MA334 Final Project

Ahmad Afzaal Cheema (2111168)

Introduction

In this project, the dataset we are using is a bike buyer dataset. We get this dataset from GitHub, link of the dataset is mentioned below,

https://github.com/Peterstangolis/EDA---R/blob/53e011f4c530c3ab217ea78a4ee95ff6efb7fbbd/bike_buyers.csv.

It is categorical data that shows either the person will buy bike or not. We have two categories yes or no. The dataset contain 1000 rows and 13 columns. It includes the detail of a person like his/her income, marital status number of cars they have etc. After reading and analyzing the dataset we observed that there is no need of two columns that are id and distance so we drop them. Now we have 1000 observations with 11 variables that are: Marital status, gender, income, no. of children, Education, Occupation, Home owner or not, number of cars , region, age and last one the dependent variable Purchased bike or not.

By using the function of checking null, we noticed that there is no missing data in this dataset. After dropping the non-useful columns, the data types of remaining columns are:

Characters data type	Numerical data type
Marital. Status	Income
Gender	Children
Education	Cars
Occupation	Age
Home owner	
Region	
Purchased. Bikes	

Table 1: Data Type of columns

Analyzing data

First of all, we made box plots to show the data. For this we plot two box-plots one for showing the ratio of having number of cars according to age and other is buying bikes according to the income

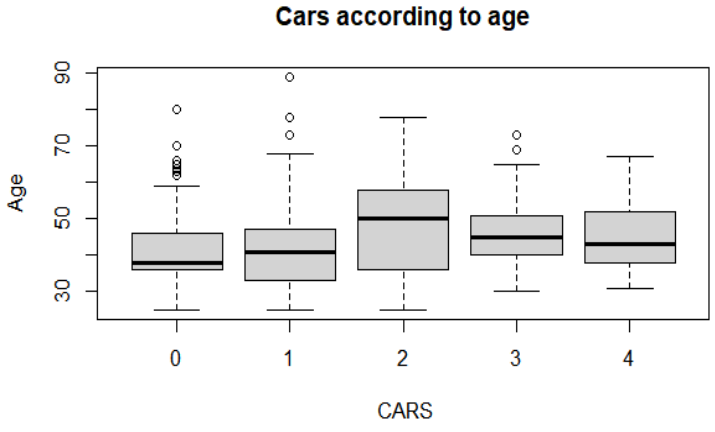

Figure 1: Boxplots of number of cars according to Age

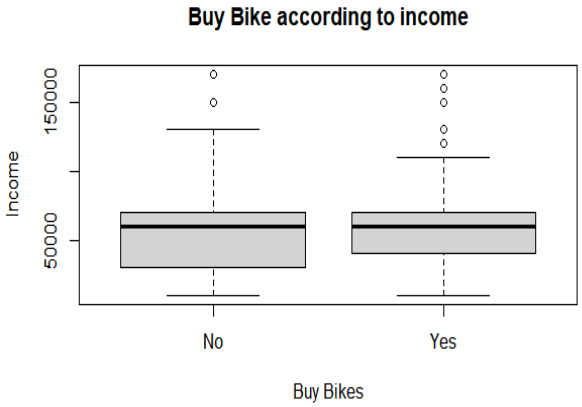

Figure 2: Boxplot of bike purchasing according to income

In first box plot, number of cars are showing according to the range of age. This graph is showing that the people whose age is between 40 to 60 have 2 cars. Range of 20 to 50 have 4 cars and we can see that the people in the age of 70 to 90 having one car.

The second box plot, showing the ratio of having bike or not according to their income. We can observed that mostly people don't have bikes and the mostly people whose salary is greater than 150000 have a bike. We can also see that the ratio of having no bike is greater than having a bike.

Summary

By using the function of summary I computed the summary of all columns including character data type and numeric. Character data types' summary shows the total rows, its data type and class. Each character data type columns have 1000 rows and data types and class are character.

Column names	Min	Max	1st Quartile	Median	3rd Quartile	Mean	Nans
Income	10000	170000	30000	60000	70000	56140	0
Cars	0	4	1	1	2	1.452	0
Children	0	5	0	2	3	1.908	0
Age	25	89	35	43	52	44.19	0
Purchased Bike	0	1	0	0	1	0.481	0

Table 2: Summary of columns for boxplots

Mode for All data type:

While computing summary of all data we did not get the mode of each column. When we used mode prebuilt function it gives me data type of each column. So, we made a function for computing mode for all data type columns.

In the function we only passed 3 to 4 columns to check the mode.

Mode of purchased bike is No. Means mostly have no bikes as I computed it before in boxplot.

Mode of having cars is 2, this is also computed earlier in boxplot.

Model of Marital status is married means mostly people are married in this data.

Mode of Gender is Male.

Dnorm () for car variable:

To find the probability distribution against each point of the data. So I computes the distribution of having car and we can see that the greatest probability is 0.4 on having no car and leas probability is 0 point something having 4 cars.

Figure 3: Dnorm of car variable

Dnorm () for having children

We also compute the probability against number of children. We observed that probability of having no children is greater than others. Probability of having 2 to 5 children in less than 0.1.

Figure 4: Dnorm of children variable

Pnorm()

It used a cumulative distribution function shows the distribution of randomly distributed number less than the given number. We can see that the probability of Cumulative distribution having 4 cars is 0.7 but the normal probability distribution of having 4 cars is near to 0. This is because in Pnorm we used cumulative sum.

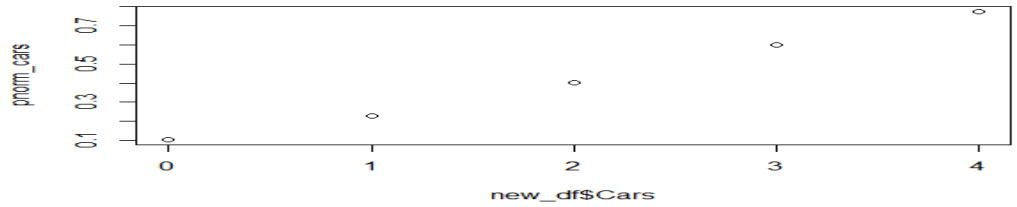

Figure 5: Pnorm of car variable

Qnorm()

It gives the numbers whose cumulative distributions same as probability distribution. So we passed the probability vector of cars that we produced in Dnorm and the results are 0.4 probability have 1.5 Qnorm probability that is greater than all.

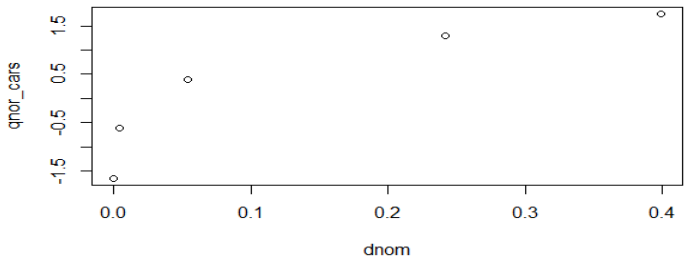

Figure 6: Qnorm of car variable

Dbinom()

We used Dbinom to find the density of each point with 100 number of trails and 0.5 probability on cars then I observed the highest probability is at having 4 cars.

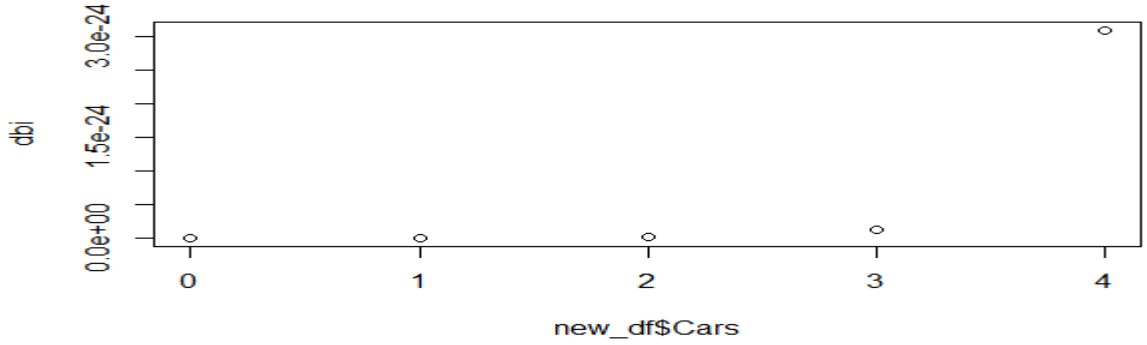

Figure 7: Dbinom of car variable

Then we also used the density function on children column too with 100 trails and 0.3 probability and we observed the highest probability at 0.00000000030.

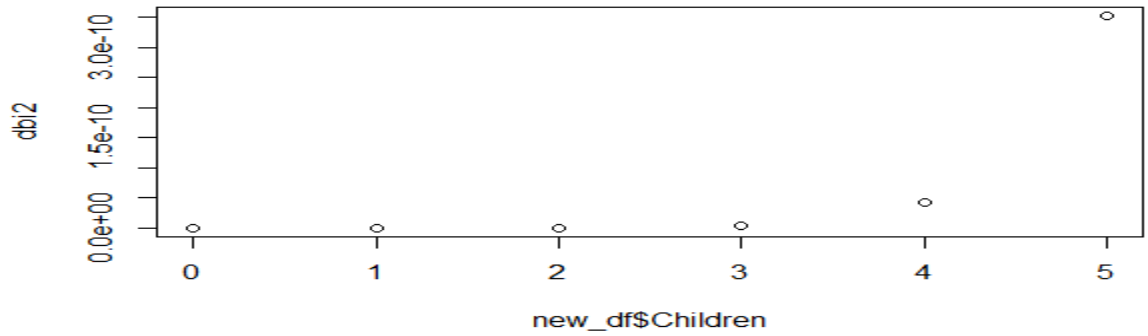

Figure 8: Dbinom of children variable

Poisson distribution

We have observed that the p values is less than 0.05 to consider an impact of the predictor variable on the response variable.

In the output the AIC means the quality of model. If it is high its mean the model quality is good and low means bad quality. The fisher score means the total iteration used to cover the algorithm, if it is high than the algorithm is not covering properly.

Output

(Dispersion parameter for poison family taken to be 1)

Null deviance: 1099.2 on 999 degrees of freedom

Residual deviance: 1095.0 on 997 degrees of freedom

AIC: 2987

Number of Fisher Scoring iterations: 5

Chi-Square:

The relation between income and purchased bike, the p value is less than 0.05 means they are correlated to each other.

Data: new_df$Purchased.Bike and new_df$Income

X-squared = 27.149, df = 15, p-value = 0.02754

ACNOVA test

To show the effect of variables on predictors, I used ancova test without interaction between categorical variable and predictor variable.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)	
Gender	1	5.5	5.458	4.323	0.0379	*
Marital.Status	1	0.7	0.699	0.554	0.4569	
Gender: Marital.Status	1	0.2	0.173	0.137	0.7112	
Residuals	996	1257.4	1.262			

P value is smaller than 0.05 for gender means we can't reject Gender column.

Test or Hypothesis test

One Sample t-test

Data: x - 9

t = -538.91, df = 999, p-value < 2.2e-16

Alternative hypothesis: true mean is not equal to 0

95 percent confidence interval: -8.55002 -8.48798

Sample estimates: mean of x -8.519

Result: P value is greater than 0.05 so there is strong evidence to reject the null hypothesis. There is overwhelming evidence against the null hypothesis.

Multivariable linear

We used both linear and multivariable linear regression. Firstly, we converted the dependent variable in numeric by using label encoding. For multiple linear we used cars children and income to predict that they will buy bike or not.

After fit the model on data we predict and test the model for having a one car 2 children and 30000 income.

$$Y = a+xcars*1+xchildren*2+xincome*30000$$

This given then intercept 0.4529762.

Conclusion

We observed that data is distributed normally and can use to predict that people having bike or not by using many ways like according to cars, salary or age. We observed that mostly males have a cars and most of them are married. Ratio of buying bike is less with having salary greater than 150000. The people having 2 cars are mostly 40 to 60 in age and they are greater than other. At the end we predict a person will buy a bike or not by given the value of car, no. of children and income.

Appendix

#load of data

```
data<- read.csv ("bike_buyers.csv")

data
```

#dimension of data

```
dim(data)
```

#Find nulls

```
sapply (data, function(x) sum (is.na(x)))
```

#non null columns

```
new_df <- subset(data, select = -c (ID, Commute. Distance))
```

#find datatype

```
sapply(new_df, typeof)
```

#find datatye in structural view

```
library(dplyr)

glimpse(new_df)
```

#Boxplot

```
boxplot(Income~ Purchased.Bike , data = new_df, xlab = "Buy Bikes",ylab = "Income", main = "Buy Bike according to income")
```

#Boxplot to show the range of ages having cars

```
boxplot(Age~ Cars , data = new_df, xlab = " CARS",ylab = "Age", main = "Cars according to age")
```

```
#Statistical Analysis
#summary of data
summary(new_df)
#Find mode for character data type columns
# Create the function.
getmode <- function(v) {
  unique_var <- unique(v)
  unique_var[which.max(tabulate(match(v, unique_var)))]
}
#mode of observing that the people purchased bikes more or not
getmode(new_df$Purchased.Bike)
getmode(new_df$Gender)
#Mostly bike buyers are male
getmode(new_df$Marital.Status)
#mostly are married
getmode(new_df$Cars)
#dnorm()
dnom<-dnorm(new_df$Cars)
dnom
plot(new_df$Cars,dnom)
#for having children
dn_child<-dnorm(new_df$Children)
dn_child
plot(new_df$Children,dn_child)
#pnorm()
pnorm_cars <- pnorm(new_df$Cars, mean = 2.5, sd = 2)
pnorm_cars
plot(new_df$Cars,pnorm_cars)
#qnorm()
qnor_cars <- qnorm(dnom, mean = 2, sd = 1)
qnor_cars
```

```
plot(dnom,qnor_cars)
#it will produace Nans
#binormal distributiion
#dbinom()
dbi=dbinom(new_df$Cars, 100, 0.5)
dbi
plot(new_df$Cars,dbi)
dbi2=dbinom(new_df$Children, 100, 0.3)
dbi2
plot(new_df$Children,dbi2)
#Poisson Regression
poisson_=glm(formula = Cars ~ Gender+Marital.Status, data = new_df,
      family = poisson)
summary(poisson_)
#chisqr
chisq.test(new_df$Purchased.Bike,new_df$Income)
#ANCOVA Analysis
result <- aov(Cars ~ Gender*Marital.Status,data = new_df)
summary(result)
#p value in both cases is less than 0.05. But the interaction between these two variables is not significant as the p-value is more than 0.05.
x=new_df$Purchased.Bike
t.test(x-9,alternative="two.sided",conf.level=0.95)
#There is strong evidence to reject the null hypothesis. There is overwhelming evidence against the null hypothesis.
#linear Regression
#label encoding
#install.packages("superml")
library(superml)
lbl = LabelEncoder$new()
new_df$Purchased.Bike = lbl$fit_transform(new_df$Purchased.Bike)
#0=no,1=yes
```

```
x=new_df$Children

y=new_df$Purchased.Bike

relation <- lm(y~x)

a <- data.frame(x = 2)

result <-  predict(relation,a)

result
```

#for Income Column

```
z=new_df$Income

relation <- lm(y~z)

a <- data.frame(z = 500000)

result <-  predict(relation,a)

result
```

#gives high accuracy upto 78%

#Multivariab;e linear regressionn

```
model <- lm(Purchased.Bike~Cars+Children+Income, data = new_df)

print(model)

a <- coef(model)[1]

print(a)

xcars <- coef(model)[2]

xchildren <- coef(model)[3]

xincome <- coef(model)[4]
```

 # predict having a car, 2 children and 30000 income

```
Y = a+xcars*1+xchildren*2+xincome*30000

Y

boxplot(Cars~ Marital.Status , data = new_df, xlab = " Marital Status",ylab = "HAVING cARS", main = "Cars according to marital status")
```