**Data Analysis – Traffic Accidents**

**INTRODUCTION**

As a part of the course work for Data Analysis, I have selected a dataset on US Accident, obtained from the website <https://smoosavi.org/datasets/us_accidents>.

This data set on US Accidents can be used for numerous applications such as real-time accident prediction, studying accident hotspot locations, casualty analysis and extracting cause and effect rules to predict accidents, or studying the impact of precipitation or other environmental stimuli on accident occurrence.

**DATASET**

 In this analysis I used data that is collected from February 2016 to June 2020 for the [Contiguous United States](https://en.wikipedia.org/wiki/Contiguous_United_States). In this dataset, we have about **3.5 million** traffic accidents. The dataset covers 49 states of the US. This dataset consists of 49 attributes. As part of the analysis some attributes have been excluded in data cleansing process.

**METHODOLOGY USED FOR ANALYSIS:**

## Step 1: Data cleaning:

Created a subset with variables using for this project. Creating factors of variable and removing the unknown and null values.

## Step 2: Creating plots:

Generating plots by using ggplot2 library.

## Step 3: Perform Regression analysis:

Selecting best subset of variables to perform regression analysis by using regularized linear regression method. Split data into two parts train and test. Creating multivariate regression model by using train data set and test this model with test dataset and finally conclusion.