Road Accidents Data Analysis and Severity Prediction

Anusha Das

August 2020

1. Introduction

1.1 Background

Road accident is most unwanted thing to happen to a road user, though they happen quite often. The most unfortunate thing is that we do not learn from our mistakes on road. Most of the road users are quite aware of the general rules and safety measures while using roads but it is only the laxity on part of road users, which cause accidents and crashes. While human errors can be corrected. There are other external factors also that play a part in accidents. If we could find these factors, and how much they influence an accident, we could predict severity of these accidents.

1.2 Problem

This project aims to find out the different factors that contribute to a severe road accident and predict the severity of accidents**.**

1.3 Interest

The target audience here are the common citizens who will be made aware of the conditions that lead to an accident so that they can adopt better road safety habits, which will lead to lesser number of casualties and severe damages. We can also incorporate this model in a real-time accident risk prediction model or develop a new real-time severe accident risk prediction. Policy implications of this project can be explored.

2. Data

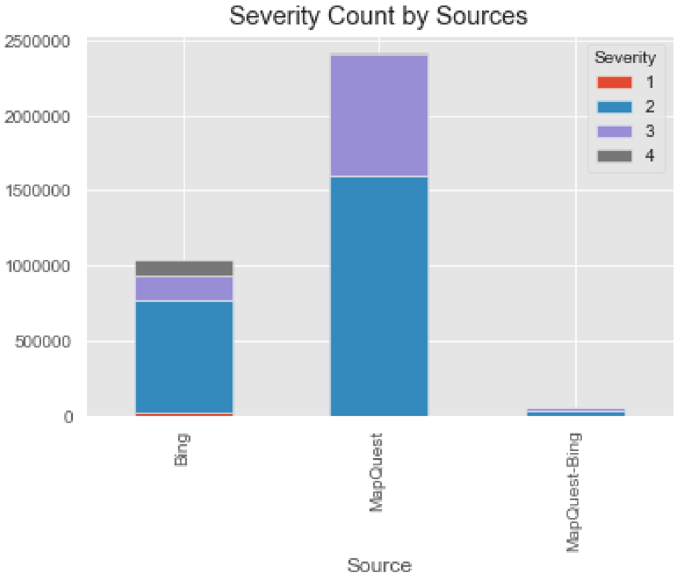
The dataset used for this analysis can be found here(Source).

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.5 million accident records in this dataset.

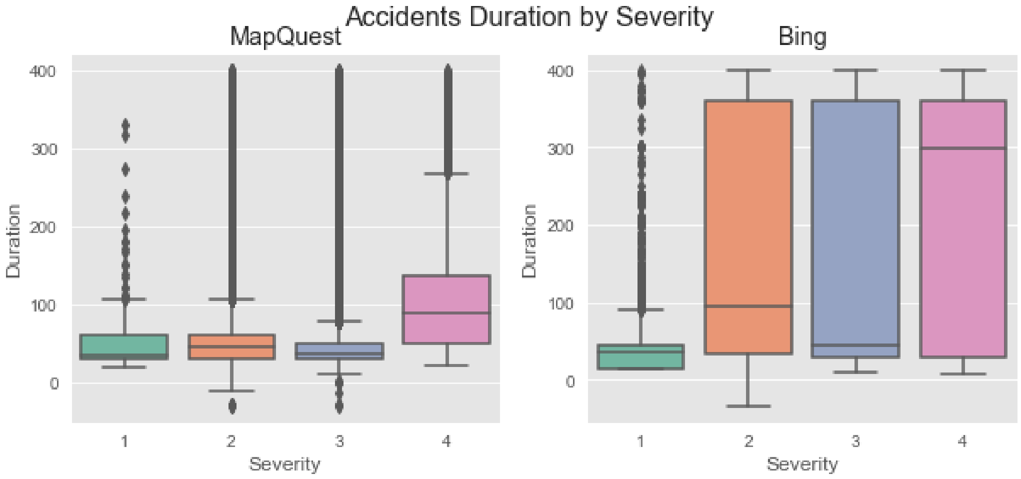
3. Methodology

3.1 Data Cleaning & Pre-processing

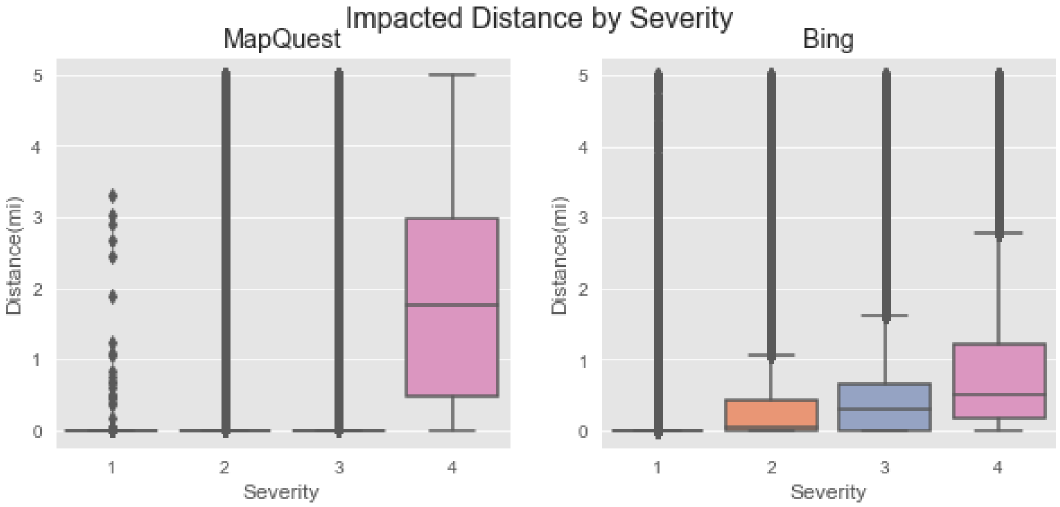
These data came from two sources, *MapQuest* and *Bing*, both of which report severity level but in a different way. *Bing* has 4 levels while *MapQuest* has 3. And according to the dataset creator, there is no way to do a 1:1 mapping between them. Since severity is what we really care about in this project, it was crucial to figure out the difference.



The stacked bar chart shows that two data providers reported totally different proportions of accidents of each level. *MapQuest* reported so rare accidents with severity level 4 which cannot even be seen in the plot, whereas *Bing* reported almost the same number of level 4 accidents as level 3. Meanwhile, *MapQuest* reported much more level 3 accidents than *Bing* in terms of proportion. These differences may be due to the different kinds of accidents they tend to collect or the different definitions of severity level, or the combination of them. If the latter is the case, we cannot use the data from both of them at the same time. To check it out, we can examine



the distribution of accidents with different severity levels across two main measures, ‘Duration’ and ‘Impacted Distance’**.**



Two differences are obvious in the above plots. The first is that the overall duration and impacted distance of accidents reported by *Bing* are much longer than those by *MapQuest*. Second, same severity level holds different meanings for *MapQuest* and *Bing*. *MapQuest* seems to have a clear and strict threshold for severity level 4, cases of which nevertheless only account for a tiny part of the whole dataset. *Bing*, on the other hand, does not seem to have a clear-cut threshold, especially regards duration, but the data is more balanced.

It is hard to choose one and we can’t use both. I decided to select *MapQuest* because serious accidents are what we really care about and the sparse data of such accidents is the reality we must confront. Hence, data reported from *Bing* and ‘Source’ column were dropped.

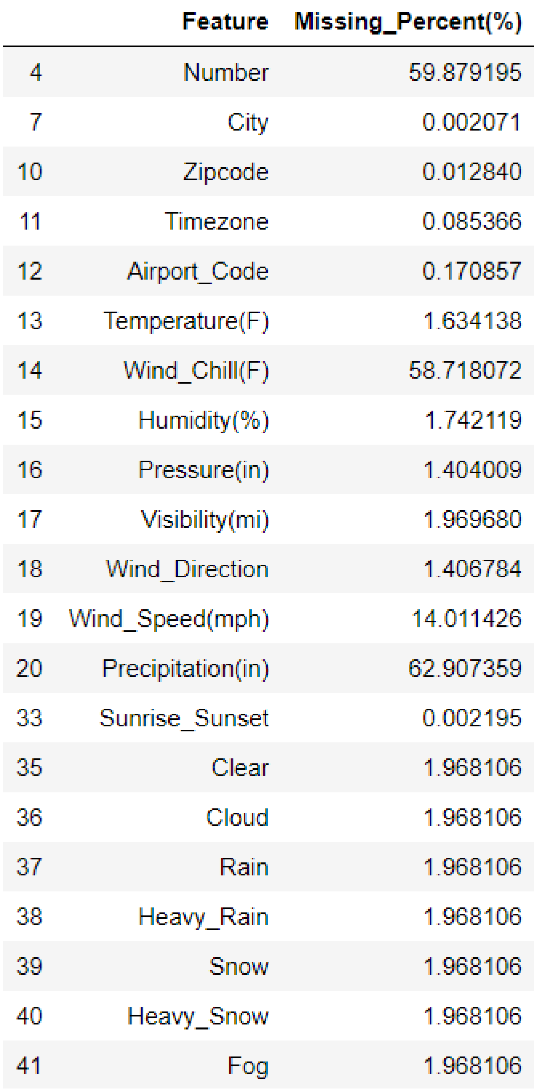
Features 'ID' does not provide any useful information about accidents themselves. 'TMC', 'Distance(mi)', 'End_Time' (we have start time), 'End_Lat', and 'End_Lng'(we have start location) can be collected only after the accident has occurred and can’t be predictors of accident. For 'Description', the POI features have already been extracted from it by dataset creators. Hence, these columns were dropped. Some other redundant features such as ‘Country’ and ‘Turning_Loop’ which had only one class were dropped. The data set mentioned 4 features that described the period of day the accident occurred, I decided to keep the most relevant one and drop the others.

Next, we see some chaos in the categorical features – ‘Wind_Direction’ and ‘Wind_Condition’. These features had more than one type of names for a condition. These were simplified to focus on main conditions.

Year, Month, Weekday, Day, Hour and Minute was extracted from ‘Start_Time’ for each accident.

The dataset also had a lot of missing data. More than 60% percent of 'Number', 'Wind_Chill(F)', and 'Precipitation(in)' was missing. Dropping NaN values and value imputation would not work for these features. 'Number' and 'Wind_Chill(F)' were dropped because they are not highly related to severity according to previous research, whereas 'Precipitation(in)' could be a useful predictor and hence can be handled by separating feature. A new feature for missing values in ‘Precipitation(in)’ was added and missing values were replaced by median.

The counts of missing values in some features were much smaller compared to the total sample. Hence, missing values were dropped in ‘City’, ‘Zipcode’, ‘Airport_Code’, ‘Sunrise_Sunset’.



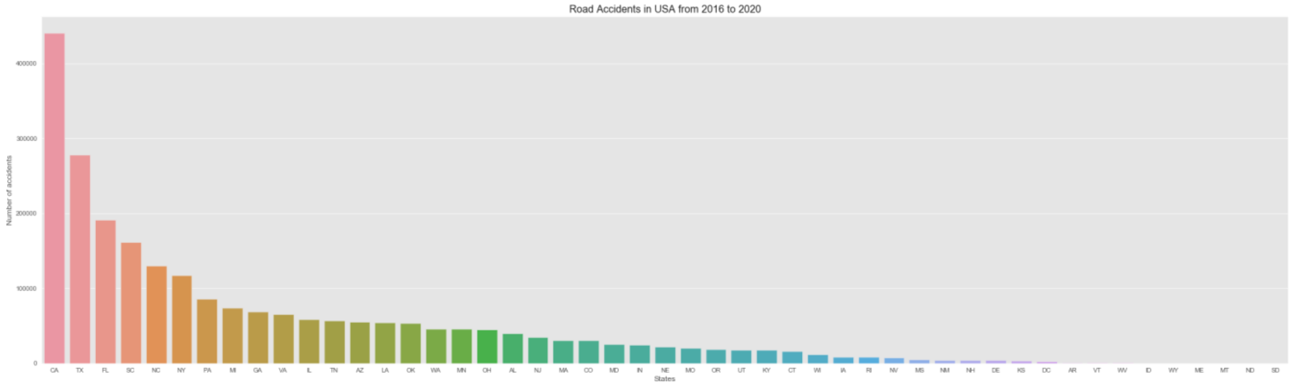
	Feature	Missing_Percent(%)
4	Number	59.879195
7	City	0.002071
10	Zipcode	0.012840
11	Timezone	0.085366
12	Airport_Code	0.170857
13	Temperature(F)	1.634138
14	Wind_Chill(F)	58.718072
15	Humidity(%)	1.742119
16	Pressure(in)	1.404009
17	Visibility(mi)	1.969680
18	Wind_Direction	1.406784
19	Wind_Speed(mph)	14.011426
20	Precipitation(in)	62.907359
33	Sunrise_Sunset	0.002195
35	Clear	1.968106
36	Cloud	1.968106
37	Rain	1.968106
38	Heavy_Rain	1.968106
39	Snow	1.968106
40	Heavy_Snow	1.968106
41	Fog	1.968106

I used Value Imputation for continuous weather features with missing values. Before imputation, weather features were grouped by location and time first, to which weather is naturally related. 'Airport_Code' was selected as location feature because the sources of weather data are airport-based weather stations. Then the data was grouped by 'Start_Month' rather than 'Start_Hour' because using the former is computationally cheaper and thus remain fewer missing values. Finally, missing values were replaced by median value of each group and remaining missing values were dropped. Same method was used for categorical weather feature ‘Wind_Direction’, but mode was used rather than median to replace missing values.

All the remaining missing values from other features were dropped, thus giving our final dataset ready for exploratory data analysis.

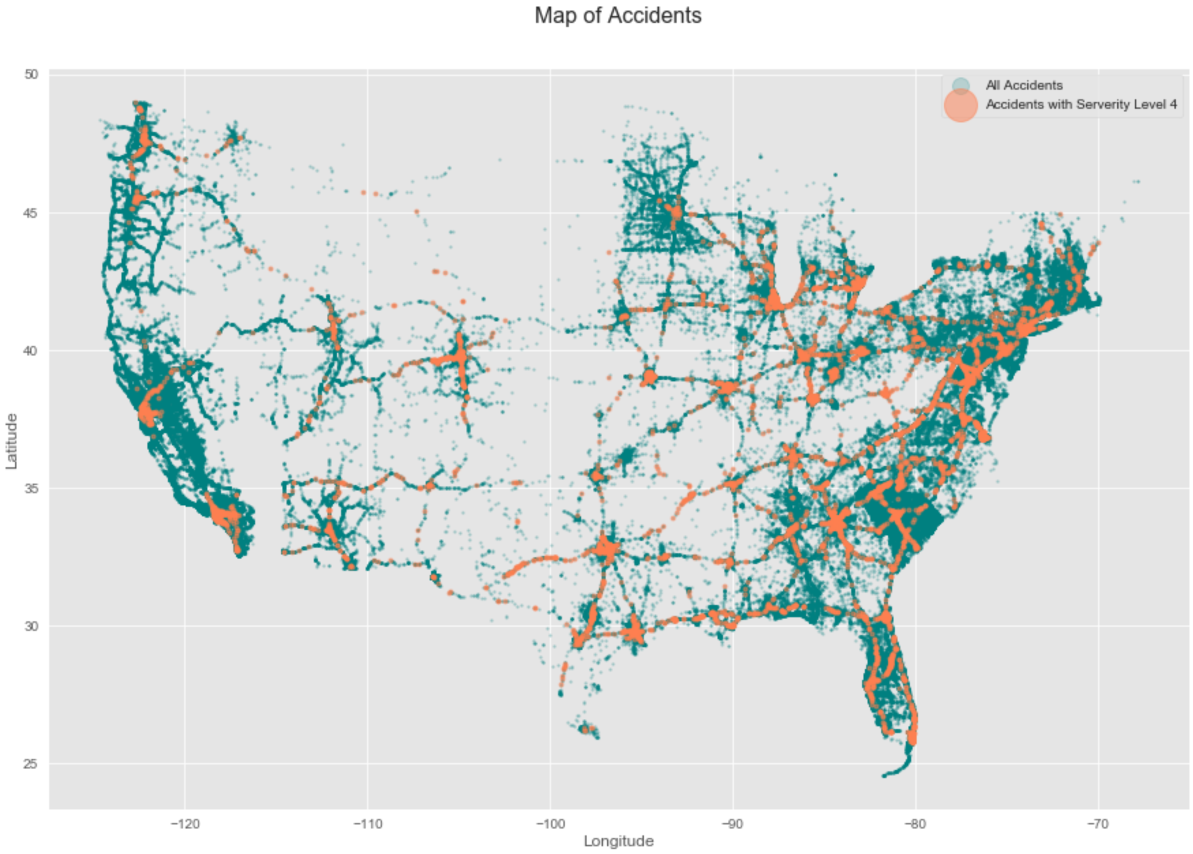
3.2 Exploratory Data Analysis

The dataset contains accident records from 08/02/2016 05:46 to 30/06/2020 23:18. In this period there have been 2.3 million total accidents, with an average of 1506 accidents per day in the USA.



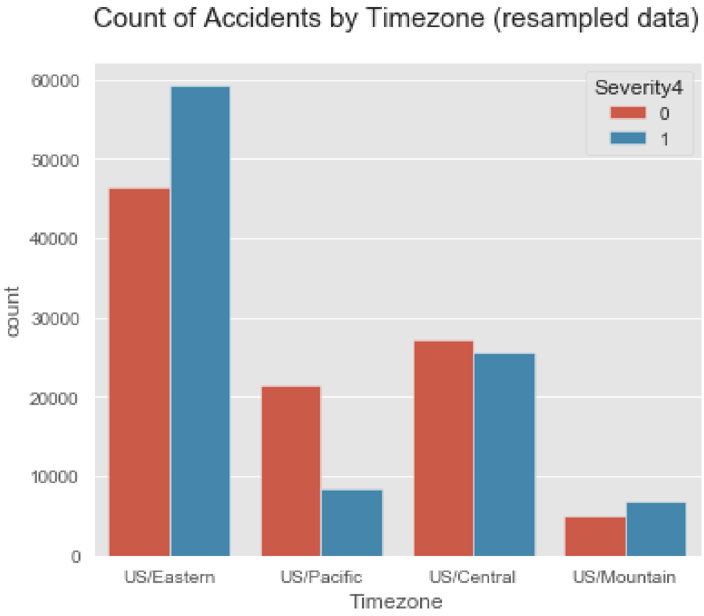
It was found that California, Texas, and Florida were the top 3 states with the highest number of accidents recorded from 2016-2020.

The data was so unbalanced that I could hardly do exploratory analysis. To address this issue, the combination of over- and under-sampling was used since the dataset is large enough. level

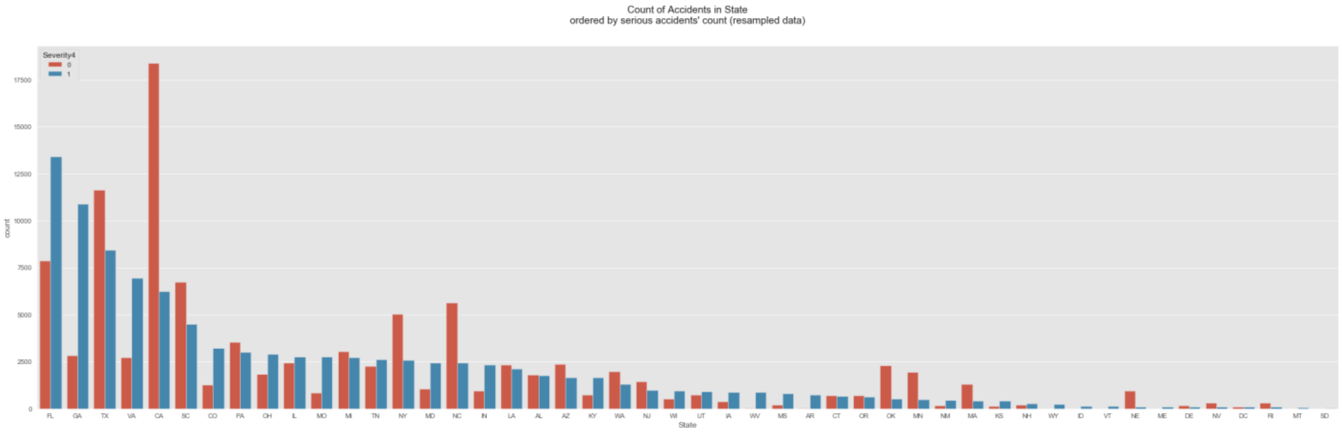


4 will be randomly oversampled to 100000 and other levels will be randomly under sampled to 100000.

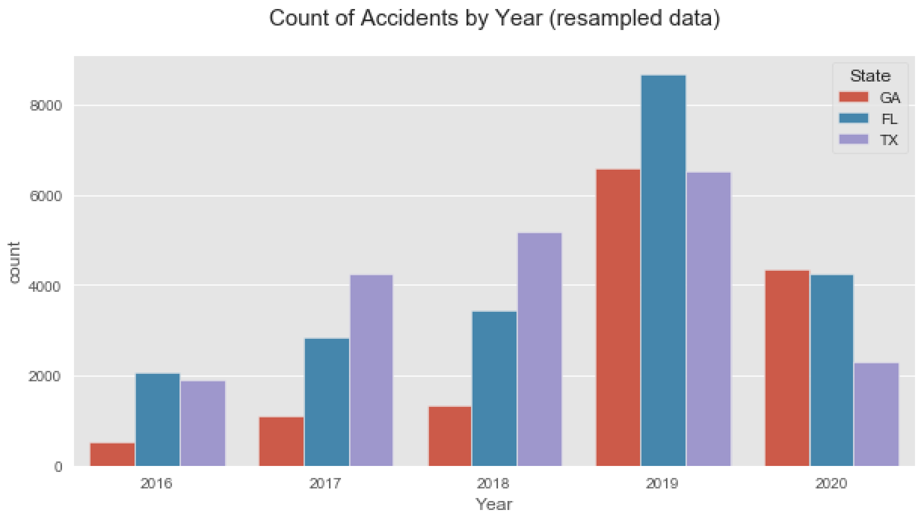
Exploratory analysis was done on resampled data.



We see that Eastern states have highest count of severe accidents followed by Central states. This could be because Eastern states of USA are more densely populated than the rest of the country.

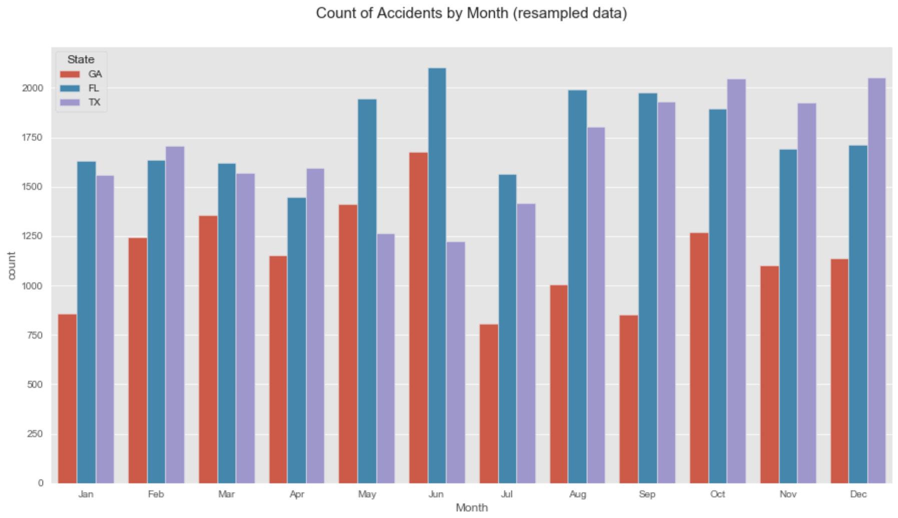


We see that Florida, Georgia and Texas are the 3 states with most severe accidents. Hence, we try to find the factors that could influence such accidents.



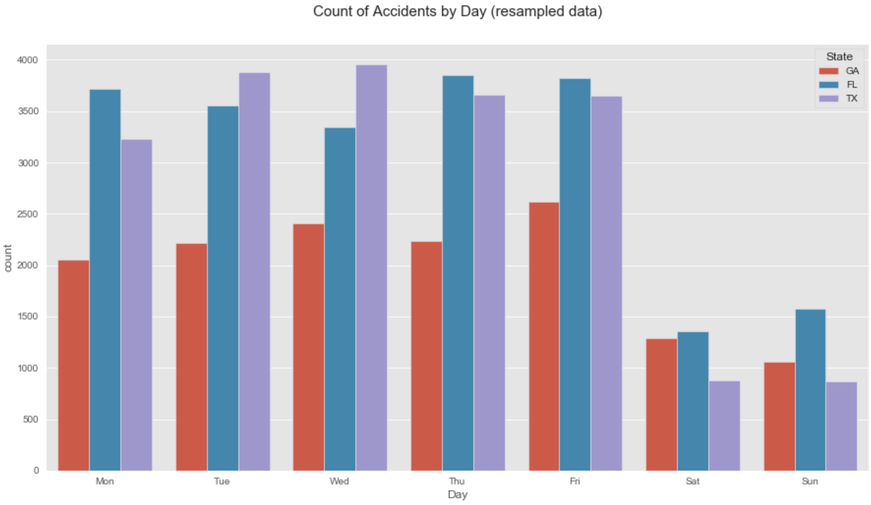
The above plot shows the number of accidents by year. The number of accidents is increasing by the year and there is not enough data for 2020 and I would guess the accidents may decrease by the end of 2020 because of current lockdown. But we also see that there is a huge amount of increase in 2019 in GA and FL.

This shows the bar plot of accidents by month.



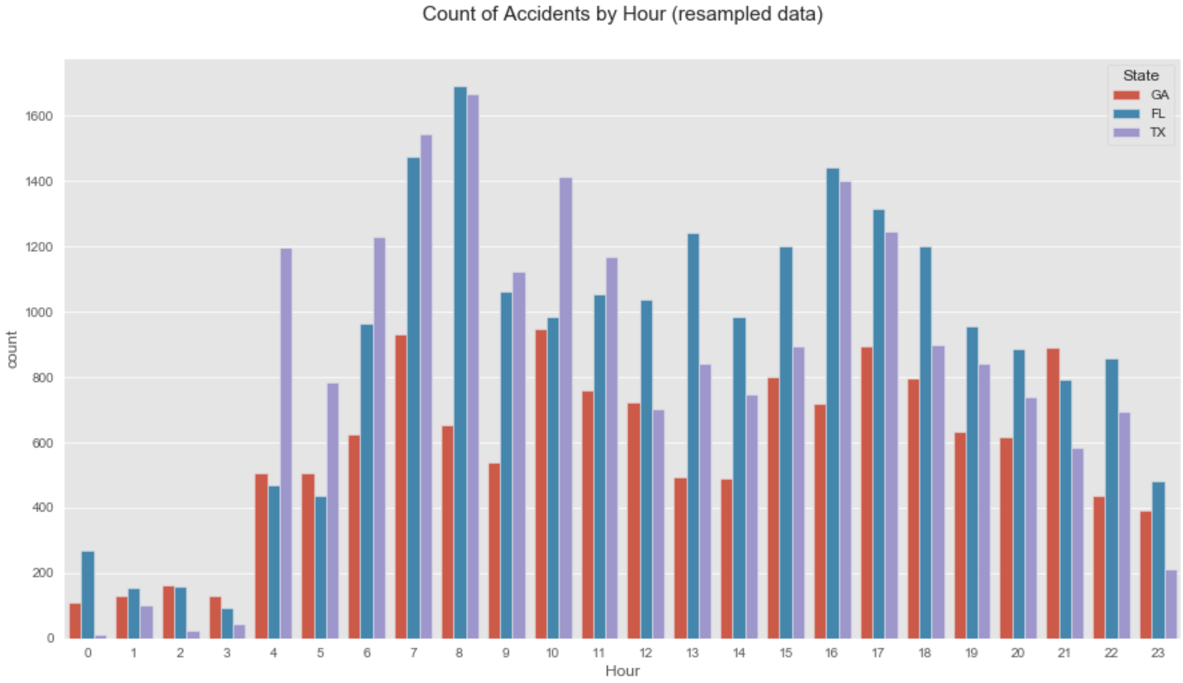
- GA - The number of accidents keeps fluctuating. Increase in summer months.
- FL - Almost uniform till April with an increase in May then sudden decrease in July and again increasing in August.
- TX - Almost uniform till April followed by a steady increase from May till October,

This shows the bar plot of accidents by Day.

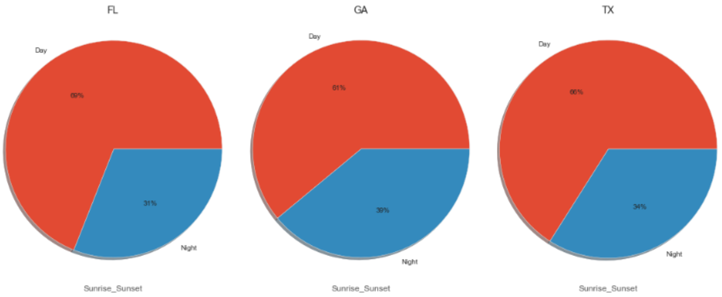


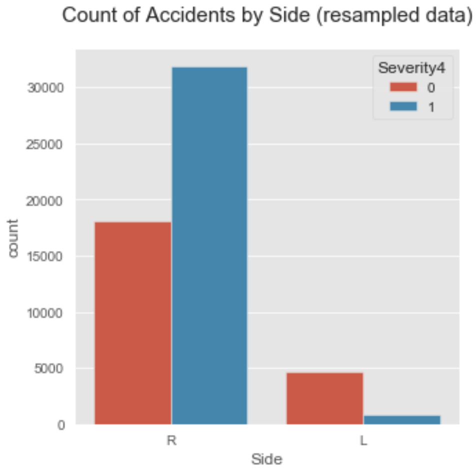
- GA - a little peak is seen on Friday.
- FL - a dip on Wednesday otherwise almost uniform
- TX - Monday has lesser accidents than other weekdays

Accidents are significantly less on weekends.

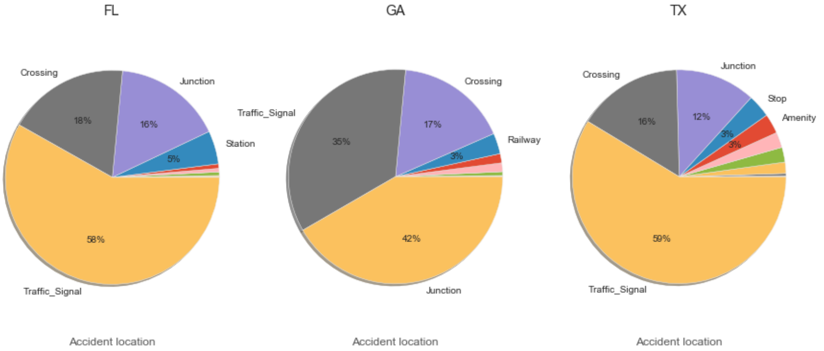


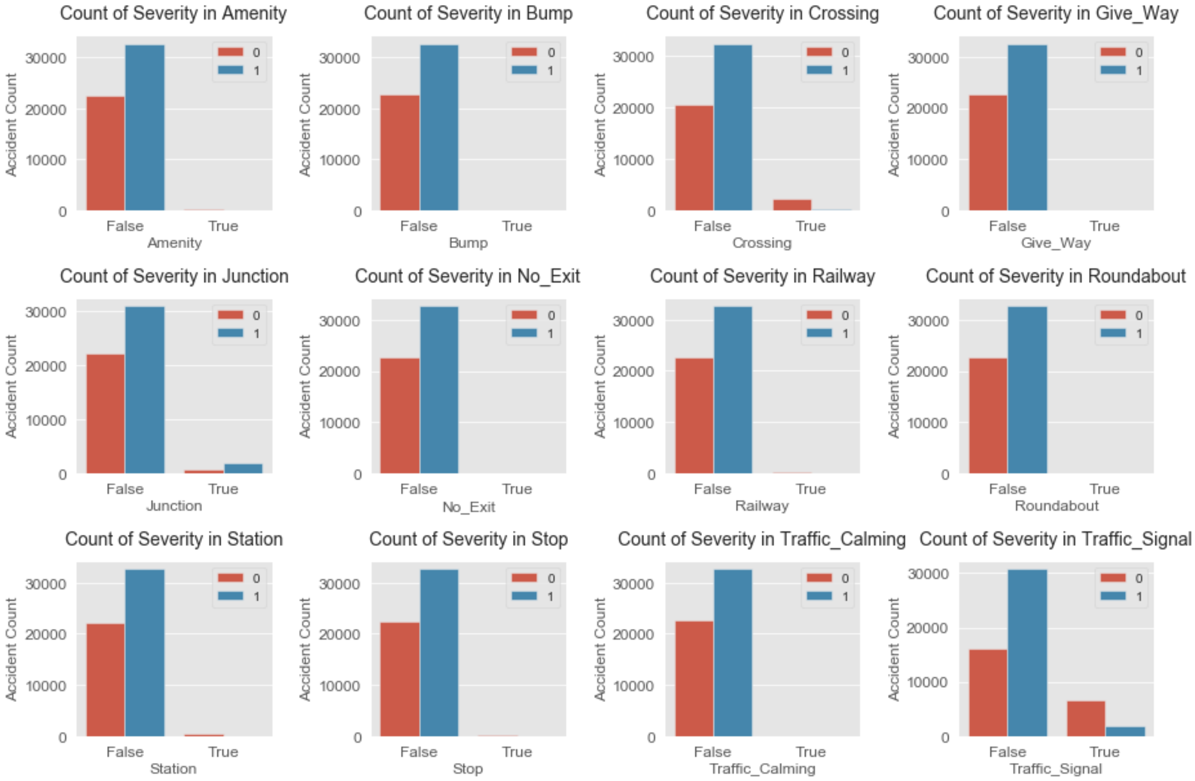
This shows the bar plot of accidents by Hour. Most of the severe accidents tend to happen during morning office hours or later in the evening. These are peak traffic periods as many people travel to offices and back to home.



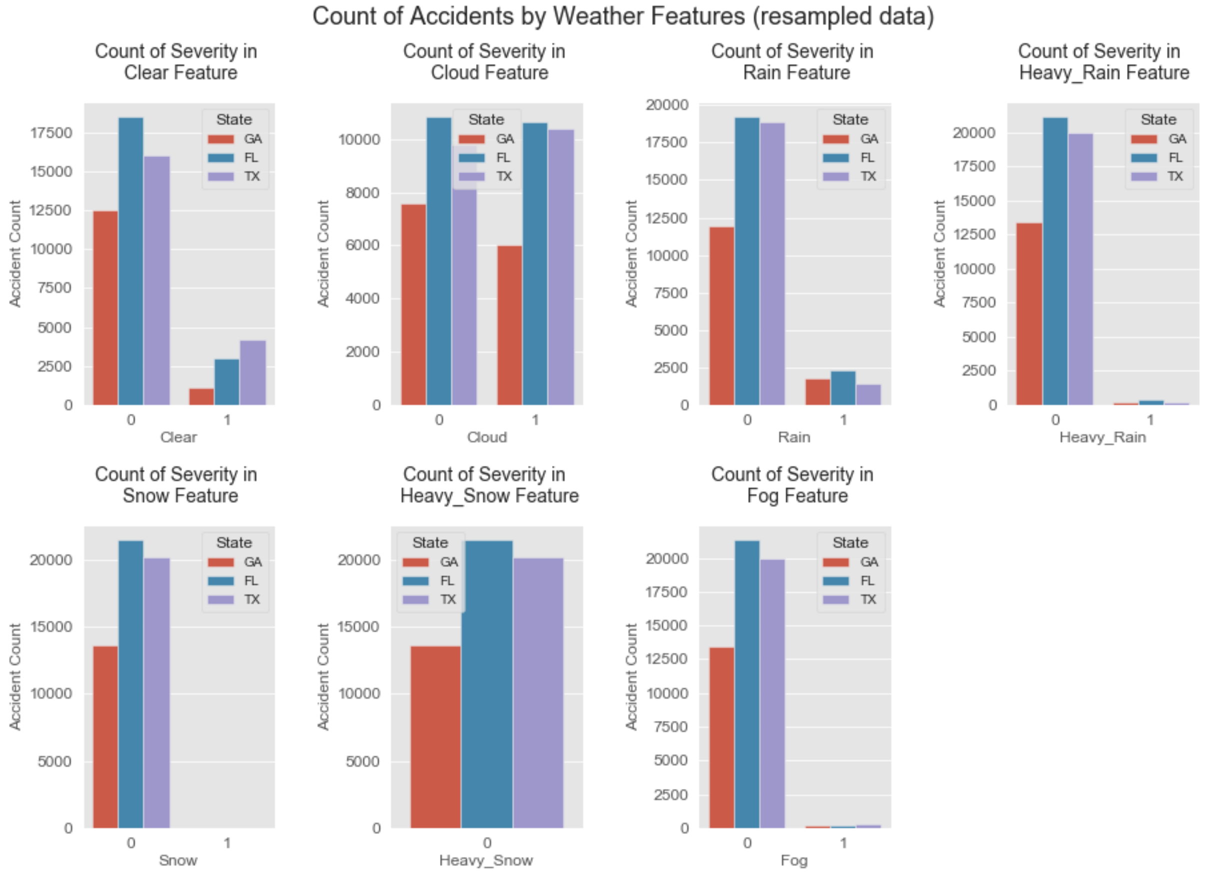


Over 60% accidents happened before sunset and on the right side of the lane.

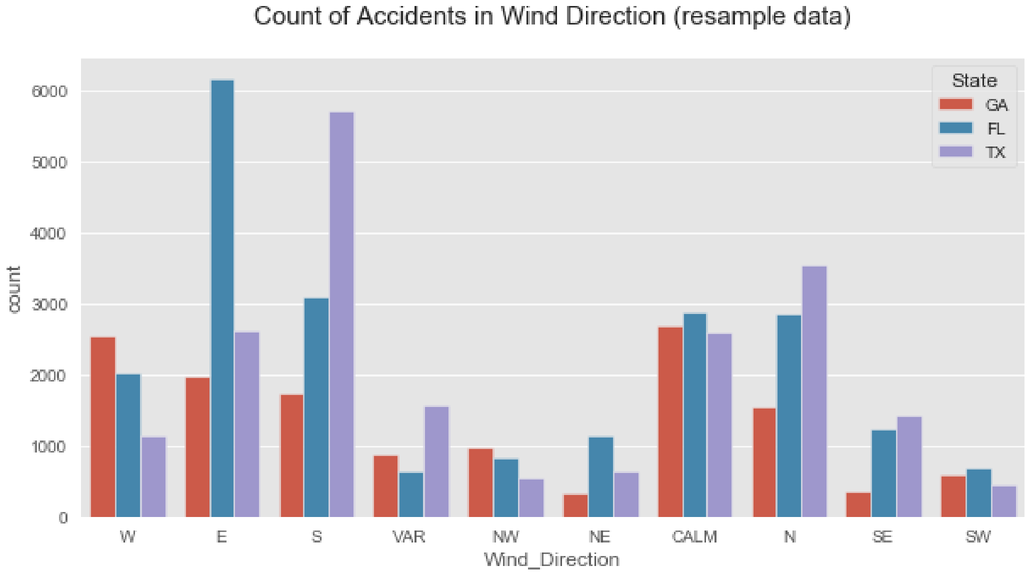




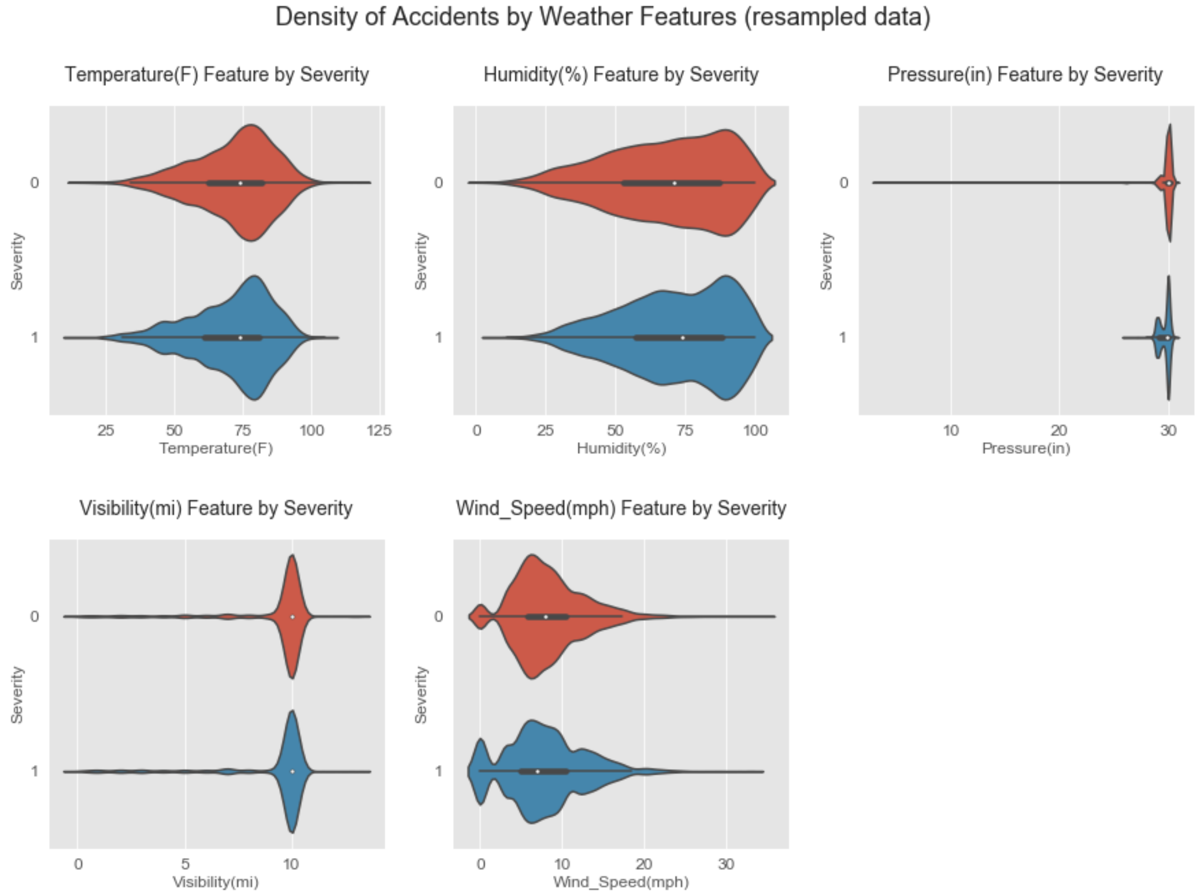
Accidents near traffic signal and crossing are much less likely to be serious. Accidents near the junction are more likely to be serious. Maybe it is because people usually slow down in front of crossing and traffic signal, but junction and severity are highly related to speed. Other POI features are so unbalanced that it is hard to tell their relationship with severity from plots.



The above plots show the count of accidents in certain weather conditions. Most accidents occurred when the weather was not clear. The presence of clouds does not seem to affect count of accidents. We would expect a high count of accidents in the presence of rain, snow, or fog, but interestingly that is not the case here.



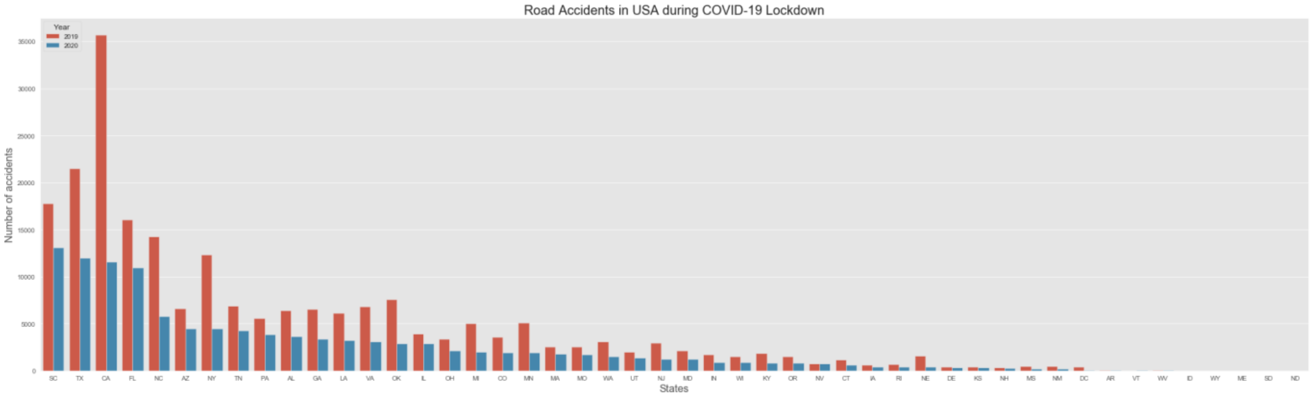
TX and FL have a steep peak in accidents when the wind blows from South and East, respectively.



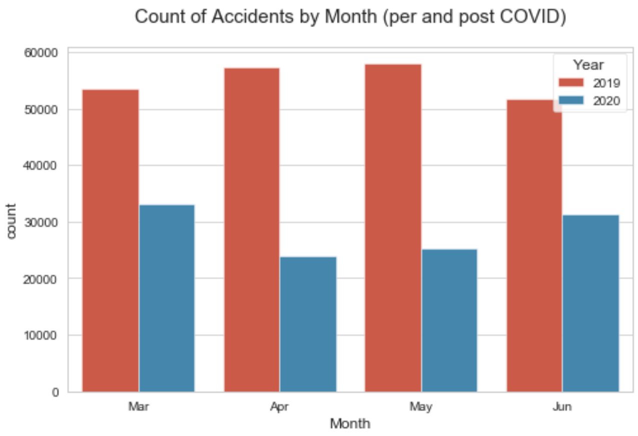
With increasing temperature and humidity, the accidents increase. This agrees with our previous observation of increase in accidents in the summer months. Statistics of visibility show that most of the values of visibility are around 10 miles.

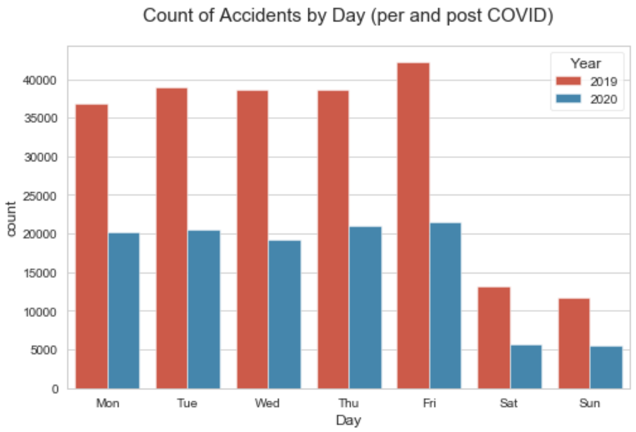
Impact of COVID-19 on accidents

I compared records of accidents from March 2020(when lockdown was imposed in most states of USA) till June 2020 with 2019 records to see difference in numbers and trends.



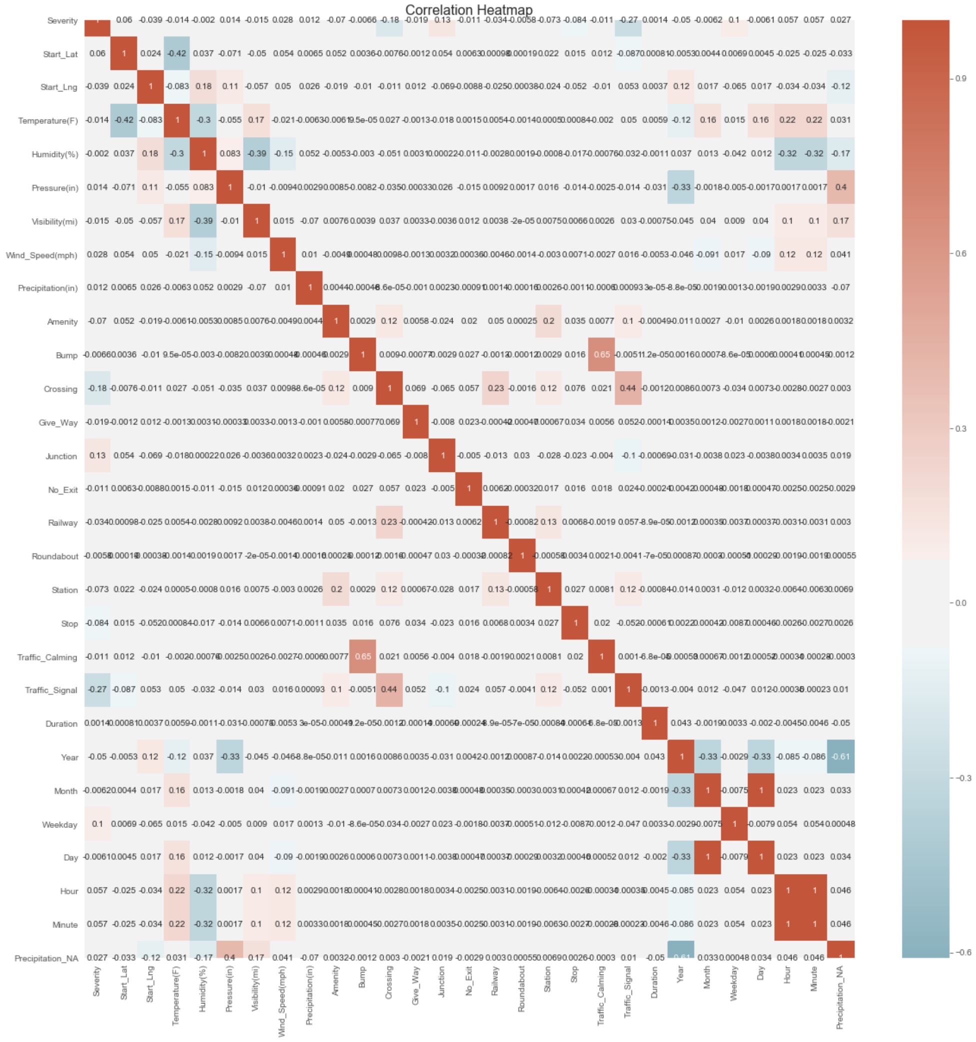
It is observed that SC, TX, and CA have taken a lead in the pandemic. But CA and TX have gone down in numbers very significantly.





As expected, post lockdown, the numbers have decreased by half. A slight increase is observed from April till June as the country stays in lockdown.

3.3 Predictive Modelling



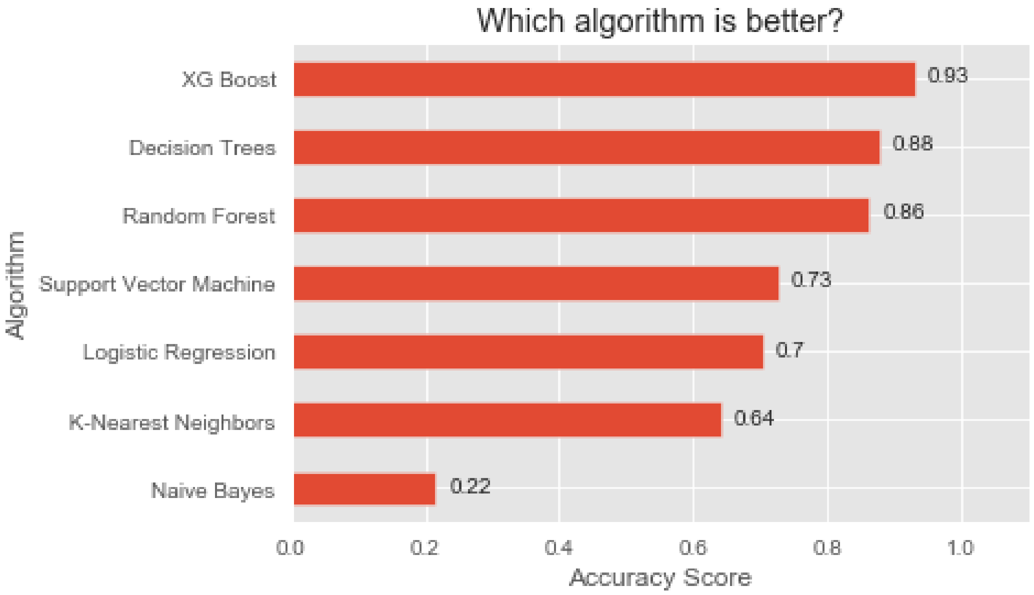
Severity does not show strong correlation with any of the feature and hence all the features are used for machine learning model.

The accident dataset is around 3 million and the computational power of my laptop is not enough to build a model on large data. From Exploratory Data Analysis (EDA), the state of

California has more accidents compared to other states. Machine learning models are built for Los Angeles, California. We choose records of California to build a model because it has the greatest number of accidents. Using records for every state won't be possible on a home computer.

Seven types of model were used to predict severity of accidents. I applied K-Nearest Neighbors, Logistic Regression, Support Vector Machines, XG Boost, Decision Trees, Random Forest and Naïve Bayes models to the dataset using accuracy score as the evaluation metric.

4. Results



It was found that XG Boost gave highest accuracy of 93% followed by Decision Trees with an accuracy of 88%.

5. Conclusion

A large dataset of around 1 GB consists of 3 million accidents is used in this project. To analyse the accidents, the dataset is cleaned systematically, and EDA is performed using several iterations. Through EDA the key findings were that Country-wide accident severity can be accurately predicted with limited data attributes (location, time, weather, and POI). An accident is much less likely to be severe if it happens near traffic signal while more likely if near junction. Time series features are also very important. A serious accident is more likely to happen during office hours. Weather features like pressure, temperature, humidity, and wind speed are also very important.

The computational power of my laptop was not sufficient to build a model on the entire dataset. From Exploratory Data Analysis (EDA), it is observed that the state California has more accidents compared to other states, and Los Angeles had about 45000 accidents in California. Hence machine learning models were developed to predict accident severity for Los Angeles, California. We find that XG Boost model with accuracy of 0.93 is best to predict severity of accidents.

6. Acknowledgements

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

7. References

- https://www.kaggle.com/sobhanmoosavi/us-accidents/discussion/113055
- https://www.kaggle.com/phip2014/ml-to-predict-accident-severity-pa-mont
- https://www.kaggle.com/suyash0010/severity-and-time-wasted-analysis