Contents

Introduction:

Road traffic accidents involving motorcycles are a significant concern globally that results in severe injuries and fatalities. Understanding the factors contributing to motorcycle accidents is crucial for developing effective preventive measures and improving road safety. This analysis is proposed to explore the dataset containing information on road traffic accidents involving motorcycles. The dataset includes various attributes related to the accidents such as location, date and time, road conditions, vehicle types, and casualty information. The aim is to identify trends, patterns, and factors contributing to motorcycle accidents. Insights derived from the analysis can inform road safety initiatives, infrastructure improvements, and policy interventions aimed at reducing the frequency and severity of motorcycle accidents. The findings will help to raise awareness among road users and stakeholders to promote safer practices and behaviors on the roads.

Dataset Overview:

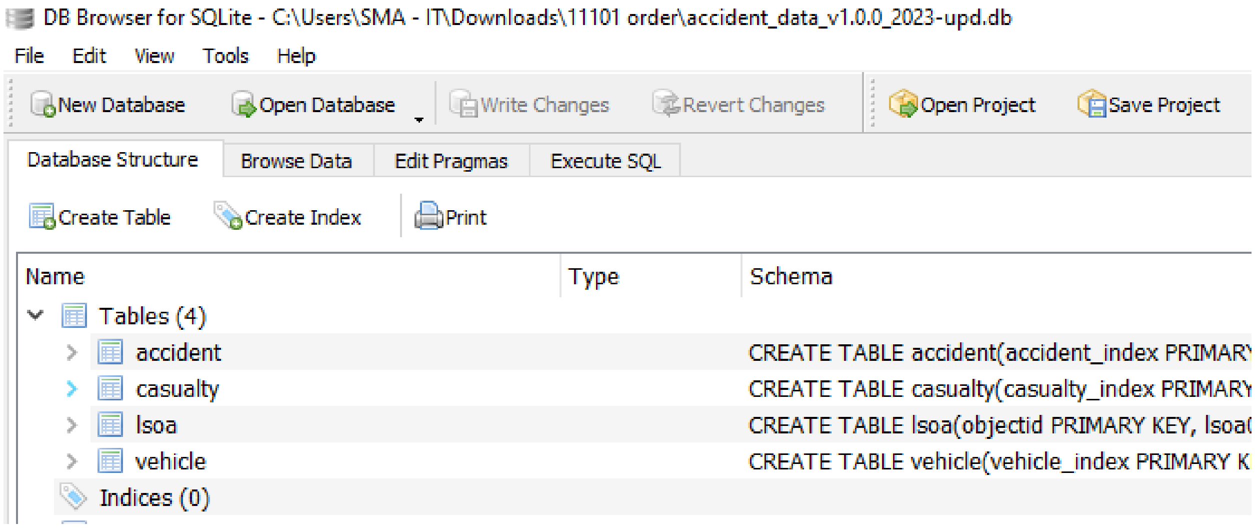
The road accident dataset provided by the Department of Transport encompasses a comprehensive array of information regarding vehicular incidents to offer valuable insights into various aspects of road safety and transportation. It comprises numerous columns such as each providing distinct details pertinent to understanding the circumstances and consequences of accidents. Key columns in the dataset include identifiers such as "vehicle_index" and "accident_index" which uniquely identify vehicles and accidents to facilitate efficient data organization and analysis. Information about the accident occurrence year and reference as well as the date and time of the incident that enables temporal analysis. The day of the week offers insights into potential patterns or trends in accident frequency. Detailed attributes concerning the vehicles involved in accidents are provided which include their types, manoeuvres preceding the incident, directions of travel, and locations relative to the road infrastructure such as junction lanes. The data factors contributing to accidents such as skidding, collisions with objects, off the carriageway, and vehicle departures from the roadway offer valuable context for understanding accident dynamics.

The dataset also includes driver-related information such as the driver sex, age, and age band, which can aid in analyzing the demographic aspects of accidents and identifying potential risk factors associated with certain age group genders. Details about the vehicles such as engine capacity, age, and model, provide insights into the characteristics of vehicles involved in accidents and their potential impact on accident severity. The dataset encompasses information about casualties that include the number and severity of injuries sustained as well as the location of the accident within the Lower Layer Super Output Area (LSOA). This geographic information allows for spatial analysis and identification of areas with higher accident frequencies and greater risk factors to aid the targeted interventions and policy formulation to enhance road safety. The development of model require thorough cleaning and preprocessing to address missing values, inconsistencies, and errors, ensuring the reliability and validity of subsequent analyses and insights derived from the data. When it is cleaned and analyzed then the model will hold the significant potential for informing evidence-based interventions, policies, and strategies aimed at reducing the incidence and severity of road accidents and promoting safer transportation systems.

Data Extraction:

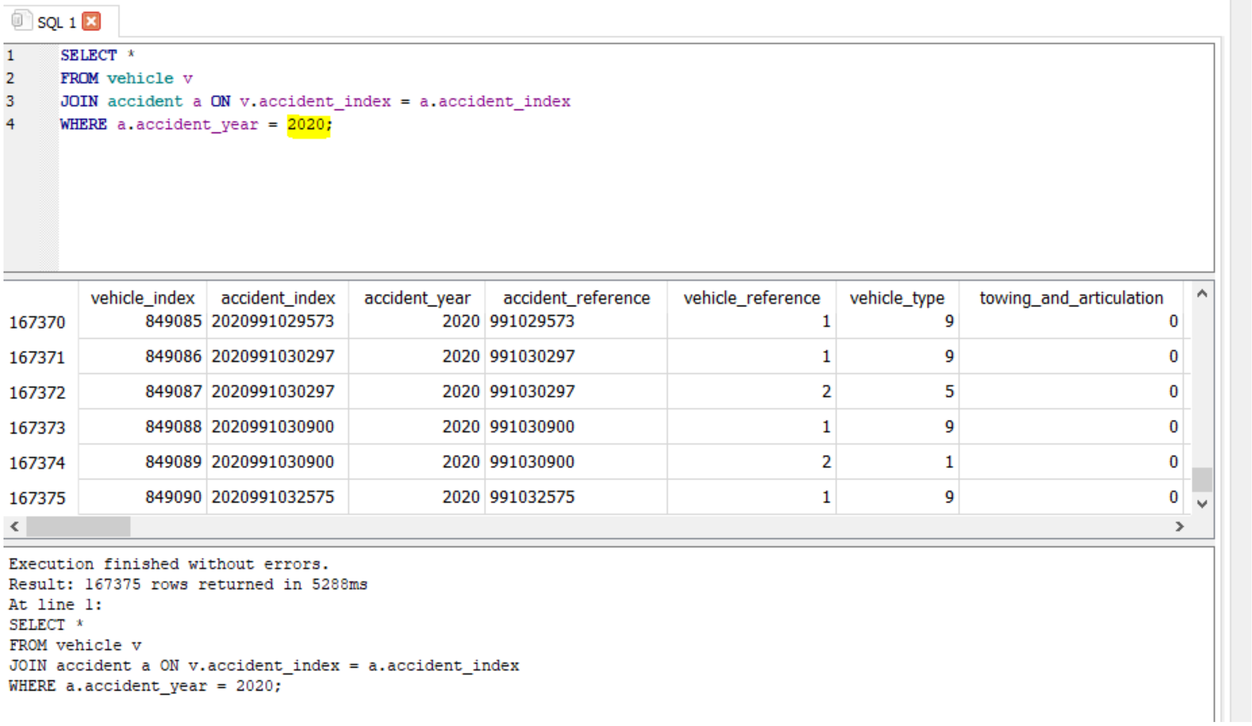
The database is loaded to extract the required data and SQL quires are performed to fetch the data from the database. The next step involves filtering the data based on specified criteria that is vehicle types provided within a certain range as per instructions of transport department (3 to 5 representing motorcycle 50cc to 500cc and over). It will require additional filtering in Python code to include only the desired vehicle types based on the instructions. It depends on the apriori algorithm and model requirements for further processing of the extracted data which involve tasks such as data cleaning, transformation, and aggregation. The model will be developed in on the transformed data that will provide facts and analysis.

Database:



Name	Type	Schema
Tables (4)		
accident		CREATE TABLE accident(accident_index PRIMARY
casualty		CREATE TABLE casualty(casualty_index PRIMARY
lsoa		CREATE TABLE lsoa(objectid PRIMARY KEY, lsoa(
vehicle		CREATE TABLE vehicle(vehicle_index PRIMARY K
Indices (0)		

Selection of data from year 2020

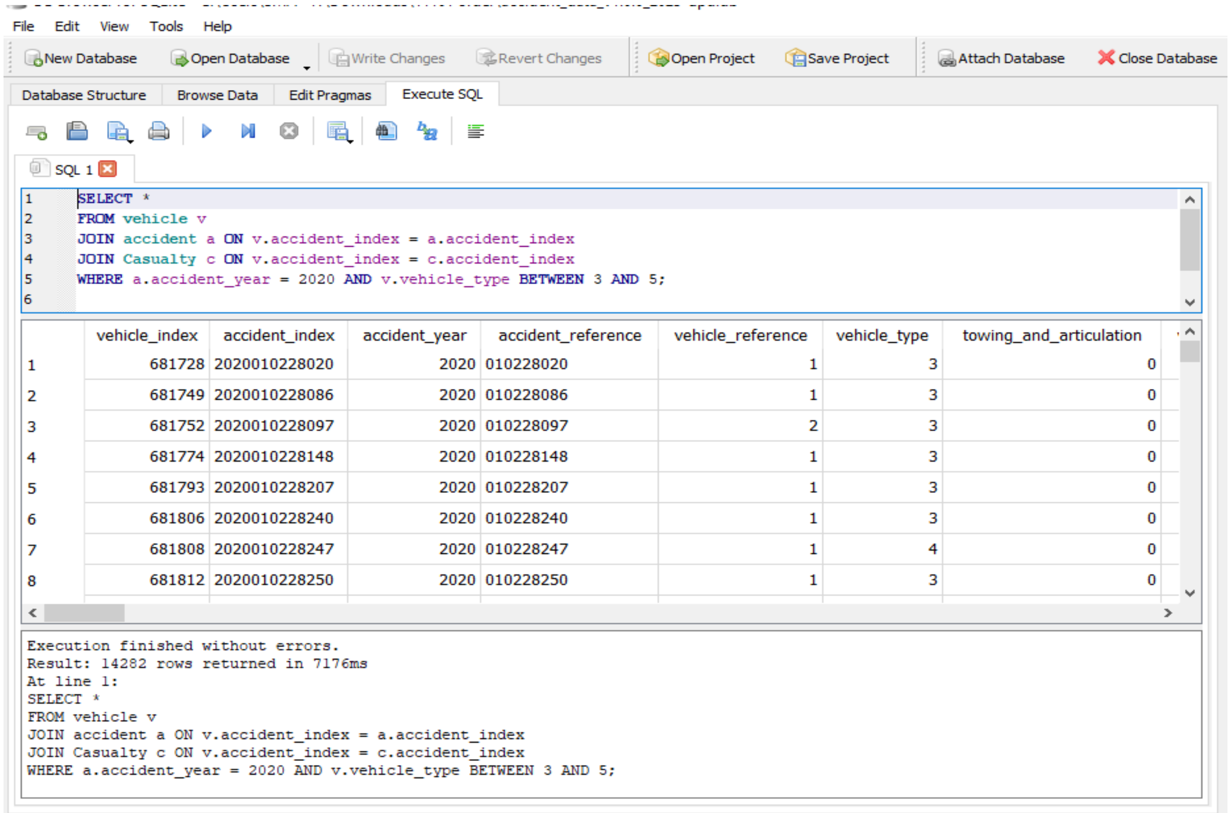


```
SELECT *
FROM vehicle v
JOIN accident a ON v.accident_index = a.accident_index
WHERE a.accident_year = 2020;
```

	vehicle_index	accident_index	accident_year	accident_reference	vehicle_reference	vehicle_type	towing_and_articulation
167370	849085	2020991029573	2020	991029573	1	9	0
167371	849086	2020991030297	2020	991030297	1	9	0
167372	849087	2020991030297	2020	991030297	2	5	0
167373	849088	2020991030900	2020	991030900	1	9	0
167374	849089	2020991030900	2020	991030900	2	1	0
167375	849090	2020991032575	2020	991032575	1	9	0

```
Execution finished without errors.
Result: 167375 rows returned in 5288ms
At line 1:
SELECT *
FROM vehicle v
JOIN accident a ON v.accident_index = a.accident_index
WHERE a.accident_year = 2020;
```

Selection of values from casualty along with vehicle-type codes.



```
SELECT *
FROM vehicle v
JOIN accident a ON v.accident_index = a.accident_index
JOIN Casualty c ON v.accident_index = c.accident_index
WHERE a.accident_year = 2020 AND v.vehicle_type BETWEEN 3 AND 5;
```

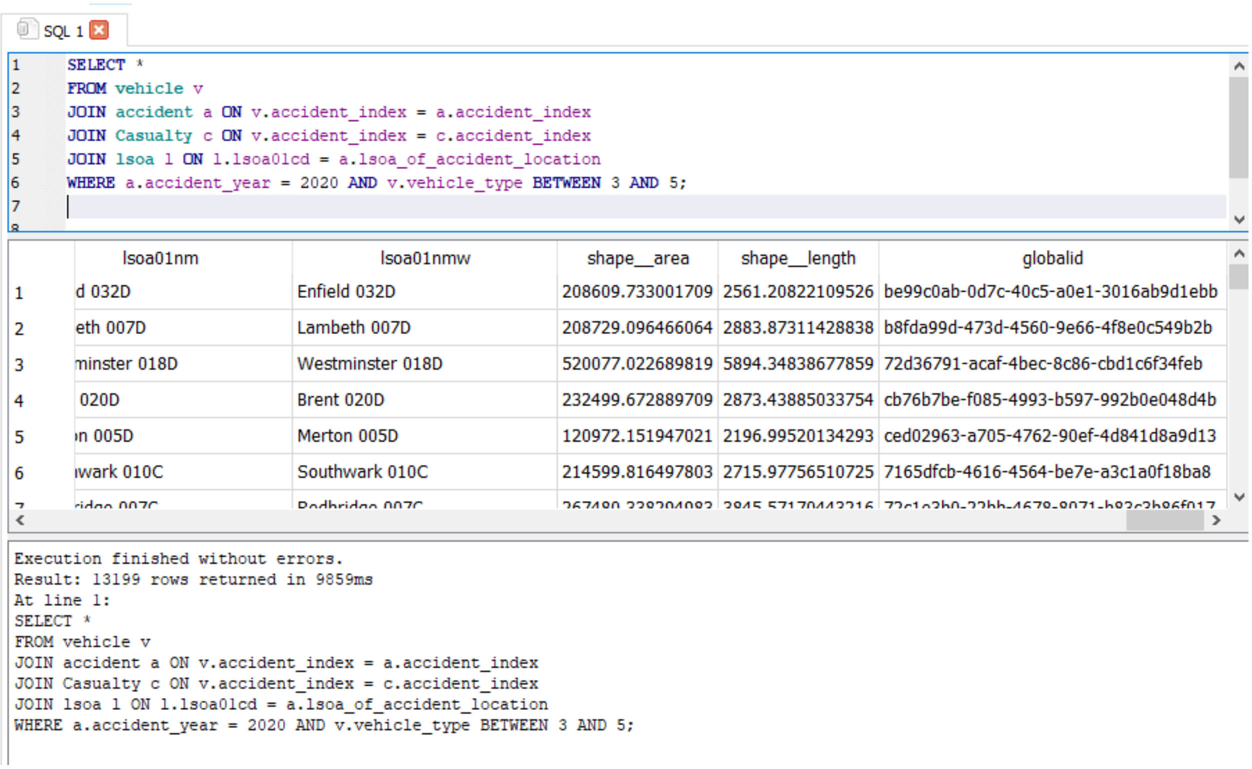
	vehicle_index	accident_index	accident_year	accident_reference	vehicle_reference	vehicle_type	towing_and_articulation
1	681728	2020010228020	2020	010228020	1	3	0
2	681749	2020010228086	2020	010228086	1	3	0
3	681752	2020010228097	2020	010228097	2	3	0
4	681774	2020010228148	2020	010228148	1	3	0
5	681793	2020010228207	2020	010228207	1	3	0
6	681806	2020010228240	2020	010228240	1	3	0
7	681808	2020010228247	2020	010228247	1	4	0
8	681812	2020010228250	2020	010228250	1	3	0

```
Execution finished without errors.
Result: 14282 rows returned in 7176ms
At line 1:
SELECT *
FROM vehicle v
JOIN accident a ON v.accident_index = a.accident_index
JOIN Casualty c ON v.accident_index = c.accident_index
WHERE a.accident_year = 2020 AND v.vehicle_type BETWEEN 3 AND 5;
```

Executed Query:

```
SELECT *

FROM vehicle v

JOIN accident a ON v.accident_index = a.accident_index

JOIN Casualty c ON v.accident_index = c.accident_index

JOIN lsoa l ON l.lsoa01cd = a.lsoa_of_accident_location

WHERE a.accident_year = 2020 AND v.vehicle_type BETWEEN 3 AND 5;
```

Following data is the actual database that fulfill the required dataset to perform analysis.



```sql
SELECT *
FROM vehicle v
JOIN accident a ON v.accident_index = a.accident_index
JOIN Casualty c ON v.accident_index = c.accident_index
JOIN lsoa l ON l.lsoa01cd = a.lsoa_of_accident_location
WHERE a.accident_year = 2020 AND v.vehicle_type BETWEEN 3 AND 5;
```

	lsoa01nm	lsoa01nmw	shape__area	shape__length	globalid
1	d 032D	Enfield 032D	208609.733001709	2561.20822109526	be99c0ab-0d7c-40c5-a0e1-3016ab9d1ebb
2	eth 007D	Lambeth 007D	208729.096466064	2883.87311428838	b8fda99d-473d-4560-9e66-4f8e0c549b2b
3	minster 018D	Westminster 018D	520077.022689819	5894.34838677859	72d36791-acaf-4bec-8c86-cbd1c6f34feb
4	020D	Brent 020D	232499.672889709	2873.43885033754	cb76b7be-f085-4993-b597-992b0e048d4b
5	ın 005D	Merton 005D	120972.151947021	2196.99520134293	ced02963-a705-4762-90ef-4d841d8a9d13
6	ıwark 010C	Southwark 010C	214599.816497803	2715.97756510725	7165dfcb-4616-4564-be7e-a3c1a0f18ba8
7	ridge 007C	Redbridge 007C	267480.338204083	3845.57170443216	72c1e3b0-22bb-4678-8071-b83c3b86f017

```
Execution finished without errors.
Result: 13199 rows returned in 9859ms
At line 1:
SELECT *
FROM vehicle v
JOIN accident a ON v.accident_index = a.accident_index
JOIN Casualty c ON v.accident_index = c.accident_index
JOIN lsoa l ON l.lsoa01cd = a.lsoa_of_accident_location
WHERE a.accident_year = 2020 AND v.vehicle_type BETWEEN 3 AND 5;
```

Development Methodology:

The analysis will be performed by using Python to explore and analyze the dataset. Data preprocessing will be done by using various libraries such as Pandas for data manipulation, Matplotlib and Seaborn for data visualization. The goal is to gain a deeper understanding of motorcycle accidents to identify patterns and trends that can inform road safety initiatives The model will be developed using the Apriori algorithm for analyzing traffic road accidents need to understand the dataset structure and how the algorithm can be applied to extract meaningful insights. The Apriori algorithm is commonly used for association rule mining, which can help us identify patterns and relationships between different attributes in the dataset.

Following steps will be taken:

- Data Pre-Processing:

Data pre-processing will involve cleaning the data, handling missing values, and encoding categorical variables. Once the data is ready then the Apriori algorithm will be applied to find frequent itemsets which represent combinations of attributes that occur together frequently in the dataset.

- Transaction Generation:

  The Apriori algorithm works with transaction data where each transaction represents a set of items. In the context of traffic road accidents will define each accident record as a transaction with the items representing different attributes such as road surface conditions, lighting conditions, and weather conditions.

- Itemset Generation:

  Once the transactions are generated then the Apriori algorithm to identify frequent itemsets. These are combinations of attributes that frequently co-occur in the dataset such as finding that accidents are more likely to occur on wet roads at night with poor lighting.

- Association Rule Mining:

  Association rules will be generaded from the frequent itemsets. These rules describe relationships between different attributes and can help us understand the factors that contribute to motorcycle accidents. For example, we may discover rules such as "If road surface is wet and lighting conditions are poor then the likelihood of an accident is higher."

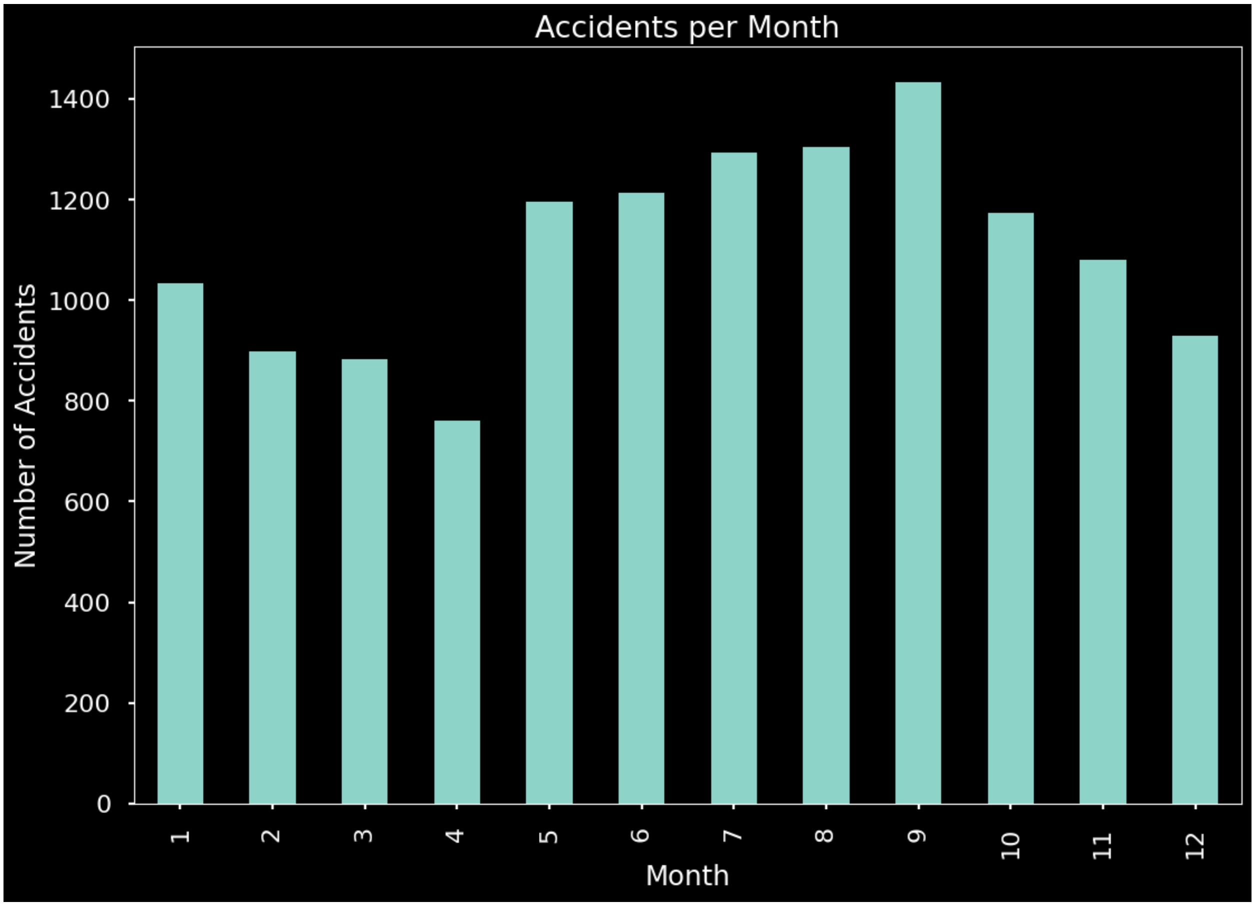
- Evaluation and Interpretation:

  The evaluation of the generated association rules will be based on metrics such as support, confidence, and lift. These metrics help to assess the strength and significance of the discovered patterns. After that, the analysis will interpret the results to gain insights into the underlying factors contributing to motorcycle accidents.

Data Analysis:

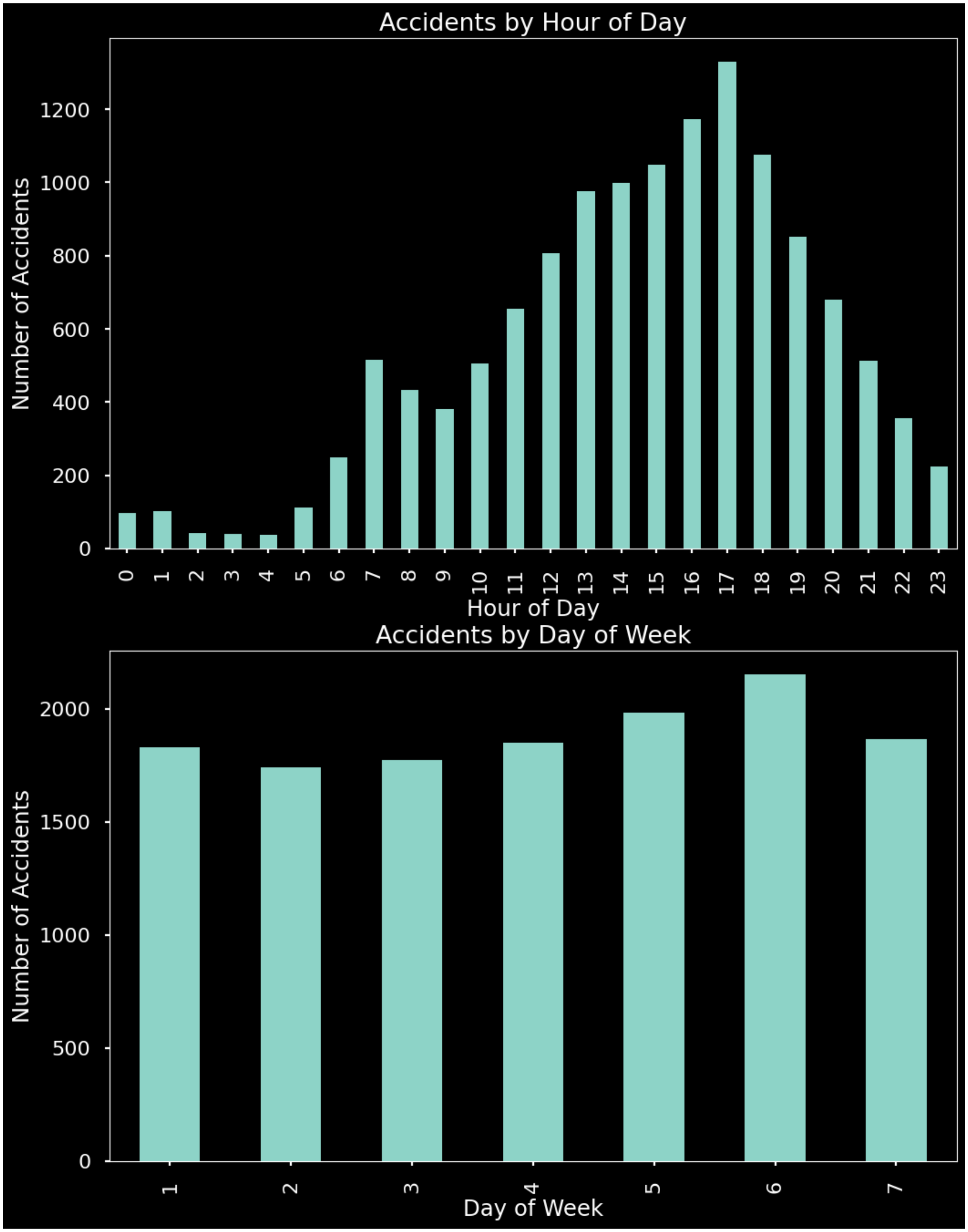
Data analysis in Python often involves utilizing popular plotting libraries like Matplotlib, Seaborn, and Plotly. These libraries offer robust functionalities for visualizing data across different formats to empower users to extract insights from the datasets. Matplotlib is widely-used tool for generating a diverse array of plots to spanning from line plots and scatter plots to bar plots and histograms. Seaborn complements Matplotlib by extending its capabilities and introducing specialized features for statistical visualization such as violin plots, box plots, and pair plots which prove invaluable for unraveling intricate relationships within complex datasets.These plotting libraries furnish data analysis with a comprehensive toolkit to proficiently explore, analyze, and convey insights gleaned from the data.

Following chart shows the total number of accident in month. It show that there are average 800 accidents occurred in each month in the year 2020. However, the minimum accidents are in the month 5 that is May, 2020 and maximum occurred in September, 2020 that are approximately 1500 accidents.



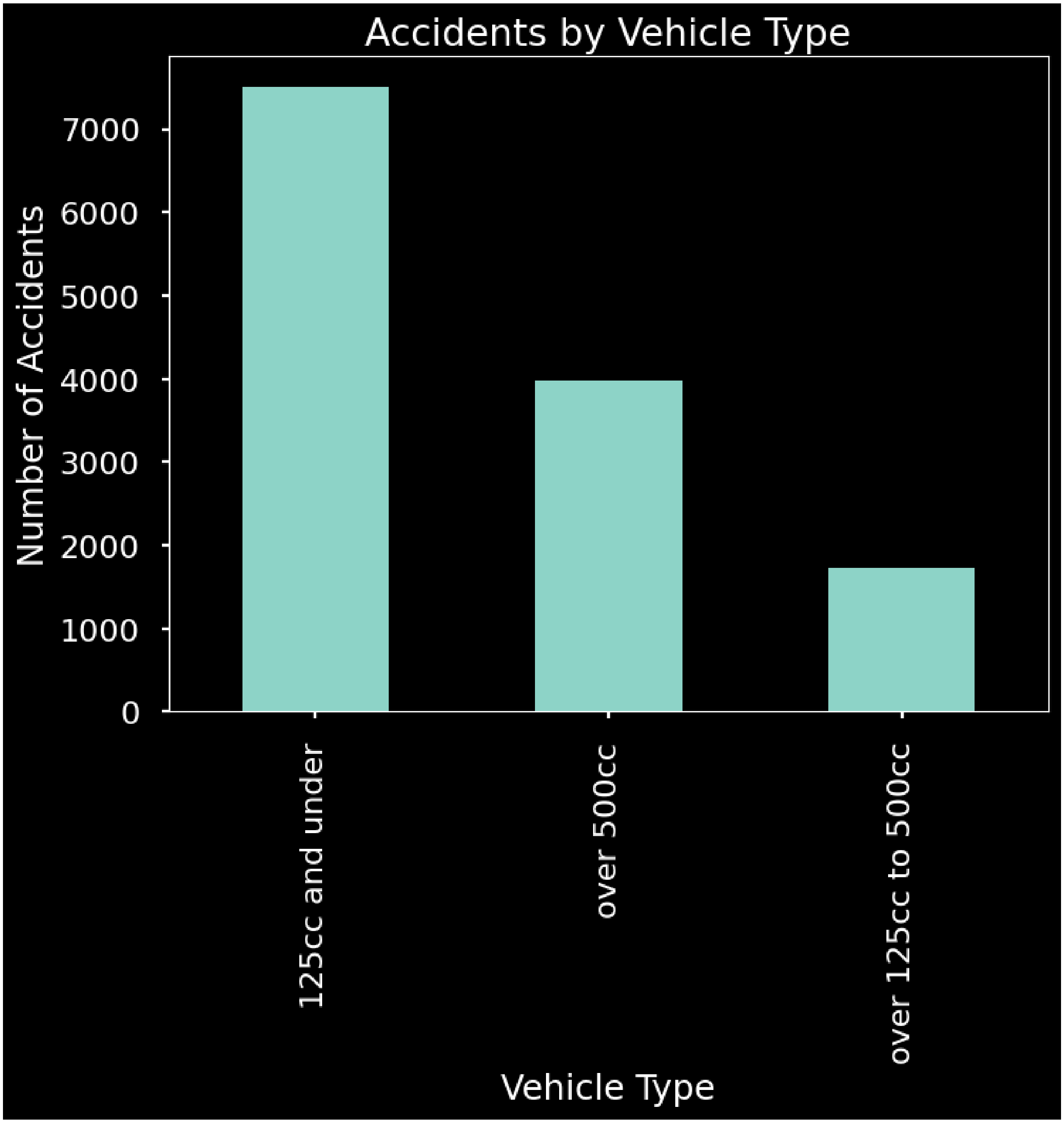
Task 1:

Following data shows the accidents occurred in each hour and the days of week.There are maximum numbers of accidents in the from 3pm till 6pm evening time. There time slots show the maximum amount ot traffic in the time that causes the accidents.



Task 2:

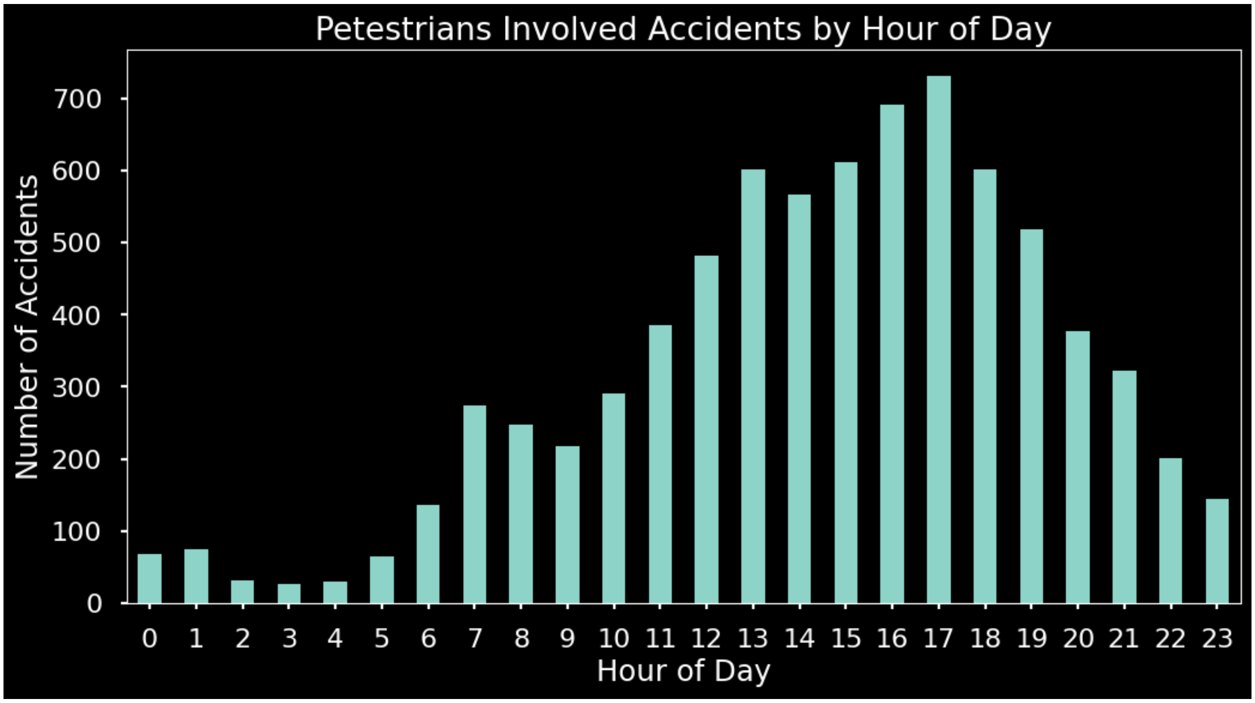
The data of different motorcycle that are 125 cc and under, over 125 cc up to 500 cc, and over 500 cc. It show that the maximum accidents occurred in the motorbikes of 125cc and under with more that 7000 accidents. While, the over 500 cc bikes are having 4000 accidents. The least number of accidents are occurred in the bikes between 125 to 500 cc that are approximately 2000 accidents.

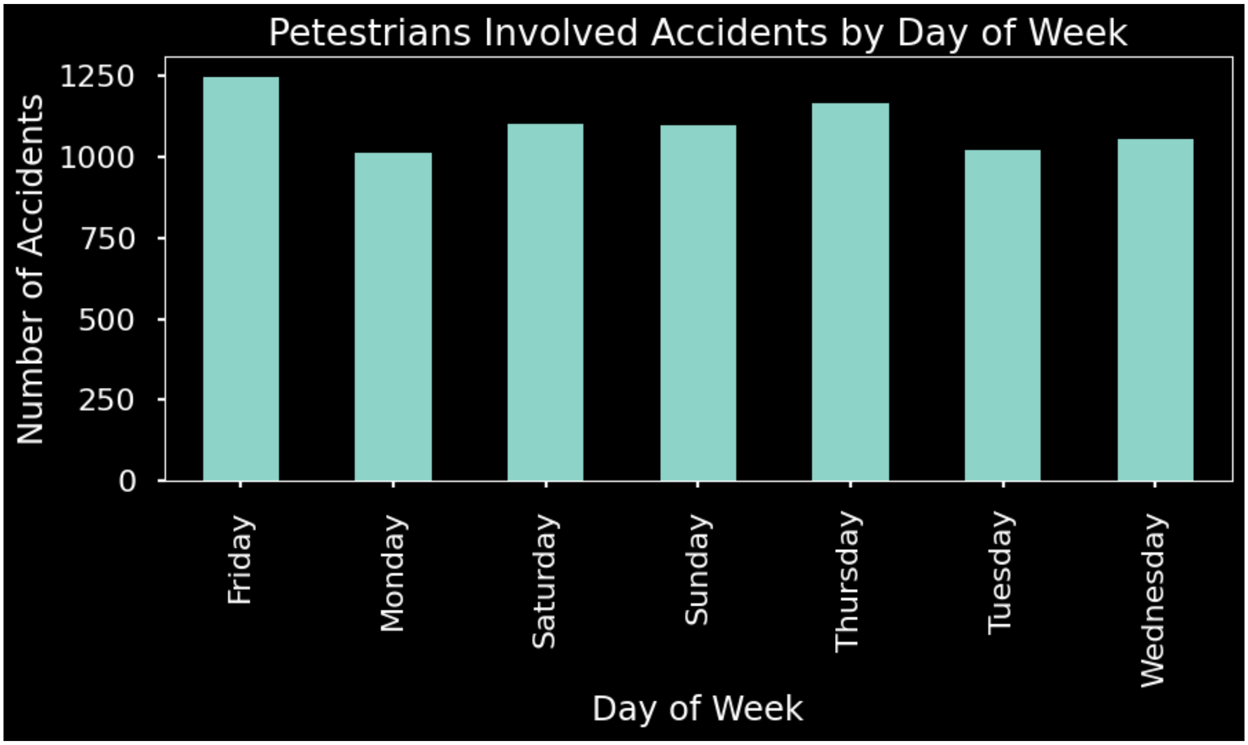


Task 3:

Following visualization shows the involvement of pedestrians in the accidents. The day of week on which accidents occurred is the Friday that got maximum accidents and other days have average of 1000 accidents.

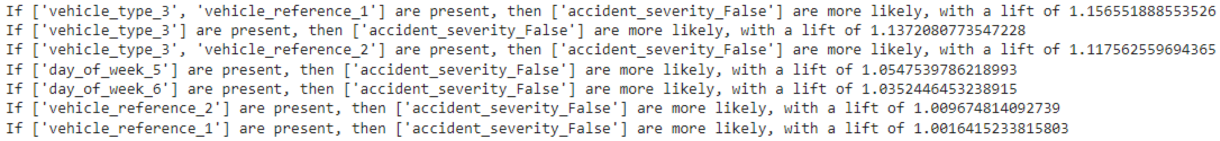
Similarly, the maximum numbers of accidents in the from 3pm till 6pm evening time. There time slots show the maximum amount ot traffic in the time that causes the accidents.





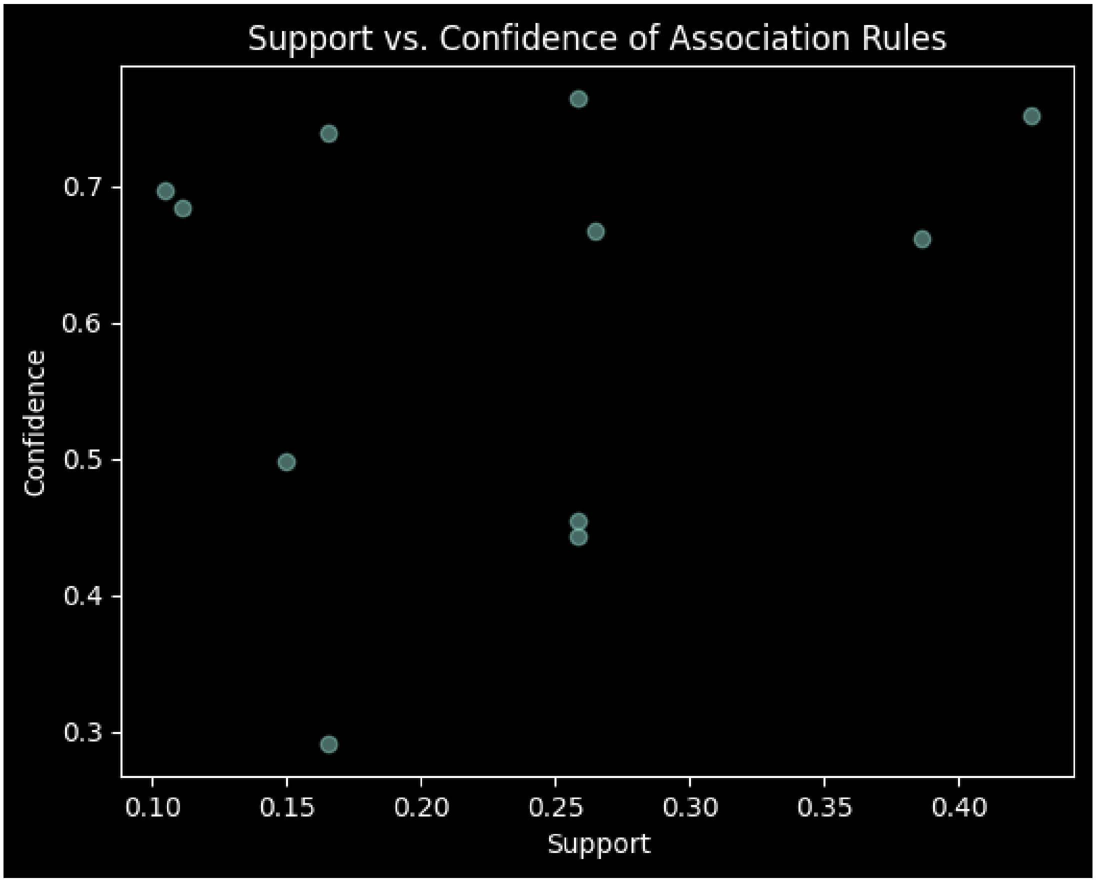
Task 4:

The following result show the association rules and item sets which are based upon the selected veriables.



```
If ['vehicle_type_3', 'vehicle_reference_1'] are present, then ['accident_severity_False'] are more likely, with a lift of 1.156551888553526
If ['vehicle_type_3'] are present, then ['accident_severity_False'] are more likely, with a lift of 1.1372080773547228
If ['vehicle_type_3', 'vehicle_reference_2'] are present, then ['accident_severity_False'] are more likely, with a lift of 1.117562559694365
If ['day_of_week_5'] are present, then ['accident_severity_False'] are more likely, with a lift of 1.0547539786218993
If ['day_of_week_6'] are present, then ['accident_severity_False'] are more likely, with a lift of 1.0352446453238915
If ['vehicle_reference_2'] are present, then ['accident_severity_False'] are more likely, with a lift of 1.009674814092739
If ['vehicle_reference_1'] are present, then ['accident_severity_False'] are more likely, with a lift of 1.0016415233815803
```

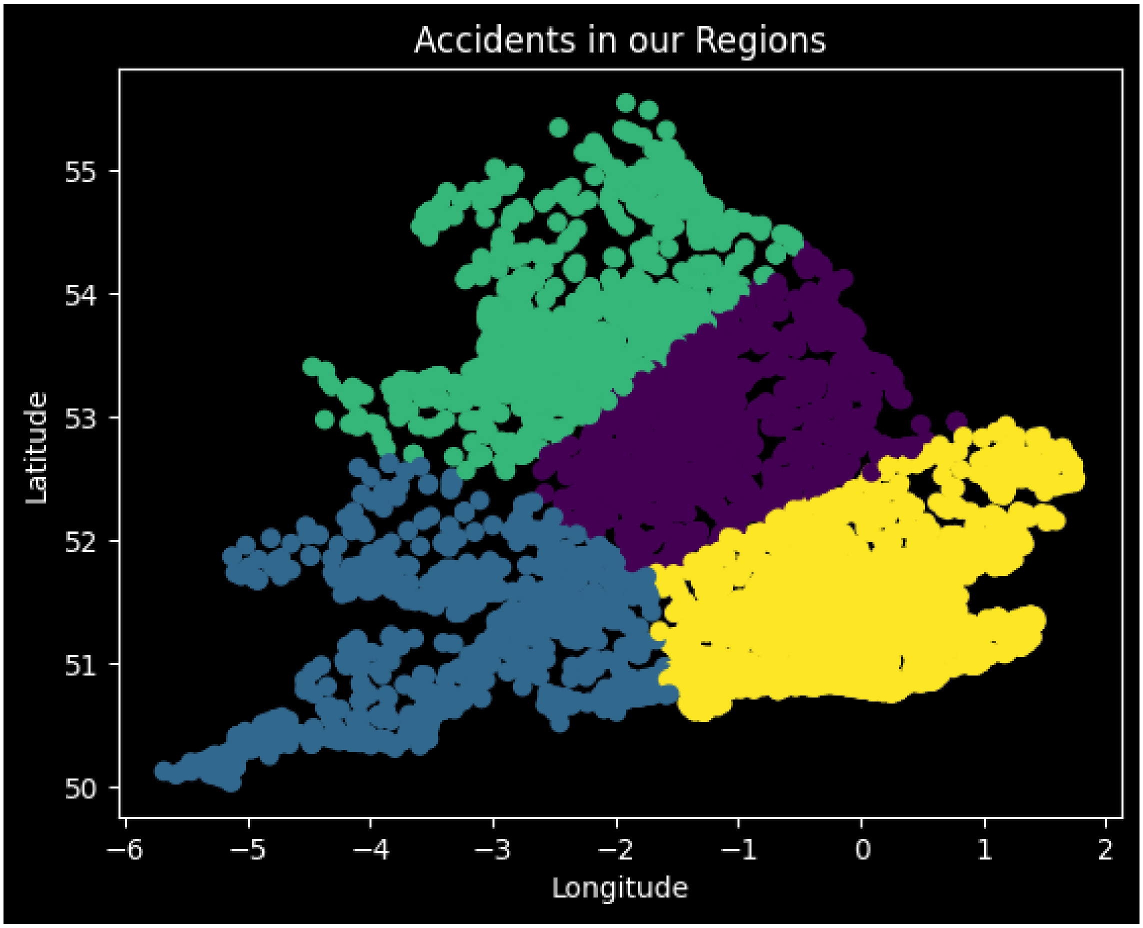
The plot show multiple values regarding confidence and support with aproiri algorithm.



Task 5:

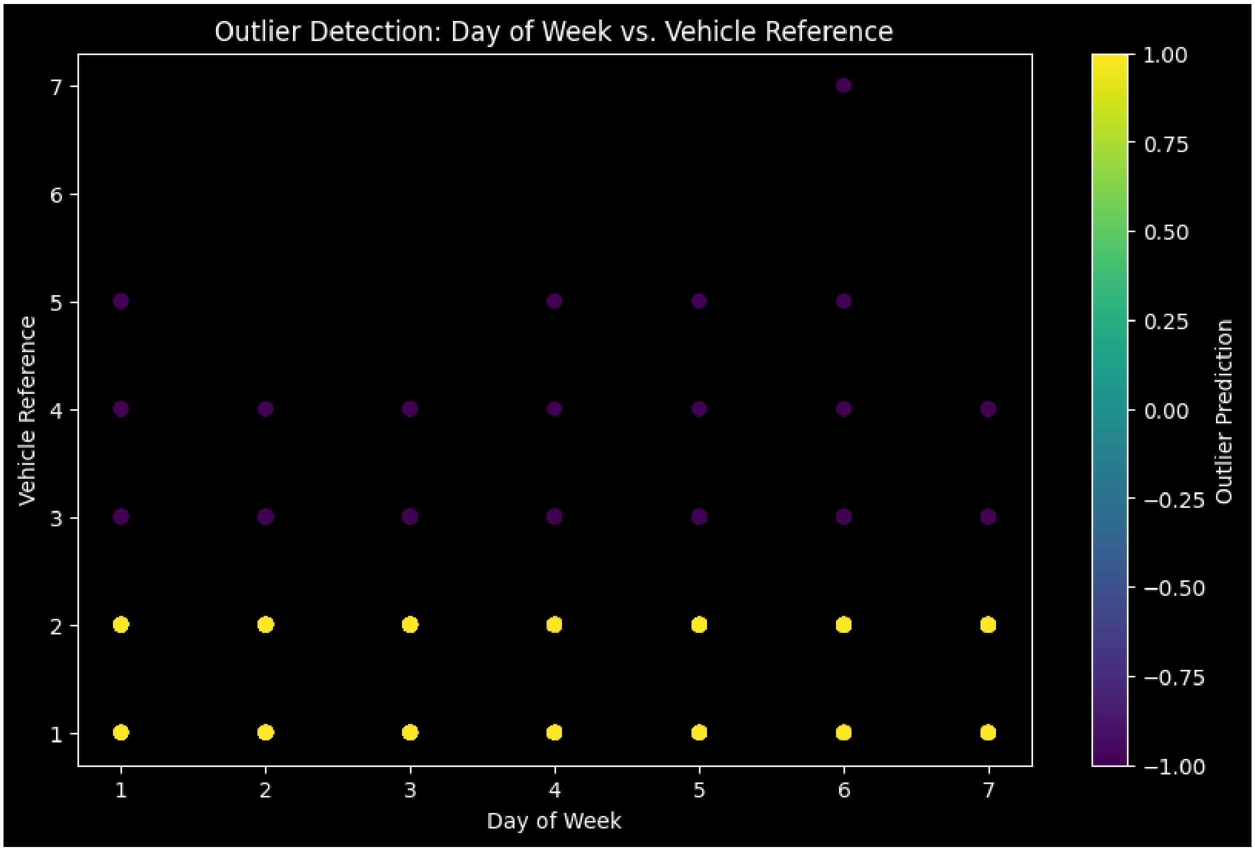
These are the accidents occurred in the regions that are mapped from the provided stat form: These code such as ['E06000010', 'E06000011', 'E06000012', 'E06000013'] represents the areas.

The names are Kingston upon Hull, Humberside, East Riding of york shire


Accidents in our Regions
Latitude
55
54
53
52
51
50
−6
−5
−4
−3
−2
−1
0
1
2
Longitude

Task 6:

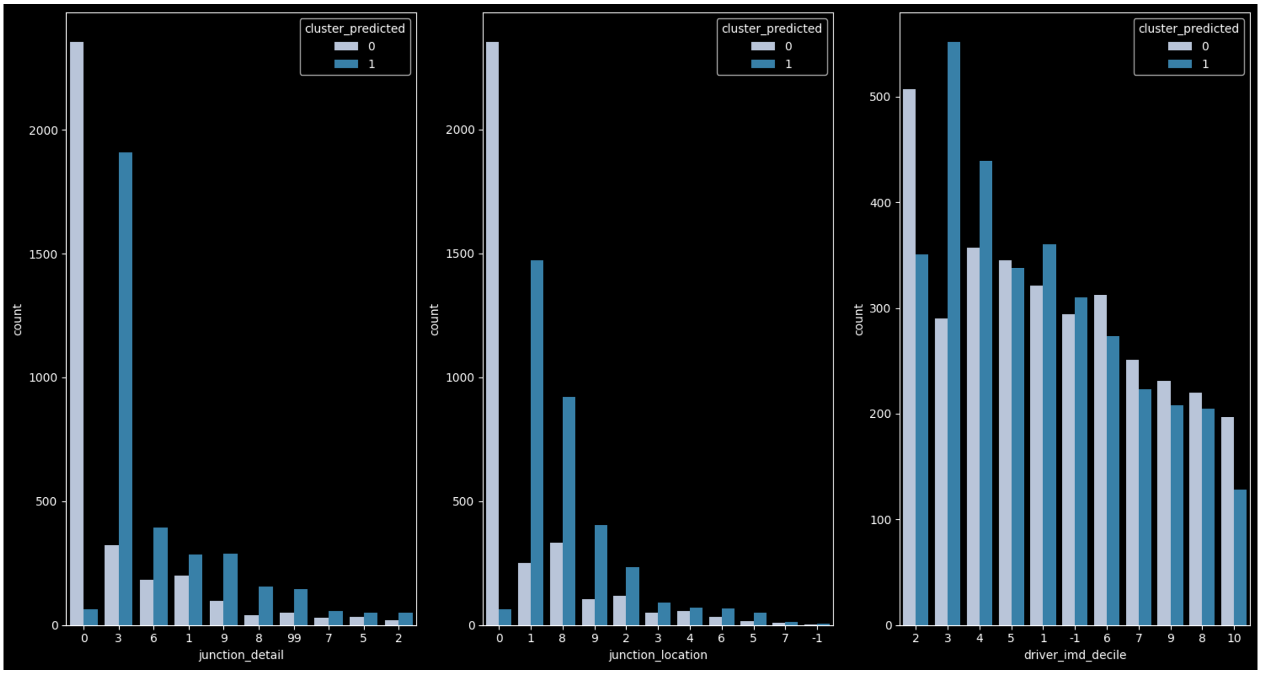
The following data shows the points as outliers that can be removed from provided dataset. To perform this analysis, the outlier are not removed due the classification model development. The model will be tested and trained upon available data then the outcome will be predicted based on outliers.

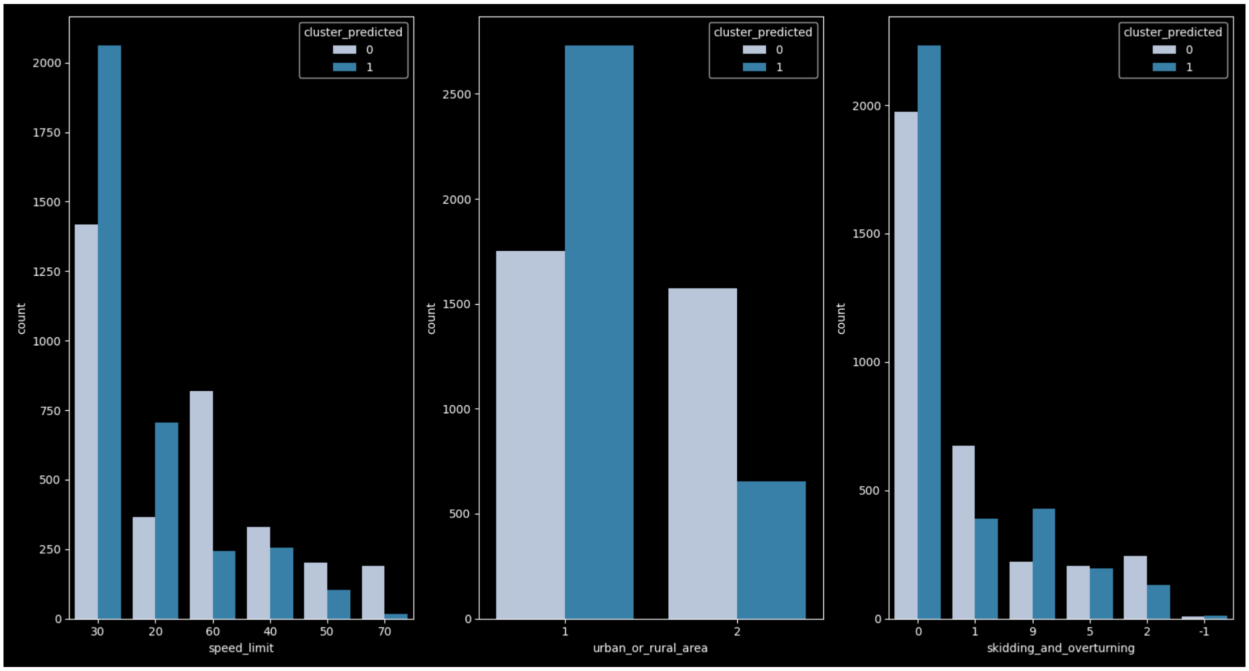


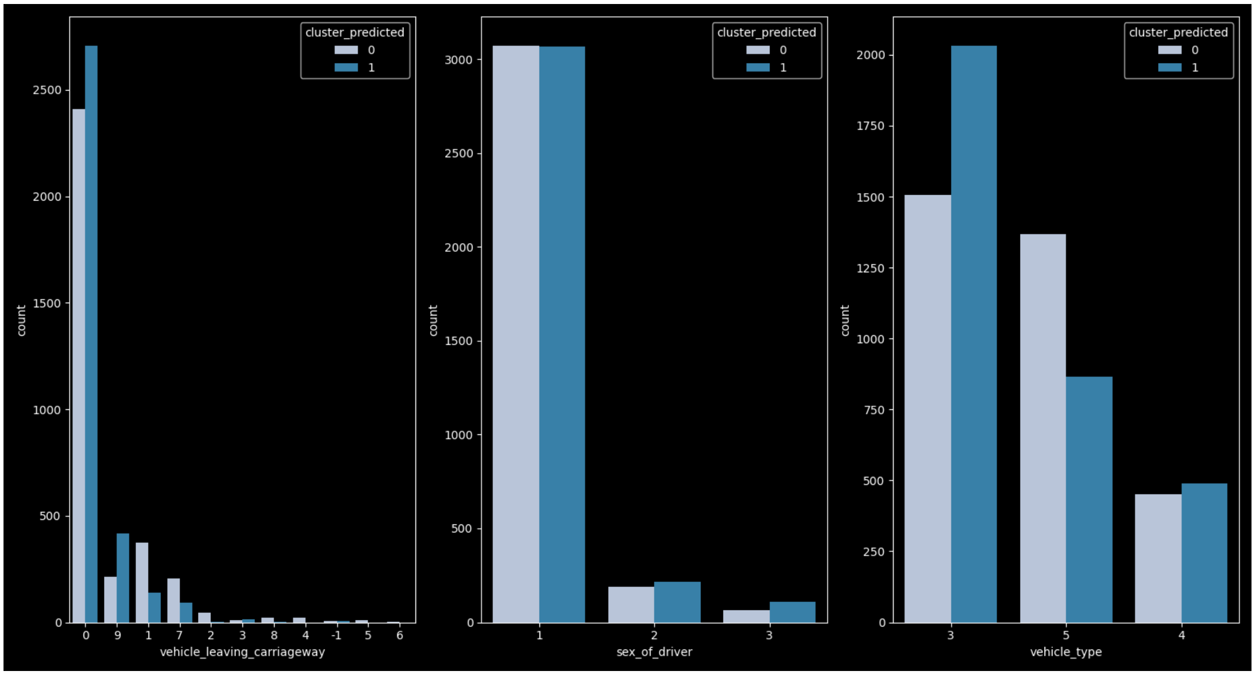
Task 7:

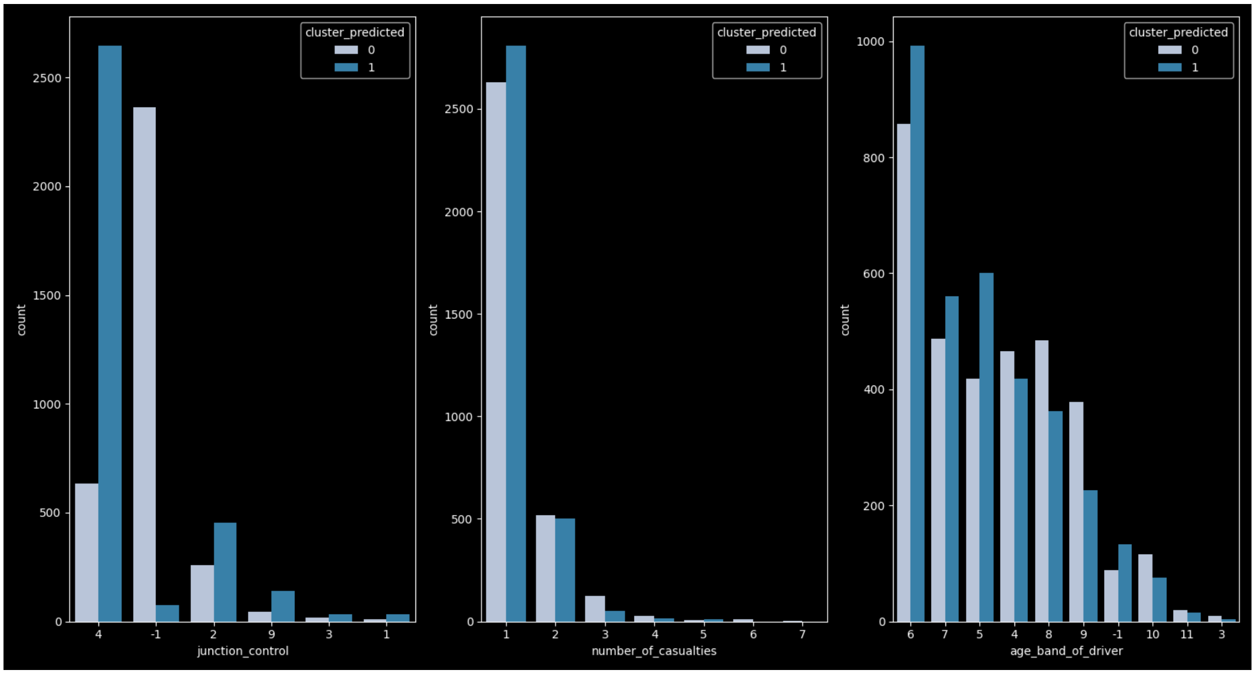
Prediction based on classification model.

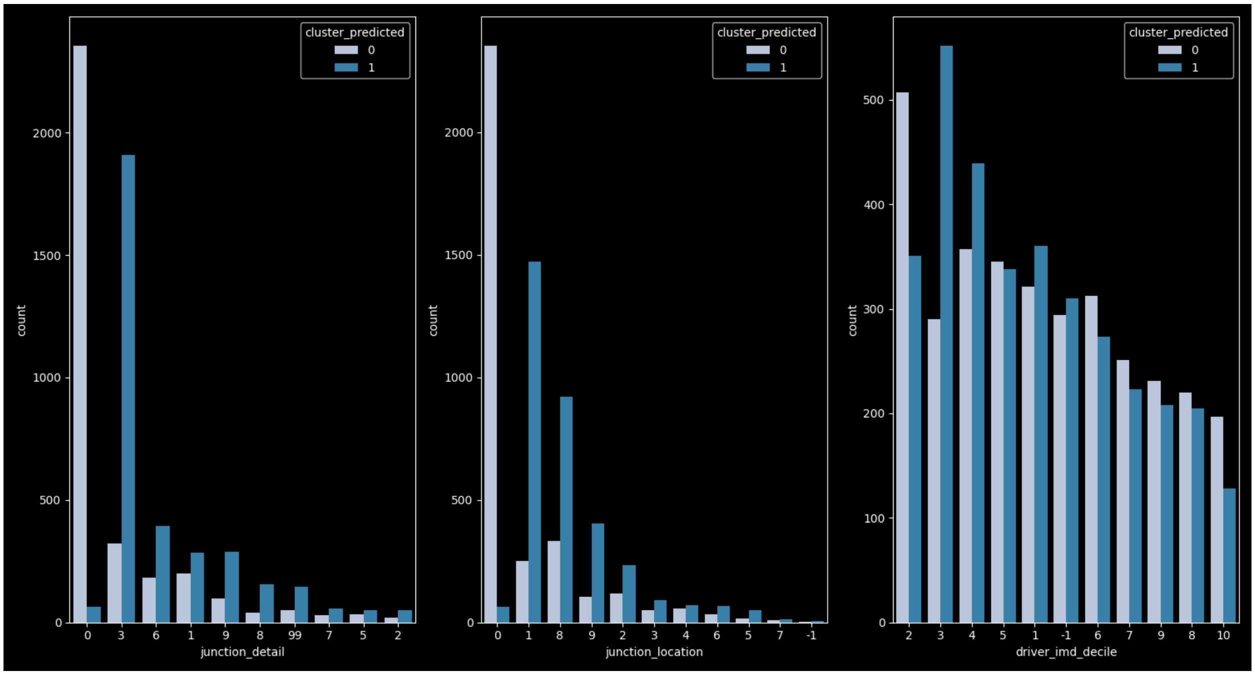
Following plots depict the occurrence of accidents based on different variables.

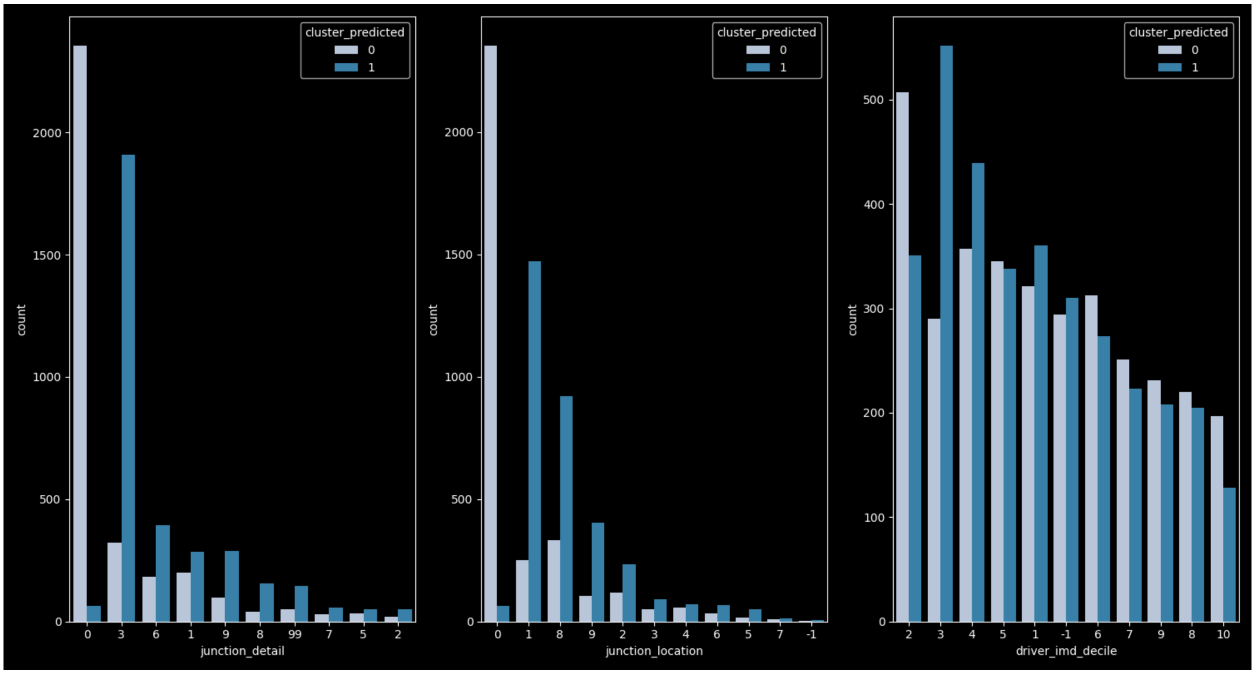

cluster_predicted
0
1
count
0
500
1000
1500
2000
0 3 6 1 9 8 99 7 5 2
junction_detail
cluster_predicted
0
1
count
0
500
1000
1500
2000
0 1 8 9 2 3 4 6 5 7 -1
junction_location
cluster_predicted
0
1
count
0
100
200
300
400
500
2 3 4 5 1 -1 6 7 9 8 10
driver_imd_decile


cluster_predicted
0
1
count
0
250
500
750
1000
1250
1500
1750
2000
30 20 60 40 50 70
speed_limit
cluster_predicted
0
1
count
0
500
1000
1500
2000
2500
1 2
urban_or_rural_area
cluster_predicted
0
1
count
0
500
1000
1500
2000
0 1 9 5 2 -1
skidding_and_overturning

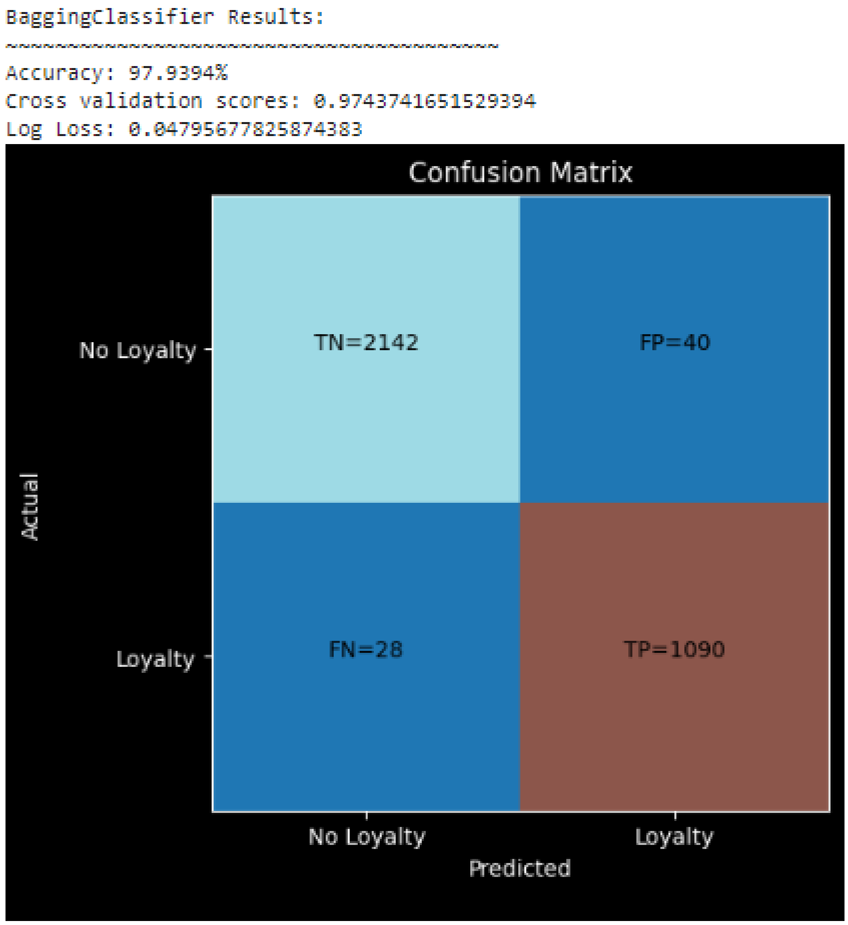

cluster_predicted
0
1
count
0
500
1000
1500
2000
2500
3000
0 9 1 7 2 3 8 4 -1 5 6
vehicle_leaving_carriageway
1 2 3
sex_of_driver
250
750
1250
1750
3 5 4
vehicle_type


cluster_predicted
0
1
count
0
500
1000
1500
2000
2500
4 -1 2 9 3 1
junction_control
1 2 3 4 5 6 7
number_of_casualties
200
400
600
800
6 7 5 4 8 9 -1 10 11 3
age_band_of_driver

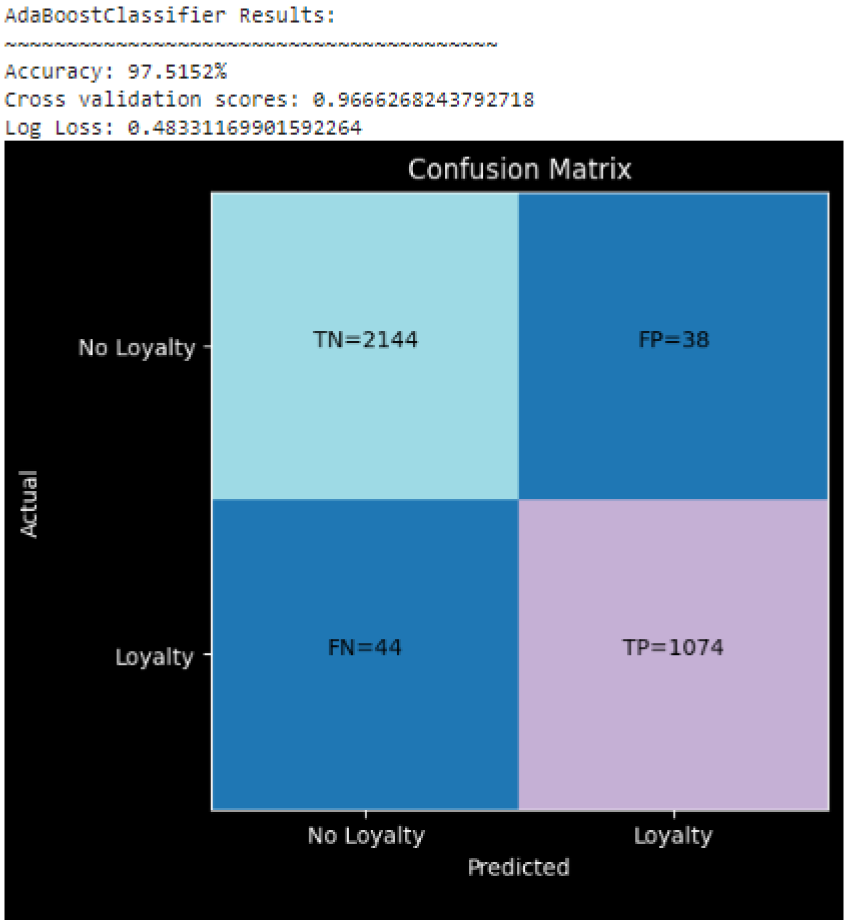

cluster_predicted
0
1
count
0
500
1000
1500
2000
0 3 6 1 9 8 99 7 5 2
junction_detail
cluster_predicted
0
1
count
0
500
1000
1500
2000
0 1 8 9 2 3 4 6 5 7 -1
junction_location
cluster_predicted
0
1
count
0
100
200
300
400
500
2 3 4 5 1 -1 6 7 9 8 10
driver_imd_decile


cluster_predicted
0
1
count
0
500
1000
1500
2000
0 3 6 1 9 8 99 7 5 2
junction_detail
cluster_predicted
0
1
count
0
500
1000
1500
2000
0 1 8 9 2 3 4 6 5 7 -1
junction_location
cluster_predicted
0
1
count
0
100
200
300
400
500
2 3 4 5 1 -1 6 7 9 8 10
driver_imd_decile

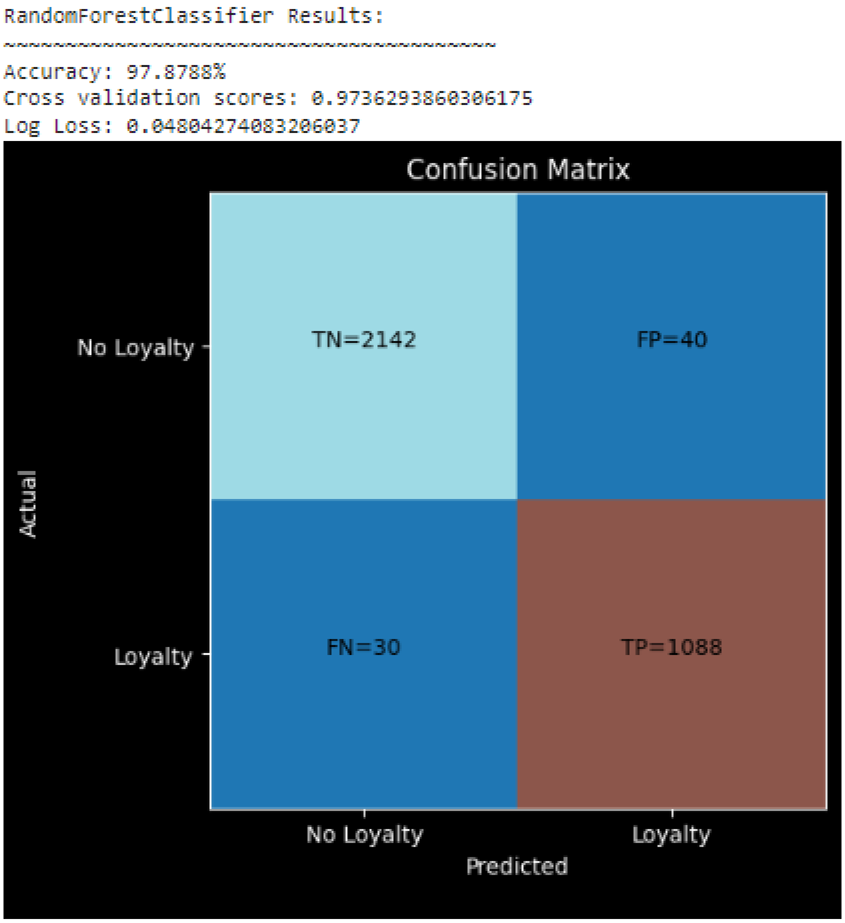
Results of different classifiers on prediction model:



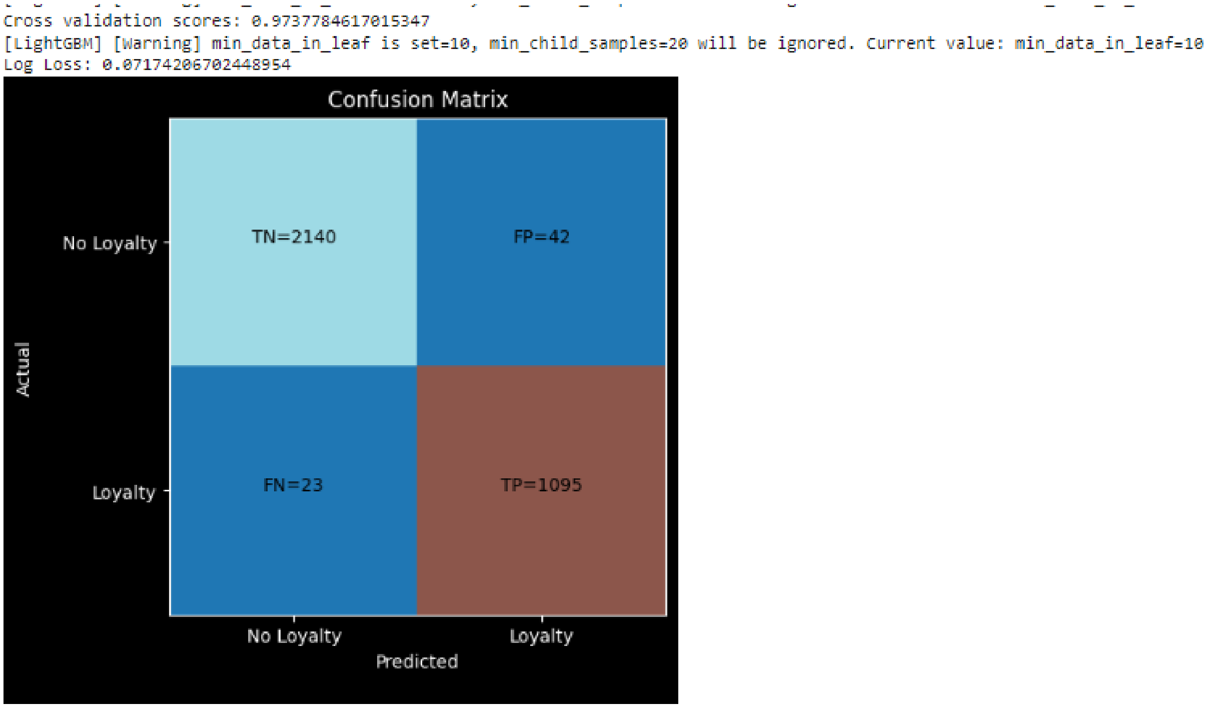
```
BaggingClassifier Results:
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Accuracy: 97.9394%
Cross validation scores: 0.9743741651529394
Log Loss: 0.04795677825874383
```



```
AdaBoostClassifier Results:
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Accuracy: 97.5152%
Cross validation scores: 0.9666268243792718
Log Loss: 0.48331169901592264
```



```
RandomForestClassifier Results:
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Accuracy: 97.8788%
Cross validation scores: 0.9736293860306175
Log Loss: 0.04804274083206037
```



```
Cross validation scores: 0.9737784617015347
[LightGBM] [Warning] min_data_in_leaf is set=10, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=10
Log Loss: 0.07174206702448954
```

Recommendation:

According to the World Health Organization mortality the global incidence of traffic accidents is a cause for concern. These accidents result in the deaths of 1.2 million individuals and injuries to 50 million others. This translates to approximately 3300 fatalities and 137000 injuries per day. With direct economic losses totaling 43 billion dollars. The frequent occurrence of traffic accidents poses a direct threat to both human lives and property safety. Predicting road accidents has become a crucial research area in traffic safety. Various factors include the geometric features of roads, traffic flow, driver characteristics, and environmental conditions, significantly influence the likelihood of road traffic accidents. Numerous studies have been undertaken to forecast accident frequencies and analyze accident characteristics, encompassing research on identifying hazardous locations or hotspots, assessing injury severities, and examining accident durations. Some studies delve into the mechanisms of accidents, while additional factors such as weather and road lighting conditions are also considered.