Investigating Contributing Factors under Rainy Weather by Association Rules Mining

Subasish Das (Corresponding author)
PhD. Student
Systems Engineering
University of Louisiana
Lafayette, LA 70504
Email: sxd1684@louisiana.edu
Phone: 225-288-9875
Fax: 337-739-6688

Xiaoduan Sun, Ph.D. & PE
Professor
Civil Engineering Department
University of Louisiana
Lafayette, LA 70504
Email: xsun@louisiana.edu

Total number of words: 7,000
Including 5 Tables and 7 Figures
Submitting to **Analytical Methods in Accident Research**

ABSTRACT

Rainy weather and wet roads are considered hazardous conditions for driving. Countermeasures should be taken to reduce the risks of driving in such conditions, but measuring the added risk and key crash contributing factors under such conditions is very challenging. With a humid subtropical climate, the annual precipitation in Louisiana is about 64 inches, twice above the national average. Approximately 11% of total crashes in Louisiana happened during rainy weather, and nearly 25% of total fatal crashes happen in rainy weather annually. Reducing the number of crashes and crash severity is critical to the state “Zero Deaths Destination” highway safety strategies.

The data mining technique is becoming immensely popular in dealing with huge dataset. It helps to identify the hidden patterns from a large and complex database, which is why these methods are being utilized in diversified areas. There are many data mining techniques that have been applied to traffic crash data analysis in the recent past. However, very little research work utilizes the association rules mining technique to discover knowledge from the traffic crash dataset. This data mining technique generates simple rules that introduce the association between different factors. This paper demonstrates how to apply this data mining methods to discover hidden patterns in rainy weather crash data with eight years of Louisiana data (2004-2011). No dependent variable is developed in this exploratory application contrary to many popular safety performance models. The findings of the research shows that ‘single vehicle run-off crashes’ is the most frequent item in rainy weather. This crash type is particularly associated with a few roadway features like ‘on grade-curve’ aligned roadways, curved roadways, and roadways with no streetlights at night.

In rainy weather, Property Damage Only (PDO) and sideswipe (same direction) crashes are also significant in numbers. Moderate injuries are dominant in single vehicle crashes. Roadways with poor illumination are associated with straight level aligned roadways in rainy weather crashes. Young drivers (15-24) are vulnerable in run-off crashes when the roadways had poor illumination and are curve-aligned. The findings will help the highway authorities to determine countermeasure selection and safety improvement.

Key words: road safety, data mining, association rules, market basket analysis, hidden patterns.

1. INTRODUCTION

Rainy weather and wet roads are considered as hazardous conditions for driving. Due to the visual obstruction from rainfall and loss of surface friction, most vehicles slow down during rainfall but crashes still occur, which may be associated with the conditions. Measuring the added risk and identifying key crash contributing factors during showery weather has been challenging. The total reported number of traffic crashes in rainy weather in Louisiana is 11,398 in 2011 and 10,204 in 2010, respectively. Approximately, 11% of yearly total crashes in Louisiana happened during rainfall and nearly 25% of total fatal crashes in Louisiana happened in rainy weather in the state. To reach "Destination Zero Deaths" set by Louisiana Highway Safety Strategies, it is critical to reduce the number of crashes and crash severities under the rainy condition.

There are several ways to identify crash risk factors. The parametric models work well if the assumptions and model format are accurate to reflect the underlying relationships between dependent and independent variables. Violation of any assumption could lead to flawed or at least inadequate estimations. Specific nonparametric data mining techniques have been receiving increased attention from researchers in traffic safety because of no-predefined assumptions. One of the data mining techniques that has not been explored fully for crash data analysis is association rules. This method is concerned with the identification of interesting patterns from massive data.

The main objective of this study is to utilize the association rules mining technique in identifying the pattern of the associated factors in traffic crashes during rainy weather.

2. LITERATURE REVIEW

Previous research on the usage of the data mining approach in traffic safety research is limited but diversified. Among the data mining methods, decision trees, neural networks, factor analysis, probabilistic graphical dependency models, association rules, correspondence analysis, cluster analysis, text mining, relational learning models and sequence mining have been used as the popular mining techniques.

In the field of traffic crash research, some studies applied different types of data mining techniques to analyze crash rates and severity problems. Dia and Rose employed real-world traffic data to develop a multi-layered neural network freeway incident detection model [1]. Mussone et al. used neural networks to examine vehicle crashes that occurred at intersections in Italy [2]. Sohn and Shin applied classification trees, neural networks and logistic regression methods to classify severity-related factors by using crash data from Korea [3]. Park and Saccomanno applied recursive partitioning method to stratify Canadian roadway-railway crossings into homogeneous classes of crossings using control variables of highway class, track type and track number [4]. Chan and Marco presented an exemplar set of data mining techniques based on real-world traffic monitoring. The methodology was illustrated by two case studies, collision warning for left-turn across-path scenarios and signal violation. [5]. In their paper, Bayam et al. explored the needs of data mining techniques in the studies on senior drivers [6]. The research of Salim et al. aimed to investigate an integration of intelligent software agents and ubiquitous data stream mining, for a novel context-aware framework to monitor an intersection to learn for patterns of crashes and elements leading to a crash and to learn to recognize potential hazards in intersections from information communicated by road infrastructures, approaching and passing vehicles, and external entities [7]. Najjar and Mandavilli utilized artificial neural network techniques and developed crash rate prediction models for Kansas road networks [8]. Six networks had been studied and crash prediction models for each network were developed. Prato et al. intended to comprehend which data mining techniques appear more suitable for the objective of providing a broad picture of the road safety situation and individuating specific problems that the allocation of resources should address first [9]. Descriptive (i.e., K-means and Kohonen clustering) and predictive (i.e., decision trees, neural networks and association rules) data mining techniques were implemented for the analysis of traffic accidents to see the advantages and limitations of the methodologies. By using the Classification and Regression Tree (CART) Kashani et al. analyzed the crash data pertaining to the last three years (2006-2008) in Iran [10]. The variable selection procedure was carried out on the basis of Variable Importance Measure (VIM) which is one of the CART method outputs. The results revealed that lack of seat belt usage, improper overtaking and speeding were the significant factors associated with injury severity.

Only a few studies examined association rules mining in traffic crash analysis in past. In his study, Marukatat applied association rules to real traffic crash data collected from local police stations. This study found out candidate rules that offered some insight into the phenomena of safety improvement [11]. Song-bai et al. conducted multidimensional association rules model of traffic crashes for the freeways of China. This study also presented preventive measures to reduce crashes [12]. Pande and Abdel-Aty used association rules mining in safety analysis by generating closely associated crash characteristics in the form of rules [13].

Although number of studies have employed various data mining approaches in traffic safety research, generating significant rules with good visualization technique is missing. The National Traffic Safety Board (NTSB) reasoned that the risk of a fatal accident, nationwide, was about 3.9 to 4.5 times greater on wet pavement than on dry pavement [14]. So, this particular case (usage of rainy weather crash data) should be investigated in-depth to find the possible risks, circumstances or combination of factors. This study serves as a starting point to demonstrate the use of association rules mining to determine significant contributing rules that could present useful insight to the potential safety and traffic operation performance.

3. METHODOLOGY

3.1 Association rules mining

Data mining is the process of identifying valid and understandable patterns in the data set. It helps in extracting and refining valuable knowledge from large data sets. Data mining involves machine learning, statistical knowledge, modeling concepts and database management. The methods can be classified into two main sections: descriptive and predictive. Association rules mining, a descriptive analytics, discovers significant rules showing variable category conditions that occur frequently together in a dataset.

Many algorithms can be used to discover association rules from data to extract useful patterns. Apriori algorithm is one of the most widely used and famous techniques for finding association rules [15]. Due to the explorative and eloquent nature, intelligible representation and visualization of the found patterns and models are essential for the successful mining process to make the results easy to understand. One important feature of the technique is that no variables are assigned as dependent or independent. The apriori algorithm for searching association rules is easy to interpret and the computations used are straightforward.

A frequent itemset generation algorithm digs out frequently occurring itemsets, subsequences, or arrangements from large data sets. Frequent itemsets mining has been applied in many branches of science, for example, social science, web mining and bioinformatics. A set of definitions are given here before demonstrating the method with an example. Let $I = \{i_1, i_2, \ldots i_m\}$ be a set of items (e.g., a set of crash categories for a particular crash record) and $C = \{c_1, c_2, \ldots, c_n\}$ be a set of database crash information (transaction) where each crash record c_i contains a subset of items chosen from I. A set of items is referred to as an itemset. An itemset that contains k items is considered as a k-itemset.

Taking the first five crash records from a crash dataset as an example shown in Table 1, there are seven variables to describe each crash characteristics, which is called a 7-itemset. The occurrence frequency of an itemset is known as 'support' count. It is the number of transactions that contain this particular itemset. Transaction width is defined as the number of items included in the transaction.

In association rules mining (also known as market basket analysis), the item-wise details, such as the quantity or cost from a particular variable category (for example crash cost of a fatal crash vs. crash cost of a 'no injury' crash) are usually ignored. Consequently, the items are represented as binary variables whose value is one when the item is present in a particular crash record and zero otherwise.

The relative minimum support threshold is fixed to 40%, which is equivalent with the minimum support count of 2 out of 5 in sample data. Any itemset with less than 2 crashes are discarded.

TABLE 1 Example of Five Crash Records

No.	Day_of_Week	Alignment	Lighting	Collision_Type	Severity	Driver_Gender	Driver_Age
1	Weekday	Straight-Level	Dark- Continuous Street Light	Rear End	No Injury	Female	35-44
2	Weekday	Straight-Level	Daylight	Rear End	Fatal	Male	15-24
3	Weekend	Curve-Level	Daylight	Sideswipe-Same Direction	No Injury	Male	25-34
4	Weekday	Curve-Level	Daylight	Right Angle	No Injury	Male	15-24
5	Weekday	Straight-Level	Dark with No Street Light	Rear End	Complaint	Female	15-24

Firstly 1-itemset from the sample data is calculated. For instance, the variable 'Day_of_Week' has value 'Weekday' in three out of five crash records, which gives 60% support. Results of 1-itemsets evaluation are listed in the Table 2. At the second iteration, 2-itemsets would be searched. According to the apriori principle all non-empty subsets of a frequent itemset must also be frequent. This means that all the 1-itemset included in 2-itemsets must satisfy the minimum support threshold. Similarly the 3rd, 4th, 5th 6th and 7th iterations (k-itemset rules) are generated subsequently.

Association rules mining extract important rules from a given frequent itemset. An association rule is an implication of the form $X => Ij$, where X is a set of some items, and Ij is a single item that is not present in X. The rule $X => Ij$ holds in the set of crash records C with confidence c if at least c percent of crash records in C that contain X also contain I_j . For example, one rule {Driver_Age=15-24} => {Alignment=Straight-Level} from the data in Table 1 is considered. The left side of the rule {Driver_Age=15-24} is antecedent and the right side {Alignment=Straight-Level} is consequent. The association rules are usually assessed in terms of support and confidence. In mathematical terms, support $= P(X \cap I)$. The combination (Driver_Age=15-24, Alignment=Straight- Level) is present twice out of five crash records. So, the support of the above rule is 0.4. Confidence of a rule is the ratio of the support count of the consequent and the support count of the antecedent. In mathematical terms, confidence $= P(X \cap I) / P(X)$. Hence the confidence for the rule {Driver_Age=15-24} => {Alignment=Straight-Level} is (2/5)/(3/5) = 0.67. Confidence measures the consistency of the inference derived from an association rule. It's important to note that the association does not necessarily mean causality between the items in the antecedent and

TABLE 2 1-itemset Table from Table 1 data

Variable Category	Count	Support	Threshold	Accept/Discard
Weekday	3 out of 5	0.60	0.40	Accept
Weekend	2 out of 5	0.40	0.40	Accept
Straight-Level	3 out of 5	0.60	0.40	Accept
Curve-Level	2out of 5	0.40	0.40	Accept
Dark- Continuous Street Light	1 out of 5	0.20	0.40	Discard
Daylight	3 out of 5	0.60	0.40	Accept
Dark with No Street Light	1 out of 5	0.20	0.40	Discard
Rear End	3out of 5	0.60	0.40	Accept
Sideswipe-Same Direction	1 out of 5	0.20	0.40	Discard
Right Angle	1 out of 5	0.20	0.40	Discard
No Injury	3 out of 5	0.60	0.40	Accept
Fatal	1 out of 5	0.20	0.40	Discard
Complaint	1 out of 5	0.20	0.40	Discard
Female	2 out of 5	0.40	0.40	Accept
Male	3 out of 5	0.60	0.40	Accept
35-44	1 out of 5	0.20	0.40	Discard
15-24	3 out of 5	0.60	0.40	Accept
25-34	1 out of 5	0.20	0.40	Discard

consequent of a rule. Depending on the data, association rules mining algorithms may produce masses of rules, for which one may need to also use other interestingness measures besides support and confidence. In this study, the rules are ranked according to lift measure, which computes the ratio between the rule's confidence and the support of the itemset appearing in the rule consequent. In mathematical terms the 'lift' is defined as (by using example data):

$$\begin{aligned} Lift &= \frac{P(Alignement = Straight-level \mid Driver\_Age = 15-24)}{P(Alignement = Straight-level)} \\ &= \frac{P(Driver\_Age = 15-24 \cap Alignement = Straight-level)}{P(Alignement = Straight-level) * P(Driver\_Age = 15-24)} \\ &= \frac{\frac{2}{5}}{\frac{3}{5} \times \frac{3}{5}} = 1.11 \end{aligned}$$

The numerator measures the observed frequency of the co-occurrence of the items in the antecedent (*X*) and the consequent (*I*) of the rule. The denominator measures the expected frequency of the co-occurrence of the items in the antecedent and the consequent of the rule under the assumption of conditional independence. The more this ratio differs from one, the stronger the dependency becomes. It is desirable for rules to have a large confidence factor, high level of support, and a lift value greater than 1. Since some events of interest in traffic safety analysis are very rare (for example, fatal crashes); the support for some rules of interest could be quite low. It essentially means that lift value is more important for determining strength of an association rule than the other two criteria. Hence, in the present application the rules should be evaluated based on the 'lift' values. The rules 'discovered' by the algorithm still need to have support greater than a minimum threshold.

3.2 Data Analysis

To identify important contributing factors for rainy weather crashes in Louisiana, a large dataset containing eight years of crash records (2004-2011) was obtained from the LADOTD. The data was stored as an unsorted and unmanageable format in Microsoft Access database tables. Every crash record has many variables and the detailed information is stored in separate data tables such as crash, driver and vehicle. Cross-table in-depth analysis would be useful to find out hidden crash patterns and association rules mining provides the best analysis tool to explore these patterns.

From the preliminary analysis, it has been found that quite a few fatalities and many injuries occurred in rainy weather as shown in Figure 1.

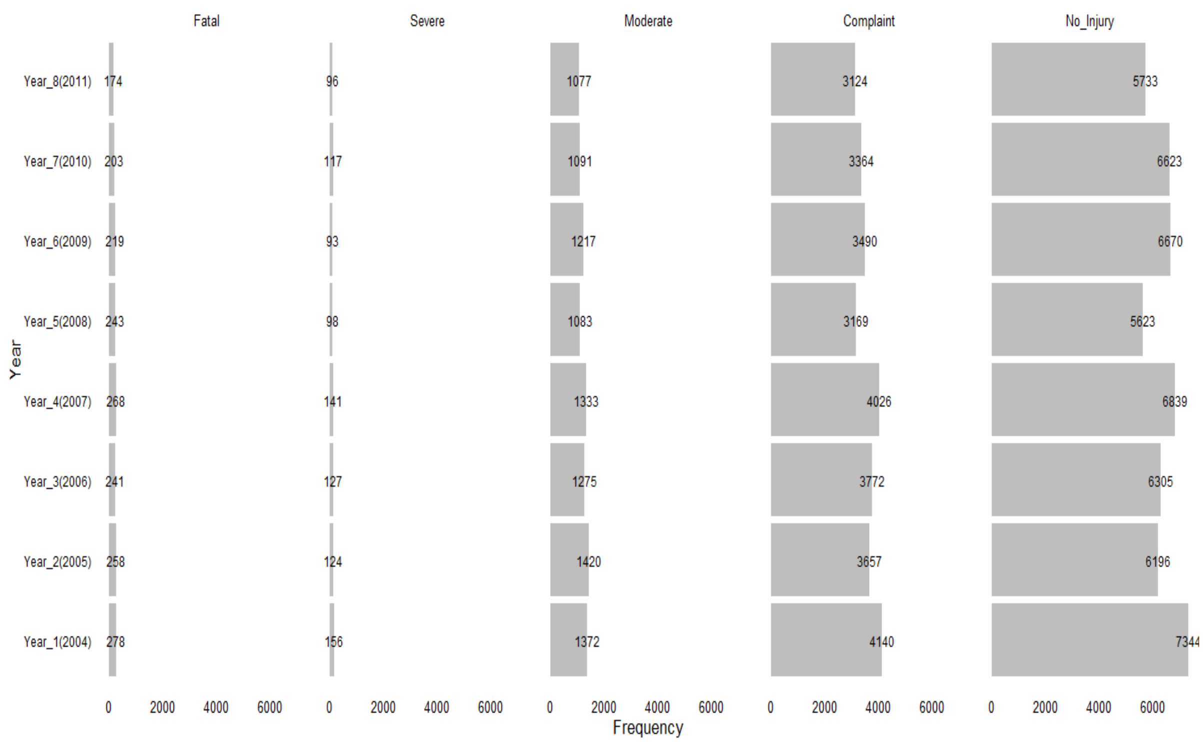


FIGURE 1 Distribution of rainy weather crash severities by year.

Traffic crash databases contain many variables which can complicate the data mining process. The crash records with missing and questionable information were removed. For example, 300 crashes have driver age listed as 200, which were removed. The final database contains 58,288 crash records with selected number of variables. The variable selection method used correlation matrix as a selection platform. The general distribution of the rainy weather crashes is listed in Table 3.

To focus on the meaningful analysis, a set of key variables are selected such as the information on crash timing (day of the week), roadway characteristics (alignment, lighting), human factor (driver gender and age) and crash characteristics (crash severity and collision type). For the association rules, there are various settings required to be altered for significant findings. The minimum support and confidence are essential to generate the important rules. After a significant number of trials and errors, the minimum support for the rules was considered as 1% with the minimum confidence of 60%. One percent of minimum support means that no item or set of items will be considered frequent for the first analysis if it does not appear in at least 583 traffic crashes (1% of total 58,288 crash records). It may be rather argued that the choices for the values of these parameters are subjective, which is partly true. However a trial and error experiment indicates that setting minimum support too low will result in exponential growth of the number of items in the frequent item sets. In contrast, by choosing a support parameter too high, the algorithm will be capable of generating a small number of rules. The minimum confidence value of 60% indicates that a rule is considered reliable when the consequent of the rule occurs at least six out of ten times that the antecedent appears. By choosing different confidence values, a trial and error experiment showed that this parameter value gives rather stable results concerning the amount of rules generated by the algorithm. The purpose of post-processing the association rules set is to identify the subset of interesting rules in a generated set of noteworthy rules.

TABLE 3 Distribution of Rainy Weather Crashes by Key Variables

Categories	Frequency	Percentage	Categories	Frequency	Percentage
Alignment			**Collision Type**		
Straight-Level	39,546	67.85%	Rear End	11,350	19.47%
Curve-Level	11,835	20.30%	Right Angle	5,096	8.74%
Straight-Level-Elevated	750	1.29%	Sideswipe- Same Direction	1,962	3.37%
On Grade-Straight	1,999	3.43%	Single Vehicle	32,357	55.51%
On Grade_Curve	2,231	3.83%	Left Turn- Opposite Direction	1,316	2.26%
Curve-Level-Elevated	783	1.34%	Left Turn- Angle	1,869	3.21%
Hillcrest-Straight	801	1.37%	Left Turn- Same Direction	755	1.30%
Hillcrest-Curve	230	0.39%	Head-on	1,108	1.90%
Dip, Hump-Straight	94	0.16%	Right turn- Opposite Direction	127	0.22%
Dip, Hump-Curve	19	0.03%	Right Turn- Same Direction	382	0.66%
Lighting			Sideswipe- Opposite Direction	1,966	3.37%
Daylight	34,660	59.46%	**Day of Week**		
Dark- Continuous Street Light	1,473	2.53%	Weekend	26,141	44.85%
Dark- No Street Light	19,170	32.89%	Weekday	32,147	55.15%
Dark- Street Light at Intersection Only	974	1.67%	**Gender**		0.00%
Dusk	957	1.64%	Male	37,542	64.41%
Dawn	1,054	1.81%	Female	20,746	35.59%
Driver Age			**Severity**		0.00%
15-24	18,977	32.56%	No Injury	32,613	55.95%
25-34	13,107	22.49%	Complaint	18,983	32.57%
35-44	9,501	16.30%	Moderate	5,913	10.14%
45-54	8,121	13.93%	Severe	345	0.59%
55-64	4,780	8.20%	Fatal	434	0.74%
64-75	2,289	3.93%			
75 plus	1,513	2.60%			

4. RESULTS AND DISCUSSION

The association rules were generated in this study by using ‘arules’ package in software R [16, 17]. The primary analysis demonstrates that the dataset has 58,288 rows with 43 items. The frequency of the items, generated by association mining, is shown in Figure 2. The top five frequent items in the dataset are Alignment=Straight-level, Driver Gender=Male, Lighting=Daylight, Severity= No Injury, and Collision Type=Single Vehicle. The frequency of the rules generated for different itemsets and the statistics of support, confidence and lift

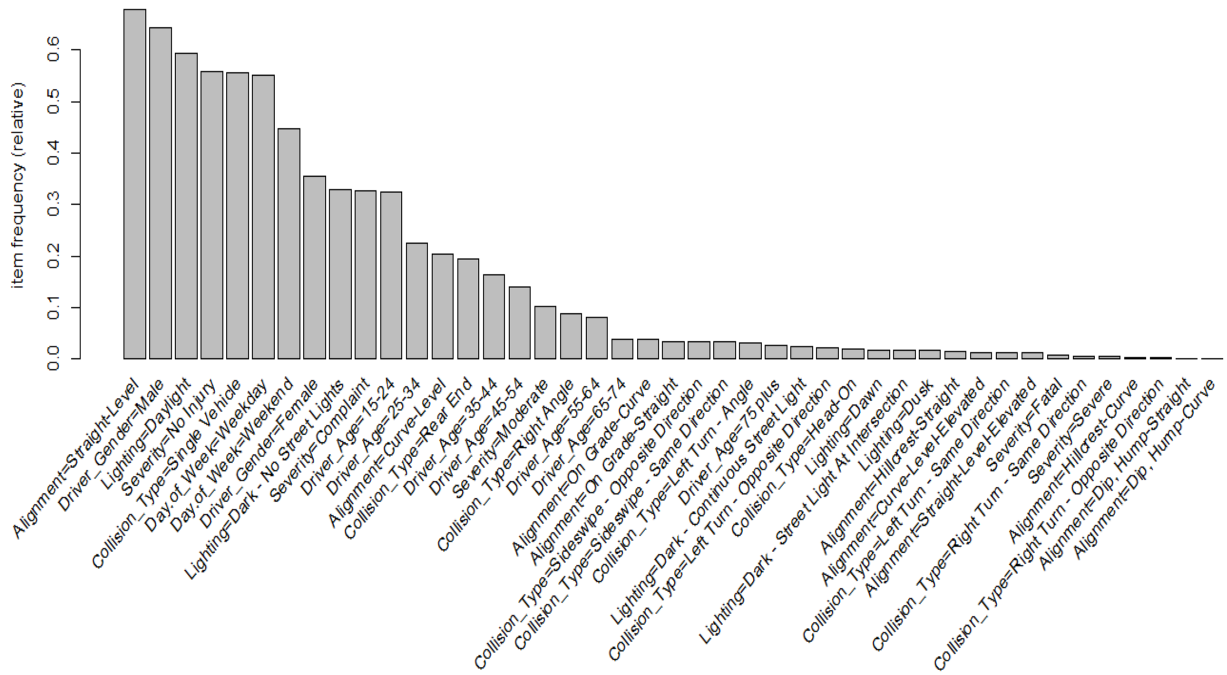


FIGURE 2 Item frequency plot.

TABLE 4 Summary Chart of the Association Rules Mining

Support$_{min}$=1%, Cofidence$_{min}$=60%				Lift ≥1								
				Support			Confidence			Lift		
Minlen	**Maxlen**	**Rules (all)**	**Rules(Lift≥1)**	**Min.**	**Mean**	**Max.**	**Min.**	**Mean**	**Max.**	**Min.**	**Mean**	**Max.**
1	1	2	2	0.644	0.661	0.678	0.644	0.661	0.678	1.000	1.000	1.000
2	2	81	59	0.010	0.109	0.412	0.600	0.715	0.849	1.002	1.147	1.482
3	3	423	330	0.010	0.054	0.257	0.601	0.722	0.922	1.001	1.176	1.660
4	4	871	692	0.010	0.033	0.152	0.600	0.721	0.939	1.000	1.190	1.691
5	5	767	622	0.010	0.022	0.092	0.601	0.723	0.939	1.001	1.207	1.693
6	6	217	179	0.010	0.015	0.036	0.600	0.720	0.928	1.000	1.217	1.934
7	7	6	6	0.011	0.011	0.012	0.631	0.740	0.843	1.072	1.249	1.471
All	All	2,367	1,890	0.010	0.034	0.678	0.600	0.722	0.939	1.000	1.194	1.934

are displayed in Figure 2 and Table 4, respectively. Rules are created for two phases. When the minimum threshold of the lift is selected as one, the total counts of the rules are lowered. For 1-itemset value, only two rules are produced; for 7-itemsets value, total counts of the rules are 6. The maximum number of rules is generated for 4-itemsets. In most cases, it's very difficult to interpret itemsets with higher numbers. In this study, the result analysis was limited to 4-itemsets for easier interpretation. The first fifty rules of 2-itemsets are listed in

TABLE 5 First Fifty Rules for 2-itemsets

No.	Antecedent		Consequent	Supp.	Cofid.	Lift	Count
1	Alignment=On Grade-Curve	>	Collision_Type=Single Vehicle	0.031	0.823	1.482	1807
2	Lighting=Dark - No Street Lights	>	Collision_Type=Single Vehicle	0.268	0.816	1.47	15621
3	Alignment=Curve-Level-Elevated	>	Collision_Type=Single Vehicle	0.011	0.803	1.447	641
4	Alignment=Curve-Level	>	Collision_Type=Single Vehicle	0.162	0.799	1.439	9443
5	Collision_Type=Rear End	>	Lighting=Daylight	0.16	0.822	1.383	9326
6	Driver_Age=75 plus	>	Lighting=Daylight	0.021	0.811	1.364	1224
7	Collision_Type=Left Turn - Same Direction	>	Lighting=Daylight	0.01	0.803	1.35	583
8	Collision_Type=Sideswipe - Same Direction	>	Severity =No Injury	0.025	0.745	1.332	1457
9	Collision_Type=Left Turn - Angle	>	Lighting=Daylight	0.025	0.786	1.322	1457
10	Collision_Type=Sideswipe - Same Direction	>	Lighting=Daylight	0.026	0.778	1.308	1515
11	Collision_Type=Right Angle	>	Lighting=Daylight	0.068	0.773	1.301	3964
12	Collision_Type=Left Turn - Opposite Direction	>	Lighting=Daylight	0.017	0.76	1.278	991
13	Driver_Age=65-74	>	Lighting=Daylight	0.029	0.746	1.255	1690
14	Collision_Type=Left Turn - Angle	>	Alignment=Straight-Level	0.027	0.85	1.252	1574
15	Collision_Type=Right Angle	>	Alignment=Straight-Level	0.074	0.848	1.25	4313
16	Collision_Type=Rear End	>	Alignment=Straight-Level	0.165	0.847	1.249	9618
17	Collision_Type=Left Turn - Same Direction	>	Alignment=Straight-Level	0.011	0.846	1.247	641
18	Collision_Type=Left Turn - Opposite Direction	>	Alignment=Straight-Level	0.019	0.843	1.243	1107
19	Collision_Type=Sideswipe - Same Direction	>	Alignment=Straight-Level	0.028	0.837	1.234	1632
20	Lighting=Dawn	>	Collision_Type=Single Vehicle	0.012	0.65	1.171	699
21	Driver_Age=75 plus	>	Alignment=Straight-Level	0.02	0.778	1.147	1166
22	Collision_Type=Sideswipe - Opposite Direction	>	Lighting=Daylight	0.023	0.68	1.144	1341
23	Collision_Type=Sideswipe - Opposite Direction	>	Driver_Gender=Male	0.025	0.732	1.137	1457
24	Collision_Type=Rear End	>	Day.of_Week=Weekday	0.122	0.627	1.137	7111
25	Severity=Moderate	>	Collision_Type=Single Vehicle	0.064	0.626	1.128	3730
26	Driver_Age=55-64	>	Lighting=Daylight	0.055	0.668	1.124	3206
27	Lighting=Dark - Continuous Street Light	>	Alignment=Straight-Level	0.019	0.76	1.12	1107
28	Lighting=Dark - Street Light At Intersection	>	Alignment=Straight-Level	0.013	0.753	1.109	758
29	Driver_Age=65-74	>	Alignment=Straight-Level	0.029	0.747	1.101	1690
30	Day.of_Week=Weekday	>	Lighting=Daylight	0.355	0.643	1.081	20692
31	Collision_Type=Rear End	>	Severity=No Injury	0.118	0.604	1.08	6878
32	Driver_Gender=Female	>	Lighting=Daylight	0.228	0.64	1.077	13290
33	Collision_Type=Head-On	>	Driver_Gender=Male	0.013	0.692	1.075	758
34	Lighting=Dark - No Street Lights	>	Driver_Gender=Male	0.226	0.687	1.066	13173
35	Lighting=Dark - Continuous Street Light	>	Driver_Gender=Male	0.017	0.685	1.064	991
36	Lighting=Dawn	>	Driver_Gender=Male	0.012	0.682	1.059	699
37	Severity=Moderate	>	Driver_Gender=Male	0.069	0.677	1.052	4022
38	Severity=No Injury	>	Alignment=Straight-Level	0.398	0.712	1.049	23199
39	Driver_Age=45-54	>	Driver_Gender=Male	0.094	0.674	1.047	5479
40	Severity=Complaint	>	Lighting=Daylight	0.011	0.674	1.046	641
41	Lighting=Dark - Street Light At Intersection	>	Driver_Gender=Male	0.203	0.622	1.046	11832
42	Driver_Age=55-64	>	Alignment=Straight-Level	0.058	0.709	1.045	3381
43	Driver_Age=55-64	>	Driver_Gender=Male	0.055	0.672	1.043	3206
44	Alignment=On Grade-Straight	>	Lighting=Daylight	0.021	0.618	1.04	1224
45	Collision_Type=Sideswipe - Same Direction	>	Driver_Gender=Male	0.022	0.667	1.036	1282
46	Alignment=Curve-Level	>	Driver_Gender=Male	0.134	0.661	1.027	7811
47	Day.of_Week=Weekend	>	Driver_Gender=Male	0.296	0.661	1.026	17253
48	Alignment=On Grade-Curve	>	Driver_Gender=Male	0.412	0.694	1.022	24015
49	Lighting=Daylight	>	Alignment=Straight-Level	0.025	0.658	1.022	1457
50	Alignment=Straight-Level	>	Lighting=Daylight	0.412	0.608	1.022	24015

Table 5. From the first four rules, it is seen that rainy weather crashes involving either curve-aligned roadways or dark with no street lights are mostly single vehicle Run-Off-Road (ROR) crashes. When looking at the interpretation of the association rules, the top 13 rules listed in Table 5 express that the top rules with higher lift mostly relate to collision type and crash temporal characteristics (daylight). The variable category single vehicle crash is found as ‘consequent’ for top four rules. This means that most of the crashes in rainy weather directs to run-off crashes due to visual obstruction. Rule number 2 {Lighting= Dark- No Street Lights} => {Collision Type= Single Vehicle} has a high lift and higher support values by generating 15,621 crashes happened under this rule. The consequent of this rule verifies that single vehicle crashes under rainy weather are closely associated with roadway segments without light at night. Rules 14 to 19 in Table 5 indicate the relationship between collision type and straight-level crashes. Nearly 68% of the rainy weather crashes happened in straight-level roadways. As the variable ‘alignment’ is skewed to a particular category (straight-level), more rules are associated with this category. The lift values for different types of collisions are almost the same. Among these rules, {Collision Type= Rear End} => {Alignment= Straight-level} displays a higher support value (0.165) with 9,618 frequency count. Middle age (45-54) to older drivers (55 plus) are seen to be crash-prone in daytime rather than nighttime in rainy weather. Female drivers are seen in fewer rules than the male drivers. Male drivers are associated with various variable categories with higher lift value than the female drivers. In hazardous, rainy weather, one particular rule {Lighting= Dark- No Street Light} => {Driver_Gender= Male} shows a higher support value of 0.203 with a frequency of 11,832.

Many interesting rules are observed in Table 6 for 3-itemsets rules. Young male drivers are almost absent in 2-itemsets rules. But this age group appears frequently at 3-itemsets rules. Rule 1 {Alignment=On Grade-Curve, Lighting=Dark - No Street Lights} => {Collision Type= Single Vehicle} has the highest lift value. Curve-aligned roadways and roadways with poor illumination are frequently visible in the antecedents of the rules. Young drivers associated with curve-aligned roadways and poorly illuminated roadways results in single vehicle run-off crashes in rainy weather. The antecedents {Lighting=Dark - No Street Lights, Driver_Gender=Female} and {Lighting=Dark - No Street Lights, Driver_Gender=Male} both resulted in single vehicle run-off crashes. The rule associated with female driver has a higher lift value than the male drivers, but the support value is higher in the rules for male drivers with a frequency count of 10,712.

For the 4-itemsets rules (Table 7), the rule {Alignment=Curve-Level, Lighting=Dark - No Street Lights, Driver_Age=35-44} => {Collision Type= Single Vehicle} has the highest lift value. The next rule is associated with {Driver_Age=15-24} in place of {Driver_Age=35-44}. The second rule has higher support value with frequency of 1,626 than the first rule. Another interesting rule is {Alignment=Curve-Level, Lighting=Dark - No Street Lights, Driver_Gender=Female} => {Collision Type= Single Vehicle}. The similar rule associated with male drivers has a lower lift value with higher support. These rules indicated that curve-level and poorly illuminated roadways are crash-prone areas in rainy weather. The age-group vulnerable for the combination of these two characteristics is young male drivers. The consequent {Collision Type= Single Vehicle} is present in all top 35 rules for 4-itemsets. Although out of 692 rules, consequent {Alignment= Straight-level} is the most frequent (nearly 22% in the rules).

TABLE 6 Association Rules for 3-itemsets

No.	Antecedent		Consequent	Supp.	Conf	Lift	Count
1	Alignment=On Grade-Curve, Lighting=Dark - No Street Lights	>	Collision_Type=Single Vehicle	0.013	0.922	1.660	752
2	Alignment=Curve-Level, Lighting=Dark - No Street Lights	>	Collision_Type=Single Vehicle	0.073	0.921	1.659	4,237
3	Day.of_Week=Weekend, Alignment=On Grade-Curve	>	Collision_Type=Single Vehicle	0.016	0.851	1.533	907
4	Alignment=On Grade-Curve, Driver_Age=15-24	>	Collision_Type=Single Vehicle	0.012	0.848	1.528	675
5	Lighting=Dark - No Street Lights, Driver_Age=35-44	>	Collision_Type=Single Vehicle	0.048	0.836	1.506	2,807
6	Alignment=Curve-Level, Severity=Moderate	>	Collision_Type=Single Vehicle	0.022	0.832	1.499	1,311
7	Alignment=Curve-Level, Driver_Age=15-24	>	Collision_Type=Single Vehicle	0.059	0.832	1.499	3,418
8	Lighting=Dark - No Street Lights, Driver_Age=25-34	>	Collision_Type=Single Vehicle	0.067	0.830	1.495	3,921
9	Alignment=On Grade-Curve, Driver_Gender=Male	>	Collision_Type=Single Vehicle	0.021	0.829	1.494	1,218
10	Lighting=Dark - No Street Lights, Driver_Gender=Female	>	Collision_Type=Single Vehicle	0.085	0.822	1.481	4,935
11	Lighting=Dark - No Street Lights, Driver_Age=15-24	>	Collision_Type=Single Vehicle	0.092	0.820	1.477	5,353
12	Lighting=Dark - No Street Lights, Driver_Gender=Male	>	Collision_Type=Single Vehicle	0.184	0.814	1.466	10,712
13	Lighting=Dark - No Street Lights, Driver_Age=45-54	>	Collision_Type=Single Vehicle	0.037	0.811	1.461	2,172
14	Alignment=On Grade-Curve, Driver_Gender=Female	>	Collision_Type=Single Vehicle	0.011	0.811	1.461	618
16	Alignment=Curve-Level, Driver_Age=35-44	>	Collision_Type=Single Vehicle	0.028	0.809	1.457	1,606
17	Alignment=Curve-Level, Driver_Age=25-34	>	Collision_Type=Single Vehicle	0.038	0.807	1.454	2,216
18	Alignment=Curve-Level, Driver_Gender=Male	>	Collision_Type=Single Vehicle	0.108	0.801	1.443	6,266
21	Collision_Type=Right Angle, Driver_Gender=Female	>	Alignment=Straight-Level	0.030	0.858	1.264	1,743
22	Collision_Type=Left Turn - Angle, Driver_Gender=Male	>	Alignment=Straight-Level	0.017	0.852	1.256	1,017
23	Collision_Type=Rear End, Driver_Age=15-24	>	Alignment=Straight-Level	0.059	0.851	1.255	3,438
24	Collision_Type=Rear End, Driver_Age=25-34	>	Alignment=Straight-Level	0.037	0.851	1.255	2,160
25	Collision_Type=Left Turn - Opposite Direction, Driver_Gender=Male	>	Lighting=Daylight	0.010	0.746	1.255	588
26	Collision_Type=Right Angle, Severity=Complaint	>	Alignment=Straight-Level	0.029	0.851	1.254	1,685
27	Collision_Type=Right Angle, Driver_Gender=Male	>	Lighting=Daylight	0.039	0.745	1.254	2,284
30	Collision_Type=Rear End, Driver_Gender=Female	>	Alignment=Straight-Level	0.064	0.849	1.252	3,731
31	Collision_Type=Right Angle, Severity=No Injury	>	Alignment=Straight-Level	0.036	0.849	1.252	2,096
32	Day.of_Week=Weekend, Collision_Type=Rear End	>	Alignment=Straight-Level	0.062	0.849	1.251	3,595
33	Collision_Type=Rear End, Severity=No Injury	>	Alignment=Straight-Level	0.100	0.846	1.247	5,806
34	Day.of_Week=Weekday, Collision_Type=Rear End	>	Alignment=Straight-Level	0.103	0.846	1.247	6,021
35	Lighting=Daylight, Collision_Type=Right Angle	>	Alignment=Straight-Level	0.057	0.846	1.247	3,334

TABLE 7 Association Rules for 4-itemsets

No.	Antecedent	Consequent	Supp.	Conf	Lift	Count
1	Alignment=Curve-Level, Lighting=Dark - No Street Lights, Driver_Age=35-44	Collision_Type=Single Vehicle	0.013	0.939	1.691	749
2	Alignment=Curve-Level, Lighting=Dark - No Street Lights, Driver_Age=15-24	Collision_Type=Single Vehicle	0.028	0.930	1.675	1,626
3	Alignment=Curve-Level, Lighting=Dark - No Street Lights, Severity=No Injury	Collision_Type=Single Vehicle	0.037	0.928	1.673	2,128
4	Alignment=Curve-Level, Lighting=Dark - No Street Lights, Driver_Gender=Female	Collision_Type=Single Vehicle	0.021	0.926	1.668	1,209
5	Alignment=Curve-Level, Lighting=Dark - No Street Lights, Driver_Gender=Male	Collision_Type=Single Vehicle	0.052	0.919	1.656	3,028
6	Alignment=Curve-Level, Lighting=Dark - No Street Lights, Severity=Complaint	Collision_Type=Single Vehicle	0.025	0.919	1.655	1,467
7	Alignment=Curve-Level, Lighting=Dark - No Street Lights, Driver_Age=25-34	Collision_Type=Single Vehicle	0.018	0.917	1.652	1,026
8	Lighting=Dark - No Street Lights, Severity=No Injury, Driver_Age=35-44	Collision_Type=Single Vehicle	0.030	0.872	1.572	1,738
9	Lighting=Dark - No Street Lights, Driver_Gender=Female, Driver_Age=35-44	Collision_Type=Single Vehicle	0.016	0.860	1.549	957
10	Lighting=Dark - No Street Lights, Severity=No Injury, Driver_Age=25-34	Collision_Type=Single Vehicle	0.040	0.856	1.541	2,317
11	Day.of_Week=Weekend, Alignment=On Grade-Curve, Driver_Gender=Male	Collision_Type=Single Vehicle	0.011	0.854	1.538	618
12	Day.of_Week=Weekend, Lighting=Dark - No Street Lights, Severity=No Injury	Collision_Type=Single Vehicle	0.082	0.852	1.535	4,751
13	Day.of_Week=Weekend, Alignment=Curve-Level, Driver_Age=15-24	Collision_Type=Single Vehicle	0.030	0.851	1.533	1,762
14	Lighting=Dark - No Street Lights, Severity=No Injury, Driver_Age=45-54	Collision_Type=Single Vehicle	0.024	0.850	1.532	1,380
15	Lighting=Dark - No Street Lights, Severity=No Injury, Driver_Gender=Female	Collision_Type=Single Vehicle	0.051	0.848	1.527	2,977
16	Lighting=Dark - No Street Lights, Severity=No Injury, Driver_Gender=Male	Collision_Type=Single Vehicle	0.108	0.846	1.523	6,320
17	Day.of_Week=Weekend, Lighting=Dark - No Street Lights, Driver_Age=35-44	Collision_Type=Single Vehicle	0.024	0.842	1.518	1,428
18	Day.of_Week=Weekday, Lighting=Dark - No Street Lights, Severity=No Injury	Collision_Type=Single Vehicle	0.078	0.840	1.514	4,546
19	Alignment=Curve-Level, Severity=Complaint, Driver_Age=15-24	Collision_Type=Single Vehicle	0.022	0.839	1.512	1,294
20	Lighting=Dark - No Street Lights, Driver_Gender=Female, Driver_Age=25-34	Collision_Type=Single Vehicle	0.021	0.838	1.510	1,213
21	Lighting=Dark - No Street Lights, Severity=No Injury, Driver_Age=15-24	Collision_Type=Single Vehicle	0.051	0.837	1.508	2,957
22	Alignment=Curve-Level, Severity=Complaint, Driver_Age=25-34	Collision_Type=Single Vehicle	0.014	0.837	1.507	829
23	Day.of_Week=Weekend, Alignment=Curve-Level, Driver_Age=35-44	Collision_Type=Single Vehicle	0.013	0.836	1.507	782
24	Day.of_Week=Weekend, Lighting=Dark - No Street Lights, Driver_Age=25-34	Collision_Type=Single Vehicle	0.036	0.836	1.506	2,102
25	Day.of_Week=Weekend, Alignment=Curve-Level, Severity=Moderate	Collision_Type=Single Vehicle	0.012	0.835	1.504	694
26	Lighting=Dark - No Street Lights, Severity=No Injury, Driver_Age=55-64	Collision_Type=Single Vehicle	0.011	0.833	1.500	642
27	Alignment=Curve-Level, Driver_Gender=Female, Driver_Age=15-24	Collision_Type=Single Vehicle	0.022	0.832	1.499	1,264
28	Alignment=Curve-Level, Driver_Gender=Male, Driver_Age=15-24	Collision_Type=Single Vehicle	0.037	0.832	1.499	2,154
29	Day.of_Week=Weekday, Alignment=Curve-Level, Severity=Moderate	Collision_Type=Single Vehicle	0.011	0.829	1.494	617
30	Day.of_Week=Weekday, Lighting=Dark - No Street Lights, Driver_Age=35-44	Collision_Type=Single Vehicle	0.024	0.829	1.494	1,379
31	Day.of_Week=Weekend, Alignment=Curve-Level, Severity=Complaint	Collision_Type=Single Vehicle	0.030	0.828	1.492	1,726
32	Alignment=On Grade-Curve, Severity=No Injury, Driver_Gender=Male	Collision_Type=Single Vehicle	0.010	0.827	1.490	589
33	Lighting=Dark - No Street Lights, Driver_Gender=Male, Driver_Age=25-34	Collision_Type=Single Vehicle	0.046	0.826	1.488	2,708
34	Day.of_Week=Weekend, Lighting=Dark - No Street Lights, Driver_Age=45-54	Collision_Type=Single Vehicle	0.019	0.826	1.487	1,099
35	Alignment=Curve-Level, Severity=Moderate, Driver_Gender=Male	Collision_Type=Single Vehicle	0.016	0.826	1.487	932

In general, the rules presented above indicate the possible associations for crashes in rainy weather. By analyzing the rules of 2, 3 and 4-itemsets, few variable categories were discovered as the dominant factors for rainy weather crashes. The factors should be carefully investigated by the traffic agencies for considering appropriate countermeasures to avoid crashes and crash severities in hazardous roadways.

An excellent visual representation of the results is also an integral part of the data mining process. Visualization of the results of the association rules is facilitated by the use of ‘arulesViz’ package [18]. Inspecting all 1,890 rules manually is not a viable option. A straight-forward visualization of association rules involves a scatter plot (Figure 3) with two interest measures on the axes: the support values of the rules are on the x-axis and the lift values are on the y-axis. The color of the points is used to indicate the confidence value of each rule. The rules with lower support and lower lift values are higher in frequencies than the rules of higher support and higher lift values.

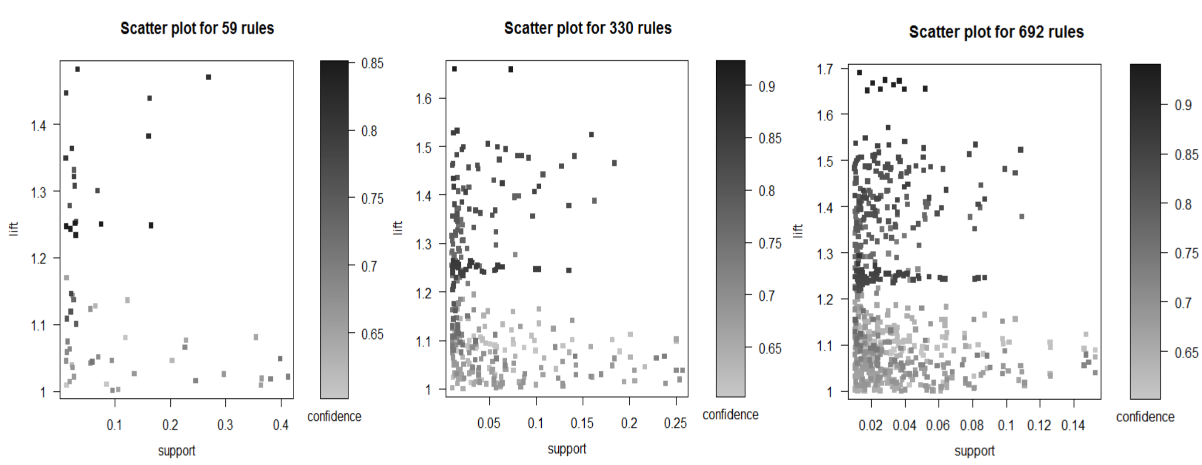


FIGURE 3 Scatter plot of the generated rules (a) 2-itemsets, (b) 3-itemsets, (c) 4-itemsets.

Matrix-based visualization methods organize the antecedent and consequent itemsets on the x and y axes, respectively. Selected interest measures can be displayed at the joint of the antecedent and consequent for a given rule. An antecedent/consequent combination with no rule keeps the joint area blank. By considering the set of association rules

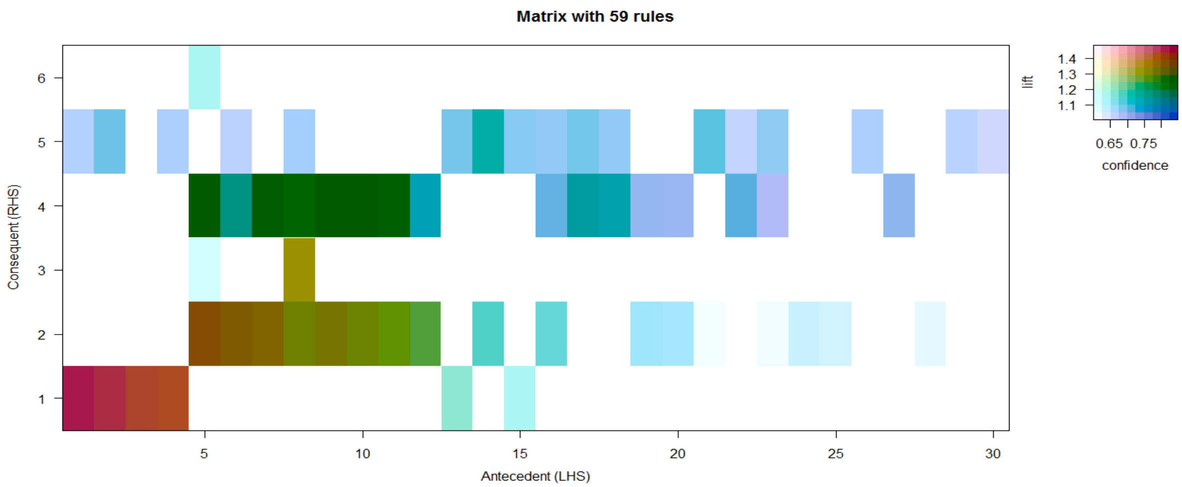
$$\mathrm{M} = \{(a_1, c_1, q_1), \ldots (a_i, c_i, q_i), \ldots (a_n, c_n, q_n)\}$$

where, a_i is the antecedent, c_i is the consequent and q_i is the selected interest measure for the i-th rule for i = 1,…, n.

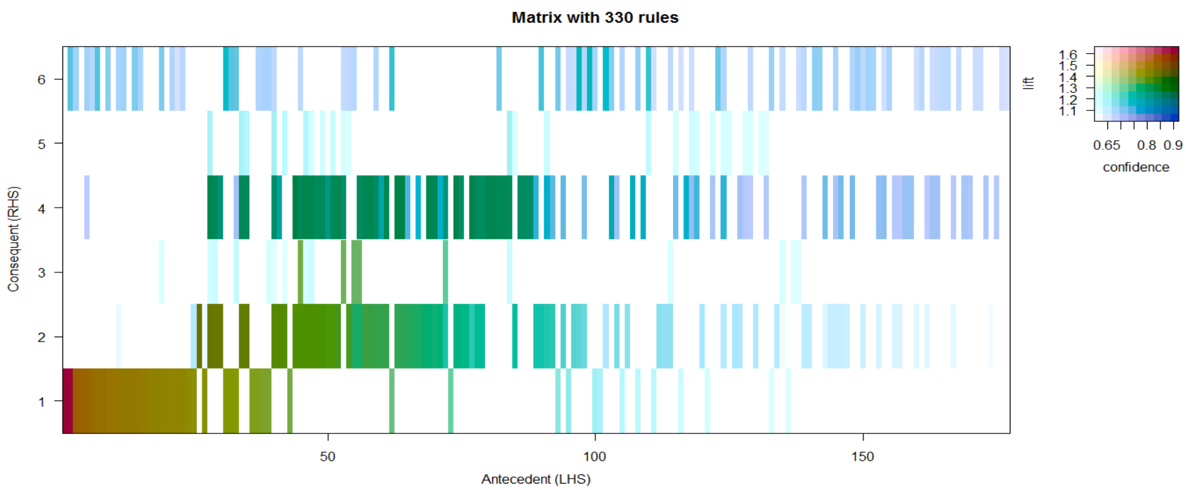
Suppose in the visualization matrix **M**, the set of K unique antecedents and L unique consequents are identified. A '$L \times K$' matrix **M** with one column for each unique antecedent and one row for each unique consequent was also done. Finally, the matrix was populated by setting $M_{ik}= m_i$ for i = 1, …, n and l and k corresponding to the position of a_i and c_i in the matrix. **M** also contains blank cells because many potential association rules will not meet the minimum thresholds for support and confidence. The matrix-based visual plots are illustrated in Figure 4. This figure gives a general idea of how the consequents are varied based on the lift and confidence value with respect to the antecedents.

Figure 4 reveals the easier visualization of large sets of association rules from the grouped matrix. Balloon plots are drawn with antecedent groups as columns and consequents as rows (for 2, 3 and 4-itemsets). The color of the balloons represents the lift value and the size of the balloon shows the aggregated support. The number of antecedents and the most important (frequent) items in the group are displayed as the labels for the columns. The dominant consequents for different itemsets are single vehicle, daylight, straight-aligned roadways, and male drivers. As the dominant rules were discussed before, similar rules easily identified from Figure 4 are not discussed again.

The study shows that the application of association rules mining in a specific environmental condition can help to reveal how drivers' behavior, roadway conditions and



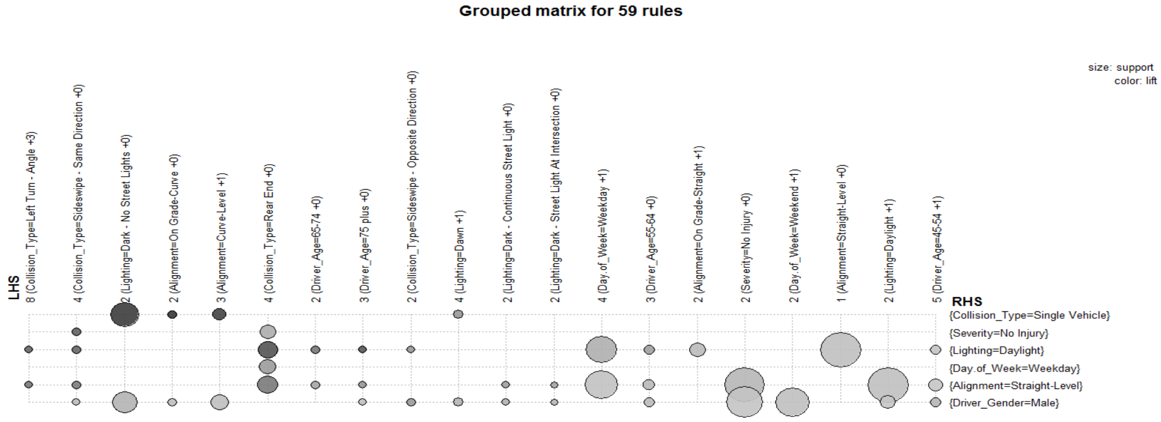
(a) 2-itemsets



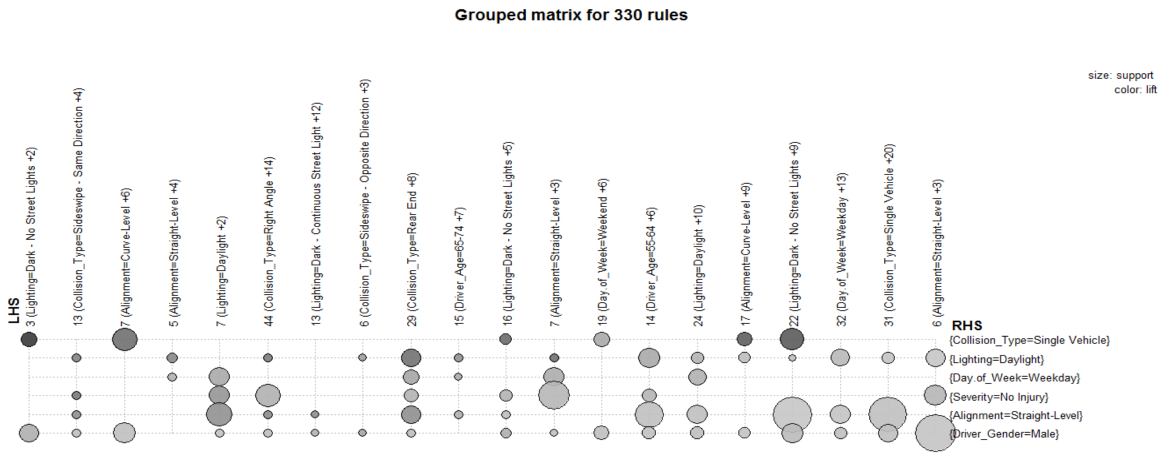
(b) 3-itemsets

(c) 4-itemsets

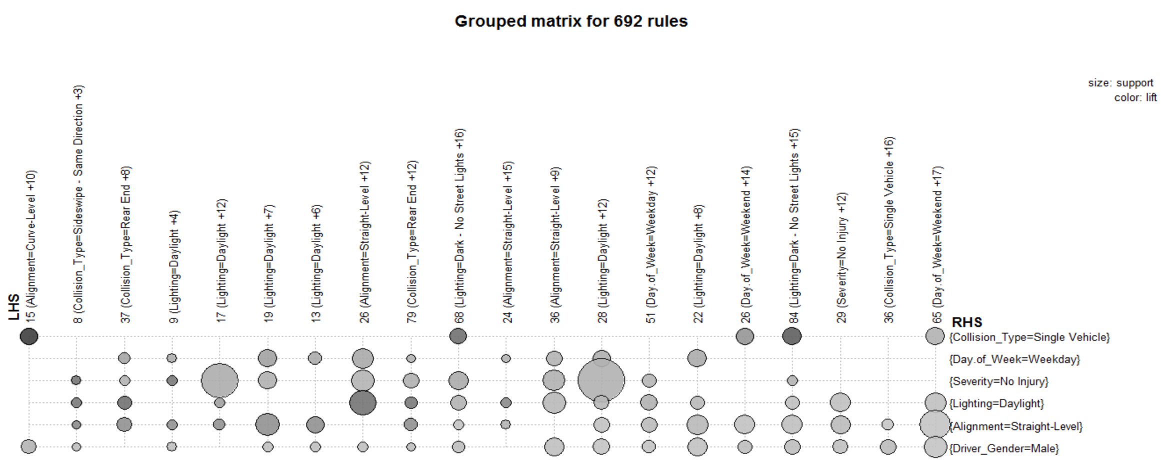
FIGURE 4 Matrix plot for the generated rules.



(a) 2-itemsets



(b) 3-itemsets



(c) 4-itemsets

FIGURE 5 Grouped matrix plot for the generated rules.

crash temporal characteristics are associated with different collision types and injury severity. These findings are expected to be useful for policy makers to develop better safety policies.

5. CONCLUSION

This paper presented a preliminary investigation on how association rules mining techniques can be used to extract knowledge from the traffic crash data under a particular environmental condition. Some interesting findings are observed for crashes in rainy weather. Some of the findings verify the general perceptions on such types of crashes and a few findings are quite surprising.

The most significant single variable category for the situation is found as single vehicle ROR crashes. This crash type is predominant for the rainy weather crashes in the presence of other roadway features such as on grade-curve aligned roadways, curved roadways, and roadways with no streetlights at night. In rainy weather, PDO and sideswipe (same direction) crashes are significant in numbers. For drivers age 55 and above, most of the crashes during rainy weather happen in daylight. Moderate injuries are also dominant in single vehicle crashes. Roadways with poor illumination are associated with straight level aligned roadways for many rainy weather crashes. Young drivers (15-24) are vulnerable in ROR crashes when the roadways have poor illumination and are curve-aligned.

By observing the potential patterns in the discovered rules, the results can provide valuable insights into the underlying relationships between risk factors and crashes under particular conditions. The exploration on the association rules mining might provide a better understanding of the risk factors. The results from this study is considered to be in use in future investigations with data mining researches in roadway safety.

6. REFERENCES

[1] Dia, H., and Rose, G. Development and Evaluation of Neural Network Freeway Incident Detection Models Using Field Data. Transportation Research Part C, Vol. 5(5), pp. 313-331, 1997.

[2] Mussone, L., Ferrari, A., and Oneta, M. An analysis of urban collisions using an artificial intelligence model. Accident Analysis & Prevention, Vol. 31(6), pp. 705-718, 1999.

[3] Sohn, S.Y., and Shin, H. Pattern recognition for road traffic accident severity in Korea. Ergonomics, Vol. 44 (1), pp. 107–117, 2001.

[4] Park, Y.J., and Saccomanno, F.F. Collision Frequency Analysis Using Tree-Based Stratification. Journal of the Transportation Research Record, Vol. 1908, National Research Council, Washington, D.C., pp. 121-129, 2005.

[5] Chan, C. and Marco, D. Traffic Monitoring at Signal-Controlled Intersections and Data Mining for Safety Applications. 2004 IEEE Intelligent Transportation Systems Conference, Washington, D.C., USA. October 36, 2004.

[6] Bayam, E., Liebowitz, J. and Agresti, W. Older drivers and accidents: A meta-analysis and data mining application on traffic accident data. Expert Systems with Applications. Volume 29, pp. 598–629, 2005.

[7] Salim, F.D., Krishnaswamy, S., Loke, S.W., and Rakotonirainy, A. Context-Aware Ubiquitous Data Mining Based Agent Model for Intersection Safety. Embedded and Ubiquitous Computing (EUC) Workshops, pp. 61 – 70, 2005.

[8] Najjar, Y. and Mandavilli, S. Data Mining the Kansas Traffic-Crash Database. Kansas Department of Transportation. Report No. K-TRAN: KSU-05-6. August, 2009.

[9] Prato, C.G., Bekhor, S., Galtzur, A., Mahalel, D., and Prashker, J. N. Exploring the potential of data mining techniques for the analysis of accident patterns. The 12th World Conference on Transportation Research. Lisbon, Portugal, July 11-15, 2010.

[10] Kashani, A.T., Shariat-Mohaymany, A., and Ranjbari, A. A Data Mining Approach to Identify Key Factors of Traffic Injury Severity. PROMET – Traffic & Transportation. Volume 23, No. 1, pp. 11-17, 2011.

[11] Marukatat, R. Structure-Based Rule Selection Framework for Association Rule Mining of Traffic Accident Data. Computational Intelligence and Security: Lecture Notes in Computer Science, Vol. 4456, pp. 231-239, 2007.

[12] Song-bai, H., Wang Y., Sun, and Y., Gao, W. The Research of Multidimensional Association Rule in Traffic Accidents, 4th International Conference on Wireless Communications, Networking and Mobile Computing, 12-14 October, 2008.

[13] Pande, A., and Abdel-Aty, M. Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. Safety Science, Vol. 47, pp. 145–154, 2009.

[14] Broadsky, H., and Hakkert, A.S. Risk of Road Accident in Rainy Weather. Accident Analysis & Prevention, Vol. 20(3), pp. 161-176, 1988.

[15] Agrawal, R., Imielinski, T., and Swami, A. N. Data mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering, Vol. 5(6), pp. 914–925, 1993.

[16] R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-

[17] Hahsler, M., Buchta, C., Gruen, B., and Hornik, K. arules: Mining Association Rules and Frequent Itemsets. http://cran.r-project.org/web/packages/arules/arules.pdf. Accessed July 18, 2013.

[18] Hahsler, M., and Chelluboina, S. arulesViz: Visualizing Association Rules and Frequent Itemsets. http://cran.r-project.org/web/packages/arulesViz/arulesViz.pdf. Accessed July 18, 2013.