ACCIDENT DATA REPORT

INTRODUCTION

This report provides a detailed analysis of accidents within the UK in 2020. The analysis is based on information derived from an accident database containing records of each accident with geographic co-ordinates, casualties, vehicle details and a lower support output area record. The objective is to answer specific questions regarding accident occurrence rate, make insightful recommendations on best actions to improve road safety and build a classification model to help classify the severity of an accident.

To ensure accuracy, errors ranging from missing values, wrong entries, etc were treated and details of treatment method used are documented within this report.

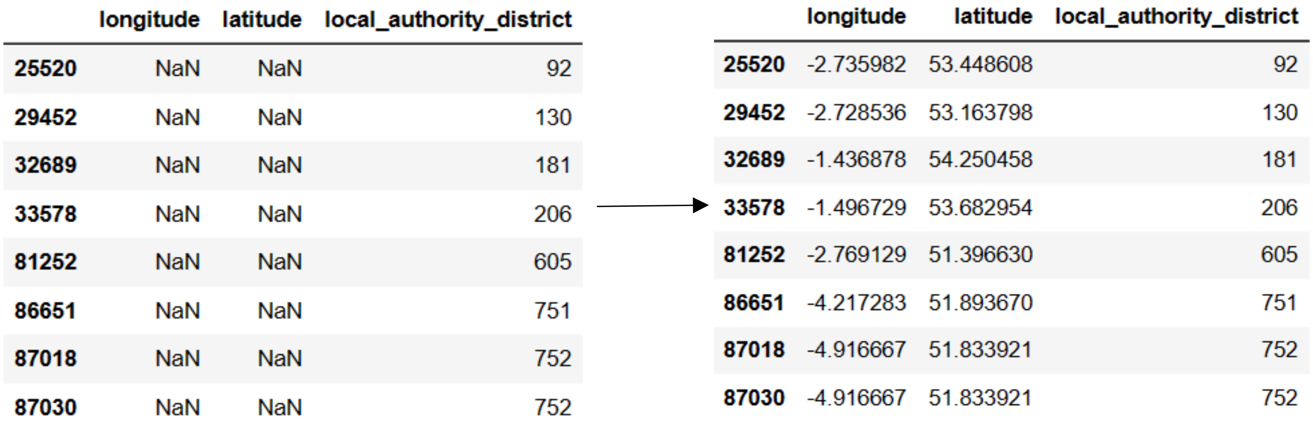
Detailed visualization and analysis will be used to arrive at recommendations in the latter part of this report.

ANALYSIS

Data Cleaning and Preprocessing

Latitude and Longitude:

Region names corresponding to the local authority district were obtained the official road safety data catalogue (Gov.uk, 2021). The null values in longitude and latitude were derived using python geocode library to determine the real-time geographical co-ordinates for the corresponding local authority district value.



	longitude	latitude	local_authority_district
25520	NaN	NaN	92
29452	NaN	NaN	130
32689	NaN	NaN	181
33578	NaN	NaN	206
81252	NaN	NaN	605
86651	NaN	NaN	751
87018	NaN	NaN	752
87030	NaN	NaN	752

	longitude	latitude	local_authority_district
25520	-2.735982	53.448608	92
29452	-2.728536	53.163798	130
32689	-1.436878	54.250458	181
33578	-1.496729	53.682954	206
81252	-2.769129	51.396630	605
86651	-4.217283	51.893670	751
87018	-4.916667	51.833921	752
87030	-4.916667	51.833921	752

Null values for entries with missing local authority district values were inferred from their local authority ONS district value (both the co-ordinates and local authority district were filled using this approach).

Entries with Missing (-1) values:

For speed limit, the type of road impact on the speed limit assigned to it (Aura Insurance, 2022). Hence, the speed limits were filled with reference to the UK government's assigned speed limit for various road type (Gov.uk, no date). Where speed limits vary for specific or unknown road types, the "-1" values were filled with the most occurring speed limit for same road type in the data.

For light condition, value of 7, corresponding to the time of accident (01:15am), was inputted according to stat-19 document. The weather condition for the region at 01:15am on 24/06/2020 was found to be "1" (Fine no wind), and filled accordingly (Timeanddate, no date).

Since age band and vehicle details of the entries with missing (-1) values are unknown, it's almost impossible making inference and inputting a wrong age value would constitute a high bias due to the number of missing values. Moreover, a driver's and vehicle's age have no impact on the accident severity, hence they were filled with NaN.

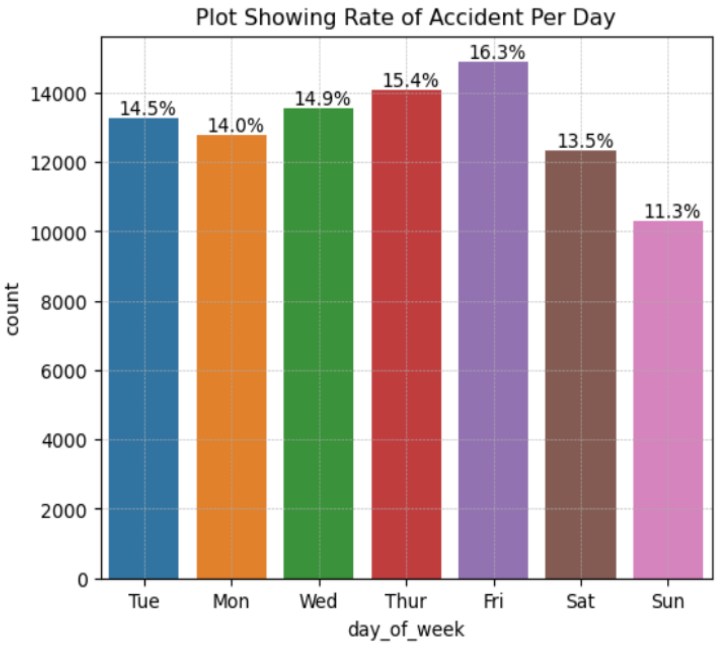
Dates were converted into pandas datetime, and the day of week was converted from the week number representation into the name of weekday (Mon, Tue, etc.).

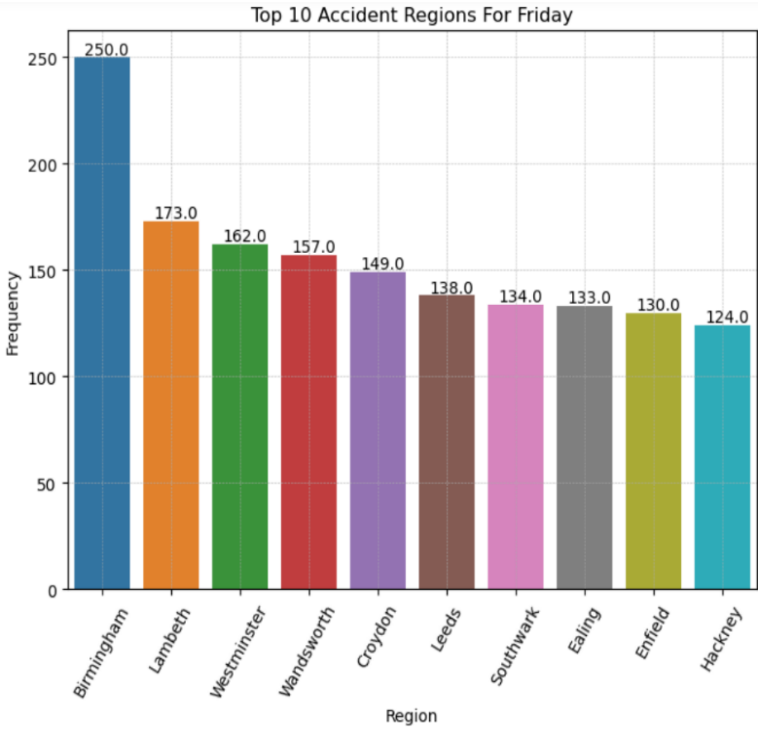
The region names were extracted from the "lsoa" table and joined to the accident table for easy identification of accident regions.

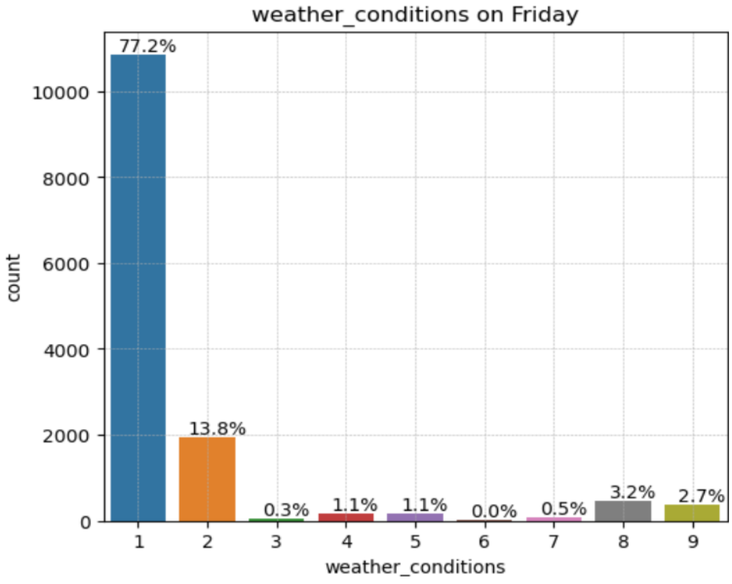
To determine time and conditions of accident, new data containing accident time and frequencies were deduced. The top regions where accidents occurred and weather conditions (on weekday with highest accident rate) were also obtained.

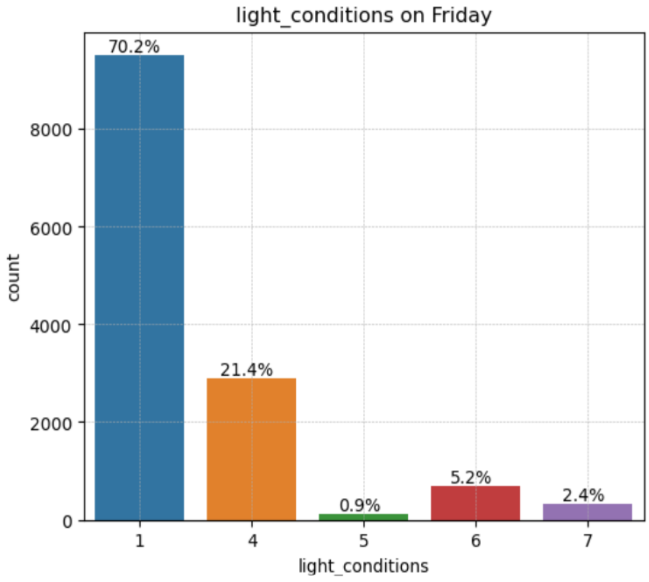
Question 1:

ACCIDENT SIGNIFICANT DAYS OF THE WEEK


Plot Showing Rate of Accident Per Day

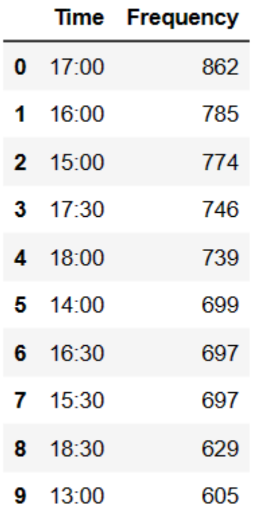

Top 10 Accident Regions For Friday


weather_conditions on Friday

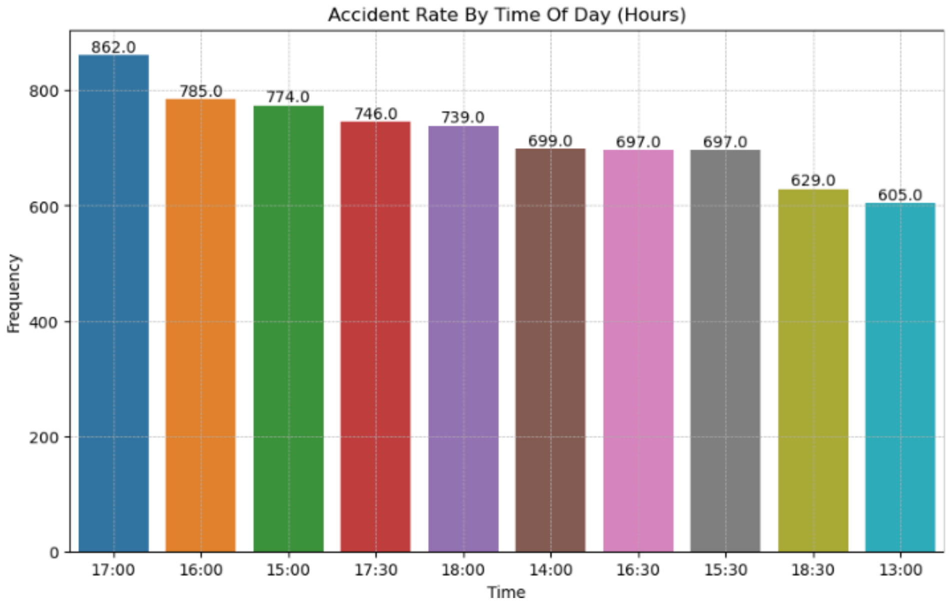

light_conditions on Friday

Although the weather and light conditions on Friday were mostly fine with no winds (77.2%) and daylight (70.2%) respectively, a larger percent (16.3%) of the accident occurred on Friday and Birmingham had the highest accident rate on Fridays than other regions in the UK with a total of 250 accidents.

ACCIDENT SIGNIFICANT HOURS:



	Time	Frequency
0	17:00	862
1	16:00	785
2	15:00	774
3	17:30	746
4	18:00	739
5	14:00	699
6	16:30	697
7	15:30	697
8	18:30	629
9	13:00	605

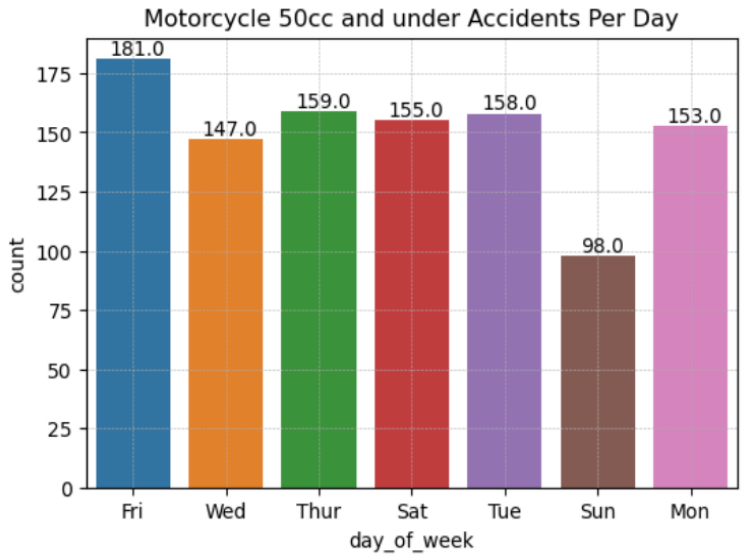


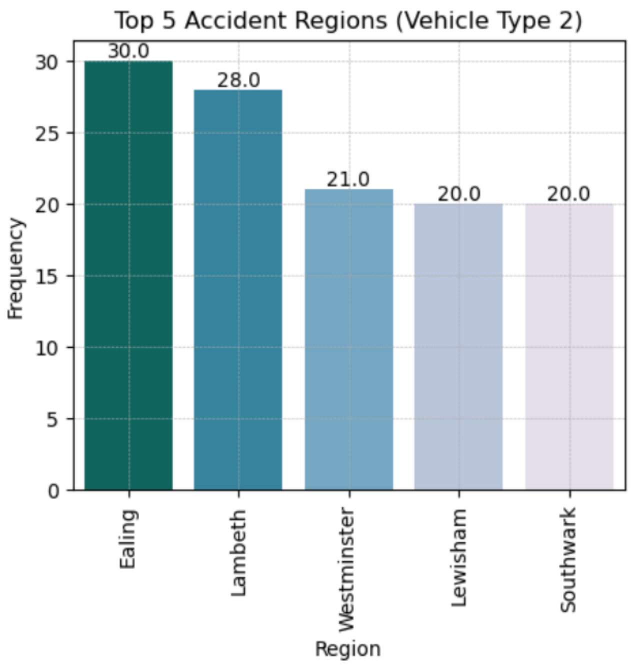
From the top ten accident occurrence rate, most of the accidents recorded in the data happened between 13:00 and 18:30 hours of the day with the highest at 17:00 hours (5pm) with 862 accidents.

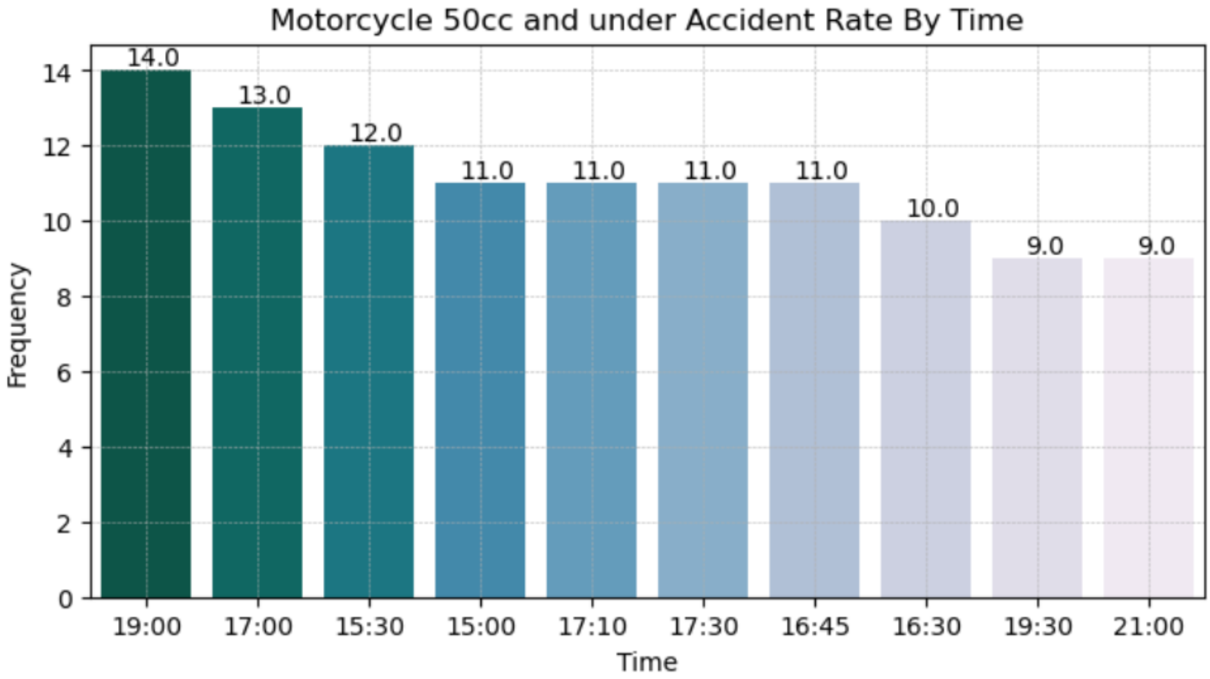
QUESTION 2

SIGNIFICANT DAYS AND HOURS FOR MOTORBIKES:

Motor 50cc and under:

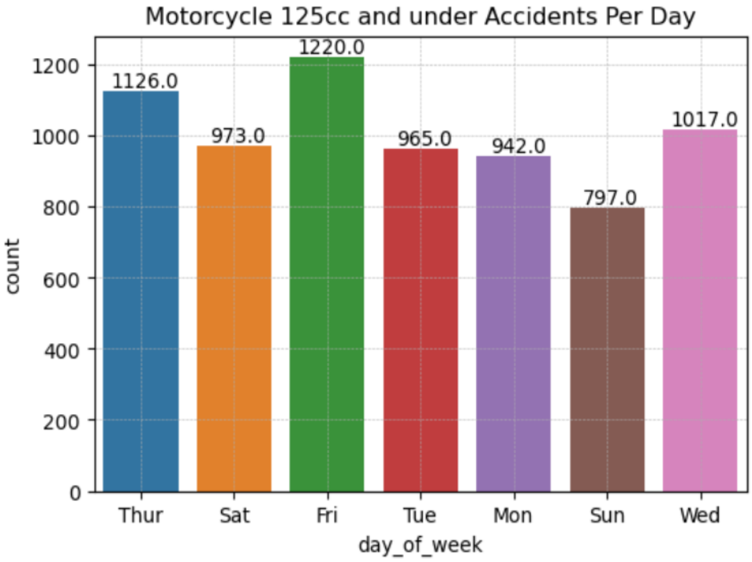


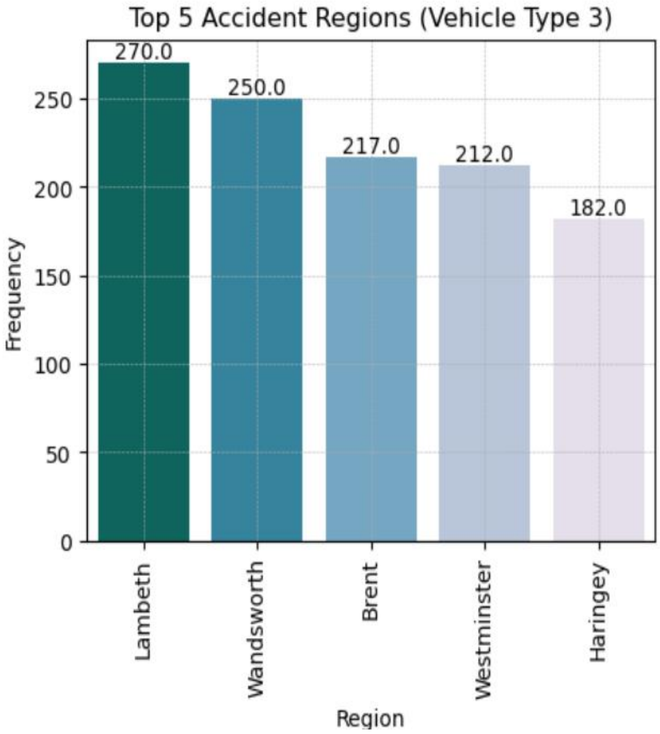


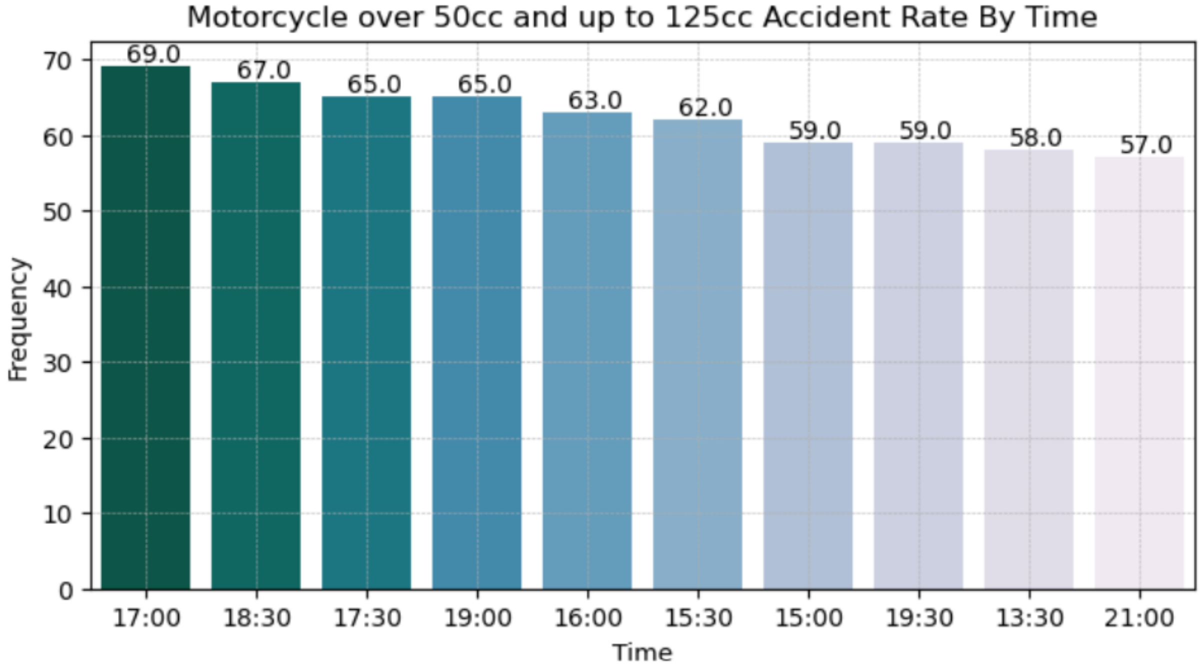


Most accidents happened on Friday (181) with the highest occurrence region in Ealing. Also, the top 5 accident hourly occurrence rates were between 15:00 and 19:00 hours with the highest at 19:00 hours (8pm) with 14 accidents.

Motorcycle over 50cc up to 125cc:

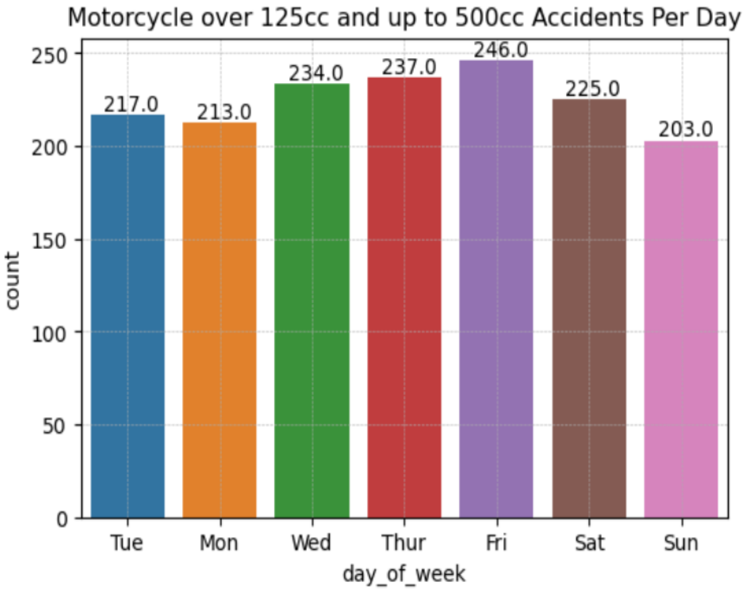


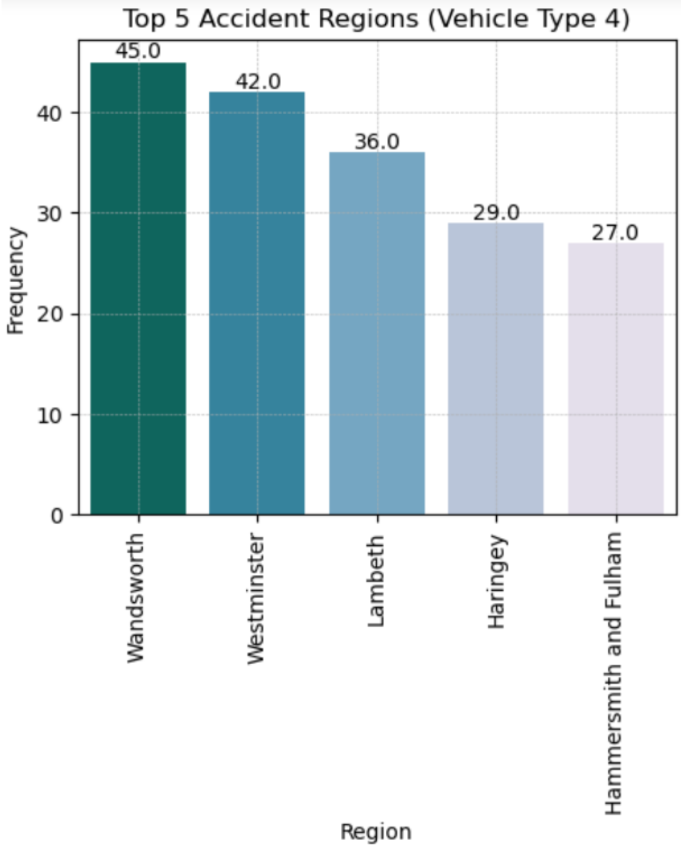


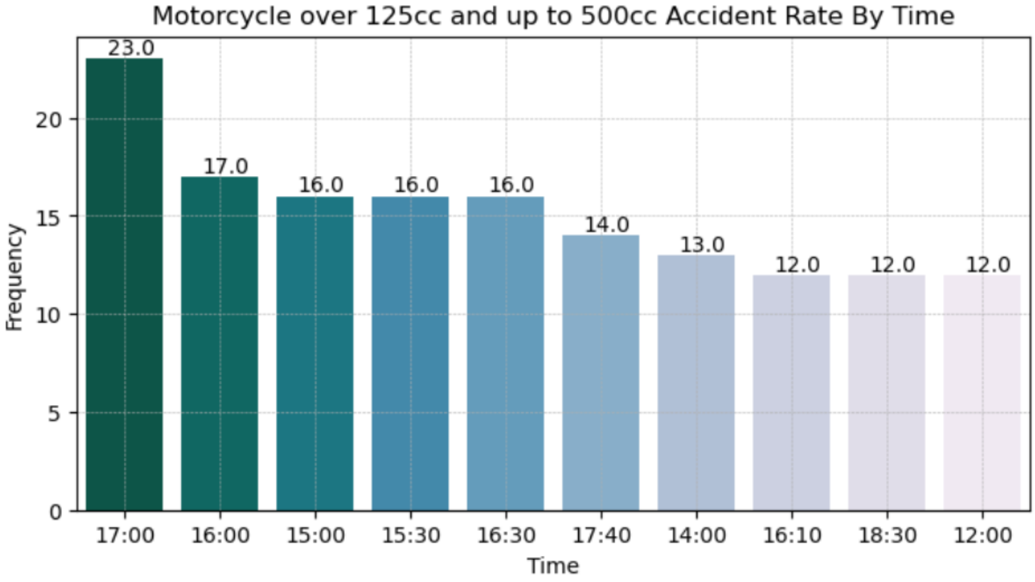


Most accidents by “motorcycle over 50cc up to 125cc” happened on Friday (1220) with the highest occurrence in Lambeth. Also, the top 5 accident hourly occurrence rates were between 16:00 and 19:00 hours with the highest at 17:00 hours (5pm) with 69 accidents.

Motorcycle over 125cc and up to 500cc:

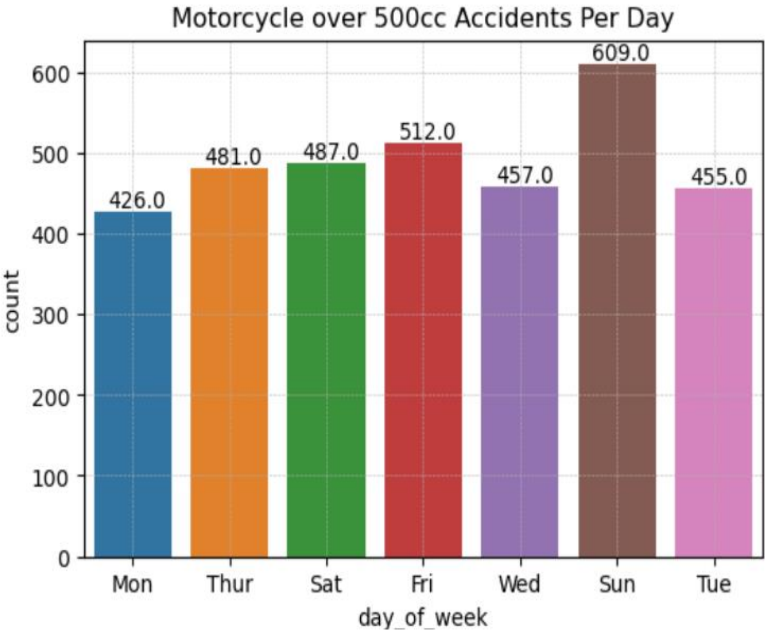


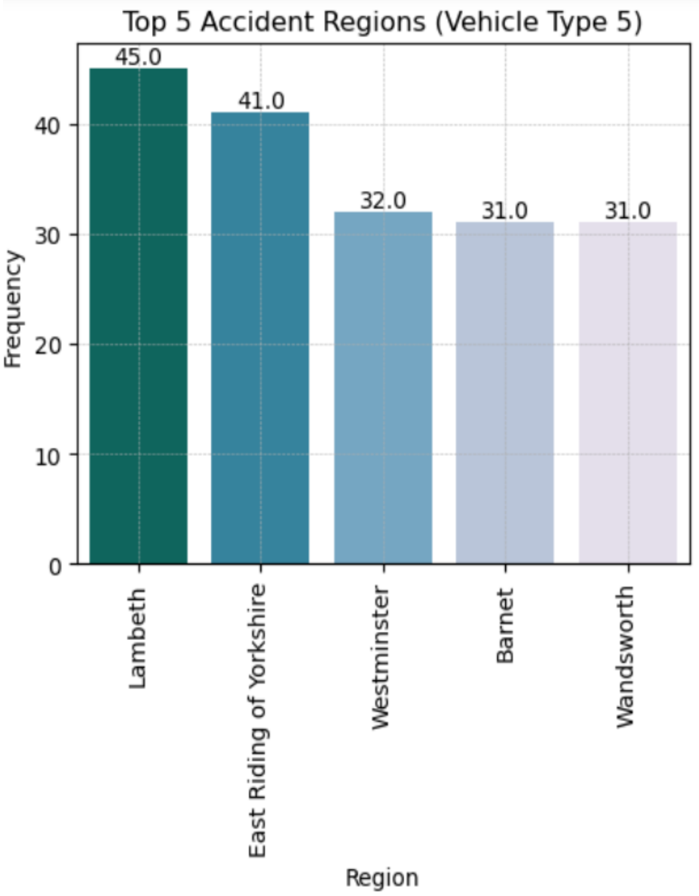


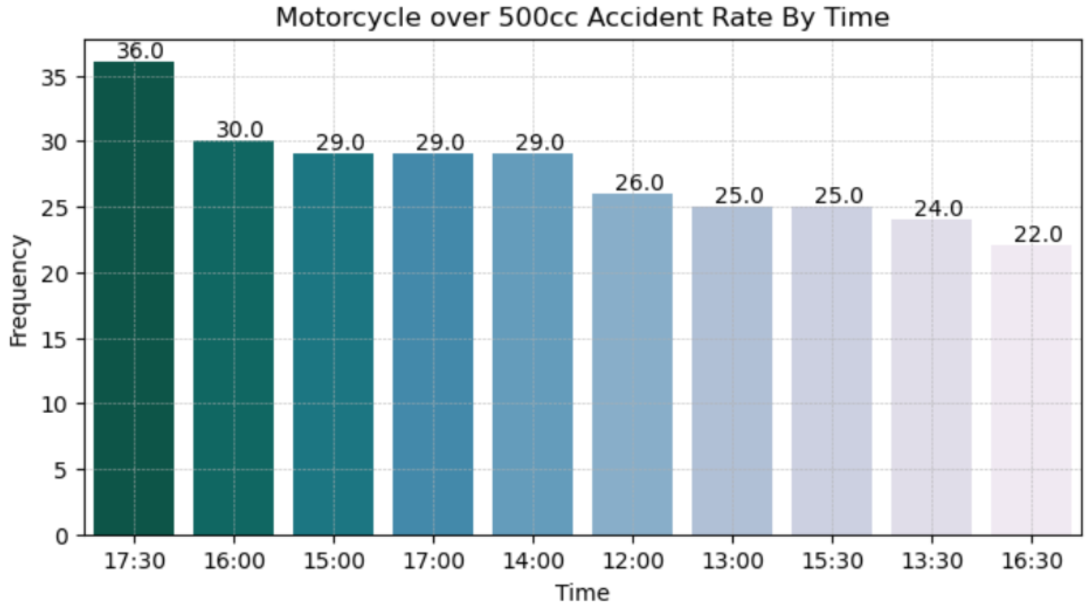


From the image analysis above, most of the accidents by vehicle type “Motorcycle over 125cc and up to 500cc” also happened on Friday (246) with the highest occurrence region in Wandsworth. Also, the top 5 accident hourly occurrence rates were from 15:00 and 17:00 hours with the highest at 17:00 hours (5pm) with 23 accidents.

Motorcycle over 500cc:

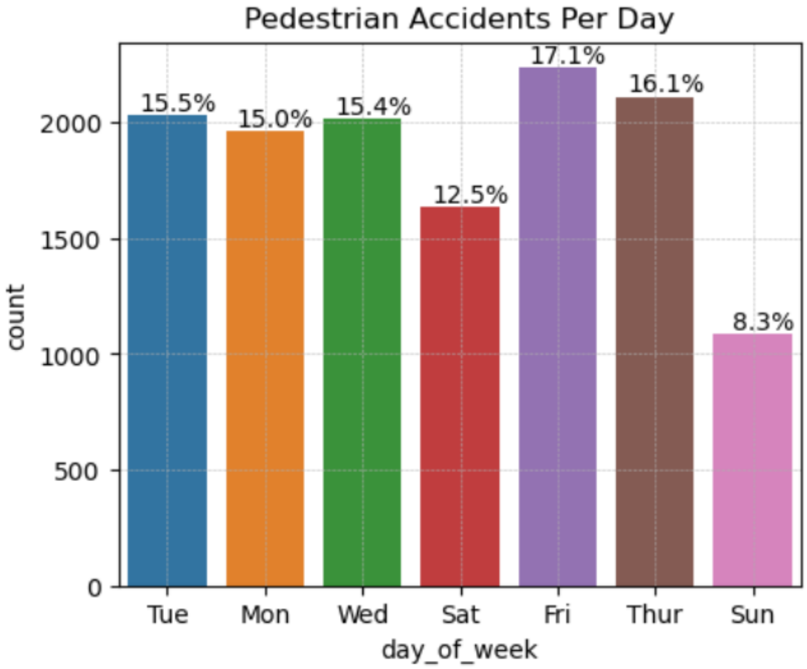


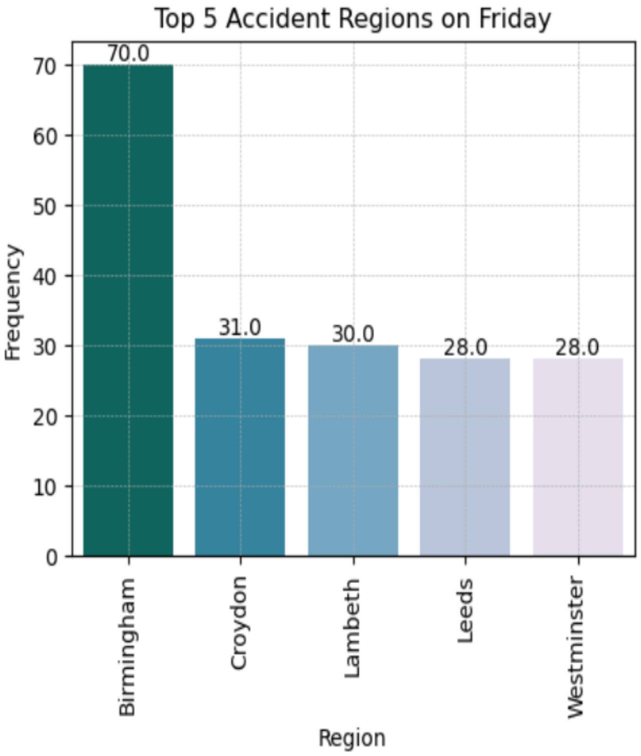


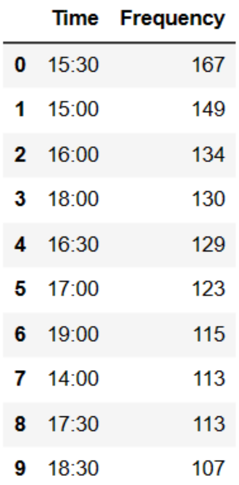


Most of the accidents by vehicle type “Motorcycle over 500cc” happened on Sunday (609) with Lambeth as the highest occurrence region. The top 5 accident hourly occurrence rates were from 14:00 and 17:30 hours with highest at 17:30 hours (5:30pm) with 36 accidents.

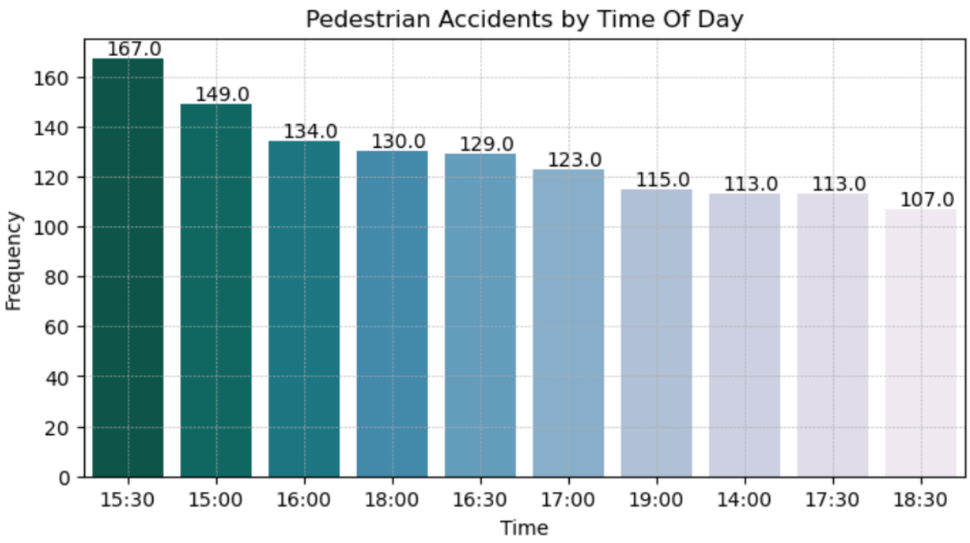
QUESTION 3







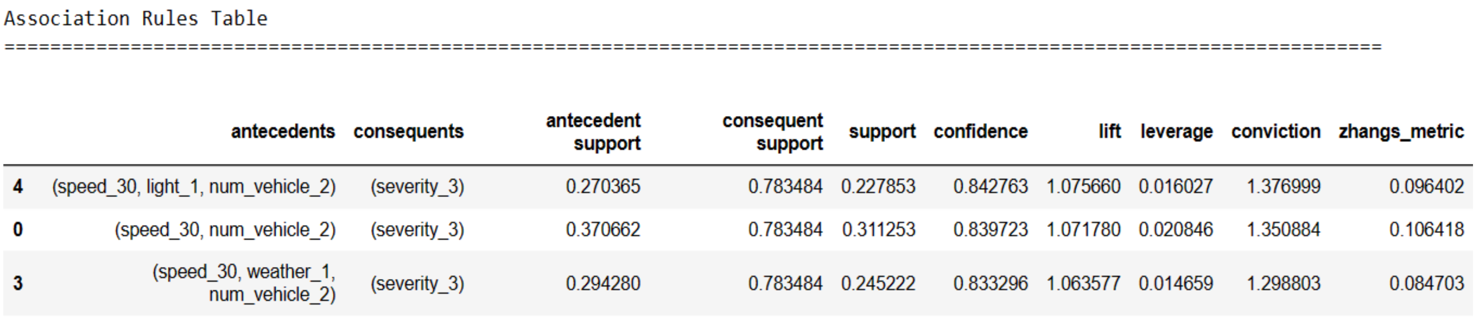
	Time	Frequency
0	15:30	167
1	15:00	149
2	16:00	134
3	18:00	130
4	16:30	129
5	17:00	123
6	19:00	115
7	14:00	113
8	17:30	113
9	18:30	107



Pedestrians had more accidents on Friday (17.1% of accidents) compared to other days of the week and most of these accidents happened in Birmingham (70 accidents). Also, most accidents involving pedestrians occurred at 15:30 hours of the day.

QUESTION 4

Speed limit, weather, and light conditions, amongst others, were chosen due to their impact on accident severity (Gedlawyers, no date). The number of vehicles involved also affect the severity of an accident (Paraskevi et al, 2015). Support threshold of 0.2 was used to reduce the number of unreasonable associations due to the data size. Similarly, association rules for each itemset were generated using a confidence metric of 0.83. Associations with accident severity as the consequence were selected. The resulting association rules are represented below:



```
Association Rules Table
===================================================================================================
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
4	(speed_30, light_1, num_vehicle_2)	(severity_3)	0.270365	0.783484	0.227853	0.842763	1.075660	0.016027	1.376999	0.096402
0	(speed_30, num_vehicle_2)	(severity_3)	0.370662	0.783484	0.311253	0.839723	1.071780	0.020846	1.350884	0.106418
3	(speed_30, weather_1, num_vehicle_2)	(severity_3)	0.294280	0.783484	0.245222	0.833296	1.063577	0.014659	1.298803	0.084703

Summarized association rules from the table above:

Rule 1 = {speed_30, light_1, num_vehicle_2} ➔ {severity = 3}

Rule 2 = {speed_30, num_vehicle_2} ➔ {severity = 3}

Rule 3 = {speed_30, weather_1, num_vehicle_2} ➔ {severity = 3}

Lift: The lift for each rule is slightly above 1, indicating both antecedents and consequents are most likely to occur together (Braham, 2020, p. 267). Although the support (0.228) for rule 1 is lower, its confidence rate (0.843) is higher compared to other rules. Further analysis based on support and confidence and the impact of these associations on accident severity is done below:

Rule 1: {speed_30, light_1, num_vehicle_2} ➔ {severity = 3}

Support = 0.228: This indicates that the association rule occurs 22.8% of the time (lower compared to other rules). This means that rule 1 only has about23% probability of occurrence based on data entries.

Confidence = 0.843: This means 84.3% of the time when the light condition is 1 (daylight) and speed limit is 30, with 2 vehicles involved, the accident severity is 3 (slight). This value also explains that the associations are 84.3% reliable and accurate. This makes sense because during the day, a driver can clearly see oncoming danger and minimize collisions, most especially when driving at a low-speed limit of 30.

Rule 2: {speed_30, num_vehicle_2} ➔ {severity = 3}

Support = 0.311: This indicates that the association rule occurs 31.1% of the time, that is how often the association occur in the data.

Confidence = 0.8397: This means that 83.97% of the time when there're 2 vehicles and speed limit is 30, accident severity is 3 (slight). This sounds reasonable because during the day, an accident impact can also be avoided when a driver can clearly see oncoming danger and minimize collisions, most especially when driving at a low-speed limit of 30 and under fine weather conditions.

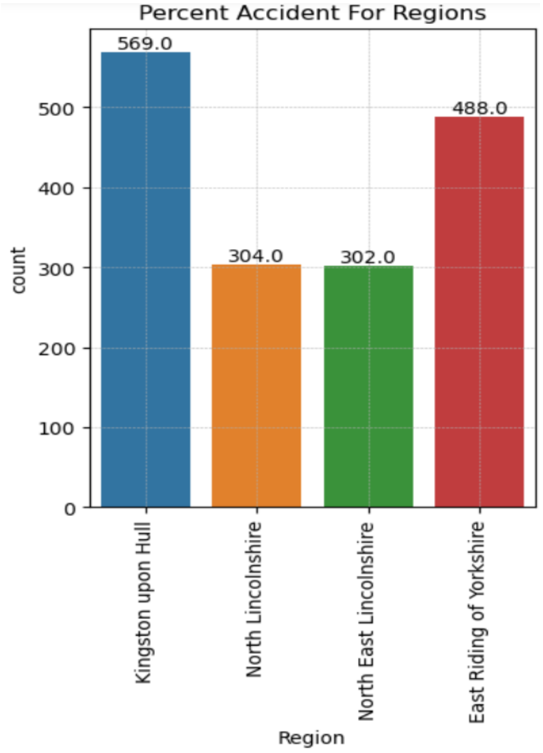
Rule 3: {speed_30, weather_1, num_vehicle_2} ➔ {severity = 3}

Support = 0.2452: This means the association rule happens 24.52% of the time, i.e., out of all data entries, a combination of both antecedent and consequence occurs about 24% of the time.

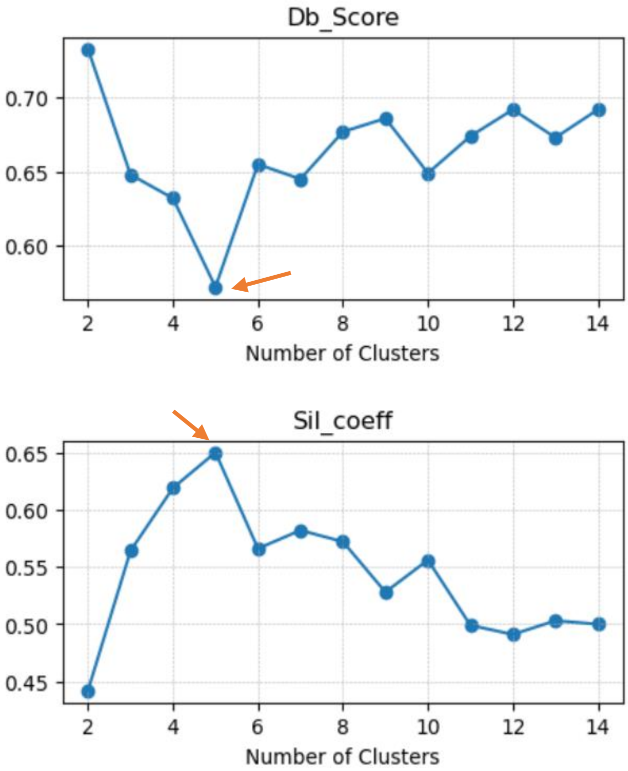
Confidence = 0.8332: This means that 83.3% of the time when weather condition is 1 (fine – no wind) with speed limit of 30 and number of vehicles is 2, accident severity is 3 (slight). In a perfect weather condition, drivers tend to drive safely and can avoid colliding with opposing traffic. A driver's visibility also improves during fine weather.

Overall, the light conditions (daytime), speed limit (30) and number of vehicles (2) involved in an accident has a huge impact (84.3%) on the likelihood of accident severity being slight compared to other item combinations in the data.

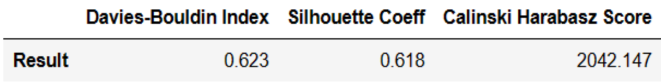
QUESTION 5



Visualizing KMeans David-Boulden index and Silhouette coefficient (Optimum cluster no. = 5):

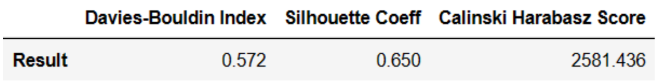


Kingston upon Hull had more accidents (569) compared to other regions in 2020. KMeans and KMedoids were implemented and KMeans performed best with lower David-Boulden index and higher silhouette. coefficient.



	Davies-Bouldin Index	Silhouette Coeff	Calinski Harabasz Score
Result	0.623	0.618	2042.147

KMeans

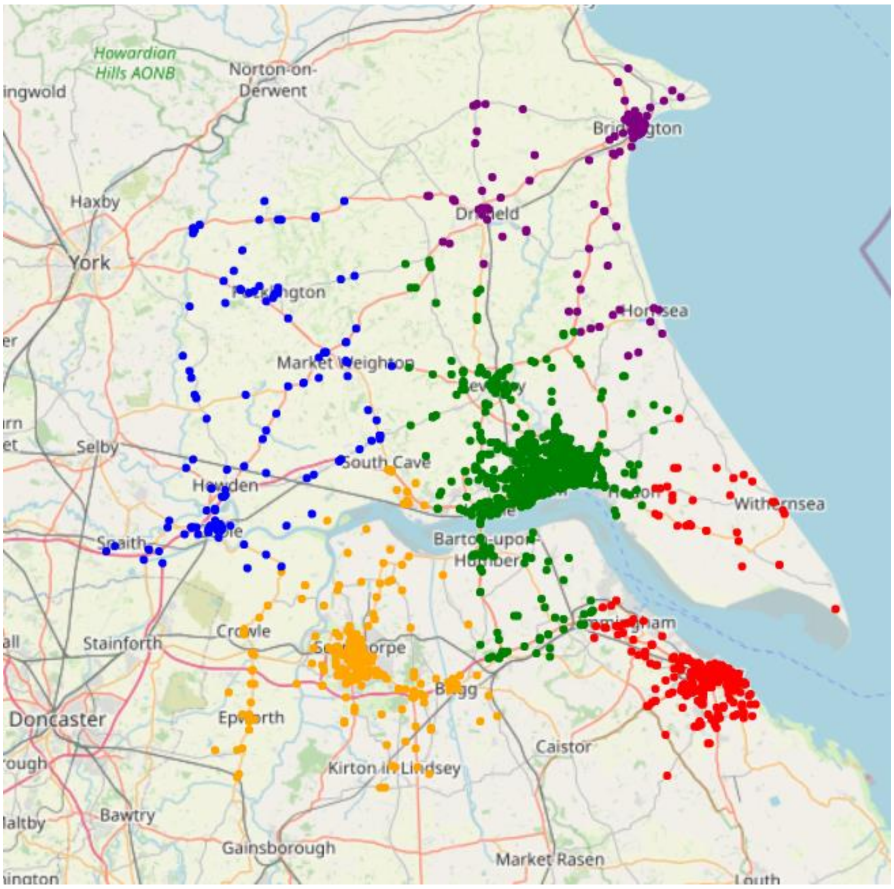


	Davies-Bouldin Index	Silhouette Coeff	Calinski Harabasz Score
Result	0.572	0.650	2581.436

KMedoid

The optimum kmeans number of clusters was determined to be 5 by obtaining the performance scores from 2 to 14 clusters.

Visualizing Cluster Result:



Cluster key:

🟩 Cluster 4

🟥 Cluster 3

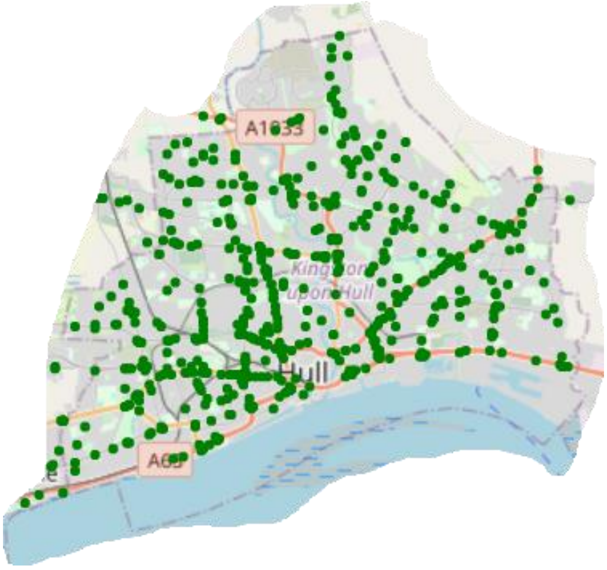
🟨 Cluster 2

🟪 Cluster 1

🟦 Cluster 0

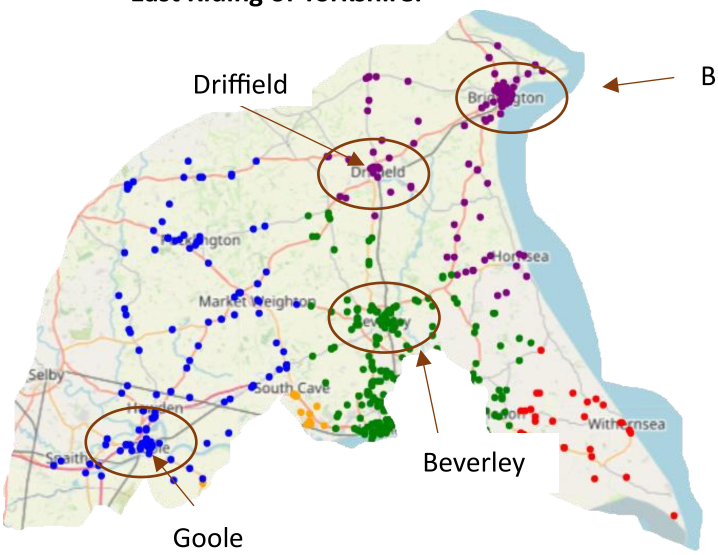
Although accidents occurred at various locations across each region, accidents in each cluster were more concentrated in certain arears (Hull, Scunthorpe, Grimsby, Goole, Beverley, and Bridlington) compared to others within same region.

Kingston upon Hull:



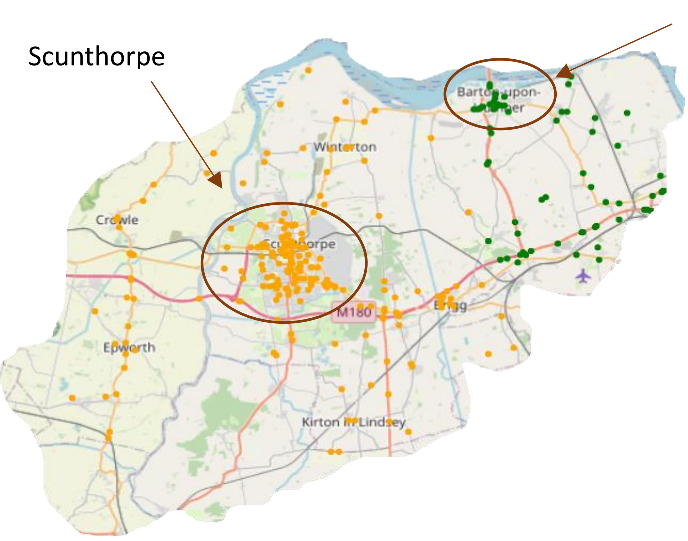
All accidents in Hull were from a single cluster (4).

East Riding of Yorkshire:



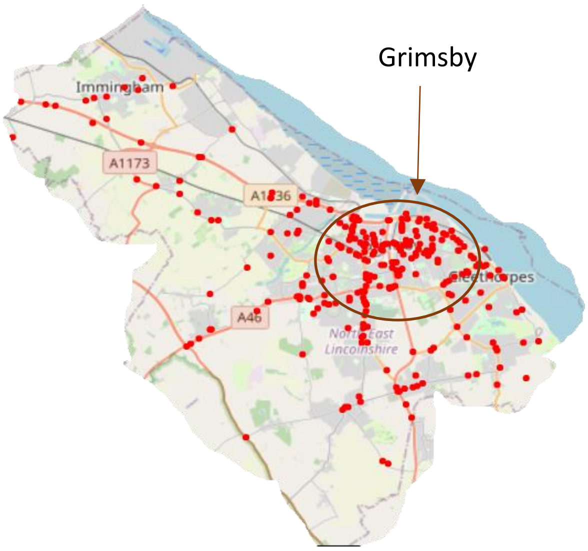
Accidents in this region were mostly concentrated in Goole, Bridlington, Beverley and Driffield.

North Lincolnshire:



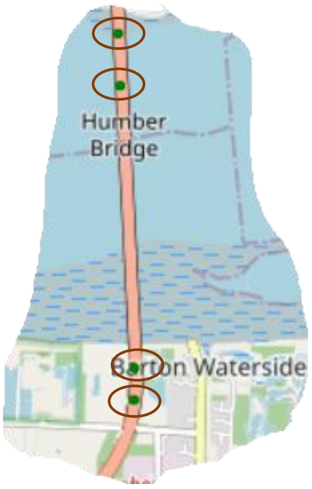
Most accidents in the region were concentrated in Scunthorpe (cluster 2) while fewer accidents were also concentrated at Barton upon Humber.

Northeast Lincolnshire



Most accidents in this region happened at Grimsby and had properties related to cluster 3.

Humber Bridge:



Four accidents, belonging to cluster 4, happened on the Humber Bridge. This is a critical place for any accident to occur, hence, it should be avoided.

QUESTION 6

Boxplot Showing Outliers in Age of Driver and Age of Vehicle

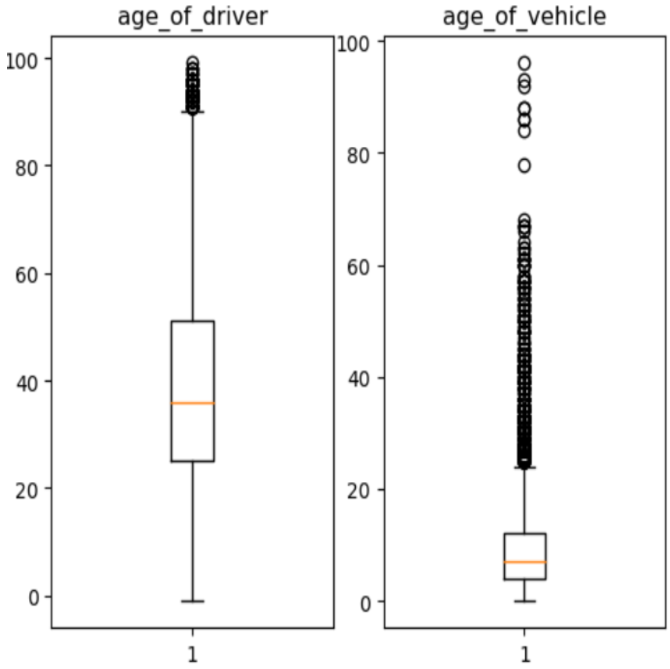
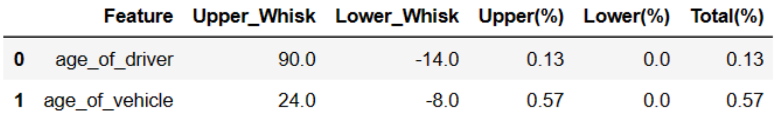


Table Showing Percent Distribution of Outlier



	Feature	Upper_Whisk	Lower_Whisk	Upper(%)	Lower(%)	Total(%)
0	age_of_driver	90.0	-14.0	0.13	0.0	0.13
1	age_of_vehicle	24.0	-8.0	0.57	0.0	0.57

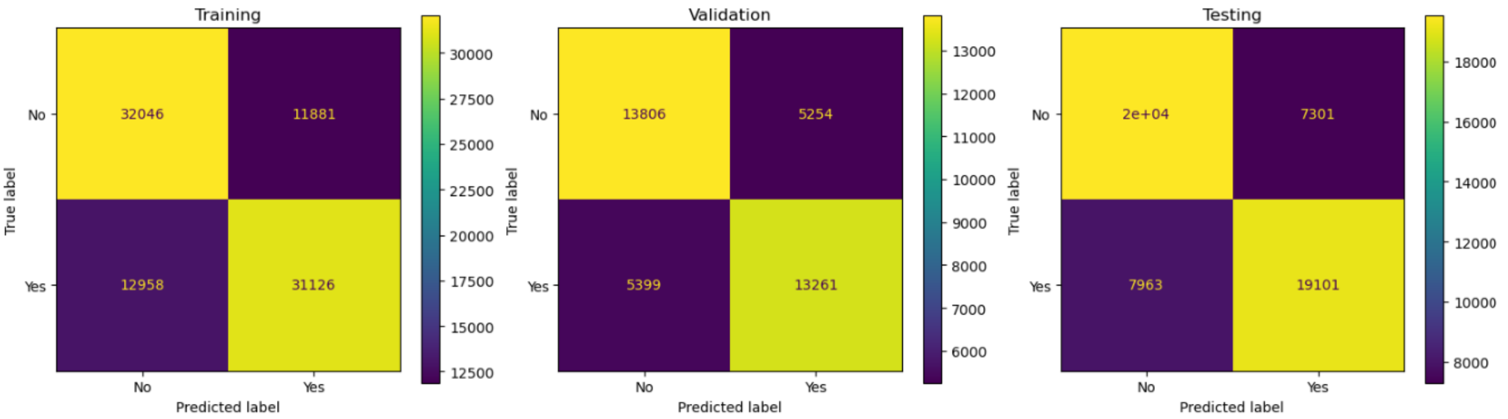
The ages of driver and vehicle beyond 90 and 25 years old were classified as outliers. However, they only represent 0.13% and 0.57% of the entire data.

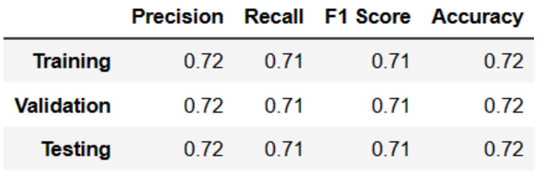
For purpose of analysis and for accurate recommendations in subsequent part of this report, these values will be kept. However, in developing accident fatality classification model, since they less than 1% of the data, a comparison of model performance with and without treating these outliers will be done. The best performing model will be adopted for deployment.

QUESTION 7

SMOTENC was used to treat imbalance because of its ability to identify categorical and continuous entries in the data (Imblearn, no date). Recursive feature elimination method was implemented to obtain top eight features for model building. Algorithms implemented were: Logistic Regression, Decision-Tree, and K-Nearest Neighbor Classifier. To optimize performance, hyperparameter tuning was done on the best performing algorithm using GridSearchCV method.

Logistic Regression:

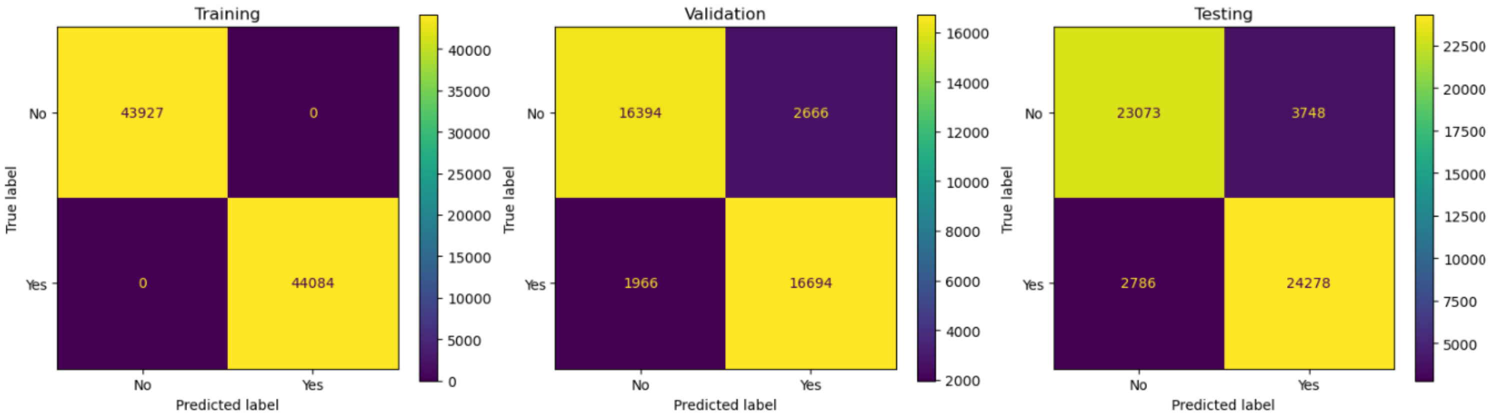


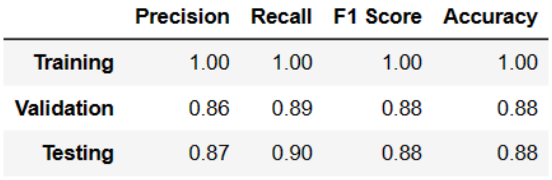


	Precision	Recall	F1 Score	Accuracy
Training	0.72	0.71	0.71	0.72
Validation	0.72	0.71	0.71	0.72
Testing	0.72	0.71	0.71	0.72

- Generalizes well with precision, recall and accuracy of 72%, 71% and 72% respectively.

Decision Tree Classifier

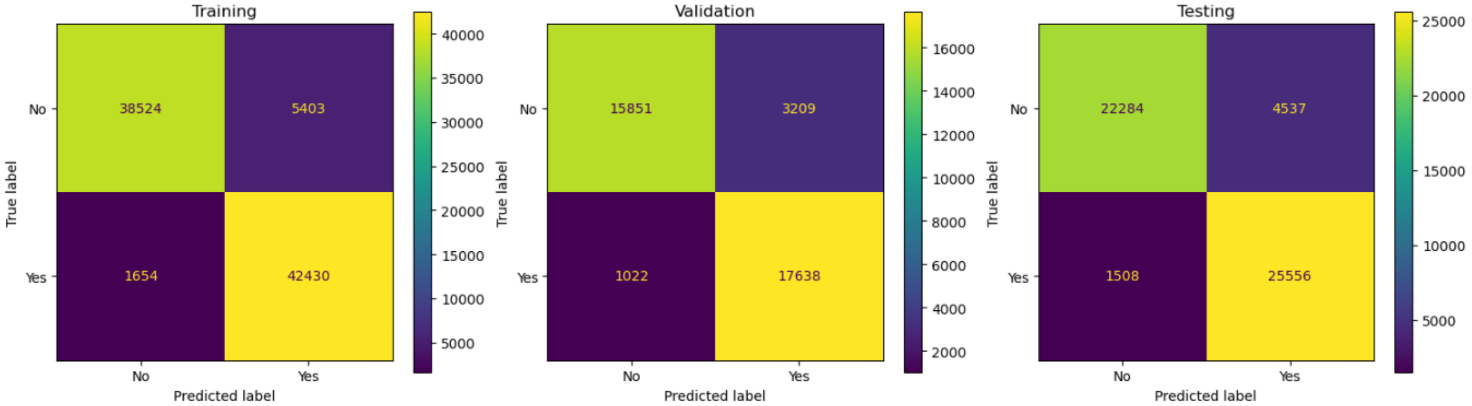


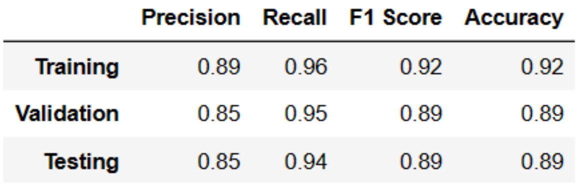


	Precision	Recall	F1 Score	Accuracy
Training	1.00	1.00	1.00	1.00
Validation	0.86	0.89	0.88	0.88
Testing	0.87	0.90	0.88	0.88

Model overfitted with 100% on training accuracy, precision and recall compared to the validation and testing.

K-Nearest Neighbors (KNN)



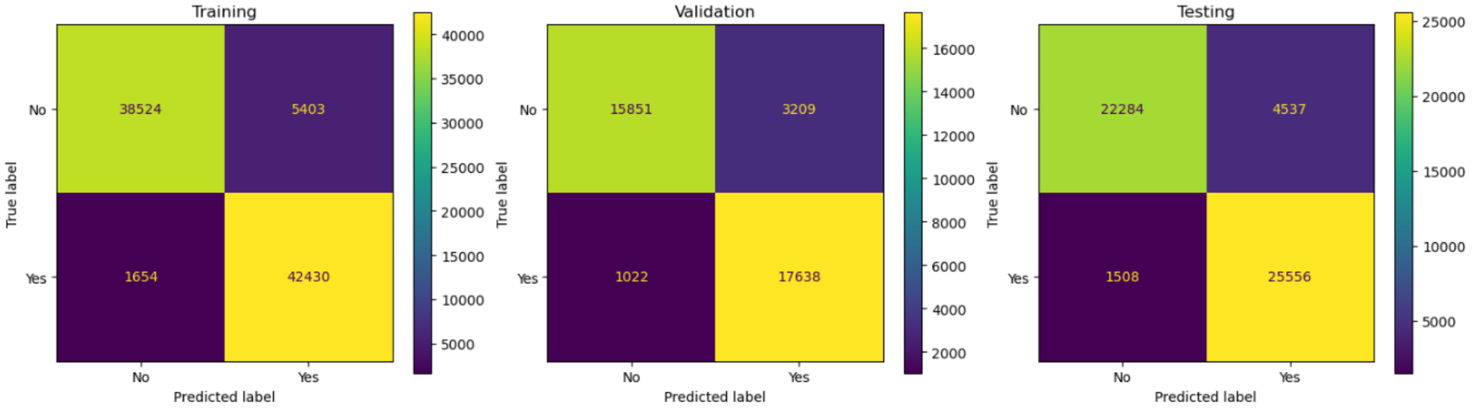


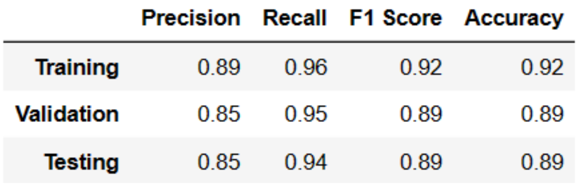
	Precision	Recall	F1 Score	Accuracy
Training	0.89	0.96	0.92	0.92
Validation	0.85	0.95	0.89	0.89
Testing	0.85	0.94	0.89	0.89

- KNN had the highest precision, recall and accuracy of 89%, 96% and 92% respectively compared to logistic regression and decision tree classifier.

KNN Hyperparameter Tuning with GridSearchCV

With recall as scoring metric, an optimal model was obtained at n_neighbor value of 5.





	Precision	Recall	F1 Score	Accuracy
Training	0.89	0.96	0.92	0.92
Validation	0.85	0.95	0.89	0.89
Testing	0.85	0.94	0.89	0.89

Model performance remained unchanged after hyperparameter tuning.

KNN model performed best with a training precision, recall and accuracy of 89%, 96% and 92%. It also generalizes well with precision, recall and accuracy of 85%, 95% and 89% on validation and 85%, 94% and 89% on test data.

A high recall value of over 94%, indicates that the model can effectively distinguish between fatal and non-fatal accidents, hence can be deployed to determine the severity of an accident.

RECOMMENDATIONS

Several age-related visual impairments could hinder a driver's ability to see clearly (aoa, no date). Hence, after license renewal, all drivers (especially above 90 years old) should be mandated to undergo periodical eye test to ascertain their visual impairment status.

Traffic on the Humber bridge should be critically monitored by road traffic and safety marshals, especially between 16:00 and 19:00 hours, as accidents occurred mainly within this period.

Road safety awareness campaign should be targeted at areas like Scunthorpe, Hull, Grimsby, and Bridlington with high accidence rate.

Vehicles in UK have an average lifespan of ten years (Garagewire, no date). Hence, vehicles above ten years old should undergo frequent road worthiness test to determine their safety status. This would prevent old cars, with poor safety conditions, from public roads.

REFERENCE

1. Timeanddate (No date) *Past Weather in Somerset, England, United Kingdom – June 2020*
Available online:
https://www.timeanddate.com/weather/@2637532/historic?month=6&year=2020
[Accessed : 30/07/2023]

2. Statista (No date). *Distribution of contributing factors leading to road accidents in Great Britain in 2020*
Available online: https://www.statista.com/statistics/323079/contributing-factors-leading-to-road-accidents-in-great-britain-uk/ [Accessed : 02/08/2023]

3. Garagewire (2023). *UK's average car age is 10 years old, new report finds.*
Available online: https://garagewire.co.uk/news/must-read/uks-average-car-age-is-10-years-old-new-report-finds/ [Accessed : 03/08/2023]

4. Bramer, M.A (2020). *Principles of data mining*. Springer, pp. 267

5. Imblearn (no date). *SMOTENC*
Available online: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTENC.html
[Accessed: 10/08/2023]

6. Gov.uk (no date). *Speed limits*
Available online: https://www.gov.uk/speed-limits [Accessed: 05/08/2023]

7. Aura Insurance (2022). *UK Speed limits: Here's what you need to know*
Available online: https://www.aurainsurance.co.uk/driving-laws/uk-speed-limits/
[Accessed: 09/08/2023]

8. Stat-19. *Accident Statistics*
Available online:
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995422/stats19.pdf [Accessed: 26/08/2023]

9. Department for Transport (2011). *Instructions for the Completion of Road Accident Reports from non-CRASH Sources.*
Available online: https://canvas.hull.ac.uk/courses/66553/files/4661442?wrap=1
[Accessed: 26/07/2023]

10. Gedlawyers (no date) *Six Car Collision Contributing Factors*
Available online: https://www.gedlawyers.com/faqs/car-collision-contributing-factors/
[Accessed: 10/08/2023]

11. Paraskevi M., Mohammed A.Q., David P., Andrew H. (2015*) Exploring the factors affecting motorway accident severity in England using the generalised ordered logistic regression model*. Journal of Safety Research, pp. 89–97

12. Gov.uk (2021) *Road Safety Data: Supporting documents*
Available online: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data [Accessed: 30/07/2023]