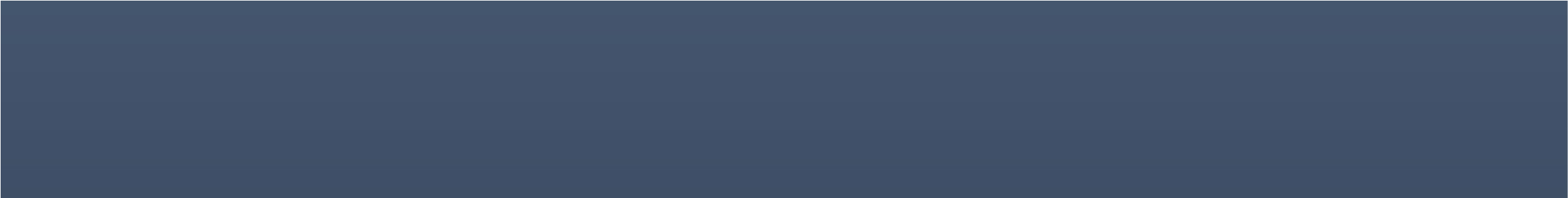
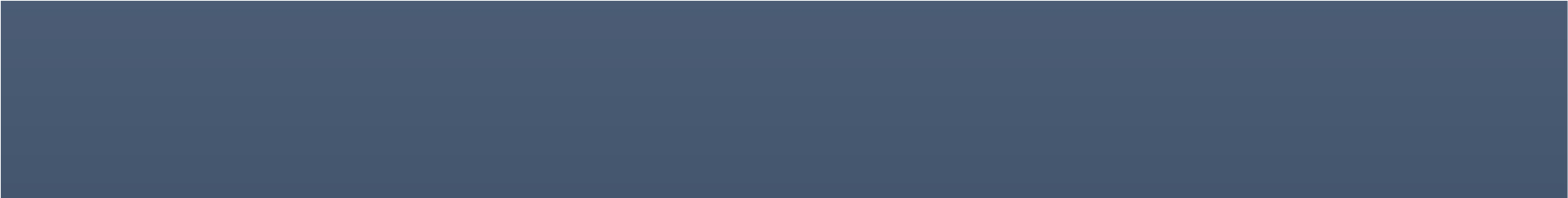
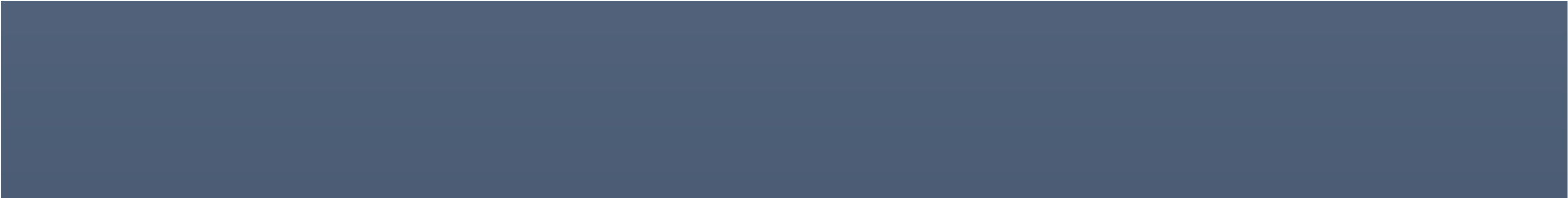
SUBMITTED BY- SANKALP RAJ MATTA



Advanced Statistics on Hair

Products



[

]

Year

# CONTENTS

1. Project Objective
2. Exploratory Data Analysis
   1. Basic Data Summary
   2. Univariate Data Analysis
   3. Bivariate Data Analysis
3. Check for the integrity of the data
   1. Outlier Detection
   2. Missing value Detection
   3. Basic Summary
4. Multicollinearity detection
   1. First method- Correlation Matrix
   2. Second method- Multiple regression
5. Simple Linear Regression for each variable
6. To perform Factor Analysis
   1. Find the eigen values
   2. Make a scree plot
   3. Kaiser Rule for number of factors
7. Interpretation of the factors chosen by Kaiser Rule
8. Creation of a 5 column data set after performing PCA
9. Multiple Linear Regression using the new Factors

10.Interpretation of the summary of Multiple Linear Regression

## 1.Project Objective

The objective of this project is to study the dataset Factor-hair. This dataset contains the quantitative aspects of a hair product selling firm . there are various independent factors on which the customer’s satisfaction depends. We are supposed to scrutinize this dataset and see how these factors are associated with each other and how they affect the customer satisfaction aspect and to what extent.

By finding the correlation between each variable and drawing out regression analysis, we will coagulate the similar factors into some generalized factors which share commonalities.

Our concepts of advanced statistics will thus, be put to practicality.

### 2.EXPLORATORY DATA ANALYSIS

2.1 Basic Data Summary

RCODE:

hair= read.csv("Factor-Hair-Revised.csv") str(hair)

Output:

'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 ...

attach(hair) names(hair) dim(hair) stat.desc(hair) head(hair,n=10) tail(hair,n=10)

Names- will show all the column names

> names(hair)

[1] "ID" "ProdQual" "Ecom" "TechSup" "CompRes" "Advertising

" "ProdLine"

[8] "SalesFImage" "ComPricing" "WartyClaim" "OrdBilling" "DelSpeed" "Satisfactio n"

Dim- will tell the dimension of our dataset

> dim(hair)

[1] 100 13

Stat.desc – will show the various statistical aspects of the dataset

stat.desc(hair)

ID ProdQual Ecom TechSup CompRes Advertising Pro dLine SalesFImage

nbr.val 100.000000 100.0000000 100.00000000 100.0000000 100.0000000 100.0000000 100.00

00000 100.0000000

nbr.null 0.000000 0.0000000 0.00000000 0.0000000 0.0000000 0.0000000 0.00

00000 0.0000000

nbr.na 0.000000 0.0000000 0.00000000 0.0000000 0.0000000 0.0000000 0.00

00000 0.0000000

min 1.000000 5.0000000 2.20000000 1.3000000 2.6000000 1.9000000 2.30

00000 2.9000000

max 100.000000 10.0000000 5.70000000 8.5000000 7.8000000 6.5000000 8.40

00000 8.2000000

range 99.000000 5.0000000 3.50000000 7.2000000 5.2000000 4.6000000 6.10

00000 5.3000000

sum 5050.000000 781.0000000 367.20000000 536.5000000 544.2000000 401.0000000 580.50

00000 512.3000000

median 50.500000 8.0000000 3.60000000 5.4000000 5.4500000 4.0000000 5.75

00000 4.9000000

mean 50.500000 7.8100000 3.67200000 5.3650000 5.4420000 4.0100000 5.80

50000 5.1230000

SE.mean 2.901149 0.1396279 0.07005164 0.1530457 0.1208403 0.1126943 0.13

15285 0.1072320

CI.mean.0.95 5.756509 0.2770521 0.13899765 0.3036758 0.2397734 0.2236099 0.26

09811 0.2127715

var 841.666667 1.9495960 0.49072323 2.3422980 1.4602384 1.2700000 1.72

99747 1.1498697

std.dev 29.011492 1.3962793 0.70051640 1.5304568 1.2084032 1.1269428 1.31

52850 1.0723198

coef.var 0.574485 0.1787810 0.19077244 0.2852669 0.2220513 0.2810331 0.22 65780 0.2093148

Head- will fetch the first 10 records of the dataset

> head(hair,n=10)

ID ProdQual Ecom TechSup CompRes Advertising ProdLine SalesFImage ComPricing WartyClaim O rdBilling DelSpeed

1. 1 8.5 3.9 2.5 5.9 4.8 4.9 6.0 6.8 4.7

5.0 3.7

1. 2 8.2 2.7 5.1 7.2 3.4 7.9 3.1 5.3 5.5

3.9 4.9

1. 3 9.2 3.4 5.6 5.6 5.4 7.4 5.8 4.5 6.2

5.4 4.5

1. 4 6.4 3.3 7.0 3.7 4.7 4.7 4.5 8.8 7.0

4.3 3.0

1. 5 9.0 3.4 5.2 4.6 2.2 6.0 4.5 6.8 6.1

4.5 3.5

1. 6 6.5 2.8 3.1 4.1 4.0 4.3 3.7 8.5 5.1

3.6 3.3

1. 7 6.9 3.7 5.0 2.6 2.1 2.3 5.4 8.9 4.8

2.1 2.0

1. 8 6.2 3.3 3.9 4.8 4.6 3.6 5.1 6.9 5.4

4.3 3.7

1. 9 5.8 3.6 5.1 6.7 3.7 5.9 5.8 9.3 5.9

4.4 4.6

1. 10 6.4 4.5 5.1 6.1 4.7 5.7 5.7 8.4 5.4

4.1 4.4

Satisfaction

1. 8.2
2. 5.7
3. 8.9
4. 4.8
5. 7.1
6. 4.7
7. 5.7
8. 6.3
9. 7.0
10. 5.5

Tails- will fetch the last 10 records of the dataset

> tail(hair,n=10)

ID ProdQual Ecom TechSup CompRes Advertising ProdLine SalesFImage ComPricing WartyClaim OrdBilling DelSpeed

1. 91 9.1 3.7 7.0 4.1 4.4 6.3 5.4 7.3 7.5

4.4 3.3

1. 92 7.1 4.2 4.1 2.6 2.1 3.3 4.5 9.9 5.5

2.0 2.4

1. 93 9.2 3.9 4.6 5.3 4.2 8.4 4.8 7.1 6.2

4.4 4.2

1. 94 9.3 3.5 5.4 7.8 4.6 7.5 5.9 4.6 6.4

4.8 4.6

1. 95 9.3 3.8 4.0 4.6 4.7 6.4 5.5 7.4 5.3

3.6 3.4

1. 96 8.6 4.8 5.6 5.3 2.3 6.0 5.7 6.7 5.8

4.9 3.6

1. 97 7.4 3.4 2.6 5.0 4.1 4.4 4.8 7.2 4.5

4.2 3.7

1. 98 8.7 3.2 3.3 3.2 3.1 6.1 2.9 5.6 5.0

3.1 2.5

1. 99 7.8 4.9 5.8 5.3 5.2 5.3 7.1 7.9 6.0

4.3 3.9

1. 100 7.9 3.0 4.4 5.1 5.9 4.2 4.8 9.7 5.7

3.4 3.5

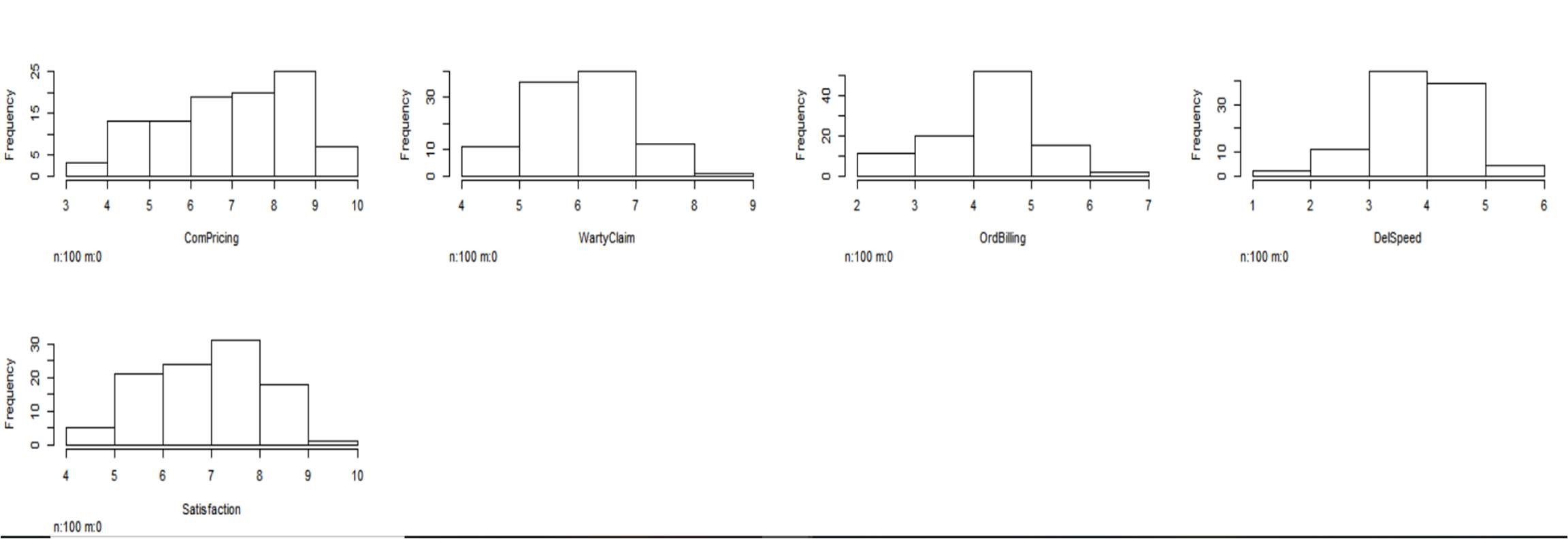
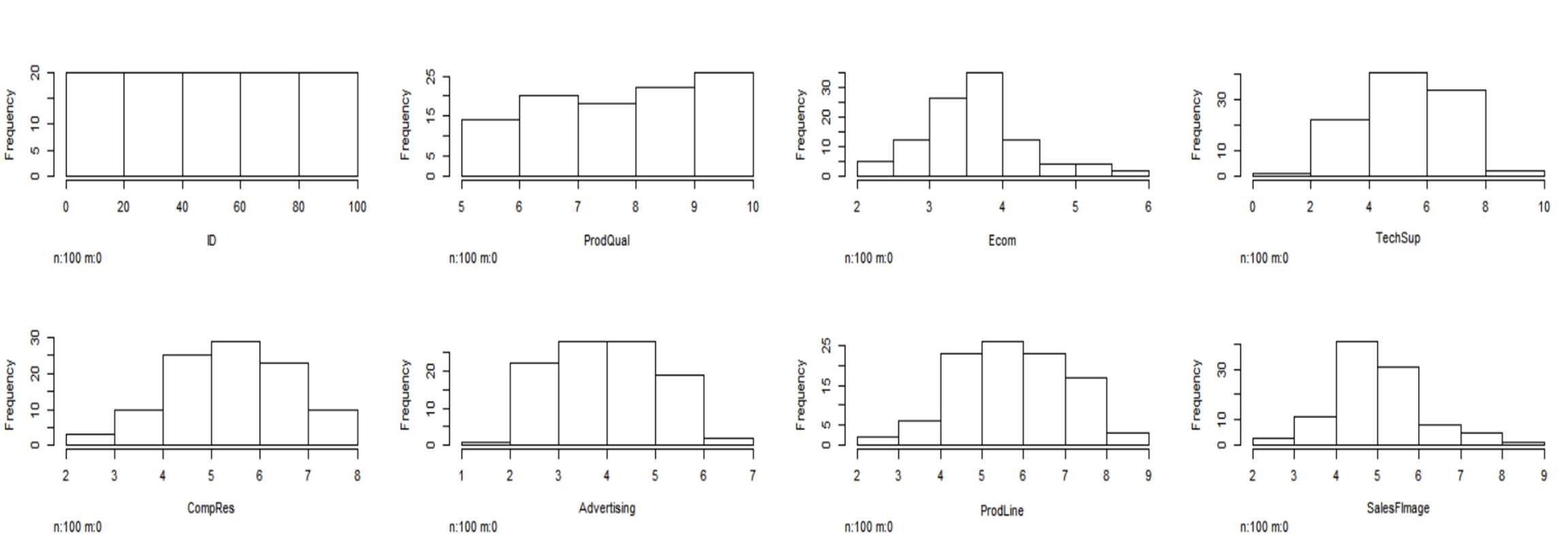
Satisfaction

1. 7.4
2. 4.8
3. 7.6
4. 8.9
5. 7.7
6. 7.3
7. 6.3
8. 5.4
9. 6.4
10. 6.4

2.2UNIVARIATE DATA ANALYSIS

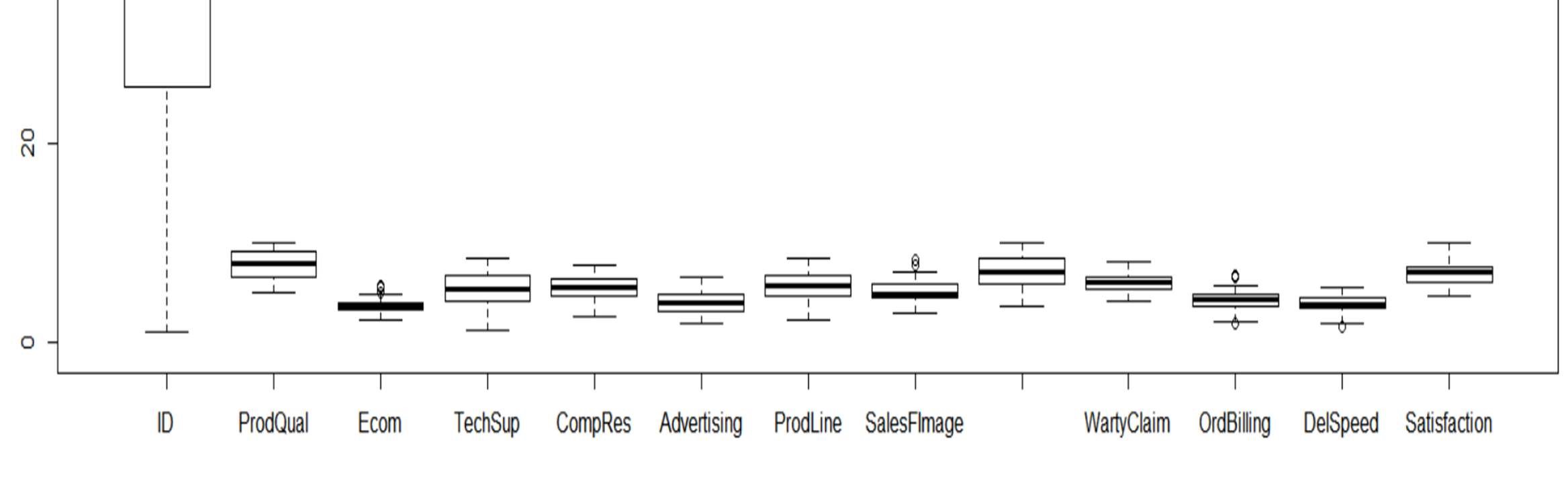
hist(hair)

This function will give the histogram of all the variables individually.



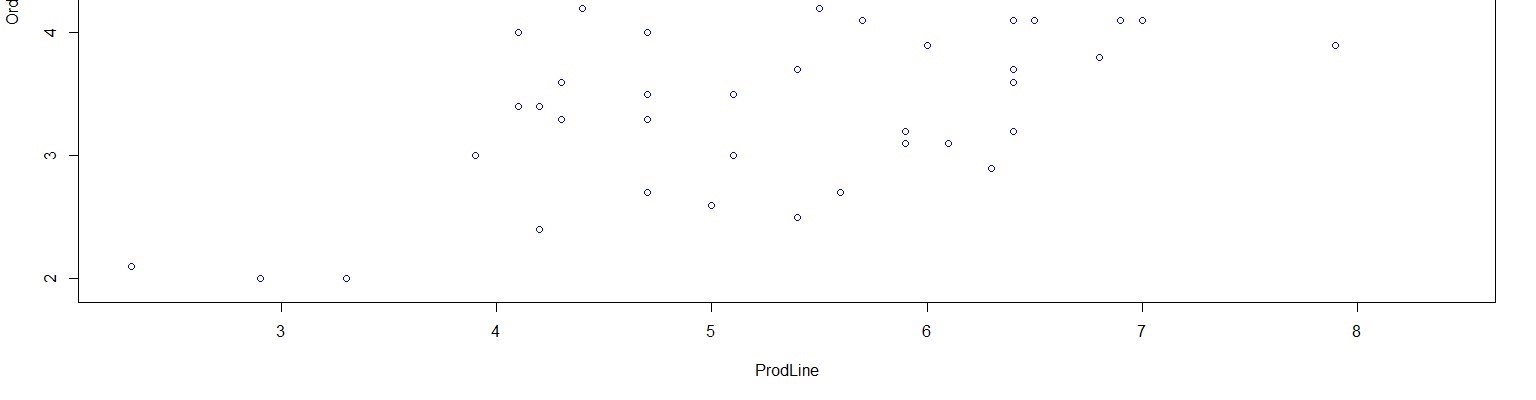
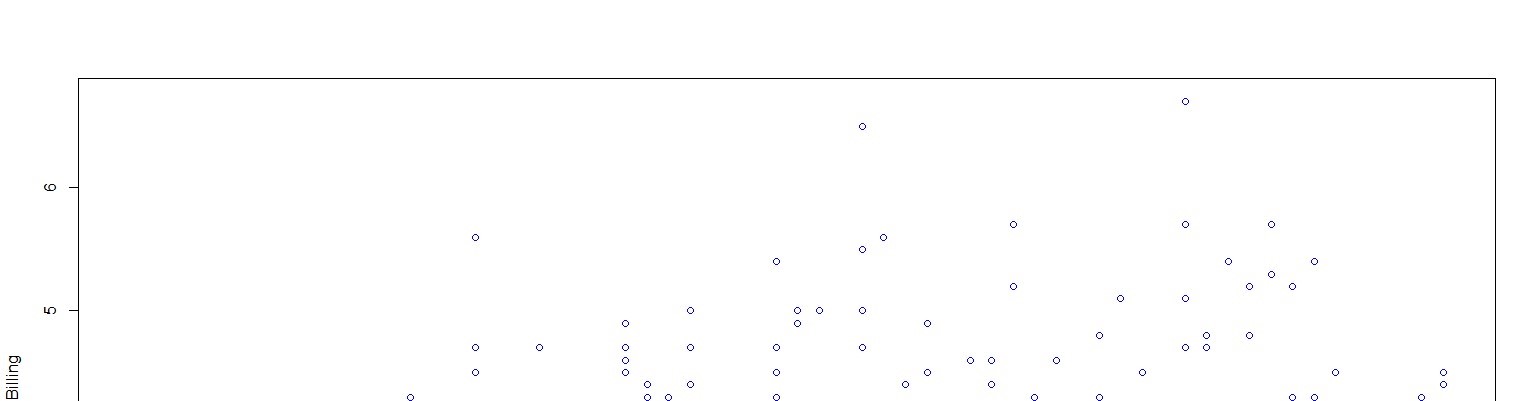
boxplot(hair)

This function will make the boxplots of all the variables individually on a single chart.

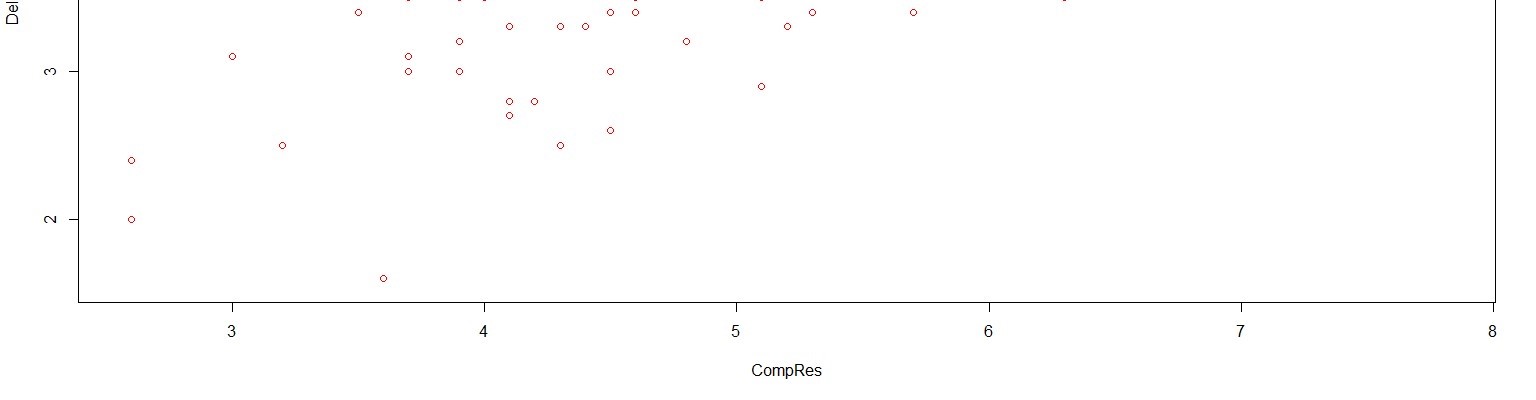
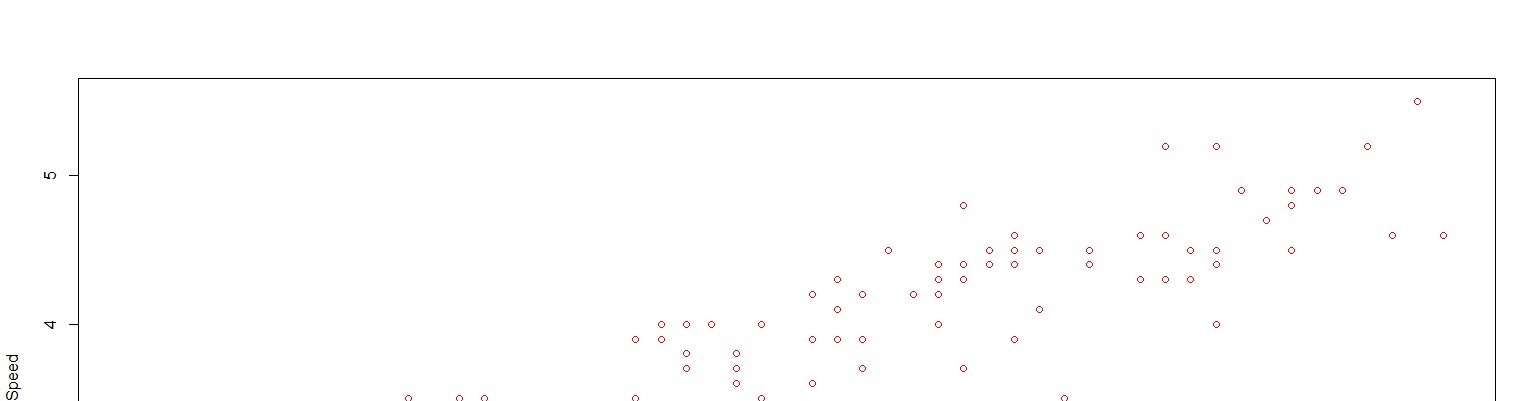


2.3BIVARIATE ANALYSIS

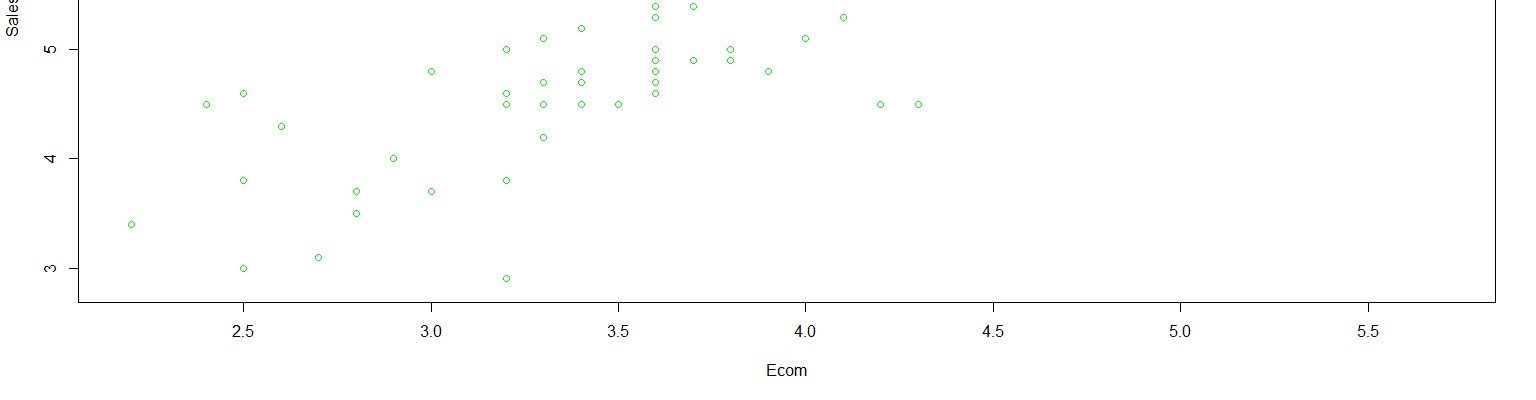
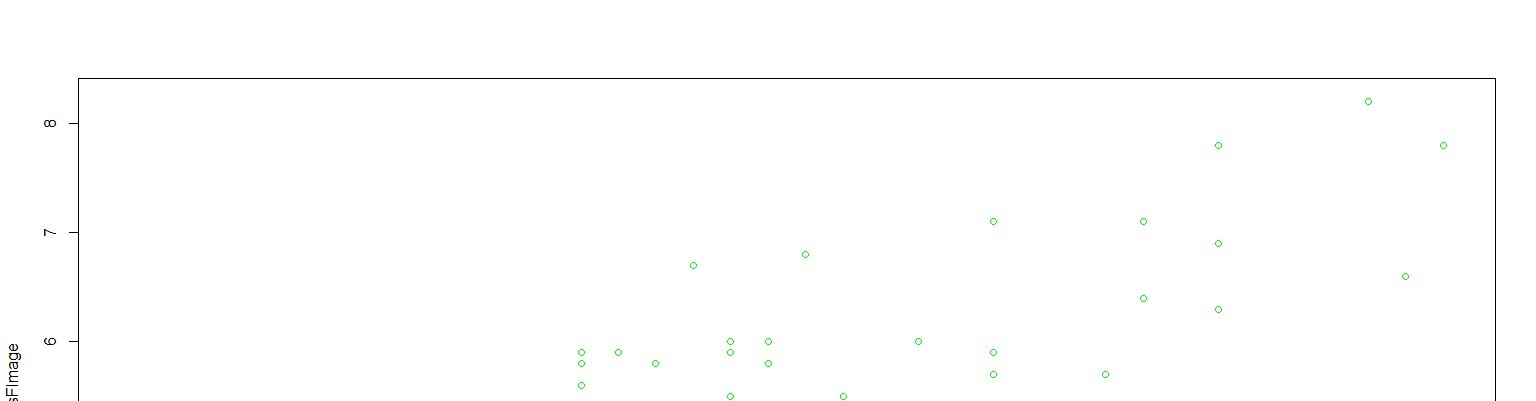
By default these plots will be scatterplots(for bivariate numerical variables). plot(ProdLine,OrdBilling)



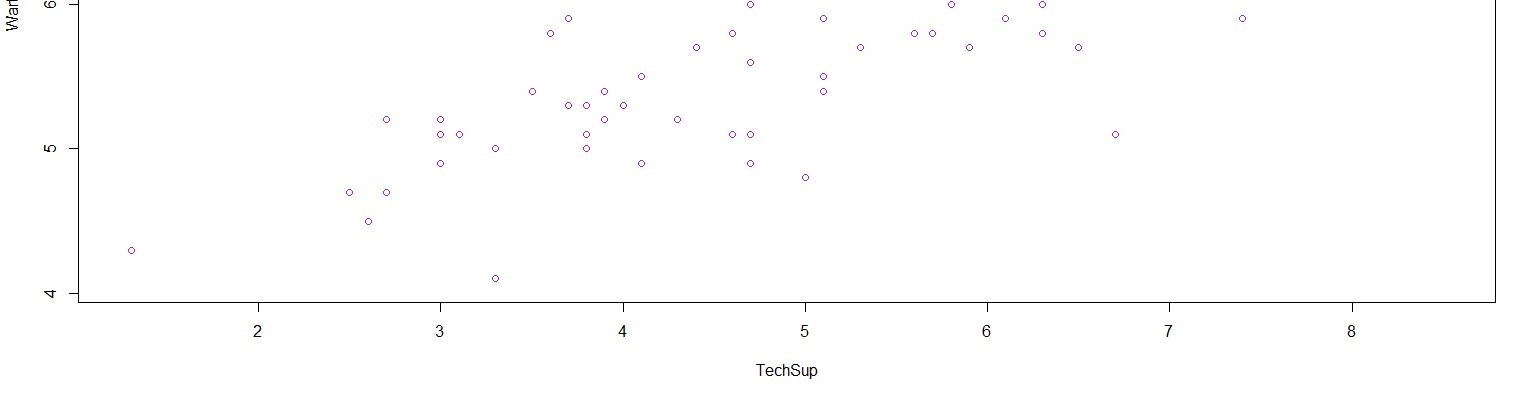
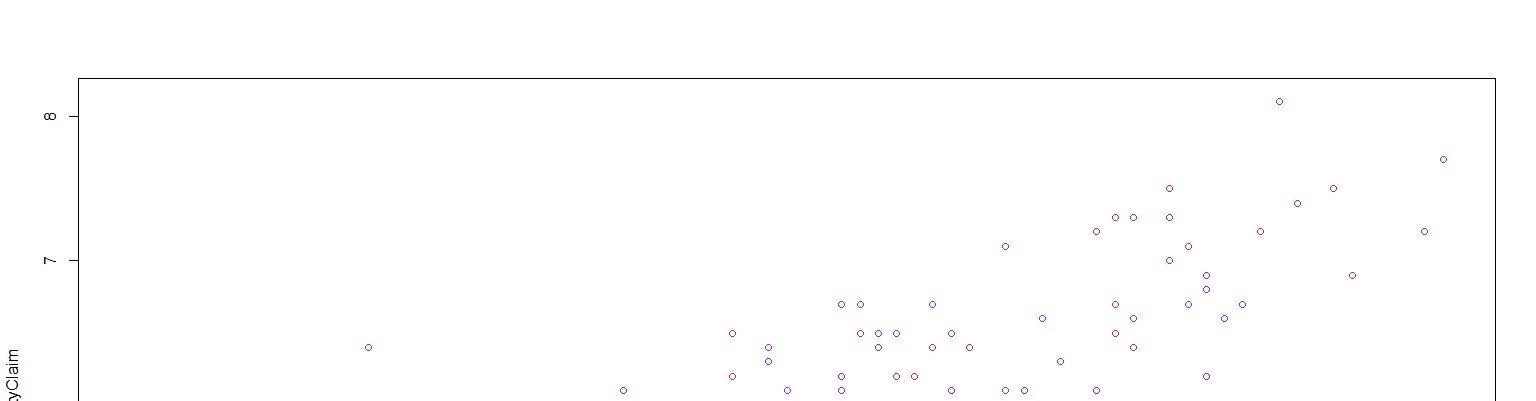
plot(CompRes,DelSpeed)



plot(Ecom,SalesFImage)



plot(TechSup,WartyClaim)

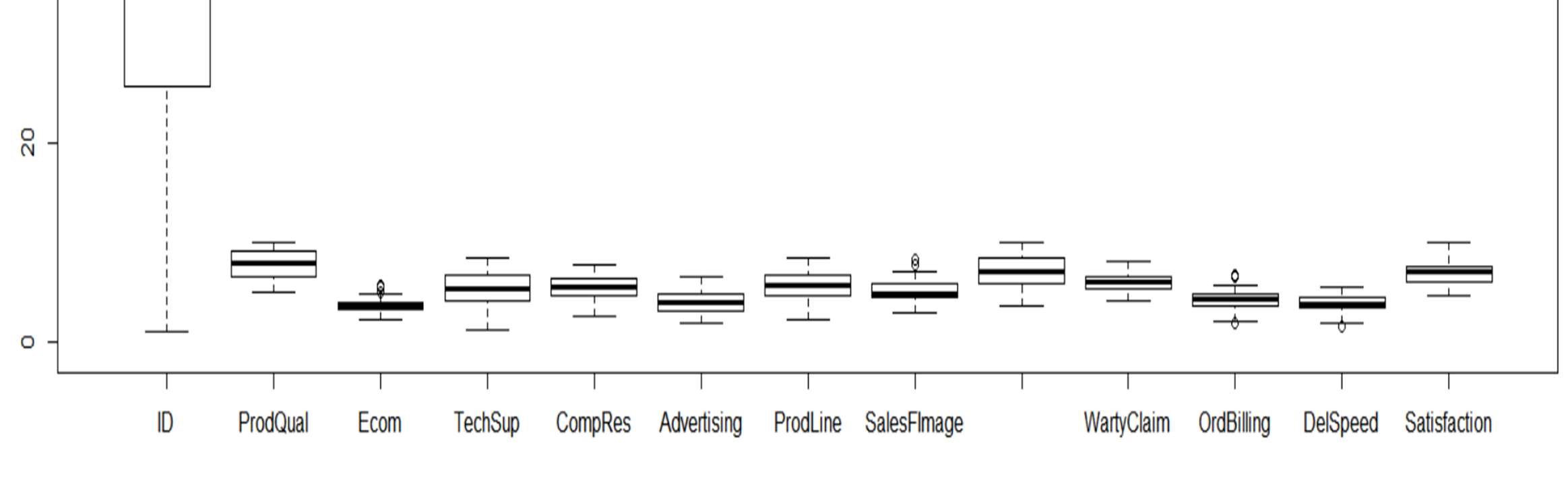


### 3.CHECKING FOR THE INTEGRITY OF THE DATA

3.1 Outlier Detection using Boxplot

Boxplot depicts that there are outliers in our data.

These plots are being made after seeing the correlation matrix and thus, figuring out which variables are to an extent related to each other and then we can visualize their relationship.



3.2 Missing Value Detection

null=is.na(hair) summary(null)

There are no missing values

> summary(null)# Shows there are no null values

ID ProdQual Ecom TechSup CompRes Advertising ProdLine

Mode :logical Mode :logical Mode :logical Mode :logical Mode :logical Mode :logic al Mode :logical

FALSE:100 FALSE:100 FALSE:100 FALSE:100 FALSE:100 FALSE:100 FALSE:100

SalesFImage ComPricing WartyClaim OrdBilling DelSpeed Satisfactio n

Mode :logical Mode :logical Mode :logical Mode :logical Mode :logical Mode :logic al

FALSE:100 FALSE:100 FALSE:100 FALSE:100 FALSE:100 FALSE:100

3.3 Summary of the dataset

All the variables are numeric except for the ID which is a number.

summary(hair)

ID ProdQual Ecom TechSup CompRes Advertis ing ProdLine

Min. : 1.00 Min. : 5.000 Min. :2.200 Min. :1.300 Min. :2.600 Min. :1

.900 Min. :2.300

1st Qu.: 25.75 1st Qu.: 6.575 1st Qu.:3.275 1st Qu.:4.250 1st Qu.:4.600 1st Qu.:3

.175 1st Qu.:4.700

Median : 50.50 Median : 8.000 Median :3.600 Median :5.400 Median :5.450 Median :4

.000 Median :5.750

Mean : 50.50 Mean : 7.810 Mean :3.672 Mean :5.365 Mean :5.442 Mean :4

.010 Mean :5.805

3rd Qu.: 75.25 3rd Qu.: 9.100 3rd Qu.:3.925 3rd Qu.:6.625 3rd Qu.:6.325 3rd Qu.:4

.800 3rd Qu.:6.800

Max. :100.00 Max. :10.000 Max. :5.700 Max. :8.500 Max. :7.800 Max. :6

.500 Max. :8.400

SalesFImage ComPricing WartyClaim OrdBilling DelSpeed Satisfacti on

Min. :2.900 Min. :3.700 Min. :4.100 Min. :2.000 Min. :1.600 Min. :4.7

00

1st Qu.:4.500 1st Qu.:5.875 1st Qu.:5.400 1st Qu.:3.700 1st Qu.:3.400 1st Qu.:6.0

00

Median :4.900 Median :7.100 Median :6.100 Median :4.400 Median :3.900 Median :7.0

50

Mean :5.123 Mean :6.974 Mean :6.043 Mean :4.278 Mean :3.886 Mean :6.9

18

3rd Qu.:5.800 3rd Qu.:8.400 3rd Qu.:6.600 3rd Qu.:4.800 3rd Qu.:4.425 3rd Qu.:7.6

25

Max. :8.200 Max. :9.900 Max. :8.100 Max. :6.700 Max. :5.500 Max. :9.9 00

1. MULTICOLLINEARITY DETECTION
   1. First method- Using Correlation Matrix ggcorr(hair[,2:13],label= TRUE)

We see that there are so many factor-pairs which exhibit a high correlation with each other. Thus, the presence of multiple corelated variables proves the presence of Multicollinearity.

We see that the dependent variable 'Satisfcation' is variably dependent on the various other independent variables. There is high co-relation amongst various pairs of variables showing multicollinearity.



* 1. Second Method- Multiple Regression

We'll test for multicollinearity using multiple-regression too.

RCODE:

model=

lm(Satisfaction~ProdQual+Ecom+TechSup+CompRes+Advertising+ProdLine+SalesFIm age+ComPricing+WartyClaim+OrdBilling+DelSpeed) summary(model)

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

We see that the amalgamated effect of each one upon the other makes the explanatory power extremely difficult. There are 3 variables which are significantly affecting the satisfaction - Product Quality , Ecommerce, Salesforce Image and we cannot say if Product Quality is more significant than Ecommerce or Salesforce Image.

This shows that there is multicollinearity because there are multiple variables which are affecting the dependent variable Satisfaction. However, Multicollinearity is not a problem.

### 5. SIMPLE REGRESSION FOR EACH VARIABLE

RCODE:

model1= lm(Satisfaction~ProdQual)

summary(model1) # 23.65% of the changes in Satisfcation are explained by Product Quality

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 \*\*\*

model2= lm(Satisfaction~Ecom)

summary(model2) # 7.99% of the changes in Satisfcation are explained by ECommerce

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 \*\*

model3= lm(Satisfaction~TechSup)

summary(model3) # 1.26% of the changes in Satisfcation are explained by Technical Support

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

model4= lm(Satisfaction~CompRes)

summary(model4) # 36.39% of the changes in Satisfcation are explained by Complaint Resolution

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 \*\*\*

model5= lm(Satisfaction~Advertising)

summary(model5) # 9.2% of the changes in Satisfcation are explained by Advertising

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 \*\*

model6= lm(Satisfaction~ProdLine)

summary(model6) # 30.31% of the changes in Satisfcation are explained by Product Line

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 \*\*\*

model7= lm(Satisfaction~SalesFImage)

summary(model7) # 25.02% of the changes in Satisfcation are explained by Salesforce Image

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 \*\*\*

model8= lm(Satisfaction~ComPricing)

summary(model8) # 4.33% of the changes in Satisfcation are explained by Competitive Pricing

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 \*

model9= lm(Satisfaction~WartyClaim)

summary(model9) # 3.15% of the changes in Satisfcation are explained by Warranty and Claims

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 .

model10= lm(Satisfaction~OrdBilling)

summary(model10) # 27.22% of the changes in Satisfcation are explained by Order and billing

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 \*\*\*

model11= lm(Satisfaction~DelSpeed)

summary(model11) # 33.3% of the changes in Satisfaction are explained by Delivery Speed

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 \*\*\*

Thus, we have performed Simple linear Regression for each variable.

### 6. TO PERFORM FACTOR ANALYSIS

6.1 To find Eigen Values RCODE:

Hair\_factor= hair[,2:12] str(Hair\_factor)

ev= eigen(cor(Hair\_factor)) print(ev,digits= 5)

eigenvalue= ev$values eigenvalue

factor= c(1,2,3,4,5,6,7,8,9,10,11)

# These eigen values are in descending order

eigen() decomposition

$values

[1] 3.426971 2.550897 1.690976 1.086556 0.609424 0.551884 0.401518 0.246952 0.203553 0.1328 42 0.098427

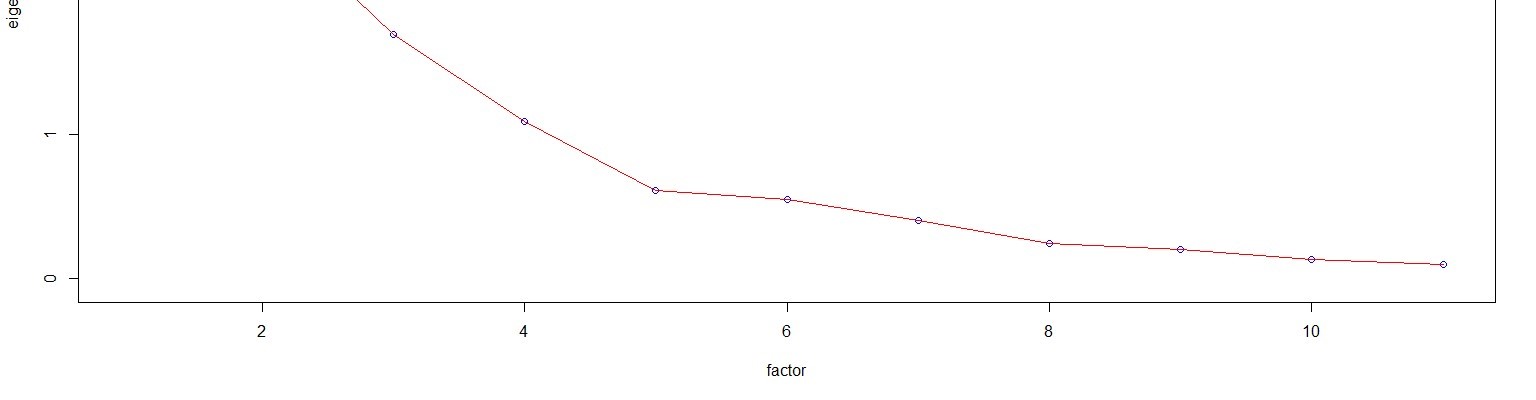
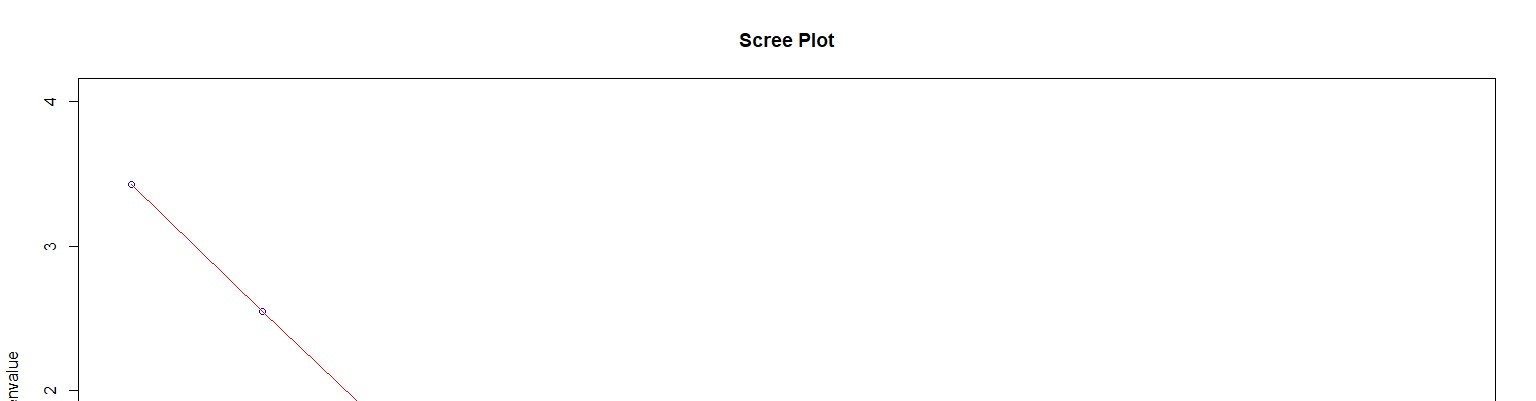
6.2 Make a Scree Plot

RCODE:

scree= data.frame(factor,eigenvalue)

plot(scree, main= "Scree Plot", col= "Blue", ylim= c(0,4)) lines(scree, col="Red")

# We observe that there are 4 significant factors according to Kaiser Rule, rest factor will be discarded



First, we will create an UNROTATED MATRIX and scrutinize the factor loadings and the communality.

RCODE:

unrotate= principal(Hair\_factor, nfactors = 4, rotate="none") print(unrotate)

PC1 PC2 PC3 PC4 h2 u2 com

ProdQual 0.25 -0.50 -0.08 0.67 0.77 0.232 2.2

Ecom 0.31 0.71 0.31 0.28 0.78 0.223 2.1

TechSup 0.29 -0.37 0.79 -0.20 0.89 0.107 1.9

CompRes 0.87 0.03 -0.27 -0.22 0.88 0.119 1.3

Advertising 0.34 0.58 0.11 0.33 0.58 0.424 2.4

ProdLine 0.72 -0.45 -0.15 0.21 0.79 0.213 2.0

SalesFImage 0.38 0.75 0.31 0.23 0.86 0.141 2.1

ComPricing -0.28 0.66 -0.07 -0.35 0.64 0.359 1.9

WartyClaim 0.39 -0.31 0.78 -0.19 0.89 0.108 2.0

OrdBilling 0.81 0.04 -0.22 -0.25 0.77 0.234 1.3

DelSpeed 0.88 0.12 -0.30 -0.21 0.91 0.086 1.4

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

# The picture isn't very clear, we will make a ROTATED MATRIX now.

rotate= principal(Hair\_factor,nfactors = 4, rotate = "verimax") print(rotate)

PC1 PC2 PC3 PC4 h2 u2 com

ProdQual 0.25 -0.50 -0.08 0.67 0.77 0.232 2.2

Ecom 0.31 0.71 0.31 0.28 0.78 0.223 2.1

TechSup 0.29 -0.37 0.79 -0.20 0.89 0.107 1.9

CompRes 0.87 0.03 -0.27 -0.22 0.88 0.119 1.3

Advertising 0.34 0.58 0.11 0.33 0.58 0.424 2.4

ProdLine 0.72 -0.45 -0.15 0.21 0.79 0.213 2.0

SalesFImage 0.38 0.75 0.31 0.23 0.86 0.141 2.1

ComPricing -0.28 0.66 -0.07 -0.35 0.64 0.359 1.9

WartyClaim 0.39 -0.31 0.78 -0.19 0.89 0.108 2.0

OrdBilling 0.81 0.04 -0.22 -0.25 0.77 0.234 1.3

DelSpeed 0.88 0.12 -0.30 -0.21 0.91 0.086 1.4

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

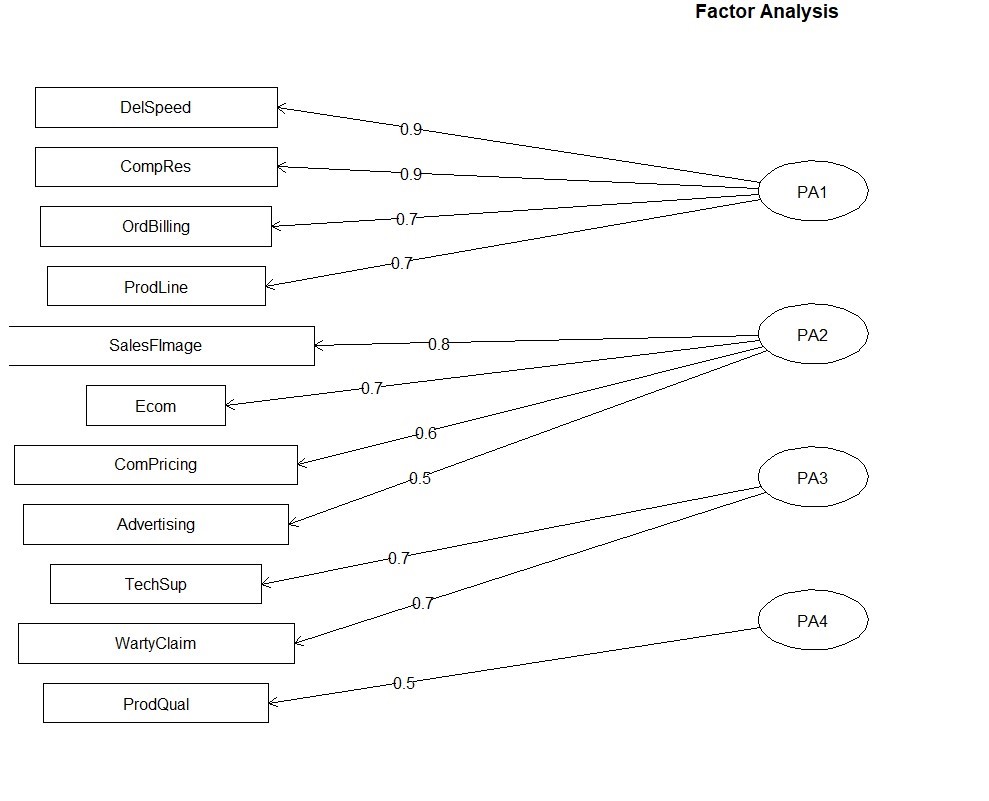
Cumulative Proportion 0.39 0.68 0.88 1.00

# Communality H2 tells the ability of all the 4 factors to capture as much of the variance of the variables as possible

### 7.INTERPRETATION OF THE FACTORS CHOSEN BY KAISER RULE

RCODE:

fa= fa(Hair\_factor, nfactors = 4, rotate = "none",fm= "pa") print(fa) fa.diagram(fa)



Thus, we know know the variables which belong to a generalized component factor.

# DelSpeed + CompRes + OrdBilling + ProdLine belong to PA1 and by pondering about the characteristics of these variables, we understand that we can rename that component as = Orders\_and\_Service

# Salesforce Image + E-Commerce + Advertising + Competitive Pricing belong to PA2 and by contemplating about the characteristics of these variables, we understand that we can rename that component as= Marketising\_Quality

# TechSup + WartyClaim belong to PA3 and we understand that we can rename that component as= Support\_Assistance

# ProdQual is an indepdent factor and will remian as it is and doesn't belong to any of the generalized groups

### 8. CREATION OF A 5 COLUMN DATASET AFTER PERFORMING PCA/FA

# Now , we will combine the columns of the original dataset and this PCA\_hair dataset and create

# a new dataset by the name= new\_hair containing 5 columns.

RCODE:

new\_hair= cbind(PCA\_hair, hair[,13]) new\_hair= as.data.frame(new\_hair) str(new\_hair)

colnames(new\_hair)[1]<- 'Orders\_and\_Service' colnames(new\_hair)[2]<- 'Marketising\_Quality' colnames(new\_hair)[3]<- 'Support\_Assistance' colnames(new\_hair)[4]<- 'Product\_Quality' colnames(new\_hair)[5]<- 'Customer Satisfaction'

str(new\_hair)

attach(new\_hair) # This is the new dataset with 5 columns and 100 rows

> str(new\_hair)

'data.frame': 100 obs. of 5 variables:

$ Orders\_and\_Service : num -0.237 0.77 1.012 -1.093 -0.416 ...

$ Marketising\_Quality : num 1.24 -1.7015 -0.0967 -0.4988 -0.5765 ...

$ Support\_Assistance : num -1.1289 -1.8403 0.0637 1.3867 -0.0133 ... $ Product\_Quality : num 0.978 -0.778 1.299 -0.629 0.372 ...

$ Customer Satisfaction: num 8.2 5.7 8.9 4.8 7.1 4.7 5.7 6.3 7 5.5 ...

### 9.MULTIPLE LINEAR REGRESSION USING NEW REDUCED FACTORS

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

RCODE:

multimodel=lm(Satisfaction~Orders\_and\_Service+Marketising\_Quality+Support\_Assi stance+Product\_Quality,data = new\_hair)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.918000 0.066959 103.317 < 2e-16 \*\*\*

Orders\_and\_Service 0.902499 0.068701 13.137 < 2e-16 \*\*\* Marketising\_Quality 0.129861 0.069265 1.875 0.0639 .

Support\_Assistance 0.002653 0.070735 0.038 0.9702 Product\_Quality 0.515861 0.076533 6.740 1.22e-09 \*\*\*

summary(multimodel) anova(multimodel)

# Yhat= 6.918 + .902x1 + .129x2 + .0026x3 + .5158x4

# multipled R Squared : .6971 implies 69.71% of the variations in Customer

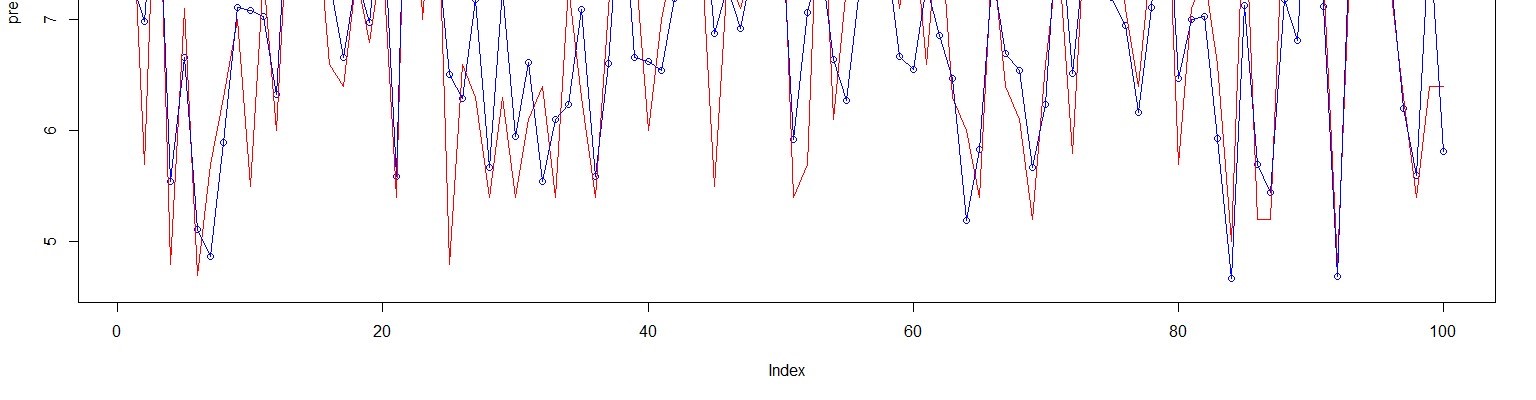
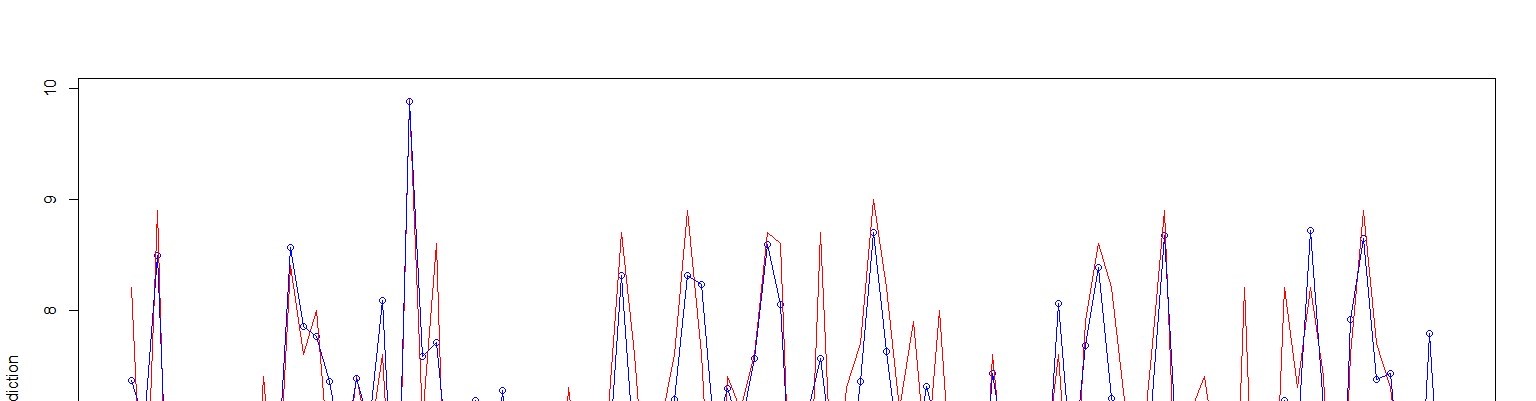
Satisfaction is explained the independent variables namely : Orders\_and\_Service , Marketising\_Quality , Support\_Assistance , and Product\_Quality

The Significant independant factors are : Orders\_and\_Service , Marketising\_Quality and Product\_Quality and the non-significant factor is Support\_Assistance because it only explains 2.65% of the changes and it can also be neglected leaving us with 3 independant factors derived from factor analysis

prediction= predict(multimodel)

actual= Satisfaction

backtrack= data.frame(actual,prediction) plot(actual, col= 'red') lines(actual, col='red') plot(prediction, col='blue') lines(prediction, col='blue')



We can clearly infer that the prediction falls mostly in line with the actual data.

# Now we'll check for the confidence level

?confint()

confint(multimodel, level = 0.95)

2.5 % 97.5 %

(Intercept) 6.785069371 7.0509306

Orders\_and\_Service 0.766110020 1.0388882

Marketising\_Quality -0.007647819 0.2673707

Support\_Assistance -0.137773264 0.1430792

Product\_Quality 0.363923251 0.6677992

Therefore we will not take the slopes which are extracted by Multiple Regression because we are conservative and will take only the slopes corresponding to the 2.5% of the predicted levels.

### 10.INTERPRETATION OF THE SUMMARY OF MLR

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.918000 0.066959 103.317 < 2e-16 \*\*\*

Orders\_and\_Service 0.902499 0.068701 13.137 < 2e-16 \*\*\* Marketising\_Quality 0.129861 0.069265 1.875 0.0639 .

Support\_Assistance 0.002653 0.070735 0.038 0.9702

Product\_Quality 0.515861 0.076533 6.740 1.22e-09 \*\*\* ---

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

Significance of Rsquared= By MLR , the value is coming to be .6971. This means that 69.71 % of the changes in Satisfaction are a explained by the four factors we have.

Degrees of freedom= There are total 100 records and the number of factors are

4, therefore, it’s explainable that the Dof is = 100-4-1= 95

F-statistic= This value of F-statistic is fairly high and thus, the result is significant.

P-Values= The P-Values of every factor is fairly low and thus, being very less than alpha of .5, we consider that the null hypothesis is rejected and the factors are considerable and significantly contributing.

\*\*As we concluded that the Support\_Assistance didn't contribute much to the predictions we can neglect it and still the model will be valid.multimodel2= lm(Satisfaction~Orders\_and\_Service+Marketising\_Quality+Product\_Quality, data = new\_hair) summary(multimodel2)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.91800 0.06661 103.858 < 2e-16 \*\*\*

Orders\_and\_Service 0.90249 0.06834 13.205 < 2e-16 \*\*\* Marketising\_Quality 0.12991 0.06889 1.886 0.0624 .

Product\_Quality 0.51600 0.07604 6.786 9.5e-10 \*\*\* ---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6661 on 96 degrees of freedom

Multiple R-squared: 0.6971, Adjusted R-squared: 0.6876

F-statistic: 73.65 on 3 and 96 DF, p-value: < 2.2e-16

# We see that in absence of the factor Support\_Assistance doesn't affect the variance explaining feature

# of model and thus , we can also suffice with the three factors.

# These 3 factors will explain the variance in the satisfaction 69.71 % .