Capstone project Analyzing Accidents Severity

Patricia A. White

Problem

- Traffic accidents and fatalities are inevitable. Many factors could influence the severity of accidents and fatalities.
- Factors includes weather conditions, driver age and experience, intoxication, speeding, infrastructure, traffic signals, road congestion, reckless driving, locations, and many others.
- It is imperative to mitigate accident impact and improve traffic safety efficiently. Accurate predictions of the severity of accidents, locations, and other causes can provide crucial information for first responders to evaluate the severity of these accidents and implement efficient treatment. The Department of Transportation, World Health Organization, National Safety Council, MapQuest, AAA Foundation for Traffic Safety, The Bureau of Transportation Statistics, and others are organizations that are interested in major road safety. Research is being performed on a regular basis with the intent to improve traffic safety.
- It is the intent to analyze data and use various models and graphs. The models are employing the use of supervise machine learning algorithms such as decision trees, Linear regression, SVM, and graphs using accident data. This presentation intends to discuss the impact of factors that affect the severity of accidents

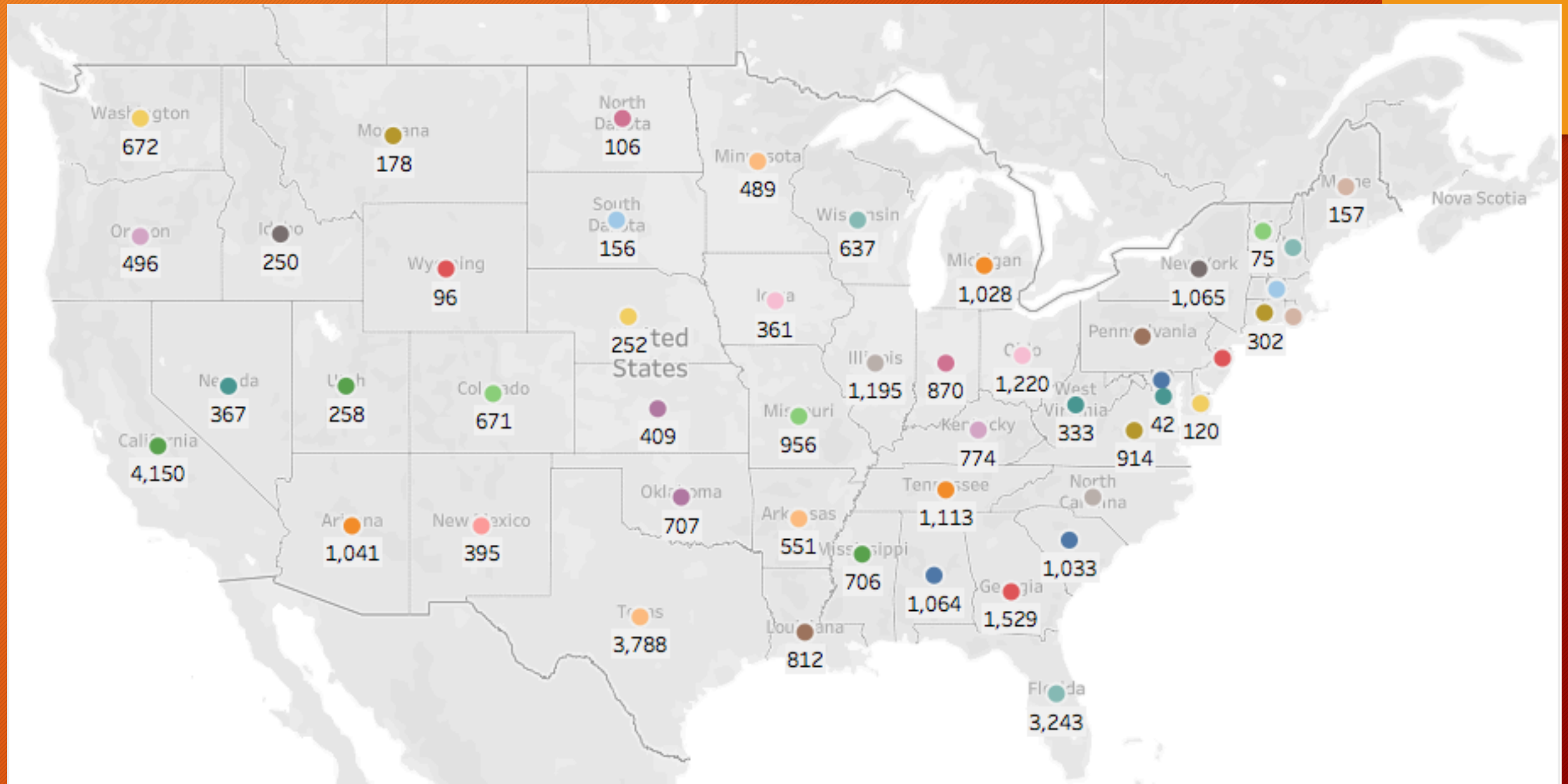
Traffic Accident Statistics

Traffic accidents is always a major concern for public health and traffic safety. Within the United States, there are approximately 228 million licensed drivers and these drivers all risk their lives every time they are behind the wheel of a vehicle. There are 277 million registered vehicles in the US.

Based on statistics from years 2007 to 2016, statistics resulted in more than 5.8 million auto crashes that occur yearly. Approximately 21% of those, a little over 1.2 million, involves hazardous weather. These vehicle crashes have killed an average of 5,376 people annually; equating to about 16% of all vehicular deaths. "More than 418,000 others were injured each year during that same period". (NSC,2020)

	2018 motor-vehicle crash highlights
Deaths	39,404
Medically consulted injuries	4.5 million
Cost	$445.6 billion
Motor-vehicle mileage	3.240 billion
Registered vehicles in the United States	277 million
Licensed drivers in the United States	228 million
Death rate per 100 mil-	1.22

Choropleth Map of 2018 Accidents throughout United States



Data Preprocessing

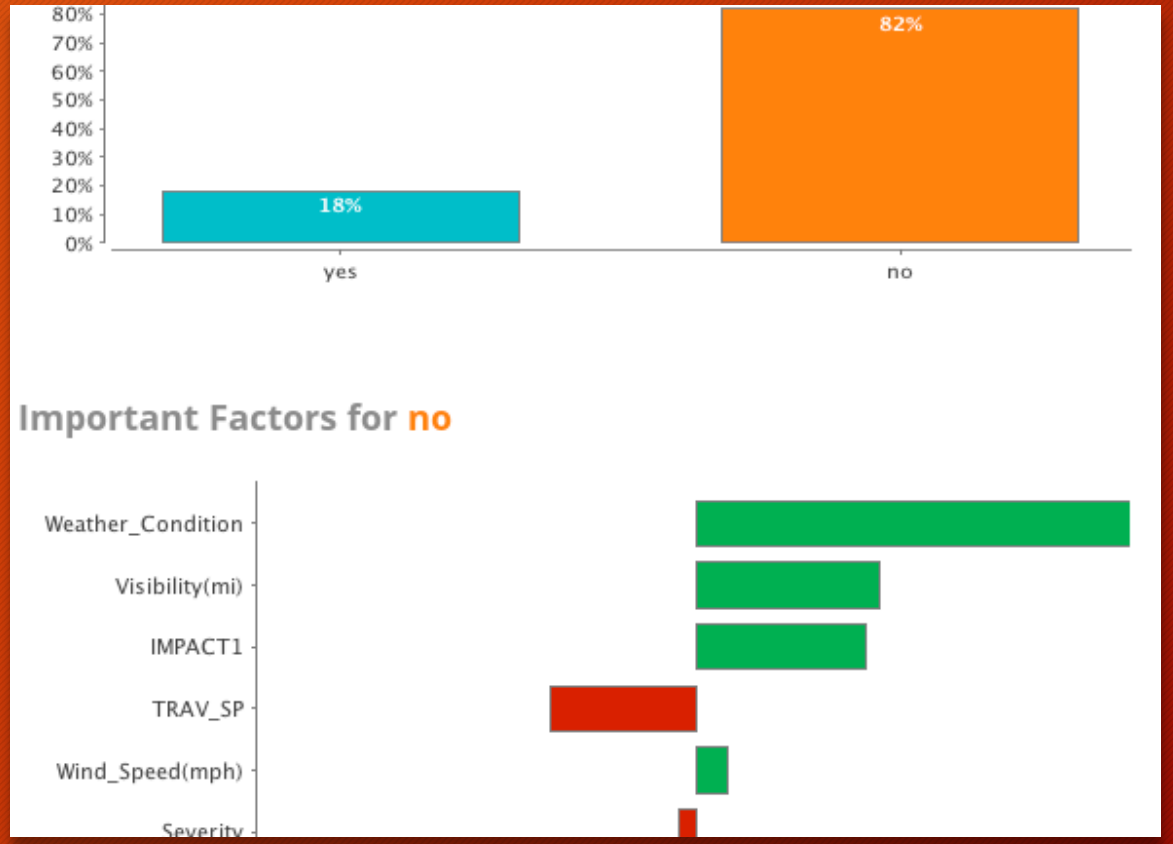
- Accident Data, Collision Data (2013 – 2018) from Coursera, Kaggle, NSC, and DOT.
- The total number of examples included in Raw Dataset 1 was 164,179 with 48 attributes
- The total number of examples included in Raw Dataset 2 was 14,786 with 29 attributes
- The total number of examples included in Raw Dataset 3 was 469 examples with 4 attributes
- Dataset 1 was extracted down to 10,000 examples and 14 attributes
- Dataset 2 was extracted down to 5030 examples and 9 attributes
- Dataset 3 was extracted down to 52 examples with 2 attributes
- Highly correlated features were dropped and not included in predictions
- A few cells were manipulated for learning purposes
- Tools used for this presentation includes Tableau, RapidMiner, and Google Colab.
- Used several machine learning models for simplicity and handling large datasets

Gradient Boosted Tree Model

- Impact of 23.12
- Severity of 2.39
- Traveling Speed at 46mph
- Visibility of 9.38
- Weather Condition being Clear
- Wind Speed of 9..61

Based on Gradient boosting machine learning for regression, it was predicted that 82% of the time, there will most likely not be a fatality from auto accident; with 18% representing yes there will be. Weather condition, visibility, impact, and wind speed supports the model. Traveling speed and severity contradicts the model.

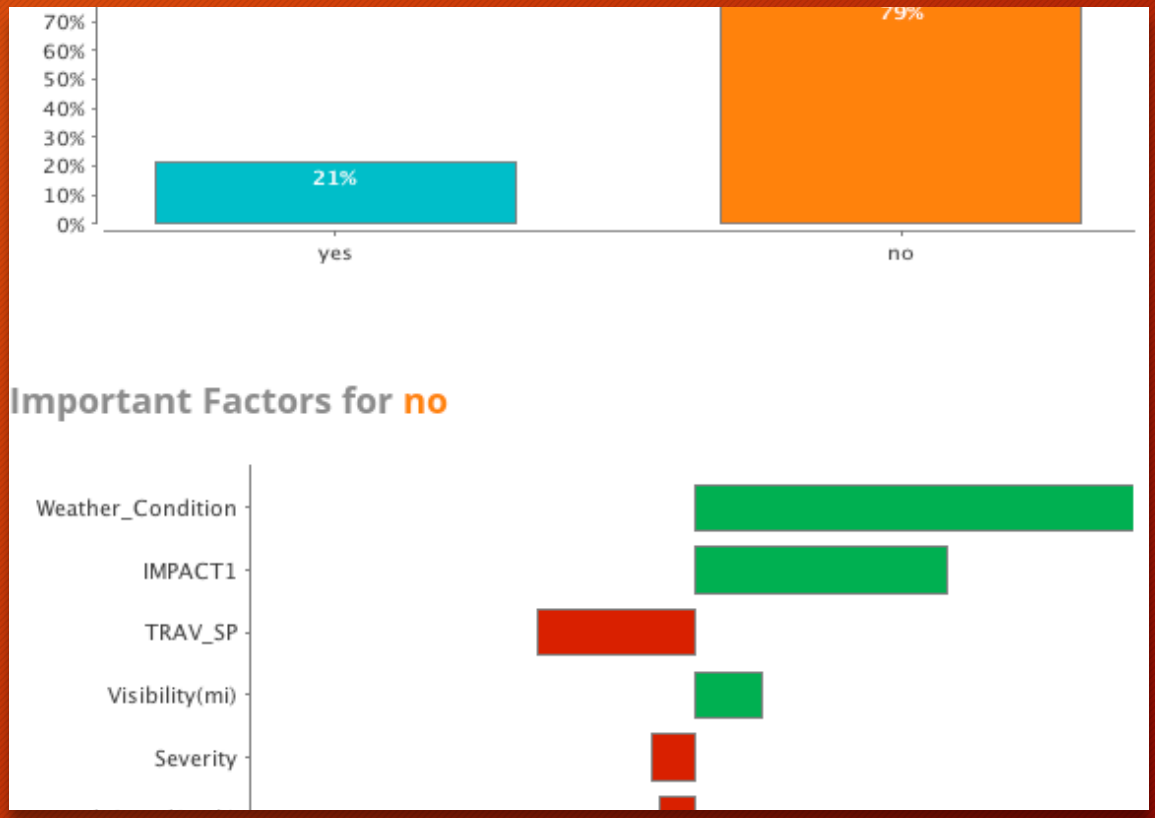
(Red contradicts, Green Supports)



Logistic Regression

- Impact of 23.2
- Severity of 2.3
- Traveling Speed of 46
- Visibility of 9.38
- Weather Condition being Clear
- Wind Speed of 9.061

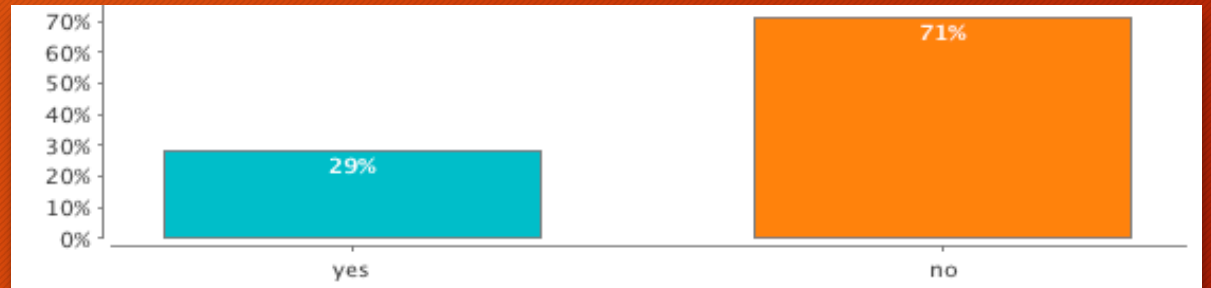
Based on the Logistic Regression Model, it predicted that 79% of the time there would not be a fatality and 21% of time there will be a fatality. Weather Condition, Impact, and Visibility supports the model. Traveling Speed, Severity, and Wind contradicted the model and had no bearing on the prediction.



Support Vector Machine Model

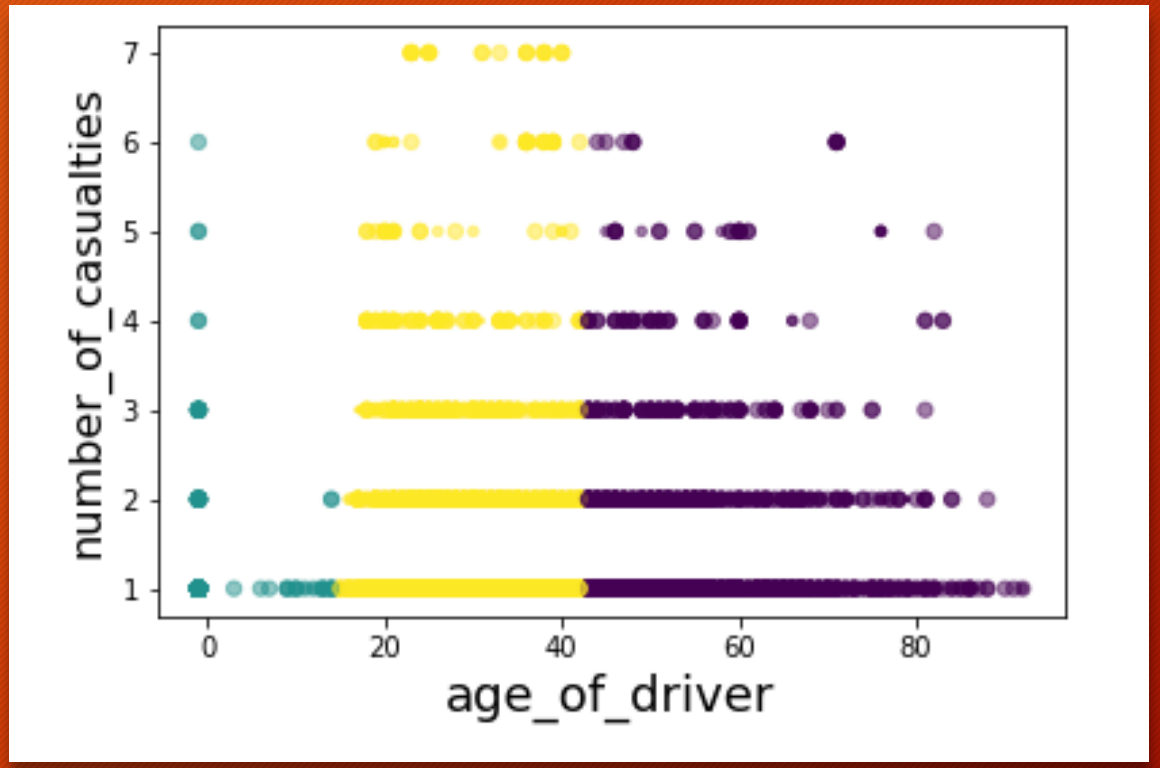
- Impact of 23.2
- Severity of 2.39
- Traveling Speed of 46
- Visibility of 9.38
- Weather Condition of 0.23
- Wind Speed of 9.061

Based on the Support Vector Machine model, it predicted that 71% of the time there will most likely not be a fatality from an accident, but Wind Speed and traveling speed supports this model; visibility, impact, severity, and weather conditions contradicts the prediction. This is a total opposite of Logistic Regression and Gradient Boosted Tree models.



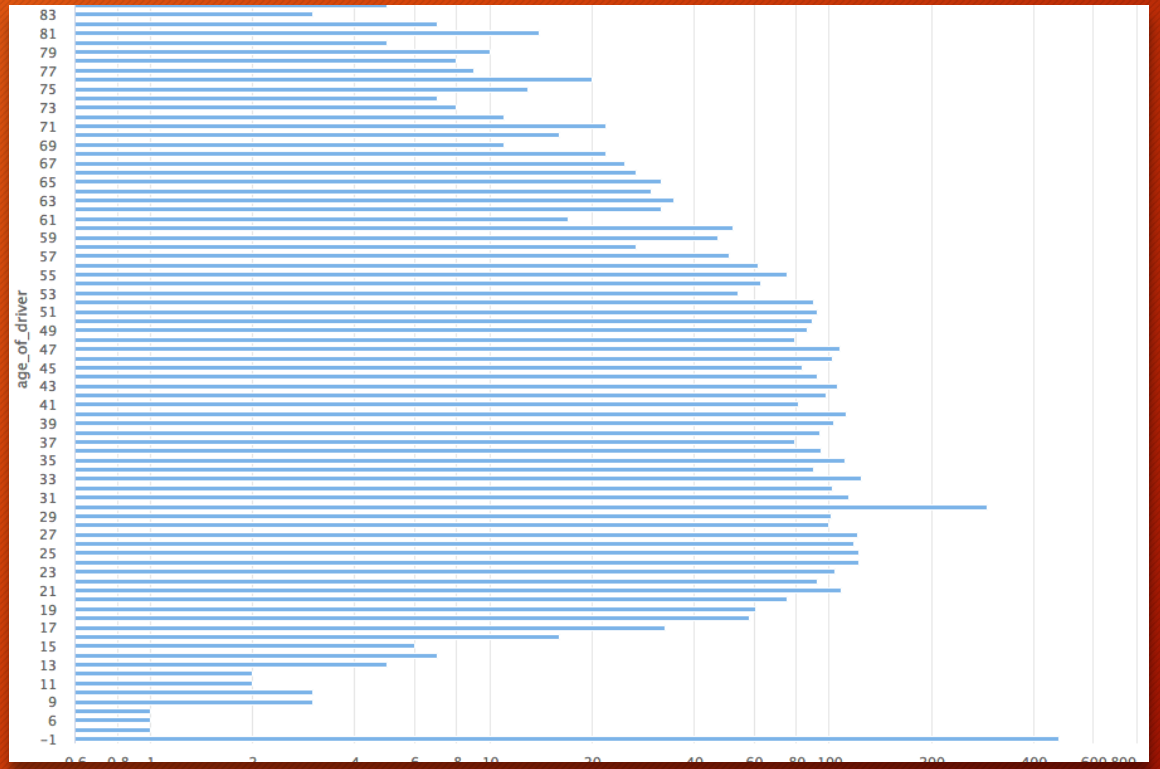
Casualties

Segmentation Chart with K means showing the number of Casualties based on the person age! The older the age, less casualties occurred.



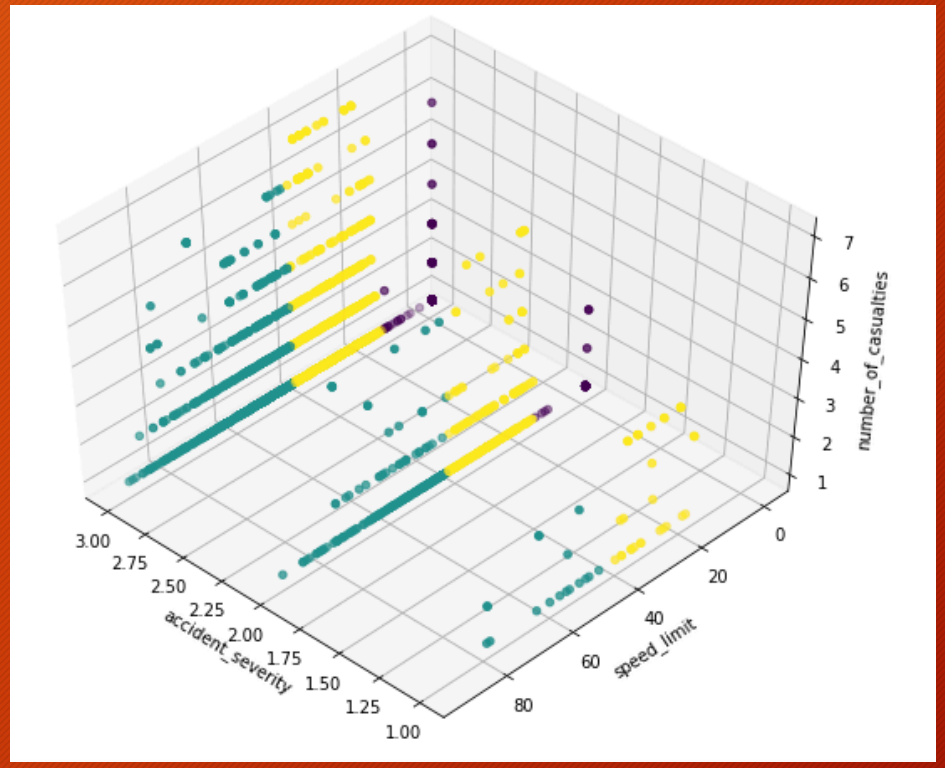
Age of Driver/Number of casualties

- Accidents can be caused and due to any number of reasons such as age
- Age of driver with the highest number of casualties is 30 years old with 349 fatalities.
- Age of driver with 2nd highest number of casualties is 33 years old.
- Least age is 3 years old with 1 casualty, but this age does not have a license.
- Oldest age is 92 with 2 casualties.
- This proves age plays a critical factor in auto accidents.



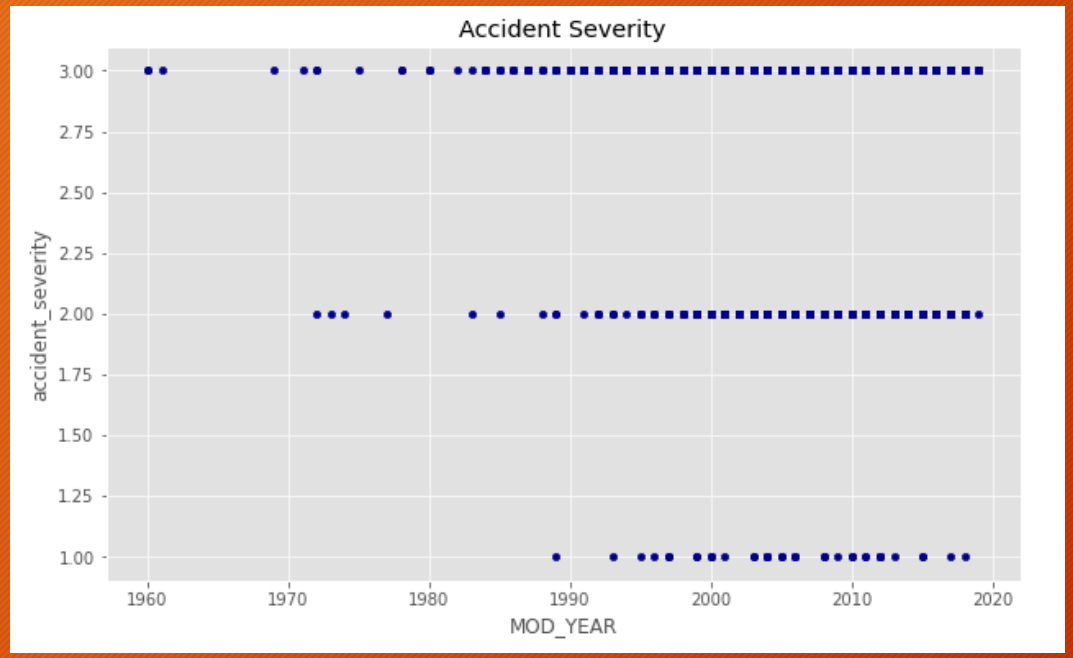
K-means Partition

- Speeding is a factor in causing severity of accidents.
- K-means partitioned data in 3 clusters.
- Clusters are created according to the severity of the accident, the speed limit, and the number of casualties.
- Based on speeding, the more severe the accident, casualties are caused.
- This proves speeding plays a critical factor in the severity of auto accidents.



Casualties based on the model year of vehicle.

It can be observed that there is a trend in the data; based on the latest year that the vehicle was made, there seems to be an increase in casualties.



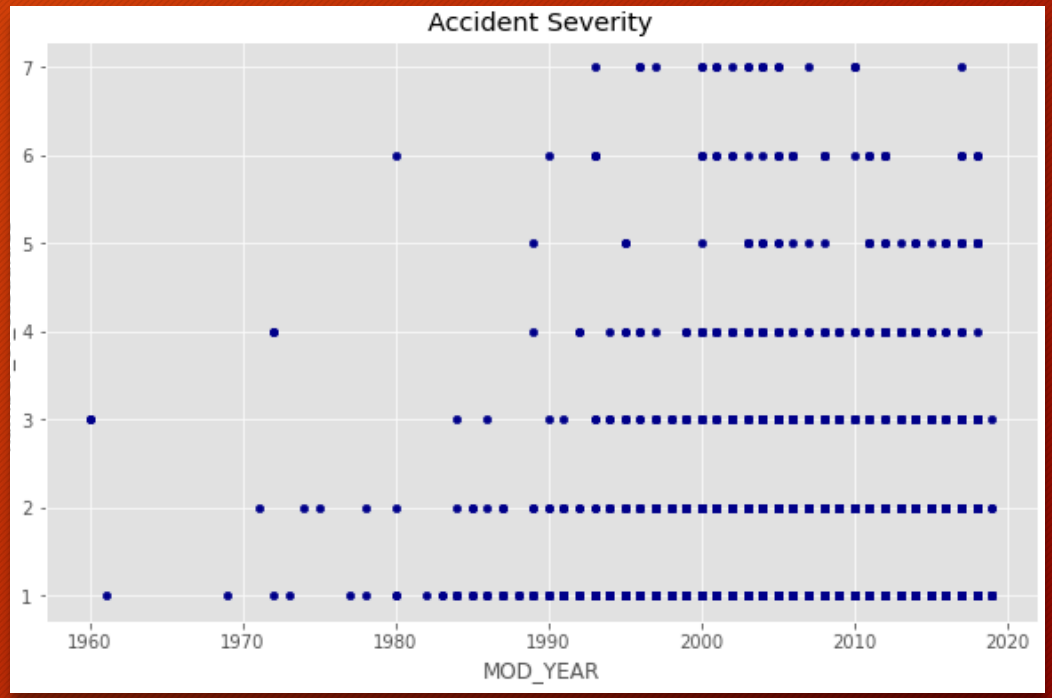
Severity of accident based on the model year of vehicle.

1 = Project Damage

2 = Injury

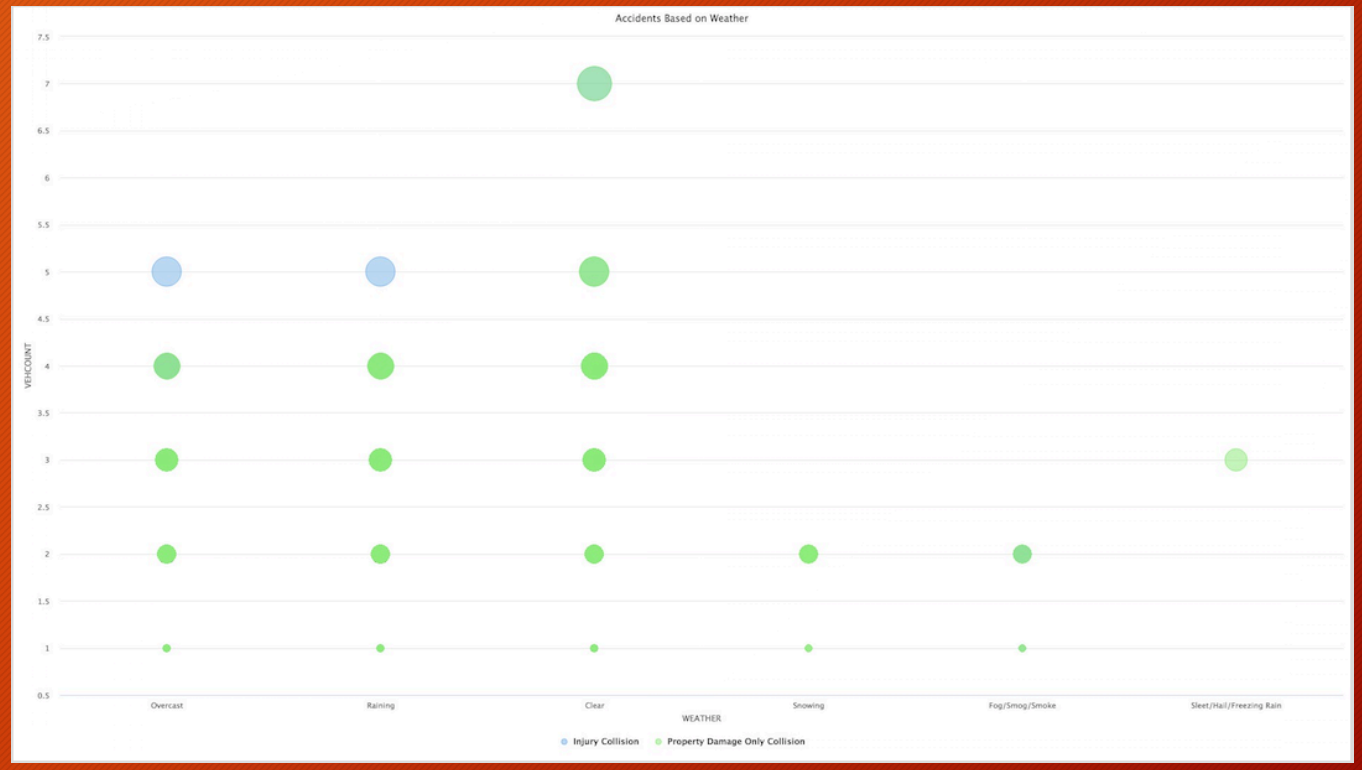
3 = Fatalities

Based on observation, there were more injuries and fatalities than property damage.



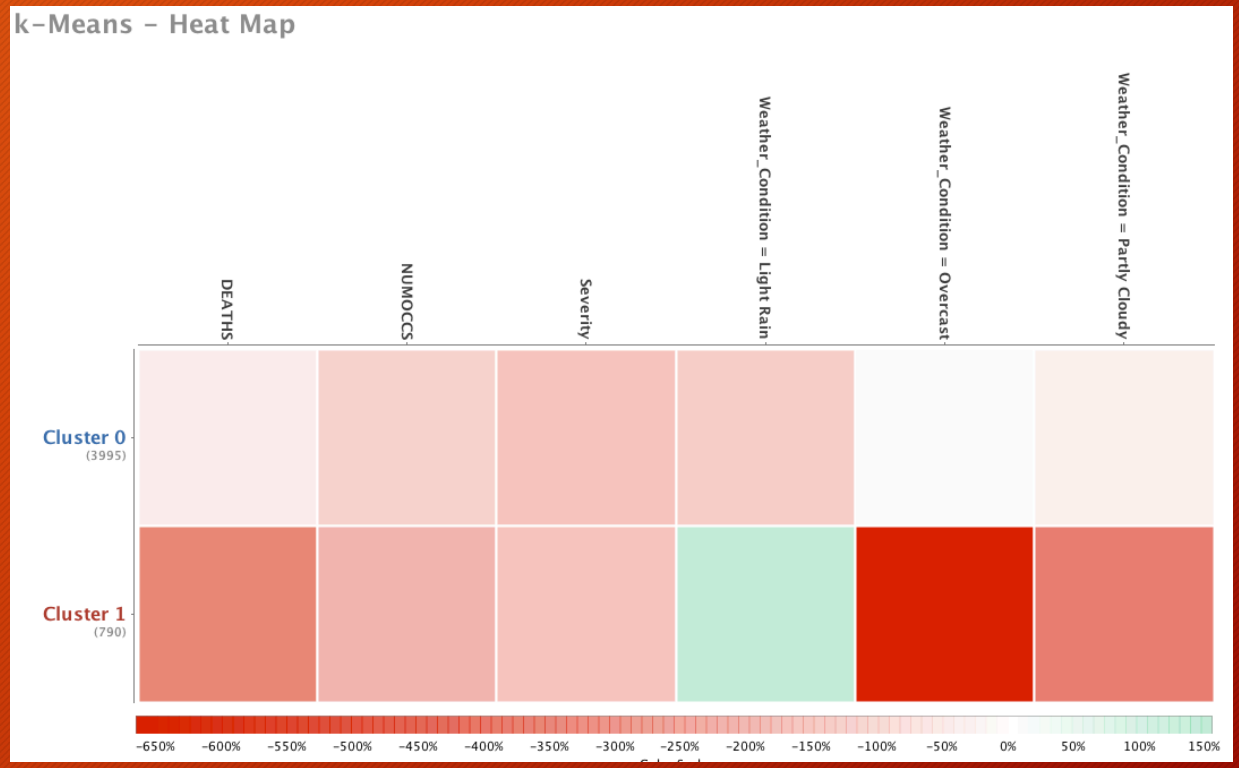
Accidents Based on Weather Conditions

- Accidents based on weather condition and vehicle count. Green represent property damage and blue represent injury collision. It is observed that when the weather is clear, there are more property damage collisions than injury collisions.



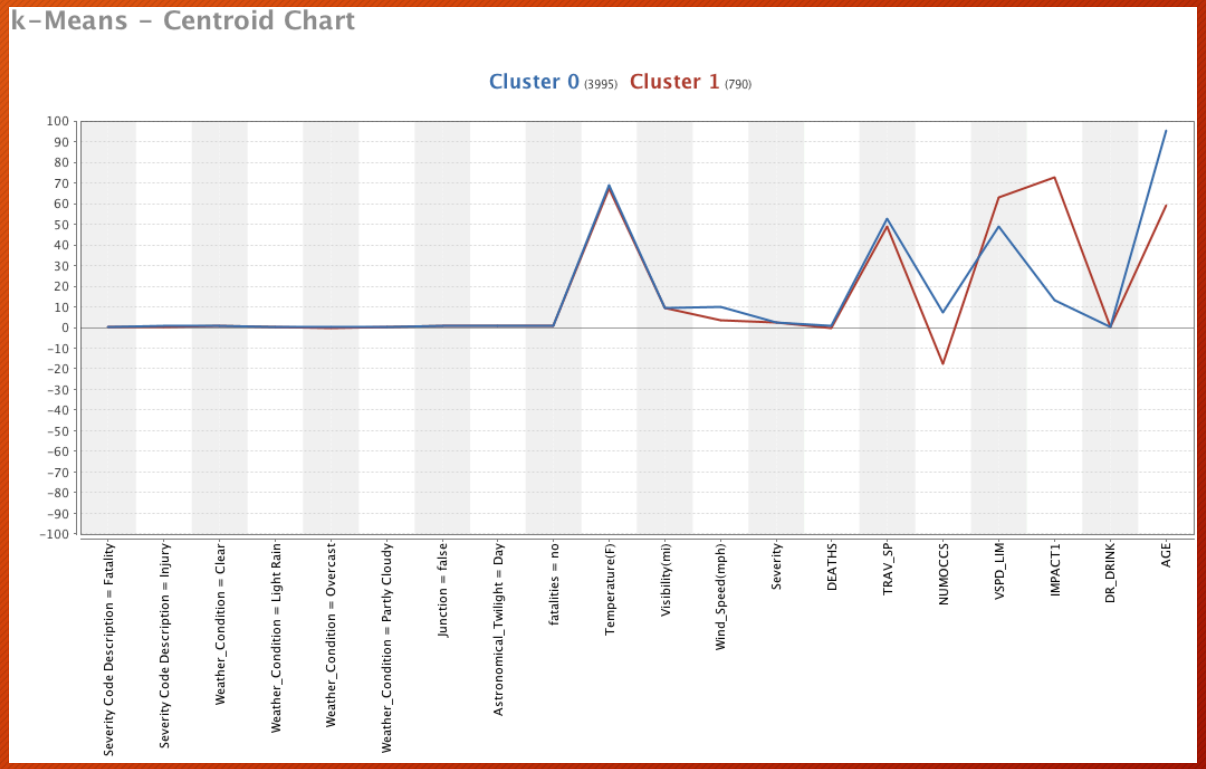
K-Means Heat Map

- Overview of clusters, there were two clusters for k-means.
- Cluster 0 is the biggest cluster with 3,995 items, with Cluster 1 having 790 items.
- The distance measure was Squared Euclidean Distance and the average cluster distance was 66287.33.
- The Heatmap has a higher value with weather conditions for Light Rain. The heatmap show similarities based on the distance between them. The clusters can be grouped together with other clustering models for further research.



K-Means Centroid Chart

This k-means centroid chart show another view of the same thing as the heat map!



Conclusions

- Created various models to predict and show the severity of auto accidents. Speeding, weather conditions, types of vehicles, played a factor in the severity of accidents.
- A small amount of the data was manipulated for learning purposes. These models have room for improvement.