**CASE STUDY OF FACTOR HAIR REVISED**

**1 Project Objective:**

What is multicollinearity and how it affects the regression model? Multicollinearity occurs when the independent variables of a regression model are correlated and if the degree of collinearity between the independent variables are high, it becomes difficult to estimate the relationship between each independent variable and the dependent variable and the overall precision of the estimate coefficients .Even though the regression models with high multicollinearity can give you a high R squared but hardly any significant variables.

The objective of the project is to use the dataset “Factor Hair Revised” .csv to build a regression model to **predict the satisfaction.**

**Project Approach:**

* Data Exploration
* Collinearity of the variables
* Initial Regression analysis
* Factor Analysis & Principal Component Analysis
* Labelling and Interpreting of the Factors
* Regression analysis using the factors as Independent Variable
* Model Performance Measures
* Finding out the Predicted and Actual satisfaction
* Insights from the dataset

**2 Assumptions:**

Satisfaction of the hair product is depending on available variables like:

* Product Quality
* E-Commerce
* Technical Support
* Complaint Resolution
* Advertising
* Product Line
* Salesforce Image
* Competitive Pricing
* Warranty & Claims
* Order & Billing
* Delivery Speed
* Customer Satisfaction

**3 Exploratory Data Analysis – Step by step approach**

A Typical Data exploration activity consists of the following steps:

1. Basic data summary, Univariate, Bivariate analysis, graphs
2. Check for Outliers and missing values and check the summary of the dataset
3. Check for Multicollinearity - Plot the graph based on Multicollinearity
4. Simple Linear Regression (with every variable)
5. Create a data frame with a minimum of 5 columns, 4 of which are different factors and the 5th column is Customer Satisfaction
6. Output Interpretation Tell why only 4 factors are being asked in the questions and tell whether it is correct in choosing 4 factors. Name the factors with correct explanations.
7. Perform Multiple Linear Regression with Customer Satisfaction as the Dependent Variable and the four factors as Independent Variables
8. MLR summary interpretation and significance (R, R2, Adjusted R2, Degrees of Freedom, f-statistic, coefficients along with p-values)
9. Output Interpretation (Conclusion)

**3.1 Environment Set up and Data Import**

**3.1.1 Install necessary Packages and Invoke Libraries**

Use this section to install necessary packages and invoke associated libraries. Having all the packages at the same places increases code readability.

Below are the Packages used in this project:

library(lattice)

library(ggplot2)

library (MASS)

library(dplyr)

library(car)

library(nFactors)

library(corrplot)

library(psych)

library(kableExtra)

library(boot)

**3.1.2 Set up 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 Read & Import Dataset**

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

Please refer Appendix A for Source Code.

**4.1 Basic data summary, Univariate, Bivariate analysis, graphs**

**Summary**

summary(mydata)

ID ProdQual Ecom TechSup CompRes

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

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

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

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

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

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

Advertising ProdLine SalesFImage ComPricing WartyClaim

Min. :1.900 Min. :2.300 Min. :2.900 Min. :3.700 Min. :4.100

1st Qu.:3.175 1st Qu.:4.700 1st Qu.:4.500 1st Qu.:5.875 1st Qu.:5.400

Median :4.000 Median :5.750 Median :4.900 Median :7.100 Median :6.100

Mean :4.010 Mean :5.805 Mean :5.123 Mean :6.974 Mean :6.043

3rd Qu.:4.800 3rd Qu.:6.800 3rd Qu.:5.800 3rd Qu.:8.400 3rd Qu.:6.600

Max. :6.500 Max. :8.400 Max. :8.200 Max. :9.900 Max. :8.100

OrdBilling DelSpeed Satisfaction

Min. :2.000 Min. :1.600 Min. :4.700

1st Qu.:3.700 1st Qu.:3.400 1st Qu.:6.000

Median :4.400 Median :3.900 Median :7.050

Mean :4.278 Mean :3.886 Mean :6.918

3rd Qu.:4.800 3rd Qu.:4.425 3rd Qu.:7.625

Max. :6.700 Max. :5.500 Max. :9.900

**Structure**

str(mydata)

'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 ...

**Dimension**

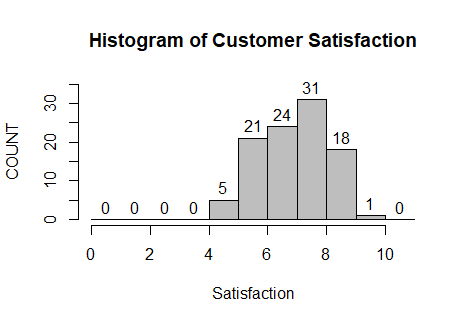
dim(mydata)

[1] 100 13

From summary and structure, we learned that the given data is scaled already and so no need to scale it further and We also know that first column named "ID" is just a column number

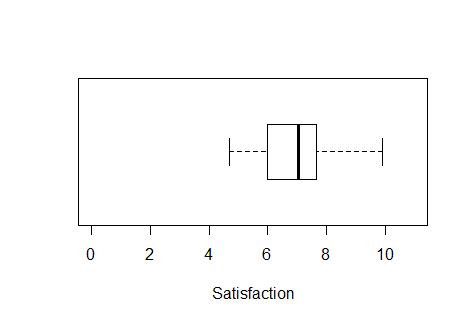
and we won't be needing in the for process.

Histogram of Satisfaction



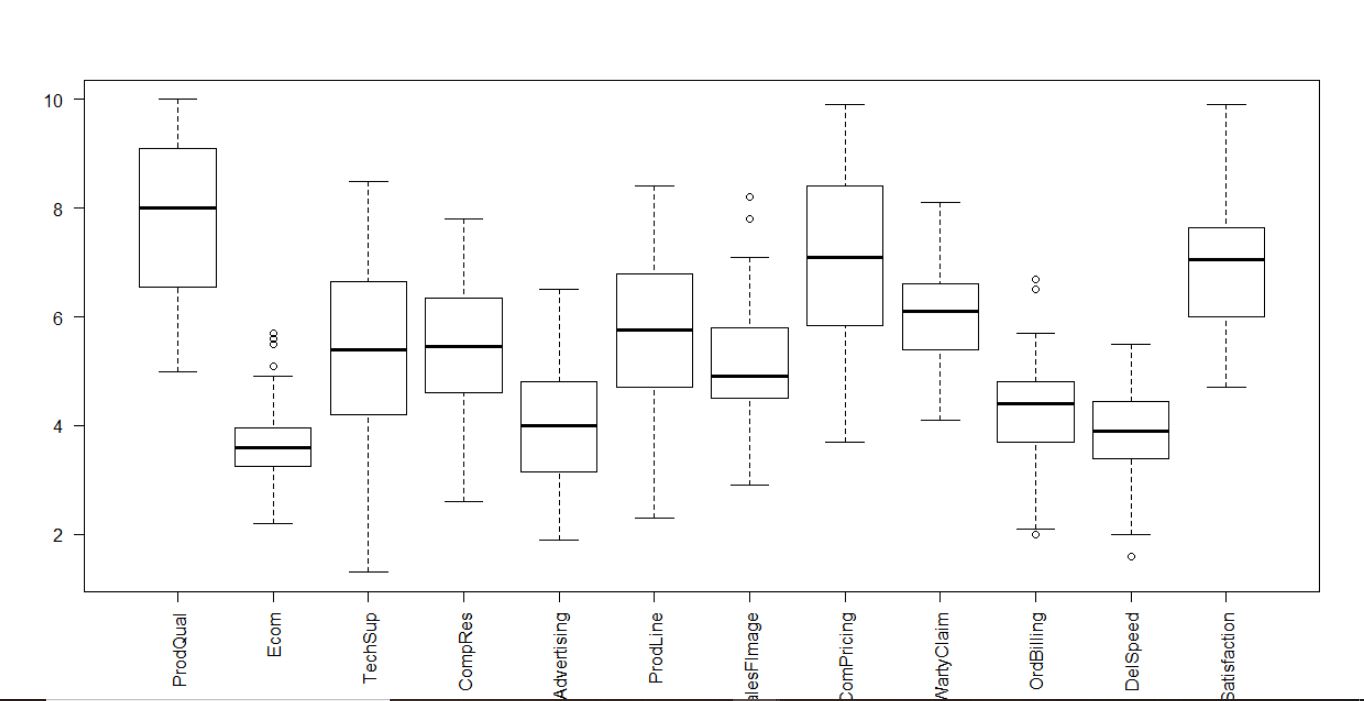
From this histogram we can conclude in 10 scale unit satisfaction rate of customer.

**Box plot of Customer Satisfaction**

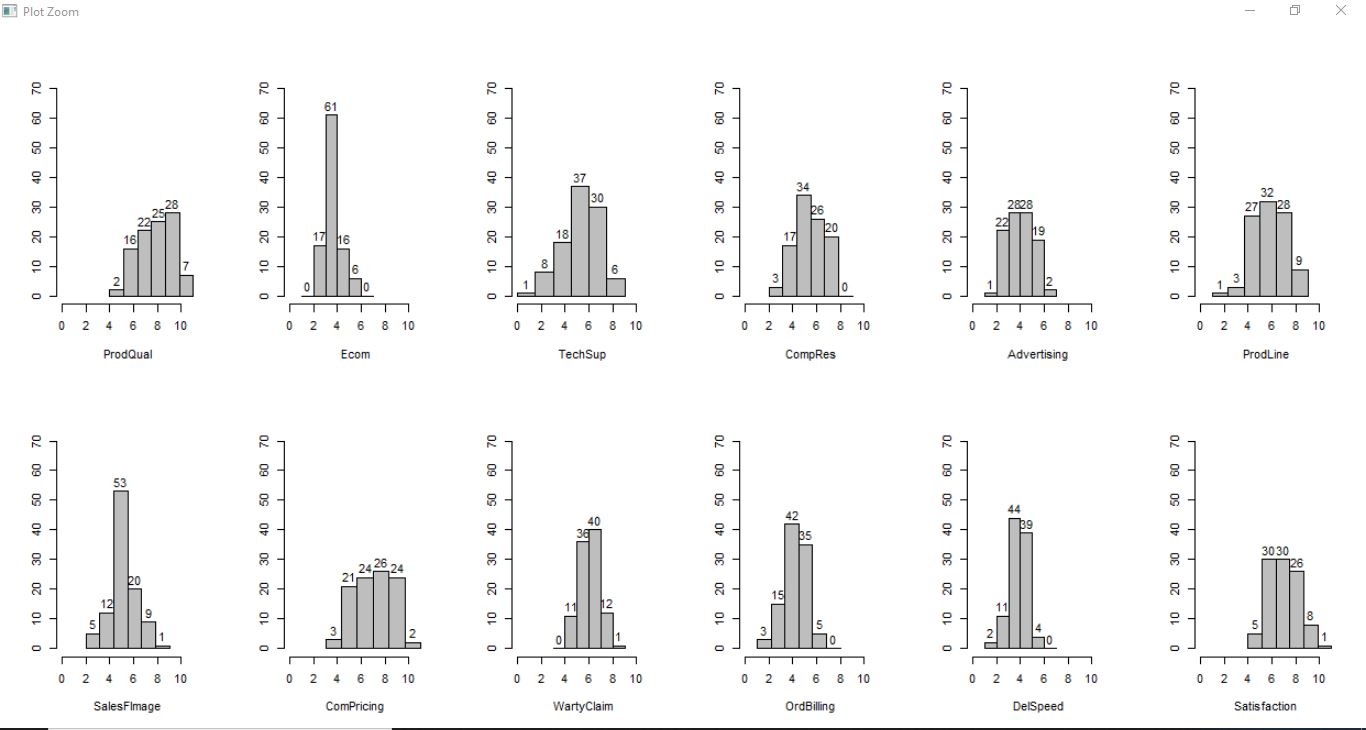


As per the box plot, we can conclude that most of the customer satisfaction is showing as below 8.

**Box Plot**

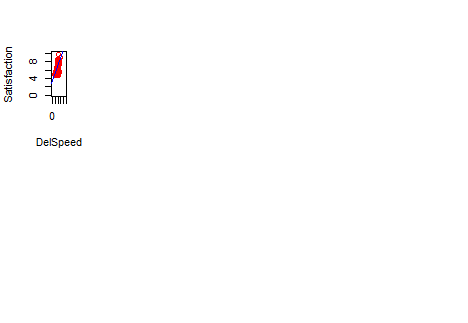
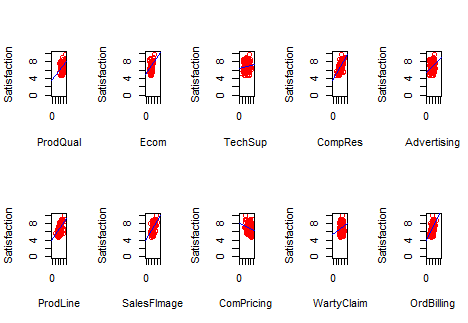


**Histogram of all variables**



Each variable here are reflecting scores given by customers on a graded scale.

**Scatter Plot**



This is the graphical representation of Independent variable V/s Target Variable (Customer Satisfaction)

**Outliers and Missing Valve of Dataset**

Outlier: From the box plot we have figured out the outlier.

Outliers which is there in dataset, we have named it as “outlier.csv”

**Missing Valve**

> sum(is.na(rithu))

[1] 0

From the Dataset we have got there is no missing valve .

**Summary of the Revised Dataset without Satisfaction**

summary(rithu)

ProdQual Ecom TechSup CompRes Advertising

Min. : 5.000 Min. :2.200 Min. :1.300 Min. :2.600 Min. :1.900

1st Qu.: 6.575 1st Qu.:3.275 1st Qu.:4.250 1st Qu.:4.600 1st Qu.:3.175

Median : 8.000 Median :3.600 Median :5.400 Median :5.450 Median :4.000

Mean : 7.810 Mean :3.672 Mean :5.365 Mean :5.442 Mean :4.010

3rd Qu.: 9.100 3rd Qu.:3.925 3rd Qu.:6.625 3rd Qu.:6.325 3rd Qu.:4.800

Max. :10.000 Max. :5.700 Max. :8.500 Max. :7.800 Max. :6.500

ProdLine SalesFImage ComPricing WartyClaim OrdBilling

Min. :2.300 Min. :2.900 Min. :3.700 Min. :4.100 Min. :2.000

1st Qu.:4.700 1st Qu.:4.500 1st Qu.:5.875 1st Qu.:5.400 1st Qu.:3.700

Median :5.750 Median :4.900 Median :7.100 Median :6.100 Median :4.400

Mean :5.805 Mean :5.123 Mean :6.974 Mean :6.043 Mean :4.278

3rd Qu.:6.800 3rd Qu.:5.800 3rd Qu.:8.400 3rd Qu.:6.600 3rd Qu.:4.800

Max. :8.400 Max. :8.200 Max. :9.900 Max. :8.100 Max. :6.700

DelSpeed

Min. :1.600

1st Qu.:3.400

Median :3.900

Mean :3.886

3rd Qu.:4.425

Max. :5.500

**Evidence of Multicollinearity**

vif(com1)

ProdQual Ecom TechSup CompRes Advertising ProdLine SalesFImage

1.635797 2.756694 2.976796 4.730448 1.508933 3.488185 3.439420

ComPricing WartyClaim OrdBilling DelSpeed

1.635000 3.198337 2.902999 6.516014

As per the data set VIF for Delivery speed found to be 6.516014 (Greater than 4)

**Simple Linear Regreesion**

lm.Prodqual

Call:

lm(formula = Satisfaction ~ ProdQual, data = rithu)

Coefficients:

(Intercept) ProdQual

3.6759 0.4151

* Beta-naught or intercept coefficient is equal to 3.6759
* Beta-slope or the variable coefficient product quality to 0.4151

As per the above summary we have applied for all the below variables.

> lm.Ecom=lm(Satisfaction~Ecom,rithu)

> lm.Ecom

Call:

lm(formula = Satisfaction ~ Ecom, data = rithu)

Coefficients:

(Intercept) Ecom

5.1516 0.4811

> lm.Techsup=lm(Satisfaction~TechSup,rithu)

> lm.Techsup

Call:

lm(formula = Satisfaction ~ TechSup, data = rithu)

Coefficients:

(Intercept) TechSup

6.44757 0.08768

> lm.compres=lm(Satisfaction~CompRes,rithu)

> lm.compres

Call:

lm(formula = Satisfaction ~ CompRes, data = rithu)

Coefficients:

(Intercept) CompRes

3.680 0.595

> lm.Advertising=lm(Satisfaction~Advertising,rithu)

> lm.Advertising

Call:

lm(formula = Satisfaction ~ Advertising, data = rithu)

Coefficients:

(Intercept) Advertising

5.6259 0.3222

> lm.Prodline=lm(Satisfaction~ProdLine,rithu)

> lm.Prodline

Call:

lm(formula = Satisfaction ~ ProdLine, data = rithu)

Coefficients:

(Intercept) ProdLine

4.0220 0.4989

> lm.SalesFImage=lm(Satisfaction~SalesFImage,rithu)

> lm.SalesFImage

Call:

lm(formula = Satisfaction ~ SalesFImage, data = rithu)

Coefficients:

(Intercept) SalesFImage

4.070 0.556

> lm.Compricing=lm(Satisfaction~ComPricing)

> lm.Compricing

Call:

lm(formula = Satisfaction ~ ComPricing)

Coefficients:

(Intercept) ComPricing

8.0386 -0.1607

> lm.WartyClaim=lm(Satisfaction~WartyClaim,rithu)

> lm.WartyClaim

Call:

lm(formula = Satisfaction ~ WartyClaim, data = rithu)

Coefficients:

(Intercept) WartyClaim

5.3581 0.2581

> lm.OrdBilling=lm(Satisfaction~OrdBilling,rithu)

> lm.OrdBilling

Call:

lm(formula = Satisfaction ~ OrdBilling, data = rithu)

Coefficients:

(Intercept) OrdBilling

4.0541 0.6695

> lm.DelSpeed=lm(Satisfaction~DelSpeed,rithu)

> lm.DelSpeed

Call:

lm(formula = Satisfaction ~ DelSpeed, data = rithu)

Coefficients:

(Intercept) DelSpeed

3.2791 0.9364

**Principal Component Analysis**

> ###Eigen valve computation###

> corlnMatrix=cor(rithu[,-12])

> corlnMatrix

ProdQual Ecom TechSup CompRes Advertising ProdLine

ProdQual 1.00000000 -0.1371632174 0.0956004542 0.1063700 -0.05347313 0.47749341

Ecom -0.13716322 1.0000000000 0.0008667887 0.1401793 0.42989071 -0.05268784

TechSup 0.09560045 0.0008667887 1.0000000000 0.0966566 -0.06287007 0.19262546

CompRes 0.10637000 0.1401792611 0.0966565978 1.0000000 0.19691685 0.56141695

Advertising -0.05347313 0.4298907110 -0.0628700668 0.1969168 1.00000000 -0.01155082

ProdLine 0.47749341 -0.0526878383 0.1926254565 0.5614170 -0.01155082 1.00000000

SalesFImage -0.15181287 0.7915437115 0.0169905395 0.2297518 0.54220366 -0.06131553

ComPricing -0.40128188 0.2294624014 -0.2707866821 -0.1279543 0.13421689 -0.49494840

WartyClaim 0.08831231 0.0518981915 0.7971679258 0.1404083 0.01079207 0.27307753

OrdBilling 0.10430307 0.1561473316 0.0801018246 0.7568686 0.18423559 0.42440825

DelSpeed 0.02771800 0.1916360683 0.0254406935 0.8650917 0.27586308 0.60185021

SalesFImage ComPricing WartyClaim OrdBilling DelSpeed

ProdQual -0.15181287 -0.40128188 0.08831231 0.10430307 0.02771800

Ecom 0.79154371 0.22946240 0.05189819 0.15614733 0.19163607

TechSup 0.01699054 -0.27078668 0.79716793 0.08010182 0.02544069

CompRes 0.22975176 -0.12795425 0.14040830 0.75686859 0.86509170

Advertising 0.54220366 0.13421689 0.01079207 0.18423559 0.27586308

ProdLine -0.06131553 -0.49494840 0.27307753 0.42440825 0.60185021

SalesFImage 1.00000000 0.26459655 0.10745534 0.19512741 0.27155126

ComPricing 0.26459655 1.00000000 -0.24498605 -0.11456703 -0.07287173

WartyClaim 0.10745534 -0.24498605 1.00000000 0.19706512 0.10939460

OrdBilling 0.19512741 -0.11456703 0.19706512 1.00000000 0.75100307

DelSpeed 0.27155126 -0.07287173 0.10939460 0.75100307 1.00000000

> cortest.bartlett(corlnMatrix,100)

$chisq

[1] 619.2726

$p.value

[1] 1.79337e-96

$df

[1] 55

> A=eigen(corlnMatrix)

> EV=A$values

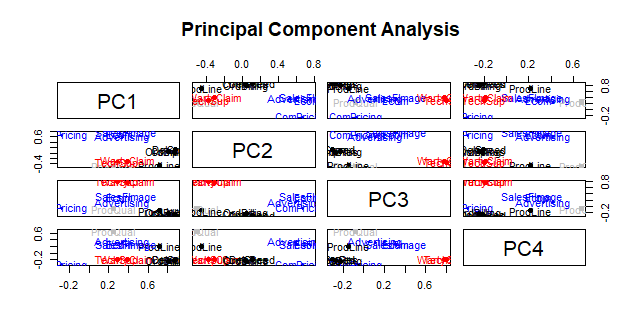
> EV

[1] 3.42697133 2.55089671 1.69097648 1.08655606 0.60942409 0.55188378 0.40151815

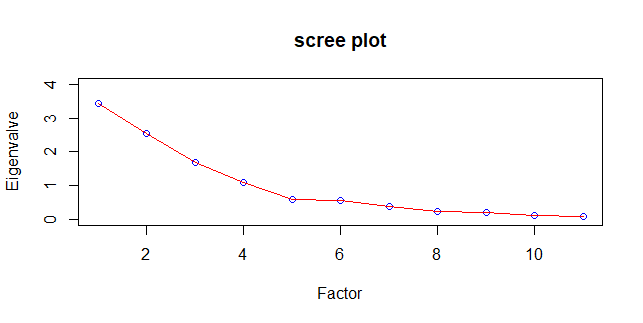
[8] 0.24695154 0.20355327 0.13284158 0.0984270

By Conducting a Barlett Sphericity test to check whether PCA can be done on the predictor variables of dataset.

Since the p value for the test is quite less significance level of **alpha = 0.001** so we reject the null hypothesis Ho (that PCA cannot be conducted implying that there is no correlation amongst the predictor variables)



**Output Interpretation**



As per the above scree plot extracting 4 factors from 11 variables

FourFactor1 = fa(r= rithu[,-12], nfactors =4, rotate ="none", fm ="pa")

> print(FourFactor1)

Factor Analysis using method = pa

Call: fa(r = rithu[, -12], nfactors = 4, rotate = "none", fm = "pa")

Standardized loadings (pattern matrix) based upon correlation matrix

PA1 PA2 PA3 PA4 h2 u2 com

ProdQual 0.20 -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 2.0

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 0.80 0.200 2.3

SalesFImage 0.39 0.80 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 2.0

OrdBilling 0.75 0.02 -0.18 -0.18 0.62 0.378 1.2

DelSpeed 0.90 0.10 -0.30 -0.20 0.94 0.058 1.4

PA1 PA2 PA3 PA4

SS loadings 3.21 2.22 1.50 0.68

Proportion Var 0.29 0.20 0.14 0.06

Cumulative Var 0.29 0.49 0.63 0.69

Proportion Explained 0.42 0.29 0.20 0.09

Cumulative Proportion 0.42 0.71 0.91 1.00

Mean item complexity = 1.9

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 0.97 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 0.90 0.82 0.56

PA1 , PA2, PA3, PA4 are the 4 factors

As per the Kyser Normalisation Rule We have taken only these 4 factors

Whichever is above 1 valve we have to consider.

**Data Frame**

> scores=round(rithu.pca$scores,2)

> scores

RC1 RC2 RC3 RC4

[1,] 0.13 0.77 -1.88 0.37

[2,] 1.22 -1.65 -0.61 0.81

[3,] 0.62 0.58 0.00 1.57

[4,] -0.84 -0.27 1.27 -1.25

[5,] -0.32 -0.83 -0.01 0.45

[6,] -0.65 -1.07 -1.30 -1.05

[7,] -2.63 -0.25 -0.56 -1.23

[8,] -0.28 -0.16 -0.75 -1.01

[9,] 1.05 -0.17 -0.09 -1.66

[10,] 0.43 0.76 -0.45 -0.89

[11,] -0.14 -0.77 -0.46 0.61

[12,] -1.45 1.36 0.44 -1.07

[13,] 0.62 2.11 -0.17 0.87

[14,] 0.43 -0.40 0.43 0.90

[15,] 1.44 0.66 -0.27 -1.04

[16,] 0.92 -1.06 -0.56 1.17

[17,] 0.52 -0.32 1.11 -1.03

[18,] 1.71 -0.16 0.25 -1.48

[19,] 1.16 -0.42 -0.38 -1.76

[20,] 0.29 1.78 -0.95 0.24

[21,] -0.62 -0.18 1.53 -1.83

[22,] -0.11 2.83 0.63 2.24

[23,] 0.08 -0.35 1.14 1.33

[24,] 1.95 -1.67 -0.86 0.50

[25,] 0.12 -0.02 0.47 -1.25

[26,] 0.57 -0.24 0.62 -1.35

[27,] 0.83 -0.99 1.04 0.92

[28,] 0.12 -1.11 0.38 -1.36

[29,] 1.16 -1.61 -0.06 0.80

[30,] -0.51 0.16 -1.55 -0.31

[31,] -0.81 -0.18 2.26 0.22

[32,] -1.07 -1.60 1.19 -0.07

[33,] -0.50 0.31 0.16 -0.97

[34,] 0.28 0.07 -0.03 -0.66

[35,] -1.21 0.61 0.28 -0.69

[36,] -1.38 -1.06 0.28 1.03

[37,] -0.62 -0.24 0.31 0.66

[38,] 1.36 0.04 0.11 0.58

[39,] 0.60 0.47 -1.29 -0.45

[40,] -0.59 1.48 -1.18 -1.04

[41,] 0.19 -0.39 -1.98 -0.60

[42,] 0.04 0.09 -1.17 0.54

[43,] 0.41 1.96 -1.09 0.99

[44,] 0.78 1.61 1.51 -1.15

[45,] 1.27 -1.77 -0.98 0.74

[46,] 1.06 0.68 0.32 -1.10

[47,] -0.12 -0.09 1.00 1.42

[48,] 2.10 0.46 0.84 -1.68

[49,] 0.16 0.88 -0.84 1.30

[50,] 0.23 0.50 -0.88 1.04

[51,] -0.94 -0.38 0.19 -0.65

[52,] 1.56 -1.91 -1.18 0.72

[53,] 0.86 -1.09 -0.24 0.87

[54,] -0.82 -0.53 0.54 0.33

[55,] 0.54 -0.68 -1.06 -0.81

[56,] -0.37 0.28 0.92 0.60

[57,] 1.98 1.43 -0.09 -0.84

[58,] 0.21 0.52 0.35 0.86

[59,] -1.34 0.55 0.33 1.94

[60,] 0.85 -1.58 0.57 0.74

[61,] 0.99 -1.26 1.70 0.79

[62,] -1.10 0.71 -0.15 0.40

[63,] -0.76 0.26 -1.19 0.78

[64,] -1.09 -1.95 0.43 -0.15

[65,] -1.21 0.15 0.58 -0.52

[66,] 1.34 0.54 -1.04 -1.25

[67,] 0.90 -0.59 2.06 -1.32

[68,] 0.42 -0.25 -0.30 -0.85

[69,] -0.87 -0.60 -1.00 -0.53

[70,] 0.14 -0.15 -1.28 -1.00

[71,] 0.34 2.06 0.69 0.09

[72,] -1.16 -0.18 -1.21 0.71

[73,] 0.93 1.32 -1.87 -0.56

[74,] -0.57 1.40 1.23 1.35

[75,] -0.30 0.87 -0.29 0.30

[76,] -0.89 0.23 1.04 1.61

[77,] -0.36 0.14 2.06 -0.63

[78,] 0.21 0.34 1.07 0.31

[79,] 1.13 0.64 0.44 1.47

[80,] -1.53 0.29 0.03 -0.31

[81,] -0.85 -0.25 0.45 1.53

[82,] 0.03 -0.92 0.49 0.40

[83,] -1.39 -0.98 0.21 0.63

[84,] -2.49 -0.74 1.63 -1.44

[85,] 1.00 -1.78 0.80 -0.01

[86,] -0.83 -0.42 -1.08 -0.45

[87,] -1.43 -0.30 -2.16 -1.27

[88,] 1.07 -1.30 1.40 0.04

[89,] 0.09 -0.06 0.13 0.24

[90,] 1.08 2.38 1.89 -1.01

[91,] -0.78 0.46 1.39 0.61

[92,] -2.35 -0.26 -0.53 -1.19

[93,] 0.30 0.21 -0.37 1.21

[94,] 1.11 0.37 0.05 1.45

[95,] -0.80 0.71 -1.09 1.06

[96,] -0.11 0.40 0.05 0.35

[97,] -0.21 -0.25 -1.88 -0.32

[98,] -1.59 -1.12 -1.34 1.24

[99,] -0.33 1.90 0.14 -0.12

[100,] -0.63 0.21 -0.75 -0.70

> as.data.frame(scores)

RC1 RC2 RC3 RC4

1 0.13 0.77 -1.88 0.37

2 1.22 -1.65 -0.61 0.81

3 0.62 0.58 0.00 1.57

4 -0.84 -0.27 1.27 -1.25

5 -0.32 -0.83 -0.01 0.45

6 -0.65 -1.07 -1.30 -1.05

7 -2.63 -0.25 -0.56 -1.23

8 -0.28 -0.16 -0.75 -1.01

9 1.05 -0.17 -0.09 -1.66

10 0.43 0.76 -0.45 -0.89

11 -0.14 -0.77 -0.46 0.61

12 -1.45 1.36 0.44 -1.07

13 0.62 2.11 -0.17 0.87

14 0.43 -0.40 0.43 0.90

15 1.44 0.66 -0.27 -1.04

16 0.92 -1.06 -0.56 1.17

17 0.52 -0.32 1.11 -1.03

18 1.71 -0.16 0.25 -1.48

19 1.16 -0.42 -0.38 -1.76

20 0.29 1.78 -0.95 0.24

21 -0.62 -0.18 1.53 -1.83

22 -0.11 2.83 0.63 2.24

23 0.08 -0.35 1.14 1.33

24 1.95 -1.67 -0.86 0.50

25 0.12 -0.02 0.47 -1.25

26 0.57 -0.24 0.62 -1.35

27 0.83 -0.99 1.04 0.92

28 0.12 -1.11 0.38 -1.36

29 1.16 -1.61 -0.06 0.80

30 -0.51 0.16 -1.55 -0.31

31 -0.81 -0.18 2.26 0.22

32 -1.07 -1.60 1.19 -0.07

33 -0.50 0.31 0.16 -0.97

34 0.28 0.07 -0.03 -0.66

35 -1.21 0.61 0.28 -0.69

36 -1.38 -1.06 0.28 1.03

37 -0.62 -0.24 0.31 0.66

38 1.36 0.04 0.11 0.58

39 0.60 0.47 -1.29 -0.45

40 -0.59 1.48 -1.18 -1.04

41 0.19 -0.39 -1.98 -0.60

42 0.04 0.09 -1.17 0.54

43 0.41 1.96 -1.09 0.99

44 0.78 1.61 1.51 -1.15

45 1.27 -1.77 -0.98 0.74

46 1.06 0.68 0.32 -1.10

47 -0.12 -0.09 1.00 1.42

48 2.10 0.46 0.84 -1.68

49 0.16 0.88 -0.84 1.30

50 0.23 0.50 -0.88 1.04

51 -0.94 -0.38 0.19 -0.65

52 1.56 -1.91 -1.18 0.72

53 0.86 -1.09 -0.24 0.87

54 -0.82 -0.53 0.54 0.33

55 0.54 -0.68 -1.06 -0.81

56 -0.37 0.28 0.92 0.60

57 1.98 1.43 -0.09 -0.84

58 0.21 0.52 0.35 0.86

59 -1.34 0.55 0.33 1.94

60 0.85 -1.58 0.57 0.74

61 0.99 -1.26 1.70 0.79

62 -1.10 0.71 -0.15 0.40

63 -0.76 0.26 -1.19 0.78

64 -1.09 -1.95 0.43 -0.15

65 -1.21 0.15 0.58 -0.52

66 1.34 0.54 -1.04 -1.25

67 0.90 -0.59 2.06 -1.32

68 0.42 -0.25 -0.30 -0.85

69 -0.87 -0.60 -1.00 -0.53

70 0.14 -0.15 -1.28 -1.00

71 0.34 2.06 0.69 0.09

72 -1.16 -0.18 -1.21 0.71

73 0.93 1.32 -1.87 -0.56

74 -0.57 1.40 1.23 1.35

75 -0.30 0.87 -0.29 0.30

76 -0.89 0.23 1.04 1.61

77 -0.36 0.14 2.06 -0.63

78 0.21 0.34 1.07 0.31

79 1.13 0.64 0.44 1.47

80 -1.53 0.29 0.03 -0.31

81 -0.85 -0.25 0.45 1.53

82 0.03 -0.92 0.49 0.40

83 -1.39 -0.98 0.21 0.63

84 -2.49 -0.74 1.63 -1.44

85 1.00 -1.78 0.80 -0.01

86 -0.83 -0.42 -1.08 -0.45

87 -1.43 -0.30 -2.16 -1.27

88 1.07 -1.30 1.40 0.04

89 0.09 -0.06 0.13 0.24

90 1.08 2.38 1.89 -1.01

91 -0.78 0.46 1.39 0.61

92 -2.35 -0.26 -0.53 -1.19

93 0.30 0.21 -0.37 1.21

94 1.11 0.37 0.05 1.45

95 -0.80 0.71 -1.09 1.06

96 -0.11 0.40 0.05 0.35

97 -0.21 -0.25 -1.88 -0.32

98 -1.59 -1.12 -1.34 1.24

99 -0.33 1.90 0.14 -0.12

100 -0.63 0.21 -0.75 -0.70

> colnames(scores)=c("Pchexp","Bdrecog","Aftsvc","Prodt")

> print(head(scores))

Pchexp Bdrecog Aftsvc Prodt

[1,] 0.13 0.77 -1.88 0.37

[2,] 1.22 -1.65 -0.61 0.81

[3,] 0.62 0.58 0.00 1.57

[4,] -0.84 -0.27 1.27 -1.25

[5,] -0.32 -0.83 -0.01 0.45

[6,] -0.65 -1.07 -1.30 -1.05

Table for Meaningful names of Principal Components

Components Meaningful Names Column Name

RC1 Purchasing Experience Pchexp

RC2 Brand Recognition Bdrecog

RC3 After Sales Service Aftsvc

RC4 Product Prodt

Explanation

RC1 - Purchasing Experienceexplains about variables affecting Complaint resolution, Order and Billing and delivery speed to customers

RC2 - Brand recognition handles Ecommmerce, image of Sales force , Advertising which is face of the product

RC3 - After Sales Servicegives information about Technical support, and Warranty and claims if there is any problem to customer after he has bought the item

RC4 - Producttalks about the qualities of product like its varities and types, prices its quality i.e all tangible aspects about the very existence of company

Score matrix was converted into a data frame and its variables which are nothing but PCA components were given meaningful names for further analysis We achieved a dimensionality reduction where just 4 factors can explain the complete 11 predictor variables of the hair dataset through PCA analysis

**Multiple Linear Regresion**

> hair\_s=mydata%>%select(c("ID","Satisfaction"))

> hair\_new=cbind(hair\_s,scores)

> print(head(hair\_new))

ID Satisfaction Pchexp Bdrecog Aftsvc Prodt

1 1 8.2 0.13 0.77 -1.88 0.37

2 2 5.7 1.22 -1.65 -0.61 0.81

3 3 8.9 0.62 0.58 0.00 1.57

4 4 4.8 -0.84 -0.27 1.27 -1.25

5 5 7.1 -0.32 -0.83 -0.01 0.45

6 6 4.7 -0.65 -1.07 -1.30 -1.05

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

m.linear.RegModel = lm(Satisfaction ~ Pchexp + Bdrecog + Aftsvc + Prodt, hair\_new)

> summary(m.linear.RegModel)

Call:

lm(formula = Satisfaction ~ Pchexp + Bdrecog + Aftsvc + Prodt,

data = hair\_new)

Residuals:

Min 1Q Median 3Q Max

-1.6346 -0.5021 0.1368 0.4617 1.5235

Coefficients:

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

(Intercept) 6.91813 0.07087 97.617 < 2e-16 \*\*\*

Pchexp 0.61799 0.07122 8.677 1.11e-13 \*\*\*

Bdrecog 0.50994 0.07123 7.159 1.71e-10 \*\*\*

Aftsvc 0.06686 0.07120 0.939 0.35

Prodt 0.54014 0.07124 7.582 2.27e-11 \*\*\*

---

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

Residual standard error: 0.7087 on 95 degrees of freedom

Multiple R-squared: 0.6607, Adjusted R-squared: 0.6464

F-statistic: 46.25 on 4 and 95 DF, p-value: < 2.2e-16

**Summary Explained**

Looking at the Pr(t) values of Coefficients like Intercept (constant beta-naught) we see that it is significant even at 0.001 level. so it definitely not zero and contributes to regression model

Similarly predictor variables like Purchase experience, Brand Recognition and Product have significant betas implying that Response variable Satisfaction is linearly associated with them

After sales service is the only variable which has some high p-value implying that its beta cofficient may not be contributing that significantly to the model or may be zero

All together Adj-R^2 explains that these predictors explains the 64.6 % of the variability in the dataset which is still good enough (may not fall in execellent category)

Overall p-value (extremely less e raise to minus 16) of Model given by F-statistic gives evidence against the null-hypothesis. Model is significantly valid at this point

**Output Interpretation**

pred.Satisfn = predict(m.linear.RegModel)

> as.data.frame(pred.Satisfn)

pred.Satisfn

1 7.465282

2 7.227410

3 8.445062

4 5.671074

5 6.539523

6 5.316752

7 4.463536

8 6.067824

9 6.577679

10 7.060607

11 6.737694

12 6.167033

13 8.835800

14 7.494763

15 7.564792

16 7.540669

17 6.594171

18 7.110605

19 6.444772

20 8.071152

21 5.557027

22 9.545299

23 7.583692

24 7.484183

25 6.338339

26 6.460261

27 7.492680

28 5.717078

29 7.242097

30 6.413479

31 6.595699

32 5.482739

33 6.253981

34 6.768365

35 6.127453

36 6.099842

37 6.789812

38 8.099623

39 7.199285

40 6.667590

41 6.380216

42 7.202198

43 8.632843

44 7.700948

45 7.134571

46 7.347193

47 7.631931

48 7.599196

49 8.111773

50 7.818146

51 5.805062

52 7.218221

53 7.347644

54 6.355465

55 6.396708

56 7.217849

57 8.411216

58 7.800993

59 7.440426

60 7.075531

61 7.427782

62 6.806428

63 6.922795

64 5.197880

65 6.004763

66 7.276893

67 6.598196

68 6.571026

69 5.721393

70 6.302445

71 8.273456

72 6.412083

73 7.738474

74 8.091208

75 7.319033

76 7.424562

77 6.564483

78 7.460264

79 8.766233

80 5.955058

81 7.121857

82 6.716345

83 5.913722

84 4.333171

85 6.676514

86 5.875764

87 5.051047

88 7.031663

89 7.081478

90 8.380018

91 7.093087

92 4.655084

93 7.839444

94 8.579314

95 7.285471

96 7.246518

97 6.362337

98 5.944592

99 7.627616

100 6.207648

> hair\_new = cbind(hair\_new, pred.Satisfn)

> hair\_new$pred.Satisfn = round(hair\_new$pred.Satisfn, 1)

> print(head(hair\_new))

ID Satisfaction Pchexp Bdrecog Aftsvc Prodt pred.Satisfn

1 1 8.2 0.13 0.77 -1.88 0.37 7.5

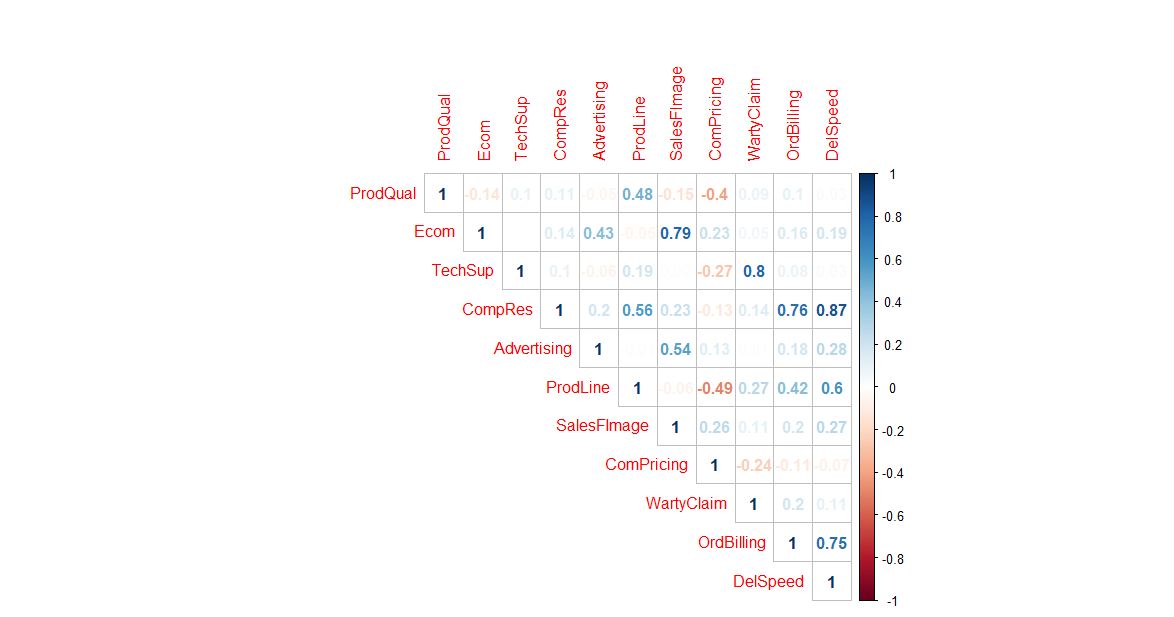
2 2 5.7 1.22 -1.65 -0.61 0.81 7.2

3 3 8.9 0.62 0.58 0.00 1.57 8.4

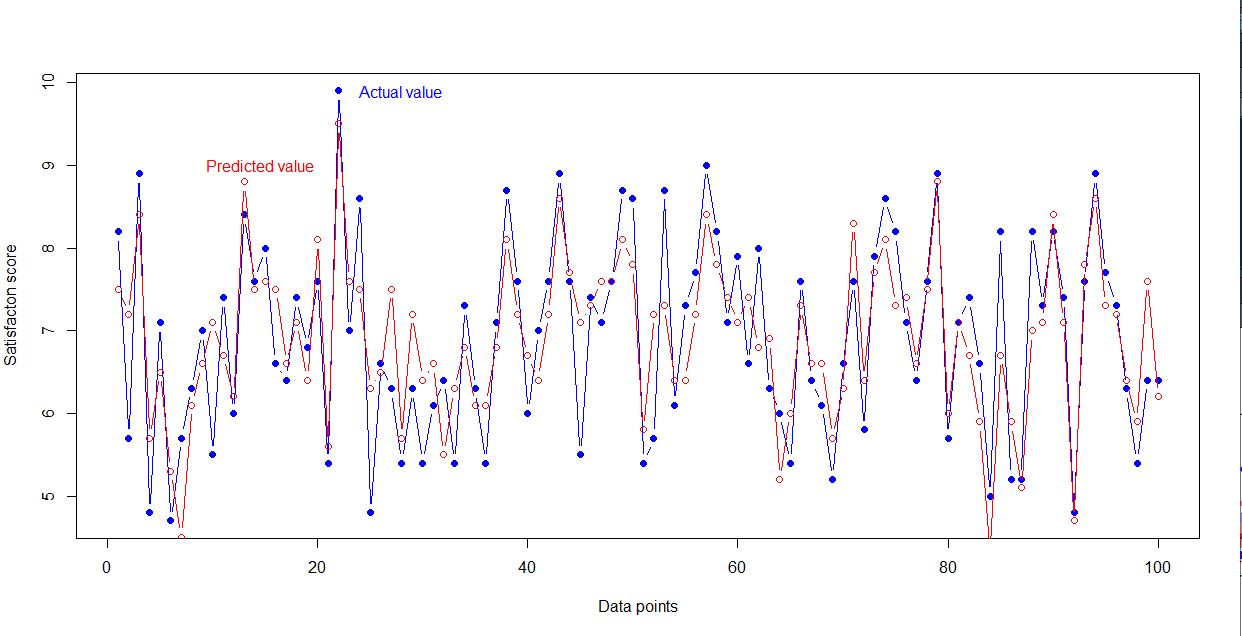
4 4 4.8 -0.84 -0.27 1.27 -1.25 5.7

5 5 7.1 -0.32 -0.83 -0.01 0.45 6.5

6 6 4.7 -0.65 -1.07 -1.30 -1.05 5.3



**Predicted & Actual Satisfaction**



Plot analysis revealed that our new MLR Regression model is quite good and close to actual Satisfaction scores Blue dots represent Actual Satisfaction ratings Red dots represent Predicted satisfaction scores derived from multiple linear regression model

**Conclusion**

Nutshell we can say that the " Satisfaction" ratings of hair product depends very highly on the overall Purchasing experience of the Customer i.e. how quickly his product is delivered, its billed and if there are complaints are resolved in shortest possible time

Brand Recognition or products adverstising comes in second in mind. Product itself comes in third in order to satisfy the customer (though statistically it variance explaination capacity has been ranked 4th ) and After sales contribution in fourth

There can be differences in the real operating world and these statistical models but this model comes closest to explaining the data provided for deduction.

In this article, we saw how Factor Analysis can be used to reduce the dimensionality of a dataset and then we used multiple linear regression on the dimensionally reduced columns/Features for further analysis/predictions.

THANK YOU