**Coursework 2: Group analysis case study**

**Group members:**

|  |  |  |
| --- | --- | --- |
| 1) | Kritin Senaprame | 201684900 |
| 2) | Mohana Sangeetha Vallinayagam | 201688572 |
| 3) | Shashidhara Kudakuru Koppal Brahme Gowda | 201665469 |
| 4) | Shreyansh Parihar | 201684933 |

**Introduction:**

The purpose of this report is to analyse the road safety data of England for the 12 months from 01/01/2019 to 31/12/2019 to understand 3 patterns majorly. They are, the patterns exist in the demographics of casualties, the patterns are there between the local authorities by considering the accidents which included a killed or seriously injured (KSI) casualty and the patterns are there in pedestrians who were actually KSI casualties. In this study, the complete process includes various steps, starts from data collection, followed by data cleaning and processing, data visualization and at the end conclusion. The raw data collected from UK Government website (<https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>) and was in csv format, the applications R and Python were used for data cleaning and analysis purpose, for data visualization Power BI and Tableau were adopted. Packages used in R: dplyr, VIM, dlookr, and corrgram.

**Data quality:**

In order to perform analysis potentially, good quality of data is necessary. For this report, it focuses on datasets of road safety data known as STATS19 in the year 2019 for only England country. These datasets include three data files (The accident data, casualty data, and Guidance of Road Safety Open Dataset Data) and contains a lot of data quality problems which are needed to address namely:

* Invalid value
* In each file, there are certain invalid values that should not be in attribute according to guide file as Table 1:

Table 1: Invalid value

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **File** | **Attribute name** | **Value** |
| 1 | Accident | ‘lsoa\_of\_accident\_location’ | -1 |
| 2 | Accident | ‘second\_road\_class’ | -1 |
| 3 | Casualty | ‘casualty\_type’ | -1 |

* These issues are addressed by replacing its field by mode average value.
* Incorrect reference
* In guide file, there are some inconsistent information in attribute as Table 2:

Table 2: Incorrect reference

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **File** | **Attribute name** | **Value** | **Remark** |
| 1 | Guide | ‘local\_authority\_highway’ | 0 | Should not have |
| 2 | Guide | ‘second\_road\_class’ | 0 | The description should mention second\_road\_class instead of first\_road\_class |

These issues are addressed by changing manually to be appropriate meaning.

* Wrong data type

1. For ‘accident\_index’ and ‘accident\_reference’ columns in accident and casualty file, these contain the unique value which combines with numbers and/or characters for each casualty. This kind of data should be assigned as string type whereas for some records which contain only numbers, they are defined as integer type.
2. For ‘date’ and ‘time’ columns in accident file, it is defined as integer type instead of datetime type.

* These issues are addressed by changing into the right format by a function of each program.
* Noted that: this problem doesn’t cause by the collected data, but it may occur in some specific software or tools like pandas’ data frame in Python.
* Missing & unknown values
* Missing and unknown values take place largely in these datasets due to following many reasons and showing in Figure 1 and 2:

1. Negative value (-1):

According to guidance in guide file, if each data value is -1, it means that data missing or out of range. The number of this error accounts for about 50,000 records in accident file and around 15,000 records in casualty file.

1. Unknown data:

According to guidance in guide file, some field names are defined that there is a specific value (2, 9, 10, and 99) meaning unknown value. For example, ‘casualty\_type’ in casualty file assigns value 99 as unknown.

1. Nan:

In accident file, there are some columns which contain the Nan. This unknown data has less than 100 records.

Timeline, bar chart

Description automatically generated

Figure 1: The number of missing & unknown values in accident file

Chart, bar chart

Description automatically generated

Figure 2: The number of missing & unknown values in accident file

* These issues are addressed by replacing its value by median average value for location data like latitude and longitude.
* Numeric outlier
* These datasets have 3 columns which contain the outlier data such as ‘number\_of\_vehicles’, ‘number\_of\_casualties’, and ‘age\_of\_casualty’.
* These issues are addressed by replacing its value by mean average value.

**Data definition and characterization:**

Overall, the dataset consists of 54 columns which are in both quantitative and qualitative in nature, to be precise majority of them are Categorical, followed by Numerical and few are Ordinal data types. In the data cleaning stage, we had to remove and replace some of the rows due to quality issue. The data variables which we considered for analysis are elaborated pointwise as below.

The Accident dataset:

Table 3: The characterization of accident file

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Columns** | **Type** | **Subtype** | **Definition** |
| 1 | accident\_reference | Categorical | nominal | The unique identification of an accident in that particular year |
| 2 | location\_easting\_osgr | numerical | Spatial | The easting location of collision as per Ordnance Survey |
| 3 | location\_northing\_osgr | numerical | Spatial | This column gives the northing location of collision as per Ordnance Survey |
| 4 | longitude | numerical | Spatial | Longitude of accident location |
| 5 | latitude | numerical | Spatial | Latitude of accident location |
| 6 | police\_force | categorical | nominal | Police force number |
| 7 | accident\_severity | categorical | Ordinal | How severe the accident is |
| 8 | number\_of\_vehicles | numerical |  | Count of vehicles involved in an accident |
| 9 | number\_of\_casualties | numerical |  | Count of casualties involved in an accident |
| 10 | date | categorical |  | Date of accident |
| 11 | day\_of\_week | nominal |  | day number in a week of accident |
| 12 | time | categorical |  | Time of accident |
| 13 | accident\_index | categorical |  | Unique number, which is combination of year, reference number |
| 14 | accident\_year | numerical |  | Year of accident |
| 15 | local\_authority\_district | Categorical | Nominal | district of local authority from where accident is reported |
| 16 | local\_authority\_ons\_district | Categorical | Nominal | the office reference for National Statistics unique code |
| 17 | local\_authority\_highway | Categorical | Nominal | Code of local authority where the highway is |
| 18 | first\_road\_class | Categorical | Nominal | UK roads (excluding motorways) fall into the following 4 categories: A road, B road, C road, unclassified, Motorways, and A road (M) (first road) |
| 19 | first\_road\_number | Categorical | Nominal | the road number (1-9999) |
| 20 | road\_type | Categorical | Nominal | the lane type of the road |
| 21 | speed\_limit | Categorical | Ordinal | the maximum speed that is allowed to drive on the road |
| 22 | junction\_detail | Categorical | Nominal | The description at the junction |
| 23 | junction\_control | Categorical | Nominal | The things that control the traffic at the junction |
| 24 | second\_road\_class | Categorical | Nominal | UK roads (excluding motorways) fall into the following 4 categories: A road, B road, C road, unclassified, Motorways, and A road (M) (second road) |
| 25 | second\_road\_number | Categorical | Nominal | the road number (second road , 1-9999) |
| 26 | pedestrian\_crossing\_human\_control | Categorical | Nominal | the person who control human crossing the road |
| 27 | pedestrian\_crossing\_physical\_facilities | Categorical | Nominal | crossing facilities which helps people cross the road |
| 28 | light\_conditions | Categorical | Nominal | The brightness of that situation |
| 29 | weather\_conditions | Categorical | Nominal | Unique number is assigned to weather type |
| 30 | road\_surface\_conditions | Categorical | Nominal | Unique number is assigned to condition of the road |
| 31 | special\_conditions\_at\_site | Categorical | Nominal | Unique number is assigned to condition of the location where accident happened |
| 32 | carriageway\_hazards | Categorical | Nominal | Hazards Type |
| 33 | urban\_or\_rural\_area | Categorical | Nominal | Accident zone area type (Rural or Urban) |
| 34 | did\_police\_officer\_attend\_scene\_of\_accident | Categorical | Nominal | Yes or No type |
| 35 | trunk\_road\_flag | Categorical | Nominal | The accident zone area type (Rural or Urban) |
| 36 | lsoa\_of\_accident\_location | Categorical | Nominal | LSOA (Lower Super Output Area) code of the accident location |

The casualty dataset:

Table 4: The characterization of casualty file

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Columns** | **Type** | **Subtype** | **Definition** |
| 1 | accident\_index | categorical |  | the unique number which is combination of year and reference number |
| 2 | accident\_year | Numerical |  | the year of accident |
| 3 | accident\_reference | Numerical |  | unique identification of an accident in that particular year |
| 4 | vehicle\_reference | Categorical | Nominal | unique value for each vehicle in a singular accident. Can be used to join a Casualty to a vehicle |
| 5 | casualty\_reference | Categorical | Nominal | unique value for each casualty in a singular accident |
| 6 | casualty\_class | Categorical | Nominal | the position of the victim while travelling during accident. |
| 7 | age\_of\_casualty | Categorical | Nominal | the age of the victim when accident happened. |
| 8 | sex\_of\_casualty | Categorical | Nominal | the gender of the victim |
| 9 | age\_band\_of\_casualty | Categorical | Ordinal | the age of the victim into one of 10 groups with equal frequency of 5, ranging from 0-5 to 66-75 and 75+. |
| 10 | casualty\_severity | Categorical | Ordinal | the seriousness of harm suffered by the victim. |
| 11 | pedestrian\_location | Categorical | Nominal | the location of pedestrian during accident. |
| 12 | pedestrian\_movement | Categorical | Nominal | the involvement of pedestrian in accident. |
| 13 | car\_passenger | Categorical | Nominal | the spot of the victim in car during accident. |
| 14 | bus\_or\_coach\_passenger | Categorical | Nominal | state of the victim in bus during accident. |
| 15 | pedestrian\_road\_maintenance\_worker | Categorical | Nominal | Possibility of the victim being Maintenance Worker. |
| 16 | casualty\_type | Categorical | Nominal | the details of the transport in which victim was |
| 17 | casualty\_home\_area\_type | Categorical | Nominal | Type of the area where the victim lives in. |
| 18 | casualty\_imd\_decile | Categorical | Nominal | the relative deprivation level of the area the casualty lives in; the relative deprivation level of the area the driver involved in the crash lives in; or in casualty reports, it is also possible to see the relative deprivation level of the area in which the related driver lives. (reference from http://mast.roadsafetyanalysis.org/wiki/index.php?title=Driver\_IMD\_Decile). |

**Detail analysis:**

1. Demographics Pattern of casualties

As a Figure 3, it can be clearly seen that the highest number of casualties occurs at Birmingham (2,623 accidents) which almost double of the second rank’s city. This might because Birmingham is one of the biggest cities and locate in the middle of England where people tend to use it as a destination for resting during the long journey. These reasons make this city overcrowd by traffic and lead to more accidents. Moreover, both Westminster and Leeds are also important location that the casualties happen the most, accounting for 1,521 and 1,451 accidents respectively. It’s quite sensible since both of those are big city in different region. For the other cites, casualty numbers are high and slightly different from Leeds.

A picture containing application

Description automatically generated

Figure 3: The casualties location divided by district in England

Another common insight of casualties, Figure 4, is that the graph shape of male and female is quite similar when focusing on age band. However, the number of men casualties is much higher than that of women accidents in the same age band and one possible cause can be that  women behavior is much calmer than that of men or it is possible that  The age between 26 and 35 is the most important because it contains the highest number of casualties in both male (20,548 cases) and female (12,305) in this age band, one possible reason can be, the majority of people are likely to buy a new car by themselves and less careful due to unfamiliarity with driving. Thus, there are more cars with drivers in age group 26-35.

Chart, bar chart

Description automatically generated

Figure 4: The number of casualties divided by different gender and age

1. Patterns for accidents involving killed or seriously injured (KSI) casualties as per different local authority levels:

Figure 5 shows KSI casualties in different districts in England. Graph shows top 30 counties with KSI casualties.  As per the analysis done and shown in graph in Figure 5, local authority of Birmingham reported highest number of KSI casualties, thus, we can infer that Birmingham roads are most dangerous to drive to on, as in case of most of its accidents someone is either killed or seriously injured.  As per graph every district who has KSI casualties above 268 are a cause of concern, and necessary measures should be taken to by local highway authorities of these districts to reduce casualties.

Chart

Description automatically generated

Figure 5: The number of KSI casualties occurring in different district

1. Patterns for Pedestrian Movement and Pedestrian location who were of KSI casualties

Looking at Figure 6, it’s obvious that the almost three of four of all accidents occur at the place where vehicle are usually on high speed. For this data, that location of pedestrian is in carriageway or on verge. The sum of casualties occurring in carriageway is more than half of the total, especially when pedestrian cross the road showing 46.23 per cents. In marked contrast, the percentage of accidents is rarely seen at zig-zag way. This might because it’s the caution zone which everyone always concerns, and it can be noticed in the long distance, the total number of this type of records are less than 1 per cent.

Chart, sunburst chart

Description automatically generated

Figure 6: The proportion of KSI casualties divided by pedestrian location

Turning to Figure 7, not only the location of pedestrian is important for analysis the pattern, but the pedestrian movement is also one of the main factors. Almost all the number of accidents occur when the pedestrian crosses the road at nearside or offside of driver. Moreover, the highest number of pedestrians movement is crossing from driver’s nearside being 47.2 per cents because it’s a blind view of the driver when driving the vehicles. However, it makes sense that pedestrian who just stands in the stationary rarely suffer the accidents.

Chart, sunburst chart

Description automatically generated

Figure 7: The proportion of KSI casualties divided by pedestrian movement

**Conclusion:**

As seen in demographics pattern we found that the greatest number of casualties were from age group of 25-35 and Birmingham has most number of casualties out of any districts or city. After looking into casualty severity, we see that the accidents where the victim was killed or seriously injured (KSI) were also from Birmingham. Thus, not only there are most number of casualties in Birmingham city many of those casualties were KSI casualties too. Hence local authorities in Birmingham should take necessary actions to bring down the casualty count.

Then we investigated casualty caused to pedestrians based on 2 parameters – their location and when the accident happened. Almost 55% of pedestrian where on carriageway when accident happened and in almost 54% of time pedestrians where on nearside of driver at the time of accident.