Analysis of Road Safety Data for England (12 Months to 31st December – 2019)

Group Information:

Name	Student ID
Achyuth Jayakumar Maneesha	201679567
Nikhil Shivasankar	201671058
Rohini Krishna Preetha	201669138
Unnikrishnan Radhakrishnan	201644452

INTRODUCTION

This report examines data on road safety in England for the 12 months leading up to December 31, 2019.The dataset we have at our disposal contained data for the entire United Kingdom, but our goal requires information specific to England only. This annual report brings out insights into the quality of data and provides detailed road safety data about the circumstances of personal injury road collisions in England. Since the data for accidents and casualties may be accessed individually, multiple transformations were carried out to obtain the necessary information for the analysis. After an appropriate combination of the dataset is available, this report emphasises some key data characterisation tasks and data quality issues. Finally, the report examines the patterns for pedestrians who were killed or seriously injured (KSI) casualties, demographics of casualties and between the local authorities. The following points indicate the data source and the number of files used in this report with their references.

Data Controller: Government of the United Kingdom, Department for Transport

Data Source: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

Data Files:

The following are the different files used in the analysis:

- dft-road-casualty-statistics-accident-2019 (CSV) (~18 MB)
- dft-road-casualty-statistics-casualty-2019 (CSV) (~9.2 MB)
- Road-Safety-Open-Dataset-Data-Guide (XLSX) (~55 KB)

DATA QUALITY

To transmit meaningful insights, it is crucial to guarantee that the data's overall quality has correctness and completeness. As a result, this report presents the results of some of the most widely used data quality tasks. Both completeness and correctness are addressed in this report. This section is divided into two subsections that discuss the results of the correctness and completeness tasks, respectively.

COMPLETENESS

Completeness guarantees that adequate quantity of data is used for analyses. The tasks are as follows:

Coverage

- There are 322 different local authority districts
- There are 10 distinct IMD decile and pedestrian locations
- There are 11 different age band of casualties
- There are 9 kinds of pedestrian movements.
- Number of distinct pedestrian’s crossing physical facilities accounts to 7
- There are 3 different types of casualty severity, sex of casualty and home area types

Duplicates

In accident and casualty dataset, there were no duplicate records. However, values for columns like "police force," "accident severity," etc. use code/format as depicted in the file ‘Road-Safety-Open-Dataset-Data-Guide (XLSX)’. For instance, even though the records in the accident index of casualty dataset have 28204 duplicates, repetition of those values is to be expected since one accident index corresponds to one or more casualties.

Missing values or records in the source data

- 1266 records for ‘pedestrian_road_maintenance_worker’ are unknown.
- The number of ‘pedestrian_movement’ that are unknown or come from other sources is 5860.
- There are 2076 ‘pedestrian_location’ records with unknown values.
- There are 23 unidentified ‘bus_or_coach_passenger’ records.
- There are 207 unknown or self-reported records of ‘car_passenger’.

As all of these records were cleaned, the data set we used had no missing data.

Rate of Recording

Road safety causality data till 2021 was published on October 2022. Similarly, Road safety accident data till 2021 was also published on October 2022. Both of these data were therefore published 10 months after the end of corresponding year.

Recency

These files contain information about the circumstances of personal injury road collisions in Great Britain from 1979 that is detailed in terms of road safety. We are only using data on road safety in England for the 12 months ending on December 31, 2019, despite the fact that the road data was last updated on November 30, 2022. The data set on the gov.uk website is current and valid, and it is constantly updated. Hence the data used is relevant.

CORRECTNESS

Correctness ensures that analysis always relies on data of good quality. The overall accuracy of our merged data is excellent as it follows a standard pattern of capturing and storing information. The data across all the datasets are consistent as it reflects the presence of all 11 different age bands, three different casualty severity, and ten distinct IMD decile. Hence, there is no consistency issue with the data. No outliers were found when analysing the distribution of casualties among the various age groups.

DATA CHARACTERISATION

This section of the report focuses on highlighting significant statistics from some of the standard data characterization tasks. It is broken into the following three sections:

CARDINALITIES

Prior to doing complicated analysis, cardinalities give early insights into the data structure's framing and aid in gaining control over the right use of the data. The cardinality tasks that are frequently carried out in data projects are as follows:

Total Number of Rows in our Merged Dataset: 109694 Rows * 32 Columns

Total Number of Distinct Values in the Main Attributes:
Figure 1 shows how many distinct values there are in our main attributes and how they compare to one another.

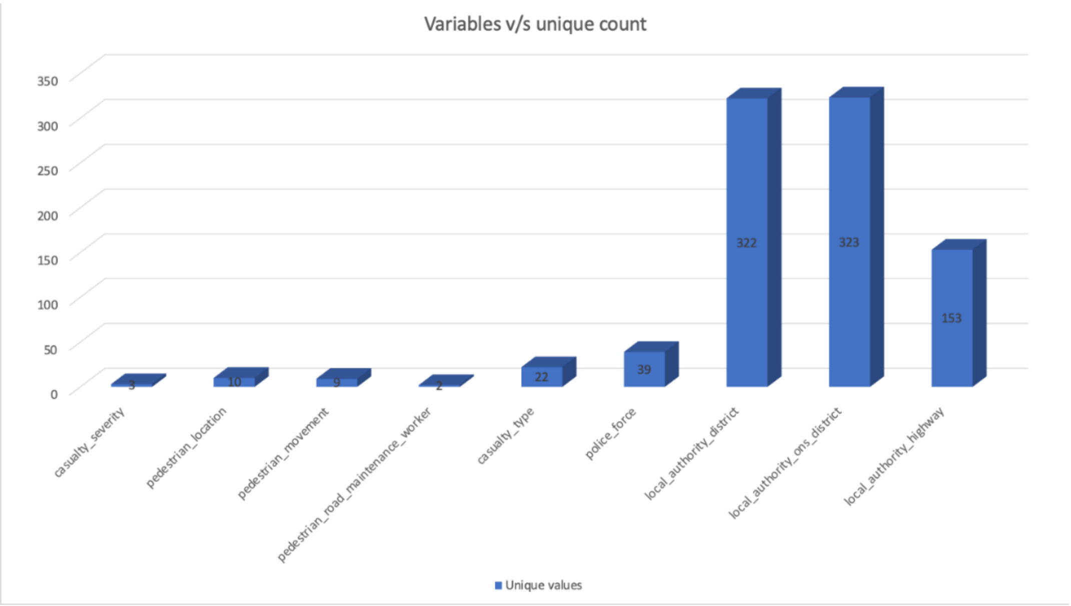


Figure 1: Number of unique values in the main attributes

DISTRIBUTION

The distributions of the number of casualty and age of casualty are measured. In addition, it gives outliers, quantiles, range, and the mean. Figure 2 and Table 1 provide explanations for this section.

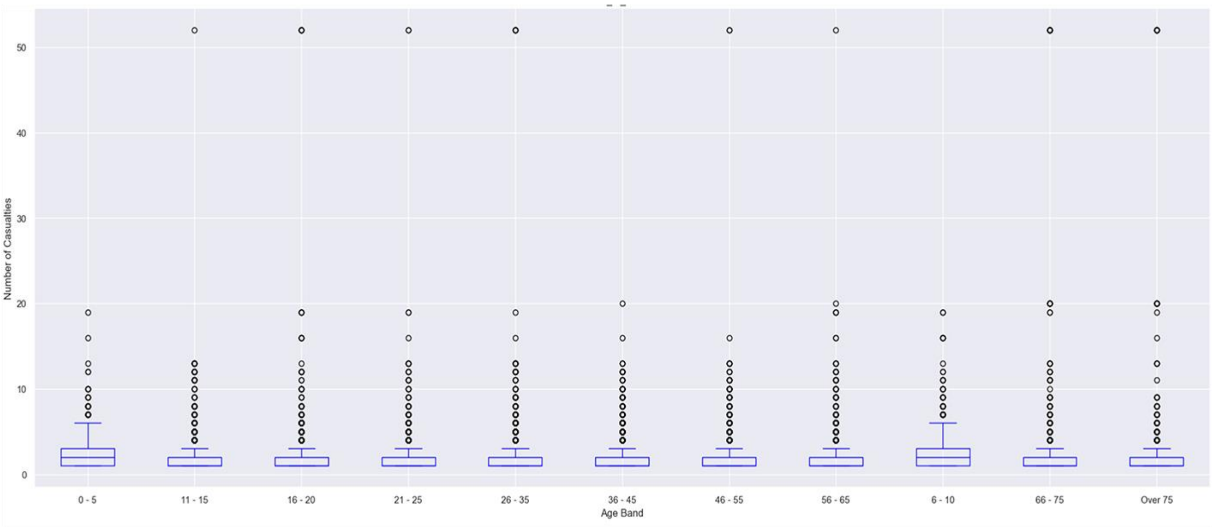
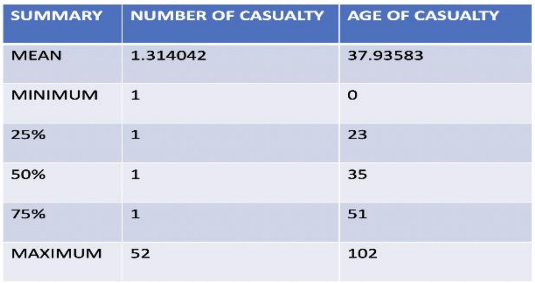


Figure 2: Box plot for the Age band v/s number of

The maximum number of casualty is 52 and its mean is 1.314. The maximum age of casualty is 102 and the mean age is 37.93.



SUMMARY	NUMBER OF CASUALTY	AGE OF CASUALTY
MEAN	1.314042	37.93583
MINIMUM	1	0
25%	1	23
50%	1	35
75%	1	51
MAXIMUM	52	102

Table 1: Descriptive Statistics for number of casualty and age of casualty

PATTERNS

Through cross-tabulation, correlation, trend, and analysis of various data types and examples, this subsection focuses on identifying significant patterns within the overall dataset. Each heading is supported with a table or a pertinent chart, depending on the information in the source. The patterns as shown in Figure 3 below can be used to determine the severity of injuries.

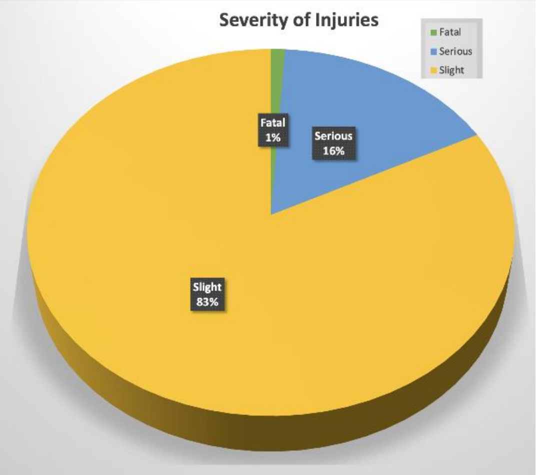


Figure 3: Severity of Injuries

Quite a few features from our combined dataset, which combines the accident and casualty records, are employed in the data analysis. Only a few of these attributes are significant to our main aims. Table 2 lists the specifics for those attributes, including data types, descriptions, etc.

Column Name	Data Type	Description	Example
casualty_imd_decile	Int64	Indices of multiple deprivation (IMD) decile-shows the relative index of deprivation	1, means Most deprived 10%
casualty_severity	Int64	Severity of the injury/casualty	1, means fatal
pedestrian_location	Int64	Provides details on where the pedestrian is	6, means on footway
pedestrian_movement	Int64	Provides details on how a pedestrian will cross a road.	1, crossing from driver’s nearside
pedestrian_road_maintenance_worker	Int64	Reveals whether a pedestrian is a worker performing road maintenance.	1, Yes
casualty_type	Int64	Identifies the type of injured individual	0, Pedestrian
accident_index	Object	Combination of the accident year and the accident_ref_no that creates a unique ID for each accident.	2019010128300
local_authority_ons_district	Object	Provides the district's local authority code	E06000014, York
local_authority_district	Int64	Identifies the Local Authority District	204, Leeds

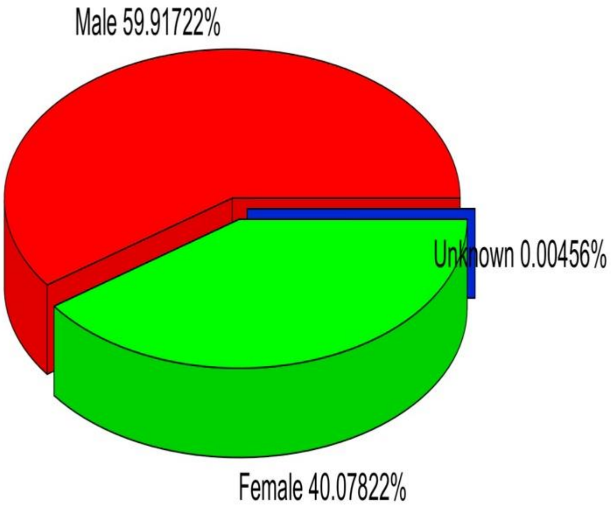
Table 2: Data attributes and their types with example

DETAILED ANALYSIS

As per the catalogue, three different analyses are performed in detail with the complete data for road safety data (called STATS19) for England for the 12months to 31/12/2019. The section is classified into three categories with each representing the outcome of a specific target.

Patterns in the demographics of casualties

Analysis of the general population's age, sex, birth rate, income, and other parameters is included in demographic studies. As a result, analysis was conducted using three criteria: home area kinds, causality sex, and age band. The percentage distribution of casualties based on sex is shown in the following pie chart (Figure 4).



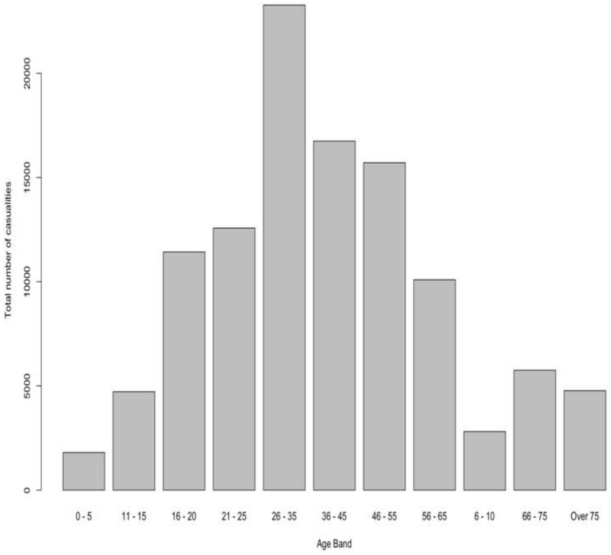


Figure 4: Percentage of causalities by gender.

Figure 5: Age band distribution

The pie chart shows that there are roughly 59.91722% males and 40.07822% females, with unknown values being less than 0.00456%, indicating that males make up most of the population when it comes to causalities. A figure showing the age band distribution from 0 to 75 and over is shown in Figure 5. According to the bar plot, the number of causalities is lowest among children in the age range of 0 to 5 years and continuously rises until 26 to 35 years, when it reaches its peak.

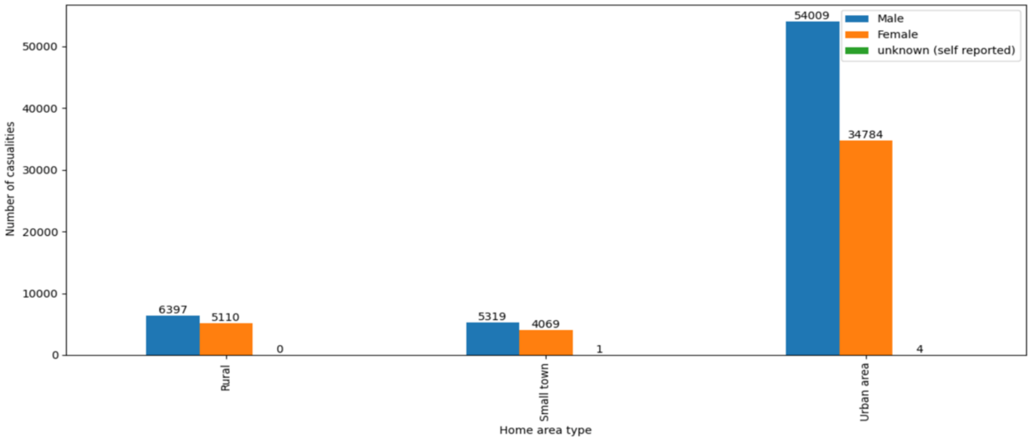


Figure 6: Bar plot of home area type vs number of causalities

Based on the home area type of causalities, as shown in Figure 6, another bar plot was created. The bar plot demonstrates that, when compared to rural and small-town locations, the number of causalities is significantly higher in metropolitan areas. Male causalities are more prevalent in all three of the home area types under consideration.

Patterns between local authorities for accidents which included KSI casualties

Analysis of local authority district-level trends in accidents that resulted in fatalities or serious injuries to victims was performed. According to the data, killed and seriously injured are defined as ‘Fatal’ and ‘Serious’ respectively. Figure 7 shows the top 10 local authority districts in terms of accidents recorded with fatalities or serious injuries. With 7 fatalities and 246 serious injuries, Birmingham is the district with the highest number of accidents. East Riding of Yorkshire, in contrast, has more deaths than Birmingham and is second highest among the top 10.

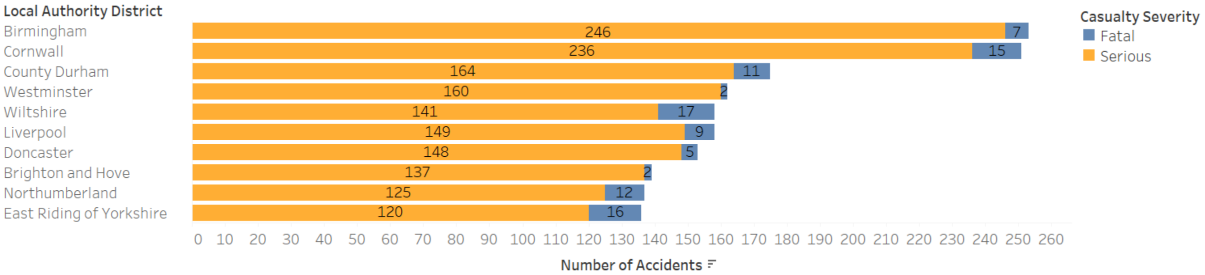


Figure 7: Top ten districts for accidents resulting in fatalities or serious injuries

Figure 8 illustrates further investigation to identify the districts with the highest toll of casualties. With 727 victims, of which 99.45% were seriously injured, South Hams appeared to be in first place. The county with the most fatalities among the top ten was Huntingdonshire. Birmingham, Cornwall, County Durham, Doncaster, Wiltshire, and Northumberland are the districts with the highest number of accidents and casualties.

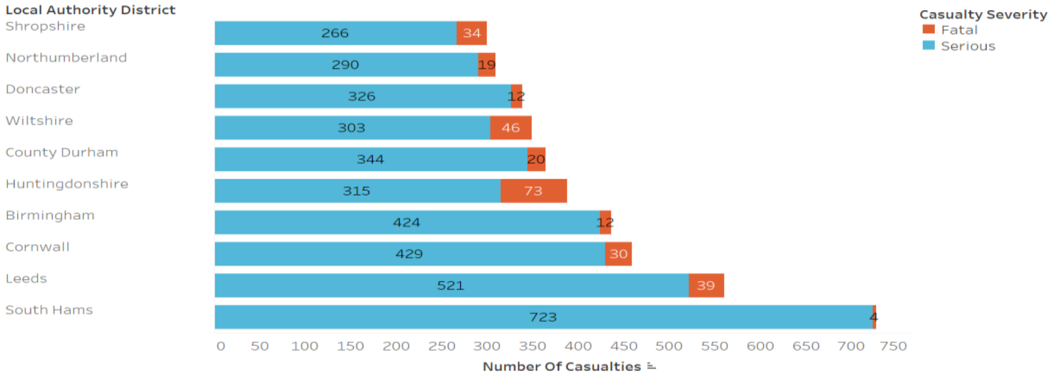
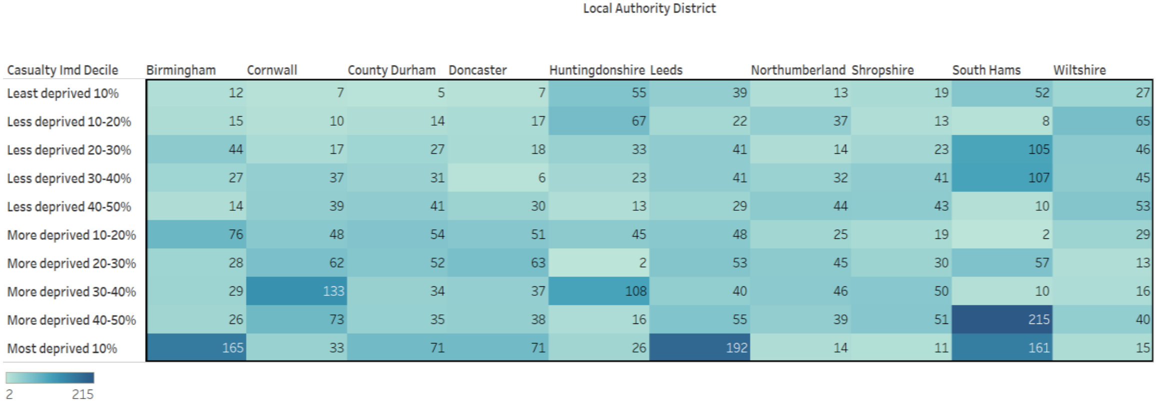


Figure 8: Top ten districts having the highest number of KSI casualties


Local Authority District
2
215

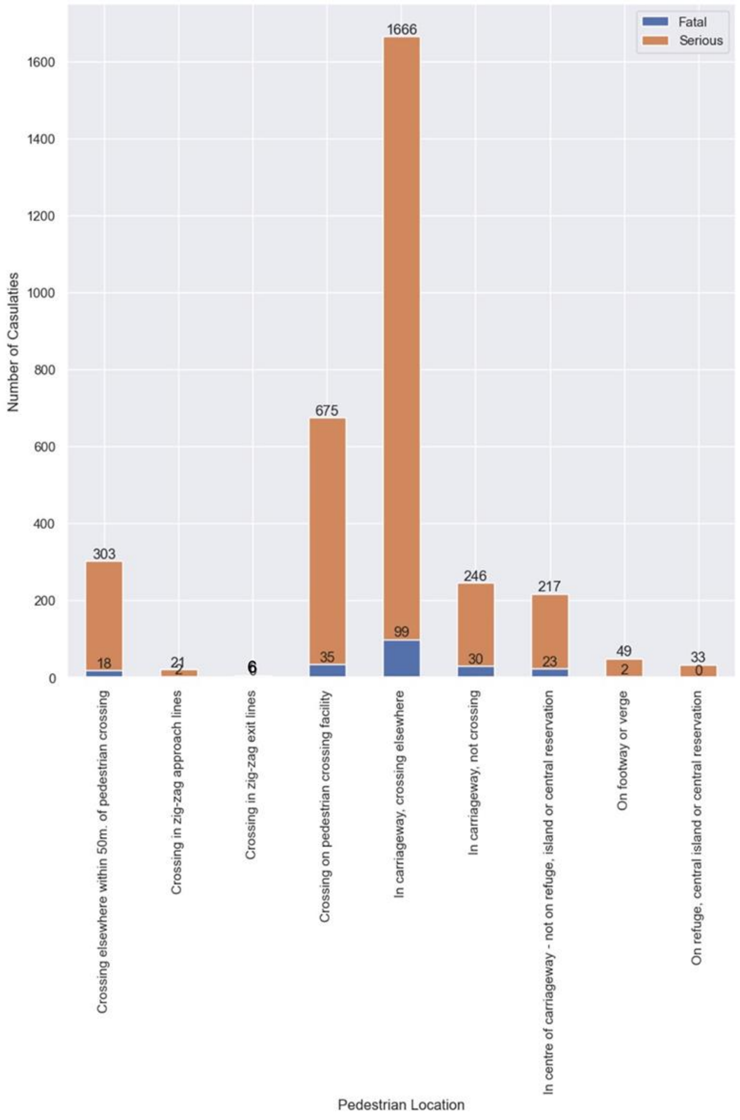
Casualty Imd Decile	Birmingham	Cornwall	County Durham	Doncaster	Huntingdonshire	Leeds	Northumberland	Shropshire	South Hams	Wiltshire
Least deprived 10%	12	7	5	7	55	39	13	19	52	27
Less deprived 10-20%	15	10	14	17	67	22	37	13	8	65
Less deprived 20-30%	44	17	27	18	33	41	14	23	105	46
Less deprived 30-40%	27	37	31	6	23	41	32	41	107	45
Less deprived 40-50%	14	39	41	30	13	29	44	43	10	53
More deprived 10-20%	76	48	54	51	45	48	25	19	2	29
More deprived 20-30%	28	62	52	63	2	53	45	30	57	13
More deprived 30-40%	29	133	34	37	108	40	46	50	10	16
More deprived 40-50%	26	73	35	38	16	55	39	51	215	40
Most deprived 10%	165	33	71	71	26	192	14	11	161	15

Figure 9: Casualties by IMD decile for the top ten districts

Casualties for the top 10 districts, broken down by IMD decile, as seen in Figure 9, reveals that 40-50% deprived IMD decile under South Hams had the most number of casualties. Leeds had the largest number of casualties among the IMD decile categories with the greatest levels of deprivation. Darker hues common among more destitute deciles indicate that the more deprived the decile, the higher the number of casualties.

Patterns in pedestrians who were KSI casualties

Here analysis was done on pedestrians (sample size 3216) who were met with an accident that were killed or seriously injured. From the data, killed and seriously injured are represented as Fatal and Serious respectively. Figure 10 shows the analysis of pedestrian location and the distribution of Fatal and Serious casualties, and it is evident that most pedestrians were in the carriageway, crossing elsewhere (1666 Serious and 99 Fatal). Less number of people met with an accident when pedestrians were crossing in zig-zag lines (6 injured). Overall, there is a general trend of Fatalities being less when compared to serious cases for every pedestrian location. Another analysis was done based on the pedestrian physical facilities (Figure 11) and it can be seen from the analysis that most of the pedestrian casualties occurred since there were no physical crossing facilities within 50 meters. If there was a footbridge or subway the number of pedestrians who were KSI are almost negligible.



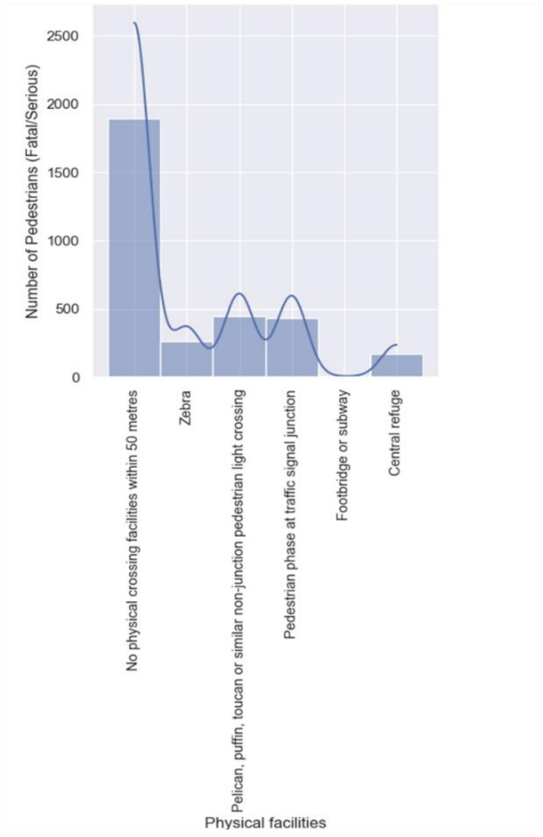


Figure 10: Pedestrian location vs Number of casualties.

Figure 11: Physical facilities vs Number of Pedestrians

,

Another pie plot was done based on pedestrian movement who were KSI as shown in Figure 12. It is evident again from the plot that the number of fatalities (3007 serious cases and around 209 cases fatal) from an accident is less and both the killed and seriously injured pedestrians are dominated by people crossing from the driver’s nearside (1410 serious and 96 fatal). The least number of KSI cases are there for pedestrian movement ‘In carriageway, stationary - not crossing (standing or playing) - masked by parked or stationary vehicle’ (27 Serious and only 1 Fatal case). Another important point to be noted from the plot is that both serious and fatal cases are dominated by people crossing from the driver’s offside and nearside and hence if people become more conscious while crossing most of the pedestrian accidents can be avoided.

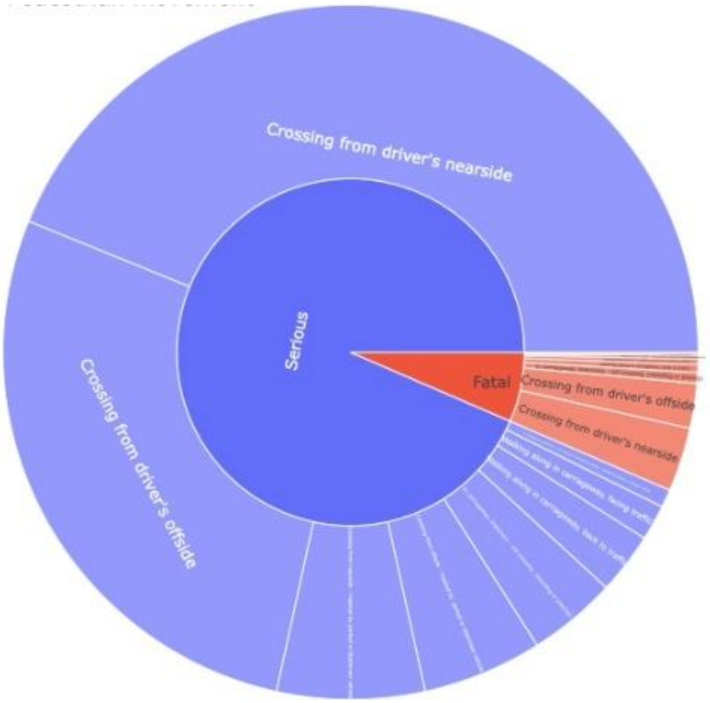


Figure 12: Pie plot of pedestrian movement with classification of fatal or serious cases.

CONCLUSION

The analysis carefully examined the data quality by verifying its completeness and correctness. Found the patterns, distribution and cardinality of data during data characterisation and used the relevant indexes to perform the analysis. Unwanted observations such as negative values, unknowns, and irrelevant columns were removed to aid the detailed analysis. Males were more likely than females to be involved in accidents, according to analysis based on demographics of casualties, and there were instances where the gender of the victims was left undocumented. Compared to young and middle-aged adults, children under the age of 15 and seniors beyond the age of 66 were safe. Based on residential area type, male residents of large cities came out on top in the analysis. When comparing the patterns between local authorities in incidents that resulted in fatalities or seriously injured victims, South Hams came out on top despite Birmingham recording the highest number of accidents. It indicates there are more KSI fatalities in sparse areas. The majority of pedestrian deaths occurred because there were no physical facilities within 50 metres and because people crossed the street carelessly, often on the driver's side or offside. According to the data, pedestrians can avoid accidents by choosing safe crossing methods like footbridges, subways, and zebra crossings.