Mini Project – Hair Factor Analysis

Numer P

Table of Contents

Sl. No.	Contents	Page No.
1	Project Objective	03
2	Assumptions	03
3	Exploratory Data Analysis – Step by step approach	03
3.1	Environment Set up and Data Import	04
3.1.1	Install necessary packages and Invoke Libraries	04
3.1.2	Set up Working Directory	04
3.1.3	Import and Read the Dataset	04
3.2	Variable Identification	05
3.2.1	Variable Identification – Inferences	06
3.3	Univariate Analysis	08
3.4	Bivariate Analysis	11
3.5	Outlier Identification	31
3.6	Variable Transformation/ Feature Creation	32
4	Conclusion	33
5	Appendix A – Source Code	34

1. Project Objective

The main objective of the report is to explore the Factor Hair Dataset ("Factor-Hair-Revised.csv") in R and generate insights about the data set. This exploration report will consist of the following,

- Importing dataset in R
- Understanding the structure of Dataset
- Graphical exploration
- Descriptive Statistics

2. Assumptions

Hair Factor Analysis is performed to understand the product quality, E-commerce facility, Technical support, complaint resolution, advertising, product line, Salesforce, competitive pricing, warranty, order and billing, delivery speed of a particular brand based on the customer satisfaction.

The analysis was to predict the satisfaction based on the regression models. It is to be predicted in the customer satisfaction is based on higher on the Product Quality.

3. Exploratory Data Analysis – Step by Step Approach

1. Environment Set up and Data Import
2. Variable Identification
3. Univariate Analysis
4. Bivariate Analysis
5. Outlier Identification
6. Variable Transformation
7. Feature Exploration

3.1 Environment Setup and Data Import

3.1.1 Install necessary packages and Import Libraries

This section is used to install packages and invoke the associated libraries. Having all packages at the same places increase code readability.

3.1.2 Setup Working Directory

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

Please refer Appendix A for source code.

3.1.3 Import and read the dataset

The given dataset is in .csv format. Hence the command ‘read.csv’ is used for import the dataset.

Please refer Appendix A for source code.

3.2 Variable Identification

R variable

- Working directory – Setwd
- library(dplyr) – datasets
- library(ggplot2) – plotting various graphs
- library(readr) – read the r code
- library(psych) – performing the correlation test
- library(lattice) – xy plot in the graphs
- library(mice) – cbind and rbind function
- library(car) – for performing regression
- library(knitr) – for the dynamic report preformation
- library(visreg) – for identify the regression models
- library(scatterplot3d) – plot the scatter plot
- library(corrplot) – plot the correlation values
- library(DataExplorer) – Explores all variables in html format
- library(nFactors) – factor loadings based on the n factors
- head – get the head values from the variables
- tail – provides the end values form the variables
- class – identifies the datasets type like data frame or in table format
- summary – summaries the variables with mean, median, min values, max values, quartile ranges.
- str – provides the structure of the dataset variables.
- dim – provides the total observations and the variables.
- Names – provides the variables names
- Boxplot – plots the quartile and outlier identification.
- cor – used for the correlation analysis
- corrplot – plot the correlation values
- lm – performs linear model regression and the multiple regression models
- eigenvalues – used to identifies the factors
- principal – performs the principal component analysis
- factor – performs the factor analysis
- fa.diagram – shows the components association with the factors.

3.2.1 Variable Identification – Inferences

- Working directory – Setwd
  - Set the working directory for the R
- library(dplyr) – datasets
  - Returns the datasets values in the exists format
- library(ggplot2) – plotting various graphs
  - plotting various graphs like histogram and boxplot
- library(readr) – read the r code
  - read the basic r commands
- library(psych) – performing the correlation test
  - activates files like correlation and regression values and KMO Test
- library(lattice) – xy plot in the graphs
  - plotting the values in the graphs
- library(mice) – cbind and rbind function
  - binding the datasets in desired formats in combine and merge.
- library(car) – for performing regression
  - corrplot and other graphs are studied
- library(knitr) – for the dynamic report preformation
  - produce the datasets and the output in dynamic reports.
- library(visreg) – for identify the regression models
  - scree plot usage in the multiple regression model
- library(scatterplot3d) – plot the scatter plot
  - plot the graphs in 3d
- library(corrplot) – plot the correlation values
  - plotting the correlated values
- library(DataExplorer) – Explores all variables in html format
  - produce the html file of the PCA and FA analysis
- library(nFactors) – factor loadings based on the n factors
  - factors used in the eigen values to rotation
- head – get the head values from the variables
  - gives the header datasets.

- tail – provides the end values form the variables
  - produces the end of the datasets values
- class – identifies the datasets type like data frame or in table format
  - the file is in data.frame
- summary – summaries the variables with mean, median, min values, max values, quartile ranges.
  - Values are produced in the following content.
- str – provides the structure of the dataset variables.
  - The variables are numeric and integers
- dim – provides the total observations and the variables.
  - 100 observations, 13 variables
- Names – provides the variables names
  - 13 Names are provided in the following content.
- Boxplot – plots the quartile and outlier identification.
  - Outlier identifies in 4 variables.
- cor – used for the correlation analysis
  - correlation are interpreted
- corrplot – plot the correlation values
  - plots are given in the following contents.
- lm – performs linear model regression and the multiple regression models
  - linear and multiple regression values are returned with 60% correlations.
- eigenvalues – used to identifies the factors
  - 4 factors are selected based on Kaiser method
- principal – performs the principal component analysis
  - PC1, PC2, PC3, PC4 are analysed in linear combination.
- factor – performs the factor analysis
  - Factor reduced to 3
- fa.diagram – shows the components association with the factors.
  - Factors are designed in the plots on Factor Analysis.

3.3 Univariate Analysis

Univariate analysis is the analysis of data of one variable at time and it involves whether the datasets are descriptive or inferential statistics.

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

```
> project2=read.csv("Factor-Hair-Revised.csv",header = T)
```

```
> head(project2,10)
```

The first 10 data are listed below from the Hair Factor.

ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
1	8.5	3.9	2.5	5.9	4.8	4.9	6	6.8	4.7	5	3.7	8.2
2	8.2	2.7	5.1	7.2	3.4	7.9	3.1	5.3	5.5	3.9	4.9	5.7
3	9.2	3.4	5.6	5.6	5.4	7.4	5.8	4.5	6.2	5.4	4.5	8.9
4	6.4	3.3	7	3.7	4.7	4.7	4.5	8.8	7	4.3	3	4.8
5	9	3.4	5.2	4.6	2.2	6	4.5	6.8	6.1	4.5	3.5	7.1
6	6.5	2.8	3.1	4.1	4	4.3	3.7	8.5	5.1	3.6	3.3	4.7
7	6.9	3.7	5	2.6	2.1	2.3	5.4	8.9	4.8	2.1	2	5.7
8	6.2	3.3	3.9	4.8	4.6	3.6	5.1	6.9	5.4	4.3	3.7	6.3
9	5.8	3.6	5.1	6.7	3.7	5.9	5.8	9.3	5.9	4.4	4.6	7
10	6.4	4.5	5.1	6.1	4.7	5.7	5.7	8.4	5.4	4.1	4.4	5.5

```
> tail(project2,5)
```

The last 5 data is shown below

ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
96	8.6	4.8	5.6	5.3	2.3	6	5.7	6.7	5.8	4.9	3.6	7.3
97	7.4	3.4	2.6	5	4.1	4.4	4.8	7.2	4.5	4.2	3.7	6.3
98	8.7	3.2	3.3	3.2	3.1	6.1	2.9	5.6	5	3.1	2.5	5.4
99	7.8	4.9	5.8	5.3	5.2	5.3	7.1	7.9	6	4.3	3.9	6.4
100	7.9	3	4.4	5.1	5.9	4.2	4.8	9.7	5.7	3.4	3.5	6.4

```
> names(project2)
```

The total variables of the dataset are displayed for the factor selection.

```
[1] "ID"          "ProdQual"    "Ecom"        "TechSup"     "CompRes"     "Advertising" "ProdLine"    "SalesFImage" "ComPricing"
[10] "WartyClaim"  "OrdBilling"  "DelSpeed"    "Satisfaction"
```

> str(project2)

'data.frame': 100 obs. of 13 variables:

$ ID : int 1 2 3 4 5 6 7 8 9 10 ...

$ 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(project2a)

ProdQual		Ecom		TechSup		CompRes		Advertising	
Min.	5	Min.	2.2	Min.	1.3	Min.	2.6	Min.	1.9
1st Qu	6.575	1st Qu	3.275	1st Qu	4.25	1st Qu	4.6	1st Qu	3.175
Median	8	Median	3.6	Median	5.4	Median	5.45	Median	4
Mean	7.81	Mean	3.672	Mean	5.365	Mean	5.442	Mean	4.01
3rd Qu	9.1	3rd Qu	3.925	3rd Qu	6.625	3rd Qu	6.325	3rd Qu	4.8
Max.	10	Max.	5.7	Max.	8.5	Max.	7.8	Max.	6.5

ProdLine		SalesFImage		ComPricing		WartyClaim		OrdBilling		DelSpeed	
Min.	2.3	Min.	2.9	Min.	3.7	Min.	4.1	Min.	2	Min.	1.6
1st Qu	4.7	1st Qu	4.5	1st Qu	5.875	1st Qu	5.4	1st Qu	3.7	1st Qu	3.4
Median	5.75	Median	4.9	Median	7.1	Median	6.1	Median	4.4	Median	3.9
Mean	5.805	Mean	5.123	Mean	6.974	Mean	6.043	Mean	4.278	Mean	3.886
3rd Qu	6.8	3rd Qu	5.8	3rd Qu	8.4	3rd Qu	6.6	3rd Qu	4.8	3rd Qu	4.425
Max.	8.4	Max.	8.2	Max.	9.9	Max.	8.1	Max.	6.7	Max.	5.5

>dim(project2a)

This shows that the observed variables after removing the dependent variable.

[1] 100 12

>any(is.na(project2))

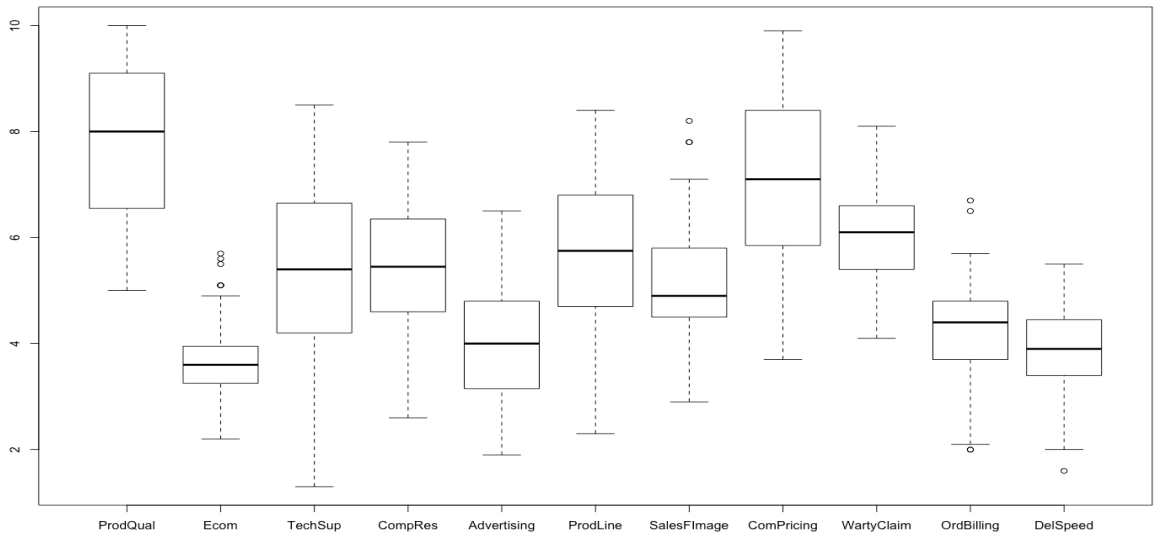
This shows that there are no missing values in the datasets.

[1] FALSE

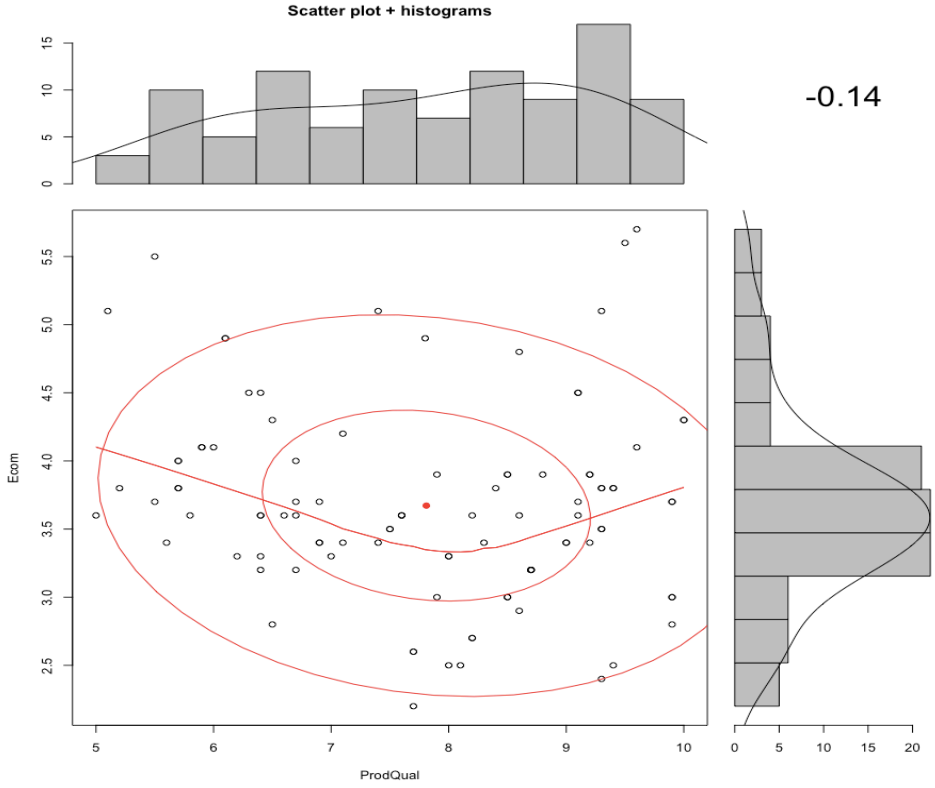
>boxplot(project2a[,1:11])

Outliers are found in the E-Commerce, Sales Images, Order and Billing, Delivery Speed.

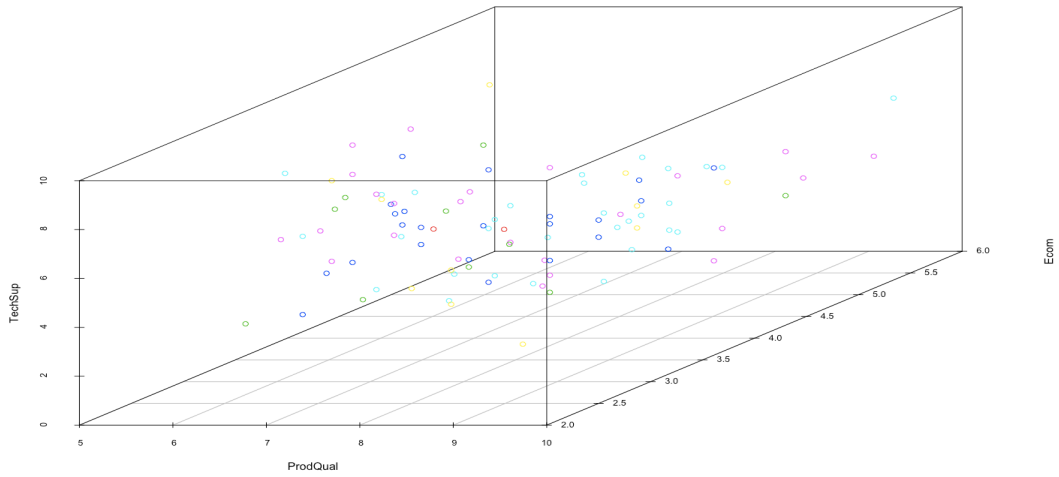
a. E-Commerce – max. 5.7
b. Sales Images – max. 8.2
c. Order and Billing – min. 2 and max. 6.7
d. Delivery Speed – min. 1.6



> scatter.hist(project2a)



> scatterplot3d(project2a)



The scatter histogram and the scatter plot 3d shows the arrangement of the data variables in the 3D formation and in circular path of datasets. This interprets the variables are arranged and to be correlated within the identified structures of the datasets.

3.4 Bivariate Analysis

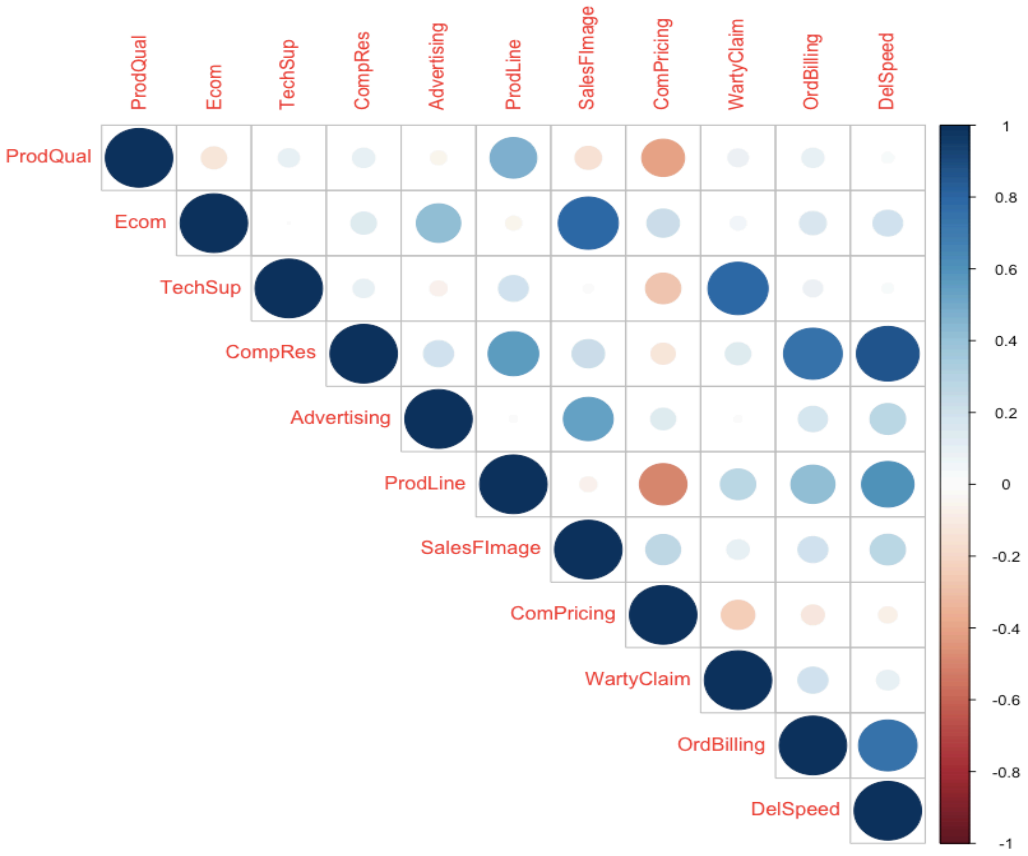
Bivariate Analysis is used to analyse the two variables and find the relationship between them. This analysis will help in identifying the association and strength of the variables. The analysis is used find the correlations, distributions and scatter plot.

2. Is there evidence of multicollinearity ? Showcase your analysis

>cor(project2a[,1:11])

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.000	-0.137	0.096	0.106	-0.053	0.477	-0.152	-0.401	0.088	0.104	0.028
Ecom	-0.137	1.000	0.001	0.140	0.430	-0.053	0.792	0.229	0.052	0.156	0.192
TechSup	0.096	0.001	1.000	0.097	-0.063	0.193	0.017	-0.271	0.797	0.080	0.025
CompRes	0.106	0.140	0.097	1.000	0.197	0.561	0.230	-0.128	0.140	0.757	0.865
Advertising	-0.053	0.430	-0.063	0.197	1.000	-0.012	0.542	0.134	0.011	0.184	0.276
ProdLine	0.477	-0.053	0.193	0.561	-0.012	1.000	-0.061	-0.495	0.273	0.424	0.602
SalesFImage	-0.152	0.792	0.017	0.230	0.542	-0.061	1.000	0.265	0.107	0.195	0.272
ComPricing	-0.401	0.229	-0.271	-0.128	0.134	-0.495	0.265	1.000	-0.245	-0.115	-0.073
WartyClaim	0.088	0.052	0.797	0.140	0.011	0.273	0.107	-0.245	1.000	0.197	0.109
OrdBilling	0.104	0.156	0.080	0.757	0.184	0.424	0.195	-0.115	0.197	1.000	0.751
DelSpeed	0.028	0.192	0.025	0.865	0.276	0.602	0.272	-0.073	0.109	0.751	1.000

>corrplot(project2a_corr, method="circle", type="upper")



- The larger circles, indicates the higher correlation among the variables. The Product quality is the highly correlated with the product line and complaint resolution.
- The higher correlation is observed in the Delivery and the complaint issues refers that more customers avail the compliance against the delivery of the products.
- The negative correlation are competitive pricing and product line and the product quality. The indications show that the prices are different based on the highly correlated advertisement and sales images and not to the quality and line of the product.
- The correlation analysis **shows the evidence of multicollinearity** is found in the given dataset and proves there is different independent variables.

3. Perform simple linear regression for the dependent variable with every independent variable

```
> reg1=lm(Satisfaction~ProdQual)
> summary(reg1)
```

Residuals

Min	1Q	Median	3Q	Max
-1.88746	-0.72711	-0.01577	0.85641	2.2522

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.67593	0.59765	6.151	1.68E-08
ProdQual	0.41512	0.07534	5.51	2.90E-07

The R squared values interpret that the Product quality is only 23% related with the customer satisfaction. Dof = 98.

```
> reg2=lm(Satisfaction~Ecom)
> summary(reg2)
```

Residuals

Min	1Q	Median	3Q	Max
-2.372	-0.78971	0.04959	0.68085	2.3458

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.1516	0.6161	8.361	4.28E-13
Ecom	0.4811	0.1649	2.92E+00	0.00437

The R squared values interpret that the E Commerce is about 79.9% related to the customer satisfaction and shows the products may be sold in online. Dof = 98.

```
> reg3=lm(Satisfaction~TechSup)
> summary(reg3)
```

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	7.82E-02	1.122	0.265

The R Squared values of Technical support is 1% relates with customer satisfaction. Dof= 98.

> reg4=lm(Satisfaction~CompRes)

> summary(reg4)

Residuals

Min	1Q	Median	3Q	Max
-2.4045	-0.66164	0.04499	0.63037	2.70949

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.68005	0.44285	8.31	5.51E-13
CompRes	5.95E-01	0.07946	7.488	3.09E-11

The complaint resolution is 36% relates the customer satisfaction with dof = 98.

> reg5=lm(Satisfaction~Advertising)

> summary(reg5)

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	3.17E+00	0.00206

Advertising is 9% related the customer satisfaction with the degrees of freedom = 98

> reg6=lm(Satisfaction~ProdLine)

> summary(reg6)

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	7.64E-02	6.529	2.95E-09

The Product line is 30% related to the customer satisfaction with dof = 98

> reg7=lm(Satisfaction~SalesFImage)

> summary(reg7)

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	2.54E-12
SalesFImage	5.56E-01	0.09722	5.72E+00	1.16E-07

The Sales Factor Images is 25% relates to the customer satisfaction with dof = 98.

> reg8=lm(Satisfaction~ComPricing)

> summary(reg8)

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	7.62E-02	-2.11E+00	0.0376

The Competitive Pricing shows that the 4% relates the customer satisfaction with the dof=98

> reg9=lm(Satisfaction~WartyClaim)

> summary(reg9)

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	2.58E-01	1.45E-01	1.786	0.0772

The Warranty claims shows that the product is 3% related to the customer satisfaction with dof=98.

> reg10=lm(Satisfaction~OrdBilling)

> summary(reg10)

Residuals

Min	1Q	Median	3Q	Max
-2.4005	-0.7071	-0.0344	0.734	2.9673

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.0541	0.484	8.377	3.96E-13
OrdBilling	6.70E-01	0.1106	6.054	2.60E-08

The Order and Billing is 66% related to the customer satisfaction and dof = 98.

> reg11=lm(Satisfaction~DelSpeed)

> summary(reg11)

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	6.99E+00	3.30E-10

The Delivery speed is highly correlated 33% with the customer satisfaction with the dof = 98.

>reg12=lm(Satisfaction~ProdQual+Ecom+TechSup+CompRes+Advertising+ProdLine+SalesFImage+ComPricing+WartyClaim+OrdBilling+DelSpeed)

> summary(reg12)

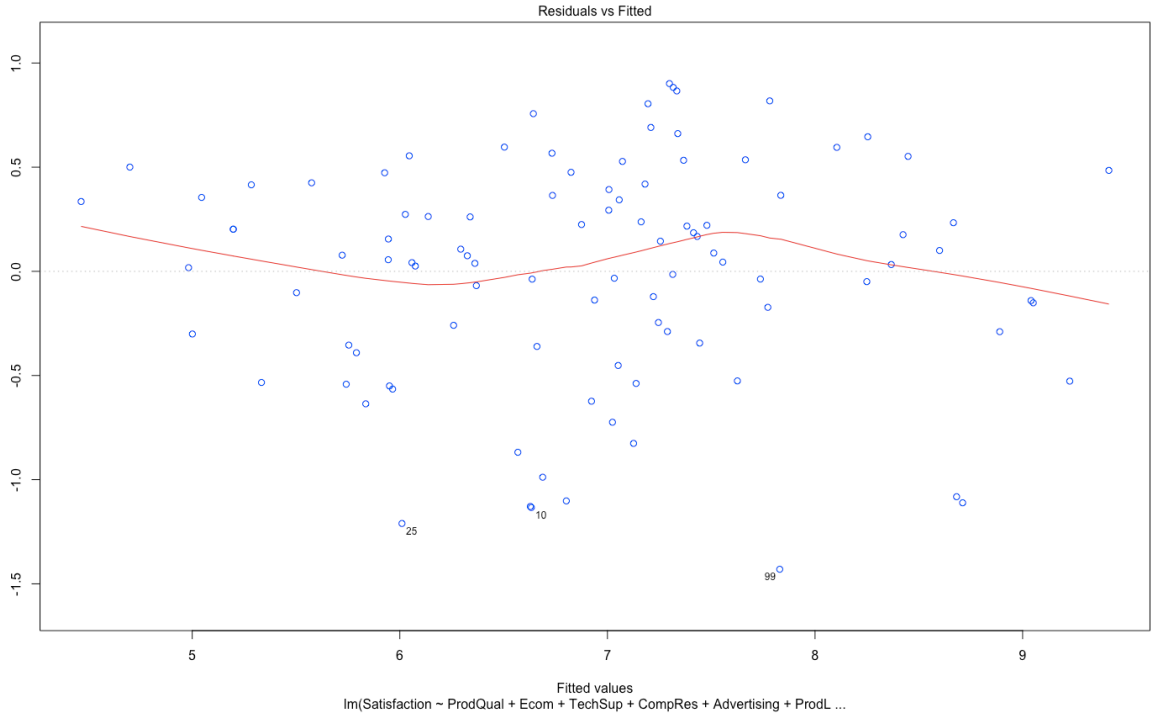
Residuals

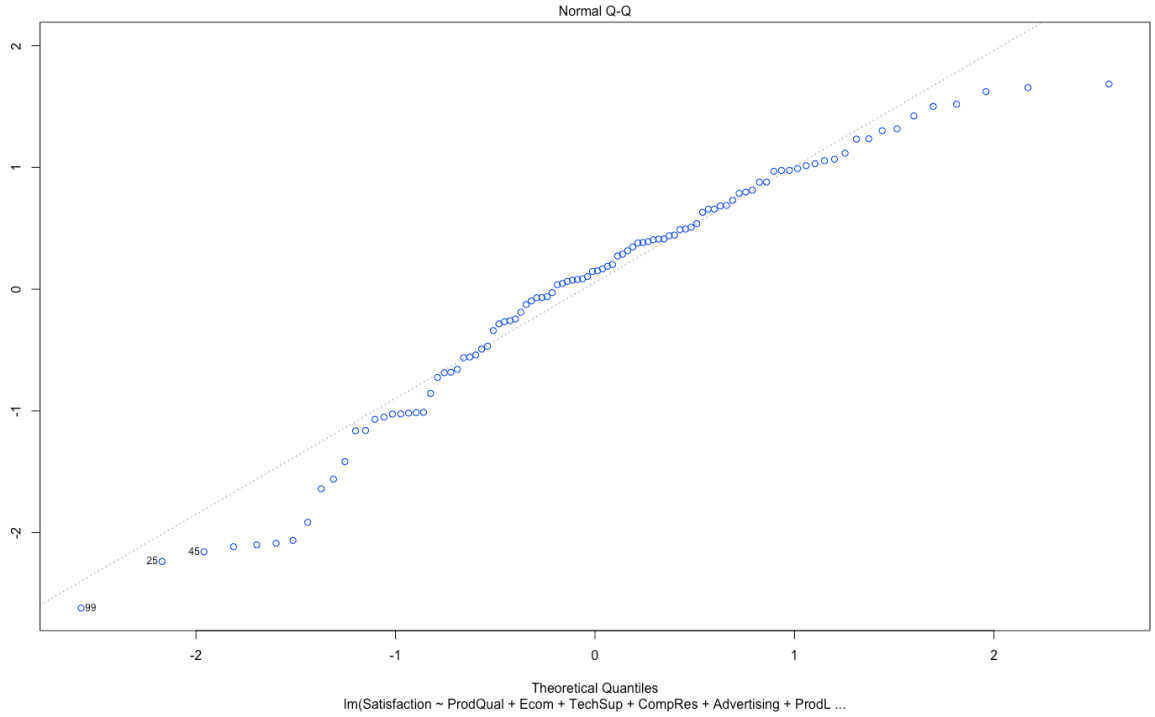
Min	1Q	Median	3Q	Max
-1.43005	-0.31165	0.07621	0.3719	0.9012

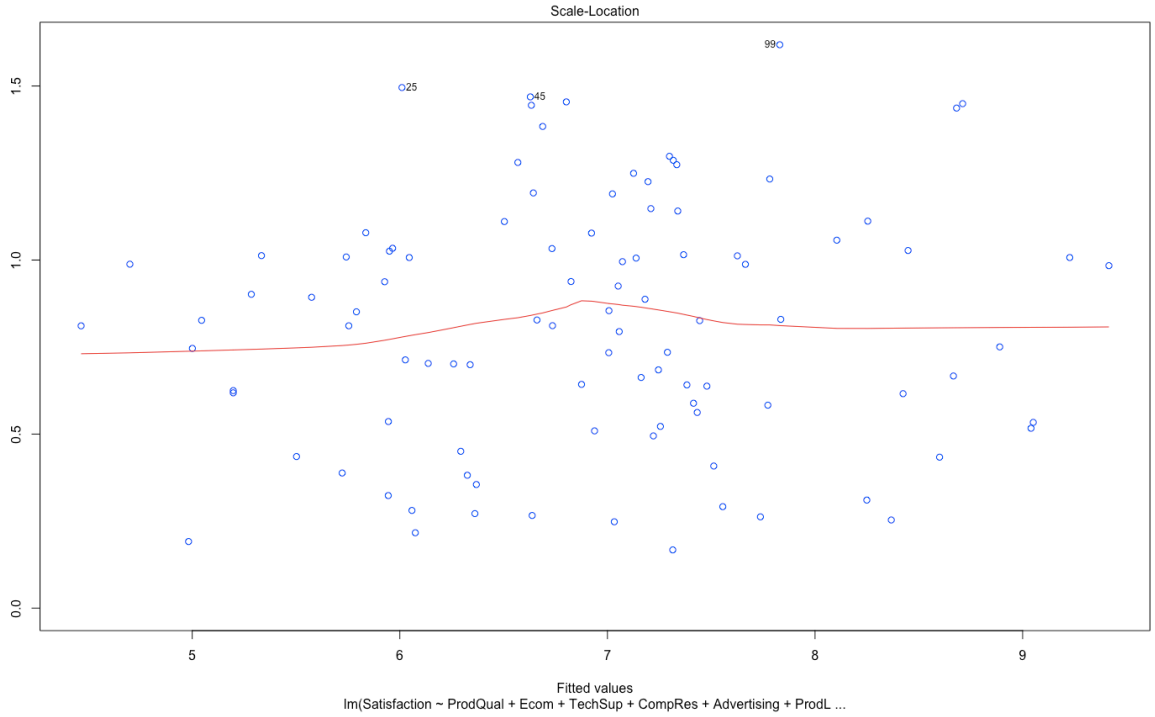
Coefficients

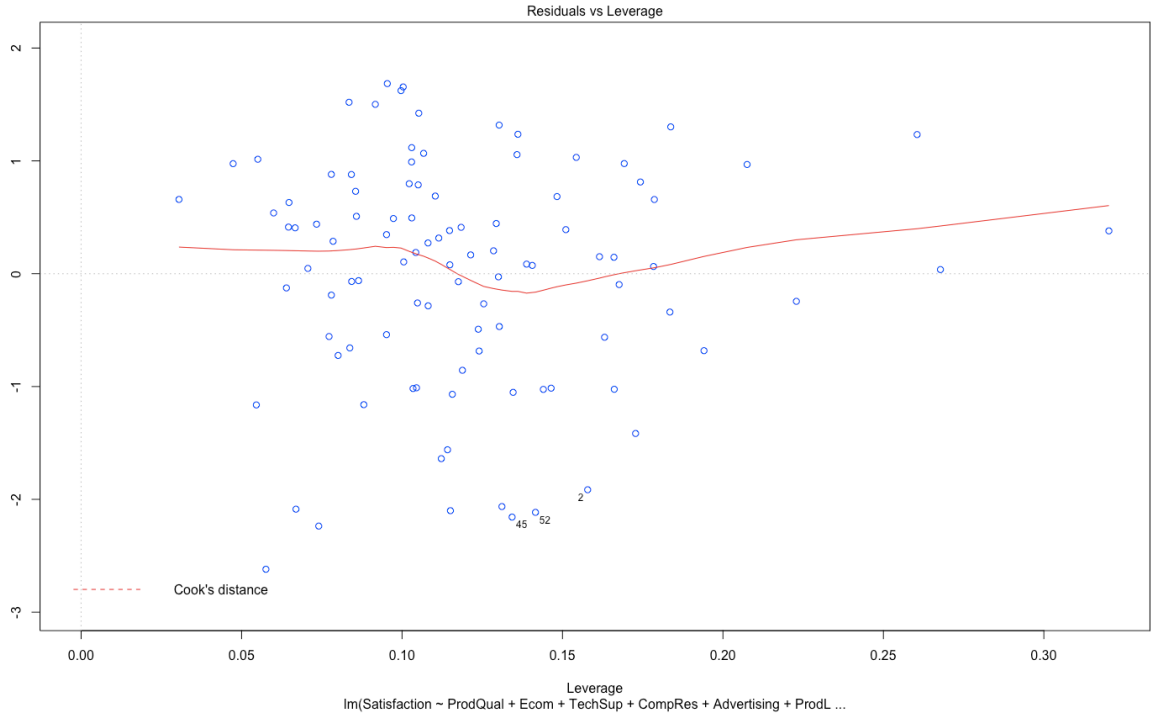
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.66961	0.81233	-0.824	4.12E-01
ProdQual	0.37137	5.18E-02	7.17E+00	2.18E-10
Ecom	-0.44056	0.13396	-3.289	0.00145
TechSup	0.03299	0.06372	0.518	0.60591
CompRes	0.16703	0.10173	1.642	0.10416
Advertising	-0.02602	0.06161	-0.422	0.67382
ProdLine	0.14034	0.08025	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	0.1233	-0.835	0.40587
OrdBilling	0.14635	0.10367	1.412	0.1616
DelSpeed	0.1657	0.19644	0.844	0.40124

The multiple regression for the independent variables is 80% correlates with the customer satisfaction with degrees of freedom is 88. The Adjusted R square value shows 77% correlation with the customer satisfaction.


Residuals vs Fitted
1.0
0.5
0.0
-0.5
-1.0
-1.5
25
10
99
5
6
7
8
9
Fitted values
lm(Satisfaction ~ ProdQual + Ecom + TechSup + CompRes + Advertising + ProdL ...


Normal Q-Q
2
1
0
-1
-2
25
45
99
-2
-1
0
1
2
Theoretical Quantiles
lm(Satisfaction ~ ProdQual + Ecom + TechSup + CompRes + Advertising + ProdL ...





The plots show that the residuals, fitted, scaled and leverage values of the variables. This interprets the values are plotted in various dimensions. The customer satisfaction is scaled thru the normal Q-Q values showing in the graphs. The highly correlated scaled values interpret the customer is satisfied with the product. The graphs identify the some of degrees to interpreted with the regressions.

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

```
> pro2=eigen(project2a_corr)
> pro2
```

This shows eigen decomposition of values and vectors and helps to find the factors that are independent to the variables.

```
> eigenvalues=pro2$values
> eigenvalues
```

[1] 3.42697133 2.55089671 1.69097648 1.08655606 0.60942409 0.55188378 0.40151815

[8] 0.24695154 0.20355327 0.13284158 0.09842702

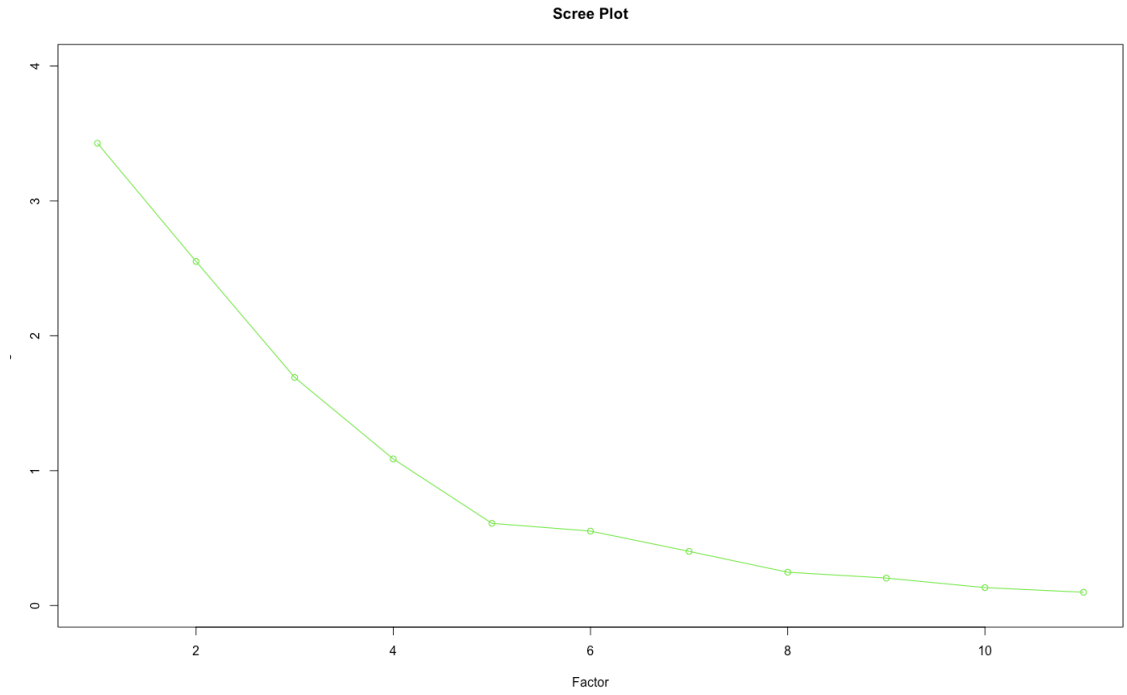
```
> eigenvectors=pro2$vectors
> eigenvectors
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]
[1,]	-0.13379	0.31350	0.06227	0.64314	0.23167	0.56457	-0.19164	0.13547	0.03133	-0.06660	-0.18279
[2,]	-0.16595	-0.44651	-0.23525	0.27238	0.42229	-0.26326	-0.05963	-0.12203	-0.54251	-0.28156	-0.06234
[3,]	-0.15769	0.23097	-0.61095	-0.19339	-0.02396	0.10877	0.01720	0.46471	-0.35930	0.38817	0.05193
[4,]	-0.47068	-0.01944	0.21035	-0.20632	0.02866	0.02815	0.00850	0.51340	0.09325	-0.53467	0.36253
[5,]	-0.18374	-0.36366	-0.08810	0.31789	-0.80387	0.20057	0.06307	-0.05348	-0.15468	-0.03716	0.08119
[6,]	-0.38677	0.28478	0.11628	0.20290	0.11667	-0.09820	0.60815	-0.33321	-0.08416	0.23480	0.38508
[7,]	-0.20367	-0.47070	-0.24134	0.22218	0.20437	-0.10497	-0.00144	0.16911	0.64490	0.35341	0.08470
[8,]	0.15169	-0.41346	0.05305	-0.33354	0.24893	0.70974	0.30825	-0.09883	-0.09414	0.04518	0.10296
[9,]	-0.21293	0.19167	-0.59856	-0.18530	-0.03293	0.13984	0.03064	-0.44354	0.31757	-0.43535	-0.12893
[10,]	-0.43722	-0.02640	0.16893	-0.23685	0.02675	0.11948	-0.65932	-0.36602	-0.09907	0.30387	0.19415
[11,]	-0.47309	-0.07305	0.23262	-0.19733	-0.03543	-0.02980	0.23424	0.06539	-0.02189	0.12010	-0.77563

```
> Factor=c(1:11)
> scree=data.frame(Factor,eigenvalues)
> plot(scree,main="Scree Plot",col="Green",ylim=c(0,4))
> lines(scree,col="green")
```

The Principal Component Analysis is performed with the factors that are independent to be analysed.

The Scree Plot values are taken based on the Kaiser Normalization Rule which defines the eigen values more than 1 is the factors to be analysed in the datasets. The eigen values are meant to be first 4 factors – **Product Quality, E-Commerce, Technical Support, Competitive Resolution.**



> unrotate=principal(project2a,nfactors = 4,rotate = "none")

> print(unrotate,digits = 3)

	PC1	PC2	PC3	PC4	h2	u2	com
ProdQual	0.319	-0.5	-0.096	0.678	0.82	0.1798	2.37
Ecom	0.334	0.704	0.308	0.217	0.749	0.2512	2.08
TechSup	0.252	-0.381	0.802	-0.198	0.89	0.1098	1.8
CompRes	0.85	0.002	-0.256	-0.309	0.883	0.1172	1.46
Advertising	0.363	0.571	0.117	0.227	0.523	0.4768	2.16
ProdLine	0.709	-0.476	-0.145	0.11	0.763	0.2372	1.91
SalesFImage	0.438	0.743	0.313	0.22	0.89	0.1095	2.24
ComPricing	-0.271	0.667	-0.067	-0.266	0.594	0.4058	1.69
WartyClaim	0.352	-0.321	0.788	-0.209	0.891	0.1088	1.92
OrdBilling	0.78	0.014	-0.202	-0.338	0.764	0.236	1.52
DelSpeed	0.849	0.087	-0.284	-0.32	0.911	0.0886	1.55
Satisfaction	0.83	0.038	-0.037	0.365	0.826	0.1739	1.38

	PC1	PC2	PC3	PC4
SS loadings	4.043	2.553	1.692	1.218
Proportion Var	0.337	0.213	0.141	0.101
Cumulative Var	0.337	0.55	0.691	0.792
Proportion Explained	0.425	0.269	0.178	0.128
Cumulative Proportion	0.425	0.694	0.872	1

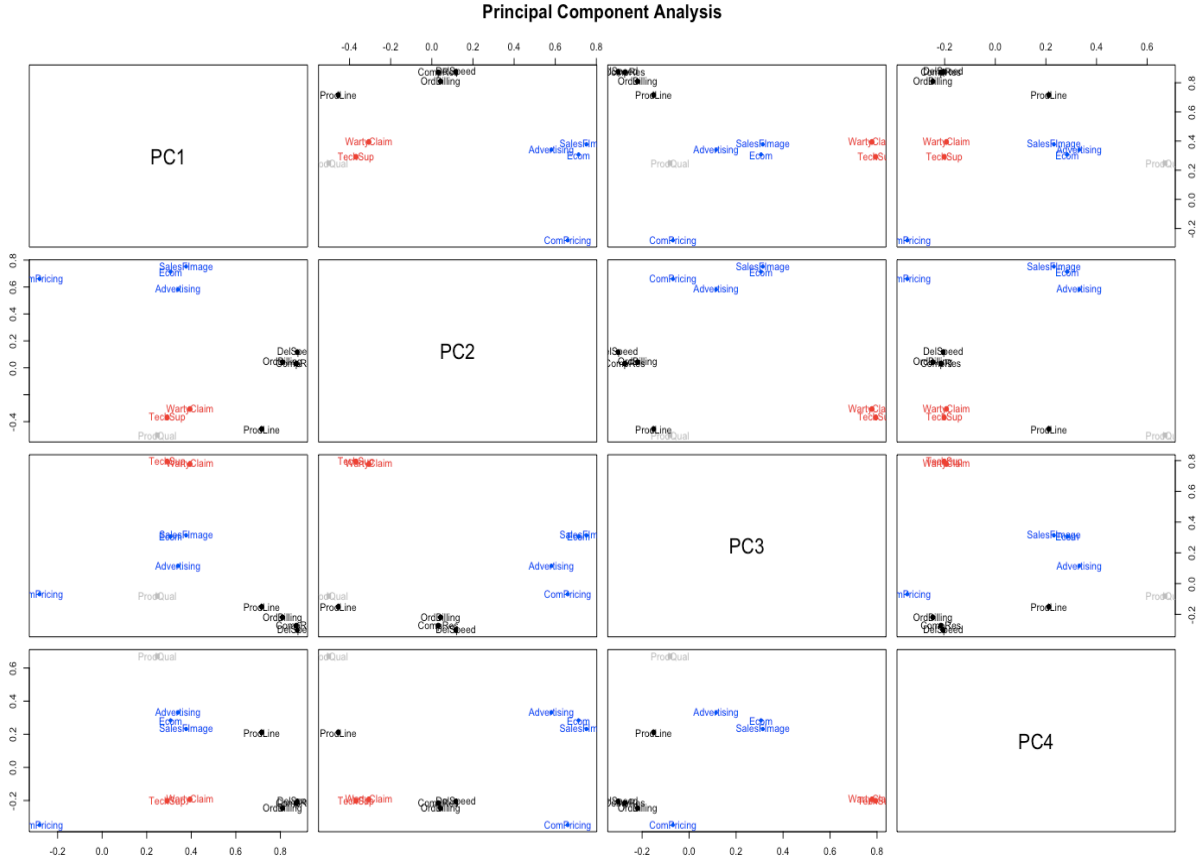
The principal component analysis shows that 97.5% is correlated with the component for the linear dependent for the factors.

The rotation showing the PC1, PC2, PC3, PC4 values are correlated in the independent variables and values are differentiated in the linear combination to promotes the products in competing with other products.

The values are analysed in the values to promotes the factors and the values are maintained in the chi square of 0.055 to shows the relationships among the variables.

Unrotated analysis is performed in the communality of the variables Proportion and the cumulative variance of 0.337 and the values are reflecting the common variance of the factors.

The components are analysed in the captured sense of the common variance produced by the PC1. Hence, the rotation is performed.



The principal component analysis shows the various plots which the components are separated from the other independent variables for the component analysis. The cluster are organised with the factors to other factors.

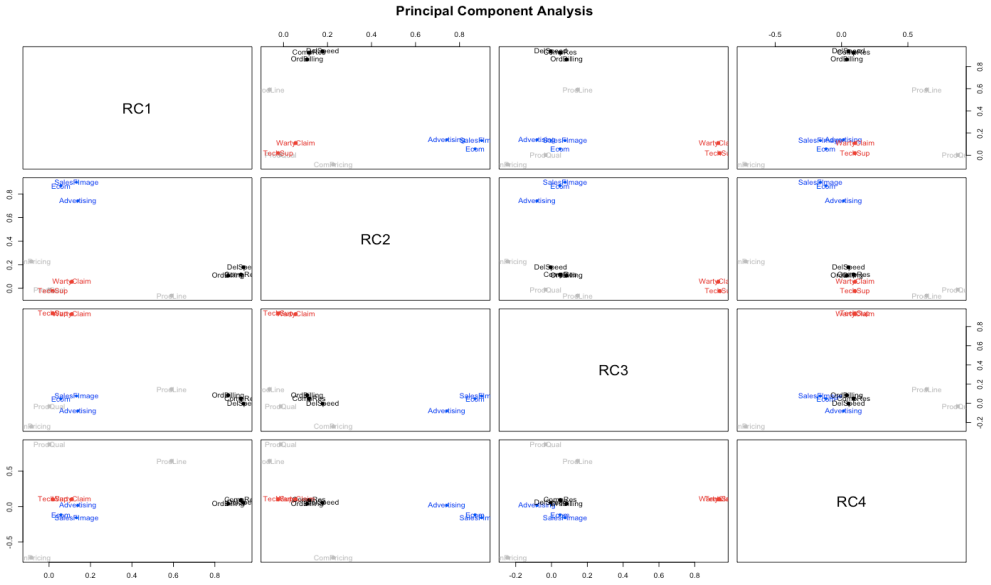
The Principal components are the Salesforce Images, Advertising, Competitive Pricing and the E-Commerce.

```
> rotate=principal(project2a,nfactors = 4,rotate = "varimax")
> print(rotate,digits = 3)
```

	RC1	RC2	RC4	RC3	h2	u2	com
ProdQual	-0.007	-0.026	0.905	-0.029	0.82	0.1798	1
Ecom	0.051	0.855	-0.116	0.049	0.749	0.2512	1.05
TechSup	0.018	-0.019	0.095	0.938	0.89	0.1098	1.02
CompRes	0.925	0.121	0.106	0.047	0.883	0.1172	1.07
Advertising	0.14	0.706	-0.011	-0.07	0.523	0.4768	1.1
ProdLine	0.588	-0.099	0.618	0.16	0.763	0.2372	2.19
SalesFImage	0.131	0.928	-0.095	0.063	0.89	0.1095	1.07
ComPricing	-0.088	0.284	-0.658	-0.271	0.594	0.4058	1.77
WartyClaim	0.109	0.056	0.094	0.931	0.891	0.1088	1.06
OrdBilling	0.862	0.109	0.047	0.083	0.764	0.236	1.06
DelSpeed	0.938	0.172	0.053	-0.003	0.911	0.0886	1.07
Satisfaction	0.522	0.479	0.568	0.04	0.826	0.1739	2.95

	RC1	RC2	RC4	RC3
SS loadings	3.155	2.47	2.012	1.87
Proportion Var	0.263	0.206	0.168	0.156
Cumulative Var	0.263	0.469	0.636	0.792
Proportion Explained	0.332	0.26	0.212	0.197
Cumulative Proportion	0.332	0.592	0.803	1

The communality variance exists the RC1 and RC2 which are highly independent in the factors to promotes the rotated cumulative variance and the values returns the 97.5% correlation with the components and organise the factors to independent the linear combination for the capture of four factors.



Factor Analysis

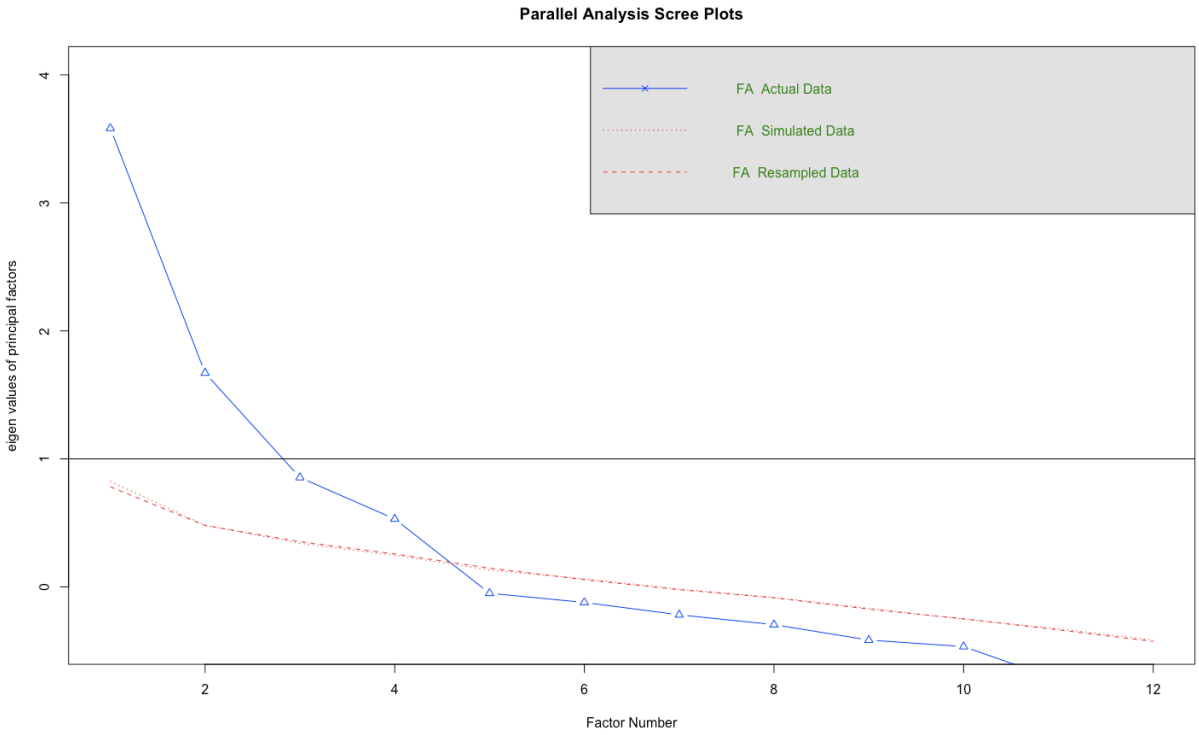
> KMO(project2a_corr)

ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
0.51	0.63	0.52	0.79	0.78	0.62	0.62	0.75	0.51	0.76	0.67

The KMO Test is performed for the suitability of the data for the factor analysis.

- The middling values are Complaint Resolution, Competitive Pricing, Order and Billing and Advertising.
- The mediocre are E-Commerce, Product Line, Salesforce Image and Delivery Speed
- The variables which are miserable are Product Quality, Technical Support, Warranty Claim

The variables are changed in the correlation for the factors on the values for the highest correlation on factor analysis.



The parallel analysis scree plot identifies the principal component analysis is varies in the factor analysis for the which lies below 1 and the factor variables are separated in the analysis to shows any regression values for the control of the factors with the particular components.

> solution1=fa(r=project2a_corr,nfactors = 4,rotate = "none",fm = "pa")

> solution1

	PA1	PA2	PA3	PA4	h2	u2	com
ProdQual	0.2	-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
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	0.8	0.2	2.3
SalesFImage	0.39	0.8	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
OrdBilling	0.75	0.02	-0.18	-0.18	0.62	0.378	1.2
DelSpeed	0.9	0.1	-0.3	-0.2	0.94	0.058	1.4

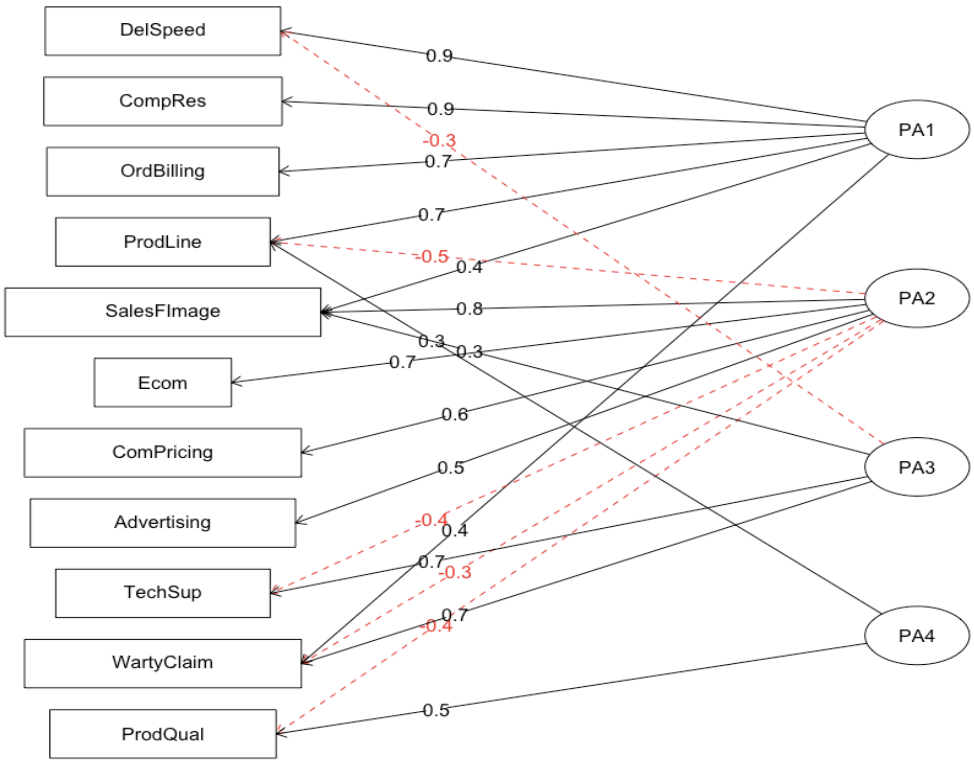
- The variables are diversely in the PA1 analysis the factors are lower from the principal analysis and performs the higher correlation in Delivery Speed and communicates the satisfaction is based on the delivery speed of the product. The factors are dependent on the independent component.

	PA1	PA2	PA3	PA4
SS loadings	3.21	2.22	1.5	0.68
Proportion Var	0.29	0.2	0.14	0.06
Cumulative Var	0.29	0.49	0.63	0.69
Proportion Explained	0.42	0.29	0.2	0.09
Cumulative Proportion	0.42	0.71	0.91	1

- The n factors are explained in 4 to get the total cumulative variance and the proportion variance which show common variance and the variance can be reduced to the simple factors for the best correlations.

	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	0.9	0.82	0.56

- The factor analysis shows highest correlation for the principal analysis factor with other independent variables to get the highest correlation form the component in the factors for the highest relations.



> solution2=fa(r=project2a_corr,nfactors = 3,rotate = "none",fm = "pa")

> solution2

The second factor analysis shows the n factors are reduced by 1 and the interpretation for independent variables for the total variance and the cumulative variance for the regression of the factors.

	PA1	PA2	PA3	h2	u2	com
ProdQual	0.18	-0.37	-0.06	0.17	0.83	1.5
Ecom	0.3	0.65	0.29	0.59	0.41	1.8
TechSup	0.27	-0.4	0.71	0.74	0.26	1.9
CompRes	0.86	0	-0.26	0.81	0.19	1.2
Advertising	0.29	0.46	0.1	0.31	0.69	1.8
ProdLine	0.65	-0.43	-0.14	0.64	0.36	1.9
SalesFImage	0.39	0.76	0.35	0.86	0.14	1.9
ComPricing	-0.22	0.54	-0.03	0.34	0.66	1.3
WartyClaim	0.38	-0.35	0.72	0.79	0.21	2
OrdBilling	0.75	0.01	-0.18	0.59	0.41	1.1
DelSpeed	0.89	0.09	-0.3	0.89	0.11	1.2

The factors are lowered by the one factor and the variables correlated. The delivery speed variable is showing 89% of component correlation for the dependent variables.

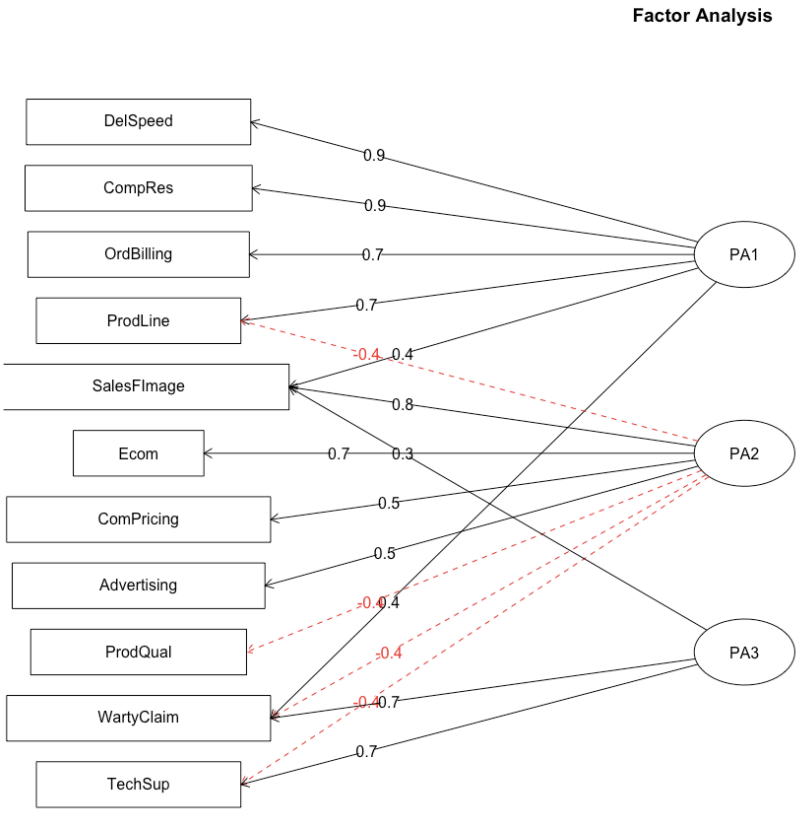
	PA1	PA2	PA3
SS loadings	3.15	2.13	1.45
Proportion Var	0.29	0.19	0.13
Cumulative Var	0.29	0.48	0.61
Proportion Explained	0.47	0.32	0.22
Cumulative Proportion	0.47	0.78	1

The cumulative and proportion variance are increased in the factors and the variable is explained by the last component analysis as 1.

	PA1	PA2	PA3
Correlation of (regression) scores with factors	0.97	0.95	0.93
Multiple R square of scores with factors	0.94	0.89	0.87
Minimum correlation of possible factor scores	0.89	0.79	0.75

The regression factors are analysed with the variables are associated with the correlation of 97% to the other variables and the variance is explained by the correlation for the possible factors.

The factor analysis plots are shown the highest correlation between factors and variables for the principal analysis.



> solution3=fa(r=project2a_corr,nfactors = 3,rotate = "varimax",fm="pa")

> solution3

	PA1	PA2	PA3	h2	u2	com
ProdQual	0.23	-0.3	0.16	0.17	0.83	2.4
Ecom	0.09	0.76	0.05	0.59	0.41	1
TechSup	0.03	-0.03	0.86	0.74	0.26	1
CompRes	0.89	0.13	0.04	0.81	0.19	1
Advertising	0.18	0.52	-0.03	0.31	0.69	1.2
ProdLine	0.7	-0.27	0.26	0.64	0.36	1.6
SalesFImage	0.14	0.91	0.09	0.86	0.14	1.1
ComPricing	-0.25	0.42	-0.32	0.34	0.66	2.6
WartyClaim	0.11	0.04	0.88	0.79	0.21	1
OrdBilling	0.75	0.14	0.07	0.59	0.41	1.1
DelSpeed	0.92	0.2	-0.02	0.89	0.11	1.1

The factor analysis of the rotation varimax show the variables are lesser correlation with compared to the other factors and the variables are correlated with the other dependent variables.

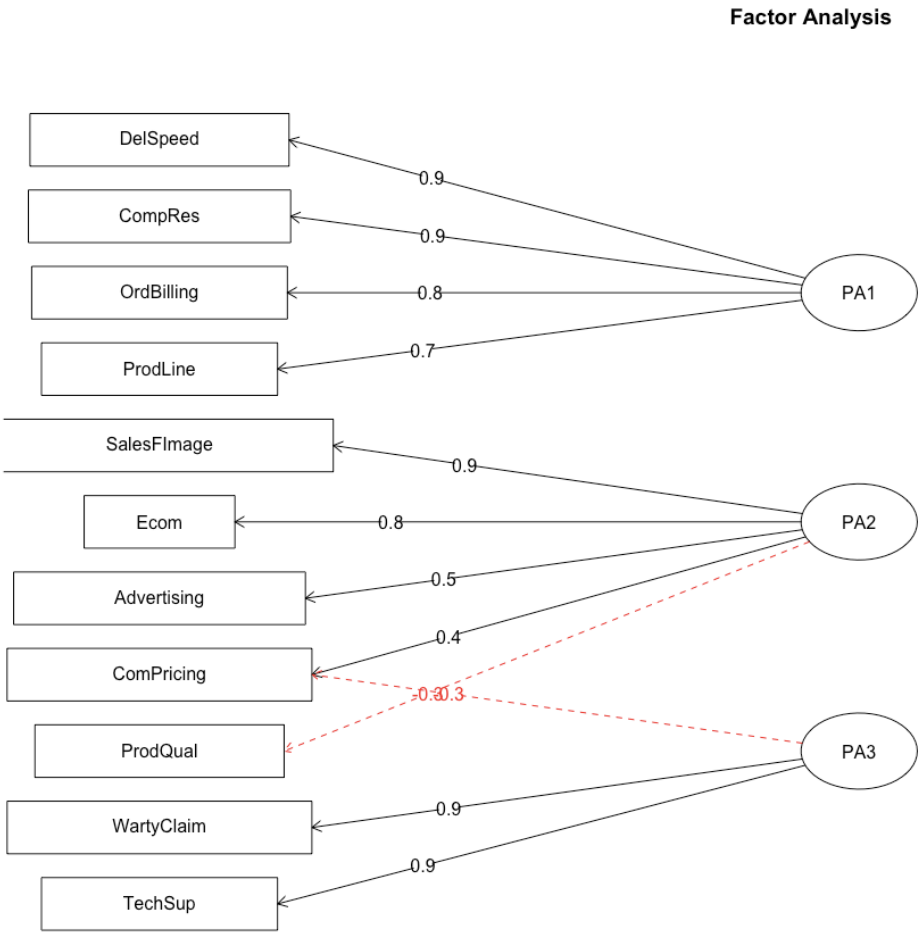
The factor analysis shows the total variance are separated to perform the general analysis in each sections for the particular variance in the cumulative and proportion for the analysis in the each factors.

	PA1	PA2	PA3
SS loadings	2.89	2.11	1.73
Proportion Var	0.26	0.19	0.16
Cumulative Var	0.26	0.45	0.61
Proportion Explained	0.43	0.31	0.26
Cumulative Proportion	0.43	0.74	1

The common variance in the PA1 and PA2 are analysed to the total factor loading for the correlation in each value.

	PA1	PA2	PA3
Correlation of (regression) scores with factors	0.97	0.95	0.94
Multiple R square of scores with factors	0.93	0.9	0.87
Minimum correlation of possible factor scores	0.87	0.81	0.75

The max. and min. correlation are continuously decreasing according the factor dependent scores and the variations are analysed with the factor for the highest possible factors for the variations in the correlation for the principal components.



The correlation are reduced in the variables for the components for the factors and the values are differed for the customer satisfaction.

> solution1$communality

ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
0.4242958	0.6381735	0.7946147	0.84281	0.314209	0.8002906	0.9792432	0.4432708	0.8135338	0.6218211	0.9420396

The common variance is shown in the factor analysed variables for the correlations between the values in each independent variable.

> solution2$communality

ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
0.1715846	0.5938321	0.742862	0.8106781	0.307625	0.636877	0.858444	0.3382443	0.7890468	0.5917181	0.8895876

The common variance is analysed and showing the values in predicting the values for the independent variables for the dependent factors.

5. Perform Multiple linear regression with customer satisfaction as dependent variables and the four factors as independent variables.

```
> model13=lm(Satisfaction~ProdQual+Ecom+TechSup+CompRes)
> summary(model13)
```

Residuals

Min	1Q	Median	3Q	Max
-1.8868	-0.4543	0.0156	0.5085	1.471

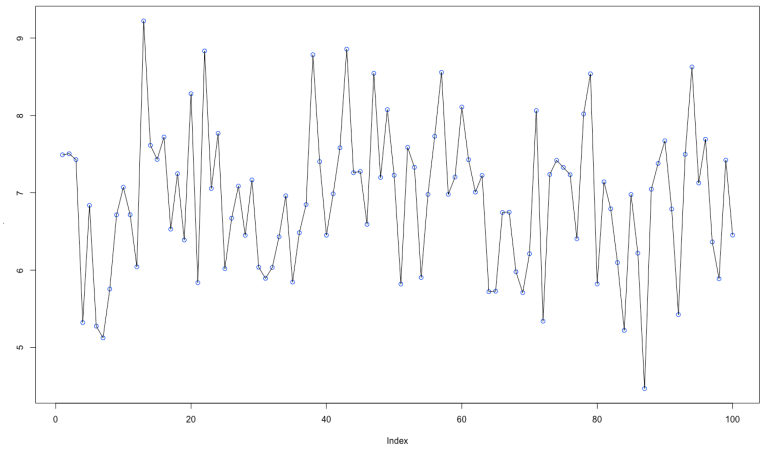
Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.74777	0.69859	-1.07	0.287
ProdQual	0.39924	0.05538	7.209	1.35E-10
Ecom	0.46777	0.11045	4.235	5.28E-05
TechSup	0.01405	0.04987	0.282	0.779
CompRes	0.50619	0.06403	7.906	4.75E-12

- The residuals/ errors in the datasets for the correlation of the independent variables selected form the basis of the quartile ranges and the maximum and minimum values.
- The quartile ranges are based on the identification of the first 25% values are about negative correlation to the dataset. The median values are about acquiring the 0.01 and the 75% of the errors are 0.5% in the dataset.
- The coefficients are stated by the p values acquire in the regression models. The p-values should be greater than 0.05 to predict the values in each factor. The product quality and the Technical support are the factors greater than p-values. The values showing in the p-values are correlated in the regression. Hence, the regression analysis is purely based on the dependent variables.
- The multiple regression on the factors are named as Product Description to produce the values mentioned in the total variance to occur the values that are predicted to the multiple ranges.
- The overall standard error is about 75% in the degrees of freedom is 95.
- The R Squared values are the factors with value 60% correlated with the satisfaction. The customer satisfaction on the Product Description is basically average when it is compared with the independent variable on Salesforce Image.
- This shows that the Salesforce Image and the delivery speed of the product enhanced the customers to be more satisfied in the brands.

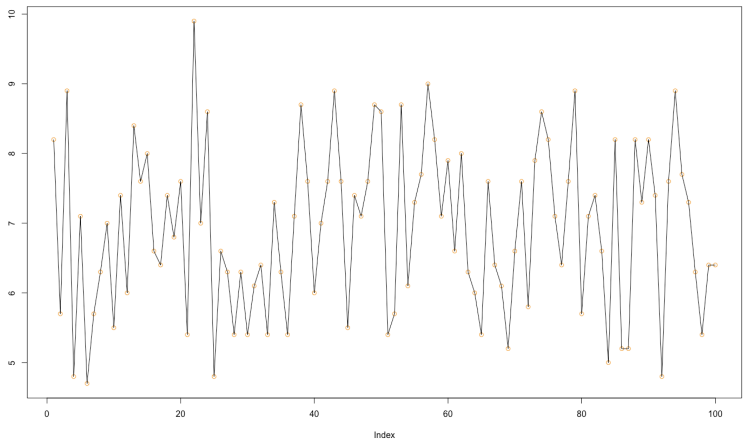
```
> prediction=predict(model13)
> plot(prediction,col = "blue")
> lines(prediction,col = "black")
```

The prediction is the multiple regression model to predict the values that are correlated in each variable and the plots are explained in the 4 factors and the distribution happens in the 100 observations of four different independent variables.




```
> actual=Satisfaction
> plot(actual,col="orange")
> lines(actual,col="black")
```

The actual plot is the dependent variable and what to be analysed based on the prediction with the variables to the products in the correlation for the factors for the values are plotted in accordance with the variables to be mentioned in the values for the highest correlation.



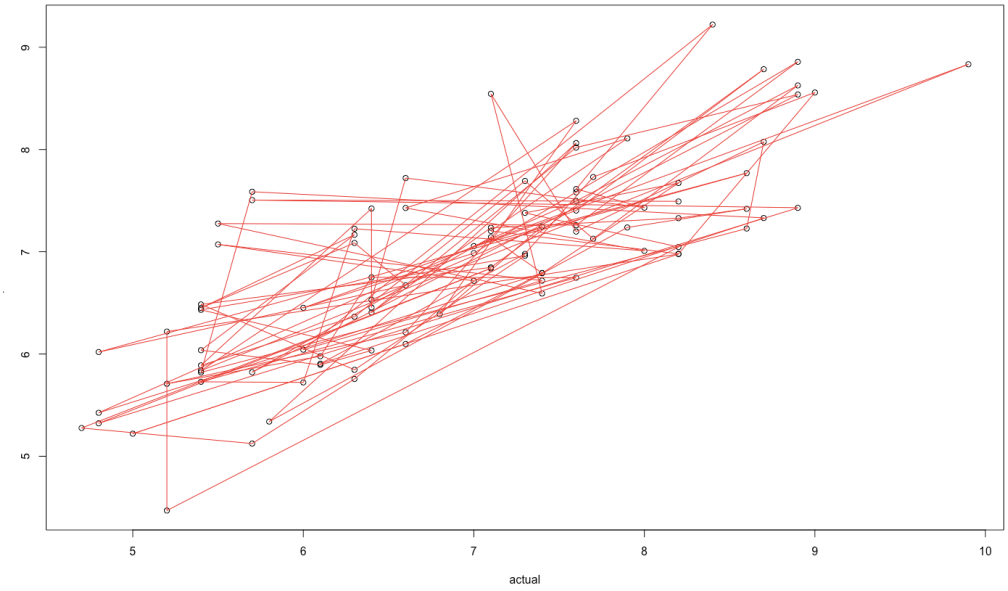
The multiple regression is allowed in the backtracking variables to check the total dimensions are correlated with each other for the multiple analysis between the variables.

> backtrack=data.frame(actual,prediction)

> backtrack

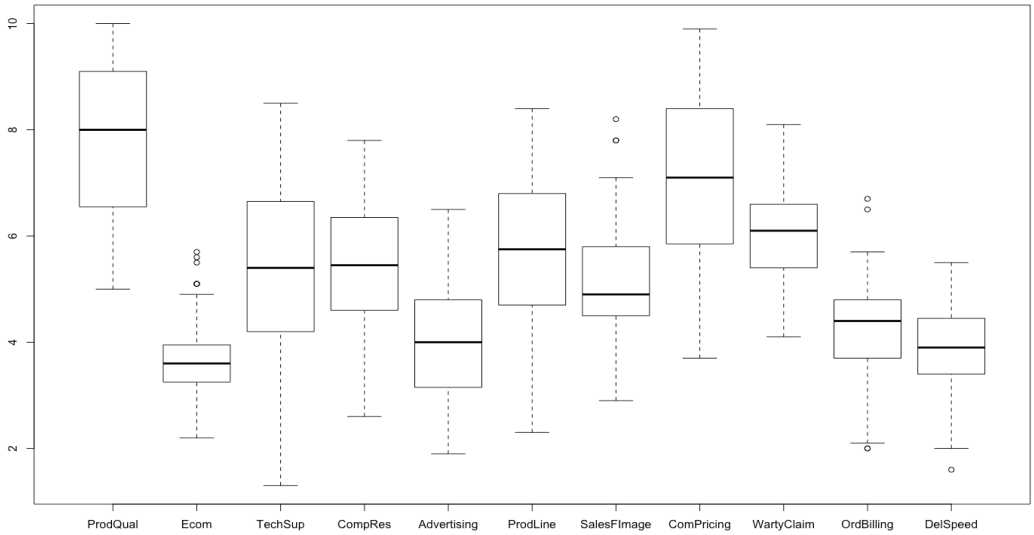
The backtrack is the values are related to the dependent and independent variables for the total regression in each value.

The variables are predicted in the accordance of the continuous plots to correlates the factors. The multiple regression shows the backtracking of variables and the values are squared to maintain at the possibility values of 60%.



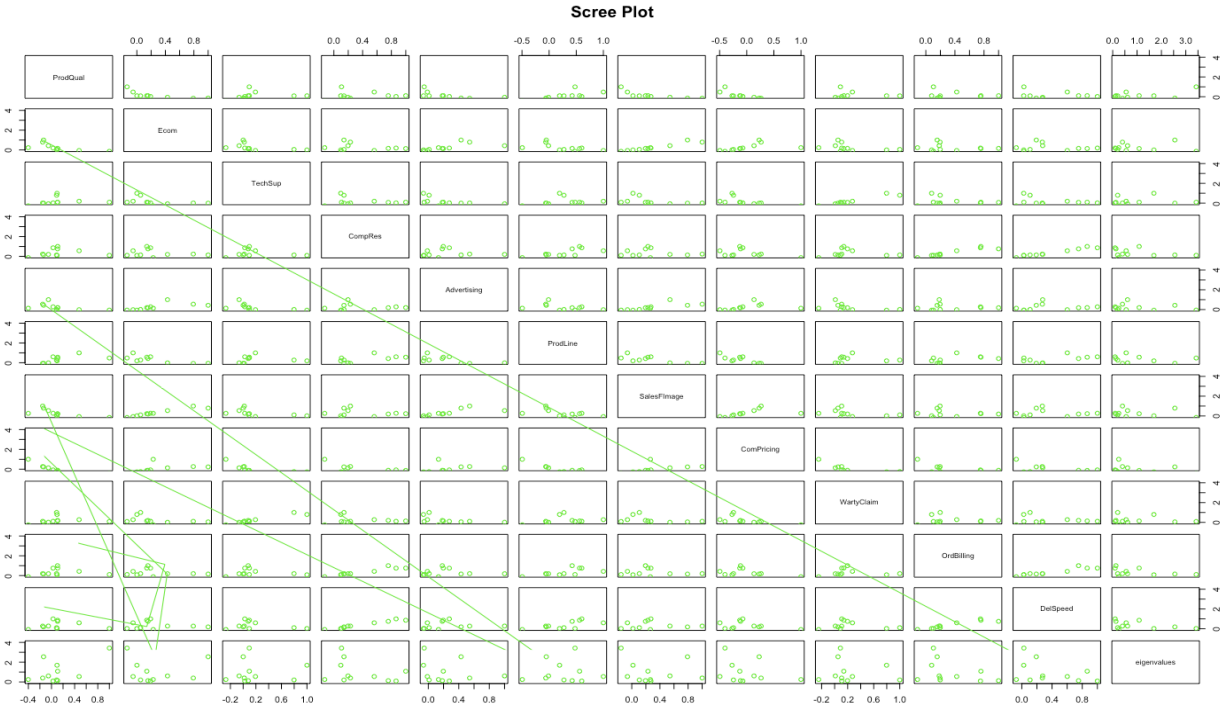
3.5 Outlier Identification

E-commerce – 5.7, Salesforce Image – 8.2, Ordered and Billing – min. 2 and max. 6.7, Delivery Speed – 1.6



3.6 Variable Transformation/ Feature creation

The values are interpreted in the principal component analysis and plotted as,



The values are variable transformed for the factors analysed in the principal component analysis. The correlation values are mainly distributed in the variance correlation of the factors below 0.4 and the variables are internally interpreted in the identification of values to the particular distribution of factors.

The variable transformation are removed the satisfaction variable to interpret the values for the independent variables and the ID variable is completely eliminated as it contains the observation values for the datasets.

The variable transformation helps in identifying the factors with regression, correlation and the PCA and FA analysis for the Hair Factor Analysis.

4. Conclusion

Hair Factor Analysis is the datasets in which the particular shampoo bottle brands scaled values are produced with the customer satisfaction range of 100 identifiers. The brands hold the customer satisfaction is based on the delivery speed and the salesforce images displayed in the advertisement is brings more attraction among the customers. The brand must be look into the higher satisfied factors to maintain the brand name. The product quality is not high possible factor for the customer satisfaction on interpreting the factor in correlation. This shows that people are more often satisfied in the pricing techniques of the brands and the values are interpreted in the customer satisfaction shows that people more often good with the pricing. Finally, the people are averagely satisfied with the product and the satisfaction has to be increased in achieve the market leader position for the brands.

5. Appendix

```
setwd("/users/numerp/documents/PGP-BABI/Module 3 Advanced Statistics/Project 2 (AS)")
getwd()
library(dplyr)
library(ggplot2)
library(readr)
library(psych)
library(lattice)
library(mice)
library(car)
library(knitr)
library(visreg)
library(scatterplot3d)
library(corrplot)
library(DataExplorer)
library(nFactors)
project2=read.csv("Factor-Hair-Revised.csv",header = T)
head(project2,10)
tail(project2,5)
names(project2)
str(project2)
summary(project2)
dim(project2)
attach(project2)
boxplot(project2)
any(is.na(project2))
project2a=project2[,c(-1)]
class(project2a)
dim(project2a)
str(project2a)
summary(project2a)
names(project2a)
boxplot(project2a[,1:11])
scatter.hist(project2a)
scatterplot3d(project2a)
cor(project2a[,1:11])
project2a_corr=cor(project2a[,1:11])
project2a_corr
corrplot(project2a_corr,method = "number",type = "upper")
corrplot(project2a_corr,method = "circle",type = "upper")
reg1=lm(Satisfaction~ProdQual)
summary(reg1)
reg2=lm(Satisfaction~Ecom)
summary(reg2)
reg3=lm(Satisfaction~TechSup)
summary(reg3)
reg4=lm(Satisfaction~CompRes)
summary(reg4)
reg5=lm(Satisfaction~Advertising)
summary(reg5)
reg6=lm(Satisfaction~ProdLine)
summary(reg6)
reg7=lm(Satisfaction~SalesFImage)
summary(reg7)
reg8=lm(Satisfaction~ComPricing)
summary(reg8)
reg9=lm(Satisfaction~WartyClaim)
summary(reg9)
reg10=lm(Satisfaction~OrdBilling)
summary(reg10)
reg11=lm(Satisfaction~DelSpeed)
```

```
summary(reg11)
reg12=lm(Satisfaction~ProdQual+Ecom+TechSup+CompRes+Advertising+ProdLine+SalesFImage+ComPricing+WartyClaim+
OrdBilling+DelSpeed)
summary(reg12)
plot(reg12,col = "blue")
prediction=predict(reg1)
plot(prediction,col = "blue")
lines(prediction,col = "black")
actual=Satisfaction
backtrack=data.frame(actual,prediction)
backtrack
plot(actual,col="orange")
lines(actual,col="black")
pro2=eigen(project2a_corr)
pro2
eigenvalues=pro2$values
eigenvalues
eigenvectors=pro2$vectors
eigenvectors
prop.var=eigenvalues/sum(eigenvalues)*100
prop.var
cumvar=cumsum(prop.var)
cumvar
Factor=c(1:11)
scree=data.frame(Factor,eigenvalues)
plot(scree,main="Scree Plot",col="Green",ylim=c(0,4))
lines(scree,col="green")
plot(eigenvalues,type = "line",xlab = "Principal Components",ylab = "Eigen Values")
unrotate=principal(project2a,nfactors = 4,rotate = "none")
print(unrotate,digits = 3)
unrotatedprofile=plot(unrotate,row.names(unrotate$loadings))
rotate=principal(project2a,nfactors = 4,rotate = "varimax")
print(rotate,digits = 3)
rotatedprofile=plot(rotate,row.names(rotate$loadings),cex=1.0)
rotate$r.scores
rotate$scores
model13=lm(Satisfaction~ProdQual+Ecom+TechSup+CompRes)
summary(model13)
pro2a=eigen(project2a_corr)
pro2a
eigenvalues=pro2a$values
eigenvalues
eigenvectors=pro2a$vectors
eigenvectors
factor.scores(project2a_corr,f = rotate$loadings,method = "Harman")
parallel=fa.parallel(project2a_corr,fm="miners",fa="fa")
solution1=fa(r=project2a_corr,nfactors = 4,rotate = "none",fm = "pa")
solution1
fa.diagram(solution1,simple = F)
solution2=fa(r=project2a_corr,nfactors = 3,rotate = "none",fm = "pa")
solution2
solution1$communality
solution2$communality
fa.diagram(solution2,simple = F)
solution3=fa(r=project2a_corr,nfactors = 3,rotate = "varimax",fm="pa")
solution3
```