# **What insight does the UK accidents data give us?**

By Shivani Sanjay Mahaddalkar

**Area:** The analysis examines the UK’s traffic dataset of all major traffic accidents reported from 2000 to 2016.

**Data:** The data set is sourced by user Dave-Fisher Hickey on Kaggle. The data is originally sourced from the UK Department of Transport website.

**Importance of the problem:** Between 1951 and 2006 a total of 309,144 people were killed and 17.6 million were injured in accidents on British roads. With increased emphasis on road safety and awareness, these fatalities could be avoidable. Therefore, it is necessary to determine what improvements can be made.

**Objective:** The projects aims to assess the gaps in the traffic data. It will try to predict which factors increase the accuracy of prediction in the number of casualties. It will visualize the frequency of various factors over time.

**Tasks:**

1. Comparing frequency distributions of the different factors in the dataset over time.
2. Comparing frequency distributions of the different factors against the number of casualties in each cases.
3. Creating predictive models(using models like decision trees, k-means clustering) to predict the number of casualties.
4. Using n-fold cross validation to test the predictive power of the models.

**About The Data:** The data is sourced by user Dave-Fisher Hickey on Kaggle. The data is originally sourced from the UK Department of Transport website. It contains major traffic accidents reported from 2000 to 2016, the data from the year 2008, however, is missing. The data set has 33 columns, including columns like ‘Road Type’, ‘Speed Limit’, ‘Light Conditions’, ‘Weather Conditions’, etc. which on preliminary glance seem like factors that could have a major impact on an accident occurring and its severity.

**Tools:**

Data-set: 1.6 million UK traffic accidents from Kaggle, originally sourced from the UK Department of Transport website.  
Tool: R studio with R markdown files  
Models: Decision Trees, K means clustering and Naïve Bayes for prediction.  
Evaluation criteria: 10-fold cross validation