Insight to Factor Analysis and PCA

Project -2

Advanced Statistics

Baijayanti Chakraborty

1 Contents

1. Project Objective

The objective of the project is to use the dataset “Factor-Hair-Revised.csv” to build an optimum regression model to predict satisfaction. We are expected to

1. Perform exploratory data analysis on the dataset. Showcase some charts, graphs. Check for outliers and missing values
2. Evidence and analysis of multicollinearity?
3. Perform simple linear regression for the dependent variable with every independent variable
4. Perform PCA/Factor analysis by extracting 4 factors. Interpret the output and name the Factors
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.

2. Exploratory Data Analysis

2.1. Environment Set up

```
setwd("~/Desktop/PGP-BABI/Project 2")

install.packages("DataExplorer")

install.packages("corrplot")

install.packages("ppcor")

install.packages("nFactors")

install.packages("psych")

install.packages("caTools")

install.packages("Metrics")


library(psych)

library(Hmisc)

library(DataExplorer)

library(corrplot)

library(ppcor)

library(ggplot2)

library(nFactors)

library(caTools)

library(car)

library(Metrics)
```

```
mydata= read.csv("Factor-Hair-Revised.csv" , header = TRUE)
mydata = mydata[,-1]
```

2.2. Data Exploration

```
#clear the global environment
rm(list = ls())

#read the data
mydata = read.csv("Factor-hair-Revised.csv",header = TRUE)
mydata = mydata[,-1]

#basic analysis of the data
View(mydata)
nrow(mydata) # 100 rows

## [1] 100

ncol(mydata) # 13 columns

## [1] 12

colnames(mydata) #[1]  "ProdQual"     "Ecom"         "TechSup"      "CompRes
"       "Advertising"

##  [1] "ProdQual"     "Ecom"         "TechSup"      "CompRes"
##  [5] "Advertising"  "ProdLine"     "SalesFImage"  "ComPricing"
##  [9] "WartyClaim"   "OrdBilling"   "DelSpeed"     "Satisfaction"

                 # [7] "ProdLine"     "SalesFImage"  "ComPricing"    "WartyCl
aim"   "OrdBilling"   "DelSpeed"  "Satisfaction"
summary(mydata) # basic stats of the columns

##     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
```

```r
str(mydata) #different data types of the columns
```

```
## '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 ...
```

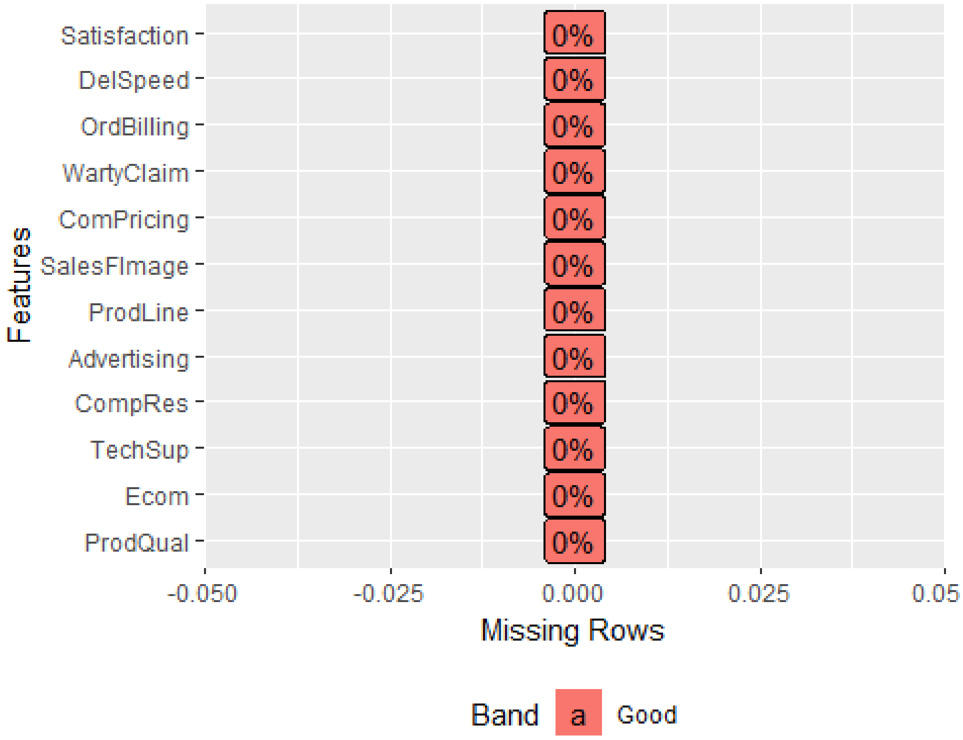
```r
describe(mydata)
```

```
##                vars   n mean   sd median trimmed  mad min  max range  skew
## ProdQual          1 100 7.81 1.40   8.00    7.85 1.78 5.0 10.0   5.0 -0.24
## Ecom              2 100 3.67 0.70   3.60    3.63 0.52 2.2  5.7   3.5  0.64
## TechSup           3 100 5.37 1.53   5.40    5.40 1.85 1.3  8.5   7.2 -0.20
## CompRes           4 100 5.44 1.21   5.45    5.46 1.26 2.6  7.8   5.2 -0.13
## Advertising       5 100 4.01 1.13   4.00    4.00 1.19 1.9  6.5   4.6  0.04
## ProdLine          6 100 5.80 1.32   5.75    5.81 1.56 2.3  8.4   6.1 -0.09
## SalesFImage       7 100 5.12 1.07   4.90    5.09 0.89 2.9  8.2   5.3  0.37
## ComPricing        8 100 6.97 1.55   7.10    7.01 1.93 3.7  9.9   6.2 -0.23
## WartyClaim        9 100 6.04 0.82   6.10    6.04 0.89 4.1  8.1   4.0  0.01
## OrdBilling       10 100 4.28 0.93   4.40    4.31 0.74 2.0  6.7   4.7 -0.32
## DelSpeed         11 100 3.89 0.73   3.90    3.92 0.74 1.6  5.5   3.9 -0.45
## Satisfaction     12 100 6.92 1.19   7.05    6.90 1.33 4.7  9.9   5.2  0.08
##                kurtosis   se
## ProdQual          -1.17 0.14
## Ecom               0.57 0.07
## TechSup           -0.63 0.15
## CompRes           -0.66 0.12
## Advertising       -0.94 0.11
## ProdLine          -0.60 0.13
## SalesFImage        0.26 0.11
## ComPricing        -0.96 0.15
## WartyClaim        -0.53 0.08
## OrdBilling         0.11 0.09
## DelSpeed           0.09 0.07
## Satisfaction      -0.86 0.12
```

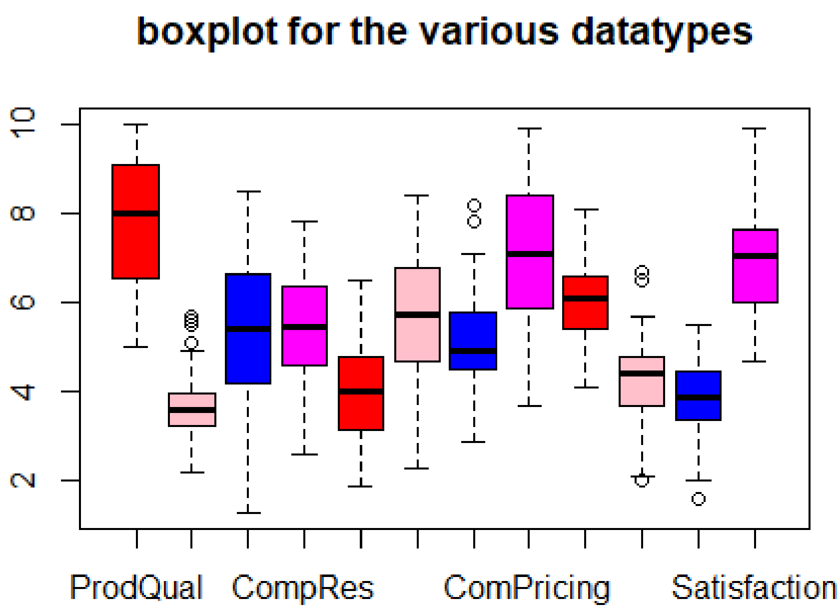
```r
#to check if any null values are present
is.na(mydata) #the data has no null values.hence the data is a clean one.
```

```
##        ProdQual  Ecom TechSup CompRes Advertising ProdLine SalesFImage
##   [1,]    FALSE FALSE   FALSE   FALSE       FALSE    FALSE       FALSE
##   [2,]    FALSE FALSE   FALSE   FALSE       FALSE    FALSE       FALSE
##   [3,]    FALSE FALSE   FALSE   FALSE       FALSE    FALSE       FALSE
##   [4,]    FALSE FALSE   FALSE   FALSE       FALSE    FALSE       FALSE
##   [5,]    FALSE FALSE   FALSE   FALSE       FALSE    FALSE       FALSE
##   [6,]    FALSE FALSE   FALSE   FALSE       FALSE    FALSE       FALSE
```

```
plot_missing(mydata)
```



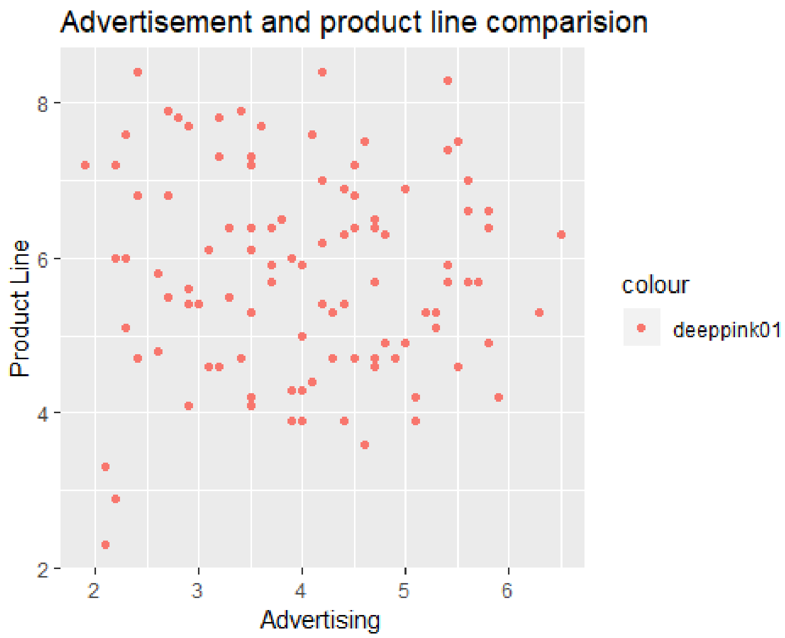

```
#to check the outliers
box = boxplot(mydata , color = "blue" , main = "boxplot for the various data types" , col = c("red","pink","blue","magenta"))
```



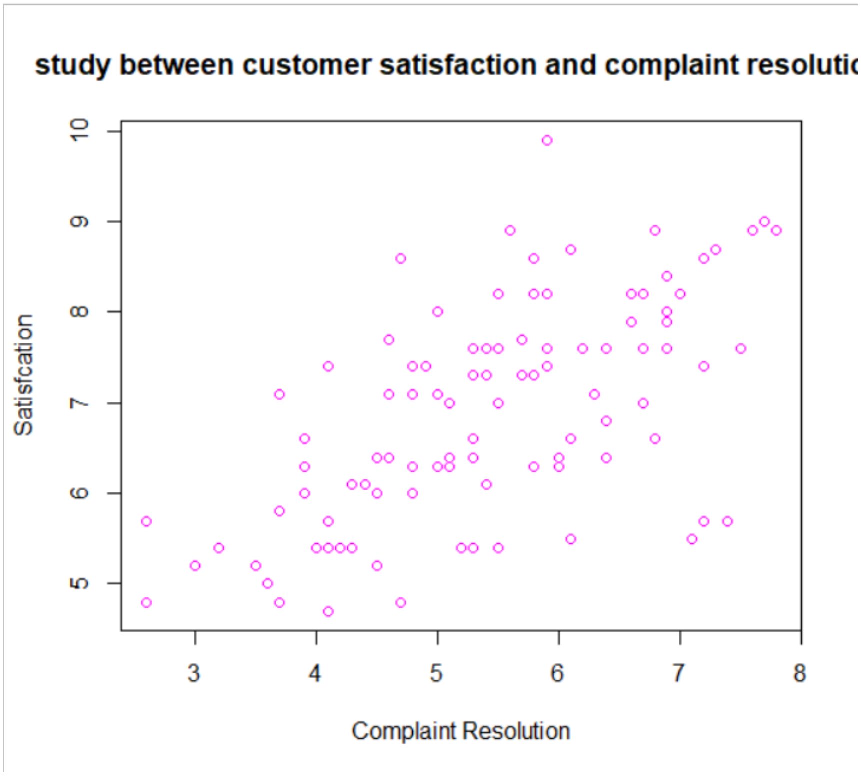

```
outlier = box$out
```

```
#the customer satisfaction rate
hist1 = hist(mydata$Satisfaction,col = "red" , main = "Customer Satisfaction
" , breaks = 15 ,xlab = "Satisfaction" ,ylab = "Frequency")
```

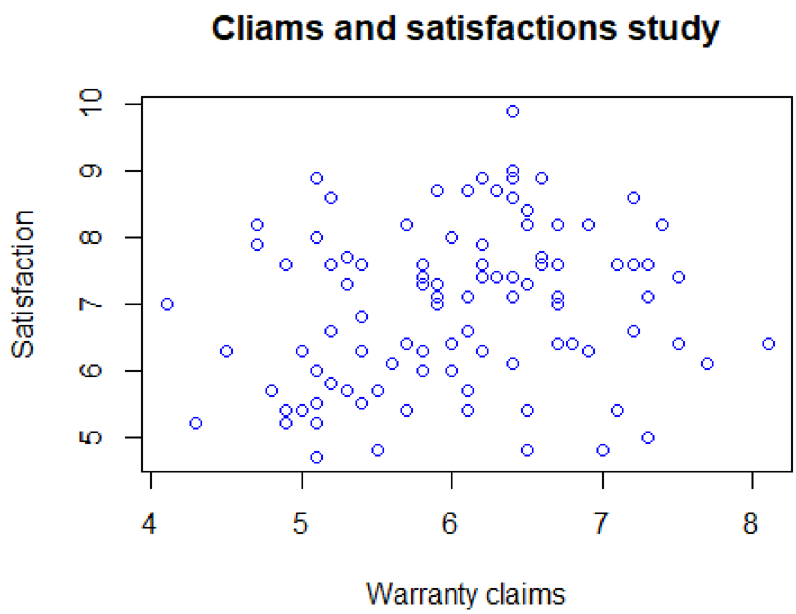

```
#a comparitive study on advertising and pricing.
plot = qplot(mydata$Advertising,mydata$ProdLine,xlab = "Advertising" , ylab
= "Product Line",main = "Advertisement and product line comparision" , margi
ns = TRUE , col = "deeppink01")
#The graph hence shows that as advertisement increases the product line also
increases.
print(plot)
```




```
#a comparitive study between customer satisfaction and complaint resolution
plot2 = plot(mydata$CompRes,mydata$Satisfaction , xlab = "Complaint Resoluti
on" , ylab = "Satisfcation" , main = "study between customer satisfaction an
d complaint resolution" ,col = "magenta")
```



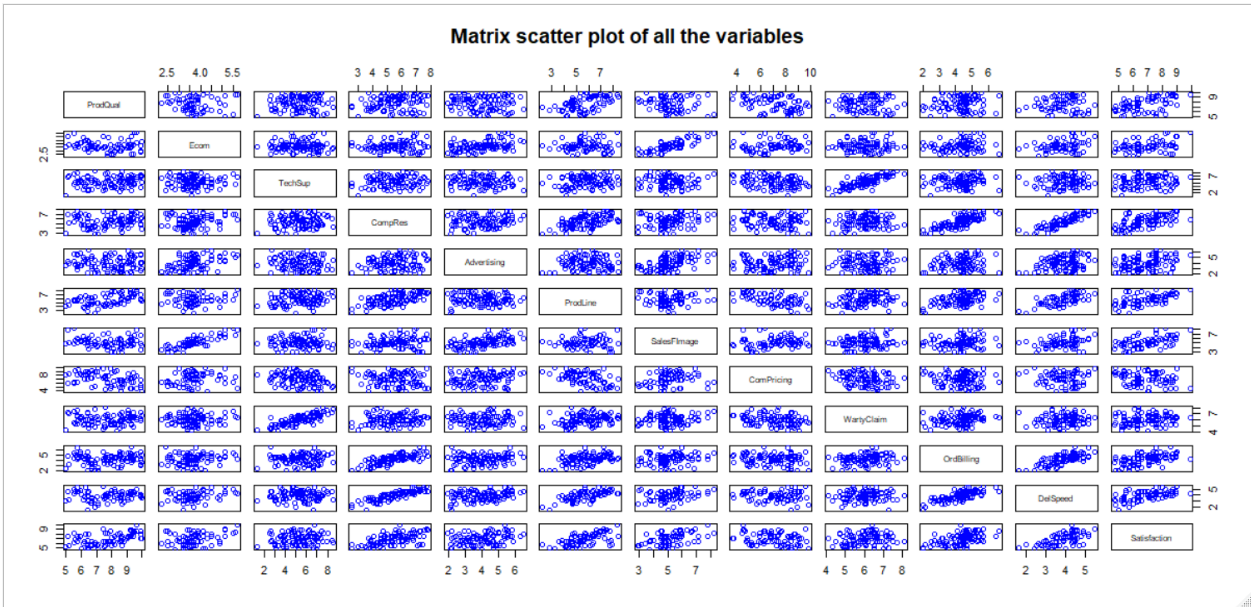

```
#The plot shows that complains resolved are satisfactory for the customers.

#a study of warranty claims and customer satisfcation
plot3 = plot(mydata$WartyClaim,mydata$Satisfaction , col = "blue",xlab = "Wa
rranty claims",ylab = "Satisfaction" , main = "Cliams and satisfactions stud
y")
```




```
#The plot shows that the relation is quite scattered.Though much of the scat
ter is between the center area.
```

3. Multi-collinearity Evidence

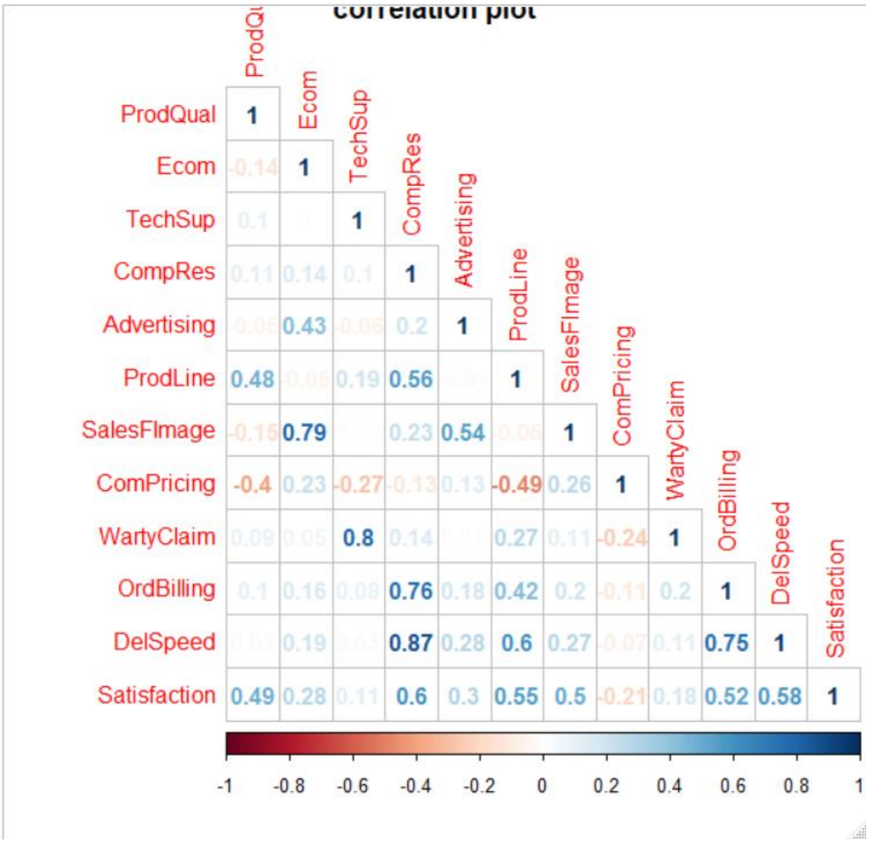
```
#correlation of the variables
plot(mydata, main = "Matrix scatter plot of all the variables", col = "blue"
)
```




```
corr = cor(mydata , method = "pearson")
corr

##                ProdQual          Ecom       TechSup    CompRes
## ProdQual      1.00000000 -0.1371632174  0.0956004542  0.1063700
## Ecom         -0.13716322  1.0000000000  0.0008667887  0.1401793
## TechSup       0.09560045  0.0008667887  1.0000000000  0.0966566
## CompRes       0.10637000  0.1401792611  0.0966565978  1.0000000
## Advertising  -0.05347313  0.4298907110 -0.0628700668  0.1969168
## ProdLine      0.47749341 -0.0526878383  0.1926254565  0.5614170
## SalesFImage  -0.15181287  0.7915437115  0.0169905395  0.2297518
## ComPricing   -0.40128188  0.2294624014 -0.2707866821 -0.1279543
## WartyClaim    0.08831231  0.0518981915  0.7971679258  0.1404083
## OrdBilling    0.10430307  0.1561473316  0.0801018246  0.7568686
## DelSpeed      0.02771800  0.1916360683  0.0254406935  0.8650917
## Satisfaction  0.48632500  0.2827450147  0.1125971788  0.6032626
##              Advertising    ProdLine SalesFImage  ComPricing  WartyClaim
## ProdQual     -0.05347313  0.47749341 -0.15181287 -0.40128188  0.08831231
## Ecom          0.42989071 -0.05268784  0.79154371  0.22946240  0.05189819
## TechSup      -0.06287007  0.19262546  0.01699054 -0.27078668  0.79716793
## CompRes       0.19691685  0.56141695  0.22975176 -0.12795425  0.14040830
## Advertising   1.00000000 -0.01155082  0.54220366  0.13421689  0.01079207
## ProdLine     -0.01155082  1.00000000 -0.06131553 -0.49494840  0.27307753
## SalesFImage   0.54220366 -0.06131553  1.00000000  0.26459655  0.10745534
## ComPricing    0.13421689 -0.49494840  0.26459655  1.00000000 -0.24498605
## WartyClaim    0.01079207  0.27307753  0.10745534 -0.24498605  1.00000000
## OrdBilling    0.18423559  0.42440825  0.19512741 -0.11456703  0.19706512
## DelSpeed      0.27586308  0.60185021  0.27155126 -0.07287173  0.10939460
## Satisfaction  0.30466947  0.55054594  0.50020531 -0.20829569  0.17754482
##               OrdBilling    DelSpeed Satisfaction
```

```
## ProdQual       0.10430307  0.02771800    0.4863250
## Ecom           0.15614733  0.19163607    0.2827450
## TechSup        0.08010182  0.02544069    0.1125972
## CompRes        0.75686859  0.86509170    0.6032626
## Advertising    0.18423559  0.27586308    0.3046695
## ProdLine       0.42440825  0.60185021    0.5505459
## SalesFImage    0.19512741  0.27155126    0.5002053
## ComPricing    -0.11456703 -0.07287173   -0.2082957
## WartyClaim     0.19706512  0.10939460    0.1775448
## OrdBilling     1.00000000  0.75100307    0.5217319
## DelSpeed       0.75100307  1.00000000    0.5770423
## Satisfaction   0.52173191  0.57704227    1.0000000
```

```
#plot the correlation
corrplot(corr , type = "lower" , title = "correlation plot" , method = "numb
er")
```



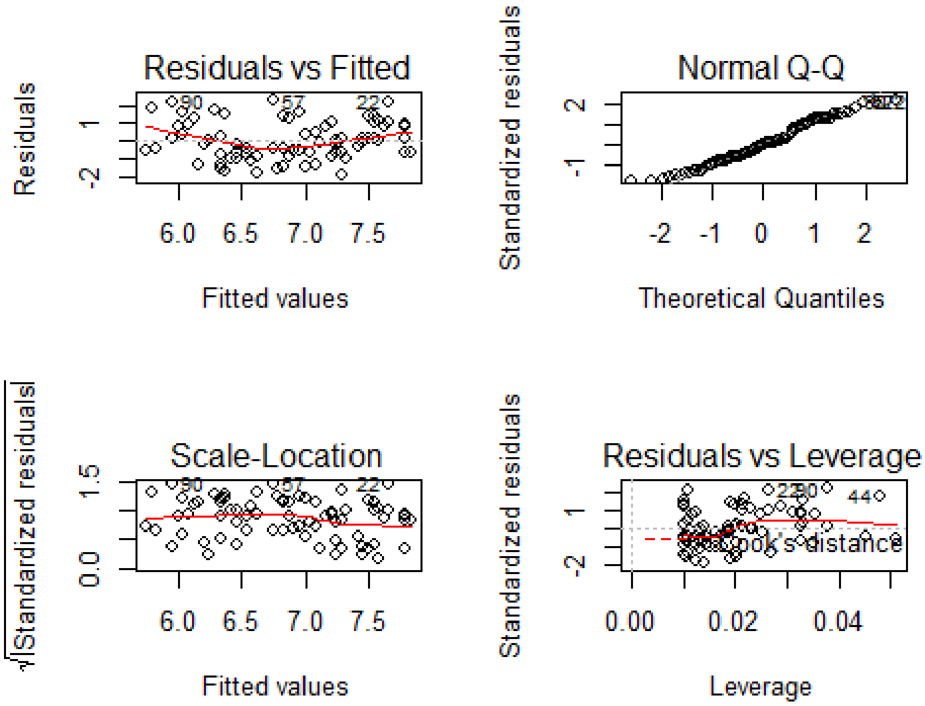

```
#As expected the correlation between sales force image and ecommerce is high
ly significant;
#so is the correlation between delivery speed and order billing with complai
nt resolution. Also,
#the correlation between order & billing and delivery speed. We can safely a
ssume that there
#is a high degree of collinearity between the independent variables
```

Observation:

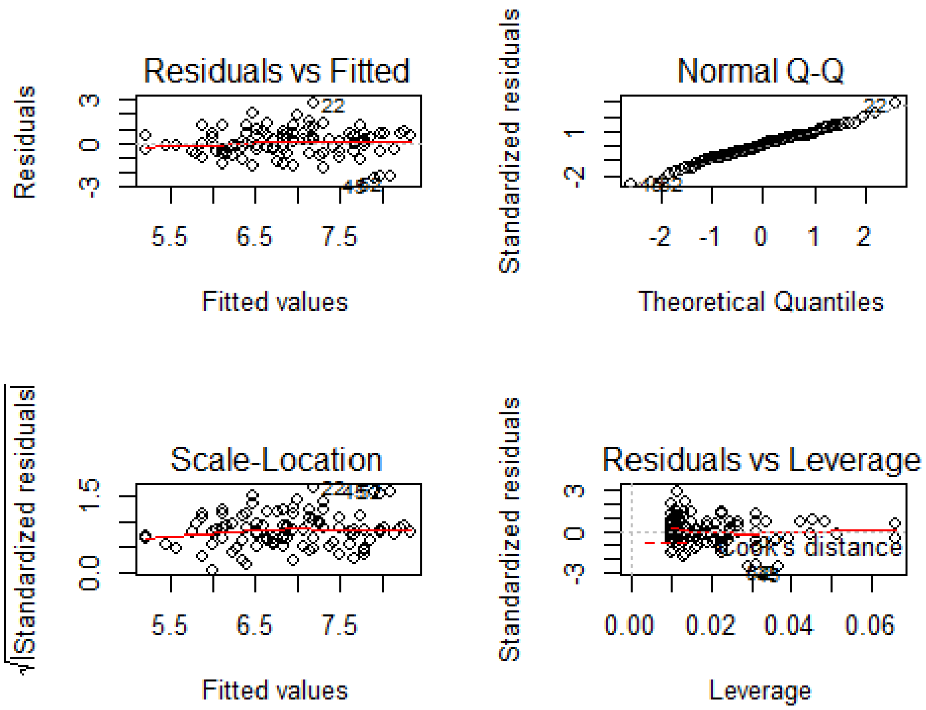
E-commerce and Salesforce Image	Highly Correlated
Technical Support and Warranty Claim	Highly Correlated
Complaint Resolution and Order Billing	Highly Correlated
Complaint Resolution and Delivery Speed	Highly Correlated
Product Line and Delivery Speed	Highly Correlated

4. Simple Linear Regression with all independent variables

```
#Building the initial linear model of dependent variable with all the indepe
ndent model
model1 = lm(Satisfaction~ProdQual,data = mydata)
summary(model1)

##
## Call:
## lm(formula = Satisfaction ~ ProdQual, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88746 -0.72711 -0.01577  0.85641  2.25220
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.67593    0.59765   6.151 1.68e-08 ***
## ProdQual     0.41512    0.07534   5.510 2.90e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.047 on 98 degrees of freedom
## Multiple R-squared:  0.2365, Adjusted R-squared:  0.2287
## F-statistic: 30.36 on 1 and 98 DF,  p-value: 2.901e-07

par(mfrow = c(2,2))
plot(model1)
```




```
#Inference from the model be that the model is having confidence of 22% appr
ox which is not a good sign, also p value is not less then 0.05 so null hypo
thesis is accepted.

#Our next model can be between CompRes and satisfaction
model2 = lm(Satisfaction~CompRes ,data = mydata)
summary(model2)

##
## Call:
## lm(formula = Satisfaction ~ CompRes, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.40450 -0.66164  0.04499  0.63037  2.70949
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.68005    0.44285   8.310 5.51e-13 ***
## CompRes      0.59499    0.07946   7.488 3.09e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9554 on 98 degrees of freedom
## Multiple R-squared:  0.3639, Adjusted R-squared:  0.3574
## F-statistic: 56.07 on 1 and 98 DF,  p-value: 3.085e-11

par(mfrow = c(2,2))
plot(model2)
```




```
#Inference from the model be that the model is having confidence of 35% appr
ox which is not a good sign, also p value is not less then 0.05 so null hypo
thesis is accepted.

#One more model say between DelSpeed and Satisfaction
model3 = lm(Satisfaction~DelSpeed ,data = mydata)
summary(model3)

##
## Call:
## lm(formula = Satisfaction ~ DelSpeed, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22475 -0.54846  0.08796  0.54462  2.59432
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2791     0.5294   6.194 1.38e-08 ***
## DelSpeed      0.9364     0.1339   6.994 3.30e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9783 on 98 degrees of freedom
## Multiple R-squared:  0.333,  Adjusted R-squared:  0.3262
## F-statistic: 48.92 on 1 and 98 DF,  p-value: 3.3e-10

#Inference from the model be that the model is having confidence of 32% appr
ox which is not a good sign, also p value is not less then 0.05 so null hypo
thesis is accepted.
```

```r
#We can also have a model between TechSup and Satisfaction
model4 = lm(Satisfaction~TechSup,data = mydata)
summary(model4)

##
## Call:
## lm(formula = Satisfaction ~ TechSup, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.26136 -0.93297  0.04302  0.82501  2.85617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.44757    0.43592  14.791   <2e-16 ***
## TechSup      0.08768    0.07817   1.122    0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.19 on 98 degrees of freedom
## Multiple R-squared:  0.01268,    Adjusted R-squared:  0.002603
## F-statistic: 1.258 on 1 and 98 DF,  p-value: 0.2647

#Inference from the model be that the model is having confidence of 2% appro
x which is not a good sign, also p value is not less then 0.05 so null hypot
hesis is accepted.
#Model4 is not in radar of acceptance.

model5 = lm(Satisfaction~Ecom,data = mydata)
summary(model5)

##
## Call:
## lm(formula = Satisfaction ~ Ecom, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.37200 -0.78971  0.04959  0.68085  2.34580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.1516     0.6161   8.361 4.28e-13 ***
## Ecom          0.4811     0.1649   2.918  0.00437 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.149 on 98 degrees of freedom
## Multiple R-squared:  0.07994,    Adjusted R-squared:  0.07056
## F-statistic: 8.515 on 1 and 98 DF,  p-value: 0.004368

#Inference from the model be that the model is having confidence of 7% appro
x which is not a good sign, but p value is less then 0.05 so null hypothesis
is rejected here.
```

```r
model6 = lm(Satisfaction~Advertising , data = mydata)
summary(model6)

##
## Call:
## lm(formula = Satisfaction ~ Advertising, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34033 -0.92755  0.05577  0.79773  2.53412
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.6259     0.4237  13.279  < 2e-16 ***
## Advertising   0.3222     0.1018   3.167  0.00206 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.141 on 98 degrees of freedom
## Multiple R-squared:  0.09282,    Adjusted R-squared:  0.08357
## F-statistic: 10.03 on 1 and 98 DF,  p-value: 0.002056

#Inference from the model is that the p-value is less than 0.05 but confiden
ce is only 8%.Null Hypothesis gets accepted.

model7 = lm(Satisfaction~ProdLine , data = mydata)
summary(model7)

##
## Call:
## lm(formula = Satisfaction ~ ProdLine, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3634 -0.7795  0.1097  0.7604  1.7373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.02203    0.45471   8.845 3.87e-14 ***
## ProdLine     0.49887    0.07641   6.529 2.95e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 98 degrees of freedom
## Multiple R-squared:  0.3031, Adjusted R-squared:  0.296
## F-statistic: 42.62 on 1 and 98 DF,  p-value: 2.953e-09

#Inference:p-value is less than 0.05.Confidence is only 29%

model8 = lm(Satisfaction~SalesFImage , data = mydata)
summary(model8)

##
## Call:
## lm(formula = Satisfaction ~ SalesFImage, data = mydata)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2164 -0.5884  0.1838  0.6922  2.0728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.06983    0.50874   8.000 2.54e-12 ***
## SalesFImage  0.55596    0.09722   5.719 1.16e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.037 on 98 degrees of freedom
## Multiple R-squared:  0.2502, Adjusted R-squared:  0.2426
## F-statistic:  32.7 on 1 and 98 DF,  p-value: 1.164e-07
```

```
#Inference : p-value is less than 0.05 and confidence is 24%

model9 = lm(Satisfaction~ComPricing , data = mydata)
summary(model9)
```

```
##
## Call:
## lm(formula = Satisfaction ~ ComPricing, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9728 -0.9915 -0.1156  0.9111  2.5845
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.03856    0.54427  14.769   <2e-16 ***
## ComPricing  -0.16068    0.07621  -2.108   0.0376 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.172 on 98 degrees of freedom
## Multiple R-squared:  0.04339,    Adjusted R-squared:  0.03363
## F-statistic: 4.445 on 1 and 98 DF,  p-value: 0.03756
```

```
#p-value is less than 0.05 and confidence is only 3%

model10 = lm(Satisfaction~WartyClaim , data = mydata)
summary(model10)
```

```
##
## Call:
## lm(formula = Satisfaction ~ WartyClaim, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36504 -0.90202  0.03019  0.90763  2.88985
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   5.3581     0.8813   6.079 2.32e-08 ***
## WartyClaim    0.2581     0.1445   1.786   0.0772 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.179 on 98 degrees of freedom
## Multiple R-squared:  0.03152,    Adjusted R-squared:  0.02164
## F-statistic:  3.19 on 1 and 98 DF,  p-value: 0.0772
```

```r
#p-value is not less than 0.05 and confidence is only 2%

model11 = lm(Satisfaction~OrdBilling , data = mydata)
summary(model11)
```

```
##
## Call:
## lm(formula = Satisfaction ~ OrdBilling, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4005 -0.7071 -0.0344  0.7340  2.9673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.0541     0.4840   8.377 3.96e-13 ***
## OrdBilling    0.6695     0.1106   6.054 2.60e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.022 on 98 degrees of freedom
## Multiple R-squared:  0.2722, Adjusted R-squared:  0.2648
## F-statistic: 36.65 on 1 and 98 DF,  p-value: 2.602e-08
```

```r
#p-value is less than 0.05 and confidence is only 26%
```

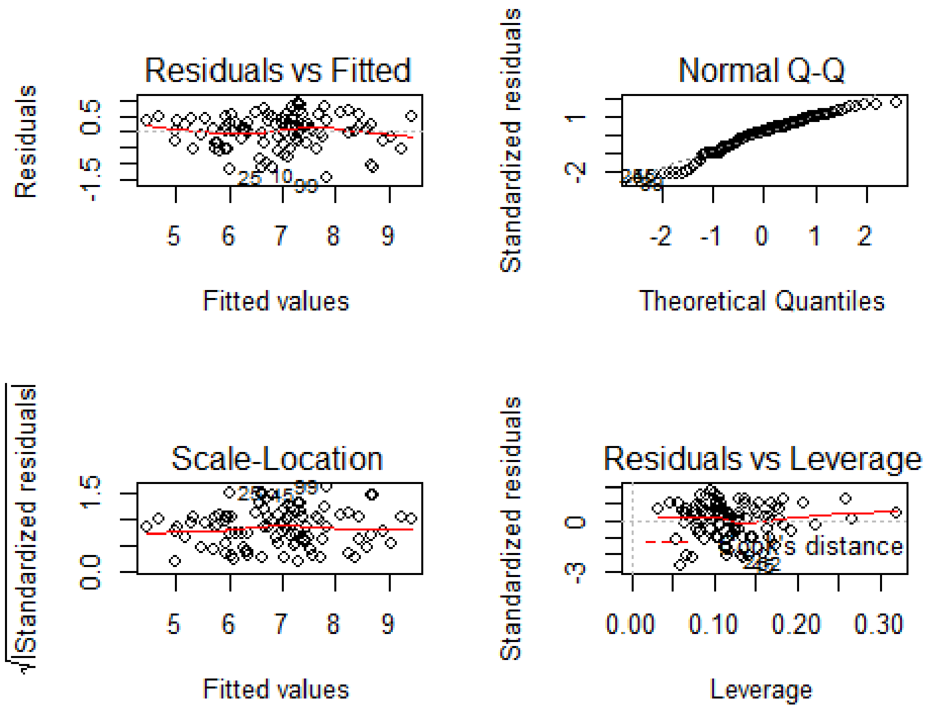
5. Principle Component Analysis

```r
#Lets now create a new model for the satisfaction including all the factors on
which it would depend like product quality, tech support etc.
model13 = lm(Satisfaction~.,data = mydata)
summary(model13)
```

```
##
## 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
```

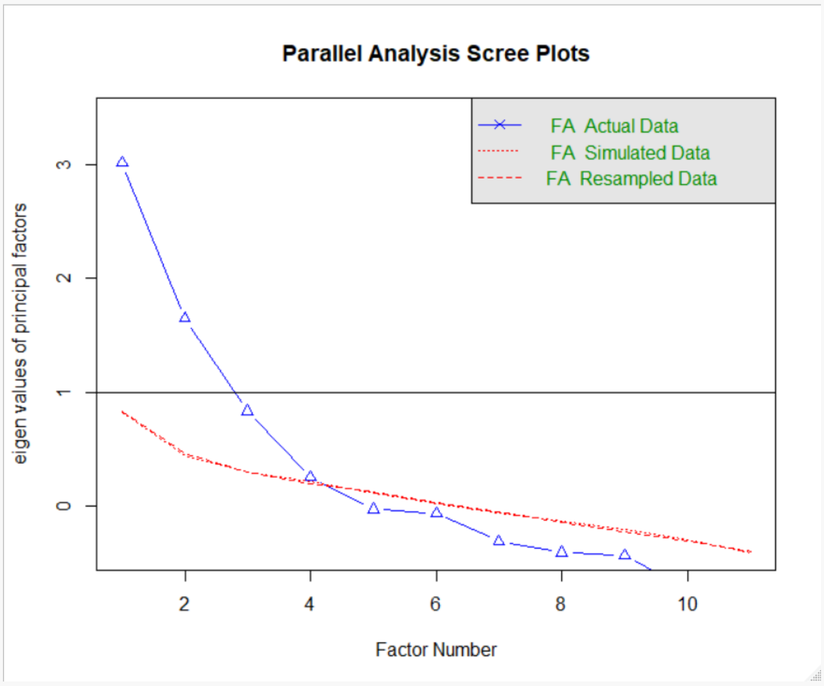
```
par(mfrow = c(2,2))
plot(model13)
```




```
vif(model13)
```

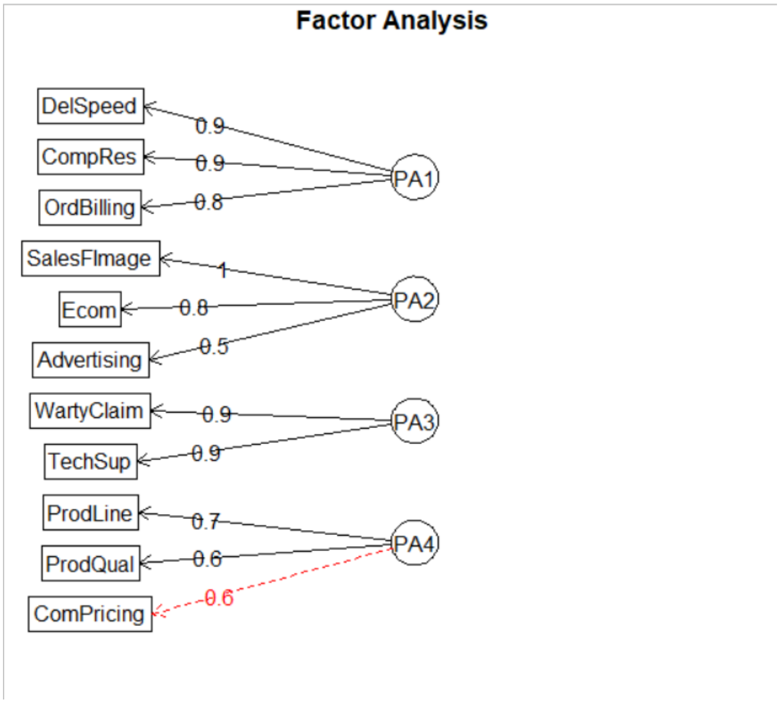
```
##    ProdQual        Ecom     TechSup     CompRes Advertising    ProdLine
##    1.635797    2.756694    2.976796    4.730448    1.508933    3.488185
## SalesFImage  ComPricing  WartyClaim  OrdBilling    DelSpeed
##    3.439420    1.635000    3.198337    2.902999    6.516014
```

```
#From the above linear model13 we see that out of 11 factors only 3 of them are
highly significant namely "ProdQual","Ecom"andSalesFImage"


#Lets implement Factor analysis on the dataset
#To do the factor analysis lets first create a subset of the dataset containing
only the independent variables
subset_mydata = subset(mydata,select = c(-12))
corr3 = cor(subset_mydata)
KMO(r=corr3)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = corr3)
## 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

#since MSA > 0.5 we can go ahead with the factor analysis
ev = eigen(corr)
Eigen_Values = ev$values
Eigen_Values

##  [1] 4.04285997 2.55292440 1.69222417 1.21754639 0.63596293 0.56853132
##  [7] 0.40282774 0.32448016 0.23613948 0.14422355 0.09913845 0.08314143

#We have the eigen values we can now find the factors which should be appropiat
e for the test.
#Eigen values > 1 can be considered as number of factors to consider for PCA (K
aizer principle)

p = fa.parallel(subset_mydata,fm="miners",fa="fa")
```



```r
#from the graph we see that number of factors which can be considered for PCA i
s 4.

factor_analysis1 = fa(r = subset_mydata,nfactors = 4,rotate = "varimax" , fm =
"pa")
print(factor_analysis1)

## Factor Analysis using method =  pa
## Call: fa(r = subset_mydata, 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 functio
n 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

#Let us see the module of grouping of the factors.
fa.diagram(factor_analysis1)

#Diagram suggests 4 Principle Components
```

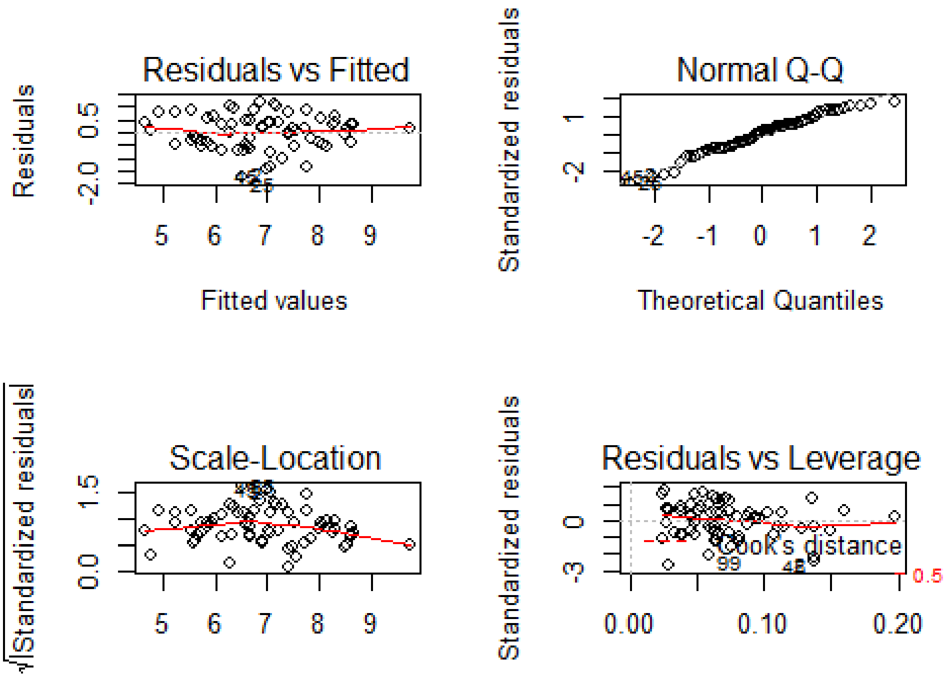


Sr. No.	Factors	Variables	Label	Short Description
1	PA1	DelSpeed, CompRes & OrdBilling	Purchase	Variables related to Order Placing, Order Delivery and Complaints
2	PA2	SalesFImage, Ecom & Advertising	Marketing	Variables are related to Website experience, Advertising, etc
3	PA3	TechSup & WartyClaim	Support	Variables are related to product support experience
4	PA4	ProdQual, ComPricing & ProdLine	Product	Variables are related to the product variety and pricing

6. Multiple Linear Regression after PCA

```
#Let us perform the regression analysis
regression_data = cbind(mydata[12],factor_analysis1$scores)
head(regression_data)

##   Satisfaction        PA1        PA2          PA3         PA4
## 1          8.2 -0.1338871  0.9175166 -1.719604873  0.09135411
## 2          5.7  1.6297604 -2.0090053 -0.596361722  0.65808192
## 3          8.9  0.3637658  0.8361736  0.002979966  1.37548765
## 4          4.8 -1.2225230 -0.5491336  1.245473305 -0.64421384
## 5          7.1 -0.4854209 -0.4276223 -0.026980304  0.47360747
## 6          4.7 -0.5950924 -1.3035333 -1.183019401 -0.95913571

names(regression_data) = c("Satisfaction","Purchase","Marketing","Support","
Product")
head(regression_data)

##   Satisfaction   Purchase  Marketing      Support     Product
## 1          8.2 -0.1338871  0.9175166 -1.719604873  0.09135411
## 2          5.7  1.6297604 -2.0090053 -0.596361722  0.65808192
## 3          8.9  0.3637658  0.8361736  0.002979966  1.37548765
## 4          4.8 -1.2225230 -0.5491336  1.245473305 -0.64421384
## 5          7.1 -0.4854209 -0.4276223 -0.026980304  0.47360747
## 6          4.7 -0.5950924 -1.3035333 -1.183019401 -0.95913571

str(regression_data)

## 'data.frame':    100 obs. of  5 variables:
##  $ Satisfaction: num  8.2 5.7 8.9 4.8 7.1 4.7 5.7 6.3 7 5.5 ...
##  $ Purchase    : num  -0.134 1.63 0.364 -1.223 -0.485 ...
##  $ Marketing   : num  0.918 -2.009 0.836 -0.549 -0.428 ...
##  $ Support     : num  -1.7196 -0.59636 0.00298 1.24547 -0.02698 ...
##  $ Product     : num  0.0914 0.6581 1.3755 -0.6442 0.4736 ...

#Divide the data into test set and train set
set.seed(1)
sample_data = sample(1:nrow(regression_data),0.7*nrow(regression_data))
train_data = regression_data[sample_data,]
test_data = regression_data[-sample_data,]
str(test_data)

## 'data.frame':    30 obs. of  5 variables:
##  $ Satisfaction: num  8.9 4.8 7.1 6.3 7 5.5 6 8 6.6 6.8 ...
##  $ Purchase    : num  0.364 -1.223 -0.485 -0.113 0.958 ...
##  $ Marketing   : num  0.836 -0.549 -0.428 -0.131 0.348 ...
##  $ Support     : num  0.00298 1.24547 -0.02698 -0.69924 -0.14226 ...
##  $ Product     : num  1.375 -0.644 0.474 -1.366 -0.935 ...

#New regression model be
model14 = lm(Satisfaction~.,data = train_data)
summary(model14)

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

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7214 -0.4681  0.0869  0.3945  1.1392
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.92697    0.07987  86.724  < 2e-16 ***
## Purchase     0.58291    0.07732   7.539 1.92e-10 ***
## Marketing    0.59318    0.07580   7.826 5.95e-11 ***
## Support      0.02175    0.08450   0.257    0.798
## Product      0.59345    0.08916   6.656 7.00e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6616 on 65 degrees of freedom
## Multiple R-squared:  0.7396, Adjusted R-squared:  0.7236
## F-statistic: 46.16 on 4 and 65 DF,  p-value: < 2.2e-16

par(mfrow = c(2,2))
plot(model14)
```




```
vif(model14)

##  Purchase Marketing   Support   Product
##  1.016649  1.023243  1.005772  1.001513

## R-squared for train dataset
summary(model14)$r.squared

## [1] 0.7396184
```

```r
#predict the train model
pred_train = predict(model14,train_data)

#train model performance
mse_train_perf = mse(train_data$Satisfaction,pred_train)
rmse_train_perf = sqrt(mse_train_perf)
print(rmse_train_perf)

## [1] 0.6375662

pred_test = predict(model14,test_data)

#prediction model performance
mse_test_perf = mse(test_data$Satisfaction , pred_test)
rmse_test_perf = sqrt(mse_test_perf)
print(rmse_test_perf)

## [1] 0.6916997

## R-squared for Test dataset
cor(test_data$Satisfaction, pred_test)^2

## [1] 0.556584
```

Value	Train Data	Test Data
R-Squared	0.7396184	0.556584
RMSE	0.6375662	0.6916997

There is not much variation in the R-squared and RMSE values of the trained and test datasets; so it can be inferred that the model is good and not over fitting.

Customer Satisfaction is having variation because of Purchase, Marketing and Product variety. The equation here would be:

Satisfaction = 0.58291 *Purchase + 0.59318 *Marketing + 0.59348*Product