- How many rows and attributes? - How many missing data and outliers? - Any inconsistent, incomplete, duplicate or incorrect data? - Are the variables correlated to each other? - Are any of the preprocessing techniques needed: dimensionality reduction, range transformation, standardization, etc.? - Does PCA help visualize the data? Do we get any insights from histograms/ bar charts/ line plots,

Etc.?

(i) Rows and attributes in the dataset:

The dataset consists of 3 CSV files for accidents, casualties and vehicles. The accidents CSV consisted of 32 columns and 1780653 rows. The vehicles CSV has 21 columns and 3004425 rows. Furthermore, the casualties CSV has 14 columns and 2216720 rows.

Columns in accidents CSV :

['Location_Easting_OSGR', 'Location_Northing_OSGR', 'Longitude',
'Latitude', 'Police_Force', 'Accident_Severity', 'Number_of_Vehicles',
'Number_of_Casualties', 'Date', 'Day_of_Week', 'Time',
'Local_Authority_(District)', 'Local_Authority_(Highway)',
'1st_Road_Class', '1st_Road_Number', 'Road_Type', 'Speed_limit',
'Junction_Detail', 'Junction_Control', '2nd_Road_Class',
'2nd_Road_Number', 'Pedestrian_Crossing-Human_Control',
'Pedestrian_Crossing-Physical_Facilities', 'Light_Conditions',
'Weather_Conditions', 'Road_Surface_Conditions',
'Special_Conditions_at_Site', 'Carriageway_Hazards',
'Urban_or_Rural_Area', 'Did_Police_Officer_Attend_Scene_of_Accident',
'LSOA_of_Accident_Location']

Columns in vehicles CSV :

['Vehicle_Reference', 'Vehicle_Type', 'Towing_and_Articulation',
'Vehicle_Manoeuvre', 'Vehicle_Location-Restricted_Lane',
'Junction_Location', 'Skidding_and_Overturning',

'Hit_Object_in_Carriageway', 'Vehicle_Leaving_Carriageway',
'Hit_Object_off_Carriageway', '1st_Point_of_Impact',
'Was_Vehicle_Left_Hand_Drive?', 'Journey_Purpose_of_Driver',
'Sex_of_Driver', 'Age_of_Driver', 'Age_Band_of_Driver',
'Engine_Capacity_(CC)', 'Propulsion_Code', 'Age_of_Vehicle',
'Driver_IMD_Decile', 'Driver_Home_Area_Type']

Columns in Casualties CSV :

['Vehicle_Reference', 'Casualty_Reference', 'Casualty_Class',
'Sex_of_Casualty', 'Age_of_Casualty', 'Age_Band_of_Casualty',
'Casualty_Severity', 'Pedestrian_Location', 'Pedestrian_Movement',
'Car_Passenger', 'Bus_or_Coach_Passenger',
'Pedestrian_Road_Maintenance_Worker', 'Casualty_Type',
'Casualty_Home_Area_Type']

(ii) Missing Data and outliers:

All of the outlier and missing data analysis was performed mainly on the accidents CSV file. The number of missing values for each of the 32 attributes in the dataset is as follows: [0, 138, 138, 138, 138, 0, 0, 0, 0, 0, 0, 151, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 129471] (attributes are mentioned below). The records having missing values for each of these attributes will be removed to handle the missing data and to avoid inconsistencies.

Box plots were used to identify the outliers for the numerical attributes in the dataset. A box plot was constructed for the speed limit grouped by accident severity which showed no outliers since all the values were within the appropriate range. The box plot for the age of the victim showed age groups of 80+ as outliers which is logically sound.

(iii) Correlation between the variables:

To analyse the correlation between the different variables, a heat map was plotted which visualised the correlation between all pairs of variables. The following are some of the pairs of attributes found to have minimum correlation : (Police force, location), (Road Junction Details, Road Type), (Carriageway hazards, Junction Detail).

Insights from histograms/bar charts/line plots/scatter plots/pie charts/box plots:

(1) Bar graph - Number of values in each unique accident severity,
- A severity of type 3 had the highest amount of values for each unique accident severity value. Severity type 1 had a very low number of values.

(2) Line graph - Percentage of accidents resulting in at least 1 fatality per speed limit
- The percentage steadily increased till 60 mph and then started declining after that.
- This implies the most number of fatalities per accident occured in the 60 mph speed limit.

(3) Horizontal Bar Plot - Number of accidents and casualties by speed limit
- The most number of traffic accidents (records) had 30 as their speed limit.
- This implies most accidents happened at roads where the speed limit was set to 30.

(4) Horizontal Bar Plot - Road Types vs Number of Accidents
- The maximum number of accidents was recorded for single carriageway (800000) and the minimum number was recorded in Slip road
- The other types 'Roundabout', 'One way street', 'Dual carriageway' had comparable accidents with slip road.

(5) Vertical Bar Plot - Years vs Number of Accidents
- The dataset comprises data from 2005 to 2015. The bar plot showed a gradual decline in the number of accidents as the years progressed

(6) Pie chart - Percentage of accidents with speed limit
- The minimum percentage was noticed under 10 kmph
- The maximum percentage of accidents was observed for 30 kmph. The large percentage is due to the fact that the accidents include all Severity types.

(7) Vertical Bar Plot - Road Condition vs Accident Severity
- The accident severity was “fatal” for a few instances in single carriageway and dual carriageway.
- The accident severity was mostly “slight” for all the road types except

Single carriageway.

(8) Pie chart - Percentage of accidents with sex of casualty.

- The percentage of male casualties was found to be 66% and the percentage of female casualties was 34%.

(9) Box Plots - Speed limit grouped by Accident Severity

- The range of speed limit was between 0-110 kmph.
- There were no outliers since all the accidents were well within the possible range.

(10) Box Plot - Age of Casualty

- The range of age was observed to be 0-100.
- The maximum outliers were observed between 80-100 years of age

(11) Scatter plot :

- Number of Vehicles vs Number of Casualties - The scatterplot was plotted using 5000 random samples from the Accidents dataset. Sampling was done without replacement. The plot seems to show very little correlation between the 2 variables. This is backed up by a very low Pearson correlation coefficient of 0.24328361087674746