Mini Project – Factor Hair Case Study

Project Report

Jyosmitha M

Table of Contents

1. Project Objective

The objective of the report is to explore the factor hair revised data set ("Factor-Hair-revised.csv") in R and to build an optimum regression model to predict customer satisfaction. This report will consist of the following:

- Importing the dataset in R
- Understanding the structure of dataset
- Graphical exploration
- Descriptive statistics
- Insights from the dataset
- Existence of multi collinearity among the variables and how to overcome it
- Grouping the correlated factors into components
- Performing simple and multi linear regression on the variables
- Business insights.

2. Assumptions

Assumptions are as below:

- Data is normally distributed.
- There is a linear correlation between the independent and dependent variables
- there is no correlation between independent variables
- Regression exists in the population data

3. Environment Set up and Data Import

3.1 Install necessary Packages and Invoke Libraries

- library(dplyr)
- library(ggplot2)
- library(nFactors)
- library(psych)
- library(car)
- library(corrplot)
- library(ggplot2)
- library(DataExplorer)

3.2 Set up working Directory

Setting a working directory on starting of the R session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes etc. related to the project

```
setwd("C:/Users/ammu/Desktop/Great Lakes/3. Advanced Statistics/Project")
getwd()
```

3.3 Import and Read the Dataset

The given dataset is in .csv format. Hence, the command 'read.csv' is used for importing the file.

```
factorData=read.csv("Factor-Hair-Revised.csv")
```

3.4 Subsetting the required data for analysis

```
finalData=factorData[,2:12]
names(finalData)
```

```
##  [1] "ProdQual"    "Ecom"        "TechSup"     "CompRes"     "Advertising"
##  [6] "ProdLine"    "SalesFImage" "ComPricing"  "WartyClaim"  "OrdBilling"
## [11] "DelSpeed
```

4. Variable Expansion

Variable	Expansion
ProdQual	Product Quality
Ecom	E-Commerce
TechSup	Technical Support
CompRes	Complaint Resolution
Advertising	Advertising
ProdLine	Product Line
SalesFImage	Salesforce Image
ComPricing	Competitive Pricing
WartyClaim	Warranty & Claims
DelSpeed	Delivery Speed
Satisfaction	Customer Satisfaction

5. Assignment Questions and Answers with R-code and explanation

1. Perform exploratory data analysis on the dataset. Showcase some charts, graphs. Check for outliers and missing values.

1.1 EDA - Basic data summary, Univariate, Bivariate analysis, graphs

Ans: R functions like View, head, tail, dim, names, str, summary etc can be used for performing EDA.

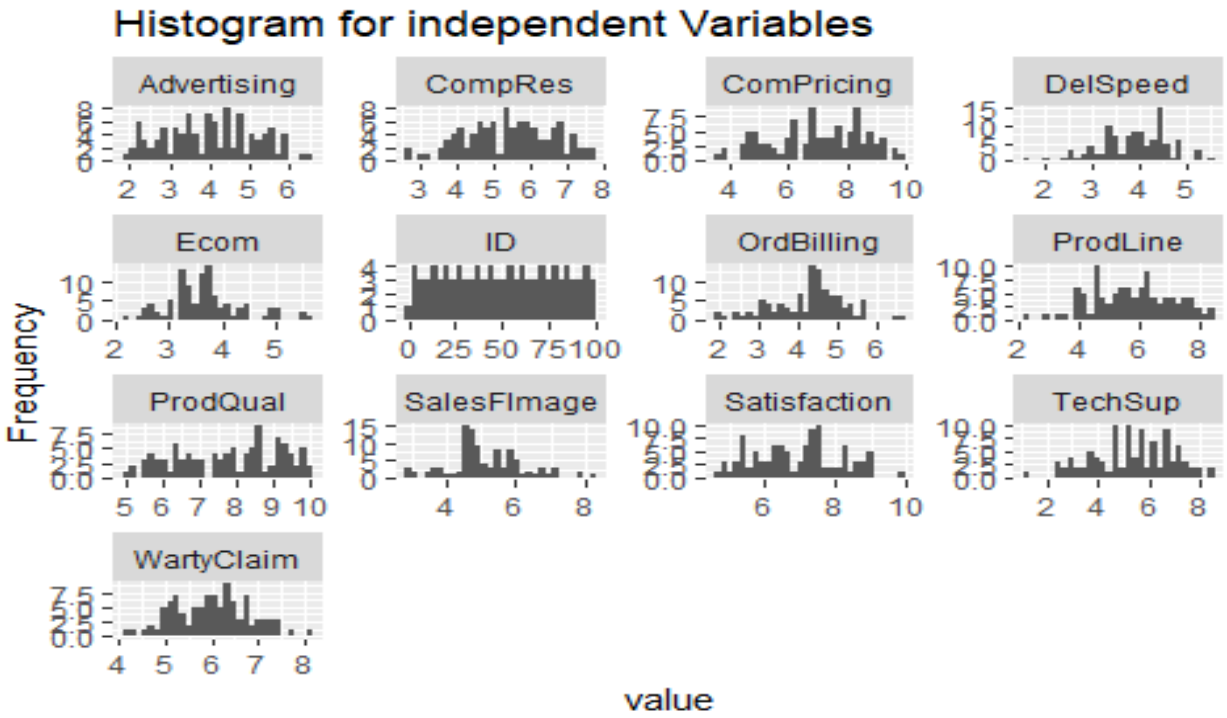
a. Univariate Analysis:

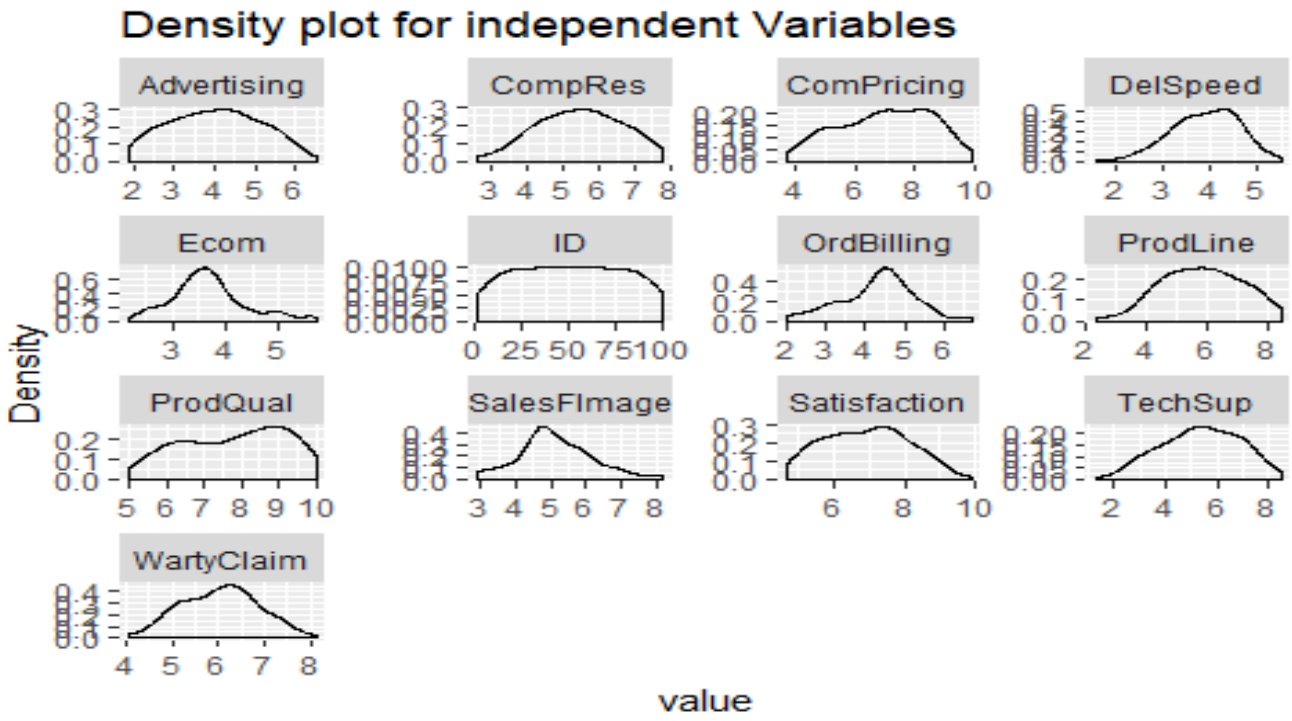
With the help of ***density plots*** and ***histograms***, the skewness in the variables of the dataset can be explained as below:

Normal distributed: advertising, complaint Resolution, competitative pricing, ID

left skew: Delivery speed, ordbilling, tech sup

right skew: ecommerce, satisfaction

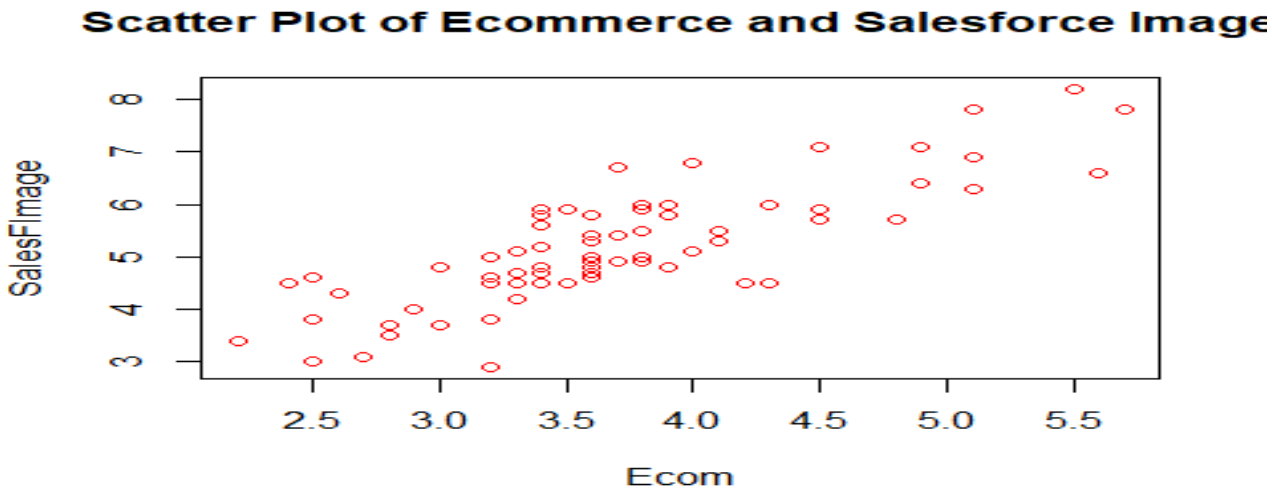


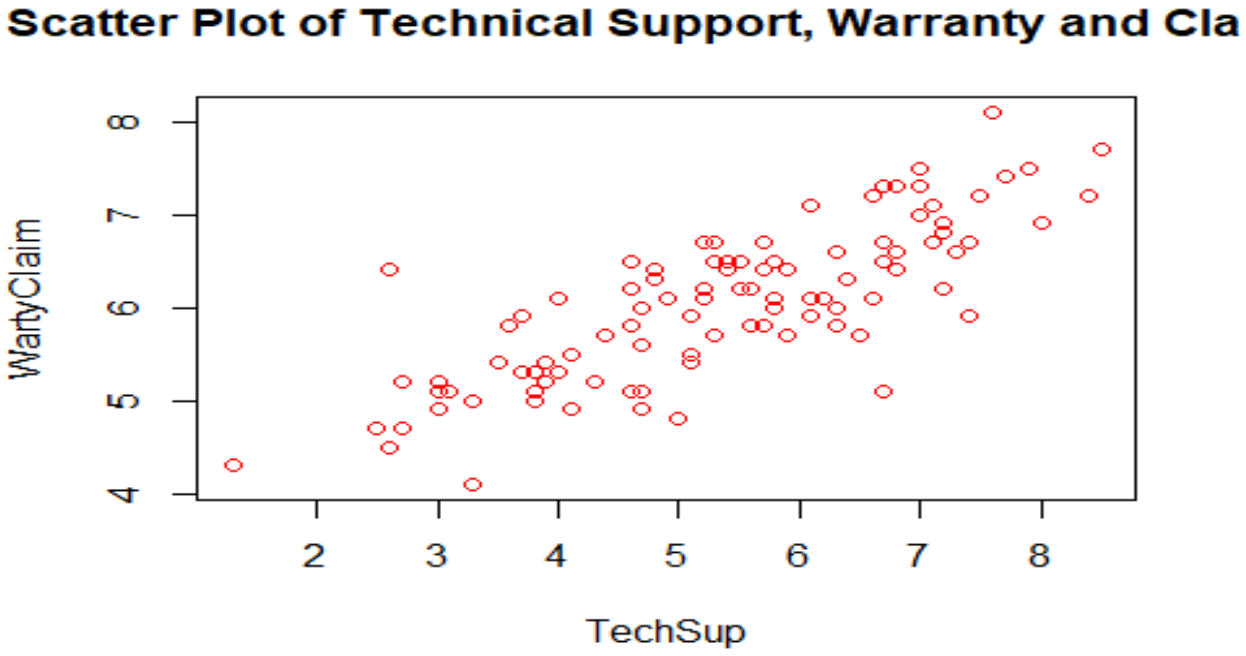


b. <u>Bivariate analysis:</u>

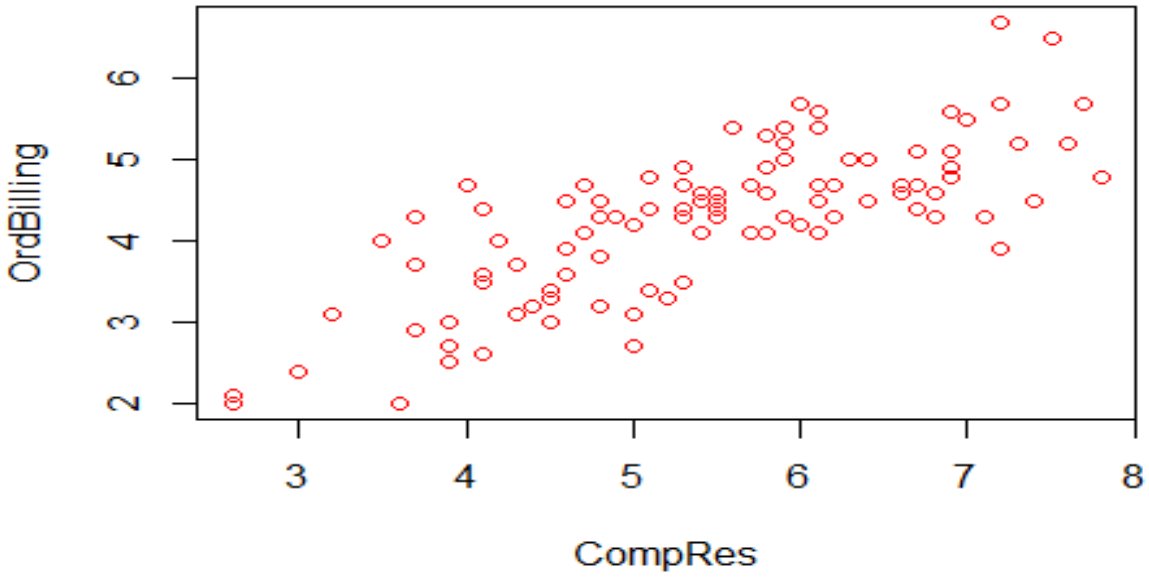
There is a high correlation existing among few variables and they are show below in the graph.

- Ecommerce and SalesforceImage
- Technical Support and Warranty & Claims
- Complaint resolution and ordered billing
- Complaint resolution and delivery speed
- Ordered billing and delivery speed

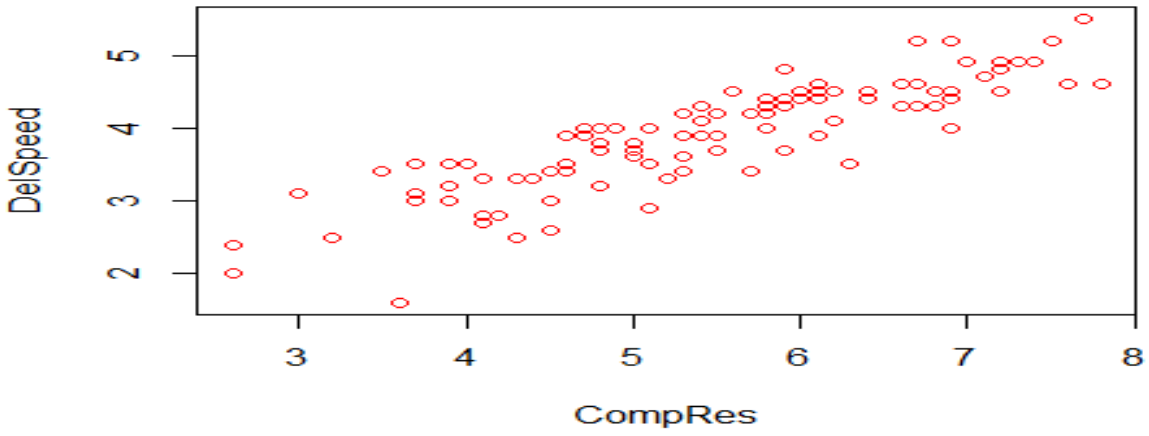

Scatter Plot of Ecommerce and Salesforce Image
SalesFImage
Ecom
3
4
5
6
7
8
2.5
3.0
3.5
4.0
4.5
5.0
5.5

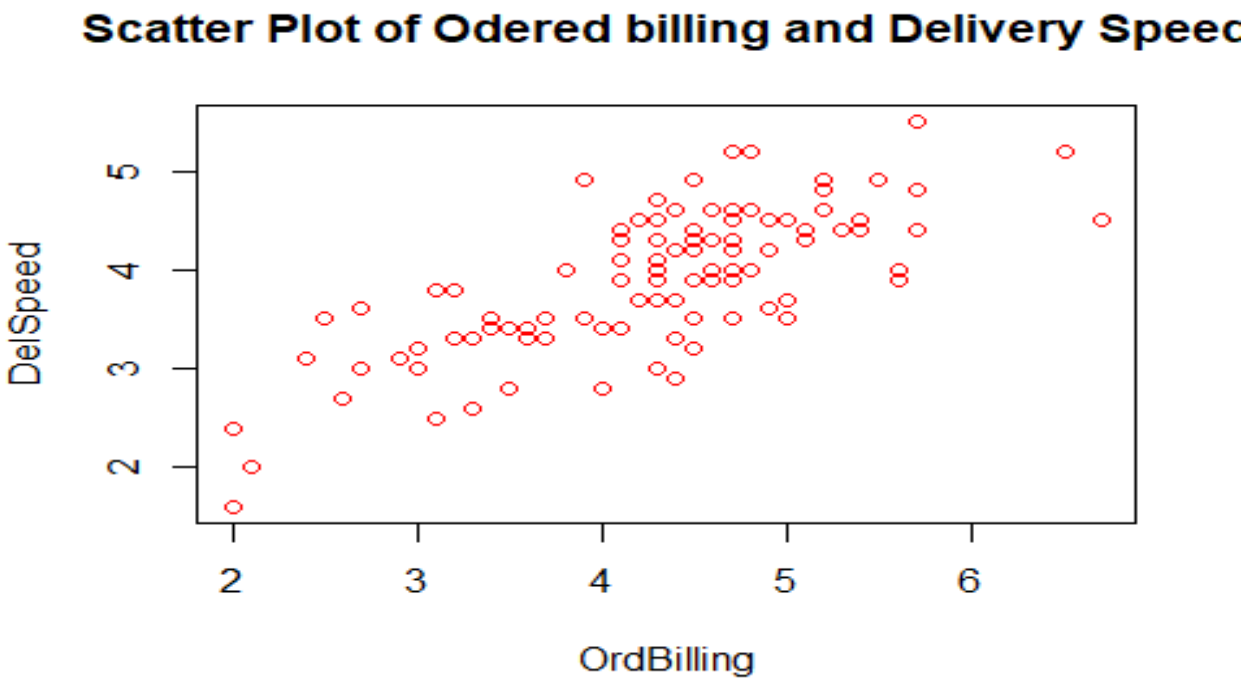

Scatter Plot of Technical Support, Warranty and Cla
WartyClaim
TechSup
4
5
6
7
8
2
3
4
5
6
7
8

Scatter Plot of Complaint Resolution and Ordered Bi



Scatter Plot of Complaint Resolution and delivery sp

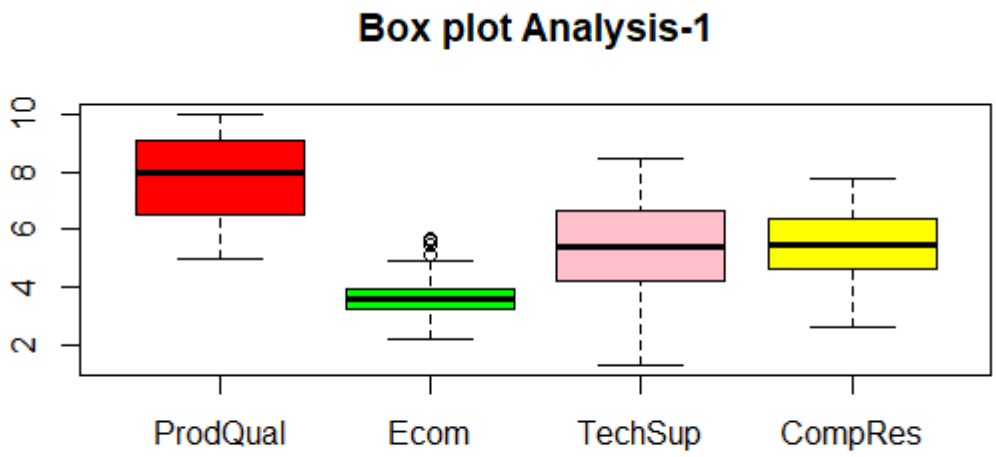


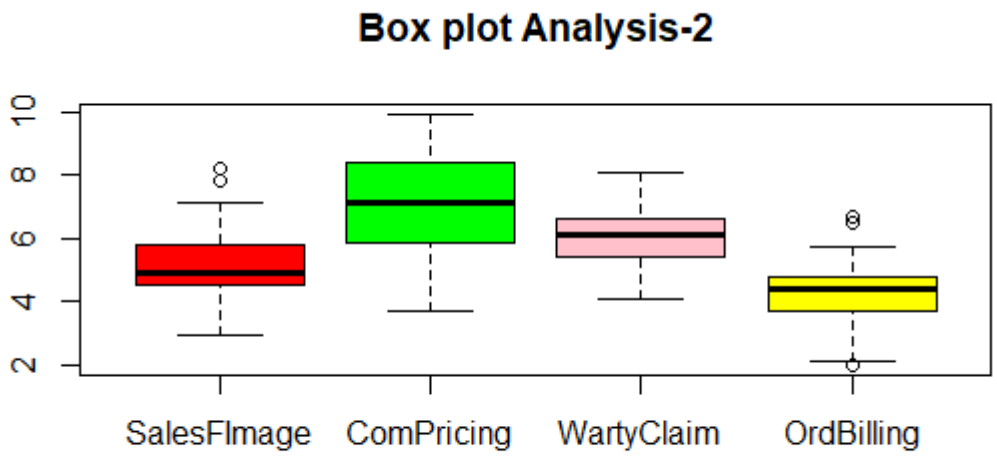


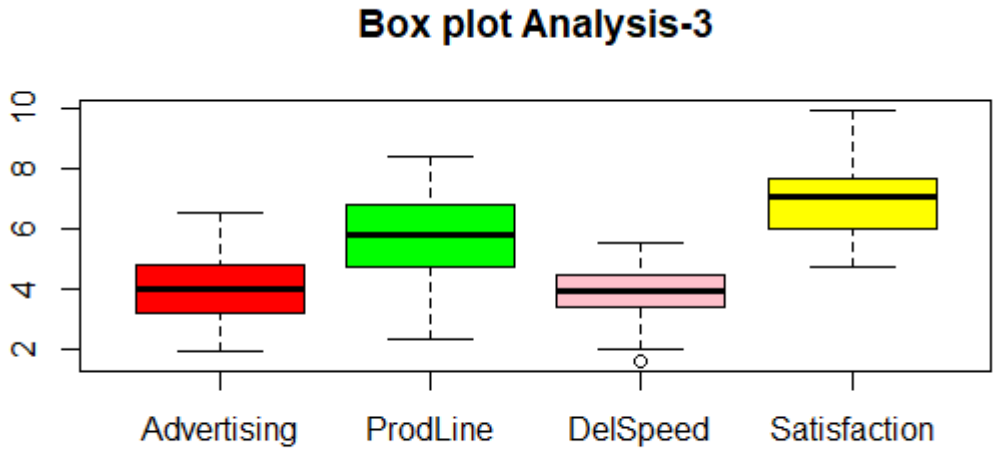
1.2 EDA - Check for Outliers and missing values and check the summary of the dataset

Ans: There are **no missing values/ Null values** in the entire data set in all the columns anyNA and is.na can be used to find out missing values.

Outliers are found in **Ecom, SalesFImage, Ordbilling variables**. Box plots can be used to identify the same.







2. Is there evidence of multicollinearity ? Showcase your analysis(6 marks)

Ans: The VIF Value of **few original variables in dataset are >2** and our dataset have variables which are having multi-collinearity

Below is the threshold values for multicollinearity check

a. <2 -> No collinearity

b. 2 and <4-5 -> Moderately collinear

c. 5 Highly collinear

A **correlation plot** can also be plotted to check the correlation among the variables.

Multi collinearity exists in the data set and below are the variables which have a high correlation.

E-commerce and salesForceImage=**0.79**

Technical Support and Warranty and claims=**0.8**

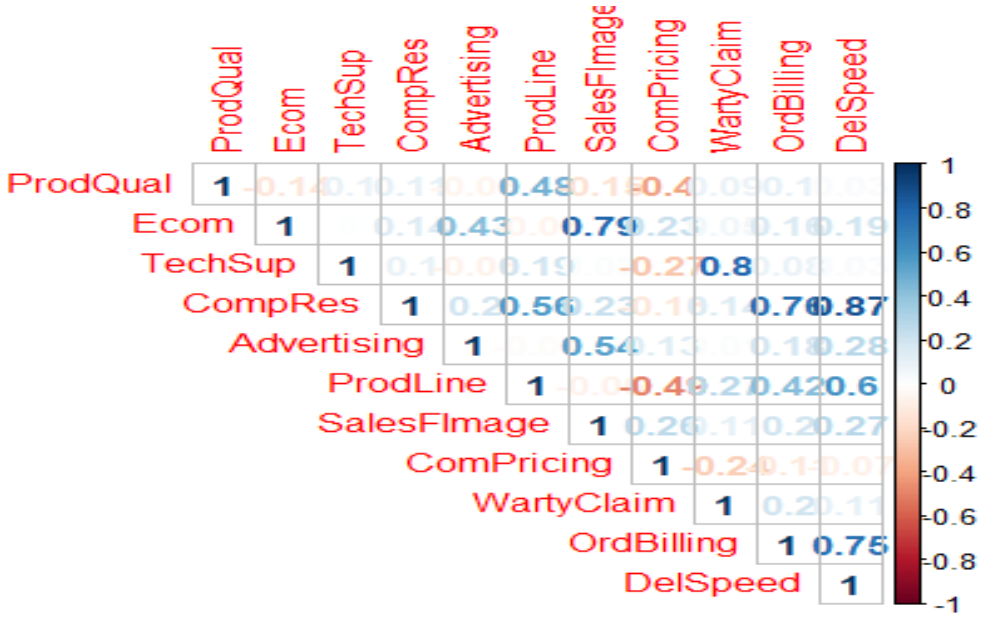
Complaint resolution and Order and billing=**0.76**

Complaint resolution and deliverySpeed=**0.87**

Order and billing and deliverySpeed =**0.75**

These collinear variables should be grouped into a single factors

If there is a collinearity among the variables, it will affect out regression model output as effect on one variable will also cause the effect on another variable.



3. Perform simple linear regression for the dependent variable with every independent variable (6 marks)

The general equation for simple linear regression is below:

yhat=beta0+beta1*X1 where

yhat = dependent variable

beta0= constant/intercept coefficient

beta1= slope/ variable coefficient/regression coefficient

X1= independent variable

The simple linear regression equations against each variable are given below:

Equation: yhat=3.67593+0.41512x**ProdQual**

For one-unit change in **ProdQual**, Satisfaction would improve by **0.41512** keeping other independent variables constant

Equation: yhat=5.1516+0.4811x**Ecom**

For one-unit change in Ecom, Satisfaction would improve by **0.4811** keeping other independent varialbes constant

Equation: yhat=6.44757+0.08768x**TechSup**

For one-unit change in TechSup, Satisfaction would improve by **0.08768** keeping other independent varialbes constant

Equation: yhat=3.68005+0.59499x**CompRes**

For one-unit change in CompRes, Satisfaction would improve by **0.59499** keeping other independent varialbes constant

Equation: yhat=5.6259+0.3222x**Advertising**

For one-unit change in Advertising, Satisfaction would improve by **0.3222** keeping other independent varialbes constant

Equation: yhat=4.02203+0.49887x**ProdLine**

For one-unit change in ProdLine, Satisfaction would improve by **0.49887** keeping other independent varialbes constant

Equation: yhat=4.06983+0.55596x**SalesFImage**

For one-unit change in SalesFImage, Satisfaction would improve by **0.55596** keeping other independent varialbes constant

Equation: yhat=8.03856-0.16068x**ComPricing**

For one-unit change in ComPricing, Satisfaction would decrease by **0.16068** keeping other independent varialbes constant

Equation: yhat=5.3581+0.2581x**WartyClaim**

For one-unit change in WartyClaim, Satisfaction would improve by **0.2581** keeping other independent varialbes constant

Equation: yhat=4.0541+0.6695x**OrdBilling**

For one-unit change in OrdBilling, Satisfaction would improve by **0.6695** keeping other independent varialbes constant

Equation: yhat=3.2791+0.9364x**DelSpeed**

For one-unit change in DelSpeed, Satisfaction would improve by **0.9364** keeping other independent varialbes constant

As per Simple Linear regression analysis, it is found that all the independent variables (ProdQual,Ecom,CompRes,Advertising,ProdLine,SalesFImage,ComPricing,OrdBilling,DelSpeed,DelSpeed) **except (TechSup,wrtClaim**)are significant on the dependent variable (Satisfaction). As per the R-Square value, a single variable cannot be taken into consideration for explaining the variation in the dependent variable.

4. Perform PCA/Factor analysis by extracting 4 factors. Interpret the output and name the Factors (20 marks)

4.1 Perform PCA/FA and Interpret the Eigen Values (apply Kaiser Normalization Rule)

Ans: We must check if PCA can be performed with **bartlett test** on the given dataset and find if variables are correlated and if the data is enough or not. Below is the null and alternate Hypothesis for performing the test.

Null Hypothesis: We cannot perform PCA on the given dataset

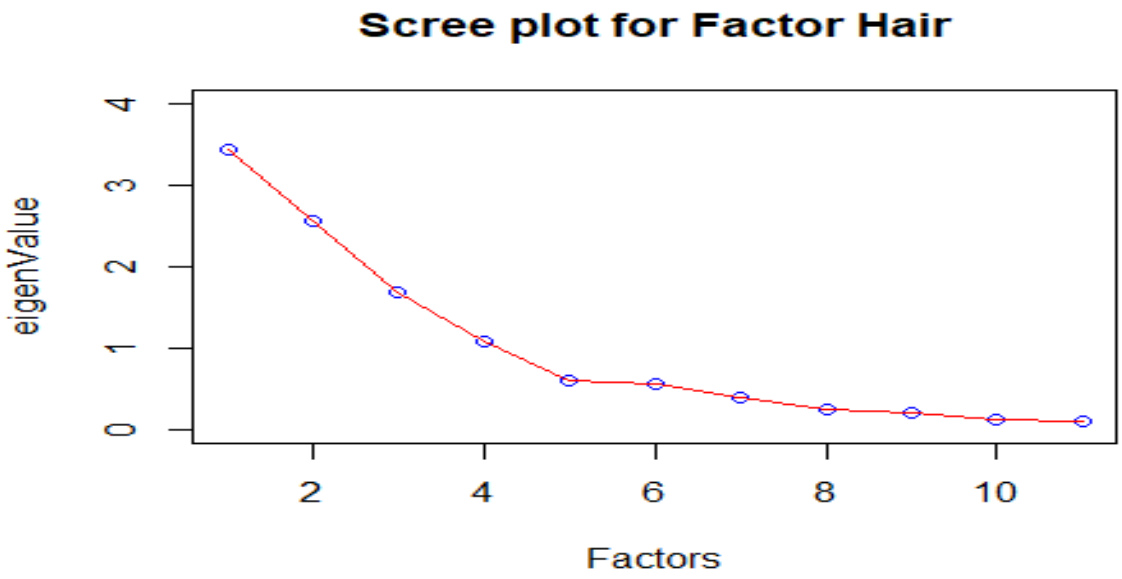
Alternate Hypothesis: We can perform PCA on the given dataset

From the below output, the P-value obtained is very low (1.79337e-96) and less significant. **Hence, we reject the null Hypothesis. So, PCA can be performed.**

We can calculate the eigen values with function “eigen”. These values helps us in choosing how many components to make without loosing the variance explaination. The eigen vectors tells us the direction

As per Kaiser Normalization Rule, eigen values >1 is considered to be as each factor. Here 4 variables are having eigen values >1. Hence, we are taking 4 factors.

As per scree plot, the factors which are below the elbow bent can be ignored. This process also leaves us with 4 factors.



4.2 Output Interpretation Tell why only 4 factors are being asked in the questions and tell whether it is correct in choosing 4 factors. Name the factors with correct explanations.

Ans: The number of factors choice is dependent on how much % of individual variation is being explained. In our output, **the first 4 eigen values (31.1542848 23.1899701 15.3725134 9.8777823) are explaning good percentage of variation.** Followed by them, there is a **drop in the eigen value percentage** and hence they are not quite useful in explaning the variation in data and also 100% of cumulative proportion is achieved with 4 components.

We can choose either more than 4 factors in our scenario to explain more variation in the data, but it **will lead to multi collinearity issue and our regression model won't be accurate** and hence not suggested.

Also as per the scree plot/kaiser normalisation rule, it is advised to take 4 factors only.

The variables which have high correlation are grouped into **4 factors** as below:

Component 1: Complaint Resolution, Order Billing, Delivery Speed (**Purchases**)

Component 2: Ecommerce, Advertising, Salesforce Image (**Marketing**)

Component 3: Technical Support, Warranty and Claims (**Customer Care**)

Component 4: Product Quality, Product Line, Competitive Pricing (**Product Position**)

Factors Explanation:

1. **Purchases**: comprises of variables which explains after the product is bought by the customer

2. **Marketing**: variables which explain about the publicity of the product in the market
3. **Customer Care:** variables which explains the support offered to customers once the product is bought
4. **Product Position**: variables which explains the basic details of the product

Output Interpretation:

cumulative variance: Around 79.59 % (approx. 80 %) of the cumulative variance can be explained by the 4 components generated in the problem

Eigen Values: the first 4 eigen values (31.1542848 23.1899701 15.3725134 9.8777823) are explaining good percentage of variation

Rotation: The rotated scores obtained by varimax rotation and helpful in grouping the variables into factors

Communality: From communality we can see that each factor is key role in explaining the common variance by each factor

RMSE: The root mean square error/residuals is `0.0596 and it is` low. The associated probability value is 0.001774.

Cumulative Proportion: 100% of cumulative proportion is achieved with 4 components

```
part.pca=eigenValue/sum(eigenValue)*100
part.pca
```

```
##  [1] 31.1542848 23.1899701 15.3725134  9.8777823  5.5402190  5.0171253
##  [7]  3.6501650  2.2450140  1.8504843  1.2076507  0.8947911
```

```
## Principal Components Analysis
## Call: principal(r = finalData, nfactors = 4, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 RC1     RC2     RC3     RC4     h2      u2   com
## ProdQual     0.0015 -0.0127 -0.0328  0.8757 0.7680 0.23197 1.003
## Ecom         0.0568  0.8706  0.0473 -0.1175 0.7771 0.22285 1.051
## TechSup      0.0183 -0.0245  0.9392  0.1005 0.8931 0.10689 1.025
## CompRes      0.9258  0.1159  0.0486  0.0912 0.8813 0.11874 1.057
## Advertising  0.1388  0.7415 -0.0816  0.0147 0.5760 0.42402 1.096
## ProdLine     0.5912 -0.0640  0.1460  0.6420 0.7871 0.21289 2.118
## SalesFImage  0.1325  0.9005  0.0756 -0.1592 0.8594 0.14055 1.122
## ComPricing  -0.0851  0.2256 -0.2455 -0.7226 0.6406 0.35944 1.471
## WartyClaim   0.1098  0.0548  0.9310  0.1022 0.8922 0.10775 1.059
## OrdBilling   0.8638  0.1068  0.0839  0.0393 0.7661 0.23391 1.054
## DelSpeed     0.9382  0.1773 -0.0046  0.0523 0.9144 0.08557 1.078
##
##                        RC1    RC2    RC3    RC4
## SS loadings           2.8927 2.2336 1.8555 1.7736
## Proportion Var        0.2630 0.2031 0.1687 0.1612
## Cumulative Var        0.2630 0.4660 0.6347 0.7959
## Proportion Explained  0.3304 0.2551 0.2119 0.2026
```

```
## Cumulative Proportion 0.3304 0.5855 0.7974 1.0000
##
## Mean item complexity =  1.2
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.0596
##  with the empirical chi square  39.0225  with prob <  0.001774
##
## Fit based upon off diagonal values = 0.9682
```

5.Perform Multiple linear regression with customer satisfaction as dependent variables and the four factors as independent variables. Comment on the Model output and validity. Your remarks should make it meaningful for everybody

5.1 Create a data frame with a minimum of 5 columns, 4 of which are different factors and the 5th column is Customer Satisfaction

Ans: we take the respective factor scores which explains the underlying variables a whole and create a dataframe alone the with predicted/dependent variable (satisfaction)

Using **cbind** data frame can be created as below and I have named the colum names of data frame in accordance with the component names as below:

"Satisfaction" "Purchases" "Marketing" "Customer Care" "Product Line"

```
factorDf=cbind(factorData$Satisfaction,Rotate$scores)
factorDf=as.data.frame(factorDf)
names(factorDf)=c("Satisfaction","Purchases","Marketing","Customer Care","Pro
duct Position")
names(factorDf)
```

```
## [1] "Satisfaction"     "Purchases"        "Marketing"        "Customer Car
e"
## [5] "Product Position"
```

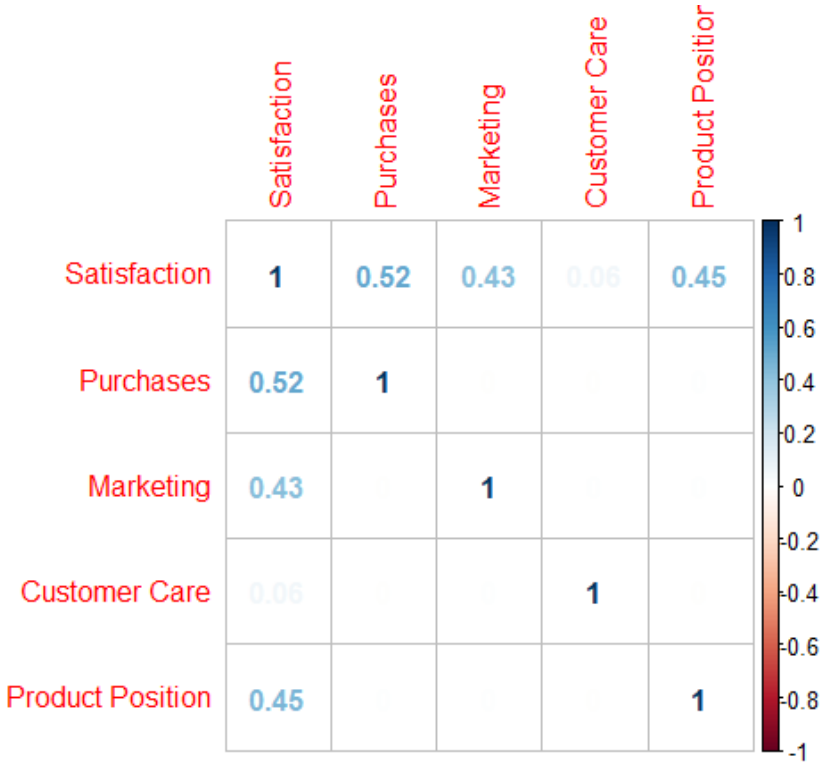
5.2 Perform Multiple Linear Regression with Customer Satisfaction as the Dependent Variable and the four factors as Independent Variables

Ans: MLR can be performed with function **lm** (linear mode) and predicted/dependent variable should be a function of factors/independent variable

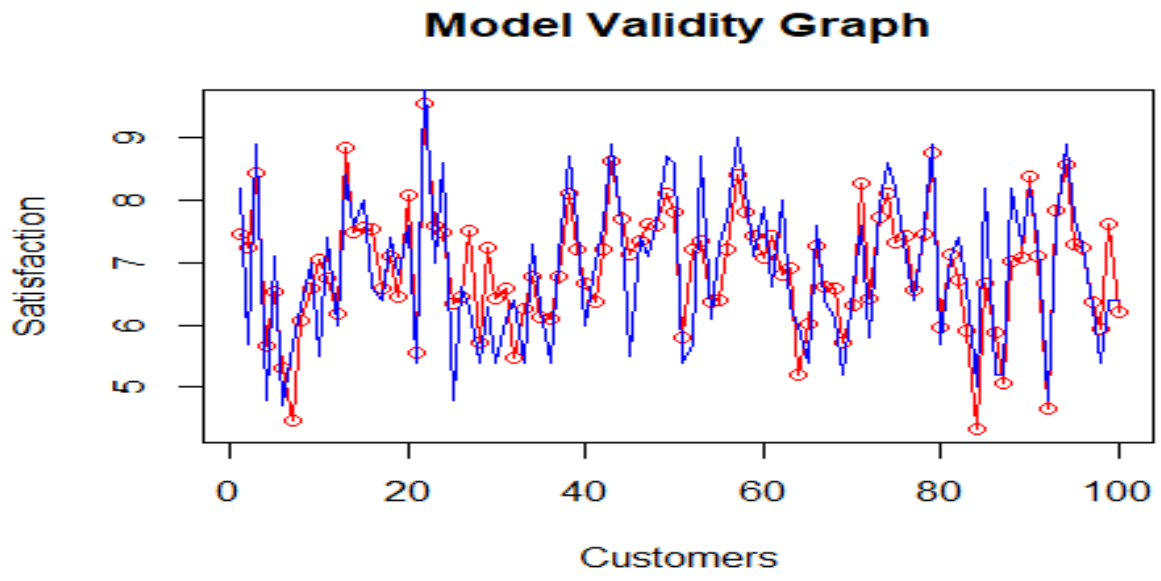
```
model=lm(Satisfaction~.,data=factorDf)
summary(model)
```

```
##
## Call:
## lm(formula = Satisfaction ~ ., data = factorDf)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6308 -0.4996  0.1372  0.4623  1.5228
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.91800    0.07089  97.589  < 2e-16 ***
## Purchases           0.61805    0.07125   8.675 1.12e-13 ***
## Marketing           0.50973    0.07125   7.155 1.74e-10 ***
## `Customer Care`     0.06714    0.07125   0.942    0.348
## `Product Position`  0.54032    0.07125   7.584 2.24e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7089 on 95 degrees of freedom
## Multiple R-squared:  0.6605, Adjusted R-squared:  0.6462
## F-statistic: 46.21 on 4 and 95 DF,  p-value: < 2.2e-16

vif(model)

##          Purchases          Marketing    `Customer Care` `Product Position
`
##                  1                  1                  1
1

finalcorPlot=corrplot(cor(factorDf),method="number")
```



```
##                   Satisfaction     Purchases     Marketing Customer Care
## Satisfaction       1.00000000  5.185657e-01  4.276870e-01  5.632963e-02
## Purchases          0.51856567  1.000000e+00 -4.511442e-16 -1.369372e-16
## Marketing          0.42768695 -4.511442e-16  1.000000e+00  2.966450e-16
## Customer Care      0.05632963 -1.369372e-16  2.966450e-16  1.000000e+00
## Product Position   0.45334892  6.517479e-16  9.043410e-17 -6.490954e-16
##                   Product Position
## Satisfaction          4.533489e-01
## Purchases             6.517479e-16
## Marketing             9.043410e-17
## Customer Care        -6.490954e-16
## Product Position      1.000000e+00

predictedSatisfaction=as.data.frame(predictedSatisfaction)
modelValidity=cbind(actualSatisfaction,predictedSatisfaction)

plot(modelValidity$predictedSatisfaction,col="Red",xlab = "Customers", ylab =
"Satisfaction", main="Model Validity Graph")
lines(modelValidity$predictedSatisfaction,col="Red")
lines(modelValidity$actualSatisfaction,col="Blue")
```



5.3 MLR summary interpretation and significance (R, R2, Adjusted R2, Degrees of Freedom, f-statistic, coefficients along with p-values)

Ans:

Regression equation: Satisfaction= 6.91800 +0.61805**Purchases** + 0.50973***Marketing**+ 0.06714* **Customer care**+ 0.54032***Product Position**

Significance: The components Purchases, Marketing and Product Positions shows more significance on satisfaction variable. **Customer care** component **doesn't** seem to have much significance.

Multiple R-square: 66.05% of customer satisfaction can be explained by the 4 independent factors (i.e Purchase, Customer care, Marketing, Product position) together.

Adjusted R-square: Helps us in suggesting how the % of variation explaination is being done when the degree of freedom is adjusted by adding more number of relevant variables. In our mode about **64.62%** is the adjusted R-square and it is calculated as: (1-R)*(n-1)/(n-k-1).

Degrees of Freedom: Here we are predicting 5 variables. Hence **Numerator DF=5-1=4**. We have 100 observation, hence total DF= 100-1=99. **Denominator DF= 99-4=95**

F-statistic: The probability of f-statistic > **46.21** is equal to 2.2e-16 which is much lesser than the alpha value significant at 0.05. (f-statistic is the ratio of ssb/ssw)

Coefficients along with p-values: Individual coefficients for the 4 factors are highly signification as evidence by tstat values and the respective p-values are very low even at 0.01 alpha significance

Model Output and validity:

The **VIF (variable Inflation Factor)** for this model is showing 1. Hence there is no multi collinearity exists in the model anymore.

Note: According to thumb rule, if VIF is <2, there regression model is assumed to be free from multi collinearity.

Also, the graph of actual Satisfaction (train dataset) vs Predicted satisfaction (test dataset) is almost similar. Hence our regression model has good validity

5.4 Output Interpretation/ Conclusion:

The satisfaction of the customer primarily influenced by how faster the product is billed and delivered along with how quickly the customer complaints are resolved by the company. Advertisements and the publicity of the product in social media and in the market also plays a descent role and followed by the product quality and its relative price with other products.

However, the services offered after the purchase of the product like warranty & claims and technical support are not quite playing a role in the customer satisfaction.

Hence the company must focus on pre-purchase activities than post-purchase activities