Highway Accidents using Association Rule

Muhammad Zain Bin Aamir

Lahore School of Economics

Contents

Introduction

Road accidents is one of the major cause of people dying every year. Every year road accidents kill more than a million people and injure more than 20 million worldwide (Peden et al). The main motive behind this paper is to come up with a predictive analysis technique that predicts what causes accidents and to supply guidance on road safety and build awareness by pinpointing the foremost causes of traffic accidents. This knowledge can be used to warn drivers about the dangers of accidents and how the consequences are worse given a specific context also the highway authorities as a whole can make sure that the necessary steps are taken to avoid accidents.

The fact that the decisions made after the use of analytical tools are better informed the ones made without the use of these tools (Klatt et al., 2011; Schläfke et al., 2012). This makes it evident that predictive analytics plays an integral part as it adds great value to it at particular levels. The companies or authorities which use these are known for their higher-rankings (relatively), strictly follow analytics (Klatt et al., 2011), whereas those with the highest rankings adopt predictive analytics methods (Daven-port, 2006; Siegel, 2013).

This paper outlines the major causes of road accidents by applying data mining algorithms to data collected from past accident records of the USA, where thousands of motor vehicle crashes occur every year. The data set consists of 36 different variables and 6568 values, it consists of inputs from the year 2012 to May 2018. After data preprocessing, the cleaned data is analyzed and mined using the data mining technique. Lastly, factors are analyzed which causes accidents and recommendations are provided according to it for the highway authority.

Literature Review

Predictive Analytics in present is constantly encountered in many organizations; a technology is that is utilized by successful organizations. Predictive analytics considered something that just came out many but the original thought goes back to DSS or Decision Support System (Negash). Negash (2004, p.178) defines Predictive Analytics as:

"Systems that combine data gathering, data storage, and knowledge man-agement with analytical tools to present complex internal and competitive information to planners and decision makers."

A lot of studies have been conducted on vehicle accidents. In one of the studies it was found out that accidents were depended on road characteristics, traffic and environmental factors, this was analyzed by using Classification and Regression Tree (CART) (Xu et al.). (Fogue et al.) in one of his studies made a model which provided automatic accident notifications and Assistance system which estimated the harshness of the accidents and it was concluded by him that over speeding was the main factor in in head-on collisions. However, the explanation for the causes of crashes was not explained in his study. In another study different models were compared, models such as Naïve Bayes, AdaBoostM1, PART, J48 and Random Forest Classifier and the model was chosen on the basis of injury harshness. (Krishnaveni et al.) found out in his study that factors like weather, age of the driver, type of vehicle are responsible for crashes.

Apart from the use of these models' text mining has also been used to identify the causes of accidents. In one of the studies of (Nayak et al.) has used text mining to provide an explanation for the cause of accidents, he analyzed around 20000 vehicle accidents in Queensland, Austria, he concluded in his study that the reason for most vehicle accidents was over speeding and most of the accidents were rear-end collisions at intersections. In another study to find out the factors which were most important for causing accidents, (Zhang and Fan) used the model decision tree on twenty years of data in Saskatchewan, Canada and they concluded that the most important factors were causing accidents were violating traffic rules, intoxicated drivers, and inexperienced drivers during poor weather conditions but conditions of vehicles were not considered in this study.

Theoretical Literature

Apriori algorithm uses all of the transactions in the database to define the market basket (Agrawal and Srikan, 1994; Agrawal et al., 1996). Association rules are used to represent the market basket with a left and a right side, Left => Right. For example, in item set {d, e, f}, the rule {e, f} => {d} should be understood as if the customer buys {e, f} he/she will also buy {d}. Initially this approach was used only in pattern recognition and with the discovery of the rule "on Thursdays, grocery store customers often purchase diapers and beer together" (Berry and Linoff, 1997) it became very popular. Two measures, namely the support measure and the confidence measure, are used to evaluate the association rules. Suppose {D, E} is an item set with D=>E as the association rule. So, in this case, support measure is equal to the relative frequency or the probability P (DnE) and the confidence measure is equal to the conditional probability of E given D, P(E|D). This is also equal to P(DnE)/P(D). According to this, the rule "biscuits => cigarettes" with support measure of 30% and confidence measure of 60% means that biscuits and cigarettes are bought together in 30% of the transactions and in these transactions, 60% of the cases are where the one who buys biscuits will also buy cigarettes, means P(cigarettes|biscuits) = 60%.

Methodology

Data

The data set that was chosen for the Machine Learning project was "Highway Accidents" which consists of accidents data put together by an international Highway Authority. The data set consists of 36 different variables and 6568 values because it consists of inputs from the year 2012 to May 2018. The data will be cleaned, prepared and analyzed in order to provide useful visualizations and insights that can be utilized to provide recommendations to the international Highway Authority regarding what actions can be taken to avoid or reduce the number of accidents.

Data Exploration and Observations

Selected Variables

For the data Exploration process, it was identified that the following 16 Variables were useful for further analysis. The reason why these variables were selected are also provided. The selected variables are:

Event ID, Event Type, Event Subtype, Year, Month, Day of the Week, Time Interval, Details of Actual Job being Done at the Time, Was there a Vehicle involved in this Event?, Weather/Visibility, Lightning Conditions, Was a Police Report Filed?, Preventable, Type, Kind of Event, Actual Lost Workdays (Total).

Removed Variables

Type of Person, Date/Time of Events, Time, Date/Time Reported, Location, Specific Location (include RCC Log no if relevant), Part of Body Affected, Injury Type, Work Related, Accident Type, Vehicle Motion, Who was Hit?, Surface Condition, Roadbed Grade, Name of

Law Enforcement Agency, Expected Lost Workdays (Total), Expected Restricted Workdays (Total), Actual Restricted Workdays (Total), Return To Work Date, Sign Off 1.

Data Cleaning and Preparation

Since the data file was in an xlxs format, the file was converted in to csv format. The data was converted to CSV because it is easier to convert the data in to transactional format. After that while observing the data it was found out that a lot of it was in the strings format and thus the unavailable data which was presented in the form of "Blanks", "Not Available" and "N/A" was not being read by R because R only reads these entries in a specific format.

The command na.strings=c(" ","NA","N/A","Not Applicable")) to convert all these respective entries to the readable format and subsequently removes them as the file is being read by R thus eradicating the need of the additional command of "na.omit".

Moving on to the within this data set it was seen that that certain columns like Expected &Actual Lost Workdays and Expected & Actual Restricted Workdays were majorly empty thus after creating a duplicate data set for HA2 by the name of "tempHA" a command ***prop.table()*** was ran. This command helped in identifying that these columns mainly consisted of zeros and hence were of no use however, it was decided to keep the Actual Lost Work Days column as it could categorize it to form useful analysis.

In addition to this, entire columns of unimportant variables that were identified initially during the data observation were removed by using the command ***tempHA<- tempHA[,-c(4,5,9,11,12,13,15,16,18,19,20,21,23,24,27,31,33,34,35,36).*** Last but not the least, for the visualization purposes, a duplicate dataset for tempHA by the name of "tempHA2".

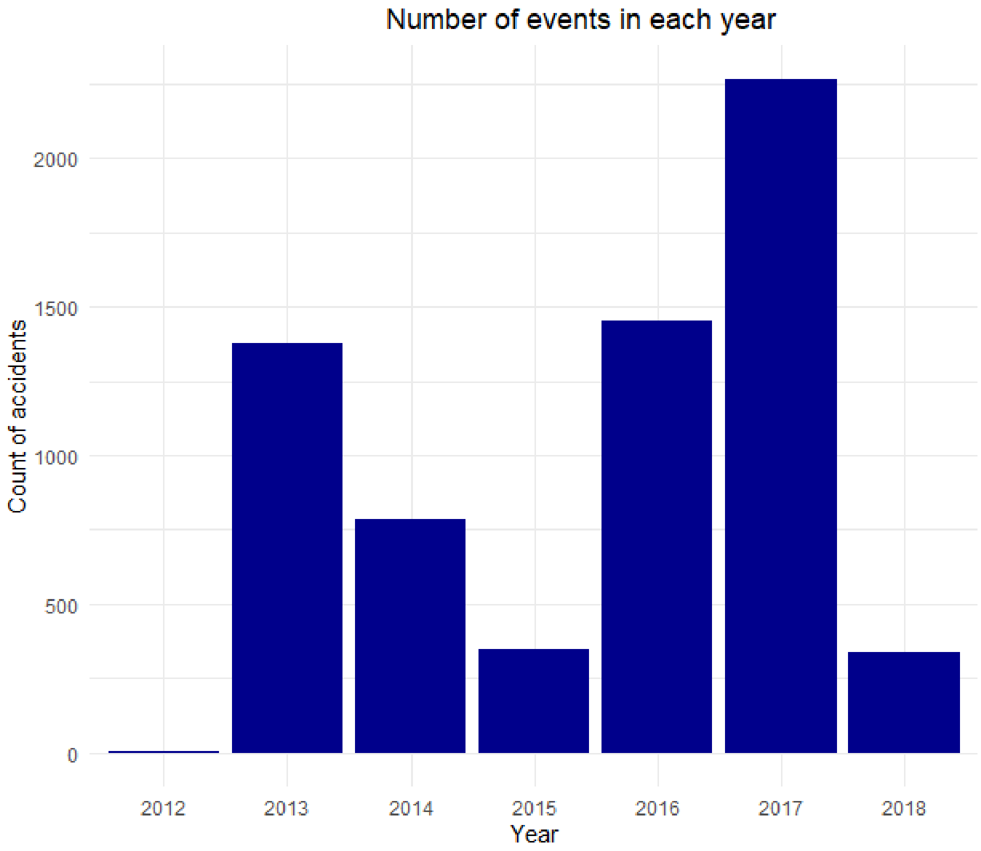
The following steps to better customize the data for further analysis:

1. Data was sorted in terms of Event ID by using the command ***tempHA2 <- tempHA2[order(tempHA2$Event.ID),]*** as to make it more organized.
2. Days of the Week column was converted in two distinct categories of “Weekdays” and ‘Weekend” by running the following command: ***tempHA2$Day.of.the.Week <- ifelse(tempHA2$Day.of.the.Week == "Mon" | tempHA2$Day.of.the.Week == "Tue" | tempHA2$Day.of.the.Week == "Wed" | tempHA2$Day.of.the.Week == "Thu" | tempHA2$Day.of.the.Week == "Fri", "Weekday", "Weekend")***
3. The Preventable column was converted from mere Yes, No to two distinct categories of “Preventable” and “Not Preventable” respectively by running the following command: ***tempHA2$Preventable <- ifelse(tempHA2$Preventable == "No", "Not Preventable", ifelse(tempHA2$Preventable == "Yes", "Preventable", "N/A"))***
4. Was a Police Report Filed? column was converted from mere Yes, No to two distinct categories of “Filed” and “Not Filed” respectively by running the following command: ***tempHA2$Was.A.Police.Report.Filed. <- ifelse(tempHA2$Was.A.Police.Report.Filed. == "Yes", "Filed", "Not Filed")***
5. The Lightening Conditions column was converted from mere 100%, Poor and Non to three distinct categories of “Morning”, “Evening” and “Night” respectively by running the following command: ***tempHA2$Lighting.Conditions <- ifelse(tempHA2$Lighting.Conditions == "100%", "Morning", ifelse(tempHA2$Lighting.Conditions == "Poor", "Evening", ifelse(tempHA2$Lighting.Conditions == "Non", "Night", "N/A")))***

6. The Actual Lost Workdays was converted from strings to a numeric variable and then categorized its values into three distinct groups of “0-4 Days”, “4-8 Days” and “Greater than 8 Days” by running the following commands: ***tempHA2$Actual.Lost.Workdays..Total. <- as.numeric(tempHA2$Actual.Lost.Workdays..Total.)tempHA2$Actual.Lost.Workdays..Total. <- ordered(cut(tempHA2$Actual.Lost.Workdays..Total., c(0,4,8,70)), labels = c("0-4 Days", "4-8 Days", "Greater than 8 Days"))***
7. Lastly, for working on Association Rules later on a duplicate dataset for tempHA2 by the name of “tempHA3” and added the command***tempHA3 <- tempHA3[,-c(1,8)]***in order to remove the columns that weren’t required.

Data Descriptive

Count of Accidents in all years

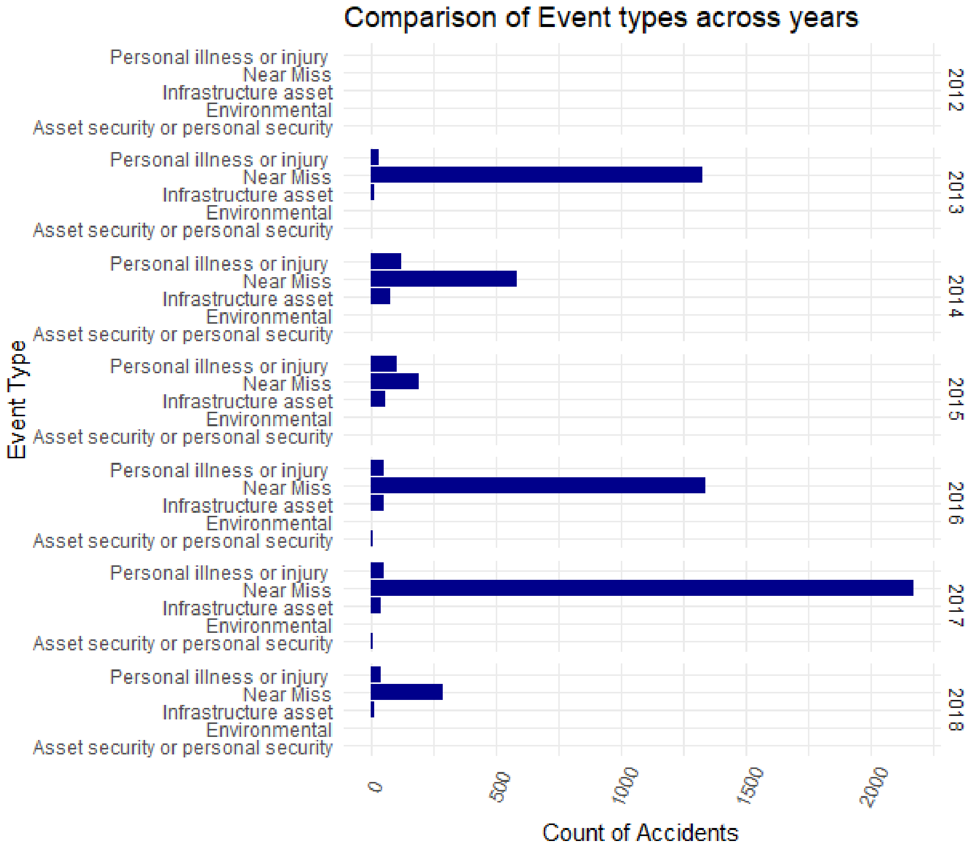


The first aspect to consider during comparison between all the years 2012 to 2018 was the number of accidents in respective years. The plot above shows that the count for year 2012 was almost zero and for 2013 it was almost around 1300 which was almost 60% more than 2014, having a count of almost 800. It can be seen that the year 2015 and 2018 had significantly less accidents as compared to rest of the years except 2012. For the year 2017 it can be seen that accidents are significantly higher than year 2016. In 2016, the number of accidents was just below 1500 and in 2017, the number of accidents were slightly above 2250. The increase is of approximately 50% in one year.

This doubling can be attributed to the fact that the traffic and infrastructure improvement projects increased on highways due to reasons such as increased mobility in area or trade activity. The world population is increasing as well as the globalization. With the transportation increasing massively, governments have had to initiate projects of road maintenances and expansions. Due to the importance of keeping the traffic in flow, the projects are conducted without road closures. Drivers are not used to narrow lanes, barriers, and newly introduced temporary speed limits. This puts the road workers at severe risks of accidents and hence explains the increasing number of total accidents between 2016 – 2017 and 2012-2013. The decrease in the accidents from the year 2013 to 2015 can be because of that projects completed and the increase afterwards can be because of that some new projects started.

However, it is still an alarming situation for any highway authority. To add into the gravity of situation, the historical data from India, Great Britain, and European Union reflects a global decrease in accidents between 2016 and 2017. This is why more variables were introduced into the picture and tried to identify any specific reason (if any) for the trend shown in this plot.

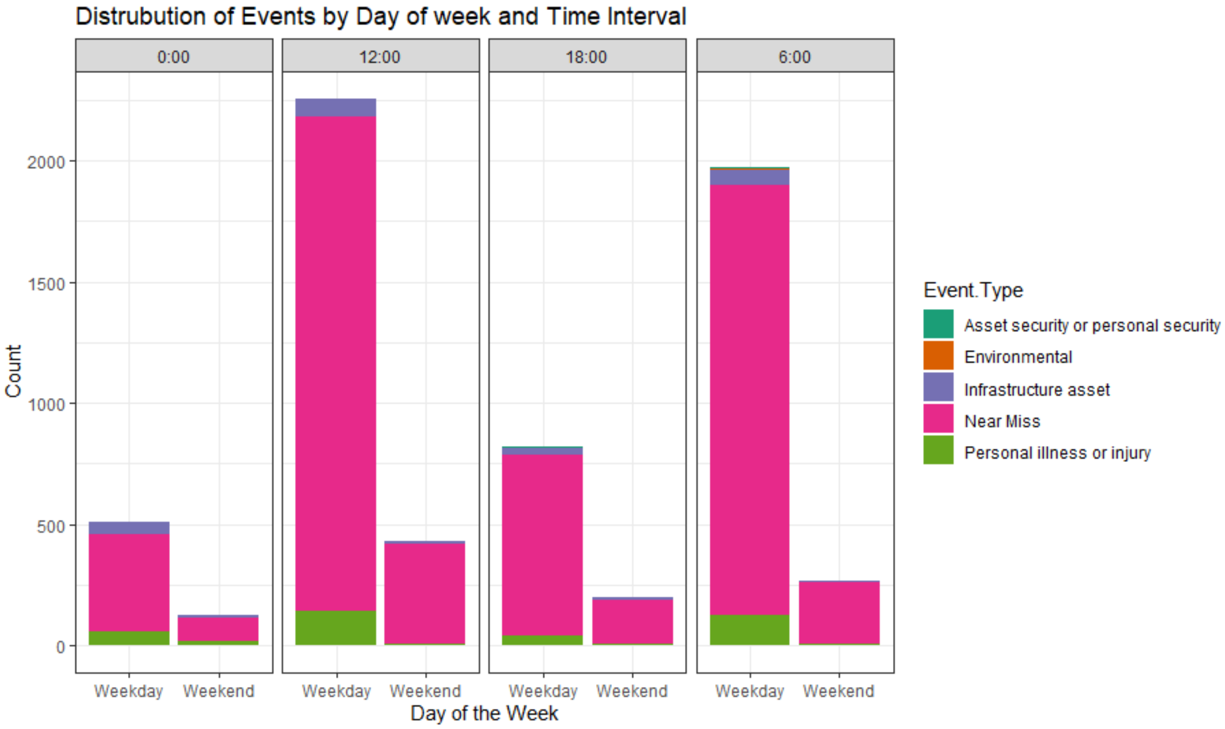
Comparison of Count of Accidents with Event Types across years



This visualization shows the annual data for number of accidents of each event type. It can be seen from the visualization that in 2012 there were no accidents and for the year 2013 there was a sharp increase in "Near Miss" category. There is a rise in overall count of accidents which is justifiable as shown in previous visualization. The major chunk of increased accidents in 2017 have gone into "“Near Miss" category. In all of the years the major category was "Near Miss" and after that it was " Personal illness or injury".

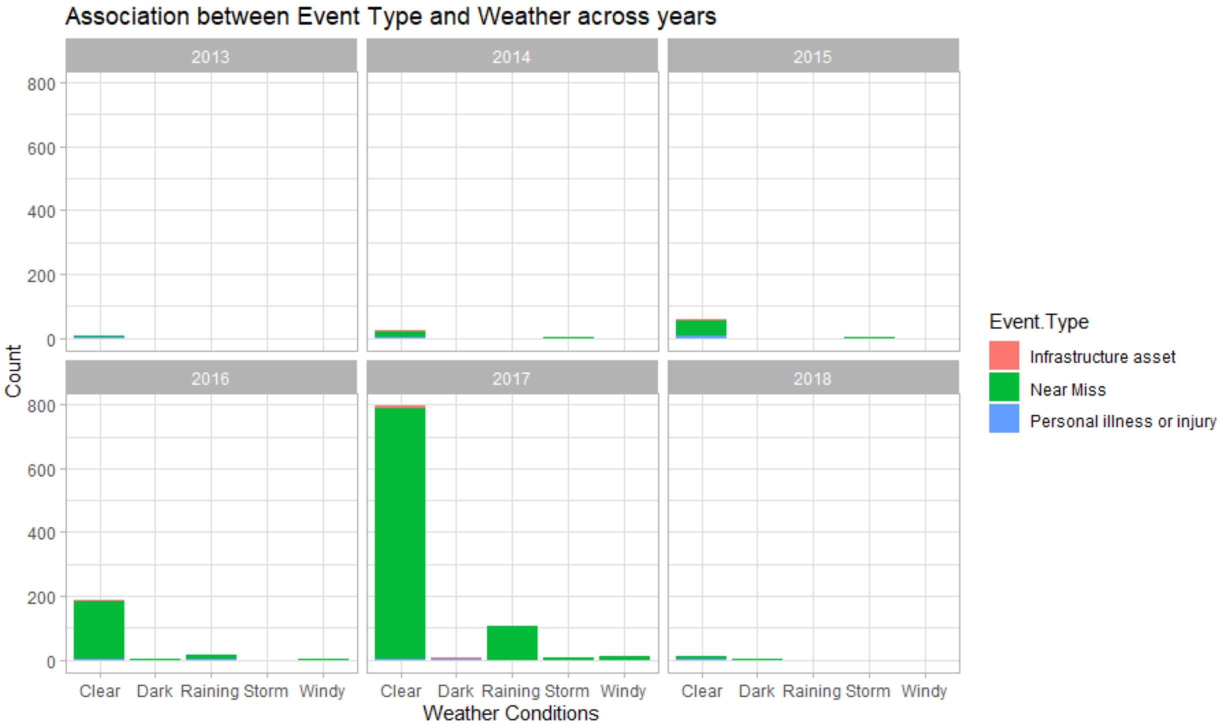
The key insight here is that it appears that precautions are undertaken at constructions and maintenances which are saving the accidents from turning into fatal ones. However, on the downside, it can be seen that there are fundamental errors which are putting the safety of roadworkers at risk even if there is not much harm done. This calls for looking into factors such as the negligence involved in setting up of appropriate warning signs to highlight the danger area, inadequate management of speed limit controls and unavailability of proper application of procedures in making precedents of mandatory precautions.

Time Intervals and Days of the Week



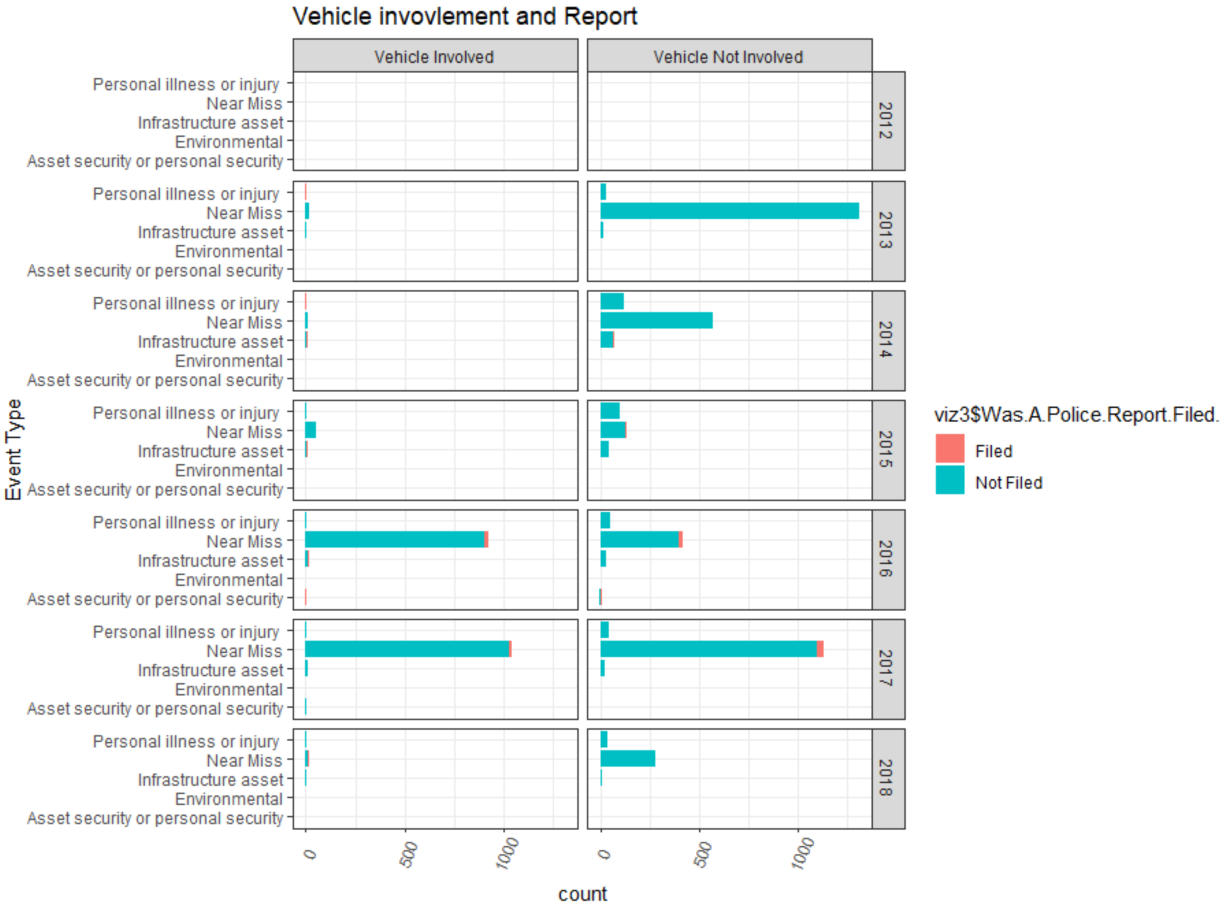
The visualization shows the number of accidents taking place on weekdays and weekends against the time intervals. It is evident that there are more number of accidents on weekdays in each interval and the interval of morning till evening has highest number of accidents. This is primarily due to the highest load of traffic in these hours of the day. Moreover, the construction activities are also taking place mainly in this time period. The weekdays serve as a higher ratio of accidents prone set which seems to be a result of most of the construction/maintenance sites to be operation then. The weekends are usually off.

Association between Event Type and Weather across Years



From the visualization it can be seen that the number of accidents increased from the year 2015 to 2017 during clear weather. Also, there was an increase in accidents during Rainy weather increased from 2016 to 2017. One possible reason for this can be the lack of weather updates on traffic info channels on radio etc. One would assume that there is a greater probability of accidents occurring during rain if it is at night. However, from the previous plot, it was found that most accidents occur during noon. There may be low levels of risk management on sites of constructions, inadequate system of water drainage system, and slipperiness on roads causing the results seen above.

The reporting to police with respect to involvement of vehicle and Event Type Annually

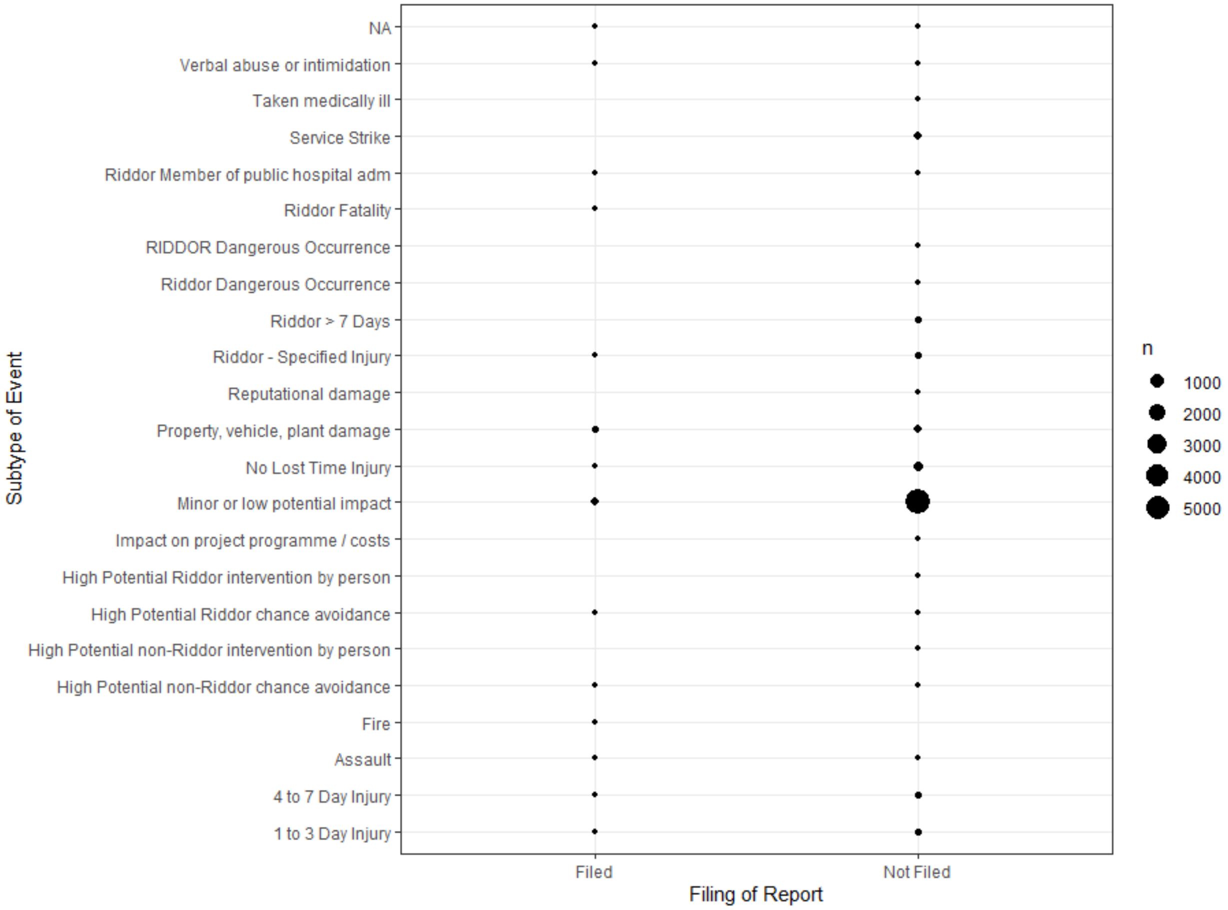


Event types were also analyzed on a deeper level to see which type of events were mostly reported or underreported. The visualization shows that for incidents even where the vehicle was involved, the reports at police haven't significantly been filed. It is not limited to near miss incidents only, but in fact it also spreads to incidents where infrastructural assets and personal injuries are involved. In 2016, incidents were relatively more reported and filed by police when asset security or personal security was there and in 2013 and 2014 the reports which were filed were those which involved personal illness or injury. In 2017 and 2018, that also seems to vanish and police is not filing reports for it either.

For this annual change, a reason can be that when police files report on asset security and personal security, they have to investigate and generate results which is an additional task for usual job description of patrolling mainly for highway authorities.

Usually, the report filing rate is very low where the vehicles are involved. It can be observed that these might include hit and run incidents largely. Construction conglomerates may also be involved in covering of incidents to maintain their contracts as reported incidents then might require them to pay compensations.

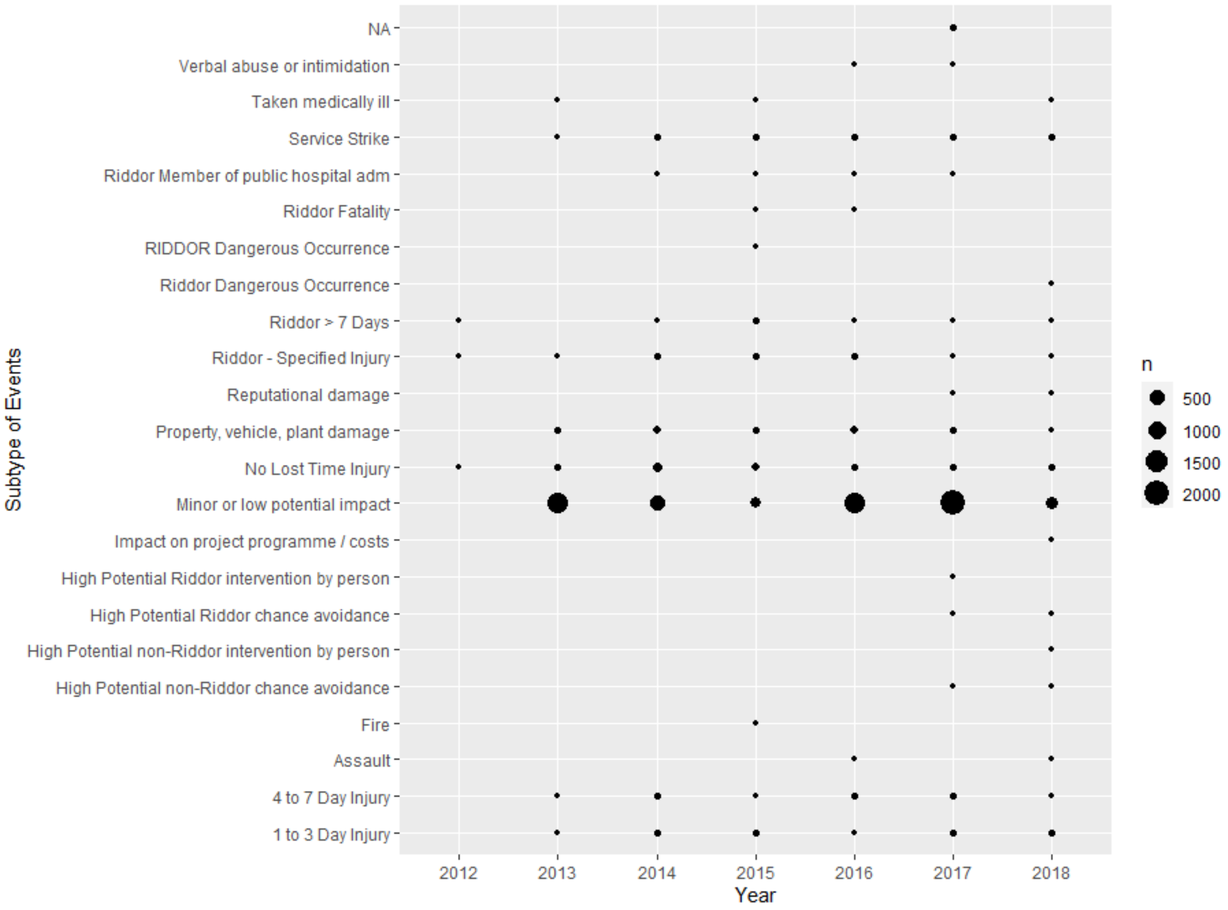
Filing of Report against each subtype of Event



It was also intended to observe that which type of accidents are considered serious enough to be reported. Underreported accidents may falsely alleviate the gravity of situation and hence, explain the un-noticed increasing number of accidents. It is observed in the visualization above that there is an overall severe lack in appropriate report filings. The incidents making an impact such as service strikes, specified injuries, reputational damage, and up to 7 days of injury are completely ignored. The reasoning behind can be divided into two segments. Firstly, the mechanism of authorities in classifying the information worthy of being reported can be problematic over here. The authorities might not have a proper channel to classify things such as reputational challenge even being reportable under the prevailing authorities record system.

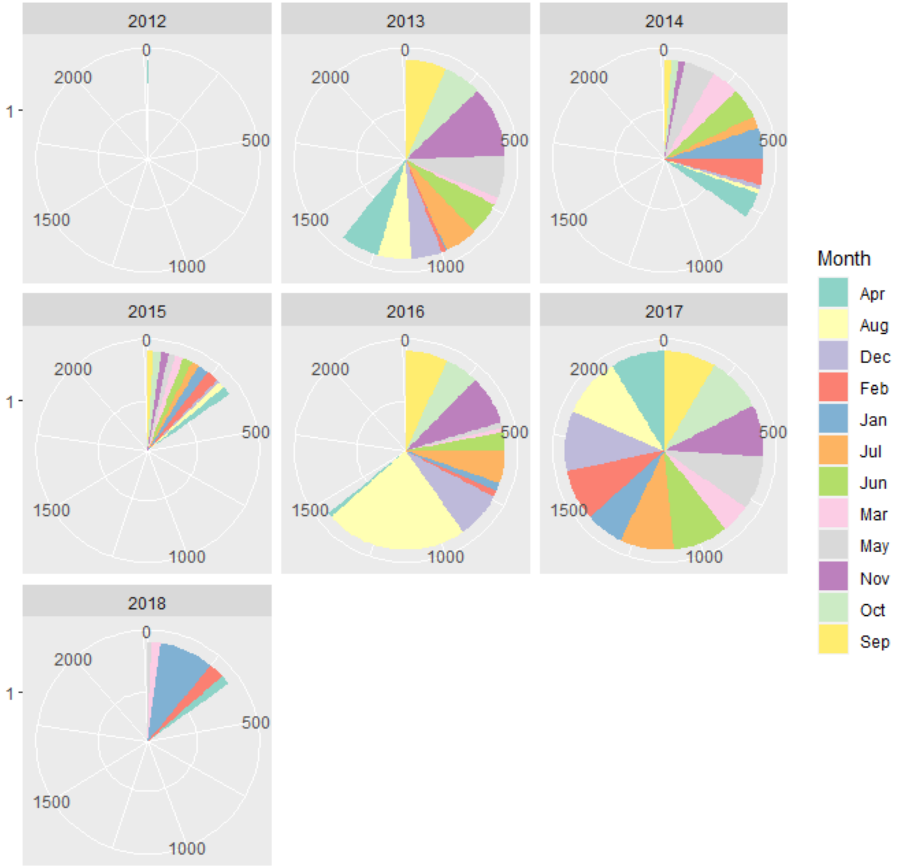
Secondly, there is a lack of properly active and transparent patrolling and performance of police apparently. Not all the cases are being recorded by them. It can be due to efforts of police to reflect lower accidents in their areas as well as escape responsibility of providing remedies and rescue operations. However, another element could be that construction companies might be involved in hiding these accidents to cover up for their negligence in providing worker safety.

Annual comparison of Subtype of Events Count



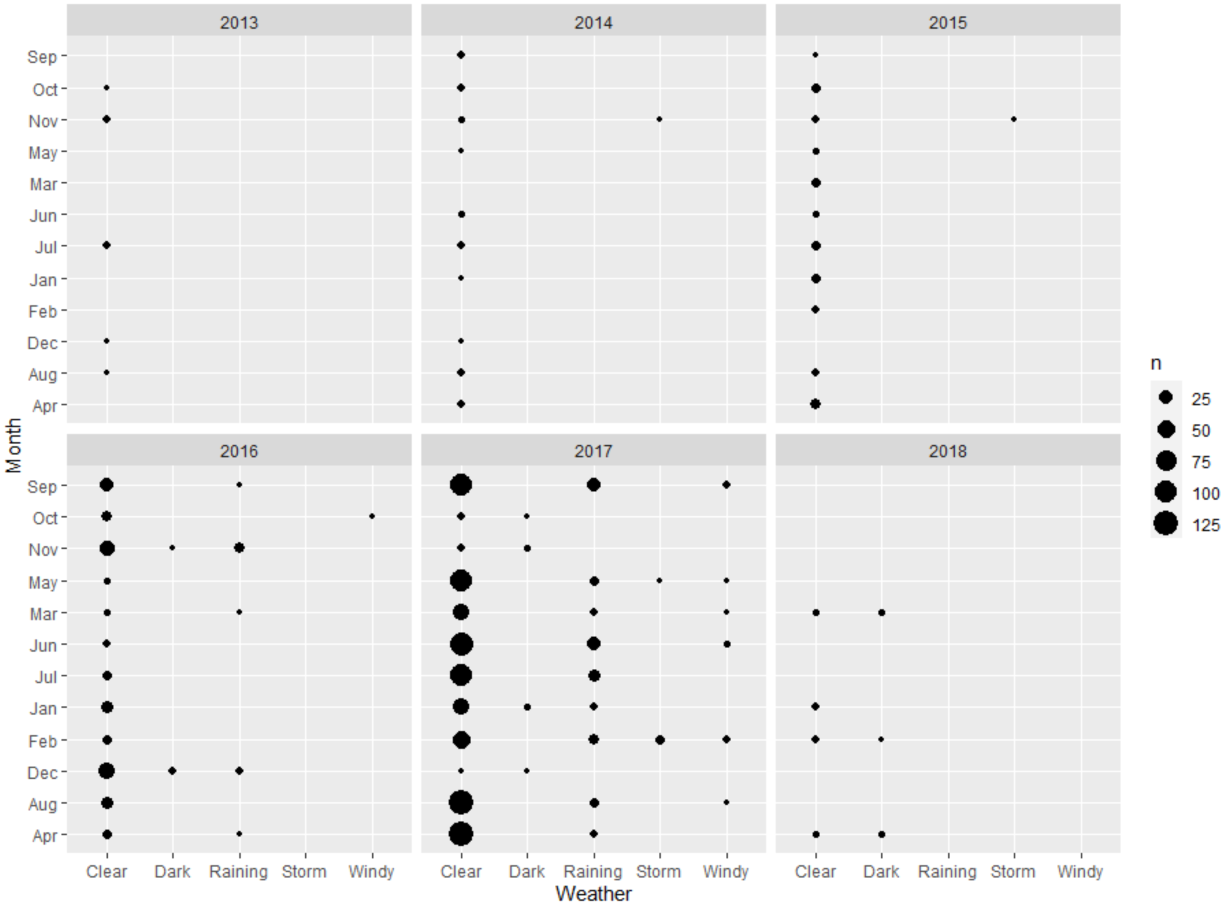
Previously it was discovered that majority of the accidents were "Near Miss", did not involve a vehicle and are not reported. From the visualization above shows the yearly comparison of occurrences of subtype of events, it also helps us to see which subtypes of events are frequently occurring. There is a sharp increase in the "Minor or low potential impact" from the year 2012 to 2013 and till the year 2015 it kept on decreasing but then here is an increase of occurrences of different events from 2016 to 2017. It is seen that majority of events are minor or low impact. This can help explain why there is trend that “Near Miss” accidents are not reported. The impact created by these accidents is perhaps not considered serious enough to illicit a reaction.

Count of Accidents with respect to months (Annual Comparison)



Similarly, a monthly comparison of accidents across years was done to see any visible patterns. From the visualization it can be seen that most of the accidents in 2013 were taken place in November and in 2014 the number of accidents were less and were mostly taken place in January. The least number of accidents according to this were taken place in 2015. The august of 2016 has a drastically higher number of accidents while it spread over in 2017. 2017 has each of the months affected almost equally, which can be explained from the fact that the overall number of accidents almost doubled this year. Comparatively more accidents occur during the latter half of the year i.e. august-December. Perhaps, it can be attributed to the hazy, dark winter weather that might affect visibility on the road. To further explore any specific reason to explain the trend above, the next visualization checks the weather's link with months.

Weather and Months



The hypothesis that weather may have something to do with the increased number of accidents during the winter months does not stand valid because this visualization shows that weather remained mostly clear at the time of accidents in August-December.

Technique Applied

Association Rules

Before the interpretation of results, it is important to understand few of the metric that are used in this because these will help us understand the interpretation, results better and it will strengthen the understanding of Association Rule.

Key Terms

Itemset

A type of basket which is a collection of one or more items.

Support

The support is basically the portion of transactions that hold an item-set. The support is calculated in the following way:

X = item

$$\text{supp}(X) \; = \frac{\text{Number of transaction in which X appears}}{\text{Total number of transactions}}$$

Confidence

This measure defines the likeliness of occurrence of consequent on the cart given that the cart already has the left-hand side (antecedents). Confidence basically answers the question that of all the transactions containing say, {LAYS}, how many also had {Eggs} on them?

Through observation, we can say that {LAYS} → {Eggs} should be a high confidence rule. Confidence is the conditional probability of occurrence of consequent given the Left-hand side.

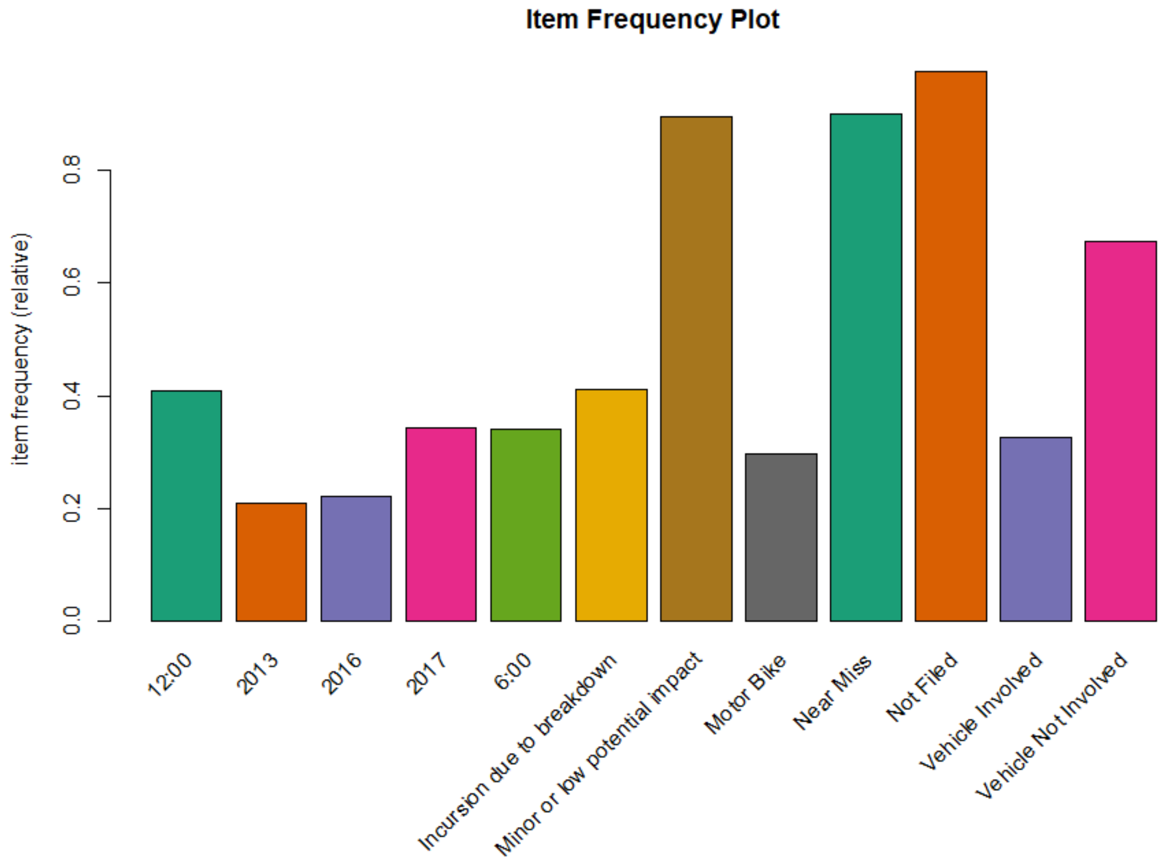
$$\text{conf}(X \longrightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$$

Lift

Lift indicates the strength of an association rule over the random co-occurrence of item C and item D. It provides information about the change in probability of item D in the presence of item A. It values greater than 1 indicates that transactions containing item D tend to contain item C more often than transactions that do not contain item A

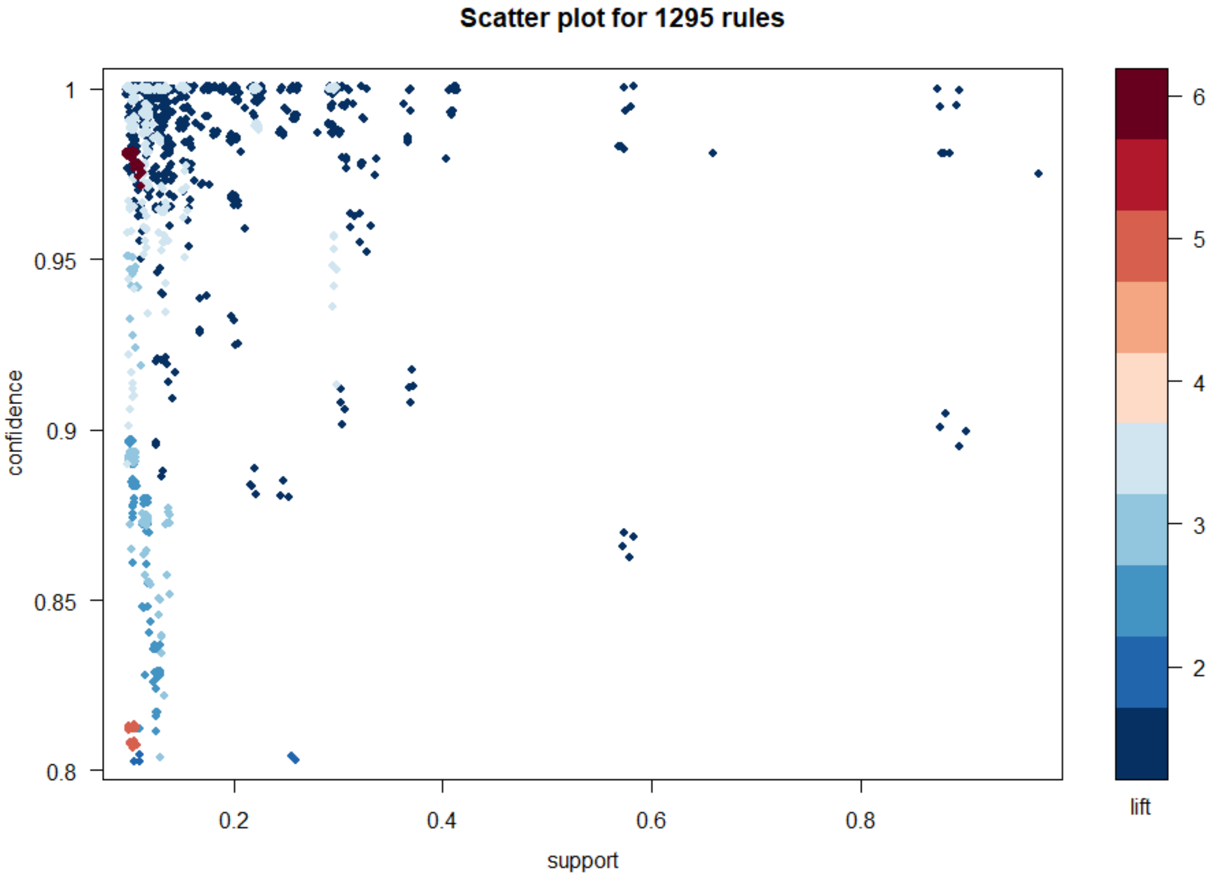
$$\mathrm{lift}(X \longrightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X) * \mathrm{supp}(Y)}$$

This means that the likelihood of the itemset X being bought when item c is bought while taking into account the popularity of D.



Using association rules as the choice of unsupervised learning technique, was to discover significant rules useful for predictive modeling. In order to use A-rules for the data was converted it into a sparse matrix and an item frequency plot was made to assess the variables with the differing support levels. The plot depicts that almost 90% of these incidents were near miss which entail that they did not result in an actual injury or harm but were considered a potential threat which is confirmed by how 90% of the accidents consisted of a minor or low potential impact evident by the plot. What is more surprising is more than 90% of all the accidents that occurred have not been reported to the police. Furthermore, this plot also gives a picture of the yearly breakdown where around 35% of incidents occurred in 2017 while the other

20% occurred in 2016, around 20% in 2013. The most important thing in this is that a lot of accidents happened at 12:00.

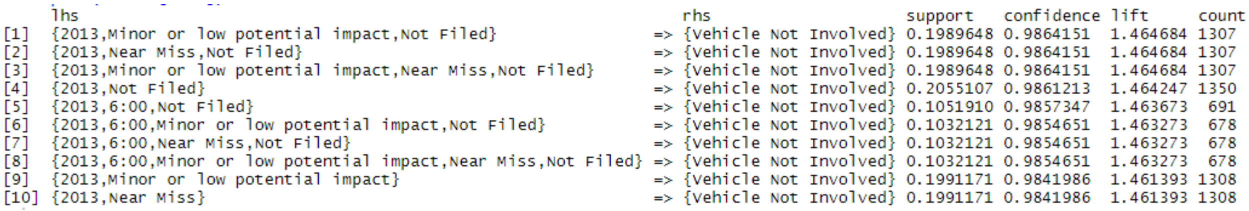


Moreover, plotting the rules as a scatter plot depicted the support, confidence and lift each rule carried in relation to the other rules. Most of the rules which had very less support (between 0 and 0.2) also had very high confidence (0.95-1). This plot was made in order to select the correct values for support and confidence for further analysis. Support and confidence level were chosen to be 0.1 and 0.8 respectively.

Analysis

From the item frequency plot, four different events which had higher levels of support were identified to be explored further. Different targeting rules were run to see what kind of situation was associated with those specific events. Support & confidence was fixed to 0.1 and 0.8 respectively for all the events. The rules were sorted with respect to lift and top ten rules were chosen.

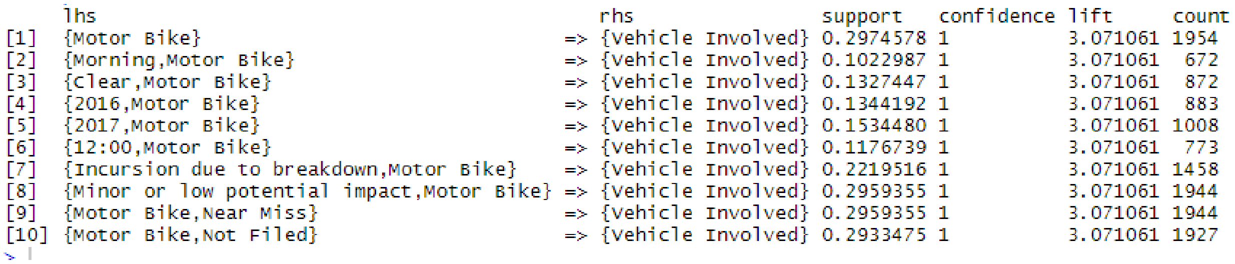
Vehicle Not Involved



```
     lhs                                                    rhs                    support   confidence lift     count
[1]  {2013,Minor or low potential impact,Not Filed}          => {Vehicle Not Involved} 0.1989648 0.9864151  1.464684 1307
[2]  {2013,Near Miss,Not Filed}                              => {Vehicle Not Involved} 0.1989648 0.9864151  1.464684 1307
[3]  {2013,Minor or low potential impact,Near Miss,Not Filed} => {Vehicle Not Involved} 0.1989648 0.9864151  1.464684 1307
[4]  {2013,Not Filed}                                        => {Vehicle Not Involved} 0.2055107 0.9861213  1.464247 1350
[5]  {2013,6:00,Not Filed}                                   => {Vehicle Not Involved} 0.1051910 0.9857347  1.463673  691
[6]  {2013,6:00,Minor or low potential impact,Not Filed}     => {Vehicle Not Involved} 0.1032121 0.9854651  1.463273  678
[7]  {2013,6:00,Near Miss,Not Filed}                         => {Vehicle Not Involved} 0.1032121 0.9854651  1.463273  678
[8]  {2013,6:00,Minor or low potential impact,Near Miss,Not Filed} => {Vehicle Not Involved} 0.1032121 0.9854651  1.463273  678
[9]  {2013,Minor or low potential impact}                    => {Vehicle Not Involved} 0.1991171 0.9841986  1.461393 1308
[10] {2013,Near Miss}                                        => {Vehicle Not Involved} 0.1991171 0.9841986  1.461393 1308
```

From the rules above it can be seen that almost all of the rules except the last two were not Filed. This tells us that the accident in which there was no vehicle involved police didn’t took it as a significant incident and didn’t file it.

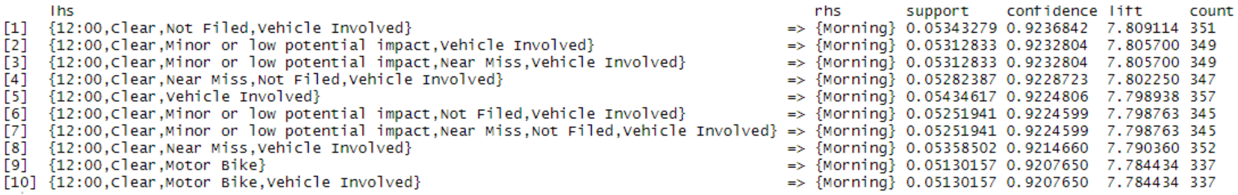
Vehicle Involved



```
     lhs                                         rhs                support   confidence lift     count
[1]  {Motor Bike}                             => {Vehicle Involved} 0.2974578 1          3.071061 1954
[2]  {Morning,Motor Bike}                     => {Vehicle Involved} 0.1022987 1          3.071061  672
[3]  {Clear,Motor Bike}                       => {Vehicle Involved} 0.1327447 1          3.071061  872
[4]  {2016,Motor Bike}                        => {Vehicle Involved} 0.1344192 1          3.071061  883
[5]  {2017,Motor Bike}                        => {Vehicle Involved} 0.1534480 1          3.071061 1008
[6]  {12:00,Motor Bike}                       => {Vehicle Involved} 0.1176739 1          3.071061  773
[7]  {Incursion due to breakdown,Motor Bike}  => {Vehicle Involved} 0.2219516 1          3.071061 1458
[8]  {Minor or low potential impact,Motor Bike} => {Vehicle Involved} 0.2959355 1          3.071061 1944
[9]  {Motor Bike,Near Miss}                   => {Vehicle Involved} 0.2959355 1          3.071061 1944
[10] {Motor Bike,Not Filed}                   => {Vehicle Involved} 0.2933475 1          3.071061 1927
>
```

For the incidents in which there was a vehicle involved it can be seen from the rules that all the rules include a bike. This tells us that motor bike drivers are more reckless as compared to other vehicles. For “Vehicle Involved”, events it was seen that motor bike was involved in its occurrence with 100% confidence and 1 > lift ratio. It is obvious that if there was a vehicle involved in an accident it has to be a motor bike.

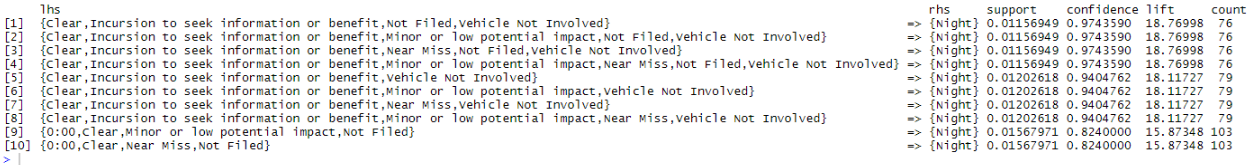
Morning



```
     lhs                                                                                 rhs       support    confidence lift     count
[1]  {12:00,Clear,Not Filed,Vehicle Involved}                                         => {Morning} 0.05343279 0.9236842  7.809114 351
[2]  {12:00,Clear,Minor or low potential impact,Vehicle Involved}                     => {Morning} 0.05312833 0.9232804  7.805700 349
[3]  {12:00,Clear,Minor or low potential impact,Near Miss,Vehicle Involved}           => {Morning} 0.05312833 0.9232804  7.805700 349
[4]  {12:00,Clear,Near Miss,Not Filed,Vehicle Involved}                               => {Morning} 0.05282387 0.9228723  7.802250 347
[5]  {12:00,Clear,Vehicle Involved}                                                   => {Morning} 0.05434617 0.9224806  7.798938 357
[6]  {12:00,Clear,Minor or low potential impact,Not Filed,Vehicle Involved}           => {Morning} 0.05251941 0.9224599  7.798763 345
[7]  {12:00,Clear,Minor or low potential impact,Near Miss,Not Filed,Vehicle Involved} => {Morning} 0.05251941 0.9224599  7.798763 345
[8]  {12:00,Clear,Near Miss,Vehicle Involved}                                         => {Morning} 0.05358502 0.9214660  7.790360 352
[9]  {12:00,Clear,Motor Bike}                                                         => {Morning} 0.05130157 0.9207650  7.784434 337
[10] {12:00,Clear,Motor Bike,Vehicle Involved}                                        => {Morning} 0.05130157 0.9207650  7.784434 337
```

For “Morning”, support level was set to 0.05 as the data did not have many entries regarding the lighting. One of the rules shows the involvement of a minor or low potential impact which might be due to a lower number of people on the highway during morning hours.

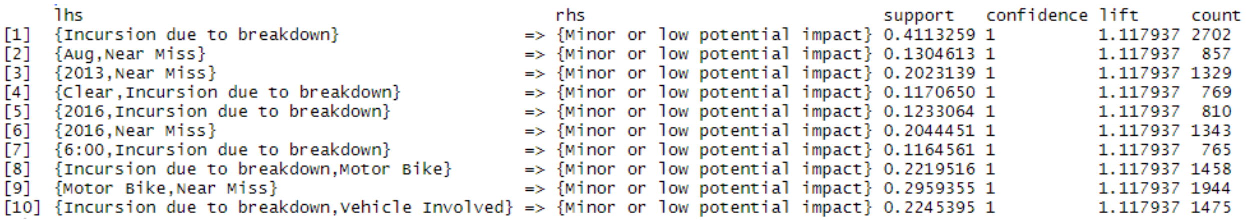
Night



```
     lhs                                                                                                                        rhs     support    confidence lift     count
[1]  {Clear,Incursion to seek information or benefit,Not Filed,Vehicle Not Involved}                                         => {Night} 0.01156949 0.9743590  18.76998  76
[2]  {Clear,Incursion to seek information or benefit,Minor or low potential impact,Not Filed,Vehicle Not Involved}           => {Night} 0.01156949 0.9743590  18.76998  76
[3]  {Clear,Incursion to seek information or benefit,Near Miss,Not Filed,Vehicle Not Involved}                               => {Night} 0.01156949 0.9743590  18.76998  76
[4]  {Clear,Incursion to seek information or benefit,Minor or low potential impact,Near Miss,Not Filed,Vehicle Not Involved} => {Night} 0.01156949 0.9743590  18.76998  76
[5]  {Clear,Incursion to seek information or benefit,Vehicle Not Involved}                                                   => {Night} 0.01202618 0.9404762  18.11727  79
[6]  {Clear,Incursion to seek information or benefit,Minor or low potential impact,Vehicle Not Involved}                     => {Night} 0.01202618 0.9404762  18.11727  79
[7]  {Clear,Incursion to seek information or benefit,Near Miss,Vehicle Not Involved}                                         => {Night} 0.01202618 0.9404762  18.11727  79
[8]  {Clear,Incursion to seek information or benefit,Minor or low potential impact,Near Miss,Vehicle Not Involved}           => {Night} 0.01202618 0.9404762  18.11727  79
[9]  {0:00,Clear,Minor or low potential impact,Not Filed}                                                                    => {Night} 0.01567971 0.8240000  15.87348 103
[10] {0:00,Clear,Near Miss,Not Filed}                                                                                        => {Night} 0.01567971 0.8240000  15.87348 103
> |
```

For “Night”, the support level was set to 0.01 as there weren’t enough observations for higher support levels. Almost all these rules had the variable “Not Filed” in them, suggesting that incidents occurring at night were most likely not reported to the police.

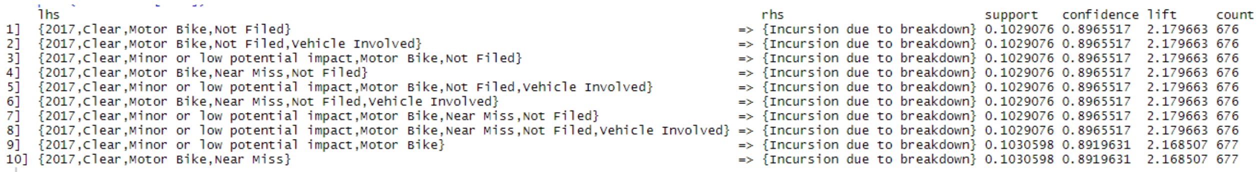
Minor or low potential impact



```
     lhs                                              rhs                              support   confidence lift     count
[1]  {Incursion due to breakdown}                  => {Minor or low potential impact} 0.4113259 1          1.117937 2702
[2]  {Aug,Near Miss}                               => {Minor or low potential impact} 0.1304613 1          1.117937  857
[3]  {2013,Near Miss}                              => {Minor or low potential impact} 0.2023139 1          1.117937 1329
[4]  {Clear,Incursion due to breakdown}            => {Minor or low potential impact} 0.1170650 1          1.117937  769
[5]  {2016,Incursion due to breakdown}             => {Minor or low potential impact} 0.1233064 1          1.117937  810
[6]  {2016,Near Miss}                              => {Minor or low potential impact} 0.2044451 1          1.117937 1343
[7]  {6:00,Incursion due to breakdown}             => {Minor or low potential impact} 0.1164561 1          1.117937  765
[8]  {Incursion due to breakdown,Motor Bike}       => {Minor or low potential impact} 0.2219516 1          1.117937 1458
[9]  {Motor Bike,Near Miss}                        => {Minor or low potential impact} 0.2959355 1          1.117937 1944
[10] {Incursion due to breakdown,Vehicle Involved} => {Minor or low potential impact} 0.2245395 1          1.117937 1475
```

For “Minor or low potential impact” (an event subtype), associations portrayed what kind of accidents were most probably leading to low impacts. Events such as incursion due to breakdown as well as near miss were the ones highly associated with minor impacts. It is interesting to note that two of these rules have motorbikes as an antecedent which makes intuitive sense as motorbikes can respond quickly to such events and manage accordingly and thus, might not be affected by the incident to an extent.

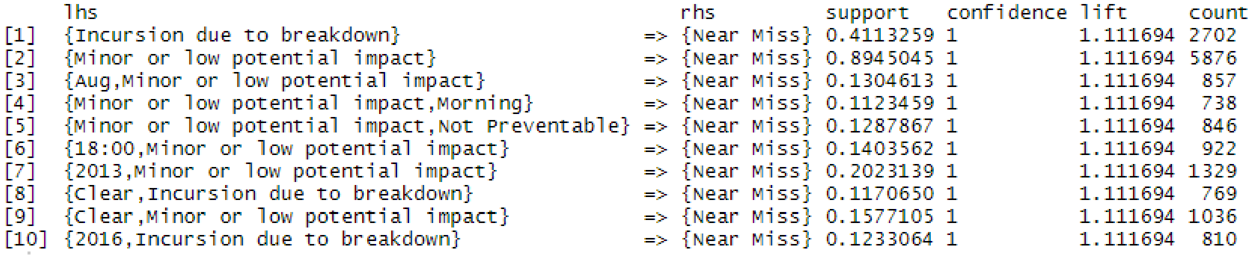
Incursion due to breakdown



```
     lhs                                                                      rhs                           support   confidence lift     count
1]   {2017,Clear,Motor Bike,Not Filed}                                     => {Incursion due to breakdown} 0.1029076 0.8965517  2.179663 676
2]   {2017,Clear,Motor Bike,Not Filed,Vehicle Involved}                    => {Incursion due to breakdown} 0.1029076 0.8965517  2.179663 676
3]   {2017,Clear,Minor or low potential impact,Motor Bike,Not Filed}       => {Incursion due to breakdown} 0.1029076 0.8965517  2.179663 676
4]   {2017,Clear,Motor Bike,Near Miss,Not Filed}                           => {Incursion due to breakdown} 0.1029076 0.8965517  2.179663 676
5]   {2017,Clear,Minor or low potential impact,Motor Bike,Not Filed,Vehicle Involved} => {Incursion due to breakdown} 0.1029076 0.8965517  2.179663 676
6]   {2017,Clear,Motor Bike,Near Miss,Not Filed,Vehicle Involved}          => {Incursion due to breakdown} 0.1029076 0.8965517  2.179663 676
7]   {2017,Clear,Minor or low potential impact,Motor Bike,Near Miss,Not Filed} => {Incursion due to breakdown} 0.1029076 0.8965517  2.179663 676
8]   {2017,Clear,Minor or low potential impact,Motor Bike,Near Miss,Not Filed,Vehicle Involved} => {Incursion due to breakdown} 0.1029076 0.8965517  2.179663 676
9]   {2017,Clear,Minor or low potential impact,Motor Bike}                 => {Incursion due to breakdown} 0.1030598 0.8919631  2.168507 677
10]  {2017,Clear,Motor Bike,Near Miss}                                     => {Incursion due to breakdown} 0.1030598 0.8919631  2.168507 677
```

For "Incursion due to breakdown", more than 3 variables were identified on the lhs which is rare considering the dataset. For such an incident, association at 89% confidence level and 676 cases involved a "Motor bike". The interesting thing here is, this rule, along with all the others, mention how these incidents were not reported to the police ("Not Filed"), which makes sense as people tend to avoid reporting events having a minor or low impact (which has again showed up in other rules associated to incursion due to breakdown)

Near Miss



```
     lhs                                                 rhs         support   confidence lift     count
[1]  {Incursion due to breakdown}                     => {Near Miss} 0.4113259 1          1.111694 2702
[2]  {Minor or low potential impact}                  => {Near Miss} 0.8945045 1          1.111694 5876
[3]  {Aug,Minor or low potential impact}              => {Near Miss} 0.1304613 1          1.111694  857
[4]  {Minor or low potential impact,Morning}          => {Near Miss} 0.1123459 1          1.111694  738
[5]  {Minor or low potential impact,Not Preventable} => {Near Miss} 0.1287867 1          1.111694  846
[6]  {18:00,Minor or low potential impact}            => {Near Miss} 0.1403562 1          1.111694  922
[7]  {2013,Minor or low potential impact}             => {Near Miss} 0.2023139 1          1.111694 1329
[8]  {Clear,Incursion due to breakdown}               => {Near Miss} 0.1170650 1          1.111694  769
[9]  {Clear,Minor or low potential impact}            => {Near Miss} 0.1577105 1          1.111694 1036
[10] {2016,Incursion due to breakdown}                => {Near Miss} 0.1233064 1          1.111694  810
```

For "Near Miss", events were seen that were involved in its occurrence with 100% confidence and 1 > lift ratio. It is evident that incursion due to breakdown was the kind of event that is associated with Near Miss. Other rules highlighted a minor or low potential impact which is obvious as near miss events would most likely pose a potential threat than lead to an actual injury.

Recommendations

From the dataset it can be seen that it is incomplete for many variables hence, one recommendation will be to have complete data for important variables especially such as weather conditions, lighting, number of actual work days lost etc. Moreover, the first priority of the highway authorities should be safety of its employees, contractors and road-users. Thus, after analyzing the data, some recommendations for the authorities are given in order to ensure safety for all.

During upgrading the motorway, instead of vehicles, the focus should be on the employees, contractors and people as this event is not associated with vehicle involvement. Thus, the road should be clear of debris and any sort of obstruction, and have warning and safety signs around the area.

While the job of construction and maintenance is being performed, traffic modeling should be done as most incidents involve a vehicle, especially a motorbike and most incidents involve incursion due to breakdowns and near miss events so it is highly imperative to have adjacent roads to direct traffic towards (these ensure that people do not follow work related vehicles to and reach the site). It is recommend shifting the job to be done more on weekends than weekdays as 70% of the accidents occur on a weekday. Furthermore, SOPs to be established for any job that happens on a weekday.

From the set of rules generated that check associations regarding incursions due to breakdown. Incursion is formally defined as when an unofficial vehicle (e.g. a public vehicle) enters the traffic management site, this depicts that there were no safety signs to begin with to direct the vehicles elsewhere. Thus, it is recommended that signs be placed at least 1 km before where the work site is, which educates people on what particular route to take.

Furthermore, presence of police during night would be preferred so as to ensure that incidents which actually have a harmful impact are reported and there should be collaboration with ambulance emergency services so they could work on the shortest routes to take in case of an emergency.

Employees should be equipped with a safety aid kit in the event of a minor or low potential impact as there is a great likelihood for a motorbike to get injured and roadside assistance providers must be present even if it’s a near miss event where there’s a low potential impact occurring due to a breakdown. This will not only help people but also the employees and contractors in the case of any injuries thus, decreasing the actual lost workdays. Furthermore, employees should themselves be wearing safety gear and complying with the safety protocol in order to avoid any harmful injuries and there should be an appointed officer who would ensure that safety protocols are being followed.

Furthermore, on the basis of the insights generated from visualizing different variables from the data, it is suggested that the following points should be followed for improvement. Firstly, the International Highway Authority should implement more specific standards for reportable accidents and ensure that even low impact accidents are recorded. The company can also maintain a record of “Near Miss” accidents that its workers get into, so that periodic review of the data can illicit prompt response.

In addition, the workers’ shifts should be altered and more work should be done during hours when traffic is light. It is seen that majority accidents take place in 2nd quarter of the day, which is the busiest in terms of traffic. Thus, most construction work should be done either early morning or late at night. Similarly, contractual workers can be hired to work for weekends instead of weekdays because the highway is busier during weekdays. Work should be done when traffic flow is less so that accidents can be avoided.

Moreover, weather forecast should be announced on the Highway Authority's radio info channel and people should be cautioned to drive carefully during such weather. This will help drivers stay more cautious and may help prevent accidents. And, even though majority accidents were of low impact, workers must have proper protective gear including hard hats so that serious damage can be avoided. Lastly, police reporting system needs to be improved in recording all accidents and running detailed checks on construction companies to identify their covering up of accidents to avoid penalties.

References

Agrawal, R., Srikan, R., 1994. Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases, pp. 478–499.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A., 1996. Fast discovery of association rules. In: Fayyad, U.M., Piatetsky Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge and Data Mining. MIT Press, Cambridge MA.

Davenport, T. H., & Harris, J. G. (2007). Competing on analytics: the new science of winning. Harvard Business Press.

Daher, J. R., Chilkaka, S., Younes, A., & Shaban, K. (2016). Association Rule Mining on Five Years of Motor Vehicle Crashes. MATEC Web of Conferences, 81, 02017. doi:10.1051/matecconf/20168102017

Klatt, T., Schläfke, M., &Möller, K. (2011). Integrating business analytics into strategic planning for better performance. Journal of business strategy, 32(6), 30-39.

Linoff, G. S., & Berry, M. J. (2011). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Negash, S. (2004).Business intelligence.The Communications of the Association for Information Systems, 13(1), 54.

Peden, M. (2005). Global collaboration on road traffic injury prevention. International Journal of Injury Control and Safety Promotion, 12(2), 85-91. doi:10.1080/15660970500086130

Schläfke, M., Silvi, R., &Möller, K. (2012). A framework for business analytics in performance management. International Journal of Productivity and Performance Management, 62(1), 110-122.

Siegel, E. (2013). Predictive analytics: The power to predict who will click, buy, lie, or die.John

Wiley & Sons.

X. Xu, Ž. Šarić, A. Kouhpanejade, Traffic&Transportation, Vol. 26, (2014), No. 3, 191- 199 191.

M. Fogue , P.Garrido, F J. Martinez, Juan-Carlos Cano, Carlos T(2012), Advances in Intelligent Systems and Computing, pages 37-46.

S.Krishnaveni , Dr.M.Hemalatha(2011), International Journal of Computer Applications, Volume 23– No.7

R. Nayak, N. iyatrapoomi, J. Weligamage, Proceedings of the 4th World Congress on Engineering Asset Management (WCEAM 2009), (28-30 September 2009).

X F Zhang; L. Fan, Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on, vol., no., pp.1,4, 5-8 (May 2013).