# Key factors that affect the severity of road accidents: A study on weather condition and road condition

Yaow Hui Chong

University of Nebraska at Omaha

Omaha, NE 68182

ychong@unomaha.edu

Abstract

The road accidents rate in the United States is constantly increasing in recent years. It remains a public health challenge to reduce the road accident cases. What are the key factors that affect the severity of road accidents? This research paper looks into the matter from three different perspectives: weather condition, road condition, and a combination of both weather and road condition. In the weather condition study, a set of weather condition factors are examined to understand their relationships with the road accident severity. Likewise, in the road condition study, a set of road condition factors are examined to understand their relationships with the road accident severity. In the study of the combination of both weather and road condition, a decision tree model is built using the weather condition and road condition factors to see what are the certain conditions that could contribute to an increase or a decrease of the road accident severity. In the weather condition study, the result shows that the snow, hail/freezing rain, and the thunderstorm conditions are the weather condition factors that significantly increase the road accident severity. The rain factor has a slightly positive impact on the road accident severity. In the road condition study, our result shows that the use of Amenity, Bump, Crossing, Give way, No Exit, Railway, Roundabout, Station, Stop, Traffic Calming, and Traffic Signal lead to a decrease in the road accident severity. However, the use of Junction leads to an increase of the road accident severity. Also, the study of the combination of both weather and road condition suggests certain combinations of weather and road conditions that would increase or decrease the road accident severity.

## 1. INTRODUCTION

Automobiles are now commonly used everywhere in the world. In 2019, there were around 17 million light vehicles sold in the U.S., accounting for about 97 percent of the roughly 17.5 million motor vehicles that were sold in the United States in 2019. (Wagner I., 2020). Each year, the motor vehicle collisions cause more than 1.2 million deaths worldwide and with millions more sustaining serious injuries and living with long-term adverse health consequences. (World Health Organization, 2015). Road traffic injuries are currently estimated to be the ninth leading cause of death across all age groups globally. (World Health Organization, 2015). Also, it is predicted to become the seventh leading cause of death by 2030. (World Health Organization, 2015). From figure 1, we can see the United States has an increased number of 13.5% of road deaths from 2010-2016. It shows the importance of studying the factors that create severe road accidents. Therefore, the purpose of this study is to analyze the underlying factors that are creating severe road accidents.

## 2. LITERATURE REVIEW

There are some existing research papers analyzing the relationships of road accidents rates and the underlying factors. The road conditions, weather conditions such as rain, snow, fog, and driver conditions such as gender, age, driver behavior, and drunken driving, and the road accident rates in different locations are some of the topics in the past research.

### 2.1 Human Factors

The characteristics of drivers is one of the main factors that are causing the road accidents. The research on young drivers' characteristics include excessive speed, driving recklessly, traffic violations, and drugs and alcohol (Gonzales et al., 2005; Lam, 2003; Bingham et al., 2008). Also, accidents related to older drivers are often related to driver errors due to age-related decline in visual, cognitive, and mobility functioning in older age (Hu et al., 1993; Janke, 1991). The driver errors likely cause accidents at intersections, make directional turns, failure to yield right of way, failure to comply with signs and signals, failure to see objects, and improper lane changes. (Hakamies-Blomqvist, 1993; Langford & Koppel, 2006; McGwin & Brown, 1999). The research of Ashraf I. et. al. (2019) suggested that drivers with 10 years or more driving experience have a higher probability of involving in a road accident. Also, male drivers have a higher accident rate than female drivers. (Ashraf I. et. al., 2019; Begg & Langley 2004; Clarke et al., 2006; Curry et al., 2012). A recent study by Klauer et al., (2014) shows that distractions can increase the risk of car accidents, and a particular example is the use of mobile phones when driving increases risk of both young and experienced drivers.

### 2.2 Weather Condition Factors

There are some research papers relating the road accident rates to the weather factors. Research of Andrey J. et al. (2003) suggested that the precipitation is associated with the increase in traffic collisions, and the snowfall effect is more pronounced than the rainfall effects in the increase of traffic collisions. Some researchers also claimed that increases in rainfall often cause the increase of road accidents. (Fridstrom and Ingebrigtsen 1991; Chang and Chen 2005; Caliendo et al. 2007, Shankar et al. 1995; Keay and Simmonds 2006; Hermans et al. 2006). Also, the extreme temperatures, such as low in winter and high in summer showed positively correlated with the road accidents rate. (Malyshkina et al., 2008). On the other hand, the number of hours of sunlight appears to increase the road accident rates. (Fridstrom et al, 1995, Hermans et al. 2006). The time-series model is often chosen to perform analysis of the influence of weather conditions on road crashes; researches are done in monthly time scale (Hermans et al., 2006) and daily time scale (Brijs et al., 2008).

### 2.3 Road Condition Factors

Many studies suggested that the road condition will affect the road accident rates. The research of Chen et al. (2018) claimed that the roadway geometric characteristics greatly affect the crash likelihood; crash likelihood decreases when the number of merging ramps per lane per mile gets higher. Also, blacktop road surface was associated with a significant increase in fatality risk for driver's gender and age groups, compared with gravel/stone and concrete surface. (Li et al., 2017). The research of Lee J. & Mannering F. (1999) suggested that the run-off roadway accident frequencies can be significantly reduced by increasing lane and shoulder widths, widening medians, expanding approaches to bridges, shielding relocating, and removing roadside hazardous objects, and flattening side slopes and medians, which means

these factors are significantly affecting the accident rates.

### 2.4 Road Accidents Heat Maps

Research of Colacino V.G. & Po L. (2017) proposed the use of road accidents heat map to identify which road sections have a higher accident rate causing high number of deaths and injuries. It uses road accident open datasets and displays the number of road accidents using the heat maps visualization. The research was done using the Polymaps, a Javascript library for making dynamic and interactive maps to show the colors for routes with different accident rates. Users could take appropriate cautions where they are driving in the high accident rate areas. Similar research is also done by Pritee K. & Garg R.D. (2017) but with the use of Quantum Geographic Information System (Q-GIS) via Kernel method feature on heatmap analysis, buffer analysis, and the nearest neighborhood analysis. The heatmaps will be published on the internet and updated periodically; they are freely accessible by anyone at zero cost.

## 3. METHODOLOGY

### 3.1 Research Question

The literature review part concludes that the road accident rates can be affected by human factors, weather conditions, and road conditions. Current research proposes that road accidents heat maps can help reduce the accident rate because the users will take appropriate cautions when they are driving in the high accident rate areas. However, there might be some certain situations that the combination of weather conditions and road conditions that affect the severity of road accidents.

Research question:

1) Will certain weather conditions affect the severity of road accidents?

2) Will certain road conditions affect the severity of road accidents?

3) Will certain combinations of weather conditions and road conditions affect the severity of road accidents?

This research aims to further study the multiple conditions affecting the severity of road accidents. The factors that cause the severity of road accidents with the combination of weather conditions and road conditions will be analyzed and identified.

### 3.2 Dataset

The dataset is retrieved from www.smoosavi.org. This dataset has been collected in real-time with two traffic APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. (Moosavi, Samavatian, Nandi, Parthasarathy, Rajiv Ramnath,2019). The dataset includes accident data in 49 states collected from February 2016 to June 2020, and it has about 3.5 million accident records. The full data dictionary table is included in the Appendix section (B1). The target variable is the severity, and it indicates the severity of the road accidents in the United States. Some important predictor variables for weather condition are temperature, humidity, pressure, visibility, wind_speed, weather_condition, Also, some important predictor variables for road condition are amenity, bump, crossing, give_way, no_exit, railway, roundabout, station, stop, traffing_calming, and traffic_signal.

## 3.3 Tools

The dataset will be analyzed with the use of RStudio. RStudio is an open-source tool for the R programming language and supports its ongoing development. (Carlsson K. et al., 2020). Forrester rated RStudio is a strong performer among notebook-based predictive analytics and machine learning (PAML) providers. It stated that the RStudio has the ability to calculate descriptive statistics and visualize with common charts and reports used for data exploration and interactive exploration conducted with external BI tools-plus the vendor includes and automatically creates descriptive statistics visualization for exploration based on data science best practices.

This research uses a lot of data exploration and interactive exploration, as well as data visualization. Thus, RStudio is very suitable to be used.

## 3.4 Data Cleaning

Data Cleaning is the first part of this research. According to tableau.com (n.d.), data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicated, or incomplete data within a dataset. Some of the data cleaning works in this research include:

### 3.4.1. Removed some irrelevant variables

Several irrelevant variables were removed, including:

ID, End_Time, End_Lat, End_Lng, Distance.mi., Description, Street, County, Zipcode, Airport_Code, Side, Timezone, Weather_Timestamp, Wind_Direction, Civil_Twilight, Nautical_Twilight, Astronomical_Twilight

These variables were removed because they are not relevant to our research and will not be used for this analysis.

### 3.4.2. Handling missing data

Some variables contain too many missing data, so I remove these variables:

TMC, Number, Wind_Chill.F., Precipitation.in.

Also, there are some variables contain an acceptable number of missing data, so I remove the certain rows that contain missing data:

Temperature.F., Humidity, Pressure.in., Visibility.mi., Wind_Speed.mph., Sunrise_Sunset

### 3.4.3. Remove some inappropriate data

I removed some observations that do not make sense, such as Temperature data that show 167 °F. I capped highest temperature at 130 °F.

In addition, I removed both Country and Turning_Loop variables because they only contain 1 level of data, which is meaningless to analyze them.

### 3.4.4. Regroup data

The original Weather_Condition data is a factor data that contains 128 levels with many duplicate conditions, such as “Rain”, “Rain shower”, “Heavy Rain shower”. To prevent having too many meaningless levels, I regrouped the factor levels. The Weather_Condition now contains:

| Factor levels | Number of Observations |
|---|---|
| Dust | 275 |
| Windy | 143 |
| Snow | 61574 |
| Clear | 1363959 |
| Cloudy | 1644606 |
| Rain | 254459 |
| Fog | 79952 |
| Hail/Freezing Rain | 3713 |
| Thunderstorm | 28891 |
| Other | 25 |

**Table 1: factor levels of Weather_condition data**

The original dataset contains 47 variables with 3,513,740 observations. After the data cleaning processing, the dataset contains 26 variables and 3,028,012 observations.

## 4. DATA ANALYSIS

### 4.1 Weather condition variables

I analyzed the data to understand the relationships between the frequency of car accidents and severity of the car accidents with the weather condition variables and road condition variables.

#### 4.1.1 Temperature

The relationship between temperature and the US accident rates looks like a normal random variable shape. There is no significant evidence showing that the accident rate is related to temperature. The data visualization of the temperature variable is available at Appendix B.1.

#### 4.1.2 Humidity

The accidents happen more frequently when the humidity is higher. It is perhaps that rain is one of the reasons for car accidents. The data visualization of the temperature variable is available at Appendix B.2.

#### 4.1.3 Weather condition

There were more road accidents recorded in the Clear and Cloudy weather condition. It makes sense since the other weather conditions such as snow, rain, fog, thunderstorm, and so on, happen less frequently than these two weather conditions. (Please refer to Appendix B.3 for the weather condition plots).

From the boxplot we observed that there is no significant relationship found between the severity of road accidents and the weather condition. However, from the table below we observed that the mean of severity is higher when the weather is in snow (mean = 2.431), Hail/Freezing Rain (mean = 2.511682), and Thunderstorm (mean = 2.425220) condition.

Table of relationship between weather condition and the mean of severity

| Weather Condition | Mean | Number of observations |
|---|---|---|
| Dust | 2.195312 | 256 |
| Windy | 2.097902 | 143 |
| Snow | 2.431099 | 57481 |
| Clear | 2.304512 | 1161385 |
| Cloudy | 2.344816 | 1481448 |
| Rain | 2.368888 | 235250 |
| Fog | 2.287831 | 61557 |
| Hail/Freezing Rain | 2.511682 | 2996 |
| Thunderstorm | 2.425220 | 27494 |
| Other | 2.000000 | 2 |

**Table 2: The means of Weather_condition data**

A correlation matrix for numeric data also attached to examined other weather condition variables. There is no significant relationship found between severity and the other numeric data such as pressure, visibility, and wind speed.

### 4.2 Road condition variables

The data visualization graphs of the road condition variables are attached at the Appendix (A.10 – A.48). The graphs do not show much of the difference if the road conditions do affect the severity of the road accidents. I analyze the variables using data tables and include them below.

Based on the result, the use of Amenity, Bump, Crossing, Give_Way, No_Exit, Railway, Roundabout, Station, Stop, Traffic Calming, and Traffic Signal does effectively reduce the severity of road accidents because they have lower mean severity than not using these road control. This makes sense because drivers usually slow down their driving speed when they

see these traffic signs. However, the use of Junction will increase the severity of road accidents. The tables of relationship between road condition variables and the mean of severity are shown below.

| Amenity | Mean | Number of observations |
|---|---|---|
| False | 2.335240 | 2991776 |
| True | 2.112761 | 36236 |

**Table 3: The severity means of Amenity data**

| Bump | Mean | Number of observations |
|---|---|---|
| False | 2.332609 | 3027481 |
| True | 2.156309 | 531 |

**Table 4: The severity means of Bump data**

| Crossing | Mean | Number of observations |
|---|---|---|
| False | 2.355272 | 2785513 |
| True | 2.071893 | 242499 |

**Table 5: The severity means of Crossing data**

| Give_Way | Mean | Number of observations |
|---|---|---|
| False | 2.332824 | 3019973 |
| True | 2.240080 | 8039 |

**Table 6: The severity means of Give_Way data**

| Junction | Mean | Number of observations |
|---|---|---|
| False | 2.322894 | 2784388 |
| True | 2.443257 | 243624 |

**Table 7: The severity means of Junction data**

| No_Exit | Mean | Number of observations |
|---|---|---|
| False | 2.332704 | 3024158 |
| True | 2.233524 | 3854 |

**Table 8: The severity means of No_Exit data**

| Railway | Mean | N |
|---|---|---|
| False | 2.333877 | 3001528 |
| True | 2.185357 | 26484 |

**Table 9: The severity means of Railway data**

| Roundabout | Mean | N |
|---|---|---|
| False | 2.332593 | 3027853 |
| True | 2.044025 | 159 |

**Table 10: The severity means of Roundabout data**

| Station | Mean | N |
|---|---|---|
| False | 2.336334 | 2967931 |
| True | 2.147035 | 60081 |

**Table 11: The severity means of Station data**

| Stop | Mean | N |
|---|---|---|
| False | 2.336143 | 2983204 |
| True | 2.095251 | 44808 |

**Table 12: The severity means of Stop data**

| Traffing_Calming | Mean | N |
|---|---|---|
| False | 2.332639 | 3026825 |
| True | 2.176074 | 1187 |

**Table 13: The severity means of Traffing_Calming data**

| Traffic Signal | Mean | N |
|---|---|---|
| False | 2.381997 | 2480149 |
| True | 2.108859 | 547863 |

**Table 14: The severity means of Railway data**

## 4.3 Combination of weather and road condition variables

I use the decision tree model to study if the combination of weather condition variables and the road condition variables affect the severity of road accidents. A regression tree is created, and the graph is shown below:

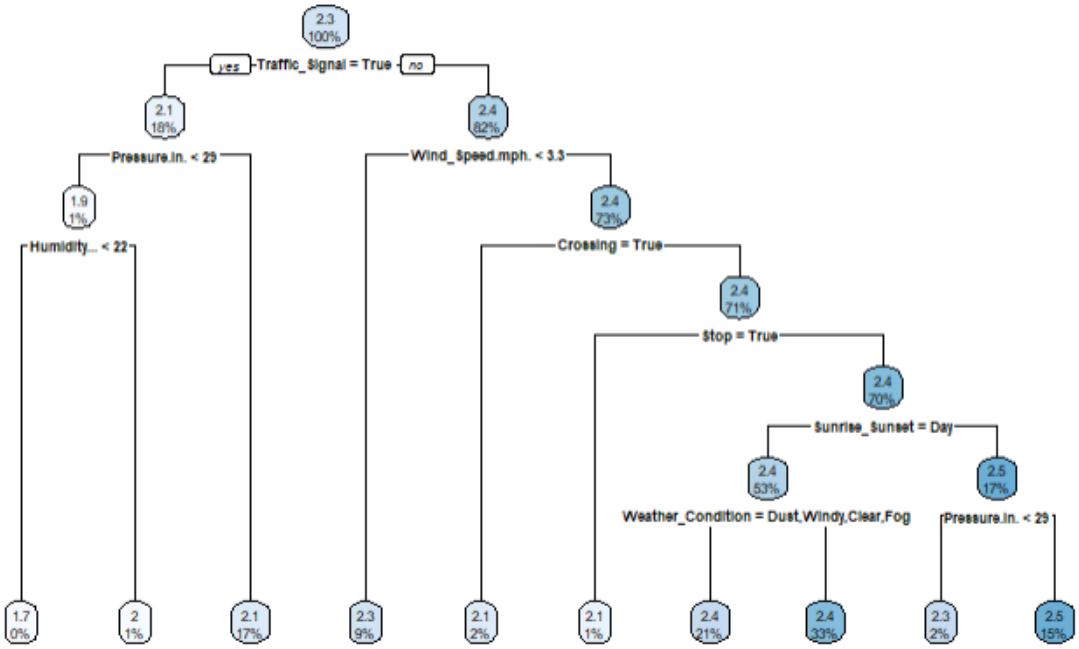



**Figure 1: Decision Tree of the combination studies of weather and road condition variables**

The decision tree created ten nodes and the result can be interpreted as below:

1. If Traffic_Signal = True, Air pressure < 29 inches, and humidity < 22%, we predict 1.7 severity.

2. If Traffic_Signal = True, Air pressure < 29 inches, and humidity >= 22%, we predict 2.0 severity.

3. If Traffic_Signal = True and Air pressure >= 29 inches, we predict 2.1 severity.

4. If Traffic_Signal = False and Wind speed < 3.3 mph, we predict 2.3 severity.

5. If Traffic_Signal = False, Wind speed >= 3.3 mph, and Crossing= True, we predict 2.1 severity.

6. If Traffic_Signal = False, Wind speed >= 3.3 mph, Crossing= False, and Stop = True, we predict 2.1 severity.

7. If Traffic_Signal = False, Wind speed >= 3.3 mph, Crossing= False, and Stop = False, Sunrise_Sunset = Day, and Weather_Condition = Dust/Windy/Clear/Fog, we predict 2.4 severity.

8. If Traffic_Signal = False, Wind speed >= 3.3 mph, Crossing= False, and Stop = False, Sunrise_Sunset = Day, and Weather_Condition = Snow/Cloudy/Rain/Hail/Thunderstorm/Other, we predict 2.4 severity.

9. If Traffic_Signal = False, Wind speed >= 3.3 mph, Crossing= False, and Stop = False, Sunrise_Sunset = Night, and Air pressure < 29 inches, we predict 2.3 severity.

10. If Traffic_Signal = False, Wind speed >= 3.3 mph, Crossing= False, and Stop = False, Sunrise_Sunset = Night, and Air pressure >= 29 inches, we predict 2.5 severity.

Therefore, we conclude that variables like traffic signal, air pressure, humidity, wind speed, crossing, stop, weather condition and sunrise_sunset affect the severity of road accidents.

5. RESULT

From the data exploratory and data analysis, we have the result for the following research questions.

1) Will certain weather conditions affect the severity of road accidents?

Based on the correlation matrix, we found no significant relationships between the variable severity and the variables temperature, humidity, pressure, visibility, and wind speed. However, we found that some weather conditions increase the severity of road accidents, these include snow, hail/freezing rain, and thunderstorms.

| Effects on Severity | Weather conditions |
|---|---|
| Increase | Snow |
| | Hail/freezing rain |
| | Thunderstorm |

2) Will certain road conditions affect the severity of road accidents?

Our findings show that the road conditions generally affect the severity of road accidents. Based on the result, the use of Amenity, Bump, Crossing, Give_Way, No_Exit, Railway, Roundabout, Station, Stop, Traffic Calming, and Traffic Signal does effectively reduce the severity of road accidents because they have lower mean severity than not using these road control. However, the use of Junction will increase the severity of road accidents.

| Effects on Severity | Road conditions |
|---|---|
| Decrease | Amenity |
| | Bump |
| | Crossing |
| | Give_Way |
| | No_Exit |
| | Railway |
| | Roundabout |
| | Station |
| | Stop |
| | Traffic Calming |
| | Traffic Signal |
| Increase | Junction |

3) Will certain combination of weather conditions and road conditions affect the severity of road accidents?

| Combination | Prediction of road accident severity | Increase/ Decrease |
|---|---|---|
| Traffic_Signal = True, Air pressure < 29 inches, and humidity < 22% | 1.7 | Decrease |
| Traffic_Signal = True, Air pressure < 29 inches, and humidity >= 22% | 2.0 | |
| Traffic_Signal = True and Air pressure >= 29 inches | 2.1 | |
| Traffic_Signal = False and Wind speed < 3.3 mph | 2.3 | |
| Traffic_Signal = False, Wind speed >= 3.3 mph, and Crossing= True | 2.1 | |
| Traffic_Signal = False, Wind speed >= 3.3 mph, Crossing= False, and Stop = True | 2.1 | |
| Traffic_Signal = False, Wind speed >= 3.3 mph, Crossing= False, and Stop = False, Sunrise_Sunset = Night, and Air pressure < 29 inches | 2.3 | |
| Traffic_Signal = False, Wind speed >= 3.3 mph, Crossing= False, and Stop = False, Sunrise_Sunset = Day, | 2.4 | Increase |

| | | |
|---|---|---|
| Weather_Condition = Dust/Windy/Clear/Fog | | |
| Traffic_Signal = False, Wind speed >= 3.3 mph, Crossing= False, and Stop = False, Sunrise_Sunset = Day, Weather_Condition = Snow/Cloudy/Rain/Hail/ Thunderstorm/Other | 2.4 | |
| Traffic_Signal = False, Wind speed >= 3.3 mph, Crossing= False, and Stop = False, Sunrise_Sunset = Night, and Air pressure >= 29 inches | 2.5 | |

With the use of a decision tree model, we can predict the severity of the road accidents under certain combination of weather and road condition. In our dataset, the mean of the severity variable is 2.332578, with a prediction of lower than the mean, we will categorize them as a decrease in the severity of road accident, however, with a prediction of higher than the mean, we will categorize them as an increase in the severity of road accident.

## 6. CONCLUSION

In the past, there were some research papers about the factors affecting the frequencies of road accidents and the severity of road accidents. Most of these studies suggest that the road accident is related to the human factors, weather conditions, and road conditions. The result of this research shows that the weather conditions and road conditions are generally affecting the severity of road accidents, which are supported by the previous studies. Our findings show that the weather conditions that affect the severity of road accidents are snow, hail/freezing rain, and thunderstorm. All three weather conditions are increasing the severity of road accidents compared to other weather conditions.

Besides that, our findings show that the road conditions that affect the severity of road accidents are the presence of amenity, bump or hump, crossing, give way, junction, no exit, railway, roundabout, station, stop, traffic calming, and traffic signal. All of these road conditions are decreasing the severity of the road accidents except with the junction road condition. The junction road condition increases the severity of road accidents, and this might be because of the confusion of road traffic with the use of junctions. Some reckless drivers might not stop and observe before crossing the junction, and this leads to the crash between vehicles. All other road conditions are controlling the driving speed of divers, and that could be one of the reasons the severity of road accidents is lower with the existence of these road conditions.

These result findings also suggest that there might be some certain combination of weather and road conditions that cause a higher severity of road accidents. In our dataset, the mean of the severity variable is 2.332578, our findings suggest seven combinations of weather and road conditions that lead to a lower severity of road accident and three combinations of weather and road conditions that lead to a higher severity of road accident. The factors that impact the prediction of road accident severity in the combination studies are Traffic_Signal, Pressure, Humidity, Wind_Speed, Crossing, Stop, and Sunrise_Sunset.

## 7. RECOMMENDATION

Some oddity is found in the data visualization in this research. From the distribution of the severity plot in Appendix A.3, we observed that the road accidents in California and Oregon are less likely to be high severity compared to the other states. This might be a good future research to examine whether the California and Oregon states have different traffic systems that lead to a lower road accident severity compared to the other states.

On the other hand, while there are other factors that affect the severity of road accidents such as human factors, there are also factors like the vehicle protection, roadside objects such as tree branches, poles, and animals that could impact the severity of road accidents. We suggest a future research that includes these factors as well as identifies whether the crash is vehicle-vehicle, vehicle-human, vehicle-animal, or vehicle-infrastructure.

## 8. REFERENCES


1. Andrey J. & Mills B. & Leahy M. & Suggett J. (2003) Weather as a chronic hazard for road transportation in Canadian cities. *Natural Hazards*, 28, 319-343

2. Begg D.J. & Langley J.D. (2004). Identifying predictors of persistent non-alcohol or drug-related risky driving behaviours among a cohort of young adults. *Accid. Anal. Prev*., 36(6), 1067-1071.

3. Bingham C.R. & Shope J.T. & Zhu J. (2008). Substance-involved driving: predicting driving after using alcohol, marijuana, and other drugs. *Traffic Inj. Prev*, 9(6), 515-526.

4. Brijs T. & Karlis D. & Wets G. (2008) Studying the effect of weather conditions on daily crash counts using a discrete time series model. *Accident Analysis and Prevention 40(3)*, 1180-1190.

5. Caliendo C. & Guida M. & Parisi A., (2007). A crash-prediction model for multilane roads. *Accident Analysis and Prevention 39*, 657-670.

6. Carlsson K. & Gualtieri M. & Dridharan S. & Perdoni R. (September 10, 2020). RStudio is a strong performer among notebook-based predictive analytics and machine learning (PAML) providers. Forrester.

7. Chang L.Y. & Chen W.C. (2005). Data mining of tree-based models to analyze freeway accident frequency. *Journal of Safety Research*, 36, 365-375.

8. Chen F. & Chen S. & Ma X. (2018). Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data. *Journal of Safety Research*, 65, 153-159

9. Clarke D.D. & Ward P. & Bartle C. & Truman W. (2006). Young driver accidents in the UK: the influence of age, experience, and time of day. *Accid. Anal. Prev*., 38(5), 871-878.

10. Colacino V.G. & Po L. (2017). Managing road safety through the use of linked data and heat maps. *WIMS '17: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*, 18, 1-8.

11. Curry A.E. & Mirman J.H. & Kallan M.J. & Winston F.K. & Durbin D.R. (2012). Peer passengers: how do they affect teen crashes? *J. Adolesc. Health*, 50(6), 588-594.

12. Fridstrom L. & Ingebrigtsen S. (1991). An aggregate accident model based on pooled, regional time-series data. *Accident Analysis and Prevention* 23(5), 363-378.

13. Fridstrøm L. & Ifver J. & Ingebrigtsen S. & Kulmala R. & Thomsen L.K. (1995). Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis & Prevention*, 27 (1), 1-20.

14. Gonzales M.M. & Dickinson L.M & DiGuiseppi L.M. & Lowenstein S.R. (2005). Student drivers: A study of fatal motor vehicle crashes involving 16 year-old-drivers. *Ann. Emerg. Med*., 45(2), 140-146.

15. Hakamies-Blomqvist L. (1993). Fatal accidents of older drivers. *Accid. Anal. Prev*., 25(1), 19-27

16. Hayat R. & Debbarh M. & Antoniou C. & Yannis G. (2013). Explaining the road accident risk : Weather effects. *Accident Analysis and Prevention, Elsevier*, 60, 456-465.

17. Hermans E. & Wets G. & Van Den Bossche F. (2006). Frequency and Severity of Belgian

Road Traffic Accidents Studied by State-Space Methods. *Journal of Transportation and Statistics*, 9(1), 63-76.

18. Hu P.S. & Young J.R. & Lu A. (1993). Highway crash rates and age related-driver limitations: literature review and evaluation of databases. *National Highway Traffic Safety Administration, Report No. ORNL:TM-12456*. Washington, D.C.

19. Janke M.K. (1991). Accidents, mileage, and the exaggeration of risk. *Accid. Anal. Prev*., 23(2), 183-188

20. Keay K., Simmonds I. (2006). Road accidents and rainfall in a large Australian city. *Accident Analysis and Prevention*, 38, 445-454.

21. Lam L.T. (2003). Factors associated with young drivers' car crash injury: comparisons among learner, provisional, and full licenses. *Accid. Anal. Prev*, 35(6), 913-920.

22. Langford J. & Koppel S. (2006). Epidemiology of older driver crashes-identifying older driver risk factors and exposure patterns. *Transp. Res. Part F: Traffic Psychol. Behav*., 9(5), 309-321.

23. Lee J. & Mannering F. (1999). Analysis of roadside accident frequency and severity and roadside safety management. *Final Research Report, Washington State Transportation Center (TRAC)*

24. Li Q. & Liang S. & Nie K. & Tu W. & Wang Z. (2017). Analyzing risk factors for fatality in urban traffic crashes: A case study of Wuhan, China.

25. McGwim G. & Brown D.B. (1999). Characteristics of traffic crashes among young, middle-aged, and older drivers. *Accid. Anal. Prev*., 31(3), 181-198.

26. Mosavi S. & Mohammed H.S. & Srinivasan P. & Rajiv R. (2019). A countrywide traffic accident dataset. *ArXiv preprint arXiv: 1906.05409*.

27. Mosavi S. & Mohammed H.S. & Srinivasan P. & Rajiv R. (2019). Accident risk prediction based on heterogeneous sparse data: new dataset and insights. *In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM*, 2019.

28. Pritee K. & Garg R.D. (2017). Cloud based spatial visualization with statistical approach for road accidents. *Spat. Inf. Res*., 25(6), 825-835.

29. Shankar V. & Mannering F. & Barfield W. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis & Prevention*, 27 (3), 371-389.

30. Tableau (n.d.). Data cleaning: The benefits and steps to creating and using clean data. Tableau.

31. Wagner I. (May 6, 2020). U.S. Automotive industry – statistic & facts. *Statista*. https://www.statista.com/topics/1721/us-automotive-industry/

32. World Health Organization (2015). Global status report on road safety 2015. *World Health Organization.* https://www.who.int/violence_injury_prevention/road_safety_status/2015/en/

8. APPENDIX

There are five sections in the Appendix, which include:

A. Additional Data Analysis (Source, Spatial, and time data)

B. Additional Data Analysis (Weather condition variable)

C. Additional Data Analysis (road condition variable)

D. Data Dictionary

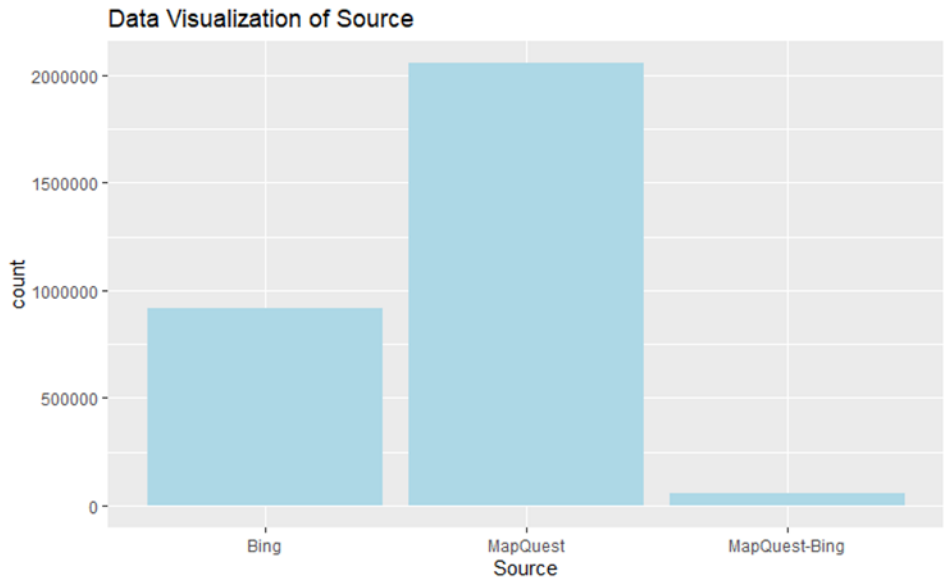
E. R code

---

A. Additional Data Analysis (Source, Spatial, and time data)

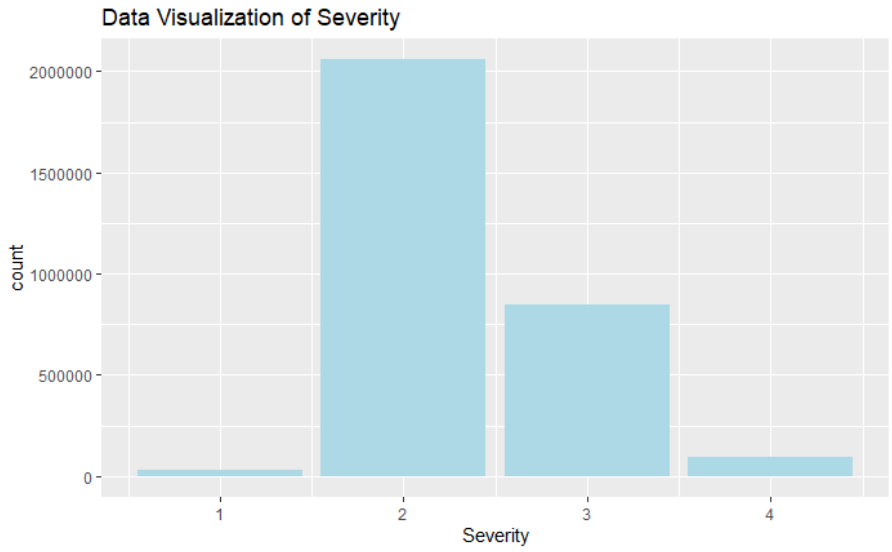
A.1 Source

There are three sources the accident report. It indicates which API reported the accident, the frequency distributions are shown below:



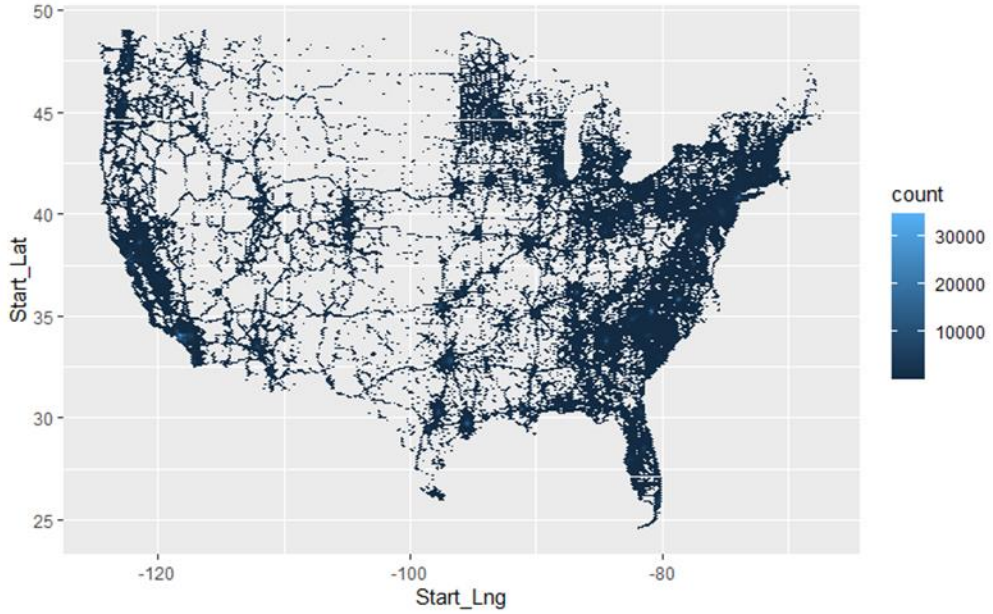

A.2 Severity

The diagram shows the severity of US Car Accident from 2016-2020. A majority of the car accidents are rated as ‘2’ and ‘3’ in severity.

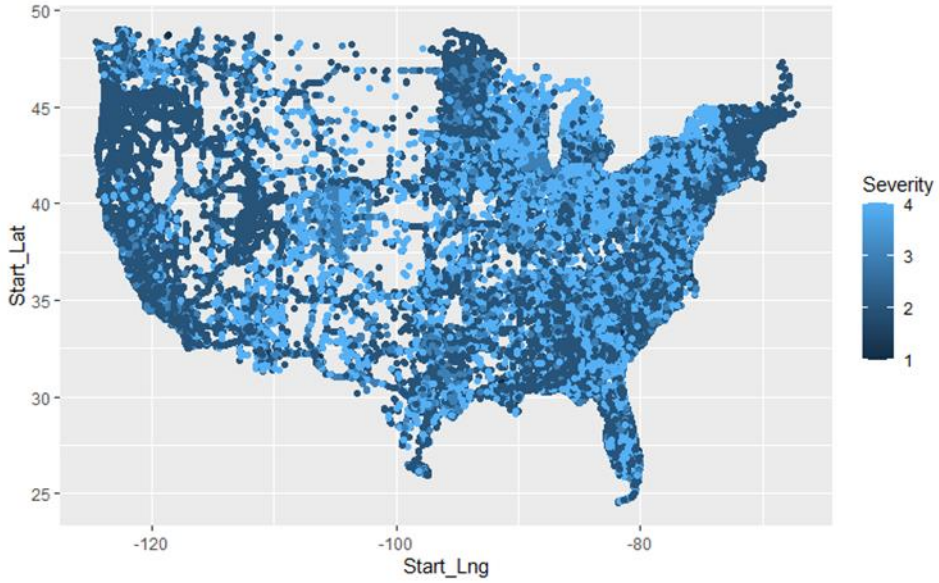



A.3 Latitude and Longitude

The diagram shows the distribution of US accident happened from 2016 to 2020. We observe that there are more car accidents happen in the East and West coast in the United States.

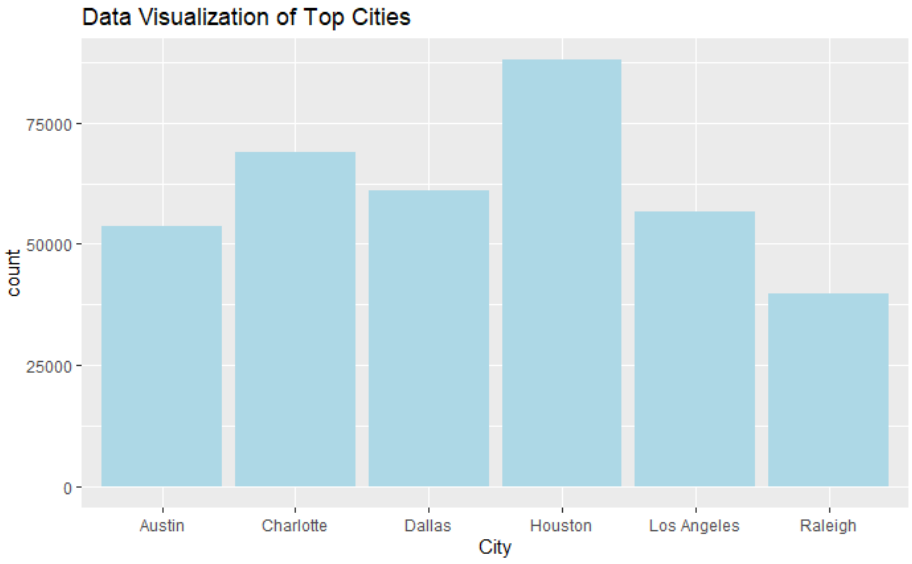


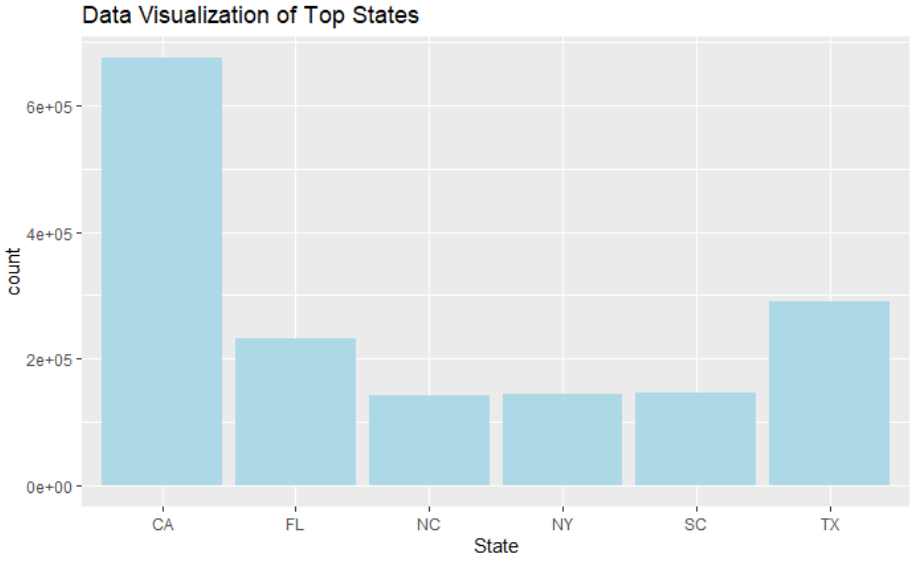

We also did analysis on the distribution of US accidents based on severity. We observed that the severity '4' accidents (the light blue dots) are more centrally distributed; they are frequently happened far from the coast. This might be something to analyze in future research.

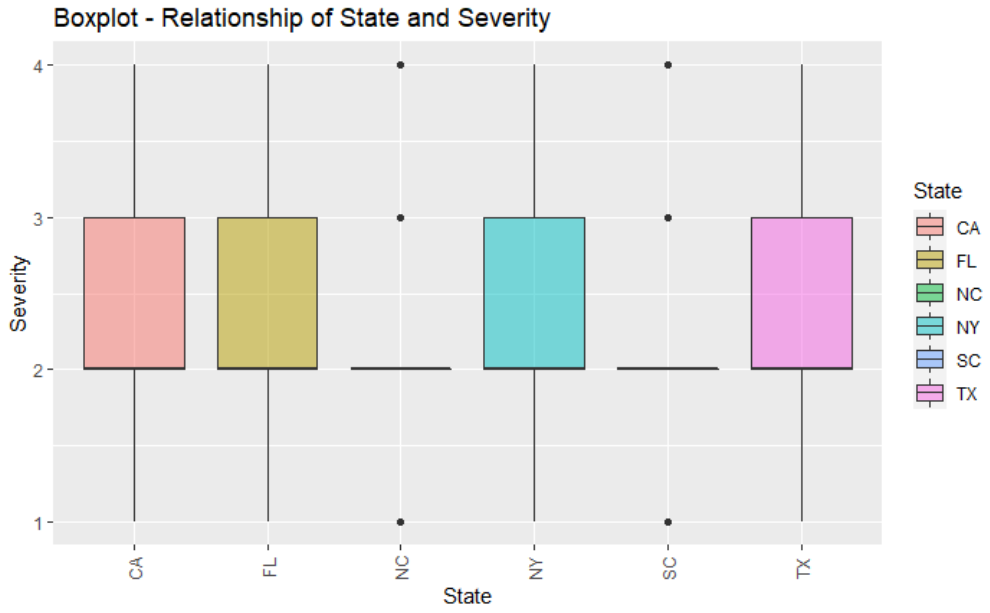




A.4 City and State

The graphs below show the cities and the states that have the highest road accident rates in United States. As shown in the graph, we can see that California has the highest accident rate, followed by Texas and Florida. However, Texas, Charlotte, and Dallas have a higher road accident rate compared to Los Angeles. In addition, the boxplot shown that the North Carolina and South Carolina have less severe road accidents happened compare to the other top states.


Data Visualization of Top Cities
75000
50000
25000
0
count
Austin
Charlotte
Dallas
Houston
Los Angeles
Raleigh
City


Data Visualization of Top States
6e+05
4e+05
2e+05
0e+00
count
CA
FL
NC
NY
SC
TX
State

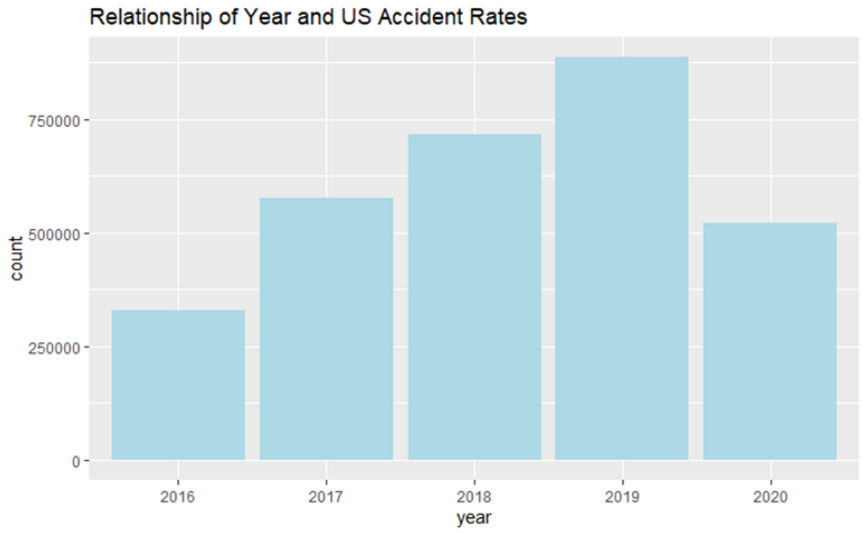


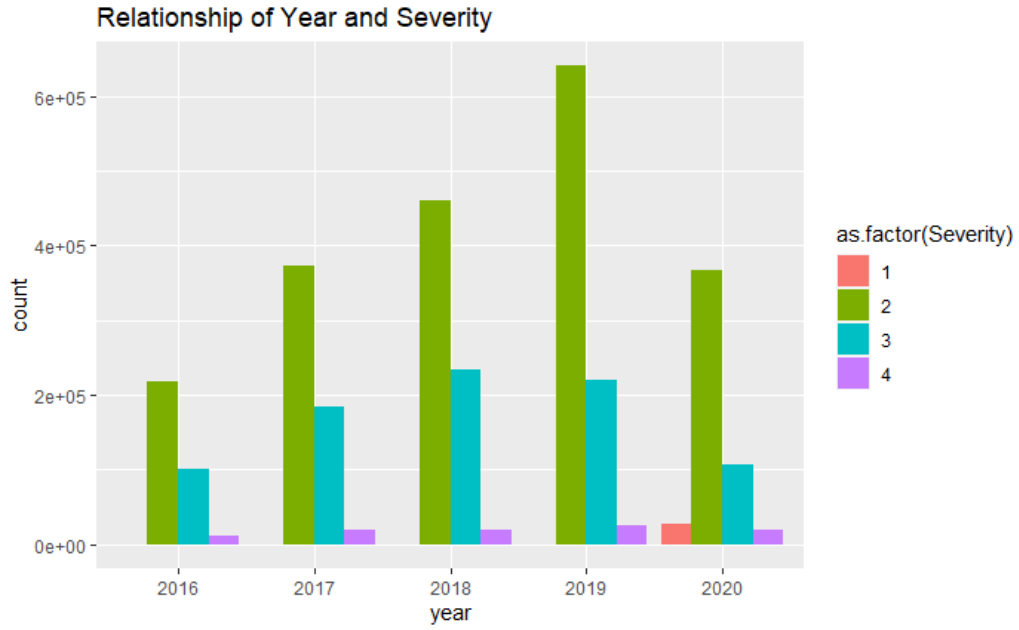

A.5 Time

I split the time variable to 'Year', 'Month', 'Day', and 'Hour'. Then, I explore each of their relationships to the frequency of car accidents and severity of the car accidents. I also included the discussion of Sunrise_Sunset variable in this section.

A.5.1 Year

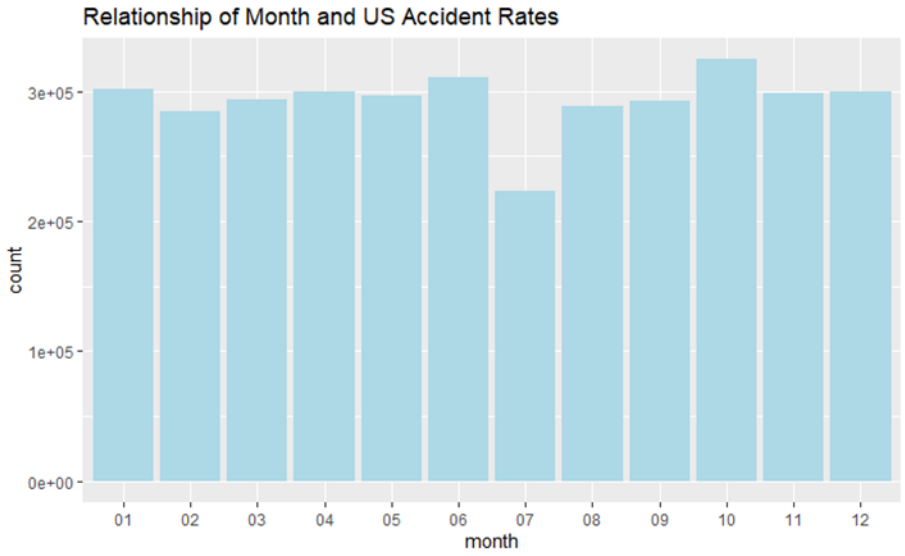
As shown in the diagram, the accident rates increase from 2016- 2019. That means there are more and more car accidents happened from 2016-2019. The accident data was collected from February 2016 to June 2020, and that explains the situation that the accident rate in 2020 is lower than the previous years. Interestingly, we observed that there were less severity '3' accidents happening in 2019 compared to 2018.

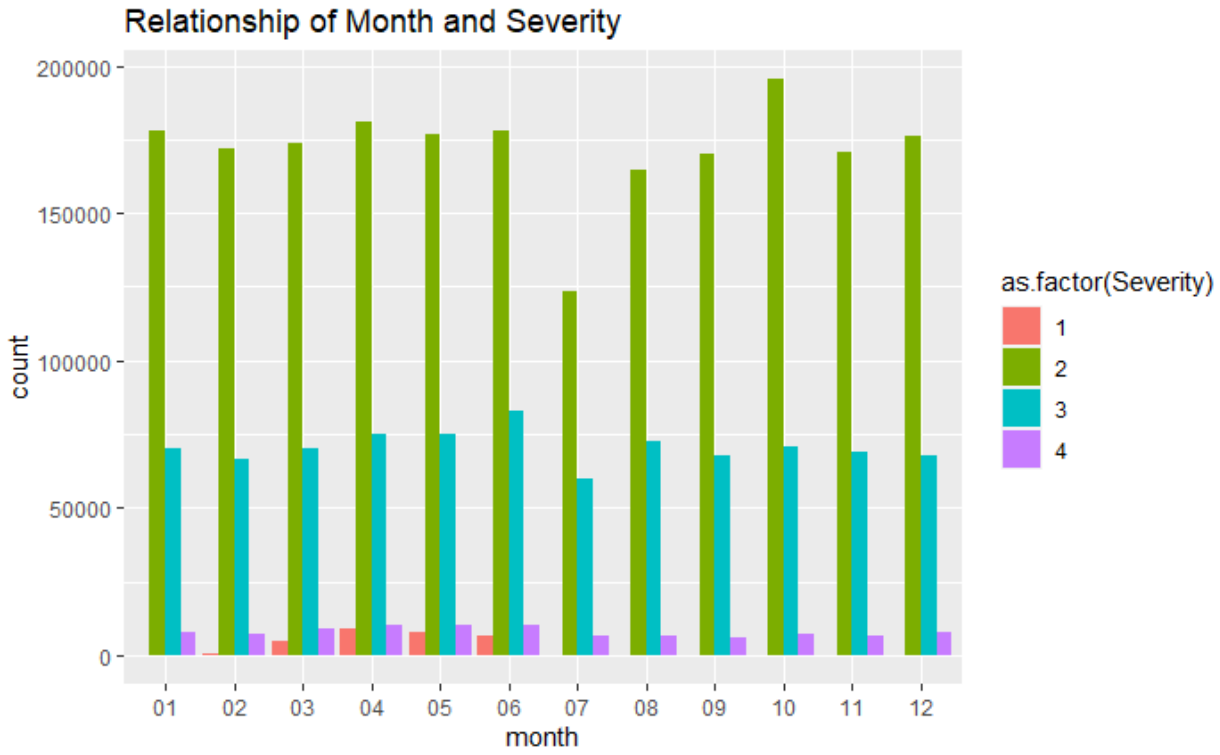



A.5.2 Month

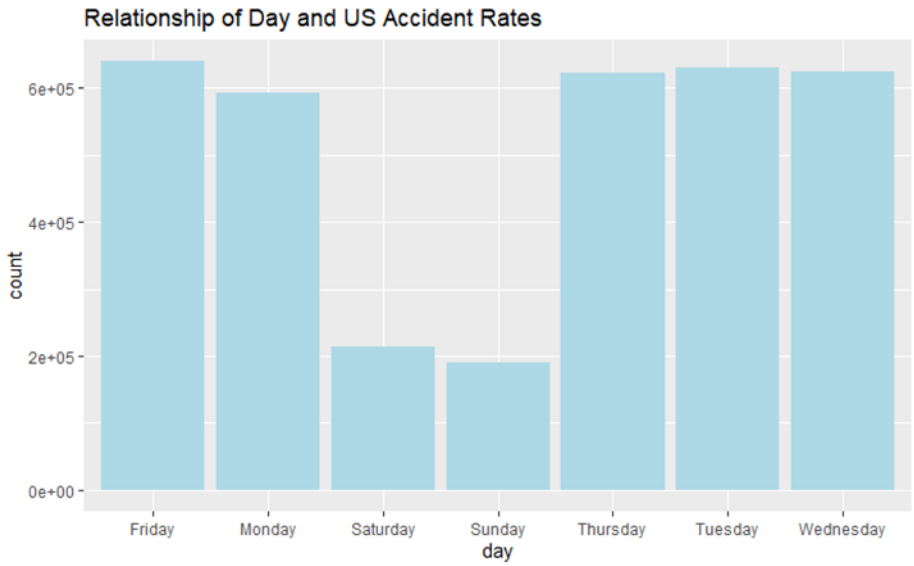
The diagram shows the relationship between months and the US accident rates. It is obvious that most of the months have the similar accident rates except that July has the lowest accident rate.

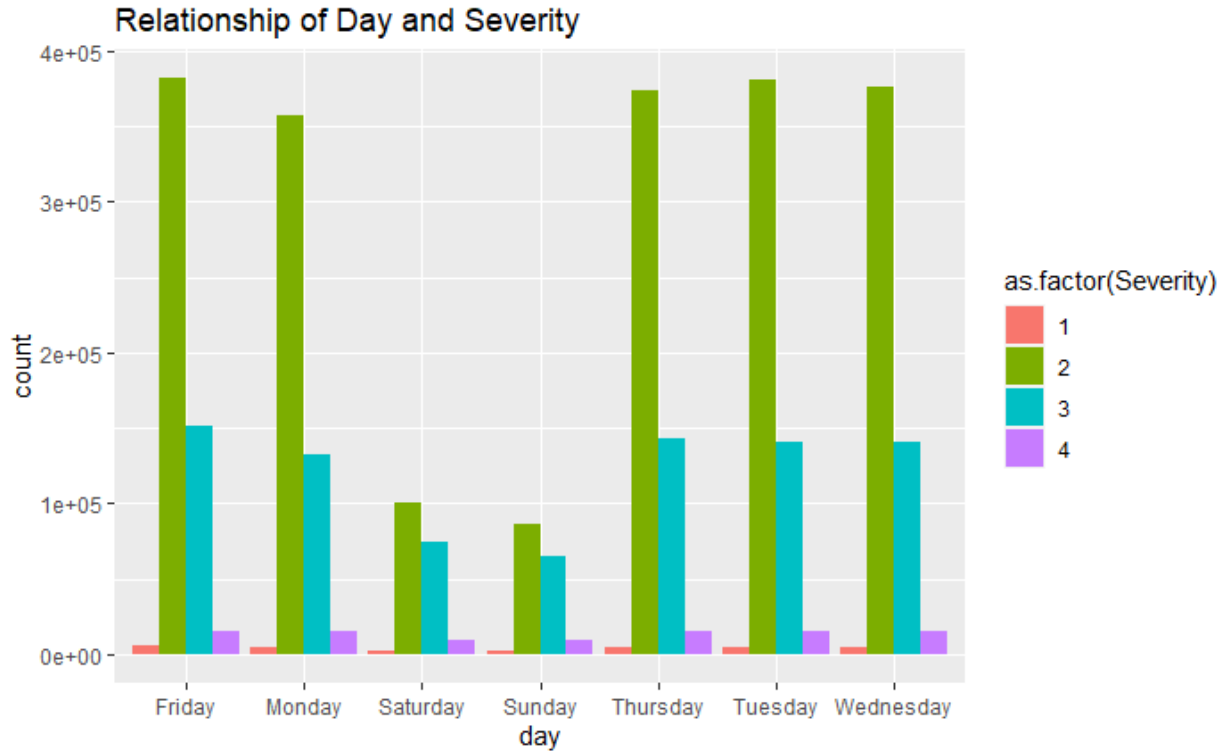



A.5.3 Day

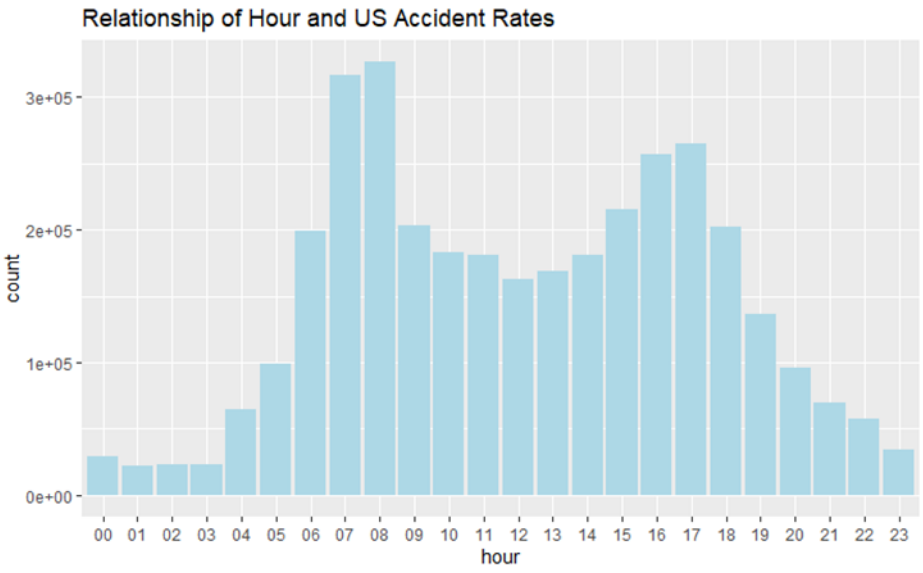
The diagram shows that the accident rates are more frequently happened in the weekdays. There are much lesser car accidents happen on Saturday and Sunday.

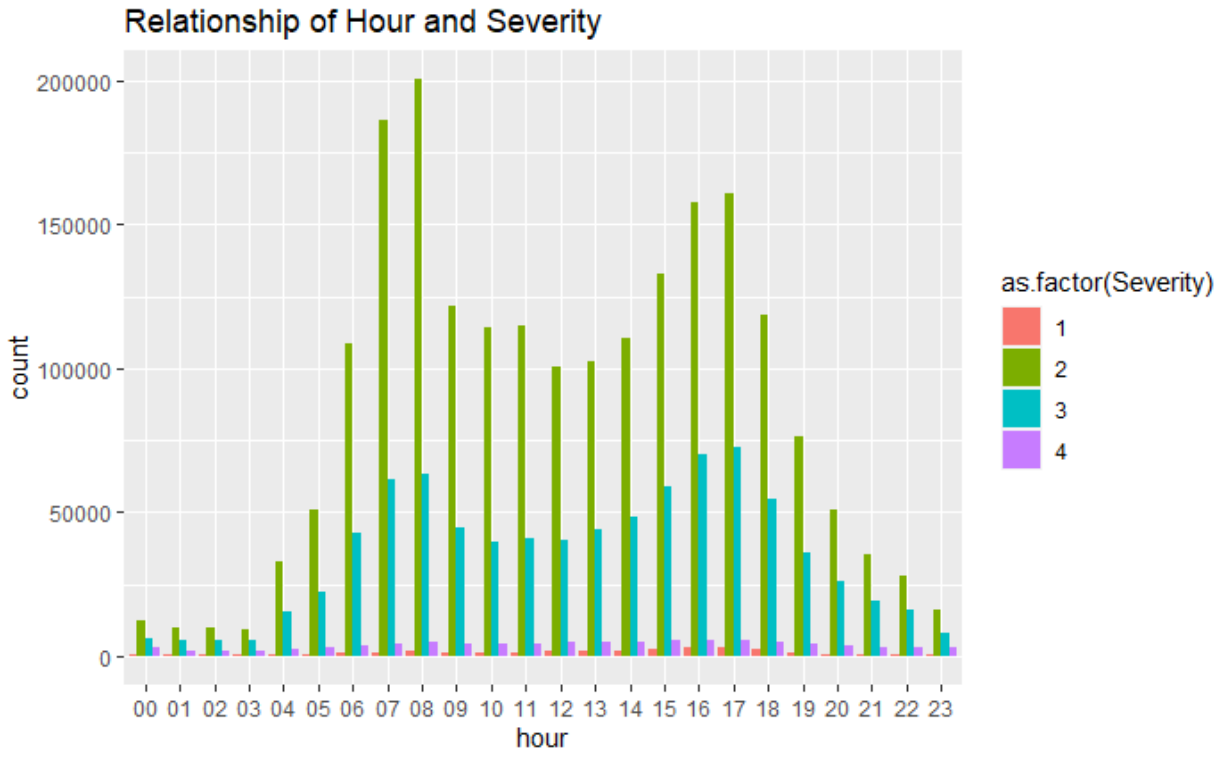



A.5.4 Hour

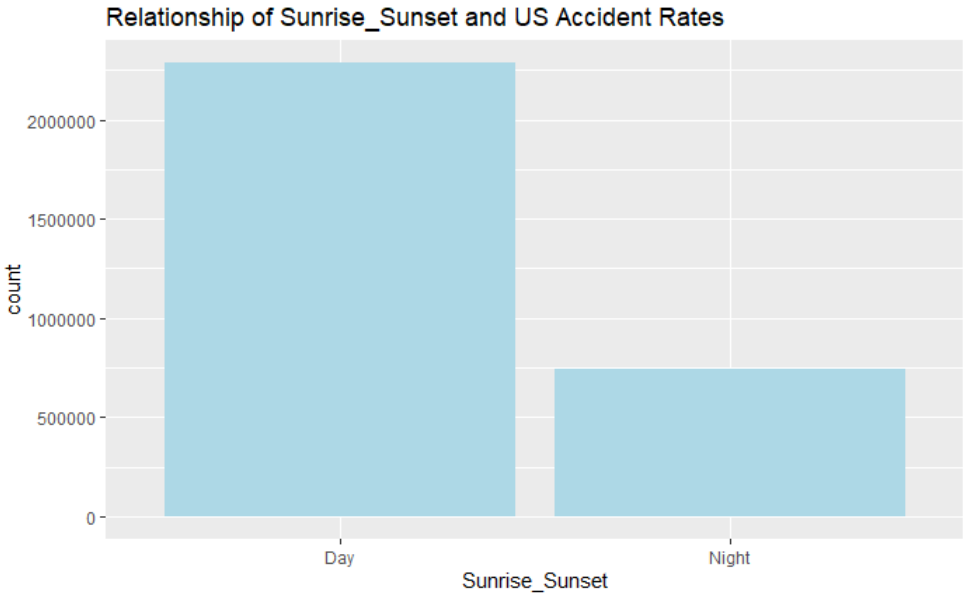
The diagram shows that car accidents are more frequently happen during the day. Perhaps we can say that people are not having many activities during the night. Also, there are more car accidents happening in around 7-9 am and 4-6pm period. This makes intuitive sense because these are the time people travel to work and people travel back to home.

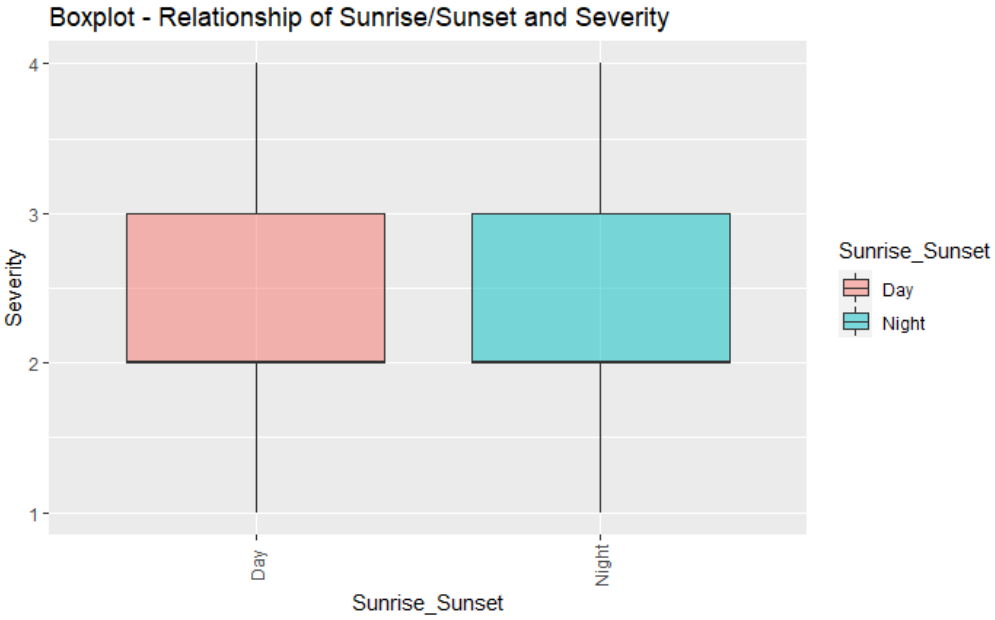



A.5.5 Sunrise_Sunset

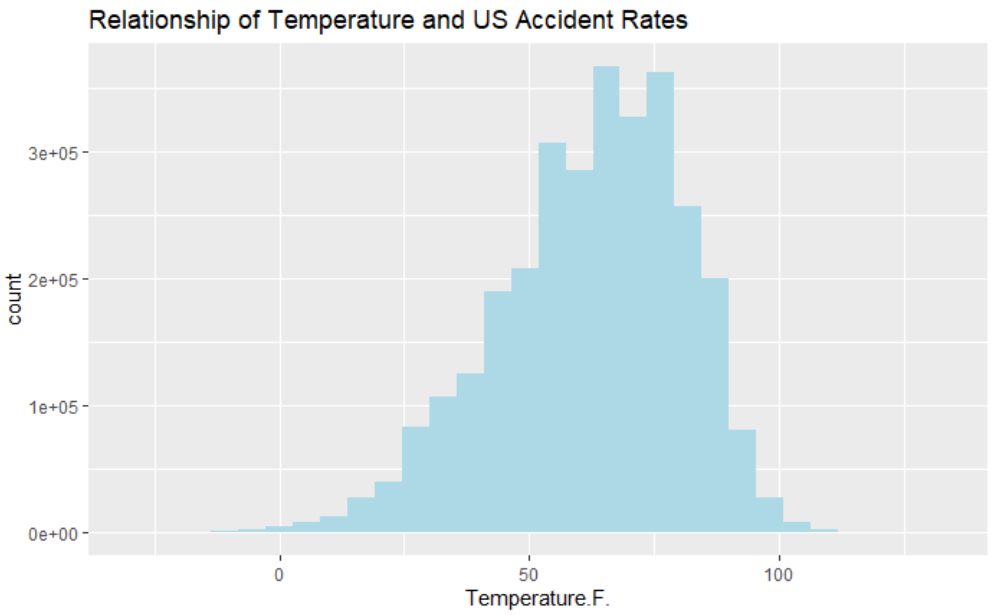
As shown in the graph, the day has more accidents than the night. This is aligning with the result from previous section. Also, the boxplot shown that the accident severity is generally same regardless of the accidents happened in day or at night.

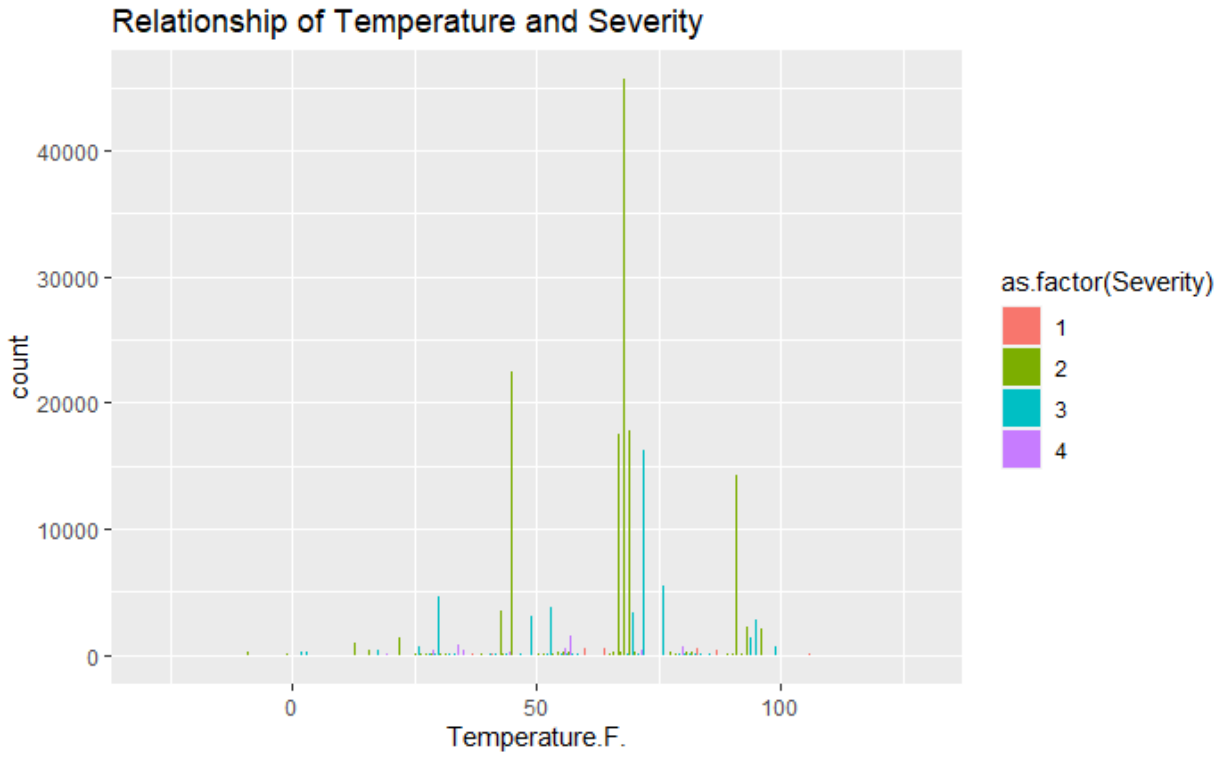


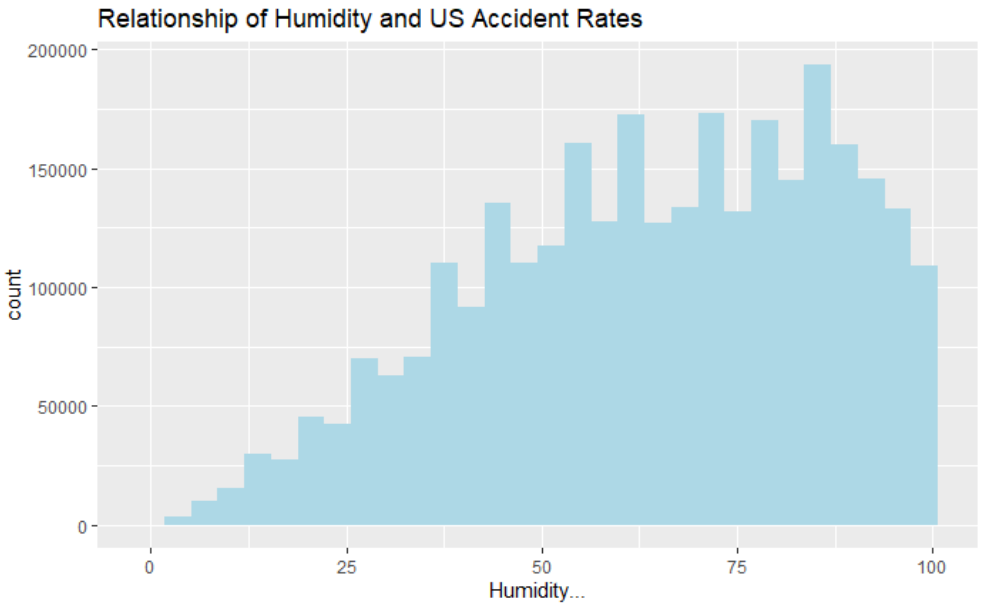

B. Additional Data Analysis (Weather condition variable)

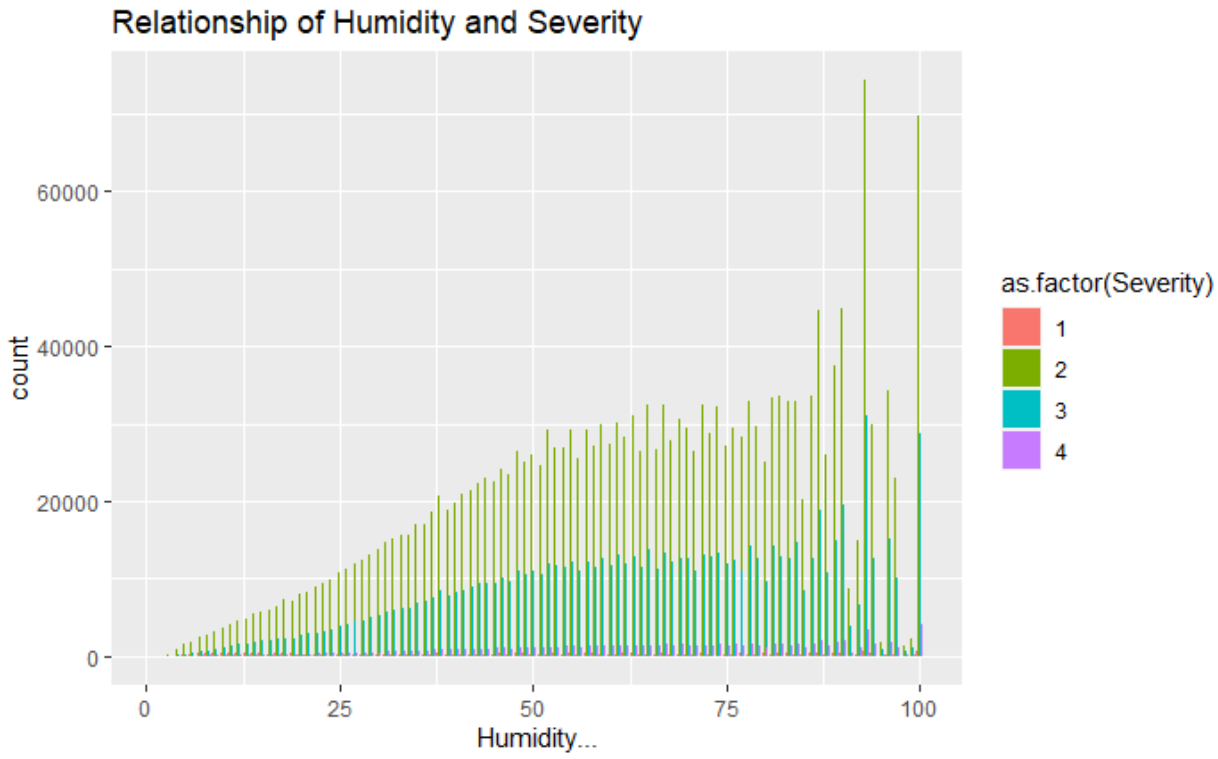
B.1 Temperature


Relationship of Temperature and US Accident Rates
count
3e+05
2e+05
1e+05
0e+00
0
50
100
Temperature.F.

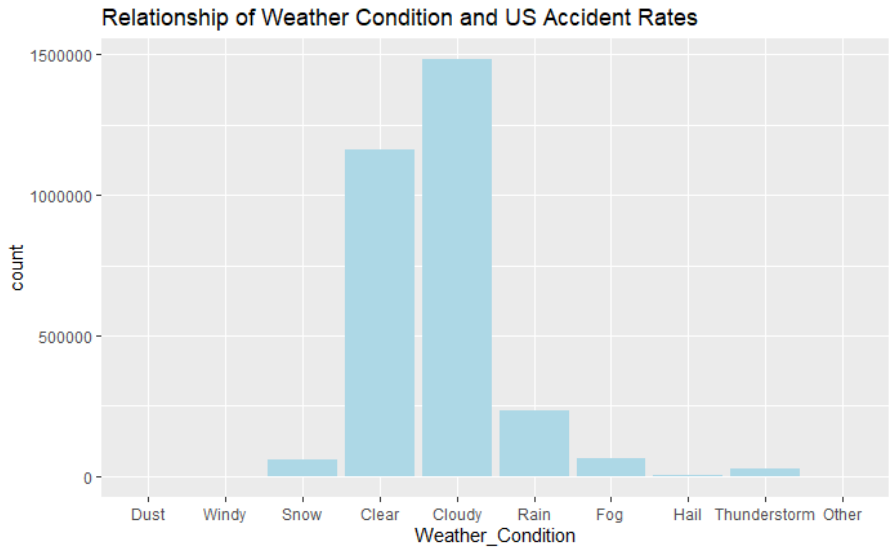

Relationship of Temperature and Severity
count
40000
30000
20000
10000
0
0
50
100
Temperature.F.
as.factor(Severity)
1
2
3
4

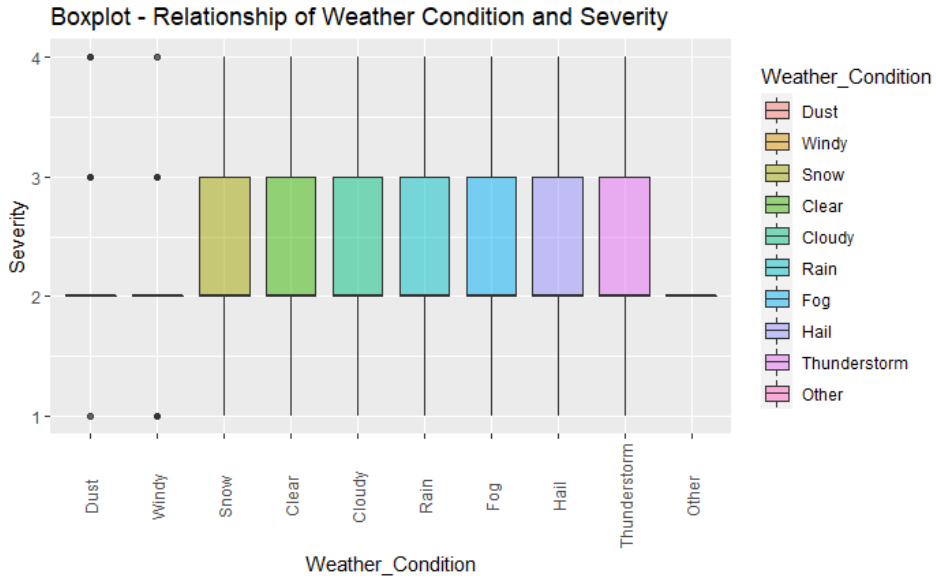
B.2 Humidity



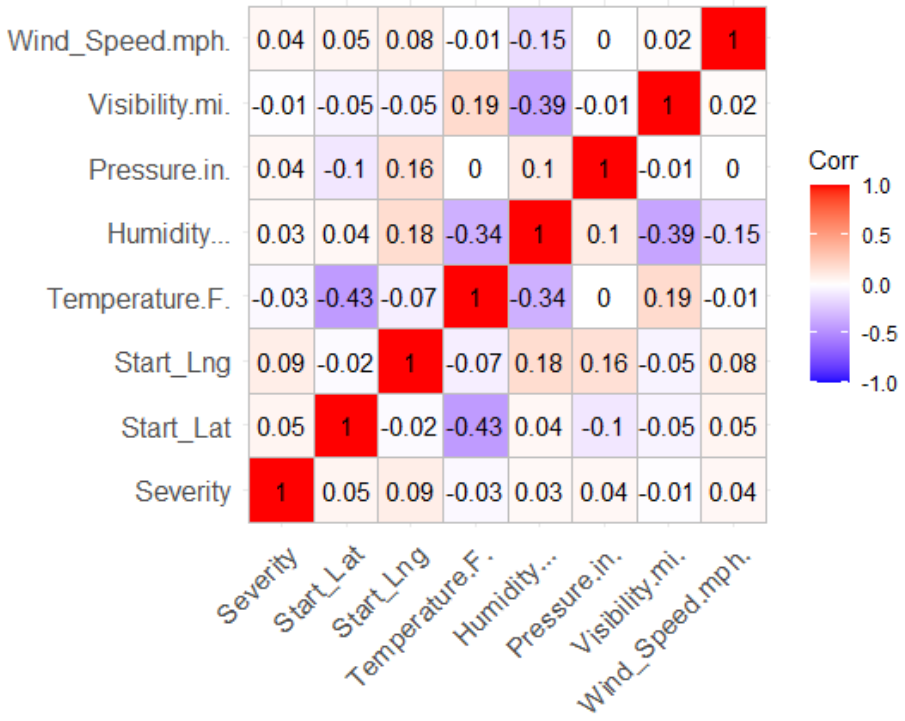



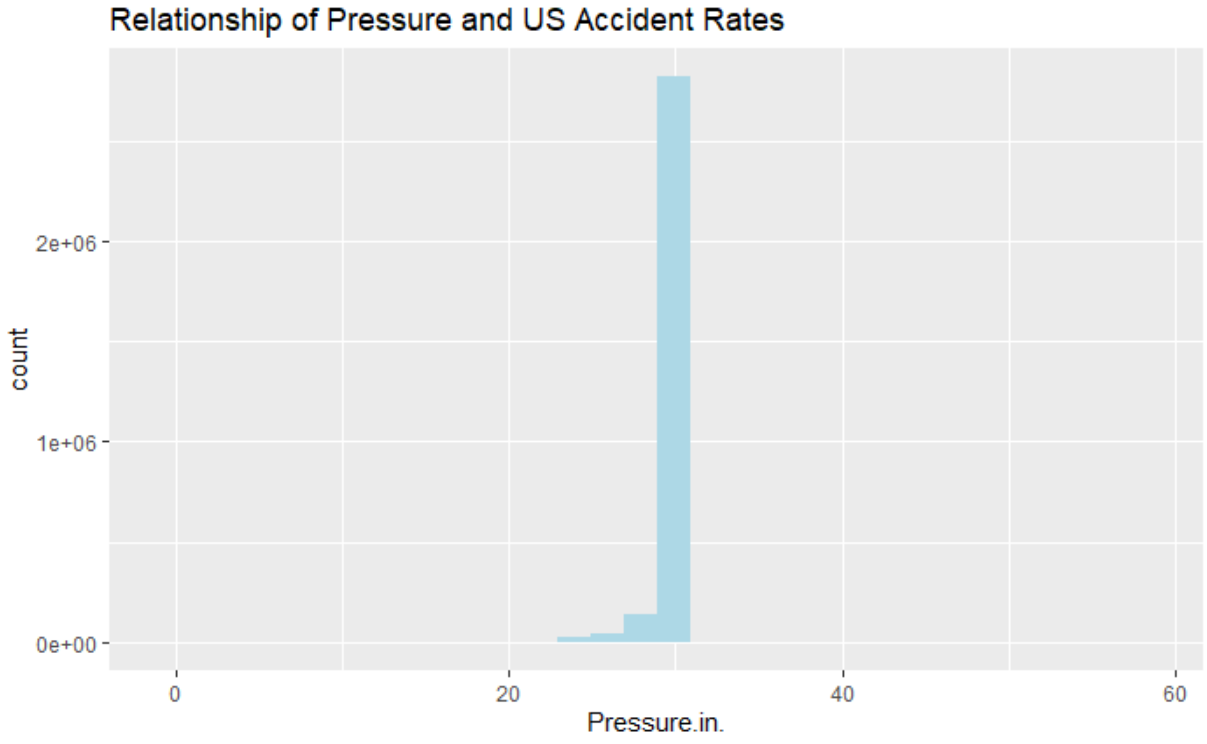

B.3 Weather condition

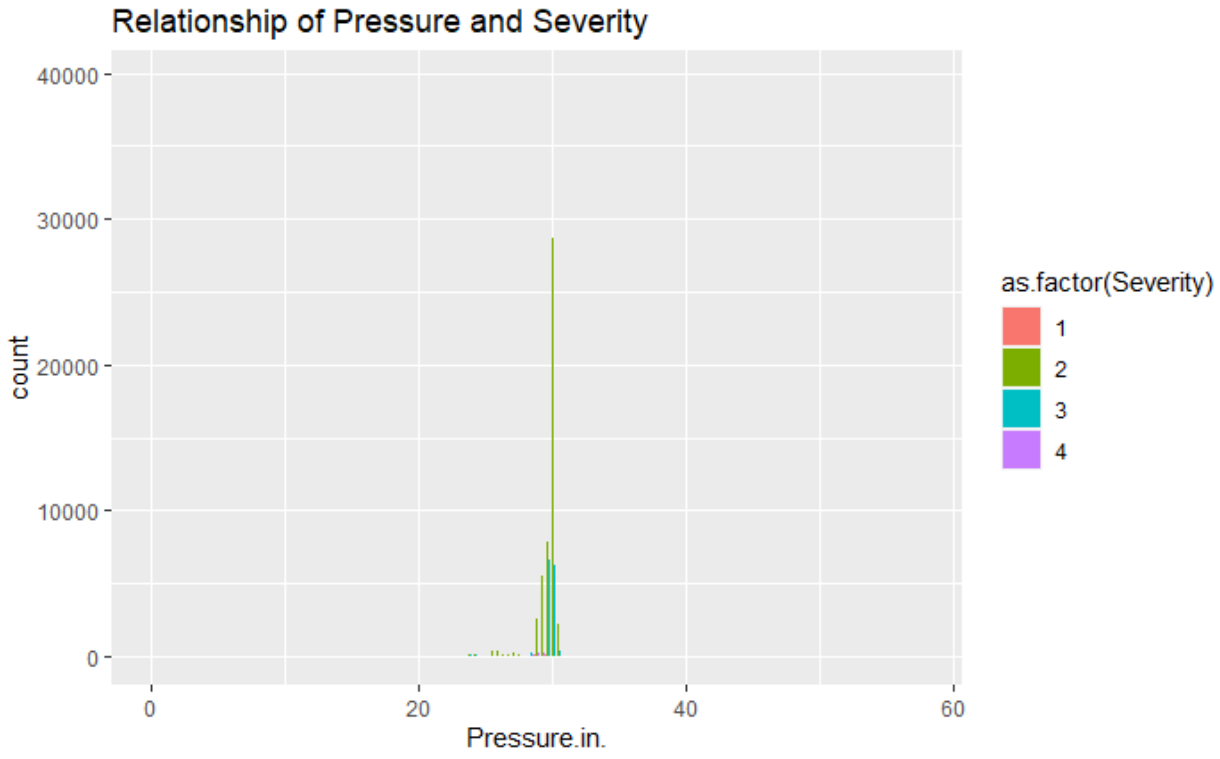


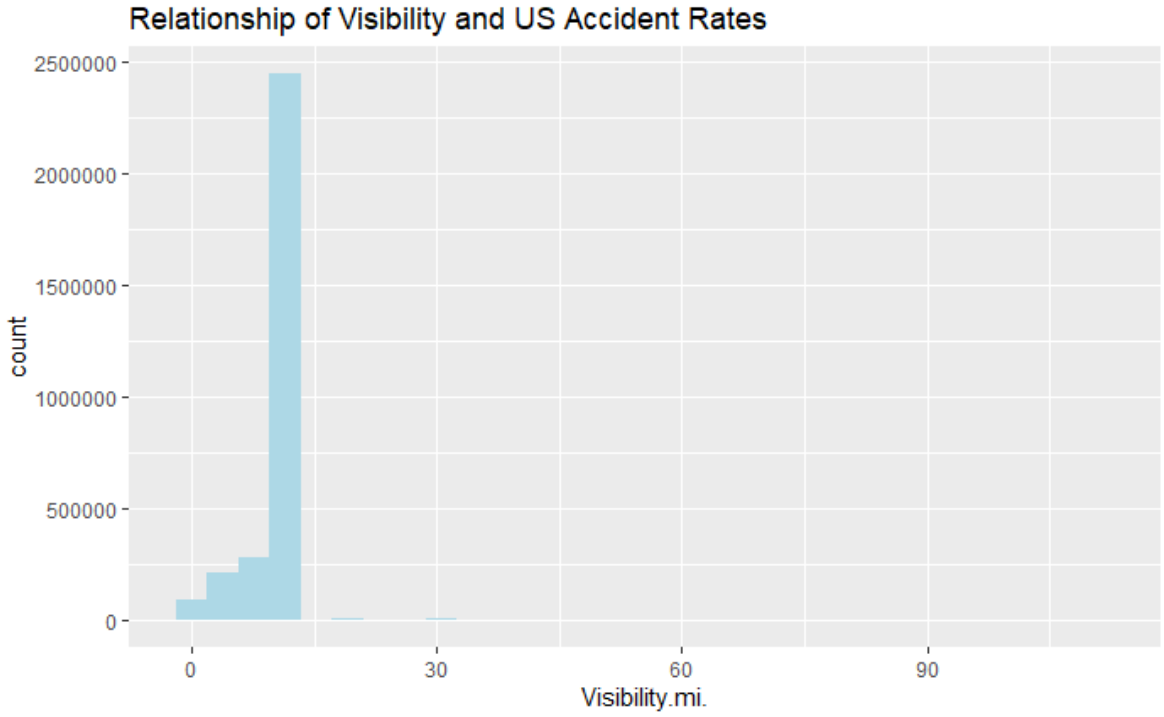



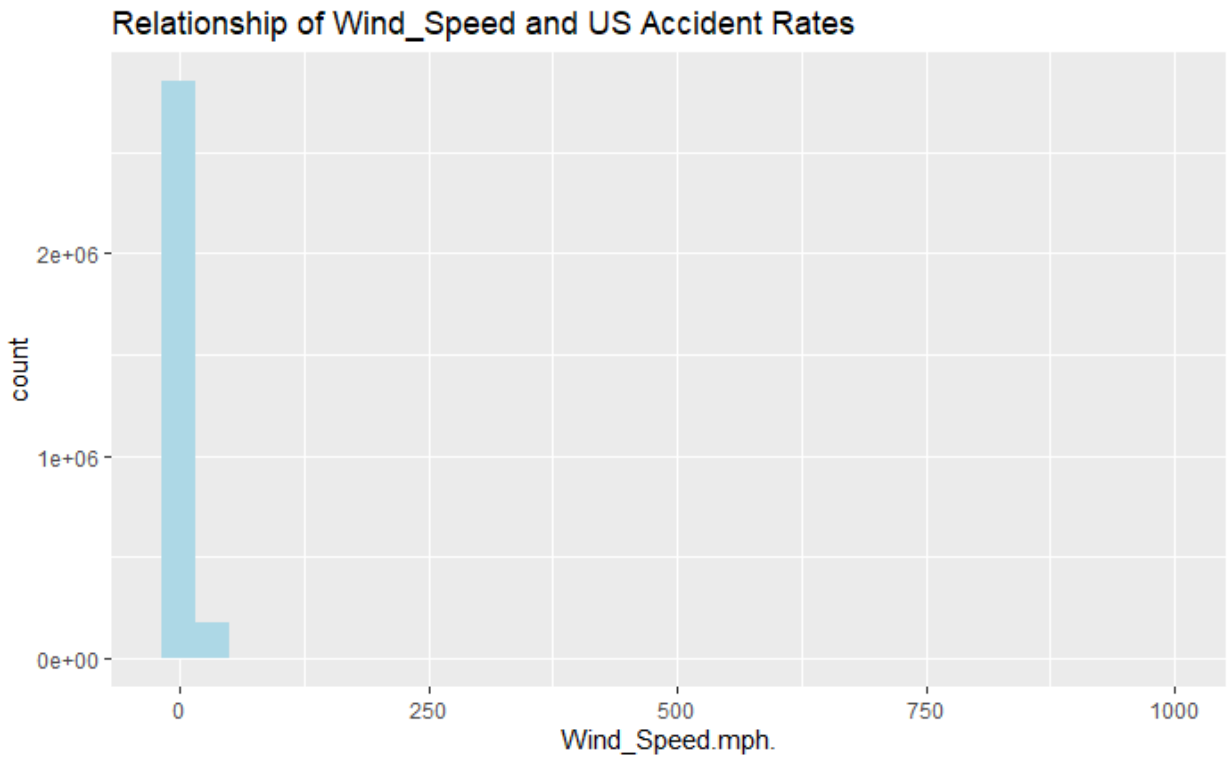

B.4 Other weather condition variables


Wind_Speed.mph. 0.04 0.05 0.08 -0.01 -0.15 0 0.02 1
Visibility.mi. -0.01 -0.05 -0.05 0.19 -0.39 -0.01 1 0.02
Pressure.in. 0.04 -0.1 0.16 0 0.1 1 -0.01 0
Humidity... 0.03 0.04 0.18 -0.34 1 0.1 -0.39 -0.15
Temperature.F. -0.03 -0.43 -0.07 1 -0.34 0 0.19 -0.01
Start_Lng 0.09 -0.02 1 -0.07 0.18 0.16 -0.05 0.08
Start_Lat 0.05 1 -0.02 -0.43 0.04 -0.1 -0.05 0.05
Severity 1 0.05 0.09 -0.03 0.03 0.04 -0.01 0.04
Severity
Start_Lat
Start_Lng
Temperature.F.
Humidity...
Pressure.in.
Visibility.mi.
Wind_Speed.mph.
Corr
1.0
0.5
0.0
-0.5
-1.0

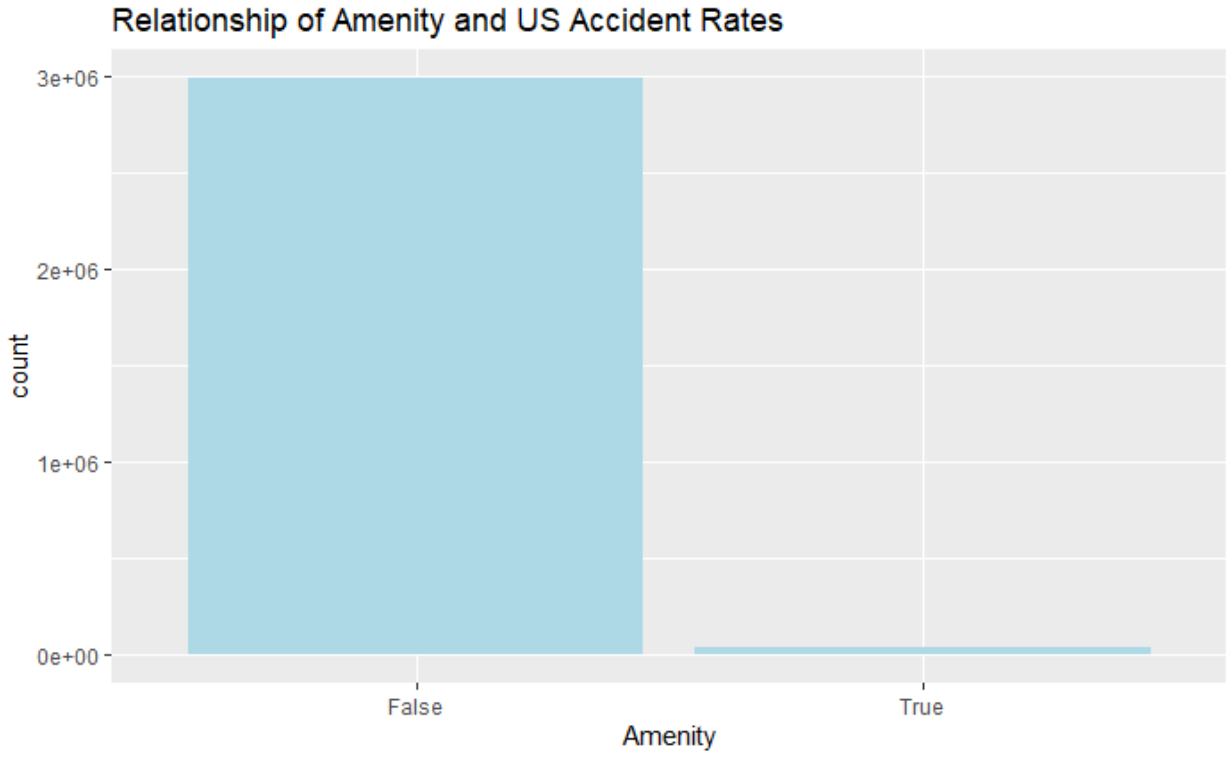

Relationship of Pressure and US Accident Rates
count
2e+06
1e+06
0e+00
0
20
40
60
Pressure.in.

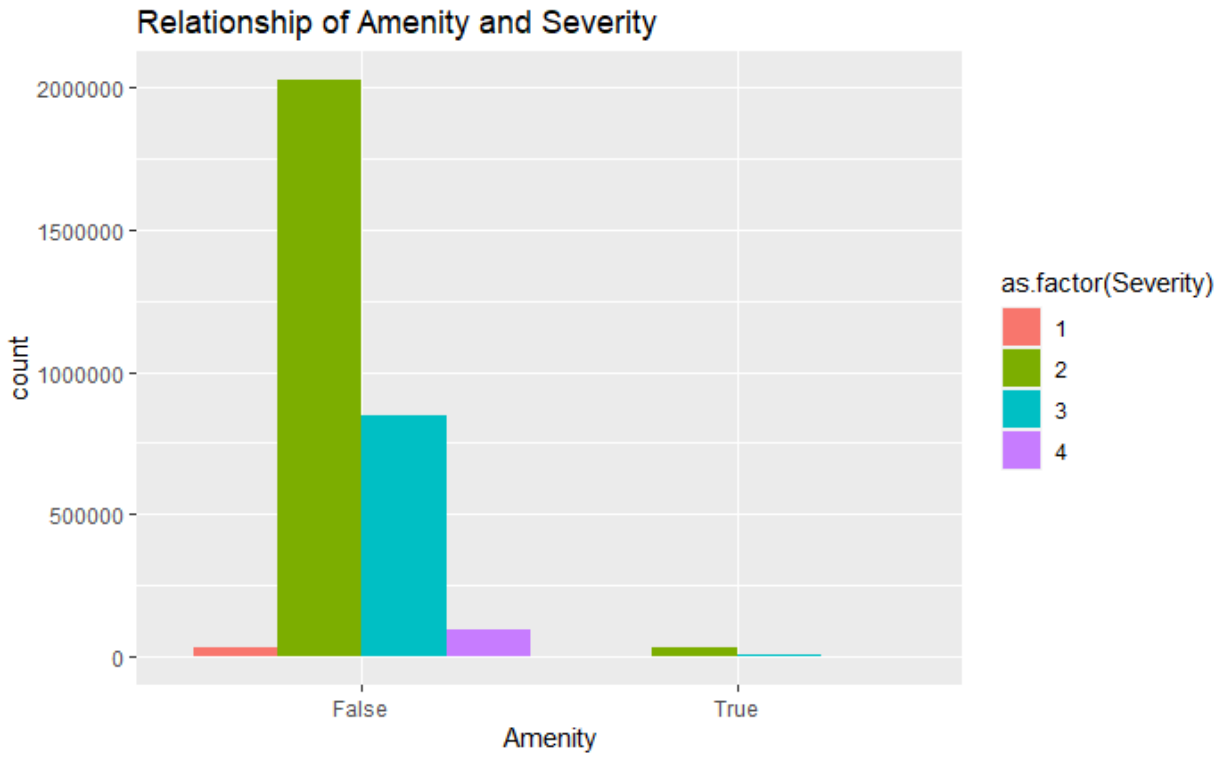

Relationship of Pressure and Severity
40000
30000
20000
10000
0
count
0
20
40
60
Pressure.in.
as.factor(Severity)
1
2
3
4

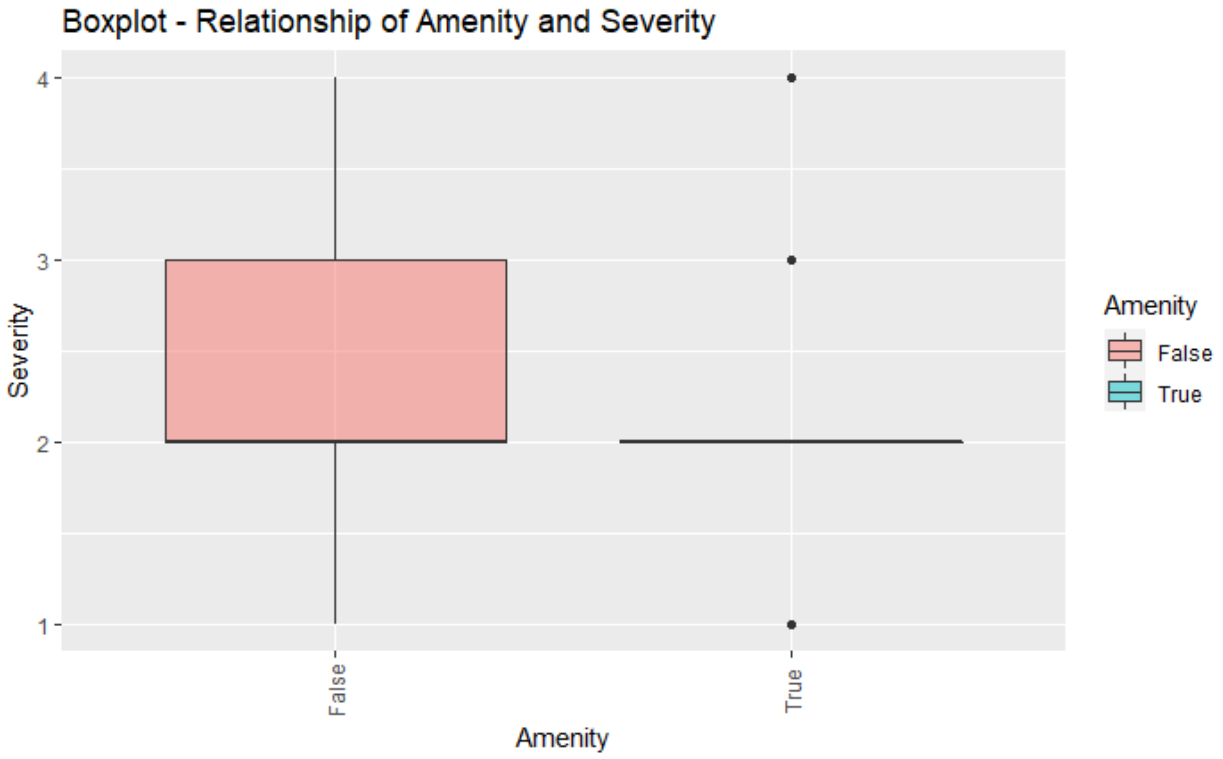

Relationship of Visibility and US Accident Rates
2500000
2000000
1500000
1000000
500000
0
count
0
30
60
90
Visibility.mi.

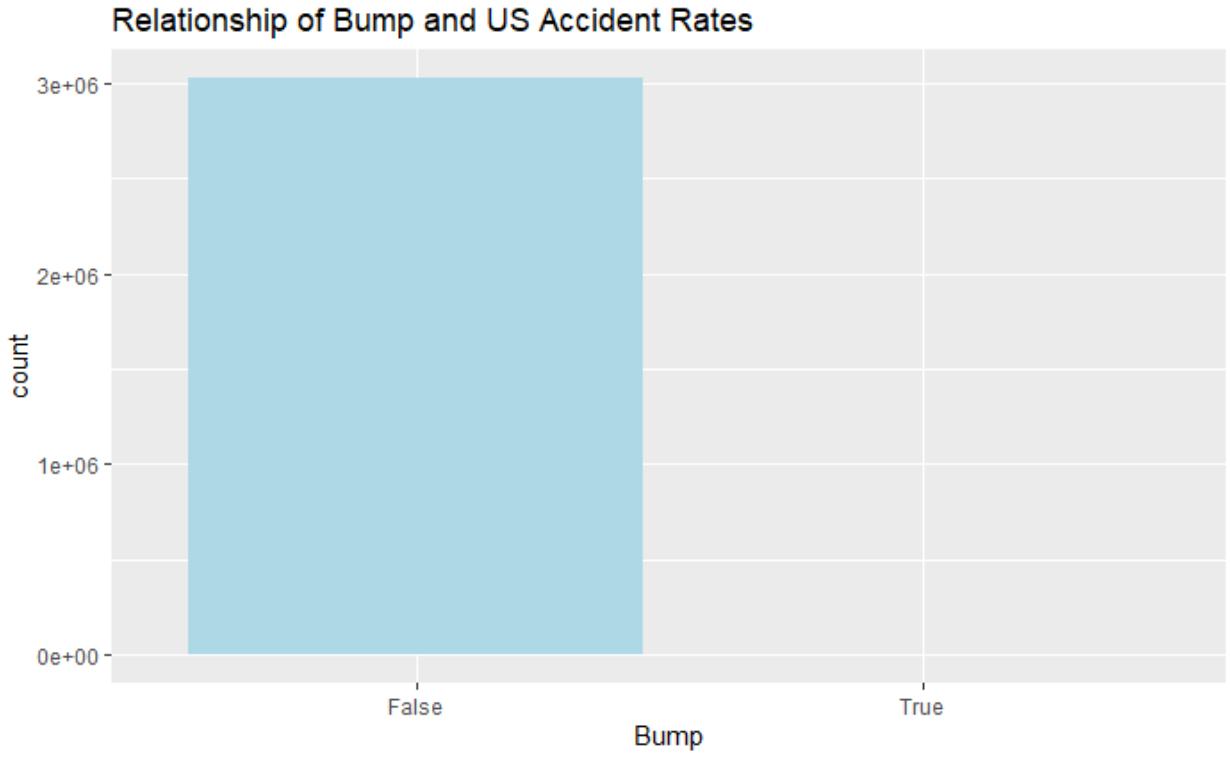

Relationship of Wind_Speed and US Accident Rates
count
2e+06
1e+06
0e+00
0
250
500
750
1000
Wind_Speed.mph.

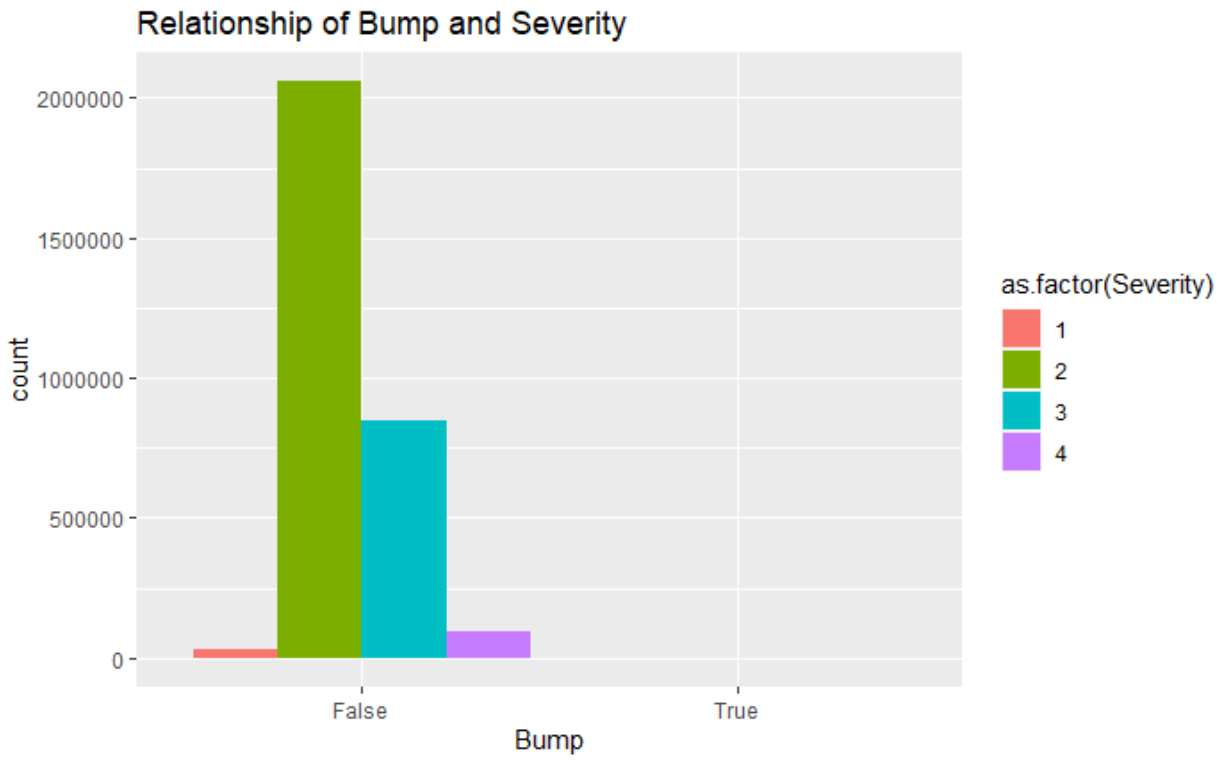
C. Additional Data Analysis (road condition variable)

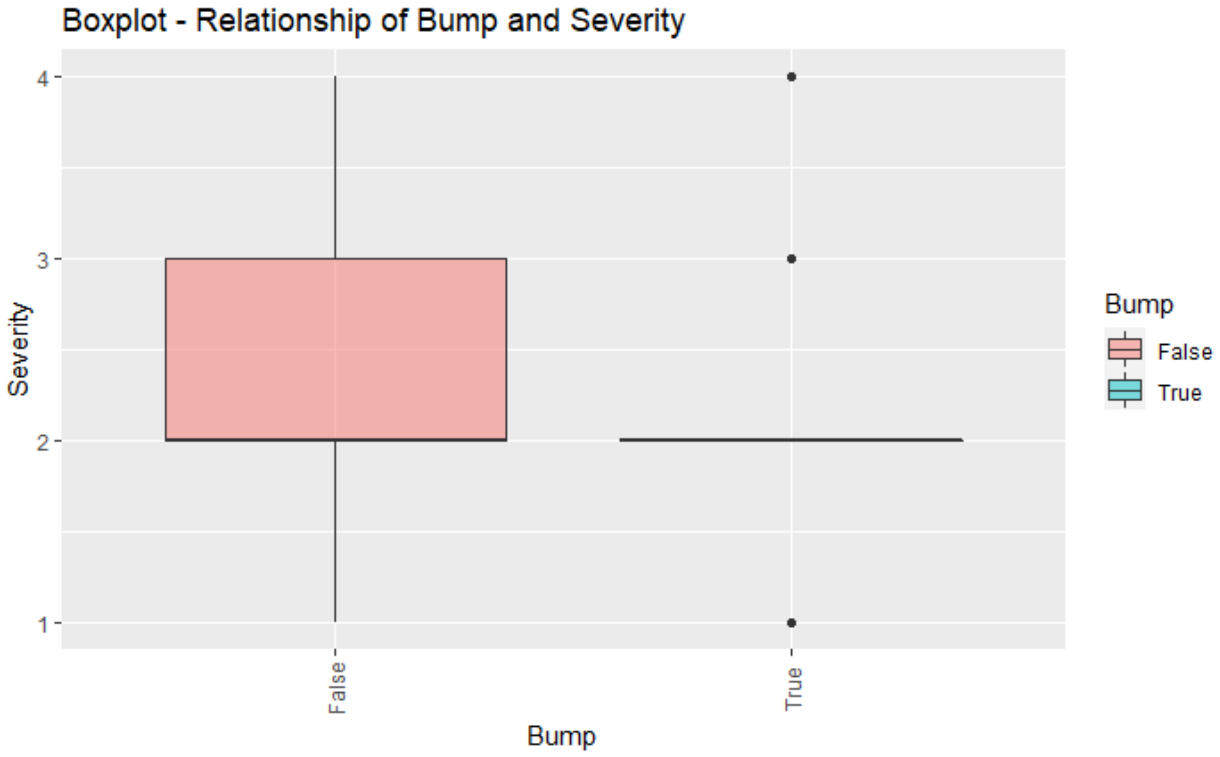


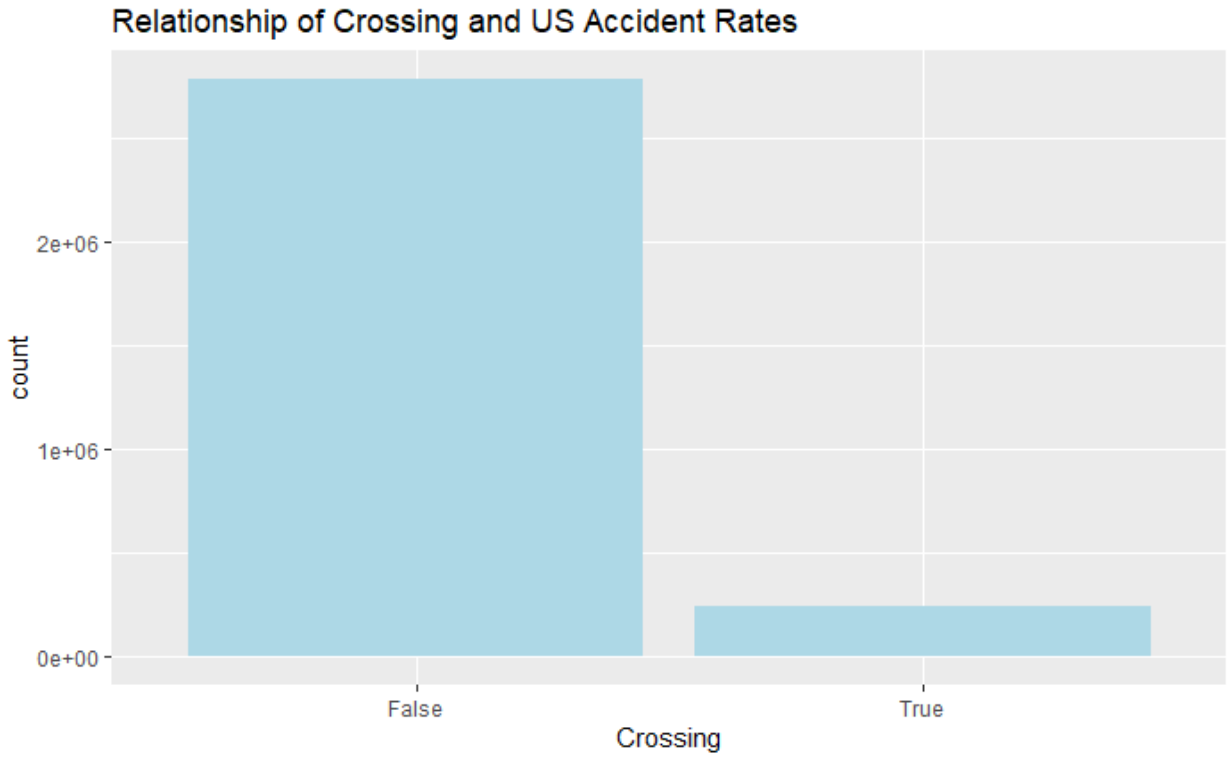



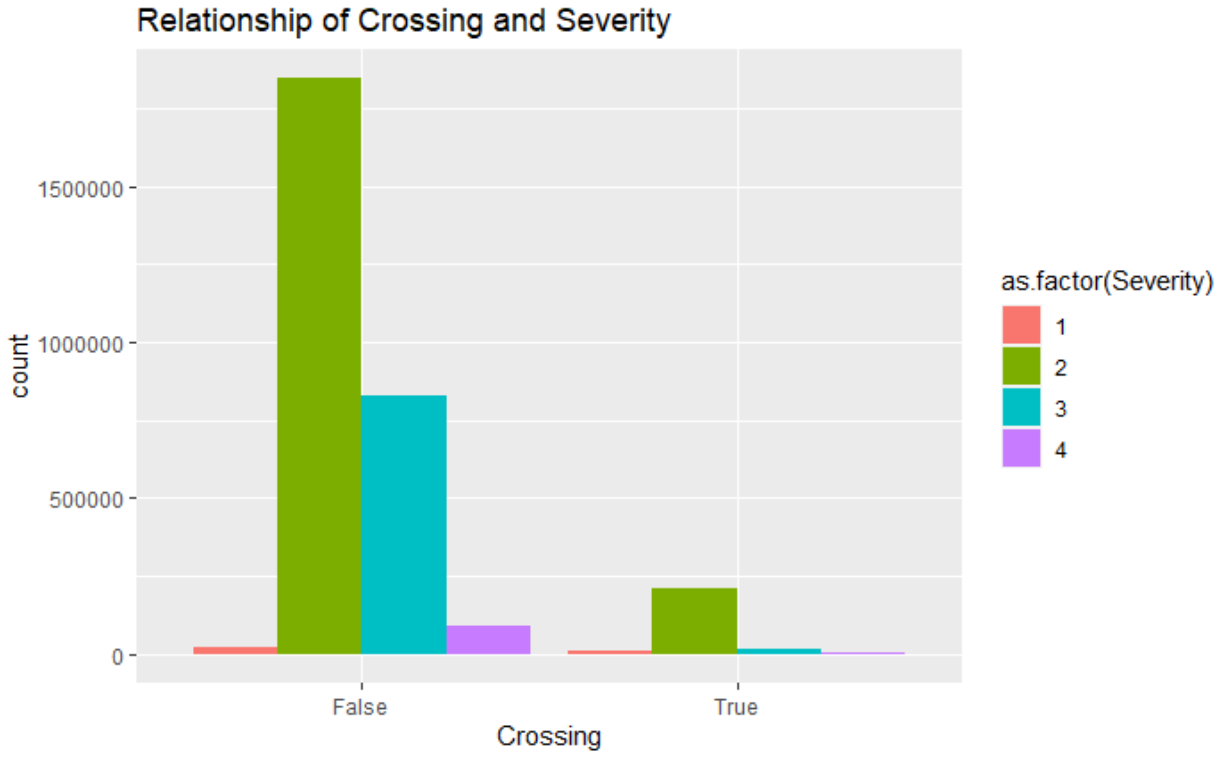

Boxplot - Relationship of Amenity and Severity
Severity
4
3
2
1
False
True
Amenity
Amenity
False
True

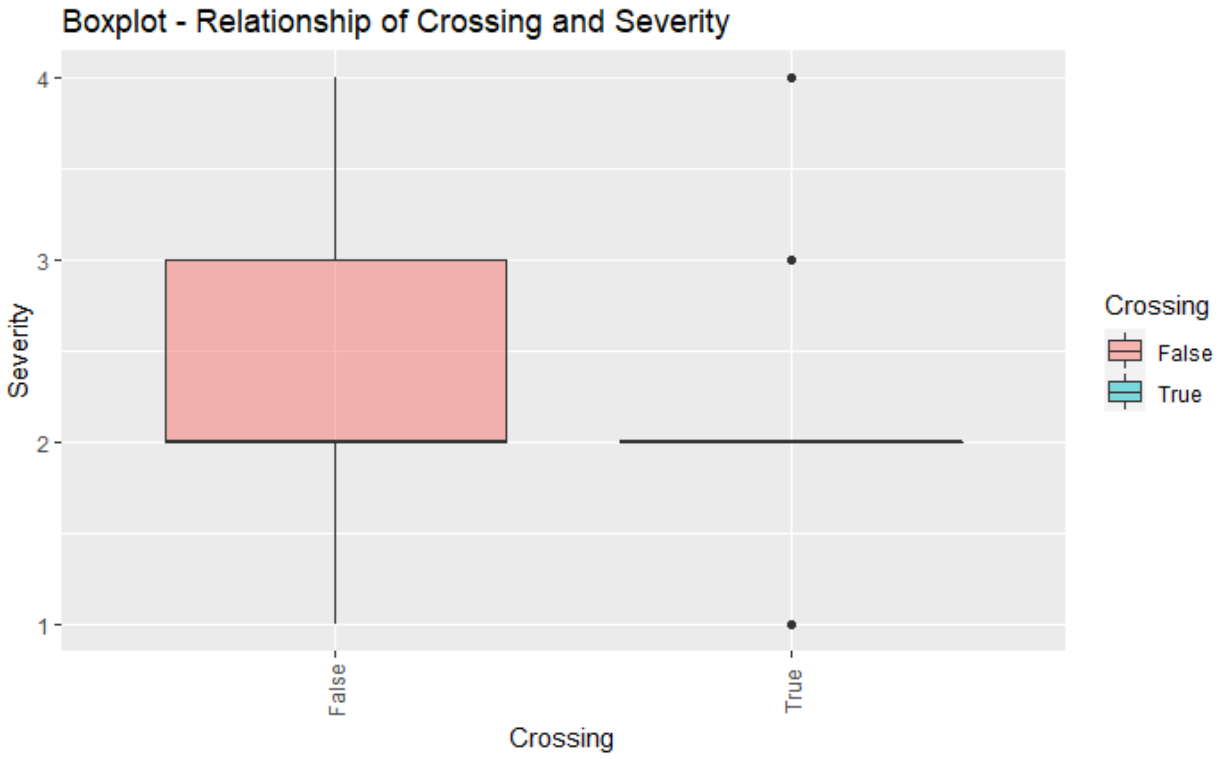

Relationship of Bump and US Accident Rates
count
3e+06
2e+06
1e+06
0e+00
False
True
Bump

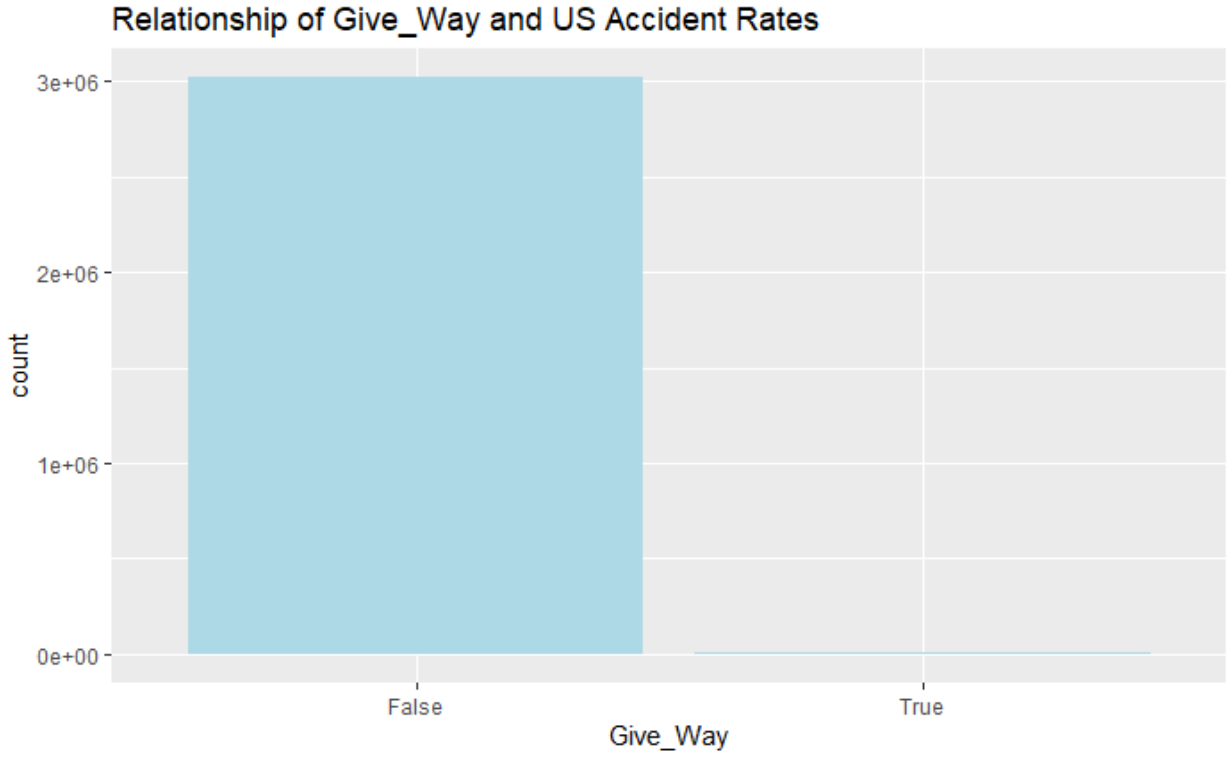

Relationship of Bump and Severity
2000000
1500000
1000000
500000
0
count
False
True
Bump
as.factor(Severity)
1
2
3
4

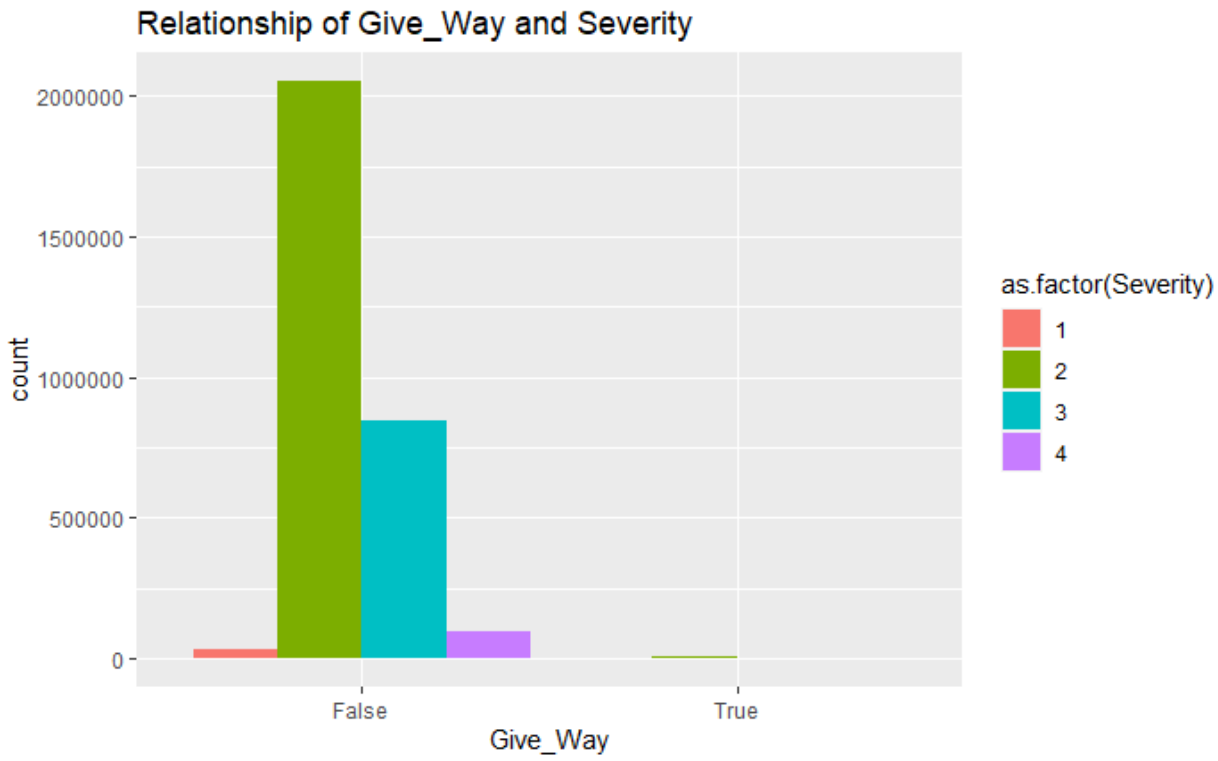

Boxplot - Relationship of Bump and Severity
4
3
2
1
Severity
False
True
Bump
Bump
False
True

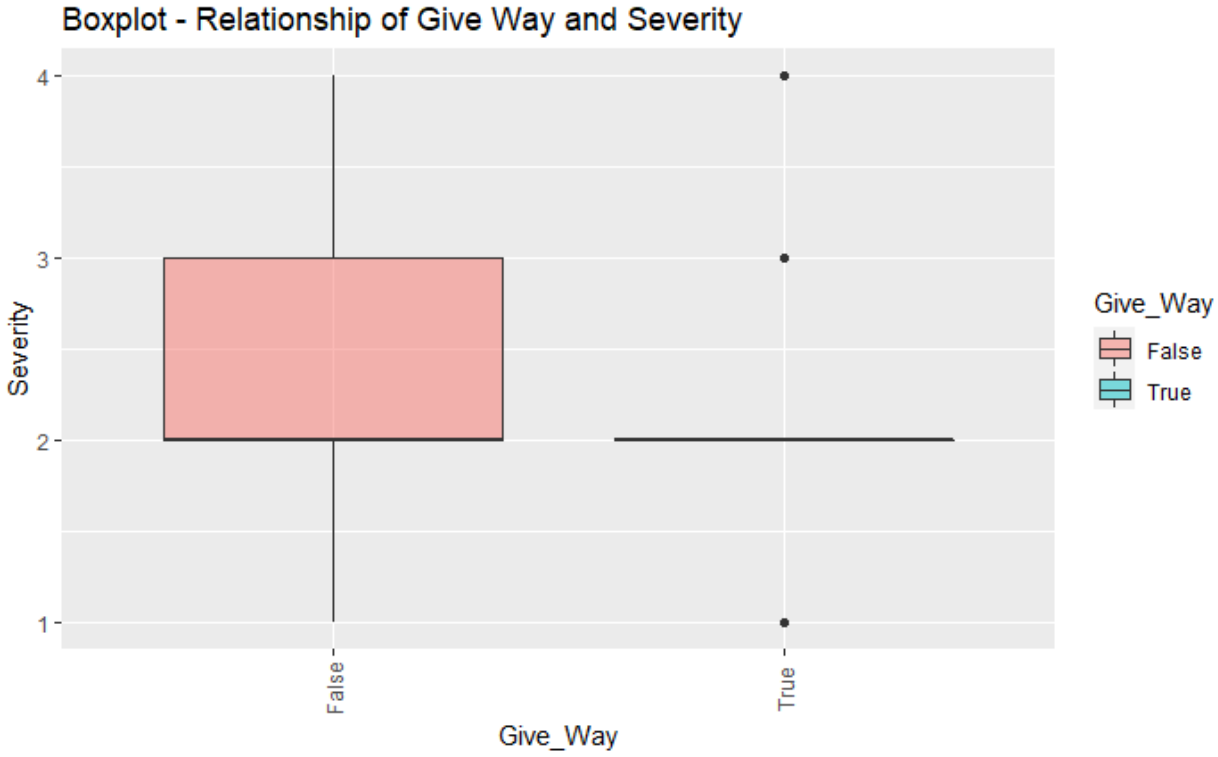

Relationship of Crossing and US Accident Rates
count
2e+06
1e+06
0e+00
False
True
Crossing

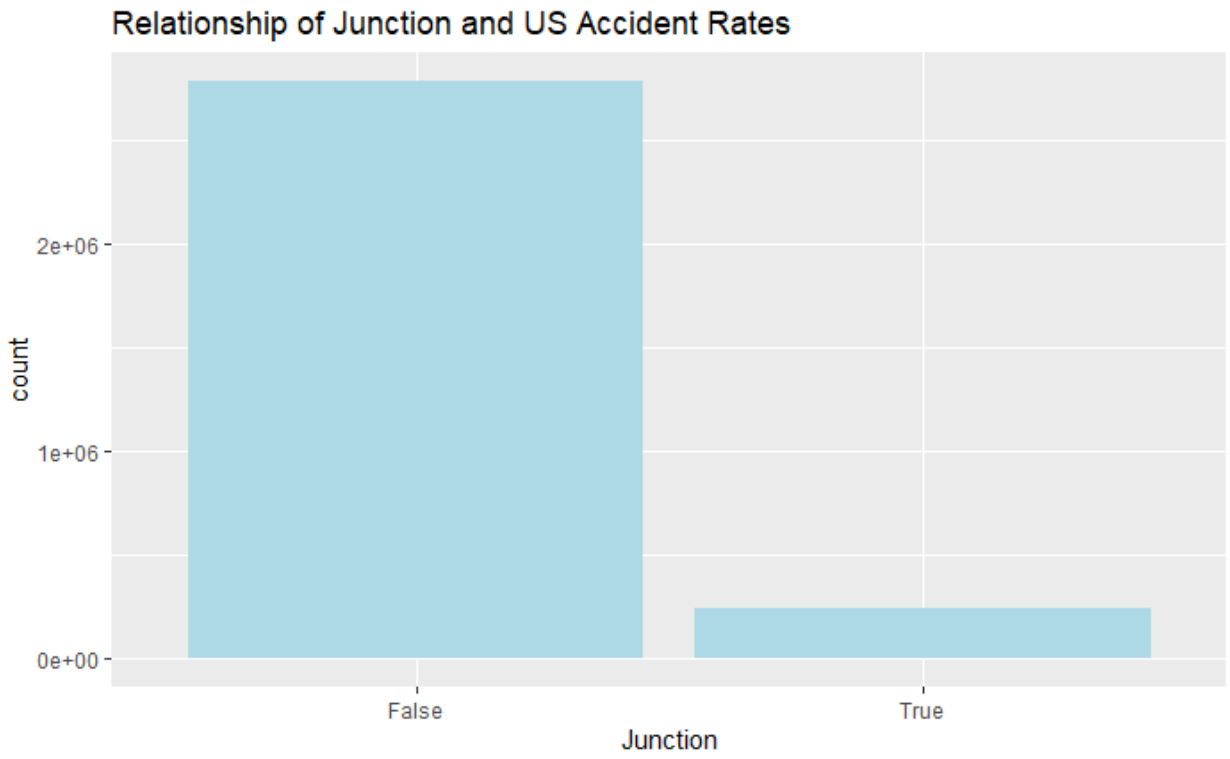

Relationship of Crossing and Severity
count
1500000
1000000
500000
0
False
True
Crossing
as.factor(Severity)
1
2
3
4

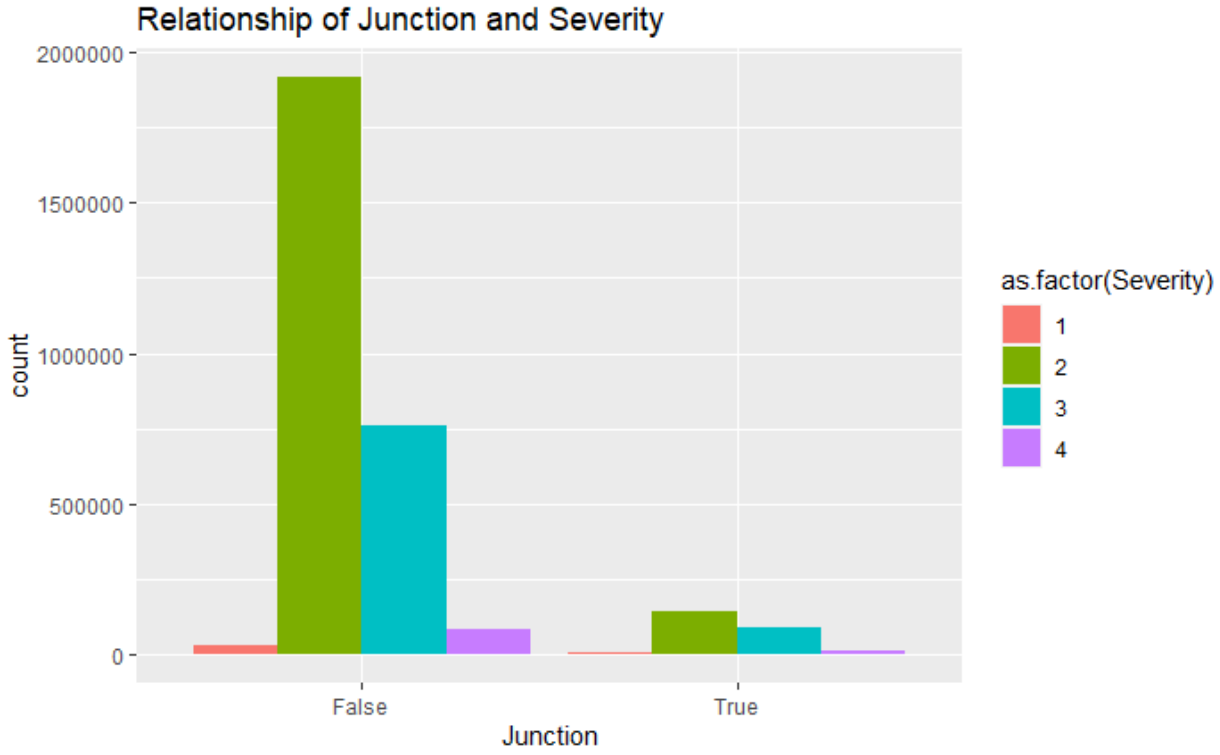

Boxplot - Relationship of Crossing and Severity
Severity
4
3
2
1
False
True
Crossing
Crossing
False
True

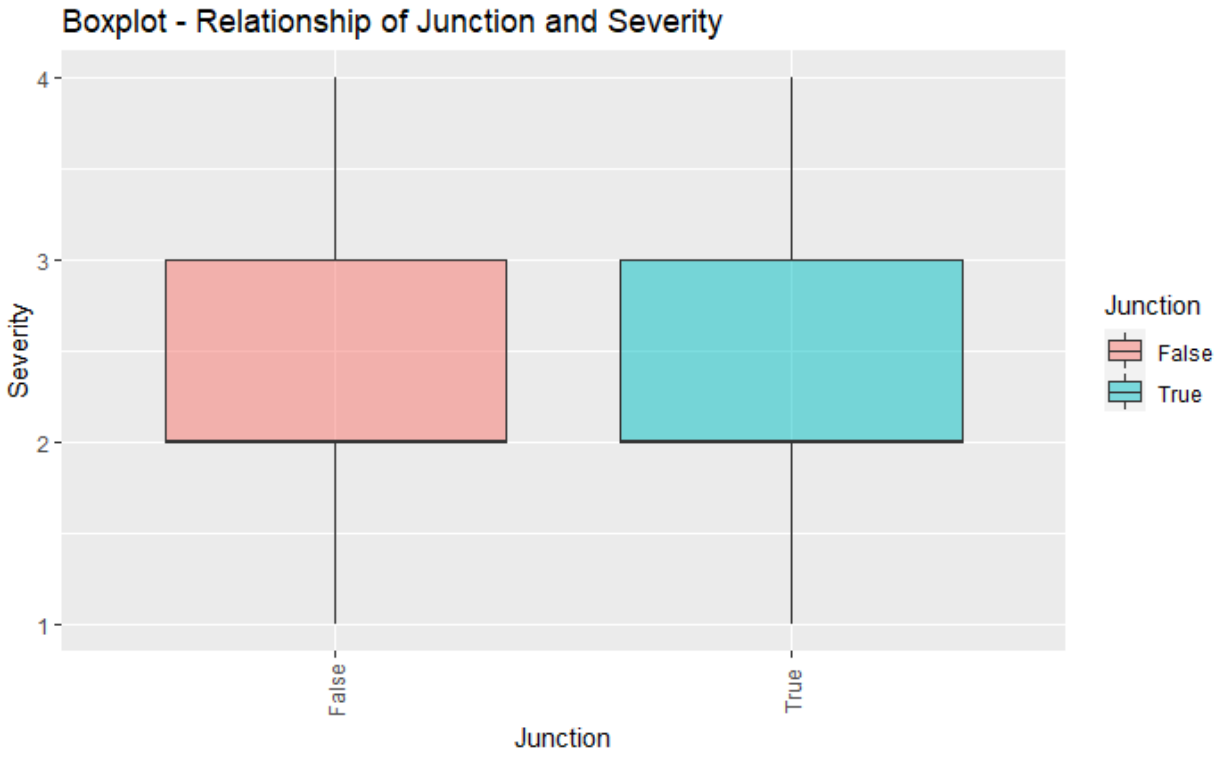

Relationship of Give_Way and US Accident Rates
count
3e+06
2e+06
1e+06
0e+00
False
True
Give_Way

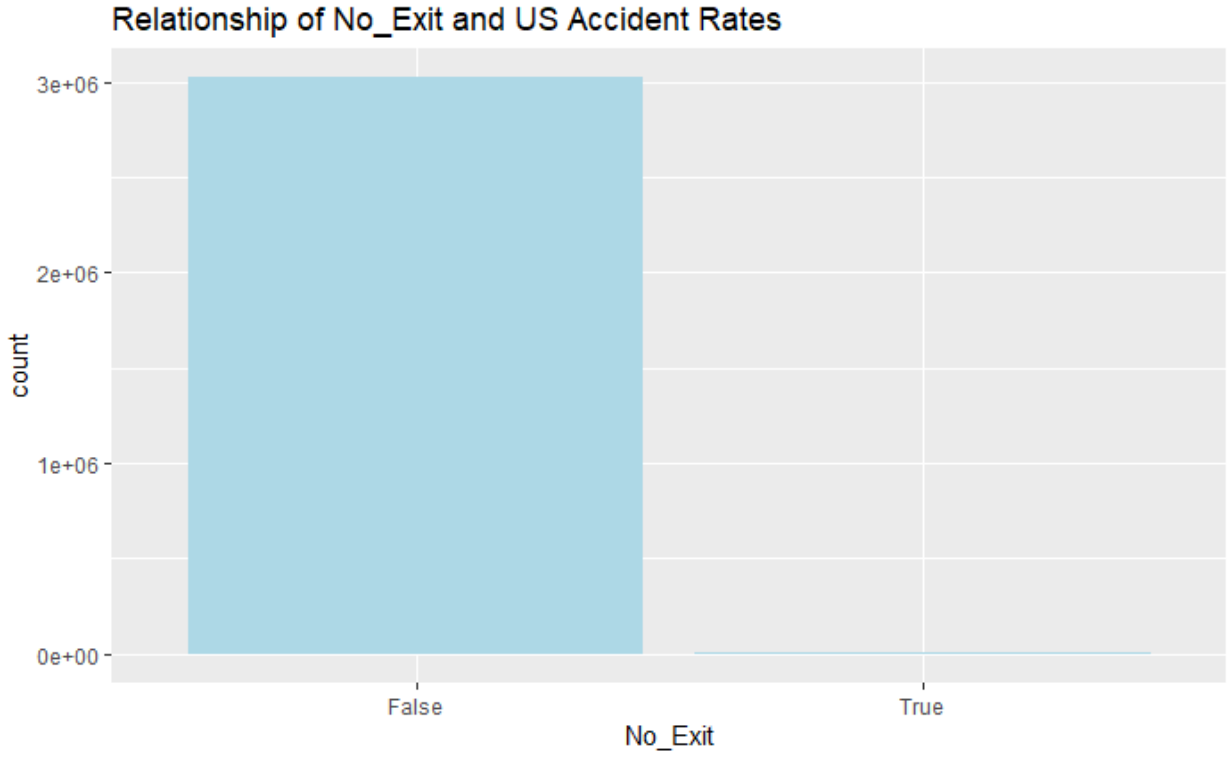

Relationship of Give_Way and Severity
2000000
1500000
1000000
500000
0
count
False
True
Give_Way
as.factor(Severity)
1
2
3
4

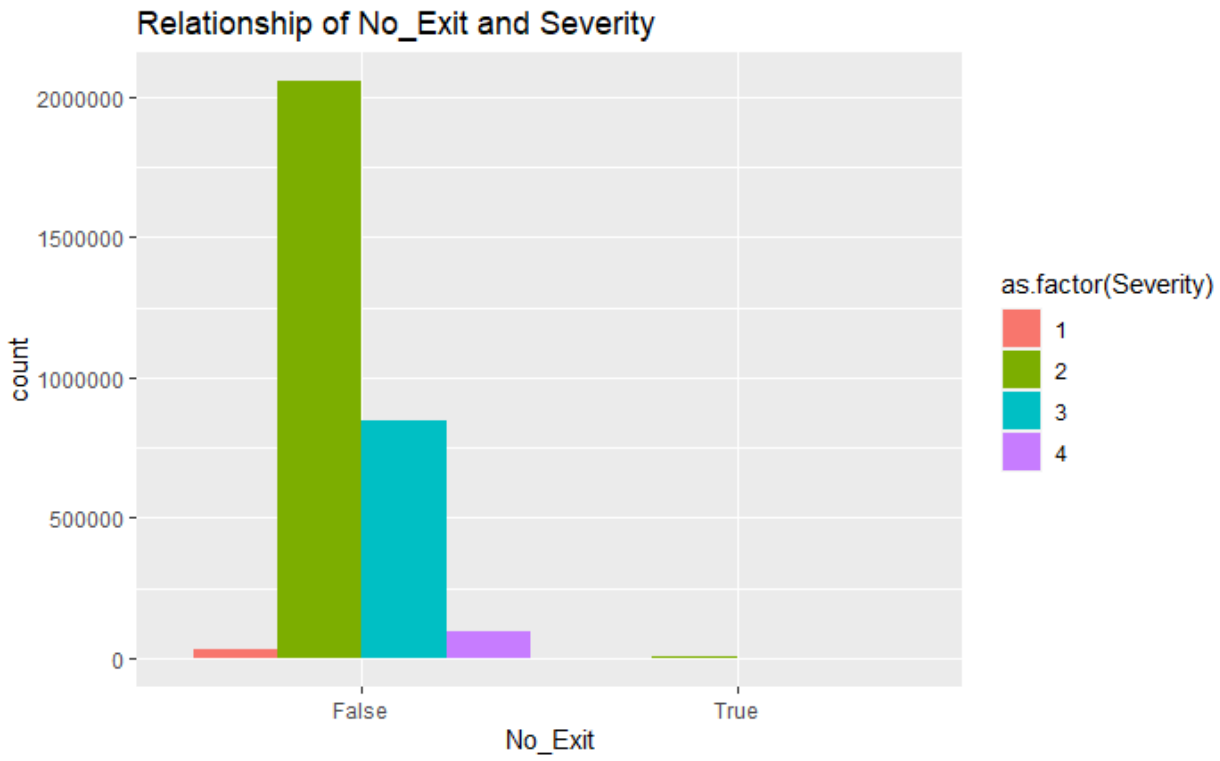

Boxplot - Relationship of Give Way and Severity
4
3
2
1
Severity
False
True
Give_Way
Give_Way
False
True

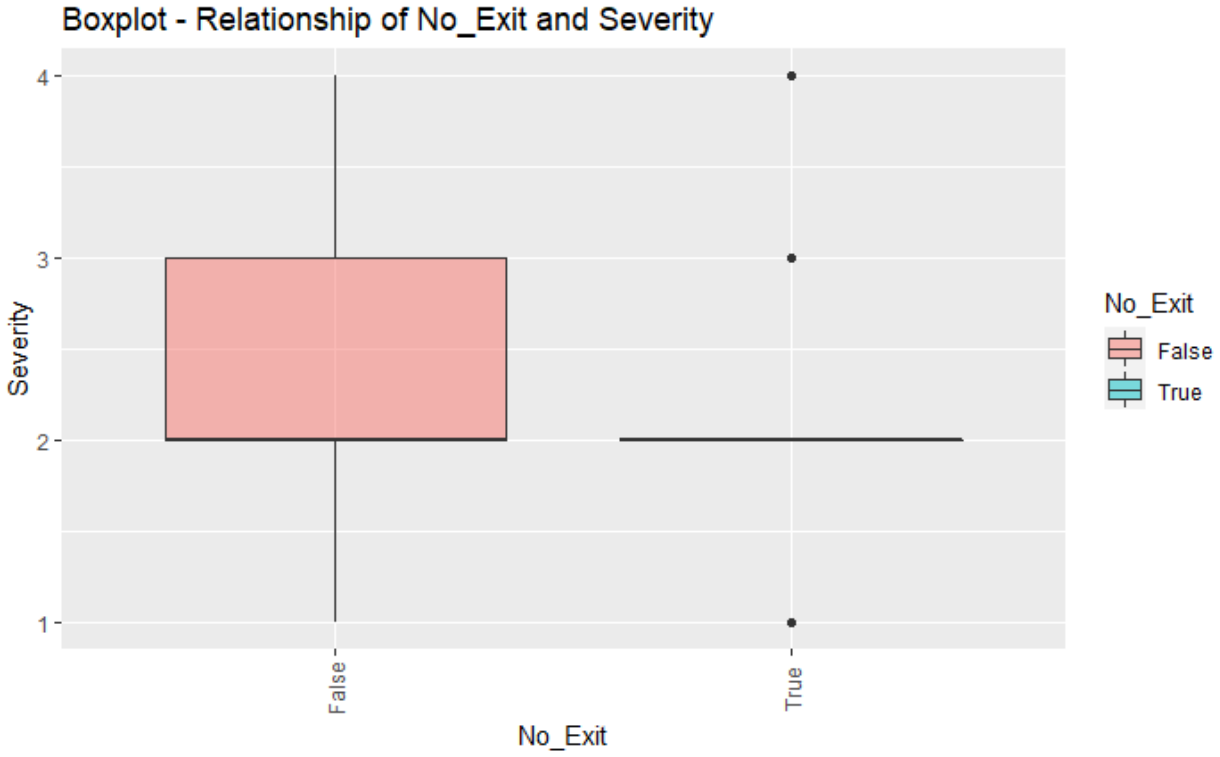

Relationship of Junction and US Accident Rates
count
2e+06
1e+06
0e+00
False
True
Junction

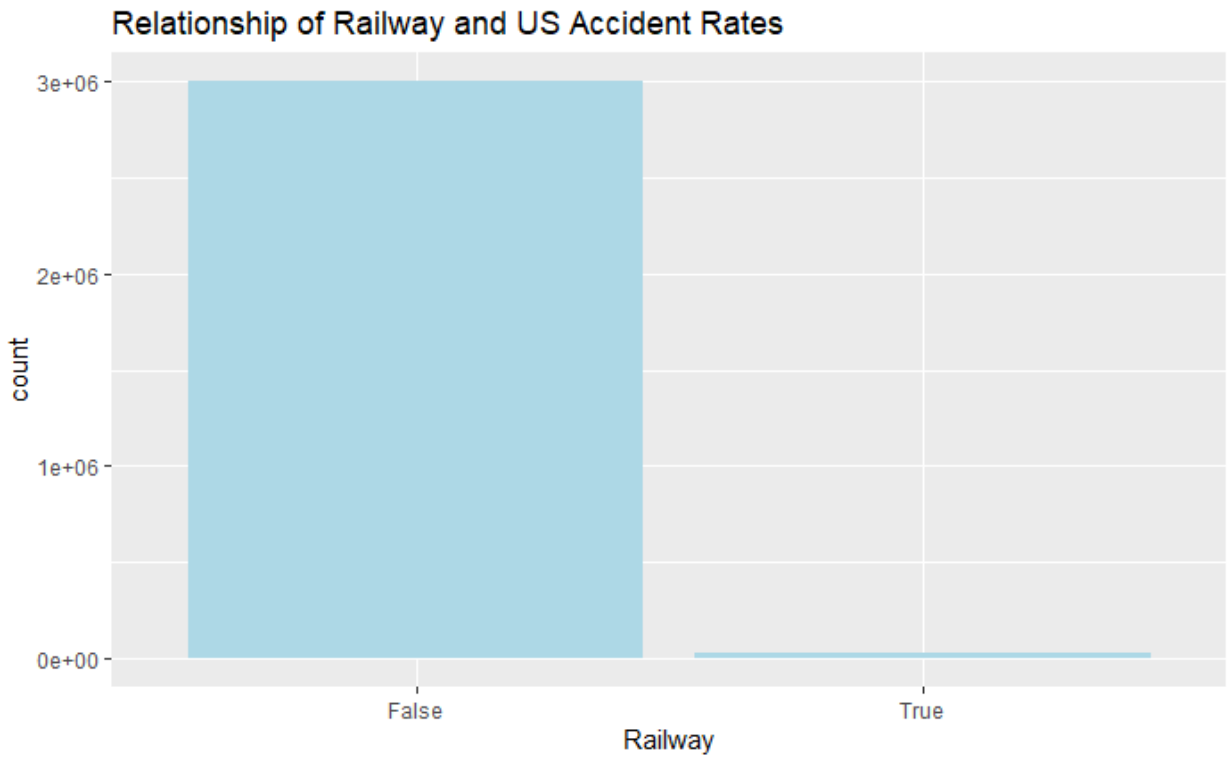

Relationship of Junction and Severity
count
2000000
1500000
1000000
500000
0
False
True
Junction
as.factor(Severity)
1
2
3
4

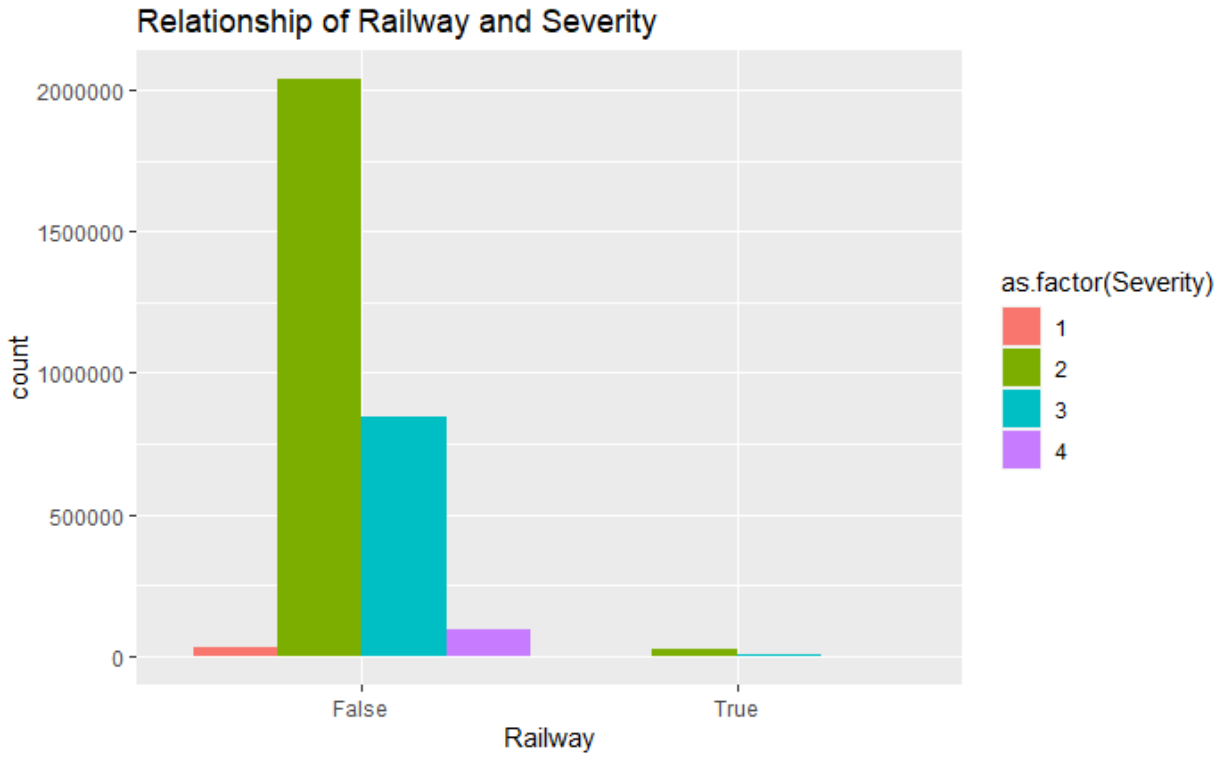

Boxplot - Relationship of Junction and Severity
4
3
2
1
Severity
False
True
Junction
Junction
False
True

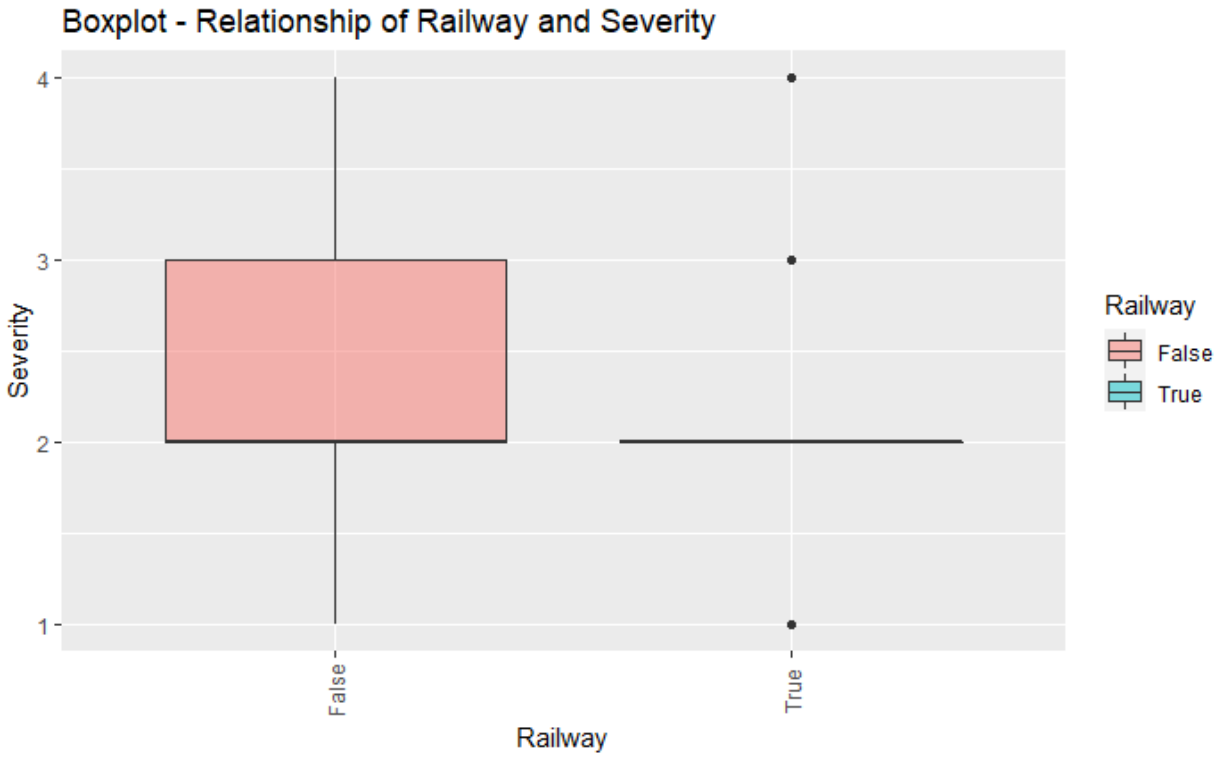

Relationship of No_Exit and US Accident Rates
3e+06
2e+06
1e+06
0e+00
count
False
True
No_Exit

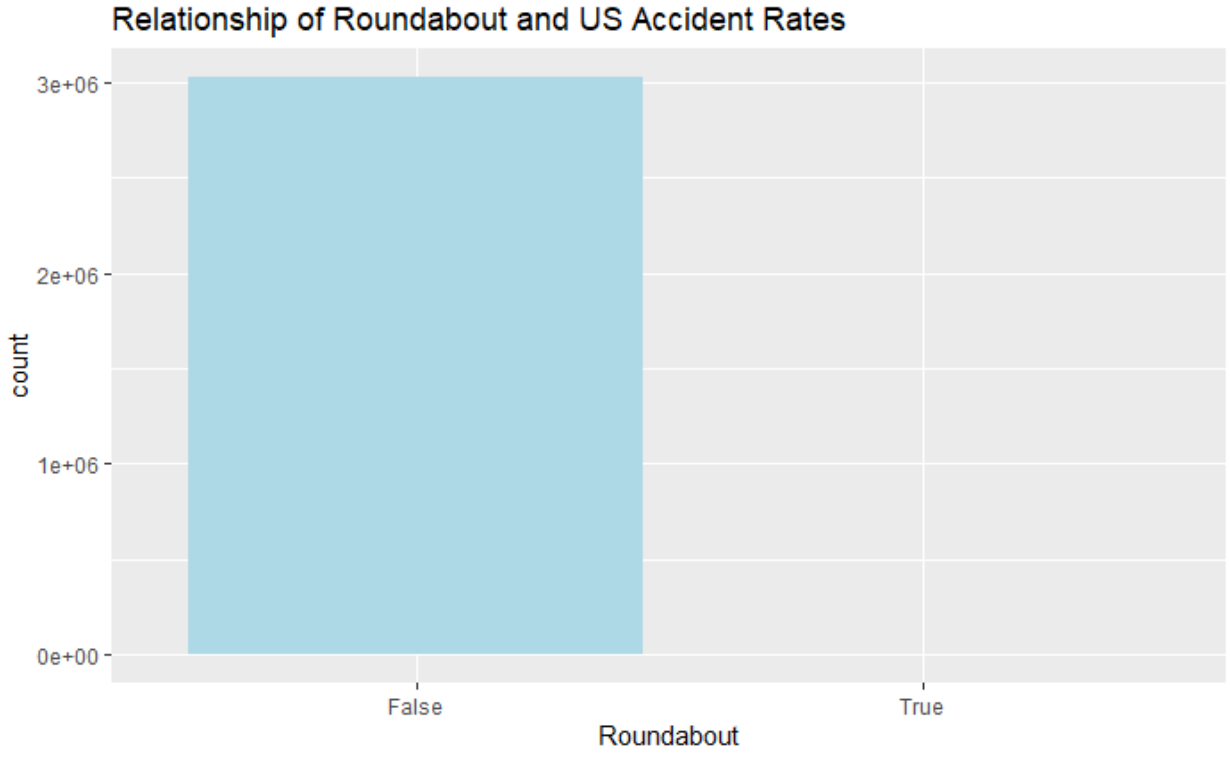

Relationship of No_Exit and Severity
2000000
1500000
1000000
500000
0
count
False
True
No_Exit
as.factor(Severity)
1
2
3
4

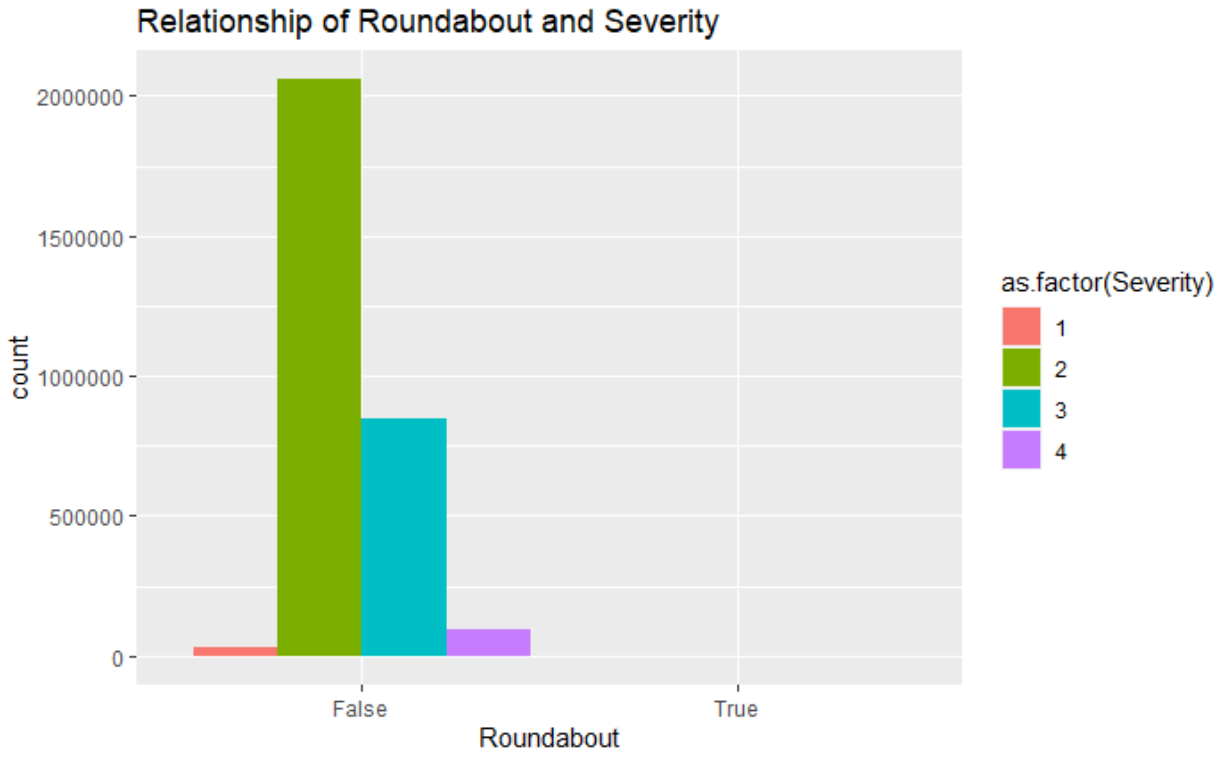

Boxplot - Relationship of No_Exit and Severity
4
3
2
1
Severity
False
True
No_Exit
No_Exit
False
True

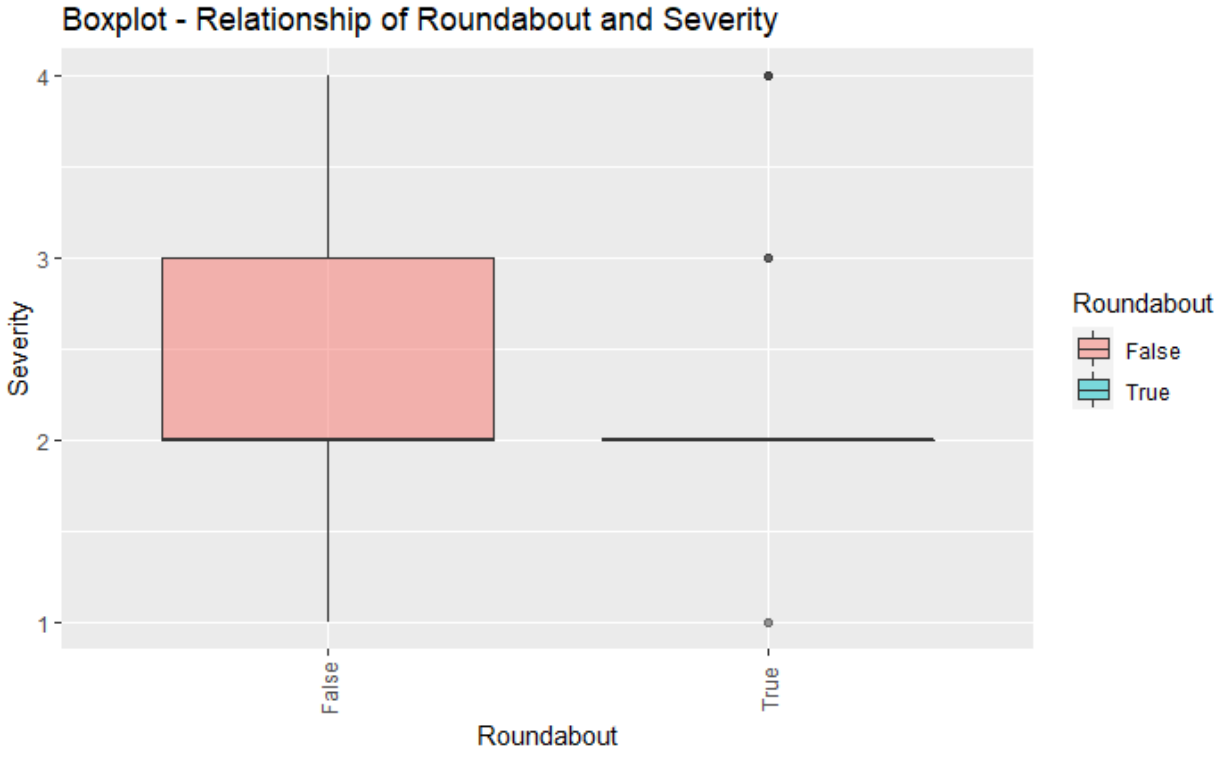

Relationship of Railway and US Accident Rates
3e+06
2e+06
1e+06
0e+00
count
False
True
Railway

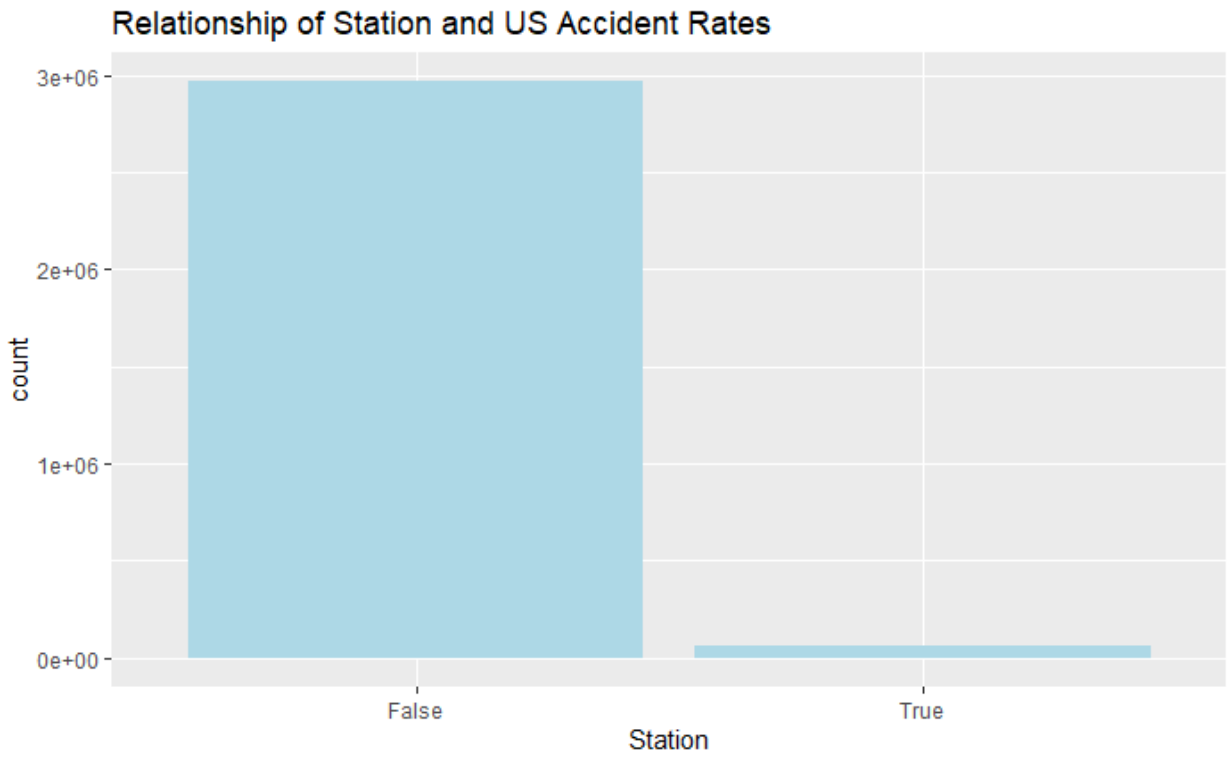

Relationship of Railway and Severity
2000000
1500000
1000000
500000
0
count
False
True
Railway
as.factor(Severity)
1
2
3
4

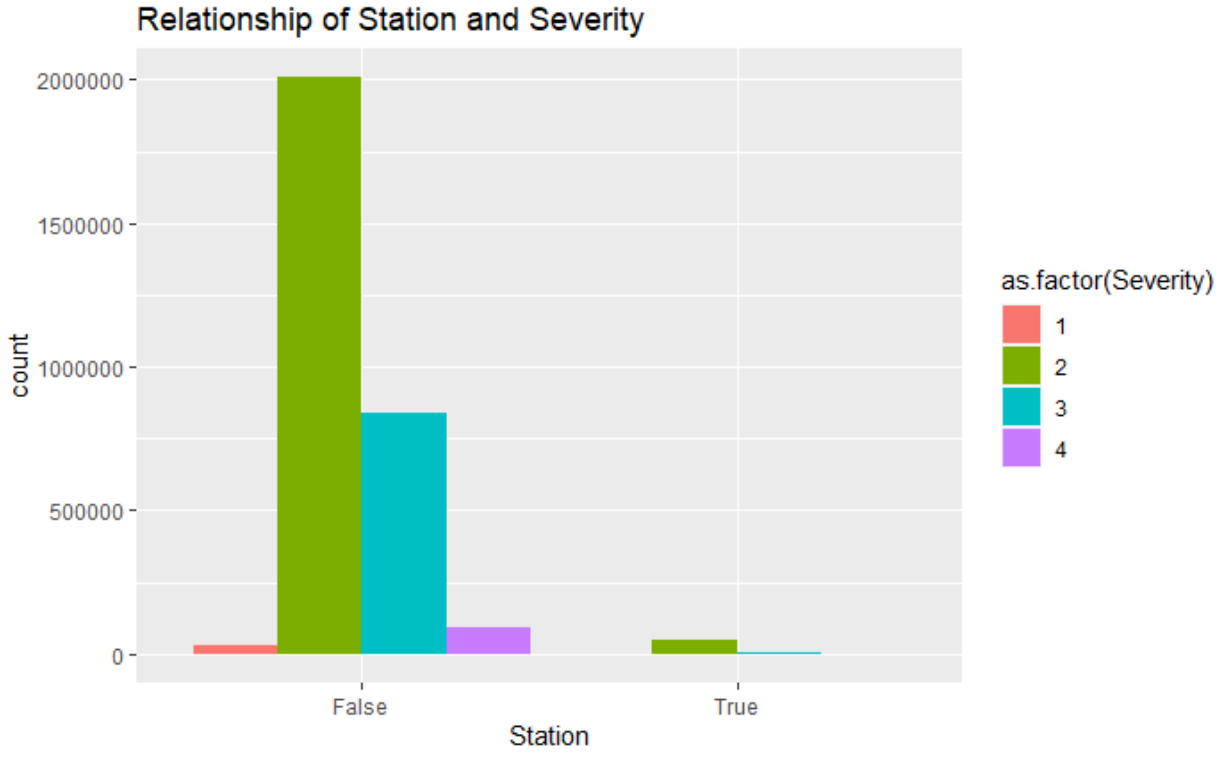

Boxplot - Relationship of Railway and Severity
Severity
4
3
2
1
False
True
Railway
Railway
False
True

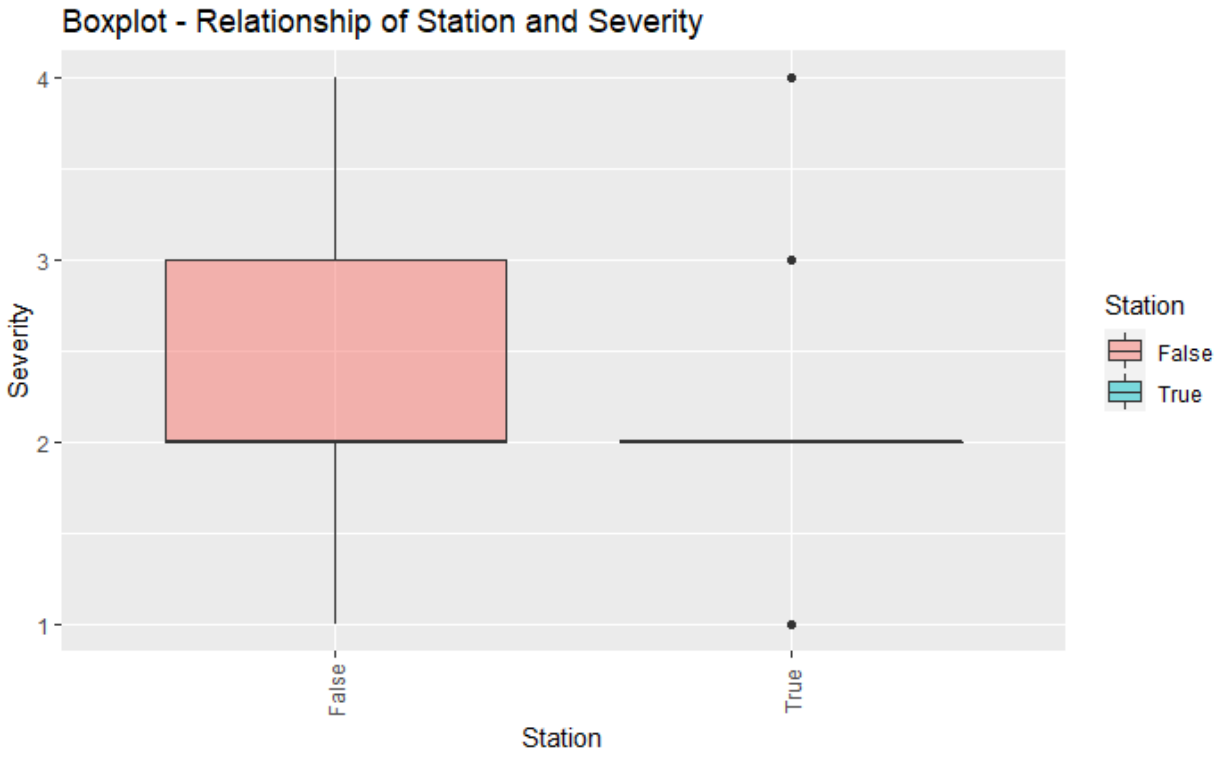

Relationship of Roundabout and US Accident Rates
count
3e+06
2e+06
1e+06
0e+00
False
True
Roundabout

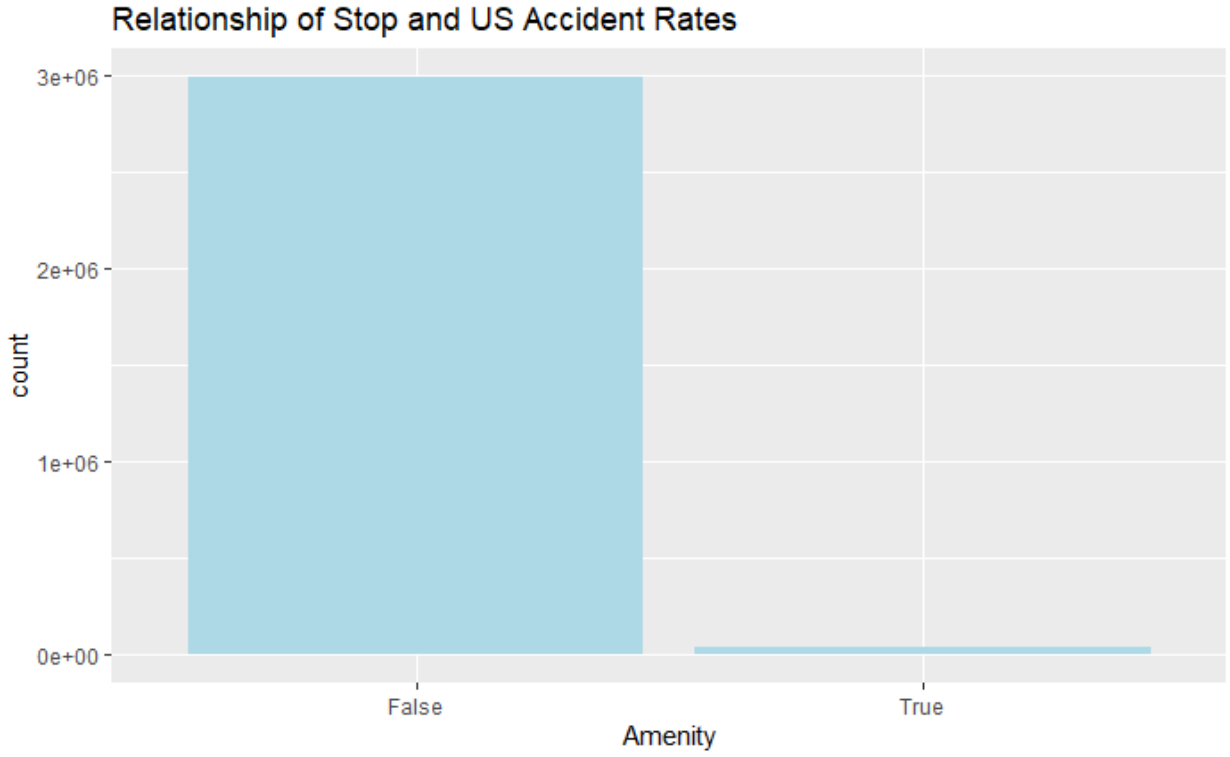

Relationship of Roundabout and Severity
2000000
1500000
1000000
500000
0
count
False
True
Roundabout
as.factor(Severity)
1
2
3
4

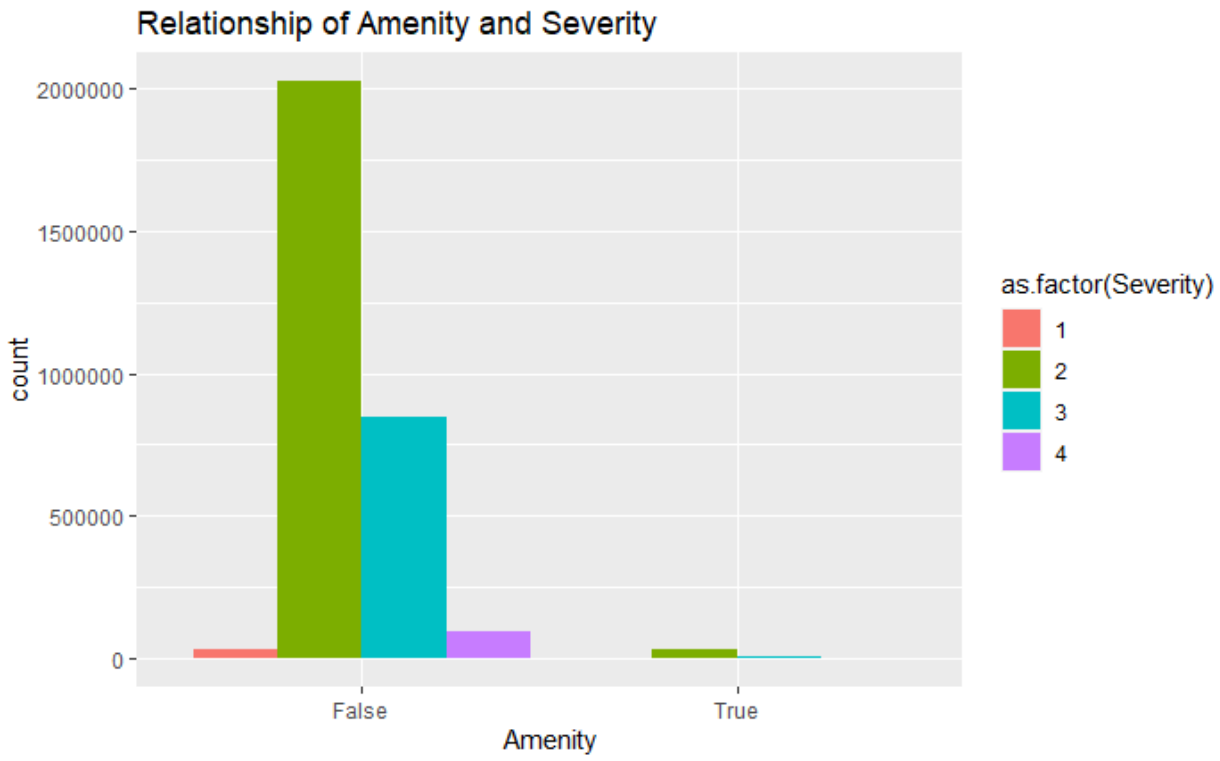

Boxplot - Relationship of Roundabout and Severity
4
3
2
1
Severity
False
True
Roundabout
Roundabout
False
True

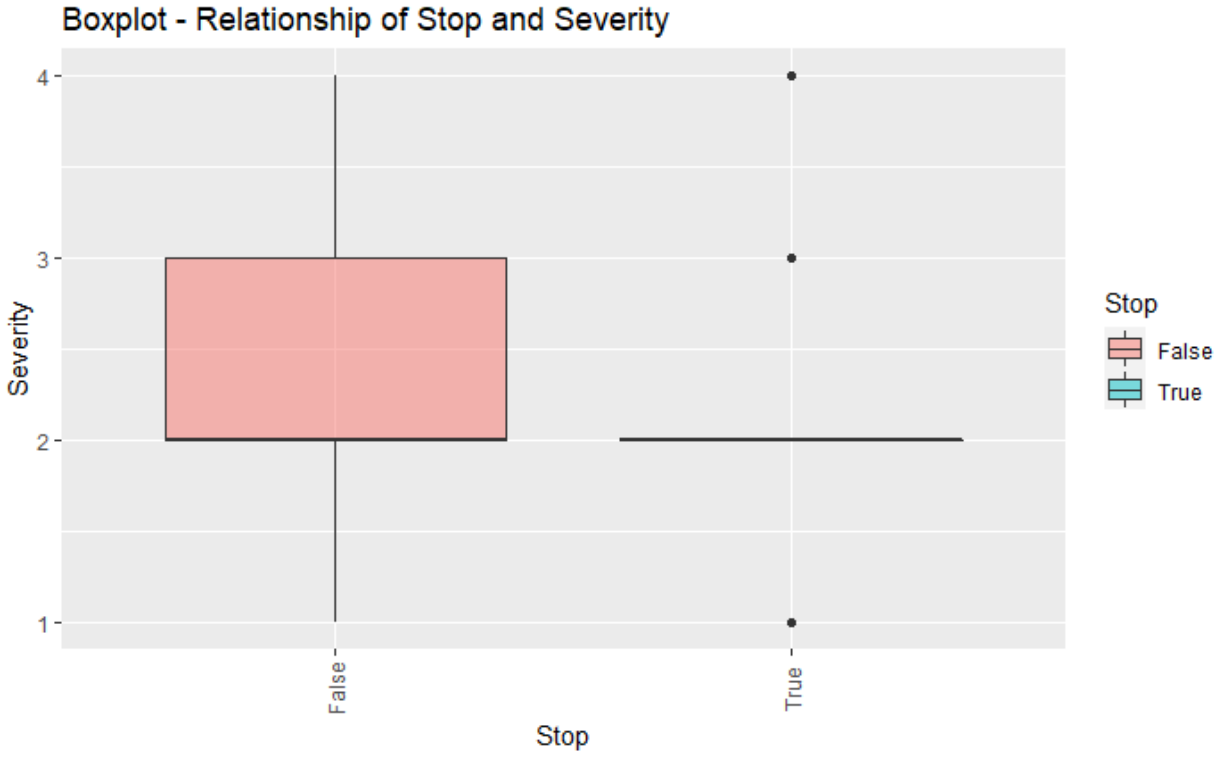

Relationship of Station and US Accident Rates
3e+06
2e+06
1e+06
0e+00
count
False
True
Station

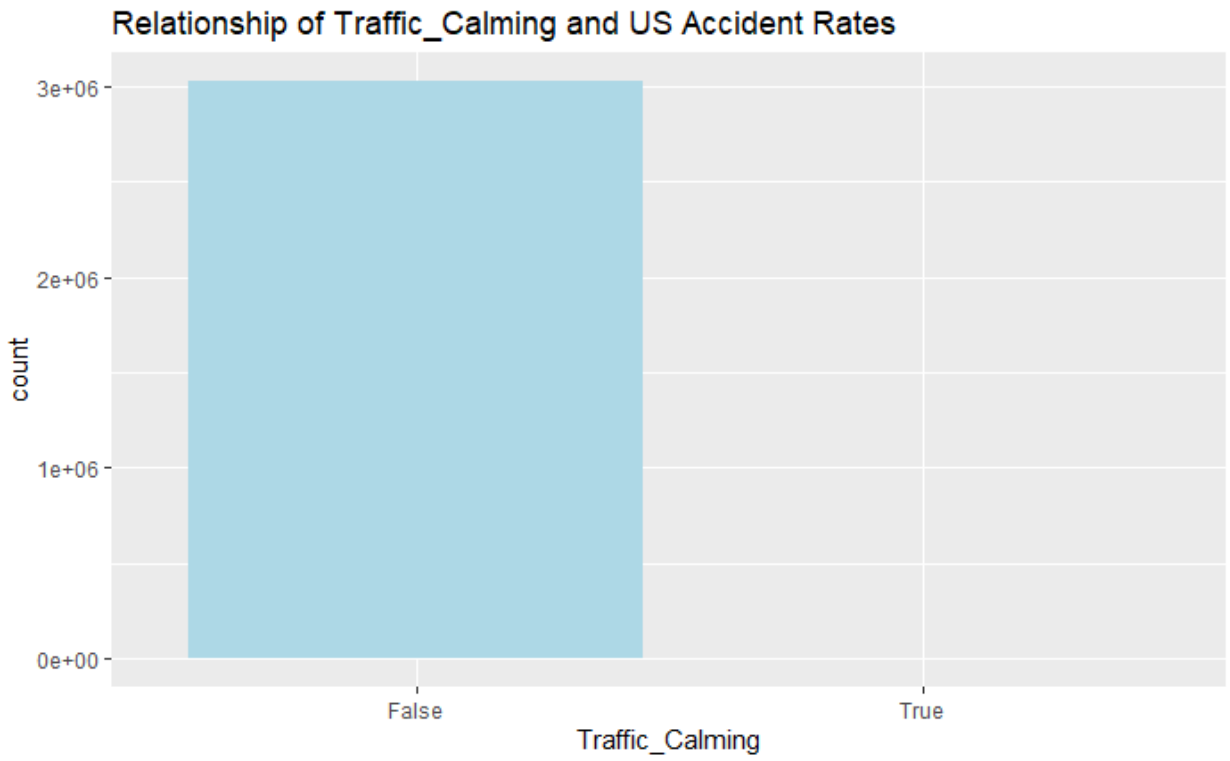

Relationship of Station and Severity
2000000
1500000
1000000
500000
0
count
False
True
Station
as.factor(Severity)
1
2
3
4

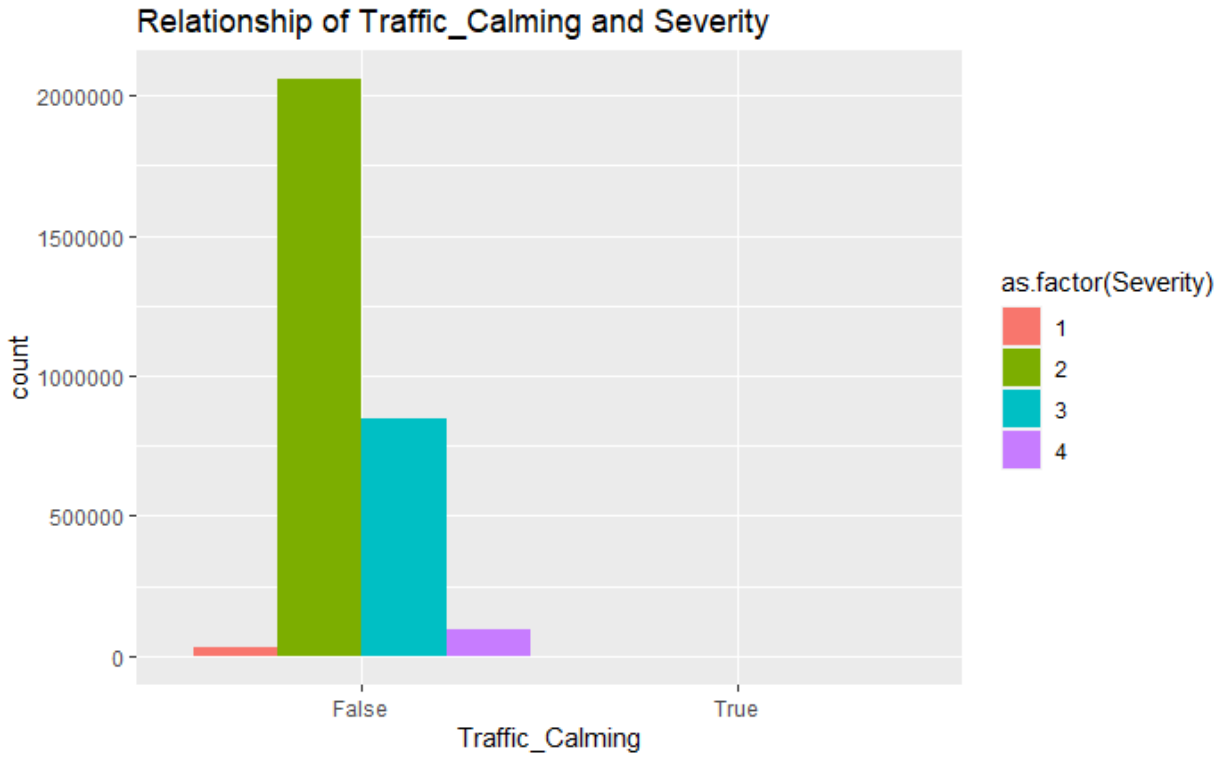

Boxplot - Relationship of Station and Severity
4
3
2
1
Severity
False
True
Station
Station
False
True

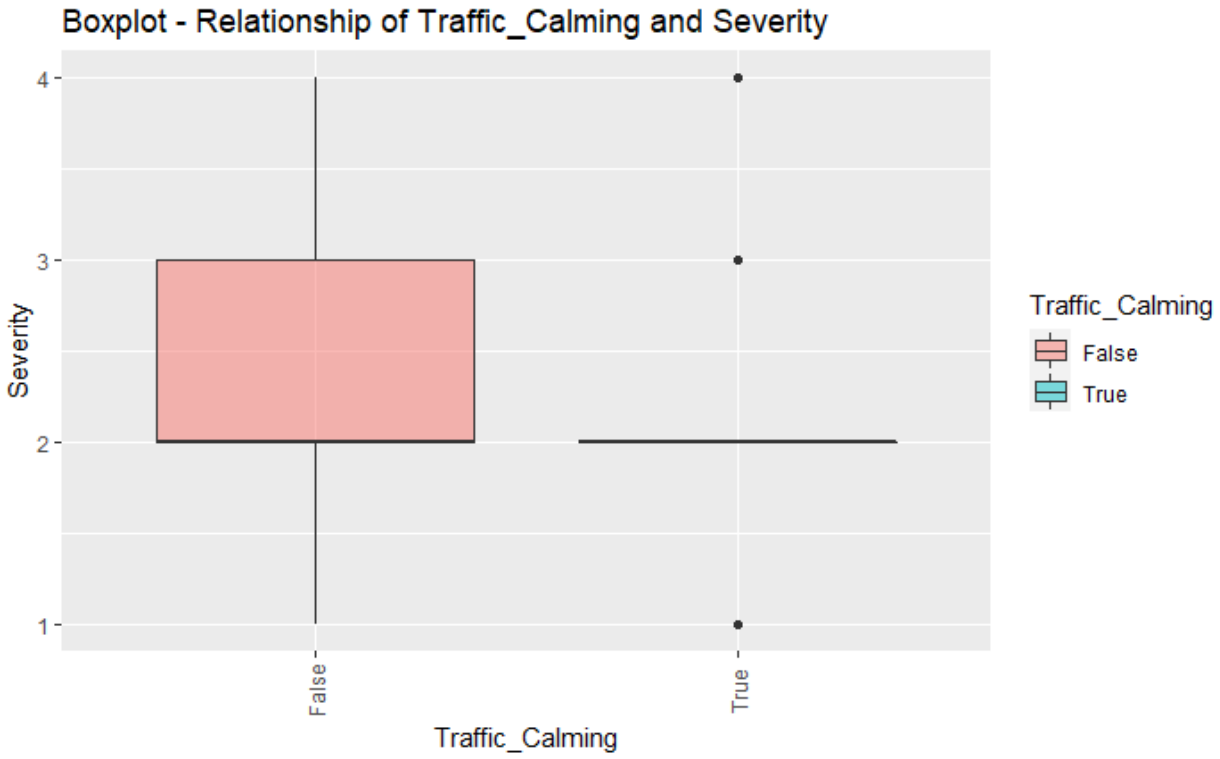

Relationship of Stop and US Accident Rates
3e+06
2e+06
1e+06
0e+00
count
False
True
Amenity

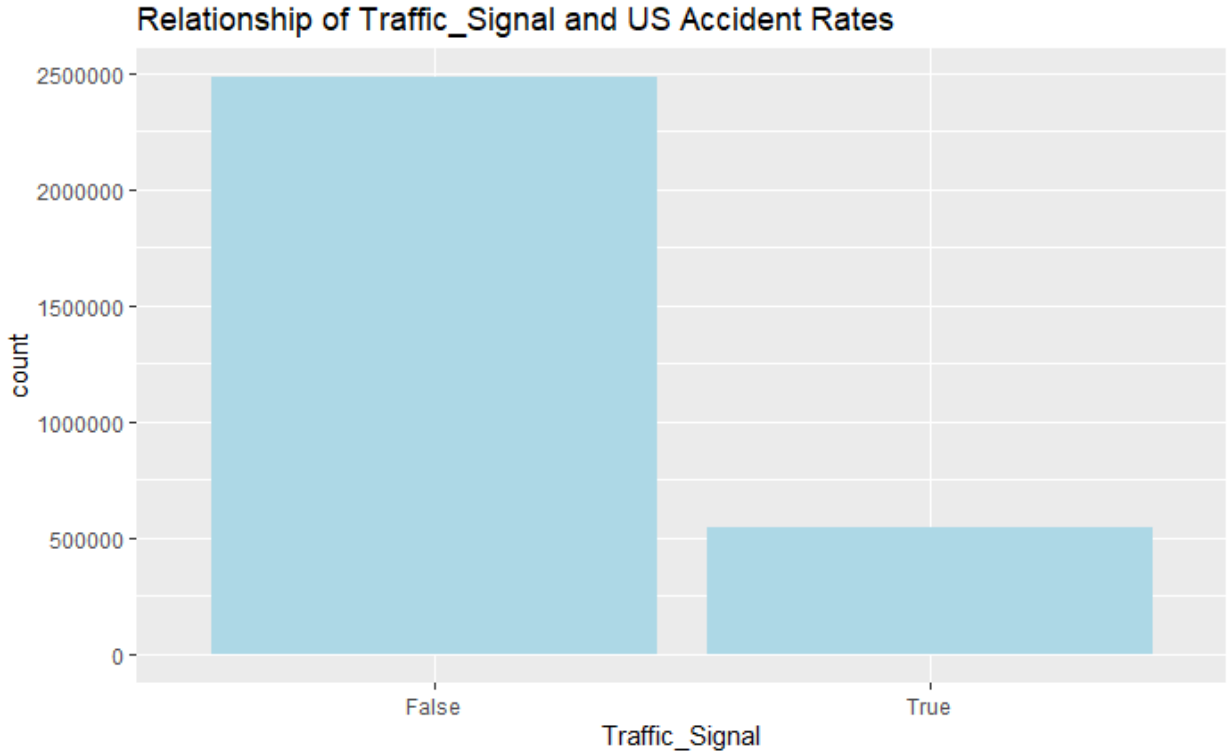

Relationship of Amenity and Severity
2000000
1500000
1000000
500000
0
count
False
True
Amenity
as.factor(Severity)
1
2
3
4

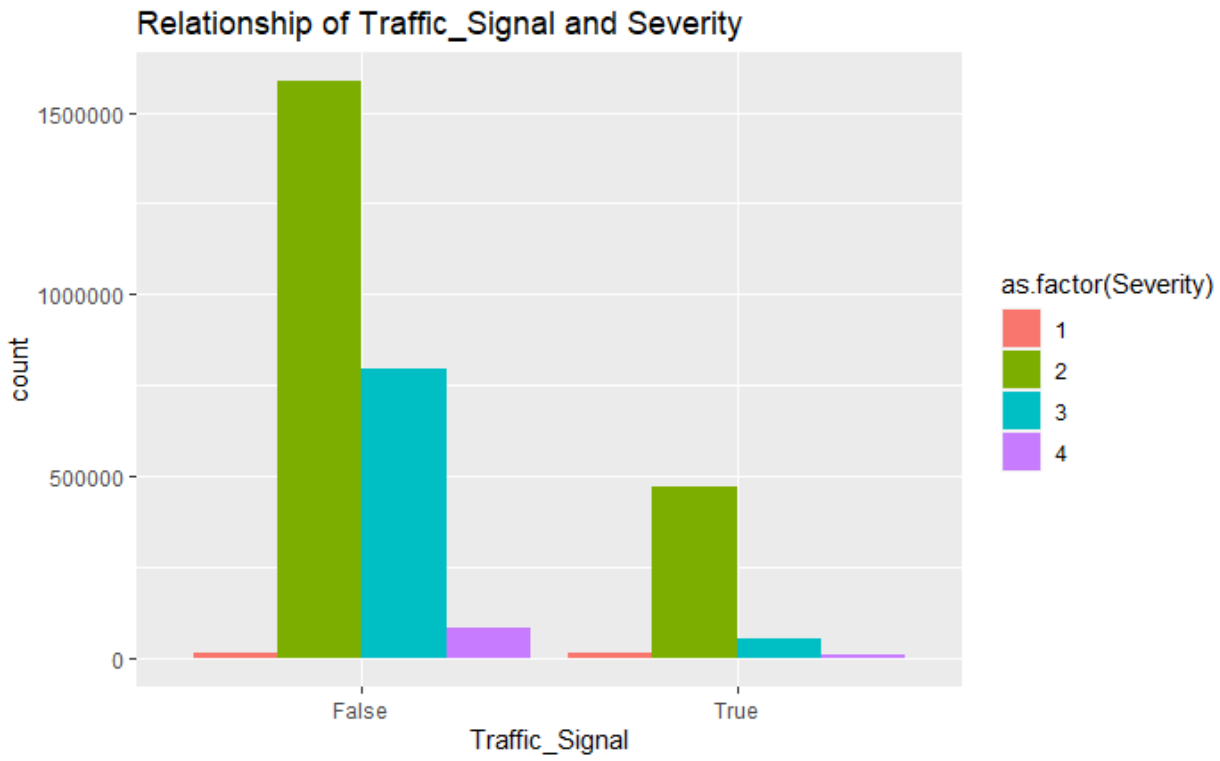

Boxplot - Relationship of Stop and Severity
4
3
2
1
Severity
False
True
Stop
Stop
False
True

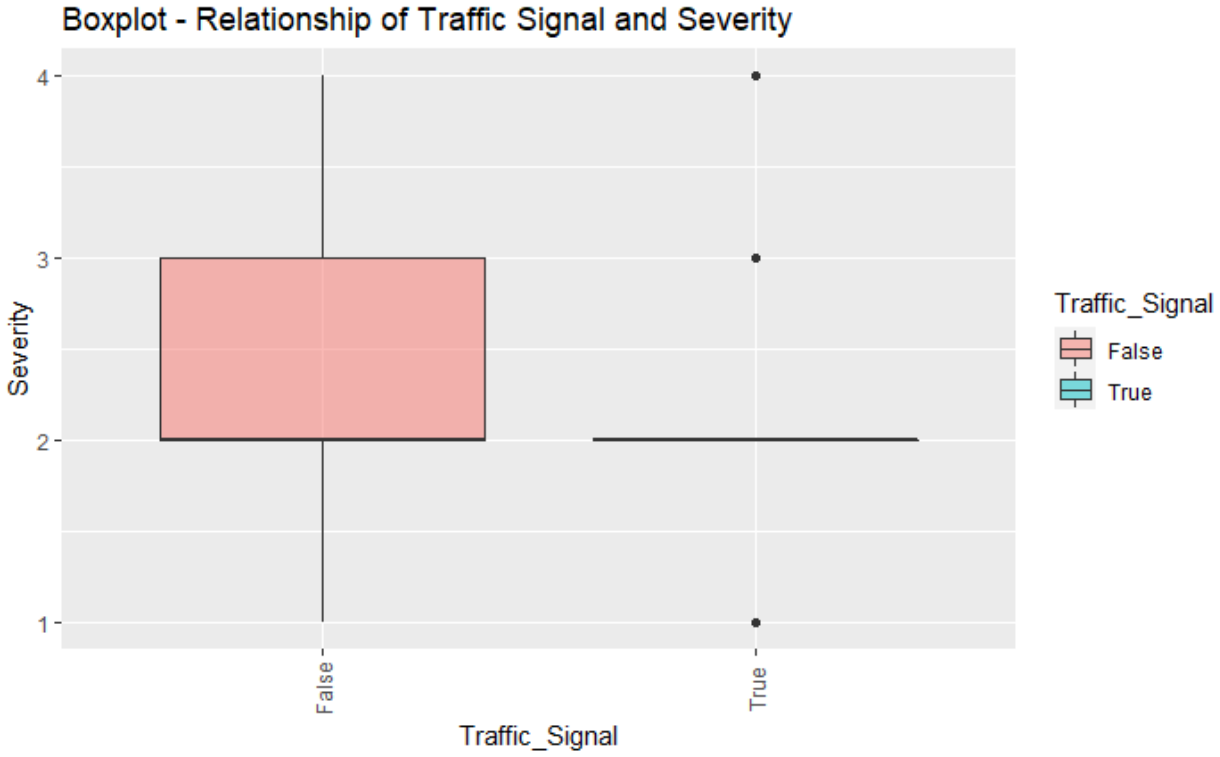

Relationship of Traffic_Calming and US Accident Rates
3e+06
2e+06
1e+06
0e+00
count
False
True
Traffic_Calming

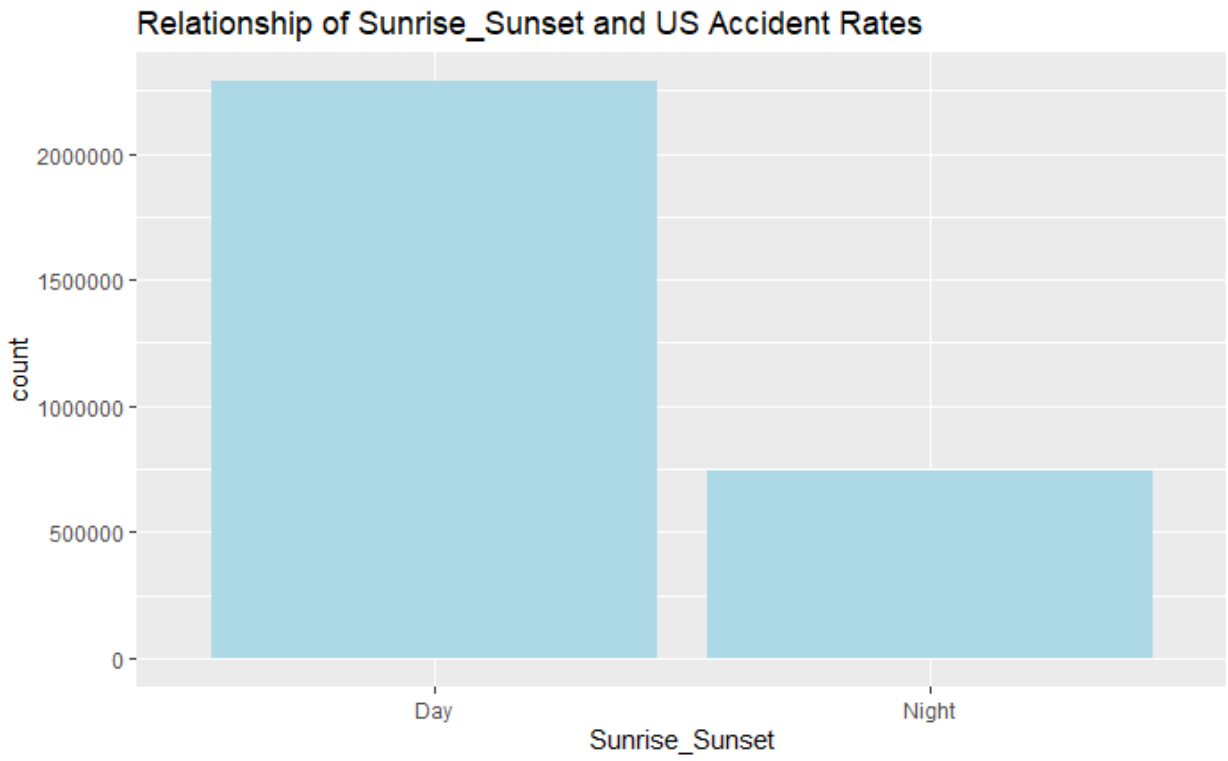

Relationship of Traffic_Calming and Severity
2000000
1500000
1000000
500000
0
count
False
True
Traffic_Calming
as.factor(Severity)
1
2
3
4

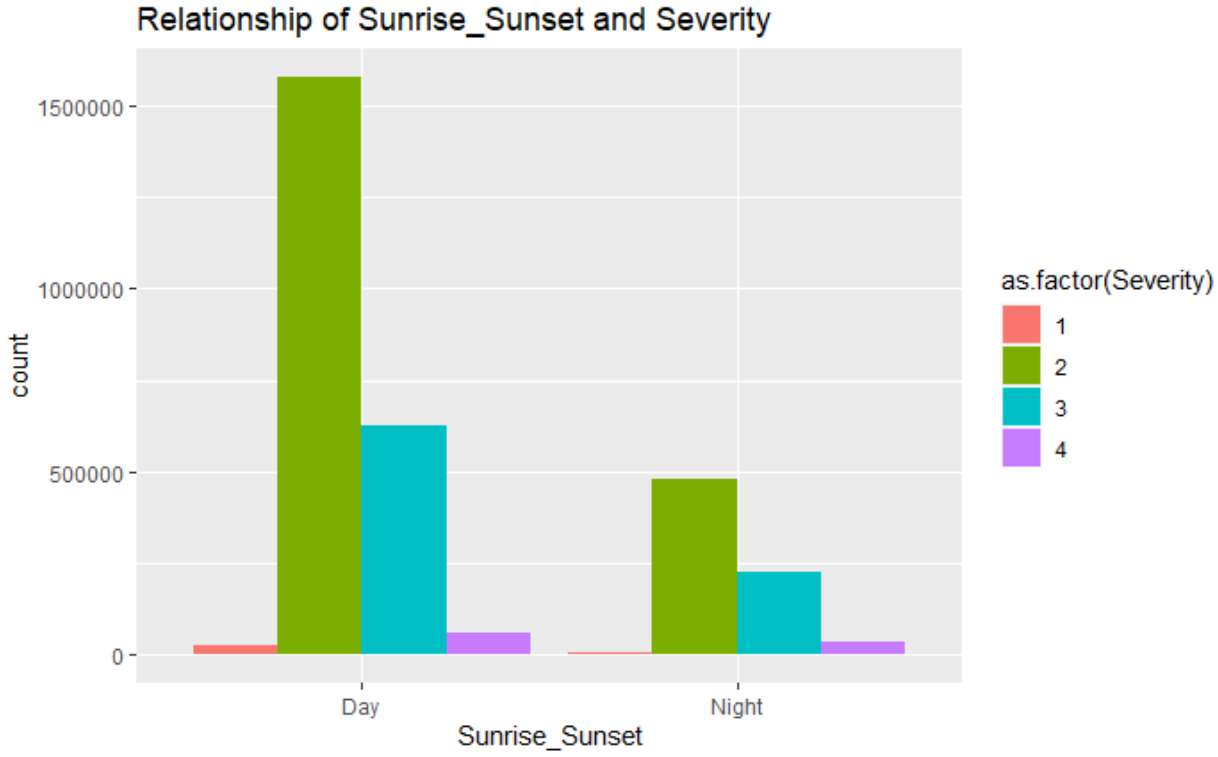

Boxplot - Relationship of Traffic_Calming and Severity
Severity
4
3
2
1
False
True
Traffic_Calming
Traffic_Calming
False
True


Relationship of Traffic_Signal and US Accident Rates
count
2500000
2000000
1500000
1000000
500000
0
False
True
Traffic_Signal

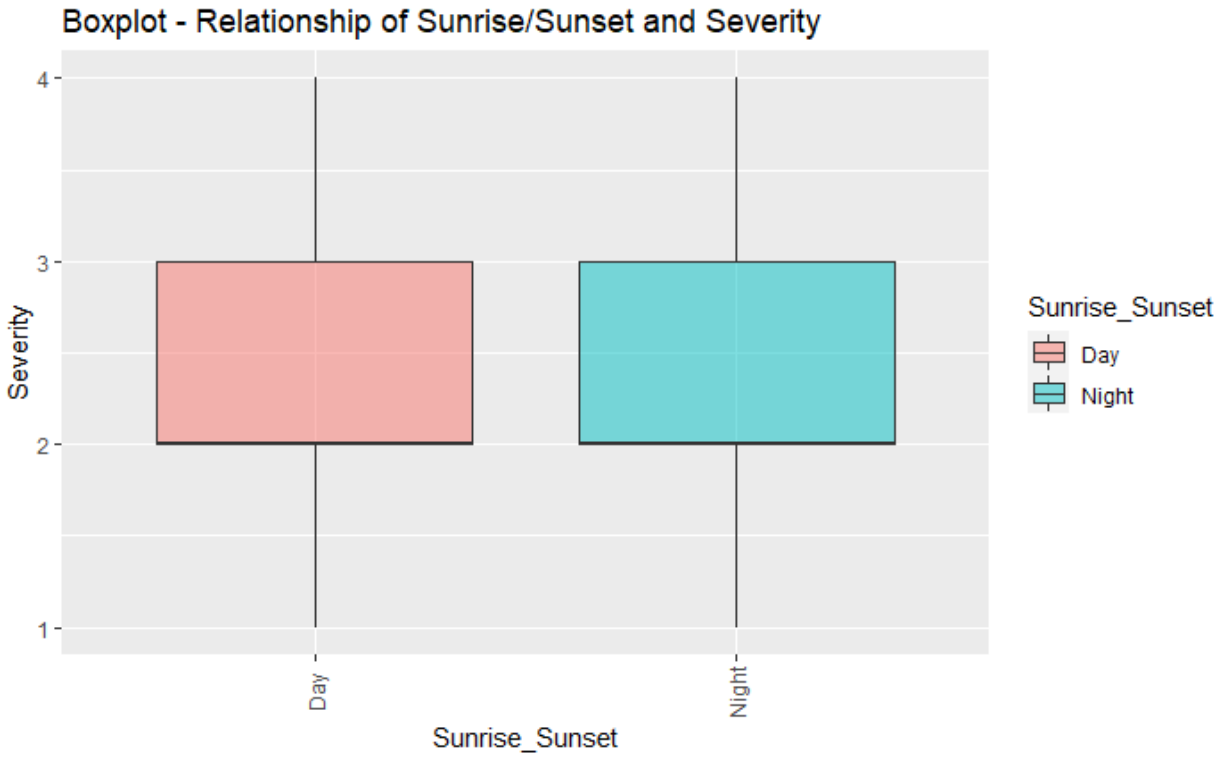

Relationship of Traffic_Signal and Severity
1500000
1000000
500000
0
count
False
True
Traffic_Signal
as.factor(Severity)
1
2
3
4


Boxplot - Relationship of Traffic Signal and Severity
4
3
2
1
Severity
False
True
Traffic_Signal
Traffic_Signal
False
True


Relationship of Sunrise_Sunset and US Accident Rates
2000000
1500000
1000000
500000
0
count
Day
Night
Sunrise_Sunset


Relationship of Sunrise_Sunset and Severity
1500000
1000000
500000
0
count
Day
Night
Sunrise_Sunset
as.factor(Severity)
1
2
3
4

C.1




D. Data Dictionary

| Attribute | Description |
|---|---|
| ID | This is a unique identifier of the accident record |
| Source | Indicates source of the accident report (ie. The API which reported the accident.) |
| TMC | A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event |
| Severity | Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (ie. Short delay as a result of the result of the accident) and 4 indicates a significant impact on traffic (ie. Long delay as a result of the accident) |
| Start_Time | Shows start time of the accident in local time zone |
| End_Time | Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed |
| Start_Lat | Shows latitude in GPS coordinate of the start point |
| Start_Lng | Shows longitude in GPS coordinate of the start point |
| End_Lat | Shows latitude in GPS coordinate of the end point |

| | |
|---|---|
| End_Lng | Shows longitude in GPS coordinate of the end point |
| Distance(mi) | The length of the road extent affected by the accident |
| Description | Shows natural language description of the accident |
| Number | Shows the street number in address field |
| Street | Shows the street name in address field |
| Side | Shows the relative side of the street (Right/Left) in address field |
| City | Shows the city in address field |
| County | Shows he county in address field |
| State | Shows the state in address field |
| Zipcode | Shows the zipcode in address field |
| Country | Shows the country in address field |
| Timezone | Shows timezone based on the location of the accident (eastern, central, etc.) |
| Airport_Code | Denotes an airport_based weather station which is the closest one to location of the accident |
| Weather_Timestamp | Shows the time-stamp of weather observation record (in local time) |
| Temprerature(F) | Shows the temperature (in Fahrenheit) |
| Wind_Chill(F) | Shows the wind chill (in Fahrenheit) |
| Humidity(%) | Shows the humidity (in percentage) |
| Pressure(in) | Shows the air pressure (in inches) |
| Visibility(mi) | Shows visibility (in miles) |
| Wind_Direction | Shows wind direction |
| Wind_Speed(mph) | Shows wind speed (in miles per hour) |
| Precipitation(in) | Shows precipitation amount in inches if there is any |
| Weather_Condition | Shows the weather condition (rain, snow, thunderstorm, fog, etc.) |
| Amenity | A POI annotation which indicates presence of amenity in a nearby location |
| Bump | A POI annotation which indicates presence of speed bump in a nearby location |
| Crossing | A POI annotation which indicates presence of crossing in a nearby location |
| Give_Way | A POI annotation which indicates presence of give_way in a nearby location |
| Junction | A POI annotation which indicates presence of junction in a nearby location |
| No_Exit | A POI annotation which indicates presence of no_exit in a nearby location |
| Railway | A POI annotation which indicates presence of railway in a nearby location |
| Roundabout | A POI annotation which indicates presence of roundabout in a nearby location |

| | |
|---|---|
| Station | A POI annotation which indicates presence of station in a nearby location |
| Stop | A POI annotation which indicates presence of stop in a nearby location |
| Traffic_Calming | A POI annotation which indicates presence of traffic_calming in a nearby location |
| Traffic_Signal | A POI annotation which indicates presence of turning_signal in a nearby location |
| Turning_Loop | A POI annotation which indicates presence of turning_loop in a nearby location |
| Sunrise_Sunset | Shows the period of day (ie. Day or night) based on sunrise/sunset |
| Civil_Twilight | Shows the period of day (ie. Day or night) based on civil_twilight |
| Nautical_Twilight | Shows the period of day (ie. Day or night) based on nautical_twilight |
| Astronomical_Twilight | Shows the period of day (ie. Day or night) based on astronomical_twilight |

E. R code

```
# Data Cleaning
# Remove unnecessary variables in data anaylsis here
```{r}
#Remove ID since unique identifier is unnecessary
US_Accidents$ID <- NULL

#Remove ID since TMC has too many missing data
US_Accidents$TMC <- NULL

#Remove End_Time since this variable is not meaningful to our analysis
US_Accidents$End_Time <- NULL

#Remove End_Lat since this variable is not meaningful to our analysis
US_Accidents$End_Lat <- NULL

#Remove End_Lng since this variable is not meaningful to our analysis
US_Accidents$End_Lng <- NULL

#Remove Distance.mi. since this variable is not meaningful to our analysis
US_Accidents$Distance.mi. <- NULL

# Remove Description since this variable is not meaningful to our analysis
US_Accidents$Description <- NULL

# Remove Number since there are too many missing data
US_Accidents$Number <- NULL

# Remove Street since this variable is not meaningful to our analysis
US_Accidents$Street <- NULL

# Remove Street since this variable is not meaningful to our analysis
US_Accidents$County <- NULL

# Remove Zipcode since this variable is not meaningful to our analysis
US_Accidents$Zipcode <- NULL

# Remove Airport_code since this variable is not meaningful to our analysis
US_Accidents$Airport_Code <- NULL

# Remove Side since this variable is not meaningful to our analysis
US_Accidents$Side <- NULL

# Remove country since there is only 1 level "US"
US_Accidents$Country <- NULL

# Remove Timezone since this variable is not meaningful to our analysis
US_Accidents$Timezone <- NULL

# Remove Airport_code since this variable is not meaningful to our analysis
US_Accidents$Airport_code <- NULL
```
```

```r
# Remove weather_Timestamp since this variable is not meaningful to our analysis
US_Accidents$Weather_Timestamp <- NULL

# Remove Wind_Direction since this variable is not meaningful to our analysis
US_Accidents$Wind_Direction <- NULL

# Remove Wind_Chill since there are too many missing data
US_Accidents$Wind_Chill.F. <- NULL

# Remove Precipitation since there are too many missing data
US_Accidents$Precipitation.in. <- NULL

# Remove Turning_Loop since there is only 1 level "False"
US_Accidents$Turning_Loop <- NULL

# Remove Civil_Twilight since this variable is not meaningful to our analysis
US_Accidents$Civil_Twilight <- NULL

# Remove Nautical_Twilight since this variable is not meaningful to our analysis
US_Accidents$Nautical_Twilight <- NULL

# Remove Astronomical_Twilight since this variable is not meaningful to our analysis
US_Accidents$Astronomical_Twilight <- NULL


summary(US_Accidents)
```

```
# Handle Weather_Condition data - combining similar groups of data
```{r}
library(plyr)
library(dplyr)
# Remove 18749 rows of missing data in Weather Condition
US_Accidents <- subset(US_Accidents, Weather_Condition != "")
US_Accidents$Weather_Condition <- droplevels(US_Accidents$Weather_Condition)
levels(US_Accidents$Weather_Condition)

# Regroup the levels of the Weather_Condition variable to reduce levels
var.levels <- levels(US_Accidents$Weather_Condition)
US_Accidents$Weather_Condition <- mapvalues(US_Accidents$Weather_Condition, var.levels,
                                           c("Dust", "Windy", "Dust",
                                             "Snow", "Snow", "Clear",
                                             "Cloudy","Cloudy", "Snow",
                                             "Rain", "Rain", "Rain",
                                             "Dust", "Clear", "Clear",
                                             "Fog", "Fog", "Rain",
                                             "Hail", "Hail", "Cloudy",
                                             "Hail", "Fog", "Fog",
                                             "Snow", "Rain", "Hail",
                                             "Hail", "Hail", "Rain",
                                             "Rain", "Rain", "Rain",
                                             "Hail", "Fog", "Snow",
                                             "Snow", "Snow", "Thunderstorm",
                                             "Thunderstorm", "Thunderstorm", "Thunderstorm",
                                             "Thunderstorm", "Hail", "Snow",
                                             "Rain", "Rain", "Fog",
                                             "Rain", "Hail", "Hail",
                                             "Hail", "Hail", "Fog",
                                             "Hail", "Rain", "Rain",
                                             "Rain", "Rain", "Rain",
                                             "Thunderstorm", "Hail", "Snow",
                                             "Snow", "Hail", "Hail",
                                             "Snow", "Snow", "Snow",
                                             "Snow", "Thunderstorm", "Thunderstorm",
                                             "Snow", "Snow", "Fog",
                                             "Cloudy","Cloudy", "Snow",
                                             "Cloudy","Fog", "Fog",
                                             "Cloudy","Cloudy", "Fog",
                                             "Fog", "Rain", "Rain",
                                             "Hail", "Rain", "Rain", "Dust",
                                             "Dust", "Dust", "Dust",
                                             "Cloudy","Fog", "Rain",
                                             "Hail", "Hail", "Fog",
                                             "Fog", "Snow", "Snow",
                                             "Hail", "Hail", "Snow",
                                             "Snow", "Snow", "Windy",
                                             "Thunderstorm", "Thunderstorm", "Thunderstorm",
                                             "Thunderstorm", "Thunderstorm", "Thunderstorm",
                                             "Hail", "Hail", "Thunderstorm",
                                             "Thunderstorm", "Thunderstorm", "Thunderstorm",
                                             "Other", "Other", "Dust",
                                             "Dust", "Snow", "Snow"
                                             ))
table(US_Accidents$Weather_Condition)
```
```

# Data Cleaning
# Handle missing and anomalies data
```{r}
# Removing missing rows
US_Accidents <- US_Accidents[!is.na(US_Accidents$Temperature.F.),]
US_Accidents <- US_Accidents[!is.na(US_Accidents$Humidity),]
US_Accidents <- US_Accidents[!is.na(US_Accidents$Pressure.in.),]
US_Accidents <- US_Accidents[!is.na(US_Accidents$Visibility.mi.),]
US_Accidents <- US_Accidents[!is.na(US_Accidents$Wind_Speed.mph.),]

# Remove 116 rows of missing data in Sunrise_Sunset
US_Accidents <- subset(US_Accidents, Sunrise_Sunset != "")
US_Accidents$Sunrise_Sunset <- droplevels(US_Accidents$Sunrise_Sunset)


# Remove the outliers from temperature data
US_Accidents <- US_Accidents[US_Accidents$Temperature.F. < 130,]


summary(US_Accidents)
str(US_Accidents)
```

# Data visualization - Source
```{r}
library(ggplot2)
ggplot(US_Accidents, aes(x=Source)) + geom_bar(fill="lightblue") + ggtitle("Data Visualization of Source")
```

# Data visualization - Severity
```{r}
ggplot(US_Accidents, aes(x=Severity)) + geom_bar(fill="lightblue") + ggtitle("Data Visualization of Severity")
```

# Data visualization - Date and Time
```{r}
# Bar chart for Severity variable
ggplot(US_Accidents, aes(x=Severity)) + geom_bar(fill="lightblue")

# The relationship of Time and Severity variable
date <- as.Date(US_Accidents$Start_Time)
year <- format(date, format ='%Y')
month <-format(date, format ='%m')
day <-format(date, format ='%A')
time <-strptime(US_Accidents$Start_Time, format = '%Y-%m-%d %H:%M:%S')
hour <- format(time, '%H')

ggplot(data=US_Accidents, aes(x=year)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Year and US Accident Rates")
ggplot(data=US_Accidents, aes(x=year, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Year and Severity")
ggplot(data=US_Accidents, aes(x=month)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Month and US Accident Rates")
ggplot(data=US_Accidents, aes(x=month, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Month and Severity")
ggplot(data=US_Accidents, aes(x=day)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Day and US Accident Rates")
ggplot(data=US_Accidents, aes(x=day, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Day and Severity")
ggplot(data=US_Accidents, aes(x=hour)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Hour and US Accident Rates")
ggplot(data=US_Accidents, aes(x=hour, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Hour and Severity")


```

# Data visualization - Latitude and Longtitude

```{r}
library(usmap)



#MainStates <- map_data("state")

#us_map <- usmap::us_map()
#usmap::plot_usmap() + geom_point(data=US_Accidents_final, aes(x=Start_Lng, y=Start_Lat)) + labs(title = "US Car Accidents",
#       size = "severity") +
#  theme(legend.position = "right")


ggplot(data=US_Accidents, aes(x=Start_Lng, y=Start_Lat)) + geom_hex(bins=300)

ggplot(data=US_Accidents, aes(x=Start_Lng, y=Start_Lat)) + geom_hex(bins=300)+ facet_wrap(~US_Accidents$Severity, ncol=2)
ggplot(data=US_Accidents, aes(x=Start_Lng, y=Start_Lat)) + geom_point(aes(color=Severity))

```

# Data visualization - City and State

```{r}
# City
City = c("Houston", "Los Angeles", "Charlotte", "Dallas", "Austin", "Raleigh")
City1 = US_Accidents[which(US_Accidents$City %in% City),]

ggplot(City1, aes(x=City)) + geom_bar(fill="lightblue") + ggtitle("Data Visualization of Top Cities")


# State
State = c("CA", "TX", "FL", "SC", "NC", "NY")
State1 = US_Accidents[which(US_Accidents$State %in% State),]

ggplot(State1, aes(x=State)) + geom_bar(fill="lightblue") + ggtitle("Data Visualization of Top States")

ggplot(data = State1, aes(x = State, y = Severity, fill = State)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of State and Severity")

```

# Data visualization - Weather condition variables

```{r}

# The relationship of Temperature and Severity variable
ggplot(data=US_Accidents, aes(x=Temperature.F.)) + geom_histogram(fill="lightblue") + ggtitle("Relationship of Temperature and US
Accident Rates")
ggplot(data=US_Accidents, aes(x=Temperature.F., group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Temperature and Severity")


# The relationship of Humidity and Severity variable
ggplot(data=US_Accidents, aes(x=Humidity...)) + geom_histogram(fill="lightblue") + ggtitle("Relationship of Humidity and US
Accident Rates")
ggplot(data=US_Accidents, aes(x=Humidity..., group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Humidity and Severity")

# The relationship of Pressure and Severity variable
ggplot(data=US_Accidents, aes(x=Pressure.in.)) + geom_histogram(fill="lightblue") + ggtitle("Relationship of Pressure and US
Accident Rates")
ggplot(data=US_Accidents, aes(x=Pressure.in., group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Pressure and Severity")


# The relationship of Visibility and Severity variable
ggplot(data=US_Accidents, aes(x=Visibility.mi.)) + geom_histogram(fill="lightblue") + ggtitle("Relationship of Visibility and US
Accident Rates")
ggplot(data=US_Accidents, aes(x=Visibility.mi., group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Visibility and Severity")


# The relationship of Wind_Speed and Severity variable
ggplot(data=US_Accidents, aes(x=Wind_Speed.mph.)) + geom_histogram(fill="lightblue") + ggtitle("Relationship of Wind_Speed and US
Accident Rates")
ggplot(data=US_Accidents, aes(x=Wind_Speed.mph., group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Wind Speed and Severity")


# The relationship of Weather Condition and Severity variable
ggplot(data=US_Accidents, aes(x=Weather_Condition)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Weather Condition and
US Accident Rates")
ggplot(data=US_Accidents, aes(x=Weather_Condition, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Weather Condition and Severity")

ggplot(data = US_Accidents, aes(x = Weather_Condition, y = Severity, fill = Weather_Condition)) + geom_boxplot(alpha = 0.5) +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Weather Condition and Severity")



```

# Data visualization - Road condition variables

```{r}

# The relationship of Amenity and Severity variable
ggplot(data=US_Accidents, aes(x=Amenity)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Amenity and US Accident Rates")
ggplot(data=US_Accidents, aes(x=Amenity, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Amenity and Severity")

ggplot(data = US_Accidents, aes(x = Amenity, y = Severity, fill = Amenity)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Amenity and Severity")

# The relationship of Bump and Severity variable

ggplot(data=US_Accidents, aes(x=Bump)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Bump and US Accident Rates")
ggplot(data=US_Accidents, aes(x=Bump, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Bump and Severity")

ggplot(data = US_Accidents, aes(x = Bump, y = Severity, fill = Bump)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Bump and Severity")

# The relationship of Crossing and Severity variable
ggplot(data=US_Accidents, aes(x=Crossing)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Crossing and US Accident Rates")
ggplot(data=US_Accidents, aes(x=Crossing, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Crossing and Severity")

ggplot(data = US_Accidents, aes(x = Crossing, y = Severity, fill = Crossing)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Crossing and Severity")

# The relationship of Give_Way and Severity variable
ggplot(data=US_Accidents, aes(x=Give_Way)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Give_Way and US Accident Rates")
ggplot(data=US_Accidents, aes(x=Give_Way, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Give_Way and Severity")

ggplot(data = US_Accidents, aes(x = Give_Way, y = Severity, fill = Give_Way)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Give Way and Severity")

# The relationship of Junction and Severity variable
ggplot(data=US_Accidents, aes(x=Junction)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Junction and US Accident Rates")
ggplot(data=US_Accidents, aes(x=Junction, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Junction and Severity")

ggplot(data = US_Accidents, aes(x = Junction, y = Severity, fill = Junction)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Junction and Severity")

# The relationship of No_Exit and Severity variable
ggplot(data=US_Accidents, aes(x=No_Exit)) + geom_bar(fill="lightblue") + ggtitle("Relationship of No_Exit and US Accident Rates")
ggplot(data=US_Accidents, aes(x=No_Exit, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of No_Exit and Severity")

ggplot(data = US_Accidents, aes(x = No_Exit, y = Severity, fill = No_Exit)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of No_Exit and Severity")
```

```
# The relationship of Railway and Severity variable
ggplot(data=US_Accidents, aes(x=Railway)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Railway and US Accident Rates")
ggplot(data=US_Accidents, aes(x=Railway, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Railway and Severity")

ggplot(data = US_Accidents, aes(x = Railway, y = Severity, fill = Railway)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Railway and Severity")

# The relationship of Roundabout and Severity variable
ggplot(data=US_Accidents, aes(x=Roundabout)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Roundabout and US Accident
Rates")
ggplot(data=US_Accidents, aes(x=Roundabout, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Roundabout and Severity")

ggplot(data = US_Accidents, aes(x = Roundabout, y = Severity, fill = Roundabout)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Roundabout and Severity")

# The relationship of Station and Severity variable
ggplot(data=US_Accidents, aes(x=Station)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Station and US Accident Rates")
ggplot(data=US_Accidents, aes(x=Station, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Station and Severity")

ggplot(data = US_Accidents, aes(x = Station, y = Severity, fill = Station)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Station and Severity")

# The relationship of Stop and Severity variable
ggplot(data=US_Accidents, aes(x=Amenity)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Stop and US Accident Rates")
ggplot(data=US_Accidents, aes(x=Amenity, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Amenity and Severity")

ggplot(data = US_Accidents, aes(x = Stop, y = Severity, fill = Stop)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x =
element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Stop and Severity")

# The relationship of Traffic_Calming and Severity variable
ggplot(data=US_Accidents, aes(x=Traffic_Calming)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Traffic_Calming and US
Accident Rates")
ggplot(data=US_Accidents, aes(x=Traffic_Calming, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") +
ggtitle("Relationship of Traffic_Calming and Severity")

ggplot(data = US_Accidents, aes(x = Traffic_Calming, y = Severity, fill = Traffic_Calming)) + geom_boxplot(alpha = 0.5) +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Traffic_Calming and Severity")
```

```
# The relationship of Traffic_Signal and Severity variable
ggplot(data=US_Accidents, aes(x=Traffic_Signal)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Traffic_Signal and US Accident Rates")
ggplot(data=US_Accidents, aes(x=Traffic_Signal, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") + ggtitle("Relationship of Traffic_Signal and Severity")

ggplot(data = US_Accidents, aes(x = Traffic_Signal, y = Severity, fill = Traffic_Signal)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Traffic Signal and Severity")

# The relationship of Sunrise_Sunset and Severity variable
ggplot(data=US_Accidents, aes(x=Sunrise_Sunset)) + geom_bar(fill="lightblue") + ggtitle("Relationship of Sunrise_Sunset and US Accident Rates")
ggplot(data=US_Accidents, aes(x=Sunrise_Sunset, group = Severity, fill = as.factor(Severity))) + geom_bar(position = "dodge") + ggtitle("Relationship of Sunrise_Sunset and Severity")

ggplot(data = US_Accidents, aes(x = Sunrise_Sunset, y = Severity, fill = Sunrise_Sunset)) + geom_boxplot(alpha = 0.5) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) + ggtitle("Boxplot - Relationship of Sunrise/Sunset and Severity")
```

```
# Correlations of the target variable to numeric variables and correlation matrix
```{r}
# caluclate the correlation matrix for numeric variables
library(ggcorrplot)
cor.matrix <- cor(US_Accidents[,sapply(US_Accidents, is.numeric)])

print("Correlation Matrix")
cor.matrix
ggcorrplot(cor.matrix, lab=TRUE)

```
```

```
# Examining the mean Severity for each variable value.

```{r}
# The mean of the target variable split by predictors
vars <- colnames(US_Accidents)
for (i in vars) {
  x <- US_Accidents %>%
    group_by_(i) %>%
    summarise(
      mean = mean(Severity),
      median = median(Severity),
      n=n()
    )
  print(x)
}


```
```

```
# Decision Tree Model
```{r}
library(rpart)
library(rpart.plot)
set.seed(123)

# Fit the model
tree1 <- rpart(Severity ~ Weather_Condition + Amenity + Bump + Crossing + Give_Way + Junction + No_Exit + Railway + Roundabout + Station + Stop + Traffic_Calming + Traffic_Signal + Sunrise_Sunset + Temperature.F. + Humidity... + Pressure.in. + Visibility.mi. + Wind_Speed.mph. ,
  data = US_Accidents,
  method = "anova",
  control = rpart.control(cp = 0.001, minbucket = 20)
)

tree1

# Plot the tree
rpart.plot(tree1)



```
```