TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 4401 - 4407

A Study on Accident Rates in Tamilnadu Using R

Dr. C. Sunitha[1], Asha Priya B[2], Lavanya S[3]

[1]Head of the Department,Sri Krishna Arts and Science College

[2,3]Student, Sri Krishna Arts and Science College

Article Info
Volume 82
Page Number: 4401 - 4407
Publication Issue:
January-February 2020

Article History
Article Received: *18 May 2019*
Revised: *14 July 2019*
Accepted: *22 December 2019*
Publication: *22 January 2020*

Abstract

In this fast-growing world, the usage of vehicles by the public will increase day by day. Due to greater traffic in developed areas such as in roads and highways, motor accidents and train accidents are growing in our state. The reason for the accidents is numerous such as traffic rules violation, drunk and drive, defective roads, non-application of protective appliances, and hindrances on the road, due to assignment of incessant driving for many hours and due to flawed mechanism in the vehicle. Few accidents are due to acts rues. Due to increasing number of accidents, there is loss of human lives, property damage, Traffic deadlock etc., and they are the main cause for social problems. So, it becomes very essential to limit the road accidents. By identifying the causes for the accidents happening, it would be easier to avoid the accidents in future. It will be useful for the whole society and even for the traffic control department. The data analyzes of road accidents is performed out in this paper using ggplot function that is available in Tidyverse package of R language. This paper aims to provide a analysis on excerpting meaningful data with the help of Data Mining concept, in order to find accident hot spots and predict accident tendencies using the techniques of data mining

Keywords: *Data mining, ggplot, dataset, accidents, R, Tidyverse package*

I. INTRODUCTION

R is a tremendously flexible statistical language and environment. R language is an Open Source language and it can be freely accessible for all probable operating systems. The flexibility of R is possibly matchless by any other statistics program, because of its object-oriented programming capability that permits for the creation of functions that perform customized procedures and/or the automation of errands that are usually completed. This flexibility, yet, has also kept some researchers away from R. The ever-increasing incredible amount of data, composed and warehoused in large and frequent data bases, has far surpassed ability for human to understand without using of tools that are powerful. Consequently, significant choices are not made based on the information that data stored in databases but on a decision maker's instincts due to the minimum tools to extract the valuable information implanted in the huge amounts of data. Therefore, data mining has received great courtesy in recent years. We hope this chapter will transport that using R is indeed a best rehearsal and can be a treasured tool in research.

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 4401 - 4407

II. DATA MINING IN R

Data mining will be done with the finding of useful, valid, unpredicted, and comprehensible knowledge from data. One of the most vital unique issues in data mining is size. c. R has limitations with management huge number of datasets because all calculation will be carried out in the main memory of the computer. Taking priority of the highly flexible database lines available in R, we can perform data mining on large problems. Data mining is used in numerous fields, such as merchandizing, economics, cable(telecommunication) and social media. The core methods for data mining contain arrangement and estimate, clustering, outlier finding, suggestion rules, sequence analysis, time series investigation and text mining and in new methods like social network analysis and sentiment analysis.

i) R

R is a free software for statistical calculating and graphics. It offers a wide variability of arithmetical and graphical techniques. R can be straightforwardly protracted with 7324 packages available on CRAN3 (as of October 20, 2015). In addition, there are many packages

1. http://www.crisp-dm.org/
2. http://www.r-project.org/
3. http://cran.r-project.org/

provided on other websites, such as Bioconductor4, and also there are huge packages still under development at R- Forge5 and GitHub6. More details about R are available in the CRAN website. R is extensively used in both education and industry. Some Task Views related to data mining are:

- Machine Learning & Statistical Learning,
- Multivariate Statistics,
- Cluster Analysis & Finite Mixture Models,
- Time Series Analysis,
- Analysis of Spatial Data,
- Natural Language Processing.

R Reference Card for Data Mining is another guide for Data Mining in R. Its latest version will be available at the following sites http://www.rdatamining.com/docs and http://www2.rdatamining.com/.

A.Data Import and Export

The data format Data frame is utmost used in R among all the formats. A data frame has a structure of a table, with each row as an observation/record and each column might be a variable or feature.

Import Data from SAS Package

To import SAS datasets into R use the function read.ssd().

- SAS essential be available on every machine, and read.ssd() function will be used to call SAS package to read those datasets and importation into R.
- The name of a SAS dataset file must be within eight characters.
- While importing from a CSV file there is no such limitation.

Read and Write EXCEL files with package xlsx

The package RODBC is used for both reading and writing EXCEL files on Windows. No additional drivers is required to write the package xlsx because it supports both reading and writing Excel 2007 and Excel 97/2000/XP/2003

B.Data Frames

For statistical work, “data.frame” objects are very convenient. They have the structure like a table with rows and columns. They can also access the row names and column names. Easiest is to create the table in a spreadsheet program and save it as comma separated values file. This csv file can then be read from within R using following command:

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 4401 - 4407

1) code:

>mydf = read.csv('filename', header=T)

The characters used for missing values and the separator (if not comma), can also be specified here:

2) code:

>mydf=read.csv('filename', header=T, na.strings="",sep=";")

III. GGPLOT2

The package GGPLOT2 is designed by Hadley Wickham, suggestions a controlling illustrations language for generating well-designed and complex plots. GGPLOT2 is built upon The Grammar Graphics by Leland Wilkinson, ggplot2 allows you to build graphs that signify both univariate as well as multivariate arithmetical and definite data in a upfront manner. Grouping in GGPLOT2 will be done based on by color, photo, size and symbol.

The ggplot2 packages is comprised in a general gathering of packages called “the tidyverse”. Take a instant to guarantee that it is installed, and that we have devoted the ggplot2 package.

To get ggplot2 package easy way is to install the tidyverse package fully using the following statement: install.packages("tidyverse")

we can also install just ggplot2: install.packages("ggplot2")

we can also use the development version from GitHub using the following command:

install.packages("devtools")devtools::install_github("tidyverse/ggplot2")

The data visualization and graphics for communication in R is called “R for data science”. R for data science is considered to give you complete walkthrough to the tidyverse package. To add components to a plot + is the key to building erudite ggplot2 graphics.

You can enhance any of the next types of objects:

•The default aesthetics will be substituted in aes(_)object.

•A new layer can be formed by a geom_ or stat_ function

•The current coordinate system will be overridden by coord.

•A scale dominates the current scale.

•A theme() alters the current theme.

•The current faceting will be overridden by facet specification. To substitute the present evasion data frame, use %+%, using S3 method precedence issues. Provide a list; in which case each element of the list will be supplementary in turn.

Examples

base<- ggplot(mpg, aes(displ, hwy)) + geom_point() base + geom_smooth()

To override the data, you must use %+% base %+% subset(mpg, fl == "p")

A.SCATTERPLOT

1.ggplot function takes the dataset as source.

2.Inside the aes() argument, add x and y axes attributes.

3.The + sign indicates R to retain the reading process of the code. The code will be broken into pieces inorder to make it readable.

4.Usegeom_point() function for the geometric object

Eg

>Library (ggplot2)

>ggplot(datasetname, aes(x=drat,y=mpg)+ geom_point()

SCATTER PLOT WITH GROUPS

January - February 2020
ISSN: 0193 - 4120 Page No. 4401 - 4407

ggplot(datasetname, aes(x=mpg, y=drat))+ geom_point(aes(color=factor(gear))

•The aes() inside the geom_

1.CHANGE AXIS

The main job of the data scientist is rescaling the dataset. Sometimes the data comes in an infrequent manner. To change the data less sensitive to outliers can be done by rescaling them.

ggplot(datasetname,aes(x=log(mpg,y=log(drat)))+point()

this above command manages the color of the group. The factor variable will act as a group here. So that you can convert the variable gear in a factor.

• The code

aes(color = factor(gear))

dots color will be changed.

geom_point(aes(color=factor(gear))

• Then transform the x and y variables in log() directly inside the aes() mapping.

2. SCATTER PLOT WITH VALUES

Another level of information can be added to the graph by plotting the fitted value of a linear regression.

My_graph<-ggplot(dataset name, aes(x=log(mpg), y=log(drat)))+
geom_point(aes(color=factor(gear)))+
stat_smooth(method="lm", col="#C42126", se=FALSE,

size=1) my_graph

Code Explanation

- graph: Store the graph into the variable graph. It will be helpful for further use or can avoid too complex line of codes
- To control the smoothing method the argument stat_smooth() is used
- method = "lm": method here is set to "Linear regression"
- col = "#C42126": the line color is fixed to red color
- se = FALSE: this specifies to not display standard errors
- size = 1: the value 1is fixed as the size of the line

IV. DATASET(Road accident in Tamil Nadu)

Road accidents in Tamil Nadu are increasing step by step, the dataset taken here is a real time dataset.[1] The most elevated street mishaps were taken place in the year 2016, the expansion of mishap were taken place the year 2016 in 2001 it was around 37 percent, and it has diminished to 26 percent in the year 2017 because of different advances and street wellbeing estimates carried by the Tamil Nadu Government. In the year 2017, the mishaps diminished and, in this way, it has decreased the casualties and non-casualties somewhat.

The decreased mishaps has made gigantic decline in the quantity of people harmed. In the year 2017, 74,572 which sums around 25 percent decline looking at the earlier year 2016. The normal every day demise in street mishaps was 47 percent in the year 2016 and it has decreased to 44 in the year 2017. The inadvertent wounds every day has additionally diminished to 204 from 272 in the year 2017 and 2016. It is fascinating to take note of that the nonstop exertion of the vehicle office has spared the life of 3 people each day during the year 2017.

STEP-WISE IMPLEMENTATION OF DATASET 1 USING R

STEP 1: Read the CSV dataset using read.csv function of R.

STEP 2: Analyse the structure of the dataset using str function.

STEP 3: Also analyse the summary of the dataset.

January - February 2020
ISSN: 0193 - 4120 Page No. 4401 - 4407

STEP 4: Install the required packages from CRAN using install function and access the package using the library function.

STEP 5: Use ggplot function specify the source, legends and specify the type of graph to be used.

After implementation and processing of the captured data of accidental rate in the above section, we got the below results. The results deal about the data analysis of accidental rate on different years from dataset 1.

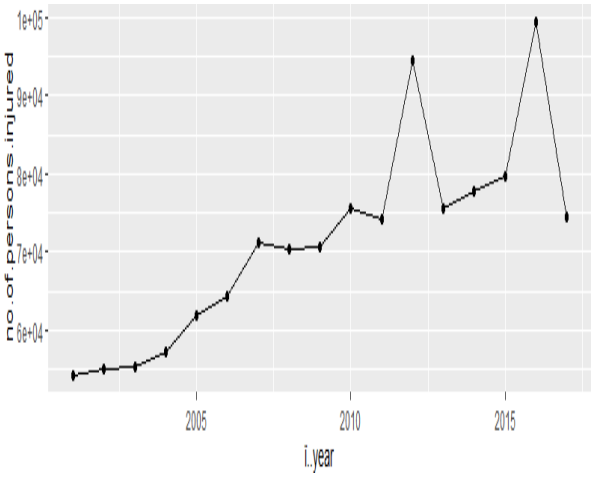

FIGURE 1

DATA ANALYSIS OF ACCIDENT DETAILS IN TAMIL NADU FROM 2001 – 2017

DATASET 1

TAMIL NADU ACCIDENT RATES FROM 2001 – 2017[1]

YEAR	ACCIDENTS	ACCIDENTS THAT CAUSED DEATHS	INJURIES	NUMBER OF PERSONS DEAD	NUMBER OF PERSONS INJURED
2001	51978	8579	43399	9571	54282
2002	53503	9012	44491	9939	55130
2003	51025	8393	42632	9275	55242
2004	52508	8733	43775	9507	57283
2005	53878	8844	45034	9760	61967
2006	55145	10055	45090	11009	64341
2007	59140	11034	48106	12036	71099
2008	60409	11813	48596	12784	70251
2009	60794	12727	48067	13746	70504
2010	64996	14241	50755	15409	75445
2011	65873	14359	51514	15422	74245
2012	67757	15072	52685	16175	94523
2013	66238	14504	51734	15563	75681
2014	67250	14165	53085	15190	77725
2015	69059	14524	54534	15642	79701
2016	71431	16092	55339	17218	99381
2017	65562	15061	50501	16157	74572

DATASET 2

CASUALITIES AND MORTALITIES OCCURRED IN THE YEAR 2017 ACCORDING TO VEHICLE TYPE[1]

January - February 2020
ISSN: 0193 - 4120 Page No. 4401 - 4407

Vehicle Type	Total accidents	Percentage of share	deaths	Percentage of share
Bus: Govt.	2796	4.26	1029	6.37
:Private	3238	4.94	827	5.12
Trucks	7373	11.25	2506	15.51
Four wheelers	18748	28.60	3994	24.72
Two wheelers	25393	38.73	5322	32.94
Three wheelers	3000	4.58	478	2.96
Others	5014	7.65	2001	12.38
Total	65562	100.00	16157	100.00

Follow the same steps for the implementation. After implementation and processing of the captured data of accidental rate based on type of vehicles as shown in the above section, we got following results as shown below. The results deal about the data Analysis of accidental rate based on type of vehicle from dataset 2 we obtained the result in form of figure 2.

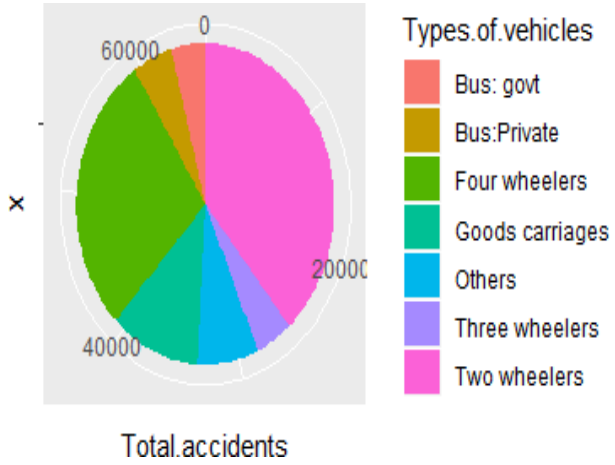

FIGURE 2

DATA ANALYSIS REPORT ON CASUALITIES AND MORTALITIES OCCURRED DURING 2017 ACCORDING TO TYPE OF VEHICLES

The most extreme mishaps are brought about by two wheelers with a percentage of 38.73% trailed by four wheelers like car, jeep, etc with a percentage of 28.60% what's more, stands second in mishaps pursued by Trucks/lorries with a percent of 11.25%. In death ratio additionally, most extreme deaths are caused because of bikes with a percent of 32.94% of all out deaths trailed by four wheelers like car, jeep, etc with a percent of 24.72% of all out deaths and by Trucks/lorries with a percentage of 15.51%.

Among the all out vehicle populace of 251.47 lakhs, the bike vehicle populace is 84.08% (211.44 lakhs). The greatest casualties are brought about by bikes due to not wearing of protective caps. These data's are analysed using R.

DATASET 3

DEATHS OCCURRED DUE TO NOT WEARING OF ANY PROTECTIVE CAPS [1]

Year	Deaths due to two wheelers	Deaths due to not wearing of helmets
2016	5666	4091
2017	5322	2956

Follow the same steps for the implementation. After implementation and processing of the captured data of deaths due to not wearing of helmets as shown in the above section, we got the following results. The results deal with the data analysis of deaths due to non-wearing helmets from dataset 3 we obtained the result in form of figure 3.

January - February 2020
ISSN: 0193 - 4120 Page No. 4401 - 4407

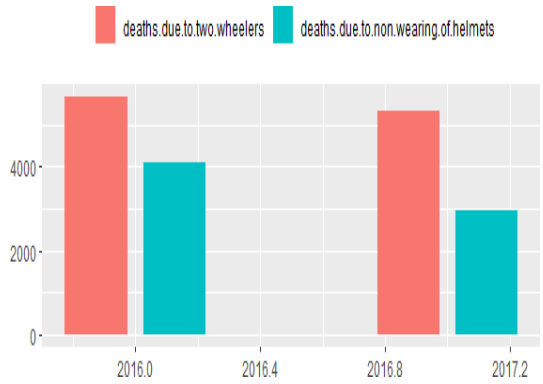

FIGURE 3

DATA ANALYSIS REPORT ON DEATHS DUE TO NON-WEARING HELMETS

V. CONCLUSION

In this work, we have applied data analysis over cognitive data reposted in open spaces, which are promptly available by means of web. These datasets are accessible in the structure as audits, surveys and so forth. In this work we determined a deliberate way to deal with concentrate and mine valuable information in the structure as assumptions to take reasonable choices. Content preprocessing is dispatched over unstructured information, with an aphorism to secretive it as organized information by utilizing highlight extraction and highlight choice procedures. As a piece of model structure over this refined information, one of the most prominent managed learning procedure, guileless bays is utilized to mine learning and perform SA on advanced psychological datasets. Besides this, the investigation is made to approve the model by adjusting measurable assessing measurements. The consequences of these assessment measurements reliable authorize the exactness of results by offering high qualities over all the intellectual datasets utilized in this work.

REFERENCE

[1] "Road accidents and road safety measures in tamilnadu-An analysis" by S.Krishnan, K.Geetha, Rabiya Basri(https://www.scribd.com/document/388733745/accident-analysis-140318-pdf).

[2] "Scatter Plot in R using ggplot2" by guru99(https://www.guru99.com/r-scatter-plot-ggplot2.html)

[3] Data Mining Definition - What is data mining (https://www.investopedia.com/terms/d/datamining.asp)

[4] Data Mining Process: Techniques and Major Issues(https://www.softwaretestinghelp.com/data-mining/")