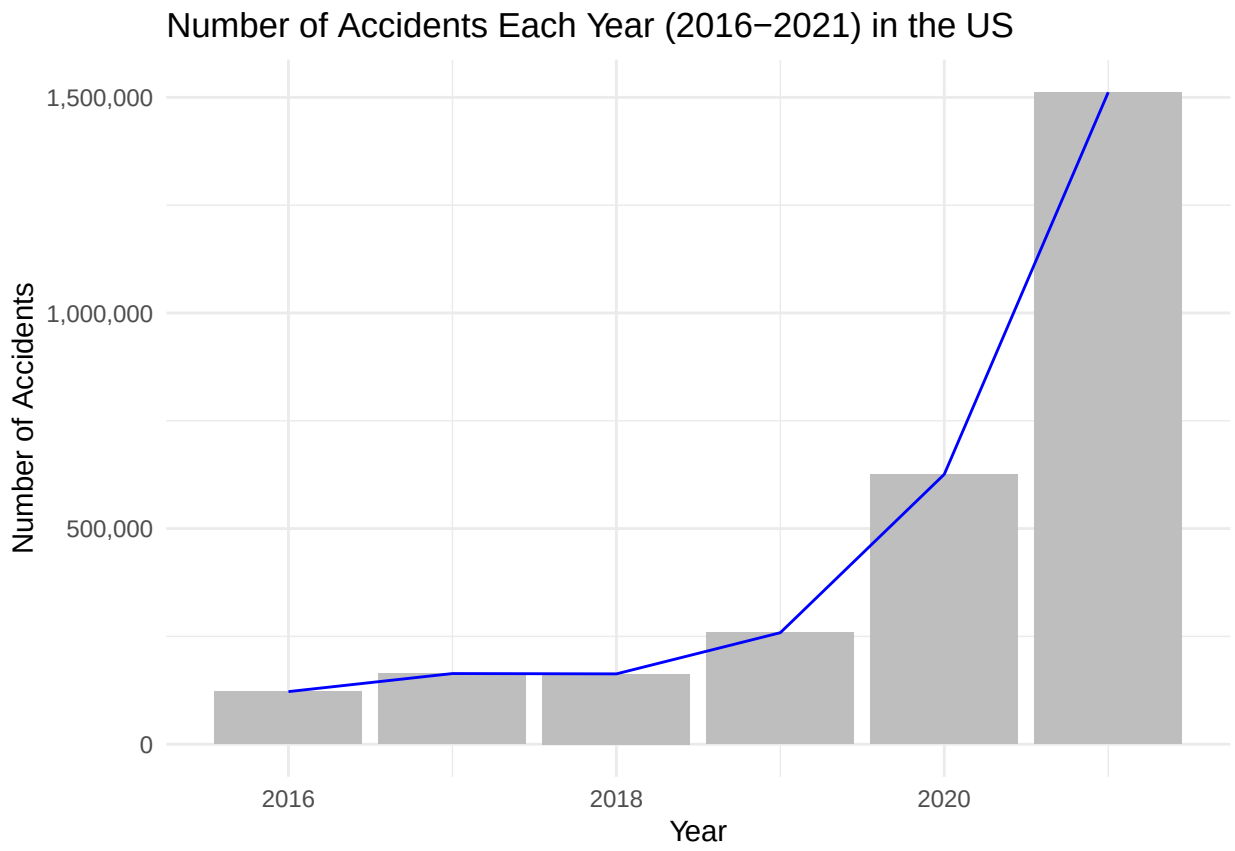


It is clearly seen that as the year comes to an end, there is an increase in the number of accidents. This can be because of the holiday season as many tourists visit many crowded places that can cause accidents.

Accidents based on year

```
# In which year most of the accident occurs?

library(ggplot2)
library(dplyr)
library(lubridate)

# Counting the number of accidents per year
year_count <- df %>%
              mutate(year = year(Start_Time)) %>%
```

```
              count(year, sort = TRUE)

ggplot(year_count, aes(x = year, y = n)) +
  geom_bar(stat = "identity", fill = "gray") +
  geom_line(aes(x = year, y = n), color = "blue") +
  labs(title = "Number of Accidents Each Year (2016-2021) in the US",
       x = "Year", y = "Number of Accidents") +
   scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```



There is a significant increase in the number of accidents in 2021 compared to the past in 2016. This shows that the accident is a huge problem and continues to grow as the years progress.

Severity analysis

```
df$Start_Time <- as_datetime(df$Start_Time)
df$End_Time <- as_datetime(df$End_Time)
df$Accident_duration <- round(abs((df$Start_Time-df$End_Time)/60))
df$Year <- as.numeric(format(df$Start_Time,format="%Y"))
df$month <- as.numeric(format(df$Start_Time,format="%m"))
df$date <- as.numeric(format(df$Start_Time,format="%d"))
str(df)
```

```
## 'data.frame':    2845342 obs. of  50 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          : POSIXct, format: "2016-02-08 00:37:08" "2016-02-08 05:56:20" ...
##  $ End_Time            : POSIXct, format: "2016-02-08 06:37:08" "2016-02-08 11:56:20" ...
##  $ 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 - Acc
##  $ 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" "20
##  $ Temperature.F.      : num  42.1 36.9 36 39 37 35.6 33.8 33.1 39 32 ...
##  $ Wind_Chill.F.       : num  36.1 59.7 59.7 59.7 29.8 ...
##  $ Humidity            : num  58 91 97 55 93 100 100 92 70 100 ...
##  $ Pressure            : num  29.8 29.7 29.7 29.6 29.7 ...
##  $ Visibility          : num  10 10 10 10 10 10 3 0.5 10 0.5 ...
##  $ Wind_Direction      : chr  "SW" "CALM" "CALM" "CALM" ...
##  $ Wind_Speed          : num  10.4 7.4 7.4 7.4 10.4 ...
##  $ Precipitation       : num  0 0.02 0.02 0.00702 0.01 ...
##  $ Weather_Condition   : chr  "Rain" "Rain" "Cloud" "Cloud" ...
##  $ 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      : int  0 0 0 0 1 1 1 1 1 1 ...
##  $ Civil_Twilight      : int  0 0 0 0 1 1 1 1 1 1 ...
##  $ Nautical_Twilight   : int  0 0 0 1 1 1 1 1 1 1 ...
##  $ Astronomical_Twilight: int  0 0 1 1 1 1 1 1 1 1 ...
##  $ Accident_duration   : 'difftime' num  6 6 6 6 ...
##   ..- attr(*, "units")= chr "mins"
##  $ Year                : num  2016 2016 2016 2016 2016 ...
##  $ month               : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ date                : num  8 8 8 8 8 8 8 8 8 8 ...
```

```
as.tibble(df)
```

```
## # A tibble: 2,845,342 x 50
##    ID    Severity Start_Time          End_Time            Start_Lat Start_Lng
##    <chr>    <int> <dttm>              <dttm>                  <dbl>     <dbl>
##  1 A-1          3 2016-02-08 00:37:08 2016-02-08 06:37:08      40.1     -83.1
##  2 A-2          2 2016-02-08 05:56:20 2016-02-08 11:56:20      39.9     -84.1
##  3 A-3          2 2016-02-08 06:15:39 2016-02-08 12:15:39      39.1     -84.5
##  4 A-4          2 2016-02-08 06:51:45 2016-02-08 12:51:45      41.1     -81.5
##  5 A-5          3 2016-02-08 07:53:43 2016-02-08 13:53:43      39.2     -84.5
##  6 A-6          2 2016-02-08 08:16:57 2016-02-08 14:16:57      39.1     -84.0
##  7 A-7          2 2016-02-08 08:15:41 2016-02-08 14:15:41      39.8     -84.2
##  8 A-8          2 2016-02-08 11:51:46 2016-02-08 17:51:46      41.4     -81.8
##  9 A-9          2 2016-02-08 14:19:57 2016-02-08 20:19:57      40.7     -84.1
## 10 A-10         2 2016-02-08 15:16:43 2016-02-08 21:16:43      40.1     -83.0
## # i 2,845,332 more rows
## # i 44 more variables: End_Lat <dbl>, End_Lng <dbl>, Distance.mi. <dbl>,
## #   Description <chr>, Street <chr>, Side <chr>, City <chr>, County <chr>,
## #   State <chr>, Zipcode <chr>, Country <chr>, Timezone <chr>,
## #   Airport_Code <chr>, Weather_Timestamp <chr>, Temperature.F. <dbl>,
## #   Wind_Chill.F. <dbl>, Humidity <dbl>, Pressure <dbl>, Visibility <dbl>,
## #   Wind_Direction <chr>, Wind_Speed <dbl>, Precipitation <dbl>, ...
```

```
library(tidyr)
df %>%
  group_by(Year, Severity) %>%
  summarise(count = n()) %>%
  pivot_wider(names_from = Severity, values_from = count)
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

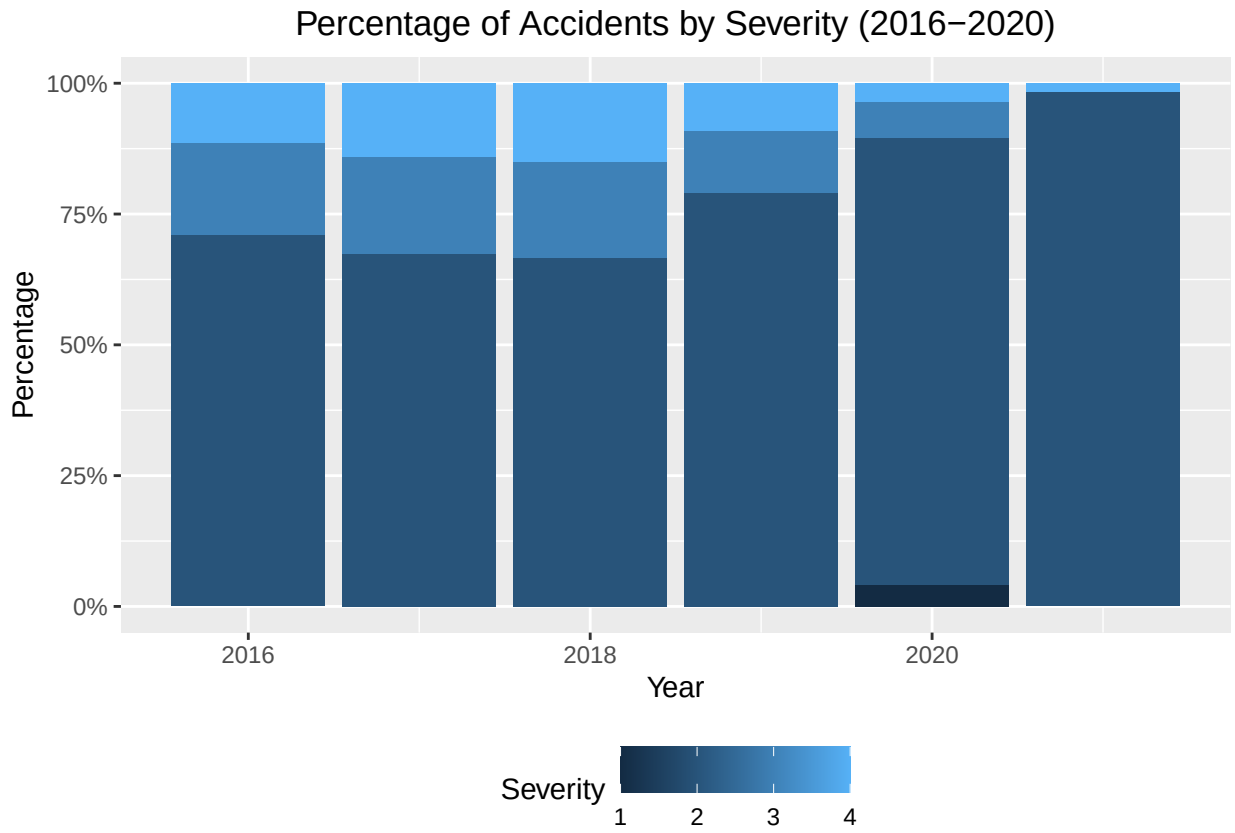
```
## # A tibble: 6 x 5
## # Groups:   Year [6]
##    Year     `2`   `3`   `4`   `1`
##   <dbl>   <int> <int> <int> <int>
## 1  2016   86758 21468 13798    NA
## 2  2017  110365 30389 23164    NA
## 3  2018  108568 30173 24435    NA
## 4  2019  204759 30269 23587    NA
## 5  2020  534828 42806 22177 26053
## 6  2021 1487713    NA 24032    NA
```

Percentage of accidents by severity

```
# create a stacked bar plot to show severity through the years
df %>%
  group_by(Year, Severity) %>%
  summarise(n = n()) %>%
  group_by(Year) %>%
  mutate(pct = n / sum(n)) %>%
  ggplot(aes(x = Year, y = pct, fill = Severity)) +
  geom_bar(stat = "identity") +
```

```
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Percentage of Accidents by Severity (2016-2020)",
       x = "Year", y = "Percentage") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = "bottom")
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

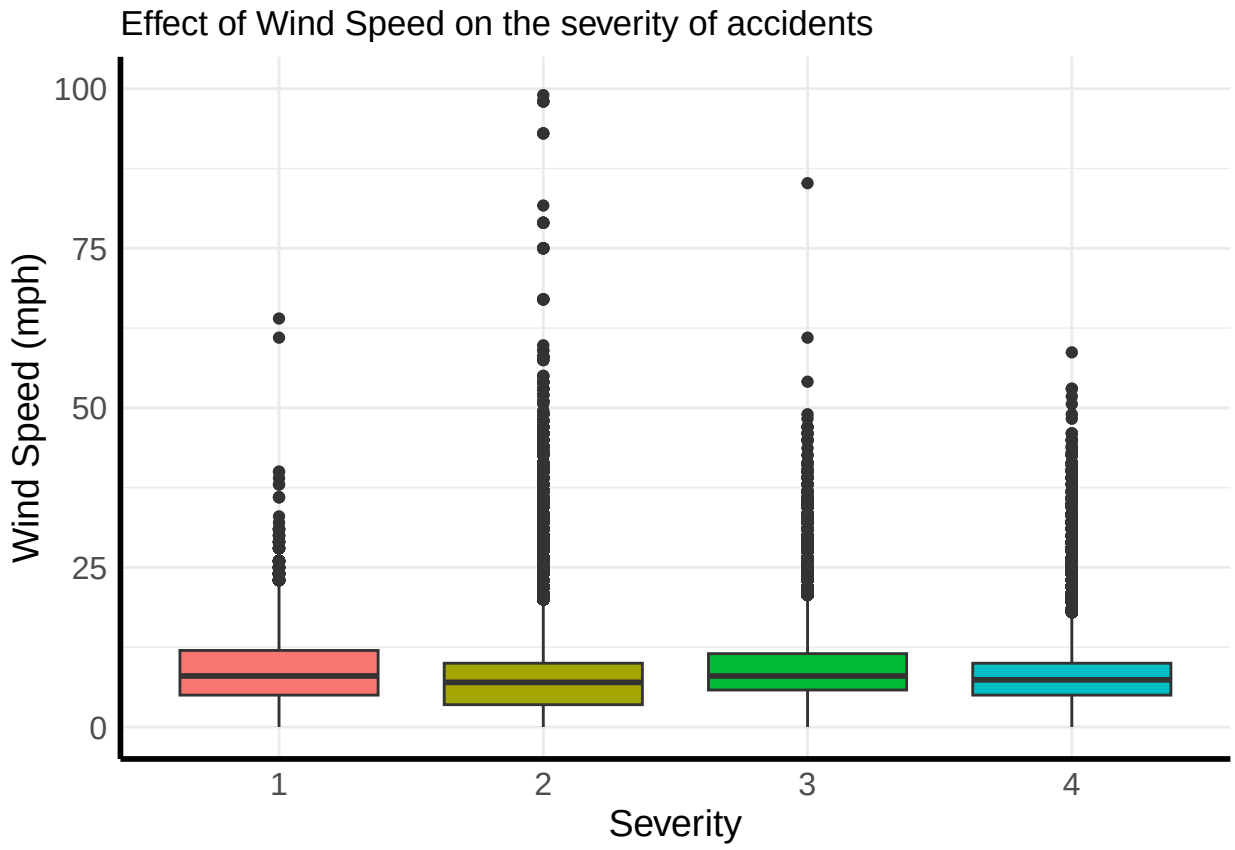


There are more number of low to moderate severity accidents in 2021 where as there were more sever accidents in the past.

Effect of wind speed on the severity

```
#Box Plot showing the effect of wind speed on severity of accidents
df$Severity <- as.factor(df$Severity)

ggplot(df, aes(x=Severity, y=Wind_Speed, fill=Severity)) +
  geom_boxplot() +
  scale_fill_manual(values=c("#F8766D", "#A3A500", "#00BA38", "#00BFC4")) +
  ylim(0, 100) +
  labs(x = "Severity", y = "Wind Speed (mph)") +
  ggtitle('Effect of Wind Speed on the severity of accidents')+
  theme_minimal() +
```

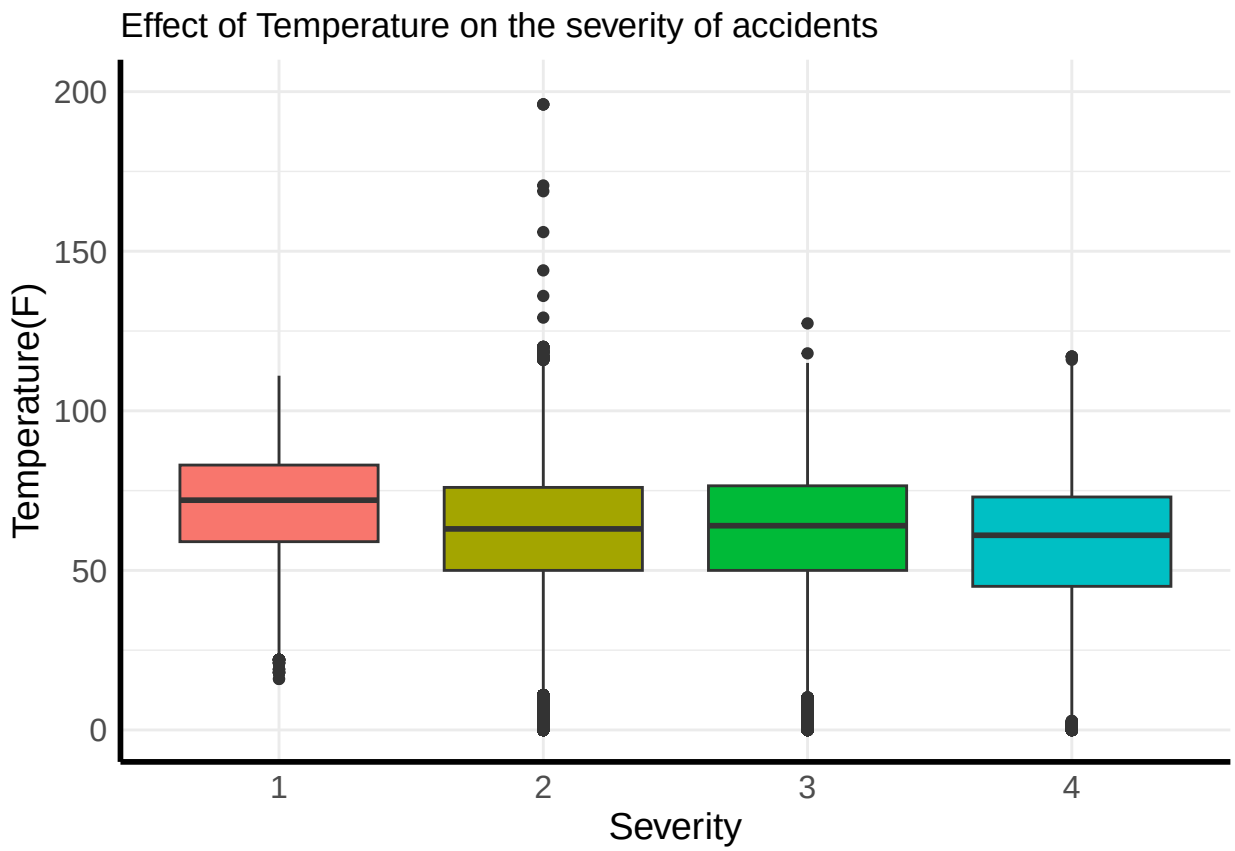
```
  theme(legend.position = "none", axis.line = element_line(size = 1),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 14))
```



There are more number of high (3) severity accidents as the wind speed increases. There are less pf a moderate severity accidents. There were some outliers found in this dataframe.

Effect of temperature on Severity

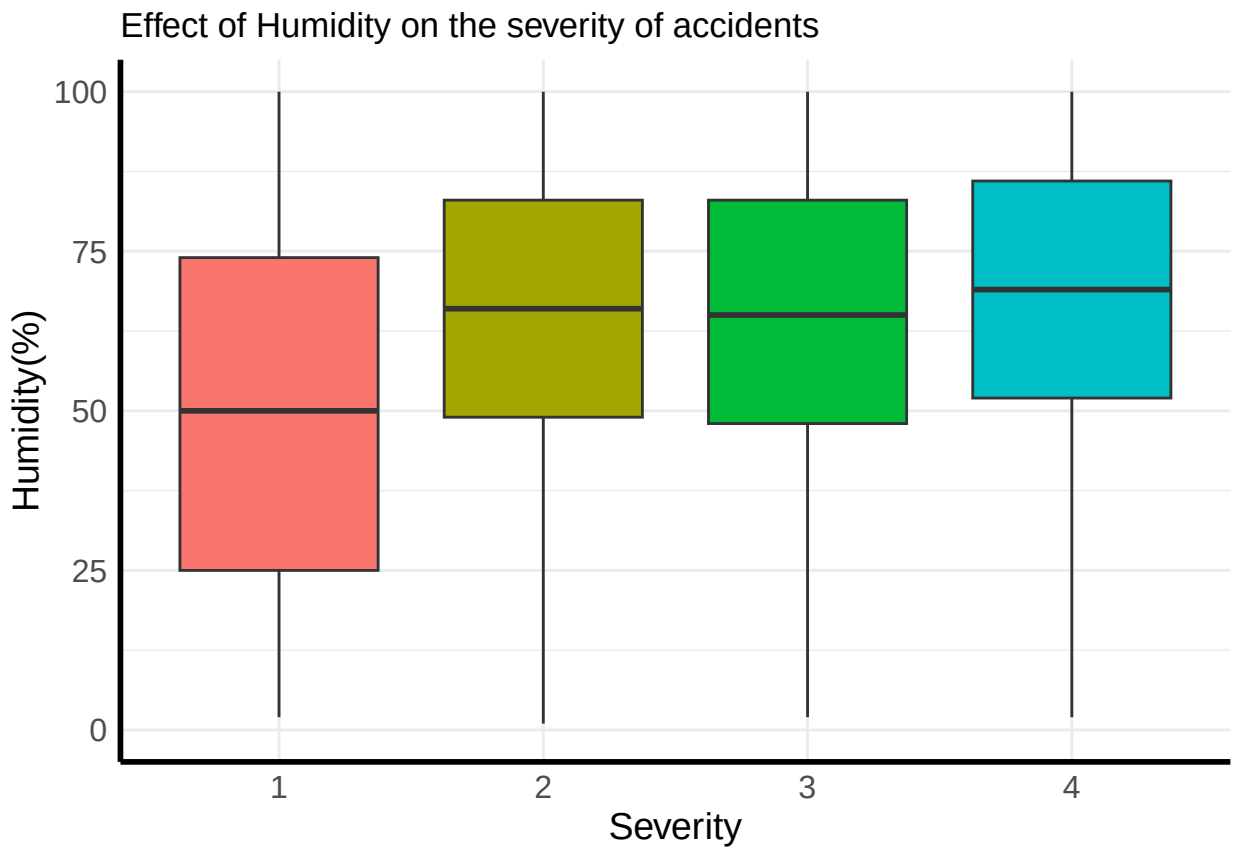
```
#Box Plot showing the effect of temperature on severity of accidents
df$Severity <- as.factor(df$Severity)

ggplot(df, aes(x=Severity, y=Temperature.F., fill=Severity)) +
  geom_boxplot() +
  scale_fill_manual(values=c("#F8766D", "#A3A500", "#00BA38", "#00BFC4")) +
  ylim(0, 200) +
  labs(x = "Severity", y = "Temperature(F)") +
  ggtitle('Effect of Temperature on the severity of accidents')+
  theme_minimal() +
  theme(legend.position = "none", axis.line = element_line(size = 1),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 14))
```



We can see that there are almost no difference in median temperature in Severity 2 and 3, while lower medium temperature in severity 4, which might indicate that lower temperature might result to more severe accidents. Whereas median temperature is slightly high for severity 1.

Effect of humidity on severity

```
#Humidity
ggplot(df, aes(x=Severity, y=Humidity, fill=Severity)) +
  geom_boxplot() +
  scale_fill_manual(values=c("#F8766D", "#A3A500", "#00BA38", "#00BFC4")) +
  ylim(0, 100) +
  labs(x = "Severity", y = "Humidity(%)") +
  ggtitle('Effect of Humidity on the severity of accidents')+
  theme_minimal() +
  theme(legend.position = "none", axis.line = element_line(size = 1),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 14))
```

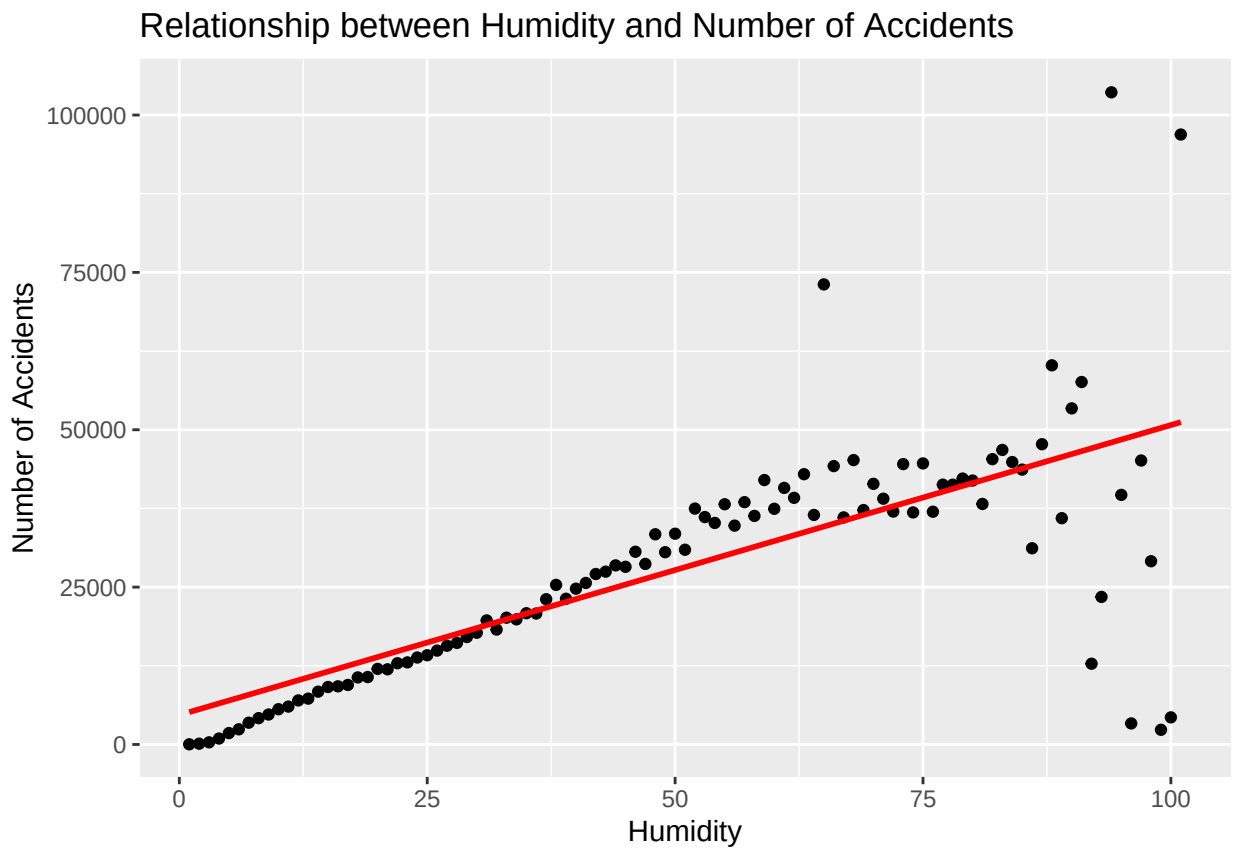


We can see that higher humidity might lead to more severe accidents.

Effect of humidity on accidents

```
library(ggplot2)

df %>%
  group_by(Humidity = as.factor(df$Humidity)) %>%
  summarise(AccidentCount = n()) %>%
  ggplot(aes(x = as.numeric(Humidity), y = AccidentCount)) +
  labs(title = "Relationship between Humidity and Number of Accidents",
       x = "Humidity",
       y = "Number of Accidents") +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE, color = "red")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



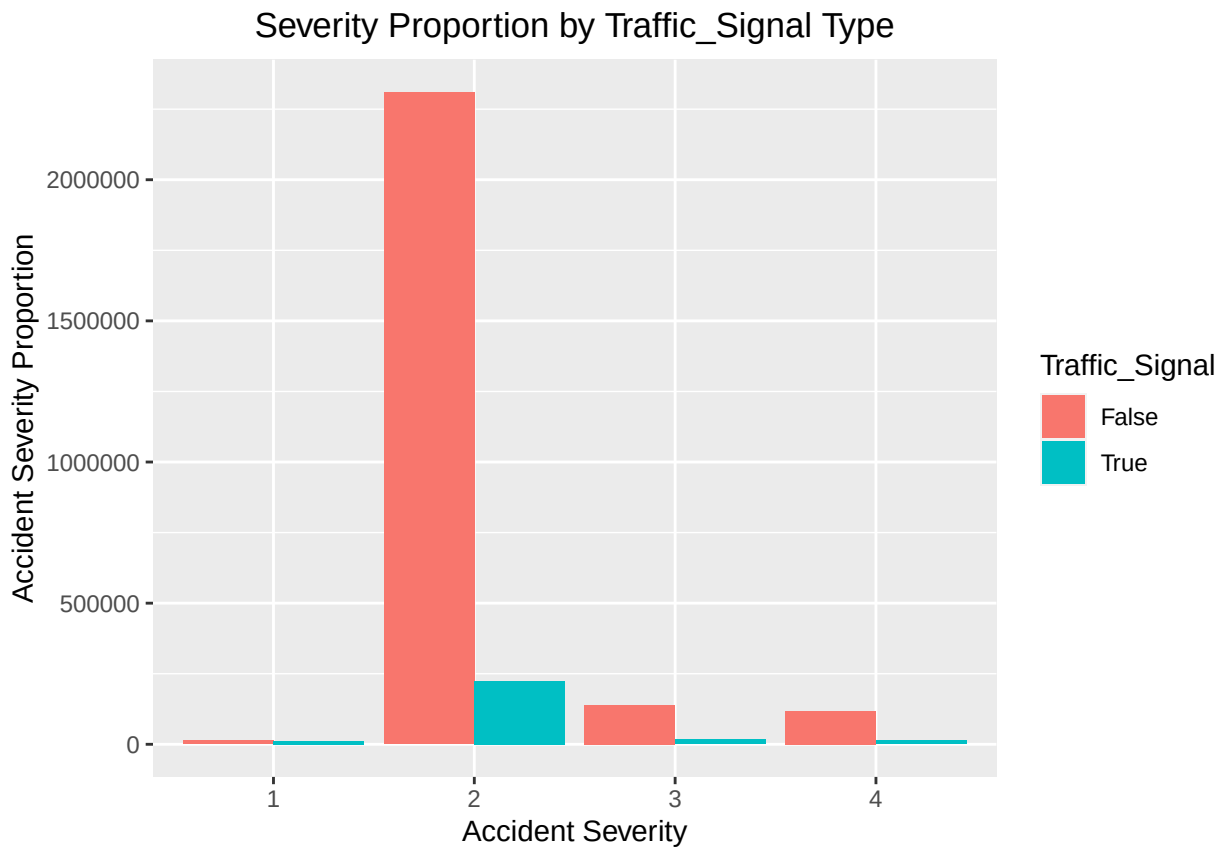
We can observe that the majority of accidents have occurred between 40 and 100 percent humidity. Also, it is clear that incidents of higher severity tend to occur more frequently at higher humidity levels. The plot also emphasizes the significance of taking humidity into account when assessing traffic accidents and creating safety measures.

Road conditions based analysis

Severity based on traffic signal

```
#Bar Plot showing Severity based on traffic signal
df %>%
  group_by(Traffic_Signal,Severity) %>%
    filter(Traffic_Signal!="Data missing or out of range") %>%
    summarize(total.count = n()) %>%
    ggplot(aes(x=Severity, y=total.count,fill=Traffic_Signal)) +
    geom_bar(stat="identity", position="dodge")+
    ggtitle("Severity Proportion by Traffic_Signal Type") +
    xlab("Accident Severity") + ylab("Accident Severity Proportion")+
    theme(plot.title = element_text(hjust = 0.5))
```

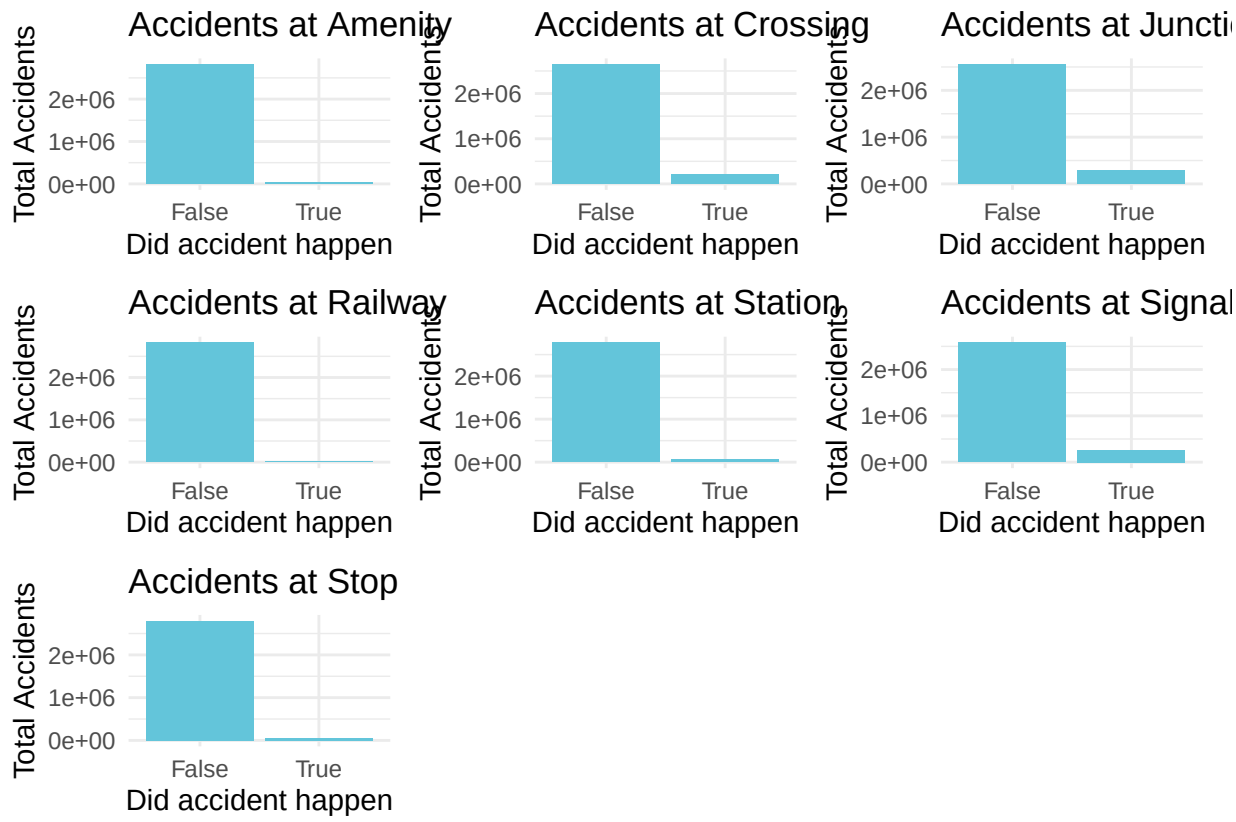
```
## `summarise()` has grouped output by 'Traffic_Signal'. You can override using
## the `.groups` argument.
```



Accidents based on differnet road conditions

```
plot1 <- df %>% group_by(Amenity) %>% count() %>%
         ggplot(aes(x=Amenity, y=n)) +
         geom_bar(stat = "identity", fill = "#63C5Da") + theme_minimal() +
         labs(title="Accidents at Amenity", x="Did accident happen", y="Total Accidents")

plot2 <- df %>% group_by(Crossing) %>% count() %>%
         ggplot(aes(x=Crossing, y=n)) +
         geom_bar(stat = "identity", fill = "#63C5Da") + theme_minimal() +
         labs(title="Accidents at Crossing", x="Did accident happen", y="Total Accidents")

plot3 <- df %>% group_by(Junction) %>% count() %>%
         ggplot(aes(x=Junction, y=n)) +
         geom_bar(stat = "identity", fill = "#63C5Da") + theme_minimal() +
         labs(title="Accidents at Junction", x="Did accident happen", y="Total Accidents")

plot4 <- df %>% group_by(Railway) %>% count() %>%
         ggplot(aes(x=Railway, y=n)) +
         geom_bar(stat = "identity", fill = "#63C5Da") + theme_minimal() +
         labs(title="Accidents at Railway", x="Did accident happen", y="Total Accidents")

plot5 <- df %>% group_by(Station) %>% count() %>%
         ggplot(aes(x=Station, y=n)) +
```

```
          geom_bar(stat = "identity", fill = "#63C5Da") + theme_minimal() +
          labs(title="Accidents at Station", x="Did accident happen", y="Total Accidents")

plot6 <- df %>% group_by(Traffic_Signal) %>% count() %>%
          ggplot(aes(x=Traffic_Signal, y=n)) +
          geom_bar(stat = "identity", fill = "#63C5Da") + theme_minimal() +
          labs(title="Accidents at Signal", x="Did accident happen", y="Total Accidents")

plot7 <- df %>% group_by(Stop) %>% count() %>%
          ggplot(aes(x=Stop, y=n)) +
          geom_bar(stat = "identity", fill = "#63C5Da") + theme_minimal() +
          labs(title="Accidents at Stop", x="Did accident happen", y="Total Accidents")


plot1 + plot2 + plot3 + plot4 + plot5 + plot6 + plot7
```



We can observe that there are more accidents for roads where there was crossing, junction or signal.

Duration of an accident

```
library(tidyverse)
library(lubridate)
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.1.2
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
accident_duration_df <- df %>%
  select(Accident_duration) %>%
  rowid_to_column("Id")

top_10_accident_duration_df <- accident_duration_df %>%
  count(Accident_duration) %>%
  top_n(10, wt = n) %>%
  sample_frac(1) %>%
  rename(Cases = n)

top_10_accident_duration_df
```

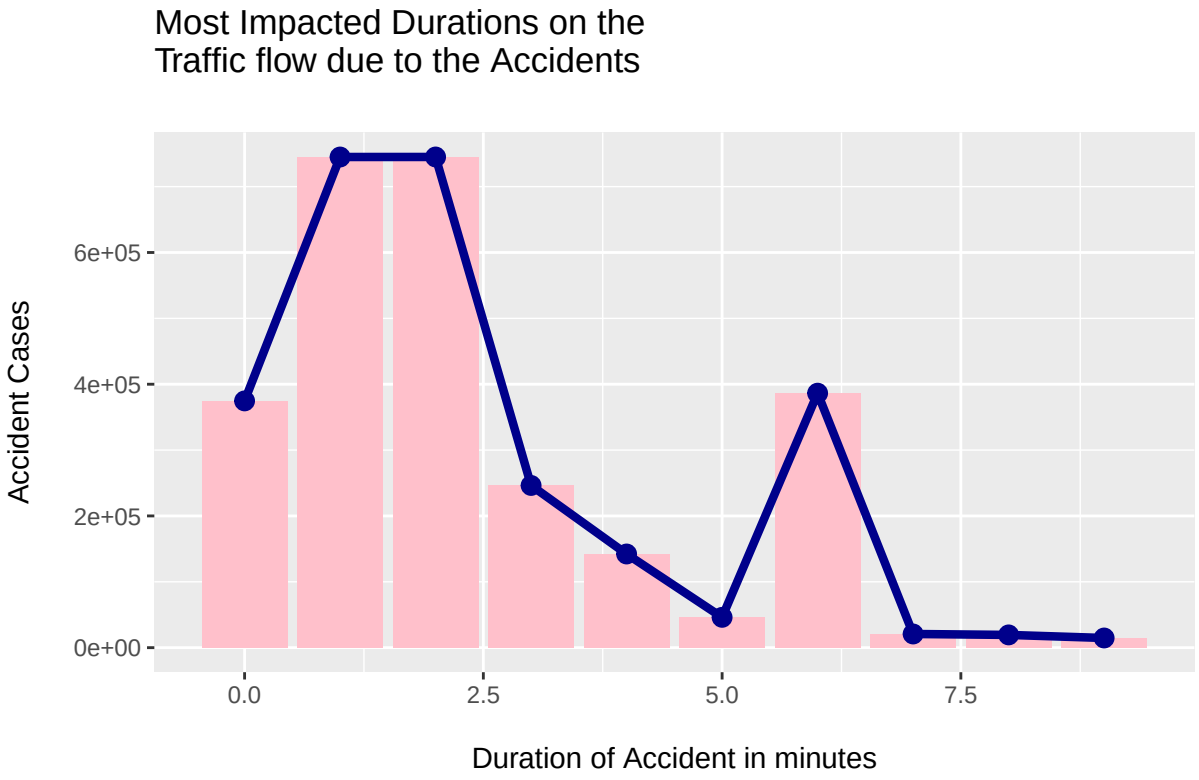
```
##    Accident_duration  Cases
## 1             9 mins  14678
## 2             8 mins  19369
## 3             6 mins 386389
## 4             5 mins  46034
## 5             4 mins 142224
## 6             7 mins  20756
## 7             0 mins 374676
## 8             3 mins 246271
## 9             1 mins 744981
## 10            2 mins 745028
```

```
#plot
fig1<- ggplot(top_10_accident_duration_df, aes(x = Accident_duration, y = Cases)) +
  geom_col(fill = "pink") +
  geom_line(aes(y = Cases), color = "darkblue", size = 1.5) +
  geom_point(aes(y = Cases), color = "darkblue", size = 3) +
  labs(
    x="\nDuration of Accident in minutes\n",
    y="\nAccident Cases\n",
    title="\nMost Impacted Durations on the \nTraffic flow due to the Accidents \n"
  )
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```
fig1
```

```
## Don't know how to automatically pick scale for object of type <difftime>.
## Defaulting to continuous.
```



Most accidents took 2 to 4 minutes to occur.

Data Modelling

Data Preparation

For this step, we decided to drop all the columns that were not relevant and had categorical type.

```
str(df)
```

```
## 'data.frame':    2845342 obs. of  50 variables:
##  $ ID                   : chr  "A-1" "A-2" "A-3" "A-4" ...
##  $ Severity             : Factor w/ 4 levels "1","2","3","4": 3 2 2 2 3 2 2 2 2 2 ...
##  $ Start_Time           : POSIXct, format: "2016-02-08 00:37:08" "2016-02-08 05:56:20" ...
##  $ End_Time             : POSIXct, format: "2016-02-08 06:37:08" "2016-02-08 11:56:20" ...
##  $ 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 - Acc
##  $ 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" "20
##  $ Temperature.F.       : num  42.1 36.9 36 39 37 35.6 33.8 33.1 39 32 ...
##  $ Wind_Chill.F.        : num  36.1 59.7 59.7 59.7 29.8 ...
##  $ Humidity             : num  58 91 97 55 93 100 100 92 70 100 ...
##  $ Pressure             : num  29.8 29.7 29.7 29.6 29.7 ...
##  $ Visibility           : num  10 10 10 10 10 10 3 0.5 10 0.5 ...
##  $ Wind_Direction       : chr  "SW" "CALM" "CALM" "CALM" ...
##  $ Wind_Speed           : num  10.4 7.4 7.4 7.4 10.4 ...
##  $ Precipitation        : num  0 0.02 0.02 0.00702 0.01 ...
##  $ Weather_Condition    : chr  "Rain" "Rain" "Cloud" "Cloud" ...
##  $ 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       : int  0 0 0 0 1 1 1 1 1 1 ...
##  $ Civil_Twilight       : int  0 0 0 0 1 1 1 1 1 1 ...
##  $ Nautical_Twilight    : int  0 0 0 1 1 1 1 1 1 1 ...
##  $ Astronomical_Twilight: int  0 0 1 1 1 1 1 1 1 1 ...
##  $ Accident_duration    : 'difftime' num  6 6 6 6 ...
##   ..- attr(*, "units")= chr "mins"
##  $ Year                 : num  2016 2016 2016 2016 2016 ...
##  $ month                : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ date                 : num  8 8 8 8 8 8 8 8 8 8 ...
```

```
modeling_data <- subset(df, select = -c(ID, Description, Street, Side, City, County, State, Zipcode, Co

as.tibble(modeling_data)
```

```
## # A tibble: 2,845,342 x 36
##    Severity Start_Lat Start_Lng End_Lat End_Lng Distance.mi. Temperature.F.
##    <fct>        <dbl>     <dbl>   <dbl>   <dbl>        <dbl>          <dbl>
##  1 3             40.1     -83.1    40.1   -83.0        3.23            42.1
##  2 2             39.9     -84.1    39.9   -84.0        0.747           36.9
```

```
##  3 2            39.1     -84.5    39.1   -84.5        0.055           36
##  4 2            41.1     -81.5    41.1   -81.5        0.123           39
##  5 3            39.2     -84.5    39.2   -84.5        0.5             37
##  6 2            39.1     -84.0    39.1   -84.1        1.43            35.6
##  7 2            39.8     -84.2    39.8   -84.2        0.227           33.8
##  8 2            41.4     -81.8    41.4   -81.8        0.521           33.1
##  9 2            40.7     -84.1    40.7   -84.1        0.491           39
## 10 2            40.1     -83.0    40.1   -83.0        0.826           32
## # i 2,845,332 more rows
## # i 29 more variables: Wind_Chill.F. <dbl>, Humidity <dbl>, Pressure <dbl>,
## #   Visibility <dbl>, Wind_Direction <chr>, Wind_Speed <dbl>,
## #   Precipitation <dbl>, Weather_Condition <chr>, Amenity <chr>, Bump <chr>,
## #   Crossing <chr>, Give_Way <chr>, Junction <chr>, No_Exit <chr>,
## #   Railway <chr>, Roundabout <chr>, Station <chr>, Stop <chr>,
## #   Traffic_Calming <chr>, Traffic_Signal <chr>, Turning_Loop <chr>, ...
```

We also chose some relevant columns with categorical datat and, replaced the values with encoding numerical values.

```
unique(modeling_data$Wind_Direction)
```

```
##  [1] "SW"   "CALM" "W"    "N"    "S"    "NW"   "E"    "SE"   "VAR"  "NE"
```

```
unique(modeling_data$Weather_Condition)
```

```
## [1] "Rain"        "Cloud"       "Snow"        "Clear"       "Fog"
## [6] "Heavy_Rain"  "Heavy_Snow"  "Dusty/Windy"
```

```
as.tibble(modeling_data)
```

```
## # A tibble: 2,845,342 x 36
##    Severity Start_Lat Start_Lng End_Lat End_Lng Distance.mi. Temperature.F.
##    <fct>        <dbl>     <dbl>   <dbl>   <dbl>        <dbl>          <dbl>
##  1 3             40.1     -83.1    40.1   -83.0        3.23            42.1
##  2 2             39.9     -84.1    39.9   -84.0        0.747           36.9
##  3 2             39.1     -84.5    39.1   -84.5        0.055           36
##  4 2             41.1     -81.5    41.1   -81.5        0.123           39
##  5 3             39.2     -84.5    39.2   -84.5        0.5             37
##  6 2             39.1     -84.0    39.1   -84.1        1.43            35.6
##  7 2             39.8     -84.2    39.8   -84.2        0.227           33.8
##  8 2             41.4     -81.8    41.4   -81.8        0.521           33.1
##  9 2             40.7     -84.1    40.7   -84.1        0.491           39
## 10 2             40.1     -83.0    40.1   -83.0        0.826           32
## # i 2,845,332 more rows
## # i 29 more variables: Wind_Chill.F. <dbl>, Humidity <dbl>, Pressure <dbl>,
## #   Visibility <dbl>, Wind_Direction <chr>, Wind_Speed <dbl>,
## #   Precipitation <dbl>, Weather_Condition <chr>, Amenity <chr>, Bump <chr>,
## #   Crossing <chr>, Give_Way <chr>, Junction <chr>, No_Exit <chr>,
## #   Railway <chr>, Roundabout <chr>, Station <chr>, Stop <chr>,
## #   Traffic_Calming <chr>, Traffic_Signal <chr>, Turning_Loop <chr>, ...
```

```
modeling_data[] <- data.matrix(modeling_data)
as.tibble(modeling_data)
```

```
## # A tibble: 2,845,342 x 36
##    Severity Start_Lat Start_Lng End_Lat End_Lng Distance.mi. Temperature.F.
##       <dbl>     <dbl>     <dbl>   <dbl>   <dbl>        <dbl>          <dbl>
##  1        3      40.1     -83.1    40.1   -83.0        3.23            42.1
##  2        2      39.9     -84.1    39.9   -84.0        0.747           36.9
##  3        2      39.1     -84.5    39.1   -84.5        0.055           36
##  4        2      41.1     -81.5    41.1   -81.5        0.123           39
##  5        3      39.2     -84.5    39.2   -84.5        0.5             37
##  6        2      39.1     -84.0    39.1   -84.1        1.43            35.6
##  7        2      39.8     -84.2    39.8   -84.2        0.227           33.8
##  8        2      41.4     -81.8    41.4   -81.8        0.521           33.1
##  9        2      40.7     -84.1    40.7   -84.1        0.491           39
## 10        2      40.1     -83.0    40.1   -83.0        0.826           32
## # i 2,845,332 more rows
## # i 29 more variables: Wind_Chill.F. <dbl>, Humidity <dbl>, Pressure <dbl>,
## #   Visibility <dbl>, Wind_Direction <dbl>, Wind_Speed <dbl>,
## #   Precipitation <dbl>, Weather_Condition <dbl>, Amenity <dbl>, Bump <dbl>,
## #   Crossing <dbl>, Give_Way <dbl>, Junction <dbl>, No_Exit <dbl>,
## #   Railway <dbl>, Roundabout <dbl>, Station <dbl>, Stop <dbl>,
## #   Traffic_Calming <dbl>, Traffic_Signal <dbl>, Turning_Loop <dbl>, ...
```

```
length(modeling_data$Severity[modeling_data$Severity == 1 |
                              modeling_data$Severity == 2])
```

```
## [1] 2559044
```

```
length(modeling_data$Severity[modeling_data$Severity == 3 |
                              modeling_data$Severity == 4])
```

```
## [1] 286298
```

Here we are dividing the target variable data from different values (1,2,3,4) to only 2 values (0,1) where 0 will be for values 1,2 and 1 is for values 3,4. This will help in creating a binary outcome after prediction.

```
modeling_data$Severity[modeling_data$Severity == 1 |
                       modeling_data$Severity == 2] <- 0
modeling_data$Severity[modeling_data$Severity == 3 |
                       modeling_data$Severity == 4] <- 1

length(modeling_data$Severity[modeling_data$Severity == 0])
```

```
## [1] 2559044
```

```
length(modeling_data$Severity[modeling_data$Severity == 1])
```

```
## [1] 286298
```

```
unique(modeling_data$Severity)
```

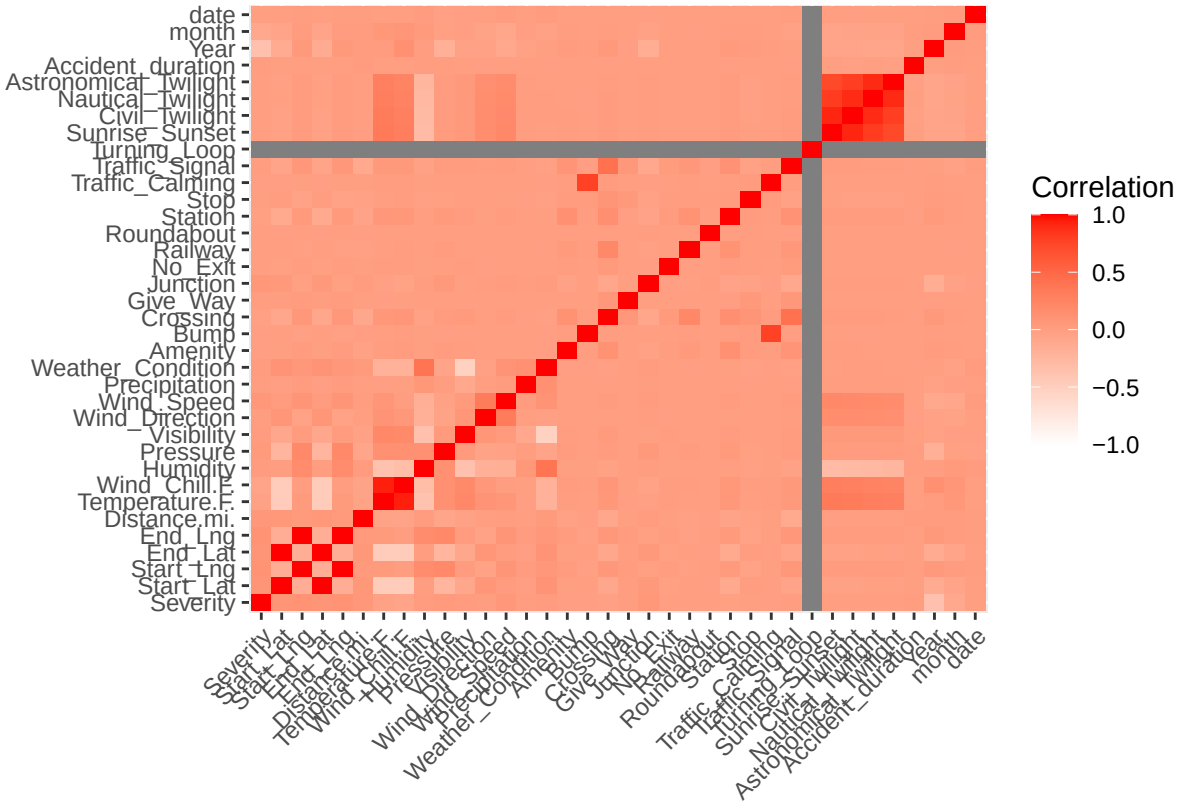
```
## [1] 1 0
```

Feature engineering correlation matrix

```
library(ggplot2)
library(reshape2)

# Compute correlation matrix
cor_matrix <- cor(modeling_data[, sapply(modeling_data, is.numeric)])
```

```
## Warning in cor(modeling_data[, sapply(modeling_data, is.numeric)]): the standard
## deviation is zero
```

```
# Plot heatmap of correlation matrix
ggplot(melt(cor_matrix), aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red",
                      limits = c(-1, 1), na.value = "grey50") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "", y = "", fill = "Correlation")
```



```r
# Identify variables strongly correlated with Severity
cor_matrix
```

```
##                          Severity     Start_Lat    Start_Lng       End_Lat
## Severity               1.0000000000  0.0885522338  0.107366932  0.0885540027
## Start_Lat              0.0885522338  1.0000000000 -0.154964772  0.9999952506
## Start_Lng              0.1073669317 -0.1549647724  1.000000000 -0.1549558953
## End_Lat                0.0885540027  0.9999952506 -0.154955895  1.0000000000
## End_Lng                0.1073674295 -0.1549619889  0.999999145 -0.1549534272
## Distance.mi.           0.0635687045  0.0715878023  0.039860621  0.0715927306
## Temperature.F.        -0.0272529937 -0.4712374828  0.031737365 -0.4712310011
## Wind_Chill.F.         -0.0645724296 -0.4670094908  0.009910927 -0.4670054252
## Humidity               0.0192429963  0.0058239287  0.168556691  0.0058170824
## Pressure               0.0369486834 -0.2334137456  0.206485748 -0.2334206111
## Visibility             0.0166258791 -0.0857783914  0.028944237 -0.0857726855
## Wind_Direction         0.0056928468  0.0786297168 -0.037878808  0.0786323791
## Wind_Speed             0.0628377670  0.0285140739  0.087849405  0.0285180586
## Precipitation          0.0136270355 -0.0026025438  0.021579291 -0.0026037645
## Weather_Condition      0.0211925636  0.1046200899  0.059066570  0.1046167833
## Amenity               -0.0039579173 -0.0058785689  0.014923442 -0.0058780082
## Bump                  -0.0038080356  0.0003383812 -0.014867478  0.0003389724
## Crossing              -0.0203697403 -0.0946983946  0.056168688 -0.0947033083
## Give_Way               0.0087839051  0.0078514142  0.018110132  0.0078402144
## Junction               0.0650959395  0.0436278611 -0.017201882  0.0436205220
## No_Exit                0.0002088153 -0.0173482971  0.006220354 -0.0173468761
## Railway                0.0018001705  0.0031840418 -0.015450124  0.0031843924
## Roundabout            -0.0009553339 -0.0033656484  0.000449983 -0.0033659454
## Station               -0.0168644594 -0.1076973796  0.038533445 -0.1076961257
## Stop                  -0.0079911271  0.0108173472 -0.040265344  0.0108274069
## Traffic_Calming       -0.0027280910 -0.0031638565 -0.007750069 -0.0031630869
## Traffic_Signal         0.0131405845 -0.0582082802  0.059567873 -0.0582097941
## Turning_Loop                     NA            NA           NA            NA
## Sunrise_Sunset         0.0039641118 -0.0349780546  0.029750373 -0.0349776218
## Civil_Twilight         0.0039898548 -0.0276071474  0.027997670 -0.0276068649
## Nautical_Twilight      0.0022040097 -0.0181150446  0.021312331 -0.0181163076
## Astronomical_Twilight  0.0032422055 -0.0154317745  0.018437064 -0.0154343673
## Accident_duration      0.0077606251 -0.0063721850  0.003268491 -0.0063817165
## Year                  -0.3547831032 -0.1295513123  0.052471277 -0.1295366738
## month                 -0.1020765713 -0.0619226104  0.027916176 -0.0619302834
## date                  -0.0057763297  0.0042246988  0.007993253  0.0042265384
##                           End_Lng  Distance.mi. Temperature.F. Wind_Chill.F.
## Severity               0.1073674295  6.356870e-02  -2.725299e-02  -0.064572430
## Start_Lat             -0.1549619889  7.158780e-02  -4.712375e-01  -0.467009491
## Start_Lng              0.9999991446  3.986062e-02   3.173736e-02   0.009910927
## End_Lat               -0.1549534272  7.159273e-02  -4.712310e-01  -0.467005425
## End_Lng                1.0000000000  3.983049e-02   3.174091e-02   0.009914016
## Distance.mi.           0.0398304884  1.000000e+00  -5.026948e-02  -0.054346391
## Temperature.F.         0.0317409091 -5.026948e-02   1.000000e+00   0.938147044
## Wind_Chill.F.          0.0099140156 -5.434639e-02   9.381470e-01   1.000000000
## Humidity               0.1685517411  2.634168e-02  -3.662865e-01  -0.319097581
## Pressure               0.2064879994 -6.804576e-02   1.371256e-01   0.129762673
## Visibility             0.0289460787 -3.334854e-02   2.085777e-01   0.194856976
## Wind_Direction        -0.0378714875 -2.693027e-04   1.049139e-01   0.057325821
```

```
## Wind_Speed             0.0878524763  1.069963e-02   7.690648e-02   0.012229670
## Precipitation          0.0215788413  2.663680e-03  -3.987457e-03  -0.006456495
## Weather_Condition      0.0590642813  3.369491e-02  -1.939362e-01  -0.198093762
## Amenity                0.0149239811 -3.271693e-02   1.327318e-02   0.015850707
## Bump                  -0.0148671953 -5.408457e-03   3.937769e-03   0.005196595
## Crossing               0.0561650111 -9.125602e-02   6.921800e-02   0.072909946
## Give_Way               0.0181103601 -6.654571e-03  -5.444802e-03  -0.006860170
## Junction              -0.0171988274  2.244226e-02  -2.010096e-02  -0.041234125
## No_Exit                0.0062200216 -1.018373e-02   1.153169e-02   0.011186544
## Railway               -0.0154494970 -2.146176e-02   3.056991e-03   0.004644219
## Roundabout             0.0004498836 -2.490619e-03   2.104201e-03   0.002342787
## Station                0.0385326988 -5.282616e-02   6.072970e-02   0.065850808
## Stop                  -0.0402642689 -2.710833e-02   2.110618e-05   0.005149196
## Traffic_Calming       -0.0077498812 -7.322018e-03   5.617476e-03   0.006806626
## Traffic_Signal         0.0595655774 -1.057223e-01   4.723965e-02   0.044651229
## Turning_Loop                     NA            NA             NA            NA
## Sunrise_Sunset         0.0297541055 -2.969153e-03   3.436516e-01   0.301026986
## Civil_Twilight         0.0280000343 -1.015904e-03   3.233456e-01   0.283980189
## Nautical_Twilight      0.0213133201 -2.873781e-04   3.007285e-01   0.265603980
## Astronomical_Twilight  0.0184366892 -1.913821e-05   2.819563e-01   0.250674694
## Accident_duration      0.0032733035  1.513935e-02  -1.079583e-03  -0.001467873
## Year                   0.0524749838  2.998376e-02   1.970025e-02   0.135103549
## month                  0.0279174636  2.039021e-02   6.552278e-02   0.073645995
## date                   0.0079958614  8.615652e-03   2.790835e-03   0.002982451
##                            Humidity      Pressure    Visibility Wind_Direction
## Severity               1.924300e-02  0.0369486834  0.0166258791   0.0056928468
## Start_Lat              5.823929e-03 -0.2334137456 -0.0857783914   0.0786297168
## Start_Lng              1.685567e-01  0.2064857482  0.0289442369  -0.0378788081
## End_Lat                5.817082e-03 -0.2334206111 -0.0857726855   0.0786323791
## End_Lng                1.685517e-01  0.2064879994  0.0289460787  -0.0378714875
## Distance.mi.           2.634168e-02 -0.0680457585 -0.0333485374  -0.0002693027
## Temperature.F.        -3.662865e-01  0.1371255559  0.2085776630   0.1049138652
## Wind_Chill.F.         -3.190976e-01  0.1297626728  0.1948569764   0.0573258212
## Humidity               1.000000e+00  0.1380955186 -0.3585900784  -0.1834417455
## Pressure               1.380955e-01  1.0000000000  0.0360939009  -0.0510692420
## Visibility            -3.585901e-01  0.0360939009  1.0000000000   0.0790533845
## Wind_Direction        -1.834417e-01 -0.0510692420  0.0790533845   1.0000000000
## Wind_Speed            -1.697380e-01 -0.0339017180  0.0350742109   0.3233799844
## Precipitation          6.988445e-02  0.0119024792 -0.1007421325   0.0018946939
## Weather_Condition      3.913514e-01 -0.0567066280 -0.5178110066  -0.0117772788
## Amenity               -5.892031e-03  0.0161147190  0.0087030264   0.0025210098
## Bump                  -7.666508e-03 -0.0041622445  0.0032800856   0.0019952553
## Crossing              -2.976436e-02  0.0157767881  0.0350668489   0.0015494595
## Give_Way               6.199890e-05 -0.0007181692  0.0024111725   0.0016897277
## Junction               6.309792e-03  0.0510790752 -0.0075949421   0.0121951480
## No_Exit               -7.045101e-03 -0.0007766759  0.0072797397  -0.0024739450
## Railway               -3.037745e-04  0.0154708357  0.0022984144  -0.0004581533
## Roundabout             8.967578e-04  0.0007392256  0.0001070491  -0.0004033724
## Station               -9.390788e-05  0.0405798489  0.0202295192  -0.0068567640
## Stop                  -1.542816e-02 -0.0156199835  0.0025368653   0.0029722562
## Traffic_Calming       -5.668323e-03  0.0002303824  0.0038501472   0.0010452428
## Traffic_Signal        -3.381006e-02  0.0152370577  0.0304955885   0.0090739772
## Turning_Loop                     NA            NA             NA            NA
## Sunrise_Sunset        -2.953784e-01  0.0234793013  0.0507693123   0.1636725647
```

```
## Civil_Twilight        -2.788259e-01  0.0228625751  0.0470318697   0.1594527267
## Nautical_Twilight     -2.626578e-01  0.0217887500  0.0457916956   0.1528143198
## Astronomical_Twilight -2.457548e-01  0.0208150375  0.0456842585   0.1445812487
## Accident_duration      6.937110e-03  0.0078558979  0.0020480603  -0.0042168817
## Year                   1.325932e-02 -0.1767900428 -0.0273576932  -0.0309398741
## month                  3.570166e-02 -0.0157871064 -0.0037318475  -0.0501804402
## date                   1.891925e-02 -0.0228321727 -0.0022866360   0.0025950460
##                         Wind_Speed Precipitation Weather_Condition
## Severity               0.0628377670  1.362704e-02       0.021192564
## Start_Lat              0.0285140739 -2.602544e-03       0.104620090
## Start_Lng              0.0878494053  2.157929e-02       0.059066570
## End_Lat                0.0285180586 -2.603764e-03       0.104616783
## End_Lng                0.0878524763  2.157884e-02       0.059064281
## Distance.mi.           0.0106996281  2.663680e-03       0.033694909
## Temperature.F.         0.0769064827 -3.987457e-03      -0.193936204
## Wind_Chill.F.          0.0122296696 -6.456495e-03      -0.198093762
## Humidity              -0.1697380329  6.988445e-02       0.391351381
## Pressure              -0.0339017180  1.190248e-02      -0.056706628
## Visibility             0.0350742109 -1.007421e-01      -0.517811007
## Wind_Direction         0.3233799844  1.894694e-03      -0.011777279
## Wind_Speed             1.0000000000  2.392275e-02       0.102265544
## Precipitation          0.0239227477  1.000000e+00       0.129163577
## Weather_Condition      0.1022655436  1.291636e-01       1.000000000
## Amenity                0.0008115082  1.621231e-03      -0.004836462
## Bump                  -0.0011433029 -9.492318e-04      -0.004407481
## Crossing               0.0183296829 -2.212723e-03      -0.019444768
## Give_Way               0.0023782638 -1.176467e-03      -0.001203065
## Junction               0.0177558810  1.232949e-02       0.014784328
## No_Exit                0.0020751753  3.179268e-04      -0.001754167
## Railway               -0.0003314099 -2.412228e-05      -0.003097058
## Roundabout             0.0002127995 -2.416063e-06      -0.000175681
## Station                0.0157163975 -1.376445e-03      -0.006302575
## Stop                  -0.0064028295 -4.208476e-03      -0.013735652
## Traffic_Calming       -0.0003983954 -1.153635e-03      -0.004171760
## Traffic_Signal         0.0164203501 -2.237374e-03      -0.023252245
## Turning_Loop                     NA            NA                NA
## Sunrise_Sunset         0.2089741854  7.753792e-03       0.002117330
## Civil_Twilight         0.1992069447  7.971903e-03       0.005391657
## Nautical_Twilight      0.1846567627  7.694115e-03       0.004795306
## Astronomical_Twilight  0.1703422997  7.557106e-03       0.003357631
## Accident_duration      0.0017241138  8.045969e-05      -0.002875700
## Year                  -0.0992337598 -3.148965e-02      -0.007816378
## month                 -0.0774924465 -4.514838e-03      -0.030714994
## date                   0.0174244681  3.320313e-03       0.020253697
##                            Amenity          Bump      Crossing      Give_Way
## Severity              -0.0039579173 -0.0038080356 -0.0203697403  0.0087839051
## Start_Lat             -0.0058785689  0.0003383812 -0.0946983946  0.0078514142
## Start_Lng              0.0149234419 -0.0148674779  0.0561686879  0.0181101320
## End_Lat               -0.0058780082  0.0003389724 -0.0947033083  0.0078402144
## End_Lng                0.0149239811 -0.0148671953  0.0561650111  0.0181103601
## Distance.mi.          -0.0327169259 -0.0054084571 -0.0912560211 -0.0066545711
## Temperature.F.         0.0132731782  0.0039377689  0.0692179967 -0.0054448025
## Wind_Chill.F.          0.0158507070  0.0051965948  0.0729099460 -0.0068601704
## Humidity              -0.0058920307 -0.0076665080 -0.0297643616  0.0000619989
```

```
## Pressure               0.0161147190 -0.0041622445  0.0157767881 -0.0007181692
## Visibility             0.0087030264  0.0032800856  0.0350668489  0.0024111725
## Wind_Direction         0.0025210098  0.0019952553  0.0015494595  0.0016897277
## Wind_Speed             0.0008115082 -0.0011433029  0.0183296829  0.0023782638
## Precipitation          0.0016212311 -0.0009492318 -0.0022127233 -0.0011764675
## Weather_Condition     -0.0048364616 -0.0044074806 -0.0194447682 -0.0012030647
## Amenity                1.0000000000  0.0054444399  0.1177566173  0.0032974613
## Bump                   0.0054444399  1.0000000000  0.0132888827 -0.0001756394
## Crossing               0.1177566173  0.0132888827  1.0000000000  0.0535067561
## Give_Way               0.0032974613 -0.0001756394  0.0535067561  1.0000000000
## Junction              -0.0264126926 -0.0017922088 -0.0802140738 -0.0070068257
## No_Exit                0.0130962442  0.0026096905  0.0420846695  0.0041763060
## Railway                0.0343715141  0.0064508101  0.2071321810  0.0039809794
## Roundabout             0.0004279049 -0.0001245714 -0.0007638848  0.0029446223
## Station                0.1243951514  0.0060265815  0.1446545322 -0.0021629812
## Stop                   0.0271373235  0.0194017857  0.0868780620  0.0476914840
## Traffic_Calming        0.0106240673  0.7721616543  0.0260873649  0.0002534085
## Traffic_Signal         0.0905549119 -0.0037135354  0.4222321100  0.0569943920
## Turning_Loop                    NA            NA            NA            NA
## Sunrise_Sunset         0.0058128553 -0.0018226855  0.0216622192  0.0001264558
## Civil_Twilight         0.0049118705 -0.0020769116  0.0189618350  0.0004091369
## Nautical_Twilight      0.0034832287 -0.0023500114  0.0156349210  0.0001017344
## Astronomical_Twilight  0.0025955373 -0.0018937930  0.0134174740 -0.0006371215
## Accident_duration     -0.0011524601 -0.0003451803 -0.0041382975 -0.0008663366
## Year                   0.0192305286  0.0071056632  0.0442481165 -0.0046199278
## month                  0.0060561999  0.0012849337  0.0065562658 -0.0013676756
## date                   0.0005390596  0.0005719152  0.0010773656  0.0010604991
##                           Junction       No_Exit       Railway    Roundabout
## Severity               0.065095939  2.088153e-04  1.800170e-03 -9.553339e-04
## Start_Lat              0.043627861 -1.734830e-02  3.184042e-03 -3.365648e-03
## Start_Lng             -0.017201882  6.220354e-03 -1.545012e-02  4.499830e-04
## End_Lat                0.043620522 -1.734688e-02  3.184392e-03 -3.365945e-03
## End_Lng               -0.017198827  6.220022e-03 -1.544950e-02  4.498836e-04
## Distance.mi.           0.022442257 -1.018373e-02 -2.146176e-02 -2.490619e-03
## Temperature.F.        -0.020100960  1.153169e-02  3.056991e-03  2.104201e-03
## Wind_Chill.F.         -0.041234125  1.118654e-02  4.644219e-03  2.342787e-03
## Humidity               0.006309792 -7.045101e-03 -3.037745e-04  8.967578e-04
## Pressure               0.051079075 -7.766759e-04  1.547084e-02  7.392256e-04
## Visibility            -0.007594942  7.279740e-03  2.298414e-03  1.070491e-04
## Wind_Direction         0.012195148 -2.473945e-03 -4.581533e-04 -4.033724e-04
## Wind_Speed             0.017755881  2.075175e-03 -3.314099e-04  2.127995e-04
## Precipitation          0.012329486  3.179268e-04 -2.412228e-05 -2.416063e-06
## Weather_Condition      0.014784328 -1.754167e-03 -3.097058e-03 -1.756810e-04
## Amenity               -0.026412693  1.309624e-02  3.437151e-02  4.279049e-04
## Bump                  -0.001792209  2.609691e-03  6.450810e-03 -1.245714e-04
## Crossing              -0.080214074  4.208467e-02  2.071322e-01 -7.638848e-04
## Give_Way              -0.007006826  4.176306e-03  3.980979e-03  2.944622e-03
## Junction               1.000000000 -3.869689e-03 -1.043554e-02  1.208336e-02
## No_Exit               -0.003869689  1.000000e+00  2.736351e-03 -2.556151e-04
## Railway               -0.010435538  2.736351e-03  1.000000e+00 -5.887262e-04
## Roundabout             0.012083359 -2.556151e-04 -5.887262e-04  1.000000e+00
## Station               -0.044454711  1.519868e-02  1.093530e-01  3.712383e-04
## Stop                  -0.034977237  1.180154e-02  7.892667e-03  6.411565e-03
## Traffic_Calming       -0.002262287  1.630574e-03  5.063596e-03  2.018567e-03
```

```
## Traffic_Signal        -0.096141901  2.337576e-02  5.372683e-02 -2.108223e-03
## Turning_Loop                  NA            NA            NA            NA
## Sunrise_Sunset         0.004556104  1.473488e-03 -9.828347e-04 -1.614062e-04
## Civil_Twilight         0.005869793  6.865042e-04 -8.840154e-04 -4.026062e-04
## Nautical_Twilight      0.007481557  5.318313e-05 -8.167130e-04 -7.582331e-04
## Astronomical_Twilight  0.007460569 -5.740548e-05 -9.989479e-04 -1.165798e-03
## Accident_duration      0.003327846 -6.386775e-04  1.672869e-04 -1.600494e-04
## Year                  -0.158444604  4.060719e-03 -1.100481e-04  2.135995e-03
## month                 -0.033211616  1.646773e-03 -3.510179e-03  6.659997e-05
## date                  -0.003192264 -9.008451e-04 -4.577596e-04  2.985394e-04
##                            Station          Stop Traffic_Calming
## Severity              -1.686446e-02 -7.991127e-03   -0.0027280910
## Start_Lat             -1.076974e-01  1.081735e-02   -0.0031638565
## Start_Lng              3.853344e-02 -4.026534e-02   -0.0077500694
## End_Lat               -1.076961e-01  1.082741e-02   -0.0031630869
## End_Lng                3.853270e-02 -4.026427e-02   -0.0077498812
## Distance.mi.          -5.282616e-02 -2.710833e-02   -0.0073220178
## Temperature.F.         6.072970e-02  2.110618e-05    0.0056174759
## Wind_Chill.F.          6.585081e-02  5.149196e-03    0.0068066265
## Humidity              -9.390788e-05 -1.542816e-02   -0.0056683233
## Pressure               4.057985e-02 -1.561998e-02    0.0002303824
## Visibility             2.022952e-02  2.536865e-03    0.0038501472
## Wind_Direction        -6.856764e-03  2.972256e-03    0.0010452428
## Wind_Speed             1.571640e-02 -6.402830e-03   -0.0003983954
## Precipitation         -1.376445e-03 -4.208476e-03   -0.0011536351
## Weather_Condition     -6.302575e-03 -1.373565e-02   -0.0041717600
## Amenity                1.243952e-01  2.713732e-02    0.0106240673
## Bump                   6.026581e-03  1.940179e-02    0.7721616543
## Crossing               1.446545e-01  8.687806e-02    0.0260873649
## Give_Way              -2.162981e-03  4.769148e-02    0.0002534085
## Junction              -4.445471e-02 -3.497724e-02   -0.0022622874
## No_Exit                1.519868e-02  1.180154e-02    0.0016305744
## Railway                1.093530e-01  7.892667e-03    0.0050635963
## Roundabout             3.712383e-04  6.411565e-03    0.0020185672
## Station                1.000000e+00  2.729486e-02    0.0097677582
## Stop                   2.729486e-02  1.000000e+00    0.0173492815
## Traffic_Calming        9.767758e-03  1.734928e-02    1.0000000000
## Traffic_Signal         1.125925e-01 -2.825807e-02    0.0086458740
## Turning_Loop                  NA            NA              NA
## Sunrise_Sunset         2.016977e-02 -9.706038e-03   -0.0012506079
## Civil_Twilight         1.849750e-02 -1.072516e-02   -0.0014869795
## Nautical_Twilight      1.632748e-02 -1.163037e-02   -0.0018091393
## Astronomical_Twilight  1.618485e-02 -1.200975e-02   -0.0015095032
## Accident_duration     -2.244536e-03 -1.173116e-03   -0.0002653585
## Year                   3.834613e-02  3.069242e-02    0.0068863331
## month                  1.373213e-02  5.923711e-03   -0.0003238649
## date                   5.355828e-04 -6.340243e-04    0.0003211162
##                       Traffic_Signal Turning_Loop Sunrise_Sunset Civil_Twilight
## Severity                  0.013140585           NA   0.0039641118   0.0039898548
## Start_Lat                -0.058208280           NA  -0.0349780546  -0.0276071474
## Start_Lng                 0.059567873           NA   0.0297503732   0.0279976705
## End_Lat                  -0.058209794           NA  -0.0349776218  -0.0276068649
## End_Lng                   0.059565577           NA   0.0297541055   0.0280000343
## Distance.mi.             -0.105722318           NA  -0.0029691529  -0.0010159035
```

```
## Temperature.F.           0.047239646           NA   0.3436516295   0.3233456174
## Wind_Chill.F.            0.044651229           NA   0.3010269858   0.2839801891
## Humidity                -0.033810058           NA  -0.2953784210  -0.2788259451
## Pressure                 0.015237058           NA   0.0234793013   0.0228625751
## Visibility               0.030495588           NA   0.0507693123   0.0470318697
## Wind_Direction           0.009073977           NA   0.1636725647   0.1594527267
## Wind_Speed               0.016420350           NA   0.2089741854   0.1992069447
## Precipitation           -0.002237374           NA   0.0077537918   0.0079719033
## Weather_Condition       -0.023252245           NA   0.0021173303   0.0053916570
## Amenity                  0.090554912           NA   0.0058128553   0.0049118705
## Bump                    -0.003713535           NA  -0.0018226855  -0.0020769116
## Crossing                 0.422232110           NA   0.0216622192   0.0189618350
## Give_Way                 0.056994392           NA   0.0001264558   0.0004091369
## Junction                -0.096141901           NA   0.0045561038   0.0058697933
## No_Exit                  0.023375763           NA   0.0014734879   0.0006865042
## Railway                  0.053726829           NA  -0.0009828347  -0.0008840154
## Roundabout              -0.002108223           NA  -0.0001614062  -0.0004026062
## Station                  0.112592530           NA   0.0201697745   0.0184975004
## Stop                    -0.028258070           NA  -0.0097060380  -0.0107251588
## Traffic_Calming          0.008645874           NA  -0.0012506079  -0.0014869795
## Traffic_Signal           1.000000000           NA   0.0181008879   0.0151387768
## Turning_Loop                      NA            1             NA             NA
## Sunrise_Sunset           0.018100888           NA   1.0000000000   0.9124056450
## Civil_Twilight           0.015138777           NA   0.9124056450   1.0000000000
## Nautical_Twilight        0.011231679           NA   0.8147999181   0.8929869680
## Astronomical_Twilight    0.008023885           NA   0.7299048265   0.8000153230
## Accident_duration       -0.004622117           NA  -0.0096380486  -0.0106934405
## Year                    -0.005837376           NA  -0.0585715835  -0.0637743790
## month                   -0.027515587           NA  -0.0562653984  -0.0523091829
## date                     0.005045334           NA  -0.0014429917  -0.0044730274
##                      Nautical_Twilight Astronomical_Twilight Accident_duration
## Severity                  2.204010e-03          3.242205e-03      7.760625e-03
## Start_Lat                -1.811504e-02         -1.543177e-02     -6.372185e-03
## Start_Lng                 2.131233e-02          1.843706e-02      3.268491e-03
## End_Lat                  -1.811631e-02         -1.543437e-02     -6.381716e-03
## End_Lng                   2.131332e-02          1.843669e-02      3.273304e-03
## Distance.mi.             -2.873781e-04         -1.913821e-05      1.513935e-02
## Temperature.F.            3.007285e-01          2.819563e-01     -1.079583e-03
## Wind_Chill.F.             2.656040e-01          2.506747e-01     -1.467873e-03
## Humidity                 -2.626578e-01         -2.457548e-01      6.937110e-03
## Pressure                  2.178875e-02          2.081504e-02      7.855898e-03
## Visibility                4.579170e-02          4.568426e-02      2.048060e-03
## Wind_Direction            1.528143e-01          1.445812e-01     -4.216882e-03
## Wind_Speed                1.846568e-01          1.703423e-01      1.724114e-03
## Precipitation             7.694115e-03          7.557106e-03      8.045969e-05
## Weather_Condition         4.795306e-03          3.357631e-03     -2.875700e-03
## Amenity                   3.483229e-03          2.595537e-03     -1.152460e-03
## Bump                     -2.350011e-03         -1.893793e-03     -3.451803e-04
## Crossing                  1.563492e-02          1.341747e-02     -4.138297e-03
## Give_Way                  1.017344e-04         -6.371215e-04     -8.663366e-04
## Junction                  7.481557e-03          7.460569e-03      3.327846e-03
## No_Exit                   5.318313e-05         -5.740548e-05     -6.386775e-04
## Railway                  -8.167130e-04         -9.989479e-04      1.672869e-04
## Roundabout               -7.582331e-04         -1.165798e-03     -1.600494e-04
```

```
## Station                    1.632748e-02          1.618485e-02     -2.244536e-03
## Stop                      -1.163037e-02         -1.200975e-02     -1.173116e-03
## Traffic_Calming           -1.809139e-03         -1.509503e-03     -2.653585e-04
## Traffic_Signal             1.123168e-02          8.023885e-03     -4.622117e-03
## Turning_Loop                         NA                    NA               NA
## Sunrise_Sunset             8.147999e-01          7.299048e-01     -9.638049e-03
## Civil_Twilight             8.929870e-01          8.000153e-01     -1.069344e-02
## Nautical_Twilight          1.000000e+00          8.958586e-01     -1.008825e-02
## Astronomical_Twilight      8.958586e-01          1.000000e+00     -1.067584e-02
## Accident_duration         -1.008825e-02         -1.067584e-02      1.000000e+00
## Year                      -6.542977e-02         -6.566058e-02     -1.005741e-02
## month                     -4.576466e-02         -4.393485e-02      6.759776e-03
## date                      -7.901891e-03         -1.036209e-02      1.954235e-03
##                                   Year         month          date
## Severity                -0.3547831032 -1.020766e-01 -0.0057763297
## Start_Lat               -0.1295513123 -6.192261e-02  0.0042246988
## Start_Lng                0.0524712768  2.791618e-02  0.0079932535
## End_Lat                 -0.1295366738 -6.193028e-02  0.0042265384
## End_Lng                  0.0524749838  2.791746e-02  0.0079958614
## Distance.mi.             0.0299837556  2.039021e-02  0.0086156520
## Temperature.F.           0.0197002492  6.552278e-02  0.0027908350
## Wind_Chill.F.            0.1351035487  7.364600e-02  0.0029824511
## Humidity                 0.0132593204  3.570166e-02  0.0189192510
## Pressure                -0.1767900428 -1.578711e-02 -0.0228321727
## Visibility              -0.0273576932 -3.731848e-03 -0.0022866360
## Wind_Direction          -0.0309398741 -5.018044e-02  0.0025950460
## Wind_Speed              -0.0992337598 -7.749245e-02  0.0174244681
## Precipitation           -0.0314896453 -4.514838e-03  0.0033203132
## Weather_Condition       -0.0078163779 -3.071499e-02  0.0202536973
## Amenity                  0.0192305286  6.056200e-03  0.0005390596
## Bump                     0.0071056632  1.284934e-03  0.0005719152
## Crossing                 0.0442481165  6.556266e-03  0.0010773656
## Give_Way                -0.0046199278 -1.367676e-03  0.0010604991
## Junction                -0.1584446043 -3.321162e-02 -0.0031922635
## No_Exit                  0.0040607194  1.646773e-03 -0.0009008451
## Railway                 -0.0001100481 -3.510179e-03 -0.0004577596
## Roundabout               0.0021359946  6.659997e-05  0.0002985394
## Station                  0.0383461286  1.373213e-02  0.0005355828
## Stop                     0.0306924185  5.923711e-03 -0.0006340243
## Traffic_Calming          0.0068863331 -3.238649e-04  0.0003211162
## Traffic_Signal          -0.0058373759 -2.751559e-02  0.0050453341
## Turning_Loop                       NA            NA            NA
## Sunrise_Sunset          -0.0585715835 -5.626540e-02 -0.0014429917
## Civil_Twilight          -0.0637743790 -5.230918e-02 -0.0044730274
## Nautical_Twilight       -0.0654297679 -4.576466e-02 -0.0079018910
## Astronomical_Twilight   -0.0656605814 -4.393485e-02 -0.0103620852
## Accident_duration       -0.0100574079  6.759776e-03  0.0019542354
## Year                     1.0000000000  2.782838e-02  0.0129044669
## month                    0.0278283796  1.000000e+00  0.0242074324
## date                     0.0129044669  2.420743e-02  1.0000000000
```

```
corr_with_target <- cor_matrix[, "Severity"]
corr_with_target[order(abs(corr_with_target), decreasing = TRUE)]
```

```
##              Severity                  Year              End_Lng
##          1.0000000000         -0.3547831032          0.1073674295
##             Start_Lng                 month              End_Lat
##          0.1073669317         -0.1020765713          0.0885540027
##             Start_Lat              Junction         Wind_Chill.F.
##          0.0885522338          0.0650959395         -0.0645724296
##          Distance.mi.            Wind_Speed              Pressure
##          0.0635687045          0.0628377670          0.0369486834
##        Temperature.F.     Weather_Condition              Crossing
##         -0.0272529937          0.0211925636         -0.0203697403
##              Humidity               Station            Visibility
##          0.0192429963         -0.0168644594          0.0166258791
##         Precipitation       Traffic_Signal              Give_Way
##          0.0136270355          0.0131405845          0.0087839051
##                  Stop     Accident_duration                  date
##         -0.0079911271          0.0077606251         -0.0057763297
##        Wind_Direction       Civil_Twilight        Sunrise_Sunset
##          0.0056928468          0.0039898548          0.0039641118
##               Amenity                  Bump Astronomical_Twilight
##         -0.0039579173         -0.0038080356          0.0032422055
##       Traffic_Calming     Nautical_Twilight               Railway
##         -0.0027280910          0.0022040097          0.0018001705
##            Roundabout               No_Exit          Turning_Loop
##         -0.0009553339          0.0002088153                    NA
```

It is evident from the plot that there are almost all variables who are correlated to target variable in similar value. There are a few which are less correlated but we will check the feature importance graph later after modeling in Python to understand which variables are highly correlated to Severity.

```
modeling_data$Severity[modeling_data$Severity == 0] <- "Low"
modeling_data$Severity[modeling_data$Severity == 1] <- "High"

write.csv(modeling_data,"modeling_data.csv", row.names = TRUE)
```

Here we have converted the 0 and 1 to “Low” and “High” for modeling and better understanding. We have then exported the data as a csv, to import it in Python Notebook for modeling and prediction.

Please refer to TEAM5_Modeling.ipynb to check out modeling and prediction steps.