Project Proposal

Title: Severity Analysis and Prediction of US Public Accidents

Authors: Mrunmayi Anchawale, Pavan Choudhari, Maanasa Kaza, Zhuocheng Lin, Yashvin Jagarlamudi

Summary:

The issue of traffic accidents is quite important as 1.25 million people meet with accidents every year. The US Accidents dataset incorporated in this project is a public dataset that is a compilation of countrywide traffic accident details spanning 49 states from 2016 to 2019. There are around 3 million records in this dataset with information regarding the location of the accident, weather conditions present, unique description of the accident using a TMC (Traffic Message code), the severity of the incident, and significant structures nearby. This traffic data is obtained through reliable sources like law enforcement websites and state-mandated motor vehicle agencies.

This project includes data analysis, visualization and modeling using R. This project aims to answer many pertinent questions regarding traffic accidents such as the states with the highest number of accidents, frequency of accidents compared between the city roads and the highways. Consequently, severity predictions can be made from the gained insights.

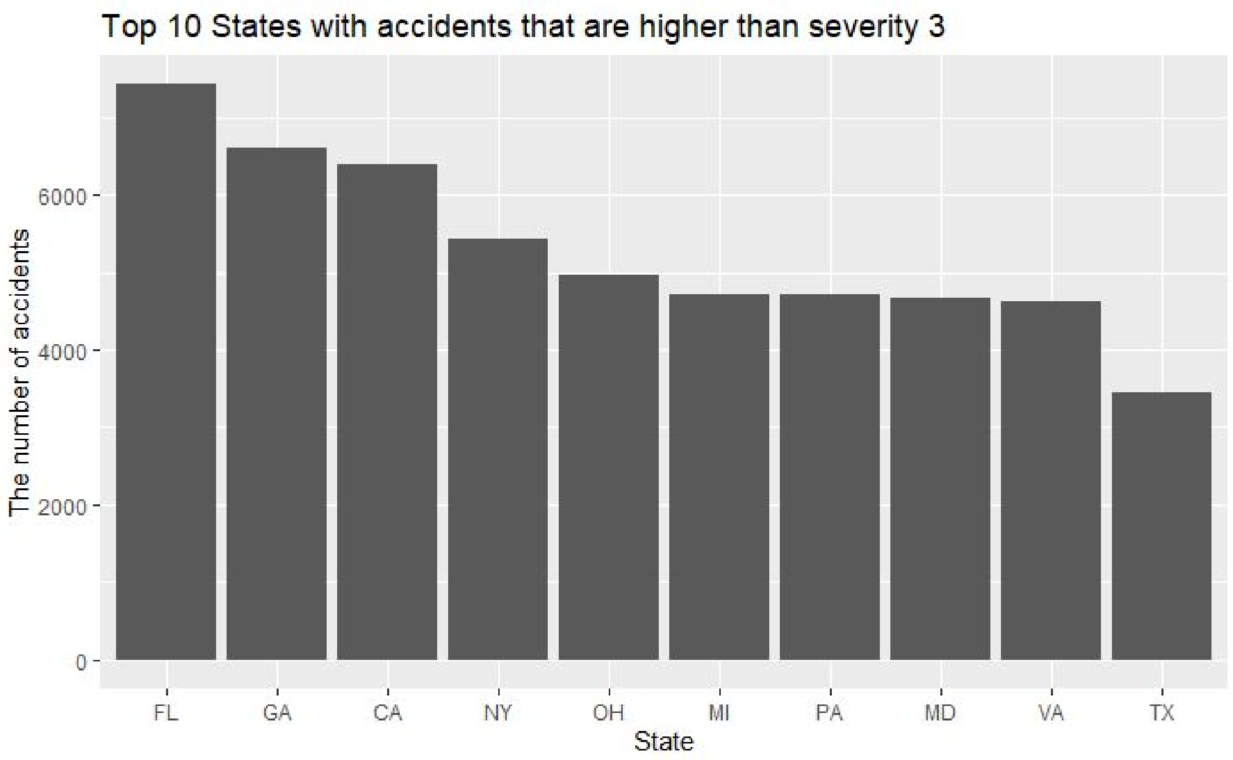
Proposed Plan of Research:

Initially, preliminary analysis of the traffic data will be done through EDA and the results will be demonstrated in the form of statistical tables, plots, or maps. The foremost objective will be to delve into the data and to find out which factors contribute to the severity of an accident. Here, severity is denoted by an index ranging from 1-5 where a 1-rated accident is inconsequential and a 5-rated accident is drastic. Conventionally, weather conditions are a prominent factor contributing to the severity of an accident. However, by using a covariance matrix and other statistical analyses, the actual factors determining the seriousness of an accident will be established.

In addition, mainly an ordinal logistic regression model will be utilized to predict the degree of severity based on factors like weather, time of the day, structures in the vicinity, location, etc. Lastly, a dashboard will be created with RShiny to succinctly display the analysis and the results of the model.

Preliminary results:

We have successfully loaded the US accident data set into R. The visualization below is created by filtering the data set with severity greater than 3 and then grouping them by state. The bar chart represents the top 10 states that have the highest number of accidents. Florida leads at around 7000 accidents that happened between 2016 and 2019, and Texas stands at the 10th position with approximately 3600 accidents.



Bibliography:

Moosavi, Sobhan. “US Accidents (3.0 Million Records).” *Kaggle*, 17 Jan. 2020, www.kaggle.com/sobhanmoosavi/us-accidents.

`