R project - US Traffic Accidents

GROUP 9 - Amber Akhtar, Amruta Bhuskute Yashwant, Darshni Vora, Sahil Raju Shah

November 25th, 2020

Executive Summary:

Traffic accidents have been a major cause of concern across the United States for many years and it comes with the horror of increasing death rates. Despite latest models of cars coming out with new safety features aimed at reducing accidents such as rearview cameras, technology to prevent skids, lane change monitors and airbags these safety measures have frankly failed to stop accidents (Boudette, 2017). According to the estimates of a non-profit organization, National Safety Council, in 2016 alone 40,200 people have died which is a sharp increase from last year, many speculations have been made that the reason could stem from lack of attention during driving due to cell phone usage of apps such as Facebook, Snapchat and Google Maps etc. (Boudette, 2017). Also many states have reduced the number of state troopers patrolling the streets to catch speeders as well as the fact that seat belts usage is also not being enforced (Boudette, 2017).

Our project is aimed at analyzing the US traffic accidents data to identify what are the causes of road accidents. This dataset includes information on weather conditions, temperature, wind_chill, visibility as well as road fixture information including any bumps, crossing signals and railroad crossings etc. during the time the accidents took place. It also shows the cities and states where the accidents are occurring. We formulated our hypothesis based on this information of whether weather conditions played a role in the accidents and in which regions of the country are the most accidents occurring. In addition to identifying the reasons for the accidents we also wanted to identify whether there are enough hospitals in the areas to accommodate the increasing number of accidents. If there are shortages we want to make recommendations to the local government to consider adding more hospitals that in turn could help save lives.

We conducted a number of exploratory analyses on the data that helped in our hypothesis testing which included identifying which US coasts had the most accidents by longitude and latitude, time zone, city and state, identifying the most severe accidents by city and state. We also explored the number of accidents by weather condition type so which condition saw the highest number of accidents as well as identifying the most severe accidents by time of day i.e. day or night.

Through our analysis of the accidents and hospital dataset our aim was to build a model that could identify what relationship Severity had with the other variables so that we could identify the most significant variables that can be used to categorize accidents as most severe to not severe. We used several algorithms such as linear regression, linear discriminant analysis and random forest regression to help identify which model can best meet our objective.

Dataset Description: US Traffic Accidents – A Countrywide Traffic Accident Dataset (2016-2020)

To conduct our analysis we used a dataset related to accidents in the United States covering all 49 states during the time of February 2016 to June 2020 ([link to the dataset])(https://www.kaggle.com/sobhanmoosavi/us-accidents).

The dataset consists of 3.5 million records with 49 variables. The dataset contains information about the various accidents, the variables describing the accidents include:

- The start and end times of the accidents
- The longitude and latitude of the location of the accidents
- The severity of the accidents ranging from 0 – 4; 4 being most severe
- Street number, name, city, county, state, zipcode and country where the accidents have occurred.
- Time zone shows the time zone in which the accidents have occurred.
- There are weather related variables including: temperature, visibility, wind chill, wind direction, pressure, humidity, weather condition i.e. rain, snow etc. that happened at the time of the accidents.
- There are road fixture related variables such as bump (speed bump), railway crossing, roundabout, traffic signal, stop sign etc. at the location of the accidents.
- Time of the day variables including sunrise_sunset, astronomical twilight, nautical twilight and civil twilight. These indicated whether the accident occurred during day or night hours.

USA hospitals – Hospitals across the USA: https://www.kaggle.com/carlosaguayo/usa-hospitals

The second dataset is the hospital dataset from Kaggle which contains location information about hospitals in each of 50 states of US. The dataset consists of 34 columns and around 1 million records. The variables of the dataset describe the hospitals around the country including:

- Hospital name, address, city, county, state, zip and telephone numbers.
- Number of beds available in that hospital
- Status of the hospital whether open or closed.
- Type of hospital – the hospitals are categorized as either general acute, children, special, psychiatric, critical access, long term care, military, women etc.
- Trauma – whether the hospital has trauma facilities or not

Data Cleaning:

So a number of challenges were discovered within the dataset and had to be rectified. This dataset contains traffic related information from 2016 to 2020 we decided to focus on the data from years 2019 to 2020. So we removed observations from 2016 to 2018. We also removed certain columns and kept those columns that were relevant to our analysis that were relevant to our hypothesis testing. Since we are analyzing the areas of the US with high traffic rates and how the weather conditions are impacting the rate of accidents. So based on this we decided to keep the following columns:

1. Severity – how the severity of accidents impacted the traffic. 1 least severe – 4 most severe.
2. Start Time – start time of the accident according to local timezone
3. End Time – end time of the accident according to local timezone
4. Start_Lat – shows the latitude in GPS of the start point
5. Start_Lang – shows the longitude in GPS of the start point
6. City
7. County
8. State
9. Time zone – shows the time zone of the location of the accident
10. Temperature (F) – shows the temperature in Fahrenheit
11. Visibility
12. Weather Condition – rain, snow, fair etc.
13. Amenity – the presence of amenity in a nearby location
14. Bump – presence of a speed bump or hump in the accident location
15. Crossing
16. Give way
17. Junction
18. No exit
19. Railway
20. Roundabout
21. Station
22. Stop
23. Traffic calming
24. Traffic signal
25. Sunrise_Sunset

```
##   Severity    Start_Time      End_Time Start_Lat Start_Lng       City
## 1        3 6/30/20 23:18 6/30/20 23:47     38.19    -85.77 Louisville
## 2        3 6/30/20 22:56 6/30/20 23:26     32.77    -96.86     Dallas
## 3        2 6/30/20 22:52 6/30/20 23:22     36.10    -86.74  Nashville
## 4        2 6/30/20 22:52 6/30/20 23:17     32.84    -96.81     Dallas
## 5        3 6/30/20 22:51 6/30/20 23:20     41.81    -71.40 Providence
##       County State   Timezone Temperature.F. Visibility.mi. Weather_Condition
## 1  Jefferson    KY US/Eastern             79             10        Light Rain
## 2     Dallas    TX US/Central             85             10     Mostly Cloudy
## 3   Davidson    TN US/Central             74             10     Mostly Cloudy
## 4     Dallas    TX US/Central             85             10     Mostly Cloudy
## 5 Providence    RI US/Eastern             67              9     Mostly Cloudy
##   Amenity  Bump Crossing Give_Way Junction No_Exit Railway Roundabout Station
## 1   FALSE FALSE    FALSE    FALSE    FALSE   FALSE   FALSE      FALSE   FALSE
## 2   FALSE FALSE    FALSE    FALSE    FALSE   FALSE   FALSE      FALSE   FALSE
## 3   FALSE FALSE    FALSE    FALSE    FALSE   FALSE   FALSE      FALSE   FALSE
## 4   FALSE FALSE    FALSE    FALSE    FALSE   FALSE   FALSE      FALSE   FALSE
## 5   FALSE FALSE    FALSE    FALSE    FALSE   FALSE   FALSE      FALSE   FALSE
##    Stop Traffic_Calming Traffic_Signal Turning_Loop Sunrise_Sunset
## 1 FALSE           FALSE          FALSE        FALSE          Night
## 2 FALSE           FALSE          FALSE        FALSE          Night
## 3 FALSE           FALSE           TRUE        FALSE          Night
## 4 FALSE           FALSE          FALSE        FALSE          Night
## 5 FALSE           FALSE          FALSE        FALSE          Night
```

We also found that various columns have missing values so we removed the NA values from those columns.

Exploratory Data Analysis:

The purpose of our analysis is to build a model that focuses on the states where the rate of accidents are highest and are the most severe. We will use this model in collaboration with the information from the hospital dataset to identify which states are lacking hospital beds so that we can make recommendations to the local government to increase resources and finance to accommodate more beds in those states with the highest severe accidents and help save precious lives. To meet this goal we need to focus on the following data requirements:

1. Data about accidents happening in the east coast and west coast
2. Data about the hospitals where the accidents are occurring.
3. Data about the weather conditions in those states/counties.
4. Data about the road conditions in those states/counties.

Based on our findings we have two null hypothesis that we are looking to analyze using this dataset:

1.Null Hypothesis: Accidents happening on the east coast are high.

2.Null Hypothesis: Accidents happening during extreme weather conditions are higher.

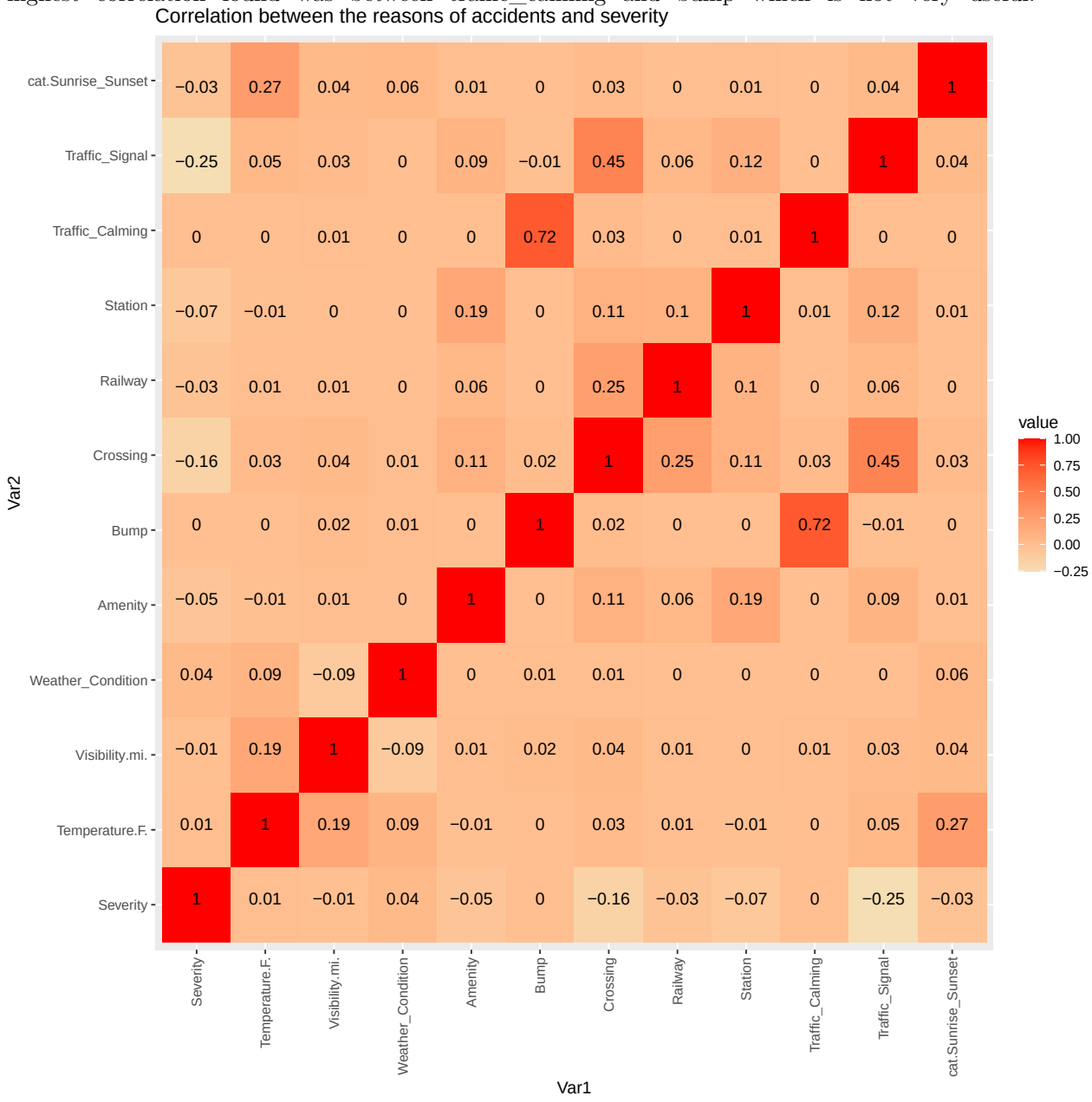
To help analyze our hypothesis we conducted a number exploratory analysis on the data.

We created a correlation matrix. In order to create the correlation matrix we first created categorical variables for the following variables:

1. Sunrise_Sunset – converted Day to 1 and Night to 0
2. Weather_Condition – converted to factor then to numeric for the various weather conditions.
3. Converted Boolean variables to factor then to numeric:
4. Amenity – True is 1 and False 0; Bump, Crossing, Give_way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop

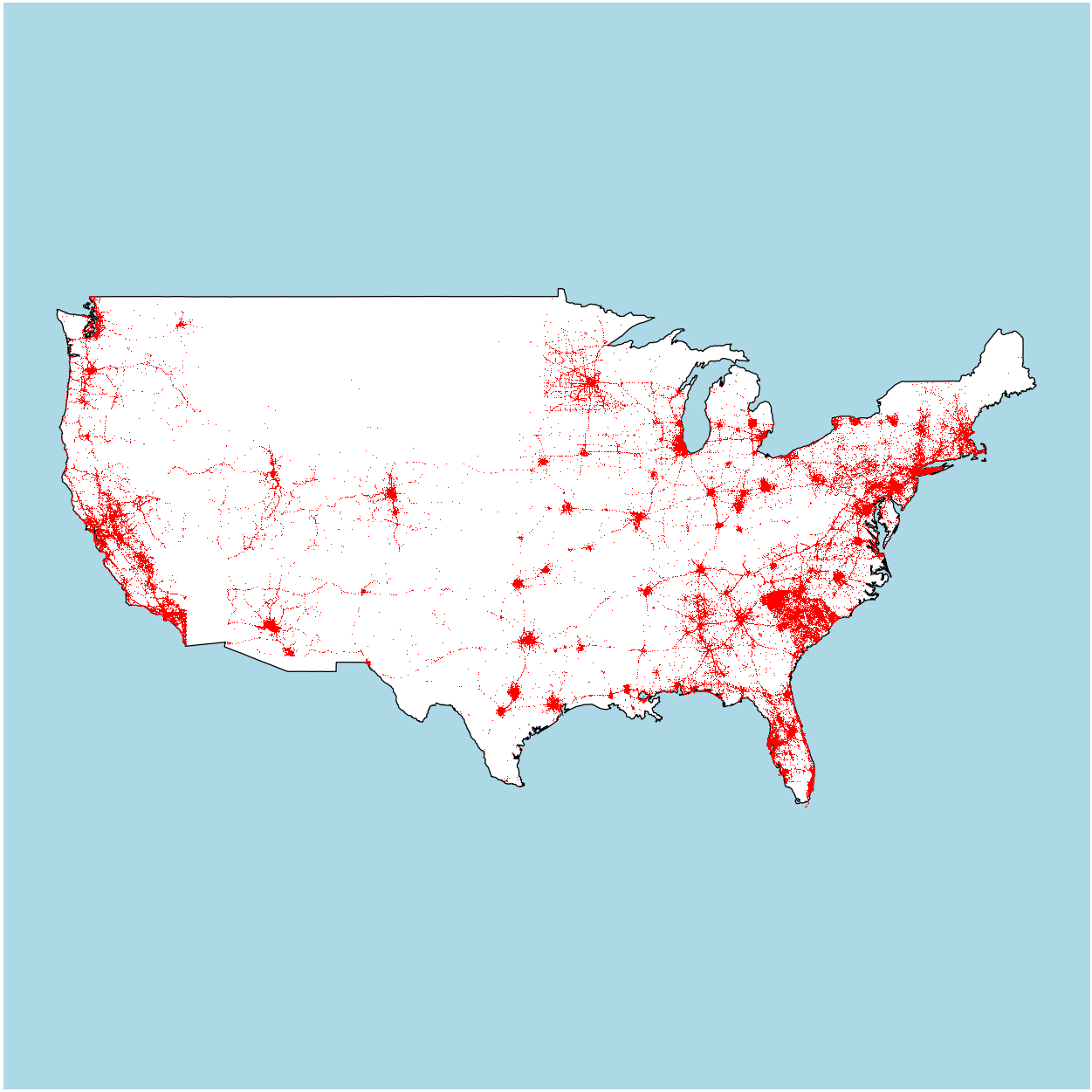
Correlation Matrix

After converting all the variables we created a correlation matrix to see whether there was any relationship or correlation between severity and the other variables that could indicate reasons for accidents. Unfortunately the correlation matrix did not show any useful relationship the highest correlation found was between traffic_calming and bump which is not very useful.

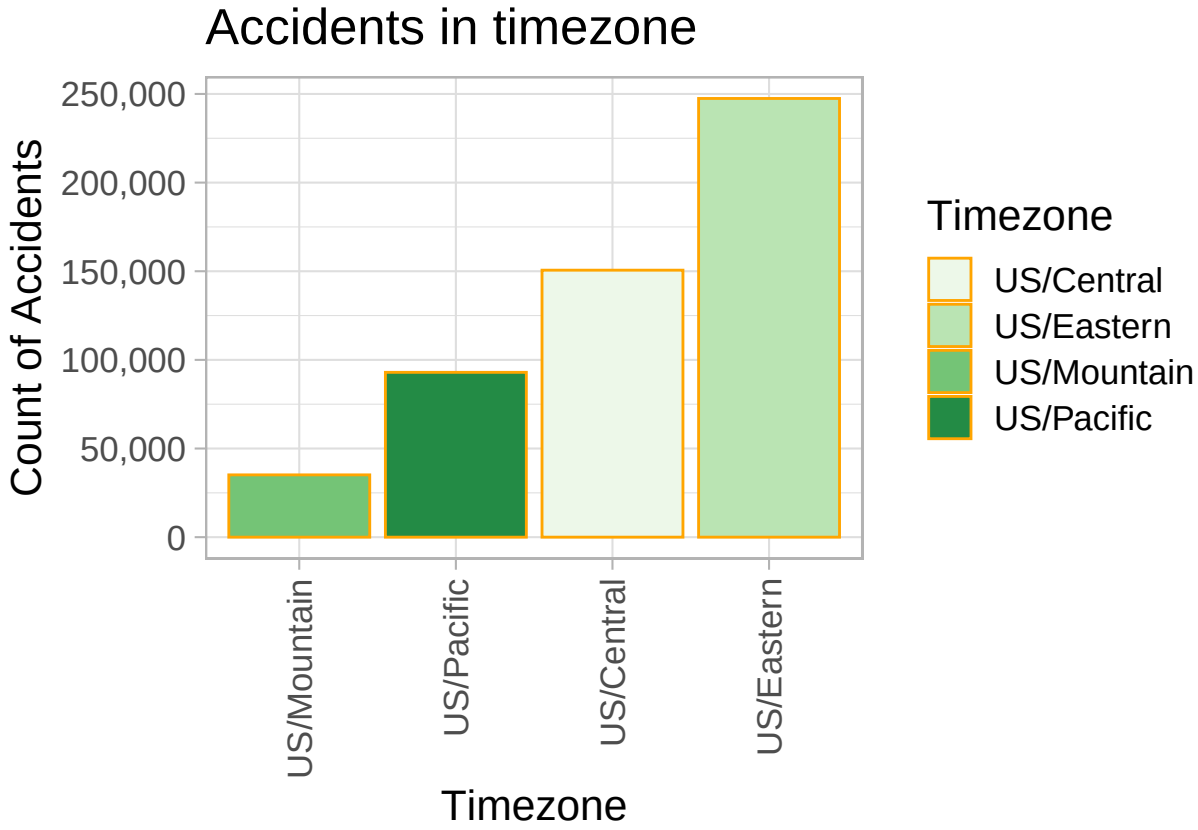


US MAP accidents locations:

Next we decided to identify which coasts had the highest rate of accidents so we created a US map using the Start_Lng and Start_Lat variables and found the most of the accidents occurring were on the east coast which we also confirmed by the bar chart we created based on time zone. This chart showed that the count of accidents in the US/Eastern time zone was highest.



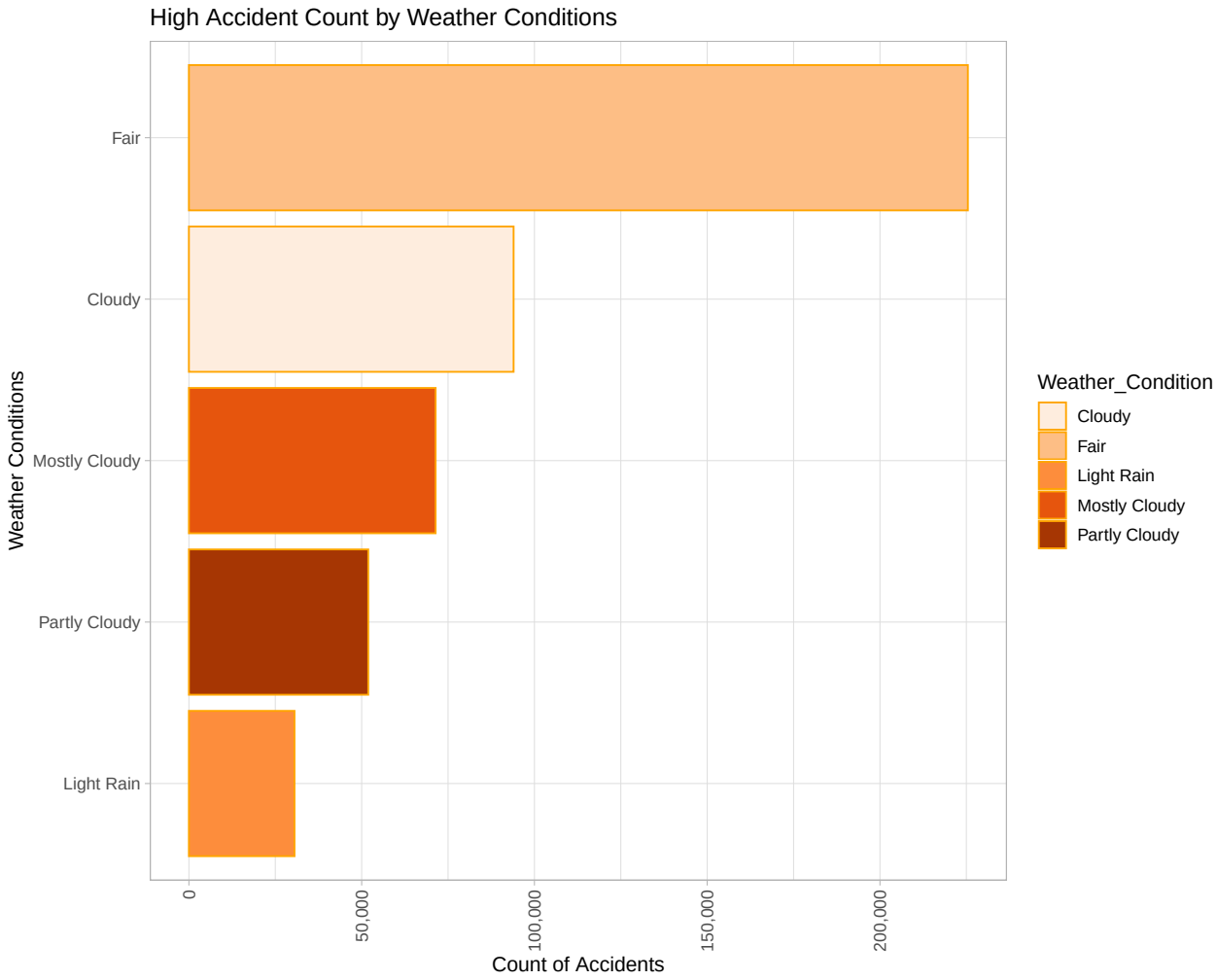
Accidents in different Timezone:

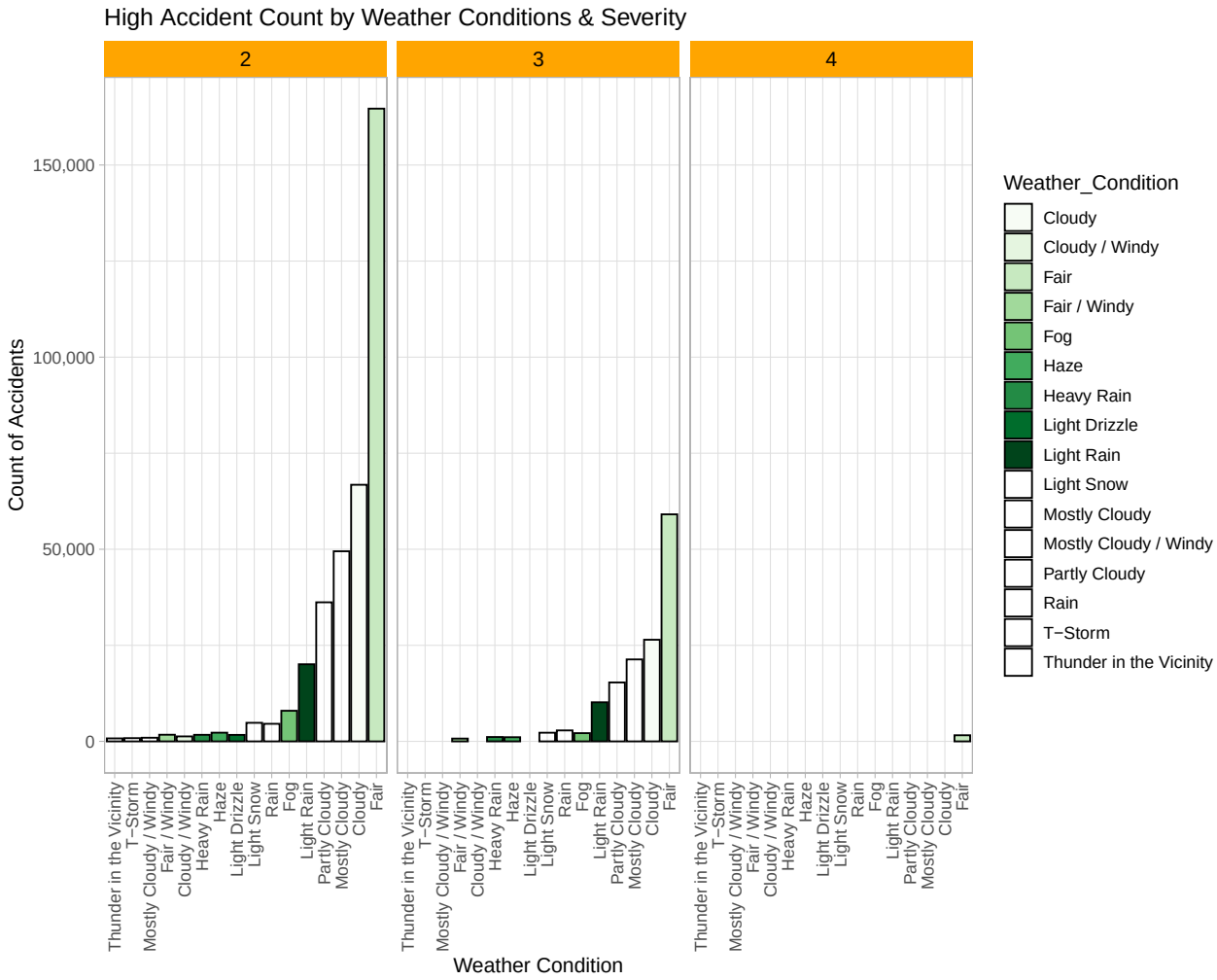


Most number of accidents happen in the Eastern time zone followed by Central, Pacific and Mountain in the time-frame of Jan 2019- June 2020.

Accidents by Weather conditions and severity:

We also wanted to analyze the effect of weather conditions on the accidents and this bar chart showed the highest number of accidents occurred during fair weather conditions. So from this chart we see that bad weather conditions didn't necessarily lead to high accidents rates.

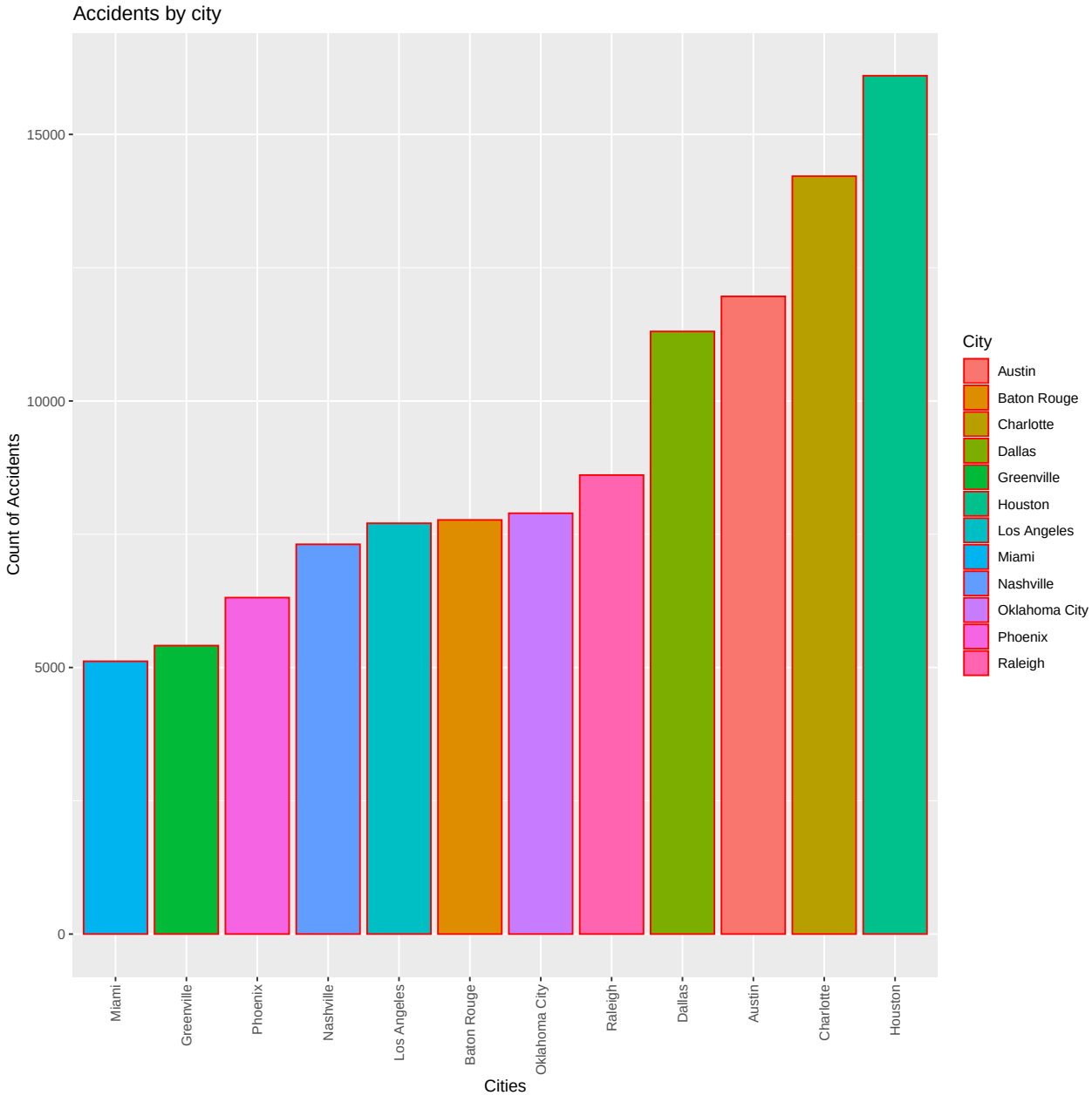


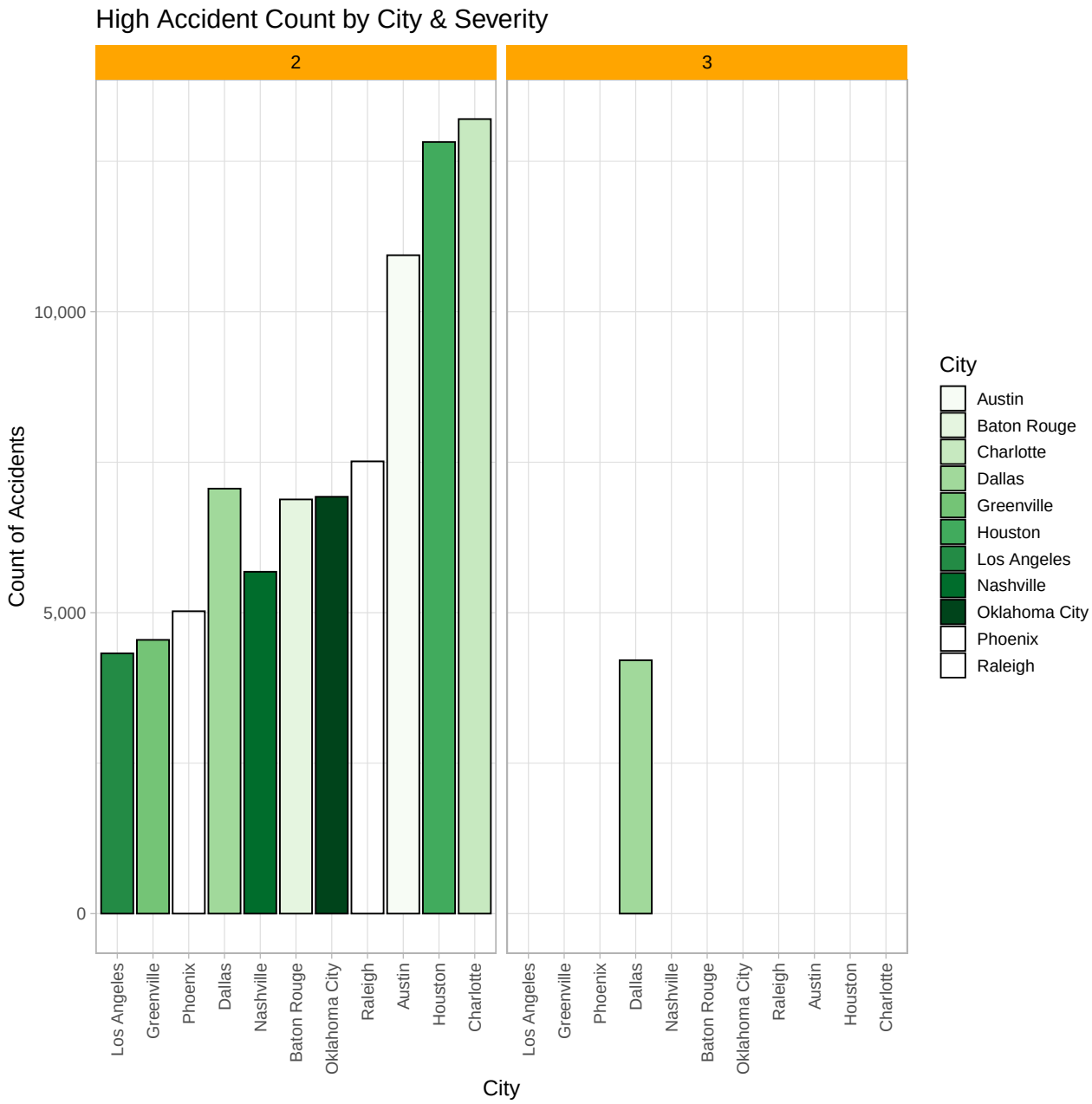


We also wanted to look at the severity of the accidents in relation to weather conditions so we created a bar chart in which we grouped by Weather_Condition and Severity variables. According to the dataset most of the accidents that happened were in severity level of 2 and 3. So we took those 2 levels and created a bar chart by weather condition and severity variables and found the count of accidents was highest in fair weather conditions for both severity levels of 2 and 3. We even noticed that severiyt 4 has very few observation in fair weather condition.

City and Severity analysis of Accidents

We also wanted to analyze the rate of accidents by city so we created a bar chart to identify which city has the highest number of accidents, which was Houston. However, our purpose of this analysis is to find those states which have the most severe accidents to identify if they have enough hospitals to accommodate these accidents.

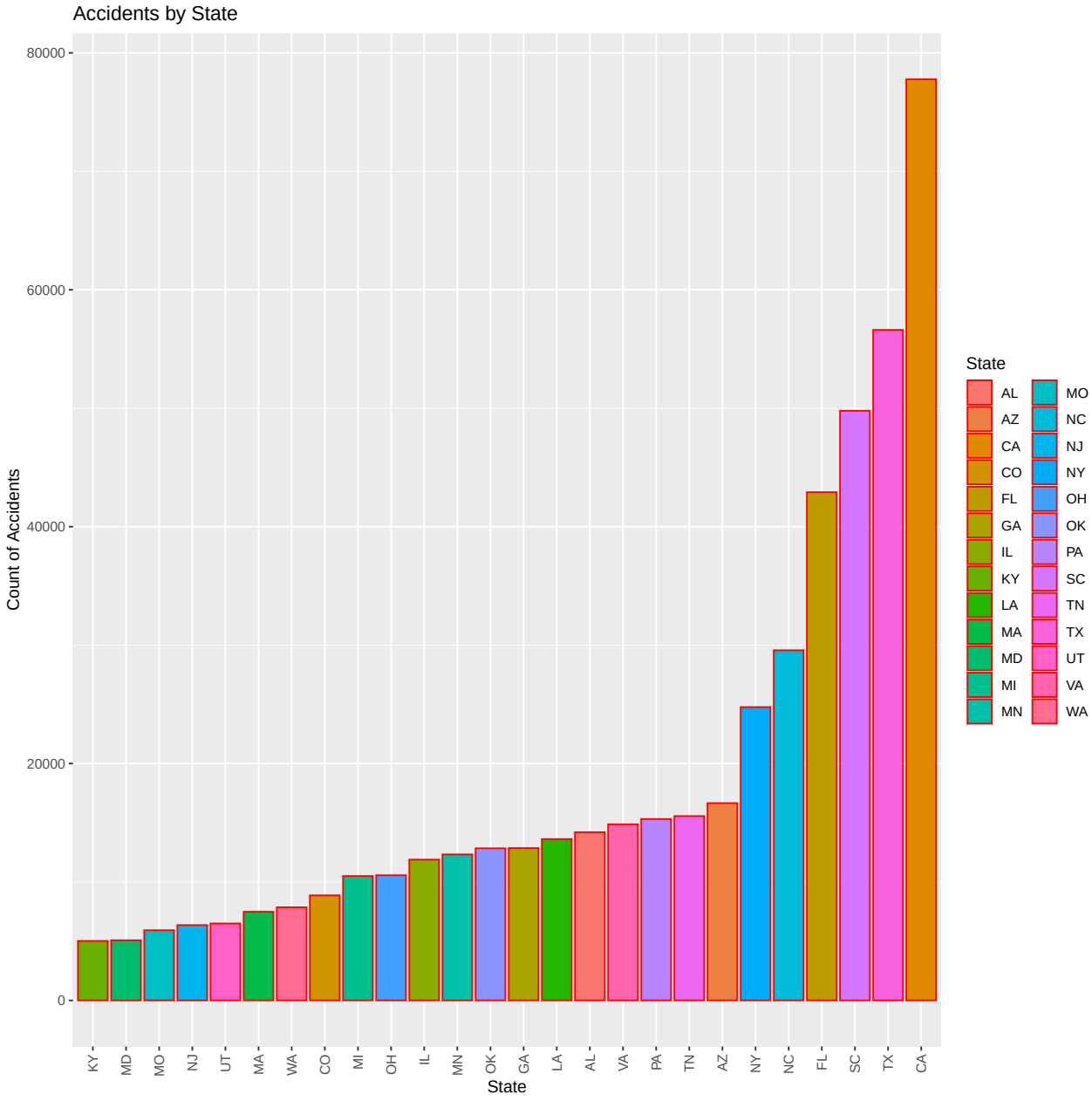


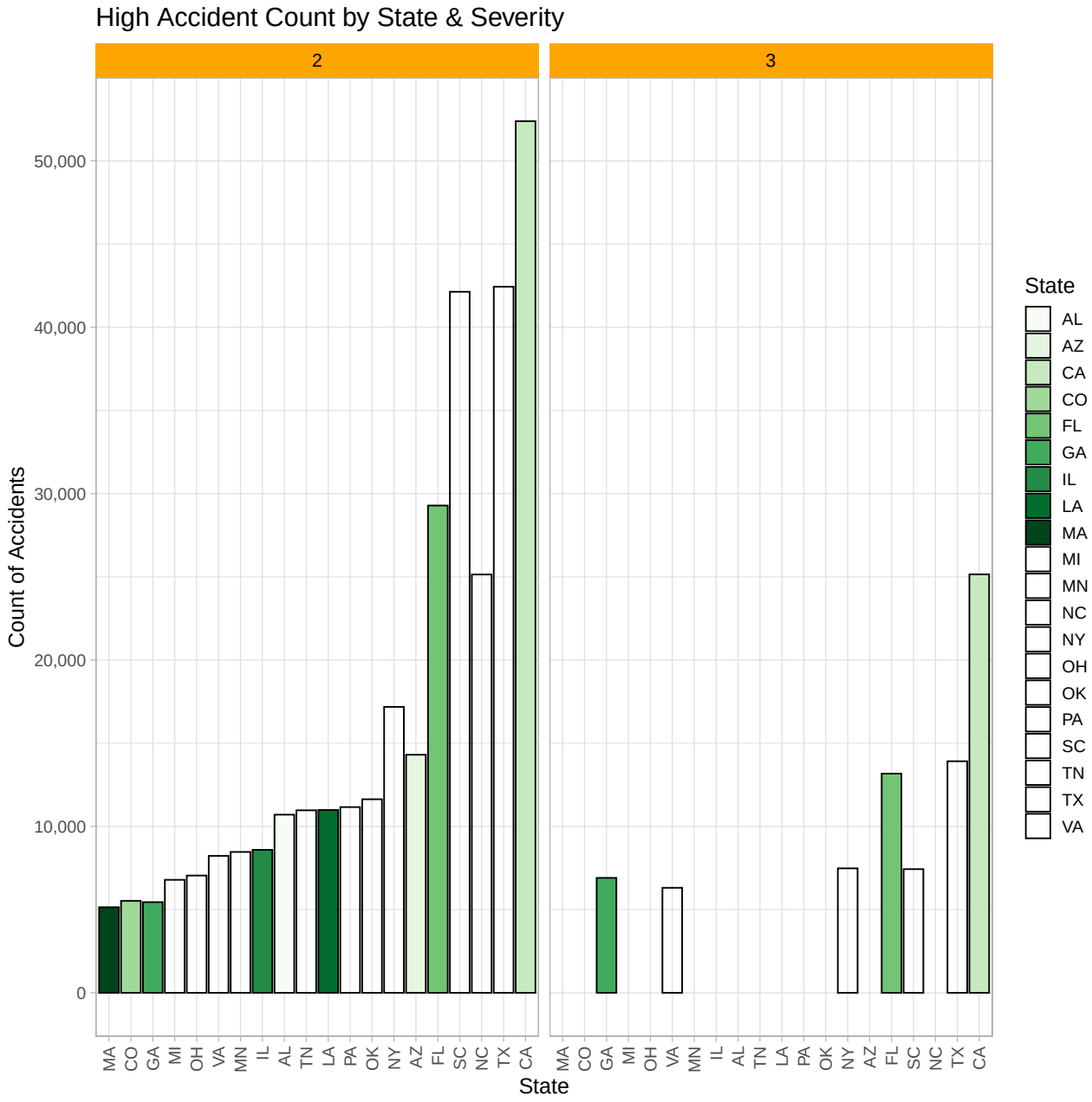


We created a bar chart using Severity and City variables to identify which city in each of severity levels 2 and 3 had the severest accidents. So in level 2 we see it is Charlotte and in level 3 it is Dallas. This can be confirmed from the US map chart above we did see that the east coast have the highest accidents in particular South Carolina.

States and Severity analysis of accidents:

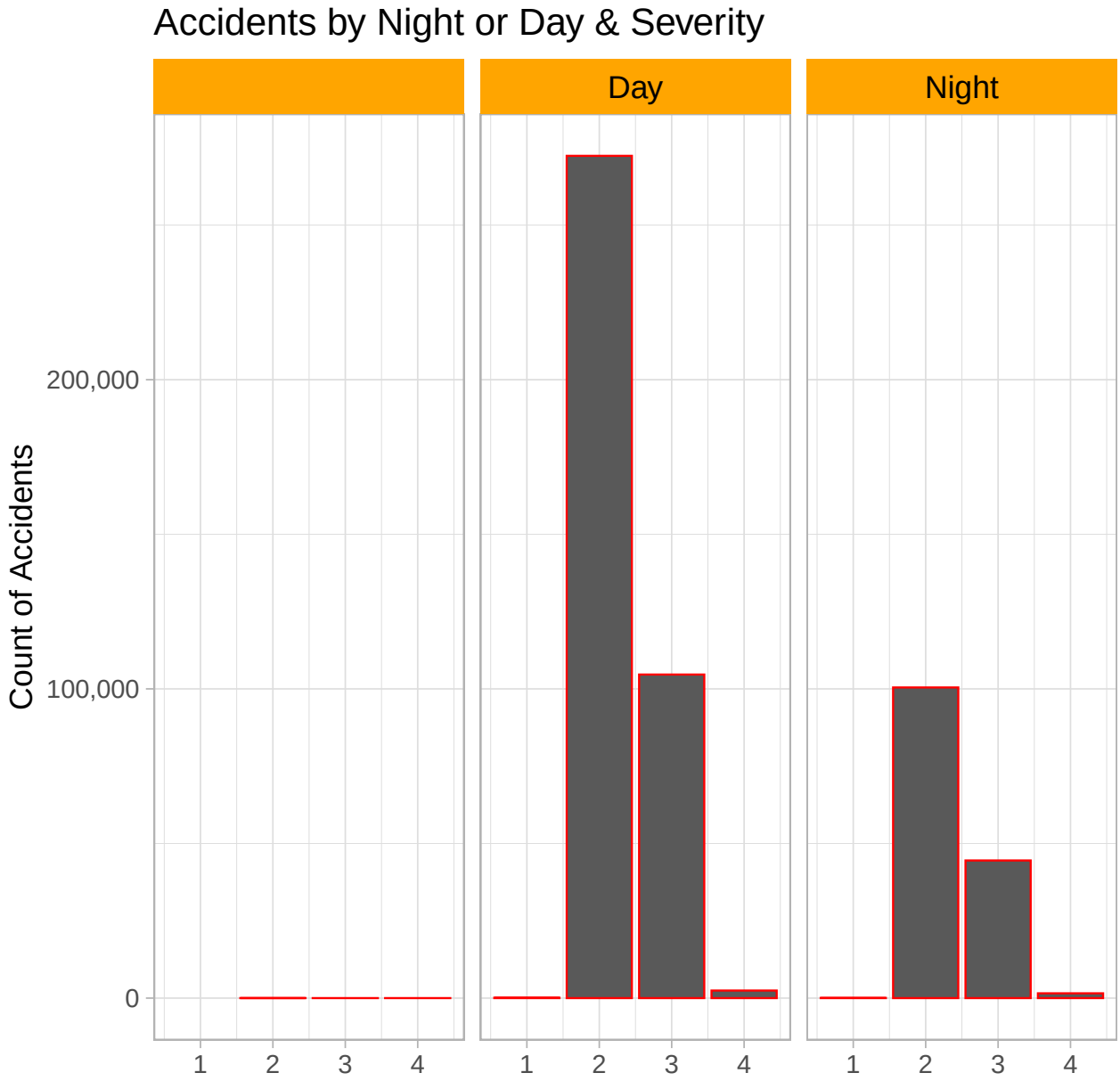
Like with the cities we also conducted the same analysis with the states and found that California followed by Florida & Texas had the highest rate of accidents.





This is further confirmed when we took the Severity variable to identify the states with the most severe accidents which in both level 2 and 3 shows California. This could be due to the fact that California is a densely populated state with high traffic output.

Analysis of accidents based on time of day: Another factor in analyzing the accidents was to check whether most of the accidents were occurring at night or day and how severe were those accidents. So we created a bar chart using Severity and Sunrise_Sunset variables to analyze which levels of severity where the most accidents were happening, according to day and night. From the chart we see that at daytime the severe accidents were level 2 and at night the most severe accident were of level 2. But day time saw the most accidents.



For the purposes of running the various algorithms we decided to drop states that had a fewer than 14500 number of accidents. So the accidents data frame now contains states with having number of accidents greater than 14500.

```
##    State     n
## 1     AZ 16655
## 2     CA 77776
## 3     FL 42909
## 4     NC 29560
## 5     NY 24763
## 6     PA 15315
## 7     SC 49780
## 8     TN 15561
## 9     TX 56609
## 10    VA 14869
```

As we can see the top ten states are Arizona, California, Florida, North Carolina, New York, Pennsylvania, South Carolina, Tennessee, Texas and Virginia.

We also wanted to filter out those weather condition that attributed to high accidents rates. So we filtered out those weather condition that had less than 2000 accident rates.

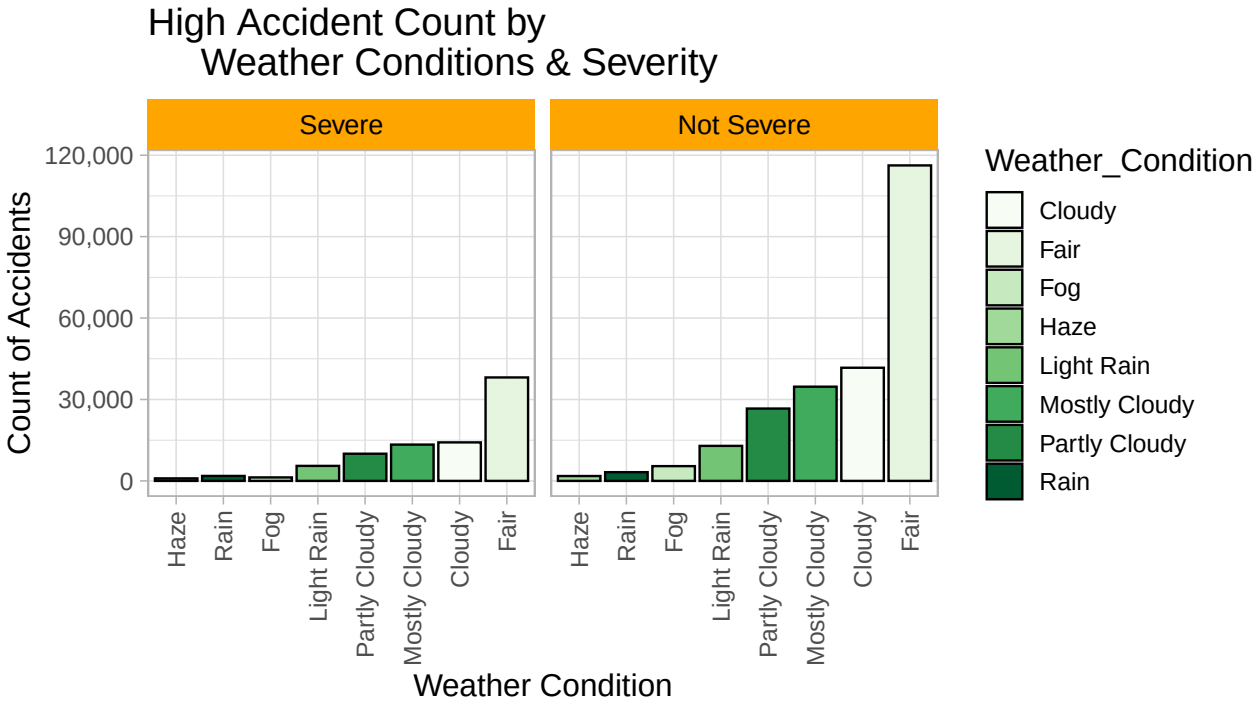
```
##   ACcidents.df$Weather_Condition      n
## 1                           Fair 154347
## 2                         Cloudy  55880
## 3                 Mostly Cloudy  48084
## 4                 Partly Cloudy  36674
## 5                    Light Rain  18434
## 6                           Fog   6726
## 7                          Rain   5042
## 8                          Haze   2721
```

So after the filter there are 8 weather condition categories that consists of high accident rates including Fair, Cloudy, Mostly Cloudy, Partly Cloudy, Light Rain, Fog, Rain and Haze.

Model Algorithms:

The purpose of our analysis of the traffic accident was to identify which states had the most severe accidents and whether those states have enough hospitals to deal with these accidents. The goal here is to build a model that can predict severe accidents occurring in the top ten states we have identified above. Before running the models, we first made the following changes to the data to ensure that our model works well:

We grouped the four severity levels of 1, 2, 3, & 4 into 2 groups: Severe (Observations from levels 3 & 4) and Not Severe (Observations from levels 1 & 2). The focus of our project is capturing severe accidents and checking if there are enough hospitals in those states to cater to the accidents. Also, we grouped the data so as to capture all the severe accidents.



```
##                  freqRatio percentUnique zeroVar  nzv
## Visibility.mi.       25.98     0.0094539   FALSE TRUE
## Amenity              86.84     0.0006099   FALSE TRUE
## Bump               6557.16     0.0006099   FALSE TRUE
## Give_Way            312.49     0.0006099   FALSE TRUE
## Junction             20.37     0.0006099   FALSE TRUE
## No_Exit             636.95     0.0006099   FALSE TRUE
## Railway              99.25     0.0006099   FALSE TRUE
## Roundabout       19287.71     0.0006099   FALSE TRUE
## Station              56.16     0.0006099   FALSE TRUE
## Stop                 56.44     0.0006099   FALSE TRUE
## Traffic_Calming    3487.38     0.0006099   FALSE TRUE
## Turning_Loop          0.00     0.0003050    TRUE TRUE
```

Some variables are near zero-variance, which means they cannot provide enough information for us because most of the data have the same values for these variables. What's worse is, when we split the dataset, the levels in training dataset and validation dataset may not match.

```
##   State Temperature.F. Visibility.mi. Weather_Condition Crossing Traffic_Signal
## 1    TX             85             10     Mostly Cloudy    FALSE          FALSE
## 2    TN             74             10     Mostly Cloudy    FALSE           TRUE
## 3    TX             85             10     Mostly Cloudy    FALSE          FALSE
## 4    SC             77             10     Partly Cloudy    FALSE          FALSE
## 5    TN             74             10     Mostly Cloudy     TRUE          FALSE
##   Sunrise_Sunset     Status
## 1          Night     Severe
## 2          Night Not Severe
## 3          Night Not Severe
## 4          Night Not Severe
## 5          Night Not Severe
```

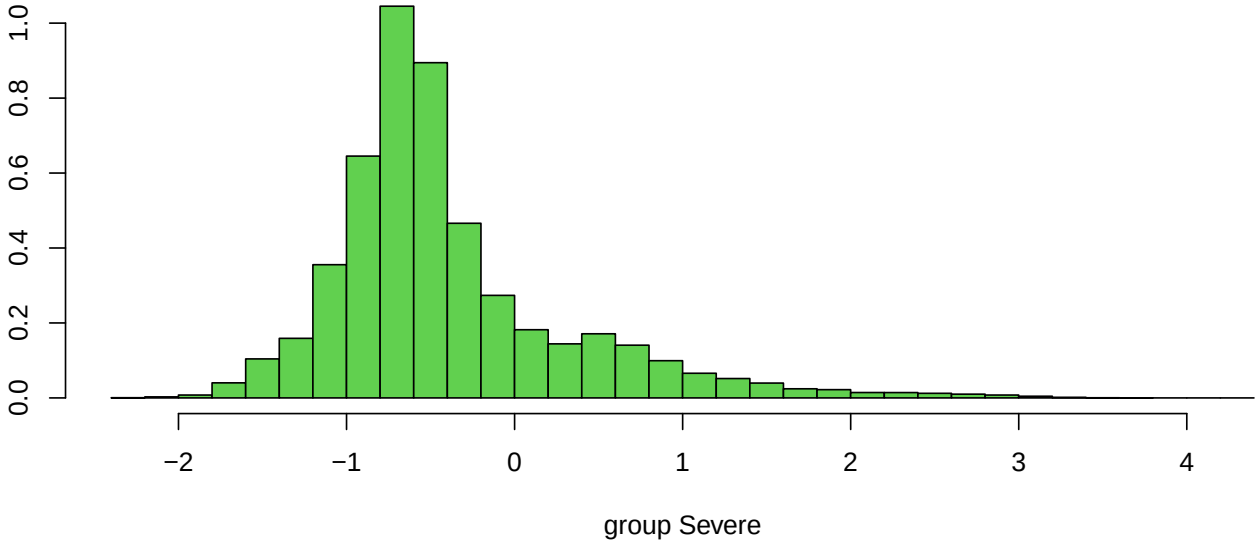
After removing the zero variance variable and analyzing the EDA we came to a conclusion that the observations in Non Severe are much more than the observation in Severe so in order to predict the severe class accurately and main focus of building the models is to predict the severe class we decided to under-sample our data set , so we took 100% of the severe observations and 50% of the total non-severe observation. After undersampling the data we decided to create a training and a validation data set, the training data set will help us to train the model and validation data set will help us to test our training model. After that we decided to preprocess our model using the training data set that will help us to make every variable unit less.

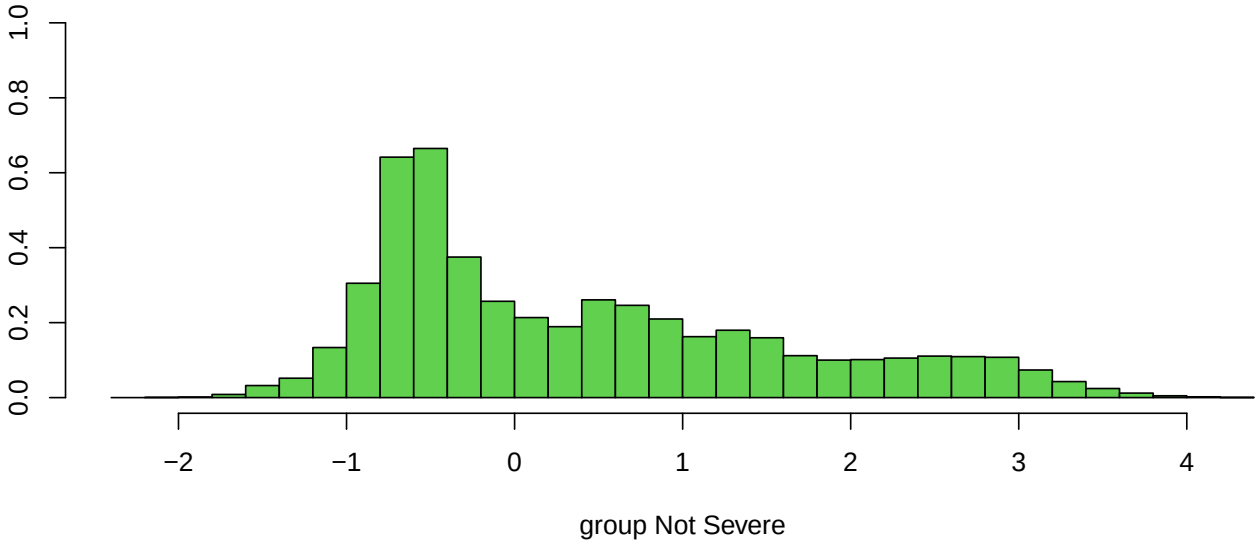
```
##        State Temperature.F. Visibility.mi. Weather_Condition Crossing
## 129480    TX         0.3412        -1.4052     Mostly Cloudy    FALSE
## 199669    SC        -2.1146         0.3545              Fair    FALSE
## 14213     NY         0.9551         0.3545     Mostly Cloudy    FALSE
## 99109     TX        -1.3779        -0.9653              Rain    FALSE
## 195471    FL        -0.2114         0.3545            Cloudy    FALSE
##        Traffic_Signal Sunrise_Sunset     Status
## 129480          FALSE          Night Not Severe
## 199669          FALSE            Day     Severe
## 14213           FALSE            Day Not Severe
## 99109           FALSE          Night     Severe
## 195471           TRUE            Day Not Severe
```

Linear Discriminant Analysis:

```
## Call:
## lda(Status ~ ., data = Acc.train.norm)
##
## Prior probabilities of groups:
##     Severe Not Severe
##        0.5        0.5
##
## Group means:
##            StateCA StateFL StateNC StateNY StatePA StateSC StateTN StateTX
## Severe      0.2911  0.1496 0.05087 0.07901 0.04439 0.08277 0.05045  0.1531
## Not Severe  0.2097  0.1139 0.10128 0.06292 0.04321 0.16788 0.04103  0.1675
##            StateVA Temperature.F. Visibility.mi. Weather_ConditionFair
## Severe     0.07272        0.03558       -0.00747                0.4467
## Not Severe 0.03338       -0.03558        0.00747                0.4822
##            Weather_ConditionFog Weather_ConditionHaze
## Severe                  0.01472              0.010891
## Not Severe              0.02240              0.006891
##            Weather_ConditionLight Rain Weather_ConditionMostly Cloudy
## Severe                         0.06545                         0.1565
## Not Severe                     0.05482                         0.1427
##            Weather_ConditionPartly Cloudy Weather_ConditionRain CrossingTRUE
## Severe                             0.1178               0.02119      0.01555
## Not Severe                         0.1083               0.01316      0.11812
##            Traffic_SignalTRUE Sunrise_SunsetDay Sunrise_SunsetNight
## Severe                 0.0485            0.6919              0.3081
## Not Severe             0.2685            0.7298              0.2702
##
## Coefficients of linear discriminants:
##                                       LD1
## StateCA                         -0.691044
## StateFL                         -0.817464
## StateNC                          0.196841
## StateNY                         -0.731257
## StatePA                         -0.547065
## StateSC                          0.645561
## StateTN                         -0.561055
## StateTX                         -0.411065
## StateVA                         -1.393033
## Temperature.F.                  -0.211822
## Visibility.mi.                   0.005741
## Weather_ConditionFair            0.115814
## Weather_ConditionFog             0.516187
## Weather_ConditionHaze           -0.336379
## Weather_ConditionLight Rain     -0.281283
## Weather_ConditionMostly Cloudy  -0.093098
## Weather_ConditionPartly Cloudy  -0.053222
## Weather_ConditionRain           -0.688703
## CrossingTRUE                     0.905009
## Traffic_SignalTRUE               2.013836
## Sunrise_SunsetDay               -0.219316
## Sunrise_SunsetNight             -0.480861
```

We ran the LDA on the training dataset and observed that it is perfectly balanced between the two severity levels with prior probability of 0.5 in each. The coefficients of linear discriminants are the weights attached to the predictors and by themselves are not interpretable. With the help of these coefficients, we can generate the linear discriminant score. We can calculate linear discriminant score by multiplying the coefficients to the actual values of the variables.





Both the training and validation dataset of which did not show a very clear separation between severe and not severe categories. We see that scores below between 1 and -2 are predicted as severe. The range of scores for the not severe class is large and ranges between -1 and 3.

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   Severe Not Severe
##   Severe      52460      35782
##   Not Severe   7040      23718
##
##                Accuracy : 0.64
##                  95% CI : (0.637, 0.643)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.28
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.882
##             Specificity : 0.399
##          Pos Pred Value : 0.595
##          Neg Pred Value : 0.771
##              Prevalence : 0.500
##          Detection Rate : 0.441
##    Detection Prevalence : 0.742
##       Balanced Accuracy : 0.640
##
##        'Positive' Class : Severe
##

## Confusion Matrix and Statistics
##
##             Reference
## Prediction   Severe Not Severe
##   Severe      22426      15364
##   Not Severe   3074      10136
##
##                Accuracy : 0.638
##                  95% CI : (0.634, 0.643)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.277
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.879
##             Specificity : 0.397
##          Pos Pred Value : 0.593
##          Neg Pred Value : 0.767
##              Prevalence : 0.500
##          Detection Rate : 0.440
##    Detection Prevalence : 0.741
##       Balanced Accuracy : 0.638
##
##        'Positive' Class : Severe
```

##

For our training and validation dataset, we have taken a cutoff of 0.4. This means that all the observations that equal to 0.4 and more, will be classified as severe. The confusion matrix for our training dataset shows us that the accuracy of the model is 63.9% with a high sensitivity of 88.5% with our class of interest being the severe class. On the other hand, we see similar results for our validation dataset, accuracy being 64.3% and sensitivity of 88.8%.

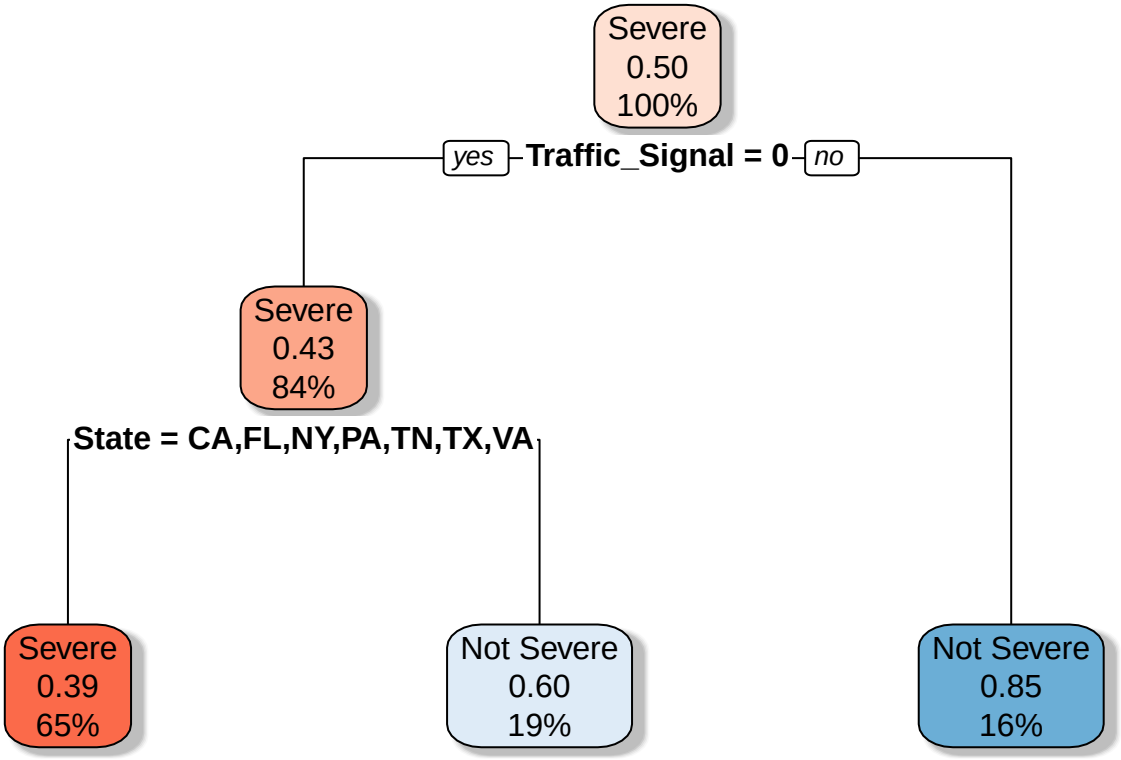
Overall, our model does a good job in capturing the severe accidents.

Decision Trees:

Decision trees are in the structure of a tree which makes them extremely easy to understand. The cp value is used to control the size of the decision tree and to select the optimal tree size. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then tree building does not continue. We could also say that tree construction does not continue unless it would decrease the overall lack of fit by a factor of cp. For our model, the cp value of 0.016 is chosen as it corresponds to the highest accuracy achieved in the model of 64%.

If there isn't the presence of a traffic signal, and if the observations belong to either of these states (CA, FL, NY, PA, TN, TX, VA), then we reach at the last node of severe accidents and it constitutes 65% of the observations.

```
## CART
##
## 119000 samples
##      7 predictor
##      2 classes: 'Not.Severe', 'Severe'
##
## No pre-processing
## Resampling: Cross-Validated (20 fold)
## Summary of sample sizes: 113050, 113050, 113050, 113050, 113050, 113050, ...
## Resampling results across tuning parameters:
##
##   cp       Accuracy  Kappa
##   0.01666  0.6427    0.2855
##   0.05662  0.6183    0.2367
##   0.21998  0.5646    0.1293
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.01666.
```



Our tree has two decision variables, presence of a traffic signal and state being one of these: CA, FL, NY, PA, TN, TX or VA.It classifies the records on that basis. The first decision variable of the presence of traffic signal has the following meaning, it can have two values of 0 and 1, with 1 indicating the presence of a traffic signal, whereas 0 indicates no traffic signal. In the tree generated by our model, it shows that when there is a traffic signal present, the accidents that are not severe constitute 16% of total observations. Whereas, when there is no traffic signal, the number of accidents that are severe are 84%. This clearly shows that the presence of a traffic signal helps in reducing the severity of accidents. The next decision variable is the state where accidents happen, if it is either one of these states, then 65% of the accidents there are severe and 19% are not severe and are not even in those states.

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   Severe Not Severe
##   Severe      51774      35316
##   Not Severe   7726      24184
##
##                Accuracy : 0.638
##                  95% CI : (0.636, 0.641)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.277
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.870
##             Specificity : 0.406
##          Pos Pred Value : 0.594
##          Neg Pred Value : 0.758
##              Prevalence : 0.500
##          Detection Rate : 0.435
##    Detection Prevalence : 0.732
##       Balanced Accuracy : 0.638
##
##        'Positive' Class : Severe
##

## Confusion Matrix and Statistics
##
##             Reference
## Prediction   Severe Not Severe
##   Severe      22104      15194
##   Not Severe   3396      10306
##
##                Accuracy : 0.635
##                  95% CI : (0.631, 0.64)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.271
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.867
##             Specificity : 0.404
##          Pos Pred Value : 0.593
##          Neg Pred Value : 0.752
##              Prevalence : 0.500
##          Detection Rate : 0.433
##    Detection Prevalence : 0.731
##       Balanced Accuracy : 0.635
##
##        'Positive' Class : Severe
```

##

For our training and validation dataset, we have taken a cutoff of 0.4. This means that all the observations that equal to 0.4 and more, will be classified as severe. The confusion matrix shows us that the accuracy of the model is 63.7% with high sensitivity of 86.8% with our class of interest being the severe class. The validation dataset increases the accuracy by a small amount to 64.2%, and the sensitivity of the model goes up to 87.2%.

Random Forest-Bagging:

We also tried modeling our dataset using Bagging. The advantage of using that is that it improves the overfitting problem by the sampling technique it uses, called "bootstrapping".

These two arguments used in the bagging model are very important:

1.Mtry- Number of variables available for splitting at each tree node

2.Ntree- Number of trees to grow

In bagging, the OOB estimate of error rate is a useful measure to distinguish between different random forest classifiers. We can, for example, vary the number of trees or the number of variables to be measured, and select the combination that produces the smallest value for this error rate. For our model, the OOB estimate of error rate is 36.19%.

```
##
## Call:
##  randomForest(formula = Status ~ ., data = Acc.train.norm, mtry = 7,      ntree = 50, importance = T
##                Type of random forest: classification
##                      Number of trees: 50
## No. of variables tried at each split: 7
##
##         OOB estimate of  error rate: 36.36%
## Confusion matrix:
##            Severe Not Severe class.error
## Severe      42339      17161      0.2884
## Not Severe  26109      33391      0.4388
```

For our training dataset, we have taken a cutoff of 0.4. This means that all the observations that equal to 0.4 and more, will be classified as severe. The confusion matrix shows us that the accuracy of the model is 63.7% with sensitivity of 86.8% with our class of interest being the severe class. The validation dataset decreases the accuracy to 63.7% and the sensitivity of the model goes down to 74.8%. Validation dataset in the Bagging model gives us the worst result, whereas the training dataset gives us the exact same result as the decision tree model.

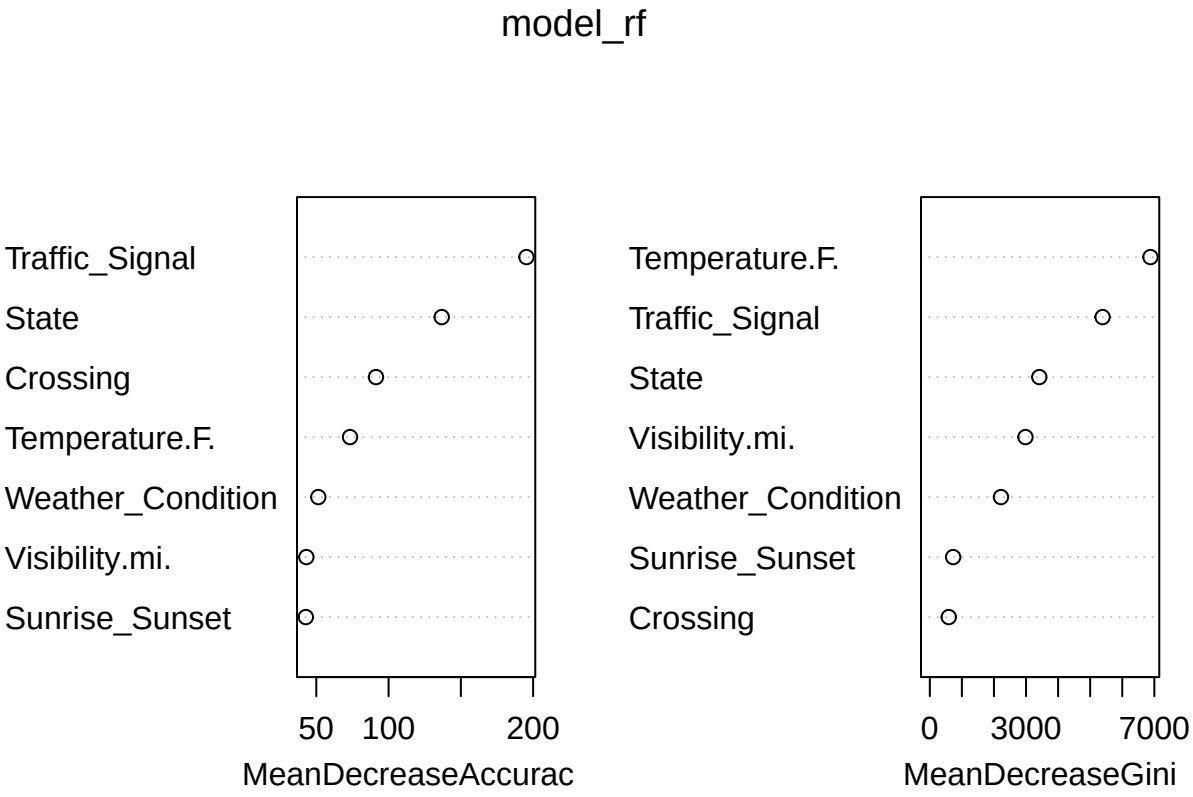
```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   Severe Not Severe
##   Severe      51774      35316
##   Not Severe   7726      24184
##
##                Accuracy : 0.638
##                  95% CI : (0.636, 0.641)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.277
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.870
##             Specificity : 0.406
##          Pos Pred Value : 0.594
##          Neg Pred Value : 0.758
##              Prevalence : 0.500
##          Detection Rate : 0.435
##    Detection Prevalence : 0.732
##       Balanced Accuracy : 0.638
##
##        'Positive' Class : Severe
##

## Confusion Matrix and Statistics
##
##             Reference
## Prediction   Severe Not Severe
##   Severe      18754      11992
##   Not Severe   6746      13508
##
##                Accuracy : 0.633
##                  95% CI : (0.628, 0.637)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.265
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.735
##             Specificity : 0.530
##          Pos Pred Value : 0.610
##          Neg Pred Value : 0.667
```

```
##              Prevalence : 0.500
##          Detection Rate : 0.368
##    Detection Prevalence : 0.603
##       Balanced Accuracy : 0.633
##
##        'Positive' Class : Severe
##
```

Importance of Variables

```
##                   Severe Not Severe MeanDecreaseAccuracy MeanDecreaseGini
## State              108.67     66.999                136.75           3414.0
## Temperature.F.      70.18      2.598                 73.33           6878.7
## Visibility.mi.      33.12      5.431                 43.11           2982.1
## Weather_Condition   43.81      5.156                 51.37           2218.5
## Crossing           169.35    -18.842                 91.29            590.5
## Traffic_Signal     256.62     60.953                195.37           5386.4
## Sunrise_Sunset      38.21     14.215                 42.74            727.2
```



Models LDA Decision Tree Bagging

	Training	Validation	Training	Validation	Training	Validation
Accuracy	63.9%	64.3%	63.7%	64.2%	63.7%	63.6%
Sensitivity	88.5%	88.8%	86.8%	87.2%	86.8%	74.8%
Specificity	39.3%	39.7%	40.7%	41.1%	40.7%	52.4%
Class of Interest	Severe	Severe	Severe	Severe	Severe	Severe

Hospitals dataset

Data Cleaning:

Like the accidents dataset, a number of challenges were discovered within the hospitals dataset as well. We only kept columns/features that could be used in conjunction with the accidents dataset. In addition, we cleaned up the data removing the null entries. Following is a list of data cleanup techniques used on this dataset:

- Only keep the relevant columns and removed the rest.
- Removed all hospitals which have non positive bed count. i.e BEDS<=0

Once the above data cleanup techniques are applied, we are left with a clean dataset with the following columns:

1. CITY – city in which the hospital is located
2. STATE – state in which the hospital is located
3. TYPE – type of hospital (children, general acute care, long term care etc)
4. STATUS – if the hospital is open or close
5. BEDS – number of beds
6. TRAUMA - level of trauma treatment available

```
##       CITY STATE               TYPE STATUS BEDS             TRAUMA
## 1  BAYTOWN    TX GENERAL ACUTE CARE   OPEN  182      NOT AVAILABLE
## 2 COLUMBUS    OH            SPECIAL   OPEN   50      NOT AVAILABLE
## 3   DAYTON    OH          CHILDREN   OPEN  155 PEDIATRIC LEVEL II
## 4 BOARDMAN    OH     LONG TERM CARE   OPEN   45      NOT AVAILABLE
## 5   DAYTON    OH       PSYCHIATRIC   OPEN   32      NOT AVAILABLE
```

Data Analysis:

Once the preliminary cleanup is completed, we analyze the data for interesting features and choose only a subset of this dataset. In this step we also correlate this dataset with the accident dataset to chose the relevant states found in our previous analysis. We filter out the data using the following conditions:

- We kept all the hospitals which are operational. i.e STATUS==OPEN
- Select entries wherein the hospital type is 'General acute care' or 'critical access'
- Select the states that correspond to the findings of the accident dataset (CA,TX,SC,FL, etc)

Once the above filtering is completed, the dataset looks like as shown below:

Once we have the filtered dataset, we count the the number of hospitals in each of the states. The count is as shown in the table below:

```
##              CITY STATE               TYPE STATUS BEDS        TRAUMA
## 1         BAYTOWN    TX GENERAL ACUTE CARE   OPEN  182 NOT AVAILABLE
## 9         DANBURY    NC GENERAL ACUTE CARE   OPEN   93 NOT AVAILABLE
## 10        HOUSTON    TX GENERAL ACUTE CARE   OPEN   16 NOT AVAILABLE
## 11        MCALLEN    TX GENERAL ACUTE CARE   OPEN  441     LEVEL III
## 13      JACKSBORO    TX GENERAL ACUTE CARE   OPEN   17      LEVEL IV
## 14        LUBBOCK    TX GENERAL ACUTE CARE   OPEN  196 NOT AVAILABLE
## 15  THE WOODLANDS    TX GENERAL ACUTE CARE   OPEN    6 NOT AVAILABLE
## 16 CORPUS CHRISTI    TX GENERAL ACUTE CARE   OPEN  557 NOT AVAILABLE
## 17 CORPUS CHRISTI    TX GENERAL ACUTE CARE   OPEN   89 NOT AVAILABLE
```

```
## 18       BIG LAKE    TX GENERAL ACUTE CARE   OPEN    7 NOT AVAILABLE
```

```
##    Hospital4$STATE   n
## 1               AZ  68
## 2               CA 434
## 3               FL 227
## 4               NC 120
## 5               NY 193
## 6               PA 173
## 7               SC  76
## 8               TN 121
## 9               TX 450
## 10              VA  88
```

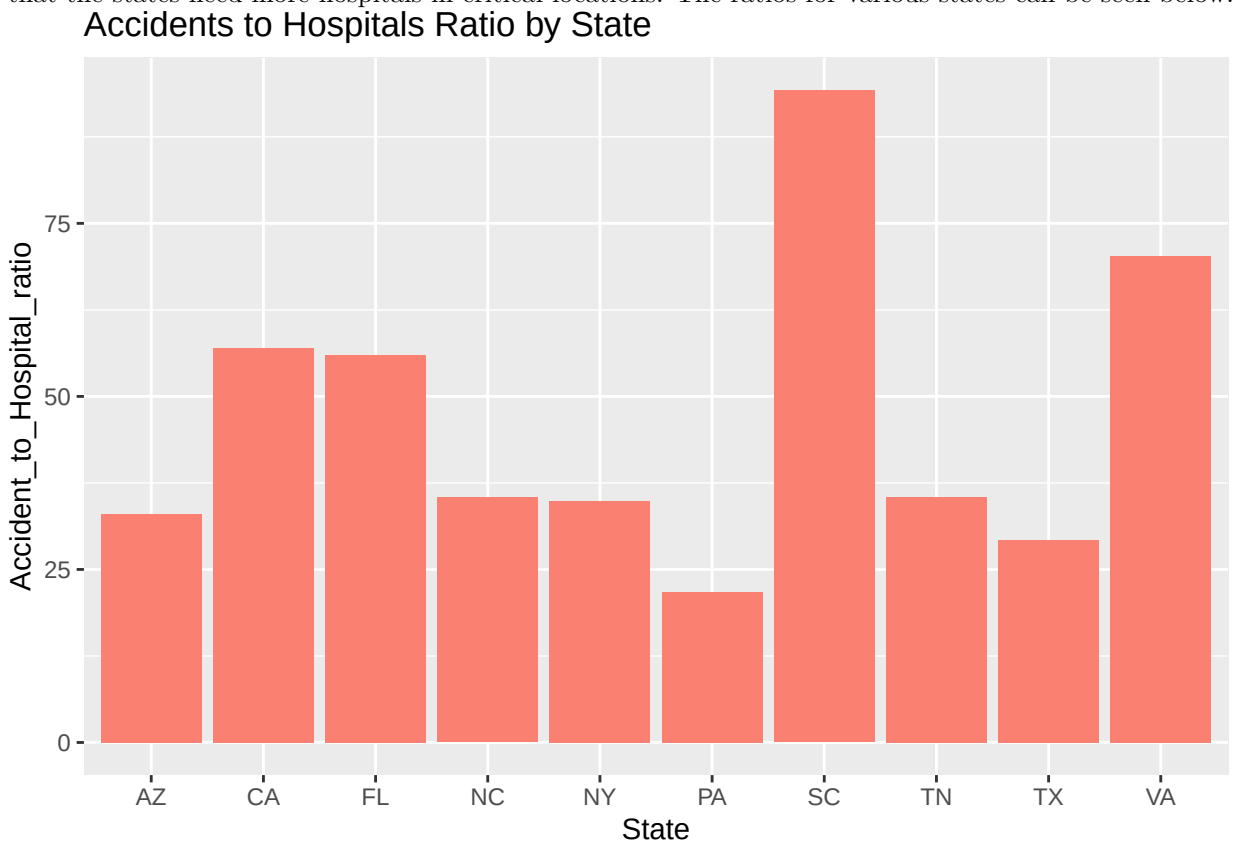
```
##    State Status severe_acc_counts
## 1     AZ Severe              2242
## 3     CA Severe             24744
## 5     FL Severe             12708
## 7     NC Severe              4247
## 9     NY Severe              6722
## 11    PA Severe              3767
## 13    SC Severe              7159
## 15    TN Severe              4295
## 17    TX Severe             13193
## 19    VA Severe              6190
```

```
##    Hospital4$STATE   n
## 1               AZ  68
## 2               CA 434
## 3               FL 227
## 4               NC 120
## 5               NY 193
## 6               PA 173
## 7               SC  76
## 8               TN 121
## 9               TX 450
## 10              VA  88
```

```
##    State Status severe_acc_counts Hospital4$STATE   n
## 1     AZ Severe              2242              AZ  68
## 3     CA Severe             24744              CA 434
## 5     FL Severe             12708              FL 227
## 7     NC Severe              4247              NC 120
## 9     NY Severe              6722              NY 193
## 11    PA Severe              3767              PA 173
## 13    SC Severe              7159              SC  76
## 15    TN Severe              4295              TN 121
## 17    TX Severe             13193              TX 450
## 19    VA Severe              6190              VA  88
##    Accident_to_Hospital_ratio
## 1                       32.97
## 3                       57.01
## 5                       55.98
## 7                       35.39
## 9                       34.83
## 11                      21.77
```

```
## 13                    94.20
## 15                    35.50
## 17                    29.32
## 19                    70.34
```

In order to make a recommendation, we use the readiness ratio as a measure of readiness of a state in case of accidents. This ratio is defined as the ratio of number of accidents in the state to the number of hospitals. The lower the ratio, more ready the state is in case of accidents. Higher the ratio indicates that the states need more hospitals in critical locations. The ratios for various states can be seen below:



Conclusion:

The analysis of a large set of data takes some extensive pre-planning as it's quite easy to get lost amongst all the potential data related questions. Data cleaning and processing is the most important aspect in any data analysis, so we first decided to analyze the data of 2019 and 2020, which rescinded our observations to 500,000, and after that in model building phase we took a sample of this data 500,000 observations.

During our exploratory data analysis, we expected clearer correlations between bad weather and the number of accidents. However, here we can see that more accidents occur when the weather is Fair/clear. This may be because people drive more carefully when the weather is bad.

After analyzing the observations and performing EDA, we came to a conclusion that majority of the accidents had a severity of degree 2 which is classified as not severe in our model, but in the model building phase our main aim was to predict the most severe accidents, so we decided to under sample our model wherein we took all the observations from the severe category i.e. severity of 3 & 4 and 50% observation from Non severe category i.e. 1 & 2.

In the model building phase we build three models Linear Discriminant Analysis, Regression Tree and Random Forest, out of these three models the best accuracy and the sensitivity was for LDA so we decided to present this model that would predict more severe accidents to the state government and Traffic law enforcement that looks after the Traffic lights.

State Government: We have analyzed the hospital data set in top 10 states that have most accidents and have come up with a severe accident to the number of hospital ratio. This analysis and with the help of the model will help the state government to build more hospitals that can cater to more severe accidents. After analyzing the data, we came to a conclusion that, South Carolina has a ratio of 94.5 accidents per hospital and Virginia has a ratio of 70, therefore, the respective state governments should take quick actions and build more number of hospitals.

Traffic law enforcement: After running the Regression Tree Model, we observed that the most important variable was Traffic signal and the most severe accidents happen where there are no traffic signals. Hence, the traffic police need to focus in these areas to reduce accidents in collaboration with the state government. If these are due to some infrastructure problems, then they need to be resolved.

Though there might be many factors such as driving under influence, texting while driving, speeding etc. that have not been considered while analyzing the US accident data set, these factors constitute to the most severe and high number of accidents, there must be strict rules laid down by the government. These all factors in our model may be hiding in the error term and may constitute to some sort of endogeneity in our model or homogeneity, which may overstate or understate the true value of our parameter but this is the best our model could predict.

Our main focus of building the model was to identify if there were enough number of hospitals in that state that can reach the accident spot on time and treat the person on time.

Our model is doing a great job by predicting the maximum severe accidents, this will help the police, ambulance and other resources to reach on time when the severity of the accident is the highest. Though our model is not very well in predicting the Non-Severe accidents which may be a concern as there would be a waste of resources, but then we would like to say is better safe than sorry!!