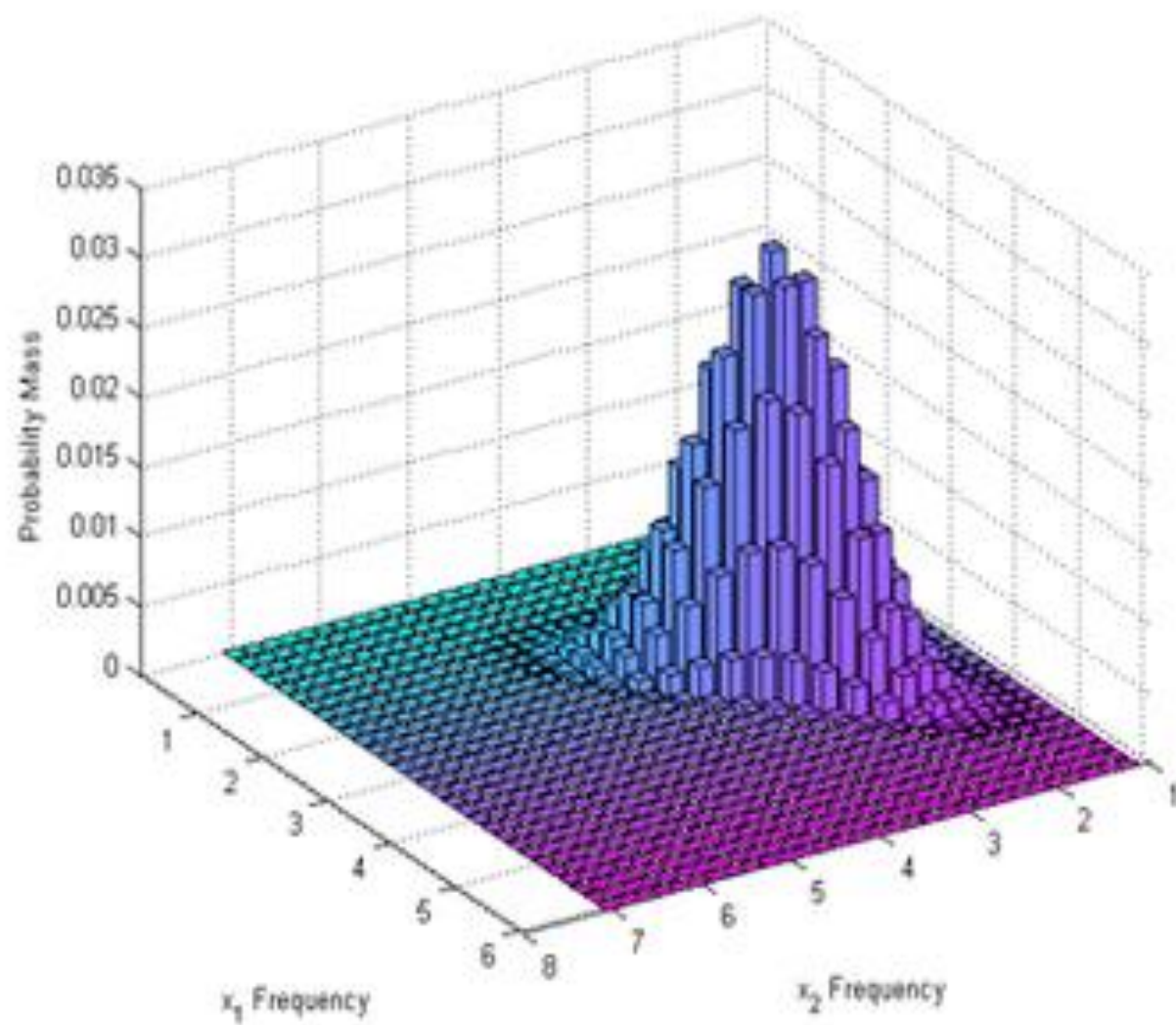


Mini Project 2 – Advance Statistics

Factor Hair PCA/Factor & Regression

Sridhar V
7/12/2020



Index

- A. Introduction.....2
 - a. Project Objective
 - b. Snippet of Data
 - i. Variable Expansion
 - c. Step by step approach
- B. Exploratory Analysis.....4
 - a. Number of Rows and Columns:
 - b. Features and their Types:
 - c. Check for Missing Values
 - d. Descriptive Statistics
 - i. Measures of Central Tendency:
 - ii. Measures of Dispersion
 - iii. Data Visualization – Histogram and Boxplot
 - iv. Bivariate analysis (Satisfaction as Y axis)
 - v. Key Observation
 - vi. Data Frame Summary
 - vii. Normality Test
 - viii. Simple linear Regression
- C. Multicollinearity.....9
 - a. Simple Linear Regression
- D. Principal Component Analysis (PCA)/Factor Analysis.....11
 - a. Scree Plot
 - b. Factor Analysis
 - i. Interpretation of the factors
 - c. Regression After Factor Analysis
- E. Conclusion.....13

1. Introduction



The data contains 12 variables used for Market Segmentation in the context of Product Service Management, we have to optimize the data, reducing the dimensionality of data and find out which aspect is the most important variable to the customer defining their satisfaction towards the product and services provided

1.1. 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.

- Exploratory data analysis on the dataset.
- Evidence for multicollinearity if any.
- Simple linear regression for the dependent variable with every independent variable.
- PCA/Factor analysis by extracting 4 factors. Interpret the output and name the Factors.
- Multiple linear regressions with customer satisfaction as dependent variables and the four factors as independent variables.

1.2. Snippet of Data

ID	Prod Qual	Ecom	Tech Sup	Comp Res	Adverti sing	Prod Line	SalesF Image	ComPri cing	WartyC laim	OrdBi lling	DelSp eed	Satisfact ion
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

1.2.1. Variable Expansion

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

In which Customer Satisfaction is the dependent variable and rest are independent Variable

1.3. Step by step approach

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

- Exploratory Data Analysis
- Descriptive Statistics
- Data Visualization
- Check for outliers and missing values
- Check Multicollinearity
- Simple Linear Regression
- PCA/FA and Interpret the Eigen values
- Creating table after minimizing the factor and running multiple linear Regression

2. Exploratory Analysis

Please refer Appendix Section 5.1 for related code.

2.1. Number of Rows and Columns:

The number of rows in the dataset is 100

The number of columns (Features) in the dataset is 13. As ID being the number suggesting the number of row in the data set we remove the variable so as in total we have 12 variables

2.2. Features and their Types:

Variable	Type	Categorical /Continuous
ProdQual	Integer	Continuous
Ecom	Integer	Continuous
TechSup	Integer	Continuous
CompRes	Integer	Continuous
Advertising	Integer	Continuous
ProdLine	Integer	Continuous
SalesFlmage	Integer	Continuous
ComPricing	Integer	Continuous
WartyClaim	Integer	Continuous
OrdBilling	Integer	Continuous
DelSpeed	Integer	Continuous
Satisfaction	Integer	Continuous

2.3. Check for Missing Values

Found with no missing Values as verified through (dfSummary)

2.4. Descriptive Statistics

2.4.1. Measures of Central Tendency:

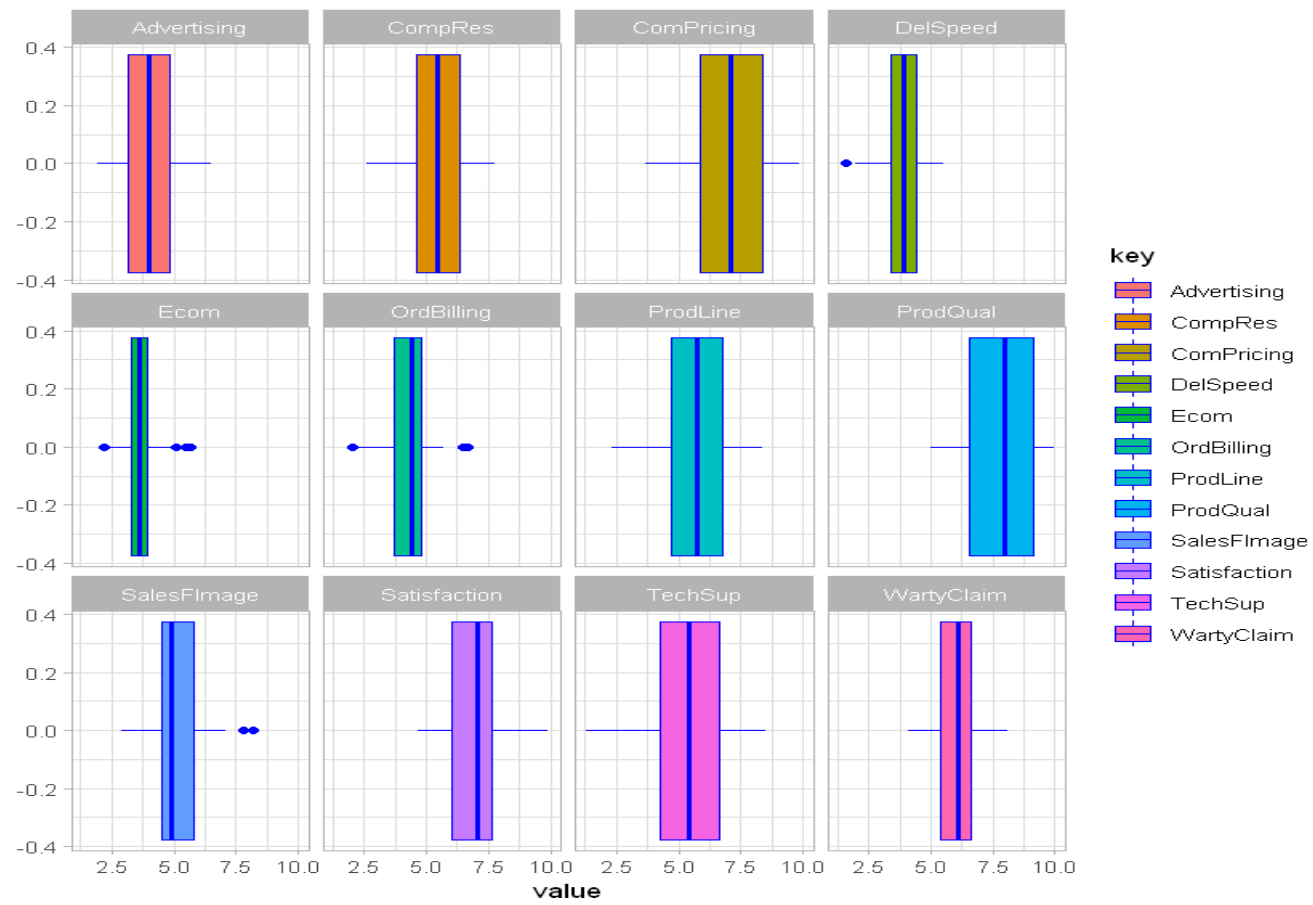
Measures of Central Tendency	Prod Qual	Ecom	Tech Sup	Com pRes	Adver tising	ProdL ine	SalesFI mage	ComP ricing	Warty Claim	OrdBill ing	DelSpe ed	Satisfacti on
Mean	7.81	3.60	5.40	5.44	4.01	5.80	5.12	6.97	6.04	4.28	3.89	6.92
Median	8.00	3.60	5.40	5.45	4.00	5.75	4.90	7.10	6.10	4.40	3.90	7.05
Mode	8.70	3.60	4.60	5.30	3.50	4.70	4.50	6.80	6.10	4.30	4.50	7.60
Min.	5.00	2.20	1.30	2.60	1.90	2.30	2.90	3.70	4.10	2.00	1.60	4.70
Max.	10.00	5.70	8.50	7.80	6.50	8.40	8.20	9.90	8.10	6.70	5.50	9.90

2.4.2. Measures of Dispersion

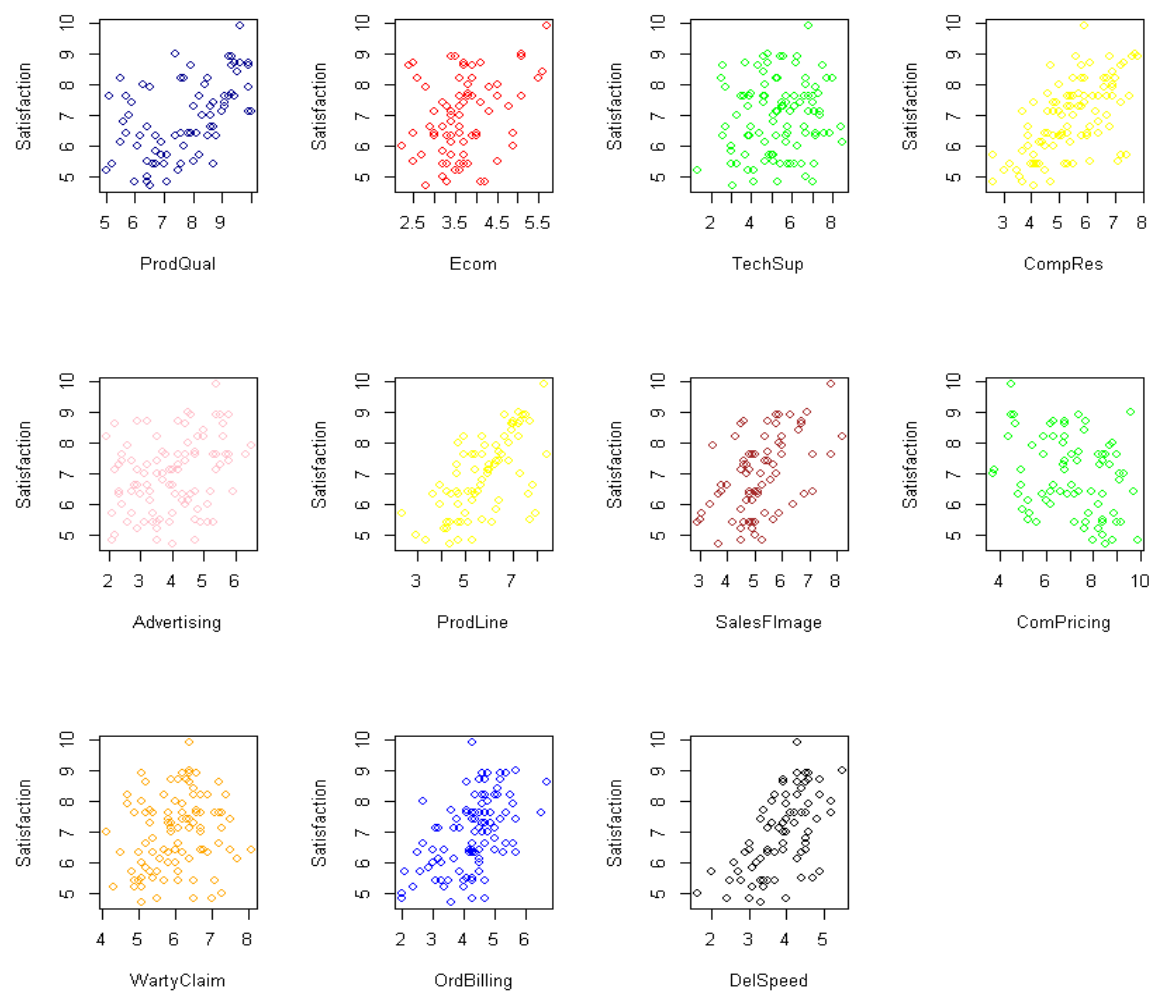
Measures of Dispersion	Prod Qual	Ecom	Tech Sup	Com pRes	Adver tising	ProdL ine	SalesFI mage	Comp Pricing	Warty Claim	OrdBill ing	DelSpe ed	Satisfacti on
Range	5.00	3.50	7.20	5.20	4.60	6.10	5.30	6.20	4.00	4.70	3.90	5.20
1st Qu.	6.58	3.20	4.25	4.60	3.17	4.70	4.50	5.88	5.40	3.70	3.40	6.00
3rd Qu.	9.10	3.90	6.62	6.32	4.80	6.80	5.80	8.40	6.60	4.80	4.42	7.62
IQR (CV)	2.5 (0.2)	0.7 (0.2)	2.4 (0.3)	1.7 (0.2)	1.6 (0.3)	2.1 (0.2)	1.3 (0.2)	2.5 (0.2)	1.2 (0.1)	1.1 (0.2)	1 (0.2)	1.6 (0.2)
Variance	1.96	0.49	2.25	1.44	1.21	1.69	1.21	2.25	0.64	0.81	0.49	1.44
Standard Deviation	1.40	0.70	1.50	1.20	1.10	1.30	1.10	1.50	0.80	0.90	0.70	1.20

2.4.3. Data Visualization – Histogram and Boxplot





2.4.4. Bivariate analysis (Satisfaction as Y axis)



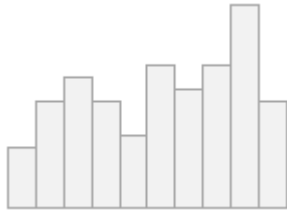
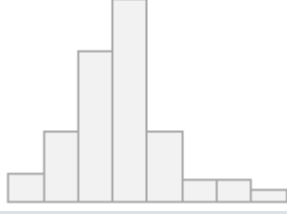
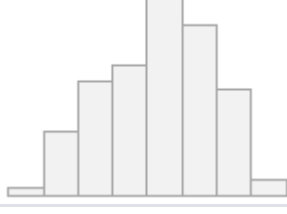
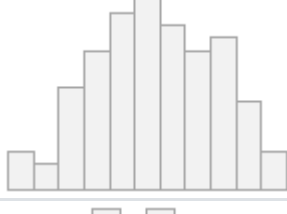
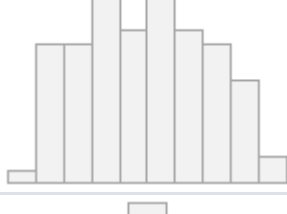
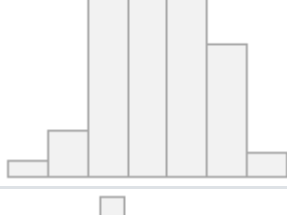
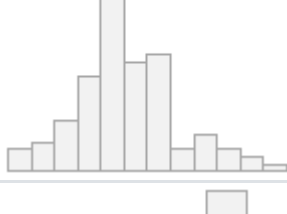
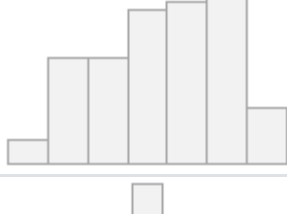
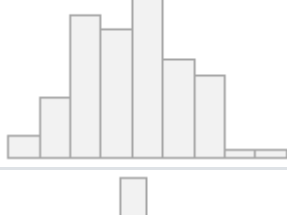
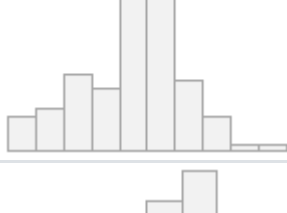
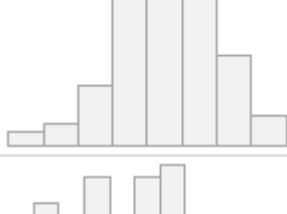
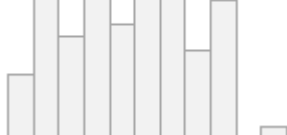
2.4.5. Key Observation

- From the boxplot we can conclude that Sales Force Image, E-commerce, Order billing, Delivery Speed have outliers.
- Product Quality rating have high variation in data distribution followed by Tech support and Competitive pricing.
- Product Quality has higher average value than other factors.

2.4.6. Data Frame Summary

Factor_Hair:

Dimensions: 100 x 13

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	SalesFIImage [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%)

Multivariate Normality

Test	Statistic	p value	Result
Mardia Skewness	461.820997074465	0.000378691957477588	NO
Mardia Kurtosis	-0.391324848847968	0.695557133670696	YES
MVN	NA	NA	NO

Univariate Normality

	Test	Variable	Statistic	p value	Normality
1	Shapiro-Wilk	ProdQual	0.9497	0.0008	NO
2	Shapiro-Wilk	Ecom	0.9585	0.0032	NO
3	Shapiro-Wilk	TechSup	0.9863	0.3900	YES
4	Shapiro-Wilk	CompRes	0.9865	0.4023	YES
5	Shapiro-Wilk	Advertising	0.9763	0.0677	YES
6	Shapiro-Wilk	ProdLine	0.9869	0.4324	YES
7	Shapiro-Wilk	SalesFImage	0.9740	0.0453	NO
8	Shapiro-Wilk	ComPricing	0.9676	0.0145	NO
9	Shapiro-Wilk	WartyClaim	0.9909	0.7404	YES
10	Shapiro-Wilk	OrdBilling	0.9741	0.0455	NO
11	Shapiro-Wilk	DelSpeed	0.9816	0.1770	YES
12	Shapiro-Wilk	Satisfaction	0.9752	0.0556	YES

Descriptives

	Skew	Kurtosis
ProdQual	-0.237215714	-1.17254306
Ecom	0.640710684	0.56725507
TechSup	-0.197201529	-0.62874790
CompRes	-0.131763526	-0.66382028
Advertising	0.042299656	-0.94496992
ProdLine	-0.089689444	-0.60412408
SalesFImage	0.365660982	0.26364425
ComPricing	-0.232782461	-0.95972556
WartyClaim	0.008120531	-0.53227071
OrdBilling	-0.323600855	0.10956856
DelSpeed	-0.449292744	0.08532059
Satisfaction	0.075851399	-0.85524249

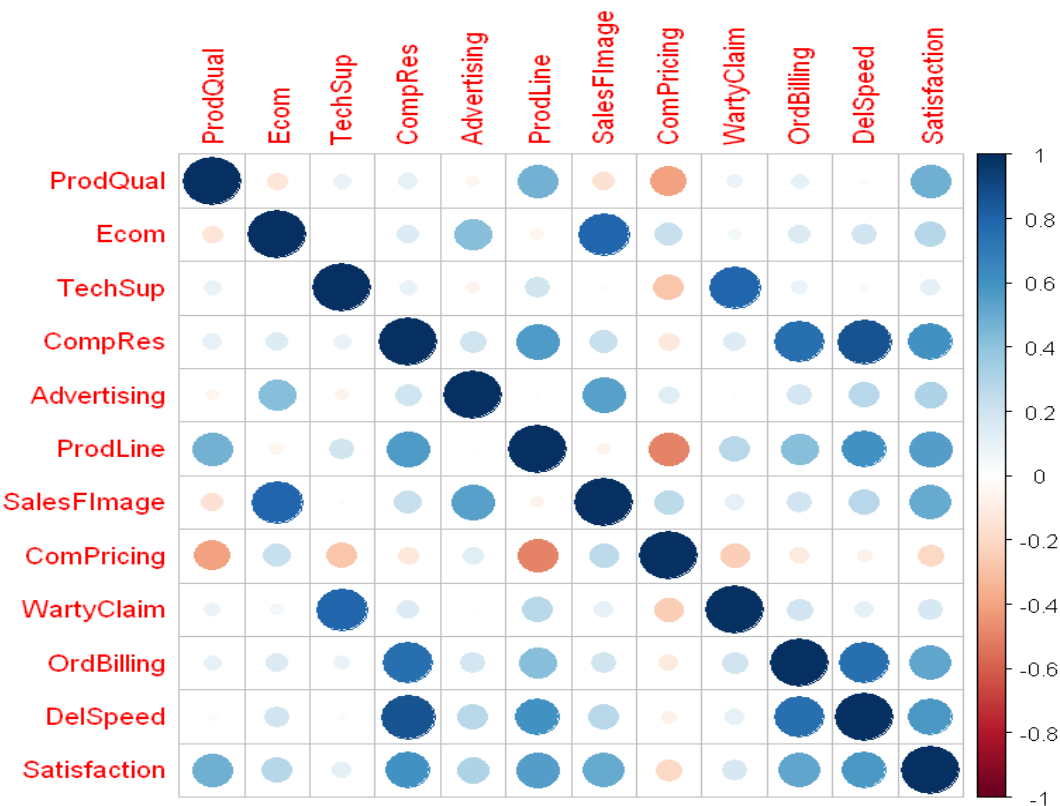
Normality Conclusion: The expected Mardia’s skewness is 0 for a multivariate normal distribution and higher values indicate a more severe departure from normality. So the Multivariate normality analysis concluded the data to be normal distributed But Univariate analysis shows that ProdQual, Ecom, SalesFImage, ComPricing, OrdBilling is not normally distributed which could be seen through there graph and Skew value. As the number of variable are not that deviated from normal distribution it is safe to move on to further test.

2.4.8. Simple linear Regression

- "Satisfaction ~ ProdQual": F-statistic: 30.36 on 1 and 98 DF, p-value: 2.901e-07
- "Satisfaction ~ Ecom": F-statistic: 8.515 on 1 and 98 DF, p-value: 0.004368
- "Satisfaction ~ TechSup": F-statistic: 1.258 on 1 and 98 DF, p-value: 0.2647
- "Satisfaction ~ CompRes": F-statistic: 56.07 on 1 and 98 DF, p-value: 3.085e-11
- "Satisfaction ~ Advertising": F-statistic: 10.03 on 1 and 98 DF, p-value: 0.002056
- "Satisfaction ~ ProdLine": F-statistic: 42.62 on 1 and 98 DF, p-value: 2.953e-09
- "Satisfaction ~ SalesFImage": F-statistic: 32.7 on 1 and 98 DF, p-value: 1.164e-07
- "Satisfaction ~ ComPricing": F-statistic: 4.445 on 1 and 98 DF, p-value: 0.03756
- "Satisfaction ~ WartyClaim": F-statistic: 3.19 on 1 and 98 DF, p-value: 0.0772
- "Satisfaction ~ OrdBilling": F-statistic: 36.65 on 1 and 98 DF, p-value: 2.602e-08
- "Satisfaction ~ DelSpeed": F-statistic: 48.92 on 1 and 98 DF, p-value: 3.3e-10

So through individual Simple linear Regression we could conclude that other than Tech Support & Warranty Claim each and every variable is significant.

3. Multicollinearity



From the above matrix we can see from the above correlation matrix:CompRes and DelSpeed,OrdBilling and Comp Res, WartyClaim and TechSupport,CompRes and OrdBilling ,OrdBilling and DelSpeed, Ecom and SalesFImage are highly correlated pairs

Now to prove the multicollinearity we would create multiple regression model

3.1. Simple Linear Regression

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

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
```

Showing only 3 variable as significant i.e. Product Quality, Ecommerce,Sales Force Image.
And the pvalue lower than 0.05 suggesting the model is significant.
VIF(Variance inflation factor) of the model which comes out to be

ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
1.63	2.75	2.97	4.73	1.50	3.48	3.43	1.63	3.19	2.90	6.51

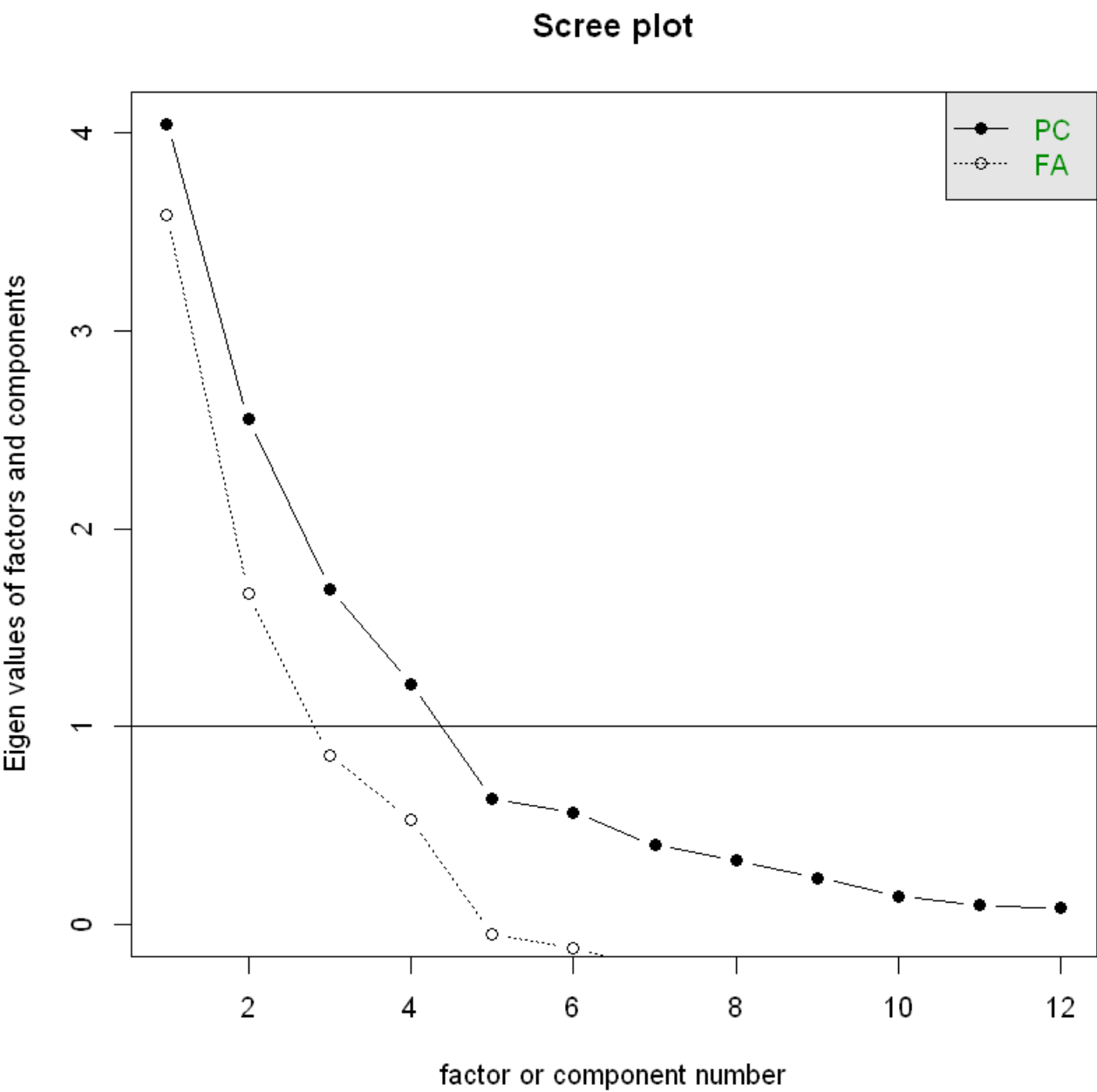
Every factor having VIF more 2 shows the presence of multicollinearity and Delspeed showing value 6.51.

4. Principal Component Analysis (PCA)/Factor Analysis

As the KMO statistic of 0.65 is also large (greater than 0.50). Hence Factor Analysis is considered as an appropriate technique for further analysis of the data.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = datamatrix)
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
```

4.1. Scree Plot



From the graph we could infer that after factor 4 there is a sharp change in the curvature of the scree plot. Which means factor 4 is accountable.

4.2. Factor Analysis

```
Parallel analysis suggests that the number of factors = 3 and the number of components = NA
Factor Analysis using method = pa
Call: fa(r = data2, nfactors = 4, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
      PA1  PA2  PA3  PA4  h2  u2 com
ProdQual 0.02 -0.07 0.02 0.65 0.42 0.576 1.0
Ecom      0.07 0.79 0.03 -0.11 0.64 0.362 1.1
TechSup   0.02 -0.03 0.88 0.12 0.79 0.205 1.0
CompRes    0.90 0.13 0.05 0.13 0.84 0.157 1.1
Advertising 0.17 0.53 -0.04 -0.06 0.31 0.686 1.2
ProdLine   0.53 -0.04 0.13 0.71 0.80 0.200 1.9
SalesFImage 0.12 0.97 0.06 -0.13 0.98 0.021 1.1
ComPricing -0.08 0.21 -0.21 -0.59 0.44 0.557 1.6
WartyClaim 0.10 0.06 0.89 0.13 0.81 0.186 1.1
OrdBilling 0.77 0.13 0.09 0.09 0.62 0.378 1.1
DelSpeed   0.95 0.19 0.00 0.09 0.94 0.058 1.1

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

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

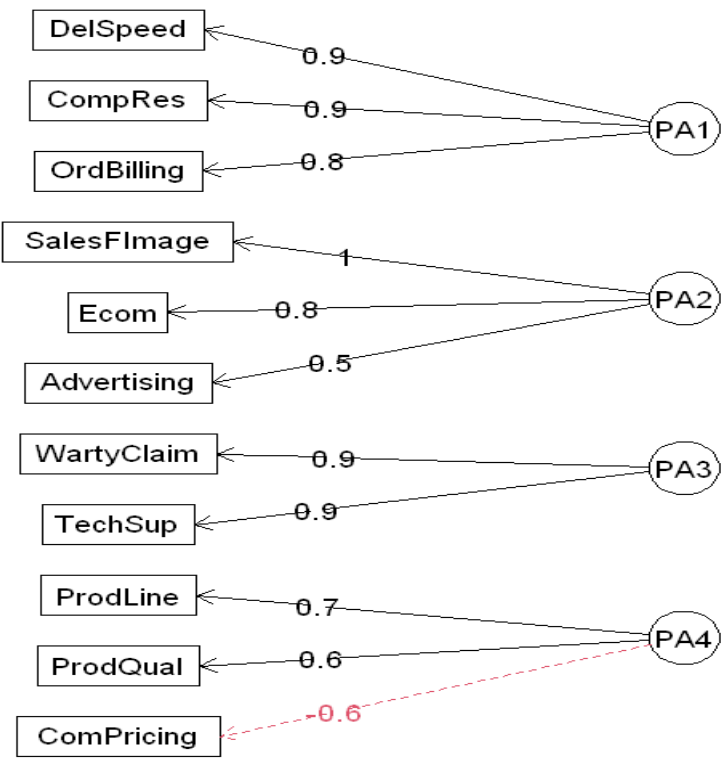
The degrees of freedom for the null model are 55 and the objective function was 6.55 with Chi Square of 619.27
The degrees of freedom for the model are 17 and the objective function was 0.33

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

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

Tucker Lewis Index of factoring reliability = 0.921
RMSEA index = 0.088 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
```

Factor Analysis



4.2.1. Interpretation of the factors

Factors	Variables	Label
PA1	DelSpeed,CompRes,OrdBilling	Purchase
PA2	SalesFImage,Ecom,Advertising	Marketing
PA3	WartyClaim,TechSup	Post Purchase
PA4	ProdLine,ProdQual,CompPricing	Product Positioning

4.3. Regression After Factor Analysis

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

Residuals:
    Min       1Q   Median       3Q      Max
-1.7125 -0.4708  0.1024  0.4158  1.3483

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.91800    0.06696  103.317  < 2e-16 ***
Purchase       0.57963    0.06857   8.453 3.32e-13 ***
Marketing      0.61978    0.06834   9.070 1.61e-14 ***
Post_purchase  0.05692    0.07173   0.794  0.429
Prod_positioning 0.61168    0.07656   7.990 3.16e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6696 on 95 degrees of freedom
Multiple R-squared:  0.6971,    Adjusted R-squared:  0.6844
F-statistic: 54.66 on 4 and 95 DF,  p-value: < 2.2e-16
```

As we could see through p value that the model is very significant and other than Post purchase variable other a ll variable are significant

4.4. Conclusion

The factors Sales, Marketing and Quality assurance are highly significant especially quality assurance and After sales service is not significant in this model. The combined factors such as Purchase ,Marketing and Product positioning plays a large role in increase in Sales.