**Reducing Traffic Mortality Through Data Science**

Saurabh Palhade

August 16,2020

**Introduction**

As the traffic related fatalities were rapidly increasing from the past few years in the USA, so in this project I used different data science techniques like data wrangling, plotting, dimensionality reduction and unsupervised learning to gain insights from the data and to understand If there are any patterns in the data and done statistical analysis in order to derive suggestions for a policy action plan for this issue of traffic mortality.

**Methods and Approaches**

After reading in and getting an overview of the data to familiarize myself with the data I produced the graphical overview of the data to get a sense for the distribution of variables within the data and also explored the pairwise relationship between all columns in the dataset by using a pairwise scatter plots. There were 3 features in the dataset. ie. percentage of drivers involved in fatal collision 1) who were speeding 2) who were alcohol impaired 3) who had not been involved in any previous accidents. After looking at the scatter plots I saw some relationship between the target variable (the no of fatal accidents) and the 3 feature variables. So to quantify the pairwise relationship I computed the pearson correlation coefficient matrix, from the correlation matrix I saw the target variable is most strongly correlated with alcohol consumption, but some other features were also correlated with each other, so to compute the association of the target with each feature while adjusting the effect of remaining features, I fitted a multivariate linear regression model on the data and checked the regression coefficients, from the regression coefficient and correlation matrix and saw that there were associations between alcohol consumption and the other two features, so I decided to split the states in a way that accounts for all the three features. Before clustering the data I used PCA to be able visualize the data in reduced dimensional space. After using PCA I saw that the first two principal components has captured high proportion of variation (99%) from all the three features. Hence to explore how the states cluster together I used the first two principal components and created a scatter plot for visualization. But it was not entirely clear from the PCA scatter plot that in how many groups the states cluster, so to get an indication of appropriate no of clusters I used KMeans clustering by creating a scree plot. Again there was not a clear elbow in the scree plot, so I assigned the states to 3 clusters. Now to get the visualization of clusters I created the PCA scatter plot colouring the states according to the cluster to which they are assigned. Moving forward in the analysis I explored how the three clusters are different in terms of the features that I used for clustering, by creating a visualization using violin plot splitting and coloured the results in plot according to KMeans clusters. From this plot it was clear that the different group of states requires different intervention. Now the question raised was which group to select first from this, so to answer this question I included the data of miles driven in each state to help me compute the total no accidents in each state, so from the new data I calculated and visualized total fatal traffic accidents within each cluster, also this marks the end of data anaysis.

**Results**

So from the visualization of total fatal traffic accidents within each cluster I suggest that the cluster no 2 should be a focus for policy intervention and further investigation as there were the highest no of fatal traffic accidents.