Introduction

According to the official statistics, there were 91,199 reported accidents and 1,472 fatalities on the roads of the United Kingdom in 2020 [1]. To store and analyse this large and complex data, SQL relational database is a suitable tool that can handle various types of information and queries. The aim of this report is to identify the key factors that contribute to road traffic accidents and to propose some solutions to improve road safety and prevent further losses of lives and resources.

Data Cleaning

The data cleaning process focused on four tables within the accident dataset, concentrating on data segments relevant for analysis this is well documented in the jupyter notebook. Initial steps involved addressing missing entries in longitude and latitude columns, totalling 14. These gaps were filled using adjacent rows with matching local authority districts and non-missing longitude and latitude values.

An additional cleaning task involved the road surface condition column, which had 316 missing entries represented as -1. These gaps were imputed using weather conditions to account for their impact on road surface conditions. Similarly, minor modifications were applied to the light conditions and weather conditions columns, each with one missing value, by replacing them with mode values. The speed limit column, with 12 entries as -1, was imputed with its mode value of 30, necessary for Apriori algorithms analysis.

The second road class column, also crucial for model creation, had 109 missing entries. These were addressed using a predefined value for unclassified road class (6). The junction control column had 38290 entries as -1, which were resolved by creating a lookup table linking different junction control types to appropriate corresponding junction details.

Additional adjustments were made for analysis purposes, including converting the date column from an object to a date representation and converting the time column from an object to a float type using decimals for improved representation.

Analysis

Initially started the exploration of the accidents by observing the days in which there is a larger number of accidents and times of the day.

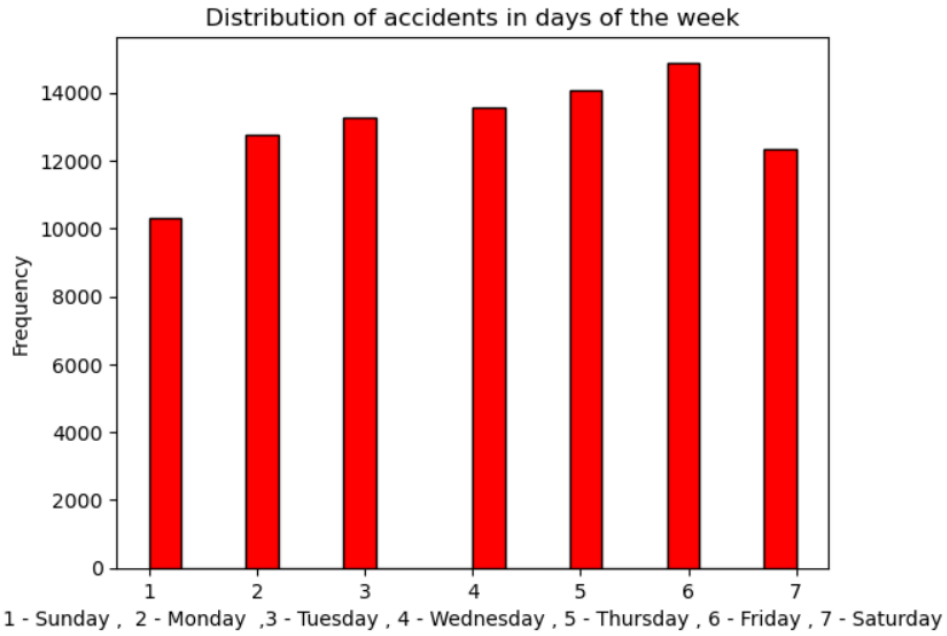

Figure 1 Histogram with the accidents during the days of the week.

The frequency of reported accidents exhibits a slight uptick on Fridays, amounting to 16.3%. Although this increase is modest in comparison to other days of the week, wherein percentages fluctuate between 11% and 15%, it nevertheless underscores a noteworthy trend [2]. Specifically, it draws attention to a surge in reported accidents towards the conclusion of the workweek. This rise could be attributed to various factors, including the onset of the weekend and individuals hastening to their respective destinations.

Analysis of the temporal distribution of recorded accidents reveals a distinct spike during the latter portion of the workday, encompassing the time span from 15:30 to 18:00, as depicted in Figure 2. Notably, the pinnacle of accident occurrences is observed at 17:30, registering a count of 862 incidents. To facilitate visualization, the timetable underwent minor adjustments, converting the data from object type to decimal format.

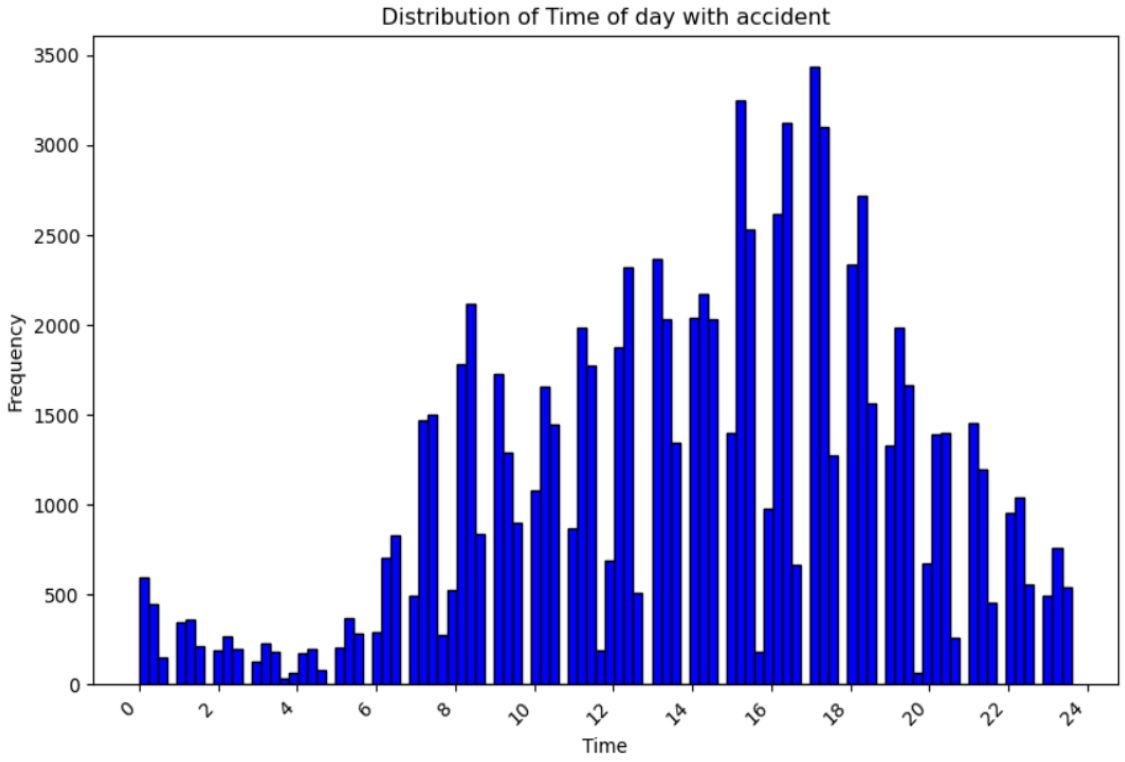

Figure 2 Accidents as they are reported throughout the day.

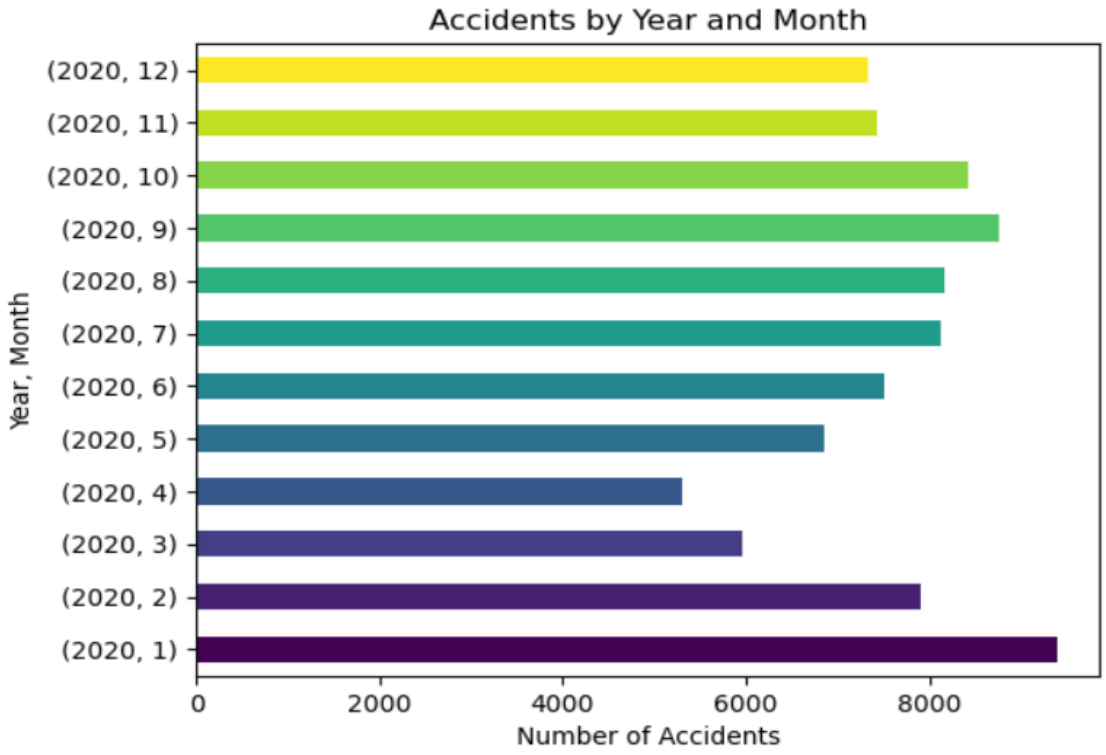

Figure 3 Accident distribution by the month during the year 2020, accidents peaks in the month of January as people are resuming their duties after the holidays.

Upon examining the different kind of categories of motorcycles listed in the data base it was clear that motorcycles over 50cc and up to 125cc were the most involve in accidents. The results also mirrored that of general accidents that occurred with majority occurring on Friday which is the last working day of the week and start of the weekend. The time of the day also replicates those results with the peak time being 18:00 these are displayed in the below figures 4,5.

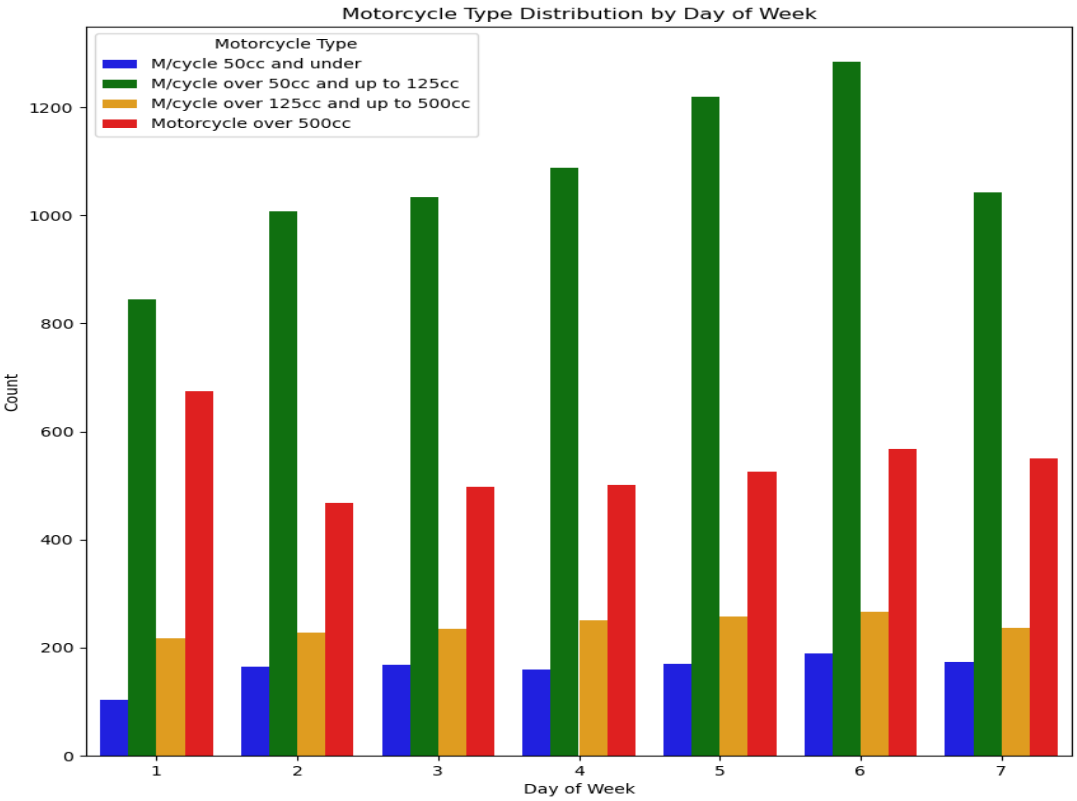

Figure 4 Visualization comparison for accidents of different types during days of the week.

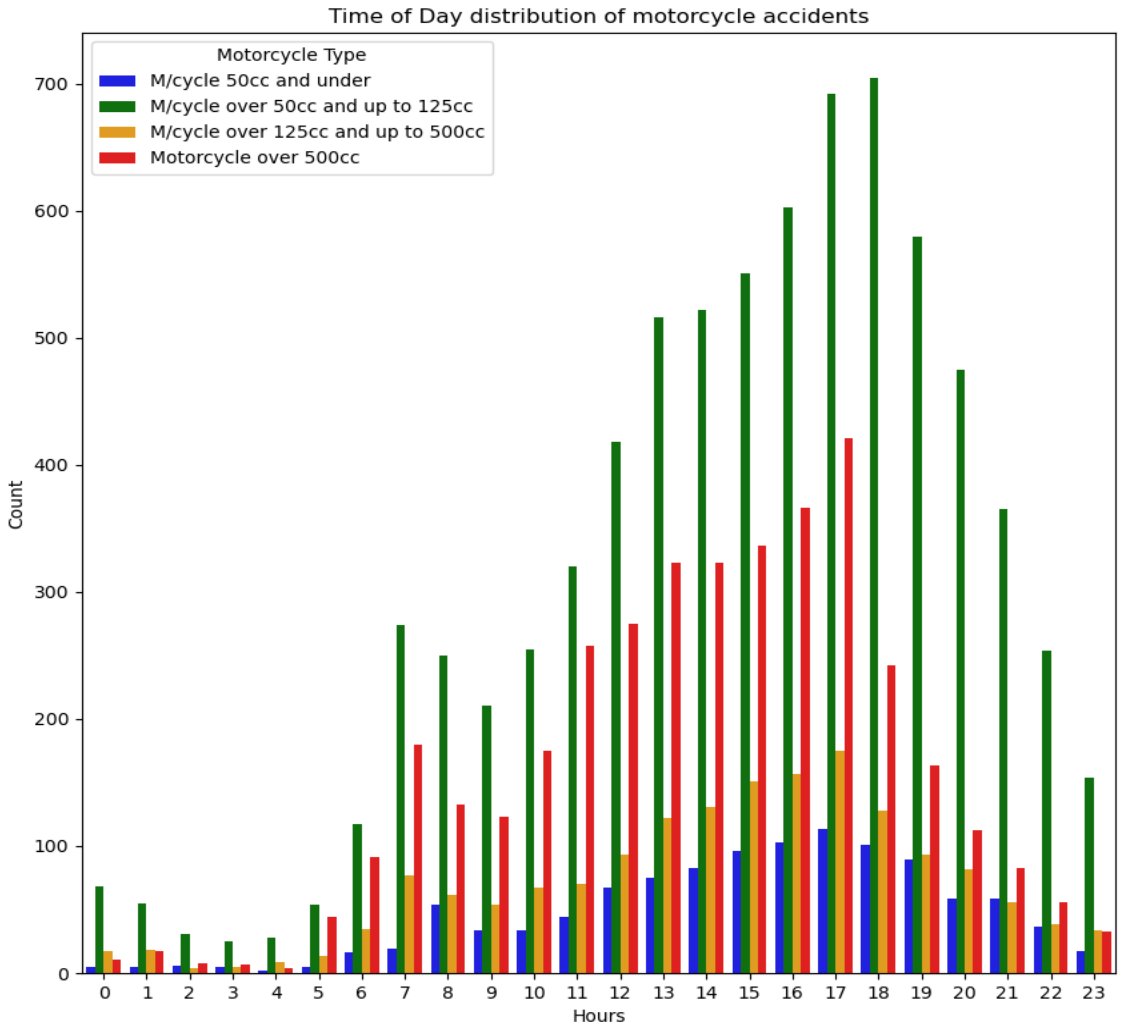

Figure 5 The time of day with different types of motorcycle accident.

In figure 6 I have explored the relationship between the motorcycles accident and the junction detail. The pattern represented indicates that type 3 motorcycles tend to be more commonly involved in accidents compared to other types. Regarding junction details related to accidents involving motorcycles, junction detail 3, identified as T or staggered junction, emerges as the most frequent. This junction configuration holds the highest frequency across all motorcycle types, barring type 5. This suggests that T or staggered junctions might be associated with greater risk or prevalence compared to other junction types this supported by the documentation released by the government statistics [3].

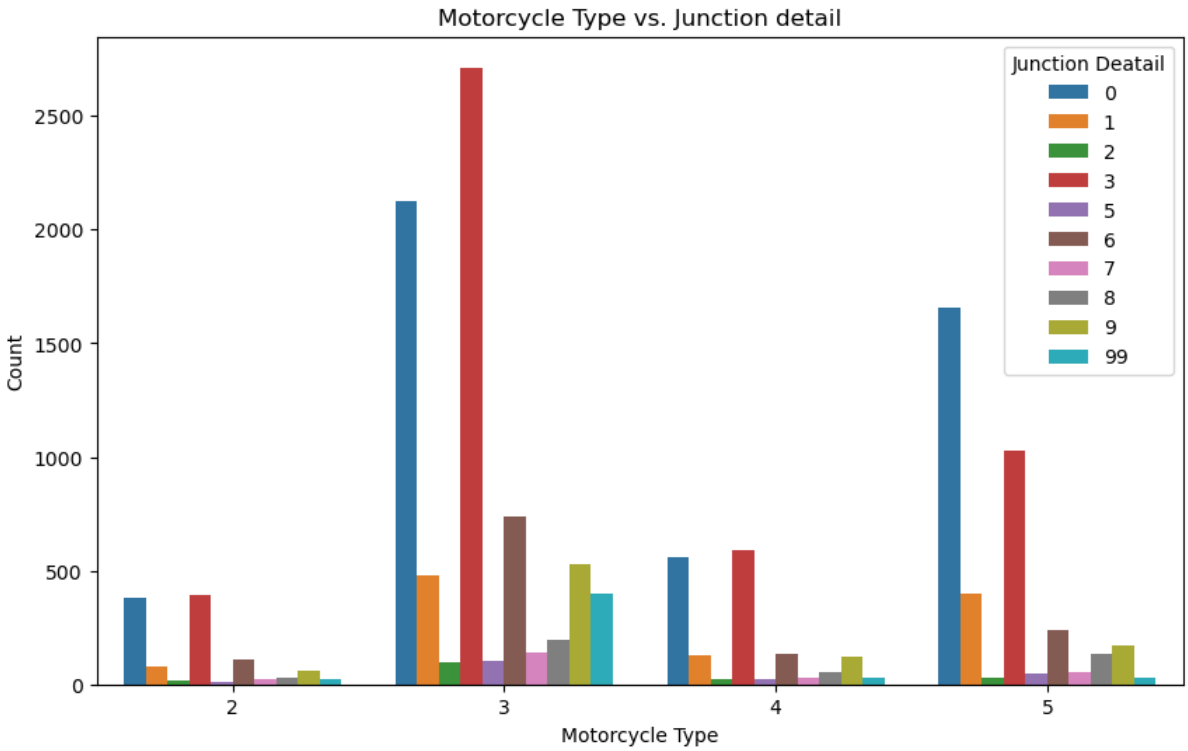

Figure 6 Representation of the different motorcycle types involved in accidents and the junction details, T or staggered junction represent the most risks

Conversely, the less common junction detail in motorcycle-related accidents is 8, characterized as a private drive or entrance. This specific junction type demonstrates the lowest frequency across all motorcycle types, except for type 5. This finding implies that private drive or entrance junctions are relatively less hazardous or less frequently encountered compared to other junction types.

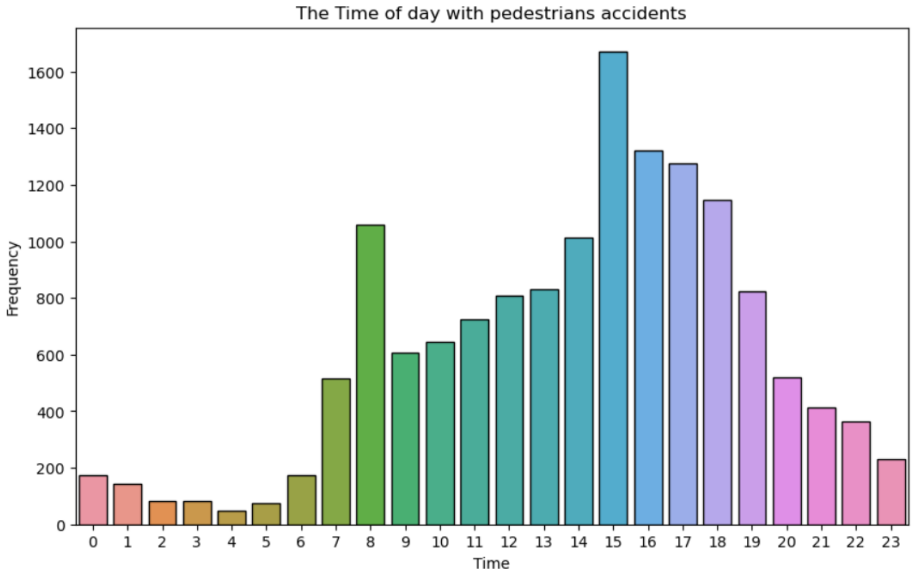

Figure 7 Time of the day with most pedestrian accidents, the peak time of recorded accidents is 15:00.

Moving on with the analysis and visualization the next part of the analysis this was finding out the association between the time of day and the day of the week where pedestrians are usually involved in accidents. The results revealed that the highest number of recorded accidents where pedestrians are involved tend to occur at 15:00 as it is shown in figure 7, and similarly the day where those occurred was slightly higher on Friday. Upon on a further investigation of the pedestrian ages revealed that those accidents occurring at 15:00 was mainly of teenagers or people below the age of 14 years this attributed to the end of the school day [4] and is represented in the figure 10 below.

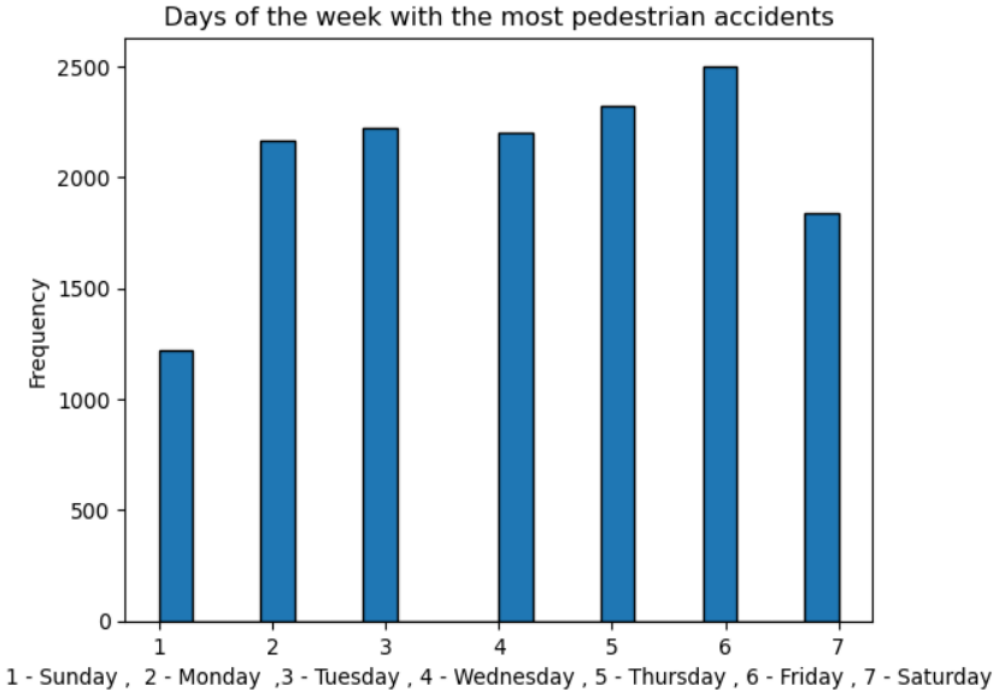

Figure 8 Representation of the day of the week with the most pedestrian accident, it shows a slight increase on Friday in comparison with the rest of the week.

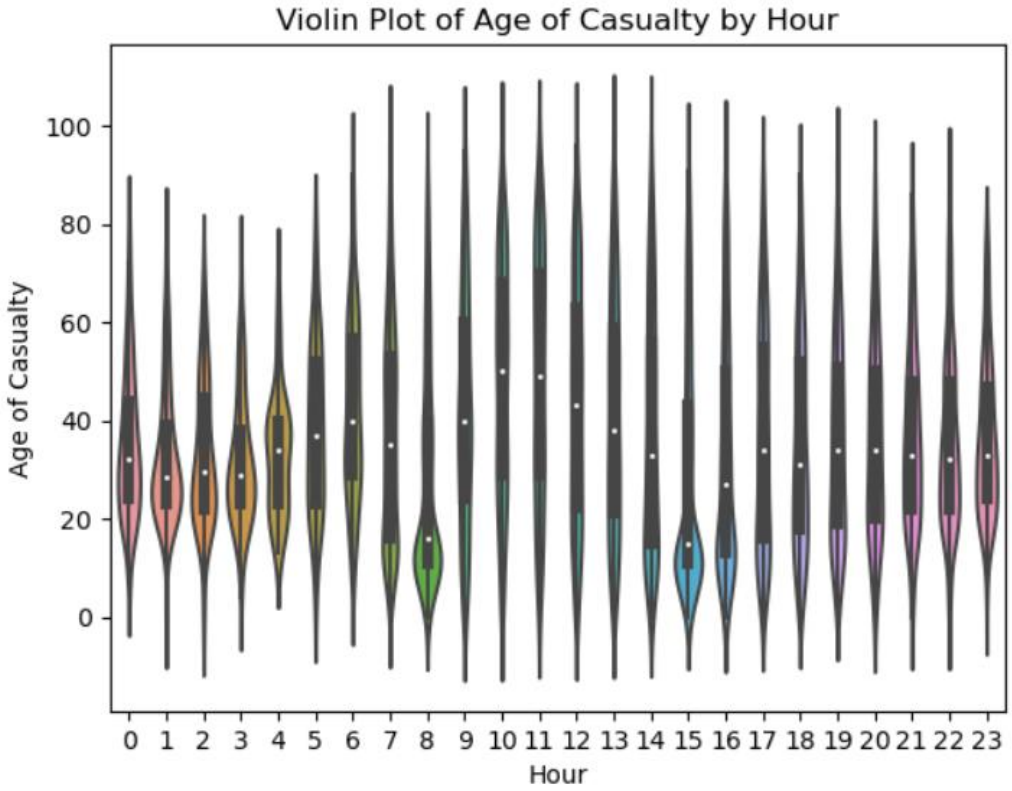

Figure 9 The age of pedestrian involved in accident throughout the day, the peak time of these accidents showed higher number of minor being involved.

In the subsequent phase of the analysis, we employed the apriori algorithm to investigate the influence of various factors within the dataset on accident severity. The findings, presented in Figure 10 with rules derived from the lift metric, reveal notable insights. The associated table indicates that accidents occurring under normal conditions predominantly are classified of slight severity. Among the most frequent itemset, the antecedents (light_1, speed limit_30, weather_1) and consequents (road_conditions_1, severity_3) exhibited the highest lift value of 1.32, with a confidence level of 0.716.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(light_1, weather_1, severity_3)	(road_conditions_1, speed limit_30)	0.453814	0.408093	0.241684	0.532563	1.305003	0.056486	1.266282	0.427910
1	(light_1, speed limit_30, weather_1)	(road_conditions_1, severity_3)	0.337314	0.539266	0.241684	0.716497	1.328651	0.059782	1.625143	0.373264
2	(road_conditions_1, severity_3)	(light_1, speed limit_30, weather_1)	0.539266	0.337314	0.241684	0.448173	1.328651	0.059782	1.200894	0.536876
3	(road_conditions_1, speed limit_30)	(light_1, weather_1, severity_3)	0.408093	0.453814	0.241684	0.592228	1.305003	0.056486	1.339442	0.394857

Figure 10 The apriori algorithm table is the result of setting the metric in the rule base to a lift of 1.3.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(road_conditions_1)	(weather_1)	0.688600	0.775577	0.646389	0.938700	1.210325	0.112327	3.661084	0.558047
1	(road_conditions_1, severity_3)	(weather_1)	0.539266	0.775577	0.504337	0.935229	1.205849	0.086095	3.464850	0.370515
2	(road_conditions_1, speed limit_30)	(weather_1)	0.408093	0.775577	0.382749	0.937896	1.209288	0.066241	3.613684	0.292390
3	(road_conditions_1, light_1)	(weather_1)	0.535439	0.775577	0.505456	0.944003	1.217162	0.090182	4.007762	0.384054
4	(road_conditions_1, speed limit_30, severity_3)	(weather_1)	0.327620	0.775577	0.306607	0.935864	1.206668	0.052513	3.499170	0.254724
5	(road_conditions_1, light_1, severity_3)	(weather_1)	0.421550	0.775577	0.396381	0.940295	1.212381	0.069437	3.758863	0.302838
6	(road_conditions_1, light_1, speed limit_30)	(weather_1)	0.317333	0.775577	0.299446	0.943634	1.216687	0.053330	3.981545	0.260882
7	(road_conditions_1, light_1, speed limit_30, s...	(weather_1)	0.256786	0.775577	0.241684	0.941192	1.213537	0.042527	3.816167	0.236759

Figure 11 The apriori algorithm table is the result of setting the metric in the rule base to a confidence of 0.9.

In order to investigate the outliers in the dataset I deployed a first I used the Local Outlier Factor (LOF) method on the UK region and tried to identify the urban and rural regions. The plot in figure 12 is the representation of outlier's detection analysis using LOF algorithm, and it classifies the outliers based on whether they belong to urban or rural areas.

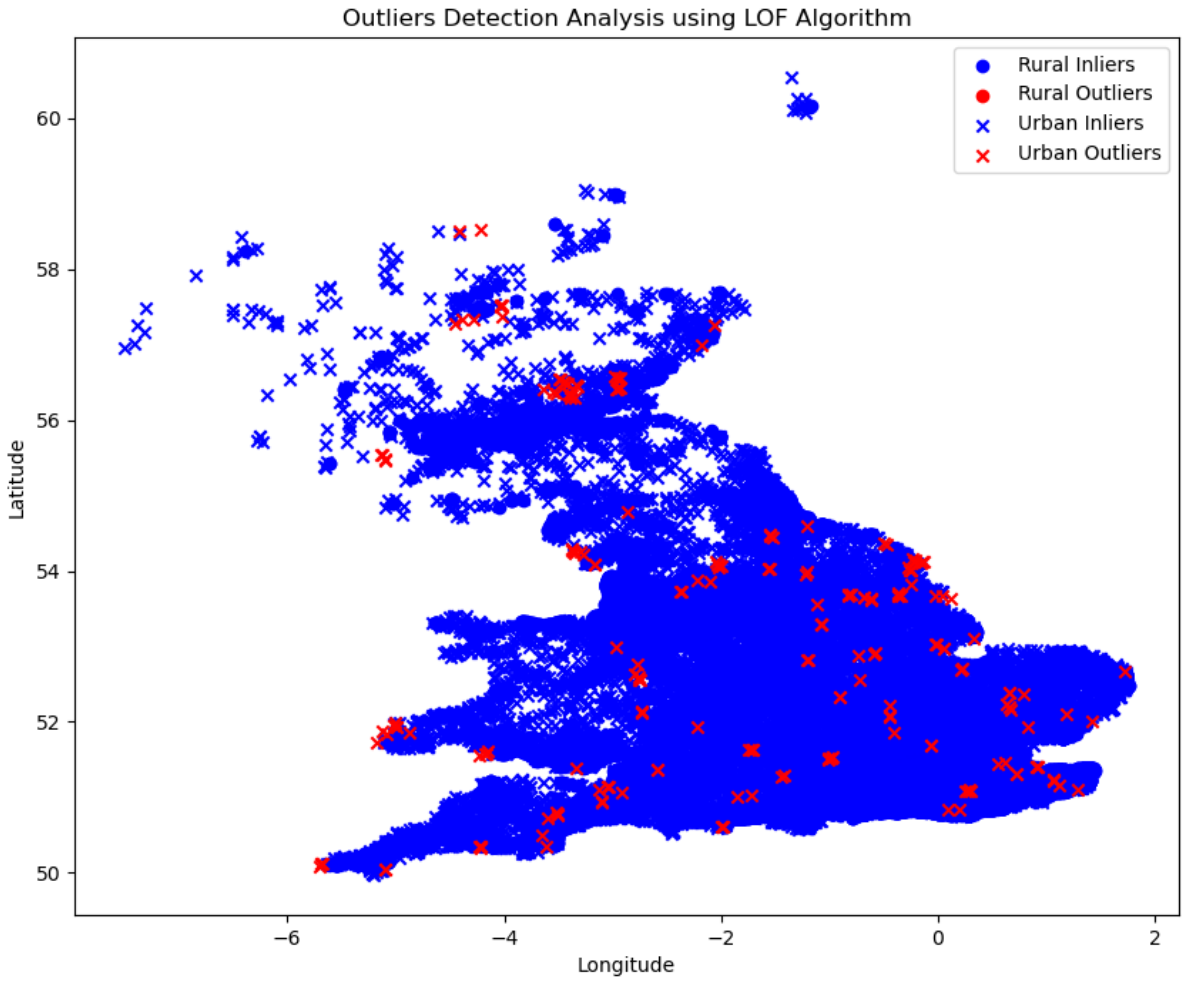

Figure 12 Representation the outlier detection in the UK, a higher number of urban outliers are represented in the plot.

It was observed from the LOF detection a higher representation of the urban outliers, and these are concentrated around major UK cities such as Greater London, Birmingham, and Greater Manchester in the North West[5]. These regions have the highest population density in the country and therefore has larger number of accidents occurrences due to traffic flow, network structure, public transportation and other factors. The plot displays a lower representation of outliers in the regions of Scotland and Wales this could be attributed to the lower population density, also fewer factors that lead to abnormal accidents in those areas. Another point is that the LOF algorithm is less sensitive to detect the low number of data points. Outliers sometimes provide insight into unique road safety challenges within a certain area. Therefore, my approach would be to retain these points as they could facilitate a more comprehensive analysis and help in designing targeted safety strategies.

The next phase of the analysis concentrates on the distribution of accidents within the Kingston upon Hull regions. This task was accomplished by utilizing k-means clustering based on the severity of accidents [6]. To narrow down the area, data from the police force was employed, specifically focusing on the entry labelled as 16, which designates the specific target region. The most suitable number of clusters was determined to be 9. This

determination was also validated using the silhouette method, yielding a result of 0.6243. This value implies a reasonable distinction between the various clusters. The clustering plot depicting this arrangement is presented in Figure 13.

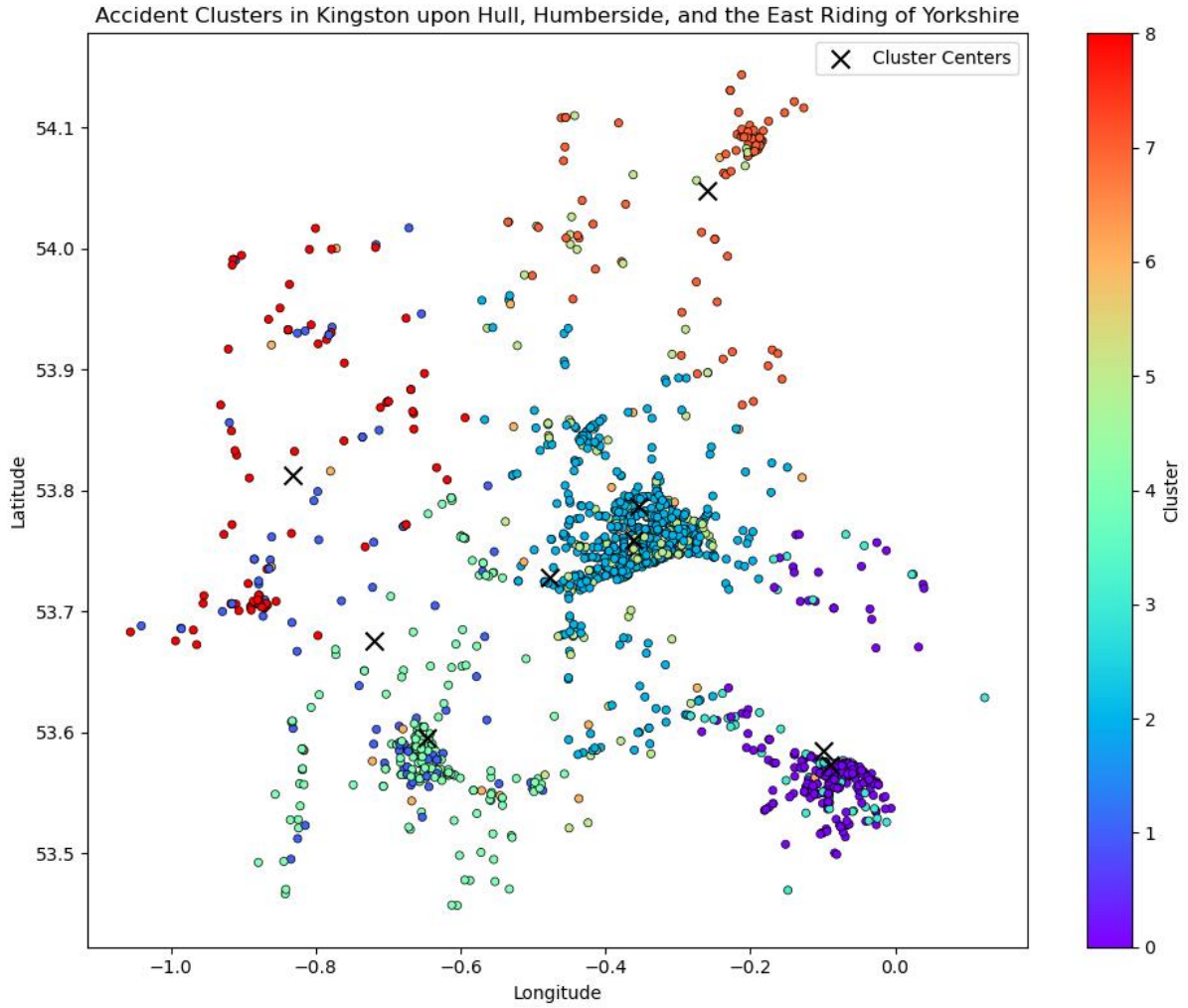

Figure 13 The clustering of accident in the region of Hull, this displayed a higher density of accidents around the city center region.

Among the clusters, the largest one is labelled as number 2. It encompasses the vicinity of Anlaby Road, Ferensway, Hessle Road, and Holderness Road. This area stands as the central hub of Hull and experiences substantial commuter traffic. However, accidents within this region are generally minor in their impact. The second most sizable cluster, identified as number 0, covers the region encompassing Princes Avenue, Spring Bank West, Queens Road, and Beverly. This area is known as a commercial district, which helps explain the higher frequency of accidents.

The third largest cluster, numbered 4, corresponds to the area of Bridlington and the southern coastal region of Hull. Notably, accidents within this cluster tend to exhibit greater severity compared to the previously mentioned regions. This could be attributed to the presence of larger motorways in this vicinity. The remaining clusters are comparatively

smaller in size when contrasted with the three aforementioned ones. They also display a more scattered pattern, lacking any discernible structure. These smaller clusters represent small towns or rural areas.

I've developed a stacking classifier for my study, which involves combining several basic classifiers and a meta classifier to enhance prediction performance. The basic classifiers include decision tree, k-nearest neighbours, Gaussian naive Bayes, support vector machine, and random forest. The meta classifier, in this case, is logistic regression [7].

Classification Report:

	precision	recall	f1-score	support
False	0.65	0.74	0.69	343
True	0.71	0.62	0.66	353
accuracy			0.68	696
macro avg	0.68	0.68	0.68	696
weighted avg	0.68	0.68	0.68	696

Figure 14 Classification Report of the Stacked model, the model produced an accuracy of 68%.

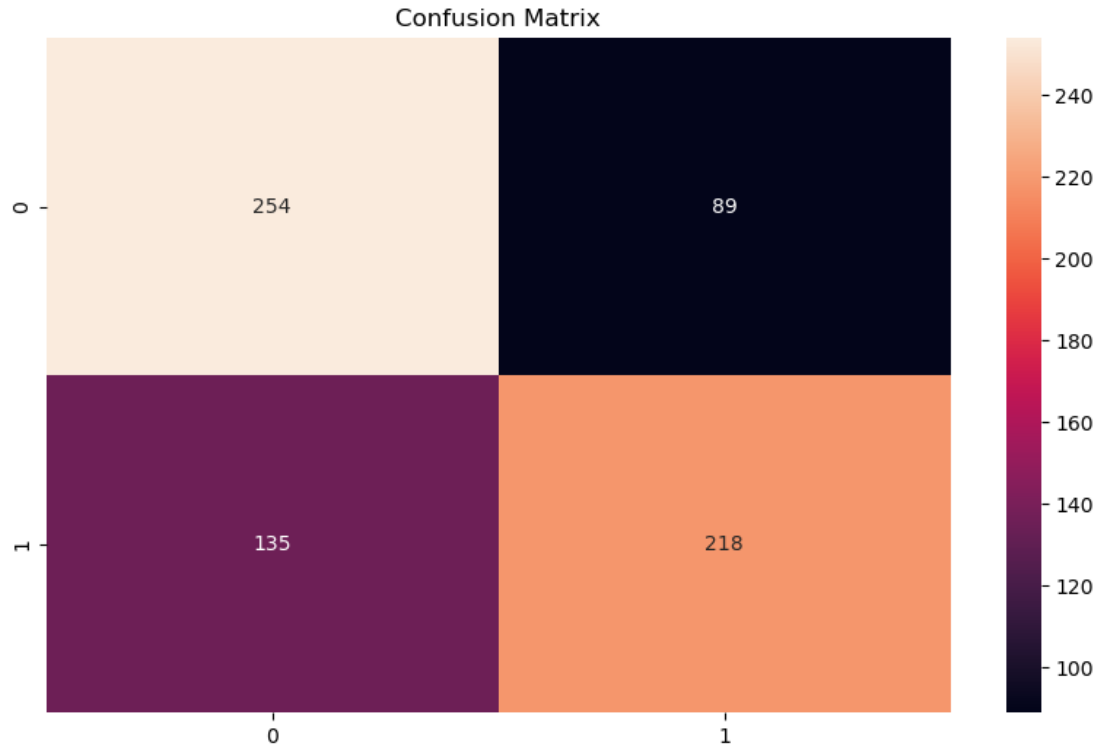

Figure 15 The confusion matrix representing the stacking model to make predictions about accident severity.

The outcomes of the model are presented through a classification report and a confusion matrix displayed in figure 14 and 15.

The model achieves an accuracy of 0.68, indicating correct predictions for 68% of the test cases. A macro average f1-score of 0.68 is attained, reflecting a well-balanced performance across both classes. The True class demonstrates higher precision (0.71) compared to the False class (0.65), indicating fewer incorrect predictions for the True class. For the False class, a greater recall (0.74) is observed in contrast to the True class (0.62), indicating fewer instances where the model incorrectly predicts the False class. Similar f1-scores are achieved for both classes (0.69 and 0.66), signifying a balanced trade-off between precision and recall for each class. The confusion matrix reveals 254 TN, 89 FP, 135 FN, and 218 TP. This implies that out of the 696 cases in the test set, the model accurately predicts 254 cases as non-fatal and 218 cases as fatal. However, it incorrectly predicts 89 cases as fatal and 135 cases as non-fatal.

Recommendations

In conclusion, based on the conducted analysis, I recommend the adoption of the following strategies to enhance road safety measures:

Given the noticeable surge in reported accidents, particularly on Fridays and during the final hours of the day, spanning from 15:30 to 18:00, it is advisable to bolster the presence of law enforcement personnel within major urban centers. These areas are important for commuting and boasting heavy traffic flux along with commercial zones, require heightened surveillance to ensure public safety.

The age demography implicated in pedestrian-related accidents underscores the involvement of younger individuals, often students commuting to and from schools. Collaborating with educational institutions to administer awareness sessions on road safety help in rectifying this predicament. Reinforcing signage regions with higher pedestrian accidents serves as an additional measure to diminish risks and uplift overall safety.

Developing a real-time system offering instant updates on traffic and weather conditions proves to be instrumental. Employing mobile notifications, this system could suggest route alterations and recommended speed limits corresponding to the prevailing weather conditions. This proactive approach ensures that individuals make informed decisions for a safer journey.

Embracing the power of machine learning advancements, notably stacking models and neural networks, is imperative for predicting accident severity. Leveraging these robust tools aids in averting critical scenarios, ultimately safeguarding lives and curtailing expenditures.

By implementing these strategies, would enhance the road safety, and help in curbing accidents.

References:

[1] Department for Transport. Reported road casualties Great Britain, provisional results: 2020 [Internet]. 2021 Sep 30 [cited 2023 Aug 9]. Available from:[https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-provisional-results-2020/reported-road-casualties-great-britain-provisional-results-2020]

[2] Simpson Millar. Friday afternoon is peak time for RTAs [Internet]. Manchester: Simpson Millar LLP Solicitors; 2019 [updated 2019 Nov 22; cited 2023 Aug 10]. Available from: [https://www.simpsonmillar.co.uk/media/road-traffic-accidents/friday-afternoon-is-peak-time-for-rtas/]

[3] Department for Transport. Reported road casualties in Great Britain: motorcycle factsheet 2020 [Internet]. 2021 Sep 30 [cited 2023 Aug 9]. Available from: [https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-motorcyclist-factsheet-2020/reported-road-casualties-in-great-britain-motorcycle-factsheet-2020]

[4] Think Student Editor. What time do primary schools start and finish in the UK? [Internet]. London: Think Student; 2021 [updated 2021 May 17; cited 2023 Aug 11]. Available from: [https://thinkstudent.co.uk/what-time-do-primary-schools-start-and-finish-in-the-uk/]

[5] The Editors of Encyclopaedia Britannica. Reference frame [Internet]. Chicago: Encyclopaedia Britannica, Inc.; 2023 [updated 2023 Jul 7; cited 2023 Aug 11]. Available from: [https://www.britannica.com/science/reference-frame]

[6] Anderson TK. Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis & Prevention. 2009 May 1;41(3):359-64.

[7] Tang J, Liang J, Han C, Li Z, Huang H. Crash injury severity analysis using a two-layer Stacking framework. Accident Analysis & Prevention. 2019 Jan 1;122:226-38.