Case Study on Determining Factors in Car Accident

Muley, Tushar

I found a large dataset of automobile accidents in the United States from Kaggle.com. I want to perform analysis to understand accident rates and what might be the driving factor that causes high accident rates in the top five state and the top five cities in those states. I believe this analysis would be of interest to individuals, cities, counties and even state officials to help understand the factors that might be contributing to accidents that happen within their boundaries. A better understanding of the different factors might give insight into leading causes of accidents that are not related to drivers but traffic controls or other factors.

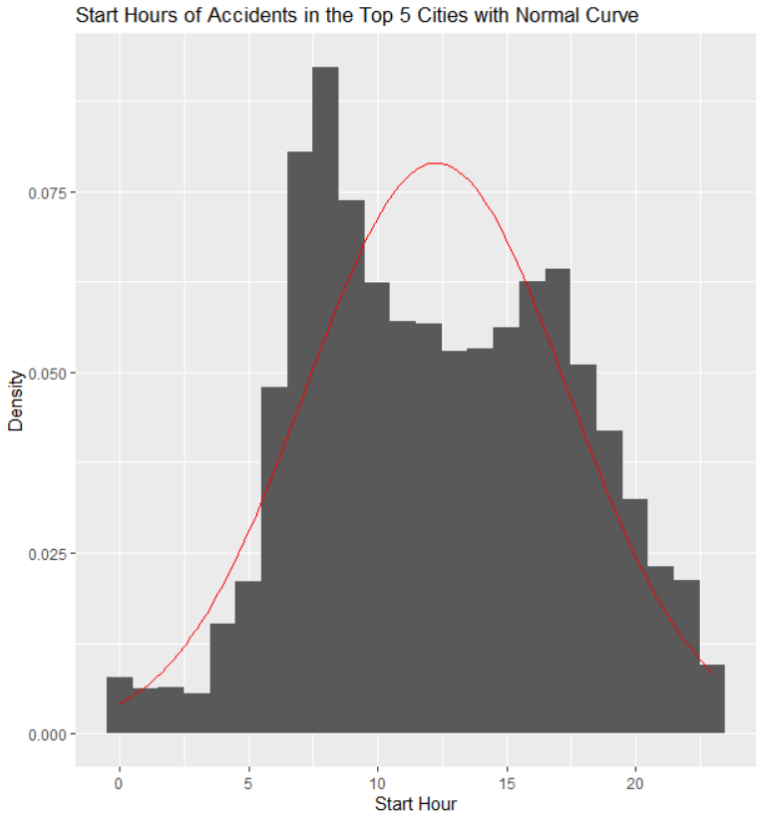
In order to arrive at the highest cities with the highest rate of accidents I tallied up the cities and select only one city from that state to provide some diversity of data. You can see this information in the Appendix section the bottom of this documentation. I also have a table defining the terms of the data.

I came up with the following list of questions I think can be addressed by the data that was available.

1. Does the time of day contribute to accident rates?
2. Where do majority of accidents occur on surface streets, interstates or state routes?
3. Do stop signs or stop lights contribute to a higher rate of accidents compared to points of interest (POI) contribution?
4. How do different points of interest contribute to accident rates?
5. What outside factors like the type of weather contribute to accident rate?
6. Using Logistic Regression to predict the severity of the accident using the contributing elements of an accident.

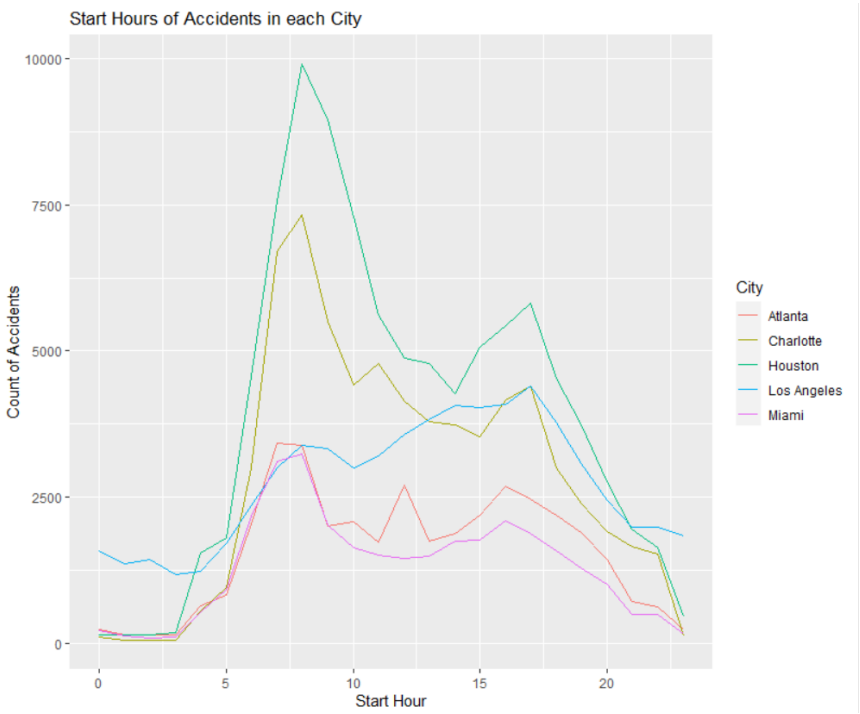
Question 1:

Does the time of day contribute to the rate of accident? In order to answer this question, I took Start Time in my data, which was represented as date time stamp. I split the date and time portions apart. After that I used the lubridate function to split the hour into a different column called ‘Start_Hour’. From here I created a histogram using ggplot with a normal curve to show which hours tend to have the most accidents.



Output1: Histogram by Start Hour

The plot is bimodal and has negative kurtosis. The data is spread out and not pointy. The histogram also shows peak between 8AM and 9AM time slot. The two almost equal peaks is between 4PM and 5PM. To see a clearer picture of which cities have the highest rates of accidents I plotted a line chart below in Output 2. As you can see peaks are similar for all cities, but for some reason Houston and Charlotte have much higher accident rate compared to cities with higher populations like Atlanta and Los Angeles.



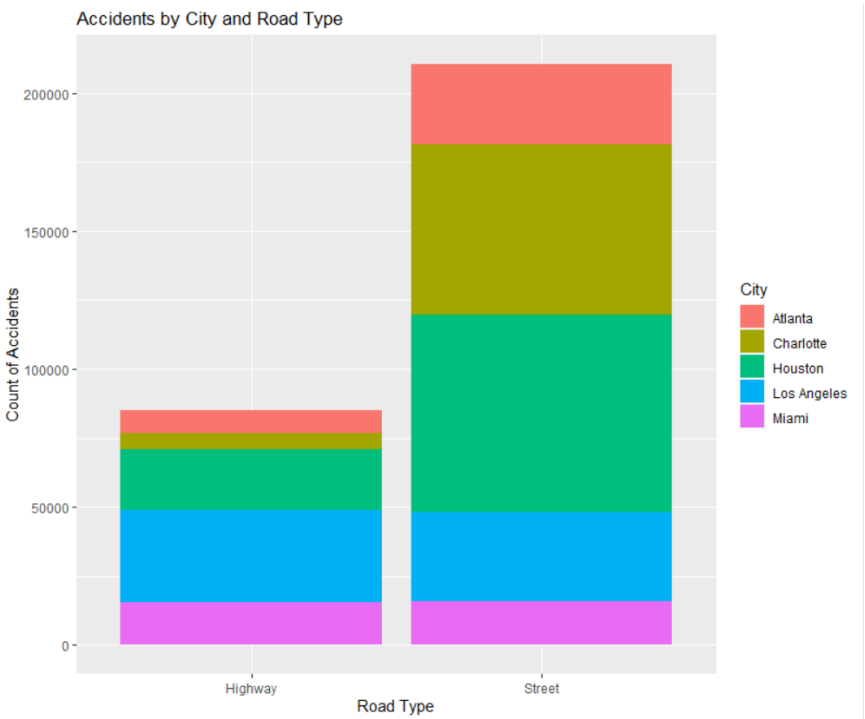
Output 2: Line plot of the cities and the start hours.

I believe more investigation as to why Houston and Charlotte have higher rate of accidents during peak hours compared to cities with higher populations. There might be data limitation or other issues that would require further investigation. The data does show the hours of 8AM to 9AM and 4PM to 5PM are the peaks for the most amount of accidents. By adding in sun rise and sun set into analysis there might be a direct correlation. Since those type of occurrence has proven to cause of higher accidents.

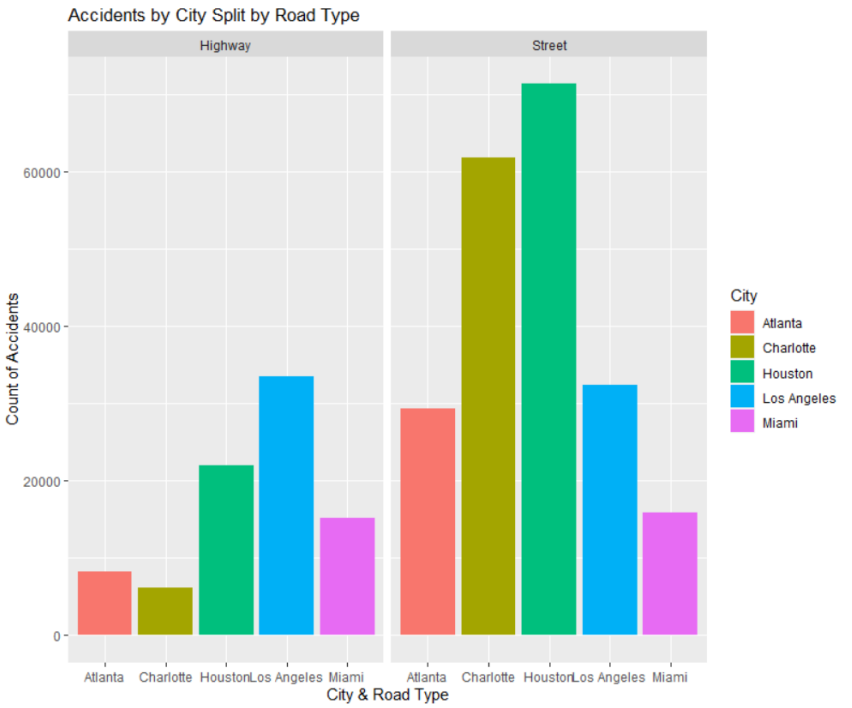
Question 2:

This leads me to my next question of where do majority of these accidents occur? Do these accidents happen on surface street or freeway or interstates? The data has detail street names and interstate names. I decided it would be best to bucket or bin this data into two types of road ways, highways and streets. I also used terms like 'Fwy', which describe Freeway a term used more on the West Coast. The term 'Expy' or Express way is used more on the East Coast. By narrowing down these different terms to two possible terms. This would help narrow down other factors that might lead to higher rates of accidents. I used the mutate function and str_detect (found in the stringr library) to bucket the data as highway which was anything with Fwy, Expy, Highway, I- or US- in the column called Street in the dataset. I created a variable called Road_Type to hold this information.

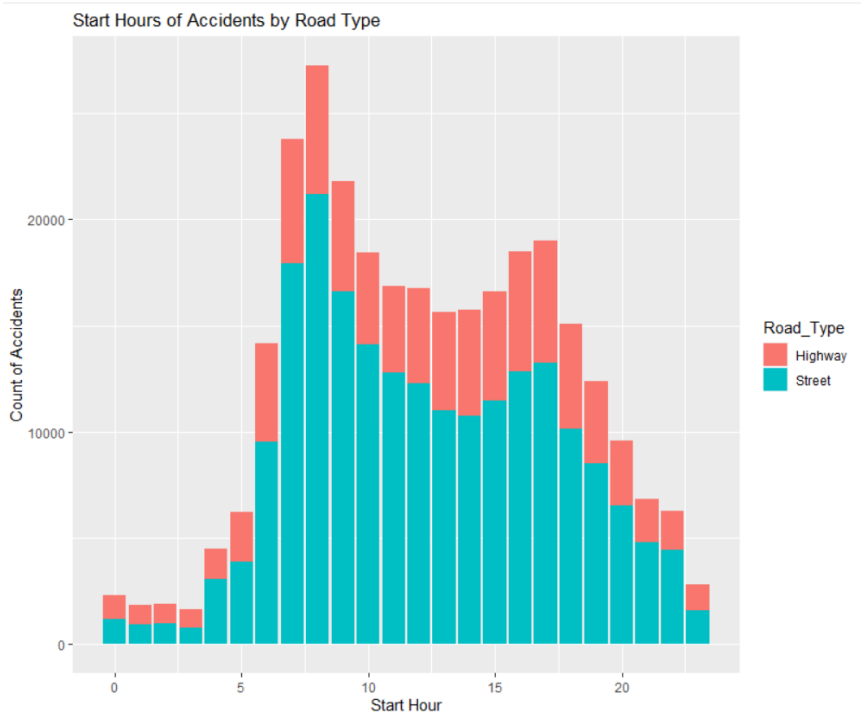
If you take a look at Output 3-5 you see accidents on the surface streets (Streets) is more common than Highways. You'll notice some difference when looking at just Road Type and number of accidents. In Los Angeles the number of Highway accidents similar to Surface Street accidents. Maim seems to have similar equal number of Highway accidents as Street accidents. The two cities that have the most accidents Houston and Charlotte have the highest number of accidents occur on the street rather than Highways. It almost seems disproportionate compared to the other cities at the number of accidents on surface streets compared to Highways.



Output3:Bar graph of how each city contributes to accident on the type of road way.



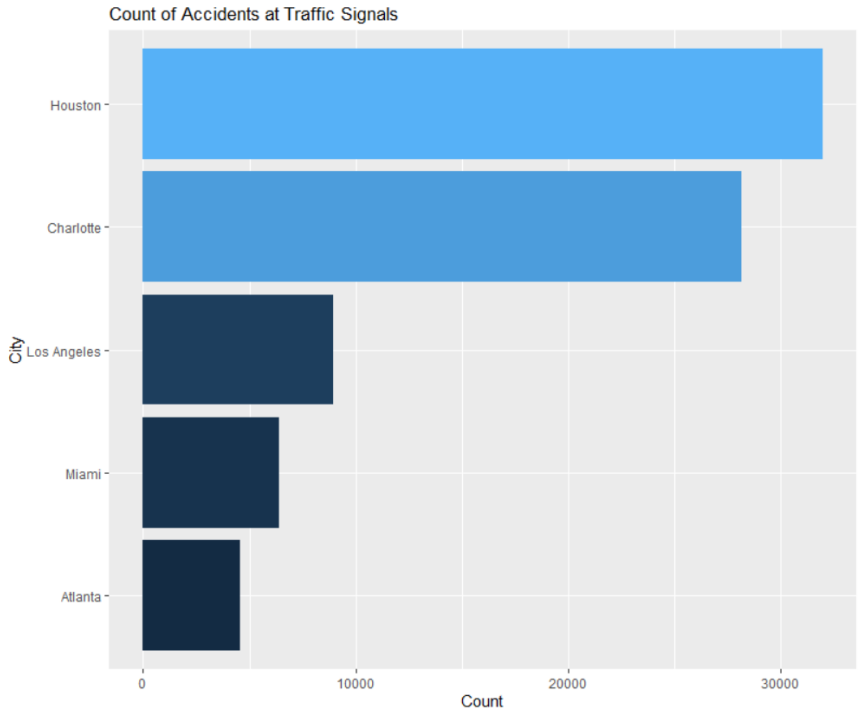
Output 4: A better view of Road Type and how each city contributes.



Output 5: Start Hour and Road Type provide as additional information.

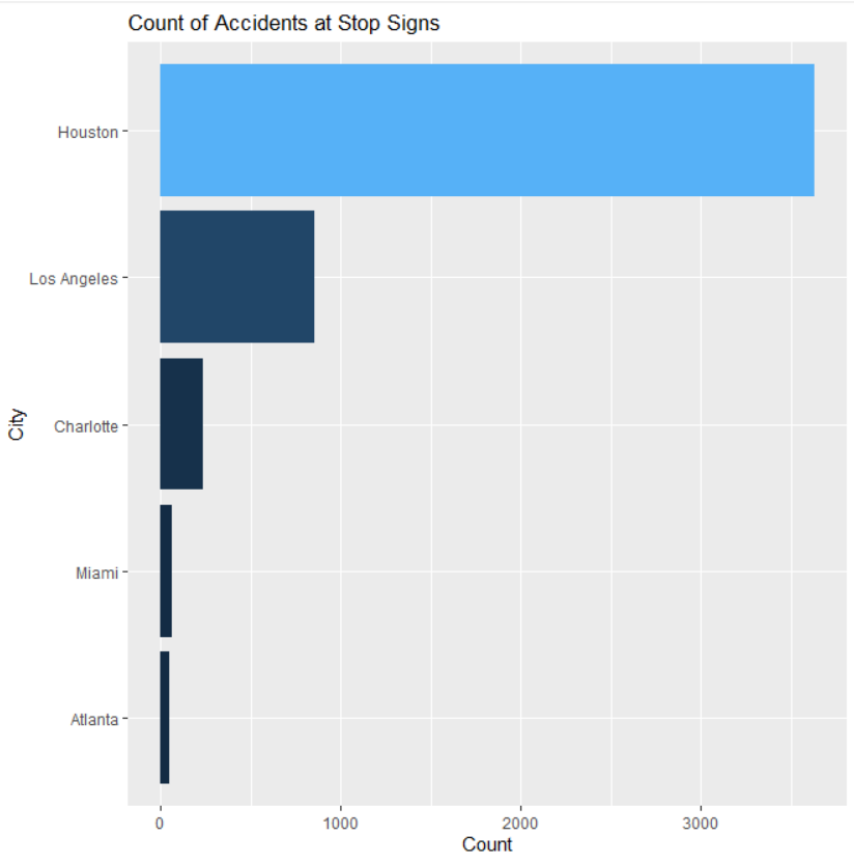
Question 3:

The next part of the analysis I wanted to know about traffic controls, more specifically traffic signals and stop signs and their contribution to number of accidents. This would focus the analysis on surface streets as there are not very many traffic lights on freeways. There are exceptions that I have seen in Tempe, Arizona and near Long Beach, California close to LAX airport. For this part of the research I was able to find a simpler way to create my graph without having to split out the data.



Output 6: Number of Accidents related to Traffic Signals

If you look at Output 6 you will notice the number of accidents associated to Traffic Signals. I took the data performed a select to isolated City and Traffic Signal variables. Perform a count on the TRUE and FALSE categorial Traffic Signal variable. This allowed me to plot the number of accidents related to Traffic Signal. Based on the data you can determine that the majority of accident that happen in Texas and North Carolina are related to Traffic Signals. The cities of Houston and Charlotte are the highest making Los Angeles and Miami seem like incredibly lower on accident rate.



Output 7: Number of Accidents related to Stop Signs.

Output 7 show the rate of accident related to Stop Signs. From this analysis Charlotte drops to third and Los Angeles takes its place. But Houston is still pretty far out in front. It makes you think that there are either very unfortunate driver in Houston or there is something else with the data. Having a contact for the data would probably help solve the issue.

```
Call:
lm(formula = Severity ~ Traffic_Signal, data = US_Accidents_Dataset_Final)

Residuals:
    Min      1Q  Median      3Q     Max
-1.39105 -0.39105 -0.09121  0.60895  1.90879

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.391045   0.001023  2336.4   <2e-16 ***
Traffic_Signal -0.299838   0.001966  -152.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4752 on 295684 degrees of freedom
Multiple R-squared:  0.0729,    Adjusted R-squared:  0.0729
F-statistic: 2.325e+04 on 1 and 295684 DF,  p-value: < 2.2e-16
```

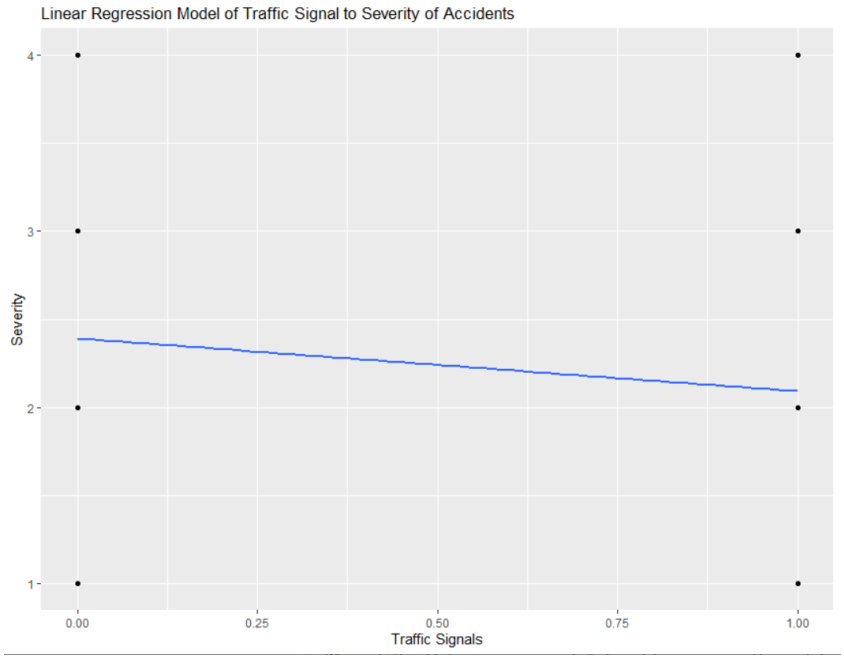
```
Call:
lm(formula = Severity ~ Traffic_Signal + Stop, data = US_Accidents_Dataset_Final)

Residuals:
    Min      1Q  Median      3Q     Max
-1.39874 -0.39874 -0.09182  0.60126  1.95491

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.398735   0.001030 2329.37   <2e-16 ***
Traffic_Signal -0.306914   0.001962 -156.39   <2e-16 ***
Stop           -0.353643   0.006882  -51.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4731 on 295683 degrees of freedom
Multiple R-squared:  0.08111,   Adjusted R-squared:  0.0811
F-statistic: 1.305e+04 on 2 and 295683 DF,  p-value: < 2.2e-16
```

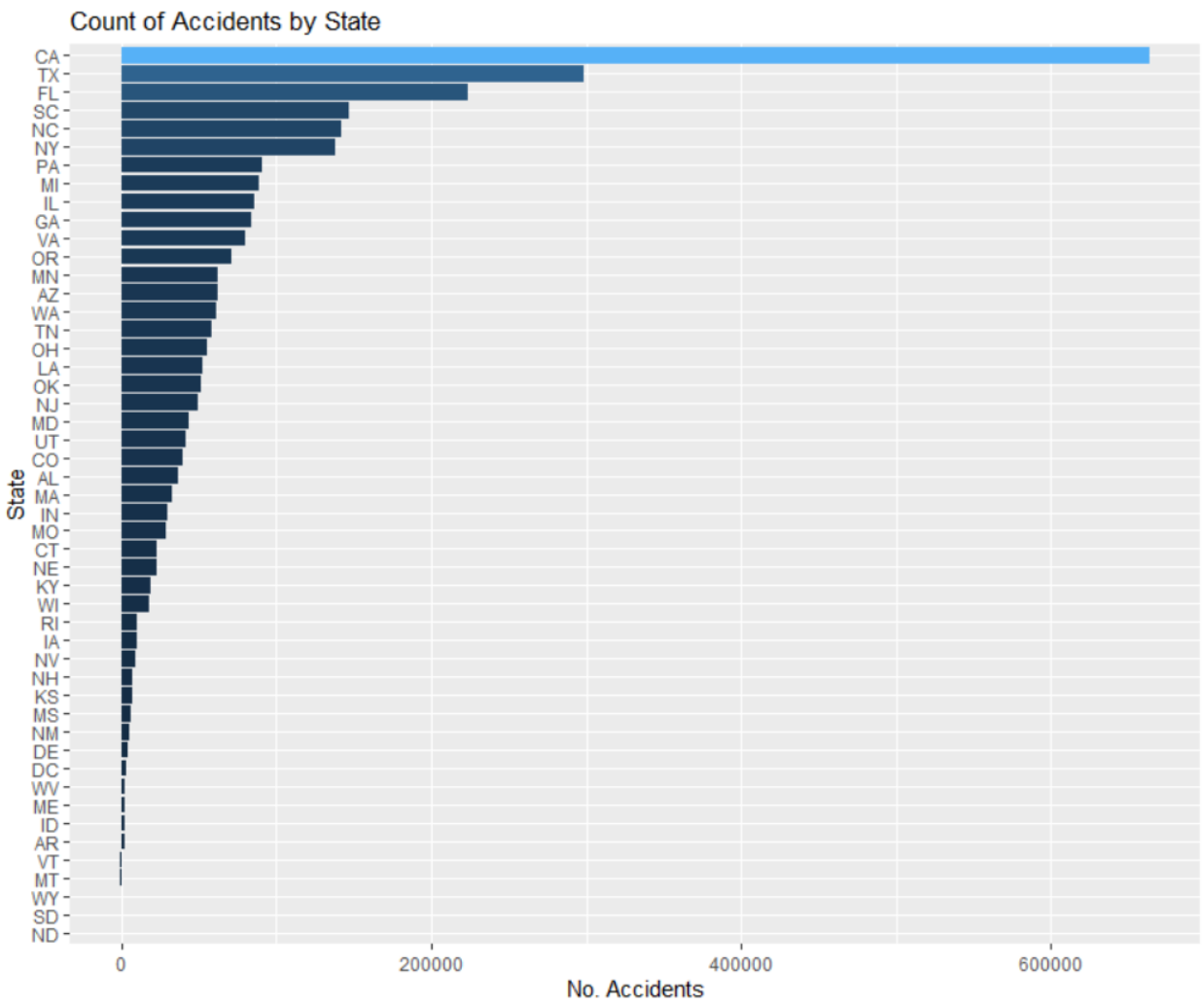
Output 8: Severity as responsible variable to Traffic Signals (on the left) and Traffic Signal and Stop signs (on the right) as Exploratory.



Output 9: Plot of Linear Regression Model with Traffic Signal.

I wanted to also show the linear regression model output for the relationship between Traffic Signal and Stop Signs. That reason being accident of those types appear to be more common but as the data shows they are not that impactful. As you can see Traffic Signal have a Multiple R-squared of .073 or 7.3% contribution rate. Leaving a very large amount of variation in the accident not contribute to Traffic Lights. If I add in Stop Signs it goes up a little to .081 or 8.1%. The relationship is negative. I don't believe traffic signal or stop signs are very high contributors to the rate of accidents or the severity of those accidents. That would be good news for Engineers and City Planner who have done a pretty good job of reducing the amount of accidents happening at that those locations.

Below in Output 10 I wanted to show how each state shows up when compared to traffic signals to provide context. If you look at the data at state level California should be top in all the above analysis. In Output 10 you will notice California is followed by Texas and then Florida. At a city level we see something very different. Without doing some further analysis it would be hard to determine what causes the state accidents rates verse city level accident rates to be very different. You will notice North Carolina is number five on the list and Georgia is even further down on the list yet the states respective cities lead in the number of accidents.



Output 10: Count of Accidents by State.

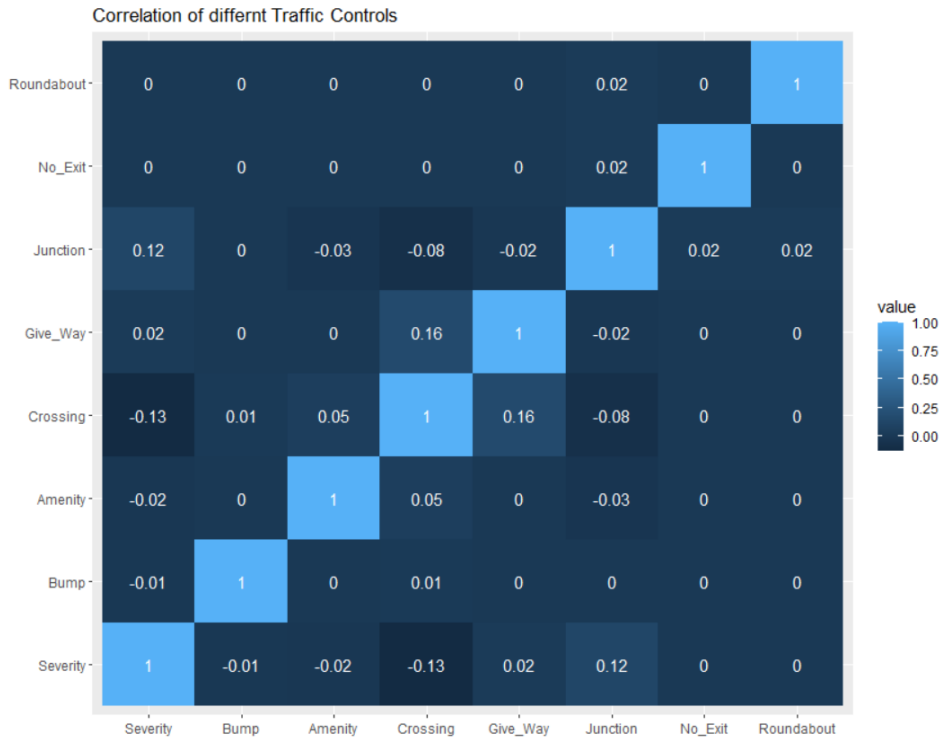
Question 4:

My fourth question related to how the individual points of interest contributed to the higher accident rates. In the dataset points of interest or POI are considered to be things like roundabouts, junctions, and railway crossings. Since most of the accident happen in clear weather there has to be other factors causing the accident rates to be so high. I took the different traffic controls and changed them from true and false to 1 and zero (0) respectively to allow me to do a correlation matrix. Below are the results from the correlation matrix and a heatmap to provide visual of what is happening.

	Severity	Bump	Amenity	Crossing	Give_Way	Junction	No_Exit
Severity	1.000000000	-0.00575928661	-0.0238990471	-0.128177234	0.0177984436	0.119310108	0.0042430446
Bump	-0.005759287	1.00000000000	-0.0011211942	0.007033774	-0.0007316673	-0.002858162	-0.0004171873
Amenity	-0.023899047	-0.00112119417	1.0000000000	0.046907019	0.0022888817	-0.030441955	-0.0029240819
Crossing	-0.128177234	0.00703377370	0.0469070194	1.000000000	0.1643922285	-0.077843279	0.0043144433
Give_Way	0.017798444	-0.00073166735	0.0022888817	0.164392228	1.0000000000	-0.015351718	-0.0017415863
Junction	0.119310108	-0.00285816192	-0.0304419549	-0.077843279	-0.0153517176	1.000000000	0.0155996979
No_Exit	0.004243045	-0.00041718726	-0.0029240819	0.004314443	-0.0017415863	0.015599698	1.0000000000
Roundabout	-0.003509490	-0.00007293391	-0.0007798909	-0.002042974	-0.0005089401	0.021850317	-0.0002901911

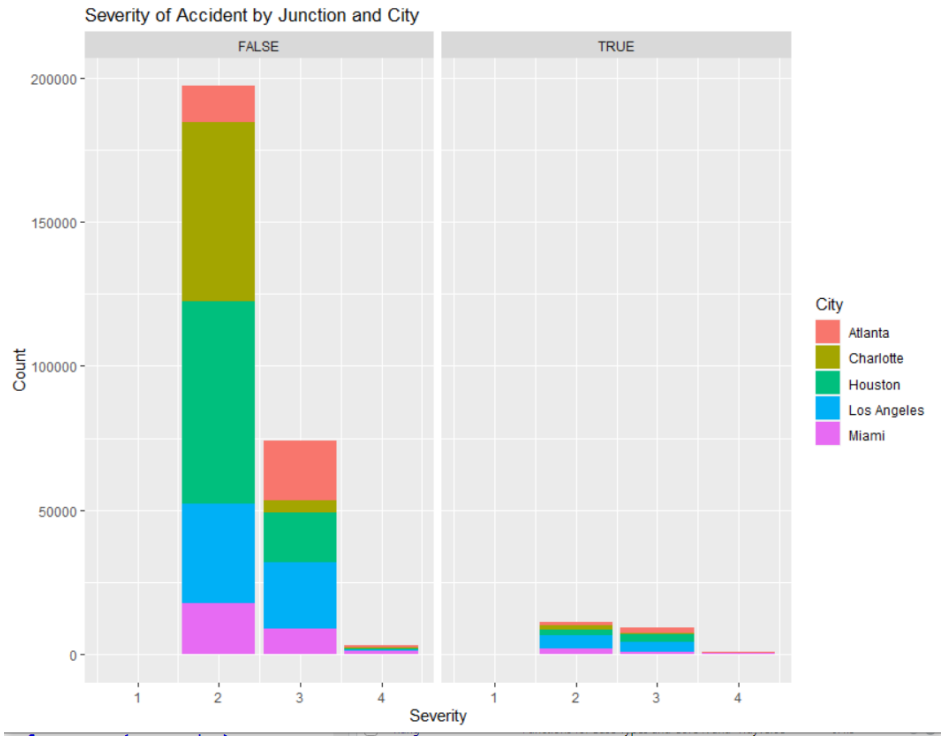
	Roundabout
Severity	-0.00350948963
Bump	-0.00007293391
Amenity	-0.00077989091
Crossing	-0.00204297446
Give_Way	-0.00050894014
Junction	0.02185031710
No_Exit	-0.00029019109
Roundabout	1.00000000000

Output11: Correlation of different traffic controls as rate of contribution to accidents.



Output12: Correlation of different Traffic Control to Severity.

Let me take a moment to explain this. The Severity variable is what really matters. Since you can't have more than one traffic control in an accident as the leading cause of the accident. There will probably be other factors. In the heatmap if you look at Severity and match its with the Traffic Control you will see a high correlation with Junction. In this case a Junction is defined as a few different things like type of road crossing like a roundabout or yield sign.

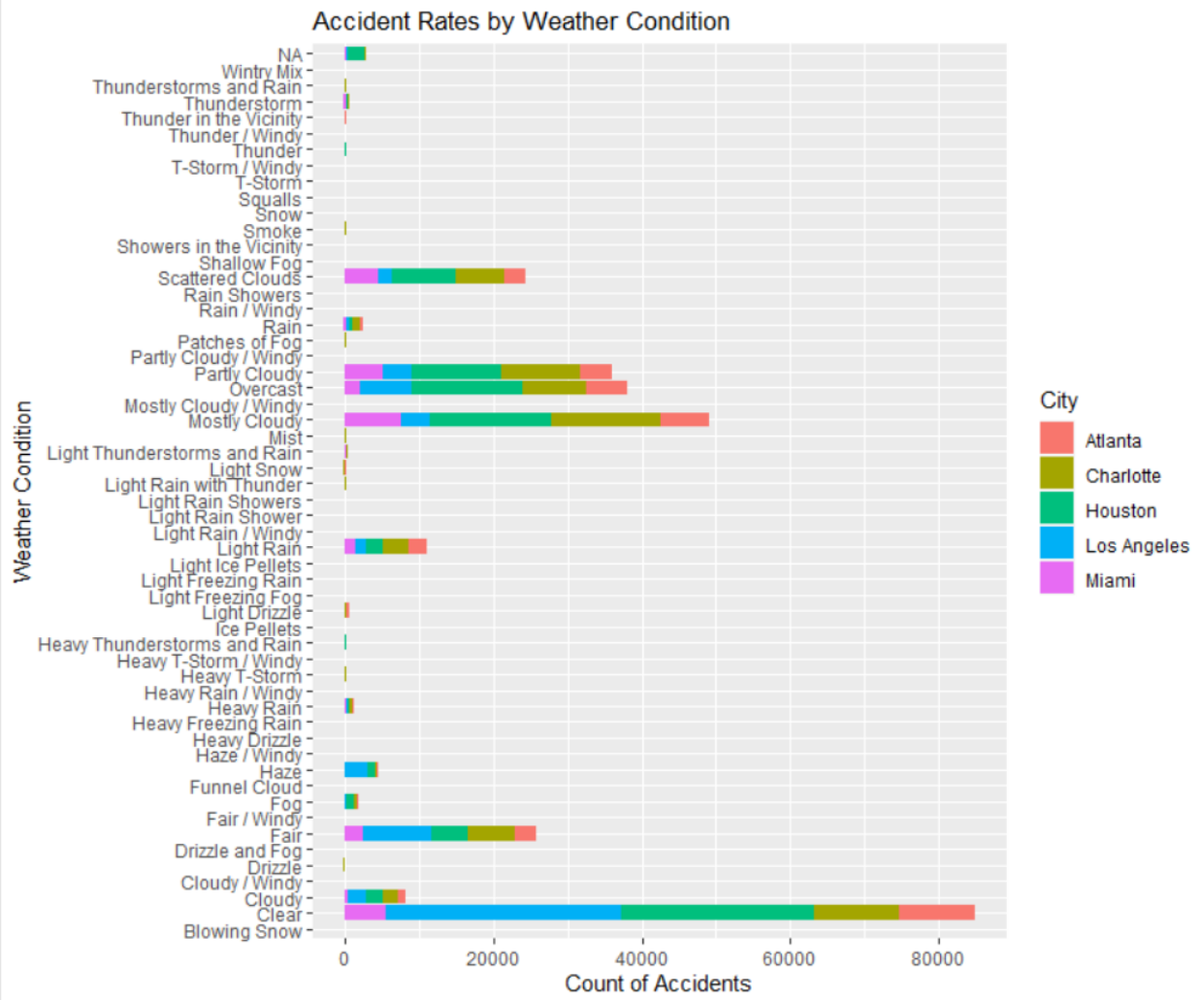


Output13: Severity of Accident by Junction and City.

This show that Junction is also not a leading contributor of accidents. At least compared to when a Junction is not involved as indicated by the False side of the grid.

Question 5:

For my fifth question I wanted to know how weather condition effected accident rates? I use the already bucketed weather condition by the source data. Unlike the one I created for Road Types. The data is already made similar across all regions and it removes the doubt that there were not standards. Your normal assumption would be bad weather would cause more accidents, but according to the data that is not the case. Since highest volume of accidents happen in California and since it never rains in California the weather conditions are clear. The other weather conditions like Mostly Cloudy, Overcast and Partly Cloudy followed clear. Based on the data it is safe to say weather is not a driving factor of accidents.



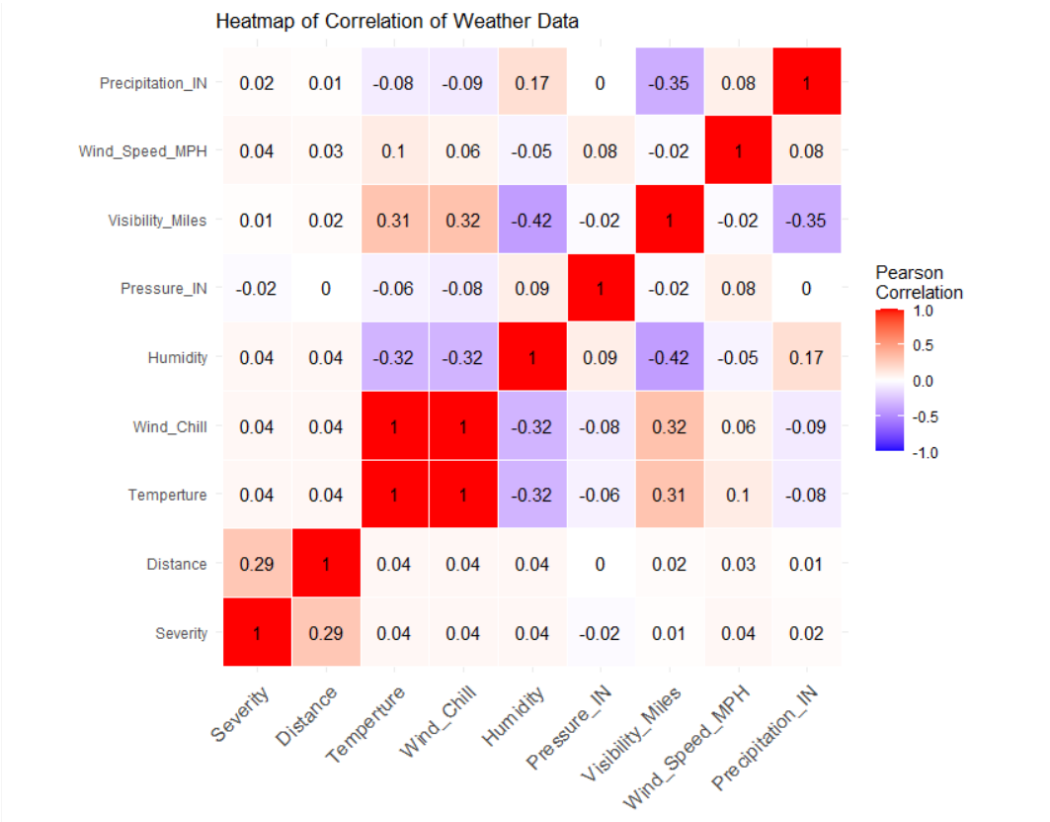
Output 14: The above show the rate of accidents by city and weather conditions.

The below part was not part of my intended analysis nor part of my questions, but I felt compelled to provide as much context as possible. I performed a correlation matrix for some of available data related to weather. I wanted to see if any of them had a high correlation to the number of accidents. Below in Output 15 you see the correlation matrix. From that you can gather that most of the items have some correlation, but it is not that strong. The one that stands out Severity and Distance. In this case Severity is the impact on traffic not the severity of the accident itself. Distance is the length of the impact of the accident. In this correlation it would prove

that these two are related which make perfect sense. If the Severity of the accident is high it would cause a larger effect on the traffic tie up. But notice how the different weather-related variable have little effect on Severity or Distance. In Output 16 I provided a heatmap to show the correlation visually.

```
> cor(usaccd_corr, use = "complete.obs", method = "pearson")
                    Severity     Distance  Temperture  Wind_Chill    Humidity
Severity          1.00000000  0.288059515  0.03668629  0.03559328  0.04122121
Distance          0.28805951  1.000000000  0.04097049  0.03893075  0.03989931
Temperture        0.03668629  0.040970492  1.00000000  0.99632528 -0.32029008
Wind_Chill        0.03559328  0.038930748  0.99632528  1.00000000 -0.31713002
Humidity          0.04122121  0.039899312 -0.32029008 -0.31713002  1.00000000
Pressure_IN      -0.01548505 -0.003573219 -0.06471939 -0.07768428  0.09198349
Visibility_Miles  0.01085008  0.019483011  0.30928012  0.31742445 -0.41834845
Wind_Speed_MPH    0.04272909  0.025403015  0.09506874  0.06398529 -0.04734977
Precipitation_IN  0.01918092  0.013193850 -0.08408937 -0.08692534  0.17084671
                  Pressure_IN Visibility_Miles Wind_Speed_MPH Precipitation_IN
Severity          -0.015485048       0.01085008     0.04272909      0.019180923
Distance          -0.003573219       0.01948301     0.02540302      0.013193850
Temperture        -0.064719392       0.30928012     0.09506874     -0.084089371
Wind_Chill        -0.077684281       0.31742445     0.06398529     -0.086925345
Humidity           0.091983490      -0.41834845    -0.04734977      0.170846706
Pressure_IN        1.000000000      -0.01568852     0.08073736      0.003861468
Visibility_Miles  -0.015688519       1.00000000    -0.01871502     -0.346986020
Wind_Speed_MPH     0.080737361      -0.01871502     1.00000000      0.084836466
Precipitation_IN   0.003861468      -0.34698602     0.08483647      1.000000000
>
```

Output 15: Correlation matrix for weather related data points in the US Accident dataset.



Output 16: Heatmap of the correlation matrix.

Question 6:

For my final question using machine learning to predict severity levels of accidents. In an earlier assignment I thought I would be able to use kNN to perform my machine learning analysis. After further analysis of the data I believe the best machine learning algorithm would be logistic regression. Why I choose logistic regression. The data is categorical and discrete in that it is true or false not a lot of numerical data.

The Severity variable which defines how bad or sever a backup or delay the accident caused has four categories. I reduce them to two Severity levels. The data contains four but Severity 1 and 4 are so small in number that they will be over shadowed by the most common, which are Severity 2 and Severity 3. It all kind of made sense to do a straight forward Logistic Regression model. The data is not a cluster type of data either. I think with the data being true or false the clustering would be pretty simple.

The method I followed base on reading some analysis and trying to understanding what I was trying to achieve as well. I wanted to understand if the severities of the traffic impacts could be predicted based off Traffic Signal, Junction, Side (which side the of the road way the accident happened occurred on) and Start Hour. I took these variables because they seemed to be ones to have the biggest impact on Severity. To get the data to a stage where I can perform the logistic regression on it. I removed a lot of 30 attributes that were in the original dataset. I was left with Severity, Traffic Signal, Junction, Side and Start Hour. I took the data and changed everything to a factor. Once that was accomplished I moved forward with splitting the data by state. This would allow me to compare each state on its own merit. Since each state have different traffic controls, rules and other factors. I still left the original cities and did not add additional data to the data set.

The accuracy of the logistic model is pretty low. I believe I needed additional time just to spend on this data and possibly having a contact who could have helped me better understand the data would have been great. For Houston, Texas the value was a mere 6.63%. I believe this requires more in-depth research as to the cause of such a low prediction rate. The North Carolina data I had a different issue. I was not able to get an expected accuracy percentage. This leads me to believe all those data points for Houston, Texas and Charlotte, North Carolina need some further cleaning or analysis to understanding what is going on with the source data.

Moving on to Los Angeles, CA I got a better expected accuracy of 65.3%. This is not very high meaning the model is just “ok” and you could probably predict at a better percentage without a model. Similar with Miami, FL I got an accuracy of 68.6%. The best I saw was from Atlanta, GA. The accuracy was 71.0%. Overall, the Logistic model is not great and given some more time I think I could find a better model to get better prediction of Severity of accidents.

In the below output of the model the accuracy is pretty much all over the place. Logistic Regression was not the best choice as a predictor model at least not for this analysis. A lot of the variables do not correlate to severity of the accident. Meaning they don’t relate to the cause of the accidents.

```
Call:
glm(formula = as.factor(Severity) ~ Junction + Traffic_Signal +
    Side + Road_Type + Start_Hour, family = "binomial", data = trainTX)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8193  -0.4830  -0.3485  -0.1775   2.9581

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.310387   0.082624 -15.860   <2e-16 ***
Junction[T.TRUE]       0.119261   0.062798   1.899   0.0575 .
Traffic_Signal[T.TRUE] -0.680754   0.041508 -16.401   <2e-16 ***
Side[T.R]              1.427508   0.067659  21.099   <2e-16 ***
Road_Type[T.Street]    -2.700522   0.035915 -75.193   <2e-16 ***
Start_Hour             0.054838   0.003498  15.678   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33894  on 32378  degrees of freedom
Residual deviance: 22697  on 32373  degrees of freedom
AIC: 22709

Number of Fisher Scoring iterations: 6
```

Output 17: Texas data. Left side the Logistic Regression Model. The right side Accuracy & Predicted Values from the Logistic Regression Model. The other for CA, NC, GA and FL are in the Appendix section.

CA		
	Predicted Value	
Actuals	FALSE	TRUE
2	7849	3893
3	2903	4946
Accuracy	0.65311	65.3106

NC		
	Predicted Value	
Actuals	FALSE	TRUE
2	19002	
3	1323	
Accuracy	??	#VALUE!

FL		
	Predicted Value	
Actuals	FALSE	TRUE
2	4400	1492
3	1293	1698
Accuracy	0.68648	68.64798

GA		
	Predicted Value	
Actuals	FALSE	TRUE
2	1524	2595
3	551	6187
Accuracy	0.71023	71.0233

TX		
	Predicted Value	
Actuals	FALSE	TRUE
2	9954	920
3	938	2066
Accuracy	0.06629	6.629197

Output 18: Predicted values of California, North Carolina, Georgia and Florida. More data is located in the Appendix section.

Conclusion:

I believe the data contained a lot of good information about rates of accidents some of the contributing factors to the accidents. Some of the insight I found from the analysis is definitely prime hours for accidents like between 8AM to 9AM block and then again the 4PM to 5PM block. The morning seems to have more accidents than in the evening block. You could use the data to avoid those times or make the case to your boss to come in earlier and leave earlier before peak accident hour starts. I also found out weather is not a big contributor to accident rates as majority of accidents occurred on clear days. Surface street have the largest amount of accidents, which I thought was due to traffic controls like Traffic Signal or Stop signs. There is no correlation between those and the severity of the impact to traffic. There is a very definitive regional fact to the data. While I assumed you could compare accidents across the entire United States I believe region naming and reporting causes issue with that analysis. In other words, everything is not the same. Which does make reporting a little more difficult.

I believe if compared on a state or city by city bases this information would provide value to cities that neighbor each other. I think City Planner and Civil Engineers who design our cities could find some insight from this data. It might help them pin point regional issues that can be resolved with minor changes like intersection or junctions or new different types of signage.

I do believe the data is lacking some information. If we had more biographical data like driver sex, age, number years of driving, last drivers test, how long they have lived in the area, type of car they are driving, speed limits, speed of vehicles involved in the accident, number of vehicles involved, technology usage and some other information the analysis could pinpoint possible reasons for rising accident rates. Future analysis should look closer as biographical data and maybe try to standardize information into smaller buckets. The one example would be weather. There are a lot of buckets for weather maybe narrow that down to five to ten.

Reference:

1. *U.S. Accidents (3.0 million records) A Countrywide Traffic Accident Dataset (2016-2019);* https://www.kaggle.com/sobhanmoosavi/us-accidents
2. Moosavi S., (2019, December) *US-Accidents: A Countrywide Traffic Accident Dataset*, https://smoosavi.org/datasets/us_accidents
3. Moosavi S., Hossein M., Parthasarathy S., Teodorescu R., & Ramnath R., (2019, November), *Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights,* https://arxiv.org/abs/1909.09638
4. Toeten, M., (2019, December) *What causes the accidents? (EDA),* https://www.kaggle.com/teoten/what-causes-the-accidents-eda
5. Zhuocheng, L., (2020,April) *US Accidents Project - Visualization and Modeling,* https://www.kaggle.com/zhuochenglin/us-accidents-project-visualization-and-modeling

Appendix:

Los Angeles, California:

```
Call:
glm(formula = as.factor(Severity) ~ Junction + Traffic_Signal +
    Side + Road_Type + Start_Hour, family = "binomial", data = trainCA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3175  -0.8892  -0.6160   1.0711   2.4263

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.786479   0.055541 -14.160   <2e-16 ***
Junction[T.TRUE]      -0.260514   0.030276  -8.605   <2e-16 ***
Traffic_Signal[T.TRUE] -0.769373   0.036801 -20.906   <2e-16 ***
Side[T.R]              1.110043   0.049195  22.564   <2e-16 ***
Road_Type[T.Street]   -1.025935   0.021843 -46.968   <2e-16 ***
Start_Hour            -0.004269   0.001734  -2.462   0.0138 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61569  on 45709  degrees of freedom
Residual deviance: 55886  on 45704  degrees of freedom
AIC: 55898

Number of Fisher Scoring iterations: 4
```

Charlotte, North Carolina:

```
> summary(glmtrainNC)

Call:
glm(formula = as.factor(Severity) ~ Junction + Traffic_Signal +
    Side + Road_Type + Start_Hour, family = "binomial", data = trainNC)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1133  -0.3611  -0.1823  -0.1483   3.4445

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.085802   0.113984 -18.299   <2e-16 ***
Junction[T.TRUE]        0.019488   0.077350   0.252    0.801
Traffic_Signal[T.TRUE] -1.401292   0.071087 -19.712   <2e-16 ***
Side[T.R]               1.807821   0.095796  18.872   <2e-16 ***
Road_Type[T.Street]    -2.456312   0.046640 -52.665   <2e-16 ***
Start_Hour              0.004603   0.004416   1.042    0.297
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23096  on 47423  degrees of freedom
Residual deviance: 16221  on 47418  degrees of freedom
AIC: 16233

Number of Fisher Scoring iterations: 7
```

Atlanta, Georgia:

```
> summary(glmtrainGA)

Call:
glm(formula = as.factor(Severity) ~ Junction + Traffic_Signal +
    Side + Road_Type + Start_Hour, family = "binomial", data = trainGA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0537  -1.0303   0.6994   0.8983   2.2739

Coefficients:
                       Estimate Std. Error z value      Pr(>|z|)
(Intercept)            -0.42138    0.07128  -5.912 0.00000000339 ***
Junction[T.TRUE]       -0.55353    0.04786 -11.566       < 2e-16 ***
Traffic_Signal[T.TRUE] -1.40827    0.04569 -30.824       < 2e-16 ***
Side[T.R]               1.33382    0.05097  26.171       < 2e-16 ***
Road_Type[T.Street]    -0.77000    0.03808 -20.222       < 2e-16 ***
Start_Hour              0.04639    0.00288  16.105       < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33592  on 25329  degrees of freedom
Residual deviance: 29921  on 25324  degrees of freedom
AIC: 29933

Number of Fisher Scoring iterations: 4
```

Miami, Florida:

```
Call:
glm(formula = as.factor(Severity) ~ Junction + Traffic_Signal +
    Side + Road_Type + Start_Hour, family = "binomial", data = trainFL)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2677  -0.8325  -0.6158   1.1302   2.3481

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.091797   0.076768  -1.196  0.23179
Junction[T.TRUE]      -0.520472   0.051044 -10.197  < 2e-16 ***
Traffic_Signal[T.TRUE] -1.576676   0.055691 -28.311  < 2e-16 ***
Side[T.R]              0.076593   0.063132   1.213  0.22505
Road_Type[T.Street]   -1.032618   0.035071 -29.443  < 2e-16 ***
Start_Hour             0.009784   0.003169   3.088  0.00202 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26482  on 20726  degrees of freedom
Residual deviance: 23614  on 20721  degrees of freedom
AIC: 23626

Number of Fisher Scoring iterations: 5
```

List of cities and states in the 295,068 rows count.

State	City	Count	Selected City/States
TX	Houston	93,245	1
NC	Charlotte	67,917	2
CA	Los Angeles	65,851	3
TX	Austin	58,553	N/S
TX	Dallas	57,823	N/S
NC	Raleigh	39,623	N/S
GA	Atlanta	37,576	4
FL	Miami	31,097	5

The green highlights were brought into R for search.

#	Attribute	Description	Nullable
1	ID	This is a unique identifier of the accident record.	No
2	Source	Indicates source of the accident report (i.e. the API which reported the accident.).	No
3	TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.	Yes
4	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	No
5	Start_Time	Shows start time of the accident in local time zone.	No

#	Attribute	Description	Nullable
6	End_Time	Shows end time of the accident in local time zone.	No
7	Start_Lat	Shows latitude in GPS coordinate of the start point.	No
8	Start_Lng	Shows longitude in GPS coordinate of the start point.	No
9	End_Lat	Shows latitude in GPS coordinate of the end point.	Yes
10	End_Lng	Shows longitude in GPS coordinate of the end point.	Yes
11	Distance(mi)	The length of the road extent affected by the accident.	No
12	Description	Shows natural language description of the accident.	No
13	Number	Shows the street number in address field.	Yes
14	Street	Shows the street name in address field.	Yes
15	Side	Shows the relative side of the street (Right/Left) in address field.	Yes
16	City	Shows the city in address field.	Yes
17	County	Shows the county in address field.	Yes
18	State	Shows the state in address field.	Yes
19	Zipcode	Shows the zipcode in address field.	Yes
20	Country	Shows the country in address field.	Yes
21	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	Yes
22	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	Yes
23	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	Yes
24	Temperature(F)	Shows the temperature (in Fahrenheit).	Yes
25	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	Yes
26	Humidity(%)	Shows the humidity (in percentage).	Yes
27	Pressure(in)	Shows the air pressure (in inches).	Yes
28	Visibility(mi)	Shows visibility (in miles).	Yes
29	Wind_Direction	Shows wind direction.	Yes
30	Wind_Speed(mph)	Shows wind speed (in miles per hour).	Yes
31	Precipitation(in)	Shows precipitation amount in inches, if there is any.	Yes
32	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Yes
33	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	No
34	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	No

#	Attribute	Description	Nullable
35	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	No
36	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	No
37	Junction	A POI annotation which indicates presence of junction in a nearby location.	No
38	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	No
39	Railway	A POI annotation which indicates presence of railway in a nearby location.	No
40	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	No
41	Station	A POI annotation which indicates presence of station in a nearby location.	No
42	Stop	A POI annotation which indicates presence of stop in a nearby location.	No
43	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	No
44	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.	No
45	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	No
46	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Yes
47	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	Yes
48	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	Yes
49	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	Yes