**Introduction**

Across the US, crashes each year result in thousands of lives lost, injured victims, and billions of dollars in property damage. The National Highway Traffic Safety Administration (NHTSA) uses data from many sources, and one of their goals is to reduce both human and property damage.

Many different factors are provided to the NHTSA, and not all factors are important in determining the severity of accidents. The goal is to identify these important factors and take a proactive approach in minimizing both the cost and the severity of accidents.

**Business**

Not all information provided to dispatch personnel is relevant to the severity of the accident. The goal is to identify and predict factors that contribute to injuries to people and the severity of the injuries. By identifying these factors, we aim to predict and minimize further serious accidents as well as giving emergency personnel the ability to prepare for the severity of the accident based on the information provided by dispatch personnel using our model.

**The Data**

* Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “[A Countrywide Traffic Accident Dataset](https://arxiv.org/abs/1906.05409).”, 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."](https://arxiv.org/abs/1909.09638) In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

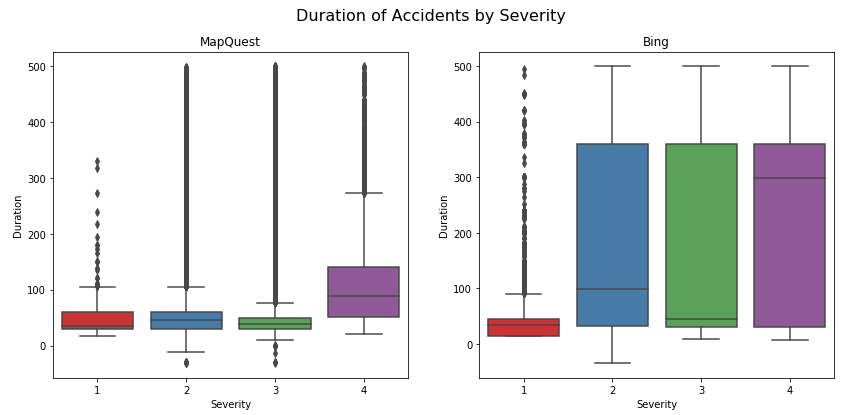
The data is for the contiguous 49 US States and is continuously being collected starting from February 2016 to June 2020, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.

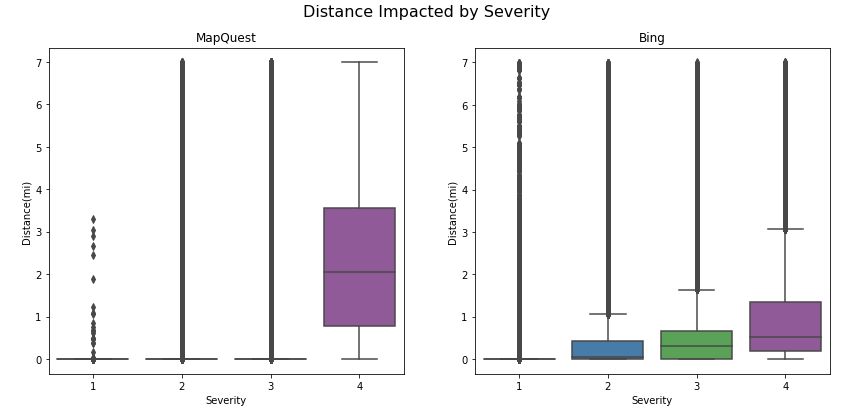
The dataset consists of 3.5 million accident data points organized into 49 columns. It has features like reporting sources of the accident, state, county, source, severity, sunrise/sunset times, weather conditions, etc.

I will be specifically looking at NY/NJ after cleaning the data since I drive mostly in these two states. The goal is to accurately identify certain factors that contribute to accident severity and to what severity degree.

**Methodology/Results/Discussion**

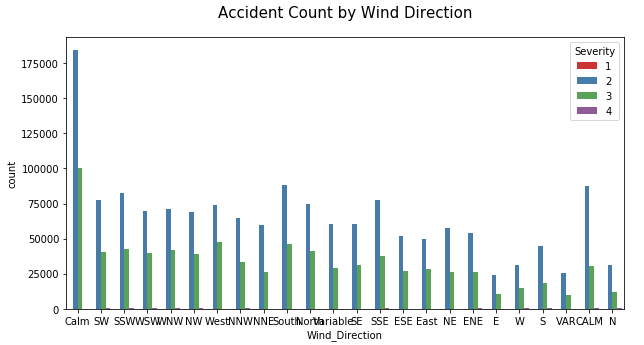
By first looking at how each source of Data classified the duration of accidents, MapQuest did a better job of having meaningful differences between each severity level. Distance showed no variability between severities 1-3 in MapQuest, and while data is more balanced in Bing, there is no clear cutoffs between distance and severity. The source of data analyzing was decided on MapQuest over Bing due to the volume of data and the superior duration of accidents in relation to severity.

There were initially 3513617 entries of data, with cleaning and focusing on NJ/NY, there were 152352 entries.

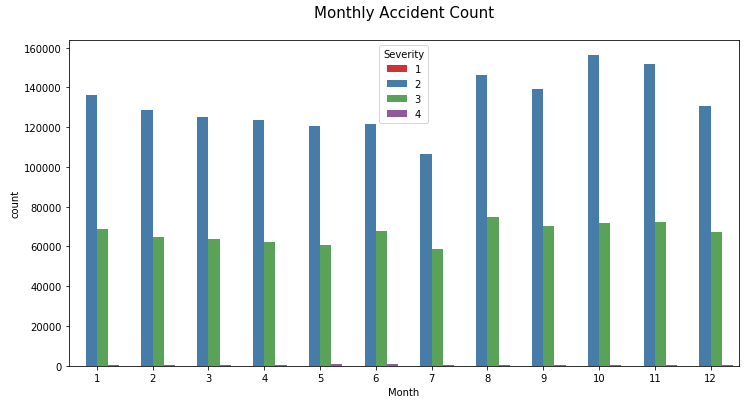


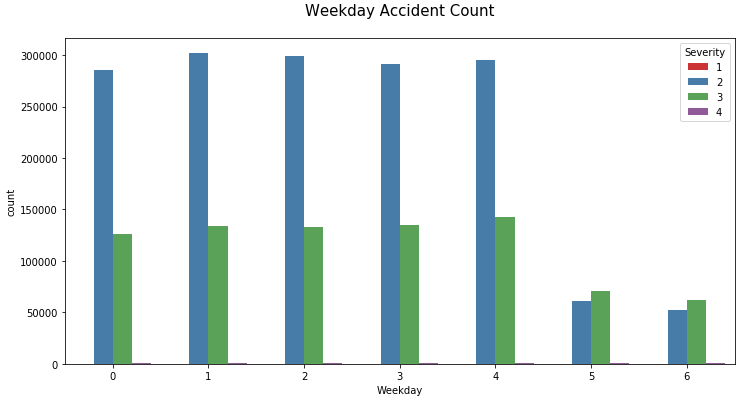
Certain categories were dropped, partly due to having no impact on accident severity (accident distance) since the accident already happened at that point. Data was cleaned by analyzing the percent missing of data according to some categories. Other categories, such as TMC, ID, Weather timestamp were removed due to having no input on severity. Certain categories were dropped (precipitation and pressure) since 75% of accidents have 0 precipitation with a pressure of 29.8 inches.

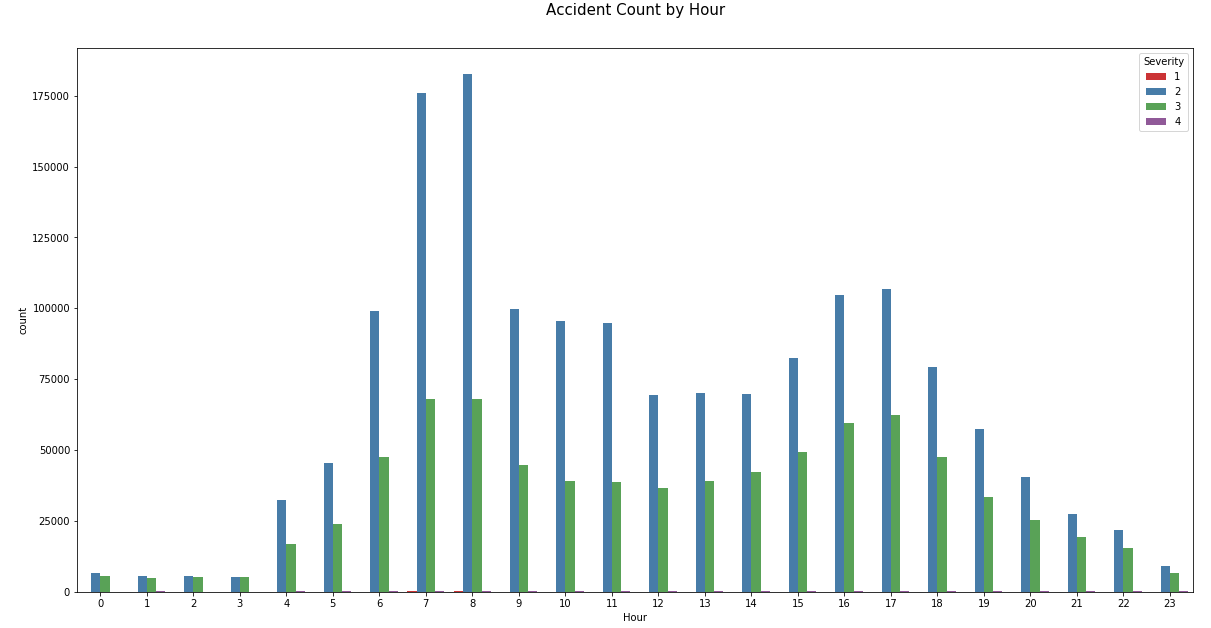
Wind direction had no clear impact on severities, as well as having 20+ different classifications of wind direction.



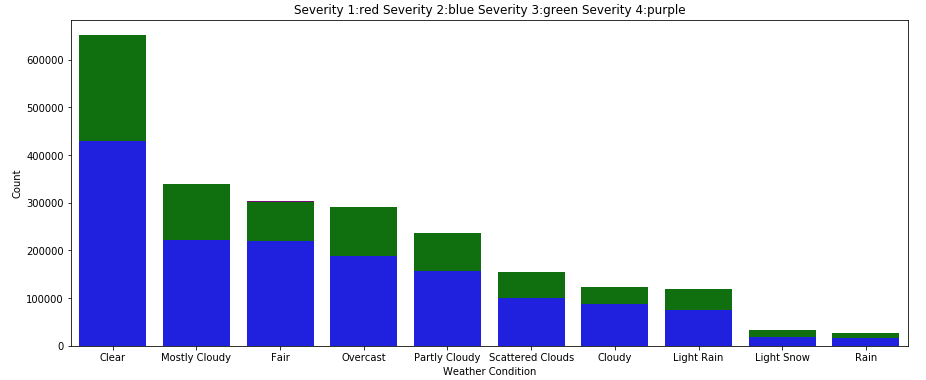
Accident severity in compared to the month showed no differences between months and the level of severities. When analyzing severity compared to weekday it became clear that many accidents happen on weekdays yet more severe accidents happen on the weekend. Accidents based on time of day show a correlation to accidents happening around times people would be commuting to work but higher ratios of severity 3 to severity 2 happened at times people would be commuting from work.



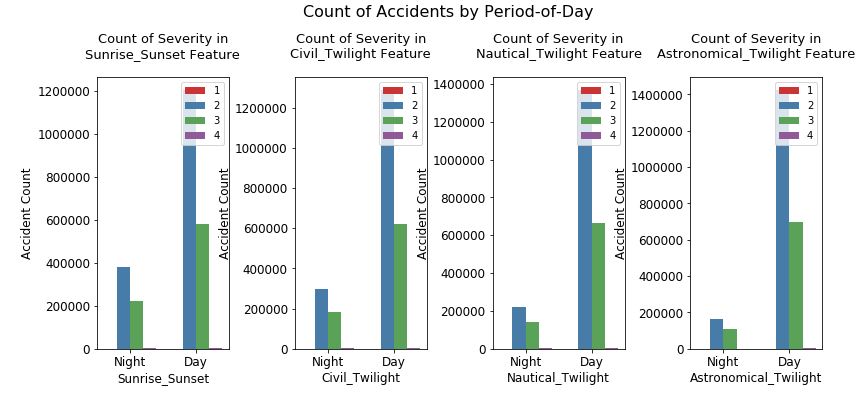


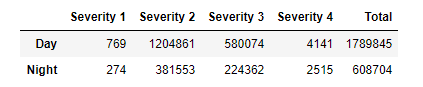


Weather conditions data was cleaned by any missing data attributed to “clear” type of day since the likelihood of weather condition as clear is high if data is missing in that category. Most accidents happened on a clear day, followed by mostly cloudy, fair, and overcast.

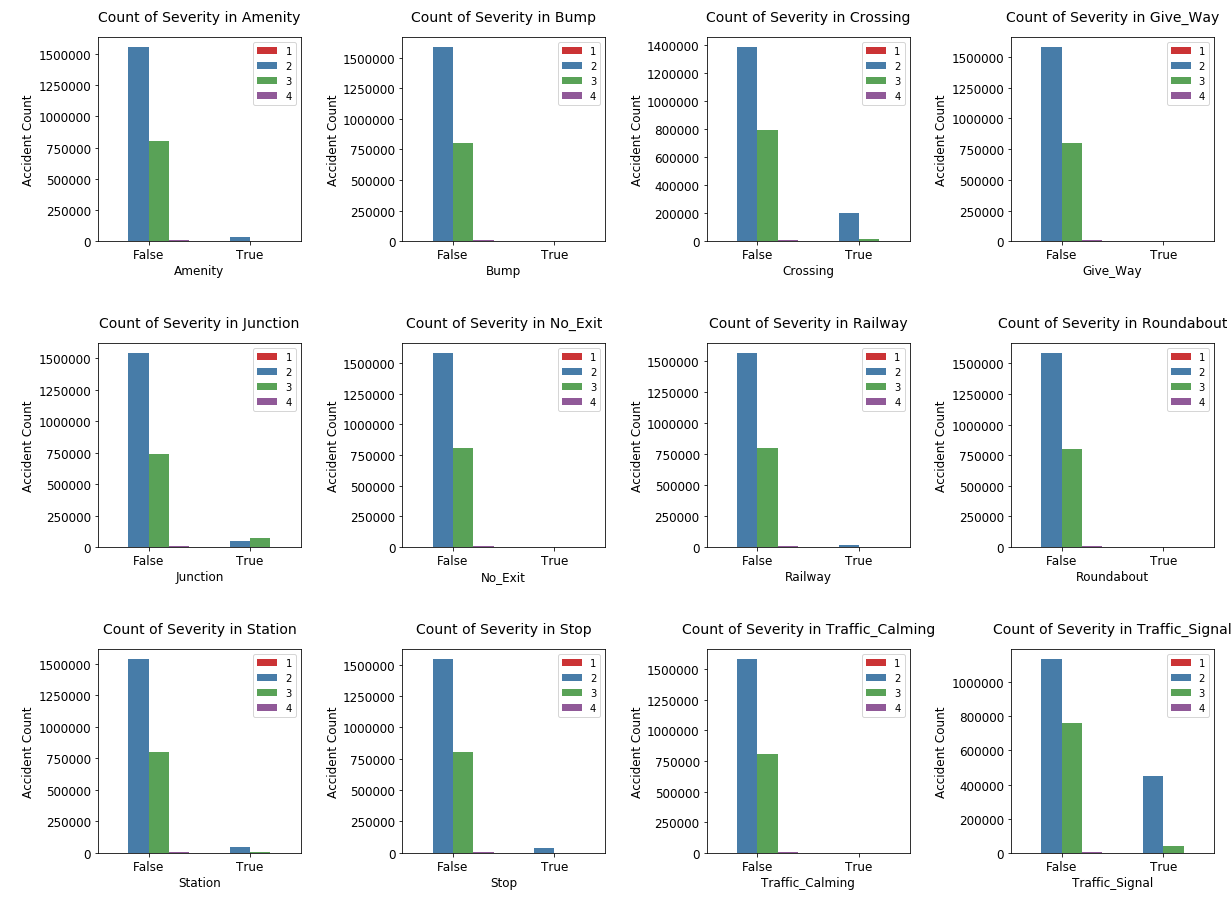


Accident severity in comparison to period of day (day/night) was analyzed. Most accidents happen during the day, which is supported previously when looking at accidents by hour earlier. No meaningful difference between severity levels and period of day when compared to each’s total. Severity 2 Day being 67% of total day accidents in comparison to 63% of Severity 2 night of total night accidents.

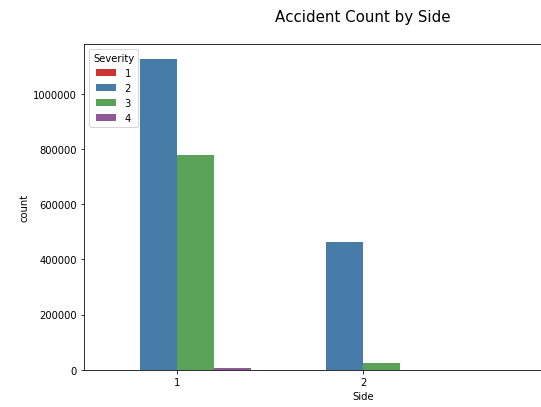




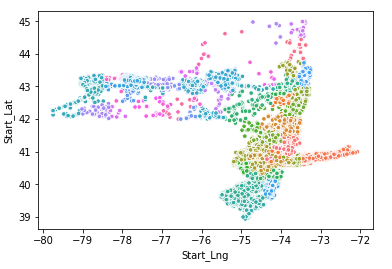
POI features were analyzed. Accidents happen mostly at traffic signals, followed by junctions and railway crossings. At junctions, accidents are prone to a higher severity compared to other POI locations. People are more likely to slow down at traffic signals and crossings while junctions and severity are highly related to speed. Most other POI features are unbalanced and dropped from further analyzation.



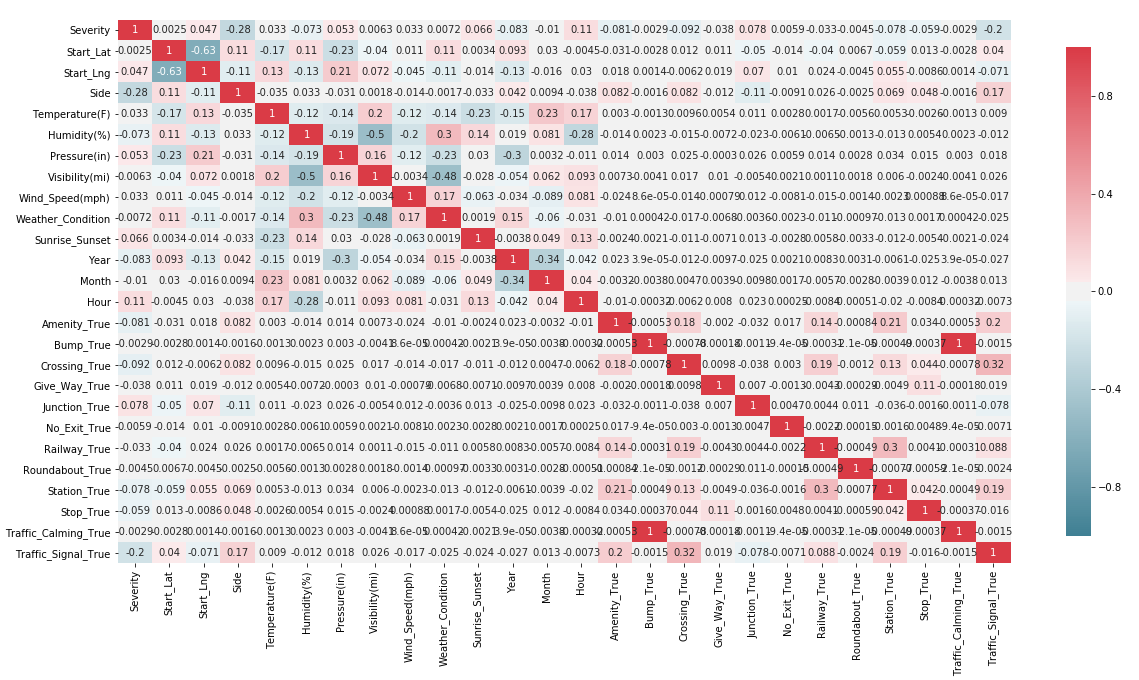
Accidents and severity were looked at based on the side reported. Side 1 being right and 2 being left. Right sided accidents were likely to be more dangerous.



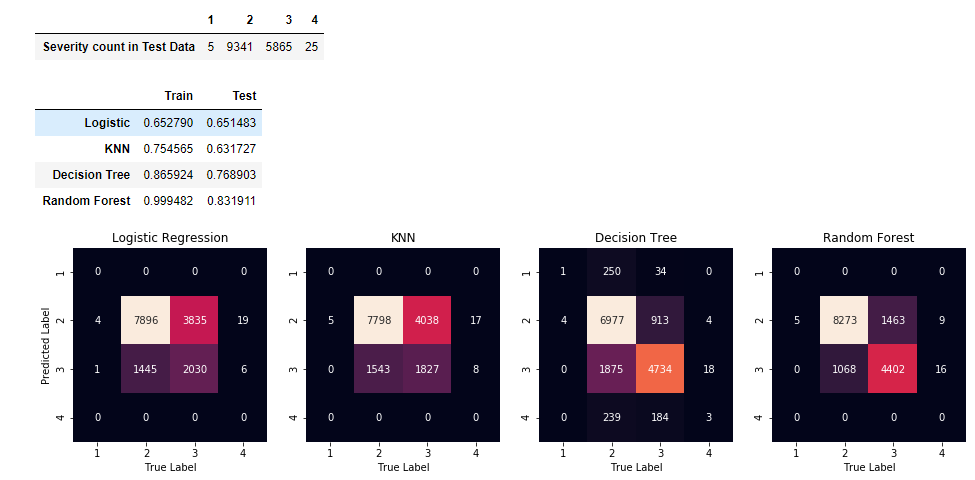
Data was narrowed down to New Jersey (NJ) and New York (NY) since I currently reside there and have a greater vested interest in analyzing these two states. A scatterplot showed a higher concentration of accidents in the northern NJ and NYC area.



A heatmap was produced on the NJ/NY states. No meaningful features related to accident severity upon first inspection, so all features listed were then used for machine learning.



Different machine learnings were used, results below.

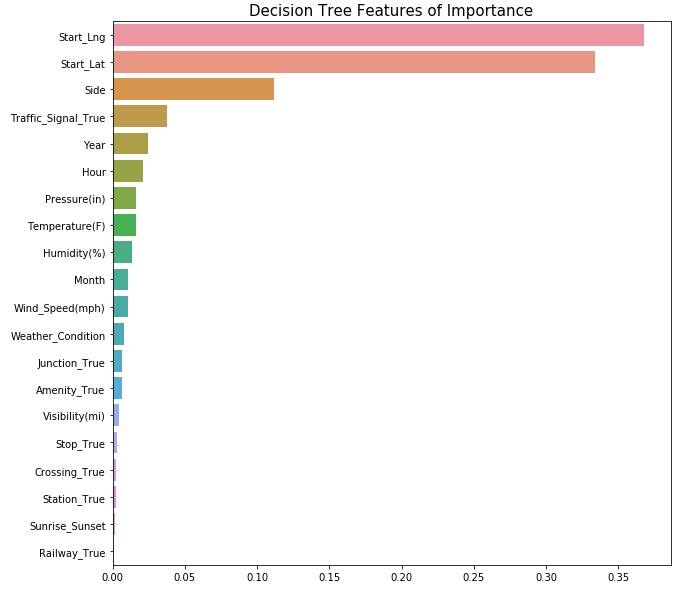


As one can see, Random Forest proved to have the highest correct labeling of severity 2. Further analysis was performed by removing outliers to assess if accuracy would improve.



As you can see, by removing outliers, the accuracy of test data improved from 83% to 85% in Random Forest and improved from 76% to 88%. Initially Logistic Regression and KNN were used as machine learnings, yet yielded poor accuracy results, which is why Random Forest and Decision Tree machine learnings were used.

Decision Tree importance features plot shows a that location is far more important than other features. Other features such as side, traffic signal, hour of day, and temperature were also important.



**Conclusion**

A very large dataset of 1+ GB consists of 3.5+ million accidents was used in this project. To analyze the accidents, the dataset is cleaned systematically and Exploratory Data Analysis (EDA) was performed using several iterations (see notebook for exact code).

The computational power of my laptop is not sufficient to build a model on the entire dataset. From EDA it is observed that New Jersey and New York make up almost half of the accidents compared to other states.

Initially, basic machine learning models such as Logistic Regression and KNN were used to build the predictive model. The models yielded poor accuracy, which is why Decision Tree Classifier, Random Forest Classifier were used. Random forest algorithm predicted severity with a training and test accuracy of 0.999482 and 0.831911, respectively.

Outliers were removed to see if it would improve the accuracy, with a train and test accuracy of 0.999415 and 0.855865 were achieved, respectively.

Based on the Decision Tree model, the most important features of accident severity were accident location, followed by \side of vehicle, traffic signals and hour of day.