Identifying potential contributory factors to fatal traffic accidents using association rule mining.

Road traffic accidents are currently one of the biggest concerns for global health organisations and national governments and are currently the leading cause of death for people ages between 5 and 29 years. Understanding the factors which contribute to fatal accidents is hugely researched. Data mining techniques are a popular choice for researching this area. This case study investigated the ability of association rule mining to identify patterns among fatal accident attributes using UK traffic accident data from 2016 and 2017. The results from this investigation contradict other study findings and seem to suggest that unusual environmental conditions do not contribute to fatal accident occurrence. Fatal accidents seem to mostly occur on single carriageways, where they are not any junctions and in urban areas. It may be interesting in the future to combine environmental factors and human factors to see if their interaction produces any interesting patterns regarding fatal road traffic accidents.

1. Introduction

The increasing numbers of motor vehicles across the globe has become a major concern for health organisations and national governments. In 2018, the global number of annual road traffic deaths had reached 1.35 million, with road traffic-related injuries being the leading cause of death for people aged between 5 and 29 years (World Health Organization, 2018). The World Health Organisation has declared that drastic action is required and in 2020, they set a target of halving road traffic deaths by 2030 as part of their resolution A/RES/74/299 "Improving global road safety". They aim to achieve this through improving road and vehicle design, laws, and immediate emergency care for injured parties. Knowing what crash attributes are important in determining a fatal crash outcome could help to meet this outcome by allowing more targeted safety designs and law changes.

Montella et al. (2012) defined a crash as “a rare, random, multi-factor event always preceded by a situation in which one or more road users fail to cope with the road environment”. Research into discovering and understanding these factors preceding a crash have been studied widely for a number of years. Initially statistical models were used to classify accident severity and determine important contributors to accident severity. Some of the most utilised methods included multinomial logit, ordered logit, ordered probit, binary logit, binary probit and nested logit (reviewed in detail by Savolainen et al. (2011)). However, statistical models are limited due to pre-defined underlying relationships between the dependent variable (accident severity) and the independent variables (potential accident risks) and model assumptions; for example, assuming independent variables are independent of one another (Chang and Wang, 2006). This has led to a rise in the popularity of applying machine learning methods, as these non-parametric models do not assume any relationship between variables or distribution (Wen et al., 2021). Many machine learning methods have been used to investigate crash severity, including association rule mining, decision trees, random forest, Bayes classifier, support vector machine and artificial neural networks (reviewed by Silva et al. (2020).

Association rule mining is a non-parametric method which is able to identify patterns within large datasets and produce rules which represent potential relationships and co-occurring nature of some variables (Tan et al., 2016). This technique has been applied to road accident data to gain insights for a number of years (Pande and Abdel-Aty, 2009, Kumar and Toshniwal, 2016, Das and Sun, 2014, Xu et al., 2018). It is particularly useful for investigating relationships regarding crash severity where more severe accidents are often rare, as association rules mining also works well on small datasets (Weng et al., 2016). Although association rules do not necessarily mean causation between variables, the rules are still useful for identifying potentially important contributing factors for road accidents. This can then inform more targeted investigation into these factors and potentially inform analysis using classification techniques, such as decision trees, where removal of redundant variables can improve model performance (de Oña et al., 2013).

The aim of this study is to apply association rules mining to identify potential relationships and patterns between attributes of fatal road accidents. The data used is from the UK Department for Road Transport and includes details of UK road accidents reported by the police in the years 2016 and 2017. Two pieces of software used to develop machine learning models are R and SAS Enterprise Miner (SAS EM). R is a free open-source software for statistical computing, with a variety of add-on packages which provide many built-in functions, including datamining techniques (R Core Team, 2021). SAS enterprise miner is a software which claims to be able to “streamline the data mining process to develop models quickly” (SAS Institute Inc., 2015). This tool aims to allow non-technical users with limited statistical skill to easily generate models and results which are easy to interpret. An additional aim of this study is to compare the results generated by these two platforms.

2. Literature Review

When investigating and modelling traffic accident severity using data mining, a number of different methods have been applied. One of the earliest methods implemented was the CART model. Chang and Wang (2006) used CART to establish any relationships between accident severity and various independent variables. They identified vehicle type as being the most important variable associated with crash severity, with motorcycle riders being more likely to be seriously injured in an accident. Another study using CART found that lack of a seatbelt, improper overtaking and speeding affected injury severity in Iran (Tavakoli Kashani et al., 2011). Supporting these findings, another study identified that human factors had the highest contribution to traffic accidents (Pakgohar et al., 2011). Using other decision tree methods has also produced similar results, with seatbelt usage, collision type and use of drugs influencing injury severity in car accidents using a C5 model (Delen et al., 2017). Li et al. (2012) compared the performance of support vector machine (SVM) models against an orbit probit model in predicting injury severity. The SVM model produced better prediction results and more reasonable results for lower frequency injury classes. SVM modelling was also applied to accident severity prediction for a mountainous freeway section. Real-time traffic and weather data were incorporated into the model, which were found to have large influence on crash severity (Yu and Abdel-Aty, 2014). Another comparative study also found that SVM models performed well, along with Random Forest, when compared to Mutinomial Logit and Nearest Neighbour Classification (Iranitalab and Khattak, 2017). Applications of Bayesian Networks identified that the type of accident, the age of the driver and lighting were associated with fatal and serious injury accidents in Spain (De Oña et al., 2011). Applying ANN to crashes among elderly drivers identified the cause of collision, average annual daily traffic, number of vehicles, age, road surface condition and gender as important contributory factors (Amiri et al., 2020). Zeng and Huang (2014) identified sex, age, wearing a seatbelt or not, vehicle age, point of impact and number of heavy vehicles as influential on crash severity in Florida.

Application of association rules mining is very useful for uncovering hidden relationships between independent factors and traffic accident severity. The earliest application analysed crashes as supermarket transactions (market basket analysis) to produce a set of rules. From the produced rules it was able to be concluded that there was a correlation between lack of light and crash severity (Pande and Abdel-Aty, 2009). For severe traffic accidents in China, complex interactions between driver behaviour, vehicle type, road characteristics and environmental factors were identified (Xu et al., 2018). Using UK data, Feng et al. (2019) suggested that road type, light, speed limit and road surface are influential factors for traffic accidents after applying association rules mining. Kumar and Toshniwal (2016) used K-means clustering to group locations according to accident frequency and then applied association rules mining to reveal the factors which contribute to accident frequency. Weng et al. (2016) used association rule mining to identify factors contributing to work zone crash casualties. Montella (2011) identified investigated the contributory factors to crashes at roundabouts.

3. Methodology

3.1 Dataset acquisition and preparation

The dataset used in this investigation was obtained from the UK’s Department for Road Transport government website. The dataset includes details of road accidents collected by the police using the STAT19 reporting system between 2016 and 2020 (Department for Public Transport, 2021a). This reporting system is used to standardise the process of collecting details regarding a reported road accident involving at least one vehicle collision or a vehicle collision with a pedestrian. The initial dataset contains 597,973 recorded incidents, each with 36 attributes recorded. Each accident is given an accident severity level of either slight, severe or fatal (Department for Transport, 2021). A slight accident includes at least one person with a slight injury (an injury of minor character, such as bruises, cuts or sprains) but no further injury or death. A serious accident includes at least one person who is seriously injured (an injury requiring hospital in-patient treatment) but no one is killed. A fatal accident includes at least one death and potentially other injuries. The first step prior to modelling requires cleaning and preparation of the dataset into an appropriate format. This process was completed using R version 4.1.2 (R Core Team, 2021), which the following section will describe, and full detail of the process and code used can be found in Appendix A.

Due to the size of the dataset and lack of computing power, only the first two years (2016 and 2017) were used for this analysis, reducing the number of accidents to 266,603. Initial brief exploration of the dataset was first conducted, using functions such as ‘summary’ and ‘head’. Each accident was attributed a unique identifier (renamed to “accident\_index”) and initial checks found there was not any duplicates. With the focus being to identify causes of different accident severities, some attributes could be removed due to irrelevance [fig.1]. For example, details such as the likely road number and local highway authority would likely not contribute to an accident’s outcome.

> #remove initially selected unnecessary variables

> #remove accident reference, police force, local\_authority\_district,

> #local\_authority\_ons\_district, local\_authority\_highway, first\_road\_number,

> #second\_road\_number, lsoa\_of\_accident\_location, did\_police\_officer\_attend\_scene\_of\_accident

> accident\_df <- accident\_df %>% select(-c(accident\_reference, police\_force, local\_authority\_district, local\_authority\_ons\_district, local\_authority\_highway, first\_road\_number, second\_road\_number, lsoa\_of\_accident\_location, did\_police\_officer\_attend\_scene\_of\_accident))

+ }

+ }

**Figure 1:** R code used to remove unnecessary variables (accident\_reference, police\_force, local\_authority\_district, local\_authority\_ons\_district, local\_authority\_highway, first\_road\_number, second\_road\_number, lsoa\_of\_accident\_location, did\_police\_officer\_attend\_scene\_of\_accident).

Dealing with missing data is hugely important for data mining and modelling, as randomly missing data can interfere with model effectiveness. Although it is likely that any missing data is due to lack of information when reporting an accident, it is possible that any missing information could not be random and therefore may be informative. In this dataset, missing data was generally given either the value ‘-1’ or ‘NULL’. These were changed to ‘NA’ to allow simpler detection of missing data. The function ‘vis\_miss’ from the package ‘visdat’ (Tierney, 2017) was used to visualise the missing data [fig.2]. Fortunately, only 3.3% of data was missing altogether. The missing data does not appear to follow a pattern. The columns ‘second\_road\_class’ and ‘junction\_control’ both have around 48% missing values each, leading to the decision to remove these columns.

A picture containing table

Description automatically generated

**Figure 2:** Plot of observations for each variable where black indicates observations where data is missing (NA).

The remaining accident attributes were all in numerical form, with nominal attributes given a number (for example, in ‘accident\_severity’ 1 was assigned to fatal accidents), allowing potential correlations between variables to be investigated. Any strong correlations could interfere with modelling. However, to reduce dimensionality and simplify analysis, ‘time’ and ‘date’ were first changed into ‘hour’ and ‘month’ respectively. ‘hour’ was produced by removing the minutes from the hour, so ‘15:57’ would become ‘15’. To produce ‘month’, the strings were converted into dates (‘%d/%m/%Y’) and then reformatted to just show the month (‘%m’). Then a correlation matrix was produced using the package ‘corrplot’ (Wei and Simko, 2021) [fig.3]. There does appear to be some correlation between ‘urban\_or\_rural’ and ‘speed\_limit’ and between ‘trunk\_road\_flag’ and ‘speed\_limit’, however it was decided that this correlation was not strong enough to warrant removal of any of these attributes.

Chart, scatter chart

Description automatically generated

**Figure 3:** Correlation matrix between variables where colour indicates correlation value and size indicates strength of correlation.

The next and final stage involves visualising the distribution of the ‘accident\_severity’ values within each individual attribute. This allows investigation of potentially interesting patterns or necessary altering of attributes to improve analysis. Numerical attributes, including ‘location\_northing\_osgr’, ‘location\_easting\_osgr’, ‘latitude’, ‘longitude’, ‘hour’, ‘month’, ‘number\_of\_vehicles’ and ‘number\_of\_casualties’, were plotted as histograms [fig.4]. Figure 4A and figure 4C show that ‘location\_northing\_osgr’ and ‘latitude’ have an identical distribution and figure 4B and figure 4D show that ‘location\_easting\_osgr’ and ‘longitude’ also have an identical distribution. This makes sense as ‘location\_northing\_osgr’ and ‘location\_easting\_osgr’ are the recorded locations of the accidents according to the Ordnance Survey National Grid reference system used by the UK. Therefore, therefore accidents at the same site will have matching longitude/latitude coordinates along with matching Ordnance Survey National Grid references. The number of accidents per each hour is interesting, as there appears to be high numbers of accidents in the morning (approximately between 7:00 and 10:00) and in the evening (approximately between 15:00 and 19:00) [fig.4E]. These times coincide with rush hour where there are often higher levels of traffic on the roads, which would explain the increased likelihood of a car accident. There does not appear to be any large differences in the accident severity or frequency of accidents between the months or any obvious patterns [fig.4F]. The number of casualties [fig.4G] and the number vehicles [fig.4H] involved in accidents are both skewed to the left, with lower values having higher frequencies. There does not appear to be any interesting differences between the accident severities of these values [fig.4G and fig.4H].

Chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

**Figure 4:** Histograms for numeric variables of accidents. **A)** Northing coordinates. **B)** Easting coordinates. **C)** Latitude. **D)** Longitude. **E)** Hour. **F)** Month. **G)** Number of causalities. **F)** Number of vehicles.

As the nominal attributes were provided in numeric format, interpreting the differences between attribute variables in any plots would be more difficult, therefore it is necessary to convert the numbers into a factor level which better implies the corresponding value. For example, the value 7 in ‘weather\_conditions’ was changed to ‘FOG’. A supporting document was provided on the same website to understand each category, including the numbers substituted for categorical data (Department for Public Transport, 2021b). Each attribute received the same treatment; checking for missing values, conversion from numbers to meaningful values, conversion to a factor and then a check to ensure each value has been successfully changed. The column ‘weather\_conditions’ received further altering. Most of the weather conditions were split into ‘condition’ and ‘condition with high winds’, such as ‘rain’ and ‘rain with high winds’. It was decided to create a new column for the presence of high winds and combine all of the same weather conditions together. It could be possible that high winds could have an increasing effect on its own or have different effects according to weather combination. Figure 5 shows the shared and novel treatments for ‘weather\_conditions’ as an example. For details of conversion of the other attributes look in Appendix A.

> #weather\_conditions

> #check NA

> sum(is.na(accident\_df$weather\_conditions))

[1] 14

> #convert to NA

> accident\_df$weather\_conditions[accident\_df$weather\_conditions==9]<-NA

> accident\_df$weather\_conditions[accident\_df$weather\_conditions==8]<-NA

> #check levels

> accident\_df %>% count(weather\_conditions)

weather\_conditions n

1 1 216951

2 2 28792

3 3 987

4 4 2503

5 5 2488

6 6 178

7 7 1369

8 NA 13335

>

> #high\_winds

> accident\_df$high\_winds<-NA

>

> for(i in 1:nrow(accident\_df)) {

+ if (is.na(accident\_df$weather\_conditions[i])) {

+ accident\_df$high\_winds[i] <- NA

+ }

+ else if ((accident\_df$weather\_conditions[i]<4) || (accident\_df$weather\_conditions[i]==7)){

+ accident\_df$high\_winds[i] <- "NO"

+ }

+

+ else if ((accident\_df$weather\_conditions[i]>3) & (accident\_df$weather\_conditions[i]<7)){

+ accident\_df$high\_winds[i] <- "YES"

>

> accident\_df %>% count(high\_winds)

high\_winds n

1 NO 248099

2 YES 5169

3 <NA> 13335

>

> #combine high winds and non high-winds

> accident\_df$weather\_conditions[accident\_df$weather\_conditions==4]<-1

> accident\_df$weather\_conditions[accident\_df$weather\_conditions==5]<-2

> accident\_df$weather\_conditions[accident\_df$weather\_conditions==6]<-3

>

> #change level names

> accident\_df$weather\_conditions[accident\_df$weather\_conditions==1]<-"FINE"

> accident\_df$weather\_conditions[accident\_df$weather\_conditions==2]<-"RAIN"

> accident\_df$weather\_conditions[accident\_df$weather\_conditions==3]<-"SNOW"

> accident\_df$weather\_conditions[accident\_df$weather\_conditions==7]<-"FOG"

> #check levels

> accident\_df %>% count(weather\_conditions)

weather\_conditions n

1 FINE 219454

2 FOG 1369

3 RAIN 31280

4 SNOW 1165

5 <NA> 13335

**Figure 5:** R code for conversion of weather condition variables (‘weather\_conditions’) into more informative values and creating the ‘high\_winds’ variable.

Bar charts were produced to visualise the spread of accident severity classes within each accident attribute [fig.6, fig.7, fig.8]. Rather than use absolute frequencies, relative frequencies were used to allow comparison across characteristics, as some characteristics may be less common than others. Code for the plots was adapted from Julia Silge’s blog (Silge, 2021). Speed limit (‘speed\_limit’) was treated as a factor as, despite being an integer, speed limits for roads are discrete variables. The frequency of the different accident severities did not appear to change between the two years [fig.6A]. The day of the week did not seem to have different levels of accident severity, except maybe a slightly higher proportion of fatal and serious accidents occurring on the weekend [fig.6B]. The junction detail, road type and speed limit have seemingly visible differences when it comes to serious and slight accidents. Vehicles not at a junction seemed have a higher proportion of serious and fatal accidents whereas vehicles on slip roads comparatively had less serious accidents but similar levels of fatal accidents [fig.6D]. Single carriageways and double carriageways have comparatively high levels of fatal and serious accidents [fig.6E]. As speed limit increases, so does the amount of serious and fatal accidents, reaching a peak at 60mph and decreasing slightly at 70mph [fig.6F]. The road class does not appear to have as obvious a relationship with accident severity compared to the previous three mentioned [fig.6C]. Areas without lighting visibly appear to have a much higher proportion of fatal accidents [fig.7A]. Roads without crossing facilities altogether seems to have more fatal accidents [fig.7B and fig.7C]. There does not appear to be a large difference between the accident severity proportions according to the different weather conditions [fig.7D]. Existing accidents acting as a road hazard appear to have a higher proportion of fatalities than other hazards and lack of road sign appears to have a higher proportion of serious accidents [fig.8A and fig.8D]. Additionally, urban areas appear to have a higher proportion of serious and fatal accidents [fig.8F]. However, the presence of high winds and the presence of special conditions at the site did not appear to have different levels of accident severities [fig.B and fig.C]. It was decided to change the day of week into whether the accident took place on a weekday (‘yes’) or the weekend (‘no’) [fig.9].

**A**

Graphical user interface, bar chart

Description automatically generated

**F**

**E**

**D**

**C**

**B**

**Figure 6:** Proportion of accident severities (fatal, serious or slight) for categorical variables. **A)** accident year **B)** day of week **C)** first road class **D)** junction detail **E)** road type **F)** speed limit

**A**

**F**

**E**

**D**

**C**

**B**

Graphical user interface, bar chart

Description automatically generated

**D**

**C**

**B**

**A**

**Figure 7:** Proportion of accident severities (fatal, serious or slight) for categorical variables. **A)** light conditions **B)** human controlled pedestrian crossing facilities **C)** physical pedestrian crossing facilities **D)** weather conditions

**A**

Graphical user interface, bar chart

Description automatically generated

**F**

**D**

**B**

**E**

**C**

**Figure 8:** Proportion of accident severities (fatal, serious or slight) for categorical variables. **A)** carriageway hazards **B)** high winds **C)** road surface conditions **D)** special conditions at site **E)** trunk road **F)** urban or rural area

> #make factor

> accident\_df$weather\_conditions<-as.factor(accident\_df$weather\_conditions)

> #change to weekday or not

> accident\_df$day\_of\_week <- as.integer(accident\_df$day\_of\_week)

> accident\_df$day\_of\_week[accident\_df$day\_of\_week==1] <- "NO"

> accident\_df$day\_of\_week[accident\_df$day\_of\_week==7] <- "NO"

> accident\_df$day\_of\_week[accident\_df$day\_of\_week==2] <- "YES"

> accident\_df$day\_of\_week[accident\_df$day\_of\_week==3] <- "YES"

> accident\_df$day\_of\_week[accident\_df$day\_of\_week==4] <- "YES"

> accident\_df$day\_of\_week[accident\_df$day\_of\_week==5] <- "YES"

> accident\_df$day\_of\_week[accident\_df$day\_of\_week==6] <- "YES"

> #make factor

> accident\_df$day\_of\_week <- as.factor(accident\_df$day\_of\_week)

> #check factor levels

> accident\_df %>% count(day\_of\_week)

day\_of\_week n

1 NO 65045

**Figure 9:** R code for converting ‘day\_of\_week’ from the day of the week into whether it was a weekday (‘YES’) or the weekend (‘NO’).

When converting the categorical attributes from numerical format to more easily interpreted factors, it was realised that some of the numbers corresponded to ‘unreported’ data and so was converted into missing values (NA). To check whether this had impacted the overall percentage of missing values, the missing data for each attribute was visualised once more [fig.10]. The number of missing data overall continued to be low (2.7%). However, the number of missing values for the class of the road where the accident first occurred (‘first\_road\_class’) was now 33.59%. Due to the high number of missing values and the possibility of other attributes, such as speed limit (‘speed\_limit’), road type (‘road\_type’) and whether the road was a trunk road or not (‘trunk\_road’), together potentially providing enough information to make up for the loss of road class, it was decided to remove ‘road\_class’ from the dataset. Additionally, the year, northing and easting coordinates, latitude and longitude were removed from the dataset as these variables are not useful for the association rules mining aims of this paper. The final dataset values can be seen in table 1.

A picture containing text, screenshot, antenna, cabinet

Description automatically generated

**Figure 10:** Plot of observations for each variable where black indicates observations where data is missing (NA).

**Table 1:** Final variables used in association rules mining with their corresponding values.

|  |  |
| --- | --- |
| **Numerical Attributes** | |
| Number of vehicles |  |
| Number of casualties |  |
| Hour | 0-24 |
| Month | January-December (1-12) |
| **Categorical Attributes** | |
| Weekday | YES |
| NO |
| Road type | Dual carriageway (DC) |
| One way street (OWS) |
| Roundabout (RB) |
| Single carriageway (SC) |
| Slip road (SR) |
| Speed limit | 20 |
| 30 |
| 40 |
| 50 |
| 60 |
| 70 |
| Junction Detail | More than 4 arms (4ARM) |
| Crossroads (CR) |
| Mini roundabout (MRB) |
| Private drive (PD) |
| Roundabout (RB) |
| Slip road (SR) |
| T-junction (TJUN) |
| Not at junction (NOT) |
| Pedestrian crossing – human controlled | YES |
| NO |
| Pedestrian crossing – physical facilities | YES |
| NO |
| Light conditions | Day |
| Lights – lit (LIT) |
| Lights – not lit (UNLIT) |
| No lights (NL) |
| Weather conditions | Fine |
| Fog |
| Rain |
| Snow |
| High winds | YES |
| NO |
| Road surface conditions | Dry |
| Flood |
| Ice |
| Snow |
| Wet |
| Special conditions at site | Auto traffic signal out (AFSO) |
| Auto traffic signal part defective (ASPD) |
| Road sign/marking obscured/defective (RSO) |
| Roadworks (RW) |
| Road surface defective (RSD) |
| Oil |
| Mud |
| None |
| Carriageway hazards | Vehicle load on road (LOAD) |
| Other object on road (OBJ) |
| Previous accident (ACC) |
| Animal in carriageway (ANI) |
| Pedestrian in carriageway (PED) |
| None |
| Urban or Rural area | Urban (URB) |
| Rural (RUR) |
| Trunk road | Trunk |
| Not |

3.2 Association Rules Mining

Association rules mining is a descriptive data mining technique which can be used to discover groups of commonly co-occurring items (itemsets) within a large dataset. The rules produced from these common itemsets can be used to infer relationships between items (Tan et al., 2016). For example, when using association rule mining to investigate patterns of traffic crashes in rainy weather, Das and Sun (2014) found the rule {Alignment=Curve-Level, Lighting=Dark - No Street Lights} => {Collision Type= Single Vehicle}, which they inferred meant that single vehicle crashes may be more likely to occur on roads with curve-level alignment but without streetlights. These types of inferences can be used to help identify common characteristics of particular crashes for further investigation and help decision-makers to improve efforts in increasing road safety.

A pioneering method for discovering these association rules in large transaction datasets was developed by Agrawal et al. (1993) known as the *Apriori* algorithm. This algorithm has been used to investigate car accident characteristics in many studies and so was used for this study. Itemsets are made from the groups of items associated with a record, such as the accident characteristics of an accident in this study. These can contain zero or more items and are known as k-itemsets, with k being the number of items in each set. The first step is to consider all possible crash attributes as 1-itemsets (containing only one attribute) and the support for each of these itemsets is calculated. The support is how frequently the itemset occurs within the dataset (1). The *Apriori* algorithm uses support-based pruning to reduce the number of produced itemsets. If an itemset’s support is below the minimum support value selected by the user, it is discarded and the remaining itemsets are then used to produce 2-itemsets, working on the assumption that if an itemset is infrequent, then all its supersets must also be infrequent. This continues until no new frequent itemset combinations can be created. Rules are then extracted from the frequent itemsets (X,Y to X=>Y or Y=>X) via confidence-based pruning. Confidence is the conditional probability of the consequent given the antecedent (2). This process is similar to the generation of frequent itemsets; combinations of a frequent itemset with one consequent are obtained. For example, itemset {abc} can generate {ab=>c} or {ac=>b} or {bc=>d}. Any of these rules which are below the confidence level set by the user are discarded and any rules which contain the same consequent can also be discarded. For example, if {ac=>b} is below the confidence threshold, then {a=>cb} and {c=>ab} can also be discarded.

(1)

(2)

Support is a useful measurement for identifying potentially interesting rules, as a rule with high support is unlikely to occur by chance. However, potentially interesting rules may not necessarily appear as often due to being rarer in incidence (Tan et al., 2016). Additionally, confidence can be misleading as it only considers the support of the antecedent itemset so can be misleading (Tan et al., 2016). A better metric is lift which computes the ratio between the rule’s consequent and the support of the itemset in the rule consequent (3). A lift above 1 implies high dependency between the items. An ideal rule is one which has high support and high lift (Tan et al., 2016).

(3)

2.1 Implementation in R

The full R code for implementation can be found in Appendix B. Initial exploration of the numeric attributes (number of vehicles, number of cars, hour and month) involved creating a table with the mean, standard deviation and median for each of the accident severities [table 2]. There does not appear to be any noticeable differences between the averages for any of the accident severities. Hour has a large standard deviation for each severity, which could be because of the peak in accidents around morning rush hour and evening rush hour [fig.4E]. This could be interesting if picked up in any association rules.

**Table 2:** Summary table for numeric variables grouped by accident severity.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Accident Severity** | **Numerical Attribute** | **Count (n)** | **Mean** | **Standard Deviation** | **Median** |
| FATAL | Number of vehicles | 2624 | 1.78 | 0.99 | 2 |
| Number of casualties | 2624 | 1.71 | 1.24 | 1 |
| Hour | 2624 | 14.04 | 6.19 | 15 |
| Month | 2624 | 6.82 | 3.46 | 7 |
| SERIOUS | Number of vehicles | 35645 | 1.74 | 0.74 | 2 |
| Number of casualties | 35645 | 1.39 | 0.92 | 1 |
| Hour | 35645 | 14.60 | 5.35 | 15 |
| Month | 35645 | 6.66 | 3.36 | 7 |
| SLIGHT | Number of vehicles | 175885 | 1.88 | 0.70 | 2 |
| Number of casualties | 175885 | 1.31 | 0.73 | 1 |
| Hour | 175885 | 14.56 | 5.03 | 15 |
| Month | 175885 | 6.54 | 3.43 | 7 |
| TOTAL | Number of vehicles | 214154 | 1.86 | 0.72 | 2 |
| Number of casualties | 214154 | 1.33 | 0.77 | 1 |
| Hour | 214154 | 14.56 | 5.10 | 15 |
| Month | 214514 | 6.56 | 3.42 | 7 |

For the categorical attributes, the percentage of accidents categorised as each variable for the different accident severities was recorded [table 3]. Surprisingly, weather conditions, high winds and road surface conditions did not appear to differ largely between fatal accidents and non-fatal accidents. This is unlike previous findings which suggest weather plays an important role in the likelihood of road accidents (Das and Sun, 2014, Yu and Abdel-Aty, 2014). It does appear that more fatal accidents occur where there are not any streetlights, on roads above 60mph and not at a junction compared to non-fatal accidents. The most interesting observation is that fatal accidents appear much more likely to occur in rural areas whereas non-fatal accidents are more common in urban areas.

**Table 3:** The percentage each categorical attribute makes up of accidents for each severity and all accidents in total.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Categorical Attribute** | | **FATAL**  **(%)** | **SERIOUS**  **(%)** | **SLIGHT**  **(%)** | **TOTAL**  **(%)** |
| Weekday | YES | 68.3 | 72.9 | 76.4 | 75.7 |
| NO | 31.7 | 27.1 | 23.6 | 24.3 |
| Road type | Dual carriageway (DC) | 20.9 | 13.8 | 16.0 | 15.7 |
| One way street (OWS) | 1.37 | 2.22 | 2.41 | 2.37 |
| Roundabout (RB) | 1.56 | 4.96 | 7.34 | 6.87 |
| Single carriageway (SC) | 75.4 | 78.3 | 73.1 | 74.0 |
| Slip road (SR) | 0.724 | 0.769 | 1.19 | 1.11 |
| Speed limit | 20 | 2.29 | 5.49 | 6.22 | 6.05 |
| 30 | 35.1 | 56.9 | 62.9 | 61.6 |
| 40 | 9.79 | 8.97 | 8.59 | 8.67 |
| 50 | 8.46 | 4.74 | 4.09 | 4.25 |
| 60 | 31.9 | 17.5 | 11.2 | 12.5 |
| 70 | 12.5 | 6.39 | 7.01 | 6.97 |
| Junction Detail | More than 4 arms (4ARM) | 0.229 | 0.735 | 0.988 | 0.936 |
| Crossroads (CR) | 5.98 | 8.91 | 10.5 | 10.2 |
| Mini roundabout (MRB) | 2.67 | 0.993 | 1.39 | 1.31 |
| Private drive (PD) | 2.48 | 3.14 | 2.93 | 2.96 |
| Roundabout (RB) | 2.74 | 6.28 | 9.46 | 8.85 |
| Slip road (SR) | 1.91 | 1.23 | 1.64 | 1.58 |
| T-junction (TJUN) | 20.8 | 32.4 | 32.6 | 32.4 |
| Not at junction (NOT) | 65.6 | 46.3 | 40.5 | 41.8 |
| Pedestrian crossing – human controlled | YES | 0.381 | 0.853 | 0.998 | 0.967 |
| NO | 99.6 | 99.1 | 99.0 | 99.0 |
| Pedestrian crossing – physical facilities | YES | 14.6 | 17.9 | 19.8 | 19.4 |
| NO | 85.4 | 82.1 | 80.2 | 80.6 |
| Light conditions | Day | 58.8 | 70.3 | 74.2 | 73.3 |
| Lights – lit (LIT) | 21.2 | 22.0 | 20.7 | 20.9 |
| Lights – not lit (UNLIT) | 0.877 | 0.746 | 24.2 | 0.644 |
| No lights (NL) | 19.1 | 7.00 | 4.53 | 5.12 |
| Weather conditions | Fine | 88.1 | 87.9 | 87.1 | 87.2 |
| Fog | 0.915 | 0.609 | 0.546 | 0.561 |
| Rain | 10.6 | 11.1 | 12.0 | 11.8 |
| Snow | 0.343 | 0.339 | 0.408 | 0.396 |
| High winds | YES | 2.71 | 1.96 | 1.89 | 1.91 |
| NO | 97.3 | 98.0 | 98.1 | 98.1 |
| Road surface conditions | Dry | 70.2 | 74.3 | 74.2 | 74.1 |
| Flood | 0.114 | 0.0898 | 0.0927 | 0.0925 |
| Ice | 1.33 | 1.14 | 1.30 | 1.27 |
| Snow | 0.191 | 0.157 | 0.219 | 0.209 |
| Wet | 28.2 | 24.3 | 24.2 | 24.3 |
| Special conditions at site | Auto traffic signal out (AFSO) | 0.152 | 0.188 | 0.219 | 0.213 |
| Auto traffic signal part defective (ASPD) | 0.000 | 0.0224 | 0.0409 | 0.0374 |
| Road sign/marking obscured/defective (RSO) | 0.152 | 0.160 | 0.113 | 0.121 |
| Roadworks (RW) | 0.991 | 0.864 | 1.08 | 1.04 |
| Road surface defective (RSD) | 0.152 | 0.387 | 0.147 | 0.187 |
| Oil | 0.114 | 0.300 | 0.186 | 0.205 |
| Mud | 0.114 | 0.267 | 0.248 | 0.250 |
| None | 98.3 | 97.8 | 98.0 | 97.9 |
| Carriageway hazards | Vehicle load on road (LOAD) | 0.191 | 0.174 | 0.171 | 0.172 |
| Other object on road (OBJ) | 1.14 | 0.957 | 0.752 | 0.791 |
| Previous accident (ACC) | 0.686 | 0.146 | 0.126 | 0.136 |
| Animal in carriageway (ANI) | 0.152 | 0.424 | 0.366 | 0.373 |
| Pedestrian in carriageway (PED) | 0.152 | 0.199 | 0.185 | 0.187 |
| None | 97.7 | 98.1 | 98.4 | 98.3 |
| Urban or Rural area | Urban (URB) | 37.8 | 60.2 | 67.9 | 66.2 |
| Rural (RUR) | 62.2 | 39.8 | 32.1 | 33.8 |
| Trunk road | Trunk | 14.6 | 7.37 | 8.35 | 8.27 |
| Not | 85.4 | 92.6 | 91.6 | 91.7 |

Association rule mining was carried out using the ‘apriori’ function from the package ‘arules’ (Hahsler et al., 2022). Firstly, the algorithm was applied with the right-hand side of the produced rules set as ‘FATAL’. This was to identify any consequents which were strongly interdependent with fatal accidents. However, even without a maximum rule length and a confidence and support of 0.1, no rules were generated. The next step was to investigate the most commonly associated items and rules within a fatal-accident-only dataset. The initial dataset was filtered to contain only fatal accidents using the ‘select’ function from ‘tidyverse’ (Wickham et al., 2019), producing a dataset with 2624 rows and 19 columns. Without modifying the ‘apriori’ function, 408,201 rules were generated. To improve the rule quality produced, the minimum support was set at 0.4 and the minimum confidence set at 0.5. Maximum itemset length was set at 3 and minimum itemset length set at 2, as any longer itemsets produced a lot of repetition of rules with the same few variables. 1097 rules with a length of between 2 and 3 items were generated. It is possible that some rules could be redundant, which is where a more general rule has the same or higher confidence. To identify any redundant rules the function ‘is.redundant’ from ‘arules’ (Hahsler et al., 2022) was used and these rules were then removed. The remaining 500 rules were plotted in figure 11 using the package (Hahsler, 2017). This plot reveals that the majority of the high-lift rules had low support (<80%) and the high support rules did not have particularly high lift. It was decided to look at the high-lift and high-support rules separately in the results section.

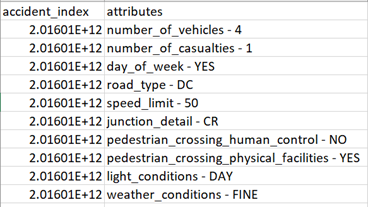
Chart, scatter chart

Description automatically generated

**Figure 11:** Scatter plot of association rules according to their confidence and support with the colour indicating lift.

3.2.2 SAS Enterprise Miner Implementation

Association rule mining in SAS Enterprise Miner requires the dataset to have each accident characteristic on a separate row (SAS Institute Inc., 2017). This required the original prepared dataset to be transformed into a longer format [fig.12]. This was achieved using R (R Core Team, 2021) as the dataset was too large to be transformed in Microsoft Excel. The function ‘pivot\_longer’ from the ‘tidyverse’ package was used to achieve this. Additionally, numeric variables are not useful in this format, so numeric attributes were converted into grouped factors, also in R. Any values of 5 and above for the number of vehicles and the number of casualties were changed to ‘5+’. The months were grouped into their corresponding seasons and the hours were grouped for every four hours (0-3, 4-7, etc.).



**A**

**B**

**Figure 12: A)** Original dataset in wide format. **B)** Transformed dataset for SAS EM into a long format.

The dataset was imported into SAS EM using the import node. For association rule mining in SAS the ‘Score’ of the data had to be set to ‘Transaction’. Additionally, an ID and a target variable are required. The ID variable is used to group each attribute to their corresponding ‘customer’ or in this case accident. The target variable contains the ‘basket item’ for each ‘customer’, or the characteristic for each accident. The target variable is used to form the rules. In this study the accident index (‘accident\_index’) was set as the ID variable and the crash attribute (‘attributes’) was set as the target variable. The association node was attached to the imported data node and similar to the implementation in R, rules containing ‘FATAL’ as the consequent were investigated first.

Similar to the implementation in R, the rules containing ‘FATAL’ as one of the consequents were filtered first using the rules selector window. This also did not generate any rules, even with the minimum confidence set as 10% and the minimum support set at 5% [fig.13]. A dataset containing only fatal accidents was then generated in a similar fashion to the full pivoted dataset used in SAS.

Graphical user interface, text

Description automatically generated

**Figure 13:** No rules were generated with Fatal as the consequent in SAS EM.

As in the R implementation, the minimum confidence was set at 50% and the minimum support was 40%. The maximum number of items was set as 3 and 200 rules were retained (those high in lift or high in support). Figure 14 shows a similar pattern to the plot of the rules produced using R, with high support rules also having high confidence values. This is reassuring that the rules produced can be seen as reliable and useful in their application across many accidents. However, similar to the rules produced by R, rules with high support appear to have low-lift values and vice versa for high-lift rules [fig.15]. High-lift and high-support rules produced by SAS EM will also be looked at separately in the results section. The full pathway used for association rule mining in SAS EM can be seen in figure 16.

Chart, scatter chart

Description automatically generated

**Figure 14:** Scatterplot of the confidence (%) and support (%) of each rule identified in SAS EM.

Chart, scatter chart

Description automatically generated

**Figure 15:** Scatterplot of the lift and support (%) of each rule identified in SAS EM with the colour of the points corresponding to the confidence (%) associated with the rule.

Graphical user interface, application

Description automatically generated

**Figure 16:** Diagram of SAS EM workflow for association rule mining.

4. Results and discussion

Although association rule mining cannot prove causation between the antecedent and consequent of a rule, it does imply strong co-occurrence of items within rules. Both SAS EM and R seem to have produced rules which either had high lift values or high support values and the items used in these rules are not necessarily the same [fig.17&18&19]. Figure 18 and figure 19 are a good visualisation of this, with the figure 18 red data points in different areas of the matrix compared to figure 19, showing different variables as valuable in SAS EM. Figure 17 also displays this for the rules produced by R, with some antecedents having predominantly small dark red points (high lift, low support) and others having predominantly pale large points (low lift, high support). Therefore, high lift and high support rules were investigated separately.

Chart

Description automatically generated

**Figure 17:** Grouped matrix plot of R rules.

Graphical user interface, application, table, Excel

Description automatically generated

**Figure 18:** Rule matrix plotting items from left hand of rule against right hand of rule with colour indicating confidence (%) for rules retained with high support.

Chart, scatter chart

Description automatically generated

**Figure 19:** Rule matrix plotting items from left hand of rule against right hand of rule with colour indicating confidence (%) for rules retained with high lift.

**Rules with High Support**

The 20 rules with the highest support are displayed in descending order for both SAS EM and R [table 4&5]. The support for the top 20 rules produced by R ranges from 0.979 to 0.867 and the support for those produced by SAS EM ranges from 97.9% to 95.3%. Additionally, the exact rules produced by each software are different with only the first three rules being identical. However, both sets of rules contain essentially the same variables in different rule combinations; {special\_conditions\_at\_site=NONE}, {pedestrian\_crossing\_human\_control=NO}, {carriageway\_hazards=NONE}, and {high\_winds=NO}. The only additional variable produced by R is {weather\_conditions=FINE} at the end. This is likely because the implementation in R involved removing redundant rules and so some rules produced by SAS EM were likely removed. These variables are combined to form different rules and so looking at individual rules to understand the specific combination of factors contributing to fatal traffic accidents is not useful. Instead, it would be more useful to consider the whole scenario these variables generate in regard to the environment fatal accidents appear to be more likely to occur in. These variables, however, do not portray a particular environment, as weather conditions are pleasant and there are not any defining characteristics of the road itself where the accident occurred.

**Table 4:** Top 20 rules ordered by support produced by R.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rule** | **Left Hand of Rule** | **Right Hand of Rule** | **Support** | **Confidence** | **Lift** | **Count** |
| 1 | {special\_conditions\_at\_site= NONE} | {pedestrian\_crossing\_human\_control=NO} | 0.979 | 0.996 | 1.00 | 2570 |
| 2 | {pedestrian\_crossing\_human\_control=NO} | {special\_conditions\_at\_site =NONE} | 0.979 | 0.983 | 1.00 | 2570 |
| 3 | {carriageway\_hazards=NONE} | {pedestrian\_crossing\_human\_control=NO} | 0.973 | 0.996 | 1.00 | 2553 |
| 4 | {pedestrian\_crossing\_human\_control=NO} | {carriageway\_hazards=NONE} | 0.973 | 0.977 | 1.00 | 2553 |
| 5 | {high\_winds=NO} | {pedestrian\_crossing\_human\_control=NO} | 0.969 | 0.996 | 1.00 | 2543 |
| 6 | {pedestrian\_crossing\_human\_control=NO} | {high\_winds=NO} | 0.969 | 0.973 | 1.00 | 2543 |
| 7 | {carriageway\_hazards=NONE} | {special\_conditions\_at\_site= NONE} | 0.963 | 0.986 | 1.00 | 2527 |
| 8 | {special\_conditions\_at\_site= NONE} | {carriageway\_hazards=NONE} | 0.963 | 0.979 | 1.00 | 2527 |
| 9 | {high\_winds=NO} | {special\_conditions\_at\_site= NONE} | 0.957 | 0.983 | 1.00 | 2510 |
| 10 | {special\_conditions\_at\_site= NONE} | {high\_winds=NO} | 0.957 | 0.973 | 1.00 | 2510 |
| 11 | {high\_winds=NO} | {carriageway\_hazards=NONE} | 0.951 | 0.978 | 1.00 | 2496 |
| 12 | {carriageway\_hazards=NONE} | {high\_winds=NO} | 0.951 | 0.984 | 1.00 | 2496 |
| 13 | {carriageway\_hazards=NONE,  high\_winds=NO} | {special\_conditions\_at\_site= NONE} | 0.938 | 0.986 | 1.00 | 2461 |
| 14 | {special\_conditions\_at\_site= NONE, high\_winds=NO} | {carriageway\_hazards=NONE} | 0.938 | 0.980 | 1.00 | 2461 |
| 15 | {special\_conditions\_at\_site= NONE, carriageway\_hazards=NONE} | {high\_winds=NO} | 0.938 | 0.974 | 1.00 | 2461 |
| 16 | {weather\_conditions=FINE} | {pedestrian\_crossing\_human\_control=NO} | 0.878 | 0.881 | 0.999 | 2303 |
| 17 | {pedestrian\_crossing\_human\_control=NO} | {weather\_conditions=FINE} | 0.878 | 0.881 | 0.999 | 2303 |
| 18 | {weather\_conditions=FINE} | {high\_winds=NO} | 0.867 | 0.881 | 1.01 | 2276 |
| 19 | {high\_winds=NO} | {weather\_conditions=FINE} | 0.867 | 0.892 | 1.01 | 2276 |
| 20 | {weather\_conditions=FINE} | {special\_conditions\_at\_site= NONE} | 0.867 | 0.984 | 1.00 | 2275 |

**Table 5:** Top 20 rules ordered by support produced by SAS EM.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rule** | **Left Hand of Rule** | **Right Hand of Rule** | **Confidence**  **(%)** | **Support**  **(%)** | **Lift** | **Count** |
| 1 | special\_conditions\_at\_site  - NONE | pedestrian\_crossing\_human\_ control - NO | 99.6 | 97.9 | 1.00 | 2570 |
| 2 | pedestrian\_crossing\_human\_ control - NO | special\_conditions\_at\_site  - NONE | 98.3 | 97.9 | 1.00 | 2570 |
| 3 | pedestrian\_crossing\_human\_ control - NO | carriageway\_hazards - NONE | 97.7 | 97.3 | 1.00 | 2553 |
| 4 | carriageway\_hazards  - NONE | pedestrian\_crossing\_human\_ control - NO | 99.6 | 97.3 | 1.00 | 2553 |
| 5 | pedestrian\_crossing\_human\_ control - NO | high\_winds - NO | 97.3 | 96.9 | 1.00 | 2543 |
| 6 | high\_winds - NO | pedestrian\_crossing\_human\_ control - NO | 99.6 | 96.9 | 1.00 | 2543 |
| 7 | special\_conditions\_at\_site  - NONE | carriageway\_hazards - NONE | 97.9 | 96.3 | 1.00 | 2527 |
| 8 | carriageway\_hazards - NONE | special\_conditions\_at\_site  - NONE | 98.6 | 96.3 | 1.00 | 2527 |
| 9 | special\_conditions\_at\_site  - NONE & pedestrian\_crossing\_human\_ control - NO | carriageway\_hazards - NONE | 97.9 | 95.9 | 1.00 | 2517 |
| 10 | special\_conditions\_at\_site  - NONE &  carriageway\_hazards - NONE | pedestrian\_crossing\_human\_ control - NO | 99.6 | 95.9 | 1.00 | 2517 |
| 11 | special\_conditions\_at\_site  - NONE | pedestrian\_crossing\_human\_ control - NO &  carriageway\_hazards - NONE | 97.6 | 95.9 | 1.00 | 2517 |
| 12 | pedestrian\_crossing\_human\_ control - NO &  carriageway\_hazards - NONE | special\_conditions\_at\_site - NONE | 98.6 | 95.9 | 1.00 | 2517 |
| 13 | pedestrian\_crossing\_human\_ control - NO | special\_conditions\_at\_site  - NONE & carriageway\_hazards - NONE | 96.3 | 95.9 | 1.00 | 2517 |
| 14 | carriageway\_hazards - NONE | special\_conditions\_at\_site  - NONE & pedestrian\_crossing\_human\_ control - NO | 98.2 | 95.9 | 1.00 | 2517 |
| 15 | special\_conditions\_at\_site - NONE | high\_winds - NO | 97.3 | 95.7 | 1.00 | 2510 |
| 16 | high\_winds - NO | special\_conditions\_at\_site  - NONE | 98.3 | 95.7 | 1.00 | 2510 |
| 17 | special\_conditions\_at\_site  - NONE & pedestrian\_crossing\_human\_ control - NO | high\_winds - NO | 97.3 | 95.3 | 1.00 | 2500 |
| 18 | special\_conditions\_at\_site  - NONE & high\_winds - NO | pedestrian\_crossing\_human\_ control - NO | 99.6 | 95.3 | 1.00 | 2500 |
| 19 | special\_conditions\_at\_site  - NONE | pedestrian\_crossing\_human\_ control - NO & high\_winds -NO | 96.9 | 95.3 | 1.00 | 2500 |
| 20 | pedestrian\_crossing\_human\_ control - NO & high\_winds - NO | special\_conditions\_at\_site  - NONE | 98.3 | 95.3 | 1.00 | 2500 |

**Rules with high lift**

The 20 rules with the highest lift are displayed in descending order for both SAS EM and R [table 6&7]. The lift for the top 20 rules produced by R ranges from 1.25 to 0.1.14 and the lift for those produced by SAS EM ranges from 1.25 to 1.17. This is also likely due to the R implementation removing redundant rules. The order of the rules produced for each software does not match, however similar to the high support rules, the same variables are used; {junction\_detail=NOT}, {urban\_or\_rural\_area=RUR}, {weather\_conditions=FINE}, {road\_surface\_conditions=DRY}, {special\_conditions\_at\_site=NONE}, {trunk\_road\_flag=NO}, {pedestrian\_crossing\_physical\_facilities=NO}, {light\_conditions=DAY}, carriageway\_hazards=NONE}, {road\_type=SC}, and {pedestrian\_crossing\_human\_control=NO}. Similar to the rules with high support, it would be more useful to use these factors to create an overview of the environment fatal accidents appear to occur in, rather than considering individual rules. The variables appear to suggest that fatal accidents are likely to occur on long (single carriage) rural roads during the daytime. This reflects initial findings during exploration of the dataset where rural environments appeared to have a higher proportion of fatal accidents compared to urban roads. Unfortunately, although above one, the lift for these rules is not particularly high and so it is likely that this is not the whole story for fatal accidents.

**Table 6:** Top 20 rules ordered by lift produced by R.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rule** | **Left Hand of Rule** | **Right Hand of Rule** | **Support** | **Confidence** | **Lift** | **Count** |
| 1 | {junction\_detail=NOT, pedestrian\_crossing\_ physical\_facilities=NO} | {urban\_or\_rural\_area=RUR} | 0.473 | 0.776 | 1.25 | 1240 |
| 2 | {light\_conditions=DAY, weather\_conditions=FINE} | {road\_surface\_conditions =DRY} | 0.470 | 0.870 | 1.24 | 1232 |
| 3 | {urban\_or\_rural\_area=RUR, trunk\_road\_flag=NOT} | {road\_type=SC} | 0.440 | 0.899 | 1.19 | 1154 |
| 4 | {pedestrian\_crossing\_ physical\_facilities=NO, urban\_or\_rural\_area=RUR } | {junction\_detail=NOT} | 0.473 | 0.780 | 1.19 | 1240 |
| 5 | {junction\_detail=NOT,  pedestrian\_crossing\_human\_ control=NO} | {urban\_or\_rural\_area=RUR} | 0.479 | 0.731 | 1.18 | 1257 |
| 6 | {special\_conditions\_at\_site =NONE,  urban\_or\_rural\_area=RUR} | {junction\_detail=NOT} | 0.471 | 0.771 | 1.17 | 1236 |
| 7 | {urban\_or\_rural\_area=RUR} | {junction\_detail=NOT} | 0.479 | 0.770 | 1.17 | 1257 |
| 8 | {junction\_detail=NOT} | {urban\_or\_rural\_area=RUR} | 0.479 | 0.730 | 1.17 | 1257 |
| 9 | {road\_surface\_conditions=DRY, trunk\_road\_flag=NO} | {light\_conditions=DAY} | 0.415 | 0.690 | 1.17 | 1089 |
| 10 | {junction\_detail=NOT, urban\_or\_rural\_area=RUR} | {pedestrian\_crossing\_ physical\_facilities=NO} | 0.472 | 0.986 | 1.15 | 1240 |
| 11 | {pedestrian\_crossing\_ physical\_facilities=NO,  trunk\_road\_flag=NOT} | {road\_type=SC} | 0.620 | 0.870 | 1.15 | 1627 |
| 12 | {junction\_detail=NOT, trunk\_road\_flag=NOT} | {road\_type=SC} | 0.468 | 0.870 | 1.15 | 1227 |
| 13 | {light\_conditions=DAY, trunk\_road\_flag=NOT} | {road\_surface\_conditions =DRY) | 0.464 | 0.808 | 1.15 | 1218 |
| 14 | {road\_surface\_conditions=DRY, carriageway\_hazards=NONE} | {light\_conditions=DAY} | 0.464 | 0.674 | 1.15 | 1217 |
| 15 | {light\_conditions=DAY, carriageway\_hazards=NONE} | {road\_surface\_conditions= DRY} | 0.464 | 0.802 | 1.14 | 1217 |
| 16 | {light\_conditions=DAY, trunk\_road\_flag=NOT} | {road\_surface\_conditions=DRY} | 0.415 | 0.802 | 1.42 | 1089 |
| 17 | {road\_type=SC,  urban\_or\_rural\_area=RUR} | {pedestrian\_crossing\_ physical\_facilities=NO} | 0.460 | 0.976 | 1.14 | 1208 |
| 18 | {weather\_conditions=FINE, road\_surface\_conditions=DRY} | {light\_conditions=DAY} | 0.470 | 0.671 | 1.14 | 1232 |
| 19 | {road\_surface\_conditions=DRY,  urban\_or\_rural\_area=RUR} | {pedestrian\_crossing\_ physical\_facilities=NO} | 0.418 | 0.974 | 1.14 | 1096 |
| 20 | {carriageway\_hazards=NONE,  urban\_or\_rural\_area=RUR} | {pedestrian\_crossing\_ physical\_facilities=NO} | 0.589 | 0.974 | 1.14 | 1546 |

**Table 7:** Top 20 rules ordered by lift produced by SAS EM.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rule** | **Left Hand of Rule** | **Right Hand of Rule** | **Confidence (%)** | **Support (%)** | **Lift** | **Count** |
| 1 | urban\_or\_rural\_area - RUR | pedestrian\_crossing\_physical\_ facilities - NO & junction\_detail - NOT | 76.0 | 47.3 | 1.25 | 1240 |
| 2 | pedestrian\_crossing\_physical\_ facilities - NO & junction\_detail - NOT | urban\_or\_rural\_area - RUR | 77.6 | 47.3 | 1.25 | 1240 |
| 3 | weather\_conditions - FINE & light\_conditions - DAY | road\_surface\_conditions - DRY | 87.0 | 47.0 | 1.24 | 1232 |
| 4 | road\_surface\_conditions – DRY | weather\_conditions - FINE & light\_conditions - DAY | 66.9 | 47.0 | 1.24 | 1232 |
| 5 | urban\_or\_rural\_area - RUR & trunk\_road\_flag – NOT | road\_type - SC | 89.9 | 44.0 | 1.19 | 1154 |
| 6 | road\_type – SC | urban\_or\_rural\_area - RUR & trunk\_road\_flag - NOT | 58.3 | 44.0 | 1.19 | 1154 |
| 7 | urban\_or\_rural\_area - RUR & pedestrian\_crossing\_physical\_ facilities - NO | junction\_detail - NOT | 78.0 | 47.3 | 1.19 | 1240 |
| 8 | junction\_detail - NOT | urban\_or\_rural\_area - RUR & pedestrian\_crossing\_physical\_ facilities - NO | 72.0 | 47.3 | 1.19 | 1240 |
| 9 | urban\_or\_rural\_area - RUR | pedestrian\_crossing\_human\_ control - NO & junction\_detail - NOT | 77.0 | 47.9 | 1.18 | 1257 |
| 10 | pedestrian\_crossing\_human\_ control - NO & junction\_detail - NOT | urban\_or\_rural\_area - RUR | 73.1 | 47.9 | 1.18 | 1257 |
| 11 | urban\_or\_rural\_area - RUR & special\_conditions\_at\_site - NONE | junction\_detail - NOT | 77.1 | 47.1 | 1.17 | 1236 |
| 12 | junction\_detail - NOT | urban\_or\_rural\_area - RUR & special\_conditions\_at\_site - NONE | 71.8 | 47.1 | 1.17 | 1236 |
| 13 | urban\_or\_rural\_area - RUR | junction\_detail - NOT | 77.0 | 47.9 | 1.17 | 1257 |
| 14 | urban\_or\_rural\_area - RUR & pedestrian\_crossing\_human\_ control - NO | junction\_detail - NOT | 77.0 | 47.9 | 1.17 | 1257 |
| 15 | junction\_detail - NOT | urban\_or\_rural\_area - RUR | 73.0 | 47.9 | 1.17 | 1257 |
| 16 | junction\_detail - NOT | urban\_or\_rural\_area - RUR & pedestrian\_crossing\_human\_ control - NO | 73.0 | 47.9 | 1.17 | 1257 |
| 17 | urban\_or\_rural\_area - RUR | junction\_detail - NOT &  high\_winds - NO | 74.6 | 46.4 | 1.17 | 1218 |
| 18 | junction\_detail - NOT &  high\_winds - NO | urban\_or\_rural\_area - RUR | 72.9 | 46.4 | 1.17 | 1218 |
| 19 | urban\_or\_rural\_area - RUR & high\_winds - NO | junction\_detail - NOT | 76.9 | 46.4 | 1.17 | 1218 |
| 20 | junction\_detail - NOT | urban\_or\_rural\_area - RUR & high\_winds - NO | 70.7 | 46.4 | 1.17 | 1218 |

The rules were plotted using a network visualisation in both SAS EM and R. Both support the variables picked up with high support are the most influential for contributing to fatal accidents out of all variables included. However, figure 20 and figure 21 further emphasises previous conclusions that the variables picked up by association rules mining are likely not the whole story for fatal accidents and it is likely that there is a large variation in the combination of factors associated with the fatal accidents in this dataset.

A picture containing map

Description automatically generated

**Figure 20:** Network visualisation of rules produced by R with the size of the point indicating support and the colour representing the lift value. Connectors indicate how variables interact with one another.

Chart

Description automatically generated

**Figure 21:** Network visualisation of rules produced by SAS EM with the size of the points indicating individual variables. Connectors indicate how variables interact with one another to form rules.

5. Conclusion

In some ways the results produced by the models are surprising, as previous research has identified that poor lighting and bad weather has been associated with traffic accident severity (Yu and Abdel-Aty, 2014, De Oña et al., 2011). This research seems to suggest that unusual environmental factors play relatively little role in contributing to fatal accidents. Other research has identified that human factors have a larger influence on accident severity outcome than environmental factors, including intoxication, seatbelt usage, vehicle type and vehicle manoeuvring (Delen et al., 2017, De Oña et al., 2011, Tavakoli Kashani et al., 2011, Pakgohar et al., 2011). The variables used in this analysis were predominantly environmental. In future research, it may be more useful to include a mixture of driver details and environmental details to see if a combination of the two produces more meaningful results. Additionally, other researchers have combined association rules mining with K-means clustering to group locations which produced more meaningful results (Kumar and Toshniwal, 2016). Research in accidents across different countries and locations has yielded varying results (Yu and Abdel-Aty, 2014, Tavakoli Kashani et al., 2011, Xu et al., 2018, Feng et al., 2019), so perhaps different locations within a country influence accident severity in different ways.

6. References

AGRAWAL, R., IMIELIŃSKI, T. & SWAMI, A. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 1993. 207-216.

CHANG, L.-Y. & WANG, H.-W. 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention,* 38**,** 1019-1027.

DAS, S. & SUN, X. Investigating the pattern of traffic crashes under rainy weather by association rules in data mining. Transportation Research Board 93rd Annual Meeting, 2014. Transportation Research Board Washington, DC.

DE OÑA, J., LÓPEZ, G. & ABELLÁN, J. 2013. Extracting decision rules from police accident reports through decision trees. *Accident Analysis & Prevention,* 50**,** 1151-1160.

DE OÑA, J., MUJALLI, R. O. & CALVO, F. J. 2011. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis & Prevention,* 43**,** 402-411.

DELEN, D., TOMAK, L., TOPUZ, K. & ERYARSOY, E. 2017. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health,* 4**,** 118-131.

DEPARTMENT FOR PUBLIC TRANSPORT 2021a. Road Safety - Accidents last 5 years.

DEPARTMENT FOR PUBLIC TRANSPORT 2021b. Road Safety Open Dataset Data Guide.

DEPARTMENT FOR TRANSPORT 2021. Reported road casualties in Great Britain, provisional estimates: year ending June 2021.

FENG, M., ZHENG, J., REN, J. & XI, Y. Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK. International Conference on Brain Inspired Cognitive Systems, 2019. Springer, 520-529.

HAHSLER, M. 2017. arulesViz: Interactive Visualization of Association Rules with R. *\_R Journal\_,* 9**,** 163-175.

HAHSLER, M., BUCHTA, C., GRUEN, B. & HORNIK, K. 2022. arules: Mining Association Rules and Frequent Itemsets. R package version 1.7-3.

IRANITALAB, A. & KHATTAK, A. 2017. Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention,* 108**,** 27-36.

KUMAR, S. & TOSHNIWAL, D. 2016. A data mining approach to characterize road accident locations. *Journal of Modern Transportation,* 24**,** 62-72.

LI, Z., LIU, P., WANG, W. & XU, C. 2012. Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention,* 45**,** 478-486.

MONTELLA, A. 2011. Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. *Accident Analysis & Prevention,* 43**,** 1451-1463.

MONTELLA, A., ARIA, M., D’AMBROSIO, A. & MAURIELLO, F. 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident Analysis & Prevention,* 49**,** 58-72.

PAKGOHAR, A., TABRIZI, R. S., KHALILI, M. & ESMAEILI, A. 2011. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. *Procedia Computer Science,* 3**,** 764-769.

PANDE, A. & ABDEL-ATY, M. 2009. Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. *Safety Science,* 47**,** 145-154.

R CORE TEAM 2021. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

SAS INSTITUTE INC. 2015. SAS® Enterprise Miner™ 14.1: Administration and

Configuration. Cary, NC: SAS Institute Inc.

SAS INSTITUTE INC. 2017. *SAS® Enterprise Miner™ 14.3: Reference Help,* Cary, NC, SAS Institute Inc.

SAVOLAINEN, P. T., MANNERING, F. L., LORD, D. & QUDDUS, M. A. 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Accident Analysis & Prevention,* 43**,** 1666-1676.

SILGE, J. 2021. Class imbalance and classification metrics with aircraft wildlife strikes. *Julia Silge* [Online]. Available from: https://juliasilge.com/blog/sliced-aircraft/.

SILVA, P. B., ANDRADE, M. & FERREIRA, S. 2020. Machine learning applied to road safety modeling: A systematic literature review. *Journal of Traffic and Transportation Engineering (English Edition),* 7**,** 775-790.

TAN, P.-N., STEINBACH, M. & KUMAR, V. 2016. *Introduction to data mining*, Pearson Education India.

TAVAKOLI KASHANI, A., SHARIAT-MOHAYMANY, A. & RANJBARI, A. 2011. A data mining approach to identify key factors of traffic injury severity. *PROMET-Traffic&Transportation,* 23**,** 11-17.

TIERNEY, N. 2017. visdat: Visualising Whole Data Frames. *Journal of Open Source Software,* 2**,** 355.

WEI, T. & SIMKO, V. 2021. R package 'corrplot': Visualization of a Correlation Matrix.

WEN, X., XIE, Y., JIANG, L., PU, Z. & GE, T. 2021. Applications of machine learning methods in traffic crash severity modelling: current status and future directions. *Transport reviews,* 41**,** 855-879.

WENG, J., ZHU, J.-Z., YAN, X. & LIU, Z. 2016. Investigation of work zone crash casualty patterns using association rules. *Accident Analysis & Prevention,* 92**,** 43-52.

WICKHAM, H., AVERICK, M., BRYAN, J., CHANG, W., D'AGOSTINO MCGOWAN , L., FRANÇOIS, R., GROLEMUND, G., HAYES, A., HENRY, L., HESTER, J., KUHN, M., LIN PEDERSEN, T., MILLER, E., MILTON BACHE, S., MÜLLER, K., OOMS, J., DAVID, R., SEIDEL, D. P., SPINU, V., TAKAHASHI, K., VAUGHAN, D., WILKE, C., WOO, K. & YUTANI, H. 2019. Welcome to the {tidyverse}. *Journal of Open Source Software,* 4**,** 1686.

WORLD HEALTH ORGANIZATION 2018. Global status report on road safety 2018. .

XU, C., BAO, J., WANG, C. & LIU, P. 2018. Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China. *Journal of safety research,* 67**,** 65-75.

YU, R. & ABDEL-ATY, M. 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Safety science,* 63**,** 50-56.

7. Appendix

**Appendix A**

R code used for data preparation and exploration.

#DATA EXPLORATION & PREP

library(tidyverse)

library(ggplot2)

library(Hmisc)

library(visdat)

rm(list=ls())

#set working directory

setwd("C:\\Users\\smann\\OneDrive\\University\\Data Science\\ASDM\\Project\\Road safety")

#ACCIDENT DATASET

df <- read.csv("dft-road-casualty-statistics-accident-last-5-years.csv")

#exploration

head(accident\_df)

tail(accident\_df)

summary(accident\_df)

str(accident\_df)

names(accident\_df)

dim(accident\_df)

#dataset large so selecting specific years to reduce size - selected 2016-2018

accident\_df <- df %>% filter(accident\_year<2018)

#change ï..accident\_index to accident\_index

accident\_df <- accident\_df %>% rename(accident\_index=ï..accident\_index)

#check for duplicated data - no duplicates

table(duplicated(accident\_df$accident\_index))

#remove initally selected unnecessary variables

#remove accident reference, police force, local\_authority\_district,

#local\_authority\_ons\_district, local\_authority\_highway, first\_road\_number,

#second\_road\_number, lsoa\_of\_accident\_location, did\_police\_officer\_attend\_scene\_of\_accident

accident\_df <- accident\_df %>% select(-c(accident\_reference, police\_force, local\_authority\_district,

local\_authority\_ons\_district, local\_authority\_highway, first\_road\_number,

second\_road\_number, lsoa\_of\_accident\_location, did\_police\_officer\_attend\_scene\_of\_accident))

names(accident\_df)

#NA variables are indicated as -1 or NULL for some variables- making NA

accident\_df[accident\_df==-1]<- NA

accident\_df[is.null(accident\_df)]<- NA

#visualising missing data

vis\_miss(accident\_df, warn\_large\_data = FALSE)

colSums(is.na(accident\_df))

# remove second\_road\_class and junction\_contol

accident\_df <- accident\_df %>% select(-c(second\_road\_class, junction\_control))

names(accident\_df)

#CONVERT TIME TO JUST HOUR

accident\_df$hour<- gsub("\\:.\*","",accident\_df$time)

#make integer

accident\_df$hour<- as.integer(accident\_df$hour)

#CONVERT DATE TO MONTH

#convert to date

accident\_df$date <- as.Date(accident\_df$date, format = "%d/%m/%Y")

#extract month and put in new variable month

accident\_df$month <- format(accident\_df$date, format = "%m")

#check

accident\_df %>% count(month)

#make integer

accident\_df$month<- as.integer(accident\_df$month)

#make speed limit integer

accident\_df$speed\_limit<- as.integer(accident\_df$speed\_limit)

accident\_df$longitude <- as.numeric(accident\_df$longitude)

accident\_df$latitude <- as.numeric(accident\_df$latitude)

accident\_df$location\_northing\_osgr <- as.numeric(accident\_df$location\_northing\_osgr)

accident\_df$location\_easting\_osgr <- as.numeric(accident\_df$location\_easting\_osgr)

accident\_df$number\_of\_vehicles <- as.numeric(accident\_df$number\_of\_vehicles)

accident\_df$number\_of\_casualties <- as.numeric(accident\_df$number\_of\_casualties)

str(accident\_df)

###CORRELATION PLOT

library(corrplot)

accident\_df.corr <- accident\_df %>%

select(-c(accident\_index, date, time)) %>% na.omit() %>% cor()

corrplot(accident\_df.corr)

heatmap(x = accident\_df.corr, symm = TRUE)

# speed limit and urban/rural area seem correlated

###VISUALISE NUMERIC DATA

accident\_df$accident\_severity <- as.factor(accident\_df$accident\_severity)

histogram <- function(attribute) {

accident\_df %>% ggplot(aes(x= attribute,fill = accident\_severity)) +

geom\_histogram(bins = 30L) +

scale\_fill\_hue(direction = 1) +

theme\_minimal()

}

nos <- histogram(accident\_df$location\_northing\_osgr) + labs(x = "locaton\_northing\_osgr")

eos <- histogram(accident\_df$location\_easting\_osgr) + labs(x = "locaton\_easting\_osgr")

long<- histogram(accident\_df$longitude) + labs(x = "longitude")

lat <- histogram(accident\_df$latitude) + labs(x = "latitude")

hour <- histogram(accident\_df$hour) + labs(x = "hour")

mon <- histogram(accident\_df$month) + labs(x = "month")

cas <- histogram(accident\_df$number\_of\_casualties) + labs(x = "number\_of\_casualties")

veh <- histogram(accident\_df$number\_of\_vehicles) + labs(x = "number\_of\_vehicles")

library(ggpubr)

ggarrange(nos, eos, lat, long,

labels = c("A", "B", "C", "D"),

ncol = 2, nrow = 2)

ggarrange(hour, mon, cas, veh,

labels = c("E", "F","G", "H"),

ncol = 2, nrow = 2)

#remove date and time as created new variables

#remove location\_easting and location\_northing as latitude and longitude similar so can be used in place

accident\_df <- accident\_df %>% select(-c(date,time,location\_easting\_osgr, location\_northing\_osgr))

###CHANGE CATEGORICAL FORMAT

#accident\_severity

#check levels

accident\_df %>% count(accident\_severity)

#check missing data (-1)

sum(is.na(accident\_df$accident\_severity))

#change value

accident\_df$accident\_severity <- as.integer(accident\_df$accident\_severity)

accident\_df$accident\_severity[accident\_df$accident\_severity==1] <- "FATAL"

accident\_df$accident\_severity[accident\_df$accident\_severity==2] <- "SERIOUS"

accident\_df$accident\_severity[accident\_df$accident\_severity==3] <- "SLIGHT"

#convert to factor

accident\_df$accident\_severity <- as.factor(accident\_df$accident\_severity)

#first\_road\_class

#check NA

sum(is.na(accident\_df$first\_road\_class))

#make unkown NA

accident\_df$first\_road\_class[accident\_df$first\_road\_class==6]<-NA

#check factor levels

accident\_df %>% count(first\_road\_class)

#change level name

accident\_df$first\_road\_class[accident\_df$first\_road\_class==1]<-"MOT"

accident\_df$first\_road\_class[accident\_df$first\_road\_class==2]<-"MOT"

accident\_df$first\_road\_class[accident\_df$first\_road\_class==3]<-"A"

accident\_df$first\_road\_class[accident\_df$first\_road\_class==4]<-"B"

accident\_df$first\_road\_class[accident\_df$first\_road\_class==5]<-"C"

#make factor

accident\_df$first\_road\_class <- as.factor(accident\_df$first\_road\_class)

#road\_type

#check NA

sum(is.na(accident\_df$road\_type))

#create NA VARIABLES

accident\_df$road\_type[accident\_df$road\_type==9]<- NA

accident\_df$road\_type[accident\_df$road\_type==-1]<- NA

accident\_df$road\_type[accident\_df$road\_type==12]<- NA

#check factor levels

accident\_df %>% count(road\_type)

#change variable names to make interpretation easier

accident\_df$road\_type[accident\_df$road\_type==1]<-"RB"

accident\_df$road\_type[accident\_df$road\_type==2]<-"OWS"

accident\_df$road\_type[accident\_df$road\_type==3]<-"DC"

accident\_df$road\_type[accident\_df$road\_type==6]<-"SC"

accident\_df$road\_type[accident\_df$road\_type==7]<-"SR"

#make factor

accident\_df$road\_type<-as.factor(accident\_df$road\_type)

#speed limit

accident\_df %>% count(speed\_limit)

#NULL didn't work earlier so make NA using string "NULL"

accident\_df$speed\_limit[accident\_df$speed\_limit=="NULL"] <- NA

#check NA

sum(is.na(accident\_df$speed\_limit))

#convert to factor

accident\_df$speed\_limit<- as.factor(accident\_df$speed\_limit)

#junction\_detail

#check NA

sum(is.na(accident\_df$junction\_detail))

#change to NA

accident\_df$junction\_detail[accident\_df$junction\_detail==99]<-NA

accident\_df$junction\_detail[accident\_df$junction\_detail==9]<-NA

#check factor levels

accident\_df %>% count(junction\_detail)

#change level names to make interpretation easier

accident\_df$junction\_detail[accident\_df$junction\_detail==0]<-"NOT"

accident\_df$junction\_detail[accident\_df$junction\_detail==1]<-"RB"

accident\_df$junction\_detail[accident\_df$junction\_detail==2]<-"MRB"

accident\_df$junction\_detail[accident\_df$junction\_detail==3]<-"TJUN"

accident\_df$junction\_detail[accident\_df$junction\_detail==5]<-"SR"

accident\_df$junction\_detail[accident\_df$junction\_detail==6]<-"CR"

accident\_df$junction\_detail[accident\_df$junction\_detail==7]<-"4ARM"

accident\_df$junction\_detail[accident\_df$junction\_detail==8]<-"PD"

#make factor

accident\_df$junction\_detail<-as.factor(accident\_df$junction\_detail)

#pedestrian\_crossing\_human\_control

#check NA

sum(is.na(accident\_df$pedestrian\_crossing\_control))

#change to NA

accident\_df$pedestrian\_crossing\_human\_control[accident\_df$pedestrian\_crossing\_human\_control==9]<-NA

#change level name - make yes or no due to low levels of not none answers

accident\_df$pedestrian\_crossing\_human\_control[accident\_df$pedestrian\_crossing\_human\_control==0]<- "NO"

accident\_df$pedestrian\_crossing\_human\_control[accident\_df$pedestrian\_crossing\_human\_control==1]<- "YES"

accident\_df$pedestrian\_crossing\_human\_control[accident\_df$pedestrian\_crossing\_human\_control==2]<- "YES"

#make factor

accident\_df$pedestrian\_crossing\_human\_control <- as.factor(accident\_df$pedestrian\_crossing\_human\_control)

#check levels

accident\_df %>% count(pedestrian\_crossing\_human\_control)

#pedestrian physical control

#check NA

sum(is.na(accident\_df$pedestrian\_crossing\_physical\_facilities))

#make unknown NA

accident\_df$pedestrian\_crossing\_physical\_facilities[accident\_df$pedestrian\_crossing\_physical\_facilities==9]<-NA

#change level name - change to yes or no for same reason as other pedestrian

accident\_df$pedestrian\_crossing\_physical\_facilities[accident\_df$pedestrian\_crossing\_physical\_facilities==0] <- "NO"

accident\_df$pedestrian\_crossing\_physical\_facilities[accident\_df$pedestrian\_crossing\_physical\_facilities==1] <- "YES"

accident\_df$pedestrian\_crossing\_physical\_facilities[accident\_df$pedestrian\_crossing\_physical\_facilities==4] <- "YES"

accident\_df$pedestrian\_crossing\_physical\_facilities[accident\_df$pedestrian\_crossing\_physical\_facilities==5] <- "YES"

accident\_df$pedestrian\_crossing\_physical\_facilities[accident\_df$pedestrian\_crossing\_physical\_facilities==7] <- "YES"

accident\_df$pedestrian\_crossing\_physical\_facilities[accident\_df$pedestrian\_crossing\_physical\_facilities==8] <- "YES"

#make factor

accident\_df$pedestrian\_crossing\_physical\_facilities <- as.factor(accident\_df$pedestrian\_crossing\_physical\_facilities)

#check levels

accident\_df %>% count(pedestrian\_crossing\_physical\_facilities)

#light conditions

#check Na

sum(is.na(accident\_df$light\_conditions))

#convert to NA

accident\_df$light\_conditions[accident\_df$light\_conditions==7]<-NA

#check factor levels

accident\_df %>% count(light\_conditions)

#change level name

accident\_df$light\_conditions[accident\_df$light\_conditions==1]<-"DAY"

accident\_df$light\_conditions[accident\_df$light\_conditions==4]<-"LIT"

accident\_df$light\_conditions[accident\_df$light\_conditions==5]<-"UNLIT"

accident\_df$light\_conditions[accident\_df$light\_conditions==6]<-"NL"

#convert to factor

accident\_df$light\_conditions<-as.factor(accident\_df$light\_conditions)

#weather\_conditions

#check NA

sum(is.na(accident\_df$weather\_conditions))

#convert to NA

accident\_df$weather\_conditions[accident\_df$weather\_conditions==9]<-NA

accident\_df$weather\_conditions[accident\_df$weather\_conditions==8]<-NA

#check levels

accident\_df %>% count(weather\_conditions)

#high\_winds

accident\_df$high\_winds<-NA

for(i in 1:nrow(accident\_df)) {

if (is.na(accident\_df$weather\_conditions[i])) {

accident\_df$high\_winds[i] <- NA

}

else if ((accident\_df$weather\_conditions[i]<4) || (accident\_df$weather\_conditions[i]==7)){

accident\_df$high\_winds[i] <- "NO"

}

else if ((accident\_df$weather\_conditions[i]>3) & (accident\_df$weather\_conditions[i]<7)){

accident\_df$high\_winds[i] <- "YES"

}

}

accident\_df %>% count(high\_winds)

#combine high winds and non high-winds

accident\_df$weather\_conditions[accident\_df$weather\_conditions==4]<-1

accident\_df$weather\_conditions[accident\_df$weather\_conditions==5]<-2

accident\_df$weather\_conditions[accident\_df$weather\_conditions==6]<-3

#change level names

accident\_df$weather\_conditions[accident\_df$weather\_conditions==1]<-"FINE"

accident\_df$weather\_conditions[accident\_df$weather\_conditions==2]<-"RAIN"

accident\_df$weather\_conditions[accident\_df$weather\_conditions==3]<-"SNOW"

accident\_df$weather\_conditions[accident\_df$weather\_conditions==7]<-"FOG"

#check levels

accident\_df %>% count(weather\_conditions)

#make factor

accident\_df$weather\_conditions<-as.factor(accident\_df$weather\_conditions)

#road\_surface\_conditions

#CHECK na

sum(is.na(accident\_df$road\_surface\_conditions))

#convert to NA

accident\_df$road\_surface\_conditions[accident\_df$road\_surface\_conditions==9]<-NA

#check levels

accident\_df %>% count(accident\_df$road\_surface\_conditions)

#change level names

accident\_df$road\_surface\_conditions[accident\_df$road\_surface\_conditions==1]<-"DRY"

accident\_df$road\_surface\_conditions[accident\_df$road\_surface\_conditions==2]<-"WET"

accident\_df$road\_surface\_conditions[accident\_df$road\_surface\_conditions==3]<-"SNOW"

accident\_df$road\_surface\_conditions[accident\_df$road\_surface\_conditions==4]<-"ICE"

accident\_df$road\_surface\_conditions[accident\_df$road\_surface\_conditions==5]<-"FLOOD"

#make factor

accident\_df$road\_surface\_conditions<-as.factor(accident\_df$road\_surface\_conditions)

#special conditions

#check NA

sum(is.na(accident\_df$special\_conditions\_at\_site))

#convert NA

accident\_df$special\_conditions\_at\_site[accident\_df$special\_conditions\_at\_site==9]<-NA

#check levels

accident\_df %>% count(special\_conditions\_at\_site)

#change levels name

accident\_df$special\_conditions\_at\_site[accident\_df$special\_conditions\_at\_site==0]<-"NONE"

accident\_df$special\_conditions\_at\_site[accident\_df$special\_conditions\_at\_site==1]<-"AFSO"

accident\_df$special\_conditions\_at\_site[accident\_df$special\_conditions\_at\_site==2]<-"ASPD"

accident\_df$special\_conditions\_at\_site[accident\_df$special\_conditions\_at\_site==3]<-"RSO"

accident\_df$special\_conditions\_at\_site[accident\_df$special\_conditions\_at\_site==4]<-"RW"

accident\_df$special\_conditions\_at\_site[accident\_df$special\_conditions\_at\_site==5]<-"RSD"

accident\_df$special\_conditions\_at\_site[accident\_df$special\_conditions\_at\_site==6]<-"OIL"

accident\_df$special\_conditions\_at\_site[accident\_df$special\_conditions\_at\_site==7]<-"MUD"

#make factor

accident\_df$special\_conditions\_at\_site<-as.factor(accident\_df$special\_conditions\_at\_site)

#carriageway hazards

#check NA

sum(is.na(accident\_df$carriageway\_hazards))

#convert NA

accident\_df$carriageway\_hazards[accident\_df$carriageway\_hazards==9]<-NA

#check levels

accident\_df %>% count(carriageway\_hazards)

#change level names

accident\_df$carriageway\_hazards[accident\_df$carriageway\_hazards==0]<-"NONE"

accident\_df$carriageway\_hazards[accident\_df$carriageway\_hazards==1]<-"LOAD"

accident\_df$carriageway\_hazards[accident\_df$carriageway\_hazards==2]<-"OBJ"

accident\_df$carriageway\_hazards[accident\_df$carriageway\_hazards==3]<-"ACC"

accident\_df$carriageway\_hazards[accident\_df$carriageway\_hazards==6]<-"PED"

accident\_df$carriageway\_hazards[accident\_df$carriageway\_hazards==7]<-"ANI"

#convert to factor

accident\_df$carriageway\_hazards<-as.factor(accident\_df$carriageway\_hazards)

#urban or rural

#check NA

sum(is.na(accident\_df$urban\_or\_rural\_area))

#check levels

accident\_df %>% count(urban\_or\_rural\_area)

#make unallocated NA

accident\_df$urban\_or\_rural\_area[accident\_df$urban\_or\_rural\_area==3]<-NA

#change level name

accident\_df$urban\_or\_rural\_area[accident\_df$urban\_or\_rural\_area==1]<-"URB"

accident\_df$urban\_or\_rural\_area[accident\_df$urban\_or\_rural\_area==2]<-"RUR"

#make factor

accident\_df$urban\_or\_rural\_area<-as.factor(accident\_df$urban\_or\_rural\_area)

#trunk\_road\_flag

#check NA

sum(is.na(accident\_df$trunk\_road\_flag))

#change level name

accident\_df$trunk\_road\_flag[accident\_df$trunk\_road\_flag==1] <- "TRUNK"

accident\_df$trunk\_road\_flag[accident\_df$trunk\_road\_flag==2] <- "NOT"

#make factor

accident\_df$trunk\_road\_flag <- as.factor(accident\_df$trunk\_road\_flag)

#check levels

accident\_df %>% count(trunk\_road\_flag)

##VISUALISE ACCIDENT SEVERITY DIFFERENCES IN NUMERICAL DATA

#histograms

accident\_df$longitude <- as.numeric(accident\_df$longitude)

#VISUALISE ACCIDENT SEVERITY DIFFERENCES IN CATEGORICAL DATA

#(modified from julia silge's blog)

str(accident\_df)

accident\_df <- accident\_df %>% mutate\_if(is.integer,as.factor)

accident\_df <- accident\_df %>% mutate\_if(is.numeric,as.factor)

accident\_df %>%

select(accident\_severity, accident\_year,day\_of\_week,first\_road\_class,

road\_type,speed\_limit,junction\_detail,

) %>% pivot\_longer(accident\_year:junction\_detail) %>%

ggplot(aes(y = value, fill = accident\_severity)) +

geom\_bar(position = "fill") +

facet\_wrap(vars(name), scales = "free", ncol = 2) +

labs(x = NULL, y = NULL, fill = NULL)

accident\_df %>%

select(accident\_severity, pedestrian\_crossing\_human\_control, pedestrian\_crossing\_physical\_facilities,

light\_conditions,weather\_conditions

) %>% pivot\_longer(pedestrian\_crossing\_human\_control:weather\_conditions) %>%

ggplot(aes(y = value, fill = accident\_severity)) +

geom\_bar(position = "fill") +

facet\_wrap(vars(name), scales = "free", ncol = 2) +

labs(x = NULL, y = NULL, fill = NULL)

accident\_df %>%

select(accident\_severity, high\_winds, road\_surface\_conditions,

urban\_or\_rural\_area,trunk\_road\_flag

) %>% pivot\_longer(high\_winds:trunk\_road\_flag) %>%

ggplot(aes(y = value, fill = accident\_severity)) +

geom\_bar(position = "fill") +

facet\_wrap(vars(name), scales = "free", ncol = 2) +

labs(x = NULL, y = NULL, fill = NULL)

accident\_df %>%

select(accident\_severity,carriageway\_hazards, special\_conditions\_at\_site

) %>% pivot\_longer(carriageway\_hazards:special\_conditions\_at\_site) %>%

ggplot(aes(y = value, fill = accident\_severity)) +

geom\_bar(position = "fill") +

facet\_wrap(vars(name), scales = "free", ncol = 2) +

labs(x = NULL, y = NULL, fill = NULL)

#change day of week to whether week day or weekend

#check NA

sum(is.na(accident\_df$day\_of\_week))

#change to weekday or not

accident\_df$day\_of\_week <- as.integer(accident\_df$day\_of\_week)

accident\_df$day\_of\_week[accident\_df$day\_of\_week==1] <- "NO"

accident\_df$day\_of\_week[accident\_df$day\_of\_week==7] <- "NO"

accident\_df$day\_of\_week[accident\_df$day\_of\_week==2] <- "YES"

accident\_df$day\_of\_week[accident\_df$day\_of\_week==3] <- "YES"

accident\_df$day\_of\_week[accident\_df$day\_of\_week==4] <- "YES"

accident\_df$day\_of\_week[accident\_df$day\_of\_week==5] <- "YES"

accident\_df$day\_of\_week[accident\_df$day\_of\_week==6] <- "YES"

#make factor

accident\_df$day\_of\_week <- as.factor(accident\_df$day\_of\_week)

#check factor levels

accident\_df %>% count(day\_of\_week)

#CHECK MISSING VALUES AGAIN

vis\_miss(accident\_df, warn\_large\_data = FALSE)

colSums(is.na(accident\_df))

#remove road\_class\_type

prep\_data <- accident\_df %>% select(-c(first\_road\_class))

prep\_data <- na.omit(prep\_data)

write.csv(prep\_data, "prepped\_data.csv",row.names = FALSE)

**Appendix B**

R code used for association rule mining implementation.

#ASSOCIATION RULES

library(tidyverse)

library(arules)

library(arulesViz)

library(psych)

rm(list=ls())

#set working directory

setwd("C:\\Users\\smann\\OneDrive\\University\\Data Science\\ASDM\\Project\\Road safety")

#import datasets

accident\_df <- read.csv("prepped\_data.csv")

###REMOVE accident\_index, accident\_year, longitude, latitude

accident\_df <- accident\_df %>% select(-c(accident\_index, accident\_year, longitude, latitude))

#change variable type

str(accident\_df)

accident\_df <- accident\_df %>% mutate\_if(is.character,as.factor)

accident\_df <- accident\_df %>% mutate\_if(is.integer,as.factor)

accident\_df$number\_of\_vehicles <- as.integer(accident\_df$number\_of\_vehicles)

accident\_df$number\_of\_casualties <- as.integer(accident\_df$number\_of\_casualties)

accident\_df$month <- as.integer(accident\_df$month)

accident\_df$hour <- as.integer(accident\_df$hour)

#fatal as consequent

rules <- apriori(accident\_df, parameter=list(minlen=2, conf=0.1, supp=0.1), appearance=list(rhs=c("accident\_severity=FATAL")))

#serious as consequent

rules <- apriori(accident\_df, parameter=list(), appearance=list(rhs=c("accident\_severity=SERIOUS")))

#slight as consequent

rules <- apriori(accident\_df, parameter=list(minlen=2, maxlen=4, conf=0.6, supp=0.4), appearance=list(rhs=c("accident\_severity=SLIGHT")))

rules.sorted<-sort(rules, by="lift")

redundant <- is.redundant(rules.sorted)

which(redundant)

rules.pruned <- rules.sorted[!redundant]

inspect(rules.pruned)

inspect(head(rules.pruned, 20))

#FATAL ACCIDENTS ONLY

rules <- apriori(fatal\_df, parameter= list(minlen=2, maxlen=3, supp=0.4, conf=0.5))

redundant <- is.redundant(rules)

which(redundant)

rules.pruned <- rules[!redundant]

rules.support<-sort(rules.pruned, by="support")

rules.lift<-sort(rules.pruned, by="lift")

inspect(rules.support)

inspect(head(rules.support, 20))

plot(rules.support, method = "graph", limit = 20)

inspect(rules.lift)

table(inspect(head(rules.lift, 20)))

plot(rules.lift, method = "graph", limit = 20)