"Analysis and visualization of bike rental data"

by

Utkarsh Brajnil (18BCE1158)
Himanshu Lohar (18BCE1138)

A project report submitted to

Dr. PATTABIRAMAN V

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

in partial fulfilment of the requirements for the course of

CSE3020 –DATA VISUALISATION

in

B.Tech. (Computer Science Engineering)

VIT UNIVERSITY, CHENNAI

Vandalur – Kelambakkam Road

Chennai – 600127

APRIL 2020

CERTIFICATE

Certified that this project report entitled "**Analysis and visualization of bike rental data"** is a bonafide work of Utkarsh Brajnil 18BCE1158, Himanshu Lohar 18bce1138 who carried out the "J"-Project work under my supervision and guidance for CSE3020-Data Visualisation .

Dr. Pattabiraman V

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT University, Chennai

Chennai – 600 127.

ABSTRACT

We have a dataset of year 2014 of a company named “Capital Bikeshare” which contains information about different bikes along with their distinct numbers that are being used by either a member or a casual user for their journey from one station to another provided with the start time and end time.

Through this project we will be basically analyzing the previous trends of the bike rental company and hence predicting the class of user (A member or a casual user).

We will analyze the trip history of the company and hence arrive at the conclusion. We made visually interactive output for better consumption of data.

.

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project faculty, **Dr. Pattabiraman V,** for his consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

Finally, we would like to thank our deemed university, VIT Chennai, for providing us with the opportunity and facilities which ensured this project's completion.

1. INTRODUCTION

We have a dataset of year 2014 of a company named “Capital Bikeshare” which contains information about different bikes along with their distinct numbers that are being used by either a member or a casual user for their journey from one station to another provided with the start time and end time.
Through this project we will be basically analysing the previous trends of the bike rental company and hence predicting the class of user (A member or a casual user).
We will analyse the trip history of the company and hence arrive at the conclusion.

We will be using linear discriminant analysis.

Linear Discriminant Analysis (LDA) is a generalization of Fisher's linear discriminant, a method used in Statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. This method projects a dataset onto a lower-dimensional space with good class-separability to avoid overfitting (“curse of dimensionality”), and to reduce computational costs.

2. DATASET USED

It contains the following columns:

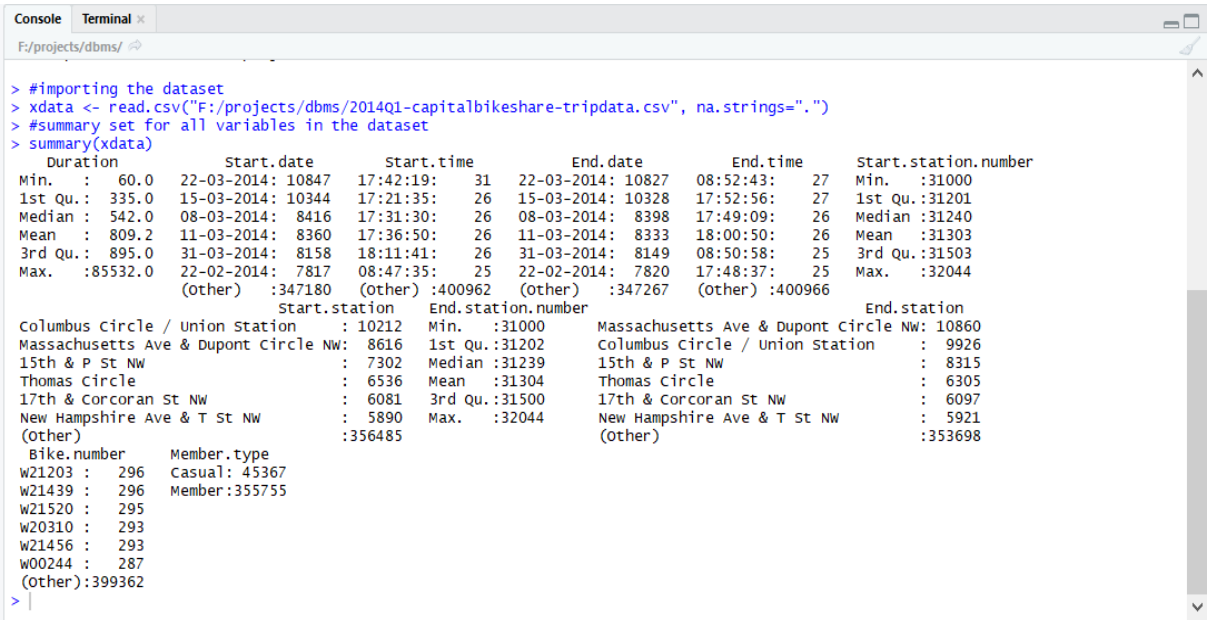
- Duration
- Start date
- Start time
- End date
- End time
- Start station number
- Start station name
- End station number
- End Station name
- Bike number
- Member type

➢ This dataset comes from a bike sharing company in US

➢ There are a total of 4,01,122 entries in the tabulation

➢ The entries are dated from 1/1/2014 to 31/3/2014 and belong to Quarter-1 of the year 2014

This picture shows the summary of the dataset.



project.R × xdata.sub1 ×

Filter

	Duration	Member.type	Start.station	Start.station.number
440	6666	Casual	15th & Crystal Dr	31003
441	6640	Casual	15th & Crystal Dr	31003
520	5582	Casual	Florida Ave & R St NW	31503
527	5517	Casual	Florida Ave & R St NW	31503
528	5534	Casual	Florida Ave & R St NW	31503
541	8791	Casual	24th & N St NW	31255
542	8769	Casual	24th & N St NW	31255
553	11641	Casual	Smithsonian-National Mall / Jefferson Dr & 12th St SW	31248
554	11624	Casual	Smithsonian-National Mall / Jefferson Dr & 12th St SW	31248
576	3638	Casual	17th & Corcoran St NW	31214
583	6924	Casual	Smithsonian-National Mall / Jefferson Dr & 12th St SW	31248
592	6600	Casual	Smithsonian-National Mall / Jefferson Dr & 12th St SW	31248
705	4272	Casual	Convention Center / 7th & M St NW	31223
707	4317	Casual	Convention Center / 7th & M St NW	31223

Showing 1 to 15 of 8,518 entries

3. IMPLEMENTATION CODE

```
#representing the data graphycally

#representing the data in histogram
par(mfrow=c(4,2))
par(mar = rep(2, 4))

hist(xdata.sub1$Duration)
hist(xdata.sub1$Start.station.number)
#hist(xdata.sub1$Member.type)

#representing as boxplot
boxplot(xdata.sub1$Duration)
boxplot(xdata.sub1$Start.station.number)

bike.freq<-table(xdata$Bike.number)
barplot(bike.freq)
hist(bike.freq)

member.freq<-table(xdata$Member.type)
barplot(member.freq[order(member.freq,decreasing = T)],
        col = "blue",
        border = NA,
        main = "preferred customer type",
        xlab = "type of customer",
        ylab = "number of passes taken")
#hist(member.freq)

stat.freq<-table(xdata$Start.station.number)
barplot(stat.freq)
barplot(stat.freq[order(stat.freq,decreasing = T)])
```

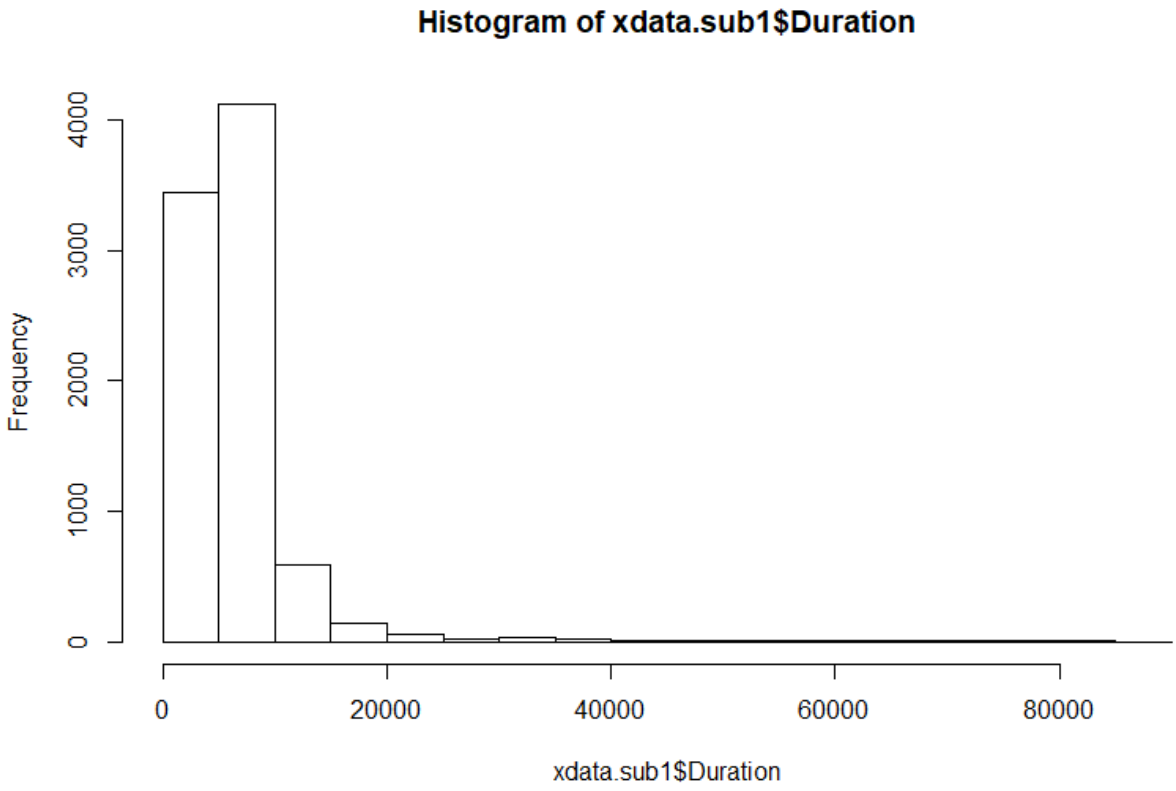
```
#PREDECTION OF CLASS OF USER USING LINEAR DISCRIMINANT ANALYSIS



#predicting the class of user(member type)
library(MASS)
ldao<-lda(Member.type~Duration,xdata.sub1)
ldapre<-predict(ldao,xdata.sub1)
ldacls<-ldapre$class
ldatbl<-table(ldacls,xdata.sub1$Member.type)
accuracy<-sum(diag(ldatbl))/sum(ldatbl)*100

#output display
#ldao
#ldapre
#ldacls
#ldatbl
#accuracy

#predicting for userd who have duration >5hours

#using subset of dataset
xdata.sub2 <- subset(xdata, Duration > 18000, select = c("Duration","Member.type","Start.station","Start.station.number"))

#predicting the class of user(member type)
library(MASS)
ldaoo<-lda(Member.type~Duration+Start.station.number,xdata.sub1)
ldapree<-predict(ldao,xdata.sub1)
ldaclss<-ldapre$class
ldatbll<-table(ldacls,xdata.sub1$Member.type)
accuracyy<-sum(diag(ldatbl))/sum(ldatbl)*100

#output display
ldaoo
ldapree
ldaclss
ldatbll
accuracyy
```

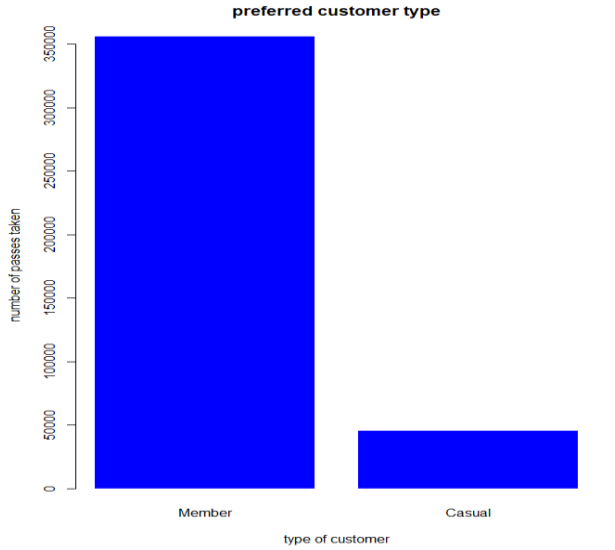
Console Terminal

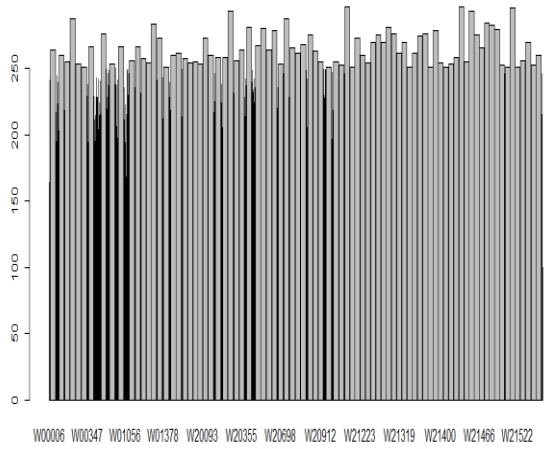
F:/projects/dbms/

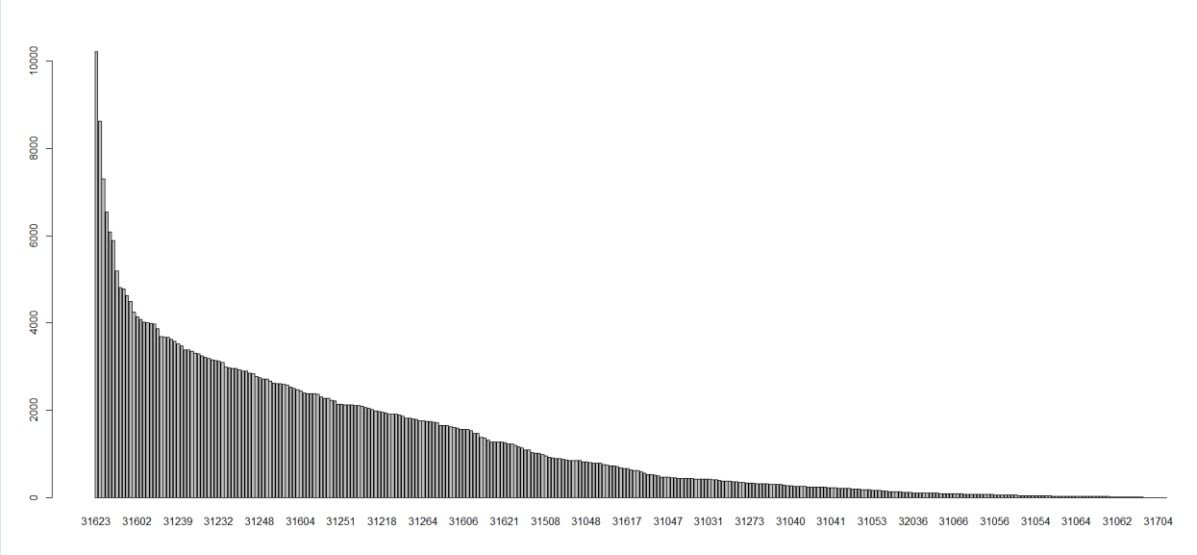
```
> #output display
> ldao
Call:
lda(Member.type ~ Duration, data = xdata.sub1)

Prior probabilities of groups:
   Casual    Member
0.8860061 0.1139939

Group means:
         Duration
Casual  6582.302
Member 11015.649

Coefficients of linear discriminants:
                    LD1
Duration 0.0001536618
> ldatbl

ldacls   Casual Member
  Casual   7500    868
  Member     47    103
> accuracy
[1] 89.25804
>
```

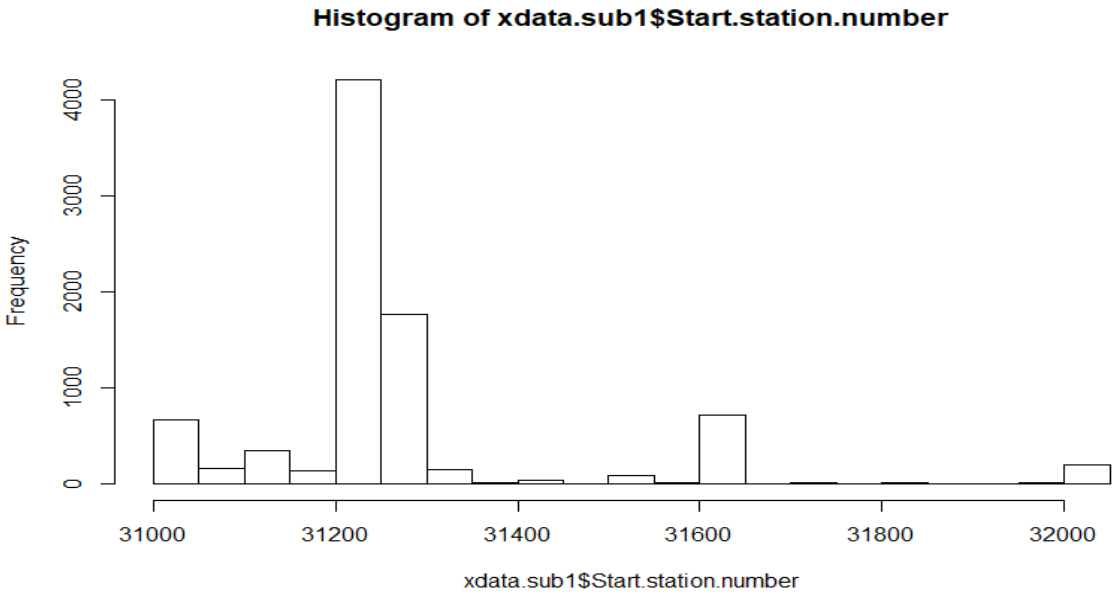
4. VISUAL OUTPUT











Console Terminal

F:/projects/dbms/

```
> #importing the dataset
> xdata <- read.csv("F:/projects/dbms/2014Q1-capitalbikeshare-tripdata.csv", na.strings=".")
> #summary set for all variables in the dataset
> summary(xdata)
    Duration            Start.date        Start.time          End.date          End.time      Start.station.number
 Min.    :   60.0   22-03-2014: 10847   17:42:19:    31   22-03-2014: 10827   08:52:43:    27   Min.   :31000
 1st Qu.:  335.0   15-03-2014: 10344   17:21:35:    26   15-03-2014: 10328   17:52:56:    27   1st Qu.:31201
 Median :  542.0   08-03-2014:  8416   17:31:30:    26   08-03-2014:  8398   17:49:09:    26   Median :31240
 Mean   :  809.2   11-03-2014:  8360   17:36:50:    26   11-03-2014:  8333   18:00:50:    26   Mean   :31303
 3rd Qu.:  895.0   31-03-2014:  8158   18:11:41:    26   31-03-2014:  8149   08:50:58:    25   3rd Qu.:31503
 Max.   :85532.0   22-02-2014:  7817   08:47:35:    25   22-02-2014:  7820   17:48:37:    25   Max.   :32044
                   (Other)   :347180   (Other) :400962   (Other)   :347267   (Other) :400966
                                  Start.station    End.station.number                                End.station
 Columbus Circle / Union Station     : 10212   Min.   :31000      Massachusetts Ave & Dupont Circle NW: 10860
 Massachusetts Ave & Dupont Circle NW:  8616   1st Qu.:31202      Columbus Circle / Union Station     :  9926
 15th & P St NW                      :  7302   Median :31239      15th & P St NW                      :  8315
 Thomas Circle                       :  6536   Mean   :31304      Thomas Circle                       :  6305
 17th & Corcoran St NW               :  6081   3rd Qu.:31500      17th & Corcoran St NW               :  6097
 New Hampshire Ave & T St NW         :  5890   Max.   :32044      New Hampshire Ave & T St NW         :  5921
 (Other)                             :356485                      (Other)                             :353698
  Bike.number     Member.type
 W21203 :   296   Casual: 45367
 W21439 :   296   Member:355755
 W21520 :   295
 W20310 :   293
 W21456 :   293
 W00244 :   287
 (Other):399362
>
```

5. CONCLUSION

- By referring to the histogram between frequency and time duration, we infer that most number of bikes (around 4100) are hired for a time duration of range 5000-10000 seconds.
- After studying the histogram between frequency and station number, we infer that most number of bikes (790) are hired from station number 31258. So we learn that availability of bikes at station number 31258 should be kept high.
- After studying the histogram we would be able to infer that which bikes are being used more than 250 times. So the bikes that are being used more than 250 times should be sent for servicing.
- We could predict from the sources that the probability that the next person visiting for rental services for more than 1 hour duration would be a member is 0.1139939 and for casual is 0.8860061.
- The accuracy of our prediction is 89.25804%.
- So by using all these data, we can maximise the profit of the company.