Group members

Deepak Bisht mm22db@leeds.ac.uk
Nilesh Kumar mm22nk@leeds.ac.uk
Julian Bara-Mason mm22jdbm@leeds.ac.uk
Sohail Suleman Zaveri mm22ssz@leeds.ac.uk

Introduction

The aim of this coursework is to communicate findings from the analysis of Road Safety Data for England in 2019. The report has three objectives it explores in terms of casualties, including Killed or Seriously Injured Casualties (KSI):

1. Patterns in the demographics of casualties
2. Patterns between local authorities of KSI accidents
3. Patterns in KSI pedestrians

The analysis works with two datasets:

1. Road Safety 2019 Casualty Data: dft-road-casualty-statistics-casualties-2019
2. Road Safety 2019 Accidents Data: dft-road-casualty-statistics-accidents-2019

Data Quality

We have encountered several data quality issues ranging from missing values to inconsistent data types in both the datasets.

1. **Unknown values:** For accident data frame we have 13 columns with unknown values. The majority of these unknown values are in *junction_control, second_road_class* and *second_road_numbe*r columns. For the casualty data frame there are 10 columns with unknowns where the majority of them can be found in *casualty_home_area_type* and *casualt_imd_decile* columns. These columns contain important information that could be useful in analysing demographics of casualties.

	column	count
0	sex_of_casualty	724
1	age_of_casualty	3255
2	age_band_of_casualty	3255
3	pedestrian_location	1
4	car_passenger	382
5	bus_or_coach_passenger	61
6	pedestrian_road_maintenance_worker	74
7	casualty_type	5
8	casualty_home_area_type	15355
9	casualty_imd_decile	15372

	column	count
0	speed_limit	80
1	junction_detail	1
2	junction_control	49366
3	second_road_class	49106
4	second_road_number	49106
5	pedestrian_crossing_human_control	161
6	pedestrian_crossing_physical_facilities	148
7	light_conditions	1
8	road_surface_conditions	299
9	special_conditions_at_site	240
10	carriageway_hazards	230
11	trunk_road_flag	10001
12	lsoa_of_accident_location	2848

Fig 1 - Counts of unknown values for respective columns, left - casualty dataframe, right - accident dataframe

2. **Missing values:** We found 25 rows with *NA* values for latitude and longitude columns in the accident data frame. There were no missing values in the casualty data frame.

```
location_easting_osgr                          25
location_northing_osgr                         25
longitude                                      25
latitude                                       25
```

Fig 2 - Columns with NaN values.

3. **Different data format:** For both the data frames all the categorical columns contained numeric values. The mappings between these numeric values and the actual categorical values are provided in a separate meta-data file. Although this approach is efficient to store and process data, for analytics it ends up being difficult to find meaningful insights. For *date* column the data type is object instead of datetime, which is a case of wrong data format quality issue.

```
df_accident["accident_severity"].value_counts()
```

```
3    92898
2    22980
1     1658
Name: accident_severity, dtype: int64
```

```
df_accident["accident_severity"].value_counts()
```

```
Slight     92898
Serious    22980
Fatal       1658
Name: accident_severity, dtype: int64
```

Fig 3 - 'accident_severity' column before and after conversion'

4. **Inconsistent data types:** Moreover, while combining accident and casualty dataframes based on *accident_index*, we found that not all *accident_index* values matched. After converting the datatypes of accident_index in both the dataframes to string, all the *accident_index* values matched and the data frames were combined properly.

Data Characterisation

For the accident dataframe there are 36 columns, 107535 rows and for casualty dataframe there are 18 columns, 153158 rows. The variables used for our analysis are characterised below:

Casualty dataframe:

Variable name	**Data type**	**Description**	**Characteristics**
accident_index	String	Unique Identifier of an accident. Concatenation of accident year and accident reference. Eg value - *2019010128300*.	Nominal - Multiple rows can have the same accident index that refer to the same accident. There are 117536 unique values of accident_index in 153158 rows.
casualty_class	String	Indicates if the casualty was **Driver or rider**, **Pedestrian** or **Passenger**.	Nominal - 64.8% of the values in this column are **Driver or rider,** 21% are **Passenger** and 14.2% are **Pedestrian**.
sex_of_casualty	String	Sex of casualty.	Nominal - Can be either **Male** or **Female.**

casualty_severity	String	Severity of casualty - **Fatal**, **Serious** or **Slight**.	Ordinal - 82.28% of the casualties are Slight, 16.56% Serious and 11.43% Fatal.
casualty_type	String	Mode of transport of casualty before an accident. Can also take the value **Pedestrian**.	Nominal - There are 21 unique values. Example values are car occupant, pedestrian, motorcycle, horse rider etc.
casualty_home_area_type	String	Indicates if casualty is from a **Small town**, **Urban** or **Rural** area.	Ordinal - Majority of casualties i.e., 80.66% are from urban area, 10.76% from rural area and 8.57% from small town
casualty_imd_decile	String	A measure of poverty used to group individual areas into one of ten groups of equal frequency	Nominal - Can fall in one of 10 ranges: Most deprived 10%, More deprived 20%-30%, More deprived 30%-40%, More Deprived 40%-50%, Less Deprived 40%-50%, Less Deprived 30%-40%, 20%-30%, Less Deprived 10%-20%, Least Deprived 10%

Accident dataframe:

Variable Name	**Data Type**	**Description**	**Characteristics**
weather_conditions	String	The weather at the time of the accident	Nominal
pedestrian_movement	String	The pedestrian's movement at the moment of the accident.	Nominal - Describes the pedestrian's movement, e.g., *Crossing from driver's nearside*
pedestrian_physical_crossing_ftc	String	Pedestrian Physical Crossing Facility that helps pedestrians cross the road	Nominal - The type of pedestrian crossing facility available at location of accident, e.g., *Pelican, Puffin, toucan, or similar non-junction pedestrian light crossing*
light_conditions	String	Indicates lighting conditions.	Nominal - Describes the lighting situation at the site of an accident. Whether it is *daylight* or *dark*.
road_surface_conditions	String	Indicates the road surface condition - **snow, wet, flood.**	Nominal - Defines the condition of the road surface. Whether the surface is covered with *snow*, *dry* or *wet*.
local_authority_district	String	The name of the local authority.	Nominal - It defines the name of the local authority of that site of accident. e.g., *Leeds, Birmingham, etc.*
road_type	String	The type of road - **highway, street.**	Nominal - It describes the type of road on which an accident happened.

junction_control	String	Indicates that the junction was controlled by any means. **Controlled, Uncontrolled.**	Nominal - Describes the presence of any kind of junction control near the site of accident.
did_police_officer_attend_scene	String	The police visited the place of the accident - **yes, no.**	Nominal - Describes the situation, whether the police visited the site of the accident.

Analysing Demographics of Casualties

We found the following patterns while analysing demographics of casualties:

1. Sex ratio of casualties based on casualty class:
   We analysed sex ratios of casualties for each casualty class and observed that
   - 67.2 % of driver or rider casualties are males.
   - 59.9 % of passenger casualties are females.
   - Number of male and female pedestrian casualties are almost equal.

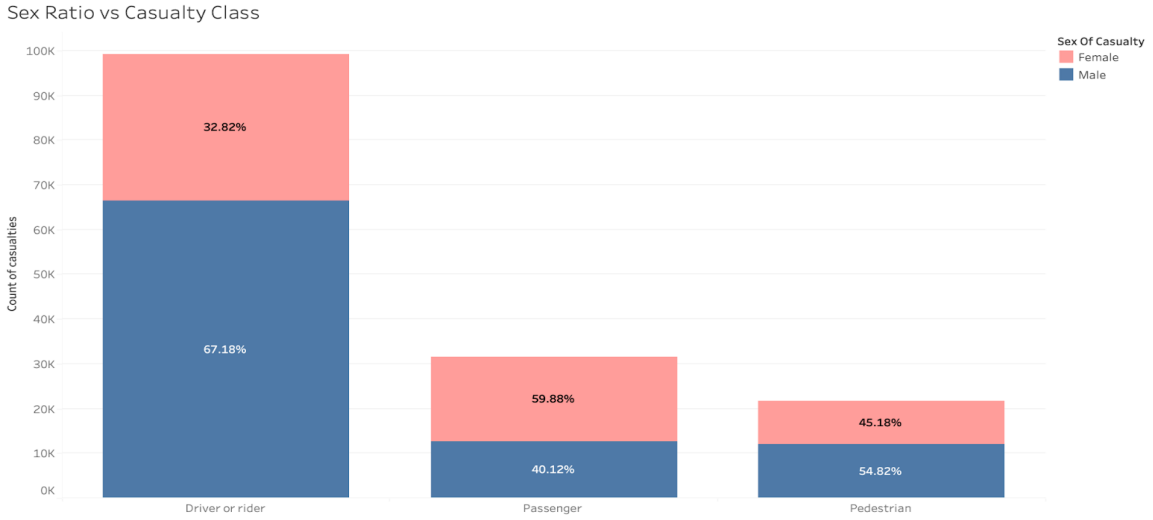

Fig 4 - Comparison of sex ratio based on casualty class.

2. Percentage of pedestrian casualties is higher in urban areas than in small towns or rural areas. 15% of urban casualties are pedestrians whereas for small towns its 9.45% and for rural its 7.18%.

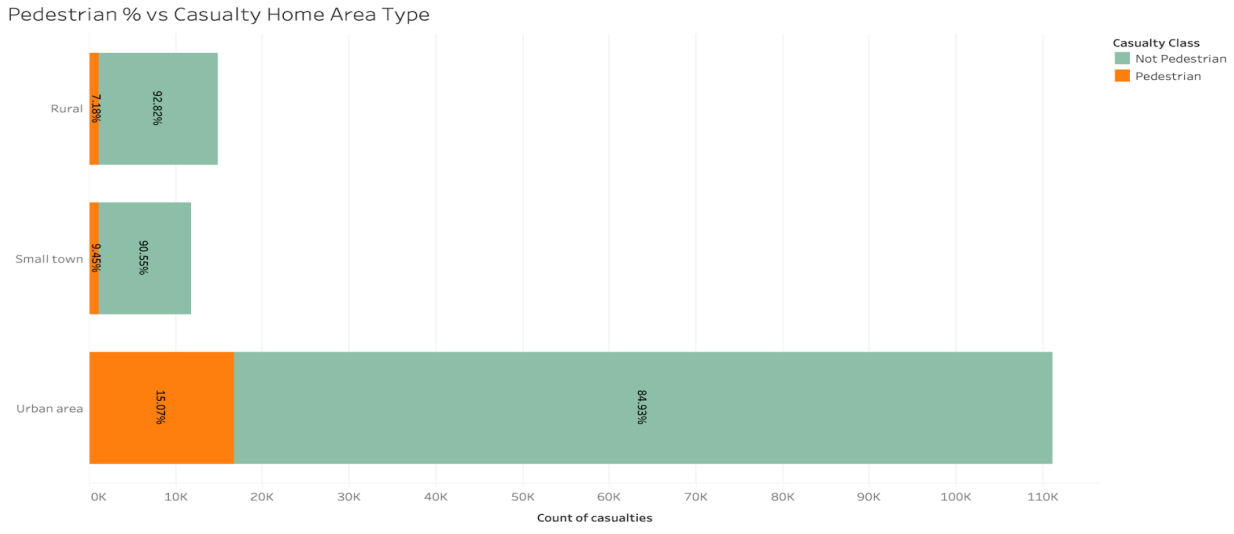

Fig 5 - Comparison of pedestrian percentages in different casualty home area type.

3. Different sex ratios are observed in different casualty types.
    - Most of the motorcycle (90.91%) and cyclist (80.16%) casualties are males.
    - 90.48% of horse riders and 62.02% of public transport casualties are females.
    - Car occupant casualties have almost equal numbers of males and females - Male(49.83%), Female(50.17).

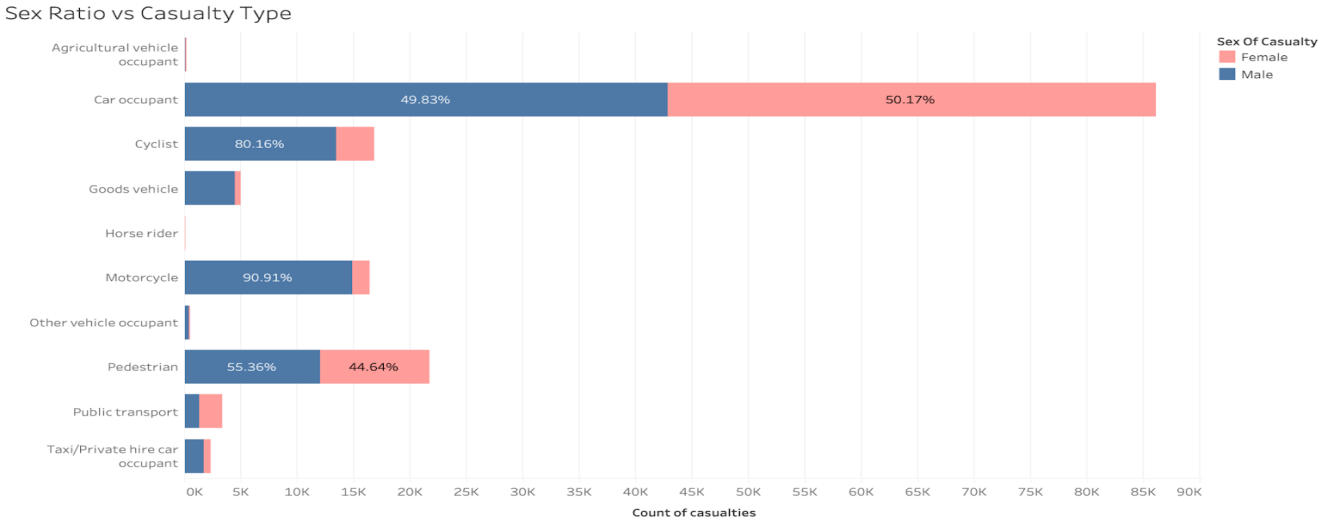

Fig 6 - Comparison of sex ratio and casualty types.

Analysing patterns between local authorities

We found the following patterns between the local authorities for killed or seriously injured casualties(KSI):

1. Accidents per day are higher on weekends than on weekdays.
    - For the majority of the local authorities, weekends saw about 41% of the accidents.
    - On a daily average, 11.8% of the accidents took place on a weekday, whereas 20.5% took place on a Saturday or Sunday.

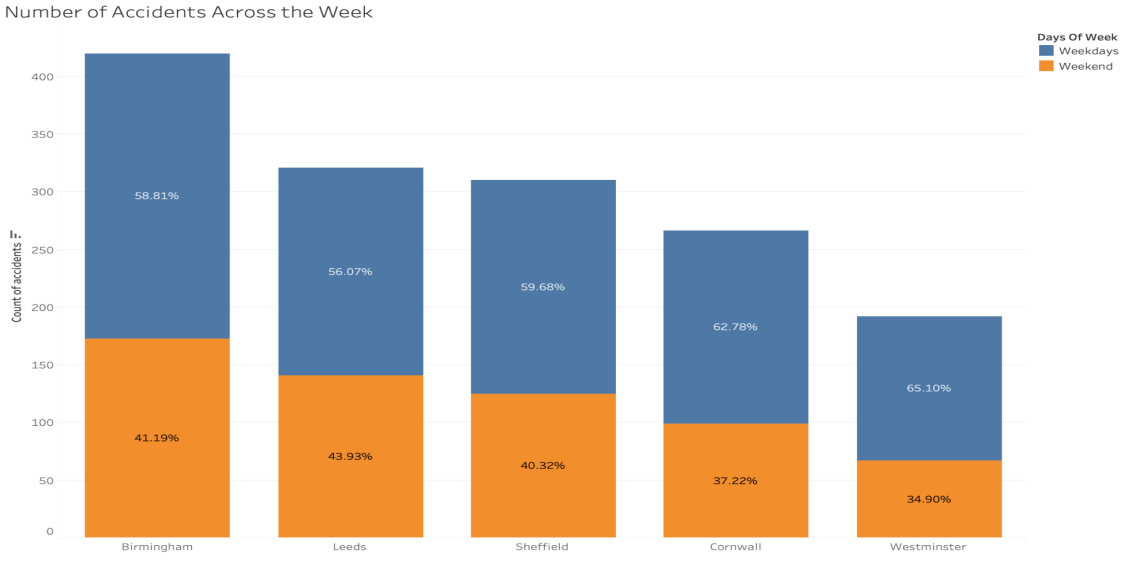

Fig 7 - Number of accidents across the week for top five local authorities.

2. More accidents on the highways when it's dry.
    - Highways account for over 92% of accidents. 62% of accidents occur even in dry conditions, where the road is ideal.
    - Accidents occur approximately twice as frequently in dry conditions as in wet ones.

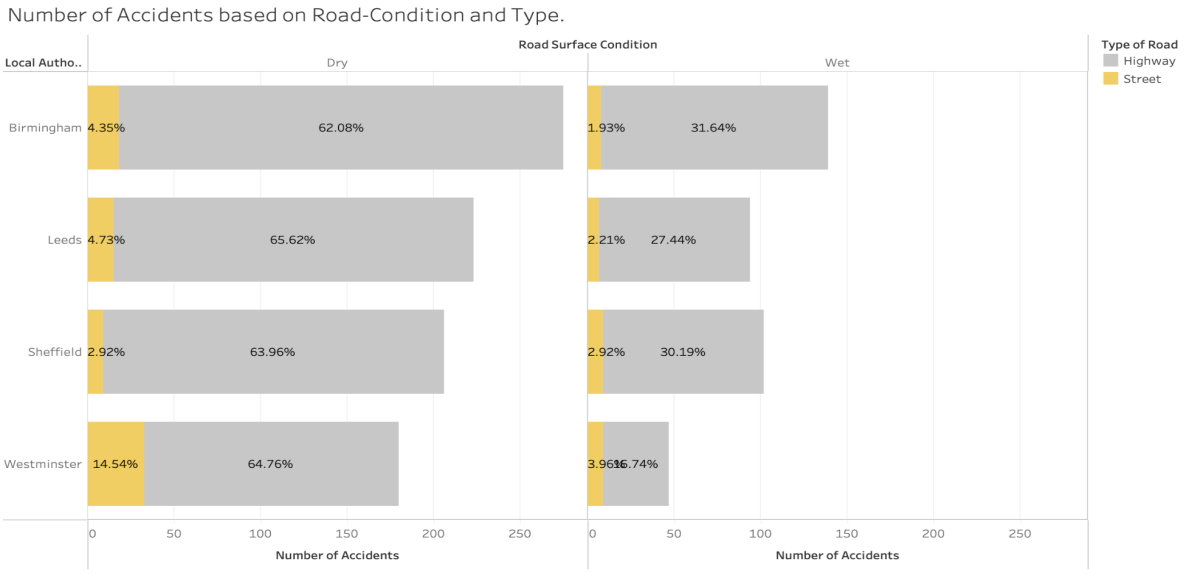

Fig 8 - Comparison of number of accidents based on road type and conditions.

3. Number of accidents based on the junction control.
    - For the majority of local authorities, uncontrolled junctions are where the majority of accidents occur.
    - Police typically showed up at the scene of the accident. In Westminster, police were always present, compared to 57% of the time in Birmingham when police arrived at an accident.

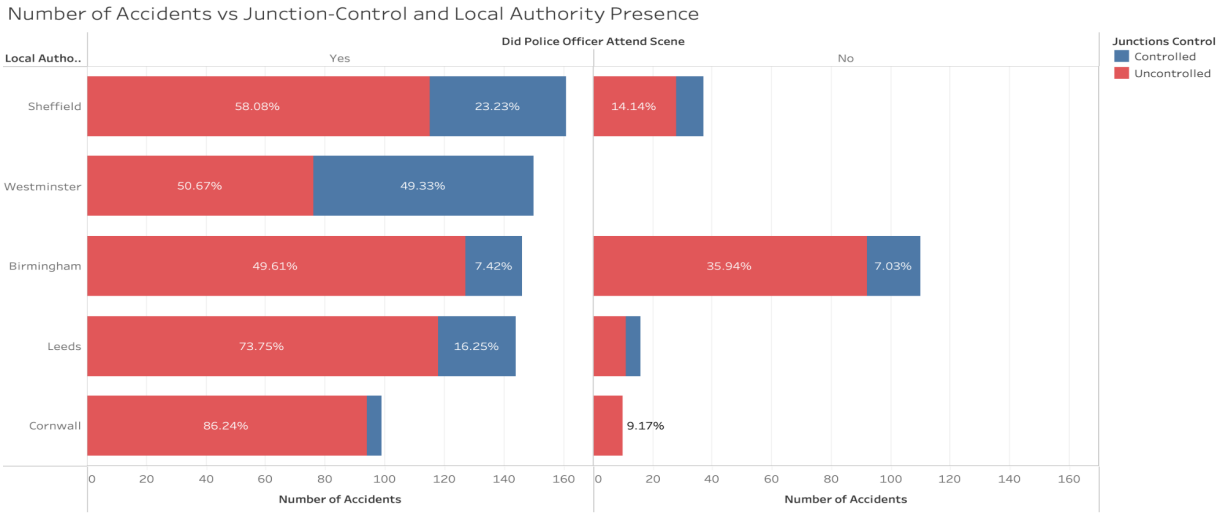

Fig 9 - Comparison of number of accidents based on junction-control.

Analysing patterns in Pedestrians who were KSI Casualties

Please note that all analysis and visualisations do not include variables that were "unknown" or "other", so some sums may not = 100%.

There were 6,200 killed or seriously injured Pedestrians in England in 2019. Of the total Pedestrian KSI casualties, 92% of them were seriously injured while only 8% were killed.

Fig 10

Casualty Severity Distribution by Age

Casualty Severity	Age Of Casualty (bin) 10	20	30	40	50	60	70	80	90
Fatal	0%	1%	1%	1%	1%	1%	2%	2%	1%
Serious	6%	8%	9%	10%	14%	12%	16%	12%	2%

The Casualty Severity Distribution by Age table above shows the Casualty Severity Distribution by Age band. It is seen that most KSI casualties happen to pedestrians that are between 50 years old and 80 years old. Half (4%) of the Fatal Casualties are between 70-80 years old.

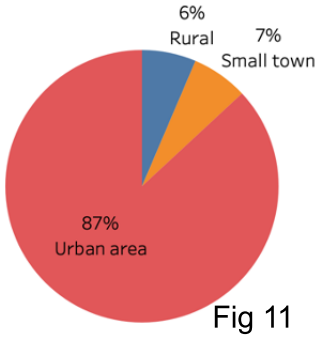

Fig 11

Figure (11) to the left is the Casualty Home Type Area ratio. 87% of Pedestrian Casualties are from an Urban Area home. 6% and 7% live in Rural and Small Town's respectively.

Figure (12) to the right is the ratio of the More to Most Deprived vs The Least-Less Deprived Casualty IMD Decile.

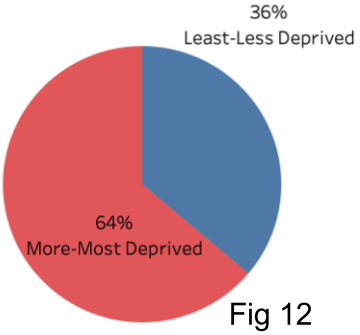

Fig 12

The Accidents by Month and Weather Conditions chart below shows the percentage of casualties each month by the weather conditions. November has the most pedestrian casualties at 10%. There is a steady rise in Raining weather conditions from Autumn (September-November). Most pedestrian accidents are in Fine weather conditions, with no rain or high winds.

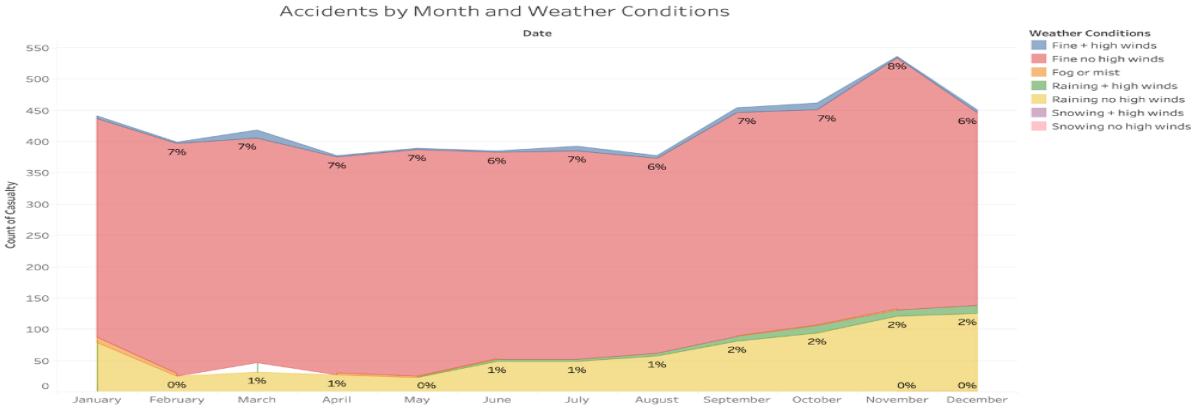

Fig 13

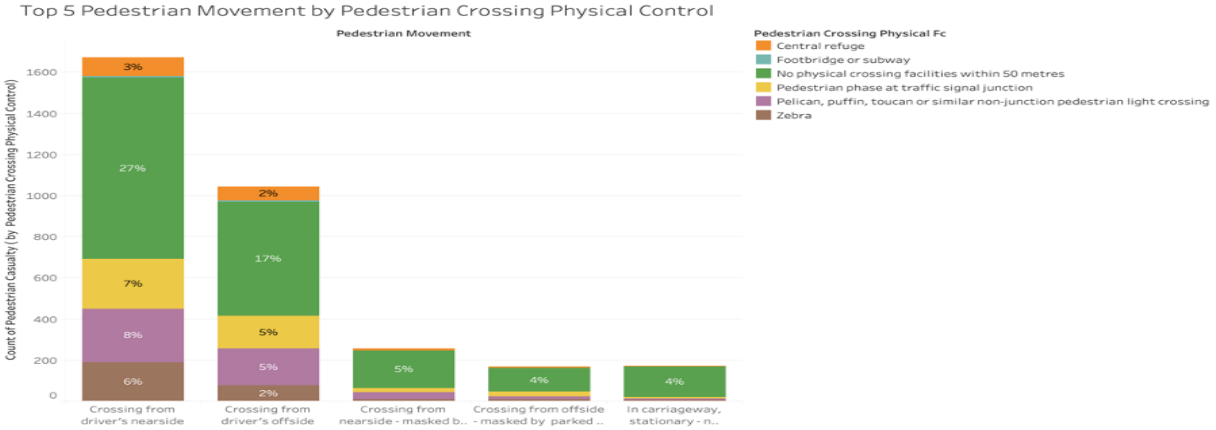

Fig 14

The Top 5 Pedestrian Movement by Light Conditions chart above shows the five highest pedestrian movement casualties, and is grouped by Pedestrian Crossing Physical Control. For every type of pedestrian movement, there was No physical crossing facilities within 50 metres for most of the accidents.

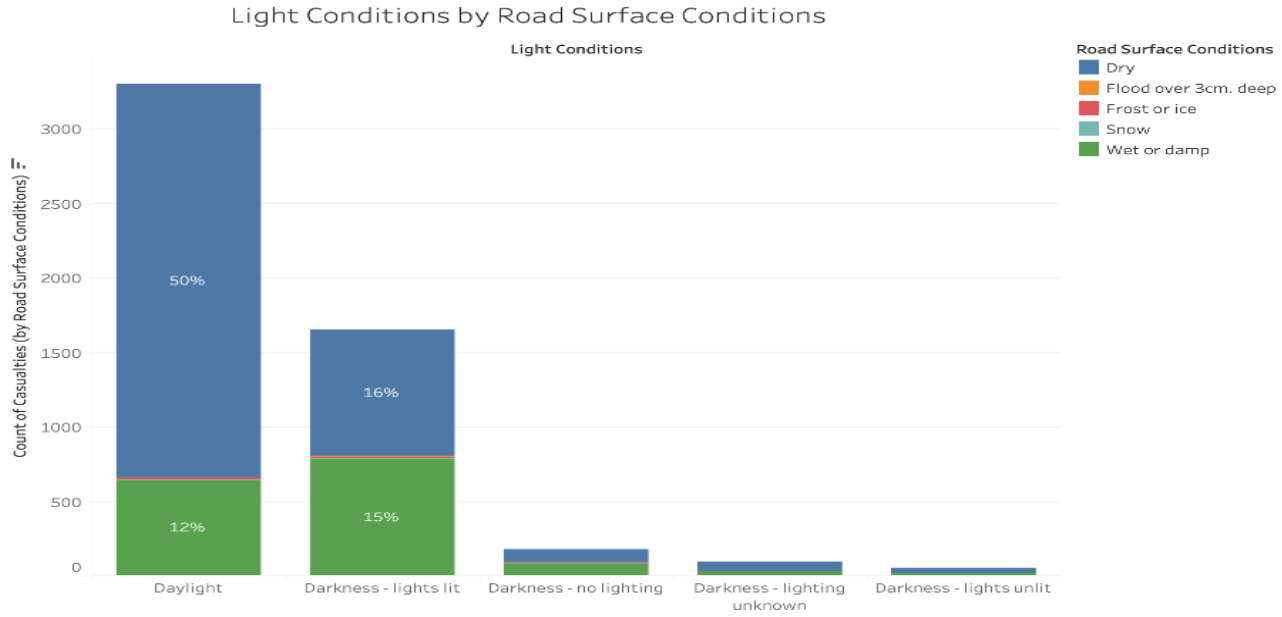

Fig 15

The Light Conditions by Road Surface Conditions chart to the left shows the Light Conditions of casualties and groups it by Road Surface Conditions. It is clear that most pedestrian accidents happen in Daylight or Darkness - lights lit and over Dry or wet roads. In Daylight, 50% of pedestrian casualties are in Dry road surface conditions, and 12% are in Wet or Damp road surface conditions.

Conclusions

Several interesting patterns were observed in the two datasets. After analysing the demographics of casualties, uneven sex ratios were discovered in different casualty classes. Majority of the driver or rider casualties were males (about 67%) while almost 60% of passenger casualties were females. Motorcycle and cycle casualties are mostly males(85%), whereas casualties caused due to horse riding and public transport are majorly females (62%). Coming to local authorities, the five local authorities with the most KSI accidents were Birmingham, Cornwall, Leeds, Sheffield, and Westminster. They made up 7.17% of total accidents. It was observed that more accidents occurred on weekends than weekdays. Additionally, more than 90% of accidents happen on highways as opposed to streets, and two-thirds of incidents occur in dry circumstances. The majority of accidents happen at unmonitored junctions.

Pedestrians who were killed or seriously injured accounted for 6,200 of the total killed or seriously injured casualties. Pedestrian casualties constitute a higher percentage (87%) in urban areas than in small towns and rural areas. Senior citizens aged between 70-80 years old accounted for half the pedestrian fatalities. Analysis shows that 69% of the pedestrian accidents happened in dry road surface conditions while 30% happened in wet or damp road surface conditions. There was also no physical crossing facility within 50 miles for 62% of the pedestrian accidents.