ROAD ACCIDENT DATA ANALYSIS AND REPORT

INTRODUCTION

The accident dataset provides comprehensive details on road traffic incidents in various regions, including Kingston upon Hull, East Riding of Yorkshire, and others. One of the critical analyses being performed involves clustering accidents to identify patterns and high-risk zones. The insight derived from this analysis aims to enable targeted safety improvements, informing road safety measures, and ultimately reducing the incidence of fatal and non-fatal accidents in the region The importance of road safety is universally recognized and affects all aspects of society (Road Safety Management, 2018). Stats20 provides detailed information about this dataset (Department for Transport, 2011).

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
accident_index	91199	91199	2020010219808	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
accident_year	91199.0	NaN	NaN	NaN	2020.0	0.0	2020.0	2020.0	2020.0	2020.0	2020.0
accident_reference	91199	91199	010219808	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
location_easting_osgr	91185.0	NaN	NaN	NaN	456487.876416	93512.711807	65947.0	392890.0	465545.0	530168.0	655138.0
location_northing_osgr	91185.0	NaN	NaN	NaN	273764.496233	147351.556104	12715.0	174569.0	208599.0	378366.0	1184351.0
longitude	91185.0	NaN	NaN	NaN	-1.189258	1.367786	-7.497375	-2.107789	-1.046912	-0.125238	1.756257
latitude	91185.0	NaN	NaN	NaN	52.351073	1.327573	49.970479	51.457237	51.763385	53.297386	60.541144
police_force	91199.0	NaN	NaN	NaN	27.488043	24.548964	1.0	4.0	22.0	45.0	99.0
accident_severity	91199.0	NaN	NaN	NaN	2.768232	0.456682	1.0	3.0	3.0	3.0	3.0
number_of_vehicles	91199.0	NaN	NaN	NaN	1.835272	0.677272	1.0	1.0	2.0	2.0	13.0
number_of_casualties	91199.0	NaN	NaN	NaN	1.267382	0.681473	1.0	1.0	1.0	1.0	41.0
date	91199	366	06/02/2020	426	NaN	NaN	NaN	NaN	NaN	NaN	NaN
day_of_week	91199.0	NaN	NaN	NaN	4.121558	1.9322	1.0	2.0	4.0	6.0	7.0
time	91199	1438	17:00	862	NaN	NaN	NaN	NaN	NaN	NaN	NaN
local_authority_district	91199.0	NaN	NaN	NaN	311.482812	253.456329	-1.0	63.0	300.0	502.0	941.0
local_authority_ons_district	91199	378	E08000025	1802	NaN	NaN	NaN	NaN	NaN	NaN	NaN
local_authority_highway	91199	206	E10000016	2964	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 1: Descriptive Statistics of the dataset before cleaning

DATA CLEANING

The following can be noted from the statistical summary:

- The presence of NaN and missing values in some columns.
- The inconsistencies in all the statistical summaries.

To address the inconsistencies in these columns, I will treat the (-1) as NaN (indicative of missing data) alongside with other missing values. Subsequently, I will employ Multiple Imputation by Chained Equations (MICE) to impute these missing values. MICE, being a multivariate imputation technique, fills in missing data based on patterns observed across

other columns. It operates under the assumption that data is missing at random (MAR). In the context of these columns, this MAR assumption aligns with what is outlined in the Stats19 form (Department for Transport, 2021). After the imputation cycle concludes, the missing values should be seamlessly substituted with predicted values that mirror the inherent patterns and relationships in the dataset (Azur et al., 2011).

Multiple imputation is a potent method for managing missing data involving multiple variables and data types. It surpasses basic imputation techniques like mean and median imputations (Aleryani et al., 2020). The method employs algorithms like K-Nearest Neighbours, Random Forest, and neural networks to predict and replace missing values, resulting in more accurate and comprehensive datasets.

	count	mean	std	min	25%	50%	75%	max
accident_year	91199.0	2020.0	0.0	2020.0	2020.0	2020.0	2020.0	2020.0
location_easting_osgr	91199.0	456481.736576	93508.08032	65947.0	392889.0	465499.0	530167.0	655138.0
location_northing_osgr	91199.0	273763.957445	147346.36241	12715.0	174572.0	208599.0	378365.5	1184351.0
longitude	91199.0	-1.18932	1.36771	-7.497375	-2.107818	-1.047244	-0.125277	1.756257
latitude	91199.0	52.351065	1.32753	49.970479	51.457249	51.763385	53.297298	60.541144
police_force	91199.0	27.488043	24.548964	1.0	4.0	22.0	45.0	99.0
accident_severity	91199.0	2.768232	0.456682	1.0	3.0	3.0	3.0	3.0
number_of_vehicles	91199.0	1.835272	0.677272	1.0	1.0	2.0	2.0	13.0
number_of_casualties	91199.0	1.267382	0.681473	1.0	1.0	1.0	1.0	41.0
day_of_week	91199.0	4.121558	1.9322	1.0	2.0	4.0	6.0	7.0
local_authority_district	91199.0	316.984638	252.78577	1.0	72.0	302.0	510.0	941.0
first_road_class	91199.0	4.22032	1.443475	1.0	3.0	4.0	6.0	6.0
first_road_number	91199.0	790.666071	1580.817743	0.0	0.0	34.0	538.0	9174.0
road_type	91199.0	5.256001	1.684878	1.0	6.0	6.0	6.0	9.0
speed_limit	91199.0	36.273863	13.889985	20.0	30.0	30.0	40.0	70.0
junction_detail	91199.0	3.934999	12.612758	0.0	0.0	2.0	3.0	99.0
junction_control	91199.0	3.686422	1.068655	1.0	4.0	4.0	4.0	9.0
second_road_class	91199.0	3.955417	2.745081	0.0	0.0	3.0	6.0	6.0

Figure 2: Descriptive Statistics of the dataset after cleaning

MICE performed well in imputing missing data and preserving the statistical properties of the dataset. The mean stayed consistent across the variables, signifying that the imputations are in line with the original data distribution. The stability in standard deviations, which measure the spread of data, further confirmed this consistency. However, the junction detail column was an exception due to an outlier (99) that skewed its statistics. After MICE addressed this outlier, the mean and standard deviation for these columns became more closely clustered,

showing that the values are now more centrally distributed. The cleaning process has thus enhanced the reliability of the data. This cleaning will be applied for **questions (1-5).**

ANALYSIS

Significant hours of the day accidents occur.

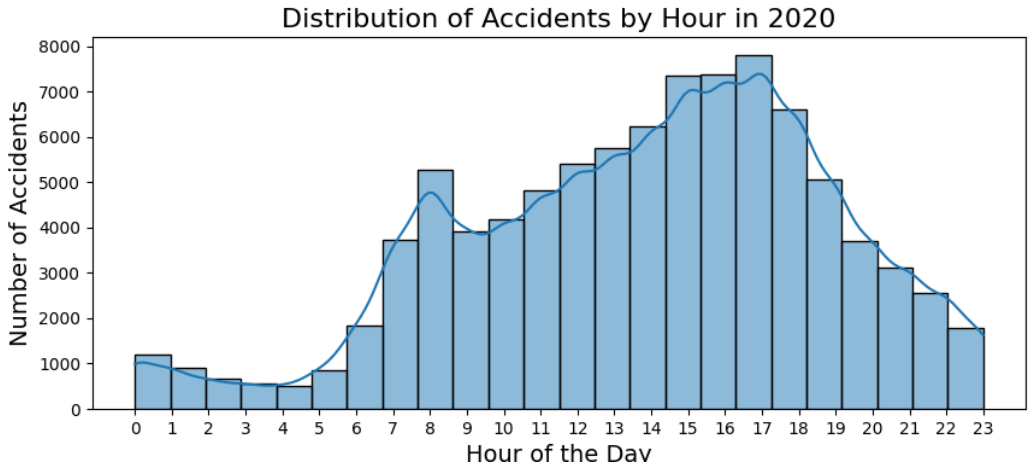

Figure 3: Distribution of Accident in hours

The analysis of accident times indicates a clear pattern, with late afternoon and early evening, particularly around 3:00PM - 5:00 PM, being the peak times for occurrences. The data, consistently shows that the hours coinciding with the end of the workday and potential rush hours are most prone to accidents.

Significant days of the week accidents occur.

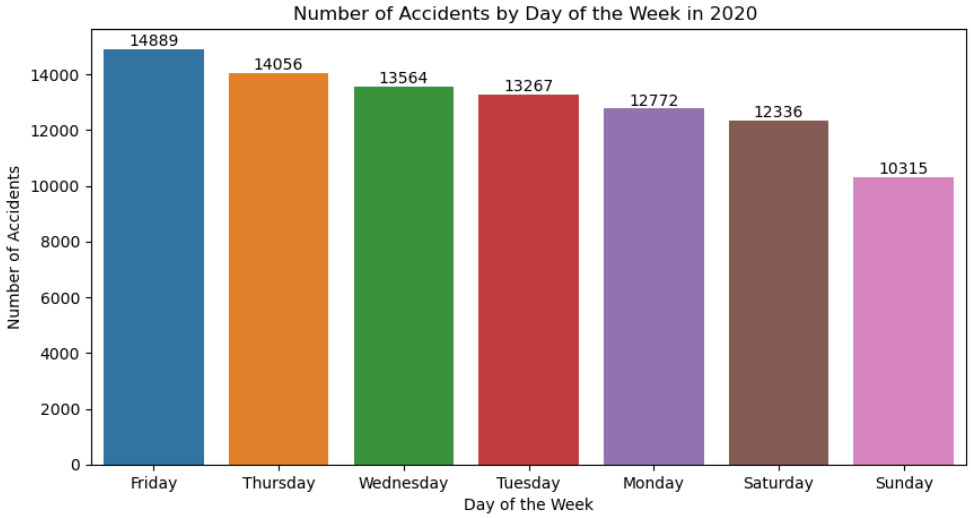

Figure 4: Distribution of Accidents across the days of the week

The frequency of accidents was found to be highest on Friday, amounting to a total of 14,889 incidents. This number gradually decreases throughout the start of the week, reaching 12,772 accidents by Monday. Interestingly, Saturday (the weekend) logged more accidents (12,336) compared to Sunday, which recorded the week's lowest at 10,315 accidents.

This analysis suggests that accident rates are most prevalent towards the tail end of the workweek, with Friday topping the list in terms of incident counts. Conversely, Sunday (the beginning of the week) registered the least number of accidents during the week.

Analysing the conditions under which these accidents occurred:

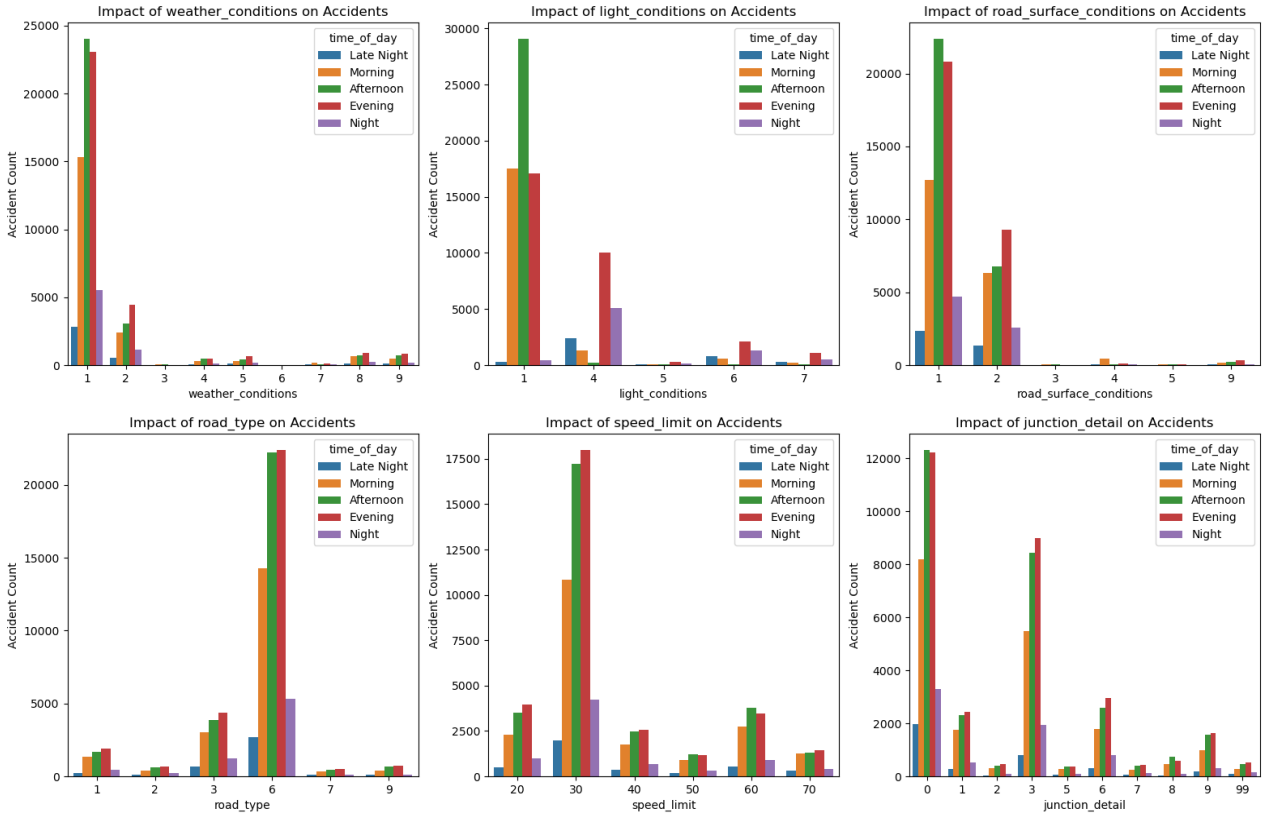

Figure 5: Conditions of the hourly accidents

These (-1) in these columns except road type were cleaned as detailed at the beginning of this report.

The analysis examines the effects of various conditions on hourly accidents, including weather, lighting, road surface and type, speed limits, and junction details. Key findings suggest that fine weather, daylight, and dry surfaces see the most accidents, challenging common assumptions. Accidents are most prevalent on single carriageways with a 30-speed limit, and straight roads away from junctions, pointing to factors like human error, road design, and speed rather than environmental conditions as primary influences. The complexity of T or staggered junctions also contributes to a higher number of accidents.

TASK 2:

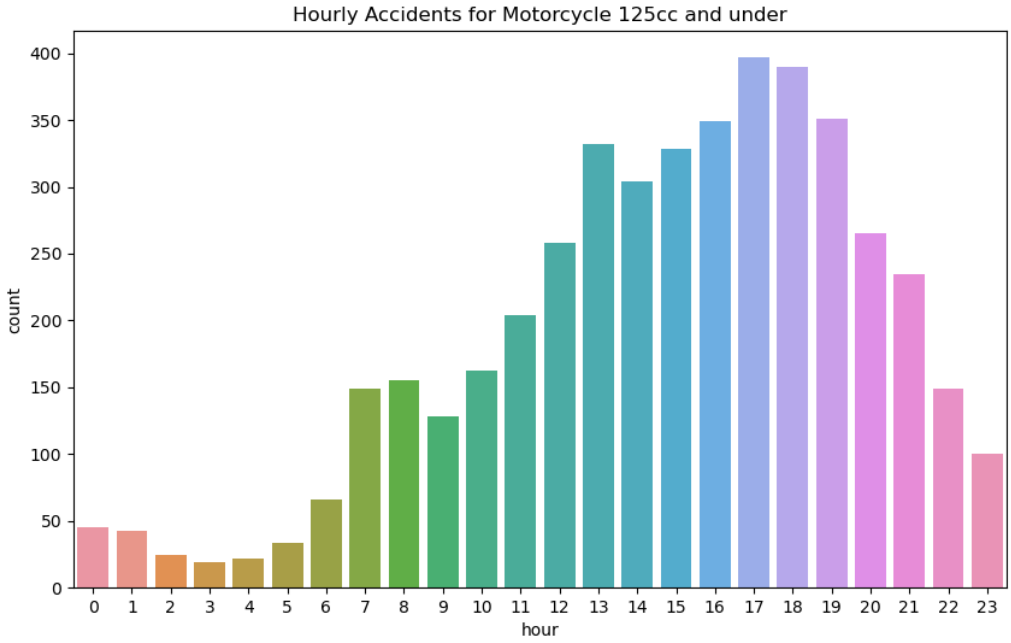

Figure 6: Motorcycle 125cc and under hourly accidents

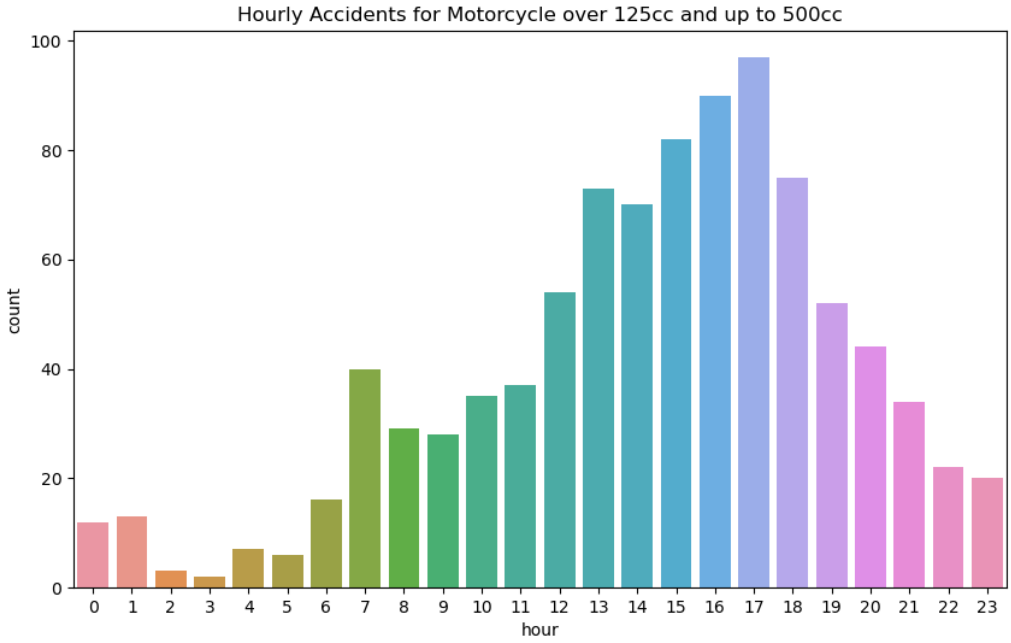

Figure 7: Motorcycle 125cc and upo 500cc hourly accidents

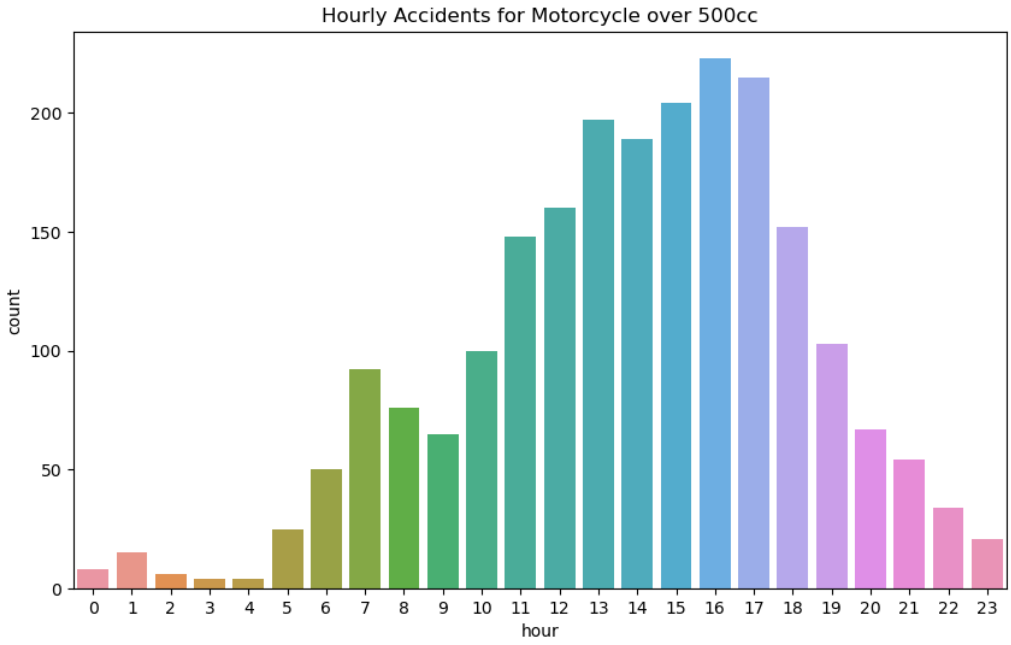

Figure 8: Motorcycle 500cc over hourly accidents

The pattern of motorcycle accidents varies with engine size. Motorcycles with 125cc and under peak in accidents at 5-6 PM, with fewest in the early hours. Motorcycles between 125cc and 500cc also experience more accidents in the late afternoon, specifically from 3-5 PM. Larger motorcycles over 500cc follow a similar trend, with heightened accident rates at 4-5 PM, and a pronounced pattern during morning rush hours, tailing off towards early morning.

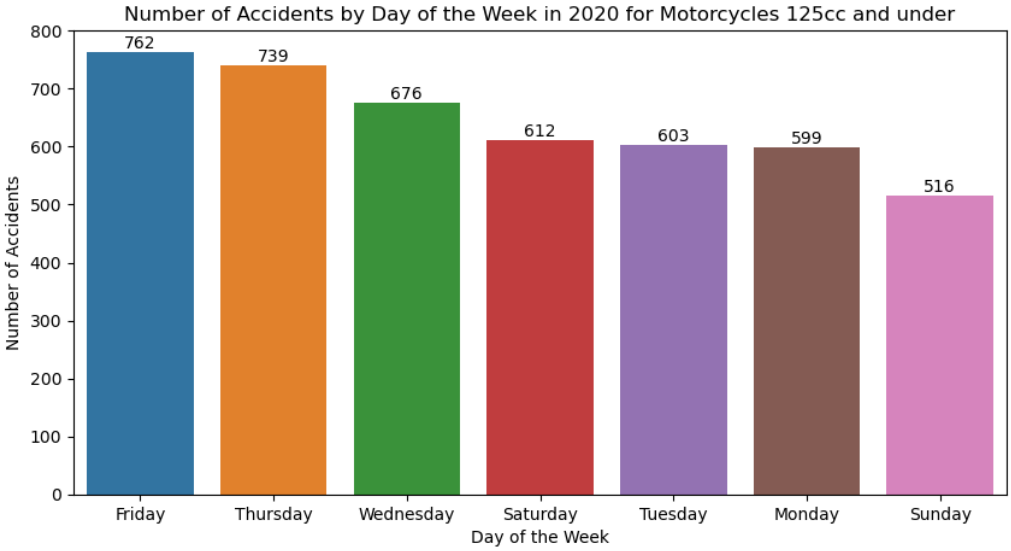

Figure 9: Motorcycle 125cc and under day of the week accidents

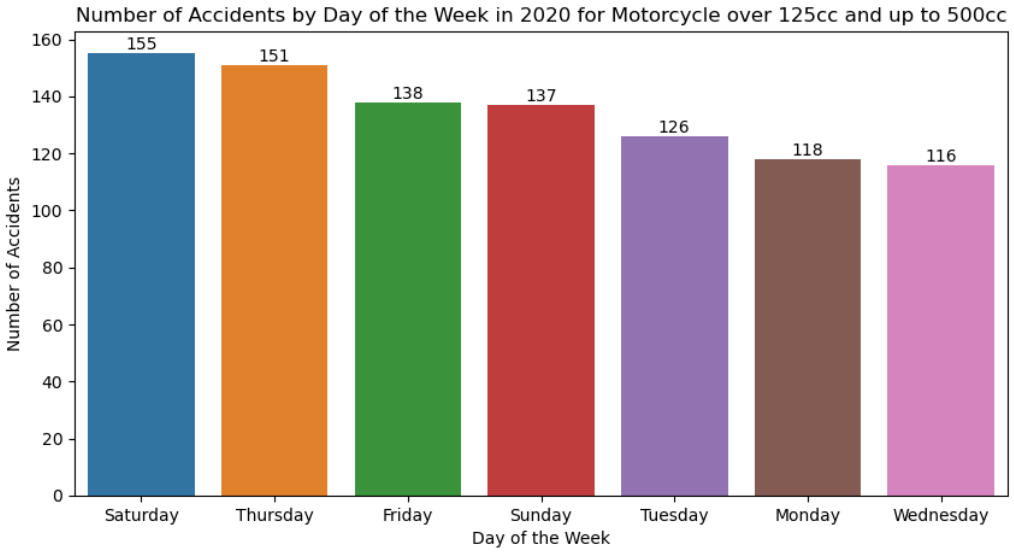

Figure 10: Motorcycle 125cc and upto 500cc day of the week accidents

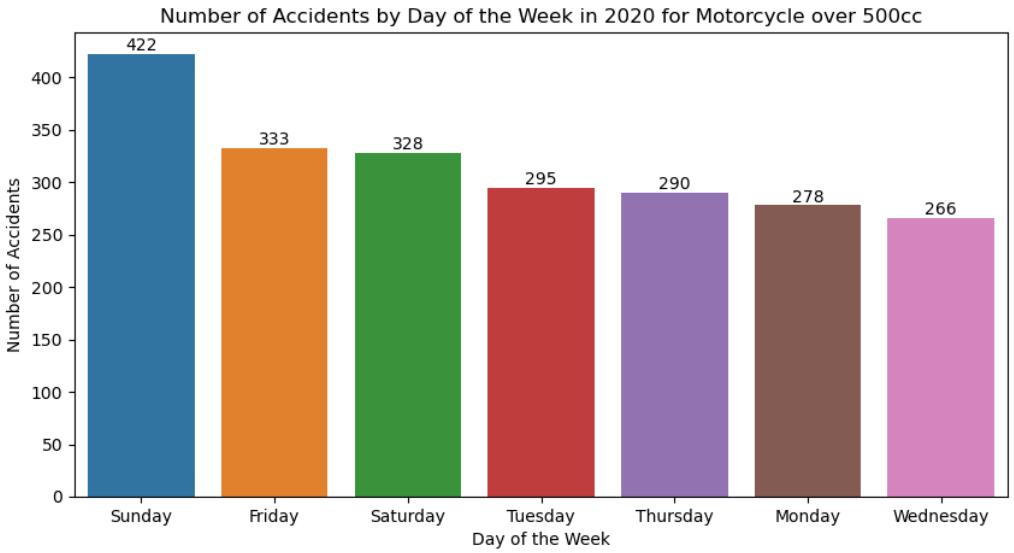

Figure 11: Motorcycle 500cc and over day of the week accidents

In 2020, motorcycle accidents showed distinctive patterns across engine sizes, with motorcycles of 125cc and under witnessing the most accidents on Fridays and least on Sundays; those over 125cc to 500cc having the most on Saturdays and least on Wednesdays; and motorcycles over 500cc experiencing the highest number of accidents on Sundays and the lowest on Wednesdays, highlighting weekends as commonly riskier across categories.

TASK 3

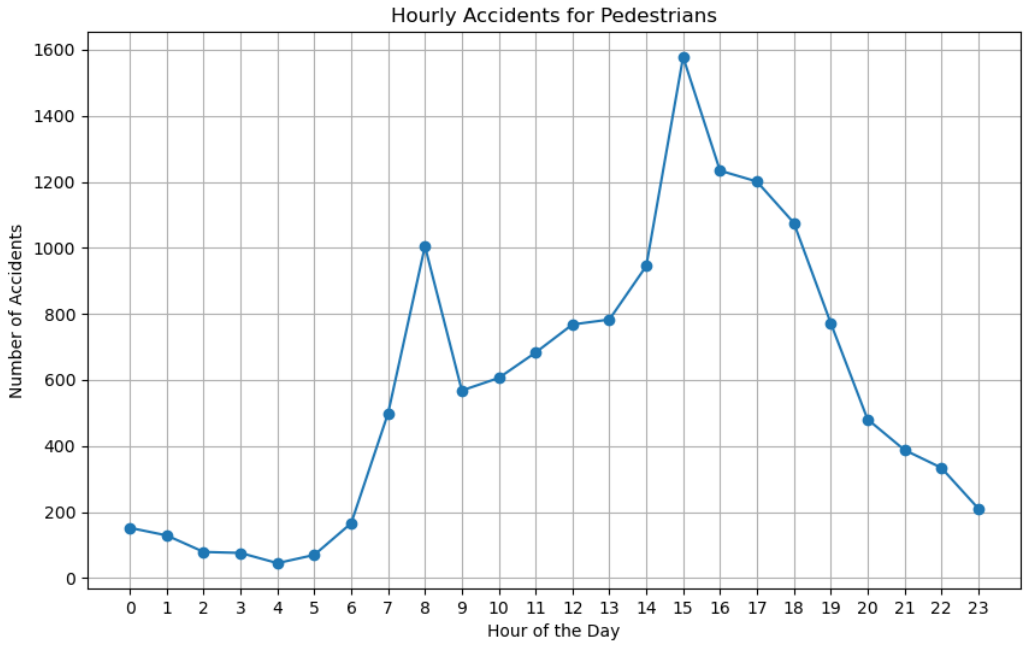

Figure 12: Pedestrians hourly accidents

The data suggests that pedestrian accidents most likely occurred at 3PM – 4PM, followed by 5PM, 6PM and 8AM, which is during the morning and afternoon rush hours. These trends is reflective of a typical daily activity patterns, where more people are on the roads during commuting times and regular business hours.

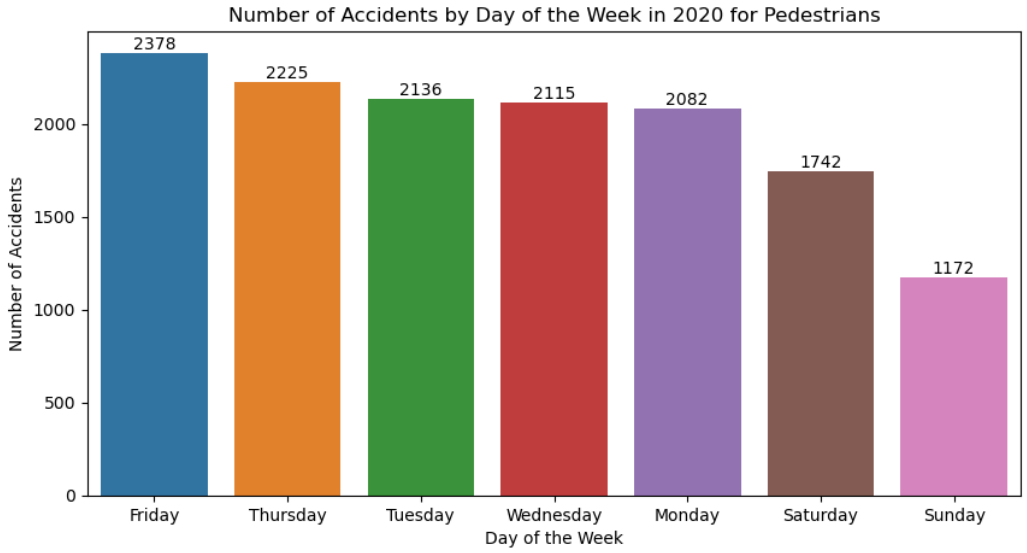

Figure 13: Pedestrians accidents by day of the week

Upon analysing the significant days of the week on which pedestrians are more likely to be involved in an accident, the resulting plot demonstrates that the number of pedestrian accidents is highest on Fridays and gradually decreases throughout the weekend, with the lowest number occurring on Sundays. The weekdays (Monday to Friday) show higher accident counts, possibly reflecting regular work and school commuting patterns. Conversely, the

weekend (Saturday and Sunday) shows a decline, likely corresponding with reduced road activity.

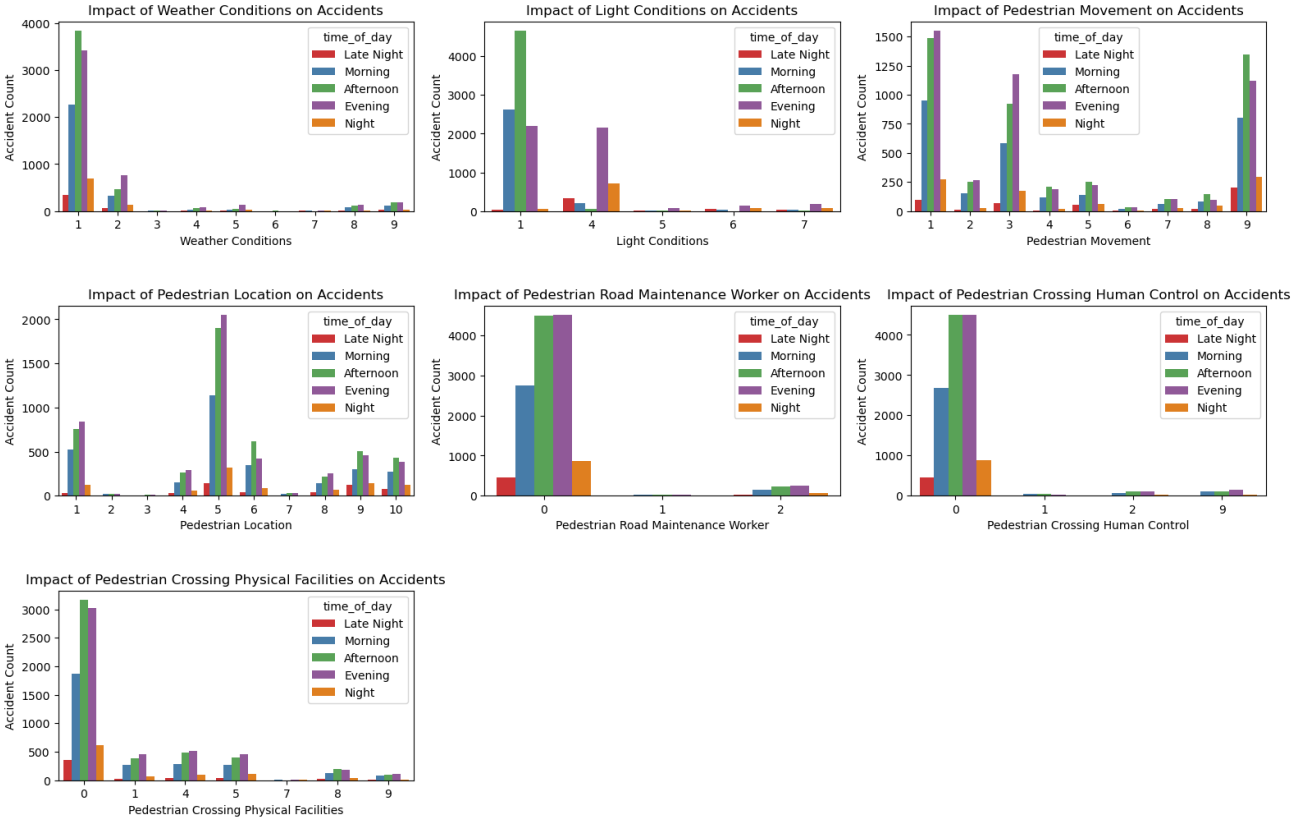

Figure 14: Conditions in which Pedestrians accidents occurred.

The analysis of pedestrian accidents shows that most incidents occur during clear weather, primarily during daylight hours. A significant number of these accidents happen when pedestrians cross the road, especially in areas without designated crossing facilities or human control within a 50-metre radius. Darkness plays a moderate role in accidents, especially during the evening and night. Notably, pedestrian road maintenance workers were not typically involved in these incidents.

TASK 4

	support	itemsets	length
0	0.984748	(accident_severity)	1
10	0.286779	(light_conditions, accident_severity)	2
11	0.923892	(road_type, accident_severity)	2
12	0.709876	(number_of_vehicles, accident_severity)	2
13	0.221559	(weather_conditions, accident_severity)	2
14	0.498416	(junction_detail, accident_severity)	2
15	0.920964	(accident_severity, vehicle_type)	2
16	0.878738	(age_of_vehicle, accident_severity)	2
17	0.331725	(sex_of_driver, accident_severity)	2
18	0.230386	(accident_severity, casualty_class)	2
42	0.268874	(light_conditions, road_type, accident_severity)	3
43	0.274477	(light_conditions, accident_severity, vehicle_...	3
44	0.257909	(light_conditions, age_of_vehicle, accident_se...	3
45	0.658472	(number_of_vehicles, road_type, accident_sever...	3
46	0.208336	(weather_conditions, road_type, accident_sever...	3
47	0.490795	(junction_detail, road_type, accident_severity)	3
48	0.864417	(road_type, accident_severity, vehicle_type)	3
49	0.824329	(age_of_vehicle, road_type, accident_severity)	3

Figure 15: Support

The support values of approximately 92.39% for (road_type, accident_severity) and 92.1% for (accident_severity, vehicle_type) indicate a strong association between these pairs in the dataset**.**

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
3	(number_of_vehicles)	(accident_severity)	0.717870	0.984748	0.709876	0.988865	1.004181	0.002956	1.369762	0.014758
4	(weather_conditions)	(accident_severity)	0.224454	0.984748	0.221559	0.987103	1.002392	0.000529	1.182636	0.003077
5	(junction_detail)	(accident_severity)	0.503240	0.984748	0.498416	0.990413	1.005753	0.002851	1.590925	0.011515
9	(sex_of_driver)	(accident_severity)	0.334499	0.984748	0.331725	0.991707	1.007067	0.002328	1.839085	0.010544
34	(light_conditions)	(accident_severity, vehicle_type)	0.293205	0.920964	0.274477	0.936126	1.016463	0.004445	1.237366	0.022915
36	(road_type, number_of_vehicles)	(accident_severity)	0.666400	0.984748	0.658472	0.988104	1.003408	0.002236	1.282105	0.010181
39	(weather_conditions, road_type)	(accident_severity)	0.211165	0.984748	0.208336	0.986603	1.001884	0.000392	1.138488	0.002384
41	(weather_conditions)	(road_type, accident_severity)	0.224454	0.923892	0.208336	0.928188	1.004650	0.000964	1.059821	0.005968
42	(junction_detail, road_type)	(accident_severity)	0.495609	0.984748	0.490795	0.990287	1.005626	0.002746	1.570368	0.011091
44	(junction_detail)	(road_type, accident_severity)	0.503240	0.923892	0.490795	0.975270	1.055610	0.025855	3.077524	0.106049
53	(road_type, sex_of_driver)	(accident_severity)	0.312438	0.984748	0.309729	0.991332	1.006686	0.002057	1.759517	0.009659
55	(sex_of_driver)	(road_type, accident_severity)	0.334499	0.923892	0.309729	0.925949	1.002227	0.000688	1.027782	0.003339
58	(casualty_class)	(road_type, accident_severity)	0.234838	0.923892	0.223116	0.950086	1.028353	0.006152	1.524800	0.036033
59	(junction_detail, number_of_vehicles)	(accident_severity)	0.392537	0.984748	0.389807	0.993044	1.008425	0.003257	2.192848	0.013754
61	(number_of_vehicles, vehicle_type)	(accident_severity)	0.664382	0.984748	0.656586	0.988266	1.003572	0.002337	1.299797	0.010606

Figure 16: Association rules

The analysis using the Apriori algorithm highlights the associations or relationships between various factors and accident severity. These factors include the number of vehicles, weather conditions, junction details, the sex of the driver, light conditions, and the combination of road type with the number of vehicles. All these elements show high confidence levels (ranging from 93.61% to 99.17%) and lift values slightly above 1, indicating more than random correlations with accident severity. These findings emphasize the multifaceted nature of road safety, where different variables interact in intricate ways, affecting accident outcomes.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
51	(light_1, n_vehicles_2)	(severity_3)	0.462450	0.783484	0.380136	0.822004	1.049164	0.017813	1.216404	0.087173
3	(n_vehicles_2)	(severity_3)	0.629305	0.783484	0.516749	0.821142	1.048065	0.023698	1.210546	0.123714
5	(sex_of_driver_2)	(severity_3)	0.251812	0.783484	0.205671	0.816765	1.042477	0.008380	1.181625	0.054460
137	(light_1, weather_1, n_vehicles_2)	(severity_3)	0.383031	0.783484	0.311736	0.813867	1.038779	0.011637	1.163230	0.060507
41	(weather_1, n_vehicles_2)	(severity_3)	0.496650	0.783484	0.403996	0.813441	1.038235	0.014878	1.160574	0.073164
205	(light_1, sex_of_driver_1, n_vehicles_2)	(severity_3)	0.304510	0.783484	0.246351	0.809009	1.032579	0.007773	1.133645	0.045365
65	(sex_of_driver_1, n_vehicles_2)	(severity_3)	0.419149	0.783484	0.338458	0.807487	1.030636	0.010061	1.124680	0.051175
192	(light_1, road_type_6, n_vehicles_2)	(severity_3)	0.336583	0.783484	0.271746	0.807369	1.030485	0.008039	1.123991	0.044592
60	(road_type_6, n_vehicles_2)	(severity_3)	0.450805	0.783484	0.363282	0.805852	1.028549	0.010083	1.115210	0.050541
307	(sex_of_driver_1, weather_1, light_1, n_vehicl...	(severity_3)	0.253917	0.783484	0.203511	0.801486	1.022976	0.004571	1.090679	0.030103
181	(sex_of_driver_1, weather_1, n_vehicles_2)	(severity_3)	0.333118	0.783484	0.266681	0.800560	1.021794	0.005688	1.085615	0.031983

Figure 17: Association rules

The impact of selected variables on accident severity reveals that light conditions and the number of vehicles, having two vehicles alone, and the sex of the driver are all associated with severity level 3. The lift values ranging from 1.042477 to 1.049164 indicate strong association between these variables and accident severity (3).

TASK 5:

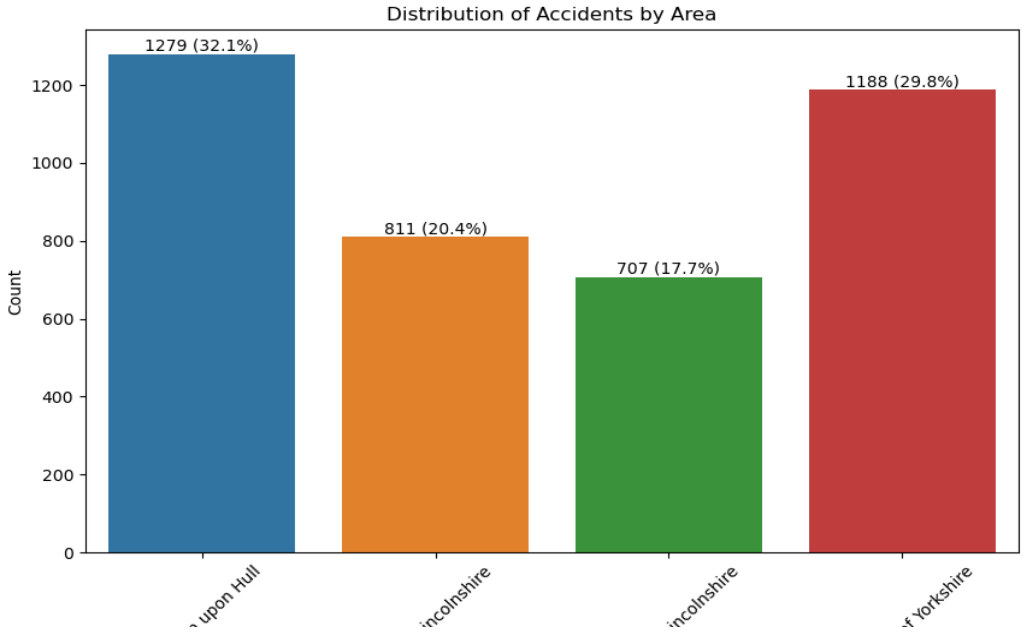

Figure 18: Humberside Accidents

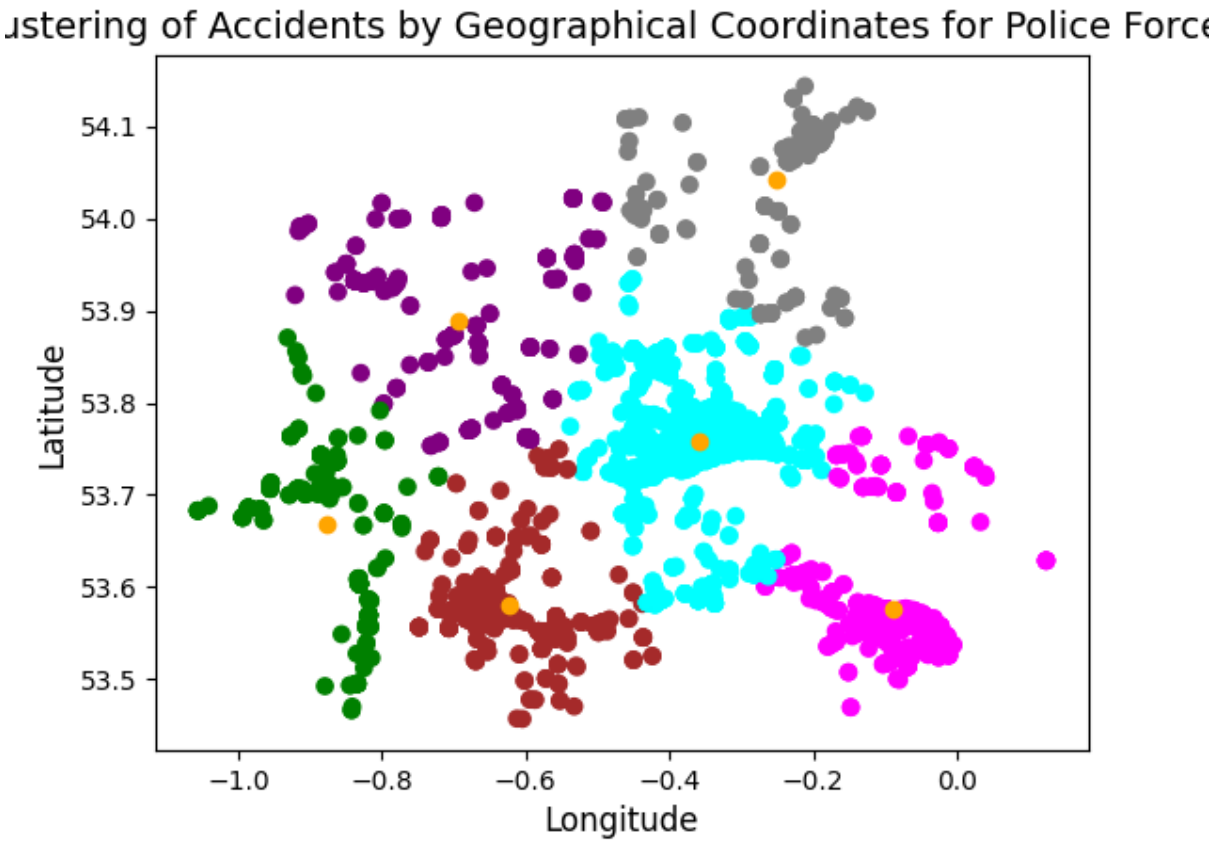

Figure 19: Humberside Accidents clusters

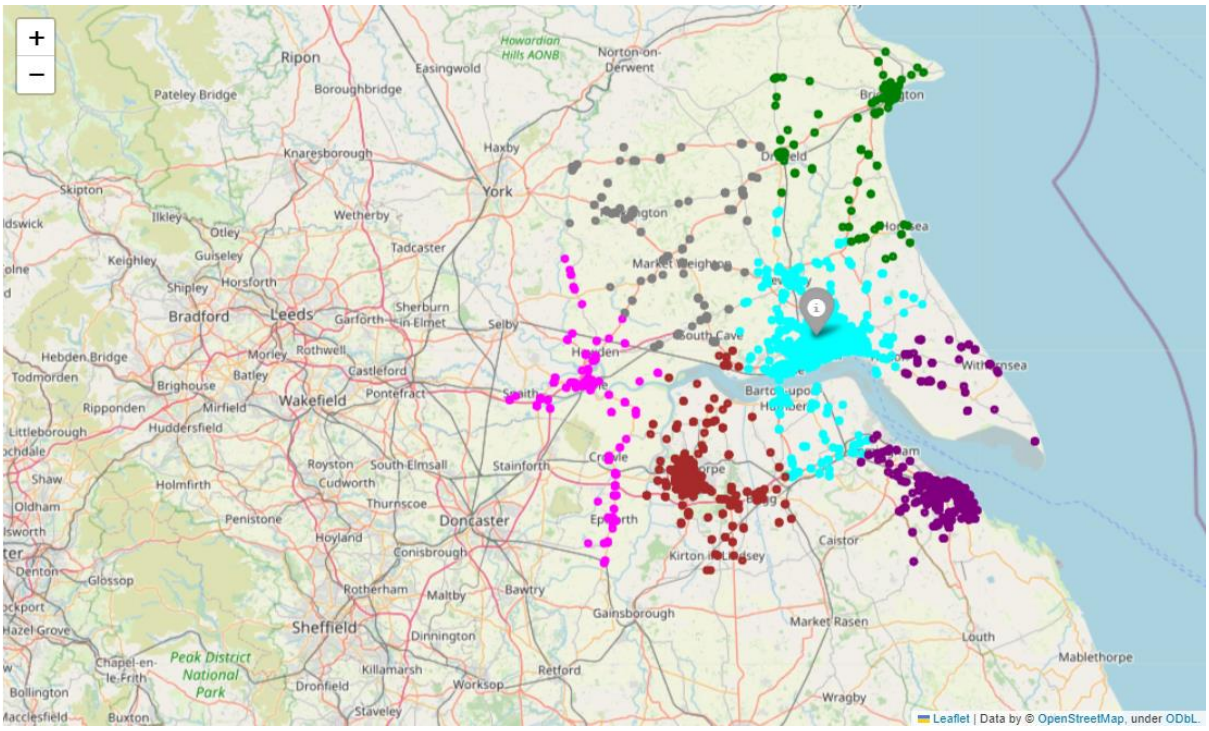

Figure 20: Clustered map of accidents in the regions

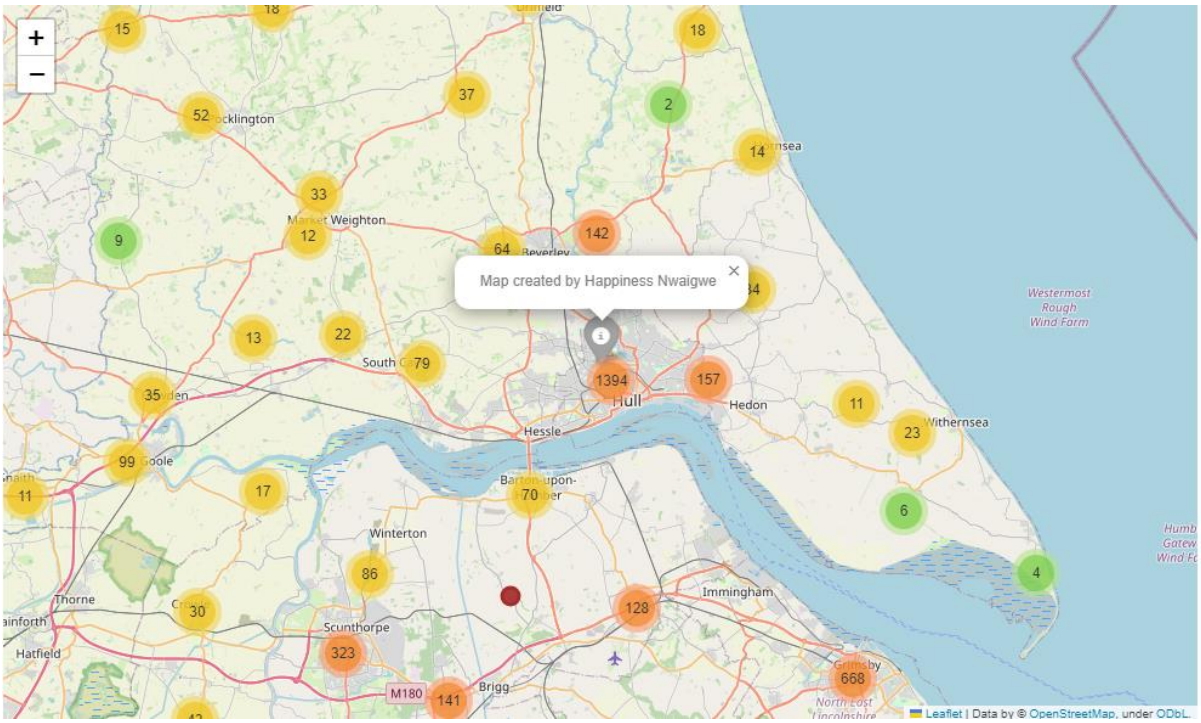

Figure 21: Accidents spatial map

This clustering map groups the accidents based on spatial proximity and not to necessarily align with the administrative boundaries of Kingston upon Hull, East Riding of Yorkshire, North Lincolnshire and Northeast Lincolnshire.

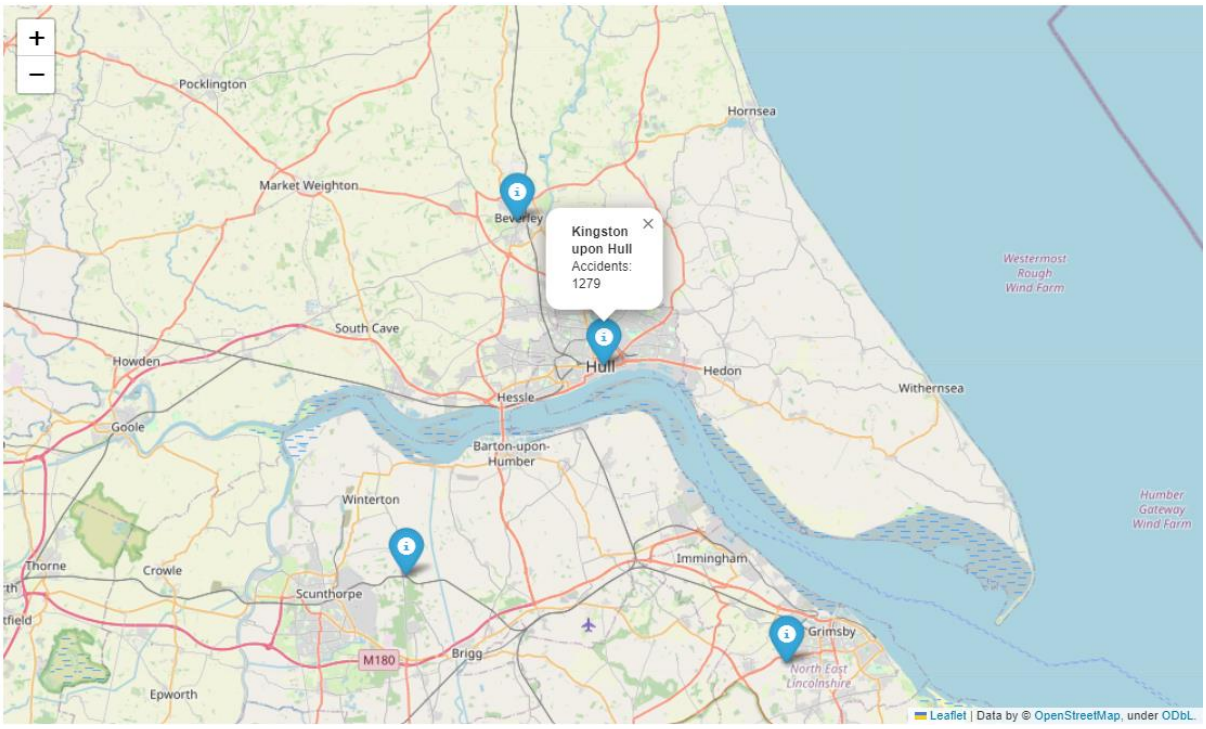

Figure 22: Total accidents by administrative boundary

From the detailed plots in (Figure 19 - Figure 22), Cluster 1 which is found to be the most spatially widespread cluster reveals that Kingston upon Hull has the highest occurrence with 1279 accidents, followed by East Riding of Yorkshire with 1188 accidents. North Lincolnshire and North East Lincolnshire report fewer accidents, with 811 and 707 respectively.

EXPLORING THE CONDITIONS OF THESE CLUSTERS

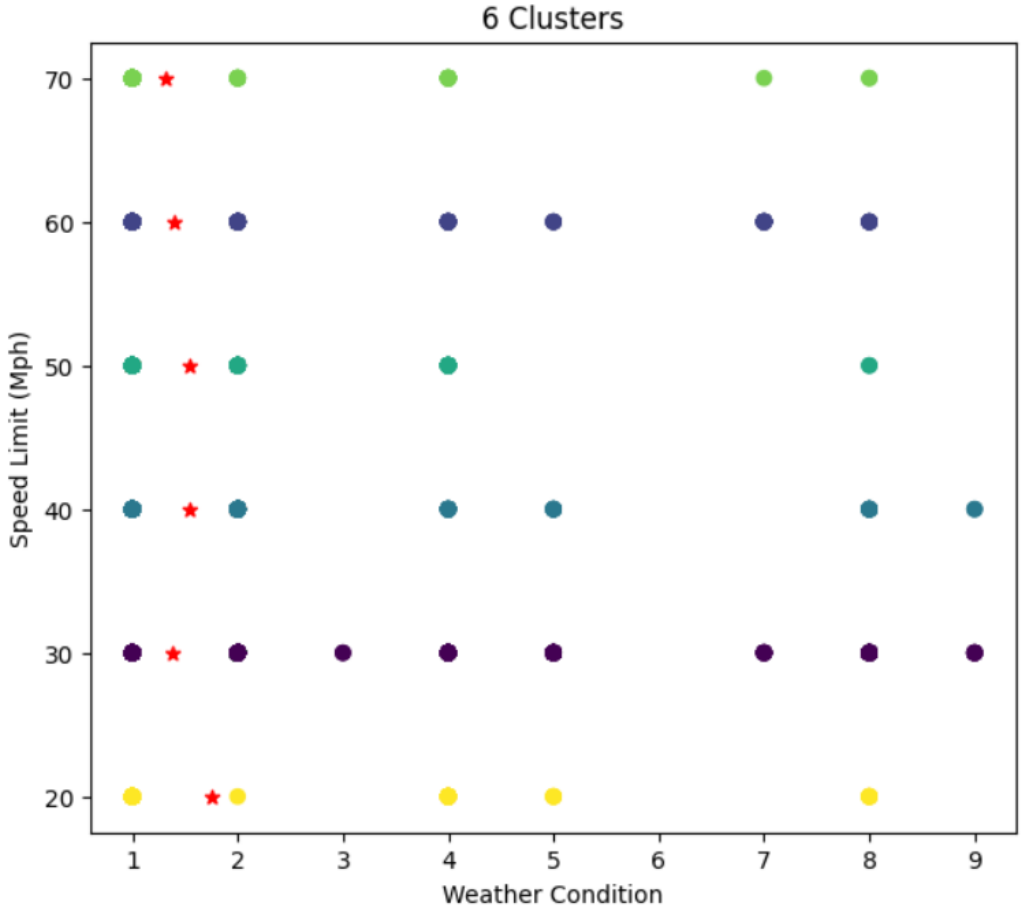

Figure 23: Weather and speed limit clusters

The clusters are between (1) and (2), which was when the weather was either fine or raining without high winds.

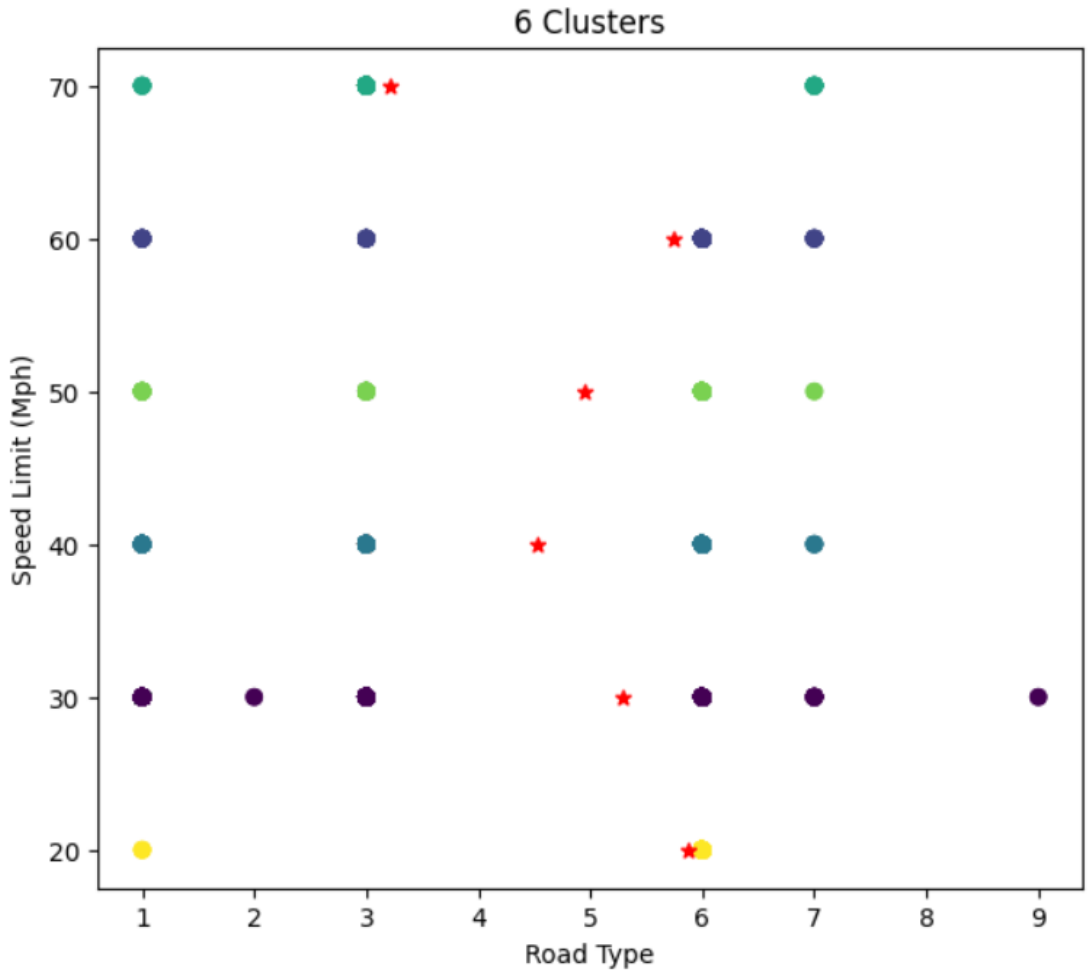

Figure 24: Road type and speed limit clusters

The clusters around Dual carriageway and Single carriageway (3 and 6) indicate the impact of 70 speed limit on Dual carriageway, and 20 and 60 speed limit impact on Single carriageway.

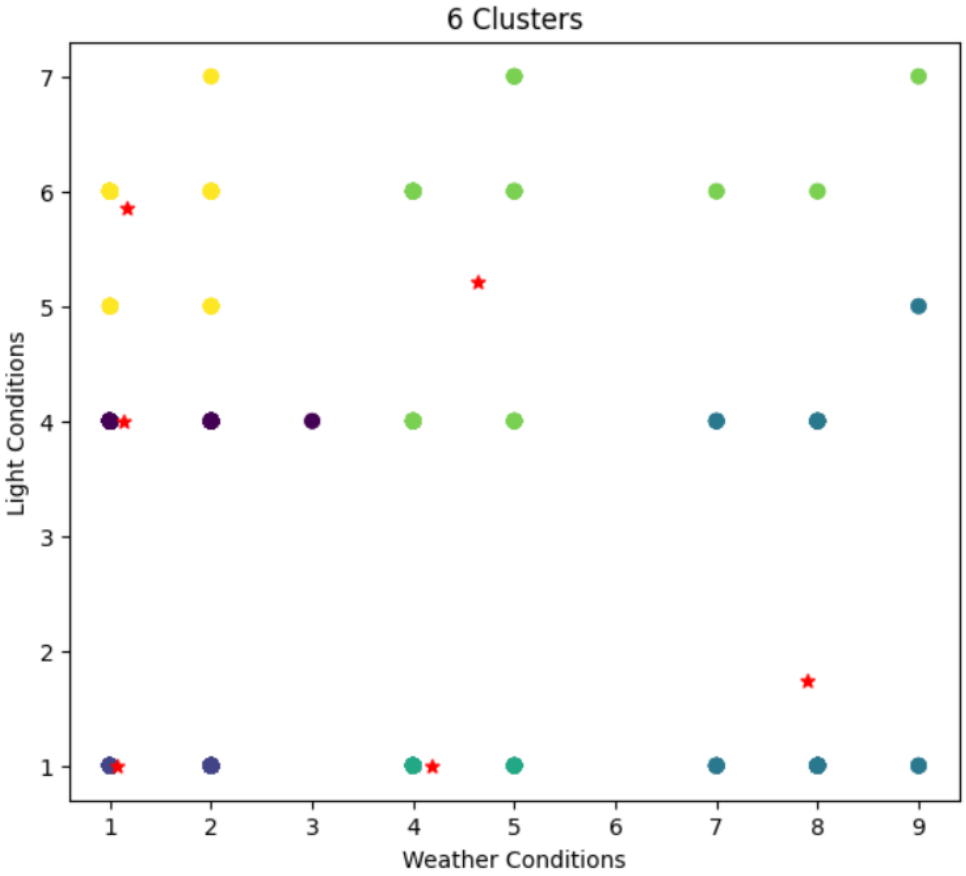

Figure 25: Light conditions and weather conditions clusters

This cluster (1) reveals that weather was fine without high winds when accidents occurred in darkness 4 and 6, either with streetlight or no street lighting.

TASK 6:
Using Isolation Forest:

	count	mean	std	min	25%	50%	75%	max
junction_detail	17383.0	19.358166	36.205680	0.0	1.0	3.0	7.0	99.0
age_band_of_driver	17383.0	3.847495	4.029182	-1.0	-1.0	6.0	7.0	11.0
vehicle_location_restricted_lane	17383.0	71.523960	43.972550	-1.0	9.0	99.0	99.0	99.0
speed_limit	17383.0	29.568832	11.365654	-1.0	20.0	30.0	30.0	70.0
number_of_casualties	17383.0	1.441638	1.231853	1.0	1.0	1.0	1.0	19.0
casualty_severity	17383.0	2.840246	0.434090	1.0	3.0	3.0	3.0	3.0
road_type	17383.0	5.611517	2.334519	1.0	3.0	6.0	6.0	9.0
weather_conditions	17383.0	2.719151	2.938805	1.0	1.0	1.0	2.0	9.0
pedestrian_crossing_human_control	17383.0	3.353621	4.331793	-1.0	0.0	0.0	9.0	9.0
age_of_vehicle	17383.0	4.715469	6.098776	-1.0	-1.0	3.0	9.0	63.0
vehicle_type	17383.0	10.832940	14.933286	1.0	8.0	9.0	9.0	98.0
second_road_number	17383.0	294.451130	1032.729265	-1.0	0.0	0.0	0.0	9174.0
engine_capacity_cc	17383.0	1355.594949	2022.394044	-1.0	-1.0	1248.0	1869.5	16400.0
age_of_casualty	17383.0	33.397515	19.062553	-1.0	24.0	32.0	44.0	99.0
special_conditions_at_site	17383.0	1.741702	3.388511	-1.0	0.0	0.0	0.0	9.0
casualty_class	17383.0	1.516712	0.793589	1.0	1.0	1.0	2.0	3.0
pedestrian_location	17383.0	1.289133	2.929883	0.0	0.0	0.0	0.0	10.0
age_of_driver	17383.0	24.151873	23.296186	-1.0	-1.0	26.0	40.0	99.0
urban_or_rural_area	17383.0	1.118967	0.323759	1.0	1.0	1.0	1.0	2.0
first_road_class	17383.0	3.995455	1.353701	1.0	3.0	3.0	6.0	6.0
number_of_vehicles	17383.0	1.936375	0.847666	1.0	2.0	2.0	2.0	13.0
light_conditions	17383.0	2.377840	2.147082	1.0	1.0	1.0	4.0	7.0

Figure 26: Isolation Forest

	count	mean	std	min	25%	50%	75%	max
accident_severity	139153.0	2.775305	0.454759	1.0	3.0	3.0	3.0	3.0
number_of_vehicles	139153.0	1.879061	0.874735	1.0	1.0	2.0	2.0	10.0
number_of_casualties	139153.0	1.521318	1.069019	1.0	1.0	1.0	2.0	13.0
first_road_class	139153.0	4.105244	1.380012	1.0	3.0	3.0	6.0	6.0
first_road_number	139153.0	827.803051	1601.146140	0.0	0.0	57.0	579.0	9174.0
road_type	139153.0	5.310133	1.798937	1.0	6.0	6.0	6.0	9.0
speed_limit	139153.0	32.637557	11.527431	-1.0	30.0	30.0	30.0	70.0
junction_detail	139153.0	6.567692	18.962642	0.0	0.0	3.0	6.0	99.0
junction_control	139153.0	1.976098	2.552733	-1.0	-1.0	2.0	4.0	9.0
second_road_class	139153.0	3.343155	2.642997	-1.0	0.0	4.0	6.0	6.0
second_road_number	139153.0	277.746466	1016.117849	-1.0	-1.0	0.0	0.0	9157.0
pedestrian_crossing_human_control	139153.0	1.149519	2.913559	-1.0	0.0	0.0	0.0	9.0
pedestrian_crossing_physical_facilities	139153.0	2.874951	3.186843	-1.0	0.0	1.0	5.0	9.0
light_conditions	139153.0	2.243890	1.813572	-1.0	1.0	1.0	4.0	7.0
weather_conditions	139153.0	2.636199	2.567238	-1.0	1.0	2.0	2.0	9.0
road_surface_conditions	139153.0	1.696341	1.414269	-1.0	1.0	1.0	2.0	9.0
special_conditions_at_site	139153.0	0.510632	1.988047	-1.0	0.0	0.0	0.0	9.0
carriageway_hazards	139153.0	0.404368	1.804154	-1.0	0.0	0.0	0.0	9.0
urban_or_rural_area	139153.0	1.185781	0.389171	1.0	1.0	1.0	1.0	3.0
trunk_road_flag	139153.0	1.765402	0.757008	-1.0	2.0	2.0	2.0	2.0
vehicle_type	139153.0	10.373488	11.797186	1.0	9.0	9.0	9.0	98.0
vehicle_manoeuvre	139153.0	26.707178	32.198440	-1.0	9.0	18.0	18.0	99.0

Figure 27: Using Inter Quantile Range (IQR)

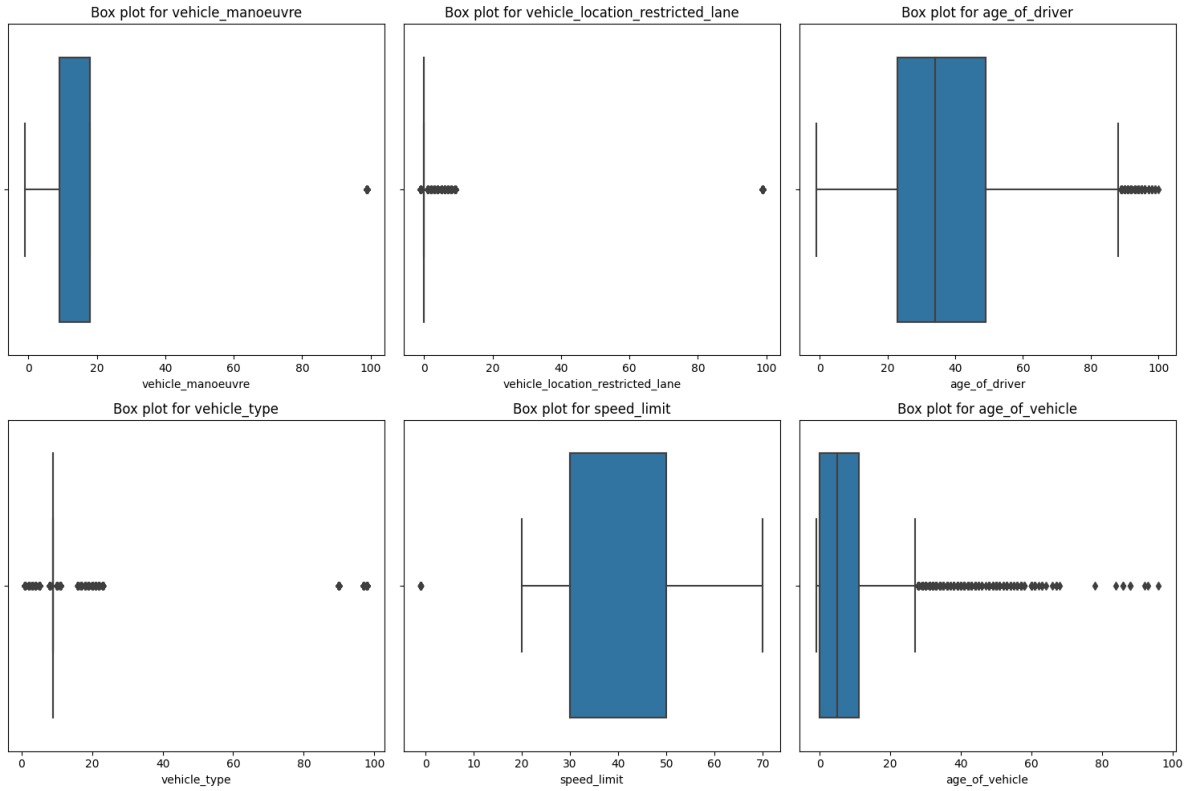

Figure 28: Outliers box plot

The box plot analysis has detected notable outliers using both the Isolation Forest and the Multiple of Interquartile Range (IQR) methods. These outliers, defined as data points more than 1.5 times the interquartile range away from the nearest quartile, include unrealistic negative values like -1 in fields such as the age of the driver and age of the vehicle. Recognizing that these are anomalies and inconsistent with logical interpretation, they will be cleaned and converted to NaN (Not a Number). A technique called Multiple Imputation by Chained Equations (MICE) will then be used to provide more accurate estimations for these missing values, taking care not to be influenced by extreme data points.

CLASSIFICATION MODEL

TASK 7:

	precision	recall	f1-score	support
Fatal	0.87	0.92	0.90	1230
Non-Fatal	0.92	0.87	0.90	1309
accuracy			0.90	2539
macro avg	0.90	0.90	0.90	2539
weighted avg	0.90	0.90	0.90	2539

Figure 29: Classification report

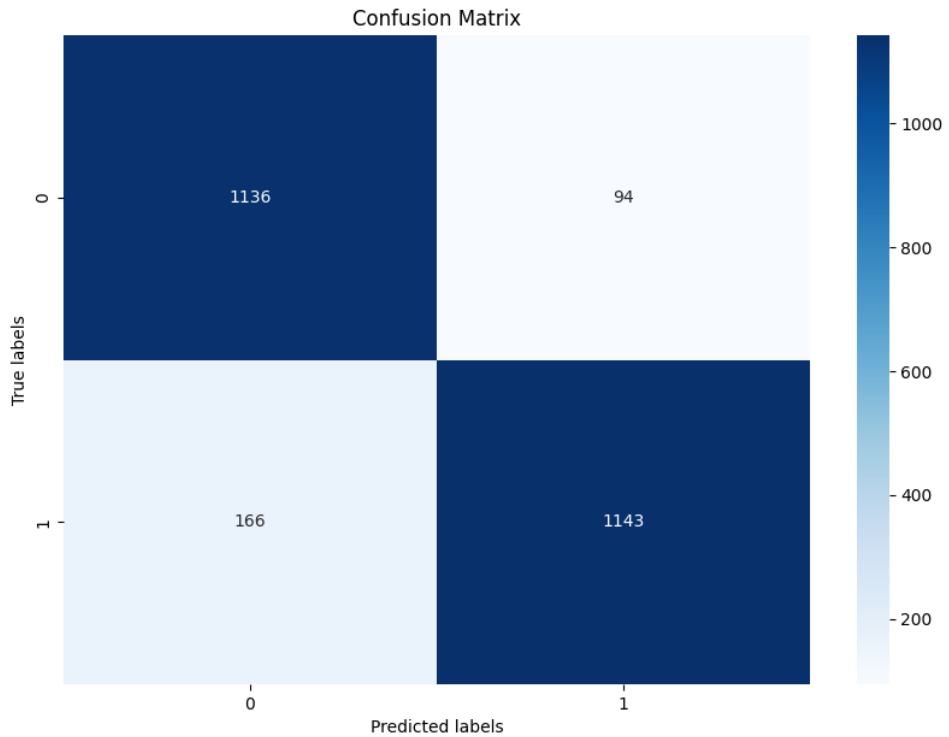

Figure 30: Confusion Matrix

The model demonstrates high accuracy, precision, recall, and F1-score values, which was all around 90%. This uniformity strongly suggests that the model is both accurate and balanced in its predictions for both "Fatal" and "Non-Fatal" injuries sustained in road traffic accidents. The confusion matrix further supports these findings, as it shows the model has made more correct predictions (1193 true positives and 1095 true negatives) than incorrect ones (135 false positives and 116 false negatives). Generally, these results indicate that the model is highly capable of informing and potentially enhancing road safety measures through accurate predictions of accident severity.

RECOMMENDATIONS

Based on the detailed accident analysis I have carried out, here are my five solid recommendations to enhance road safety and reduce accident occurrences:

- Manage Traffic During Peak Hours: Implement strategies to regulate flow and minimize collisions during high-risk times like late afternoons and early evenings.
- Improve Road Infrastructure: Focus on road types, junctions, and crossings by enhancing design, signage, and pedestrian facilities in high-risk zones.
- Consider Weather and Lighting: Address the risks in dark or wet conditions with adaptive lighting, anti-skid surfaces, and responsive traffic management.
- Encourage Safe Riding for Motorcyclists: Tailor awareness campaigns and training to different engine sizes, emphasizing safe practices, speed limits, and protective gear.
- Enhance Pedestrian Safety Measures: Prioritize pedestrian safety by implementing well-lit crosswalks, pedestrian signals, and public awareness campaigns on safe crossing practices. Monitor accident data to identify high-risk areas and invest in infrastructure that protects pedestrians, such as barriers or pedestrian bridges, especially in zones where accidents frequently occur.

REFERENCES

Aleryani, A., Wang, W. & Iglesia, B. de la (2020) *Multiple imputation ensembles (mie) for dealing with missing data - sn computer science*. SpringerLink. Available online: https://link.springer.com/article/10.1007/s42979-020-00131-0 [Accessed 1/8/2023].

Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J. (2011) Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1), 40–49.

Department for Transport (2021) *STATS19 forms and guidance*. GOV.UK. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995422/stats19.pdf [Accessed 10 / 8/ 2023].

Department for Transport (2011) *STATS20 Instructions for the Completion of Road Accident Reports from non-CRASH Sources*. GOV.UK. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995423/stats20-2011.pdf [Accessed 10 / 8/ 2023].

Road Safety Data Guide. (2022) Road Safety Data. Available online: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data [Accessed 13/8/2023].

Road Safety Management (2018) *European Commission, Road Safety Management*. European Road Safety Observatory. Available online: https://road-safety.transport.ec.europa.eu/system/files/2021-07/ersosynthesis2018-roadsafetymanagement.pdf [Accessed 10 / 8/ 2023].

Stacked bar chart — Matplotlib 3.3.4 documentation. (n.d.) matplotlib.org. Available online: https://matplotlib.org/stable/gallery/lines_bars_and_markers/bar_stacked.html [Accessed 10 / 8/ 2023].

Zhan, Q., Deng, S. and Zheng, Z. (2017) An adaptive sweep-circle spatial clustering algorithm based on gestalt. *ISPRS International Journal of Geo-Information*, [online] 6. Available online: https://www.mdpi.com/2220-9964/6/9/272.