WAYNE STATE UNIVERSITY

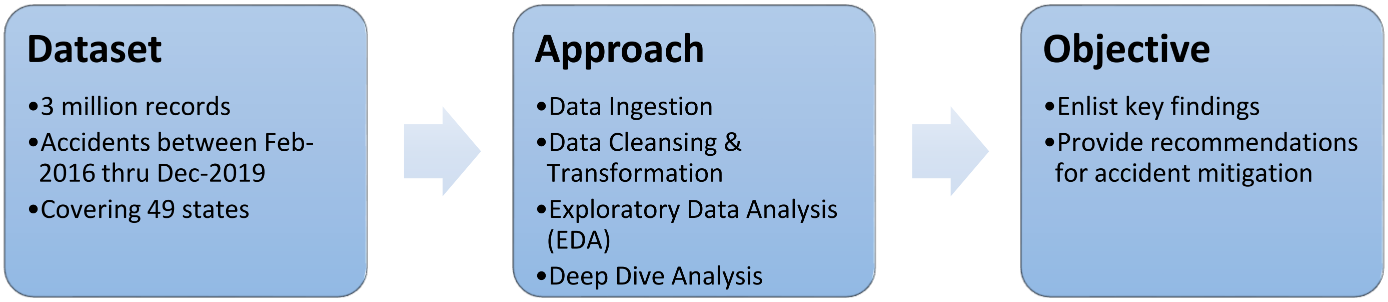
DSA 6000 Final Project Report (Group: Project 3): Insights and Inference from US Accident Dataset

Team - *Gargeyi Baipa, Michael Lupo, Naveen Donthineni, Sreejit Sivasankaran, Vidhi Shah*

Introduction

Road accidents are one of the main causes of several injuries and fatalities. They also have an economic impact on countries, considering the importance of road accidents, 'US Accident Dataset' [1] sourced from Kaggle is considered for analysis in this work. The dataset encapsulates traffic accidents across USA over a 4-year period (2016-2019). 'Vehicle population' [2] dataset with the number of vehicles registered in a state is also used. Inference from the data set will be uncovered in this report.

The following diagram provides an overview of key elements of selected dataset, approach towards data analysis and objectives to be realized from the project.



Dataset
- 3 million records
- Accidents between Feb-2016 thru Dec-2019
- Covering 49 states

Approach
- Data Ingestion
- Data Cleansing & Transformation
- Exploratory Data Analysis (EDA)
- Deep Dive Analysis

Objective
- Enlist key findings
- Provide recommendations for accident mitigation

Data Ingestion, Cleansing & Transformation

The raw data with 3 million plus records was loaded into RStudio. A 2-pronged approach was taken towards data cleansing and transformation, i.e. using R (R file referred to as [b]) and Tableau (Tableau file referred to as [a]):

Column(s)	Transformation
Temperature, Pressure, Wind Chill	Replaced NAs with mean value
Number	Replaced NAs with zero
End Latitude, End Longitude	Dropped
City, Zip code, Time zone	Replaced spaces with max value, by State
Hour, Month, Weekday	New columns created for EDA

Following the preliminary cleansing and transformation using R, data was exported out and fed into Tableau for further alterations necessary for generating visualization as part of Exploratory and Deep Dive data analysis.

Column(s)	Transformation
Infrastructure Variables (e.g. Crossing, Junction, Traffic Signal, etc.)	Created new measures (calculated fields)
Zip Code	Created new Dimension
Vhpop (from Vehicle Population table)	Dataset is joined with 'US Accident Dataset'

Exploratory Data Analysis (EDA):

An in-depth Exploratory Data Analysis (EDA) using two approaches with focus on the list of topics as given below was performed:

1. Analysis by Visualization:
   - Distribution across states by number of accidents and rate of accidents
   - Distribution of accidents by severity and weather condition(s) over the course of a day
   - Distribution of accidents by infrastructure variables

2. Analysis by ML Technique [b]
    - Identify key factors impacting severity of accidents using supervised machine learning algorithms (Linear Regression and Lasso)

Analysis by Data Visualization:

The goal of this section is to visualize **where** and **when** most accidents occurred, **what** features are present in accidents with higher severity and their contribution to the accident count.

Analysis on the pure volume of accidents over the entire time period across the US is shown in Figure 1 on left, and the accident rate which is (total number of accidents/ total number of vehicles registered in 4 years) is shown in Figure 1 on right. While the left chart shows California as the top state with the most accidents followed by Texas, this is an intuitive insight. These states are heavily populated with a great deal of traffic, so more accidents will naturally happen here. The chart on the right indicated that South Carolina has the most accidents per vehicle registration in the state – a much more telling figure.

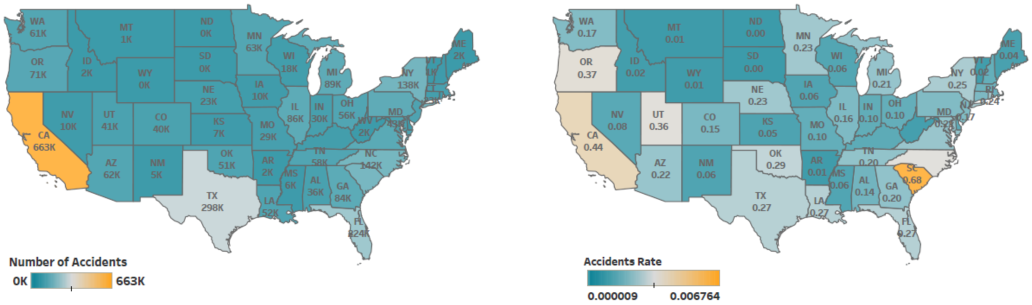


Figure 1: (Left) (Right) Total number of accidents by State (over 4 years); (Right): Number of accidents per number of vehicle registrations (over 4 years)

It was found that over the course of the day, accidents occur at various rates. As shown in Figure 2 below, majority of accidents occur during high traffic times of day – around 8am and 5pm, coinciding with daily rush hour. Also, the severity of most of the accidents were at level 2. Considering that over four years the number of clear days is higher than the non-clear days, accidents occurred on non-clear days at a higher rate than clear days.

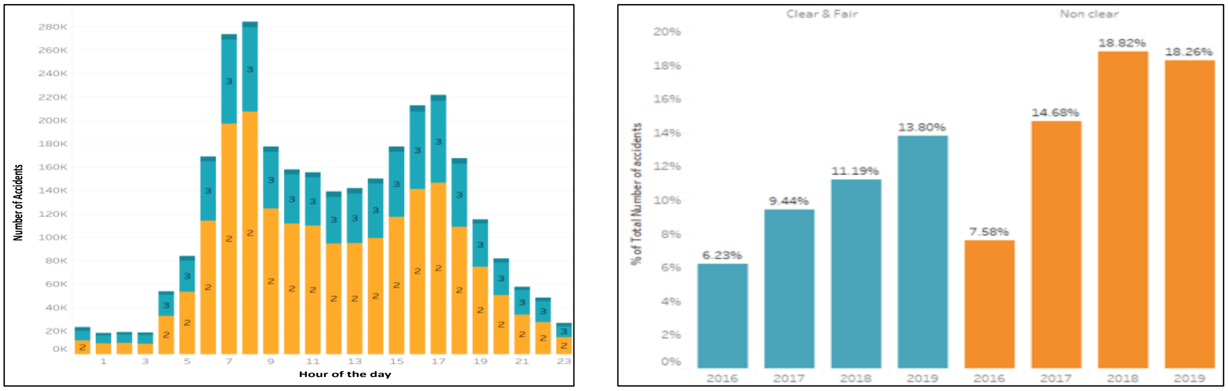


Figure 2: (Left) Accidents by severity and hour of day; (Right): Accidents by weather condition over time

It is known that accidents can be impacted by the surrounding infrastructure – this was examined in two views 1) the distribution of accidents by surrounding infrastructure 2) and the same view split out by severity. As seen in Figure 3 below, most accidents occur at Traffic (Traffic lights) in the US – almost 2x as many as the next closest variable. However, most of these accidents are level 2 in severity. The chart on the right shows how Junctions are the leading infrastructure variable related to level 3 and 4 accidents.

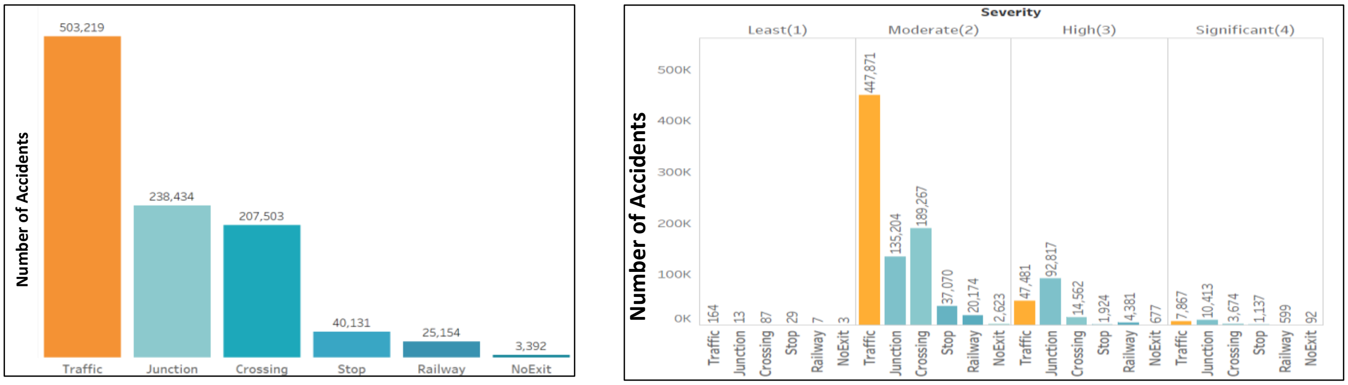


Figure 3: (Left) Number of Accidents by infrastructure type; (Right) Number of Accidents by infrastructure type and severity. Levels of severity 2-4 represent impact on traffic with 4 being the highest

The next interesting feature is the 'Description,' which is a natural language description of the accident. Text analytics [b] on two leading states in accidents (California for volume, South Carolina for rate) was performed using NLP (natural language processing) in R, which is depicted below in Figure 4. The analysis revealed the most frequently occurring words in each state. For South Carolina, **exit** was the most frequently occurring word, whereas words like "blocked" and "accident" occurred most frequently in California. One interesting insight is that accidents where vehicles travelled northbound or southbound occurred more frequently than westbound or eastbound in California.

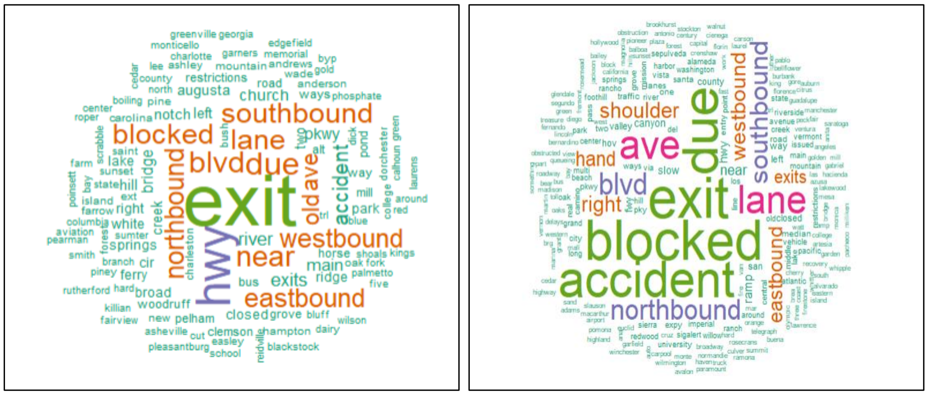


Figure 4: (Left) South Carolina word cloud; (Right) California word cloud

Analysis by Machine Learning Technique:

Two supervised machine learning techniques [b] were used in order to determine key factors impacting Severity of accidents.

Using Linear Regression and Lasso algorithms, it was found that the following predictors (esp. Weather and Infrastructure) had low p values, representing their influence to accident severity - Latitude, Side, Temperature, Humidity, Pressure, Visibility, Precipitation, Amenity, Crossing, Junction, Railway, Roundabout, Station, Stop, Traffic Signal and time of the day.

Based on the model, the following predictors were dropped from further analysis based on their high p values: Windchill, Visibility, Wind Speed, Bump, Give Way, and No Exit.

Deep Dive Analysis

Based on insights learned from EDA, the state of South Carolina was chosen for Deep Dive Analysis since it has the highest rate of accidents. The goal of this analysis is to identify hot spots across the state and explore impact of key features (e.g. weather, infrastructure, etc.) leading to an accident

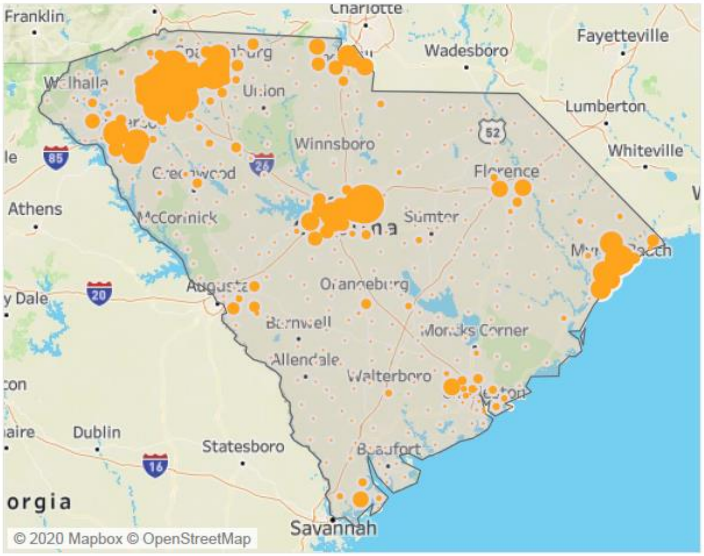


Figure 5: Accident distribution in the state of SC

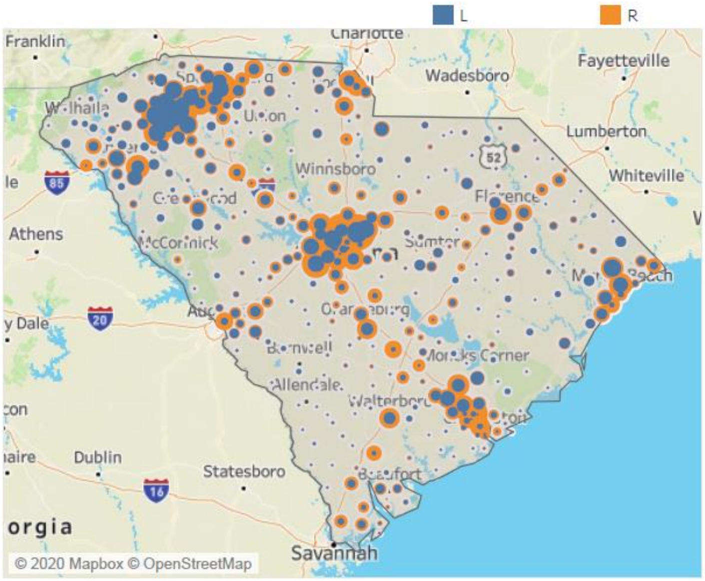
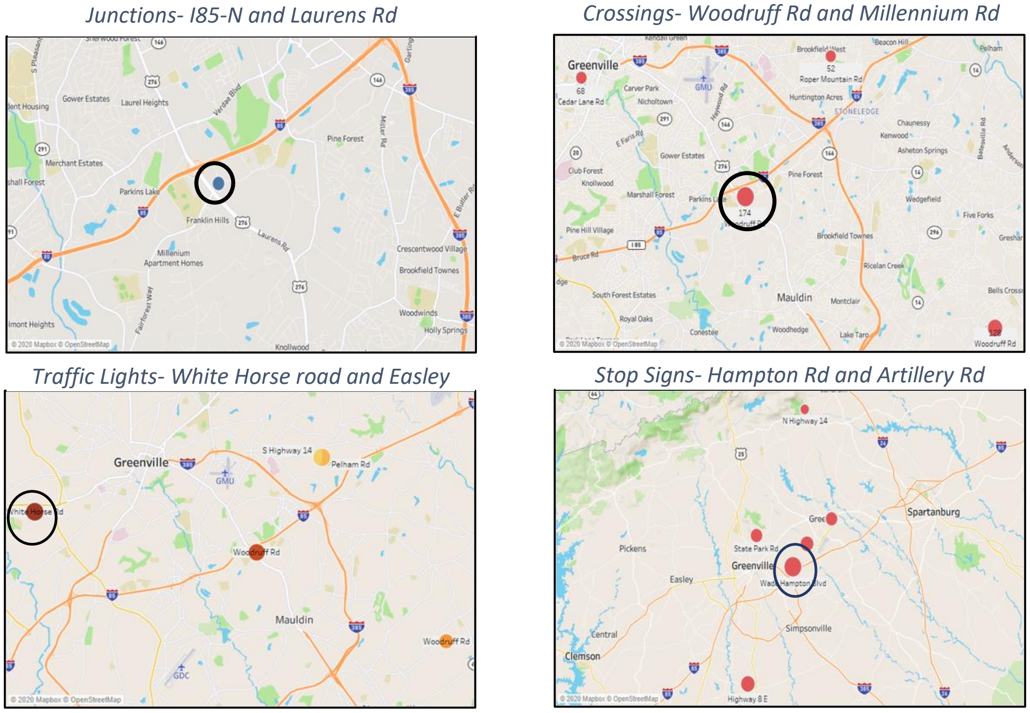


Figure 6: More accidents on right side of road. Orange- right side and blue-left side

In Figure 5, analysis of accidents in SC showed three main high concentration clusters of accidents. Greenville county had the highest concentration with over 20K accidents. This county will be drilled down for further analysis. Figure 6 reveals that most accidents occur on the right side of the road – with ride side accidents highlighted in orange.

The key features contributing to accidents are Traffic lights, Crossings, Junctions and Stops. Drill down on each of the key contributing features highlighted the roads that are the prone to most accidents namely the I85-N and Laurens Road for most accidents at junctions, White Horse Road and Easley Road had most accidents happen at a traffic light, Wade Hampton Road and Artillery Road had the most accidents at a stop sign and Woodruff Rd and Millennium Rd had the most accidents at a crossing Figure 7 below shows the hotspots with the highest accidents count at each feature. The next highest accident count at each feature can be seen in more detail using attached file [b]. The roads below represent accident prone intersections that would benefit from accident mitigation strategies.



Junctions- I85-N and Laurens Rd

Crossings- Woodruff Rd and Millennium Rd

Traffic Lights- White Horse road and Easley

Stop Signs- Hampton Rd and Artillery Rd

Figure 7: Top Left: Junction Hot Spots || Top Right: Crossings || Bottom Left: Traffic Lights || Bottom Right: Stops Signs

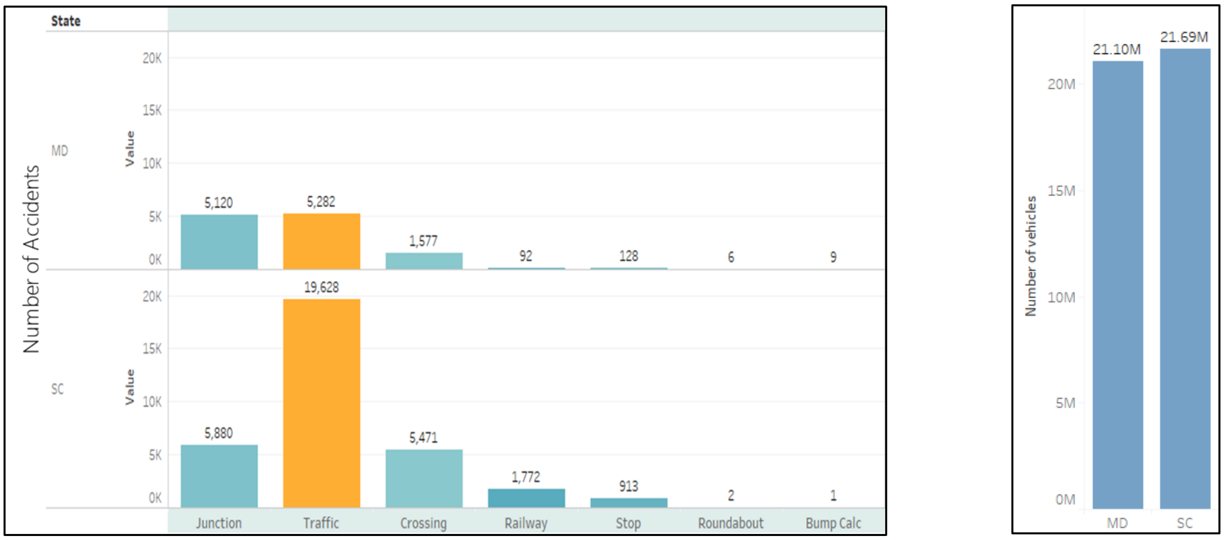


Figure 8: (Left) Comparison of Number of Accidents between Maryland and SC; (Right) Number of Vehicle Registrations per State

A comparative analysis was performed next, between the state of South Carolina and Maryland which had similar number of vehicles registered during the period of interest of this work. Figure 8 shows an interesting insight that many more accidents occur in South Carolina compared to Maryland despite their similar number of vehicle registrations. This led to further research about the traffic planning strategies [3] and it was found that Maryland has highest number of roundabouts in lieu of junctions and traffic lights. This could potentially be a solution to the state of South Carolina to mitigate the high accident rate.

Key Findings and Conclusion:

Upon meticulous and detailed analysis of accident data within state of South Carolina, the following key discoveries around features impacting severity of an accident were uncovered.

a. Across the US, accidents occur on non-clear weather days at a higher rate than on clear weather days.
b. There are a greater number of accidents on the right side of the road than the left in South Carolina.
c. The infrastructure feature that causes high accidents varied between the states [a], however traffic light is found to be most accident prone across the country.
d. Comparison of accidents between Maryland and SC showed that even though Maryland has similar number of vehicles, there are fewer accidents when compared to SC.
e. In South Carolina, most of the accidents occurred around 11 A.M and 4 P.M at crossings and junctions hotspots but this trend is slightly different for traffic lights and stop signs [7].
f. Hotspots locations varied between day and night in Greenville county [8].

This work highlighted the hotspots and locations that are accident prone in Greenville county SC (with the highest accident rate). The goal of this analysis is to highlight the roads to government agencies to come up with appropriate mitigation strategies. Various features like weather, infrastructure, side of the road, time of the day, that contribute to accident occurrence were analyzed and insights were uncovered. This analysis can be further extended to see insights of other states also using the Tableau file[a].

Recommendations

Based on analysis and comprehensive study of accident dataset for State of South Carolina, this project makes the following recommendations to policy makers and authorized personnel of respective departments (DoT, DMV etc.) towards potential reduction in number of accidents as well as severity mitigation:

- Smart signal systems instead of traditional traffic signals [4][6]
- Converting junctions to roundabouts [5]