**I have a dataset containing family information of married couples, which have around 10 variables & 600+ observations.**

**Independent variables are ~ gender, age, years married, children, religion etc.**

**I have one response variable which is number of extra marital affairs.**

**Now, I want to know what all factor influence the chances of extra marital affair.**

**Since extra marital affair is a binary variable (either a person will have or not),**

**so we can fit logistic regression model here to predict the probability of extra marital affair.**

Ans of Que 1:-

**Objective** - Affairs -> **predict the probability of extra marital affair.**

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

Data set talks about the Affairs with respect to multiple variables with 601 Observations

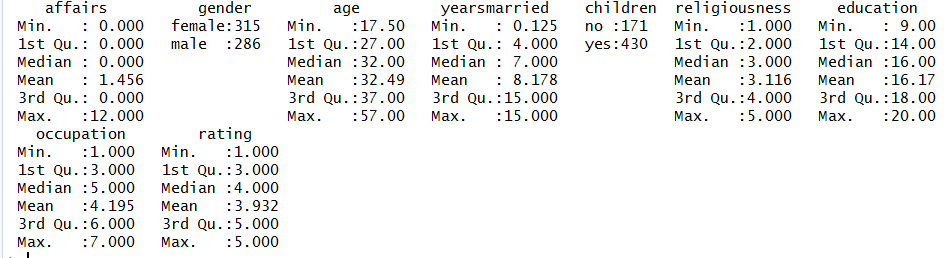
**Columns:**

"affairs" "gender" "age" "yearsmarried" "children" "religiousness" "education" “occupation" "rating"

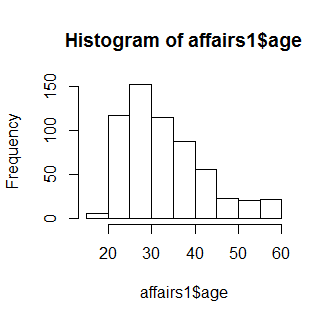
Dataset Size: 601

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.

**Affairs info.**

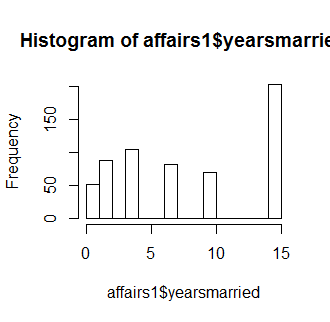


* Here we observed that minimum affair is “0” and maximum goes till “12”
* All the observation mean is greater than median except occupation and rating
* Data is skewed
* Data is normally distributed
* The age and education have positive kurtosis, it means it has thin peak and wider tails.
* Education and occupation have negative skewness.

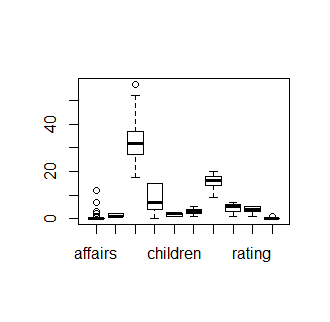


From this histogram we can notice that most of the affairs are between

20 to 45 age group.



From this histogram of yearsmarried we noticed that maximum affairs are taking place after 15 years of married



From this box plot we can notice that outliers are in affairs and in age features but those points are valid.

**Model Building**:

We use **glm() function from Base Package in R-studio** to estimate the affairs using the other variable whereas in python Logistic**Regression() is used from the sklearn package**

Call:

glm(formula = ynaffairs ~ factor(gender) + age + yearsmarried +

factor(children) + religiousness + education + occupation +

rating, family = "binomial", data = affairs1)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5713 -0.7499 -0.5690 -0.2539 2.5191

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.37726 0.88776 1.551 0.120807

factor(gender)male 0.28029 0.23909 1.172 0.241083

age -0.04426 0.01825 -2.425 0.015301 \*

yearsmarried 0.09477 0.03221 2.942 0.003262 \*\*

factor(children)yes 0.39767 0.29151 1.364 0.172508

religiousness -0.32472 0.08975 -3.618 0.000297 \*\*\*

education 0.02105 0.05051 0.417 0.676851

occupation 0.03092 0.07178 0.431 0.666630

rating -0.46845 0.09091 -5.153 2.56e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 675.38 on 600 degrees of freedom

Residual deviance: 609.51 on 592 degrees of freedom

AIC: 627.51

Number of Fisher Scoring iterations: 4

Call:

lm(formula = Salary ~ YearsExperience)

Residuals:

Min 1Q Median 3Q Max

-7958.0 -4088.5 -459.9 3372.6 11448.0

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 25792.2 2273.1 11.35 5.51e-12 \*\*\*

YearsExperience 9450.0 378.8 24.95 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5788 on 28 degrees of freedom

Multiple R-squared: 0.957, Adjusted R-squared: 0.9554

F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16

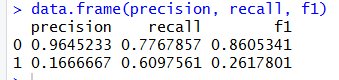
**p-values:**

values are significant as it is less than 0.5 except education, occupation, gender

now using confusion matrix we calculate the followings:

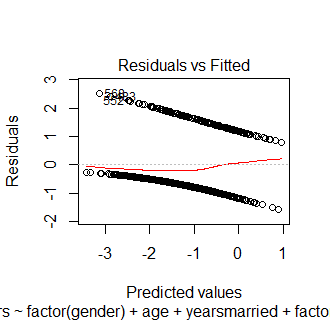
**Accuracy= 0.7653**

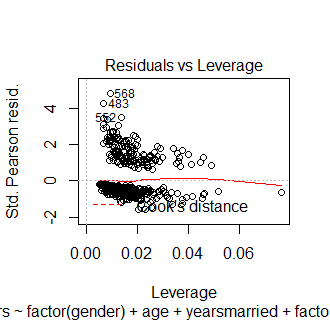
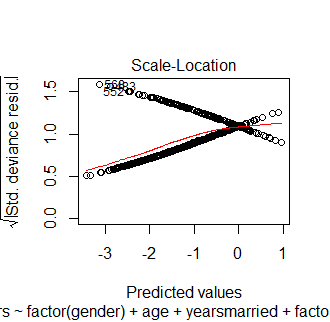
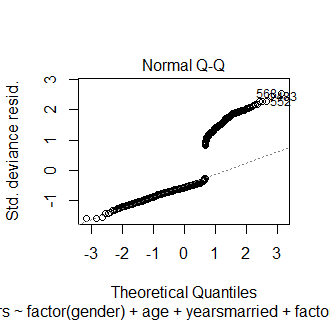
**Thus your model can predict the probability of extra marital affair with an accuracy of 0.7653.**



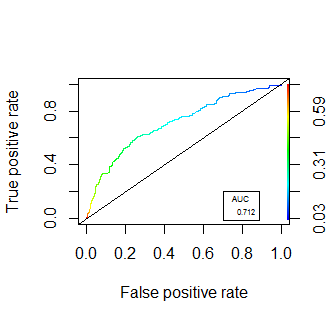
The model is able to predict 0’s (FAlSE) better than 1’s(TRUE)

Plot(model)





**ROC CURVE**

****

Area under the curve is 0.712

**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

Please refer the attached R and Python file for codes.