DATA ACQUISITION AND CLEANING

2.1 Data Sources

The secondary data used for this analysis has been taken from Kaggle which consist of accidents in UK over a period of 10 years ranging from 2004 to 2014 and which is further divided into four datasets namely Casualties, Road Accident Safety Data Guide, Accidents and Vehicles.

2.2 Data Cleaning and Feature Selection

Data downloaded or scraped from the secondary source were put together and a thorough study of data was done. The data was kept recorded and then data cleaning was done to make sure that results are significant at the end.

After loading the datasets, information about each dataset was printed to get a better look and idea about the data. The important questions that arose from this were-

1. Is there any relationship between the timing of the accident and number of fatal accidents?
2. Does the age of driver have any effect on the number of accidents happened?
3. Hoe does the weather impact the severity of accidents?
4. Are certain designs of vehicles safer than others?

To answer the above questions along with some others, I merged the casualties and accident dataset first and then again merged it with vehicle dataset on the basis of accident index. After checking for the missing values in the new dataset, I dropped some features like latitude, longitude, accident location because there were quite a few missing values in those categories and the results would not have been valid.