PREDICTION MODEL TO ANALYSE CAR PURCHASE GLM MODEL R Project

USING

Introduction:

We have taken data of 1000 customers arriving in a car showroom in a week (random sample).

The data has 6 columns :

- User ID(Nominal- labelling of a particular customer),
- Gender(Male assigned value "1" and Female assigned value "0",
- Age(in years),
- Annual Salary (in rupees) and
- AgeCatg; Categories are assigned to different age groups such as 17-24 as "1", 25-32 as "2" and so on.
- Our analysis is about predicting whether the individual will purchase a car or not based upon these characteristics.

METHODS:

- First we extract the data set and view it by writing the following code:

```
library(readxl)
Car3 <- read_excel(file.choose())
View(Car3)
```

- Then we try to get the summary of the dataset:

```
summary(Car3)
```

- Then we change the Age to a factor (categorical)

```
Car3$AgeCatg<- as.factor(Car3$AgeCatg)
summary(Car3)
```

From the summary, we can see that mean salary is greater than the median salary so the annual salary distribution is right skewed and mean very close to median so the age distribution is approximately normally distributed.

Then we attach the data : attach() function in R Language is used to access the variables present in the dataset without calling the data frame, it is done as the attach function stores the dataset in R's environment. Then we view the data.

```
attach(Car3)
View(Car3)
```

- Analysing the gender variable:

In our dataset Gender has 2 categories i.e.

Male(represented as "1") and Female(represented as"0")

```
table(cars.R$Gender)  #gives the number of males(1-516) and females (0-484)
```

writing the above code will show us the number of males in the data set and the number of females which are 516 and 484 respectively.

- The variable purchased means car purchased or not.

  i.e '1' for purchased and '0' for not purchased

```
cars.R <- data.frame(Car3)

head(cars.R)
```

- Writing the code below gives the number of persons purchased(1-402) car and not purchased(0-598) car i.e 402 and 598 respectively

  Calculating deviation in data.

```
table(cars.R$Purchased)
 sd(cars.R$Age)

sd(cars.R$AnnualSalary)
```

-> the total salary (GDP of our randomly chosen small economy)

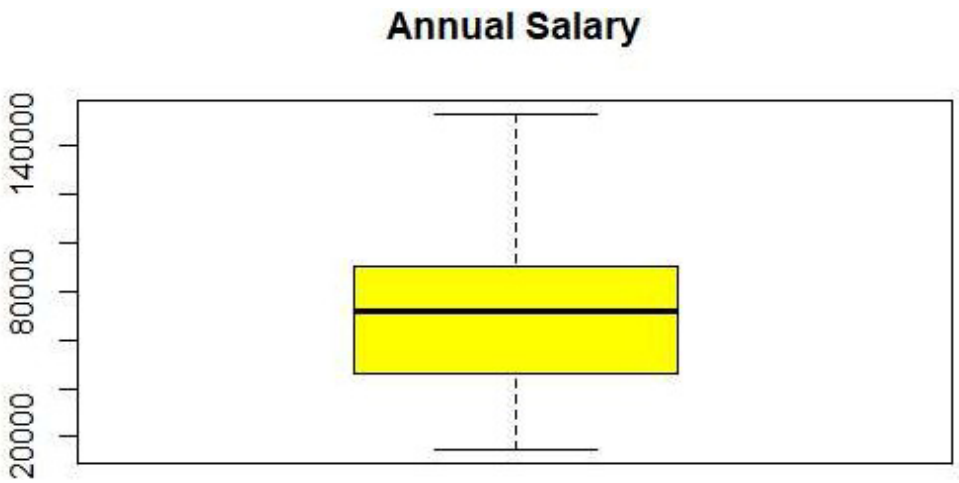
```
sum(cars.R$AnnualSalary)
```

Visualization of Data:

We make Box plots to check for the outliers

Here the box plot displays the summary of a set of data. The summary gives the minimum, first quartile, median, third quartile, maximum.

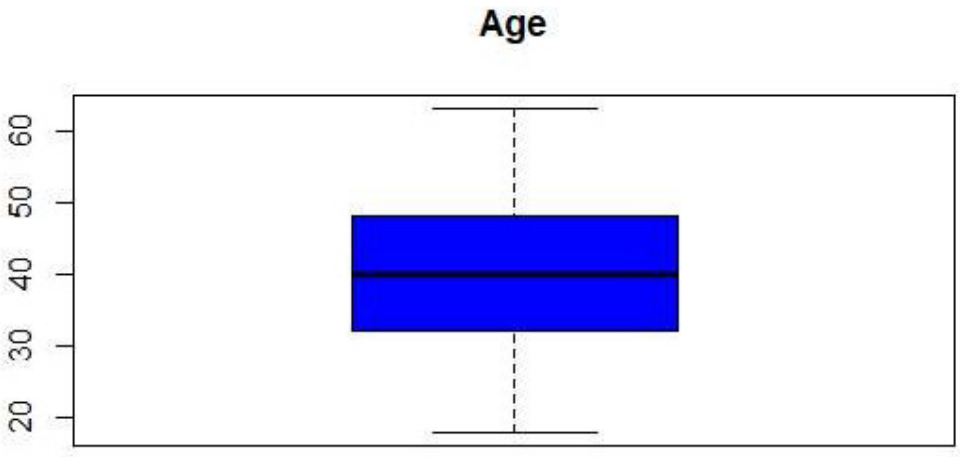
```
boxplot(cars.R$AnnualSalary, main = "Annual Salary", col = "Yellow")
```



Here most of the annual salary falls below the median salary value.

Now lets see the box plot for age:

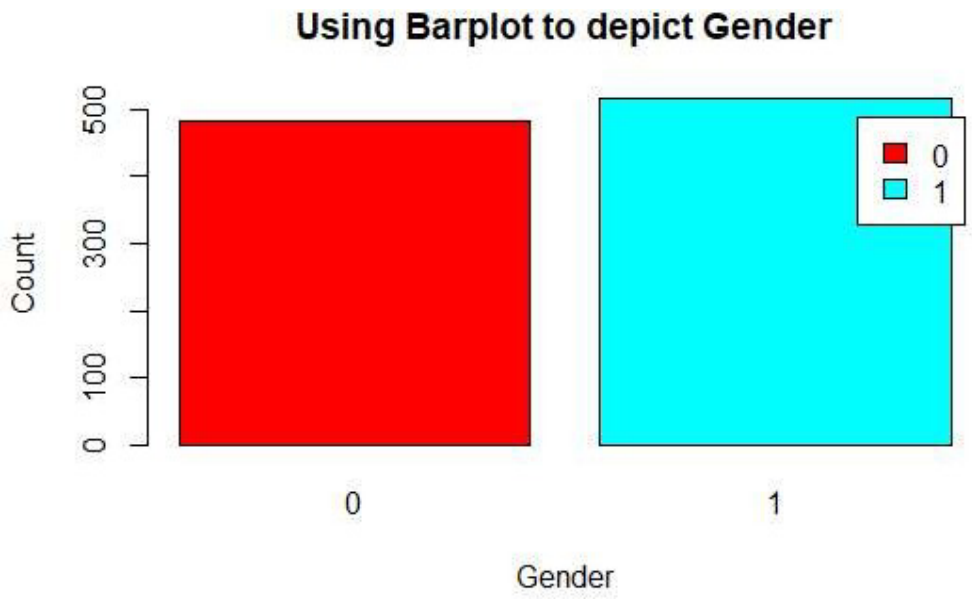
```
boxplot(cars.RSAge, main = "Age", col = "Purple")
```



We can see that the age variable is symmetrically or normally distributed. No outliers.

Now we can create the bar graph for Gender

```
a=table(cars.R$Gender)
barplot(a,main = "Using Barplot to depict Gender",
        ylab = "Count",
        xlab = "Gender",
        col=rainbow(2),
        legend=rownames(a))
```

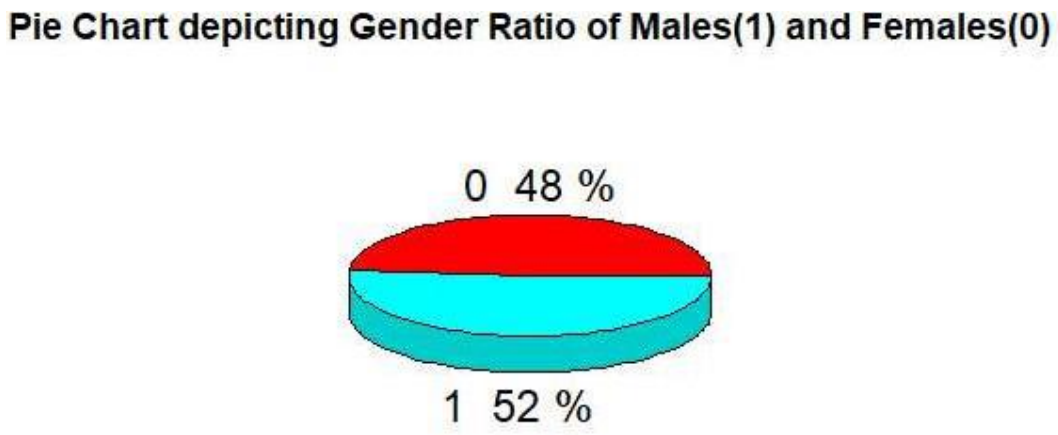


There are more males than females in the population.

 # Pie chart for Gender

```
library(plotrix)
pct=round(a/sum(a)*100)
lbs=paste(c("Females","Males"),"",pct,"%",sep = "  ")
```

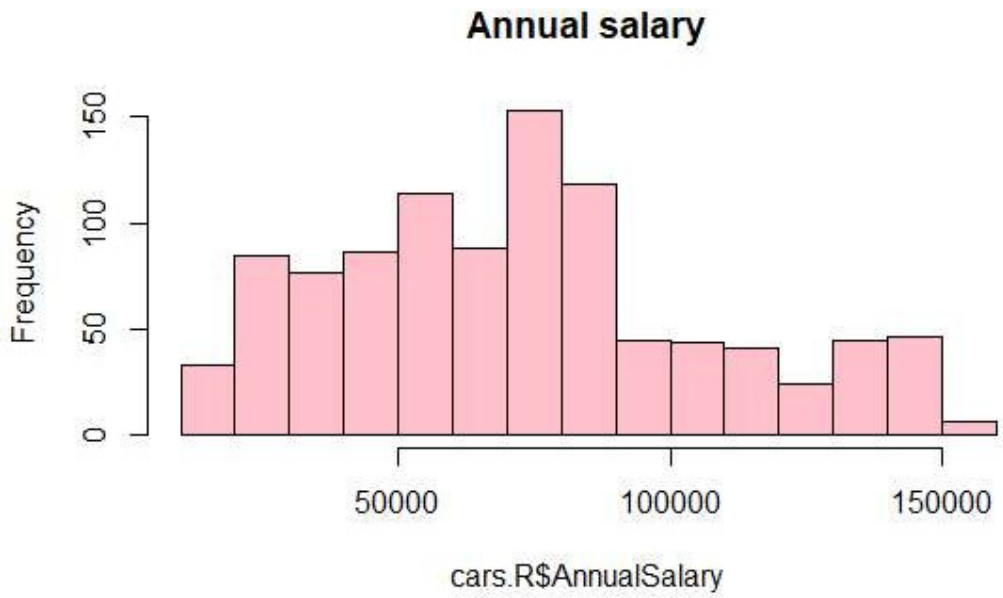
```
pie3D(a,labels = lbs,
    main="Pie Chart depicting Gender Ratio of Males(1) and Females(0)")
```

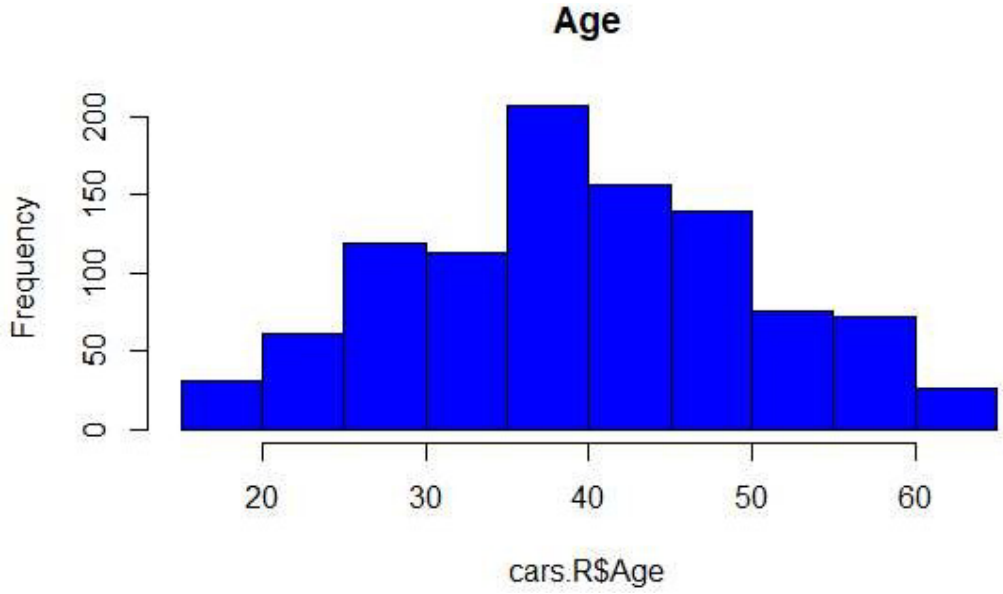


This shows us that we have 52% males in the data and 48% females.

We create Histogram to check the Distribution of data, using the following syntax:

```
hist(cars.R$AnnualSalary, main = "Annual salary", col = "pink")
hist(cars.R$Age, main = "Age", col = "blue")
```


Annual salary
Frequency
0
50
100
150
50000
100000
150000
cars.R$AnnualSalary


Age
Frequency
0
50
100
150
200
20
30
40
50
60
cars.R$Age

Creating a linear regression model:

We will create a Simple Linear Regression Model (only one independent variable) to show how age affects the car purchased.We are doing this to find out whether we can fit a linear relationship in the kind of dataset we have.

First let's take a look at plot of Y vs X:

```
plot(cars.R$Age,cars.R$Purchased, main= "Age Vs Purchased" , ylab="(Prob. of) Purchased" ,
    ylim = c(0,1) , las=1)
```

We want to model P(Purchased|Age), which is same as E(Y|X) like in Linear regression

So we segregate data using age groups for a better understanding.

We do so by adding vertical lines to separate the Age-Categories

```
abline(v=24, col="red")
abline(v=32, col="red")
abline(v=40, col="red")
abline(v=48, col="red")
abline(v=56, col="red")
abline(v=64, col="red")
```

Then we add labels to age categories:

```
mtext("1", side=1, adj=0.15, col = "Green")
mtext("2", side=1, adj=0.30, col = "Green")
mtext("3", side=1, adj=0.45, col = "Green")
mtext("4", side=1, adj=0.60, col = "Green")
mtext("5", side=1, adj=0.75, col = "Green")
mtext("6", side=1, adj=0.90, col = "Green")
```

We use AgeCategory to simplify so that we can calculate the mean Purchases for each category and look at the variable Purchased (if evidence of Purchased) for AgeCatg 1.

```
Purchased[AgeCatg=="1"]

#What proportion of people in AgeCatg =1 have evidence
mean(Purchased[AgeCatg=="1"])
```

>>no one from age category 1 purchases a car

```
#look at the variable Purchased (if evidence of Purchased) for AgeCatg 2
Purchased[AgeCatg=="2"]
#What proportion of people in AgeCatg =2 have evidence
mean(Purchased[AgeCatg=="2"])
#look at the variable Purchased (if evidence of Purchased) for AgeCatg 3
  <- c(mean(Purchased[AgeCatg=="1"]), mean(Purchased[AgeCatg=="2"]),
Purchased[AgeCatg=="3"]
    mean(Purchased[AgeCatg=="3"]), mean(Purchased[AgeCatg=="4"]),
  #What proportion of people in AgeCatg =3 have evidence
    mean(Purchased[AgeCatg=="5"]), mean(Purchased[AgeCatg=="6"]))
#Calculate means and store them in object called "p"
mean(Purchased[AgeCatg=="3"])
#this shows us the proportion of people from each age category who visit showroom and purchase a car

# Plot them Vs the mid-point of Age Categories
<- c(20.5,28.5,36.5,44.5,52.5,60.5)
```

Image

#add in the Points...

```
points(MidAge , p, pch="p", col="red")
```

Looking at the plot showing Purchased Yes = mod1

Not purchased = 0

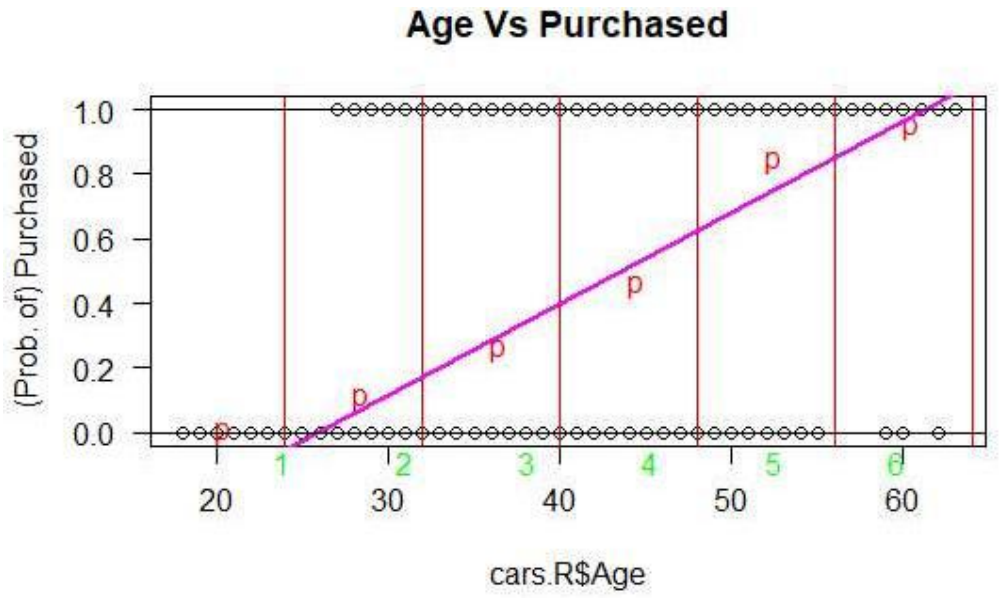
Trying to fit a linear model

```
mod1
<-lm(Purchased~Age)
```

We add in the Reg line (where we fit a linear regression)
Helping the visual, to see the problem

```
abline(mod1 , col="magenta", lwd=2)
abline(h=0, col="black") # Prob = 0

abline(h=1, col="black") # Prob = 1
```



#Problems in the model are:

1) The graph is Not linear as it should be in a linear regression, it is likely S-shaped curve usually because it is bounded between 0 and 1.

2) when the line goes below 0 and above 1, probability becomes negative or greater than 1 which is not possible as probability always lies between 0 and 1.

SO THE SOLUTION FOR THIS KIND OF DATA(WHERE DEPENDENT VARIABLE IS CATEGORICAL) IS GIVEN BY THE PREDICTION MODEL USING ***LOGISTIC REGRESSION MODEL***

#logit model (Logistic Regression Model- where Dependent variable is categorical)

We have used a Multiple Logistic Regression Model/ Generalized Linear Model for our analysis.

Taking Purchased as dependent variable and Gender,age and Annual Salary as independent variables.

```
cars_purchased <- glm(Purchased ~ Gender + Age + AnnualSalary, data = cars.R,
family = binomial(link="logit") )
```

here link ="logit" is the default function in R and it will work even if we don't specify

Binomial family is chosen as there are only two outcomes in the dependent variable

#calling for the model summary

```
cars_purchased
```

summary(cars_purchased)

```
Call:
glm(formula = Purchased ~ Gender + Age + AnnualSalary, family = binomial(link = "logit"),
    data = cars.R)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9283  -0.5945  -0.1348   0.4783   2.4760

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.186e+01  7.740e-01 -15.324   <2e-16 ***
Gender      -3.184e-01  1.855e-01  -1.716   0.0861 .
Age          2.195e-01  1.517e-02  14.471   <2e-16 ***
AnnualSalary 3.370e-05  3.232e-06  10.426   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1347.63  on 999  degrees of freedom
Residual deviance:  742.96  on 996  degrees of freedom
AIC: 750.96

Number of Fisher Scoring iterations: 6
```

THE ***purchased = -1.186e+01 +3.184e-01(Gender) + 2.195e-01(Age) 3.370e-05(Annual Salary)***

MODEL

Looking at this model we can see age and annual salary are highly significant as they have 3 asterisk sign on them.

Whereas gender is not that significant

Null deviance(1348) is similar to Explained Sum of Squares

Residual Deviance(743) is Similar to Residual Sum of Squares

#calling for ANOVA(Analysis of Variation) of the model

anova(cars_purchased)

Gender reduces the residuals by 2.23 and lost one degrees of freedom, whereas Age reduces residuals by a huge margin of 459.45 and Annual Salary reduces residuals by 142.99 by losing one degree of freedom each; therefore gender is not explaining much of variation in decision regarding purchasing a car or not, according to our dataset/sample.

HYPOTHESIS TESTING

Now we create another model by removing gender and checking whether models are similar to each other or different from each other, as we found that gender was not a significant variable in the model according to our dataset. We are testifying this fact using the hypothesis testing.

We are going to perform Hypothesis testing by Chi-Square test(Test used for non-normal data).

Our null hypothesis i.e H0 : Models are as good as each other

And alternative hypothesis is H1 : Models are different from each other

Now we run the model without gender

```
cars_purchased1 <- glm(Purchased ~  Age + AnnualSalary, data = cars.R,
family = binomial(link="logit") )

anova(cars_purchased, cars_purchased1, test = "Chi")
```

P-value here comes out to be 0.08

#UPDATING AN EXISTING MODEL

We could perform another task i.e updating an existing model. In car_purchased1 model we removed gender. So if we want to add it back we update it.

So, here we do not reject the null hypothesis (P value > 0.05) i.e both the models are as good as each other.

```
car_purchased_updated <- update(cars_purchased1,~.+Gender)

summary(cars_purchased1)

summary(car_purchased_updated)
```

Prediction using the model:

Prediction for a Male , aged 45 years having annual salary of Rs 150000

```
newdata <- data.frame(Gender= 1, Age = 45, AnnualSalary = 150000 2.753
predict.glm(cars_purchased, newdata) #1
```

We can predict that the person having above characteristics will purchase a car with 2.753 chance.

```
fitted(cars_purchased)
```

The fitted() function gives the fitted values, according to dataset gives the probability of each of the person purchasing a car, depending upon their characteristics.

Deviance Residuals : Here they are approximately Normally Distributed

```
residuals(cars_purchased)
```

Pearson Residuals : Skewed for Non Normal data

```
residuals(cars_purchased, type= "pearson")
```

Both residuals are identical in case of normal data but different for non normal data

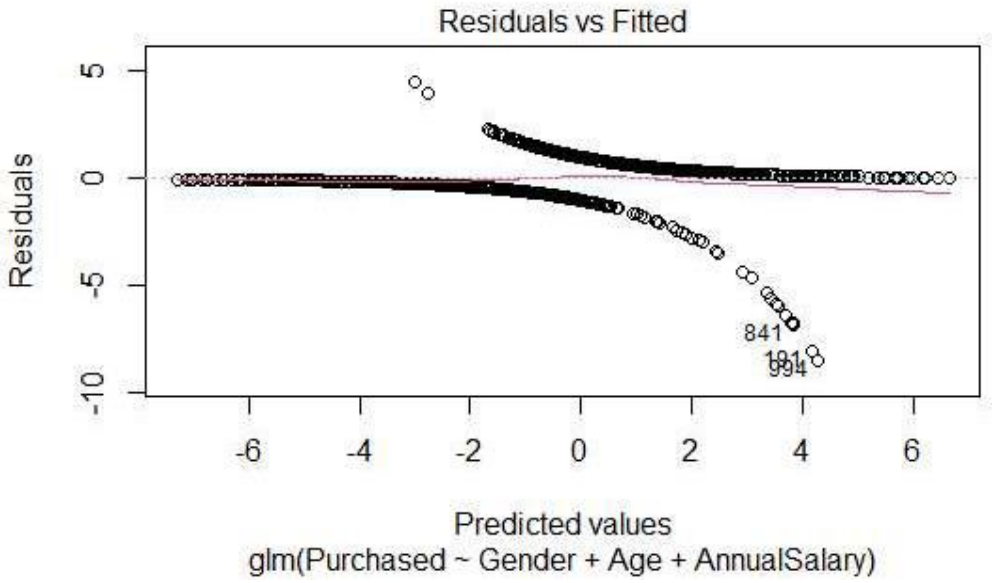
Plot of regression

To check whether Regression is correct and best fit or not, we use plots. The syntax for the function :

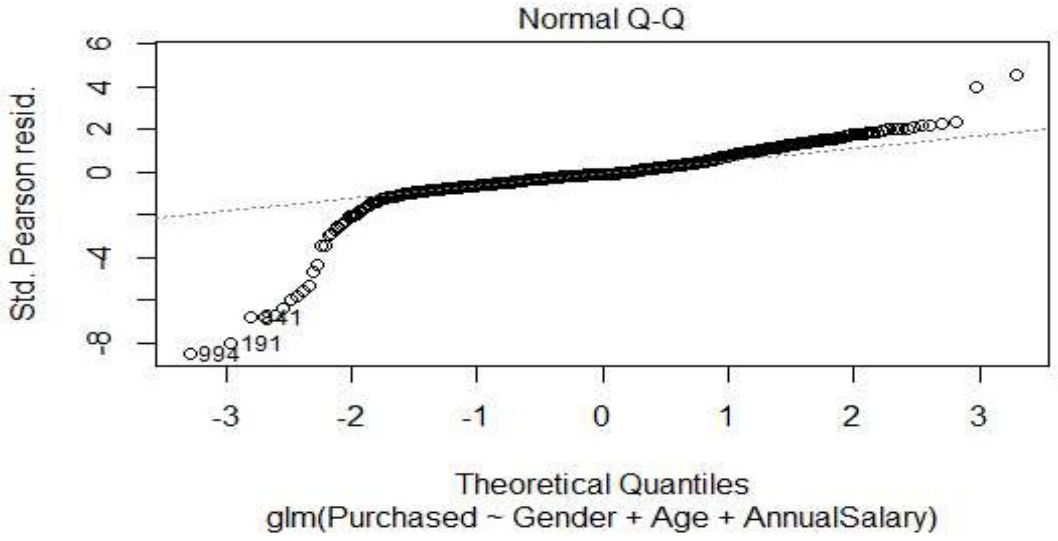
```
plot(cars_purchased)
```

R gives us following 4 graphs to understand the patterns in our regression models.

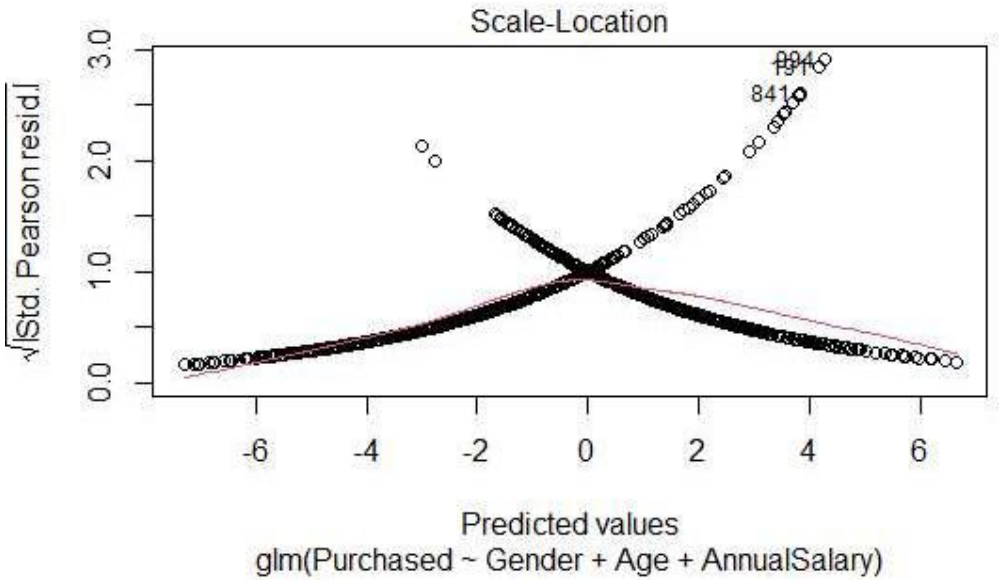
- **Residuals vs Fitted:** fitted/ predicted values given by red line and residuals plotted with black dots. The model here seems to be a nearly good fit.



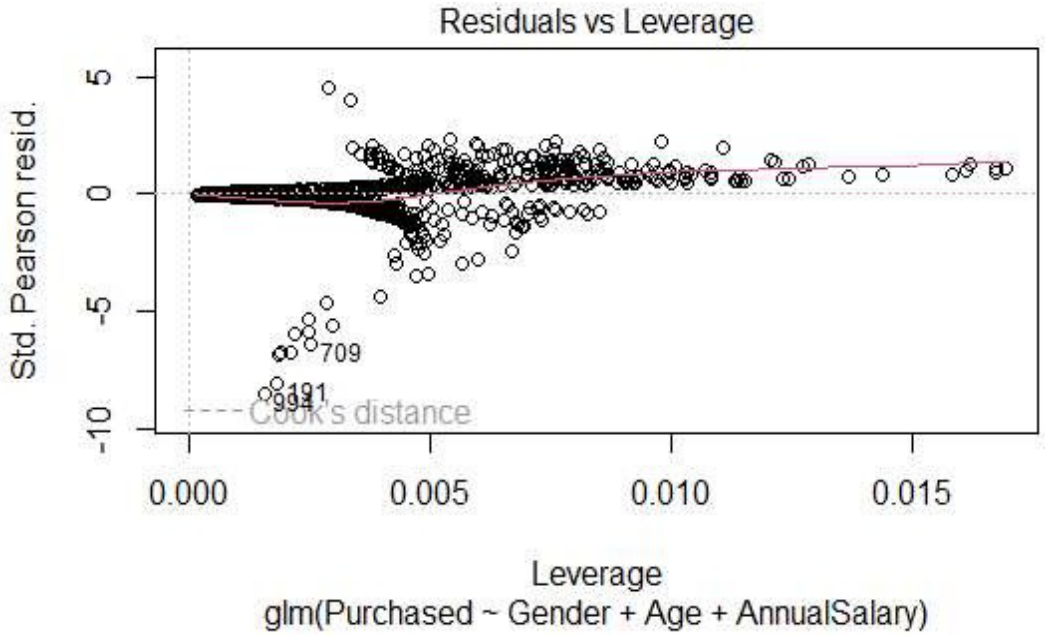
- **Normal Q-Q** : It tells us whether our standard deviation residuals follow the theoretical quantiles as per the dotted line.



- **Scale Location**: To check for the homoscedasticity assumption. The residuals are scattered around the red line with equal variability at all fitted values . Here we can see the problem of heteroscedasticity as the points increase/decrease like a funnel.



- **Residuals vs Leverage**: Helps to identify influential data points on the model

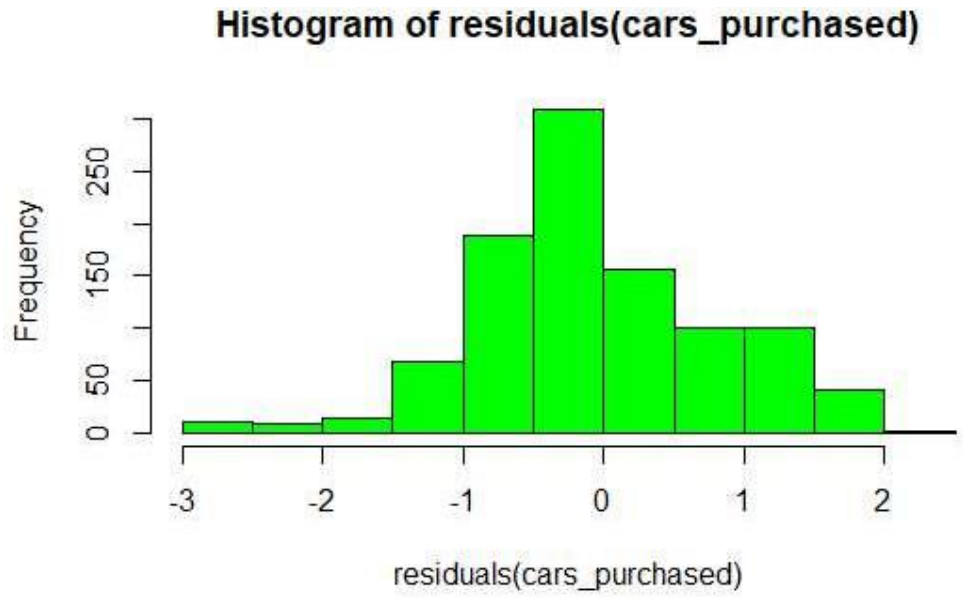


HISTOGRAM OF ERROR TERMS

To check normality of the error terms here we write to visualize the graph in the form of a histogram.

```
hist(residuals(cars_purchased))
```

We can see the residuals are near to normally distributed. When residuals are normally distributed , it means that our assumption(residuals should be normally distributed) is valid and model inference(the predictions) should be valid and meaningful.



CONCLUSION

Our study shows how to deal with a categorical dependent variable using Generalised Linear Models in R. GLM function helped us in predicting the cars purchased by people having different characteristics. Data visualisation using R helped us to visualise our data and draw some important conclusions about the data and purchasing pattern of customers.