Traffic Violation Data Analysis Using R and Hadoop

[1]G. NIRANJANA, [2]Dr. R. VIJAYABHANU
[1]Former PG Student,Department of Computer Science, Avinashilingam Institute of Home Science and Higher Education For Women , Coimbatore, India,
[2]Assistant professor, Department Of Computer Science, Avinashilingam Institute of Home Science and Higher Education For Women, , Coimbatore, India,

Abstract:Traffic violations and offences are becoming more and more serious as the traffic volume increasing, which may bring property damage and threaten personal safety. In 2013, approximately 32,000 people died in traffic crashes in the USA. Almost 90 people on average lose their lives every day and more than 250 are injured every hour. Traffic crashes can be reduced by enhancing the Road safety. Traffic data volumes are increasing very fast and huge amount of data are generated, which could not be handled with traditional tools like excel or RDBMS. Big Data analytics tools helps us to gain useful insights to enhance road safety and decrease traffic crashes. In this paper, Hadoop and R will be used to analyse and mine the data. Analyses are done using HIVE and interpolating them using R. After doing analyses, the goal is to predict the accidents from the type of violation and other factors using different regression techniques and also to find the relation/Association between different factors on violations and accidents. The aforementioned analyses will be helpful for the decision makers and practitioners to develop new traffic rules and policies, in order to prevent accidents, and increase roadway safety.

IndexTerms **- Big Data analytics, R tool, Data mining techniques, Regression techniques.**

I. INTRODUCTION

1.1 Problem Definition

The purpose of this study is to provide quick and effective method to analyse the accidents from the different type of violations and other factors. Depending upon those factors we can predict which type of violation has contributed more to accident. The association between major factors are identified. Various data mining techniques are applied for the prediction.

1.2 Overview

As the traffic volume is increasing day by day, traffic violations are becoming more serious which may bring property damage and threaten personal safety. Almost 90 people on average lose their lives every day and more than 250 are injured every hour. Road safety can be enhanced by decreasing the traffic crashes. Traffic data volumes are increasing very fast and huge amount of data are generated, which could not be handled with traditional tools like excel or RDBMS. Big Data analytics toolssuch as Hadoop can help us gain useful insights to enhance road safety and decrease traffic crashes. In this paper, Hadoop and R will be used to analyse and mine the data. Summary analysis is done in HIVE using SQL aggregations and then interpolating them using R. The chi-square method is used for finding the relationship between influencing factors on traffic violation data set. After doing analysis, the goal is to predict the accidents from the type of violation and other factors using different regression techniques like logistic regression and decision tree are applied to predict the accidents based on different violations. The above mentioned analysis will be helpful for the decision makers and practitioners to develop new traffic rules and policies, in order to prevent accidents, and increase roadway safety.

II. SYSTEM ANALYSIS

The system study is comparative study of existing system and proposed system in order to analyse the development. It helps to understand the functions of the existing system.

Existing system lacks the capability to analyse high-throughput traffic monitoring stream and detect various types of violations. It cannot analyse and compare large amount of traffic violation data and cannot predict accurate results. Right now, system is using excel to view and analyse data but it will be very slow when multiple years of data is there. Even RDBMS like Oracle can't be used for futuristic purpose.

This system can discover all the violations from the high-throughput traffic monitoring stream. This system analysis helps in understanding violations from different perspective. Hadoop system is used to load and analyse the huge amounts of data. Summary analysis is done in HIVE based on different parameters. Data exploration and finding association between different factors using chi-square method is done using R tool.By using R tool different regressions techniques like logistic regression and decision tree is used to predict the accidents in traffic violations.

2.1 Exploratory Data Analysis

The first step of any analysis is the Exploratory Data analysis i.e. exploring the dataset through box plot, bar plot, histogram and graphs. This analysis helps in finding the major violation factors influencing the accident but it cannot find whether the violation factors are dependent or independent regarding accidents. Chi-square method is used to check whether the two factors are dependent on each or not. This system uses two methods univariant and bivariant for data exploration and visualization.

Univariant

In this, the system considers only one factor for data exploration. The following points mentioned below shows how each factor has influenced the violation.

1. The system used ggplot to explore the vehicle type and its violations count and it observed that automobile vehicles are more prone to accidents are shown in **Fig.4**
2. The alcohol count has been mentioned in the following **Fig.5**
3. The gender count has been mentioned in the following **Fig.6** and the output ggplot shows that male violation count is dominating.
4. The personal injury counts have been mentioned in the following **Fig.7**
5. The race count has been mentioned in the following **Fig.8** and the output ggplot shows that white race people are leading in violation counts
6. The belt count has been mentioned in the following **Fig.9**
7. The system used ggplot to explore the manufacturer count and it is clear from the plot that Toyota is in leading in violation counts are shown in **Fig.10**
8. The top colors that make violations are shown in the **Fig.11**
9. The date wise violation counts are shown in the **Fig.12** and it is clear from the graph that 2016 is peak in violation count
10. The violation reading regarding the time are shown in **Fig.13** and the plot clearly represents that violation counts are high in noon and night
11. The following **Fig.14** represents the counts of property damage

Bivariant

In this, the system compares accident with two factors. The following points mentioned below

1. The vehicle type and color are plotted using ggplot and from the following **Fig.15** it is clear that automobile with black color are more contributed to accident
2. The following **Fig.16** shows the alcohol count based on gender and time and the plot shows that count of alcohol consumption in male is high during night time.
3. The following **Fig.17** shows the accident count based on gender and time and the plot shows that count of accident caused by male is more during night time.
4. The following **Fig.18**represents the accident count based on gender and alcohol and the plot shows that accident count of alcohol consumption is more in male
5. The following **Fig.19** indicates that the accident count is high when personal injury and alcohol consumption is yes in both.

Chi-Square Test

The chi-square method is used to find whether the two factors are fully dependent on each other or not. If p value is <0.05 then they are dependent or vice versa.

1. The accident and personal injury are tested and the p value is <0.05 so they are dependent factors.
2. The accident and vehicle type are tested and the p value is <0.05 so they are dependent factors
3. The accident and race are tested and the p value is>0.05 so they are less dependent factors
4. The accident and belts are tested and the p value is >0.05 so they are less dependent factors
5. The accident and alcohol are tested and the p value is <0.05 so they are dependent factors
6. The accident and commercial vehicle are tested and the p value is>0.05 so they are less dependent factors

2.2 Logistic Regression

The first step is to split the dataset into two chunks: training and testing set. The training set will be used to fit our model and tested over the testing set. After fitting the model, the glm score distribution is done and is shown in the **Fig.20**. The system then plots the ROC curve and calculates the AUC (area under the curve) which are typical logistic regression performance measurements. The ROC is a curve is shown in the **Fig.21** is generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings while the AUC is the area under the ROC curve. As a rule of thumb, a model with good predictive ability should have an AUC closer to 1 than to 0.5. The AUC value is 0.850283872189296, which indicates that the model prediction is good.

2.3 Decision Tree

Decision tree is a type of supervised learning algorithm that has a pre-defined target variable. The decision tree is a graph used to represent choices and their results in form of a tree. The graph has nodes and edges, nodes represent an event or choice and the edges of the graph represent the decision rules or conditions. Decision trees can be generated through the **rpart** package. The Rpart tree performance is represented in the **Fig.22** shows that the prediction is good. The decision tree has taken the alcohol as the target variable and predicted that the alcohol is the major factor which contributed to the accident is shown in the **Fig.23**.

2.4 Database Design

Database Design is a crucial factor in the performance of a system, both in terms of system timings and in the case with which the system can be maintained or modified. It permits simple retrieval of data in response to query and requests. It also supports the maintenance of data through updates, insertions and deletions. Hadoop is an open-source software framework for

storing massive data and running applications on clusters of commodity hardware. It provides enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. The traffic violation dataset is downloaded from the data.gov website and it is loaded into Hadoop for storage and processing. Hive tool is used for Database creation. The table consists of 30 fields which are therein the dataset. Size of the dataset is 12.7 MB. Filtering is also done at back end for removing duplicates and a new table is created and stored in the Hadoop by using hive tool.

2.4.1 Table Design

Table 1 Traffic violation

FIELD NAME	DATA TYPE	FIELD NAME	DATA TYPE
Dateofstop	String	Vehicletype	String
Timeofstop	String	Year	String
Agency	String	Make	String
Description	String	Model	String
location	String	Color	String
Belts	String	Violationtype	String
Personalinjury	String	Charge	String
propertydamage,	String	Article	String
Commerciallicense	String	Contributedtoaccident	String
Hazmat	String	Race	String
Commercialvehicle	String	Gender	String
Alcohol	String	Drivercity	String
Accident	String	Driverstate	String
Workzone	String	Dlstate	String
State	String	Arresttype	String

The above Table 1 consists of 30 fields or factors that contribute to the traffic violations.

III. SYSTEM DEVELOPMENT

3.1 Data Collection

The data collection is the process of gathering the data from the website. This dataset has been obtained from Data.gov, which is the home of the U.S. Government's open data. This data contains the entries of all electronic traffic violations including date, time, location details, violation description and type, detailed information about vehicle, gender and race of violator etc. The data is captured from year 2012 – 2017. To get the input data, download the dataset from the website https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q/data.The downloaded dataset is loaded and pre-processed before used for analysis.

3.2 Data Loading & Cleaning

3.2.1 Loading the data set

Initially the data is in CSV format. The collected CSV format data is loaded into the Hadoop system using Hive tool. Creating a table in hive with all the 30 fields, which are there in the dataset.

3.2.2 Pre-processing

Data cleansing or data cleaning is the process of detecting and correcting corrupt or inaccurate records from a dataset and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the unwanted data. Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data. The cleansed data can be analysed efficiently. The data which are not necessary for analysis are removed using filtering method. In Hadoop Hive tool is used for pre-processing. Checking for the duplicates and the null values is done in hive using queries and the new table is created with the proper data which are used for further analysis.

3.3 Analysis of the Summary

Finding the counts and percentages based on major factors like by state, year, violation type, injury, alcohol, race, vehicle type etc., using Hive SQL aggregations. The counts and the percentage of accidents are referred in the **Fig.2** and **Fig.3**.

3.4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analysing data. Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a dataset. Data visualization is used for data exploration because it allows users to quickly and simply view most of the relevant features of their dataset. From this step, users can identify variables that are likely to have interesting observations. The charts and plots are used for data exploration. The main goals of EDA are

1. Maximize insight into a dataset.
2. Uncover underlying structure.

Exploratory Data Analysis is done through graphs and plots for the better visualization of the data. Here the system uses two methods univariant and bivariant for data exploration and visualization. In univariant, the system considers only one factor for data exploration and in bivariant the system compares accident with two factors. This analysis helps in finding the major violation factors influencing the accident but it cannot find whether the violation factors are dependent or independent regarding accidents.

Finding Association Using Chi-Square Method

This chi-square test is done to check whether the two factors are really dependent or not. If p value is <0.05 then they are dependent or vice versa. The dependent factors are identified using this test.

3.5 Prediction

Predicting the accidents based on different factors using regression techniques like logistic regression and Decision tree. Logistic regression is mainly used for predicting the binary outcome from a set of continuous predictor variables. Like other linear models, generalized linear model is also very easy to fit in R. The generalized linear model glm() function is called and the fitting process is not so different from the one used in linear regression. Before performing logistic regression, the first step is to split the dataset into two chunks: training and testing set. The training set will be used to fit our model and tested over the testing set. After fitting the model, the system plots the ROC curve and calculates the AUC (area under the curve) which are typical logistic regression performance measurements. The ROC is a curve is shown in the Fig 25 generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings while the AUC is the area under the ROC curve. As a rule of thumb, a model with good predictive ability should have an AUC closer to 1 than to 0.5. The AUC value is 0.850283872189296 which indicates that the model prediction is good. Decision tree is a type of supervised learning algorithm that has a pre-defined target variable. Classification and regression trees - CART can be generated through the **rpart** package.

The packages that are used for analysis are:

1. Caret
2. Rpart
3. Rpart.plot

Steps are in building the decision tree are:

1. Include all variables
2. Perform the splitting
3. Cross validation
4. Traverse maximum depth of the tree
5. Plot the tree

The decision tree is a graph used to represent choices and their results in form of a tree. The graph has nodes and edges, nodes represent an event or choice and the edges of the graph represent the decision rules or conditions. The decision tree has taken the alcohol as the target variable and predicted that the alcohol is the major factor which contributed to the accident.

IV. RESULTS AND DISCUSSION

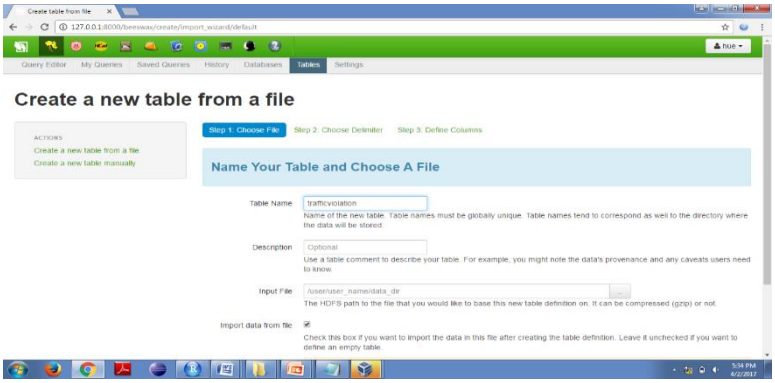


Figure 1 Loading the traffic violation dataset into Hadoop

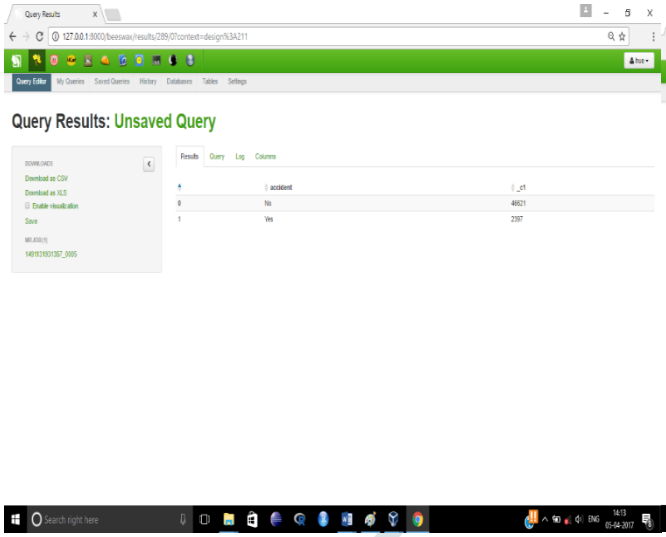


Figure 2 Finding counts of accident in summary analysis

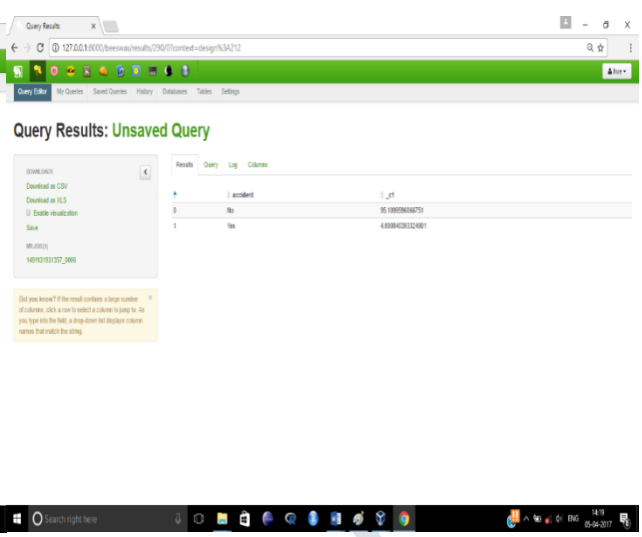


Figure 3 Finding percentage of accident in summary analysis

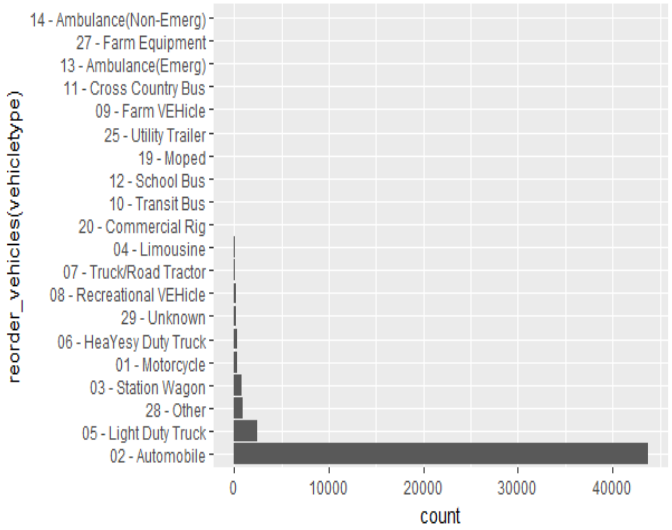


Figure 4 Vehicle type - violation count

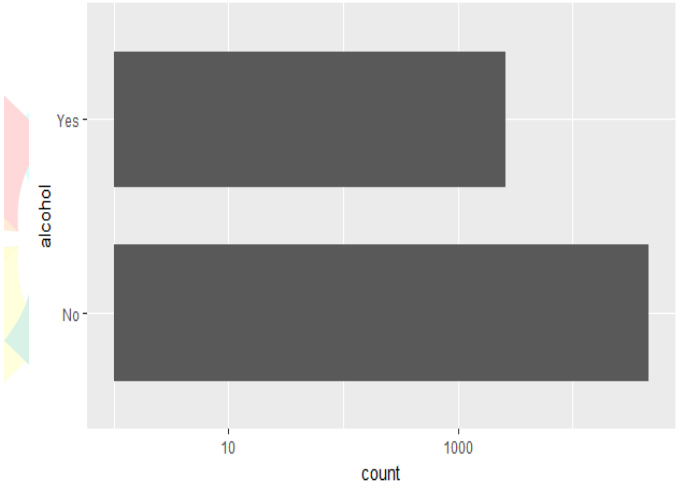
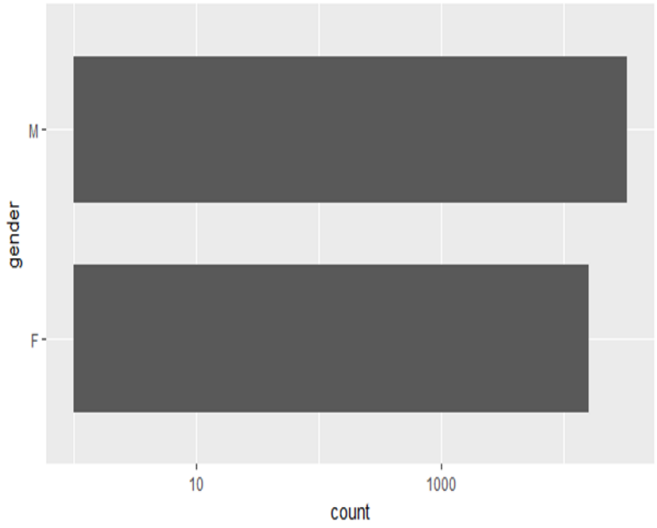


Figure 5 Alcohol - violation count



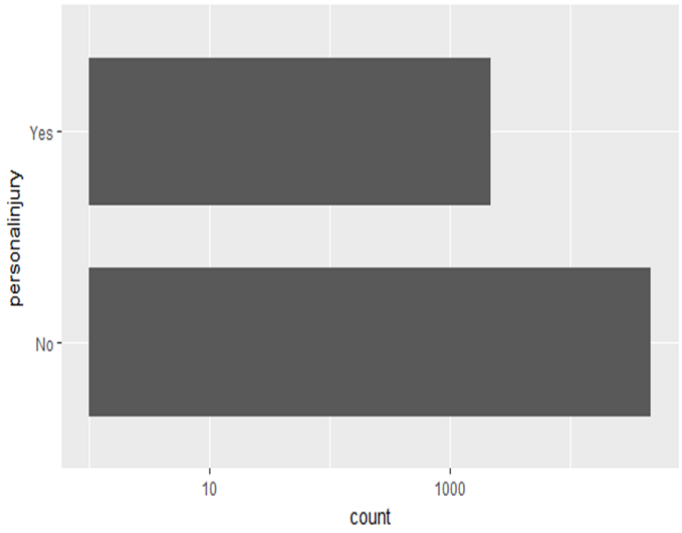


Figure 6 Gender - violation count**Figure 7** Personal injury - violation count

Figure 8 Race violation – count**Figure 9** Belt - violation count

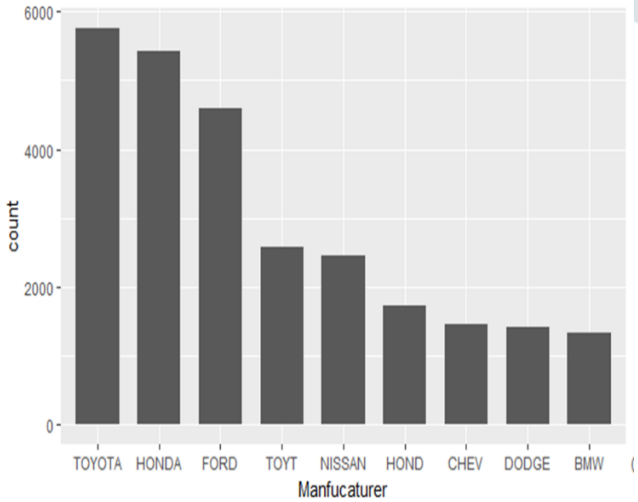


Figure10 Make - violation count

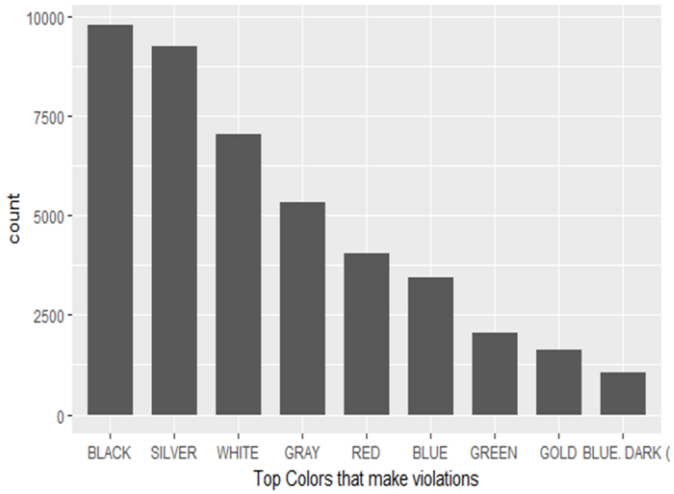


Figure 11 Top colors - violation count

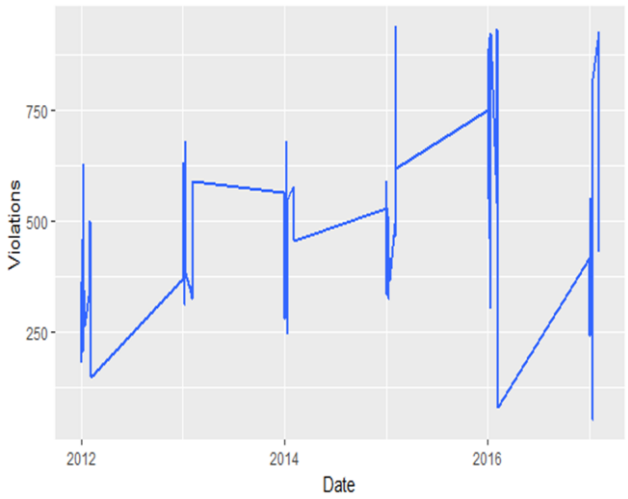


Figure 12 Date wise violation count

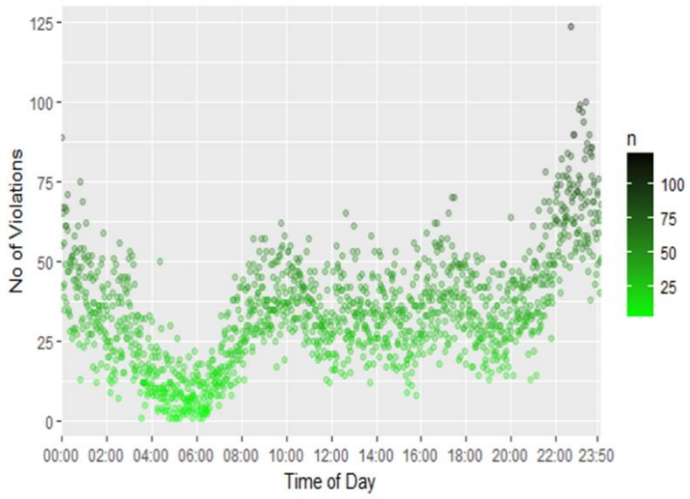


Figure 13 Violation counts regarding Time

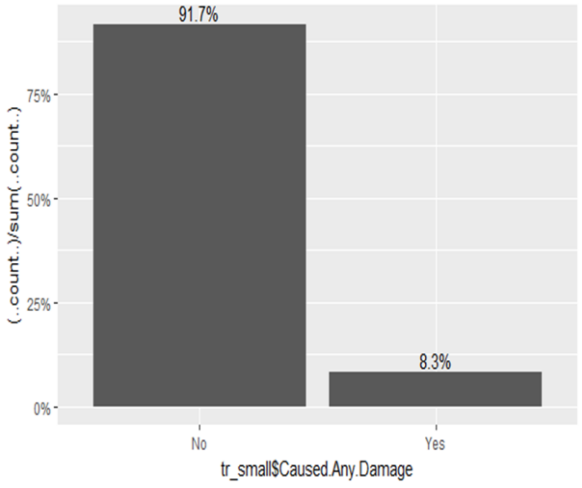


Figure 14 Property damage – violation count

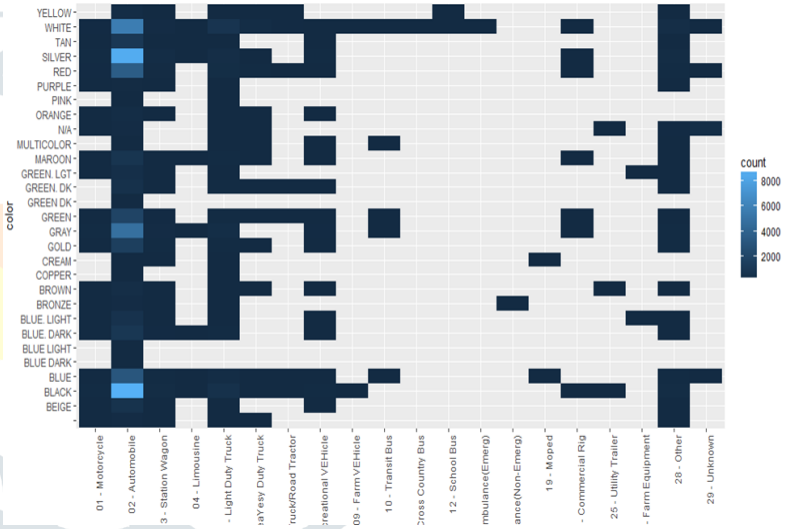
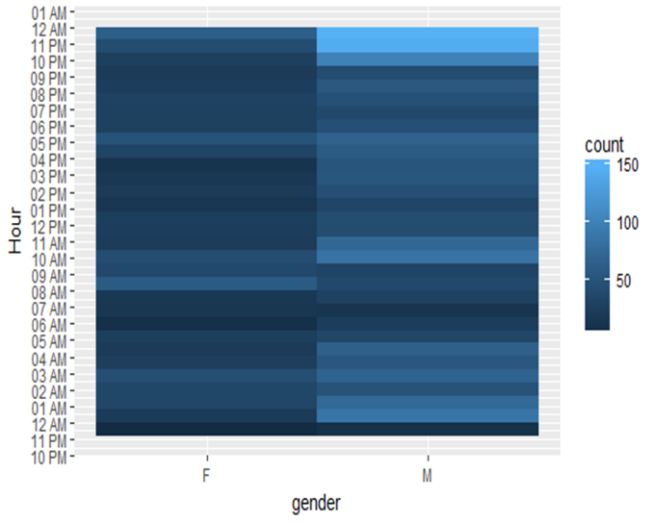


Figure 15 Vehicle Type and color are compared basedon accident



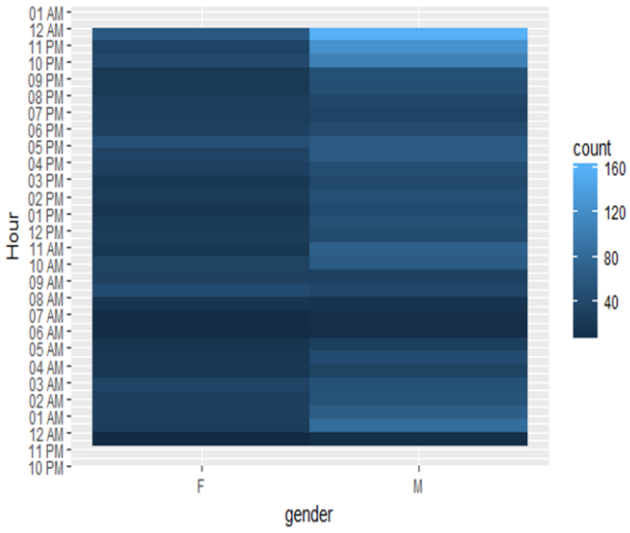


Figure 16 Gender and Time are compared based on alcohol

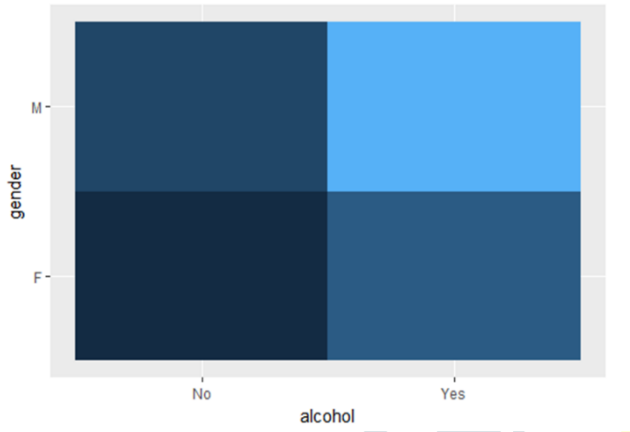


Figure 18 Gender and alcohol are compared based on accident

Figure 17 Gender and Time are compared based on accident

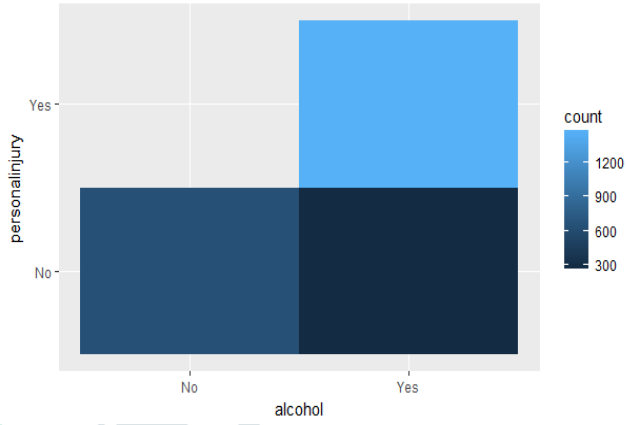


Figure 19 Personal injury and alcohol are compared based on accident

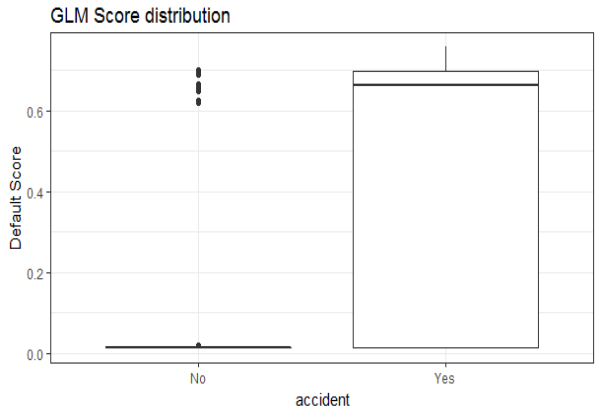


Figure 20 GLM score distribution

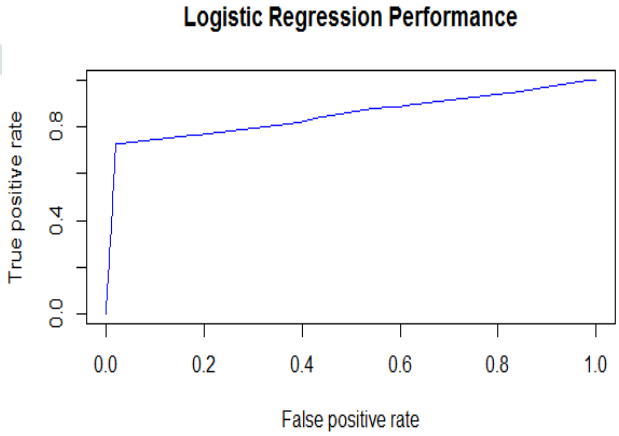


Figure 21 Logistic Regression Performance

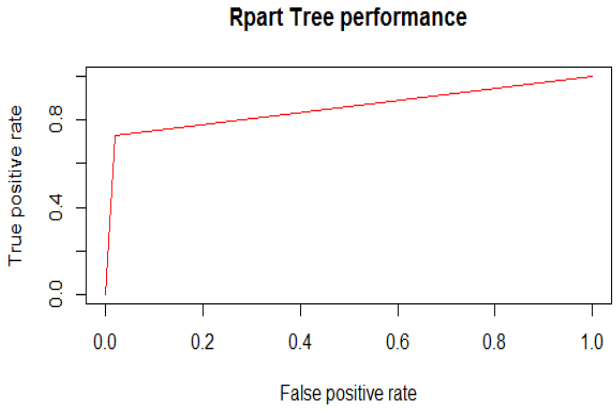


Figure 22 Rpart tree performance for Decision tree

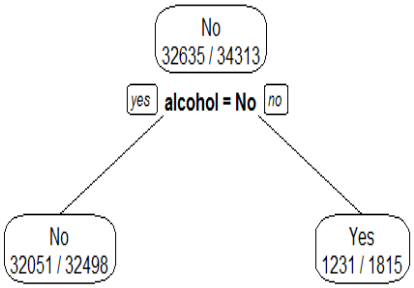


Figure 23 Decision tree predicts alcohol as the target variable

V. CONCLUSION

In general, this type of analysis will be helpful for the decision makers to develop new traffic rules and policies. The system analyse the accidents based on different violation factors and the impact over the years and also the relationship between different factors. The prediction states that the alcohol, personal injury and gender as a major violation factors that are contributed to the accident. Since the alcohol count is more, the decision tree predicted that the alcohol is more prone to accidents.

REFERENCES

[1] https://www.linkedin.com/pulse/traffic-violation-data-analysis-using-hive-azure-jongwook-woo

[2]https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q

[3]http://public.tableau.com/views/USTrafficViolationsDataAnalysis/Story1?%3Aembed=y&%3AshowVizHome=no&%3AshowTabs=y&%3Adisplay_count=y&%3Adisplay_static_image=y&%3Aretry=yes

[4]https://searchdatamanagement.techtarget.com/definition/Hadoop

[5]https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

[6]https://www.technologynetworks.com/informatics/articles/data-mining-techniques-from-preprocessing-to-prediction-307060

[7]https://data.world/data-society/montgomery-traffic-violations