Capstone Project 1 - Data Wrangling

For my Springboard Capstone Project I, I will be using the following three datasets downloaded from the website: https://data.gov.uk/dataset/road-accidents-safety-data. The data is from UK, with information about injury related accidents that were reported to the police in the year 2016. My goal is to analyze the data to understand the various factors that cause serious accidents.

The zipped data file consists of the following three datasets in CSV format:

1) Accident.csv
2) Vehicle.csv
3) Casualty.csv

Accident Data:

This dataset has 31 columns and 136621 rows. Some features have missing or 'Nan' values. In particular, LSOA_of_Accident_Location', 'Location_Northing_OSGR', 'Location_Easting_OSGR', 'Longitude', 'Latitude' have missing values. These features, which have information about location, latitude, longitude or county will not be used in the study and have been dropped. In addition to these features, 'Speed Limit' and 'Time of Accident' 37 and 2 missing values. Clearly, these are important variables for the analysis. There are several ideas that have been suggested in the literature on how to deal with missing values. We could impute them by a summary statistic like mean or median or do a simple linear regression to predict the missing value. But, that is suboptimal, since not only are we using information already in the data, but adding some noise at the same time. We cannot leave it as it is, as these features will most likely be predictive in the supervised learning algorithm, and creating separate dummy variables for a few observations, 37 and 2 respectively, will not be predictive. Since it's a small number as a percentage of the dataset, the corresponding 39 rows have been dropped. After deleting the 5 columns and 39 rows, the Accident dataset now has 26 columns and 136582 rows. Some variables like 'Road Conditions' or 'Light Conditions' have rows (few hundreds to 1) with missing information, indicated by -1. These are categorical variables, and one possibility is to simply create a separate category, and then drop them from the features dataset during modeling to avoid multicollinearity.

Vehicle Data:

The vehicle dataset has 22 columns and 252,000 rows. There are no missing values indicated by 'Nan', but several variables have missing values indicated by '-1'. The home area type of the driver (urban, rural or small town) has 47,490 missing values, while "Age Band of the Driver" has 29,418 missing values indicated as '-1'. As before, these will treated as a separate category while creating dummy variables during modeling.

Casualty Data:

The Casualty dataset has 15 columns and 181384 rows and no variables with missing values indicated by ‘Nan’. But several variables like ‘Home Area Type of Casualty’ have missing values indicated by ‘-1’.

Outliers:

There are 25 data points with number of casualties more than 12, and 13 data values where number of vehicles involved is 10 or more. Clearly, these are outliers, but a very small percentage of the overall dataset. Moreover, while some supervised learning algorithms like Logistic Regression are sensitive to outliers, there are several that are robust to outliers, like Decision Trees and Support Vector Machines.

Data Merge:

Each dataset has a ID variable, that uniquely identifies the accident. This is unique in the Accident dataset, but may have non-unique values both in the Vehicle and the Casualty datasets respectively. This is because a particular accident may involve one or more vehicles and casualties. To proceed with our analysis, we would then need to merge the three datasets. The first merge involves the Vehicle and the Casualty data sets. Its an outer join, on two indices, namely, Accident ID and the Vehicle Reference column. The second merge involves this merged dataset with the Accident data. This is also an outer join, but now on the index Accident ID. The final merged dataset has 252,000 rows.