**Problem understanding:**

Every year there is X number of accidents on the British roads out of which Z% is severe. Hospital directors that plan resources may want to predict risk and severity of accidents on streets to help people that were victims of car accidents. Heads of police departments want to know how to protect citizens and patrol the streets so to limit dangerous behaviour on the streets. Aim of this project is to predict the severity of road accidents so to limit the number of incidences on the way and optimize resources to protect victims. To understand what influence the severity of the accident we may want to look at things like: weather conditions, light conditions, road conditions or place location.

**Data source:**

To this project I have used data from British department of transportation on car accidents that happened in years 2010 to 2018 <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data> The raw data is made of two files: vehicle details and accident circumstances. Vehicle details include facts such as: vehicle type, junction location, sex and age band of driver, vehicle propulsion code and more. Accident circumstances include facts such as: police force division that attended the accident, road type, speed limit, weather conditions, urban or rural area and more. For full details on columns included please refer to the appendix.

Looking at the data I have noticed that the accident with maximum number of vehicles was not on the same data as accident with maximum number of casualities. The maximum number of casualities recorded during car accident was 93 on date 20/10/2014 and the maximum number of vehicles recorded during car accident was 67 on date 05/09/2013. I found information on the second accident in the press so it will allow me to validate the accuracy of the data.

**Data Cleaning:**

The data comes from multiple years, each year in separate csv file. Comparing columns across years we see that Age\_of\_Driver and Vehicle\_IMD\_Decile are not in all files across the years so we remove the variables in files that the columns exists so to uniform the data and we normalize the column names across the files. Files from the same category: vehicle details or accident details we append on top of each other to have one dataset for each year and then we join together both complete files on accident index column. To better understand the data types and what columns say about the incident I recode the data using variable lookup file provided. Some of variables are in too granular level so that the ratio across different options is too big. For this reason some columns are changed to Yes|No type columns such as: Vehicle\_Leaving\_Carriageway, Hit\_Object\_off\_Carriageway, Carriageway\_Hazards and more. Finally we have 37 columns of type string, 11 of type float and 5 of type integer making a total of 53 columns and nearly 3mlns observations.

Some of the missing values are left blank, some are recorded as -1 and some are recorded as unclassified or else. To better understand the missing values I calculate proportion of all missing values cases per column. For some columns such as Was\_Vehicle\_Left\_Hand\_Drive?, Propulsion\_Code, 2nd\_Road\_Class, Junction\_Control, Age\_of\_Vehicle I imput missing values by the most popular option. Time column I imput using forwardfill as I noticed that the data is in chronical order. For column Age\_Band\_of\_Driver the missing values I change on Unknown. I noticed that the data has some records for driver in the age 1-5, 5-10, 10-15 since this is illegal to drive in this age and that doesn’t make sense I simply make this as unknown as well. For columns such as Engine\_Capacity(CC) or Driver\_Home\_Area\_Type the missing values I imput using mode grouping by Vehicle\_Type or Accident\_Severity. Since I want to visualise the data on a map with accident severity but I don’t need the columns Longitude and Latitude in the general data I create temporary data set with Accident\_Index, Accident\_Severity, Longitude and Latitude and I remove those two geographical columns from the main dataset. In total I have removed 10 columns either because the number of missing values was really more than 50% or because they would not give good insights. For example Location\_Easting\_OSGR and Location\_Northing\_OSGR are geographical variables. We already stored the longitude and latitude and in the main data we have column Police\_Force that gives information what area of Britain the Police\_Force operate in. Local\_Authority\_(District), 1st\_Road\_Number, 2nd\_Road\_Number are too granular. We don’t have information on LSOA\_of\_Accident\_Location or Driver\_IMD\_Decile to know what does it mean.

I have noticed that for each vehicle taking part in the accident there is separate record so all columns with small proportion of missing values first I tried to get the missing values from the accident index as they were taking part in the same accident. In total I managed to get Y% of the total data cleaned.