1.0 Summary

Most people in the U.S. drive a vehicle, so traffic collisions are a relevant and interesting topic. More than a million people are killed on the world's roads each year. In the United States, road fatalities averaged more than 40,000 deaths annually for the past 40 years. FARS provides NHTSA (NHTSA - National Highway Traffic Safety Administration), and the American public, yearly data regarding fatal injuries suffered in motor vehicle traffic crashes. The primary goal of the system is to help reduce fatalities on U.S. roads and to aid in legislation and regulation of traffic safety programs. Fatality Analysis Reporting System (FARS) collected data from the year 1982 to 2017.

The program collects data for analysis of traffic safety crashes to identify problems, and evaluate countermeasures leading to reducing injuries and property damage resulting from motor vehicle crashes. The FARS dataset contains descriptions in a standard format of each fatal crash reported. To qualify for inclusion, a crash must involve a motor vehicle traveling a traffic-way customarily open to the public and resulting in the death of a person (occupant of a vehicle or a non-motorist) within 30 days of the crash. Each crash has more than 100 coded data elements that characterize the crash, the vehicles, and the people involved. The specific data elements may be changed slightly each year to conform to the changing user needs, vehicle characteristics and highway safety emphasis areas. The type of information that FARS, a major application, processes are therefore motor vehicle crash data. (NHTSA, 2017)

The data lets us create correlations with various measurement factors and helps us to come to statistical inferences. We would like to figure out the major factors involved in an accident scenario. We would further take the factors and create a model to determine the cause of the accident rates per location. Once the cause has been studied we can recommend implementation of safety procedures to the specific location after studying the location thoroughly.

The data is recorded and stored in categories which need to be read from the manual. To make it easier to understand the data, we will upload all the category names and make it easier for us to read it. We will in most cases not consider the missing data. But, places which are crucial we will be taking the median of the values of the column to replace the missing data. Further, we will try visualizing the data using GGplot2. Find inferences which majorly stand out and try working with them further to know the reasons behind it. Once we have determined the factors which play a crucial role in the crash we will try suggesting policies and procedures to help reduce the number of fatalities.

From the statistics of the data, nearly 35000 people have died in the year 2015 i.e. every day 100 people die due to fatality from road accidents, which is startling. After having a model, we could suggest preventive measures to the authorities. The data also gives us the location of the fatality, hence the places which are accident prone would need more attention and higher frequency of patrolling.

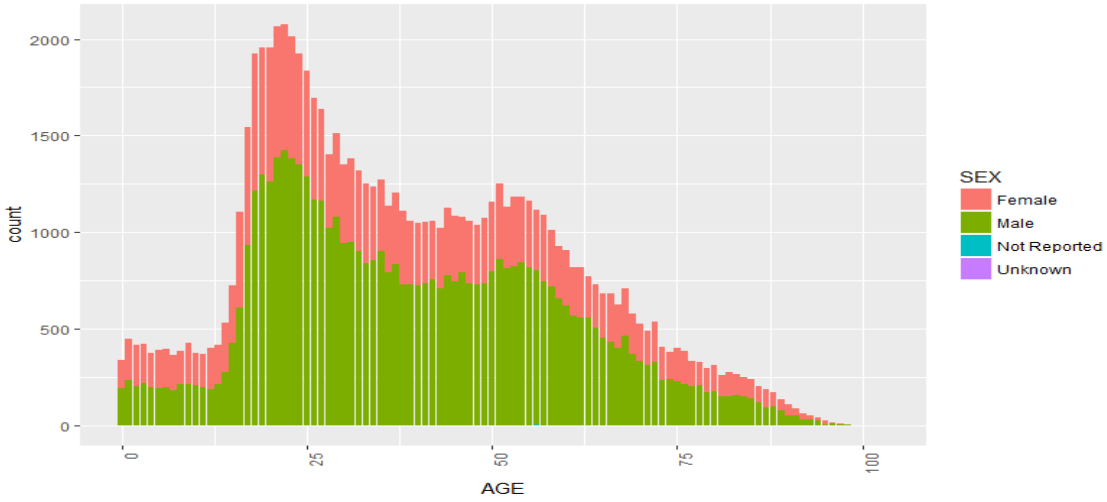
Our main aim for this project is to clean the data using different data wrangling techniques which includes data cleaning like missing data handling, normalization, transformations, outlier handling, visualization, and performing exploratory data analysis on the same. Creating visuals for a better understanding of the situation and viewing the statistical graph for various recommendations.

2.0 Findings

There were many correlations noticed some of the major factors which we notice causing the fatalities were

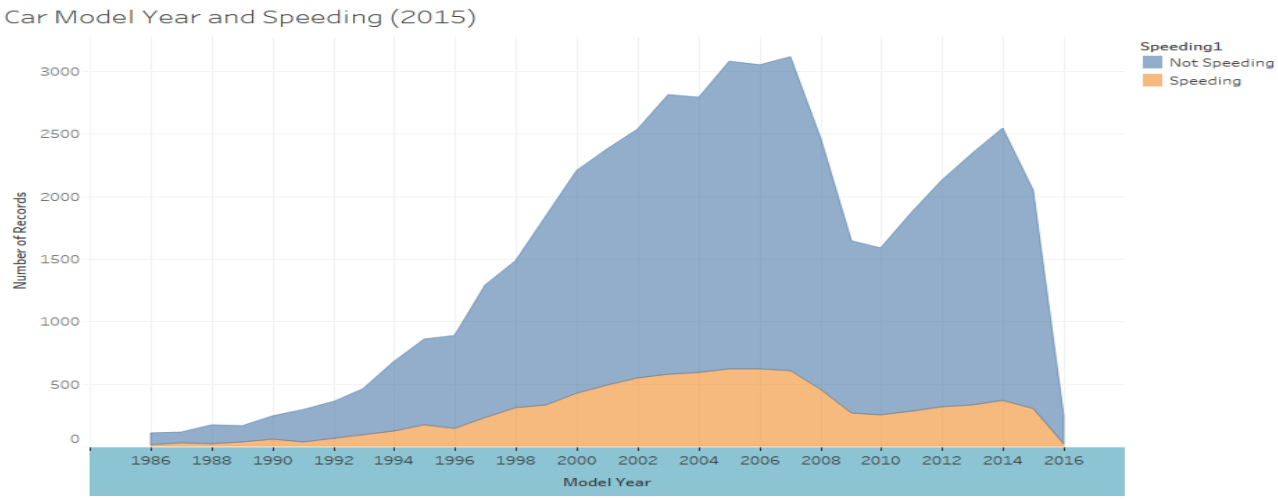
2.1 Age

We can see a general trend with the age where we have many fatalities taking place from the age group of 18-27 years. Then we notice a pattern in the age group of 60-75 years, which has been growing for the last 5 years. The aging population of the US will lead to an increase in these sections of fatalities. There must be subjected to periodical health checks.



2.2 Car Model Year

Car Model Year has been one of the interesting findings in this project. There is a trend pattern here which shows a high number of fatalities for cars which are 2-3 years old and with cars which are 8-17 years. The cars which are relatively new have a higher number of accidents due to more proportion of cars. While there is a larger number of fatalities due to the cars which are older than 8 years. There is also no proper census on the number of the car in the road today for each model year. There is a stark reduction in accidents in New England states which have periodic inspections of all vehicles. Which could be replicated in other states.

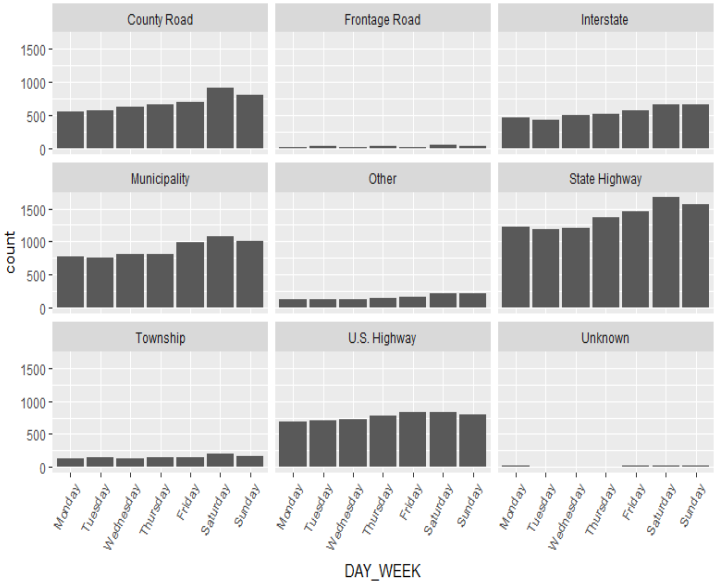


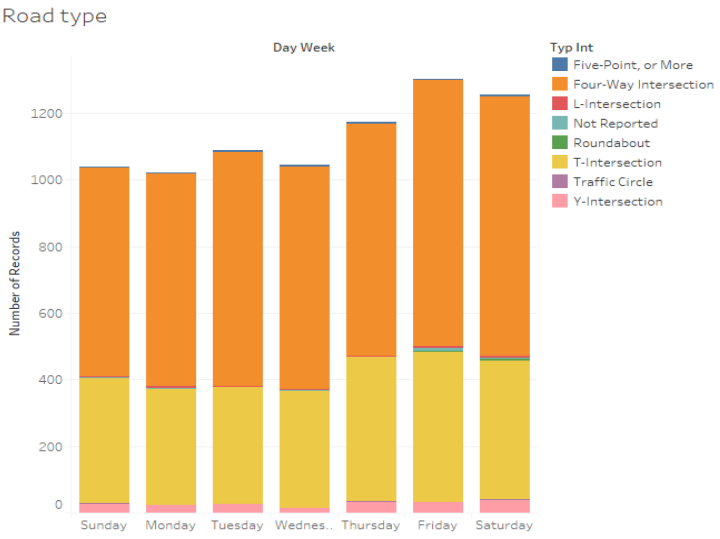
2.3 Road Type and Intersections

There are maximum fatalities is state highways. Increasing speeding fines and reducing speed limit can help in the reduction of fatalities.

The four-way intersection and T intersection curtail 90% of accidents in the intersections. There is an already existing solution to these problems which are roundabouts. These are used in countries like England. Studies have shown that roundabouts are safer than a traditional stop sign or signal-controlled intersections. Roundabouts reduced injury crashes by 75 percent at intersections where stop signs or signals were previously used for traffic control, according to a study by the Insurance Institute for Highway Safety (IIHS). Studies by the IIHS and Federal Highway Administration have shown that roundabouts typically achieve:

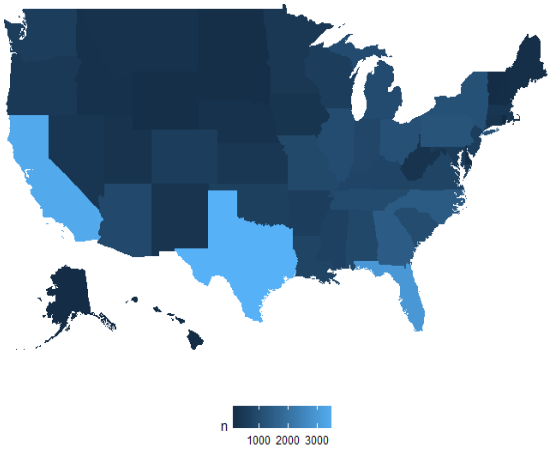
- A 37 percent reduction in overall collisions
- A 75 percent reduction in injury collisions
- A 90 percent reduction in fatality collisions
- A 40 percent reduction in pedestrian collisions

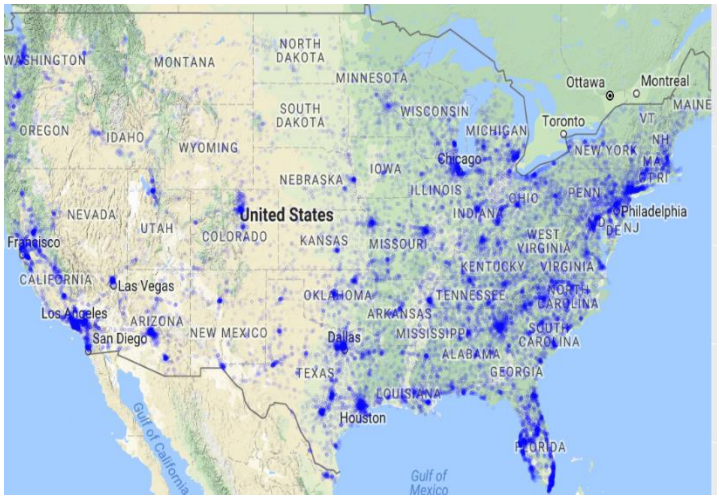




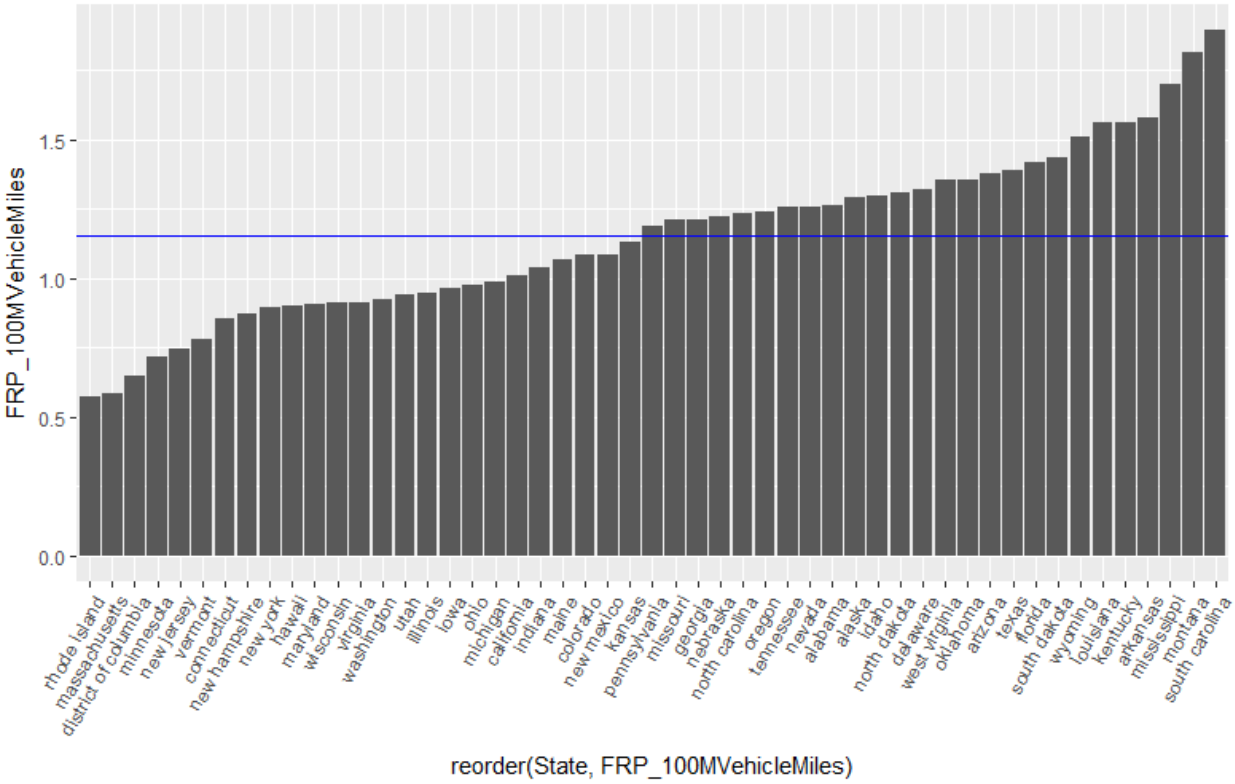
2.4 State

There is a high number of accidents in Texas, California, and Florida. Looking at the population and accident rate per 100,000 registered drivers Wyoming, South Carolina, Montana, Mississippi have a high number of fatal accidents are more unsafe. Driver educational programs in these states can reduce fatalities.

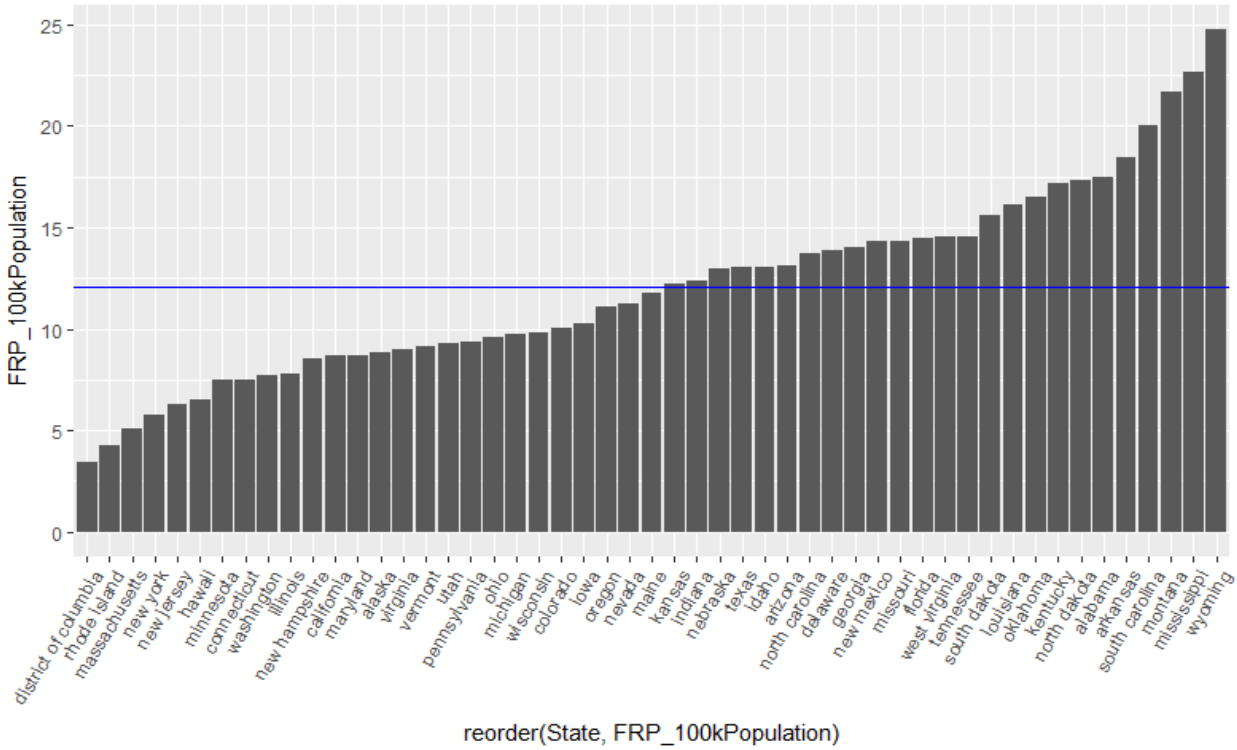




- Accident rate per 100,000 registered drivers

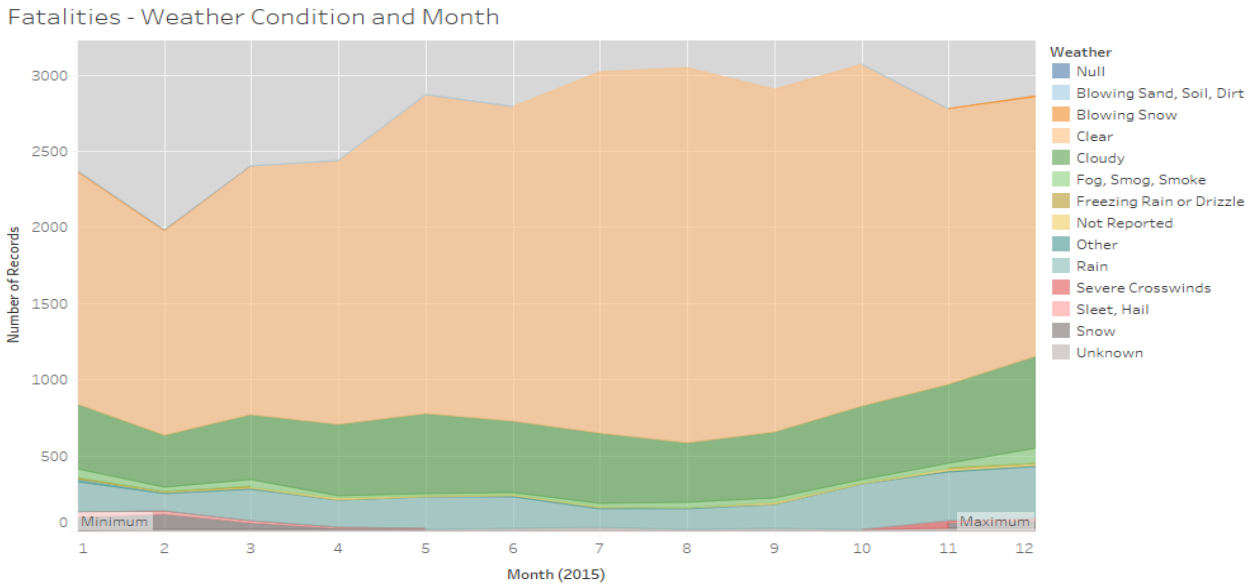


- Accident rate per 100,000 population in each state



2.5 Day of the Week, Month and, Weather

Friday and Saturdays have a high number of accidents compared to other days. August, September, and October have risen in 10-15% fatalities. Cloudy and Rainy days count for 15-20% of the fatalities. 70-75% of fatalities occur on a clear day according to the data.



The plot of sum of Number of Records for Month. Color shows details about Weather.

3.0 Solution

The US loses an average of 30,000 lives every year due to road fatalities. It impacts the US economy by 240 billion dollars. The sheer loss of life is staggering. Also, the startling facts about these deaths are most of these accidents up to 70% can be prevented. There are measures already existing. But they are constrained to few specific states helping them having half the fatalities as the national average.

The dataset is one of the most widely mapped of such kind. Even that falls short for the analysis. There could further fields added to the dataset like the reason for travel, financial status, personal life information, family information and pre-existing health conditions of the patients to give us better insight into finding the people who are prone to accidents. Also, inter-relating the fatalities dataset to a non-fatal dataset could give us more insight into making the model more reliable and target specific policies to people who might have a higher chance of being in the fatal accident and prevent it.

Although fatalities could not be avoided even with the introduction of autonomous vehicles. This dataset can be a learning curve for the technology and help it be a lot safer. We would try working on an algorithm which can highlight important locations where the accidents rates are high. When we have some high-frequency locations, we would like to investigate the detailed picture and information (with and without data) on the crash site. Coming up with measures for specific sites or locations to reduce accidents which could be implemented with immediate effect.

This would require studying various factors thoroughly for their role in the fatalities. It would be beneficial to include all the available data. The number of hours spent on sorting the data would be exponentially higher and even the book to decode the data set would be around 1,200 pages. It would certainly lead to more discoveries which will help reduce the fatalities.

Overall much more attention needs to be given to sorting the effects of highway safety policies in a rigorous manner that takes careful account of changes in external factors such as changing travel patterns and changing demographics.