PROJECT REPORT

INTRODUCTION

This study delves into 2020 road traffic accident (RTA) data to comprehend causes, risks, and solutions. By analyzing accident, casualty, vehicle, and LSOA tables, it aims to address hotspot identification, contributing factors, vulnerable road user safety, trends, and predictive modeling for accident prevention, offering insights to enhance road safety measures.

CHARACTERISTIC FEATURES OF THE RTA DATASET

The RTA dataset for year 2020 consists of the following tables:

A. Accident Table: The dataset contains 91,119 records with 36 columns, featuring diverse data types such as objects, floats, and integers. Key attributes include accident index, year, reference, and geographic coordinates, with a primary key being 'accident index'.

B. Vehicle Table: The dataset contains 167,375 records with 28 columns, including objects and integers. Notable attributes include vehicle index, accident year, and references, with a primary key 'vehicle index' and foreign key 'accident index'.

C. Casualty Table: The dataset holds 115,584 records with 19 columns, including objects and integers. Notable attributes like casualty index, accident year, and references are present, with a primary key 'casualty index' and foreign key 'accident index'.

D. Lower Layer Super Output Area (LSOA) Table: The dataset contains 34,378 records with 7 columns, featuring diverse data types including objects, floats, and integers. It encompasses relevant information in various attributes.

DATA SOURCE/COLLECTION

The RTA datasets (accident, vehicle, casualty, and LSOA) for year 2020 was queried from the Structured Query Language (SQL) database ‘accident_data_v1.0.0_2023’ where ‘accident_year = 2020’. The datasets were further converted into a Pandas DataFrame.

DATA PREPROCESSING

The 'df.isnull()' method effectively detected 14 occurrences of missing values in the datasets, primarily in columns like location easting osgr, location northing osgr, longitude, and latitude. Additionally, columns with '-1' entries indicate 'data missing or out of range,' as per the Road Safety Open Dataset Guide were also detected.

Data Cleaning

Histograms with kernel density estimation (KDE) were utilized to assess the skewness of distributions in location easting osgr, location northing osgr, longitude, and latitude columns, aiding in addressing null values. The resulting charts are displayed below.

Figure 1: Distribution of Location Easting Osgr

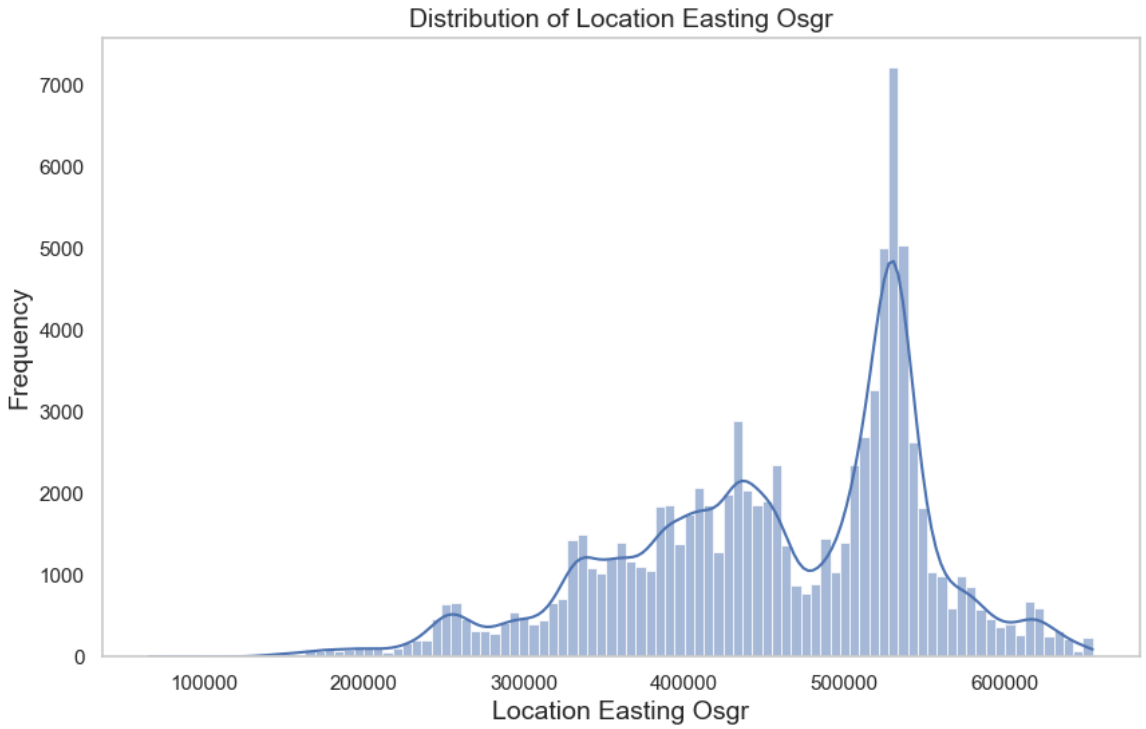


Figure 2: Distribution of Location Northing Osgr

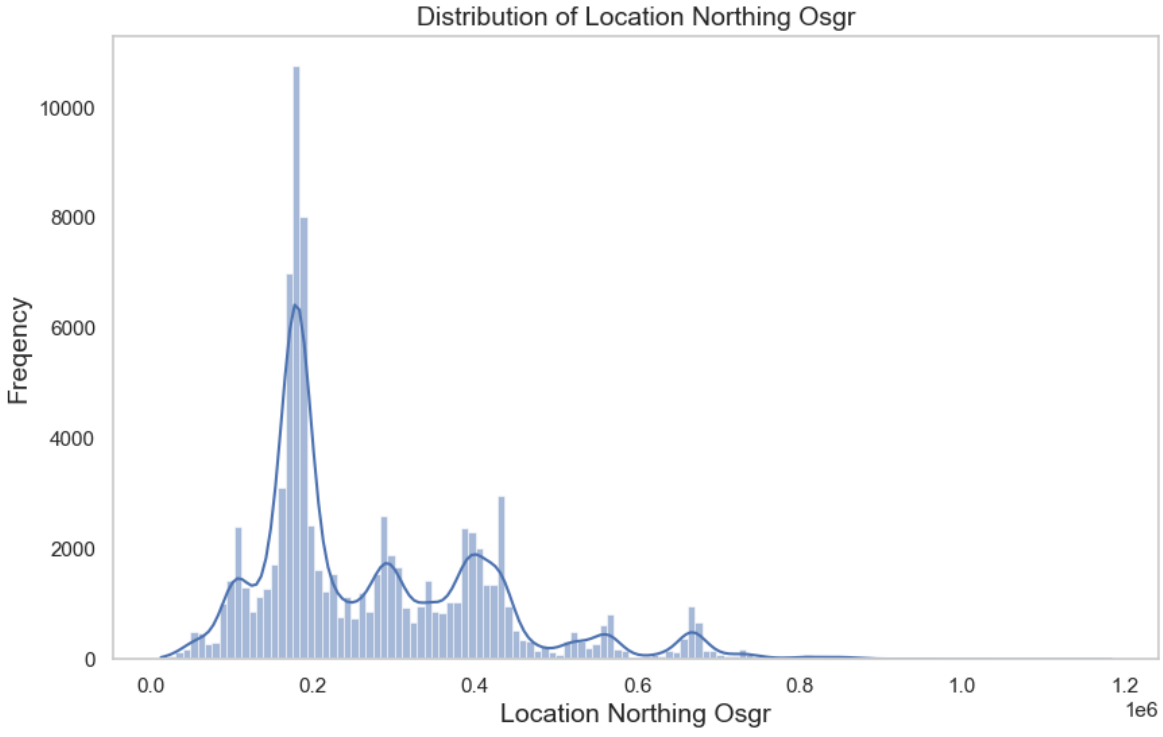


Figure 3: Distribution of Longitude

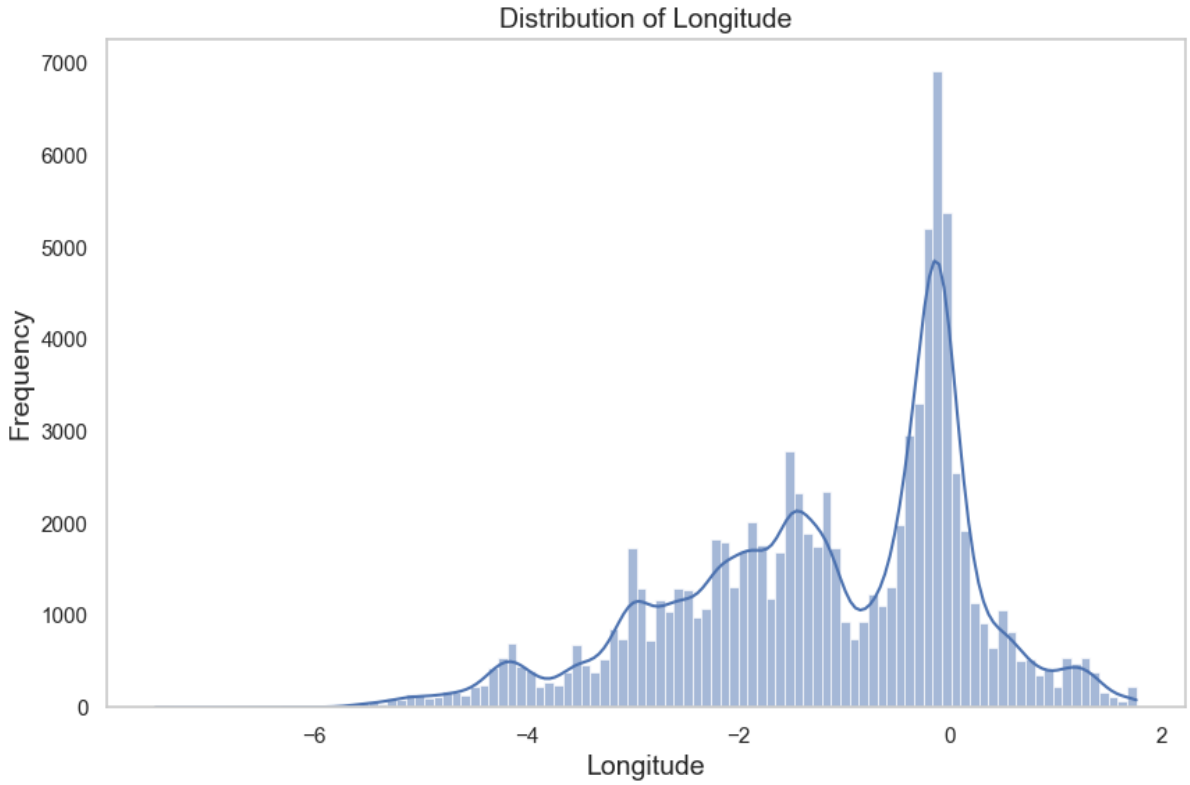
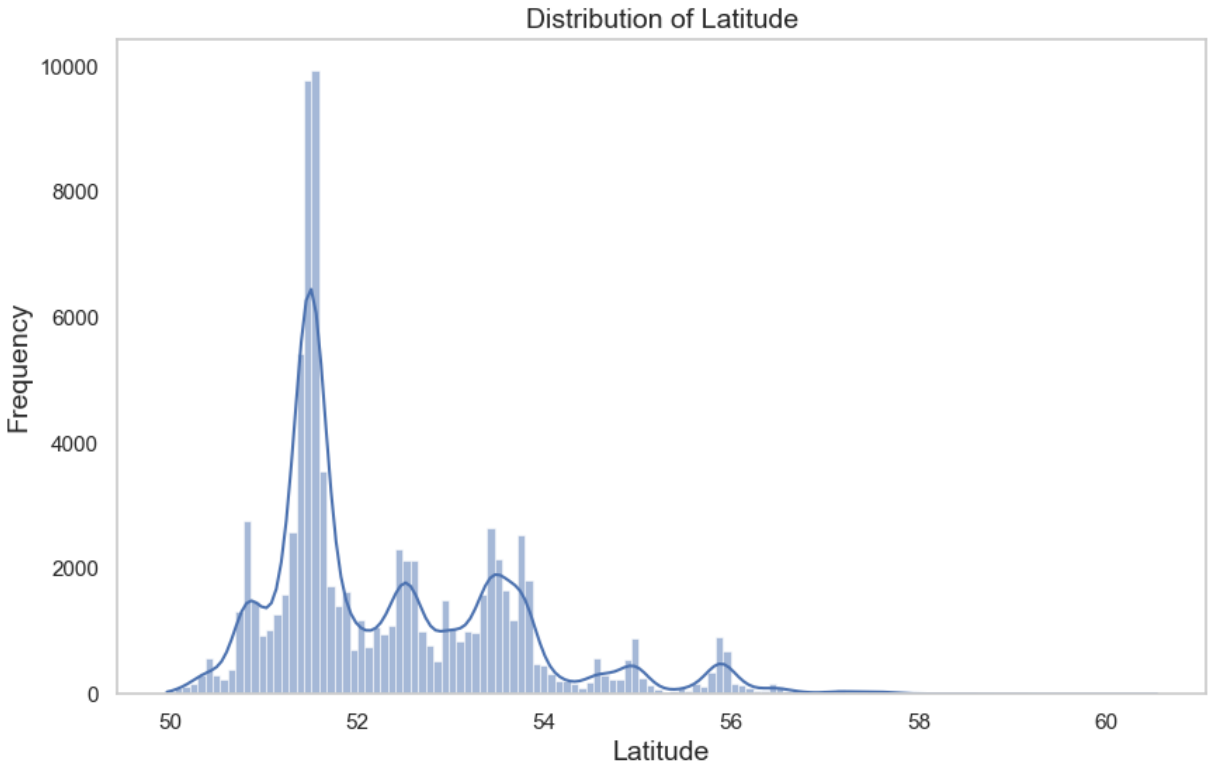


Figure 4: Distribution of Latitude



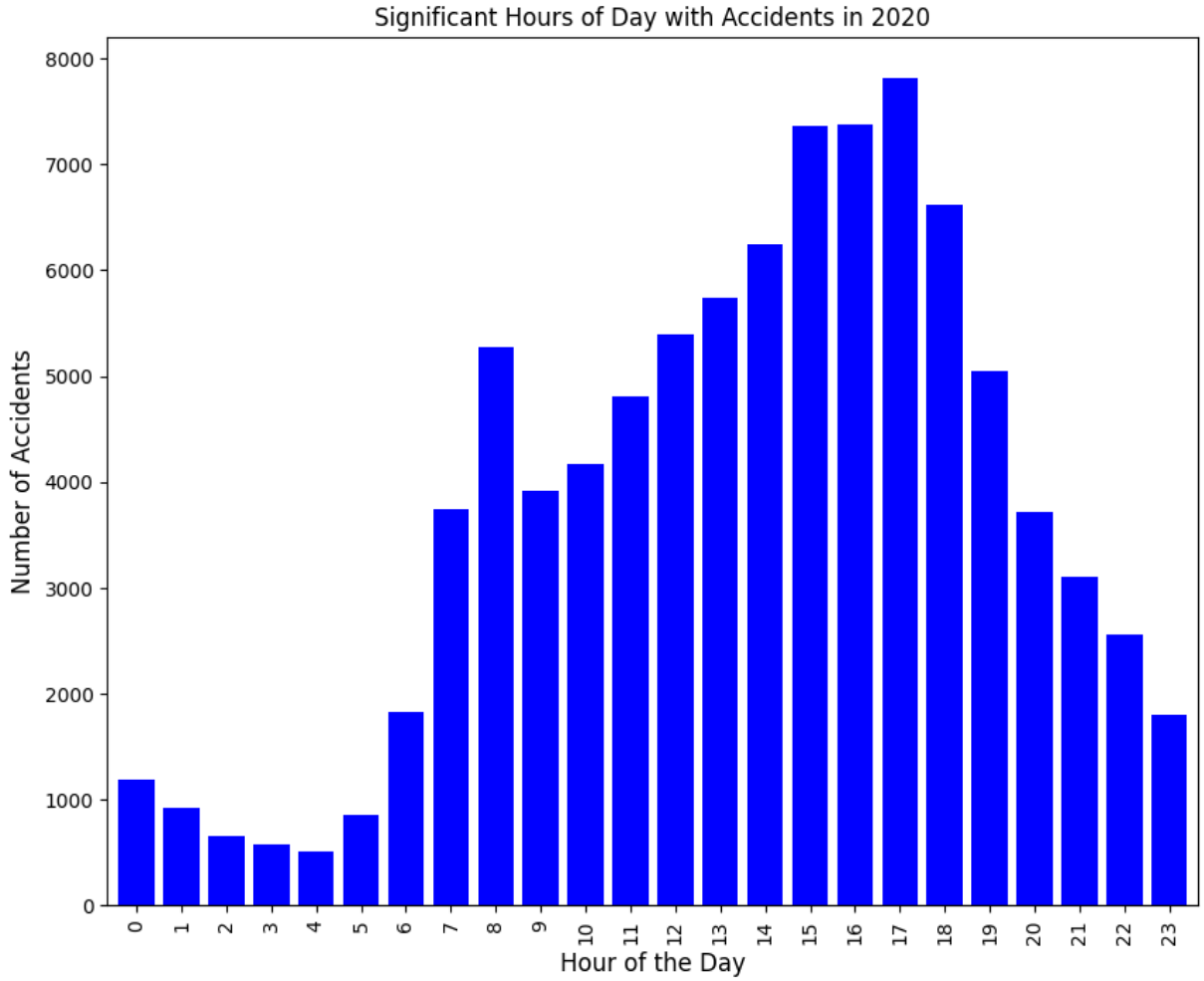
Findings: Location features exhibit skewed distributions: location easting osgr and longitude are left-skewed (Figures 1 and 3), while location northing osgr and latitude are right-skewed (Figures 2 and 4). Median imputation was preferred for handling missing values due to its resilience against skewness and outliers (Wise-Answer, 2020). Further insights available in the linked Jupyter notebook.

DATA ANALYSIS

Question 1A: Are there Significant Hours of the Day on which Accidents Occur?

Converting timestamps into hours revealed a significant pattern, with the 17th hour having the highest accident count of 7813 incidents in the year 2020. Conversely, the 4th hour had the lowest count of 508 incidents, indicating distinct peak and off-peak periods for accidents. To encapsulate the findings, a chart is provided below.

Figure 5: Significant Hours of the Day with Accidents in 2020



Question 1B: Are there Significant Days of the Week on which Accidents Occur?

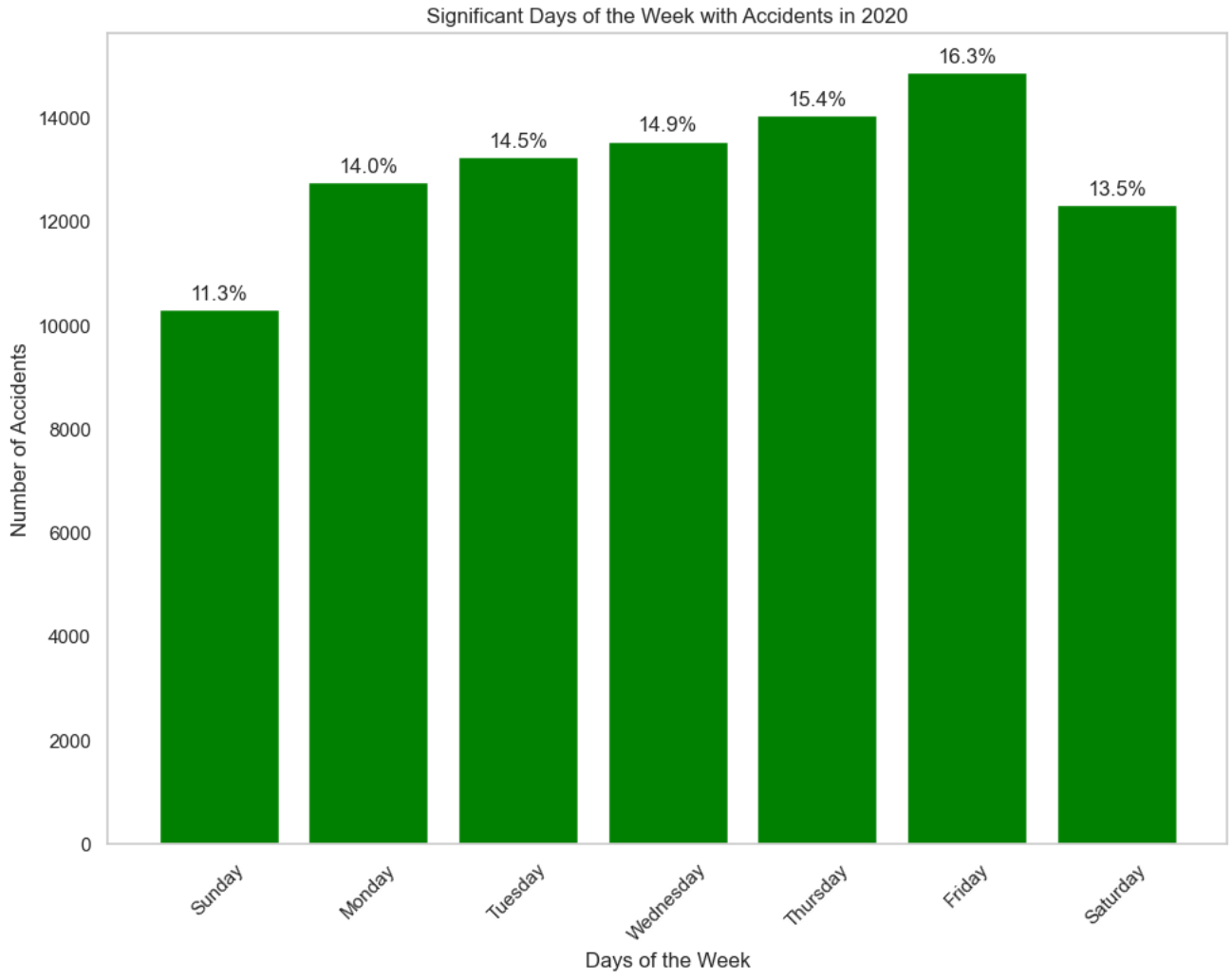
A prominent pattern emerges:

a. Fridays stand out with the highest accident count (16.3%), signifying elevated risk.

b. Accidents decline through the week, with lower counts on weekends.

c. Sundays show fewer accidents, suggesting reduced road risk. Vigilance on Fridays and potential safety measures on Sundays are highlighted.

Figure 6: Significant Days of the Week with Accidents in 2020



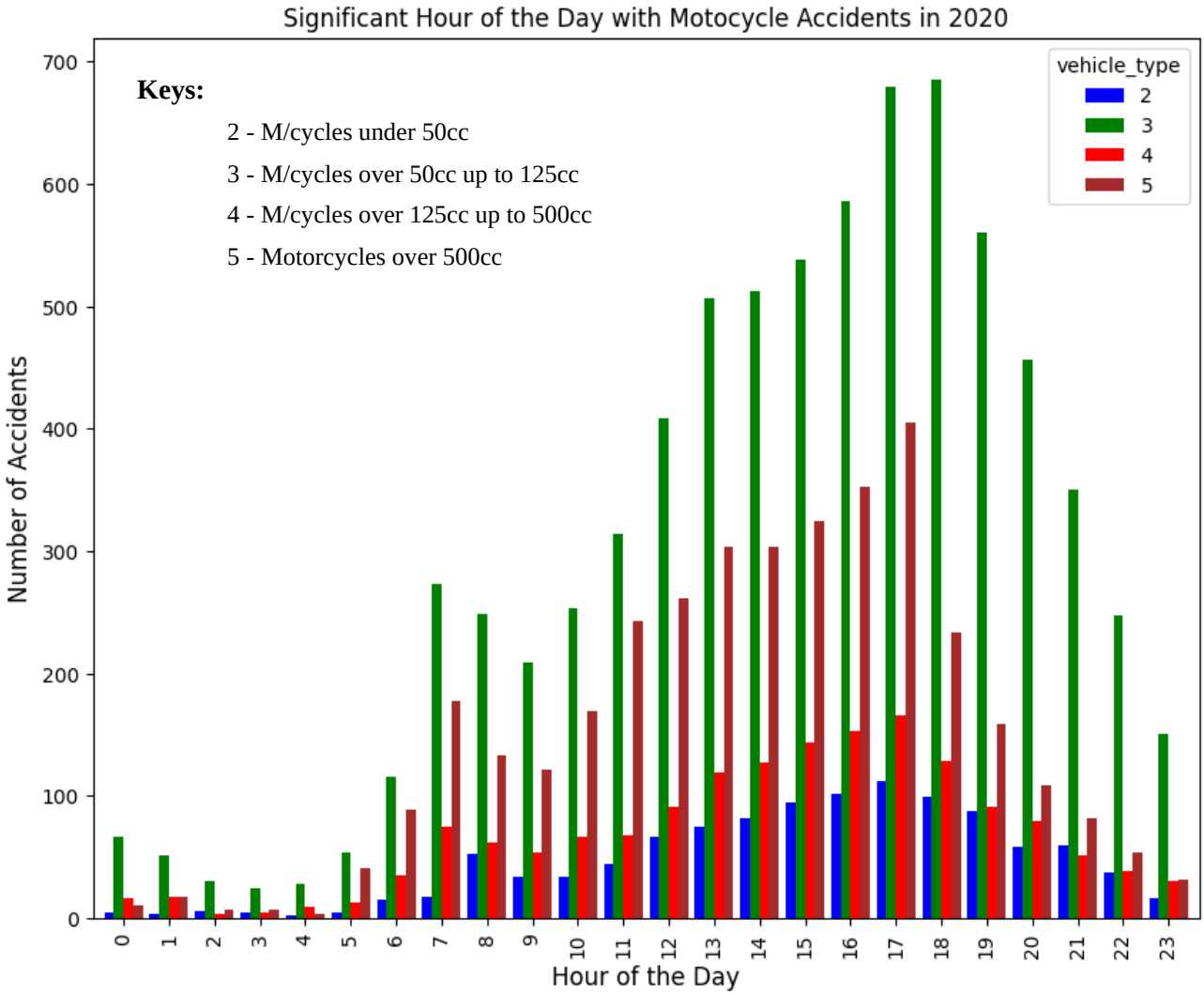
Question 2A: For Motorbikes, are there Significant Hours of the Day on which Accidents Occur? Focus on: Motorcycle 125cc and Under, Motorcycle Over 125cc and up to 500cc, and Motorcycle Over 500cc.

A distinct pattern emerges:

a. Notable motorcycle accident hours based on engine size: 17th hour for under 50cc, over 125cc up to 500cc, and over 500cc; 18th hour for over 50cc up to 125cc.

b. Different motorcycle segments show unique accident trends, highlighting critical safety hours (17th and 18th) for measures and awareness. This underscores targeted accident prevention during those times. The chart is provided below.

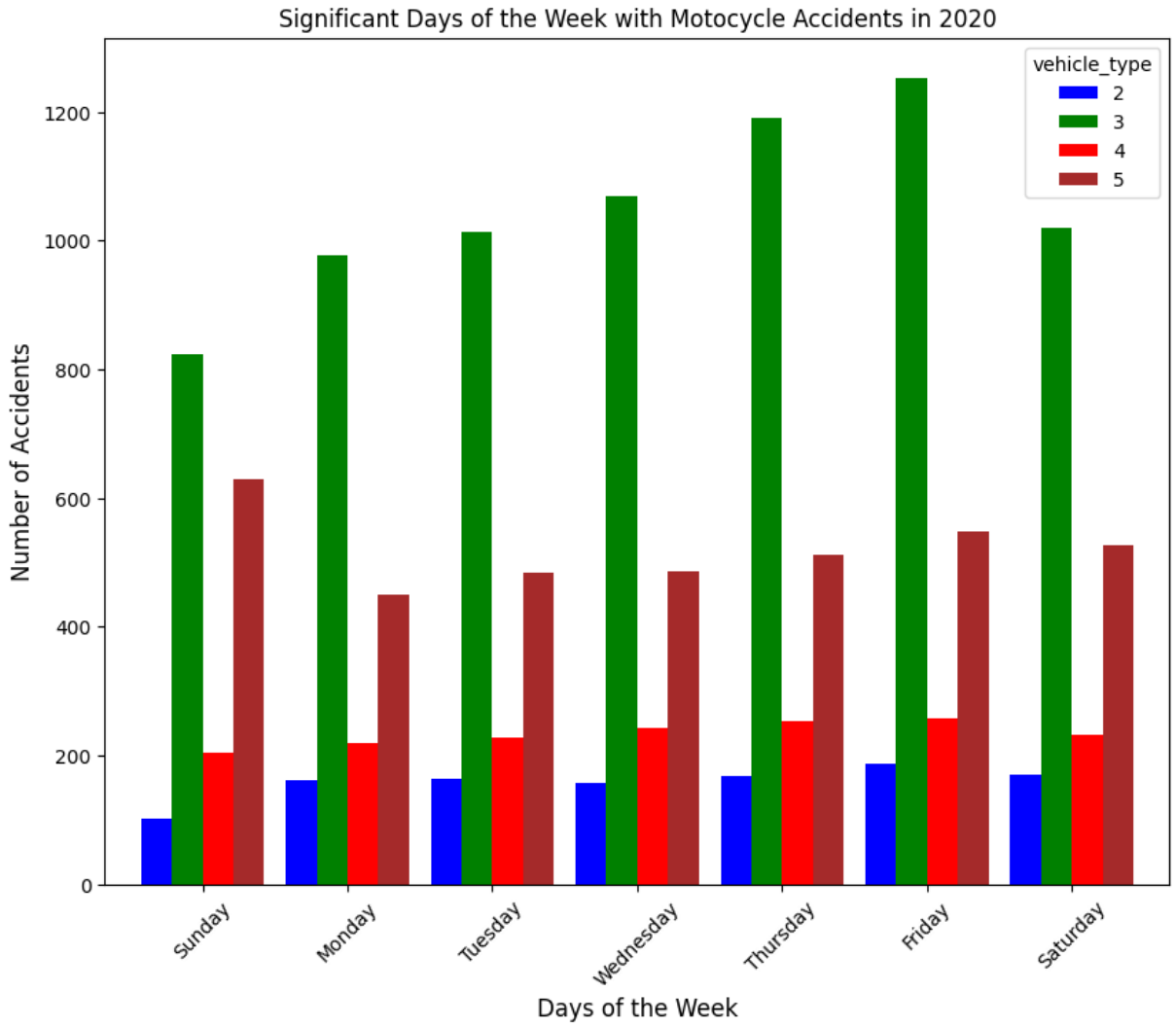
Figure 7: Significant Hours of the Day with Motorcycle Accidents in 2020



Question 2B: For Motorbikes, Are There Significant Days of the Week on Which Accidents Occur? Focus on: Motorcycle 125cc and Under, Motorcycle Over 125cc and up to 500cc, and Motorcycle Over 500cc

Approximately 15% of total accidents involve motorcycles (about 13740 incidents), revealing distinct patterns by category. Fridays show peaks for under 125cc and over 125cc up to 500cc, while Sundays stand out for 'Motorcycle over 500cc,' emphasizing the importance of targeted safety measures on specific days.

Figure 7: Significant Hours of the Day with Motorcycle Accidents in 2020



Keys:

2 - M/cycles under 50cc

3 - M/cycles over 50cc up to 125cc

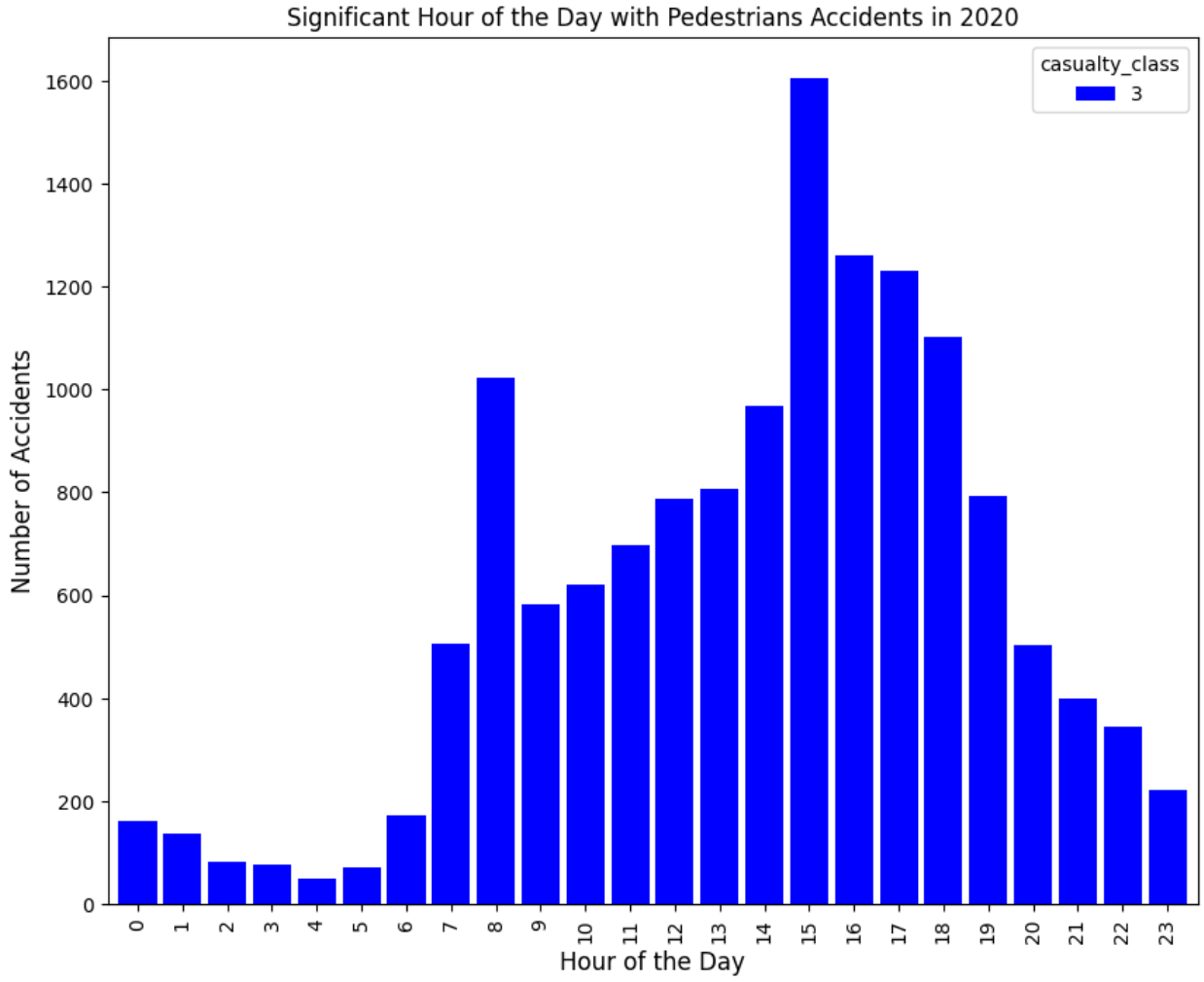
4 - M/cycles over 125cc up to 500cc

5 - Motorcycles over 500cc

Question 3A: For Pedestrians Involved in Accidents, Are There Significant Hours of the Day, on Which They Are More Likely to Be Involved?

Pedestrian accidents constituted approximately 15.5% (14193 incidents), with the analysis highlighting a notable peak at the 15th hour as the highest risk period, contrasting with a safer period between the 2nd and 3rd hours. The decreasing trend from the 16th to the 4th hour emphasizes diminishing risk during late afternoon to early morning.

Figure 8: Significant Hours of the Day with Pedestrians Accidents in 2020

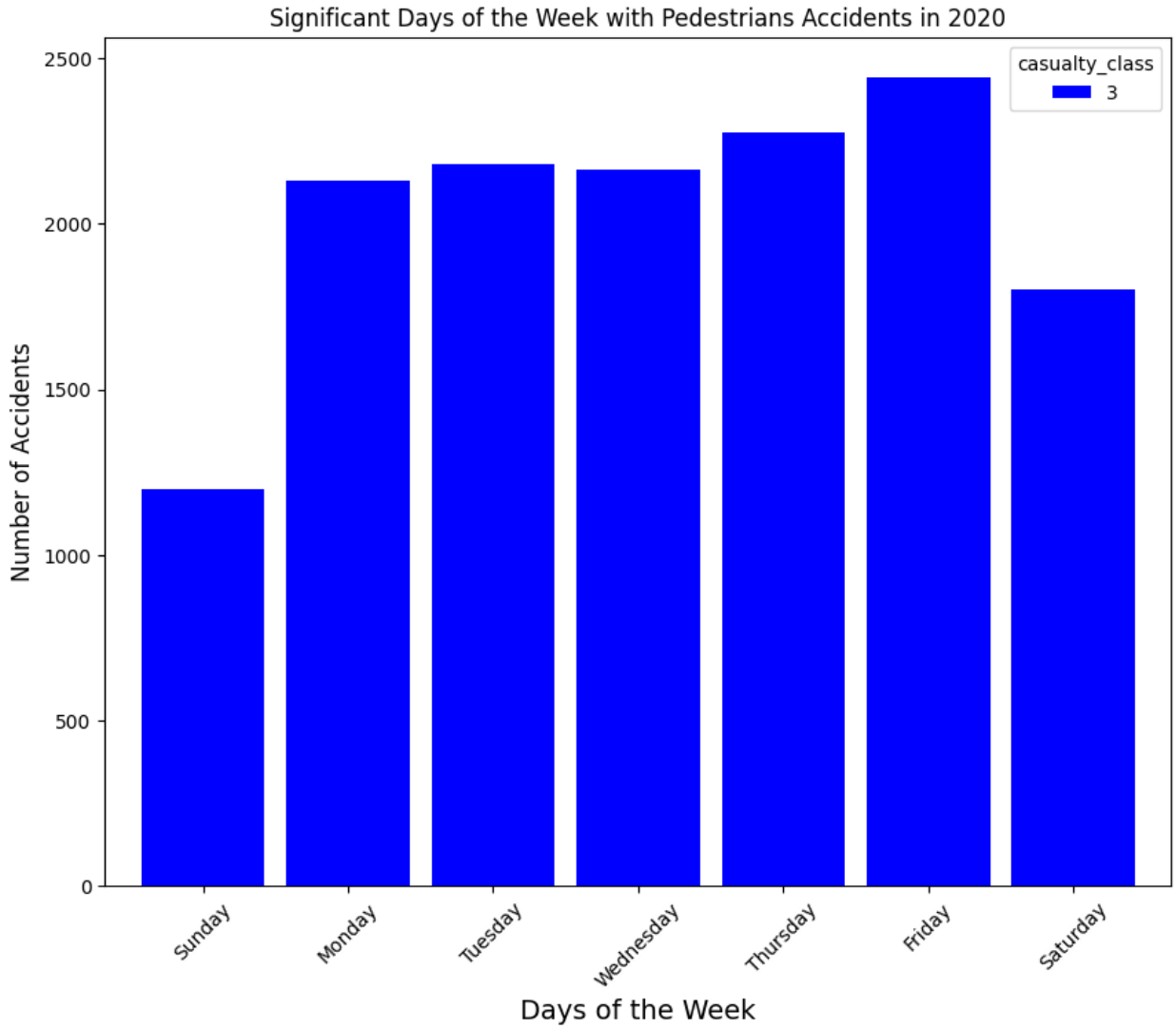


Question 3B: For Pedestrians Involved in Accidents, Are There Significant Days of the Week, on Which They Are More Likely to Be Involved?

Distinct patterns in pedestrian accidents across days of the week reveal the highest count on Fridays (2442 incidents), followed by a decline throughout the week. Safety measures should be targeted on Fridays, and opportunities exist for enhanced

pedestrian safety interventions throughout the week, as indicated by the decreasing accident counts on weekends.

Figure 9: Significant Days of the Week with Pedestrians Accidents in 2020



Question 4: Using the Apriori Algorithm to Explore the Impact of Selected Variables on Accident Severity

The Apriori algorithm was utilized to analyze the impact of factors such as speed limit, weather conditions, and light conditions on accident severity. The variables were encoded using pd.get_dummies() to generate binary values (0 and 1), and then subjected to the Apriori algorithm with specified thresholds, revealing significant associations and findings.

	support	itemsets
0	0.201263	(severity_2)
1	0.783484	(severity_3)
2	0.573033	(speed_30)
3	0.775546	(weather_1)
4	0.706784	(light_1)
5	0.208621	(light_4)
6	0.459983	(speed_30, severity_3)
7	0.603186	(weather_1, severity_3)
8	0.559337	(light_1, severity_3)
9	0.450137	(weather_1, speed_30)
10	0.406540	(light_1, speed_30)
11	0.578669	(weather_1, light_1)
12	0.359697	(weather_1, speed_30, severity_3)
13	0.330409	(light_1, severity_3, speed_30)
14	0.453810	(weather_1, light_1, severity_3)
15	0.337273	(weather_1, light_1, speed_30)
16	0.272097	(weather_1, light_1, severity_3, speed_30)

Table 1: Result of Apriori Algorithm

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(weather_1)	(light_1)	0.775546	0.706784	0.578669	0.746144	1.055688	0.030525	1.155047	0.235018
1	(light_1)	(weather_1)	0.706784	0.775546	0.578669	0.818735	1.055688	0.030525	1.238263	0.179904
2	(weather_1, severity_3)	(light_1)	0.603186	0.706784	0.453810	0.752354	1.064475	0.027487	1.184013	0.152641
3	(light_1)	(weather_1, severity_3)	0.706784	0.603186	0.453810	0.642077	1.064475	0.027487	1.108656	0.206571
4	(weather_1, speed_30)	(light_1)	0.450137	0.706784	0.337273	0.749269	1.060111	0.019124	1.169445	0.103120
5	(light_1, speed_30)	(weather_1)	0.406540	0.775546	0.337273	0.829620	1.069724	0.021983	1.317376	0.109830
6	(weather_1)	(light_1, speed_30)	0.775546	0.406540	0.337273	0.434885	1.069724	0.021983	1.050159	0.290392
7	(light_1)	(weather_1, speed_30)	0.706784	0.450137	0.337273	0.477194	1.060111	0.019124	1.051755	0.193380
8	(weather_1, speed_30, severity_3)	(light_1)	0.359697	0.706784	0.272097	0.756463	1.070288	0.017869	1.203987	0.102564
9	(light_1, severity_3, speed_30)	(weather_1)	0.330409	0.775546	0.272097	0.823516	1.061853	0.015850	1.271809	0.086994
10	(weather_1, severity_3)	(light_1, speed_30)	0.603186	0.406540	0.272097	0.451100	1.109609	0.026878	1.081181	0.248936
11	(weather_1, speed_30)	(light_1, severity_3)	0.450137	0.559337	0.272097	0.604477	1.080703	0.020319	1.114127	0.135808
12	(light_1, severity_3)	(weather_1, speed_30)	0.559337	0.450137	0.272097	0.486464	1.080703	0.020319	1.070739	0.169463
13	(light_1, speed_30)	(weather_1, severity_3)	0.406540	0.603186	0.272097	0.669301	1.109609	0.026878	1.199923	0.166450
14	(weather_1)	(light_1, severity_3, speed_30)	0.775546	0.330409	0.272097	0.350846	1.061853	0.015850	1.031482	0.259519
15	(light_1)	(weather_1, speed_30, severity_3)	0.706784	0.359697	0.272097	0.384979	1.070288	0.017869	1.041108	0.223972

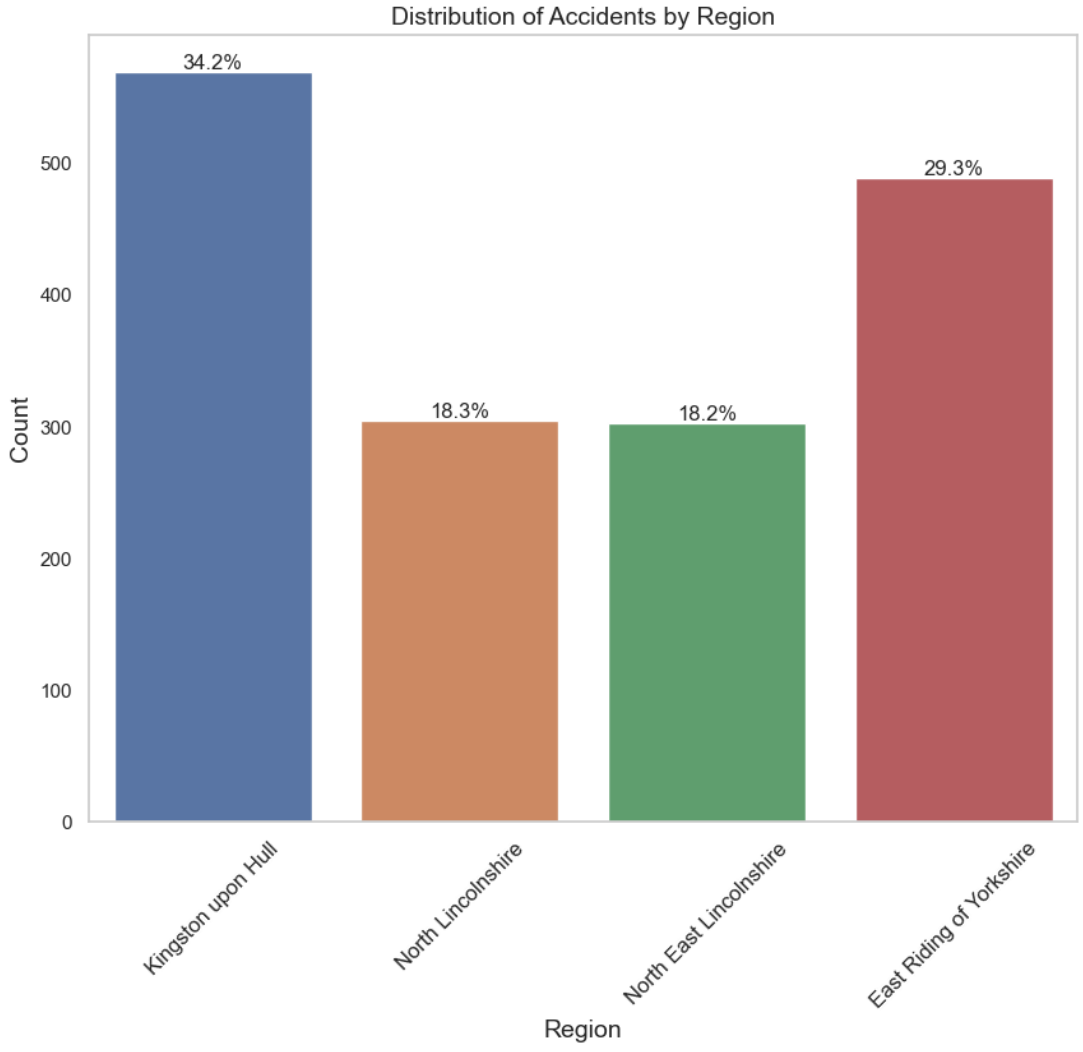
Table 2: Result of Association Rules

Table 1 reveals significant associations between 30mph speed limit and slight severity (46%) and between slight severity and 'Fine without high winds' weather condition (60%) in 2020 RTA dataset, while Table 2's lift value of 1.109609 underscores positive correlations between severity, weather, light conditions, and speed limit.

Question 5: Identify Accidents in Our Region: Kingston upon Hull, Humberside, and the East Riding of Yorkshire, etc. Run Clustering on This Data. What Do These Clusters Reveal About the Distribution of Accidents in Our Region?

Data for the analysis was obtained by merging accident and LSOA tables for the year 2020, filtering by police force '16' (Humberside). Around 34.2% (569) of the 1663 accidents occurred in Kingston upon Hull, while North Lincolnshire and North East Lincolnshire had similar occurrence rates of approximately 18.2%, as shown in the chart.

Figure 9: Distribution of Accidents by Region (Humberside) in 2020



The optimal number of clusters (5 clusters) was determined from silhouette scores shown in Figure 10, guiding the application of K-means algorithm on geographical coordinates (latitude and longitude) of regions, visualized in Figure 11.

Figure 10: Silhouette Scores for Different Number of Clusters

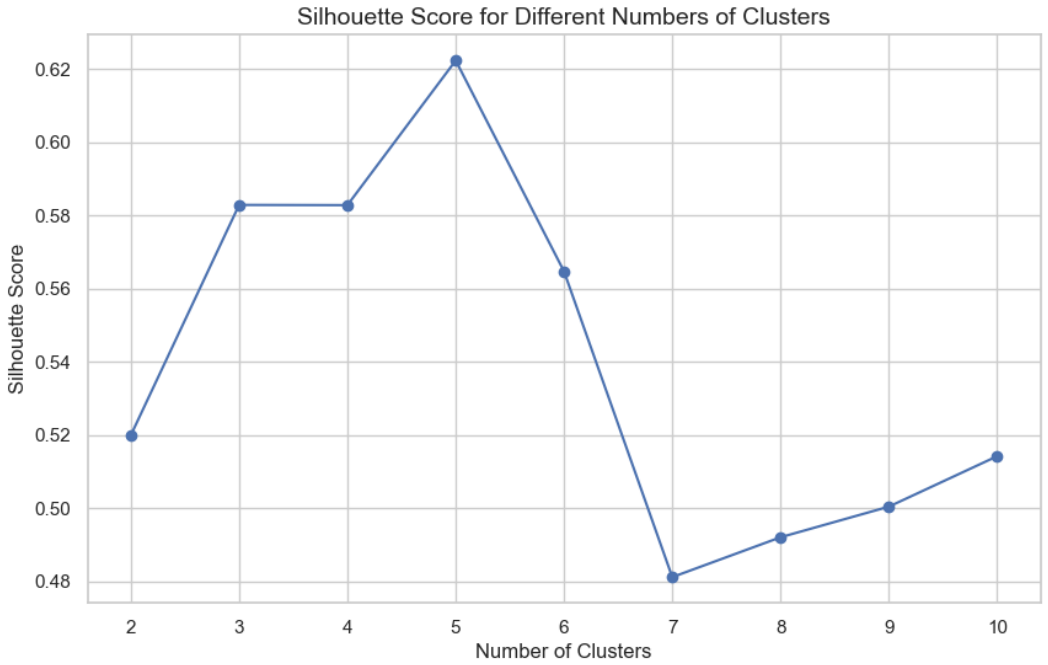
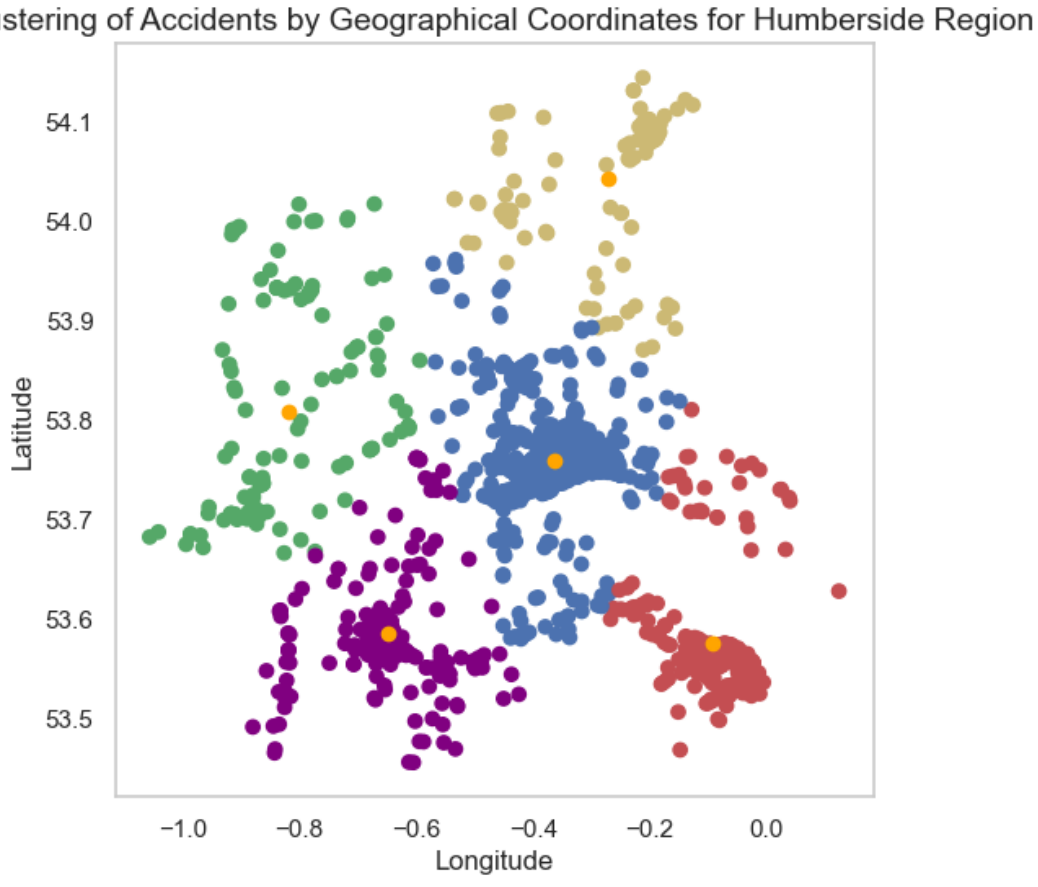
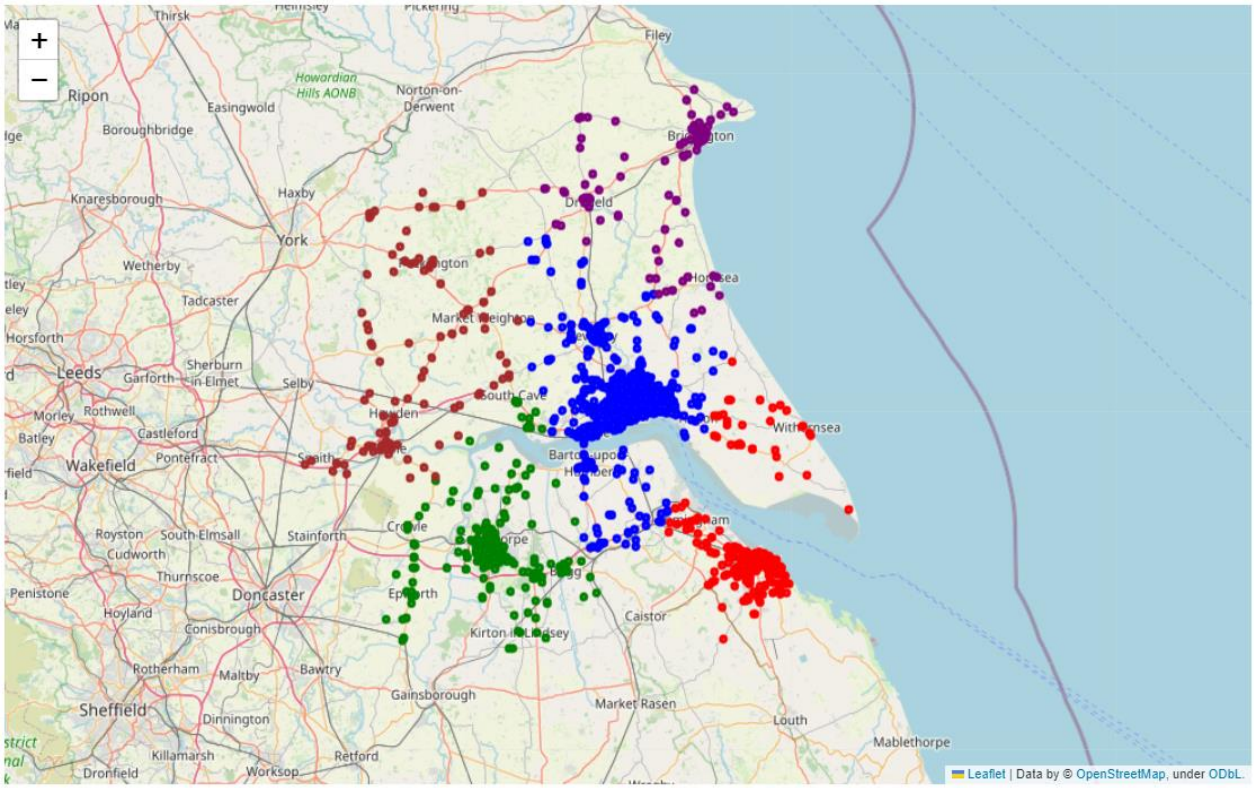


Figure 11: Clustering by Geographical Coordinates for Humberside Region



The Folium library was employed to create a map displaying clustered points (geographical coordinates), showcased in the map below.



Source: Open Street Map

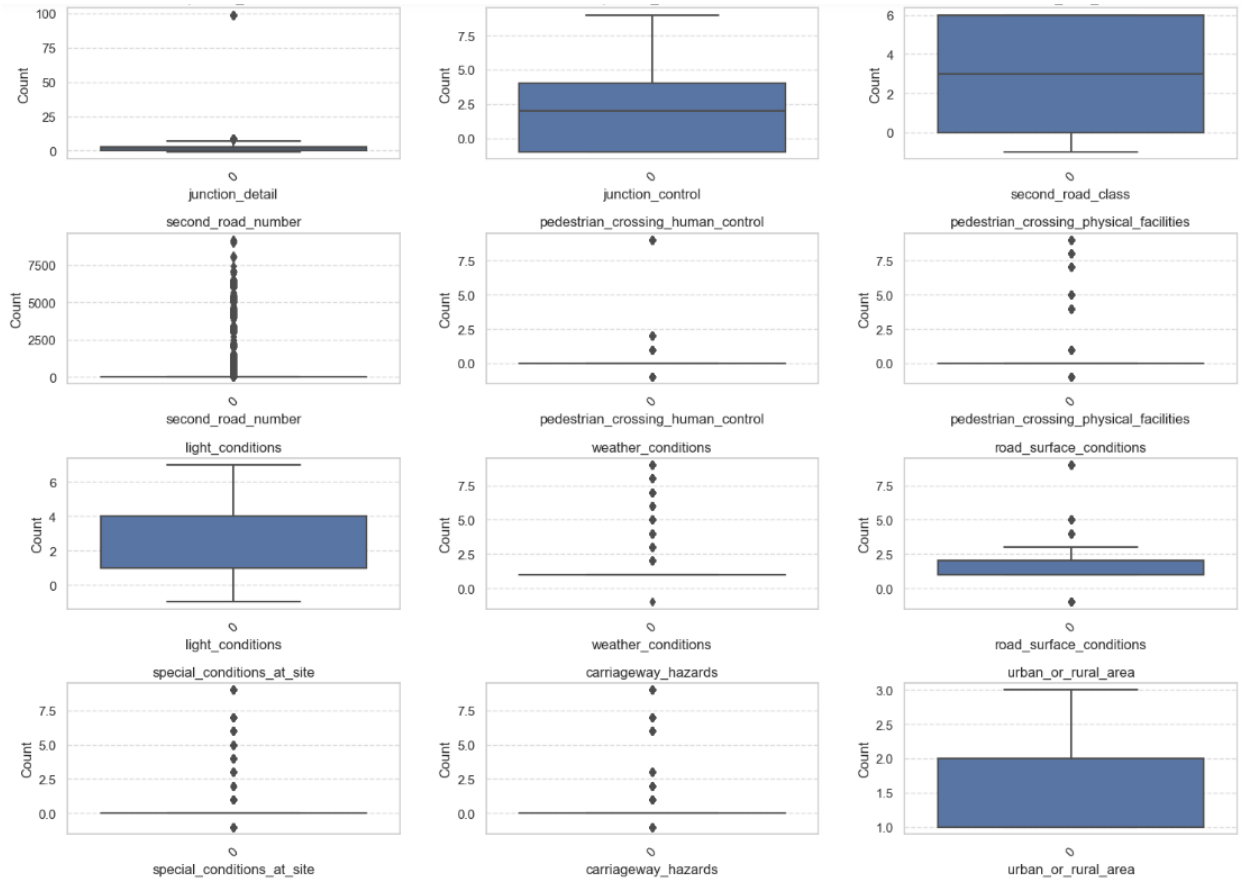
Analyzing the map depiction, it becomes apparent that Kingston upon Hull (Hull) (depicted in blue) had the highest occurrence of accidents in the year 2020, outpacing all other regions. A close second is North East Lincolnshire (Grimsby), signifying a notable concentration of accidents. Occupying the third spot is North Lincolnshire (Scunthorpe), while East Riding of Yorkshire (Bridlington) secures the fourth position. Notably, East Riding of Yorkshire (Goole) follows suit, establishing its presence. Other areas show decreasing accident frequencies and some regions are assigned to different clusters.

Question 6: Using Outlier Detection Methods, Identify Unusual Entries in Your Data Set. Should You Keep These Entries in Your Data?

Box plots, Local Outlier Factor (LOF), and Isolation Forest methods were employed in identifying anomalous entries within the dataset. The results are summarized below:

A. Boxplot: The boxplot analysis unveiled the presence of outliers which were mostly '-1' entries within our dataset. The chart is presented below.

Figure 12: Boxplot of Accident Dataset in 2020



B. Local Outlier Factor (LOF): Utilizing the LOF algorithm, 228 outlier records were identified among 90971 geographical coordinates in the 2020 accident dataset with a 0.0025 contamination level. These mild outliers, shown in blue (Figure 13), and 5 outliers within the Humberside region, displayed in green (Figure 14), reflect minor deviations from the norm. Retaining these outliers depends on domain knowledge and contextual relevance.

C. Isolation Forest (IF): Applying the Isolation Forest (IF) algorithm to the 2020 accident data, 912 outlier records were identified from 90287 data points with a contamination level of 0.01. The resulting chart (Figure 15) displays the geographical coordinates (longitude and latitude), with outliers depicted in red, indicating subtle deviations from the normal points.

Figure 13: Outlier Detection on The Accident Data (Geographical Coordinates) in 2020 by Local Outlier Factor (LOF)

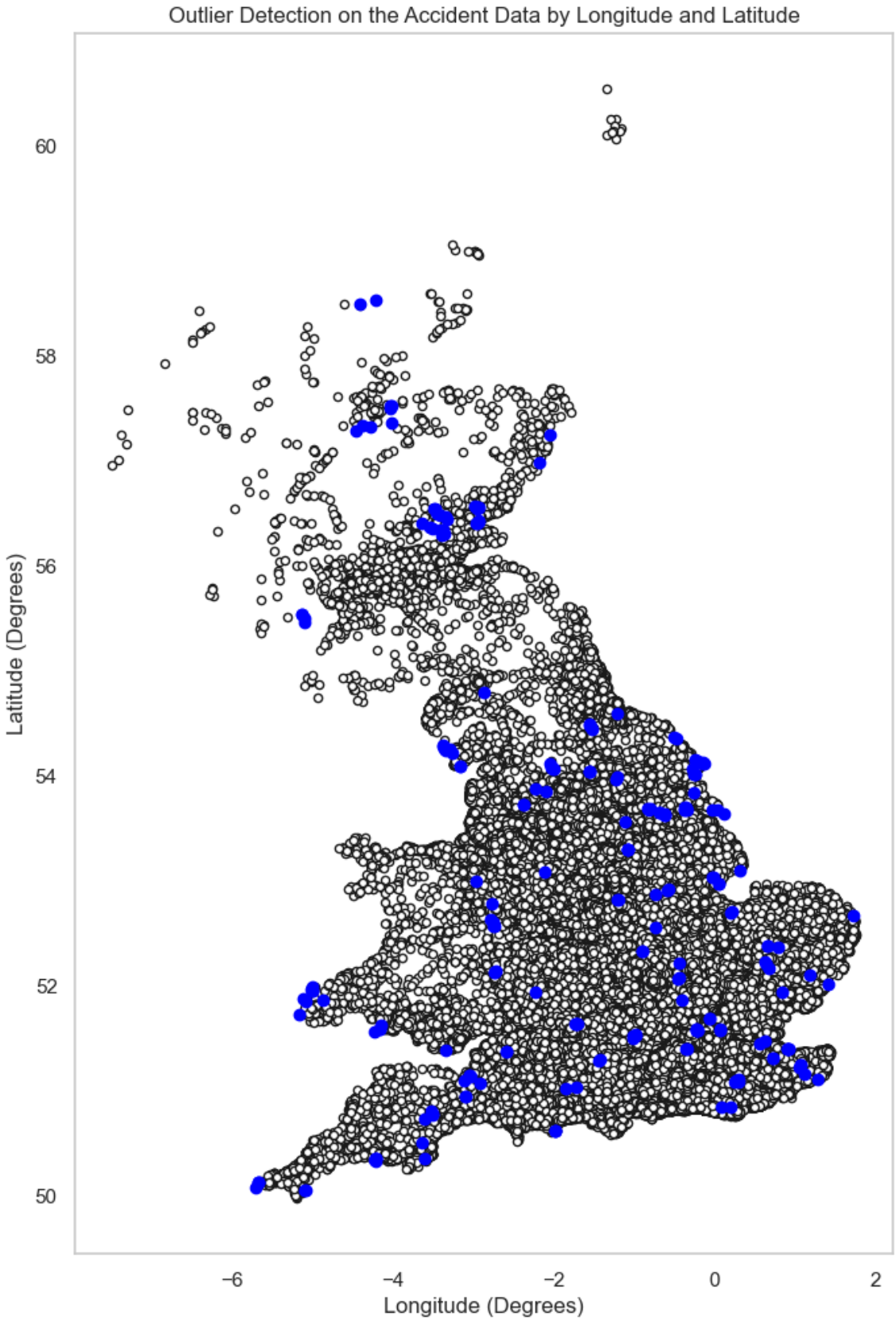


Figure 14: Outlier Detection on The Accident Data (Humberside Region Geographical Coordinates) in 2020 by Local Outlier Factor (LOF)

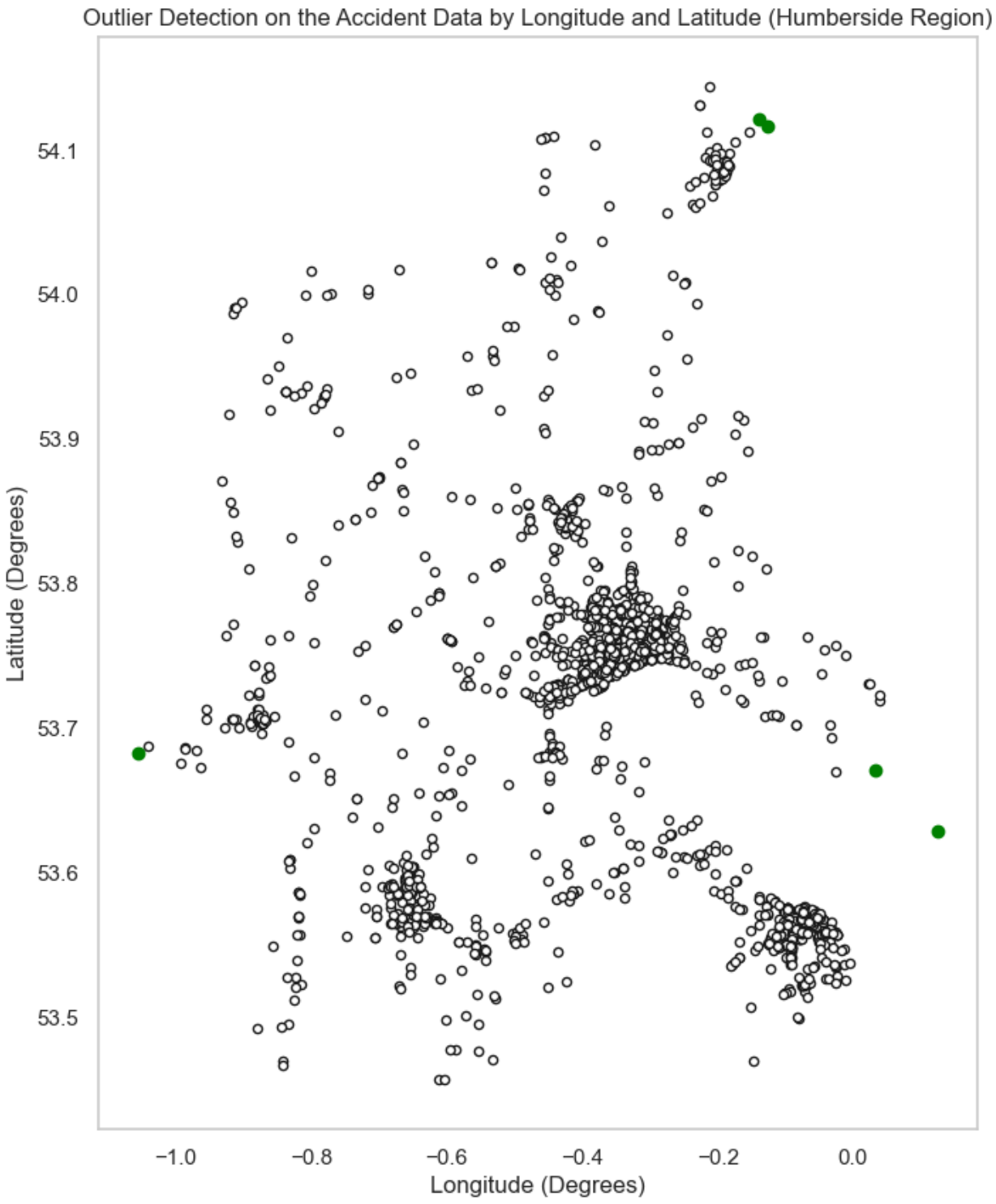
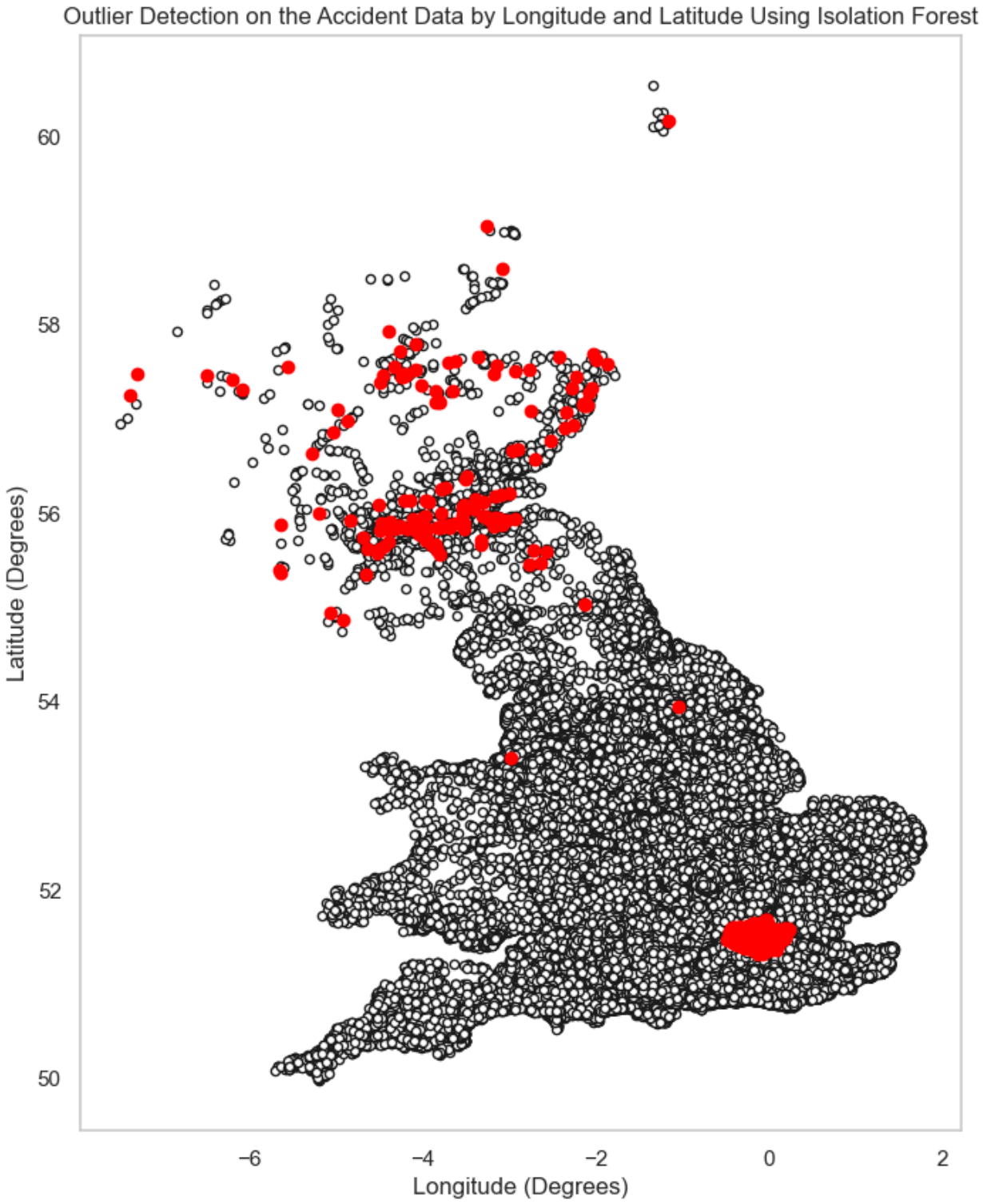


Figure 15: Outlier Detection on The Accident Data (Geographical Coordinates) in 2020 by Isolation Forest



Question 7: Can You Develop a Classification Model Using the Provided Data That Accurately Predicts Fatal Injuries Sustained in Road Traffic Accidents, With the Aim of Informing and Improving Road Safety Measures?

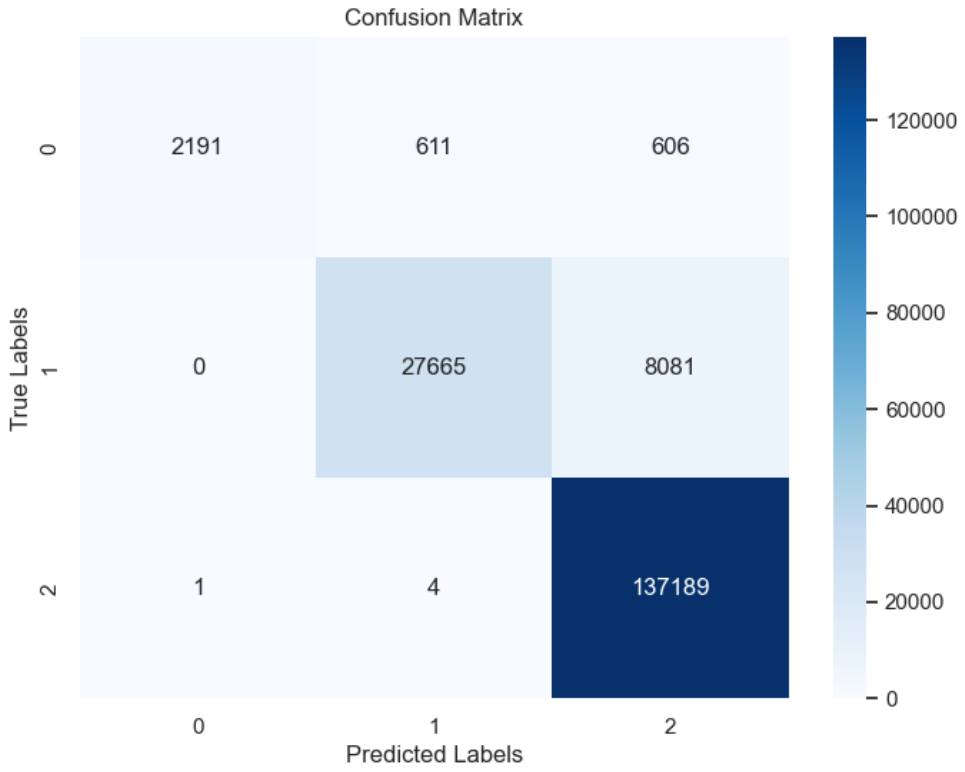
The selected features for predicting fatal injuries in RTA included speed, longitude, latitude, speed limit, road surface conditions, weather conditions, towing and articulation, hit object in carriageway, light conditions, and more. Missing values represented as '-1' were imputed using the mode for categorical features (Goyal, 2021). The data was split into training and testing sets (80:20 ratio), and a **Gradient Boosting Classifier** model was employed. The classification report and confusion matrix for the training data are provided below.

For training data:

```
Accuracy: 0.9472463538004401
Classification Report:
              precision    recall  f1-score   support

           1       1.00      0.64      0.78      3408
           2       0.98      0.77      0.86     35746
           3       0.94      1.00      0.97    137194

    accuracy                           0.95    176348
   macro avg       0.97      0.81      0.87    176348
weighted avg       0.95      0.95      0.94    176348
```

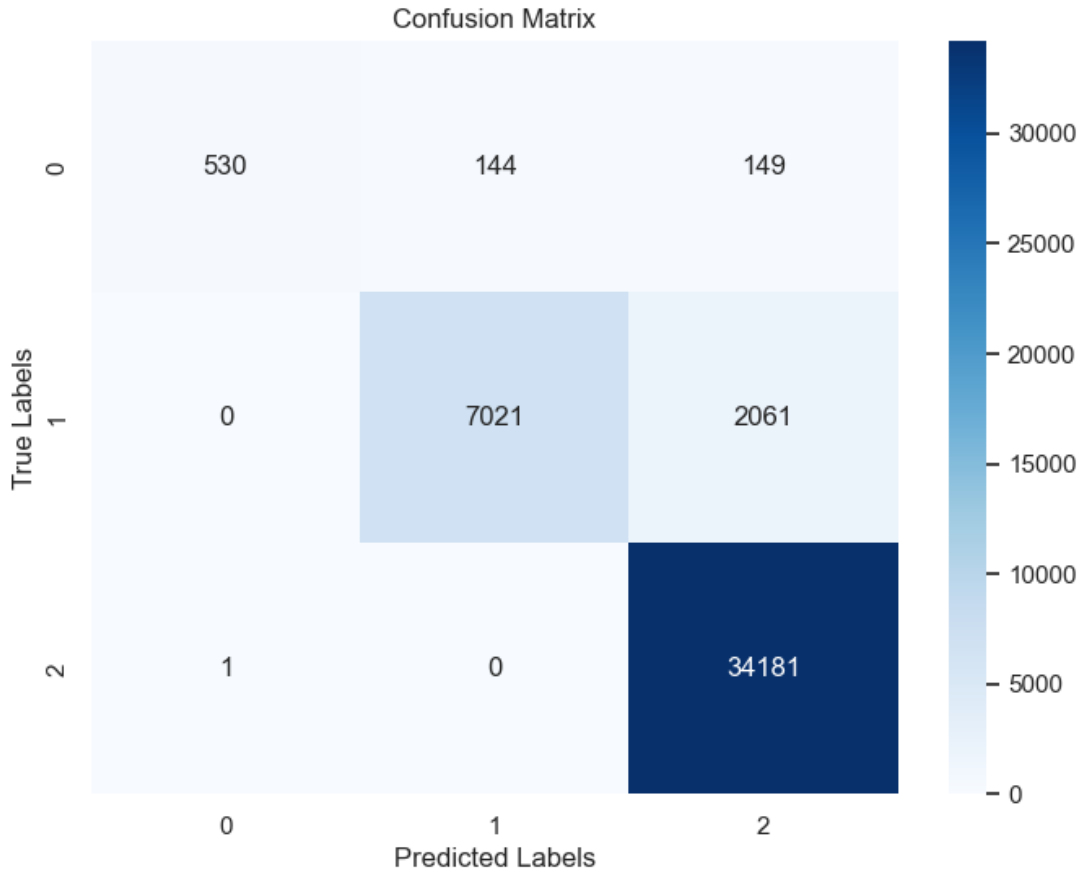


For testing data:

```
Accuracy: 0.9465828929162792
Classification Report:
              precision    recall  f1-score   support

           1       1.00      0.64      0.78       823
           2       0.98      0.77      0.86      9082
           3       0.94      1.00      0.97     34182

    accuracy                           0.95     44087
   macro avg       0.97      0.81      0.87     44087
weighted avg       0.95      0.95      0.94     44087
```



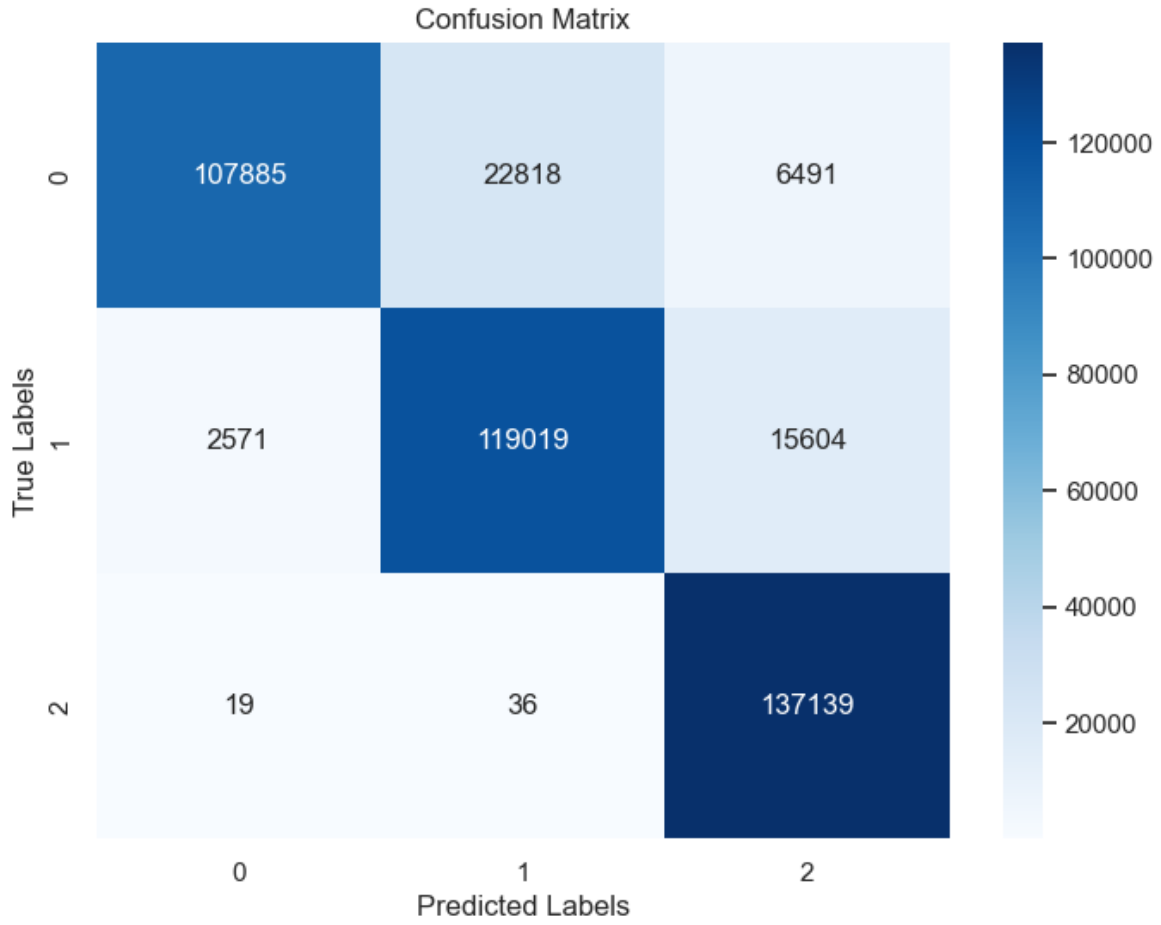
The model attained accuracy rates of 0.9472 and 0.9465 on training and testing data, respectively, misclassifying 1217 and 293 labels. The training set was balanced using SMOTE, leading to improved performance, as evidenced by the provided classification report and confusion matrix.

For training data:

```
Accuracy: 0.8844968924782911
Classification Report:
              precision    recall  f1-score   support

           1       0.98      0.79      0.87    137194
           2       0.84      0.87      0.85    137194
           3       0.86      1.00      0.93    137194

    accuracy                           0.88    411582
   macro avg       0.89      0.88      0.88    411582
weighted avg       0.89      0.88      0.88    411582
```

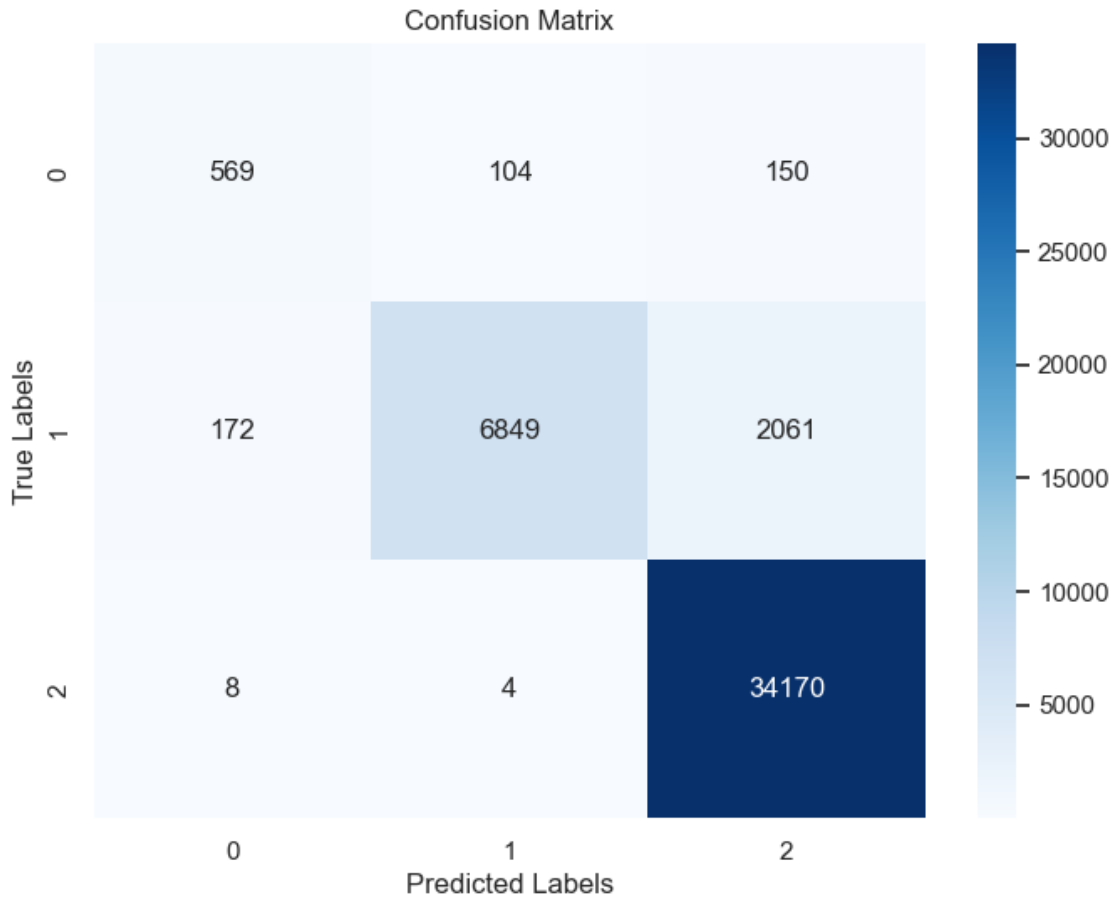


For testing data:

```
Accuracy: 0.9465828929162792
Classification Report:
              precision    recall  f1-score   support

           1       1.00      0.64      0.78       823
           2       0.98      0.77      0.86      9082
           3       0.94      1.00      0.97     34182

    accuracy                           0.95     44087
   macro avg       0.97      0.81      0.87     44087
weighted avg       0.95      0.95      0.94     44087
```



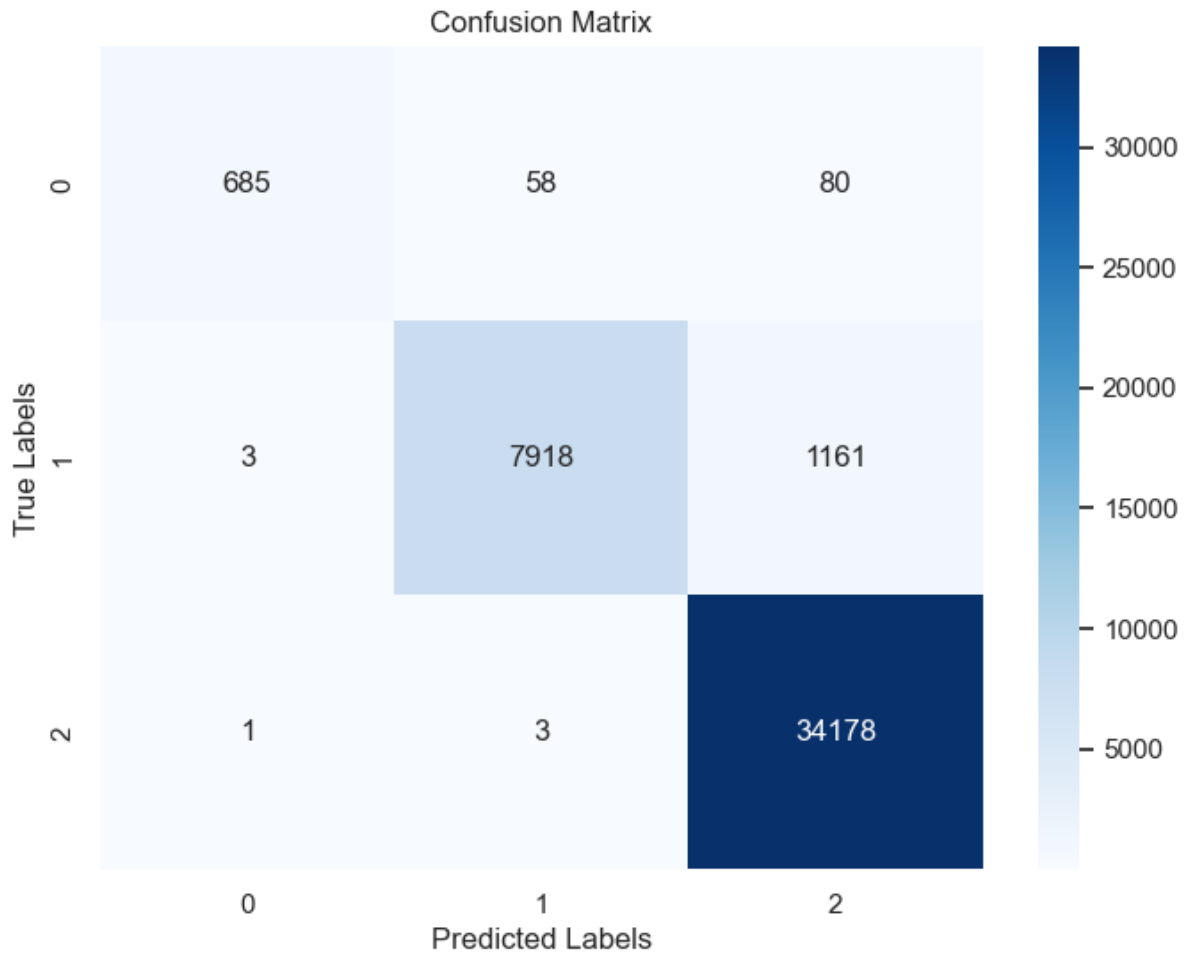
After applying SMOTE, the model achieved 0.8844 accuracy (training) and 0.9465 accuracy (testing), yet misclassified 29309 and 254 labels respectively. Subsequently, the **Random Forest Classifier** was employed to improve accuracy, with the classification report and confusion matrix provided for assessment.

For testing data:

```
Accuracy: 0.9703767550525099
Classification Report:
              precision    recall  f1-score   support

           1       0.99      0.83      0.91       823
           2       0.99      0.87      0.93      9082
           3       0.96      1.00      0.98     34182

    accuracy                           0.97     44087
   macro avg       0.98      0.90      0.94     44087
weighted avg       0.97      0.97      0.97     44087
```



The Random Forest Classifier yielded an improved accuracy of 0.9703 on testing data, though it still misclassified 138 labels. This enhancement highlights the selected variables' effective predictive capability for fatal injuries in RTAs, showcasing a strong relationship.

The significant influence of variables such as longitude, latitude, speed limit, road surface and weather conditions, road type, and lighting conditions and more, on fatal injuries emphasizes the importance of improved road safety measures, including

engineering enhancements, intelligent traffic management, advanced vehicle safety features, and timely weather alerts to mitigate risks.

RECOMMENDATIONS

The analysis underscores the need for focused road safety interventions, especially in Kingston upon Hull, to curb accidents in the Humberside region. Addressing the growing motorcycle and pedestrian accidents is crucial. Strengthening road safety measures during early hours and weekdays can help mitigate risks. Additionally, implementing regulations to prevent minors from driving, expanding dual carriageways, and addressing oil or diesel spillage on roads are essential steps for enhancing overall road safety.

REFERENCES

Department for Transport. (2022). *Reported road casualties in Great Britain: notes, definitions, symbols and conventions*. Available online: https://www.gov.uk/government/publications/road-accidents-and-safety-statistics-notes-and-definitions/reported-road-casualties-in-great-britain-notes-definitions-symbols-and-conventions [Accessed 07/08/2023].

Department for Transport. (2021). *STATS19 road accident injury statistics – report form*. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995422/stats19.pdf [Accessed 20/07/2023].

Goyal, C. (2021). *Handling missing values of categorical variables* [Blog post]. Analytics Vidhya's Blog. 27 April. Available online: https://www.analyticsvidhya.com/blog/2021/04/how-to-handle-missing-values-of-categorical-variables/ [Accessed: 11/08/2023].

Kate. (2022). *What is an acceptable silhouette score* [Blog post]. Kate's Blog. 30 November. Available online: https://thedutchladydesigns.com/what-is-an-acceptable-silhouette-score/ [Accesses 11/08/2023].

Wise-Answer. (2020). *Is median influenced by outliers.* Available online: https://wise-answer.com/is-median-influenced-by-outliers/ [Accessed 11/08/2023].