MINI PROJECT
- Factor Hair

Model Report

By – Nikhil Rawal

Project Overview

Multicollinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors.

The regression models with high multicollinearity can give you a high R squared but hardly any significant variables.

The analysis is carried out in the R environment for statistical computing and visualisation, which is an open-source dialect of the S statistical computing language. It is free, runs on most computing platforms, and contains contributions

The objective of the project is to use the dataset Factor-Hair-Revised.csv to build a regression model to predict satisfaction

Project Approach

- **Data Exploration**
- **Data Visualisation**
- **Checking the evidence of Multicollinearity**
- **Initial Regression analysis**
- **Factor Analysis**
- **Labelling the factors**
- **Regression analysis using the factors as independent variable**

- ## **Data Exploration**

#getwd()

>"F:/hair r data"

- Read Input File

```
>mydata=read.csv("Factor-Hair-Revised.csv", header = T)
```

- Names of the columns

```
>names(mydata)
"ID"          "ProdQual"     "Ecom"         "TechSup"
 "CompRes"      "Advertising"  "ProdLine"     "SalesFImage"
 "ComPricing"   "WartyClaim"   "OrdBilling"   "DelSpeed"
"Satisfaction"
```

*Hence, the column name 'ID' is just the column number, and do not have any use and explanatory power . So, we can drop it .

#mydata=mydata[,2:13]

```
>"ProdQual"     "Ecom"        "TechSup"     "CompRes"      "Advertising"  "ProdLine"
"SalesFImage" "ComPricing"   "WartyClaim"   "OrdBilling"   "DelSpeed"  "Satisfaction"
```

#Head(mydata)

```
ProdQual Ecom TechSup CompRes Advertising ProdLine SalesFImage ComPricing
1      8.5  3.9     2.5     5.9         4.8      4.9         6.0        6.8
2      8.2  2.7     5.1     7.2         3.4      7.9         3.1        5.3
3      9.2  3.4     5.6     5.6         5.4      7.4         5.8        4.5
4      6.4  3.3     7.0     3.7         4.7      4.7         4.5        8.8
5      9.0  3.4     5.2     4.6         2.2      6.0         4.5        6.8
6      6.5  2.8     3.1     4.1         4.0      4.3         3.7        8.5
  WartyClaim OrdBilling DelSpeed Satisfaction
1        4.7        5.0      3.7          8.2
2        5.5        3.9      4.9          5.7
3        6.2        5.4      4.5          8.9
4        7.0        4.3      3.0          4.8
5        6.1        4.5      3.5          7.1
6        5.1        3.6      3.3          4.7
```

- Dimension of data

```
# dim(mydata)
   100  13
```

- Structure of data

str(mydata)

```
>
data.frame'     :    100 obs. of  12 variables:
 $ ProdQual     :    num  8.5 8.2 9.2 6.4 9 6.5 6.9 6.2 5.8 6.4 ...
 $ Ecom         :    num  3.9 2.7 3.4 3.3 3.4 2.8 3.7 3.3 3.6 4.5 ...
 $ TechSup      :    num  2.5 5.1 5.6 7 5.2 3.1 5 3.9 5.1 5.1 ...
 $ CompRes      :  num  5.9 7.2 5.6 3.7 4.6 4.1 2.6 4.8 6.7 6.1 ...
 $ Advertising :   num  4.8 3.4 5.4 4.7 2.2 4 2.1 4.6 3.7 4.7 ...
 $ ProdLine     :    num  4.9 7.9 7.4 4.7 6 4.3 2.3 3.6 5.9 5.7 ...
 $ SalesFImage : num  6 3.1 5.8 4.5 4.5 3.7 5.4 5.1 5.8 5.7 ...
 $ ComPricing  :  num  6.8 5.3 4.5 8.8 6.8 8.5 8.9 6.9 9.3 8.4 ...
 $ WartyClaim  :  num  4.7 5.5 6.2 7 6.1 5.1 4.8 5.4 5.9 5.4 ...
 $ OrdBilling  :    num  5 3.9 5.4 4.3 4.5 3.6 2.1 4.3 4.4 4.1 ...
 $ DelSpeed     :    num  3.7 4.9 4.5 3 3.5 3.3 2 3.7 4.6 4.4 ...
 $ Satisfaction:    num  8.2 5.7 8.9 4.8 7.1 4.7 5.7 6.3 7 5.5 ...
```

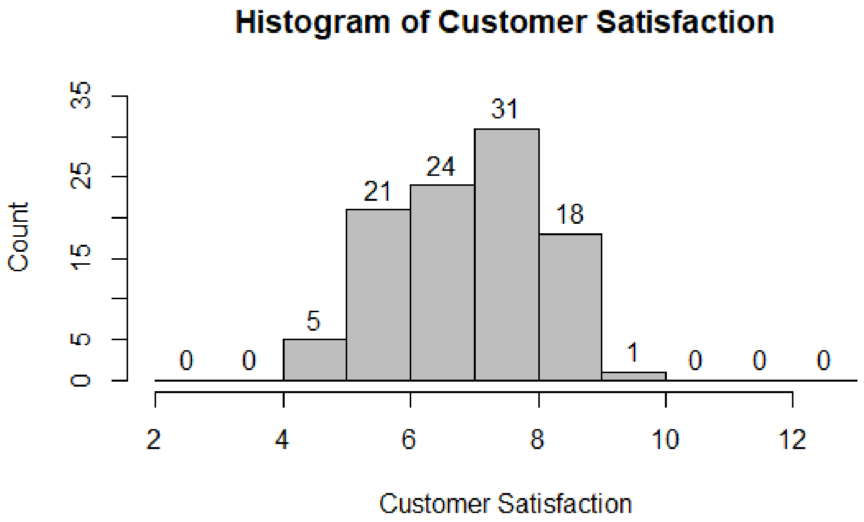
- Summary of data

#Summary(mydata)

```
>
ProdQual             Ecom              TechSup           CompRes
 Min.   : 5.000   Min.   :2.200    Min.   :1.300    Min.   :2.600
 1st Qu.: 6.575   1st Qu.:3.275    1st Qu.:4.250    1st Qu.:4.600
 Median : 8.000   Median :3.600    Median :5.400    Median :5.450
 Mean   : 7.810   Mean   :3.672    Mean   :5.365    Mean   :5.442
 3rd Qu.: 9.100   3rd Qu.:3.925    3rd Qu.:6.625    3rd Qu.:6.325
 Max.   :10.000   Max.   :5.700    Max.   :8.500    Max.   :7.800
  Advertising        ProdLine       SalesFImage      ComPricing
 Min.   :1.900   Min.   :2.300   Min.   :2.900   Min.   :3.700
 1st Qu.:3.175   1st Qu.:4.700   1st Qu.:4.500   1st Qu.:5.875
 Median :4.000   Median :5.750   Median :4.900   Median :7.100
 Mean   :4.010   Mean   :5.805   Mean   :5.123   Mean   :6.974
 3rd Qu.:4.800   3rd Qu.:6.800   3rd Qu.:5.800   3rd Qu.:8.400
 Max.   :6.500   Max.   :8.400   Max.   :8.200   Max.   :9.900
   WartyClaim       OrdBilling        DelSpeed      Satisfaction
 Min.   :4.100   Min.   :2.000   Min.   :1.600   Min.   :4.700
 1st Qu.:5.400   1st Qu.:3.700   1st Qu.:3.400   1st Qu.:6.000
 Median :6.100   Median :4.400   Median :3.900   Median :7.050
 Mean   :6.043   Mean   :4.278   Mean   :3.886   Mean   :6.918
 3rd Qu.:6.600   3rd Qu.:4.800   3rd Qu.:4.425   3rd Qu.:7.625
 Max.   :8.100   Max.   :6.700   Max.   :5.500   Max.   :9.900
```

- **Data Visulaisation**

➢ Histogram of Customer Satisfactiom

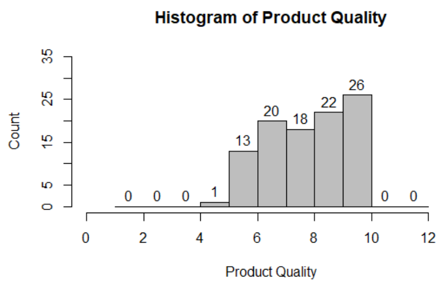


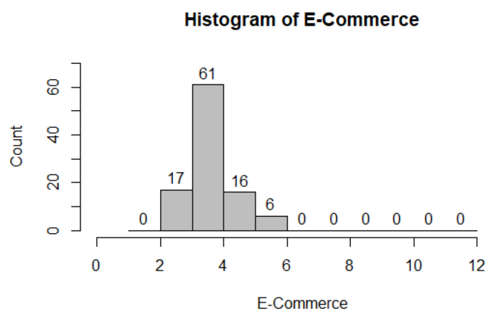
➢ Boxplot of Customer satisfaction

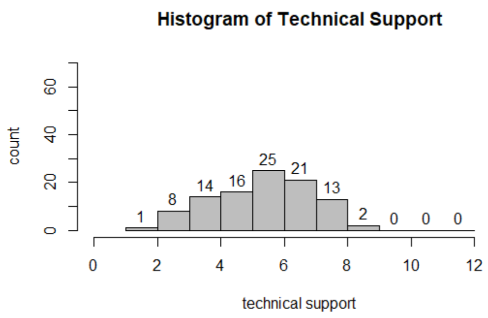
- <u>Key Observation</u>

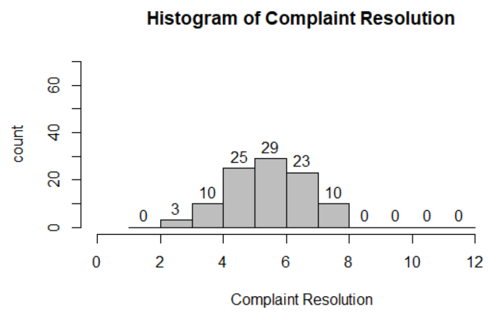
  - According to all the variables the customer satisfaction has been calculated.
  - The observation clearly shows the count of customer satisfaction on point 7 to 8 is highest.
  - Else the mean point of 6, got second highest

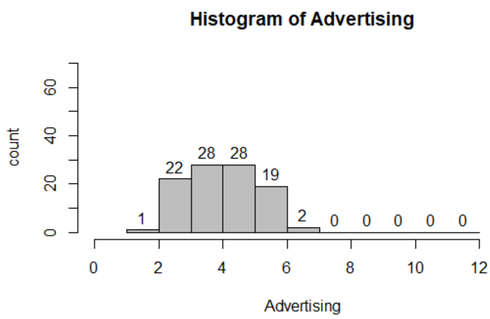
- Histogram of all Variables

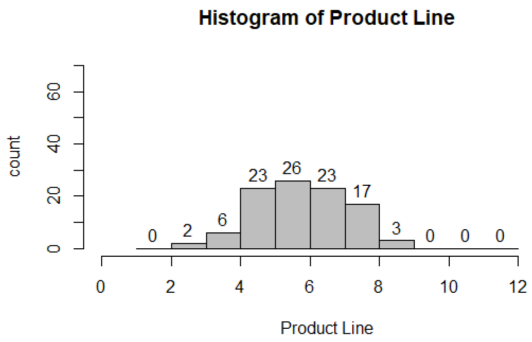


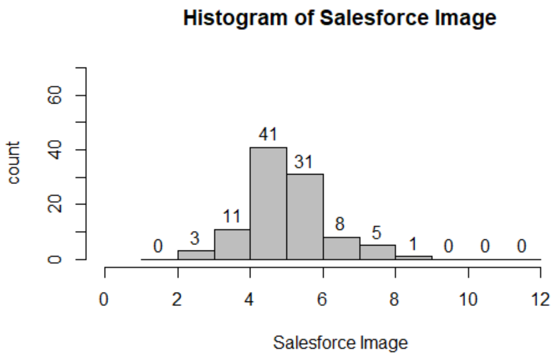


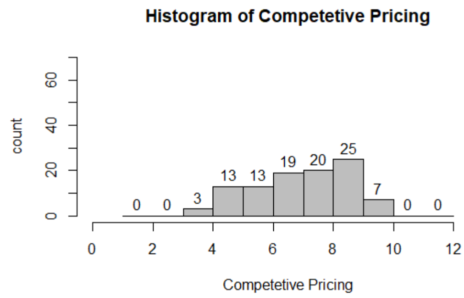


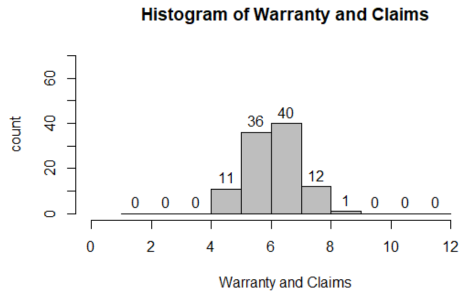


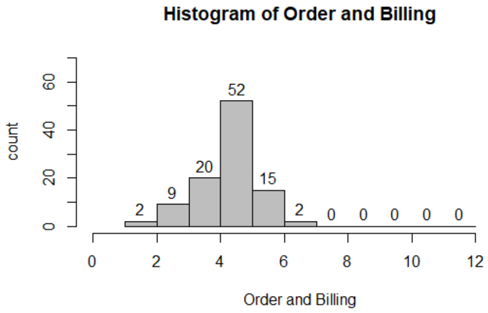


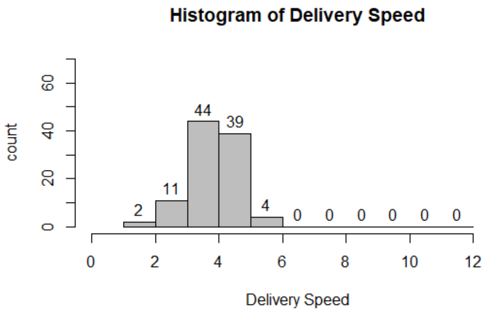




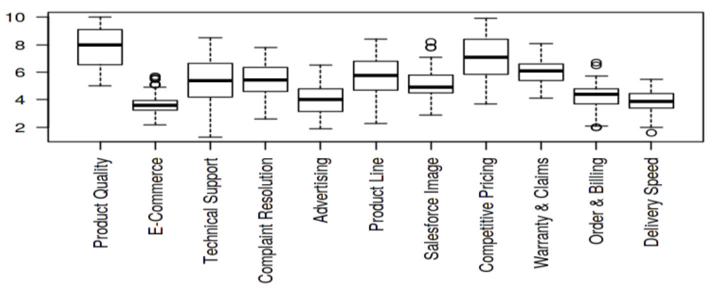






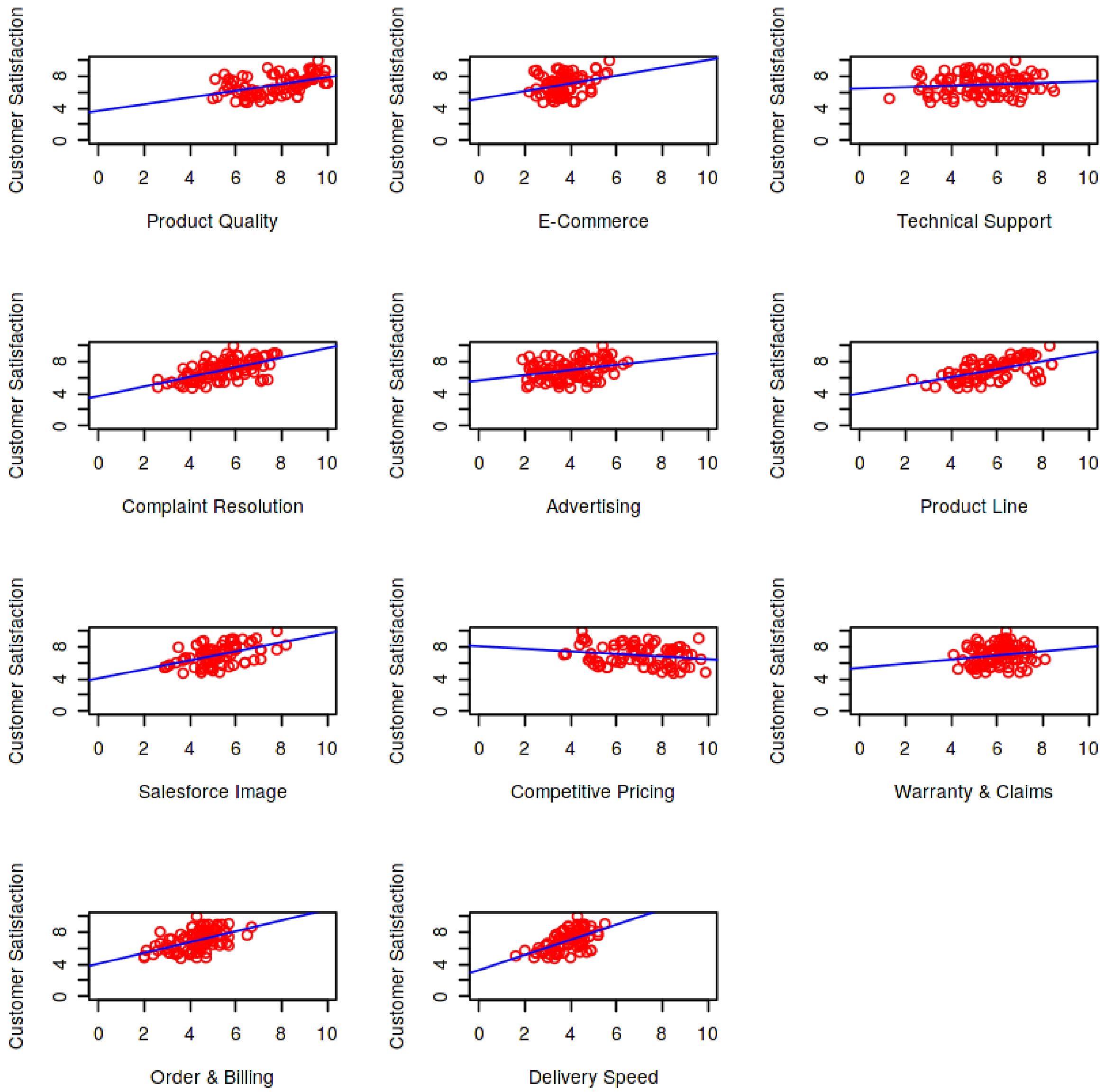


- Boxplot of All Variables

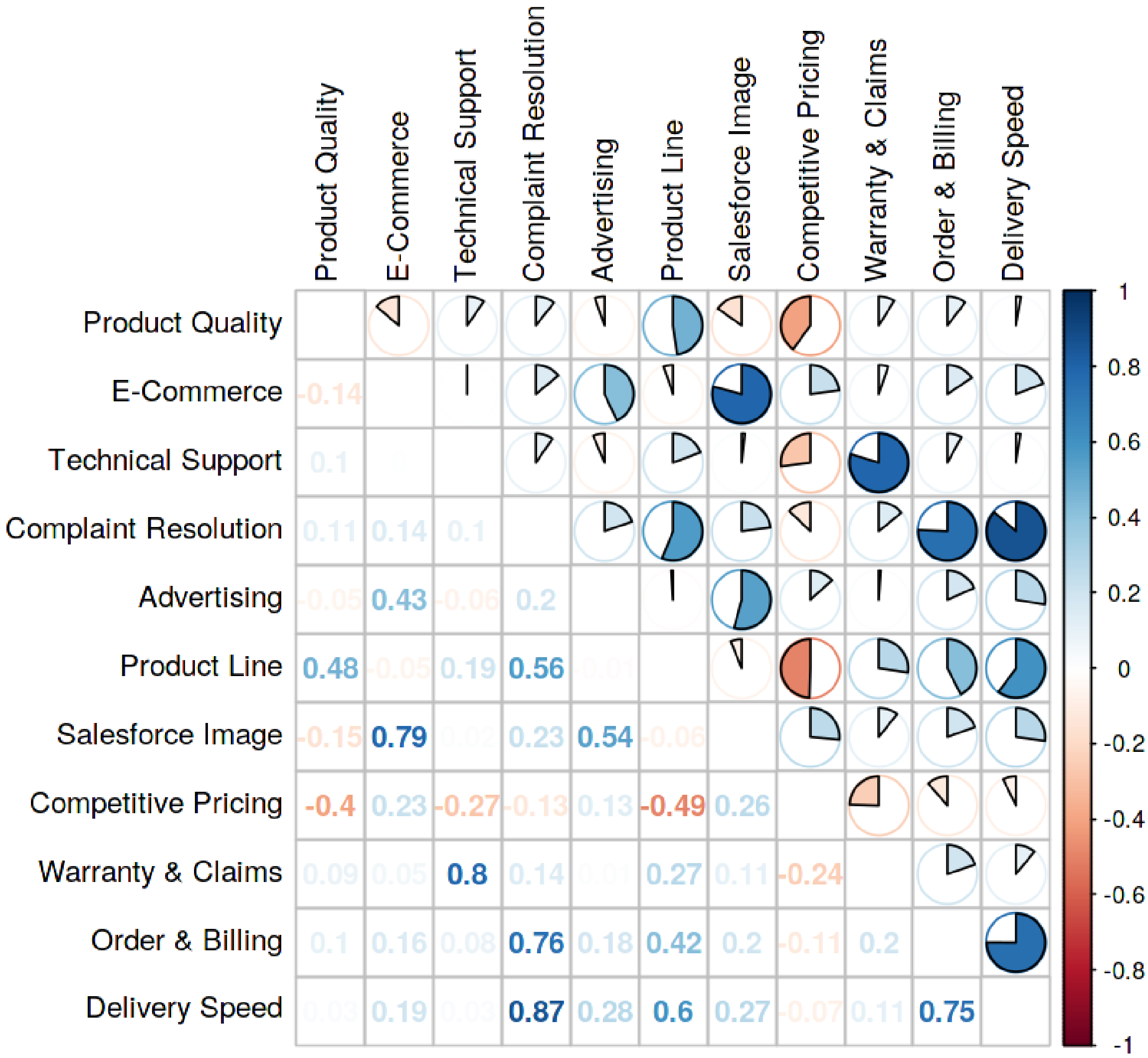


- ## Bivariate Analysis

  − To find the independent variable against the target variable



- **Correlation Mattrix**



- Observation
  - CompRes and DelSpeed are highly correlated
  - OrdBilling and CompRes are highly correlated
  - WartyClaim and TechSupport are highly correlated
  - CompRes and OrdBilling are highly correlated
  - OrdBilling and DelSpeed are highly correlated
  - Ecom and SalesFImage are highly correlated

- ## Initial Regression Analysis

- We have to prepare a model using all Independent Variable

```
#Model2=lm(Satisfaction~.,mydata)
#summary(Model2)
```

```
-------------------------------------------------------------------
Call:
lm(formula = Satisfaction ~ ., data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43005 -0.31165  0.07621  0.37190  0.90120

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.66961    0.81233  -0.824  0.41199
ProdQual     0.37137    0.05177   7.173 2.18e-10 ***
Ecom        -0.44056    0.13396  -3.289  0.00145 **
TechSup      0.03299    0.06372   0.518  0.60591
CompRes      0.16703    0.10173   1.642  0.10416
Advertising -0.02602    0.06161  -0.422  0.67382
ProdLine     0.14034    0.08025   1.749  0.08384 .
SalesFImage  0.80611    0.09775   8.247 1.45e-12 ***
ComPricing  -0.03853    0.04677  -0.824  0.41235
WartyClaim  -0.10298    0.12330  -0.835  0.40587
OrdBilling   0.14635    0.10367   1.412  0.16160
DelSpeed     0.16570    0.19644   0.844  0.40124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5623 on 88 degrees of freedom
Multiple R-squared:  0.8021,	Adjusted R-squared:  0.7774
F-statistic: 32.43 on 11 and 88 DF,  p-value: < 2.2e-16

-------------------------------------------------------------------
```

- ## Checkking MultiCollinearity in Independent variable using VIF

```
#vifmatrix <- vif(lm(Satisfaction ~., data = mydata))
#vifmatrix
```

```
ProdQual        Ecom     TechSup     CompRes Advertising    ProdLine
   1.635797    2.756694    2.976796    4.730448    1.508933
3.488185
SalesFImage  ComPricing  WartyClaim  OrdBilling    DelSpeed
   3.439420    1.635000    3.198337    2.902999    6.516014
```

❖ High Variable Inflation Factor (VIF) is a sign of multicollinearity. There is no formal VIF value for determining presence of multicollinearity; however in weaker models VIF value greater than 2.5 may be a cause of concern.

- ## Factor Analysis

- Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors

- We will go throught he KMO test,(Kaimer- Meyer- Oklin); to measure the suiting of the data for Factor Analysis

```
#KMO(corMtrx)

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = corMtrx)
Overall MSA =  0.65
MSA for each item =
   ProdQual         Ecom      TechSup      CompRes Advertising
ProdLine
       0.51         0.63         0.52         0.79        0.78
0.62
SalesFImage  ComPricing  WartyClaim  OrdBilling    DelSpeed
       0.62         0.75         0.51         0.76        0.67
```

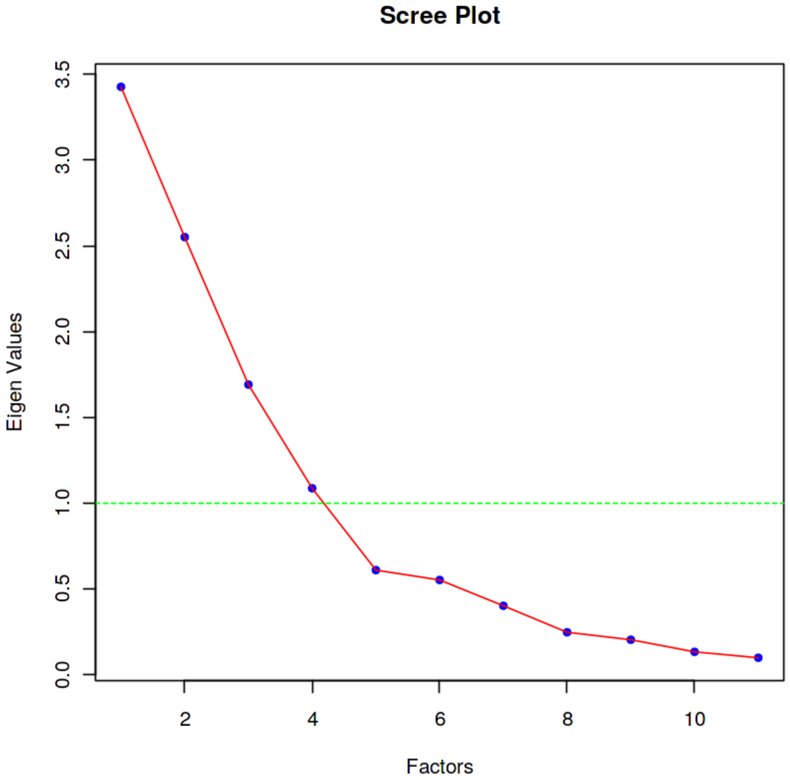
----

*MSA>0.5, since we can run Factor Analysis.

- Calculate the Eigen value

```
#ev <- eigen(corMtrx)
#ev

"3.42697133" "2.55089671" "1.69097648" "1.08655606" "0.60942409"
"0.55188378"
 [7] "0.40151815" "0.24695154" "0.20355327" "0.13284158"
"0.09842702"
```

- **Scree Plot Visulaisation**



<u>Observation</u>

- As perthe above Scree plot, 4 factors out of 11 are above 1
- Therefore, 4 Principle components should be shortlisted

- Factor Analysing Using method ‘pa’

- Without Rotating

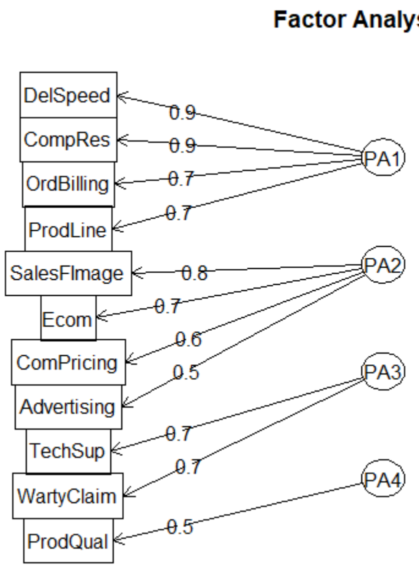
```
Factor Analysis using method =  pa
Call: fa(r = mydata[, -12], nfactors = 4, rotate = "none", fm =
"pa")
Standardized loadings (pattern matrix) based upon correlation matrix
              PA1   PA2   PA3   PA4   h2    u2 com
ProdQual     0.20 -0.41 -0.06  0.46 0.42 0.576 2.4
Ecom         0.29  0.66  0.27  0.22 0.64 0.362 2.0
TechSup      0.28 -0.38  0.74 -0.17 0.79 0.205 1.9
CompRes      0.86  0.01 -0.26 -0.18 0.84 0.157 1.3
Advertising  0.29  0.46  0.08  0.13 0.31 0.686 1.9
ProdLine     0.69 -0.45 -0.14  0.31 0.80 0.200 2.3
SalesFImage  0.39  0.80  0.35  0.25 0.98 0.021 2.1
ComPricing  -0.23  0.55 -0.04 -0.29 0.44 0.557 1.9
WartyClaim   0.38 -0.32  0.74 -0.15 0.81 0.186 2.0
OrdBilling   0.75  0.02 -0.18 -0.18 0.62 0.378 1.2
DelSpeed     0.90  0.10 -0.30 -0.20 0.94 0.058 1.4

                       PA1  PA2  PA3  PA4
SS loadings           3.21 2.22 1.50 0.68
Proportion Var        0.29 0.20 0.14 0.06
Cumulative Var        0.29 0.49 0.63 0.69
Proportion Explained  0.42 0.29 0.20 0.09
Cumulative Proportion 0.42 0.71 0.91 1.00

Mean item complexity =  1.9
Test of the hypothesis that 4 factors are sufficient.

The degrees of freedom for the null model are  55  and the objective
function was  6.55 with Chi Square of  619.27
The degrees of freedom for the model are 17  and the objective
function was  0.33

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  0.03

The harmonic number of observations is  100 with the empirical chi
square  3.19  with prob <  1
The total number of observations was  100  with Likelihood Chi
Square =  30.27  with prob <  0.024

Tucker Lewis Index of factoring reliability =  0.921
RMSEA index =  0.096  and the 90 % confidence intervals are  0.032
0.139
BIC =  -48.01
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                 PA1  PA2  PA3
PA4
Correlation of (regression) scores with factors   0.98 0.97 0.95
0.88
Multiple R square of scores with factors          0.96 0.95 0.91
0.78
Minimum correlation of possible factor scores     0.92 0.90 0.82
0.56
```

– With Rotating 'Varimax'

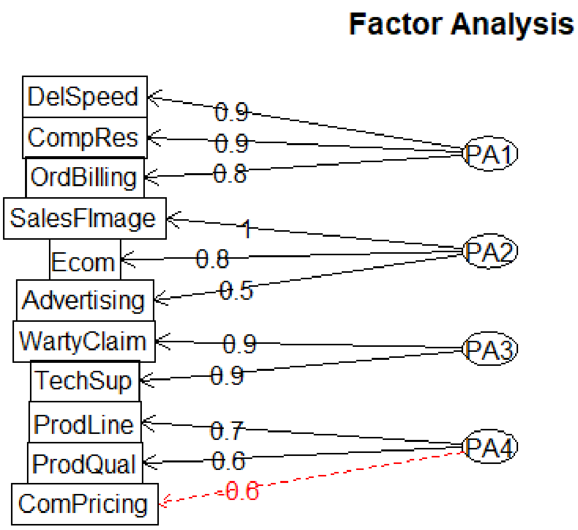
```
Factor Analysis using method =  pa
Call: fa(r = mydata[, -12], nfactors = 4, rotate = "varimax", 
    fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
              PA1   PA2   PA3   PA4   h2    u2 com
ProdQual     0.02 -0.07  0.02  0.65 0.42 0.576 1.0
Ecom         0.07  0.79  0.03 -0.11 0.64 0.362 1.1
TechSup      0.02 -0.03  0.88  0.12 0.79 0.205 1.0
CompRes      0.90  0.13  0.05  0.13 0.84 0.157 1.1
Advertising  0.17  0.53 -0.04 -0.06 0.31 0.686 1.2
ProdLine     0.53 -0.04  0.13  0.71 0.80 0.200 1.9
SalesFImage  0.12  0.97  0.06 -0.13 0.98 0.021 1.1
ComPricing  -0.08  0.21 -0.21 -0.59 0.44 0.557 1.6
WartyClaim   0.10  0.06  0.89  0.13 0.81 0.186 1.1
OrdBilling   0.77  0.13  0.09  0.09 0.62 0.378 1.1
DelSpeed     0.95  0.19  0.00  0.09 0.94 0.058 1.1

                       PA1  PA2  PA3  PA4
SS loadings           2.63 1.97 1.64 1.37
Proportion Var        0.24 0.18 0.15 0.12
Cumulative Var        0.24 0.42 0.57 0.69
Proportion Explained  0.35 0.26 0.22 0.18
Cumulative Proportion 0.35 0.60 0.82 1.00

Mean item complexity =  1.2
Test of the hypothesis that 4 factors are sufficient.

The degrees of freedom for the null model are  55  and the objective 
function was  6.55 with Chi Square of  619.27
The degrees of freedom for the model are 17  and the objective 
function was  0.33 

The root mean square of the residuals (RMSR) is  0.02 
The df corrected root mean square of the residuals is  0.03 

The harmonic number of observations is  100 with the empirical chi 
square  3.19  with prob <  1 
The total number of observations was  100  with Likelihood Chi 
Square =  30.27  with prob <  0.024 

Tucker Lewis Index of factoring reliability =  0.921
RMSEA index =  0.096  and the 90 % confidence intervals are  0.032 
0.139
BIC =  -48.01
Fit based upon off diagonal values = 1
Measures of factor score adequacy             
                                                  PA1  PA2  PA3 
PA4
Correlation of (regression) scores with factors   0.98 0.99 0.94 
0.88
Multiple R square of scores with factors          0.96 0.97 0.88 
0.78
Minimum correlation of possible factor scores     0.93 0.94 0.77 
0.55
```

- **Visualisation of the factors**

- Without Rotation



- With Roation 'Varimax'



- ## Labelling the Factor

Factors	Variables	Labels
PA 1	**DelSpeed, CompRes, OrdBilling**	**Purchase**
PA 2	**SalesFImage, Ecom, Advertising**	**Marketing**
PA 3	**WartyClaim, TechSup**	**Post Purchase**
PA 4	**ProdLine, ProdQual, CompPricing**	**Product Positioning**

- ## Regression analysis using the factors scores as the independent variable.

- Create new data structure using score of four factors and target variable.
- #hair2 <- cbind(mydata[,12], fa2$scores)
  #head(hair2)

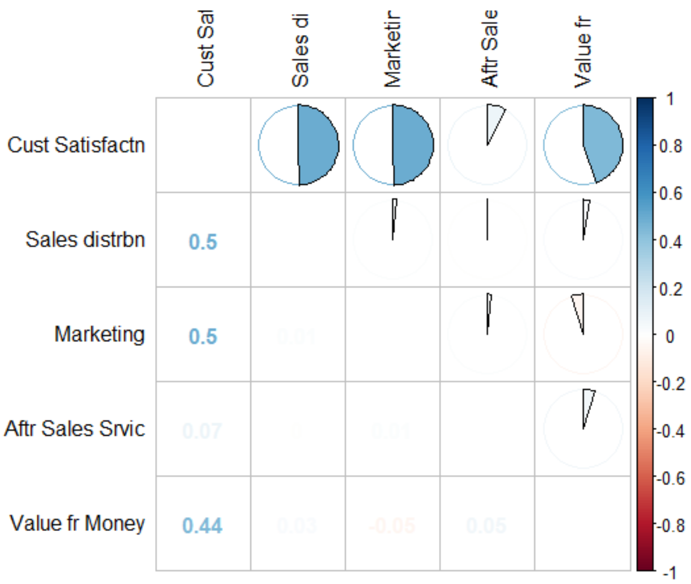
```
          PA1        PA2          PA3         PA4
8.2 -0.1338871  0.9175166 -1.719604873  0.09135411
5.7  1.6297604 -2.0090053 -0.596361722  0.65808192
8.9  0.3637658  0.8361736  0.002979966  1.37548765
4.8 -1.2225230 -0.5491336  1.245473305 -0.64421384
7.1 -0.4854209 -0.4276223 -0.026980304  0.47360747
4.7 -0.5950924 -1.3035333 -1.183019401 -0.95913571
```

- Naming the column for hair2
- # colnames(hair2) <- c("Cust.Satisfaction", "Sales distrbn", "Marketing", "Aftr Sales Srvic", "Value fr Money")
  #head(hair2)

Cust.Satisfaction	Sales distrbn	Marketing	Aftr Sales Srvic	Value fr Money
8.2	-0.1338871	0.9175166	-1.719604873	0.09135411
5.7	1.6297604	-2.0090053	-0.596361722	0.65808192
8.9	0.3637658	0.8361736	0.002979966	1.37548765
4.8	-1.2225230	-0.5491336	1.245473305	-0.64421384
7.1	-0.4854209	-0.4276223	-0.026980304	0.47360747
4.7	-0.5950924	-1.3035333	-1.183019401	-0.95913571

- Correlation Plot of data hair2

```
#corrplot.mixed(cor(hair2),
                lower = "number", upper = "pie",
                tl.col = "black",tl.pos = "lt")
```



- **Creating 2 datasets, one to train and another to test the model**

#set.seed(100)
#indices= sample(1:nrow(regdata), 0.7*nrow(regdata))
#train=regdata[indices,]
#test = regdata[-indices,]

- **Train the regression Model**

#linearModel = lm(Cust.Satisfaction~., data = train)
#summary(linearModel)

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

Residuals:     Min       1Q  Median       3Q      Max  -1.6857 -0.4018
0.1051  0.4027  1.2036

Coefficients:                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.92625    0.08263  83.827  < 2e-16 *** Purchanse
0.62022    0.08408   7.377 3.73e-10 *** Marketing         0.57735
0.08047   7.175 8.50e-10 *** Post_purchase     0.09567    0.08667
1.104    0.274      Prod_positioning  0.66562    0.09374   7.101
1.15e-09 *** --- Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6814 on 65 degrees of freedom Multiple R-
squared:  0.7079, Adjusted R-squared:   0.69  F-statistic: 39.39 on
4 and 65 DF,  p-value: < 2.2e-16
```

- **Checking of VIF Score**

#vif(linearModel)

“Sales distrbn”	“Marketing”	“Aftr Sales Srvic”	“Value fr Money”
1.005383	1.004905	1.009641	1.018316

- **Compute R-Sq for the test data**

 # SST= sum((test$Cust.Satisfaction - mean(train$Cust.Satisfaction))^2)
#SSE= sum((pred - test$Cust.Satisfaction)^2)
#SSR= sum((pred - mean(train$Cust.Satisfaction))^2)
#R.square.test <- SSR/SST
#cat("SST:", SST,"\n",
 #   "SSE:", SSE,"\n",
 # "SSR:", SSE,"\n",
 #   "R squared test:", R.square.test)

```
SST: 28.83703
 SSE: 6.858458
 SSR: 6.858458
 R squared test: 0.7109649
```

- **Hence, The R squared is not varying much for the model both in test and train dataset; so we can infer that the model is valid and also not overfit.**

- ## Conclusion

- We have came to conclusion that, the following data has the evidence of Multicollinearity in independent variables using VIF. Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other. In other words, it results when you have factors that are a bit redundant.

- In this data, We also concluded the factor Analysis. By characterising the variables verbally. And have gone through few tests and calculations, like KMO(Kaiser-Meyer-Olkin) test, and calculating the Eigen value.

- With the help of Eigen Value and Scree-plot Graph, we concluded 4 principle components for the factor analyisis. We have gone through the ‘pa’ method, with and without rotation and analyse the result of factor variables.

- 4 variables have been divided in 4 categories – Purchase, Marketing, Post – Purchase, Product Positioning.

- While performing the Multiple Linear Regression with customer satisfaction as the independent and four factor as independent variable; we found that the R squared is not varying much for the model both in test and train dataset; so we can infer that the model is valid and also not overfit.

Thank You

Source Code

```
#--------------------------------------------------------------------------------
#
# Explore Data Analysis – Factor hair
#
#--------------------------------------------------------------------------------
#
# Install Necessary Packages in this section, including libraries.
# Having all packages and libraries at one place makes the code readable. –
#
# install.packages("corrplot")
# install.packages("tidyverse")
# install.packages("ggplot2")
# install.packages("psych")
# install.packages("car")
#--------------------------------------------------------------------------------
#
# Setup Working Directory
#
 setwd("F:/hair r data")
#
# getwd()

  >"F:/hair r data"

#  Read Input File
#mydata=read.csv("Factor-Hair-Revised.csv", header = T)

# Variable names
 variables <- c("Product Quality" , "E-Commerce" , "Technical Support" ,
"Complaint Resolution" ,
```

```
+             "Advertising" , "Product Line" , "Salesforce Image", "Competitive
Pricing" ,
+             "Warranty & Claims" , "Order & Billing" , "Delivery Speed" ,
"Customer Satisfaction")
```

#

#

 # Find out Total Number of Rows and Columns

```
dim(mydata)
< 100  13
```

 # Find out the names of the columns

```
# "ID"    "ProdQual"    "Ecom"       "TechSup"
 "CompRes"     "Advertising"  "ProdLine"     "SalesFImage"
 "ComPricing"   "WartyClaim"   "OrdBilling"   "DelSpeed"
 "Satisfaction"
```

#

 # Summary of data

#

#

```
Summary(mydata)
```

```
        ID            ProdQual          Ecom          TechSup         CompRes
 Min.   :  1.00.    Min.   : 5.000.   Min.   :2.200   Min.   :1.300   Min.   :2.600
 1st Qu.: 25.75    1st Qu.: 6.575   1st Qu.:3.275   1st Qu.:4.250   1st Qu.:4.600
 Median : 50.50    Median : 8.000   Median :3.600   Median :5.400   Median
:5.450
 Mean   : 50.50   Mean   : 7.810   Mean   :3.672   Mean   :5.365   Mean   :5.442
 3rd Qu.: 75.25   3rd Qu.: 9.100   3rd Qu.:3.925   3rd Qu.:6.625   3rd Qu.:6.325
 Max.   :100.00   Max.   :10.000   Max.   :5.700   Max.   :8.500   Max.   :7.800

  Advertising      ProdLine      SalesFImage      ComPricing      WartyClaim
 Min.   :1.900   Min.   :2.300   Min.   :2.900   Min.   :3.700   Min.   :4.100
 1st Qu.:3.175   1st Qu.:4.700   1st Qu.:4.500   1st Qu.:5.875   1st Qu.:5.400
 Median :4.000   Median :5.750   Median :4.900   Median :7.100   Median
:6.100
 Mean   :4.010   Mean   :5.805   Mean   :5.123   Mean   :6.974   Mean   :6.043
 3rd Qu.:4.800   3rd Qu.:6.800   3rd Qu.:5.800   3rd Qu.:8.400   3rd Qu.:6.600
 Max.   :6.500   Max.   :8.400   Max.   :8.200   Max.   :9.900   Max.   :8.100

   OrdBilling      DelSpeed      Satisfaction
```

```
 Min.   :2.000   Min.   :1.600   Min.   :4.700  
 1st Qu.:3.700   1st Qu.:3.400   1st Qu.:6.000  
 Median :4.400   Median :3.900   Median :7.050  
 Mean   :4.278   Mean   :3.886   Mean   :6.918  
 3rd Qu.:4.800   3rd Qu.:4.425   3rd Qu.:7.625  
 Max.   :6.700   Max.   :5.500   Max.   :9.900  
```

#
#
#

 # From summary and data we learned that the given data is scaled already And no need to scale it again.
 # We also leanred that first column "ID" is just a column number.
 # So we don't need this column in further process, so we will remove it
#
#

```
mydata=mydata[,2:13]
```

#
#

```
names(mydata)
 "ProdQual"      "Ecom"          "TechSup"       "CompRes"
 "Advertising"   "ProdLine"      "SalesFImage"   "ComPricing"
 "WartyClaim"    "OrdBilling"    "DelSpeed"      "Satisfaction"
```

#
 # Summary of new data
#
#

#---
#
#Histogram of the Customer satisfaction
#
#

```
hist(mydata$Satisfaction, breaks = c(2:13), col = 8, border = 1, main = paste("Histogram of Customer Satisfaction"), xlab = "Customer Satisfaction", ylab = "Count", xlim = c(2,13), ylim = c(0,35), labels = T, include.lowest=T, right=T )
```

#
#

#Boxplot of the Customer satisfaction
#
#
boxplot(mydata$Satisfaction, horizontal = T, main ='Customer Satisfaction')
#
#
Histogram of all variables
#
#
hist(mydata$ProdQual, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of Product Quality"), xlab = "Product Quality", ylab = "Count", xlim = c(0,13), ylim = c(0,35),labels = T, include.lowest=T, right=T)

hist(mydata$Ecom, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of E-Commerce"), xlab = "E-Commerce", ylab = "Count", xlim = c(0,13), ylim = c(0,70),labels = T, include.lowest=T, right=T)

hist(mydata$TechSup, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of Technical Support"), xlab = "technical support", ylab = "count", xlim = c(0,13), ylim = c(0,70),labels = T, include.lowest=T, right=T)

hist(mydata$CompRes, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of Complaint Resolution"), xlab = "Complaint Resolution", ylab = "count", xlim = c(0,13), ylim = c(0,70),labels = T, include.lowest=T, right=T)

hist(mydata$Advertising, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of Advertising"), xlab = "Advertising", ylab = "count", xlim = c(0,13), ylim = c(0,70),labels = T, include.lowest=T, right=T)

hist(mydata$ProdLine, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of Product Line"), xlab = "Product Line", ylab = "count", xlim = c(0,13), ylim = c(0,70),labels = T, include.lowest=T, right=T)

hist(mydata$SalesFImage, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of Salesforce Image"), xlab = "Salesforce Image", ylab = "count", xlim = c(0,13), ylim = c(0,70),labels = T, include.lowest=T, right=T)

hist(mydata$ComPricing, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of Competetive Pricing"), xlab = "Competetive Pricing", ylab = "count", xlim = c(0,13), ylim = c(0,70),labels = T, include.lowest=T, right=T)
hist(mydata$WartyClaim, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of Warranty and Claims"), xlab = "Warranty and Claims", ylab = "count", xlim = c(0,13), ylim = c(0,70),labels = T, include.lowest=T, right=T)

hist(mydata$OrdBilling, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of Order and Billing"), xlab = "Order and Billing", ylab = "count", xlim = c(0,13), ylim = c(0,70),labels = T, include.lowest=T, right=T)

hist(mydata$DelSpeed, breaks = c(1:12), col = 8, border = 1, main = ("Histogram of Delivery Speed"), xlab = "Delivery Speed", ylab = "count", xlim = c(0,13), ylim = c(0,70),labels = T, include.lowest=T, right=T)

#
#
#Explore data (mydata) using boxplot method
#
boxplot(mydata[,-12], las = 2, names = variables[-12], cex.axis = 1)
#

```
#for (i in c(1:11))
  {
  plot(mydata[,i],'Customer Satisfaction',
       xlab = variables[i], ylab = NULL, col = "red",
       cex.lab=1, cex.axis=1, cex.main=1, cex.sub=1,
       xlim = c(0,10),ylim = c(0,10))
  abline(lm(formula = 'Customer Satisfaction' ~ mydata[,i]),col =
"blue")
}
#corMtrx <- cor(mydata[,-12])
#corMtrx
#cor(mydata[1:12])
#corMtrx
#library(corrplot)
#corrplot.mixed(corMtrx, lower = "number", upper = "pie", tl.col =
"black", tl.pos = "lt")
#library(car)
#Model1=lm(Satisfaction~ProdQual)
#summary(Model1)
#Model2=lm(Satisfaction~.,mydata)
#summary(Model2)
#vifmatrix <- vif(lm(Satisfaction ~., data = mydata))
#vifmatrix
#summary(vifmatrix)
#KMO(corMtrx)
#library(tidyverse)
#library(psych)
#install.packages("caTools")
#library(caTools)
#cortest.bartlett(corMtrx, 100)
#KMO(corMtrx)
#ev <- eigen(corMtrx)
#ev
#library(ggplot2)
#ev1 <- ev$values
#ev1
#plot(ev1, main = "Scree Plot", xlab = "factors", ylab =
"eigenvalue", pch = 20, col ="blue")
#lines(ev1, col = "red")
#abline(h = 1, col = "green", lty = 2)
#fa1= fa(r= mydata[,-12], nfactors = 4, rotate = "none", fm="pa")
#print(fa1)
#load1 <- print(fa1$loadings,cutoff = 0.3)
#write.csv(load1, "load.csv")
#fa.diagram(fa1)
#fa2= fa(r= mydata[,-12], nfactors = 4, rotate = "varimax", fm="pa")
#print(fa2)
#load2 <- print(fa2$loadings,cutoff = 0.3)
```

```
#write.csv(load2, "load2.csv")
#fa.diagram(fa2)
#air2 <- cbind(mydata[,12], fa2$scores)
#head(hair2)
#colnames(hair2) <- c("Cust.Satisfaction", "Sales distrbn",
"Marketing", "Aftr Sales Srvic", "Value fr Money")
#head(hair2)
#class(hair2)
#hair2 <- as.data.frame(hair2)
#corrplot.mixed(cor(hair2),
               lower = "number", upper = "pie",
               tl.col = "black",tl.pos = "lt")
#library(caTools)
#set.seed(1)
#spl= sample.split(hair2$Cust.Satisfaction, SplitRatio =0.8)
#train= subset(hair2, spl==T)
#test= subset(hair2, spl==F)
#cat("Train Dimension:", dim(train), "\n","Test Dimension",
#dim(test))
#linearModel = lm(Cust.Satisfaction~., data = train)
#summary(linearModel)
#vif(linearModel)
#pred= predict(linearModel, newdata = test)
#SST= sum((test$Cust.Satisfaction -
mean(train$Cust.Satisfaction))^2)
#SSE= sum((pred - test$Cust.Satisfaction)^2)
#SSR= sum((pred - mean(train$Cust.Satisfaction))^2)
#R.square.test <- SSR/SST
#cat("SST:", SST,"\n",
    "SSE:", SSE,"\n",
    "SSR:", SSE,"\n",
    "R squared test:", R.square.test)
```

Thank You

The END
