ADVANCED STATISTICS – LINEAR REGRESSION AND FACTOR ANALYSIS

Smitayan Nandy

[COMPANY NAME] [Company address]

Contents

Data Analysis

Summary of the data

> summary(hair[-c(1)])

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

Checking for Null values in the data:

```
> sum(is.na(hair))
[1] 0
```

Result shows that data frame doesn't have any null values.

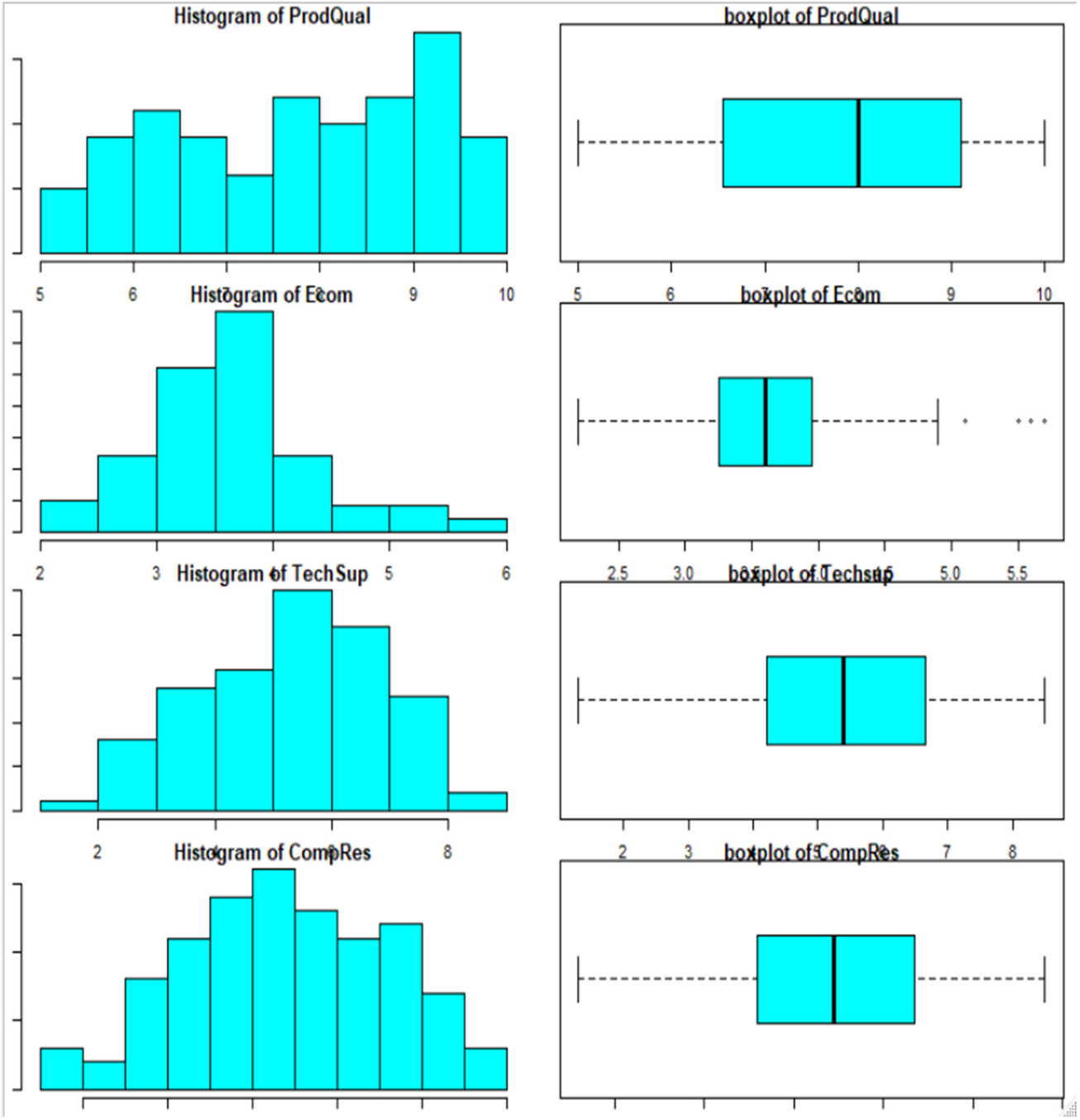
Descriptive Statistics

Test for Normality of variables:

Variable	Shapiro.pvalue
ProdQual	0.000795287
Ecom	0.003156537
TechSup	0.390038148
CompRes	0.402267237
Advertising	0.067694567
ProdLine	0.432446397
SalesFImage	0.045338957
ComPricing	0.014484137
WartyClaim	0.740365537
OrdBilling	0.045492395
DelSpeed	0.177048305
Satisfaction	0.055558217

From the above table we see that variables TechSup,Compres,ProdLine and WartyClaim are normally distributed.

Plots of variables

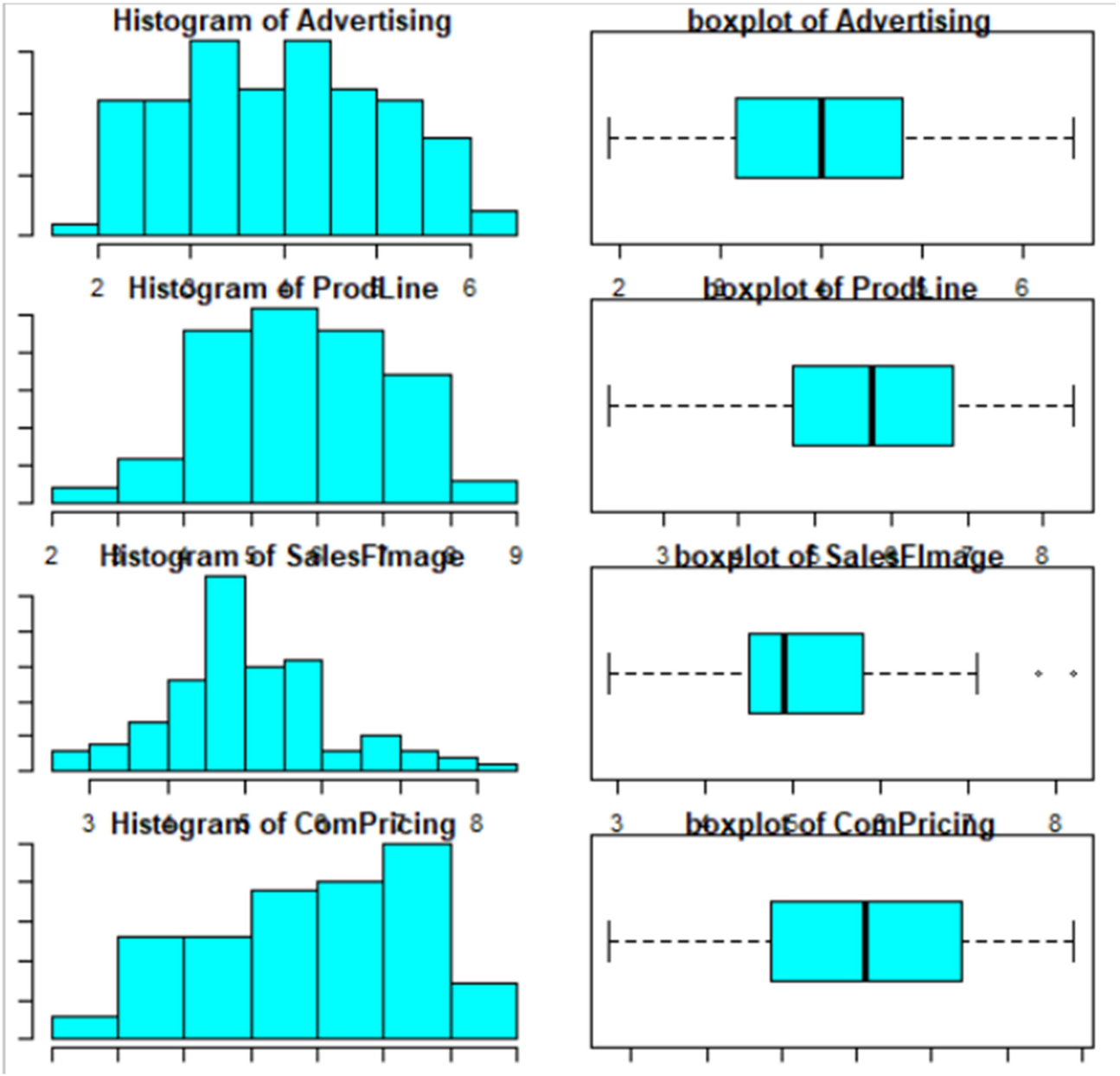


Distribution of ProdQual is left skewed and it is not a perfectly normal distribution.

Distribution of Ecom is a bit right skewed with some outliers too.

Distribution of TechSup is left skewed but it's close to a normal distribution.

Distribution of CompRes is close to a normal distribution.

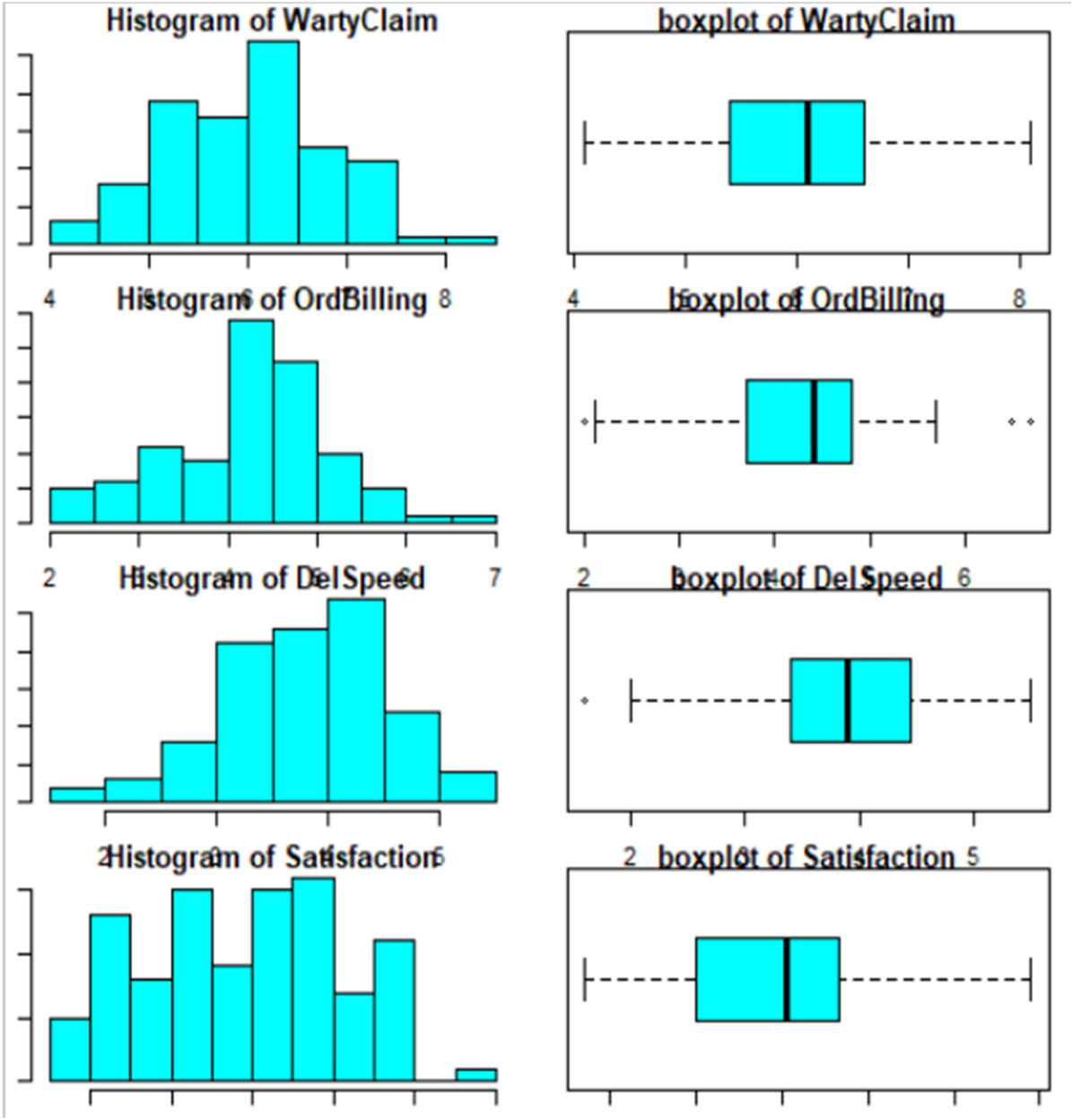


Distribution of Advertising is a bit right skewed

Distribution of ProdLine is a bit left skewed, but it's close to a normal distribution

Distribution of SalesFImage is right skewed.

Distribution of ComPricing is left skewed.



Distribution of WartyClaim is normally distributed.

Distribution of OrdBilling is left skewed.

Distribution of DelSpeed is left skewed.

Distribution of Satisfaction is left skewed.

Collinearity in Data

We see some strong positive correlation between SalesFImage and Econ, WartyClaim and TechSup, OrdBilling and CompRes, DelSpeed and CompRes.

Satisfaction has medium strength positive correlation with ProdQual, CompRes, ProdLine and DelSpeed.

There is a negative correlation between ComPricing and ProdQual, ComPricing and ProdLine.

Satisfaction is going to be our dependent variable. Since there is correlation between other independent variables, it fails the criteria of conducting a Linear Regression. We will try do reduce dimensions and see if we can derive new variables which are not correlated with each other.

Sphericity check:

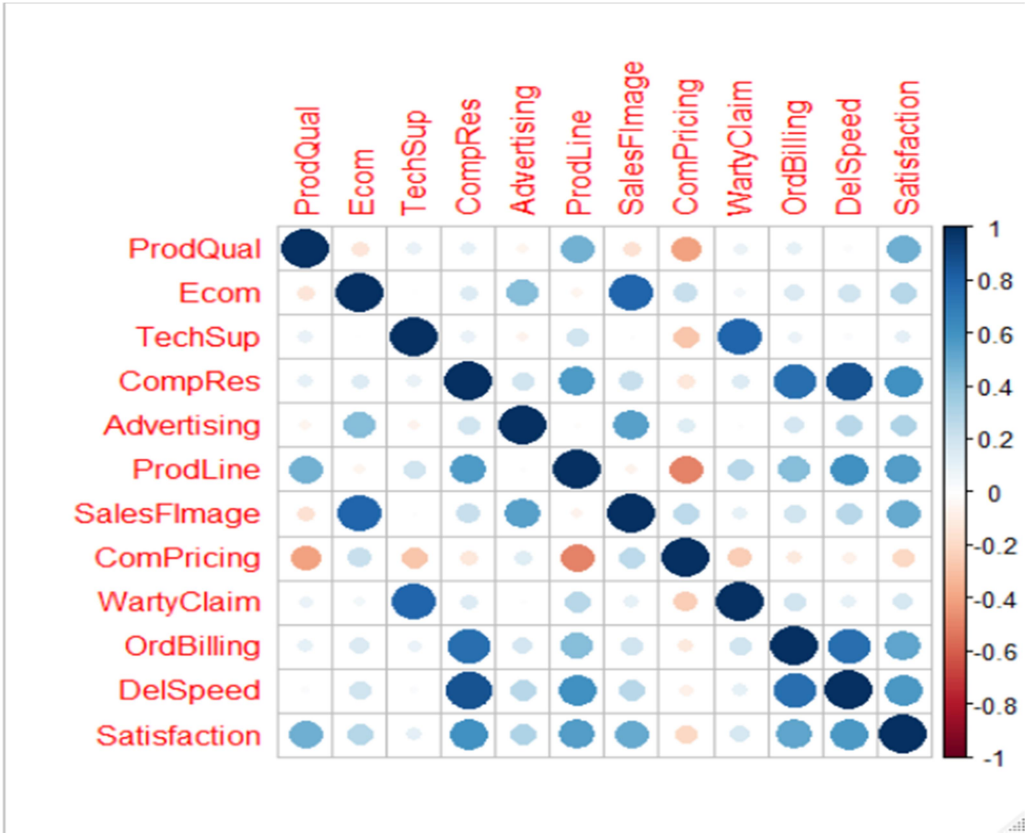
```
> cortest.bartlett(corrmatrix)
$chisq
[1] 769.6422

$p.value
[1] 1.65971e-120

$df
[1] 66
```

The p value from Bartlett sphericity check confirms that we can reject the null hypotheses and proves that there is a scope of reducing dimensions of the data.

Summary from Corrplot of the data

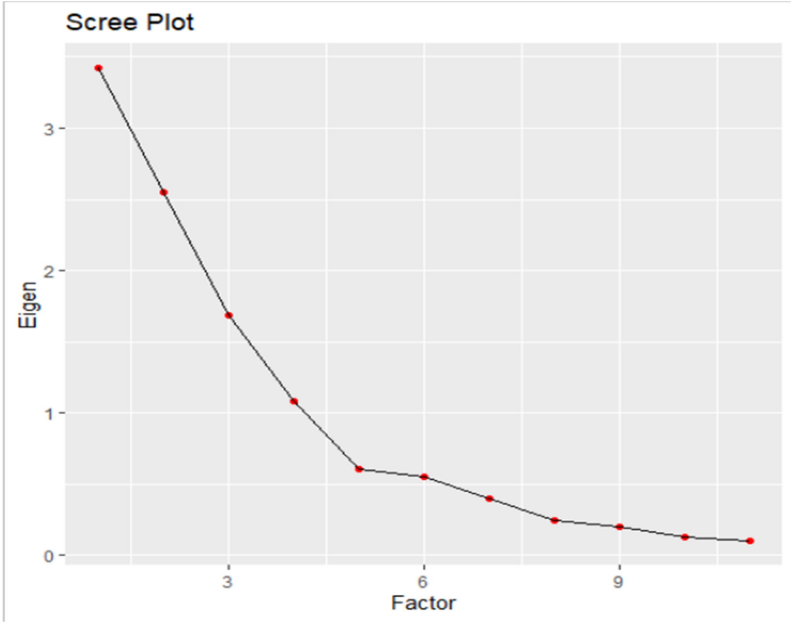


KMO test for Factor Analysis adequacy, shows an overall MSA for 0.65, which means that data is moderately adequate for Factor Analysis.

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

Factor Analysis

Scree Plot done from the independent variables is shown on the right. Using the Kaizer rule(Eigen >1), we can have four factors extracted.



Code:

```
indvariables = hair[-c(1,13)]

ev= eigen(cor(indvariables))

Eigen = ev$values

Factor = c(1:length(indvariables))

Scree = data.frame(Factor,Eigen)

ggplot(data =
Scree)+aes(x=Factor,y=Eigen)+geom_point(col='red')+geom_line()+labs(title='Scree Plot')
```

Unrotated Factor Analysis:

```
Unrotate = fa(indvariables,nfactors = 4,rotate = 'none')

print(Unrotate,digits = 5)
```

```
Factor Analysis using method =  minres
Call: fa(r = indvariables, nfactors = 4, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
                MR1      MR2      MR3      MR4      h2         u2    com
ProdQual    0.20082 -0.40771 -0.05679  0.46290 0.42406 0.5759427 2.3918
Ecom        0.29005  0.65569  0.26653  0.21036 0.62935 0.3706493 1.9911
TechSup     0.27826 -0.38188  0.74424 -0.16946 0.80587 0.1941269 1.9393
CompRes     0.86216  0.00968 -0.25540 -0.18399 0.84249 0.1575113 1.2722
Advertising 0.28652  0.45628  0.08112  0.12710 0.31302 0.6869811 1.9446
ProdLine    0.68898 -0.45429 -0.14143  0.31551 0.80063 0.1993720 2.3038
SalesFImage 0.39781  0.80674  0.34795  0.25481 0.99508 0.0049185 2.1181
ComPricing -0.23098  0.55268 -0.04596 -0.28677 0.44315 0.5568456 1.9082
WartyClaim  0.37770 -0.32211  0.73030 -0.15096 0.80254 0.1974622 2.0377
OrdBilling  0.74689  0.01909 -0.17580 -0.18112 0.62192 0.3780843 1.2349
DelSpeed    0.89516  0.09607 -0.30405 -0.19807 0.94222 0.0577812 1.3611

                       MR1     MR2     MR3     MR4
SS loadings            3.21540 2.22629 1.49984 0.67879
Proportion Var         0.29231 0.20239 0.13635 0.06171
Cumulative Var         0.29231 0.49470 0.63105 0.69276
Proportion Explained   0.42195 0.29215 0.19682 0.08908
Cumulative Proportion 0.42195 0.71410 0.91092 1.00000
```

From the loadings of unrotated factor analysis, we see that some of the variables are not clearly aligned to a particular dimension. Hence we will try rotation and check the loadings again.

VariRotate = fa(indvariables,nfactors = 4,rotate = 'varimax')

print(VariRotate,digits = 5)

```
Factor Analysis using method =  minres
Call: fa(r = indvariables, nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                MR1      MR2      MR3      MR4      h2        u2    com
ProdQual    0.02398 -0.07019  0.01571  0.64677 0.42406 0.5759427 1.0275
Ecom        0.06892  0.78147  0.02805 -0.11455 0.62935 0.3706493 1.0615
TechSup     0.01955 -0.02566  0.88968  0.11537 0.80587 0.1941269 1.0363
CompRes     0.89743  0.12973  0.05382  0.13182 0.84249 0.1575113 1.0933
Advertising 0.16636  0.52876 -0.04288 -0.06256 0.31302 0.6869811 1.2410
ProdLine    0.52543 -0.03528  0.12718  0.71214 0.80063 0.1993720 1.9211
SalesFImage 0.11360  0.98007  0.06365 -0.13261 0.99508 0.0049185 1.0726
ComPricing  -0.07557  0.21276 -0.20895 -0.59035 0.44315 0.5568456 1.5654
WartyClaim  0.10262  0.05671  0.87870  0.12916 0.80254 0.1974622 1.0797
OrdBilling  0.76827  0.12661  0.08811  0.08879 0.62192 0.3780843 1.1090
DelSpeed    0.94884  0.18513 -0.00471  0.08734 0.94222 0.0577812 1.0936

                       MR1     MR2     MR3     MR4
SS loadings            2.63452 1.97327 1.64107 1.37146
Proportion Var         0.23950 0.17939 0.14919 0.12468
Cumulative Var         0.23950 0.41889 0.56808 0.69276
Proportion Explained   0.34572 0.25895 0.21535 0.17997
Cumulative Proportion 0.34572 0.60467 0.82003 1.00000
```

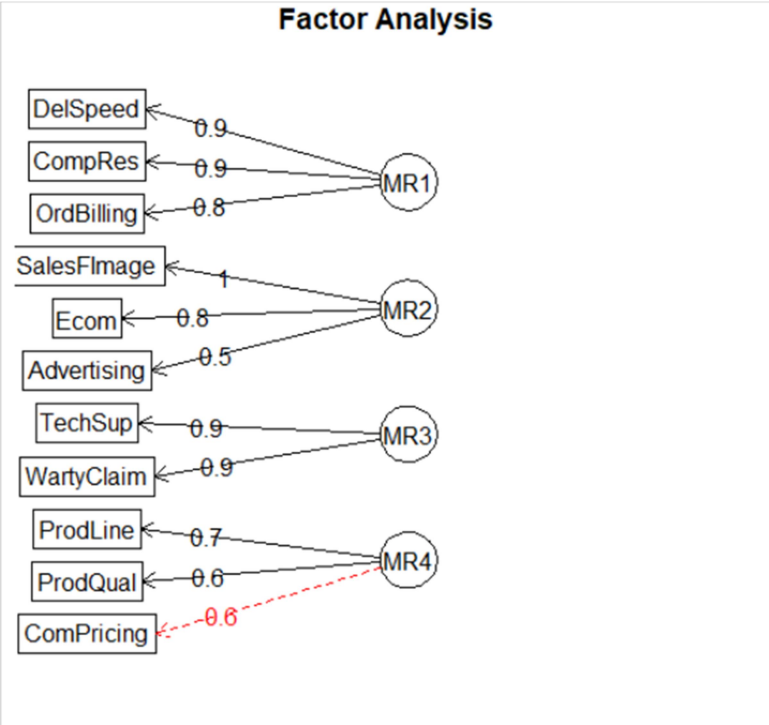
With varimax rotation we see that variables are now more clearly aligned to a particular dimension.

VariRotate$loadings

```
            MR1    MR2    MR3    MR4
ProdQual                         0.647
Ecom                0.781       -0.115
TechSup                    0.890  0.115
CompRes      0.897  0.130         0.132
Advertising  0.166  0.529
ProdLine     0.525         0.127  0.712
SalesFImage  0.114  0.980       -0.133
ComPricing          0.213 -0.209 -0.590
WartyClaim   0.103         0.879  0.129
OrdBilling   0.768  0.127
DelSpeed     0.949  0.185
```

Factor Naming

fa.diagram(VariRotate)



From the plot on left we see how variables are mapped to each factors.

MR1 comprises of Delivery Speed, Complaint Resolution and Order and Billing. This factor can be named as Customer Servicing

MR2 comprises of Salesforce Image, E-commerce and Advertising. This factor can be named as Marketing.

MR3 comprises of Tech Support and Warranty and Claim. This factor can be named as Customer Support.

MR4 comprises of Product Line, Product Quality and Competitive Pricing. This factor can be named as Product.

Hence four factors we have extracted are: **Customer Servicing, Marketing, Customer Support and Product.**

Linear Regression with derived Factors

Creation of data frame with new variables and dependent variable:

```
x = VariRotate$scores

regrdataframe = as.data.frame(cbind(x,Satisfaction))

colnames(regrdataframe) = c('Customer.Servicing','Marketing','Customer.Support','Product','Satsifaction')
```

Running Linear Regression on the new data frame

```
library(car)

LM = lm(formula = Satisfaction~Customer.Servicing+Marketing+Customer.Support+Product,data = regrdataframe)

summary(LM)
```

```
Call:
lm(formula = Satisfaction ~ Customer.Servicing + Marketing +
    Customer.Support + Product, data = regrdataframe)

Residuals:
     Min       1Q   Median       3Q      Max
-1.70781 -0.47233  0.08959  0.41930  1.35539

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.91800    0.06659 103.889  < 2e-16 ***
Customer.Servicing 0.57991    0.06816   8.508 2.53e-13 ***
Marketing          0.61811    0.06734   9.180 9.37e-15 ***
Customer.Support   0.05779    0.07135   0.810     0.42
Product            0.61117    0.07609   8.033 2.57e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6659 on 95 degrees of freedom
Multiple R-squared:  0.7004,   Adjusted R-squared:  0.6878
F-statistic: 55.54 on 4 and 95 DF,  p-value: < 2.2e-16
```

The linear model has an adjusted R square of .6878, which means that the model is able to account for 68.78% of variance in data. P-values of intercept, Customer.Servicing, Marketing and Product show that these are significant variables in predicting Satisfaction.

```
vif(LM)
```

```
Customer.Servicing          Marketing   Customer.Support            Product
          1.000880           1.002203           1.003109           1.005509
```

Vif values show that the variables are not having correlation amongst them.

Trend line between actual and predicted value