Mini Project – Advanced Statistics (PCA/FA, Regression)

By Ramprasad Mohan

Customer Satisfaction -- Hair product

Table of 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 customer satisfaction.

- To perform exploratory data analysis on the dataset. Showcase some charts, graphs. Check for outliers and missing values
- Find if there is evidence of multicollinearity
- Perform simple linear regression for the dependent variable with every independent variable
- Perform PCA/Factor analysis by extracting 4 factors. Interpret the output and name the Factors
- 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

2 Assumptions

The assumptions such as homoscedasticity, normality, linearity, multicollinearity is considered

3 Step by Step approach

We shall follow step by step approach to arrive to the conclusion as follows:

- Exploratory Data Analysis
- Descriptive Statistics
- Check Multicollinearity
- Perform Simple Linear Regression
- Perform PCA/FA, Apply Kaiser – Normalization rule
- Need of Larger Sample Size
- Conclusion

4 Solution

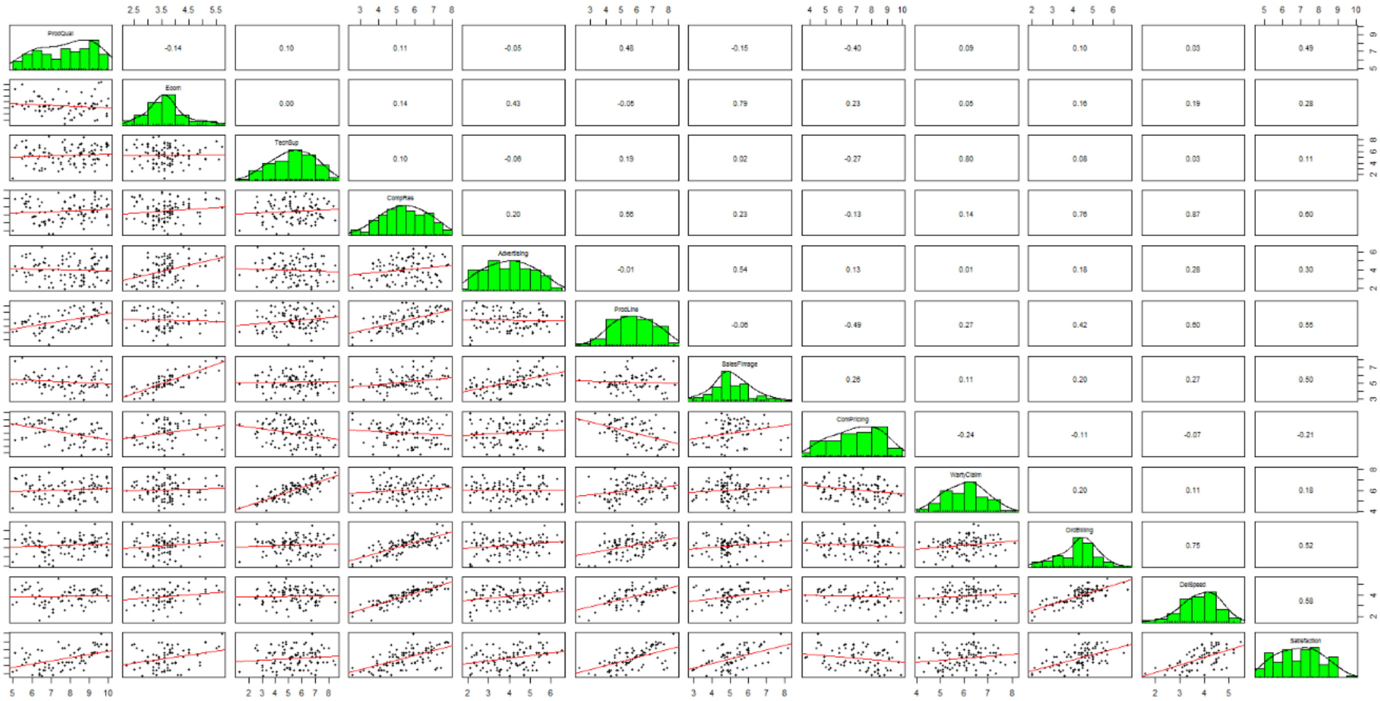
4.1 EDA - Basic data summary, Univariate, Bivariate analysis, graphs

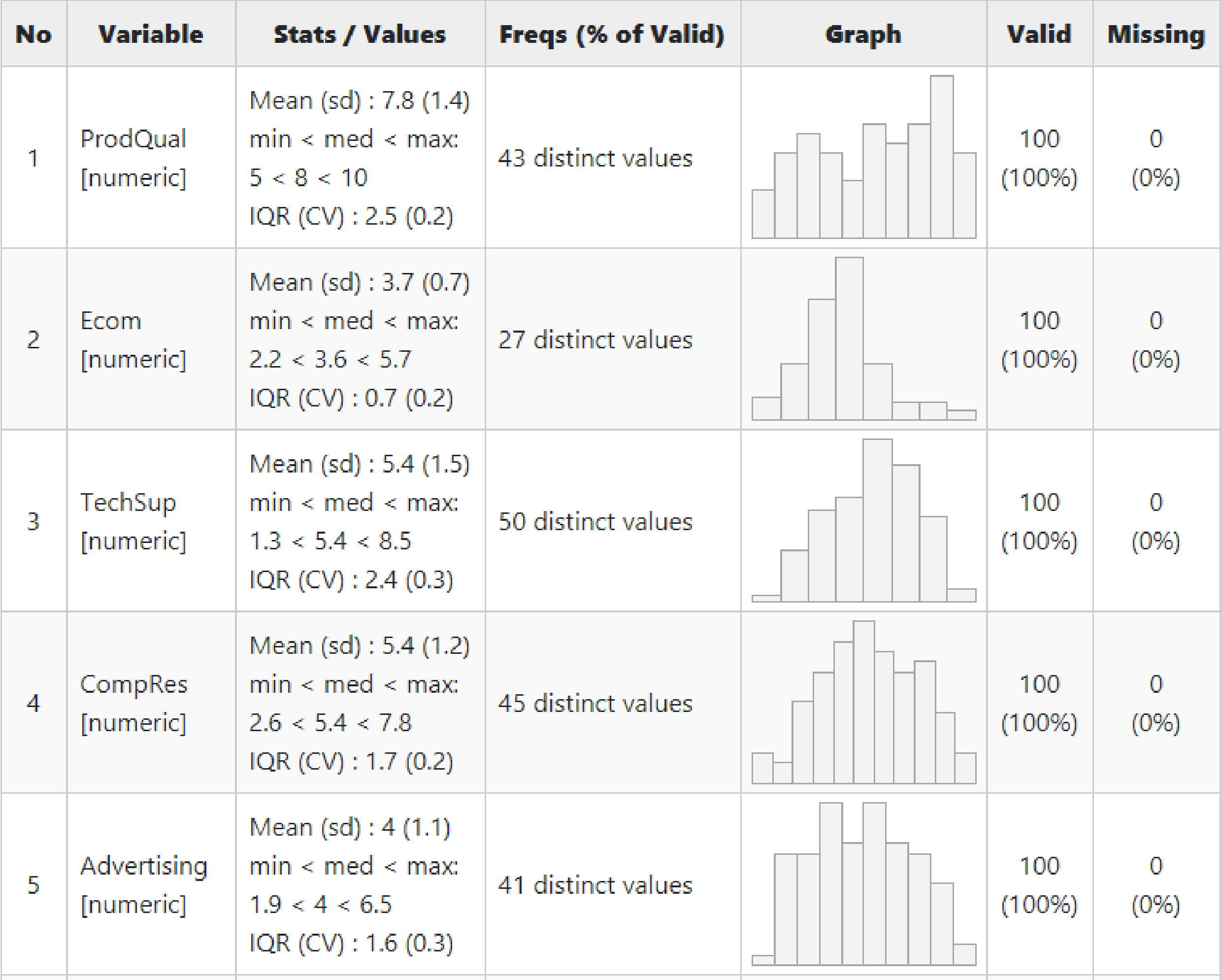
Since all the variables are continuous variables, below is the consolidated univariate and bivariate analysis using pair.panels() function in psych package.

Alternatively the summary of each variable can be done from 'summarytools' package by using the view and dfsummary function

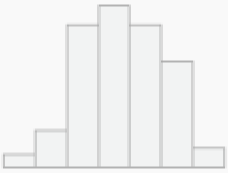
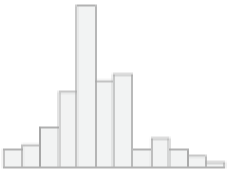
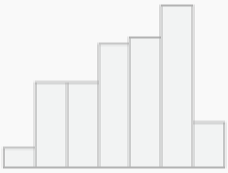
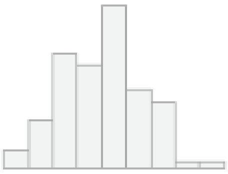
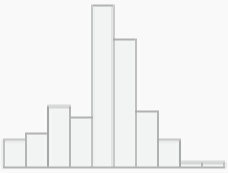
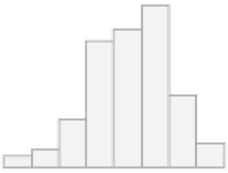
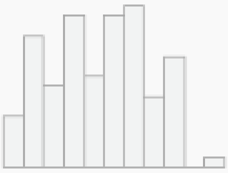
Dimensions: 100 x 12

plot_zoom_png





No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	ProdQual [numeric]	Mean (sd) : 7.8 (1.4) min < med < max: 5 < 8 < 10 IQR (CV) : 2.5 (0.2)	43 distinct values		100 (100%)	0 (0%)
2	Ecom [numeric]	Mean (sd) : 3.7 (0.7) min < med < max: 2.2 < 3.6 < 5.7 IQR (CV) : 0.7 (0.2)	27 distinct values		100 (100%)	0 (0%)
3	TechSup [numeric]	Mean (sd) : 5.4 (1.5) min < med < max: 1.3 < 5.4 < 8.5 IQR (CV) : 2.4 (0.3)	50 distinct values		100 (100%)	0 (0%)
4	CompRes [numeric]	Mean (sd) : 5.4 (1.2) min < med < max: 2.6 < 5.4 < 7.8 IQR (CV) : 1.7 (0.2)	45 distinct values		100 (100%)	0 (0%)
5	Advertising [numeric]	Mean (sd) : 4 (1.1) min < med < max: 1.9 < 4 < 6.5 IQR (CV) : 1.6 (0.3)	41 distinct values		100 (100%)	0 (0%)

6	ProdLine [numeric]	Mean (sd) : 5.8 (1.3) min < med < max: 2.3 < 5.8 < 8.4 IQR (CV) : 2.1 (0.2)	42 distinct values		100 (100%)	0 (0%)
7	SalesFImage [numeric]	Mean (sd) : 5.1 (1.1) min < med < max: 2.9 < 4.9 < 8.2 IQR (CV) : 1.3 (0.2)	35 distinct values		100 (100%)	0 (0%)
8	ComPricing [numeric]	Mean (sd) : 7 (1.5) min < med < max: 3.7 < 7.1 < 9.9 IQR (CV) : 2.5 (0.2)	45 distinct values		100 (100%)	0 (0%)
9	WartyClaim [numeric]	Mean (sd) : 6 (0.8) min < med < max: 4.1 < 6.1 < 8.1 IQR (CV) : 1.2 (0.1)	34 distinct values		100 (100%)	0 (0%)
10	OrdBilling [numeric]	Mean (sd) : 4.3 (0.9) min < med < max: 2 < 4.4 < 6.7 IQR (CV) : 1.1 (0.2)	37 distinct values		100 (100%)	0 (0%)
11	DelSpeed [numeric]	Mean (sd) : 3.9 (0.7) min < med < max: 1.6 < 3.9 < 5.5 IQR (CV) : 1 (0.2)	30 distinct values		100 (100%)	0 (0%)
12	Satisfaction [numeric]	Mean (sd) : 6.9 (1.2) min < med < max: 4.7 < 7 < 9.9 IQR (CV) : 1.6 (0.2)	29 distinct values		100 (100%)	0 (0%)

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

Below is the code to check for outliers, we could see that there are no missing /null values in the dataset

Code : `summary(is.na(hairorginal))`

Output :

```
 ProdQual         Ecom          TechSup         CompRes       
 Mode  :logical   Mode  :logical   Mode  :logical   Mode  :logical  
 FALSE:100        FALSE:100        FALSE:100        FALSE:100       

 Advertising       ProdLine        SalesFImage      ComPricing     
 Mode  :logical   Mode  :logical   Mode  :logical   Mode  :logical  
 FALSE:100        FALSE:100        FALSE:100        FALSE:100       

 WartyClaim       OrdBilling        DelSpeed        Satisfaction   
 Mode  :logical    Mode  :logical    Mode  :logical    Mode  :logical
 FALSE:100         FALSE:100         FALSE:100         FALSE:100
```

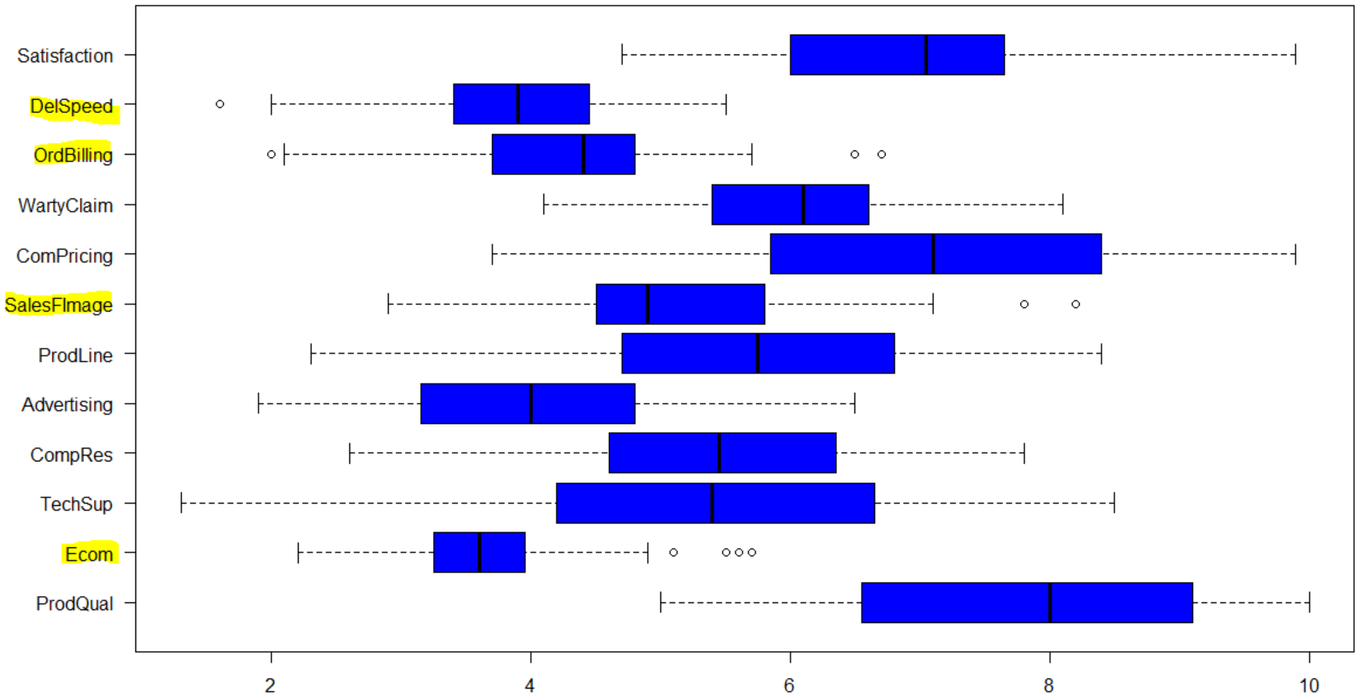
The summary of the values in the dataset are shown below:

Output :

```
    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
```

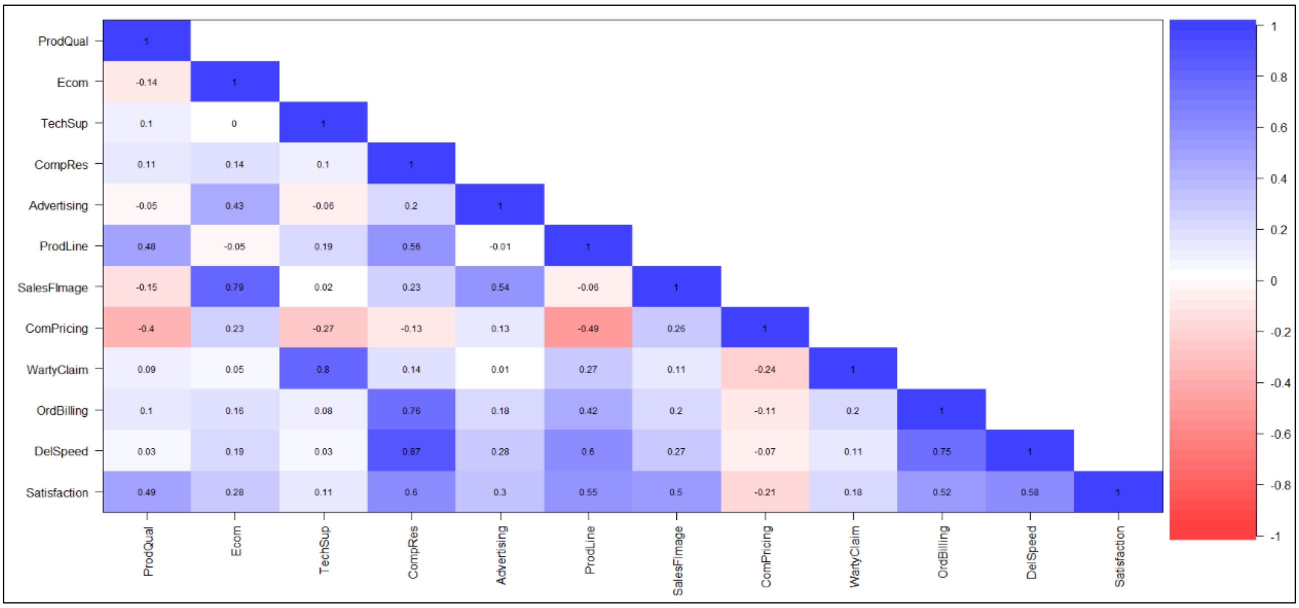
Outliers:

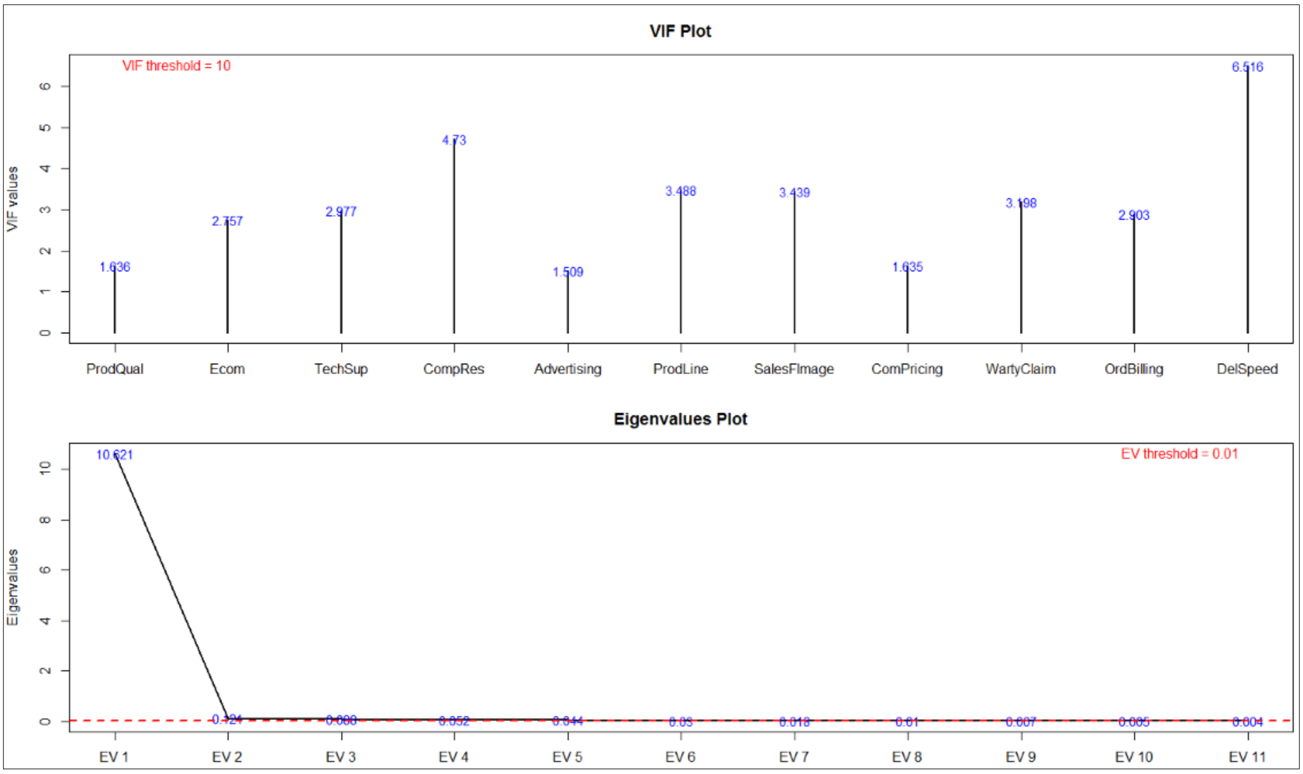
Boxplot is drawn for each variable to determine outlier values (1.5 IQR)



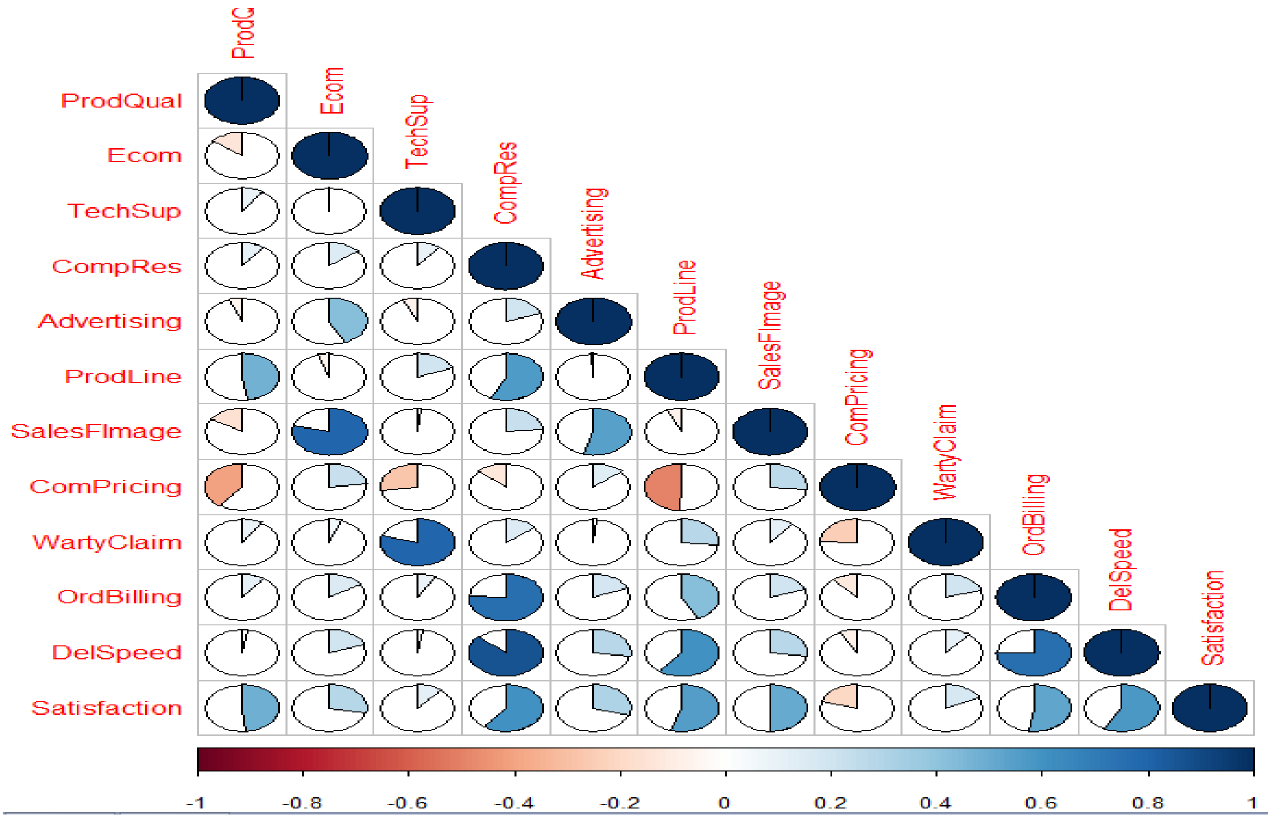
We could see that the Delivery Speed, Order & Billing , Sales Force Image and E-Commerce variables have outlier values while the rest do not have any outliers.

4.3 *Check for Multicollinearity - Plot the graph based on Multicollinearity*

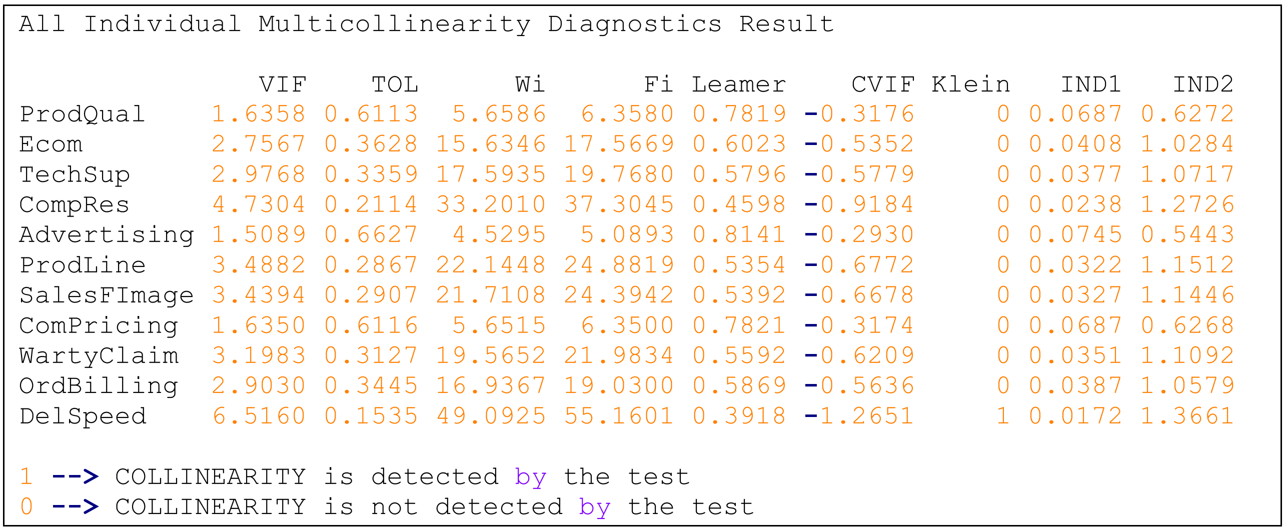




From the correlation plots and Normality check using MVN() function we could see that there is no multicollinearity with the dataset and we can proceed for further analysis.



Running diagnostic test for multicollinearity



```
All Individual Multicollinearity Diagnostics Result

               VIF    TOL      Wi      Fi Leamer    CVIF Klein   IND1   IND2
ProdQual    1.6358 0.6113  5.6586  6.3580 0.7819 -0.3176     0 0.0687 0.6272
Ecom        2.7567 0.3628 15.6346 17.5669 0.6023 -0.5352     0 0.0408 1.0284
TechSup     2.9768 0.3359 17.5935 19.7680 0.5796 -0.5779     0 0.0377 1.0717
CompRes     4.7304 0.2114 33.2010 37.3045 0.4598 -0.9184     0 0.0238 1.2726
Advertising 1.5089 0.6627  4.5295  5.0893 0.8141 -0.2930     0 0.0745 0.5443
ProdLine    3.4882 0.2867 22.1448 24.8819 0.5354 -0.6772     0 0.0322 1.1512
SalesFImage 3.4394 0.2907 21.7108 24.3942 0.5392 -0.6678     0 0.0327 1.1446
ComPricing  1.6350 0.6116  5.6515  6.3500 0.7821 -0.3174     0 0.0687 0.6268
WartyClaim  3.1983 0.3127 19.5652 21.9834 0.5592 -0.6209     0 0.0351 1.1092
OrdBilling  2.9030 0.3445 16.9367 19.0300 0.5869 -0.5636     0 0.0387 1.0579
DelSpeed    6.5160 0.1535 49.0925 55.1601 0.3918 -1.2651     1 0.0172 1.3661

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

From the diagnostic test using mctest() package we could see that Delivery Speed is causing multicollinearity. Since we are performing PCA, the multicollinearity will be removed.

4.4 Simple Linear Regression (with every variable)

Simple linear regression is performed with Customer Satisfaction as dependent variable and every other variable as Independent using for loop. From the below R output we could see that Complaint resolution, Product line and delivery speed has the highest R value that contributes to the customer satisfaction.

```
----------------------------------------------------------------------
"Satisfaction ~ ProdQual"

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

-----------------------------------------------------------------------------
"Model for combination -2"
"Satisfaction ~ Ecom"
```

```
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

-----------------------------------------------------------------------------
"Model for combination -3"
"Satisfaction ~ TechSup"

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

-----------------------------------------------------------------------------
"Model for combination -4"
"Satisfaction ~ CompRes"
```

```
Call:
lm(formula = formula, data = field)

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

------------------------------------------------------------------------------
"Model for combination -5"
"Satisfaction ~ Advertising"
```

```
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

------------------------------------------------------------------------------
"Model for combination -6"
"Satisfaction ~ ProdLine"

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

------------------------------------------------------------------------------
"Model for combination -7"
 "Satisfaction ~ SalesFImage"
```

```
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

-----------------------------------------------------------------------------
"Model for combination -8"
"Satisfaction ~ ComPricing"

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

------------------------------------------------------------------------------
"Model for combination -9"
"Satisfaction ~ WartyClaim"

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

------------------------------------------------------------------------------
"Model for combination -10"
"Satisfaction ~ OrdBilling"

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

-------------------------------------------------------------------------------
"Model for combination -11"
"Satisfaction ~ DelSpeed"

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
```

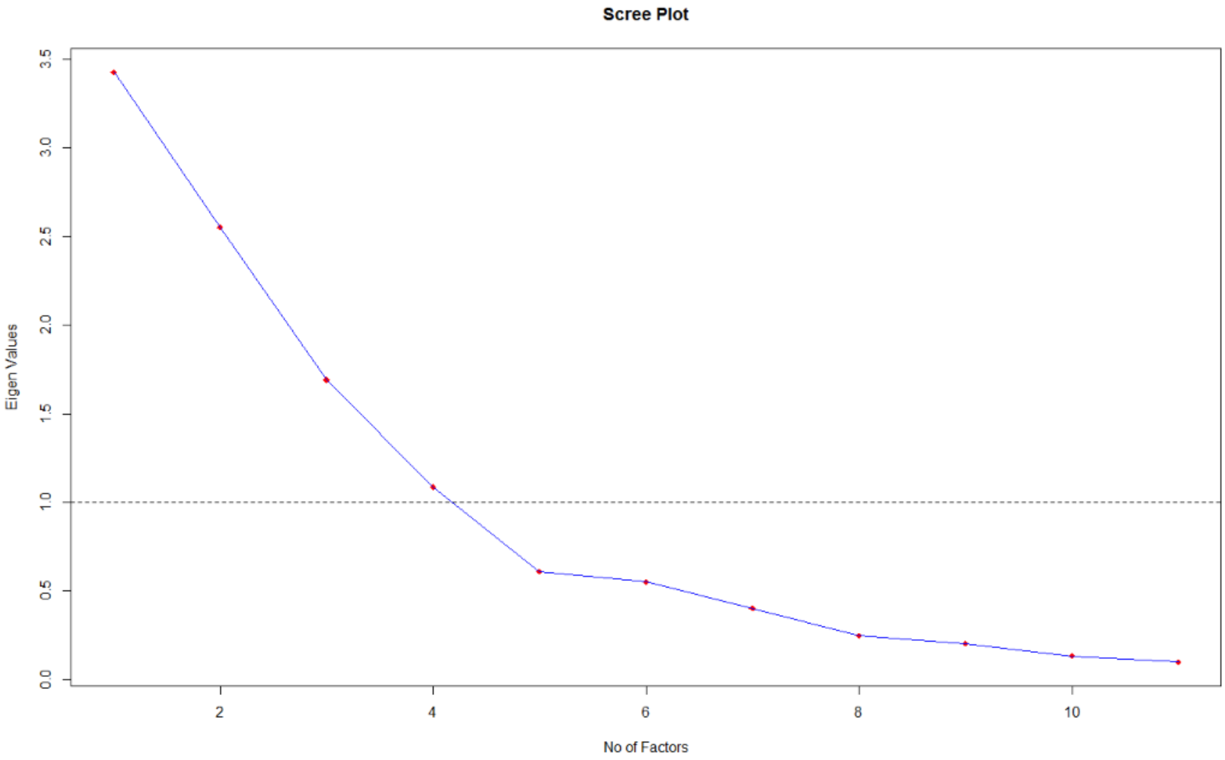
4.5 Perform PCA/FA and Interpret the Eigen Values (apply Kaiser Normalization Rule)

First we are preparing the data by removing the 'ID" column and "Customer satisfaction" column. The new data created is subjected to KMO – Barlett test to check if the data is suitable for PCA/FA analysis by using KMO() function

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

From the above R output we could see that the MSA value is 0.65 which is greater than 0.5. If the value is less than 0.5 or nearer to zero, it suggest that there are large partial correlations compared to the sum of correlations. In other words, there are widespread correlations which are a large problem for factor analysis.

Then we are running the new data with eigen() function to determine the eigen values and plotting them in Scree plot



As per Kaiser Normalization Rule, any eigen values above 1 should be considered for PCA/FA analysis hence as per rule we are considering 4 factors for PCA.

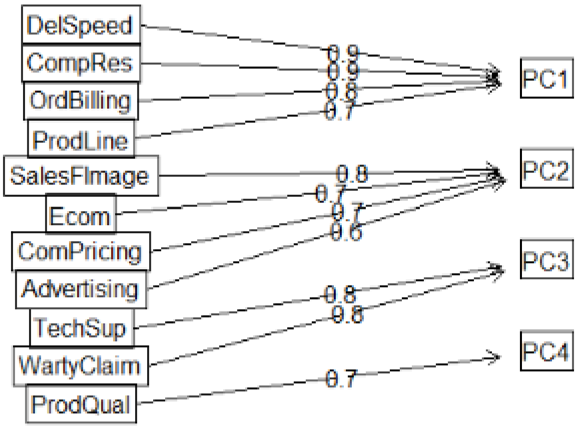
After determining the factors we are performing PCA/FA without rotation of axis and getting the factor loadings.

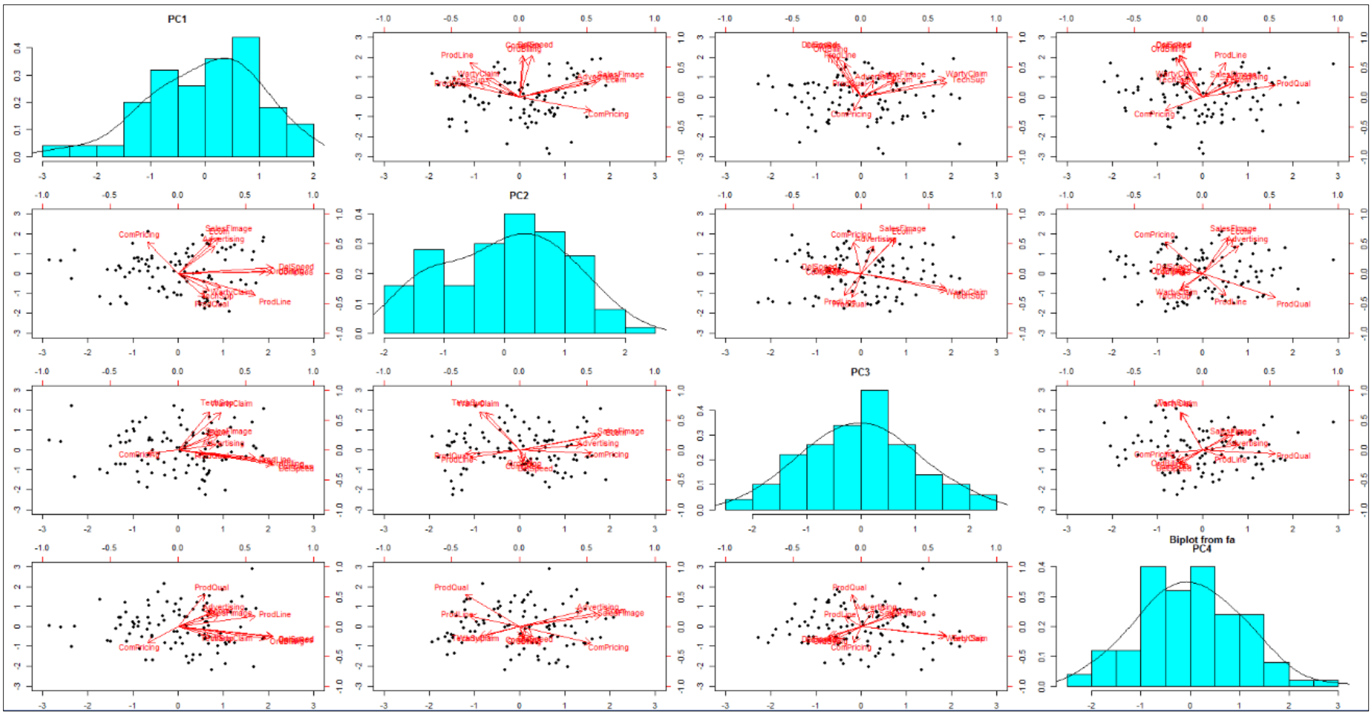
We could see that 80% of the variations is explained by the four factors. The output also shows the communality and uniqueness values, below are the results. H2 are the communalities which shows the variances of each variable explained by the four factors

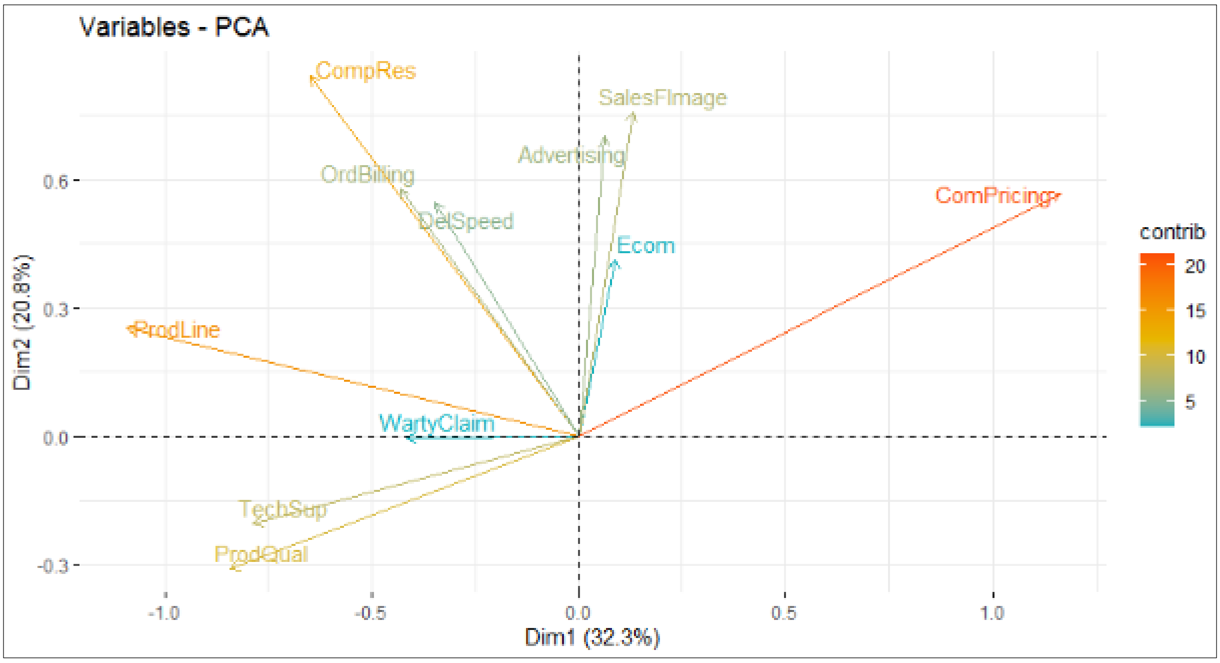
	item	PC1	PC2	PC3	PC4	h2	u2	com
DelSpeed	11	0.88	0.12	-0.30	-0.21	**0.91**	0.086	1.4
CompRes	4	0.87	0.03	-0.27	-0.22	**0.88**	0.119	1.3
OrdBilling	10	0.81	0.04	-0.22	-0.25	**0.77**	0.234	1.3
ProdLine	6	0.72	-0.45	-0.15	0.21	**0.79**	0.213	2.0
SalesFImage	7	0.38	0.75	0.31	0.23	**0.86**	0.141	2.1
Ecom	2	0.31	0.71	0.31	0.28	**0.78**	0.223	2.1
ComPricing	8	-0.28	0.66	-0.07	-0.35	**0.64**	0.359	1.9
Advertising	5	0.34	0.58	0.11	0.33	**0.58**	0.424	2.4
TechSup	3	0.29	-0.37	0.79	-0.20	**0.89**	0.107	1.9
WartyClaim	9	0.39	-0.31	0.78	-0.19	**0.89**	0.108	2.0
ProdQual	1	0.25	-0.50	-0.08	0.67	**0.77**	0.232	2.2

	PC1	PC2	PC3	PC4
SS loadings	3.43	2.55	1.69	1.09
Proportion Var	0.31	0.23	0.15	0.10
Cumulative Var	0.31	0.54	0.70	0.80
Proportion Explained	0.39	0.29	0.19	0.12
Cumulative Proportion	0.39	0.68	0.88	1.00

Since there are ambiguities in factor loadings and few variables are showing equally high factor loadings we are subjecting it to “Varimax” rotation to push all higher correlation values to 1 and lower correlation values nearer to zero. Below is the factor plot for the components.





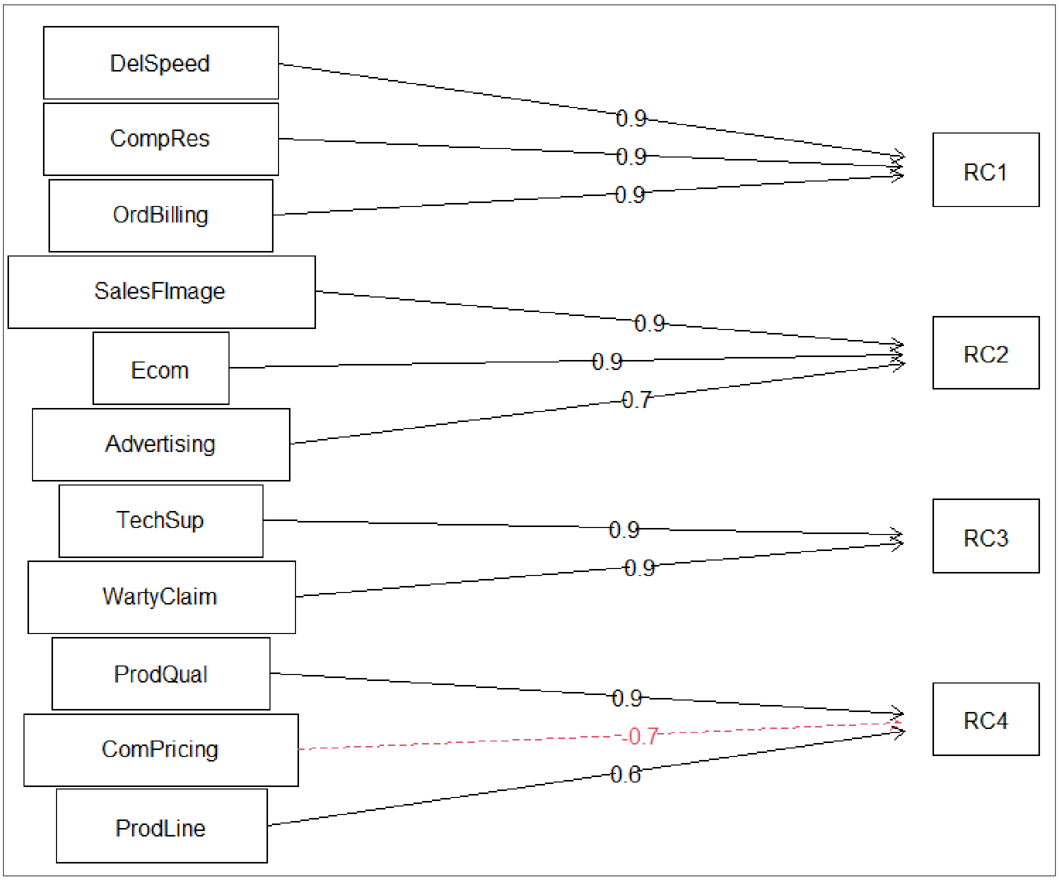


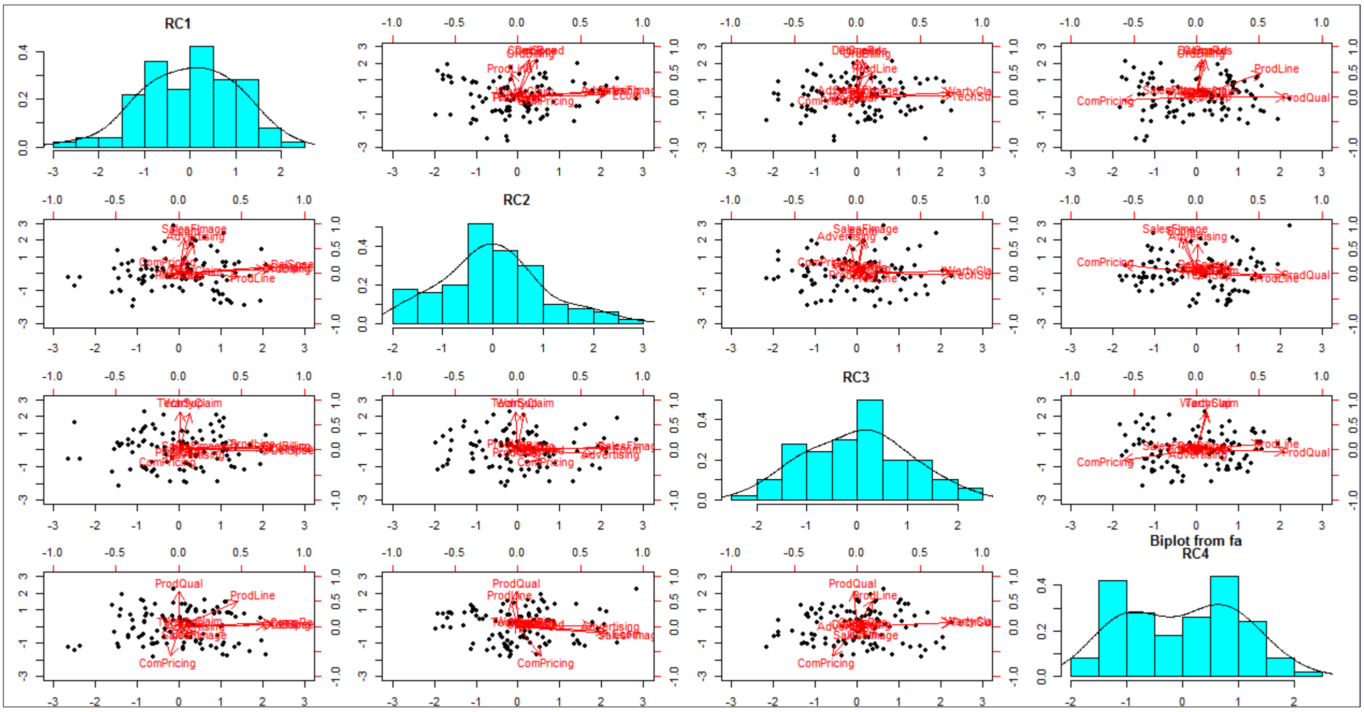
4.6 Output Interpretation Tell why only 4 factors are being asked in the questions and tell whether it is correct in choosing 4 factors. Name the factors with correct explanations.

After running PCA with "Varimax" or orthogonal rotation below are the rotated components factor loadings. We could see that the communalities value has not changed and clear high loading values for variables in each components. Thus varimax rotation is powerful to provide clear output

	item	RC1	RC2	RC3	RC4	h2	u2	com
DelSpeed	11	0.94	0.18	0.00	0.05	**0.91**	0.086	1.1
CompRes	4	0.93	0.12	0.05	0.09	**0.88**	0.119	1.1
OrdBilling	10	0.86	0.11	0.08	0.04	**0.77**	0.234	1.1
SalesFImage	7	0.13	0.90	0.08	-0.16	**0.86**	0.141	1.1
Ecom	2	0.06	0.87	0.05	-0.12	**0.78**	0.223	1.1
Advertising	5	0.14	0.74	-0.08	0.01	**0.58**	0.424	1.1
TechSup	3	0.02	-0.02	0.94	0.10	**0.89**	0.107	1.0
WartyClaim	9	0.11	0.05	0.93	0.10	**0.89**	0.108	1.1
ProdQual	1	0.00	-0.01	-0.03	0.88	**0.77**	0.232	1.0
ComPricing	8	-0.09	0.23	-0.25	-0.72	**0.64**	0.359	1.5
ProdLine	6	0.59	-0.06	0.15	0.64	**0.79**	0.213	2.1

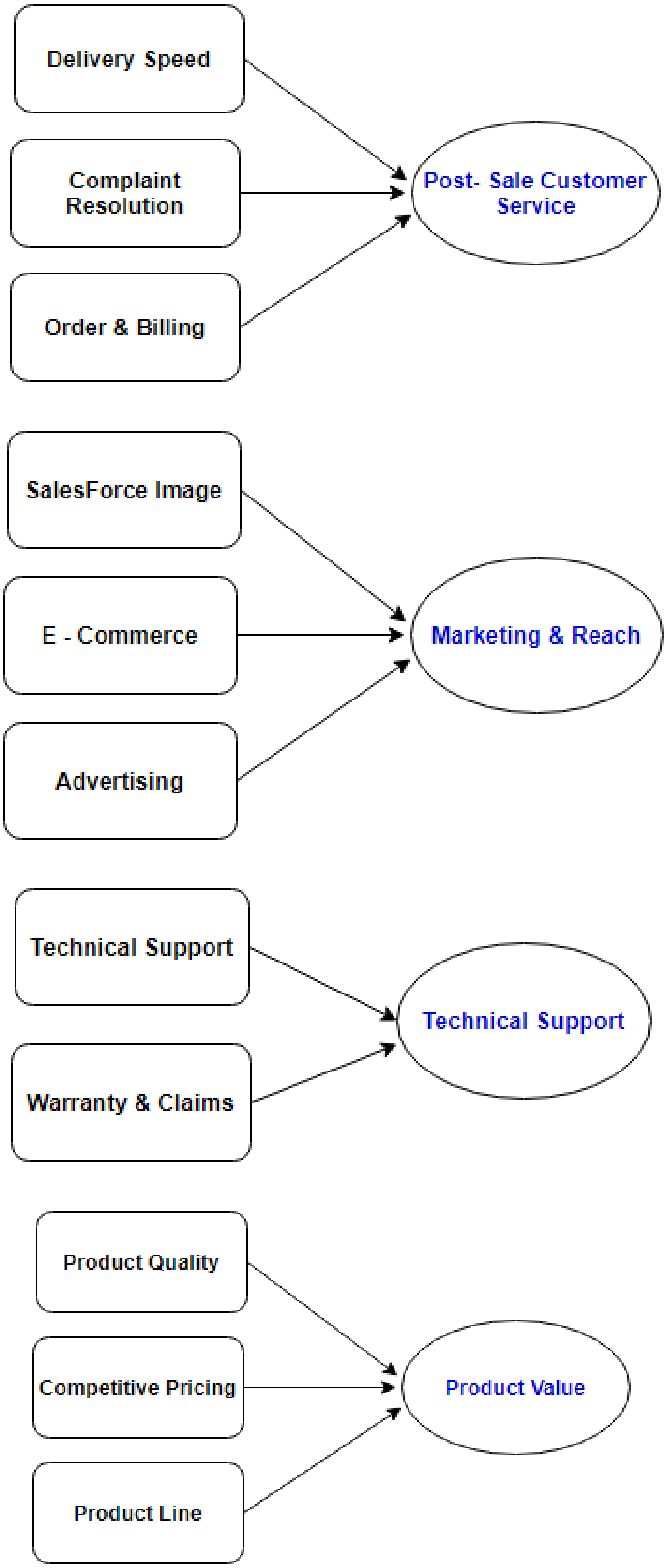
```
                       RC1  RC2  RC3  RC4
SS loadings           2.89 2.23 1.86 1.77
Proportion Var        0.26 0.20 0.17 0.16
Cumulative Var        0.26 0.47 0.63 0.80
Proportion Explained  0.33 0.26 0.21 0.20
Cumulative Proportion 0.33 0.59 0.80 1.00

Mean item complexity =  1.2
Test of the hypothesis that 4 components are sufficient
```





Only four factors are asked because, from factor loadings we could see that most of the variables are covered under 4 factors and any addition of factors will create ambiguities in result interpretation. It is also evident from eigen values in Scree plot that four factors will be the optimal number for PCA/FA analysis and the same has been proved from the output diagram above.

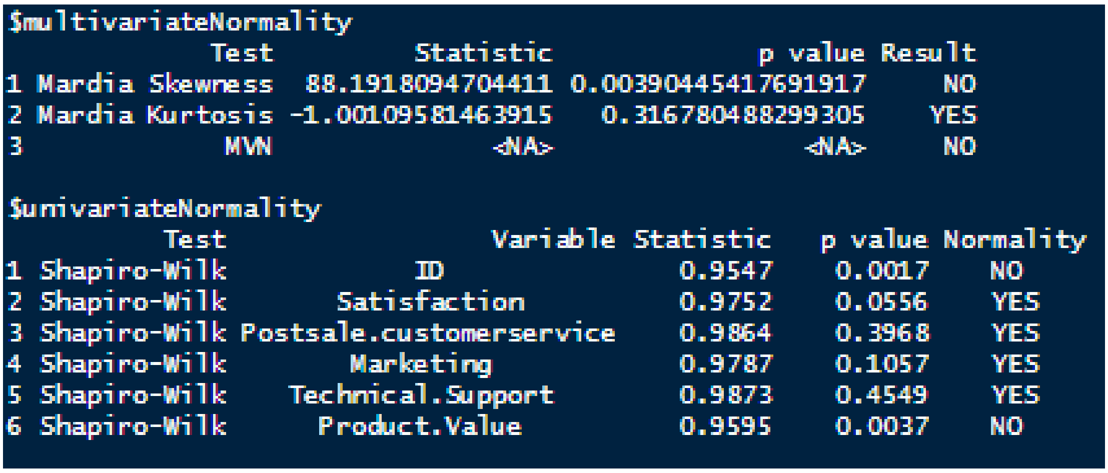
NAME OF THE FOUR FACTORS:



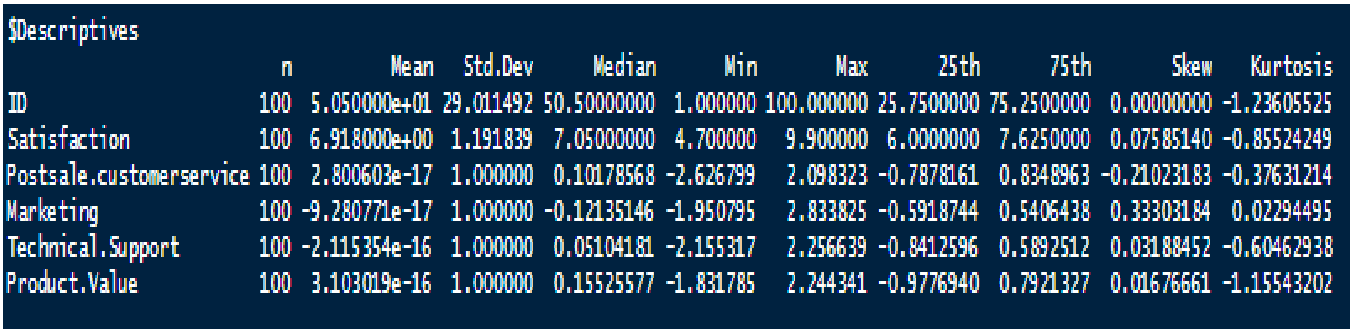
4.6 Create a data frame with a minimum of 5 columns, 4 of which are different factors and the 5th column is Customer Satisfaction

A new dataframe is prepared by selecting the factor scores from the final factor analysis performed in section 4.2 and the customer satisfaction from original data. These two separate data are merged using data.frame function (Refer code)

Also after performing multivariate test using MVN() function, we can go ahead with overall analysis. Below is the output of MVN



```
$multivariateNormality
             Test         Statistic              p value Result
1 Mardia Skewness  88.1918094704411 0.00390445417691917     NO
2 Mardia Kurtosis -1.00109581463915   0.316780488299305    YES
3             MVN              <NA>                <NA>     NO

$univariateNormality
          Test                 Variable Statistic   p value Normality
1 Shapiro-Wilk            ID                0.9547    0.0017    NO
2 Shapiro-Wilk       Satisfaction           0.9752    0.0556    YES
3 Shapiro-Wilk Postsale.customerservice     0.9864    0.3968    YES
4 Shapiro-Wilk        Marketing             0.9787    0.1057    YES
5 Shapiro-Wilk    Technical.Support         0.9873    0.4549    YES
6 Shapiro-Wilk      Product.Value           0.9595    0.0037    NO
```



```
$Descriptives
                           n          Mean  Std.Dev      Median       Min        Max       25th      75th        Skew    Kurtosis
ID                       100  5.050000e+01 29.011492 50.50000000  1.000000 100.000000 25.7500000 75.2500000  0.00000000 -1.23605525
Satisfaction             100  6.918000e+00  1.191839  7.05000000  4.700000   9.900000  6.0000000  7.6250000  0.07585140 -0.85524249
Postsale.customerservice 100  2.800603e-17  1.000000  0.10178568 -2.626799   2.098323 -0.7878161  0.8348963 -0.21023183 -0.37631214
Marketing                100 -9.280771e-17  1.000000 -0.12135146 -1.950795   2.833825 -0.5918744  0.5406438  0.33303184  0.02294495
Technical.Support        100 -2.115354e-16  1.000000  0.05104181 -2.155317   2.256639 -0.8412596  0.5892512  0.03188452 -0.60462938
Product.Value            100  3.103019e-16  1.000000  0.15525577 -1.831785   2.244341 -0.9776940  0.7921327  0.01676661 -1.15543202
```

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

Multiple linear regression is performed by taking Customer Satisfaction as dependent variable and the four factor scores generated as independent variable

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

We could see that the adjusted R square values shows 64.62% of the ---- contributing to customer satisfaction and the remaining 36.3 % is explained by other factors apart from the four independent factor scores
Below is the summary of multiple linear regression

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.91800    0.07089  97.589  < 2e-16 ***
Postsale.customerservice  0.61805    0.07125   8.675 1.12e-13 ***
Marketing                 0.50973    0.07125   7.155 1.74e-10 ***
Technical.Support         0.06714    0.07125   0.942    0.348
Product.Value             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
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Postsale Customer Service	1	37.8162171	37.8162171	75.2528601	1.118471e-13
Marketing & Reach	1	25.7230564	25.7230564	51.1879217	1.739523e-10
Technical Support	1	0.4462152	0.4462152	0.8879517	3.484231e-01
Product Value	1	28.9025221	28.9025221	57.5149399	2.244522e-11
Residuals	95	47.7395892	0.5025220	*NA*	*NA*

We could see from the P – values in the table that Post Sale customer service (Delivery Speed, Complaint resolution, Order & billing) is having the lowest P-value of 1.118471e-13 which shows that customer satisfaction is highly significant and dependent on the three variables or the above one factor.

Second lowest p –value is Product Value (Product quality, Line and Comp pricing) and the third is

Marketing and reach.

To see the relevance of the model we perform **variance inflation factor** test. For a given predictor (p), multicollinearity can assessed by computing a score called the **variance inflation factor** (or **VIF**), which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model. The smallest possible value of VIF is one (absence of multicollinearity). As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. In the output received, the VIF values are one hence its not affected by multicollinearity.

```
Postsale.customerservice                Marketing        Technical.Support 
Product.Value
                       1                        1                        1 
1
```

5 Conclusion

PCA / FA is a machine learning technique that helps us in dimension reduction. In the above example we have seen how we have reduced 12 variables into combinations of four factors and performed regression analysis and other modelling techniques to get the desired results.

6 Appendix A – Source Code

MiniProjAdvnced.R