INTRODUCTION

Accidents are events that happen unexpectedly which can result in unintended loss, injury, and even death. It can be caused by several factors. Data from traffic accident reports sheds light on the scenario that causes road accidents as well as the impact they have on society and the people. This project aims to provide recommendations and suggestions to improve road safety and to develop a model that can predict accident occurrence and related injuries and the best approach to mitigate and prevent accidents. The study employed accident data 2020 extracted from the Great Britain accident database.

METHODOLOGY

Exploratory data analysis was conducted to analyze and investigate the dataset for any abnormalities, imbalances, or variations in the data. Descriptive statistics and visualizations were also employed to make proper decisions and recommendations. The datasets were cleaned sequentially to improve accuracy. The Apriori algorithm was employed for association rule mining to explore the impact of some variables on accident severity and the data was transformed using one hot encoding. Clustering was utilized to identify and reveal the distribution of accidents. Outliers' detections to identify unusual entries using isolation forest. The classification model was built, trained with different hyperparameters to enhance accuracy, and followed by feature selection, random forest, Decision tree and XGBoot were all employed to predict road traffic accident severity.

DATA CLEANING

Column by column, data cleaning was carried out in accordance with each dataset, including the LSOA, vehicle, casualty, and accident tables. The irrelevant columns were removed because the percentage is insignificant. The accident table had 36 rows and 91199 columns; the vehicle table had 28 rows and 167375 columns; the casualty table had 19 rows and 115584 columns; and the lsoa table had 34378 rows and 7 columns. The columns for longitude, latitude, location_easting_osgr, and location_northing_osgr each contain 14 NAN values. The mean "34" was used to replace the -1 value in the vehicle table's age of driver column, and the mean "5" was used to replace the -1 value in the age of the vehicle. Since we are not speculating, the values in the journey_purpose_of_driver' column -1 were changed to '6,' as 6 in the stats20-2011 form denotes unknown.

CamScanner

ACCIDENT DISTRIBUTION

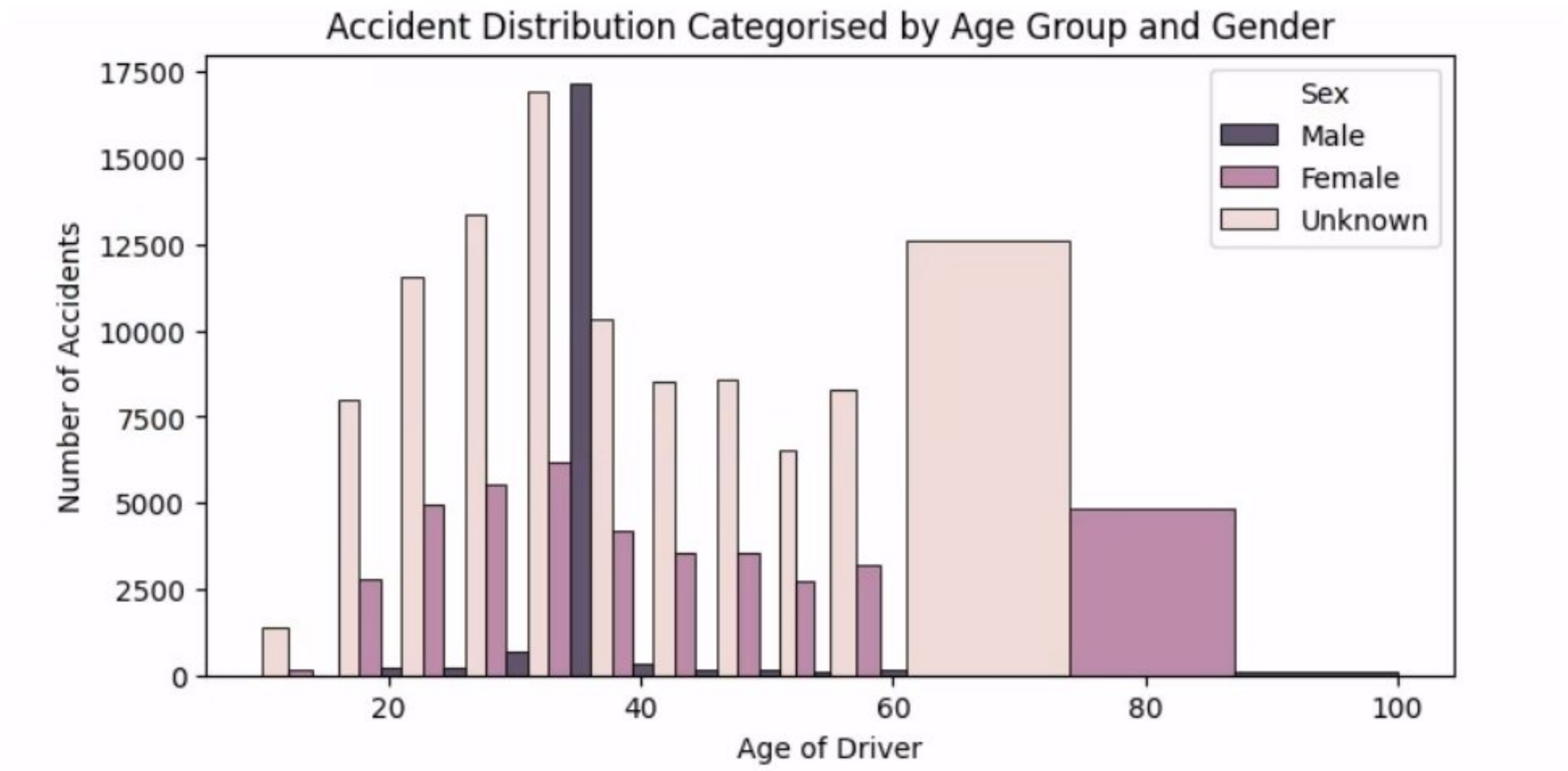

Fig.1. Accident distribution

The distribution of accidents by age group and gender is depicted in the above figure. The statistics provided indicate that there are more male accident casualties than female casualties in the 26–35 age bracket, and most accident casualties are in the 25–55 age range. This suggested that more men are driving than women.

Frequency of Accident Severity

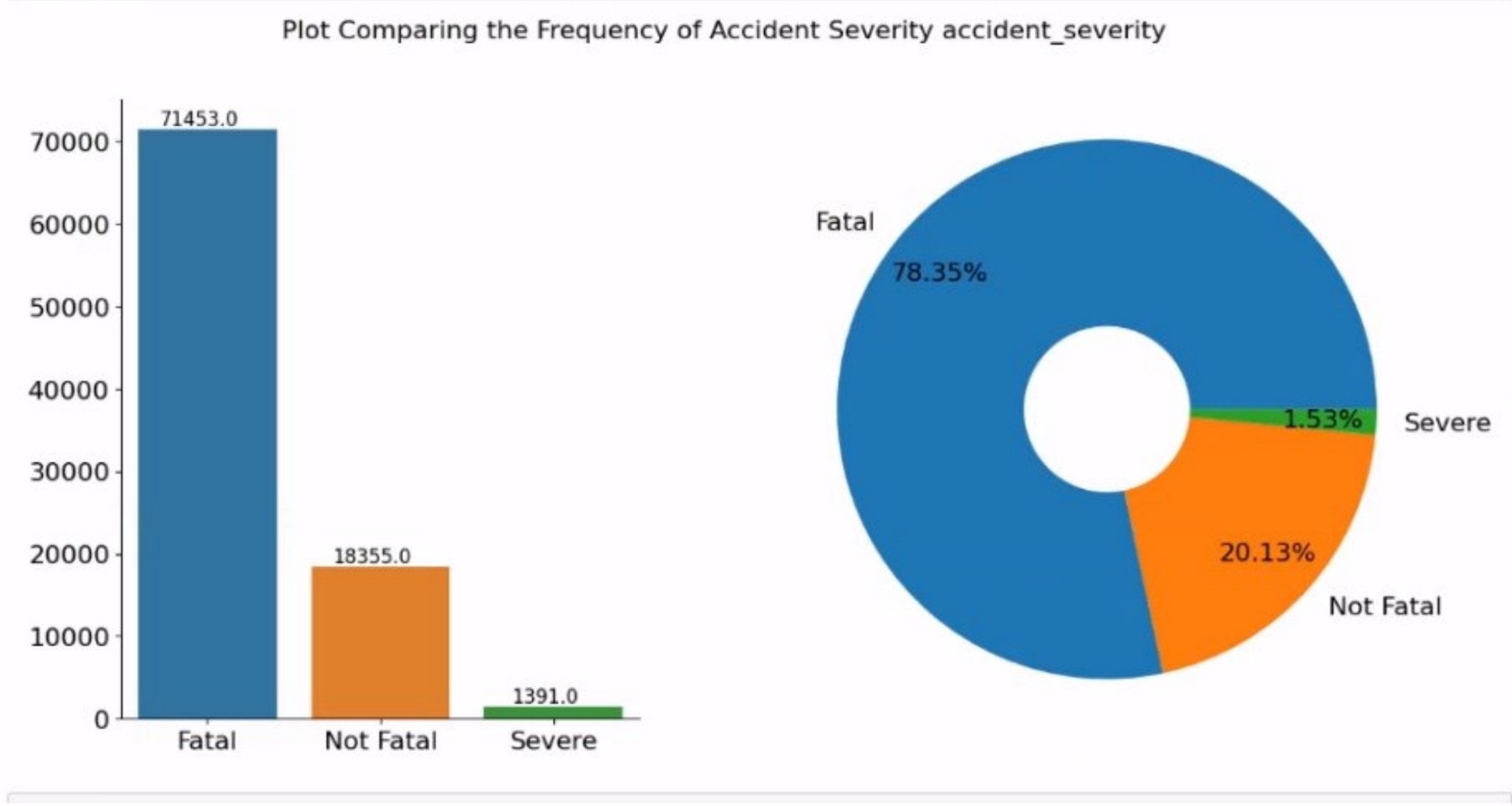

Fig. 2. Frequency of accident severity

CamScanner

Generally, fig 2 shows and compares the frequency of accident severity accident. More than 78% of accidents were recorded as fatal, 20.13% were recorded as not fatal, and less than 2% of the entire accident population was recorded as severe.

DATA INSIGHT

1. The significant hours of the day, and days of the week, when accidents occur.

The fig.3. bar chart below shows that most accidents happen between the hours of 3 and 4 pm (8.1% of all accidents occur between these times), with the highest percentage occurring around 5 pm (8.6%). Many accidents happen in the morning, around 8 am. These findings are consistent with real life, where the morning rush hour occurs around 8 a.m. when people leave for work and school, and the evening peak hour is between 3 and 5 p.m. when most people get home.

The bar chart in Figure 4 illustrates that most accidents happen during the week, with Friday having the highest percentage of accidents (16.3%). The results are consistent with reality since most people travel on Fridays to make the most of their weekends, which means there are more cars on the road on that day.

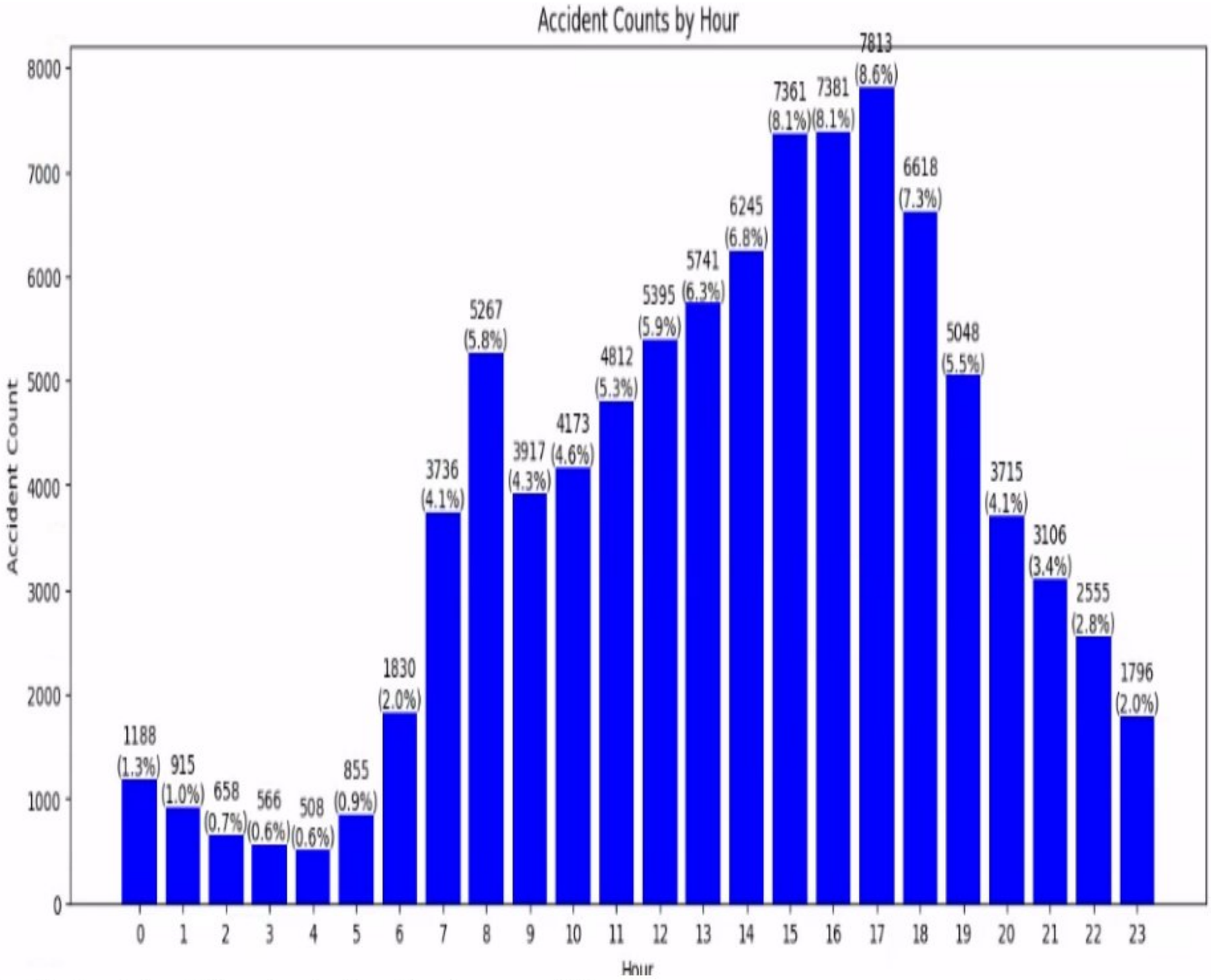

Fig. 3. Accident Counts during the hours of the day

CamScanner

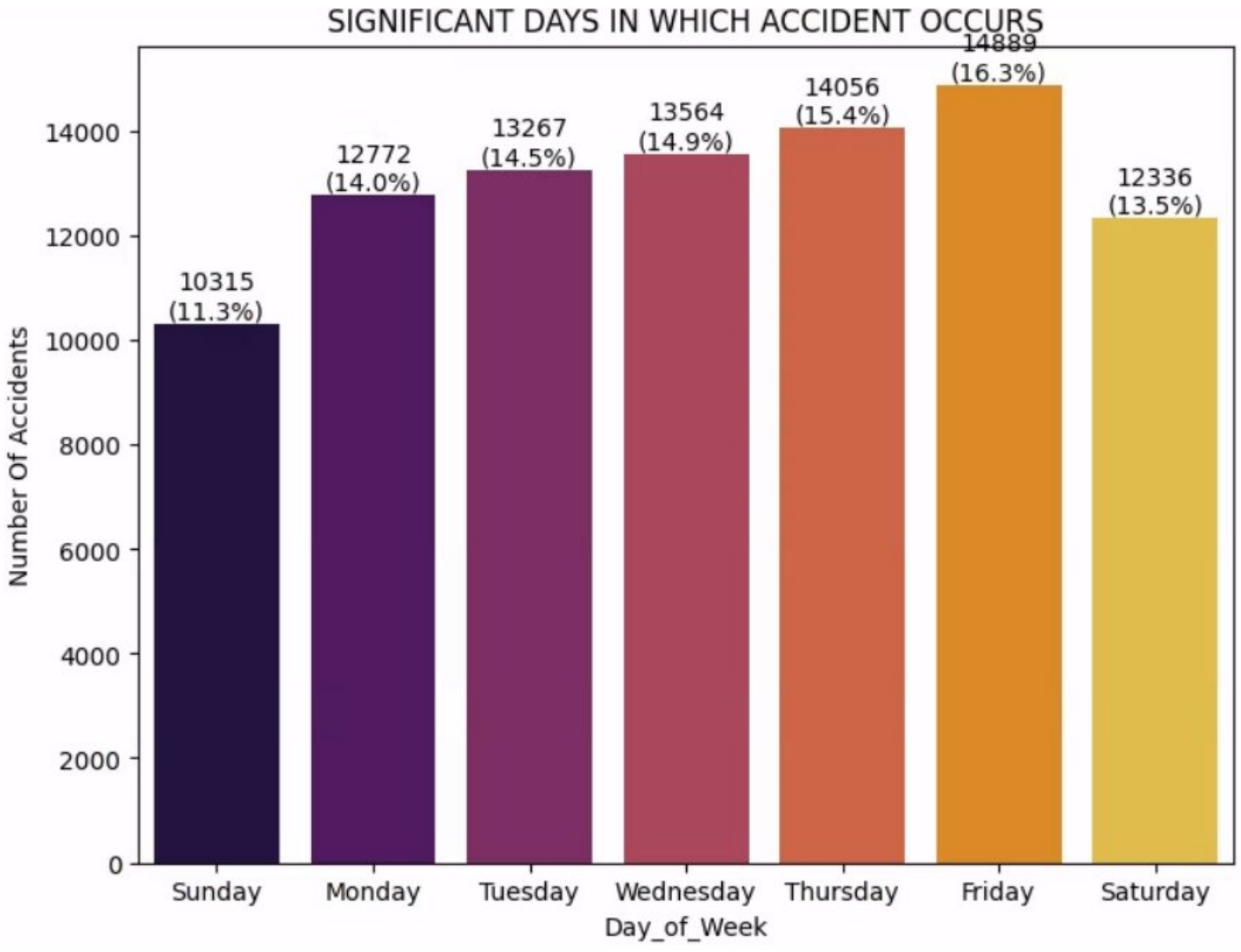

Fig. 4. Significant days in which accident occurs.

2. Significant hours of the day, and days of the week, when motorbikes accidents occur.

Motorcycle accidents also occur between 3 and 4 p.m. and at peak hour, which is 5 p.m., according to figures 5., 7., and 9. However, figure 5 shows that the accident peak hour is 6 p.m. These motorcycles fall into three categories: 125cc and under, 125cc and up to 500cc, and 500cc and above.

Figure 10 illustrates that Friday was the day with the highest accident frequency across all motorcycle categories, except for motorcycles over 500cc, which had their peak on Sunday.

CamScanner

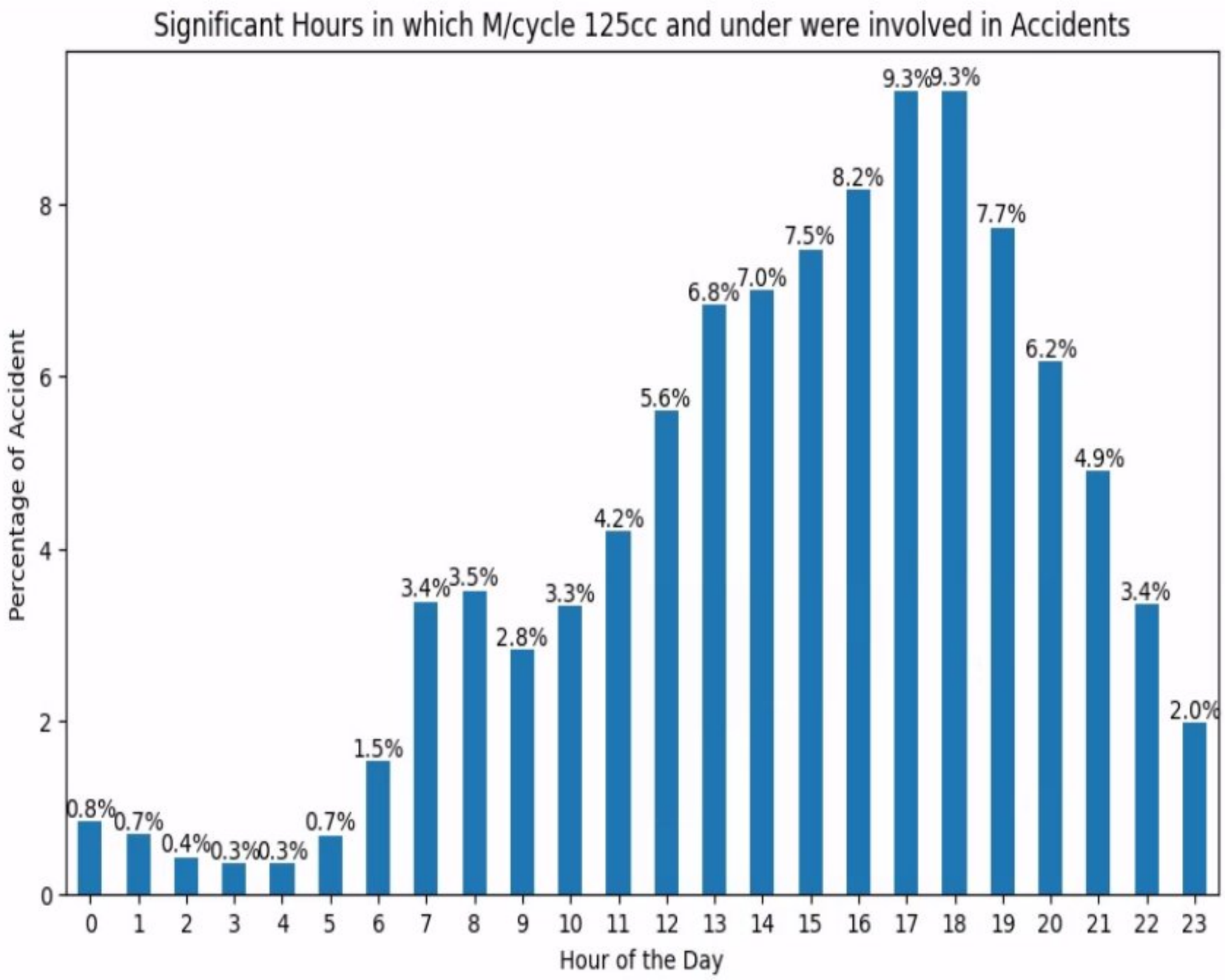

Fig. 5. Significant Hours in which Motorcycle 125cc and under were involved in accident.

CamScanner

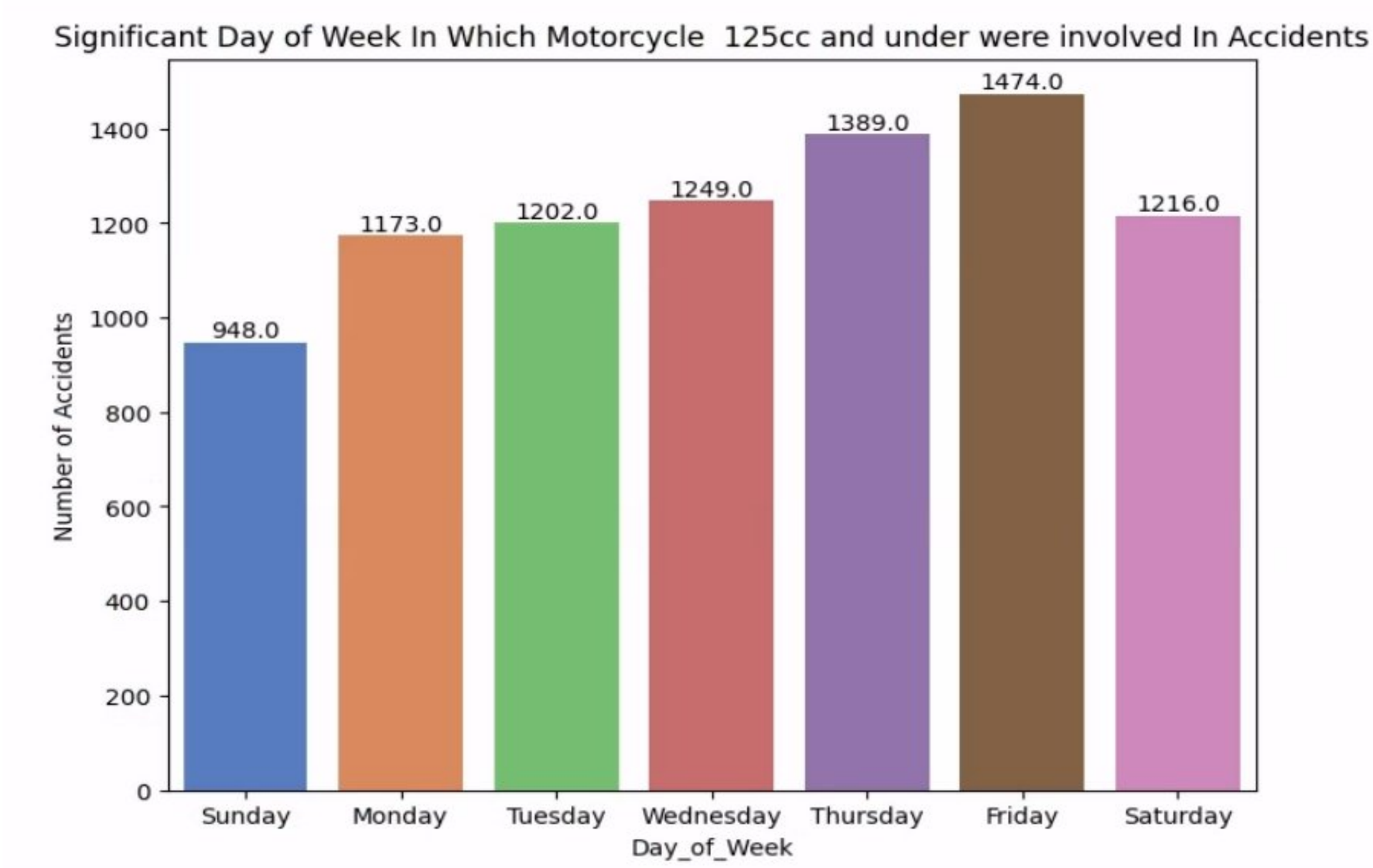

Fig. 6. A significant day of the week in when motorcycles 125cc and under are involved in accident.

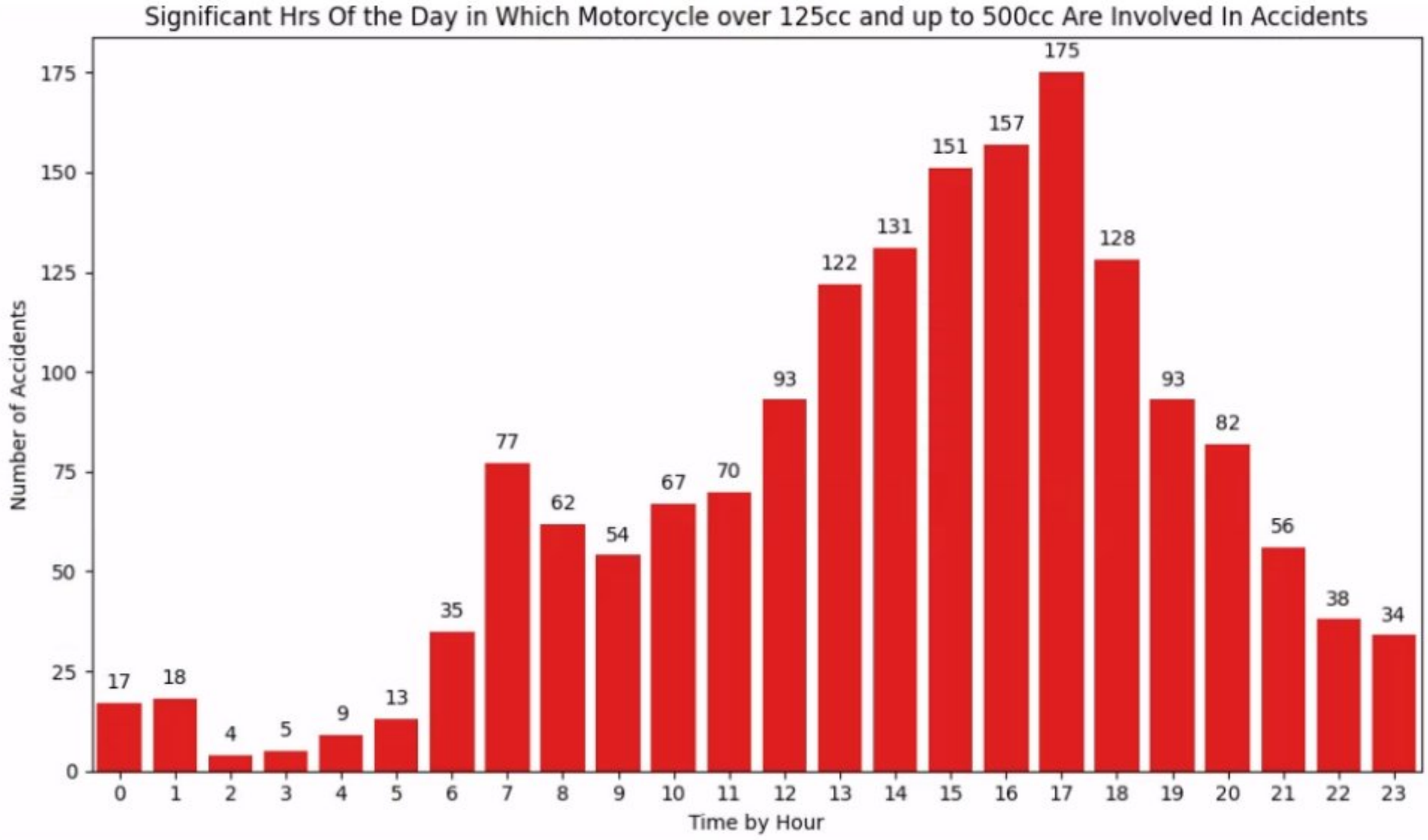

Fig. 7. Significant Hours in which Motorcycle over 125cc and up to 500cc are involved in accident.

CamScanner

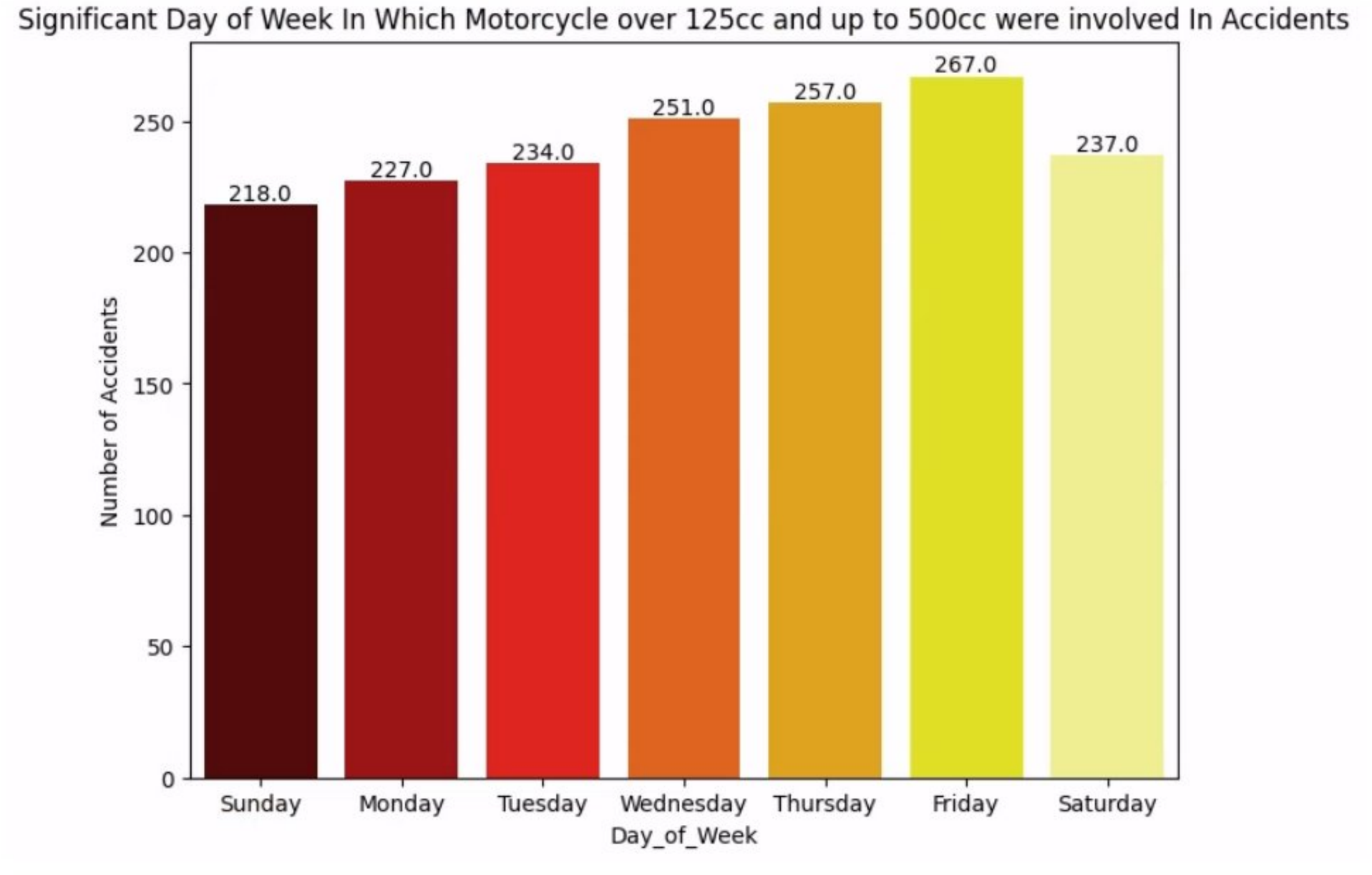

Fig. 8. A significant day of the week in which motorcycles over 125cc and up to 500cc are involved in accident

CamScanner

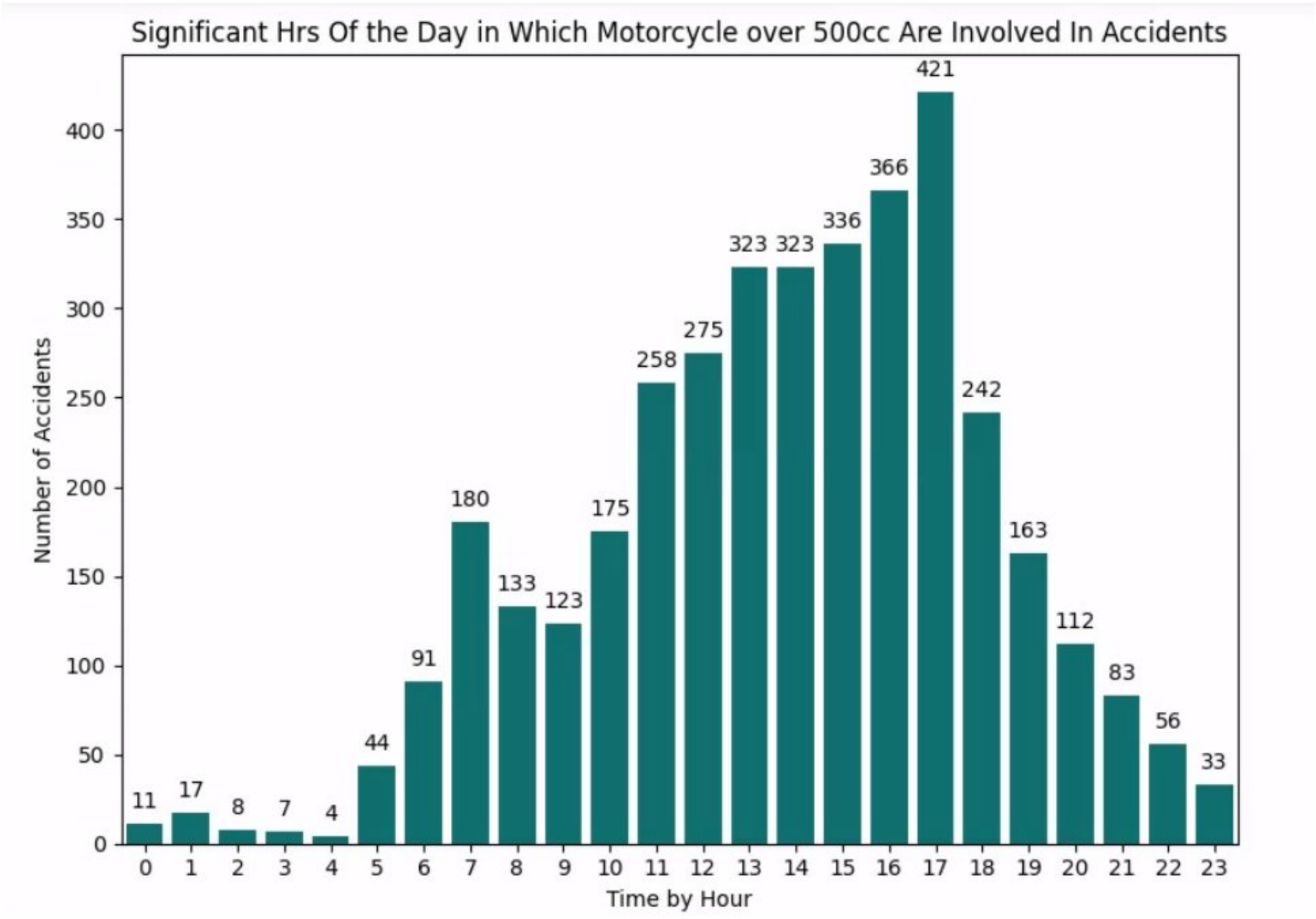

Fig. 9. Significant Hours in which Motorcycle over 125cc and up to 500cc are involved in accident.

CamScanner

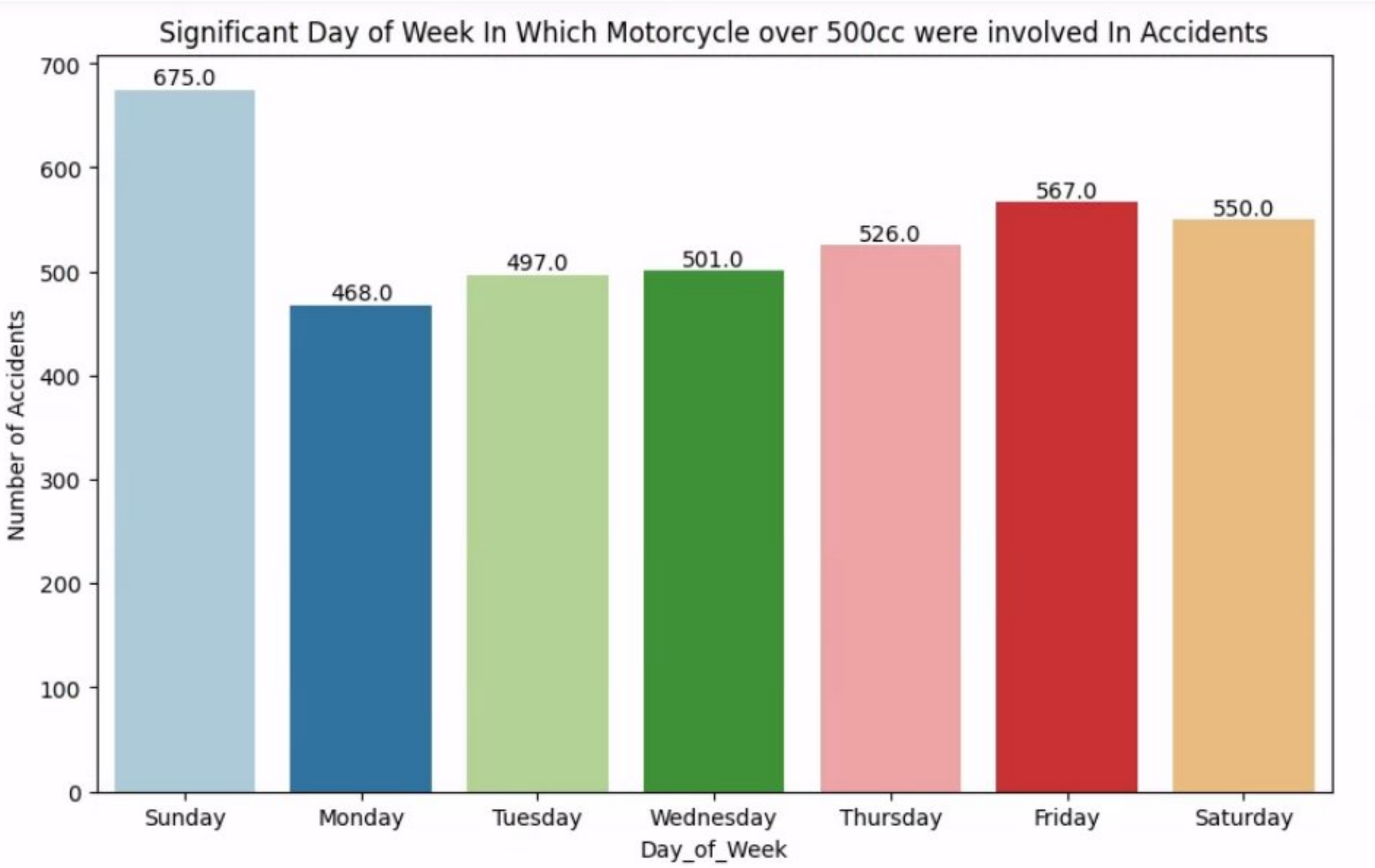

Fig. 10. A significant day of the week in which motorcycles over 125cc and up to 500cc are involved in accidents.

3. The significant hours of the day, and days of the week, when pedestrian accidents occur.

Accident starts to occur at the early hour of 8 am which is a rush hour when people either go to work or school. The peak hour when a pedestrian accident occurs is 3 pm this is the period when people start returning from work or school back to their houses. Generally, weekdays record most pedestrian accident occurrences, but Friday significantly has the highest pedestrian accidents with Sunday the lowest pedestrian accident occurrence as shown in (Fig.12)

CamScanner

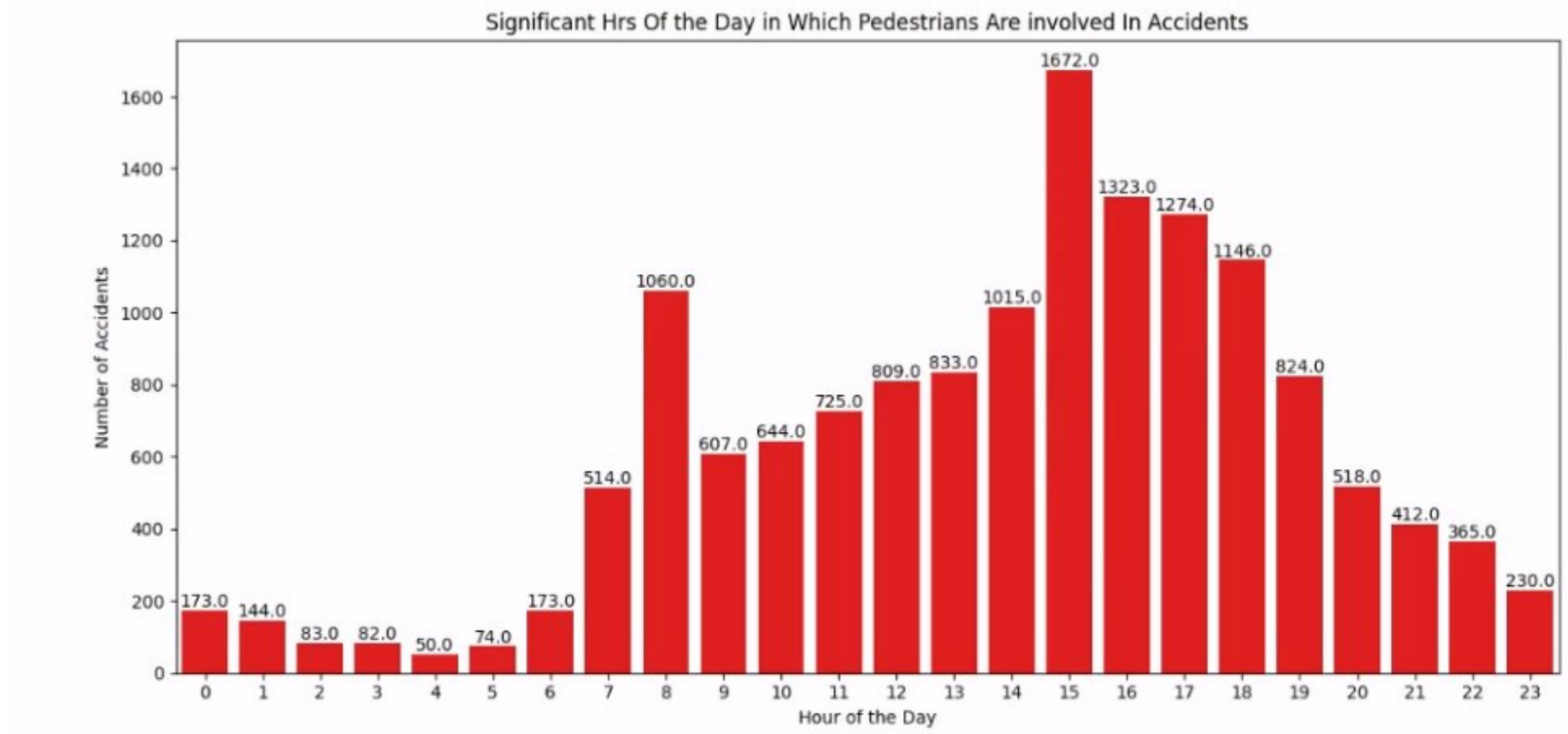

Fig.11. Significant hours of the day when a pedestrian accident occurs.

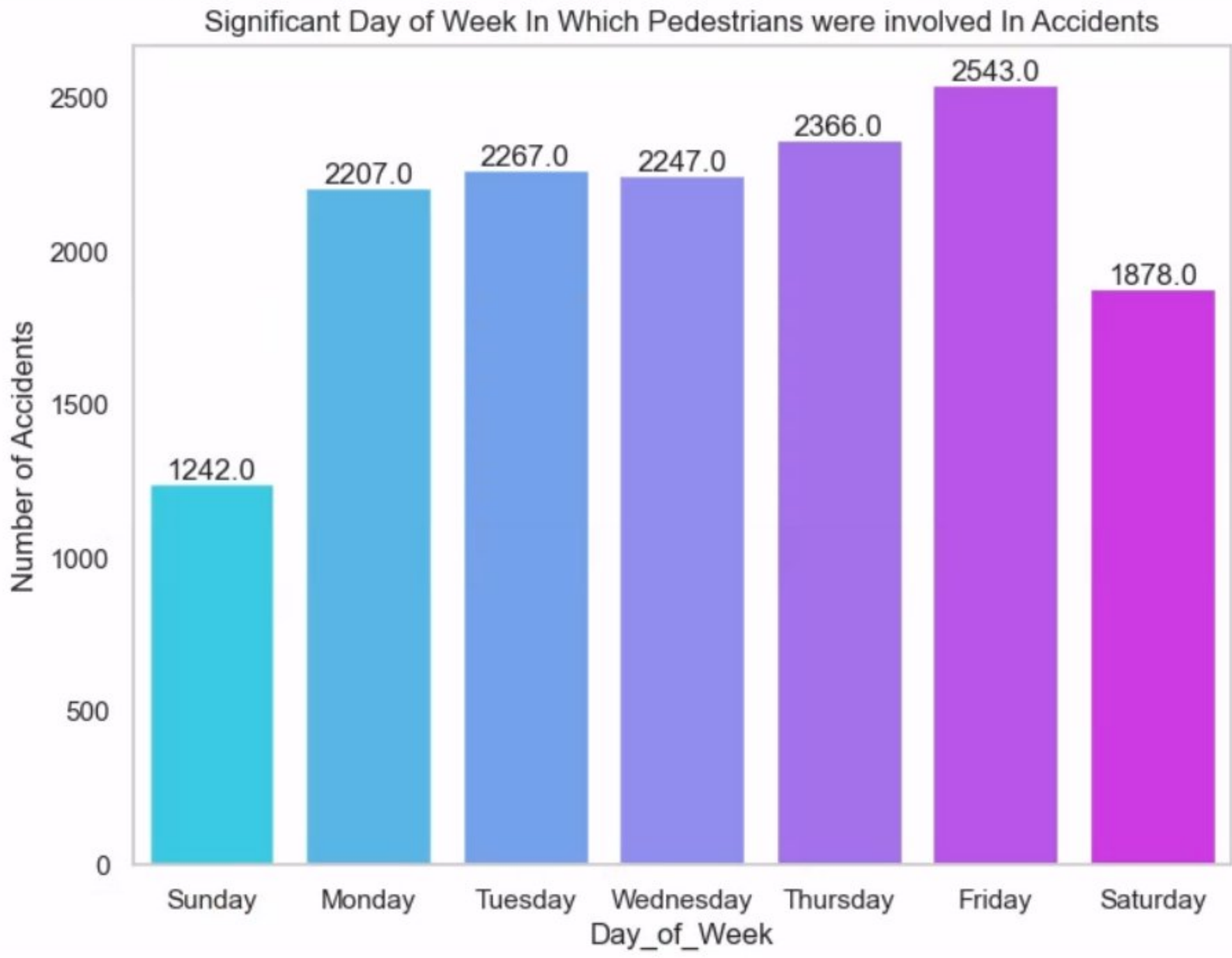

Fig.12. Significant days of the week when a pedestrian accident occurs.

4. Causes of Accidents

Since people are usually indoors during snow, ice, and floods, fewer vehicles are moving on the road, the majority of accidents caused by road surface conditions are typically caused by dry road surfaces. This suggests that dry road surfaces significantly increase the risk of accidents while snow, ice, and floods do not.

CamScanner

The relationship between weather and accidents reveals that low wind speeds are the primary cause of most weather-related accidents, whereas lower traffic volumes throughout the winter and spring reduce the number of accidents.

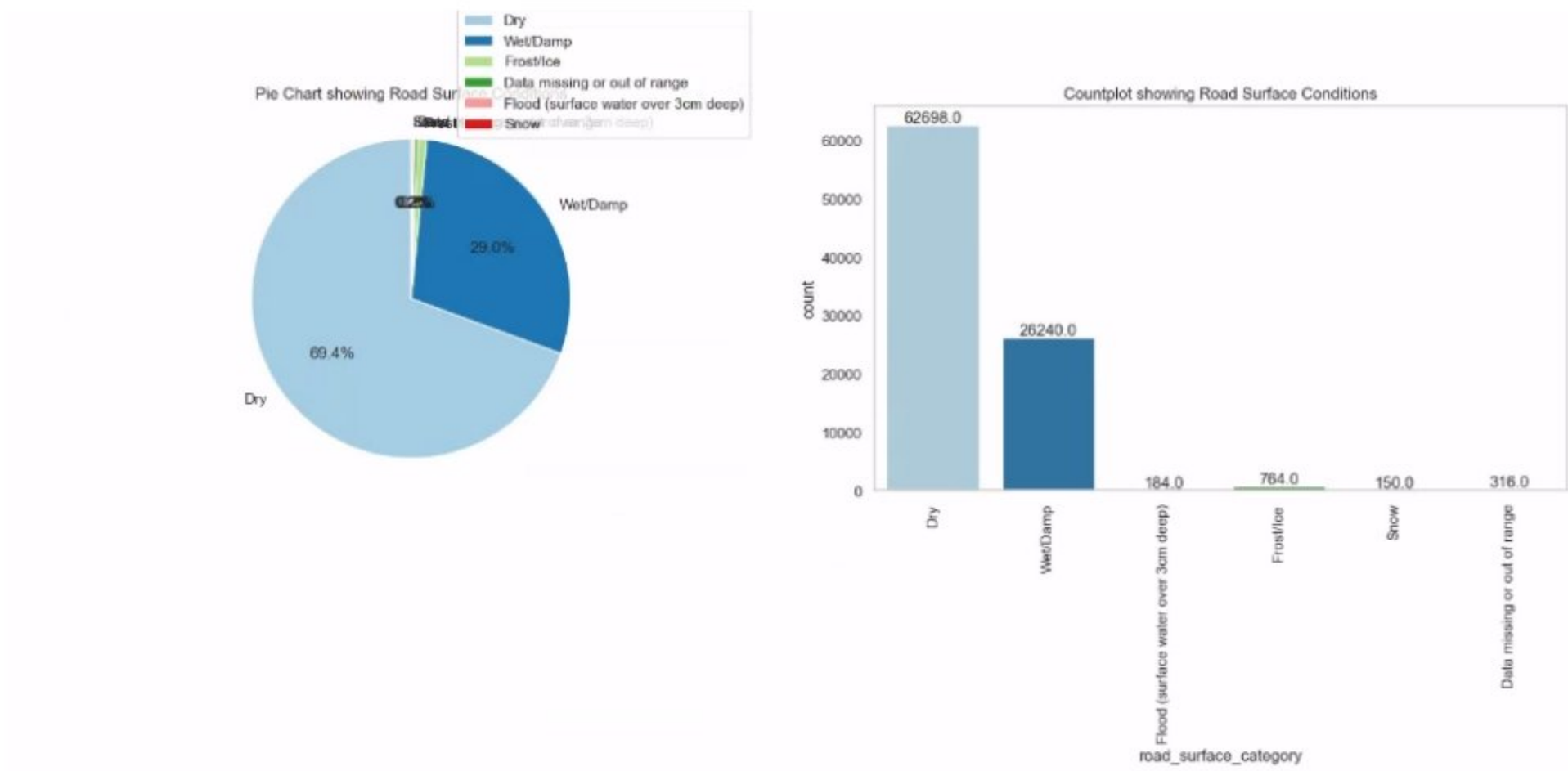

Fig.13. Road Surface Conditions on Accidents

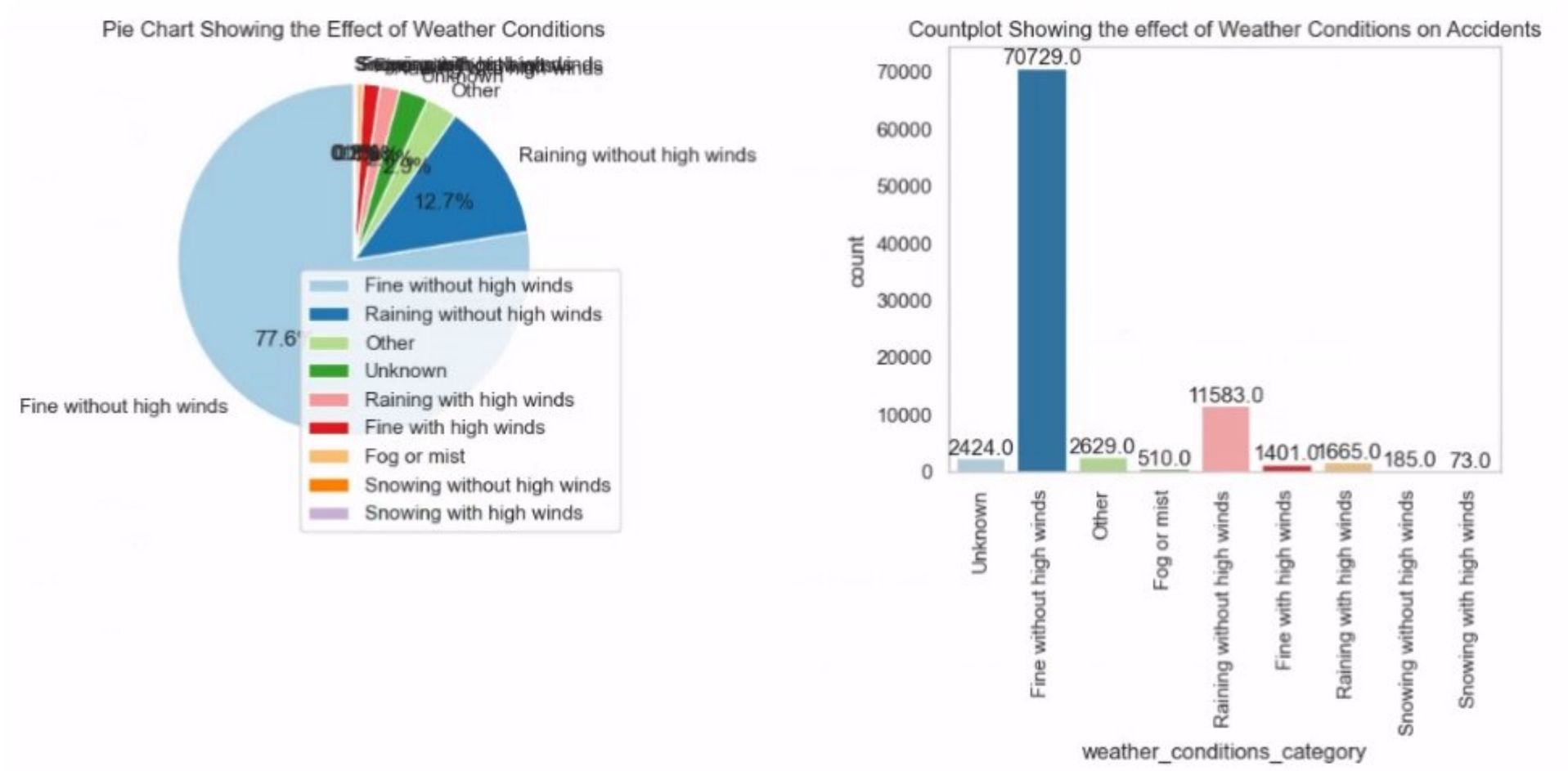

Fig.14. Effect of Weather Conditions on Accidents

Using the Apriori Algorithm, to explore the impact of selected variables on accident Severity

CamScanner

The Association Rule are used to discover relationship between variables in transaction databases. (Agrawa et al., 1993) . The support metric helps to indicate the frequency with which the antecedents and consequent of the association rules appear simultaneously. The higher the support level, the stronger the association between the selected variables while the Confidence measure how often the rule was found to be true. The higher the confidence, the likelihood for the antecedents to occur when the consequents occur. Lift tells the likelihood of one variable occurring when the other variable is present. In this context, lift will be used as a measure of association rule. According to Hussein et al. (2015) lift as a measure is more effective in discovering association rules and it also gives the user more choices in determining the type of association rules to be discovered. As a result, it helps to reduce the search space and consequently enhances performance.

Rule	Condition	Support	Confidence	Lift
1	(urban_or_rural_area_1)→(severity_Fatal)	0.550313	0.812932	1.037586
2	(urban_or_rural_area_1, road_surface_condition) →(severity_Fatal)	0.263523	0.80464	1.027002
3	(speed_limit_30) → (severity_Fatal	0.460082	0.802705	1.024532
4	(light_conditions_1) →(severity_Fatal)	0.559337	0.790994	1.010082
5	(junction_detail_3) → (severity_Fatal)	0.222261	0.790994	1.009584

Table.1. Five Association Using Lift as metric form Apriori Algorithm.

Note: severity _fatal represent moderate

Rule 1 suggests that accidents in a specific urban or rural area are slightly more likely to be moderate with the lift value of (1,03), However, support is also high (0,55) suggesting accident in this area are frequent. Rule 2 support (0.263525) value is very low compared to rule 1 while confidence (0.80464) indicates that out of the accidents that meet both conditions urban or rural area and road surface conditions 80% are severity fatal(moderate). Lift (1.027002) means accident in urban or rural areas and road surface location might be more likely to be moderate than accident elsewhere. Rule 3 implies that speed limit 30 correlate with severity fatal which is moderate accident though the confidence value is (0.802705) the lifts indicates that this is not necessary because of the 30-mph speed limit it could be related to other factors. Rule 4. lift value of (1.010082) suggests that the light conditions might not be a strong determinant of accident severity while the confidence (0.790994) value indicates a notable portion of accident under light conditions are severity fatal (moderate).

5. Clustering

To evaluate the distribution of accidents, longitude and latitude were clustered using the Kmeans and DBSCAN clusters. The right cluster number for a KMeans cluster was found using the Elbow technique, and it was found to be 5, which is the point at which inertia

CamScanner

begins to decrease. According to the map, there were many accidents near key cities in the Humber region, including Hessle, Hull, Bridlington, and Scunthorpe.

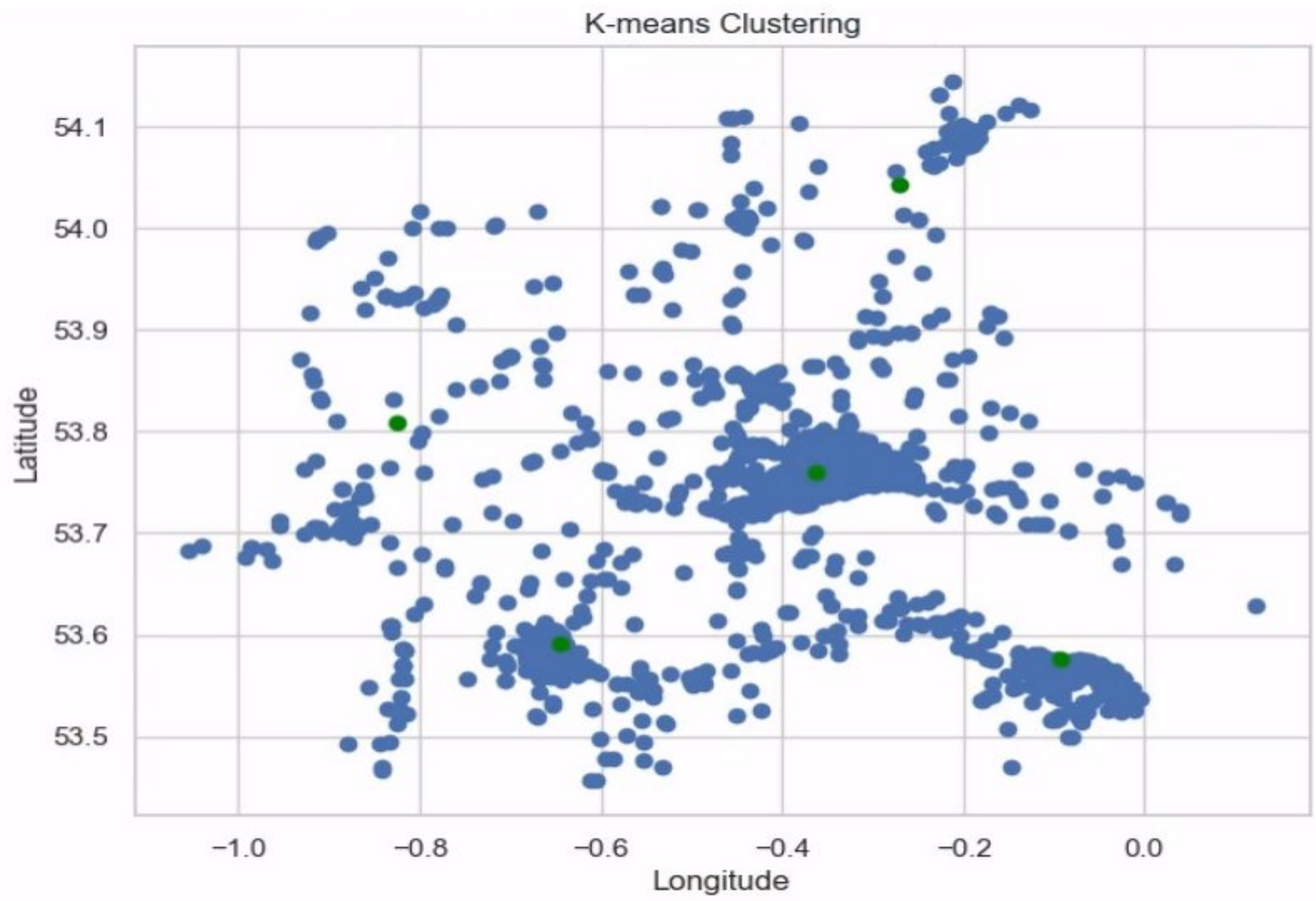

Fig.15. Kmeans Clustering

CamScanner

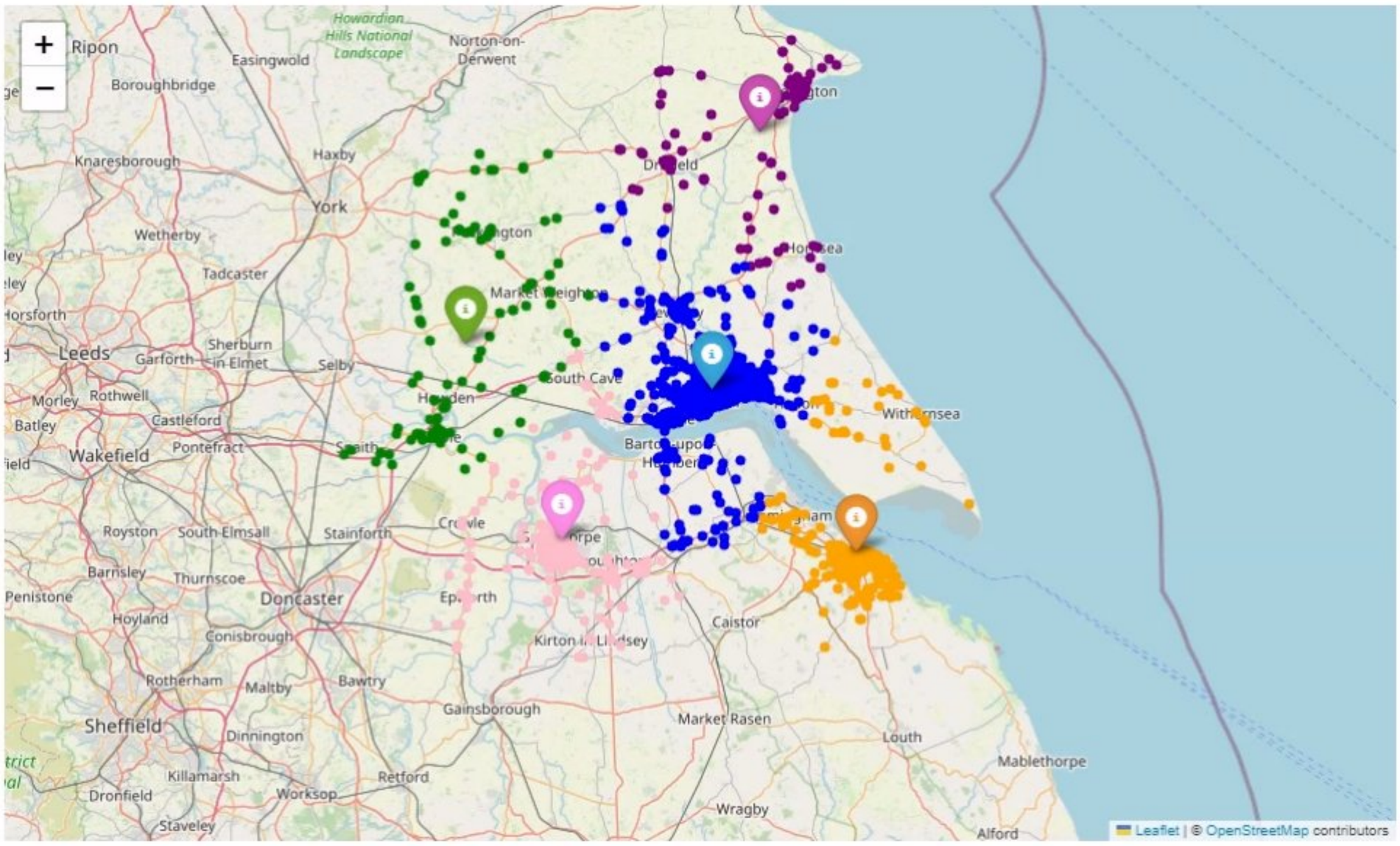

Fig.16. Kmeans Clustering of accident distribution in Humberside

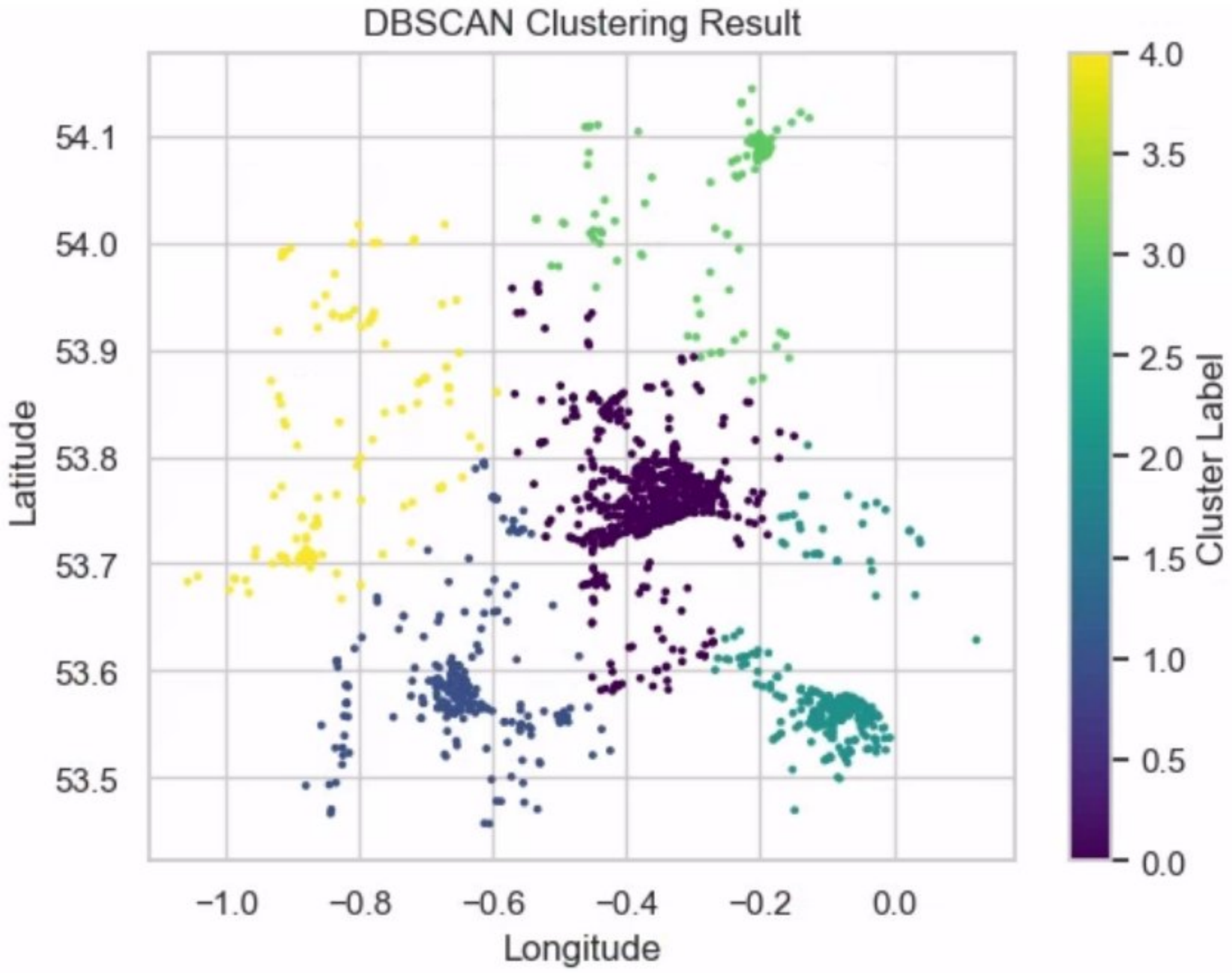

Fig.17 DBSCAN Clustering

CamScanner

6. Outliers Detection

Selected accident data features were subjected to outlier detection utilizing the IQR, Isolation Forest, and Local Outlier Factor approaches. Outliers were dispersed throughout the sample, with a small concentration close to the London region, according to geospatial analysis Fig 18 and 19. These outliers' localized grouping raises the possibility that conditions or variables played a role in these unusual occurrences. The visualization helps to understand the spatial distribution of actual data and the detected outliers. The geographic coordinates (longitude and latitude) of the data point are displayed in Fig. 18. The primary data points are represented by white dots with black edges, while the outliers that the LOF Algorithms detected are represented by teal circles. I observed mild outliers.

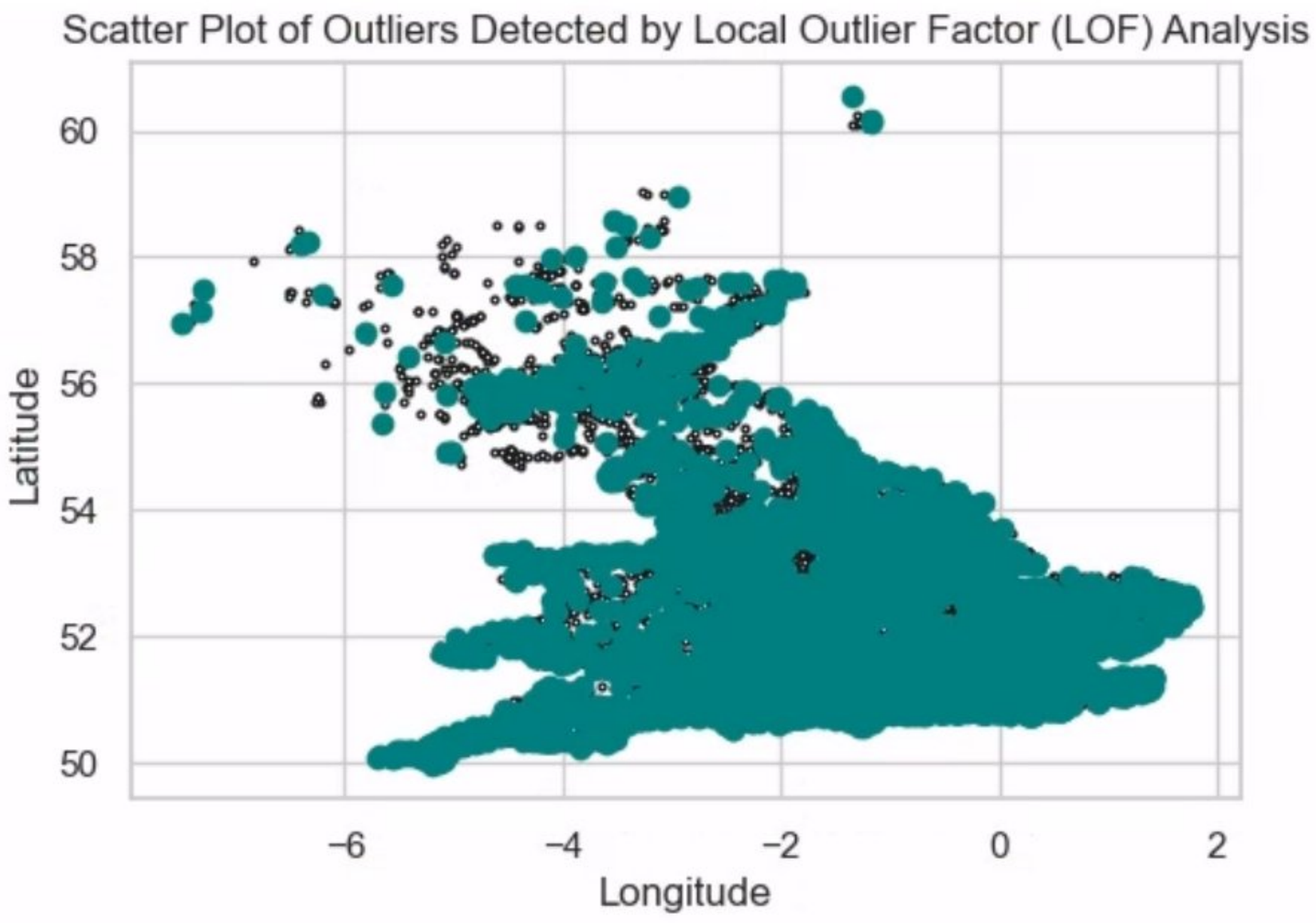

Fig.18. Outliers Detected by Local Outlier Factor

CamScanner

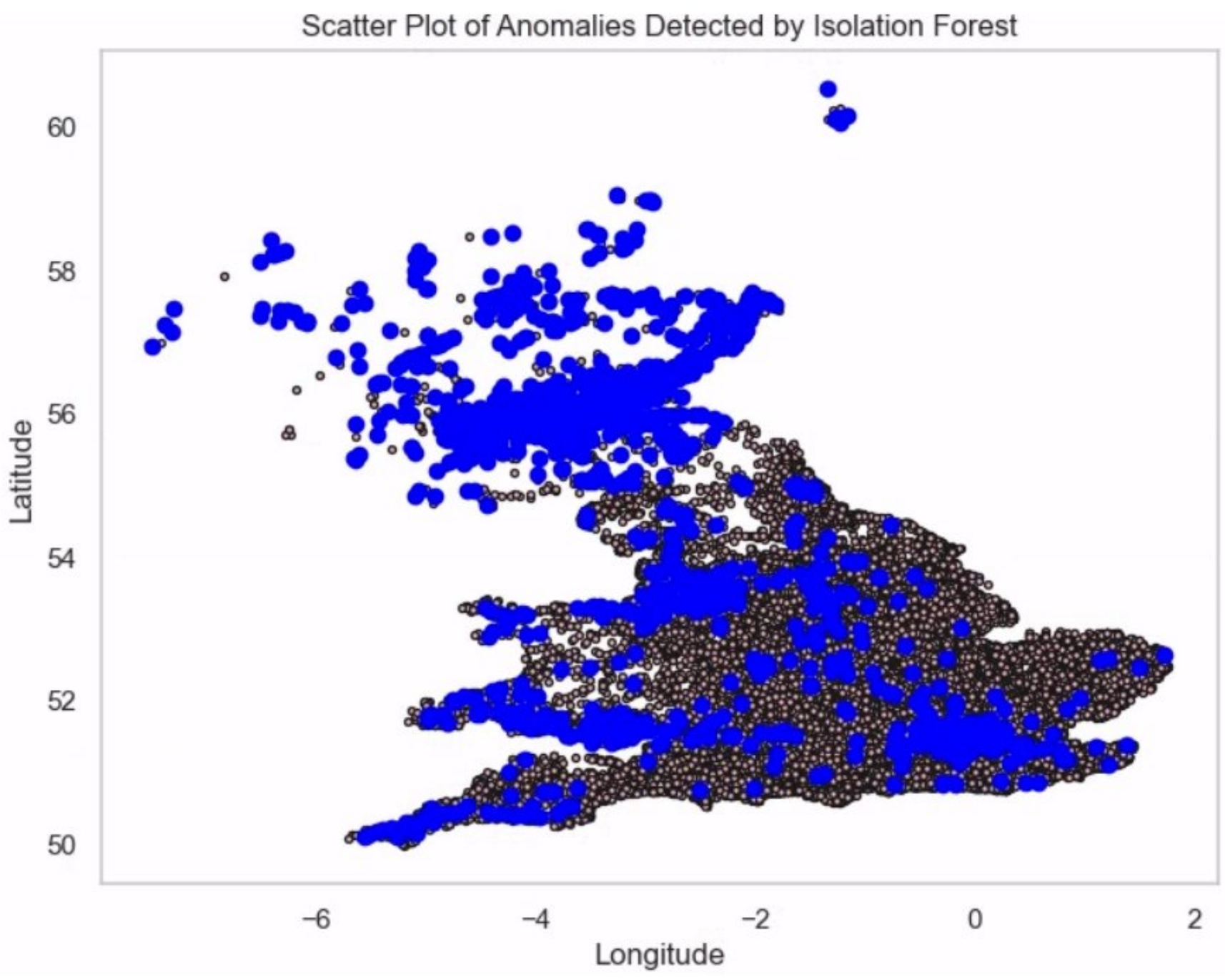

Fig.19. Anomalies Detected by Isolation Forest

7. Prediction of fatal injuries sustained in road accidents.

The important features were selected using Random Forests as depicted in Fig. 20. The visualization of the feature importance ranking derived from the Random Forests was done. The best hyperparameters for building the classification model were found via a thorough Grid search, which was conducted once the top 10 features were selected (criterion='gini', max_depth=20 min_samples_split=2, n_estimators=300). The Random Forest model had an accuracy of 61%. Overall, the model's accuracy, precision 60%, and F1 score, at 63%, showed that the model needs an improvement like selecting features that predicts fatal injuries sustained in road accident.

CamScanner

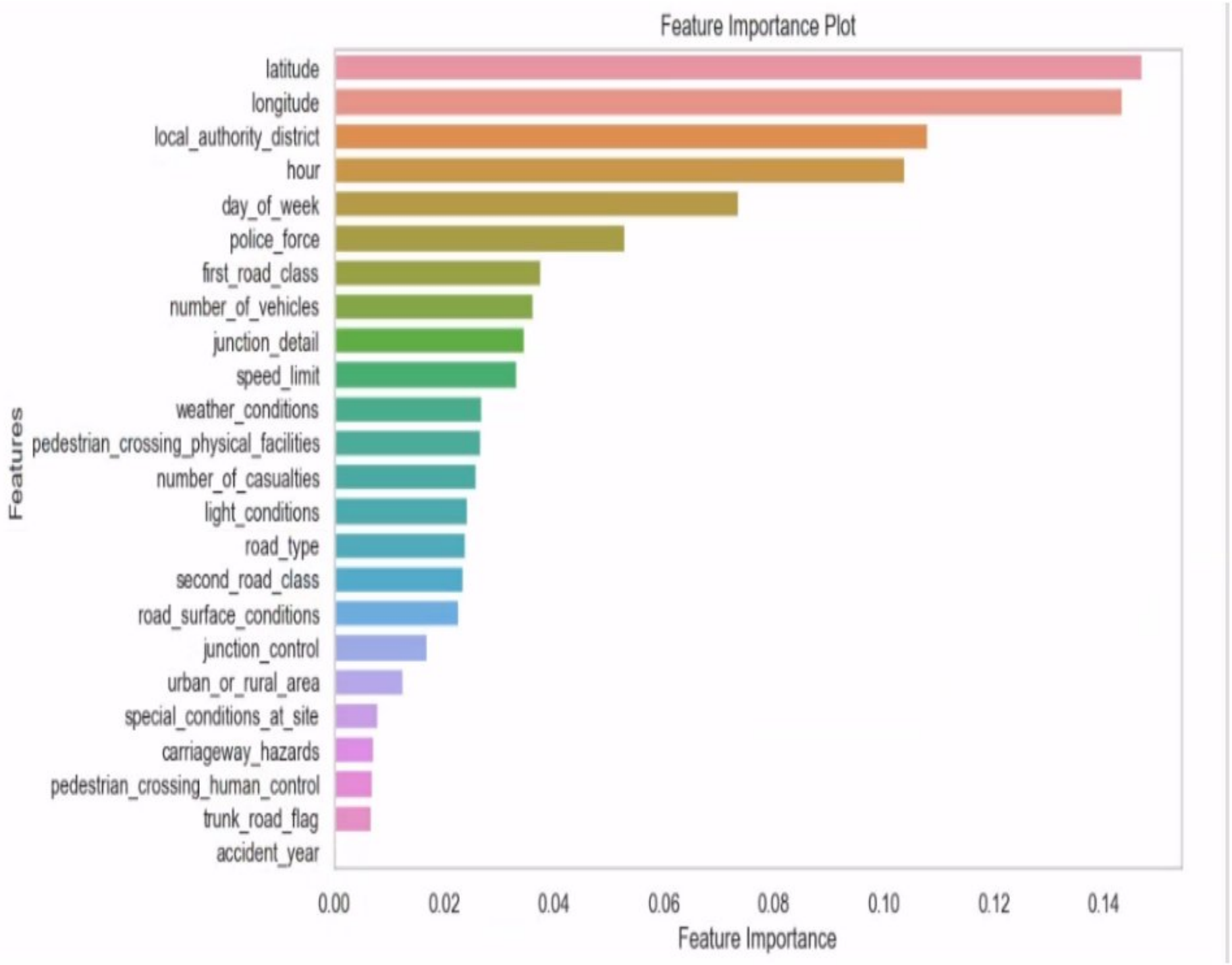

Fig.20. Feature Importance

CamScanner

Classification Report:

	precision	recall	f1-score	support
False	0.63	0.58	0.61	3988
True	0.60	0.65	0.63	3911
accuracy			0.62	7899
macro avg	0.62	0.62	0.62	7899
weighted avg	0.62	0.62	0.62	7899

Fig.20. Confusion Matrix and Classification Report of Random Forest Classifier with Hyperparameter

From the above fig. 20. The classification reports shows that the model demonstrates comparable performance for both classes, exhibiting precision, recall, and F1 scores in the range of 0.60-0.63 There is significant number of misclassifications of false positives and false negatives in the confusion matrix, and this suggests that the model needs further improvement.

Recommendations

- Create Public Traffic Awareness and Strengthen Law Enforcement Agencies: More efforts should be targeted towards creating awareness on importance of safe driving especially on Fridays during peak hours. Police presence should be increased during peak hours to serve as deterrent and to enhance enforcement.
- Infrastructure Improvement: This includes constructing and maintain appropriate sidewalks, improving the design and visibility of crosswalks, adjusting traffic lights to coincide with pedestrian crossing and reducing speed limits in places with significant pedestrian areas to reduce pedestrian accident.

CamScanner

- Motorcycle Lane markers: Clear and well lane markings for motorcycles should be implemented to enhance visibility and reduce the risks of cars entering motorcycles lane.
- Dualization of carriage ways in urban areas: single lane should be dualized to accommodate urban road congestion to reduce accident in urban areas.

REFERENCE

Agrawal, R., Imieliński, T. and Swami, A., 1993, June. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207-216).

Hussein, N., Alashqur, A. and Sowan, B., 2015. Using the interestingness measure lift to generate association rules. *Journal of Advanced Computer Science & Technology*, *4*(1), p.156.

CamScanner