Oluwafemi Olobashola

Accident Data Project

1. **Introduction**

Transportation is an important part of human existence, as it plays important roles in determining the quality of life we live. With increase in population, especially in urban areas, there is usually a commensurate increase in the need for transportation, which increases traffic congestion and road traffic accidents (Othman, A. G. and Ali, K. H., 2020). Accident occurrences and its attending fatalities is on a steady increase, making it a source of major concern to which preventive and corrective actions should be devised using available data. (Gopalakrishnan S., 2012).

According to WHO road safety report of 2018, 1.35 million deaths are recorded from road traffic accidents annually. Of these deaths, pedestrians, cyclists, and motorcycles accounts for the larger proportion, especially in developing countries. (WHO, 2018).

This report investigates the severity of recorded accidents, possible pattern in the timing of their occurrence, with the aim of predicting the occurrence and severity of future accidents to offer some more protection to the most vulnerable members of the society. We would be using the accident data from the department of transport, which is made available to them by the police, who gather some 50 data items from the scenes of reported accidents.

1. **Explorative Data Analysis**

Upon exploration the dataset was found to consist of four tables; accident, vehicle, casualty and lsoa tables., A summary of the various tables were obtained and summarized as follows.

**Accident Table:** There are 36 columns in the table; 13 contain categorical variables, while 23 contain numerical variables. It had 91,199 observations, and 56 missing values, constituting less than 0.1% of the total observations. No duplicates were found in the table.

**Vehicle Table:** There are 28 columns (21 numerical variables and 7 categorical variables), 167,375 observations and no missing values. Several columns had erroneous imputes like -1, 9 and 99.

**Casualty Table:** There are 19 columns in the casualty table; 10 categorical variables and 23 numerical variables, it contains 115,584 observations with no missing values. The sex\_of\_casualty, car\_passanger, pedestrian\_road\_maintenance\_worker and bus\_or\_coach\_passanger columns seems to have some erroneous imputes.

**Lsoa Table:** There are 7 columns and 83,154 observations and contains no missing values.

**Data Cleaning**

The missing values observed were found to be in four columns of 14 rows. The columns involved are location\_easting\_osgr, location\_northing\_osgr, longitude and latitude. Since these features are separated by distance, the missing values were filled by linear interpolation in the forward direction (Raghav, A., 2012). The data was however first sorted using the local\_authority\_highway column.

* 1. **Accident trends by hours of the day and by days of the week**

Time of various accident occurrences were recorded in hours, with the format HHMM, the hour of the day was therefore extracted using the substring method to extract the first two positions in the time string.

**Accident Trends by Hours of The Day**

A plot of accident occurrences shows that most accidents occur late morning and early evening, while the least accidents occurred between late at night and the early hours of the morning up till 0500. The highest number of accidents were recorded around 1700 hour with an occurrence of 7,813 accidents. This was followed by 1600 and 1500 hours at 7,381 and 7,361 accident occurrences respectively. The least accidents were recorded between the hours of 0200, 0300 and 0400 hours at 658, 566 and 508 respectively.

Accident occurrences progressively increased from the 0500 hours up until 1700 hours. The only exception to this is between 0800 and 0900 hours where there was a drop from 5,267 to 3,917, and begin to dip at 1800 hours, falling to the least level at 0400 hours where 508 accidents were recorded.

There is therefore a clear link between human activities peaking at around close of daily business and when traffic is usually highest, and least at periods that witness the least daily activities (Gopalakrishnan S., 2012).

A graph of accident occurrence

Description automatically generated

Figure ; Accident Occurrence by Time of Day

**Accident Trends by Days of The Week**

Observation from analysis of accident occurrences by days of the week showed that the rate increases from 10,315 on Sunday to 14,889 on Friday. The highest accident rates are therefore recorded on Fridays. Rate of accident occurrences dropped from 14,889 on Friday to 12,336 on Saturday, and a further drop on Sunday. As posited by Foster et al. (2015) who studied the correlation between alcohol consumption and alcohol related accidents in Switzerland saw an increased incident rates during weekends, corroborating the observed trend in our own analysis.

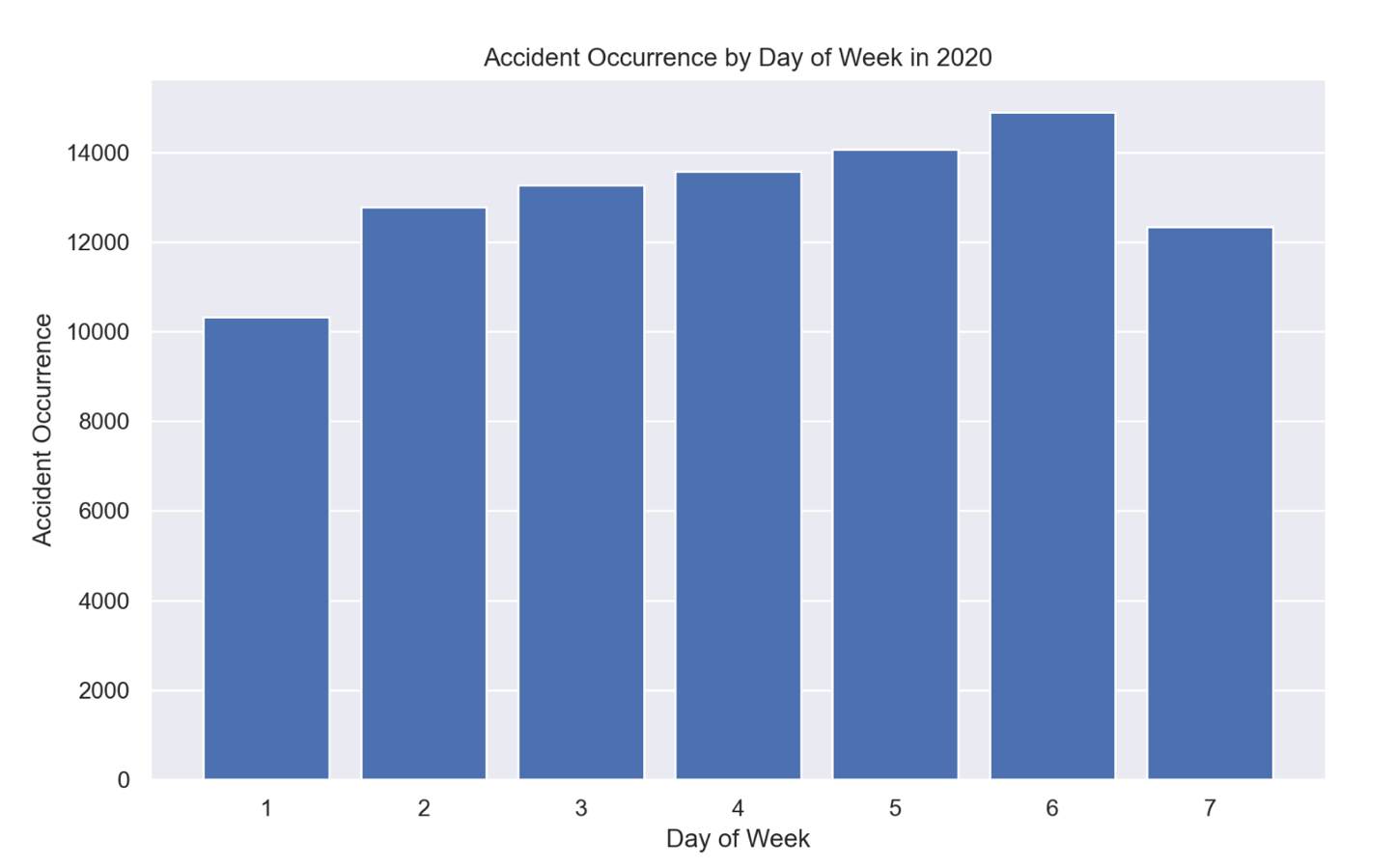


Figure ; Accident Occurrence by Days of The Week

* 1. **Motorcycle accident trends by hours of the day and by days of the week**

Motorcycles of various engine capacities were considered; under 125cc Motorcycles, 125cc – 500cc motorcycles and motorcycles with capacity over 500cc. The trend of accidents was almost consistent with all of them, having the maximum accident occurrences around 1700 hours, and the least occurrences between 0200 and 0400 hours. Under 125cc motorcycles recorded the highest hourly accident rate of 1,293, while over 500cc motorcycles recorded 421 and 125cc – 500cc motorcycles recorded the least at 265 occurrences.

For the days of the week, Friday recorded the highest occurrences for both motorcycles under 125cc and those of 125cc – 500cc, whereas over 500cc motorcycles record the highest occurrences on Sundays.

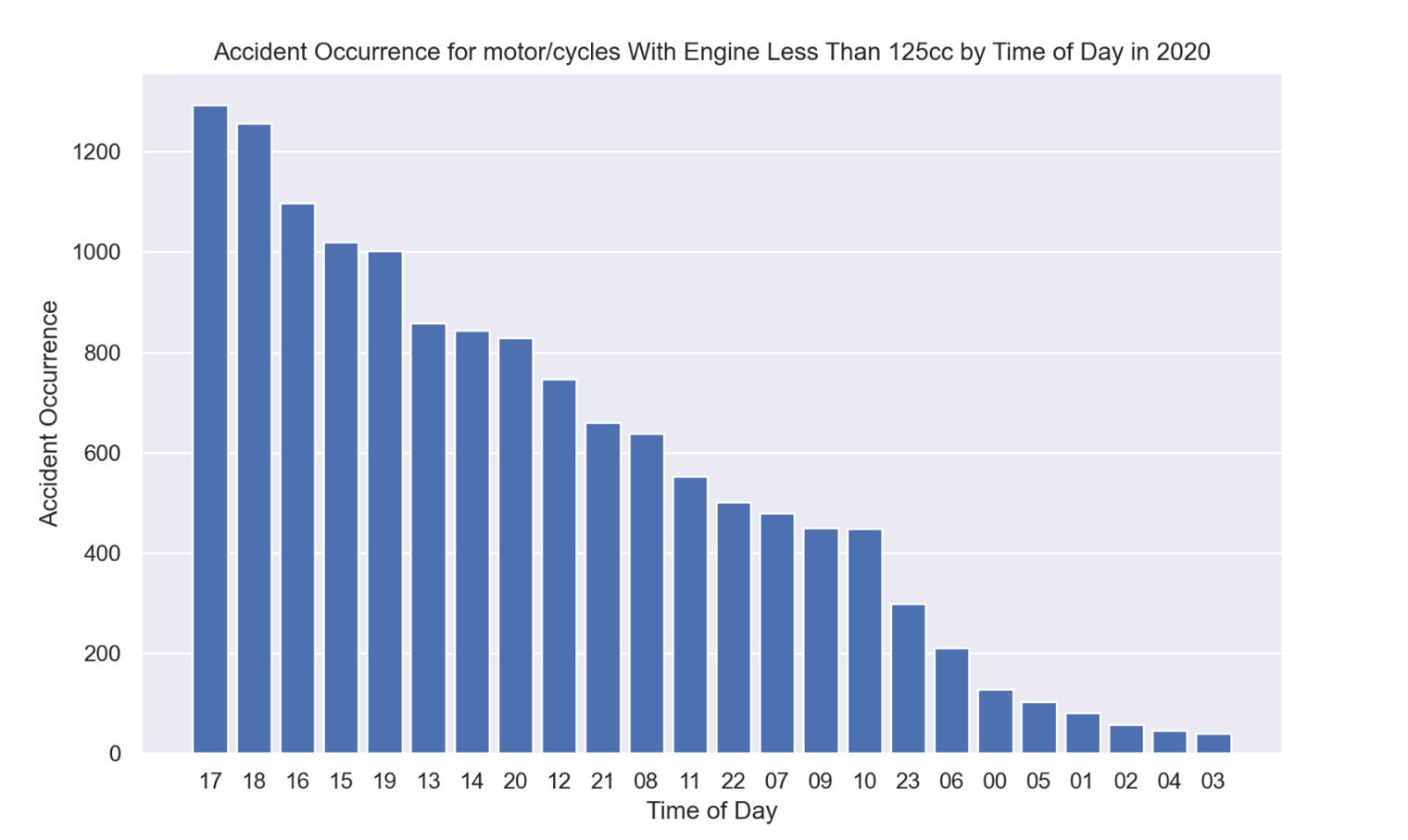


Figure ; Accident occurrence for under 125cc motorcycles

* 1. **Pedestrians’ accident trends by hours of the day and by days of the week**

Like the general accidents, accidents involving pedestrian occurs at every hour of the day, with 1500, 1600 and 1700 hours having the highest records of accidents involving pedestrians at 1,672, 1,323 and 1,274 respectively. The hours of 0400, 0500 and 0300 had the lowest records of accidents involving pedestrians at 50, 74 and 82 respectively. 0800 hours seems to be the most dangerous hour of the morning for pedestrians with an average of 1,060 accident occurrences within this hour involving pedestrians. This might as well be attributed to the peak of rush hour when individuals and school aged children are making their way to work or school.

The days with the highest occurrence of accidents involving pedestrians were Fridays and Thursday at 2,543 and 2,366 accident occurrences respectively.

The least accidents involving pedestrians happen on Sundays and Saturdays at 1,242 and 1,878 respectively. Pedestrians are evidently at high risk of injuries from road traffic accidents.

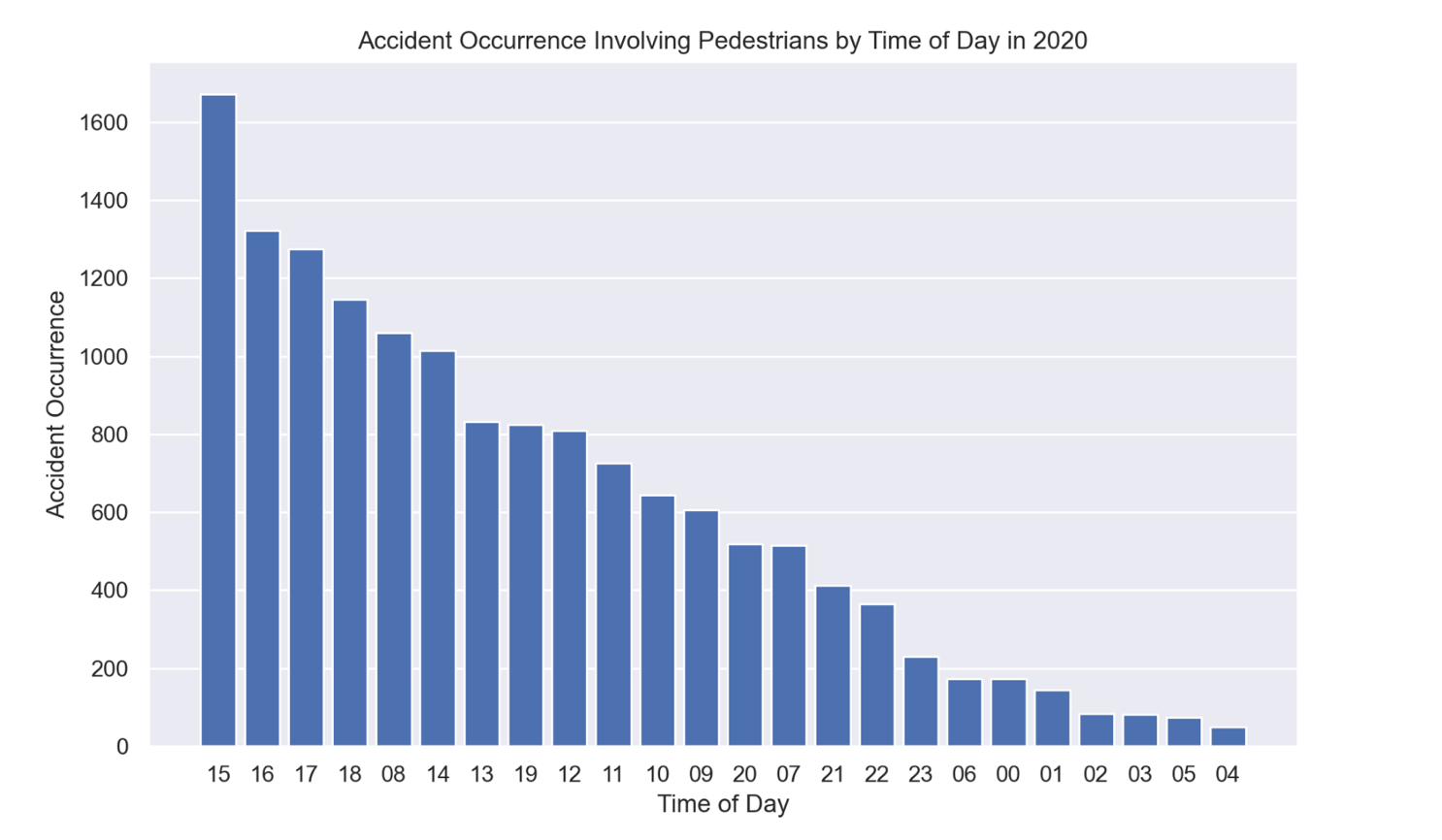


Figure ; Pedestrian Accidents by Hour of Day

* 1. **Exploring the impact of select features on the severity of accidents using Apriori Algorithm**

Apriori being an algorithm which mines for frequent items in a dataset, using association rules. Association rules look at the probability of occurrence of the relationships between various features in the dataset. The major indictors of these probabilities are support, confidence and lift.

**Support:** This measures the frequency of occurrence of each of the antecedent A and consequent B in the entire dataset. The higher the support of any rule, the more prevalent it's itemset are prevalent in the dataset. It is calculated by dividing the frequency of occurrence of the antecedent and consequence by the total number of events.

𝑆𝑃 = 𝑝(𝐴∩𝐵) 𝑁

**Confidence**

This measures the reliability or certainty of the rules in a dataset. It is calculated by dividing the frequency of occurrence of the antecedent A and consequence B occurring together by the frequency of the antecedent A. Higher confidence values indicate more reliable associations.

𝐶𝑓 = 𝑃(𝐴∩𝐵) 𝑃(𝐴)

**Lift**

Lift is an indication of the statistical dependence between the itemset of a rule. Lift of value greater than one indicates a positive relationship between the antecedent and the consequent, lift of less than one is an indication of negative relationship between the itemset, while a lift of one is an indication of no relationship between both. Lift is calculated by dividing the support of both antecedent and consequence by the multiplication of both their individual supports.

𝐿𝑡 = 𝑃(𝐴∩𝐵) 𝑃(𝐴)×𝑃(𝐵)

(Hoanca, B. and Mock, K., 2011).

Table ; Association Rules

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Antecedents | Consequents | Support | confidence | Lift |
| carriageway\_hazards\_0, casualty\_severity\_3 | accident\_severity\_3 | 0.750965 | 0.938589 | 1.207275 |
| casualty\_severity\_3, weather\_conditions\_1 | accident\_severity\_3 | 0.604246 | 0.939787 | 1.208816 |
| vehicle\_type\_9 | accident\_severity\_3 | 0.571302 | 0.795391 | 1.023083 |
| light\_conditions\_1, casualty\_severity\_3 | accident\_severity\_3 | 0.558282 | 0.942233 | 1.211961 |

The apriori algorithm was used to explore the impact of selected variables on accident severity. The most frequently occurring form of accident severity reported was accident\_severity\_3, which is associated with slight or minor injuries. The following are the most frequent conditions associated with accident\_severity\_3:

* No carriageway hazards.
* Fine weather conditions
* Daylight
* Normal passenger cars (taxis, estate cars, three- and four-wheel cars and minibuses)

These conditions portend no apparent threats or dangers to road users, which explains why most of the accidents recorded were associated with slight or minor injuries.

* 1. **Clustering of accidents in the Humber region.**

The Humber region consists of Kingston Upon Hull, East Riding of Yorkshire, North Lincolnshire and North East Lincolnshire

To visualize the distribution of accidents in the region, a scatter plot was created using the GPS coordinates of each accident, and a distinctive distribution was obtained, which gives some insight into these accidents.

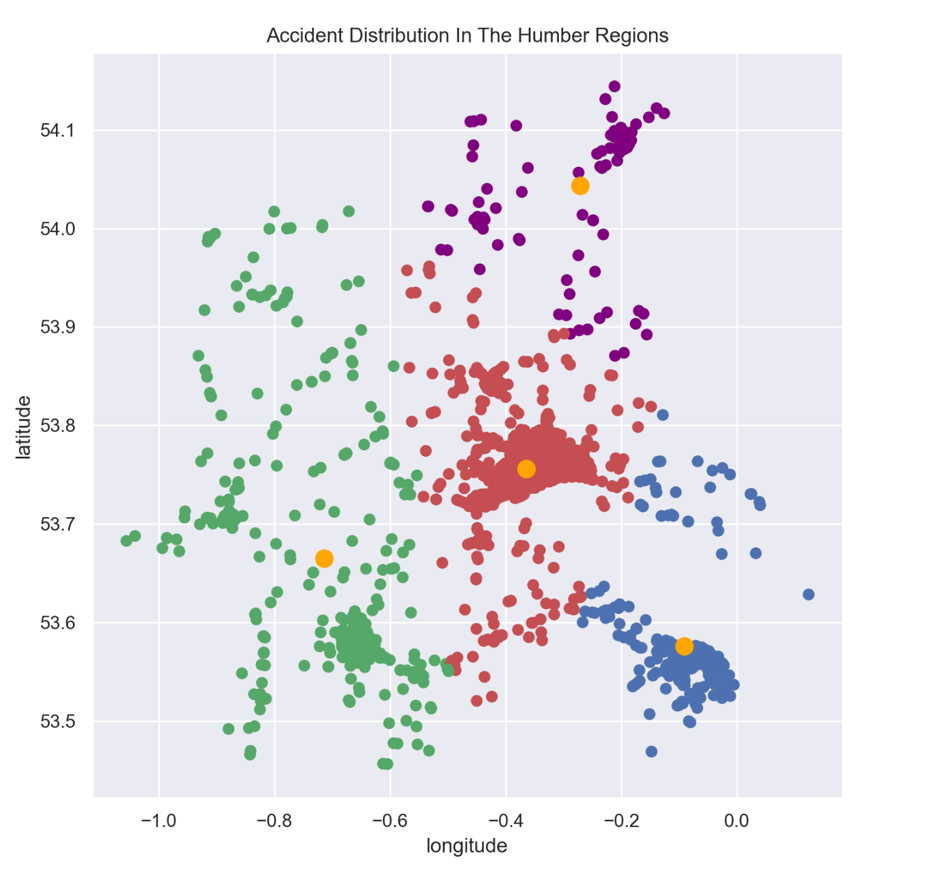


Figure ; Accident distribution in the Humber region

|  |  |  |
| --- | --- | --- |
| Local Authority | Accident Occurrence | Cluster Colour |
| Kingston upon Hull, City of | 603 | Red |
| East Riding of Yorkshire | 495 | Purple |
| North Lincolnshire | 306 | Green |
| North East Lincolnshire | 305 | Blue |

From the above accident occurrences, it can be clearly seen that there is a clear correlation between population density and accident occurrences in these towns, making preventive measures necessary to ensure public safety.

* 1. **To detect data points that are significantly different from other points in the dataset, Local Outlier Factor (LOF) was carried out on the dataset.**

Upon application of LOF, 20,195 outlying data points were identified in the dataset. Further investigation revealed some of these outliers to be values like -1, 9, and 99 in columns that do not have such classes or range of values. These outliers can potentially impact further analysis on the data and was therefore cleaned. The outliers were converted to NAN values after which continuous columns had their NAN values replaced with the median value, while the categorical columns were treated using simple imputer.

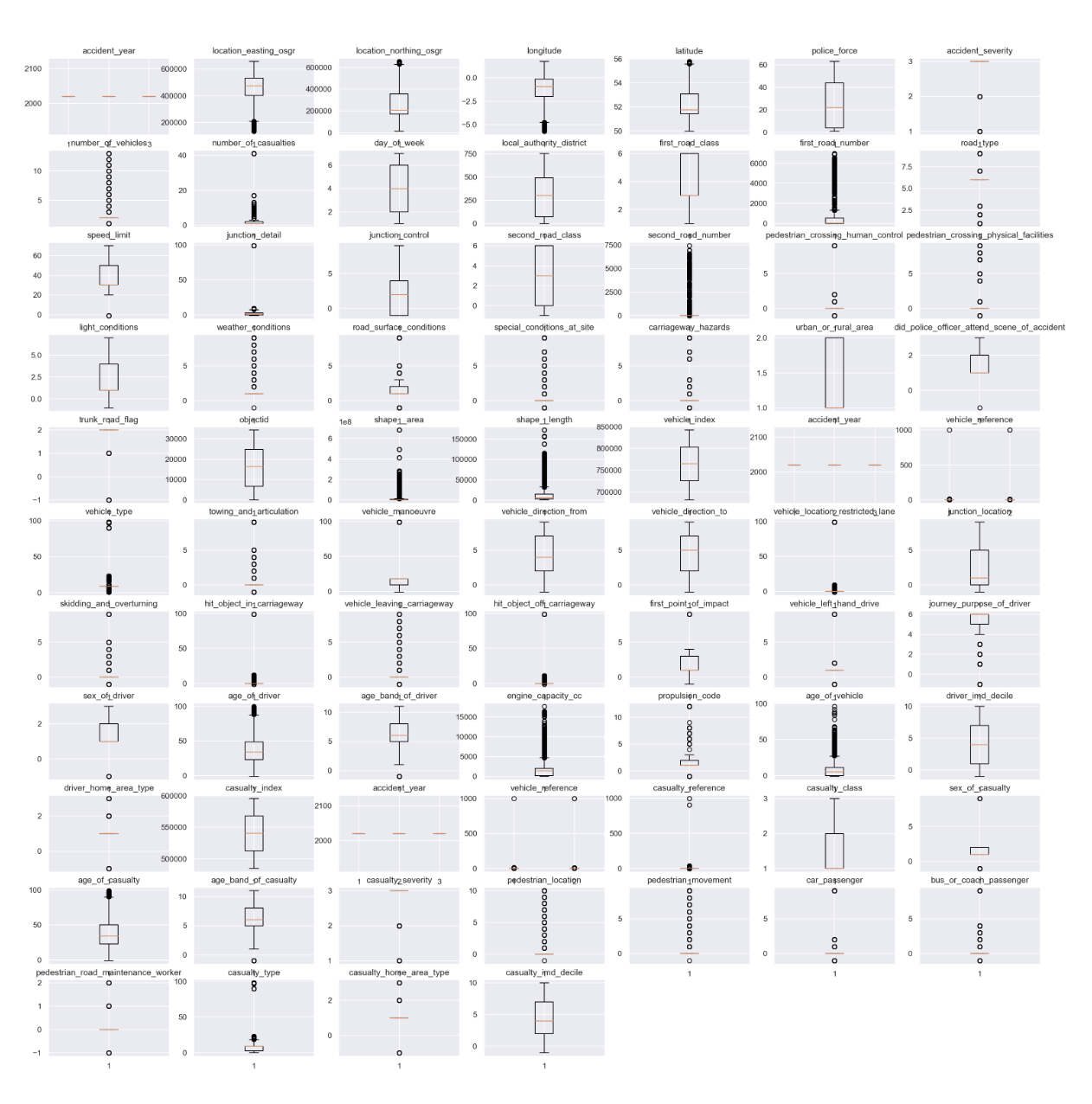


Figure ; Visualisation of the distribution of values in the variables

* 1. To accurately predict fatal injuries, feature selection was carried out to determine the most important features to use for the classification task. Random forest was used, and performance compared with stacker classifier model, K-Nearest Neighbor, Decision Tree, and Gradient Boosting. In building the stacker models, cross validation and hyperparameter tunning was carried out to give the best results, the random forest classifier however showed superior performance to other models.

Table ; Performance of various classification models

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Precision | Recall |
| Random Forest | 93% | 92% | 94% |
| Decision Tree | 87% | 87% | 88% |
| K-Nearest Neighbor | 88% | 89% | 87% |
| Gradient Boosting | 91% | 90% | 91% |
| Stacker | 90% | 90% | 91% |

The random forest model can therefore be adopted for use in predicting severe accident occurrences, with the aim of putting in place measures to ensure the safety of the public.

Further recommendation include.

* Training and enlightenment of cyclists (especially under 125cc motorcycles) on traffic rules and safety measures
* Increased presence of law enforcement on the roads physically and remotely with the use of monitored circuit cameras to promote adherence to traffic rules.
* Increased testing of drivers for alcohol especially around weekends to discourage drunken driving.
* Execution of road safety awareness campaign to sensitize not just drivers and riders, but also pedestrians and the public.

If deployed alongside the above recommendations, the random forest classifier would help to accurately predict accident fatalities, thereby making it possible to deploy targeted measures to prevent their occurrences.

**Reference:**

Ferris, J. O. (2018). Data analysis accidents in UK. Retrieved from <https://www.kaggle.com/code/joseconomy/data-analysis-accidents-in-uk> July 30, 2023.

Foster, S., Gmel, G., Estévez, N., Bähler, C., & Mohler-Kuo, M. (2015). Temporal Patterns of Alcohol Consumption and Alcohol-Related Road Accidents in Young Swiss Men: Seasonal, Weekday and Public Holiday Effects. Alcohol and Alcoholism, 50(5), 565-572. doi:10.1093/alcalc/agv037.

Gopalakrishnan S. (2012). A public health perspective of road traffic accidents. *Journal of family medicine and primary care*, *1*(2), 144–150. Retrieved from <https://doi.org/10.4103/2249-4863.104987> on July 27, 2023.

Hoanca, B. and Mock, K., 2011. Using market basket analysis to estimate potential revenue increases for a small university bookstore. In *Conference for Information Systems Applied Research* (Vol. 4, No. 1822).

Kurniawan, T. B. (2021). The Clustering Algorithm with Geolocation data. Retrieved from <https://medium.com/thelorry-product-tech-data/the-clustering-algorithm-with-geolocation-data-d6dd07ed36a> on August 4, 2023.

Othman, A. G. and Ali, K. H. (2020)."Transportation and Quality of Life." Planning Malaysia: Journal of the Malaysian Institute of Planners, vol. 18, no. 3, 2020, pp. 35-50. Retrieved from <https://www.planningmalaysia.org/index.php/pmj/article/view/774/575> on July 22, 2023.

Parashanth, J. (2018). Substringing Function Overview. Retrieved from <https://www.sqlshack.com/sql-substring-function-overview/> on July 28, 2023.

Petch, R. O., & Henson, R. R. (2000). Child road safety in the urban environment. *Journal of Transport Geography*, 8(3), 197-211. ISSN 0966-6923. doi:10.1016/S0966-6923(00)00006-5. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0966692300000065> on August 8, 2023.

Raghav, A. (2012). Interpolation Techniques Guide & Benefits | Data Analysis. Retrieved from<https://www.analyticsvidhya.com/blog/2021/06/power-of-interpolation-in-python-to-fill-missing-values/> on July 28, 2023.

WHO Publications. (2018). *Global status report on road safety, 2018*. Retrieved from <https://www.who.int/publications/i/item/9789241565684> on July 27, 2023.