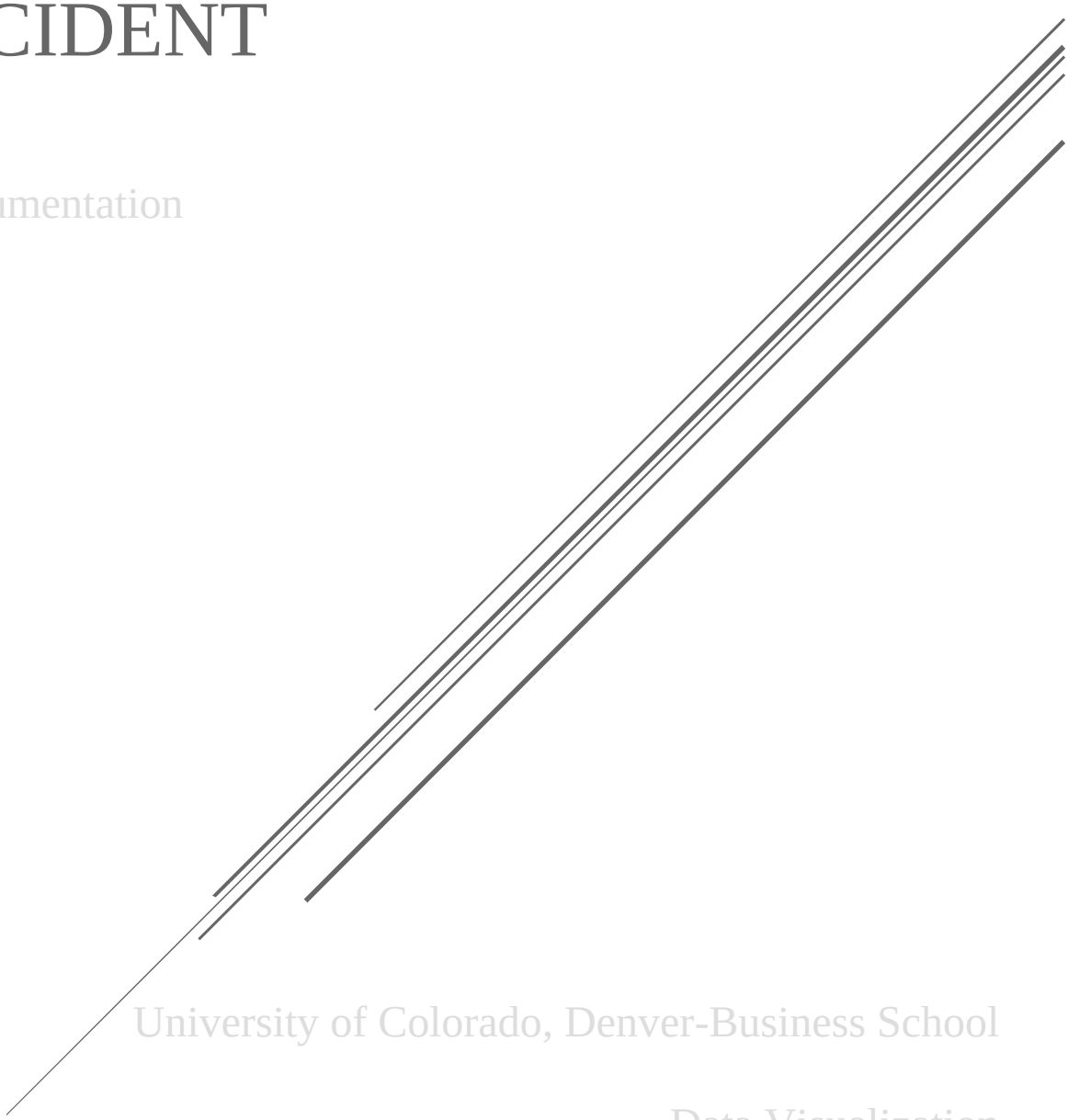
US ACCIDENTS (2016 - 2019)

COUNTRYWIDE TRAFFIC ACCIDENT

R Documentation



University of Colorado, Denver-Business School

Data Visualization

US Accidents (2016 - 2019)

Countrywide Traffic Accident

Urvish A. Patel

University of Colorado, Denver

Business School

Abstract

This document offers an analytical view on road accidents which helps the Department of Transportation, Road safety department, Hospitals & various response teams as they will be prepared for such accidents and will be knowing the exact measures to take beforehand an accident occur. Specifically, this document is a summary of findings through R-code which tells us statistics of the dataset which also answers various questions such as which State/Zip-code/County has the highest number of accidents? At what time/day/ do accidents usually occur in the US? What are the factors which causes road accidents? Predictions of accidents in future. This document lays out everything which might be helpful for safety of travelers.

Keywords: Exploratory Data Analysis, Time Series Forecasting, Regression Model, Geospatial Visualization

US Accidents (2016 - 2019)

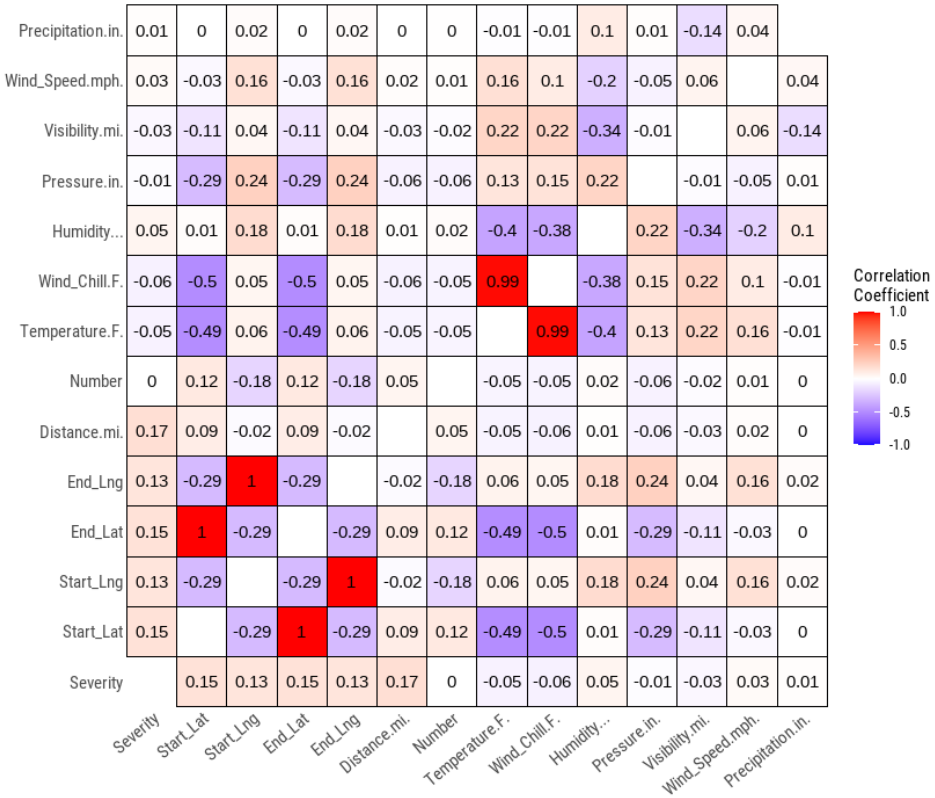
Countrywide Traffic Accident

The US Accidents dataset is a data reported by the relevant police department, US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks whenever an accident occurs. The dataset used covers 49 states of the USA. The accident data are collected from February 2016 to Aug 2019. They note down the exact location of accident and the distance of traffic which is affected after an incident occurs. The very first thing which Department of Transportation wants to know is the severity of an accident and how frequently an accident occurs during the day/week/time. How can a response team work efficiently given these circumstances? Weather plays a severe role in any type of accident. This analysis discloses the facts and relations about accidents which might be directly linked to the type of weather condition during an accident.

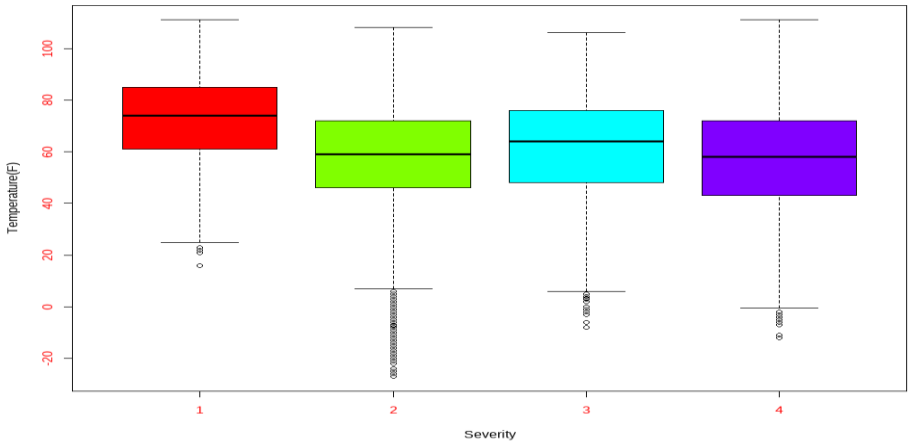
Understanding and cleaning Data

Very basic steps of EDA analysis is to first approach the data and analyze categorical, numerical, & both at the same time. After loading the dataset into R, all the values with “NA” is removed. As this may become outliers in future analysis. After removing NA from the original dataset which consisted of 1516064 rows shorted to 335552 rows. Now, there are instances

where your data rows is repeated, reason might be a machine error. So we check if we have any repeated rows. Understanding the summary of data frame helps us to know which columns are of what data type. Describe function of R gave me the statistics of each column. It shows mean, standard deviation, skewness, Kurtosis & Inter Quartile Range. R has this amazing function by which we can generate reports. After doing all the things mentioned above, I generated a normality report, which shows normality plot and along with the plots if we transform the columns into log or squared value. Correlation matrix is the best way to find out which columns is dependent in pairs. The below image shows us the same.



Now let's see if each severity is having a normal distribution or not. To prove this I created a Box Plot which is shown below.



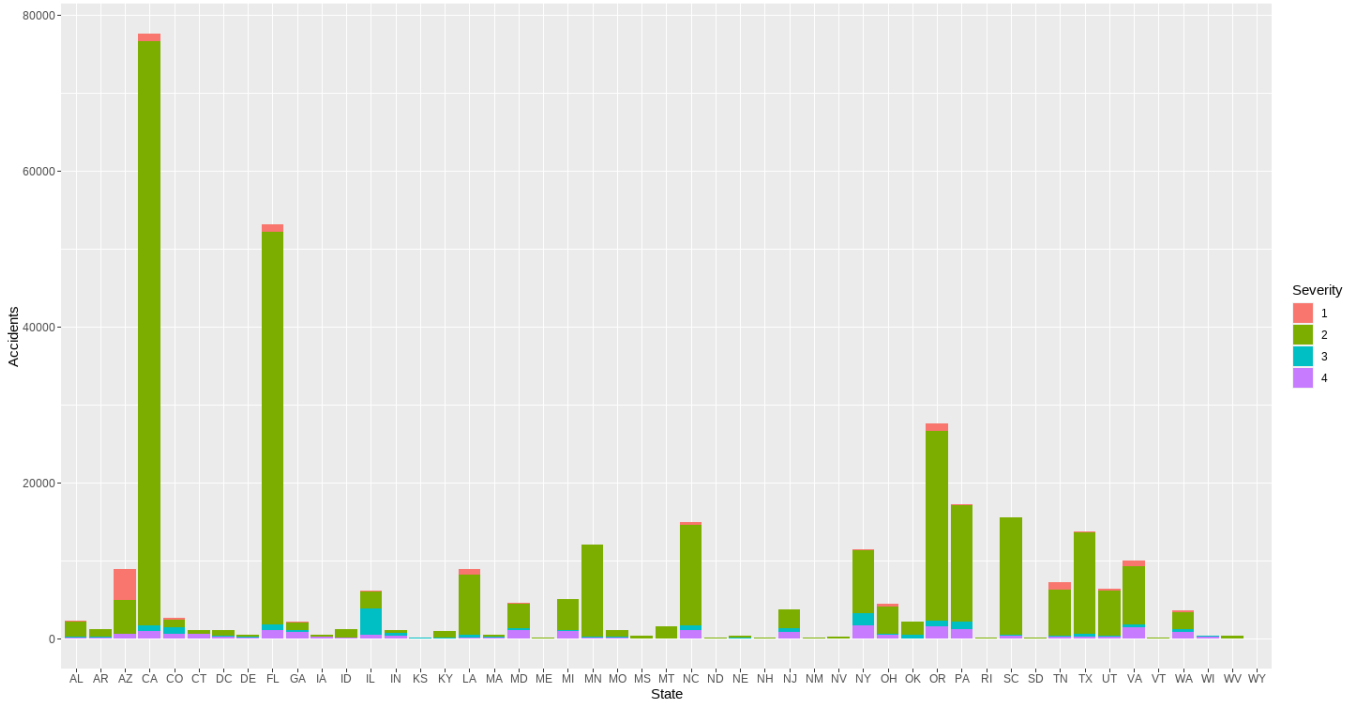
From this we can conclude that the Severities of 1, 2, 3 & 4 is normally distributed. As I can see there is hardly any skewness. The points shown is the outliers but these may not be considered as outliers as these can be useful in future analysis.

After plotting a correlation matrix I converted Boolean data type (True, False) to 1 and 0 so that I could build Logistic regression model for knowing the factors which causes accidents.

Common Visualizations.

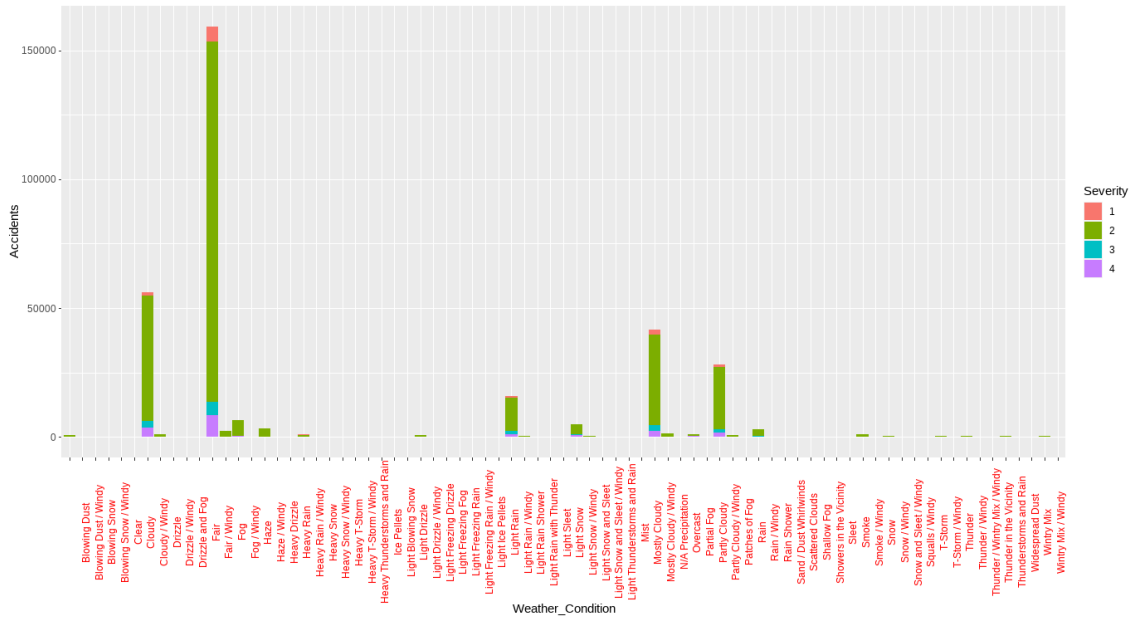
After going through cleaning part I created some visualizations which answers the following questions.

1. What is the count of accidents along with its severity in each state?



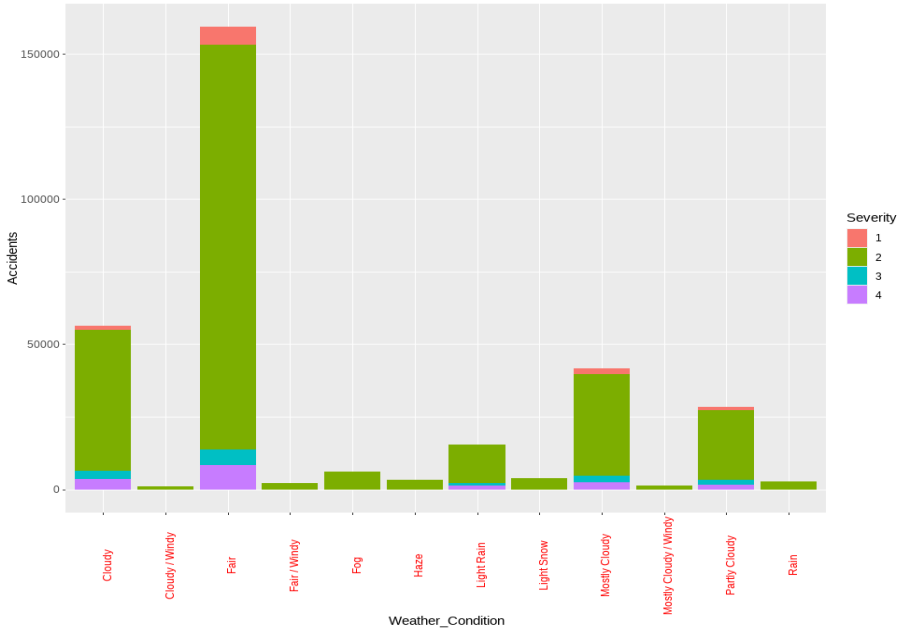
This shows CA with the highest number of road accidents with severity of 2. We can also conclude that most of the accidents are of severity 2.

2. In which Weather Condition do accidents occur most?



Most of the accidents occur when the weather is fair. Apart from that accidents also occur when the weather is cloudy, windy and having a light rain. Now this doesn't give us a clear view, this visualization is not good enough, so for that reason I created a plot with same weather conditions on x axis but where the accidents are greater than 1000. This answer's the following question.

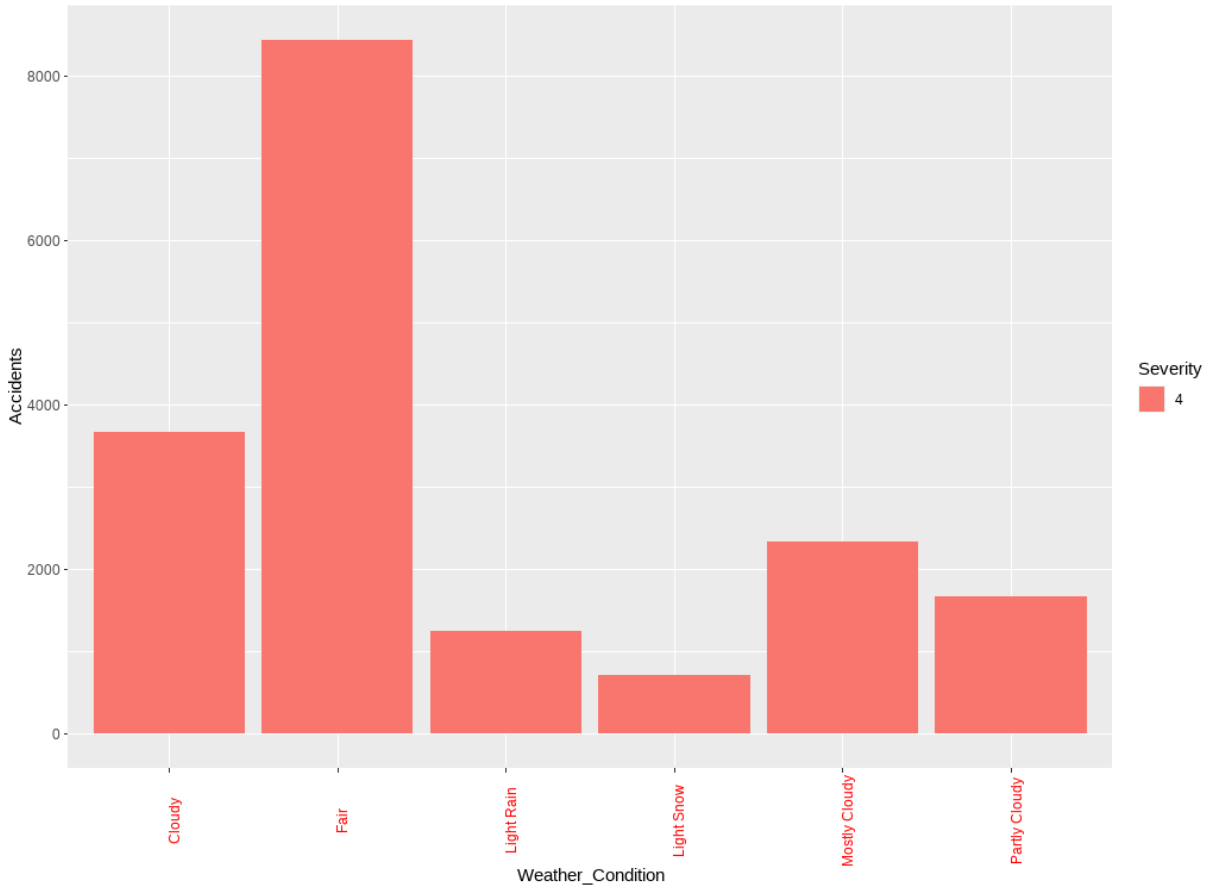
3. In which Weather Condition do accidents occur most where accidents are greater than 1000?



Now this shows us a clear picture. Sometimes it is hard to read any visualization when it contains more space.

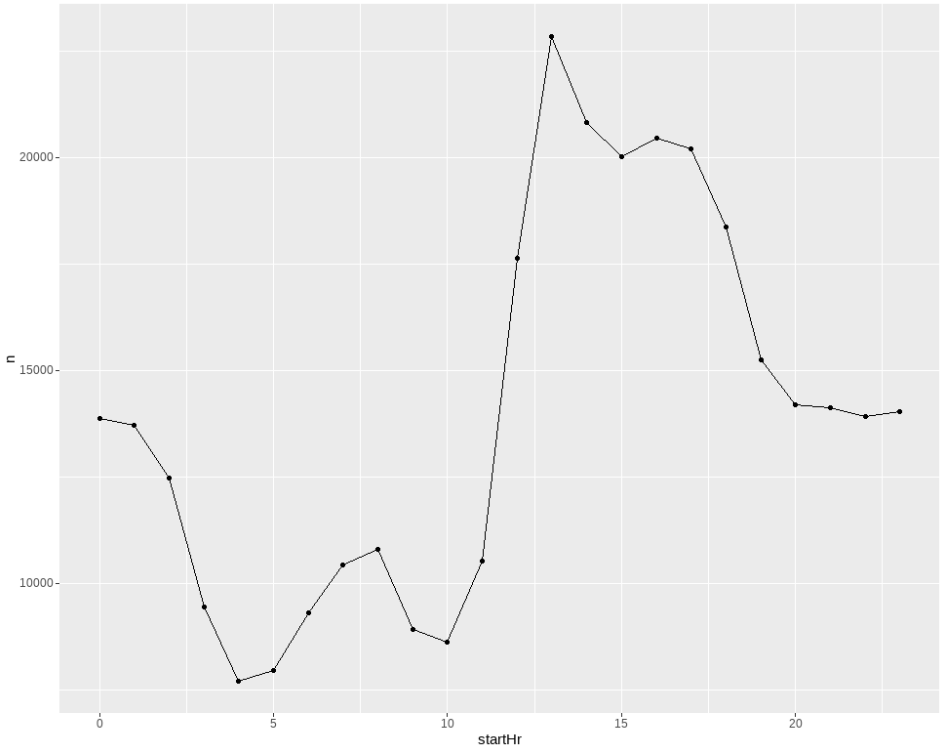
Now the area of focus is Severity 4, as this are with the most severe accidents. This answers the following question.

4. In which Weather Condition do accidents occur most where accidents are greater than 1000 and having severity of 4?



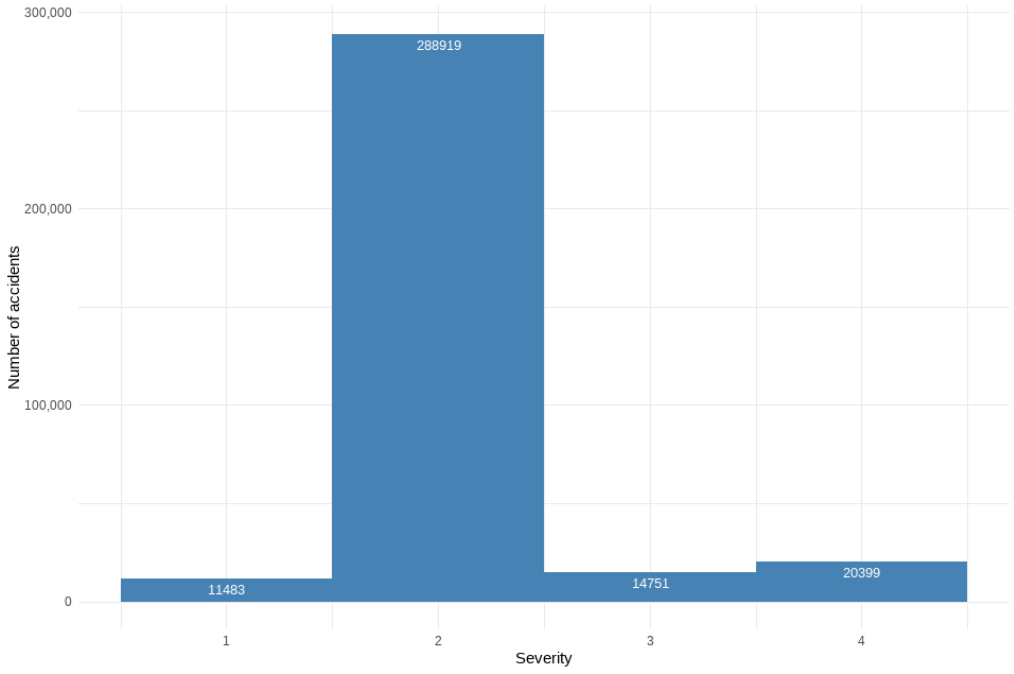
This concludes most of the accidents occur during fair weather condition. But the total sum of all others accidents with weather condition surpasses i.e. Accidents occur most when the weather is Cloudy, have a light rain & snow.

5. Usually at what hour most accidents occur in US?



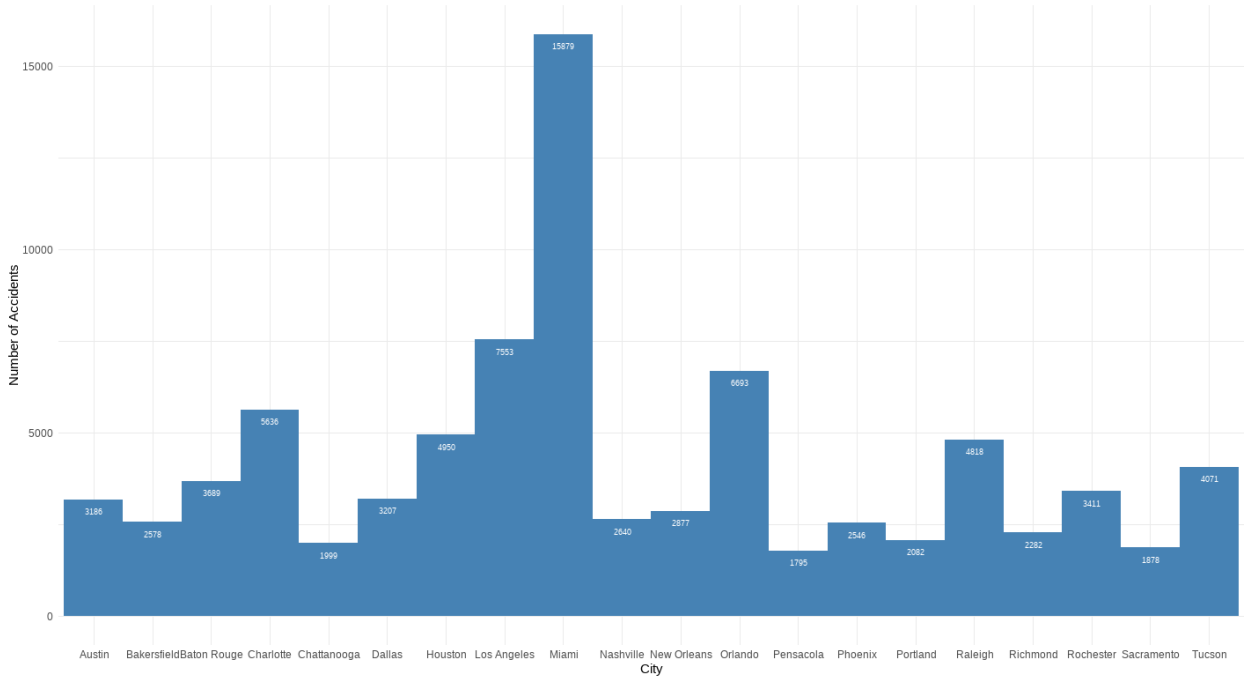
According to this plot we can conclude that most accidents occur between 11 hrs. To 15 hrs.

6. What type of severity accidents mostly occur?



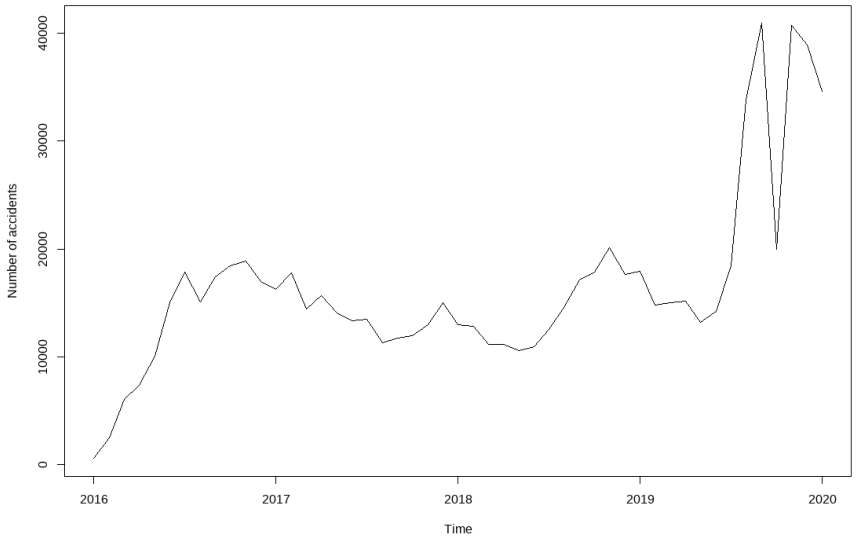
Severity of type 2 has the highest number of followed by severity 4, 3, & 1

7. What city has the highest number of accidents?

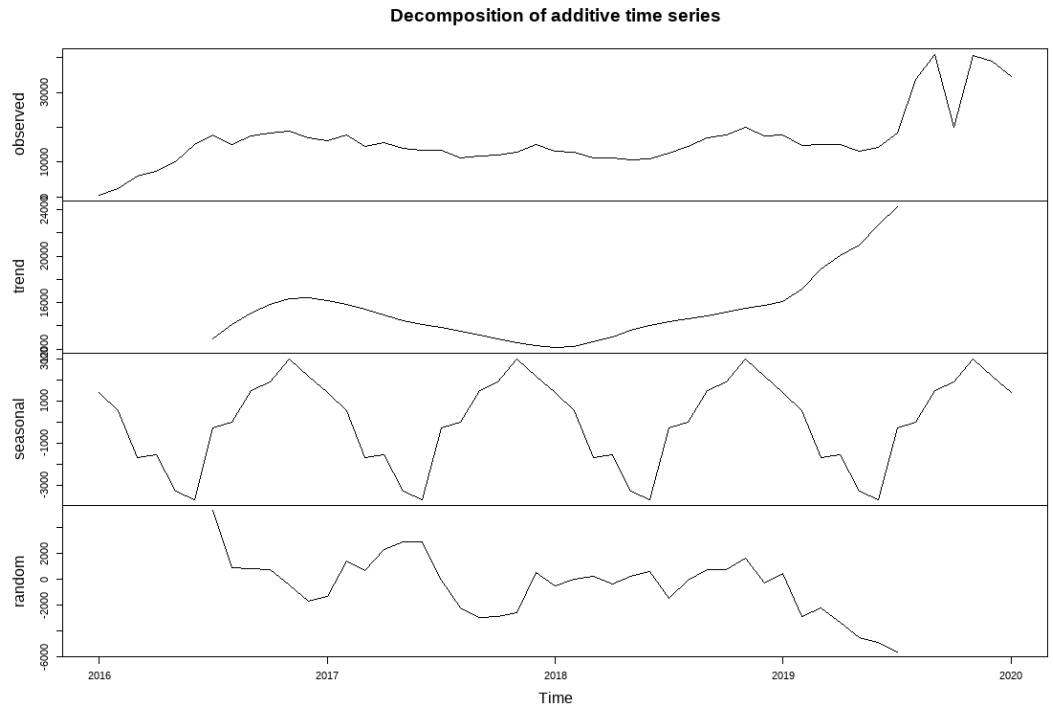


Miami has the highest number of accidents followed by Los Angeles, Orlando and many others.

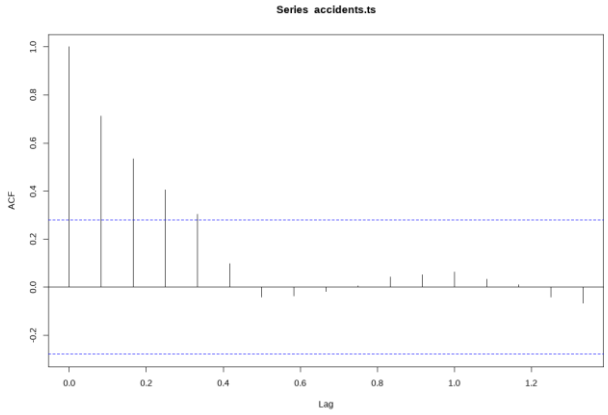
8. What would be a time series plot for number of accidents?

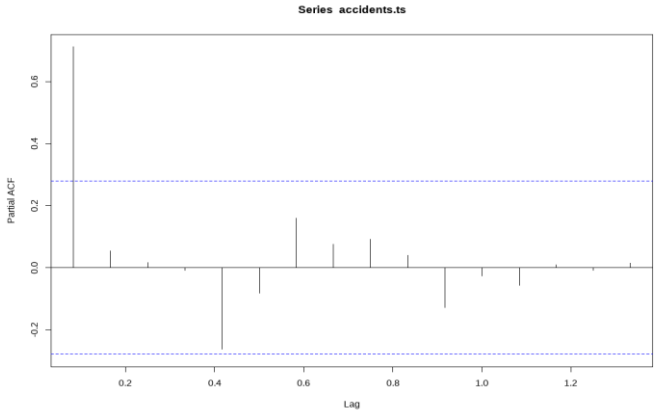


We can see that the trend of number of accidents that occur each year is in positive trend. To see clearly what type of trend is there both seasonally and randomly I plotted a decomposition plot, which looks like the image below.

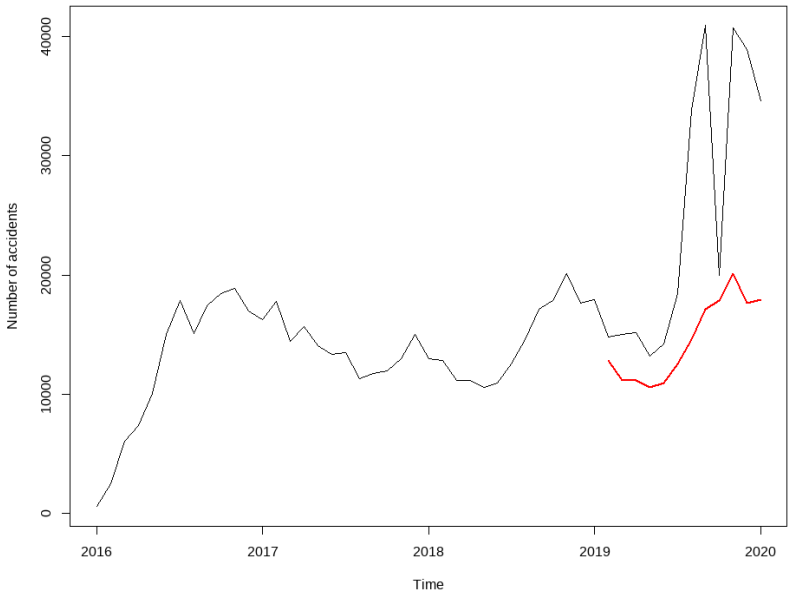


From this plot we can conclude that the trend is positive if we see the trend seasonally it is also a positive trend, but the randomly selected data is showing a negative trend. I have also created an ACF & PACF plots for knowing the autocorrelation and the type of trend.



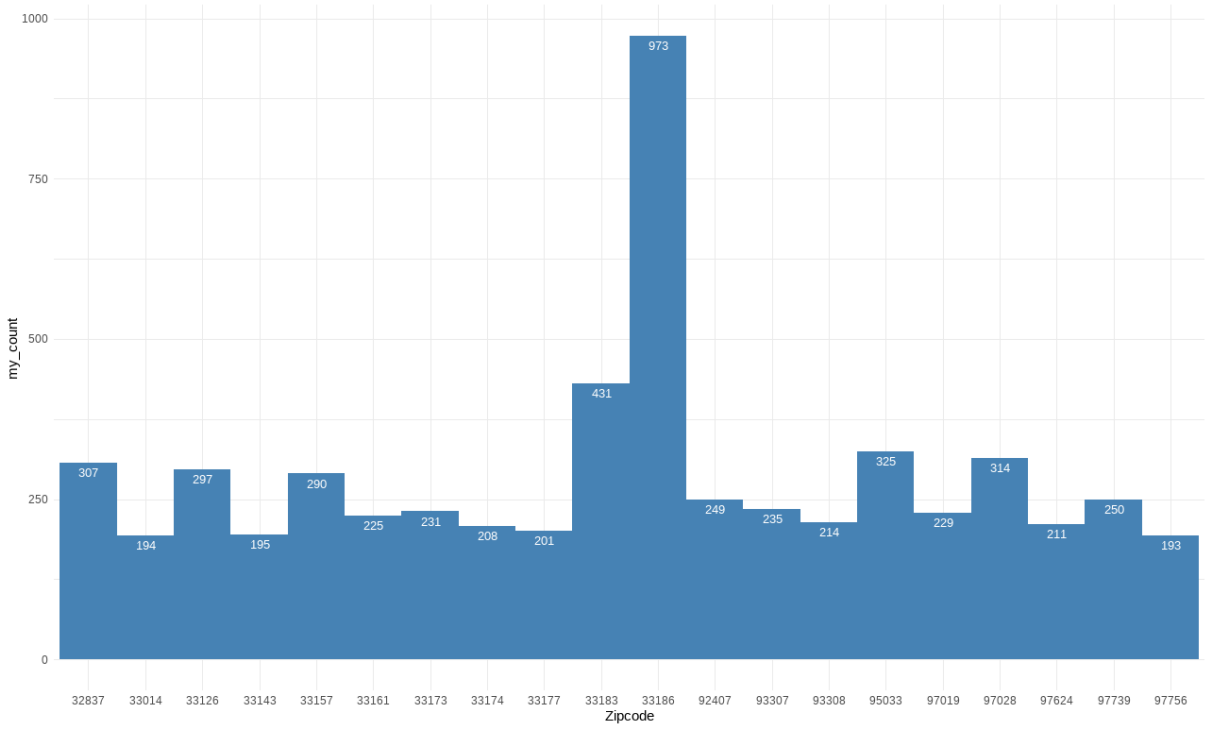


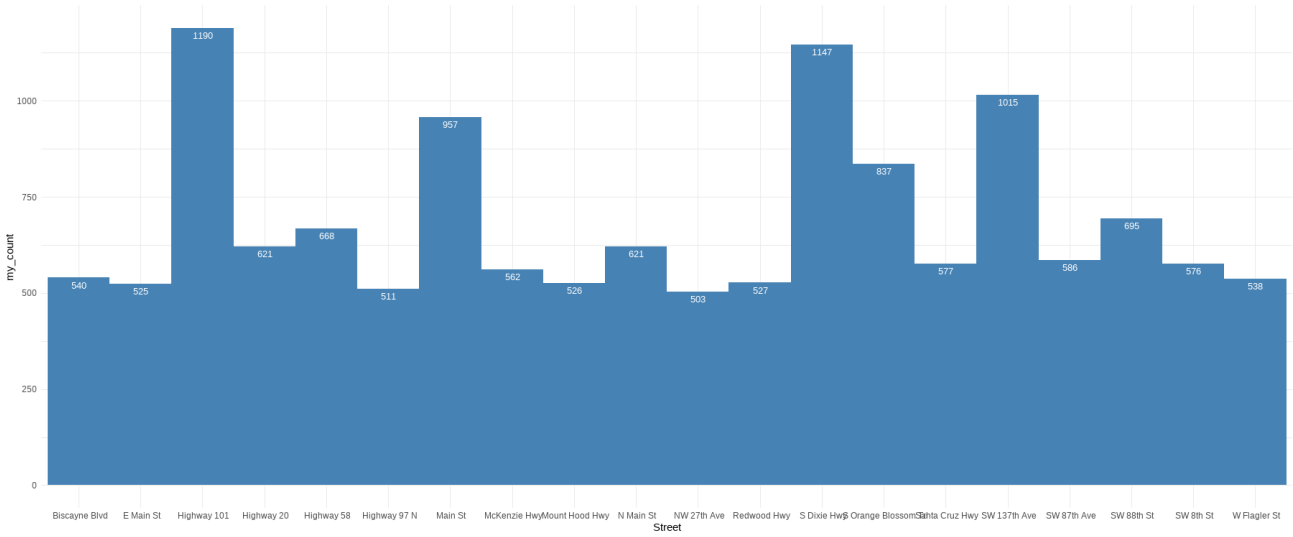
9. What are the predictions if we use naïve model if we divide the dataset into train and test data?



From this we can conclude how good our naïve model is we can see the predictions in red line and can compare with actual dataset.

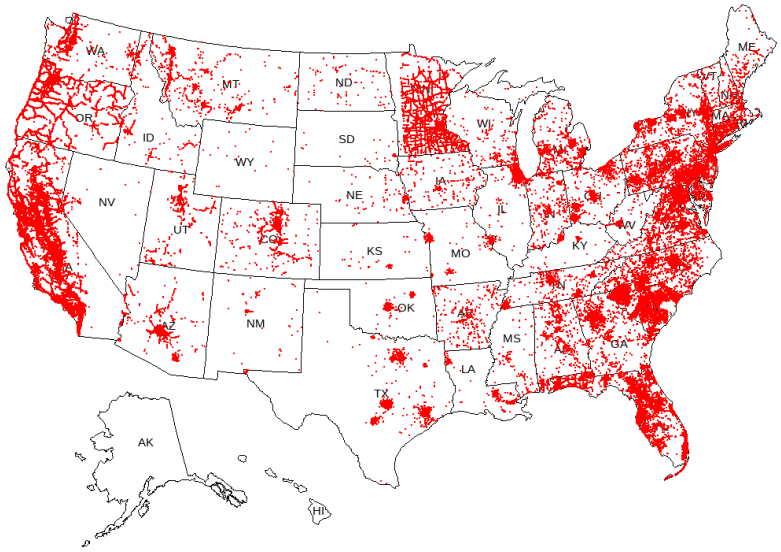
10. Which Zip code and street of US has the highest number of accidents?





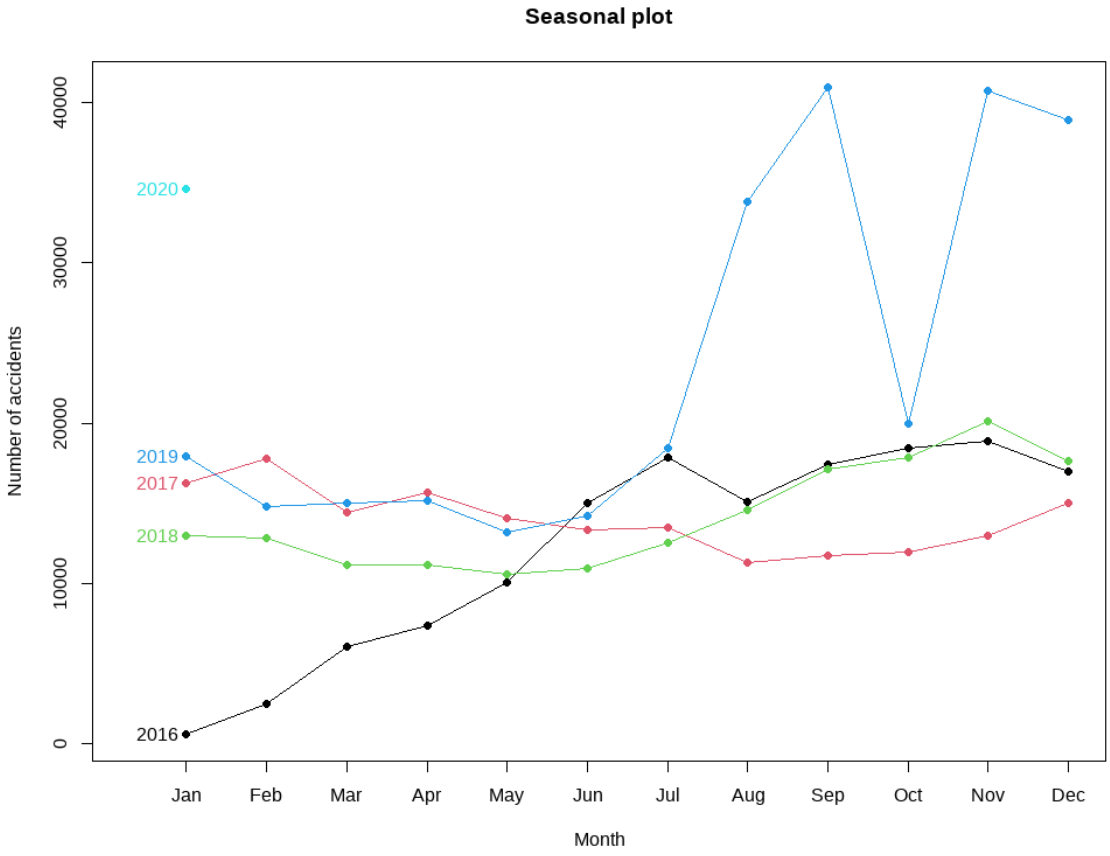
From this we can conclude that Zip code with 33183 & Highway 101 has the highest number of accidents. This area can be declared as accident prone zone and we can provide a better response team. We can improve the roads and Lights.

11. General view on number of accidents in US using map.



From this we can conclude that most of the accidents reported are from either east side or west side.

12. What is the comparison between different years with respect to count of accidents?



This shows the monthly number of accidents with respect to the year. We can see the trends of these lines. Surely, in 2016 there was a positive trend. Year 2017 may be concluded as a negative trend as the count of accidents in this year are in decreasing numbers.

Logistic Regression Model.

From this model we will be able to identify the actual causes of road accidents. The sole objective of this model is to find out factors which affect the severity of road accidents. Upon using and trying many models I found one bet model with good P-value. Following is the image of Logistic Regression model.

```
> summary(best_logistic_model)

Call:
glm(formula = as.factor(Severity) ~ Temperature.F. + Civil_Twilight +
    Nautical_Twilight + Astronomical_Twilight + Humidity... +
    Pressure.in. + Wind_Speed.mph. + Precipitation.in., family = binomial,
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5991   0.0983   0.1780   0.3018   1.0246

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.1583568  0.2064055   5.612    2e-08 ***
Temperature.F.       -0.0328198  0.0007383 -44.454  < 2e-16 ***
Civil_Twilight       -0.1880180  0.0583786  -3.221 0.001279 **
Nautical_Twilight    -0.2986898  0.0892893  -3.345 0.000822 ***
Astronomical_Twilight -0.9932627  0.0782458 -12.694  < 2e-16 ***
Humidity...           0.0186887  0.0005022  37.215  < 2e-16 ***
Pressure.in.          0.1526831  0.0073554  20.758  < 2e-16 ***
Wind_Speed.mph.       0.0062393  0.0018171   3.434 0.000595 ***
Precipitation.in.    -0.4012353  0.1155659  -3.472 0.000517 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 100077  on 335551  degrees of freedom
Residual deviance:  85670  on 335543  degrees of freedom
AIC: 85688

Number of Fisher Scoring iterations: 8
```

From this we can conclude that Severity of an accident mostly depends on Temperature, Civil Twilight (Beginning in morning or ends in evening), Humidity, Pressure, Wind speed and precipitation. So it is most like that an accident might occur if these conditions are fulfilled.

Recommendations.

After analyzing and visualizing the data I will recommend the following things which may reduce the number of road accidents.

- The response teams and hospitals must be given special provisions in the hours in which most accidents are occurred.
- Warning signs about speed limits are to be put in the accident prone streets
- The state with highest accidents must be provided with better resources and budget plans to avoid accidents and rescue the victims.
- Warnings are to be put depending on the weather conditions which cause accidents.
- A mandate of vehicles to have first aid kit should be passed.

- Online surveillance for prompt response from emergency services should be implemented.
- To have enough response teams to rescue in accident prone locations

References

https://medium.com/analytics-vidhya/visual-outputs-in-r-example-1-d71cb4be50eb

https://youtu.be/HPJn1CMvtmI

https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html

http://r-statistics.co/ggplot2-Tutorial-With-R.html

https://towardsdatascience.com/exploratory-data-analysis-in-r-for-beginners-fe031add7072

https://blog.datascienceheroes.com/exploratory-data-analysis-in-r-intro/

Dhakal, 2018. *A Naïve Approach for Comparing a Forecast Model*.

Paul S.P. Cowpertwait, 2009. Introductory Time Series with R (Use R!)

Hrishi V. Mittal, 2011. R Graphs Cookbook