Que 2.Output variable -> y

y -> Whether the client has subscribed a term deposit or not

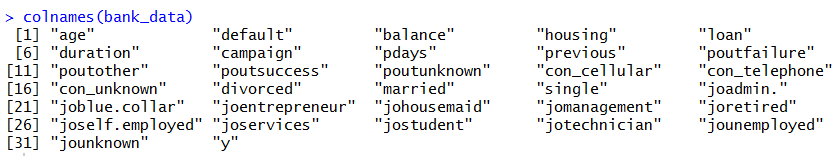
Binomial ("yes" or "no")

Sol:- **Objective** – Bank\_data -> y -> Whether the client has subscribed a term deposit or not

**Data pre-processing and Inferences from the data Set**

Data set talks about the term deposit subscription with respect to multiple variables with 45211 Observations

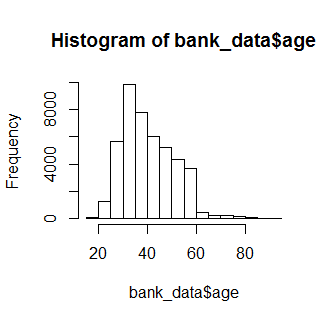
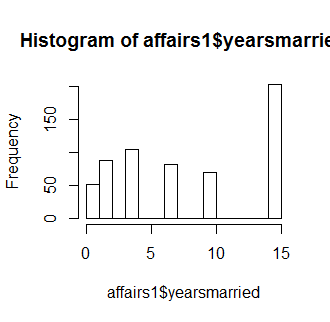
**Columns:**

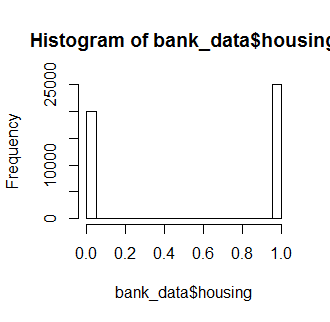
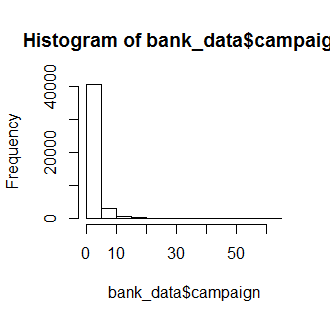
****

Dataset Size: 45211

Data given is found to be a continuous data and discrete for which a logistic regression can be performed getting deeper in to the data analysis and its behavior.

* Here we observed that 5289 people have subscribed the term deposit and 3922 had not subscribed.

hist(bank\_data$age) # we can observe that most of the subscription taken 25yrs to 60yrs

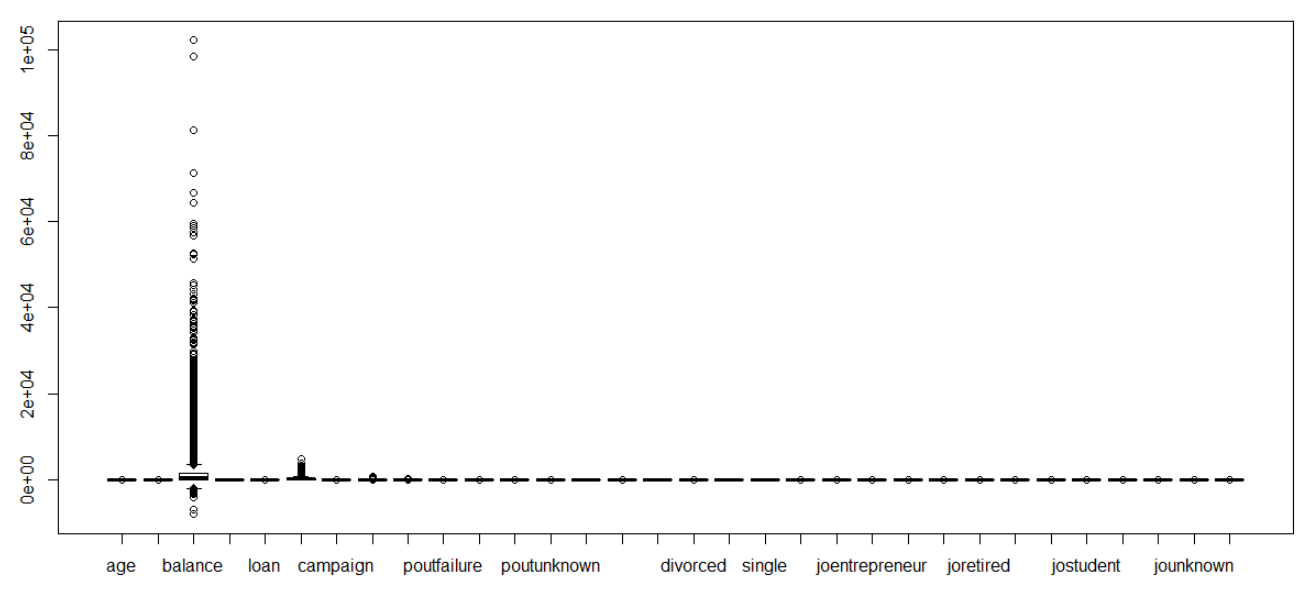
hist(bank\_data$campaign) # maximum subscription were taken in 0 to 5 campaign

hist(bank\_data$married) # there is not much difference between married and unmarried but still married people are more

hist(bank\_data$loan) # loan is not contribution much because very less people had taken loan

hist(bank\_data$housing) # housing contribution is a almost equal response of 0's and 1's

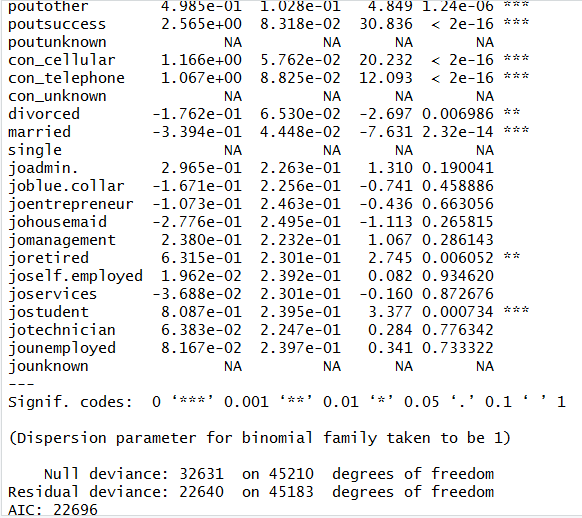
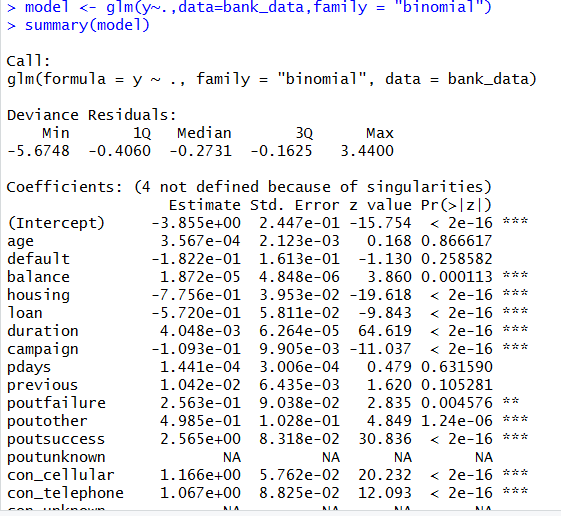
Boxplot :-

 Balance variable have more outliers but they all are valid data.

**Model Building**:

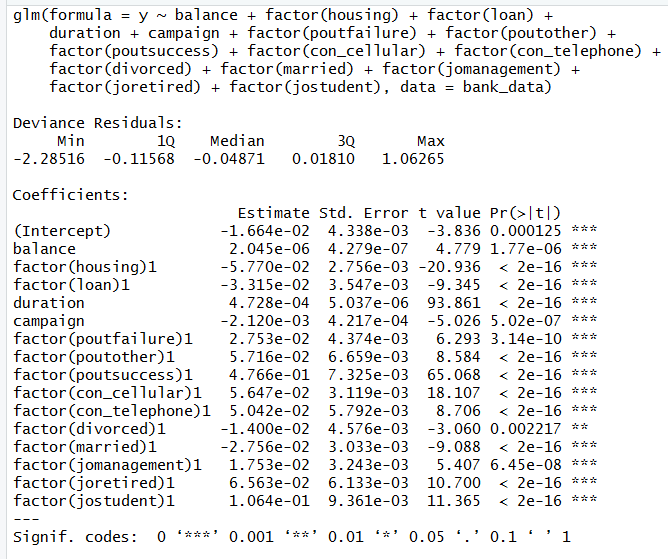
We use **glm() function from Base Package in R-studio** to estimate the years of experience using the other variable Salary whereas in python **LogisticRegression() is used from the sklearn package**

Model 1

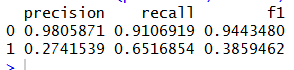


Some of the variables p-values are insignificant their contribution is very less or may not be having any contribution in subscription of term deposit. So will build another model after removing some variables to get the more accuracy with more significant variables.

**Model 2**- here we observed that all the variables are significant with less null deviance and residual deviance

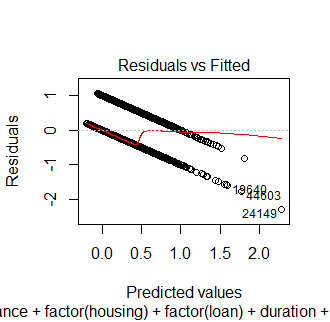
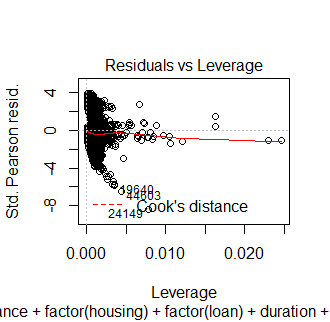
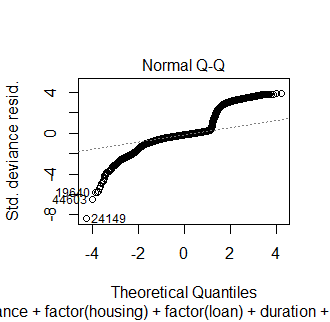
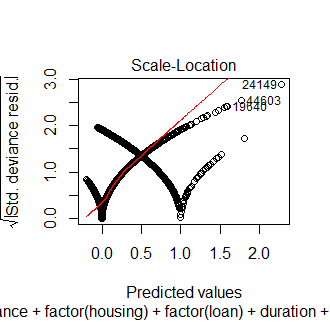


Model Accuracy = 0.8979

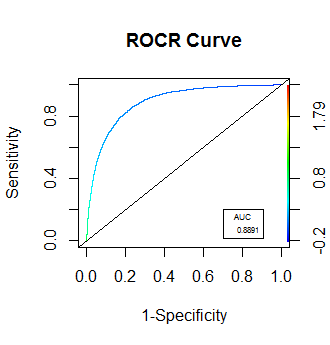


The model is able to predict 0’s (FAlSE) better than 1’s(TRUE)

Plot model 2

**Now plotting the ROCR curve to get the cutoff values**



The Area under the curve is **88.91**

**Packages used**

R studio

* Readr
* Ggplot2
* Moments
* (ROCR)

Python

* import numpy as np
* import pandas as pd
* import matplotlib.pylab as plt
* import statsmodels.formula.api as smf
* import seaborn as sns
* from sklearn.linear\_model import LogisticRegression
* from sklearn.model\_selection import train\_test\_split
* from sklearn.metrics import confusion\_matrix
* from sklearn.metrics import confusion\_matrix
* from sklearn.metrics import accuracy\_score
* from sklearn.metrics import recall\_score
* from sklearn.metrics import precision\_score
* from sklearn.metrics import f1\_score