Coursework submission for Transport Data Science (TRAN5340M)

Road safety analysis of motorists and motorcyclists in Leeds

Utkarsh Balooni (201792310)

Statement

```
TRAN5340M
Transport Data Science
Assignment Title:   Road safety analysis of motorists and motorcyclists in Leeds
Student ID: 201792310
Word Count: xxxx
Lecturer: Dr. Robin Lovelace
Submission Date: 24-05-2024 (extended)
Semester: 2
Academic Year: 2024-25
Generative AI Category: AMBER
```

Use of Generative Artificial Intelligence (Gen AI) in this assessment:

I have used Gen AI only for the specific purposes outlined in my acknowledgements

By submitting the work to which this sheet is attached you confirm your compliance with the University's definition of Academic Integrity as: "a commitment to good study practices and shared values which ensures that my work is a true expression of my own understanding and ideas, giving credit to others where their work contributes to mine". Double-check that your referencing and use of quotations is consistent with this commitment.

Acknowledgements

Gen AI in this report has been used for:

- Literature review
- Generating references
- Debugging code
- Explaining important concepts
- Correcting grammar

1. Introduction

Road accidents are a pressing public safety issue, posing significant risks to human life. Five people die every day on the road in the UK, and 82 are seriously injured on average (GOV.UK 2022). The government has been actively involved in making policies to address road safety issues and cut down road deaths (Department for Transport 2019). This report aims to provide valuable insights on road safety of motorists and motorcyclists using transport data science approaches. The techniques used are hotspot analysis, black spot analysis, classification modelling and association rule mining.

2. Background

Existing studies have analysed various risk factors relating to accident severity. (Kwon, Rhee, and Yoon 2015) inspected traffic accident reports accumulated by the California Highway Patrol (CHP) since 1973 for each accident report containing around 100 data fields. Among them, 25 fields between 2004 and 2010 were selected as the most relevant to car accidents. The analysis found five risk factors (including population) that majorly contributed to accident severity. Another study on road safety (Daniels, Thompson, and Smith 2020) analysed the reliability of freeway and state roads in the province of Cosenza (Italy) by defining the seriousness of the accidents. The report also emphasised that it is impossible to attribute the cause of an accident to a sole component of the system 'driver – vehicle – environment – infrastructure'. There have also been multiple studies on hotspot analysis. In ("Kernel Density Estimation and k-Means Clustering to Profile Road Accident Hotspots" 2009) a methodology using Geographical Information Systems (GIS) and Kernel Density Estimation was used to study the spatial patterns of injury related road accidents in London, UK. A clustering methodology was developed using environmental data and results from the spatial pattern analysis in order to create a classification of road accident hotspots. In (Kuyumcu, Aslan, and Yurtay 2023) likelihood of accidents were analysed by investigating the possible relationships of risk factors using association rule mining.

3. Scope

The report's analysis is structured into two parts. Both parts primarily use accident data pertaining to cars and motorcycles in Leeds from 2018 to 2022.

1. Part one analyses the accident data to find accident-prone areas on the LSOA and road level.
2. Part two analyses accident data to find the attributes most affecting the severity level.

4. Datasets used

The primary dataset used in this report is the collision (and casualty) data from the "stats19" package. Vector data for roads is extracted from open street maps through an overpass query built using the `osmdata` package. Boundary (LSOA) data for Leeds and population density data are obtained from the Office for National Statistics website. Road traffic flow data is obtained from the Department of Transport website.

4.1 Data preparation

Since collision data from "stats19" is the primary data set used, we will illustrate the data preparation methods used in context of combined collision data.

- We combine the data for all years using the `bind_rows()` function. Prior to combining the data sets, `lsoa_of_casualty` column of the `cas_2019` dataset had to be converted to `character` class. The variable `accident_index` contained `inf` values so we use the key `accident_reference` for data linkage. Since the `local_authority_district` variable was missing for the years 2021 and 2022, LSOA (lower super output area) data for Leeds was used to filter the records. Out of the 37 variables provided, 17 were chosen for further analysis.
- To get the casualty type of the collision, we join the casualty data sets obtained from the `stats19` package to the collision data using the `right_join()` function on the column `accident_reference`.

Data for motor cycles and cars is filtered using the `casualty_type` field. Since one accident can be linked to multiple casualties, the joined data set contained duplicate rows which were removed. Out of the 6737 total accidents in Leeds, 3876 records were filtered for final analysis.

- Only values corresponding to "Car" and "Motorcycle" are selected for the variable `casualty_type`. Date and time are converted into appropriate formats using the `lubridate` package and day of the week and hour of accident occurrence is extracted.

4.2 Exploratory data analysis

Summary statistics for the dataset reveal that the average number of casualties per accident is 1.4 and the median speed limit for the roads is 30 mph.

Figure 1 shows the number of accidents for each year from 2018-2022 grouped by severity. We can see a sharp dip in the total number of accidents for the year 2020, likely due to the lock down imposed. Overall, the number of accidents seem to rise each year with a significant increase in the count of serious injuries.

Figure 2 (Lovelace 2020) shows the proportion of accidents for each day of the week per hour. It can be seen that the highest proportion of accidents occur after Friday night and Saturday night.

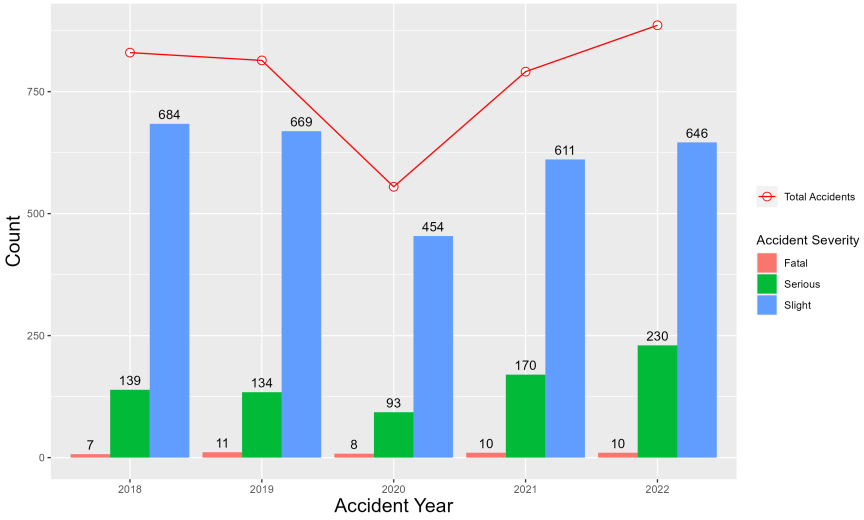

Figure 1: Accident count per year by severity

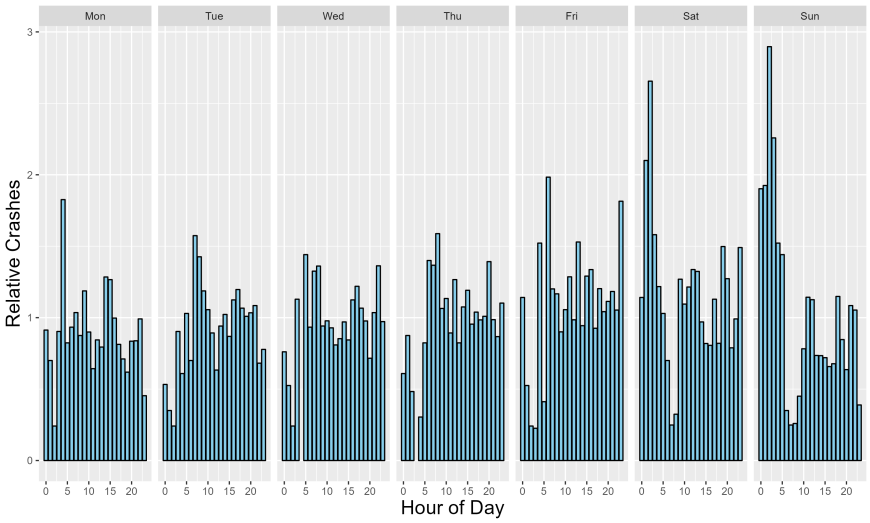

Figure 2: Proportion of accidents per day per hour

5. Hotspot Analysis

Hotspot analysis is a mapping technique used to identify clustering of spatial phenomena. A hotspot can be defined as an area that has a higher or lower concentration of events compared to the expected number of a random distribution of events. There are many techniques available for analysing hotspots. We will use the Getis Ord Gi^* statistic for our analysis.

The Getis Ord local statistic is given by

$$Gi^*(i) = \frac{\sum_{j=1}^{n} w_{ij} \cdot x_j - \bar{x} \cdot \sum_{j=1}^{n} w_{ij}}{s \cdot \sqrt{\frac{n \cdot \sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})^2}{(n-1) \cdot n}}}$$

where, x_j is the attribute value for feature j, w_{ij} is the feature weight between feature i and j, n is the number of features, $\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$, and $s = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \bar{x}^2}$ (Ord and Getis, n.d.)

The Gi^* statistic returned for each feature in the dataset is a z-score.

5.1 Methodology

We will use the accident count by population density for each LSOA as the target variable for clustering. Population density data is obtained from the ONS website and joined with the grouped LSOA accident data for Leeds. Following this, we calculate the data's global and local Gi statistics using the `spdep` and `sfdep` packages. Finally we classify the LSOAs into five categories - "Very Hot"„ "Hot", "Insignificant", "Cold", "Very Cold" based on the local Gi values and the corresponding p values. Relevant plots for each step are also shown.

5.2 Results

Figure 4 shows the accident counts by the population density (people per sq km). The LSOA with the highest proportion of car and motorcycle accidents is "E01011297", Harewood, Elmet And Rothwell, Leeds. One possible reason could be that it is located on the outskirts and experiences high vehicular traffic. There have also been multiple drunk driving cases in the area, including the Whitlam case.(Wharfedale Observer 2017)

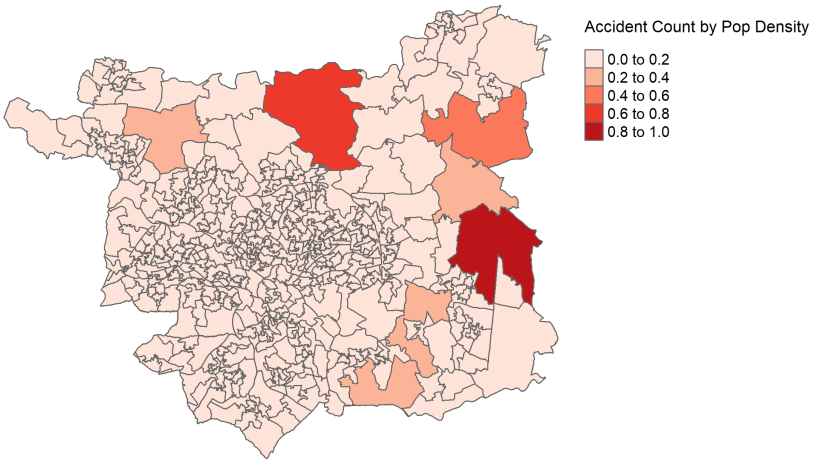

Figure 3: Accidents per person per sq Km

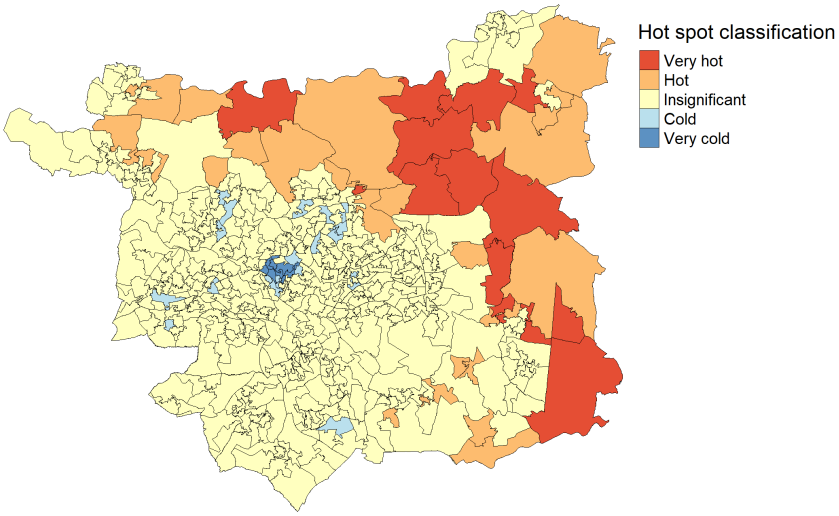

Figure 4: Accident hotspots by population density for Leeds

The global G statistic is calculated as 1.07e-02, which is statistically significant (p-value = 3.21e-11), indicating that there is some level of clustering present in the data. The results of the local G test are summarized in Figure 5.

It can be observed from Figure 5 that most of the car and motorcycle crashes occur along the northern and eastern boundary of Leeds which is evident by the presence of "Hot" and "Very Hot" regions in the area. Minor LSOAs in the central parts of the city witness little to no accidents as compared to the average.

5.4 Analyzing A6120

The major highway in the "accident-prone" region is "A6120," which is the outer ring road between Horsforth and Pudsey. Analysing the road closely, it was found that out of the 303 accidents recorded in the "Very Hot" and "Hot" regions, 101 were located within 10m of "A6120." Four of these accidents were fatal.

Leeds City Council has already introduced comprehensive safety proposals to significantly enhance 'A6120' and ensure safer travel (Leeds City Council 2014).

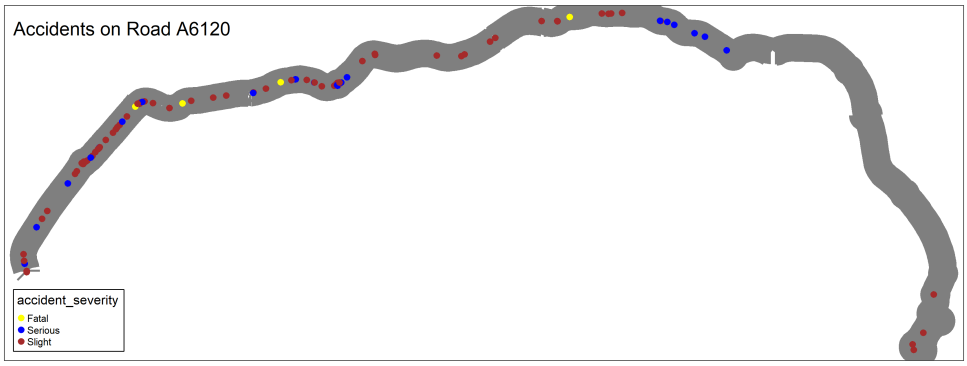

Figure 5: Accidents on A6120

6. Black spot analysis

Black spots are hazardous locations where road traffic collisions have historically been concentrated. Countries are increasingly using it to enhance road safety. We will refer to the black spot manual proposed by KGM (General Directorate of Highways in Turkey) (Turkish General Directorate of Highways 2001) to identify black spots for roads in Leeds.

The KGM black spot manual mentions a statistical method called Rate – Quality – Control for identifying black spots. It consists of calculating three different parameters for each road section: 1. Accident rate 2. Accident frequency 3. Severity index

Each of these parameters is compared to a critical value. If a certain road section shows higher values than the critical ones, it is considered a black spot. For our analysis, we will only consider the first two parameters, "Accident rate" and "Accident frequency."

Accident Rate: A road segment j is considered a black spot in terms of accident rate if:

$$(A_j/m_j) > R_c, \quad R_c = \hat{\lambda} + k_\alpha \sqrt{\hat{\lambda}/m_j} - 0.5/m_j$$

where,
A_j = Number of accidents on road j during the time period.
m_j = Number of vehicle kilometers in millions on road j during the same time period.
$\hat{\lambda} = \sum_{i=1}^{n} A_i / \sum_{i=1}^{n} m_i$, n is the total number of road sections.
k_α is a constant chosen to set the significance level.

Accident frequency: A road segment j is considered a black spot in terms of accident frequency if:

$$A_j > A_c, \quad A_c = F_{ave} + k_\alpha \sqrt{F_{ave}/L_j} - 0.5/L_j$$

where,
F_{ave} = Average accident frequency for all road sections.
L_j = Length of the j^{th} road.

6.1 Methodology

We obtain the vector data for "motorway", "primary", and "trunk" roads from Open Street Map using the `osm_data` package. Average annual daily flow data is obtained from the Department For Transport website to find the vehicle kilometres travelled on each road. Crash locations are joined to the nearest road within 50 m of the crash. Out of the 313 roads, only 163 were mapped to corresponding crashes. Traffic flow is measured using count points, which are joined to the road data using the `st_join()` function. Since count point data is unavailable for all roads, the joined dataset comprised 85 records. The significance level was chosen at 1%. Finally, the parameters for accident rate and frequency are computed and interpreted for the final dataset.

6.2 Results

There were no significant roads that can be considered black spots in the "accident rate" domain. The critical value R_c was much higher than the computed value, which assures that the accident rate is quite low, taking into account the daily traffic on the roads.

Appendix A mentions the 20 roads with a higher accident frequency parameter value than the critical value. These roads experienced a greater frequency of accidents than the other roads. The top seven significant roads, ordered by the number of accidents, are shown below in Table 1.

Table 1: Top 7 roads by accident count

name	number_of_accidents
Dewsbury Road	110
Bradford Road	92
York Road	86
Inner Ring Road	81
Wakefield Road	76
Leeds Road	66
Otley Road	64

Figure 7 shows the relationship between the average annual daily flow of cars and motorcycles (for five years in millions), the number of accidents and the road length. Most of the points in the plot are clustered near the origin, with a low accident count and AADF. The plot indicates that there is no apparent relationship between AADF and accident count, suggesting that other road attributes might influence the number of accidents on a particular road.

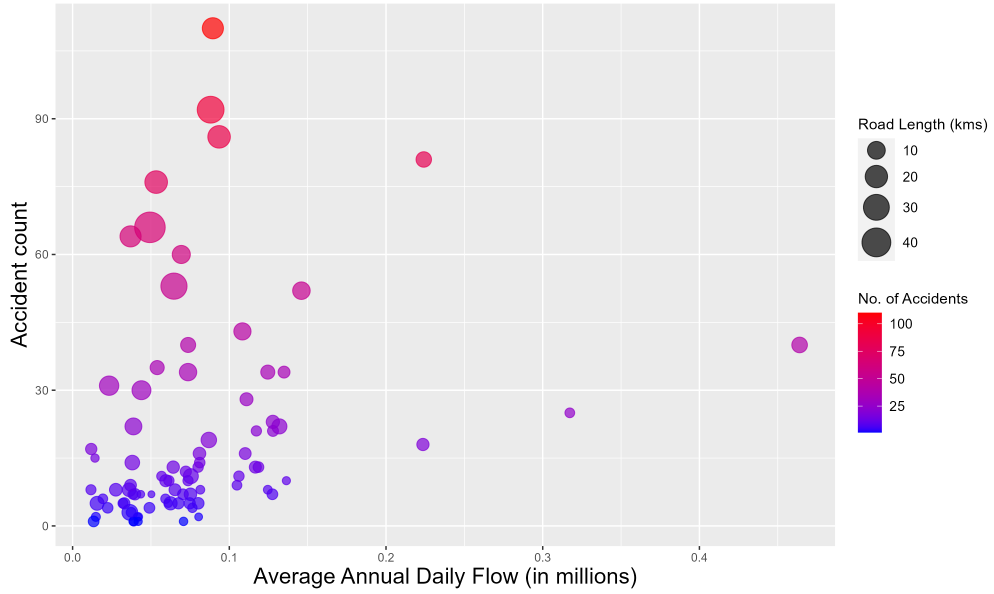

Figure 6: AADF vs Accident count by length of the road

7. Modelling accident severity

For the next part of our analysis, we will build a classification model for predicting `accident_severity` based on the geographical, temporal, and environmental factors during the collision. This will help identify the significant variables that determine the severity of an accident. Following this, we will use a technique called "Association rule mining" to uncover relationships between the important variables and the severity of the accident.

7.1 Classification Model

We will train a random forest classifier on the data using `number_of_vehicles`, `road_surface_conditions`, `road_type`, `urban_or_rural_area`, `speed_limit`, `light_conditions`, `weather_conditions`, `area`, `hour`, and `traffic_calming` as the predictor variables and `accident_severity` as the target variable.

7.1.1 Methodology

Except `traffic_calming`, all the other variables are taken from the accidents dataset. To create the `traffic_calming` variable, we extract osm data for traffic calming measures [1] using the `osm_data` package. Then, we find the intersection of all traffic calming measures in a 100 m radius of the accident site by creating buffers. Finally, we split the data into train and test samples and used a random forest classifier for modelling.

7.1.2 Results

The trained classifier achieves an accuracy of 76.87%. However, as we can see from the confusion matrix in table 2, it does not predict any fatal accidents for the test set. This is likely because of class imbalance issue (only 46 points for fatal type).

[1] we only consider traffic calming measures created by causing vertical deviation in the road

Table 2: Confusion Matrix

	Fatal	Serious	Slight
Fatal	0	2	17
Serious	0	9	209
Slight	0	22	904

Plot of feature importance[2] for the classifier is shown in figure 8.

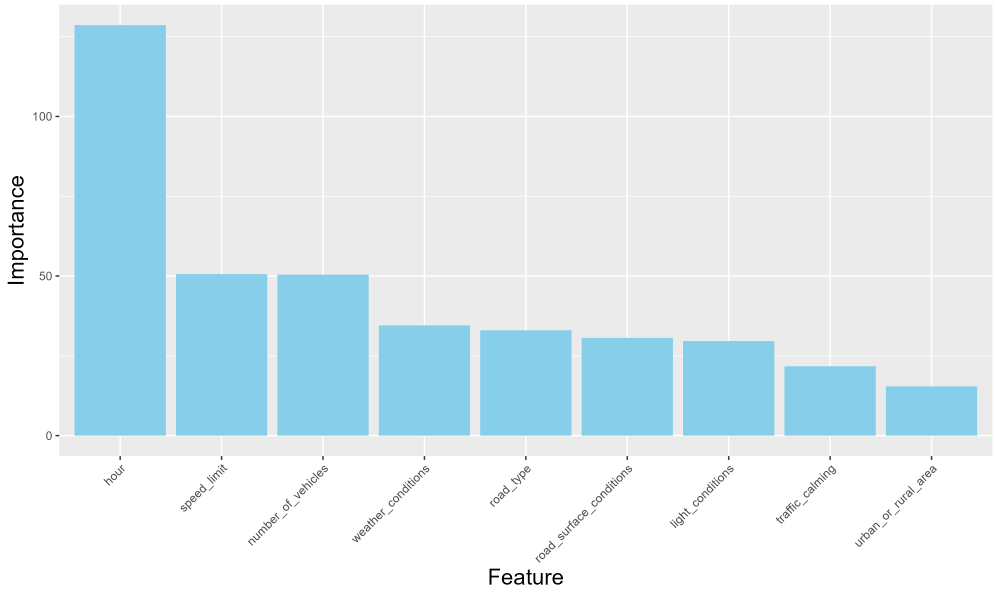

Figure 7: Feature importance of rf classifier

The hour of crash occurrence is by far the most important variable in predicting severity. Figure 9 shows the probability distribution of accident severity per hour as predicted by the model. Probability for the occurrence of serious and fatal accidents is highest at night between 11 pm to 6 am.

[2]feature importance is computed as the mean decrease in Gini index

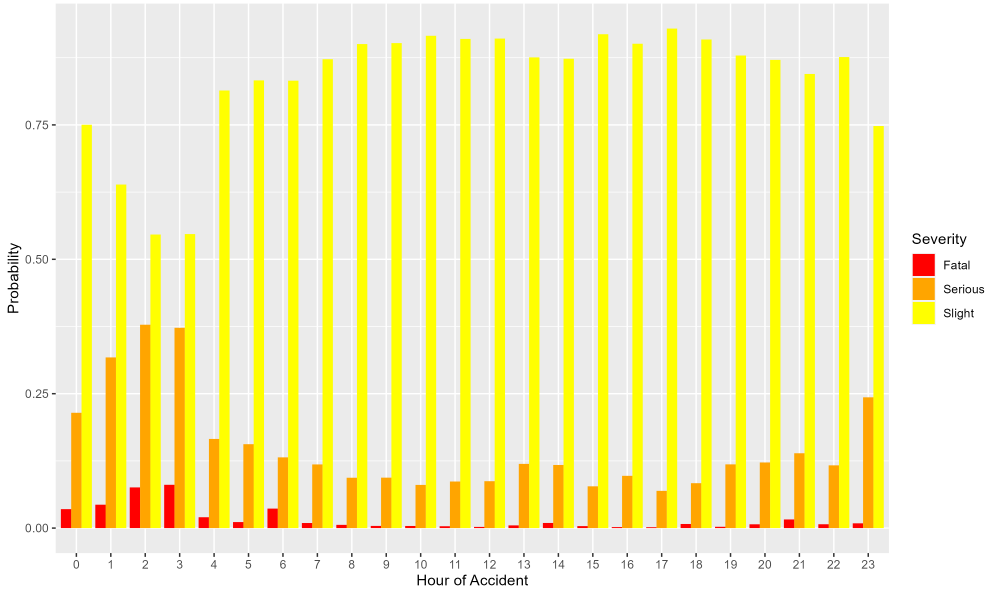

Figure 8: Distribution of predicted accident severities by hour

7.2 Association rule mining

Association rule mining is a rule based machine learning method for discovering relationships between different variables. It is primarily applied to market basket analysis and uses an "if-then" approach to uncover patterns in the data. Let "if {A}, then {B}" define a sample rule in the analysis then,

- Support: $\frac{N(A \cap B)}{N(total)}$
- Confidence: $\frac{N(A \cap B)}{N(A)}$
- Lift: $\frac{N(A \cap B)}{N(A)N(B)}$

7.2.1 Methodology

We will use the top two most important variables, `hour` and `speed_limit`. Since associate rule mining only works on categorical data, we discretise the variables into distinct categories. `hour` is divided into the categories: "morning" (5-12), "daytime" (12-19), and "night" (19-4), and `speed_limit` is divided into: "low_speed" (0-25), "medium_speed" (25-50), and "high_speed" (50-75) using appropriate threshold values. The `apriori` function of the `arules` package is used to carry out the analysis.

7.2.2 Results

Table 3: Rule mining results

rules	support	confidence	lift
{hour=morning} => {accident_severity=Slight}	0.2554180	0.8088235	1.023172
{hour=daytime} => {accident_severity=Slight}	0.3929309	0.8024236	1.015076
{speed_limit=medium speed} => {accident_severity=Slight}	0.6261610	0.8017839	1.014267
{hour=morning,speed_limit=high speed} => {accident_severity=Slight}	0.0425697	0.8048780	1.018181

rules	support	confidence	lift
{hour=morning,speed_limit=medium speed} => {accident_severity=Slight}	0.1994324	0.8128286	1.028239
{hour=daytime,speed_limit=medium speed} => {accident_severity=Slight}	0.3129515	0.8130027	1.028459

Table 3 shows the results obtained after running the `apriori` function. Only the results where `lift > 0.9` were filtered out. The statement before the $\implies$ forms the "if" part and the statement following forms the "then" part of the analysis. It can be seen that if the speed limit is "medium" (25-50 mph), then there is 62.6% support for the accident severity to be "Slight". Similarly, we can interpret the other rules and draw conclusions backed by data.

8. Policy Analysis

- The analysis carried out in the report provides an understanding of the distribution of accidents and the factors affecting accident severity. Policymakers can use it to narrow down their target areas and help in optimal resource allocation.
- Hotspot analysis in section 5 reveals that the North-Eastern boundary of Leeds is experiencing a high number of crashes. The road infrastructure and traffic management in the respective areas should be analysed in more detail to come up with effective solutions.
- To ensure the ongoing effectiveness of the safety policies implemented, it is crucial that more recent accident data for A6120 Outer Ring Road is thoroughly analysed. This will provide a clear picture of the current situation and guide future policy decisions.
- Roads with the highest accident frequency counts, as computed in section 6, should be analysed further to estimate the factors leading to more collisions.

Section 7, along with EDA plot 2, reveals that most accidents occur at night. Hence, safety policies should be more focused on enhancing road lighting, increasing night patrols, and deploying reflective markings wherever necessary.

9. Limitations and Scope for improvement

The hotspots computed in section 5 largely depended on the area under consideration. Although the data was normalised by population density, the high variation in surface area of LSOAs influenced the formation of clusters. Alternatively, MSOAs or other higher-resolution data can be chosen.

The average annual daily traffic dataset used in section 6 was only available for a few count points. This led to a 50% decrease in the data points when joining it with the crash dataset. No other datasets were available for traffic flow count by road. Also the data extracted from osm contained missing values for the names of roads which had to be removed.

Very little data (3876 data points) was available for modelling in section 7, especially for the fatal class (46 data points). Oversampling techniques from the package `ROSE` could not be applied due to the multiclass nature of the data, and `SMOTE` was not available for the current R version. No data could be fetched from the `stats19` package for the years preceding 2018.

Appendix A

Statistically significant roads in the accident frequency domain for black spot analysis

Table 4: Top 20 roads in leeds with highest accident frequency

name	number_of_accidents
Dewsbury Road	110
Bradford Road	92
York Road	86
Inner Ring Road	81
Wakefield Road	76
Leeds Road	66
Otley Road	64
Gelderd Road	60
Harrogate Road	53
Stanningley Bypass	52
Scott Hall Road	43
Aberford By-Pass	40
Easterly Road	40
Elland Road	35
Clay Pit Lane	34
Kirkstall Road	34
Selby Road	34
Whitehall Road	31
Wetherby Road	30
Broadway	28

References

Daniels, P., K. Thompson, and R. Smith. 2020. "An Integrated Approach for Studying the Safety of Road Networks." In *Proceedings of the 17th International Conference on Urban Transport and the Environment*, 553–64. https://www.witpress.com/Secure/elibrary/papers/SW10/SW10048FU1.pdf.

Department for Transport. 2019. "Road Safety Statement 2019." https://assets.publishing.service.gov.uk/media/5d2de4f1ed915d2ff003b711/road-safety-statement-2019.pdf.

GOV.UK. 2022. "Reported Road Casualties Great Britain, Annual Report: 2022." https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2022/reported-road-casualties-great-britain-annual-report-2022.

"Kernel Density Estimation and k-Means Clustering to Profile Road Accident Hotspots." 2009. *Accident Analysis & Prevention* 41 (3): 359–64. https://doi.org/https://doi.org/10.1016/j.aap.2008.12.014.

Kuyumcu, Zeliha Cagla, Hakan Aslan, and Nilufer Yurtay. 2023. "Identifying Interrelated Factors of Fatal and Injury Traffic Accidents Using Association Rules." *Turkish Journal of Civil Engineering* 34 (2): 55–79. https://doi.org/10.18400/tjce.1322965.

Kwon, Oh Hoon, Wonjong Rhee, and Yoonjin Yoon. 2015. "Application of Classification Algorithms for Analysis of Road Safety Risk Factor Dependencies." *Accident Analysis & Prevention* 75: 1–15. https://doi.org/10.1016/j.aap.2014.11.005.

Leeds City Council. 2014. "Report to the Chief Officer Highways and Transportation." https://democracy.leeds.gov.uk/documents/s121676/Report%20to%20the%20Chief%20Officer%20Highways%20and%20Transportation.pdf.

Lovelace, Robin. 2020. *Reproducible Road Safety Research with r: A Practical Introduction*. https://itsleeds.github.io/rrsrr/index.html.

Ord, J. K., and Arthur Getis. n.d. "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application." *Geographical Analysis* 27 (4): 286–306. https://doi.org/https://doi.org/10.1111/j.1538-4632.1995.tb00912.x.

Turkish General Directorate of Highways. 2001. "Black Spot Manual." https://www.kgm.gov.tr/SiteCollectionDocuments/KGMdocuments/Eng/Traffic/BlackSpotManual.pdf.

Wharfedale Observer. 2017. "Harewood's MP Cites Tragic Local Case in Push for Change to Drink-Drive Law." https://www.wharfedaleobserver.co.uk/news/15408475.harewoods-mp-cites-tragic-local-case-in-push-for-change-to-drink-drive-law/.