GROUP MEMBERS

Ameena Farooq (201666061), Neil Abraham (201588844), Rui Ci Chong (201675197), Tania John (201677655).

INTRODUCTION

The dataset mentioned below was analysed and certain conclusions were reached. The aim of this report is to present our findings and give explanations on how we reached to these conclusions.

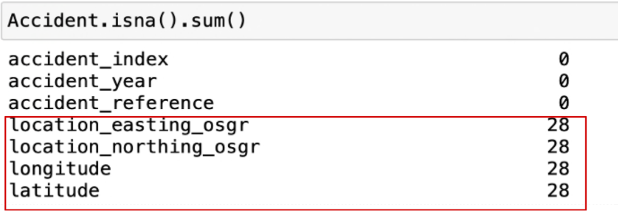
The data files that were used for analysis are Road Safety Data – Accidents 2019, Road Safety Data – Casualties 2019 and Road Safety Open Dataset Data Guide. The links for data files used are Road Safety Data - Accidents 2019, Road Safety Data - Casualties 2019, Road Safety Open Dataset Data Guide.

DATA QUALITY

Below are the data quality issues we have spotted from Accident and Casualty datasets and each issue are followed by the solutions we have made.

Completeness (Missing Values)

- There are null values in 4 columns from the Accident dataset where Casualty dataset does not have any null values.
- The columns with null values are 'location_easting_osgr', 'location_northing_osgr', 'longitude' and 'latitude'. Each column has 28 null values.
- We removed the null values of each column in Accident dataset to solve this issue.



```
Accident.isna().sum()
```

```
accident_index                    0
accident_year                     0
accident_reference                0
location_easting_osgr            28
location_northing_osgr           28
longitude                        28
latitude                         28
```

Figure 1. Null Values in the Accident dataset

Accuracy (Numeric Outlier)

- There are outliers in the Casualty dataset.
- The outliers are in the column 'casualty_reference'. There are 3 outliers, and they are 111, 256 and 991.
- Casualty_reference is a unique number representing each casualty that happened in a particular accident.
- We replaced the outliers to solve this problem. That is 'casualty_reference' was changed to a number that corresponds to the number of casualties (coloumn name:number_of_casualties) that was caused in that certain accident (accident_index was used as reference).

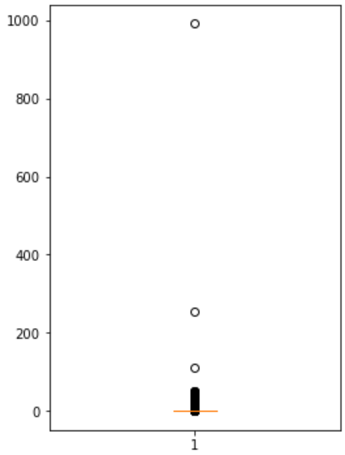
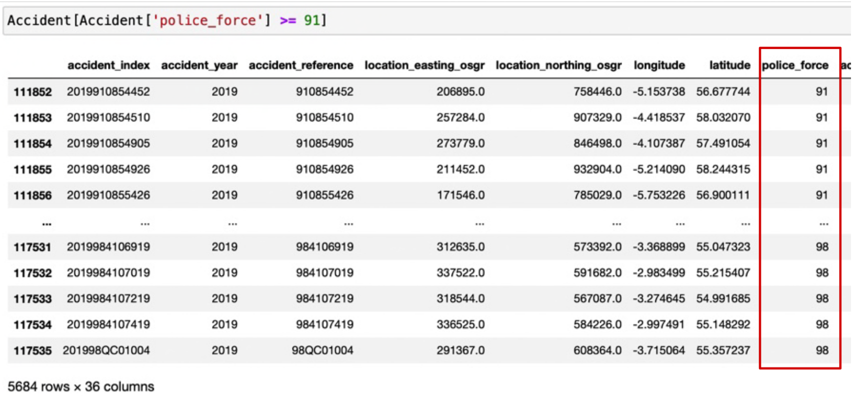


Figure 2. Boxplot for casualty_reference

Accuracy (Wrong Data Format)

- In 2019, 8 categories were discontinued, and 1 category was added for the variable ‘police_force’.
- The variable ‘police_force’ from Accident dataset has introduced a new category with code 99 to group the categories with code 91 to 98 and this should be updated in 2019 data.
- There are 5684 rows indicated by code 91-98 in the ‘police_force’ column.
- To make the data follow the updated format, we replaced code 99 to rows with code 91 to 98.



```
Accident[Accident['police_force'] >= 91]
```

	accident_index	accident_year	accident_reference	location_easting_osgr	location_northing_osgr	longitude	latitude	police_force	ac
111852	2019910854452	2019	910854452	206895.0	758446.0	-5.153738	56.677744	91	
111853	2019910854510	2019	910854510	257284.0	907329.0	-4.418537	58.032070	91	
111854	2019910854905	2019	910854905	273779.0	846498.0	-4.107387	57.491054	91	
111855	2019910854926	2019	910854926	211452.0	932904.0	-5.214090	58.244315	91	
111856	2019910855426	2019	910855426	171546.0	785029.0	-5.753226	56.900111	91	
...	...	...	...	...	...	...	...	...	
117531	2019984106919	2019	984106919	312635.0	573392.0	-3.368899	55.047323	98	
117532	2019984107019	2019	984107019	337522.0	591682.0	-2.983499	55.215407	98	
117533	2019984107219	2019	984107219	318544.0	567087.0	-3.274645	54.991685	98	
117534	2019984107419	2019	984107419	336525.0	584226.0	-2.997491	55.148292	98	
117535	201998QC01004	2019	98QC01004	291367.0	608364.0	-3.715064	55.357237	98	

5684 rows × 36 columns

Figure 3. Discontinued Categories in column ‘police_force’

Consistency (Different Data Format)

- To represent the ‘unknown’ values 9 is used as label in various fields of the datasasets- Accident and Casualty. But, in some other fields 9 represents something else.
- Also, there are other values that are used to represent the unknown fields.
- For example, in Accident dataset ‘unknown(self-reported)’ in the speed_limit and junction_detail is represented by 99. And in Casualty dataset the unknown is represented by -1 for the columns ‘first_road_number’ and ‘second_road_number’.
- To make the format more consistent, we replaced labels with 'unknown' with –2, this way uniformity is maintained across the whole dataset.
- By making the format consistent, this has also solved the problem for label -1 only indicates ‘Data missing or out of range’.

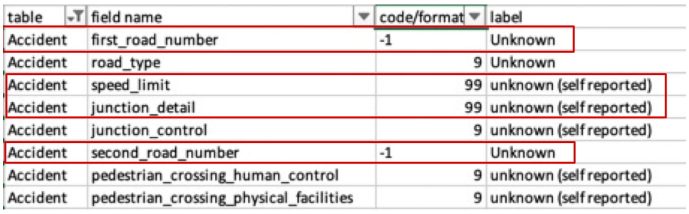


table	field name	code/format	label
Accident	first_road_number	-1	Unknown
Accident	road_type	9	Unknown
Accident	speed_limit	99	unknown (self reported)
Accident	junction_detail	99	unknown (self reported)
Accident	junction_control	9	unknown (self reported)
Accident	second_road_number	-1	Unknown
Accident	pedestrian_crossing_human_control	9	unknown (self reported)
Accident	pedestrian_crossing_physical_facilities	9	unknown (self reported)

Figure 4.1 Accident Data Guide: ‘Unknown’ denoted by code -1 and 99

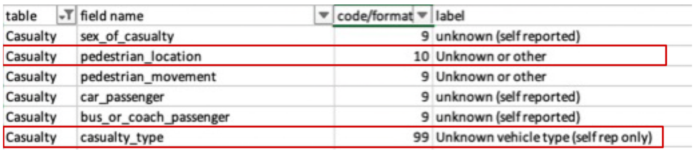


table	field name	code/format	label
Casualty	sex_of_casualty	9	unknown (self reported)
Casualty	pedestrian_location	10	Unknown or other
Casualty	pedestrian_movement	9	Unknown or other
Casualty	car_passenger	9	unknown (self reported)
Casualty	bus_or_coach_passenger	9	unknown (self reported)
Casualty	casualty_type	99	Unknown vehicle type (self rep only)

Figure 4.2 Casualty Data Guide: ‘Unknown’ denoted by code 10 and 99

Consistency

- The ‘accident_index’ column in both Accident and Casualty datasets has an issue of inconsistency.
- Some of the indices are string type but some of them are integers.

DATA CHARACTERISATION

Accident Dataset

Variable Name	Type	Description
accident_index	Nominal	Unique value for each accident. 'accident_year' and 'accident_ref' is combined to form this unique ID. It can be used to join to Vehicle and Casualty tables
accident_year	Numerical	The year in which the accident took place
accident_reference	Nominal	For a particular year, the unique id used by the police to reference a collision
location_easting_osgr	Numerical	Easting coordinates of the accident location
location_northing_osgr	Numerical	Northing coordinates of the accident location
longitude	Numerical	Longitude of the accident location
latitude	Numerical	Latitude of the accident location
police_force	Nominal	Police force name of the force that has jurisdiction in the acciddent location
accident_severity	Ordinal	Severity of an accident (1-Fatal, 2-Serious, 3-Slight)
number_of_vehicles	Numerical	Number of vehicles recorded in the accident
number_of_casualties	Numerical	Number of casualties recorded in the accident
date	Nominal	Date of the accident
day_of_week	Nominal	Day of the accident
time	Nominal	Time of the accident
local_authority_district	Nominal	Local authority in whose area the accident occurred
local_authority_ons_district	Nominal	ONS district of local authority in whose area the accident occurred
local_authority_highway	Nominal	Local authority responsible for the highway where the accident occurred
first_road_class	Nominal	Class of the first road where the accident occurred
first_road_number	Numerical	First road number
road_type	Nominal	Type of road
speed_limit	Numerical	Speed limit of the road where the accident occurred
junction_detail	Nominal	Type of junction where the accident occurred
junction_control	Nominal	Junction control that was present at the junction
second_road_class	Nominal	Class of the second road where the accident occurred
second_road_number	Numerical	Second road number
pedestrian_crossing_human_control	Nominal	Type of pedestrian crossing human control

pedestrian_crossing_physical_facilities	Nominal	Type of pedestrian crossing physical facility
light_conditions	Nominal	Light conditions of the accident location
weather_conditions	Nominal	Weather conditions at the time and location of the accident
road_surface_conditions	Nominal	Road surface condition at the time of the accident
special_conditions_at_site	Nominal	Any Special conditions on the road at the time of the accident
carriageway_hazards	Nominal	Any carriageway hazards on the road at the time of the accident
urban_or_rural_area	Nominal	Indicates whether the accident occurred in an urban or rural area
did_police_officer_attend_scene_of_accident	Nominal	Whether a police officer attend the scene of accident
trunk_road_flag	Nominal	Indicates whether the accident happened on a trunk road
lsoa_of_accident_location	Nominal	Lower Layer Super Output Area of the accident location

Casuality Dataset

Variable Name	**Type**	**Description**
accident_index	Nominal	Unique value for each accident. 'accident_year' and 'accident_ref' is combined to form this unique ID. It can be used to join to Vehicle and Casualty tables
accident_year	Numerical	The year of the accident
accident_reference	Nominal	In a particular year, the id used by the police to reference a collision
vehicle_reference	Nominal	Unique value for each vehicle in a single accident
casualty_reference	Nominal	Unique value for each casualty in a single accident
casualty_class	Nominal	Type of casuality - Driver/Passenger/Pedestrian
sex_of_casualty	Nominal	Gender of casuality
age_of_casualty	Numerical	Age of casuality
age_band_of_casualty	Ordinal	Age band of the casuality
casualty_severity	Ordinal	Severity of the casuality (1-Fatal, 2-Serious, 3-Slight)
pedestrian_location	Nominal	Location of the pedestrian during the accident
pedestrian_movement	Nominal	Movement of the pedestrian at the time of the accident
car_passenger	Nominal	Position of the passenger inside the car
bus_or_coach_passenger	Nominal	Position of the passenger inside the bus
pedestrian_road_maintenance_worker	Nominal	Whether the casualty was a pedestrian road maintenance worker
casualty_type	Nominal	Type of the casuality
casualty_home_area_type	Nominal	Home area type of the casuality
casualty_imd_decile	Ordinal	IMD decile of the casuality

DETAILED ANALYSIS

I. Patterns in the Demographics of Casualties.

1. Comparison of Age and Casualties

We use the age band and the casualty type to analyse the total casualties and their different types. We then restricted the top 5 types to analyse the likelihood of those casualties. The data is split into 11 age bands and 31 casualty types. We have listed the age ranges in the bands and the top 5 casualties. Figure 2 denotes that there is a steep increase in casualties from the ages of 16-35 and the most common casualty occurs when they are an occupant in a car. The second most common is as a pedestrian.

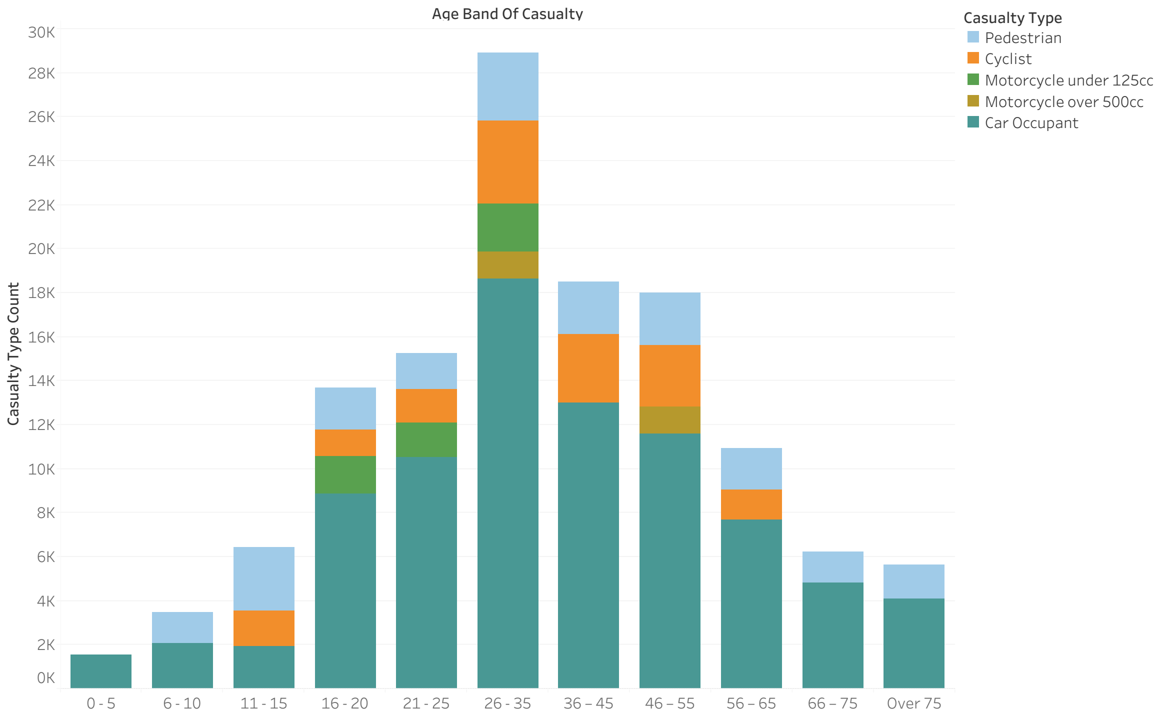


Figure 5: Visualising the Age Groups with respect to the Casualty Type

2. Probability of a gender meeting with a casualty

The likelihood of a male going through a casualty is much higher than that of a female. In 2019, the number of casualties met by males were roughly 50% higher than the total number of casualties where a female was involved.

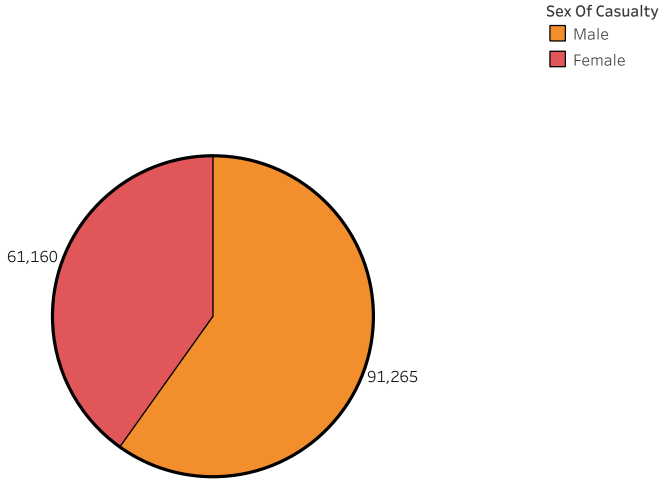


Figure 6: Visualising the males and females who met with a casualty in 2019.

II. Patterns between the local authorities and KSI Casualties.

1. The effect of road type in KSI casualties

We analysed the top 5 districts having the most KSI casualties. All the districts showed that the highest number of casualties occurred in single carriageways. The lowest number of casualties were observed in Slip roads.

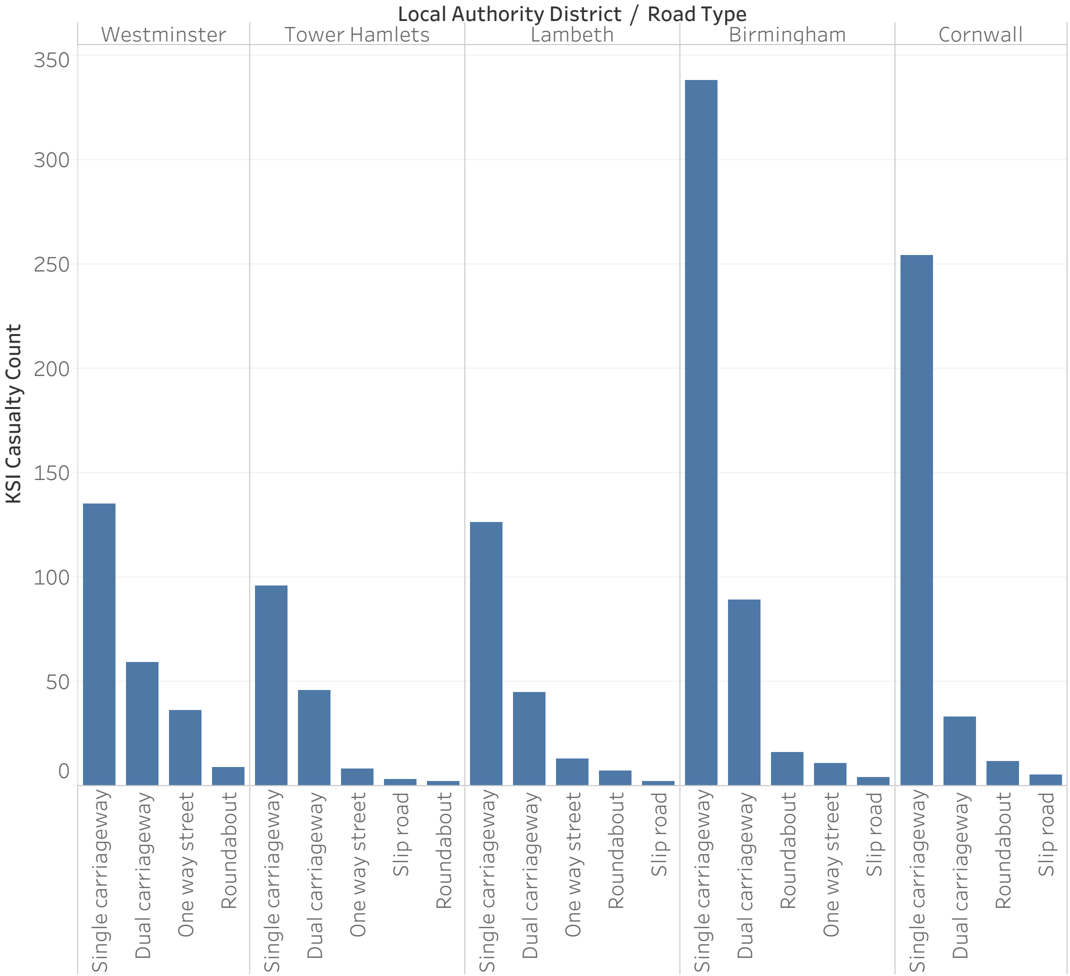


Figure 7: Plotting the Local Authority District and the casualties by Road Type.

2. Effect of junction control in casualties

Out of the top 5 districts with most KSI casualties, it is evident pattern that the highest number of casualties are caused by the Give way or uncontrolled followed by Auto Traffic Signals. And the least casualty is caused when there is an Authorised person to control the traffic at junctions.

It can be observed from the graph below that maximum number of casualties has taken place in the city of Birmingham.

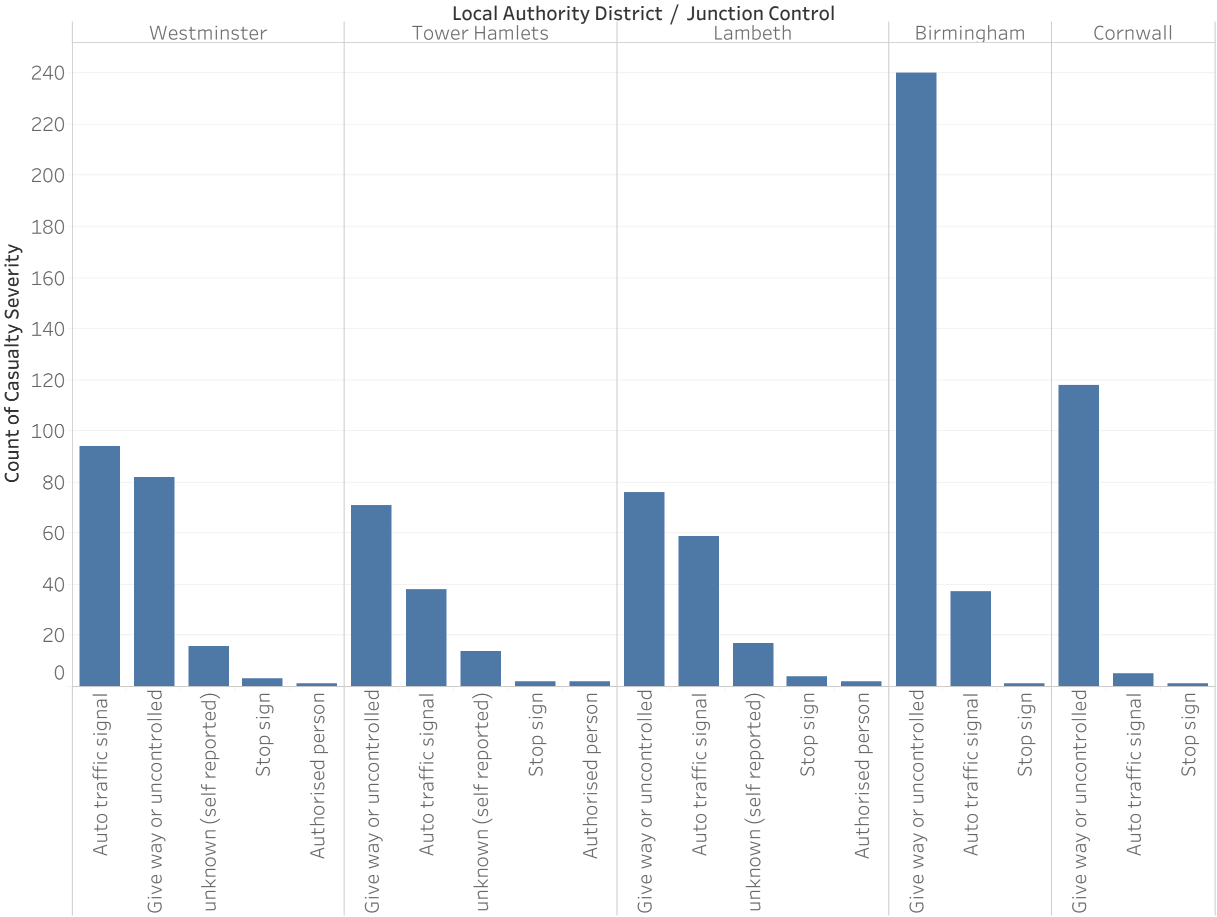


Figure 8: Plotting the casualties by district and the type of junction control.

III. Pedestrians who were KSI Casualties

1. Pedestrian Casualties and Speed Limits

As the speed limit increases, the number of casualty decreases. This might be due to increased awareness when there are vehicles around you at a faster speed. When the speed is lower, pedestrians might be less careful which leads to accidents.

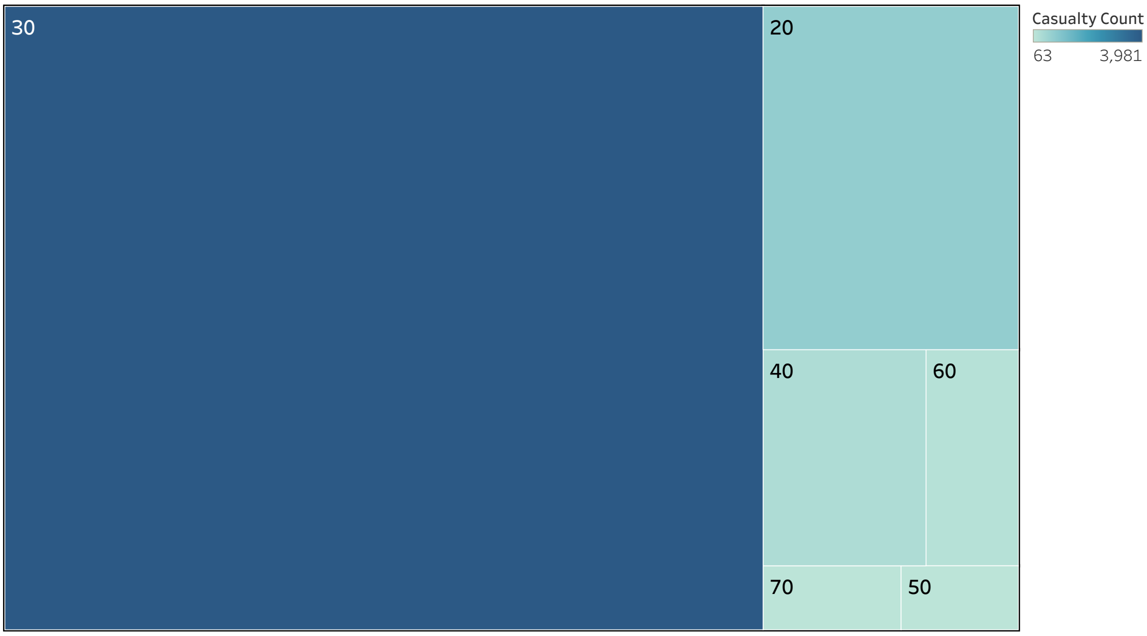


Figure 9: Using a treemap to determine pedestrian KSI casualties and the speed limit at the location.

2. Casualties with respect to their positions

Pedestrians are most likely to meet with a KSI casualty if they are in a carriageway, crossing elsewhere followed by Crossing on pedestrian crossing facility. The least number of KSI casualties were due to the stop sign at the junction.

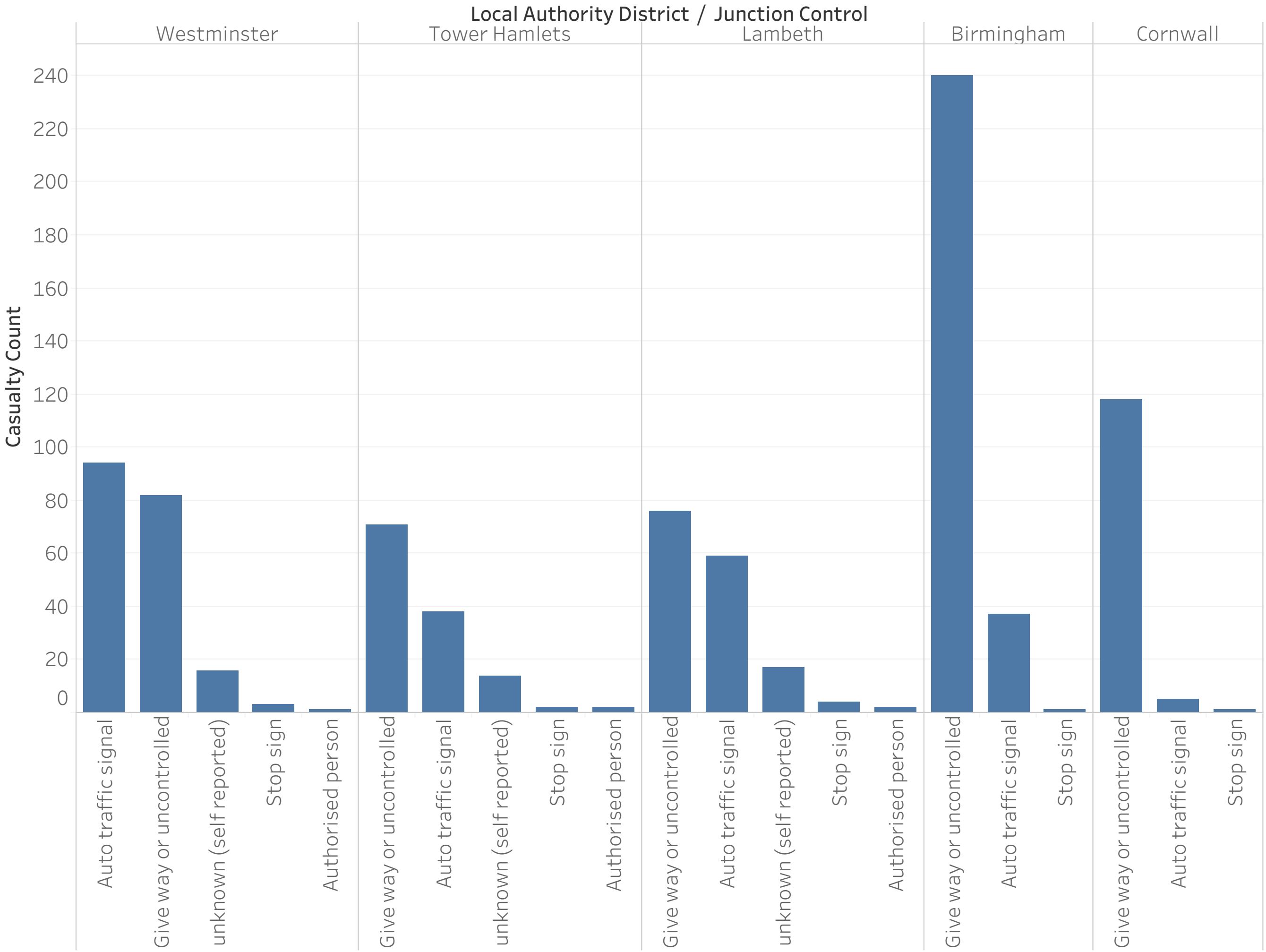


Figure 10: Bar graph representing count of casualty with respect to the location of the pedestrian.

CONCLUSION

- After initial analysis data quality issues such as missing values, outliers, etc were identified.
- The data quality issues were rectified before doing further analysis
- After analysis of data, we reached the following conclusions:
  a) People of the age group 16-35 are the most susceptible to casualties.
  b) More men are prone to accidents when compared to women.
  c) There is a higher chance of accidents to occur in single carriageway when compared to other road types.
  d) Give way or uncontrolled junctions are likely to have more accidents.
  e) The speed limit and pedestrian KSI casualties are inversely proportional to each other.
  f) Most of the accidents takes place in carriageway, crossing elsewhere.
- Thus, making reforms after studying the conclusions can help in reduction of accidents and casualties.