Mini Project – Factor-Hair

Table of Contents

1. Project Objective

The objective of the report is to explore the Factor-Hair data set (“Factor-Hair.csv”) in R and prepare a Managerial Report by explaining the following points. This Managerial report consists of the following:

- Is there evidence of Multicollinearity?
- Perform Factor Analysis by extracting four factors
- Name the factors
- Perform Multiple Liner Regression with Customer Satisfaction as dependant variable and the four factors as the independent variables. Comment on the Model validity.

The sample Factor-Hair dataset is:

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
11	8.7	3.2	4.6	4.8	2.7	6.8	4.6	6.8	5.8	3.8	4	7.4
12	6.1	4.9	6.3	3.9	4.4	3.9	6.4	8.2	5.8	3	3.2	6
13	9.5	5.6	4.6	6.9	5	6.9	6.6	7.6	6.5	5.1	4.4	8.4
14	9.2	3.9	5.7	5.5	2.4	8.4	4.8	7.1	6.7	4.5	4.2	7.6
15	6.3	4.5	4.7	6.9	4.5	6.8	5.9	8.8	6	4.8	5.2	8
16	8.7	3.2	4	6.8	3.2	7.8	3.8	4.9	6.1	4.3	4.5	6.6
17	5.7	4	6.7	6	3.3	5.5	5.1	6.2	6.7	4.2	4.5	6.4
18	5.9	4.1	5.5	7.2	3.5	6.4	5.5	8.4	6.2	5.7	4.8	7.4
19	5.6	3.4	5.1	6.4	3.7	5.7	5.6	9.1	5.4	5	4.5	6.8
20	9.1	4.5	3.6	6.4	5.3	5.3	7.1	8.4	5.8	4.5	4.4	7.6
21	5.2	3.8	7.1	5.2	3.9	4.3	5	8.4	7.1	3.3	3.3	5.4
22	9.6	5.7	6.8	5.9	5.4	8.3	7.8	4.5	6.4	4.3	4.3	9.9
23	8.6	3.6	7.4	5.1	3.5	7.3	4.7	3.7	6.7	4.8	4	7
24	9.3	2.4	2.6	7.2	2.2	7.2	4.5	6.2	6.4	6.7	4.5	8.6
25	6	4.1	5.3	4.7	3.5	5.3	5.3	8	6.5	4.7	4	4.8
26	6.4	3.6	6.6	6.1	4	3.9	5.3	7.1	6.1	5.6	3.9	6.6
27	8.5	3	7.2	5.8	4.1	7.6	3.7	4.8	6.9	5.3	4.4	6.3
28	7	3.3	5.4	5.5	2.6	4.8	4.2	9	6.5	4.3	3.7	5.4
29	8.5	3	5.7	6	2.3	7.6	3.7	4.8	5.8	5.7	4.4	6.3
30	7.6	3.6	3	4	5.1	4.2	4.6	7.7	4.9	4.7	3.5	5.4
31	6.9	3.4	8.5	4.3	4.5	6.4	4.7	5.2	7.7	3.7	3.3	6.1
32	8.1	2.5	7.2	4.5	2.3	5.1	3.8	6.6	6.8	3	3	6.4

Note: This is a sample data set of 32 rows. The actual provided data set has 100 rows.

2. Assumptions

Since, the data provided in the dataset is continuous and appears on same scale, it is not required to scale the data.

3. Step by step approach

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

1. Environment set up and Data import
2. Identifying dependant and independent variables

3. Identify Correlation between independent variables
4. Perform Dimension Reduction test (Bartlett Test)
5. Perform Sampling adequacy test (KMO Test)
6. Perform Factor Analysis
7. Name the factors
8. Perform Multiple Linear Regression
9. Interpreting the Result and Model Validity
10. Conclusion

3.1. Environment Set up and Data Import

Please refer Appendix A for Source Code.

3.2. Variable Identification

In the given data set, first column is an ID column, which is not considered as a variable.

13th Column "**Customer Satisfaction**" is the dependant variable.

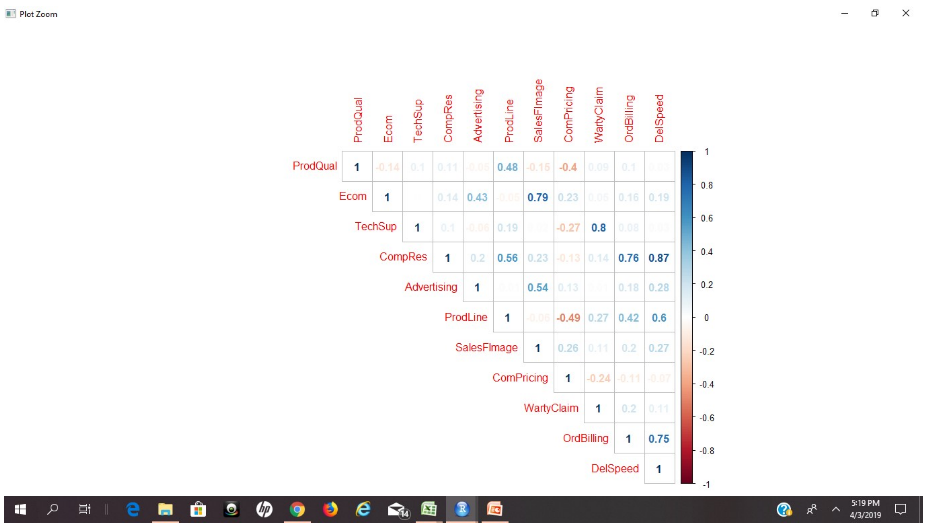
These are the Independent variables with their expansion in the given data set.

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

3.3. Identify Correlation between Independent variables

A Correlation matrix gives the correlation scores of each variable against each variable.Also, "corrplot" method of "corrplot" Package is used to obtain the correlation diagram with the correlation scores as shown below.

The highlighted values in the correlation diagram shows that some of these independent variables are highly correlated ith each other. For ex: ProdQual is correlated with ProdLine and ComPricing, Ecom is correlated with Advertising and SalesFImage, TechSup is correlated with WartyClaim etc.



Hence, the correlation scores in the above diagram confirms that the independent variables are multicollinear.

3.4. Perform Dimension Reduction Test

As there exists multicollinearity between the independent variables, a dimension reduction test Bartlett test for sphericity has been performed to confirm the scope for dimension reduction of the variables.

The Bartlett test uses p-value to confirm the scope for dimension reduction. If there is a p-value, there exists a Null Hypothesis and Alternate Hypothesis.

Null Hypothesis	There is no scope of dimension reduction
Alternate Hypothesis	There is scope of dimension reduction.

Bartlett test gives the following result:

```
> cortest.bartlett(matrix, 100)
$chisq
[1] 619.2726

$p.value
[1] 1.79337e-96

$df
[1] 55
```

In the above result, the extremely low value of p-value < 0.05, hence the null hypothesis is rejected and the Alternate hypothesis is accepted.

Please refer Appendix A for Source Code.

3.5. Perform Sampling Adequacy Test

Since, the independent variables are multicollinear and there exists scope for dimension reduction, the next step is to test if the given sample data size is adequate for performing Factor Analysis to reduce the dimensions.

A KMO Test is performed to test the sampling adequacy and the result is as shown below:

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

The KMO test shows an Overall MSA value and if the Overall MSA > 0.5, the sample size is good enough to perform Factor Analysis.

In the above result, the Overall MSA is 0.65 which is > 0.5, hence, the given sample size is good enough to perform Factor Analysis.

Please refer Appendix A for Source Code.

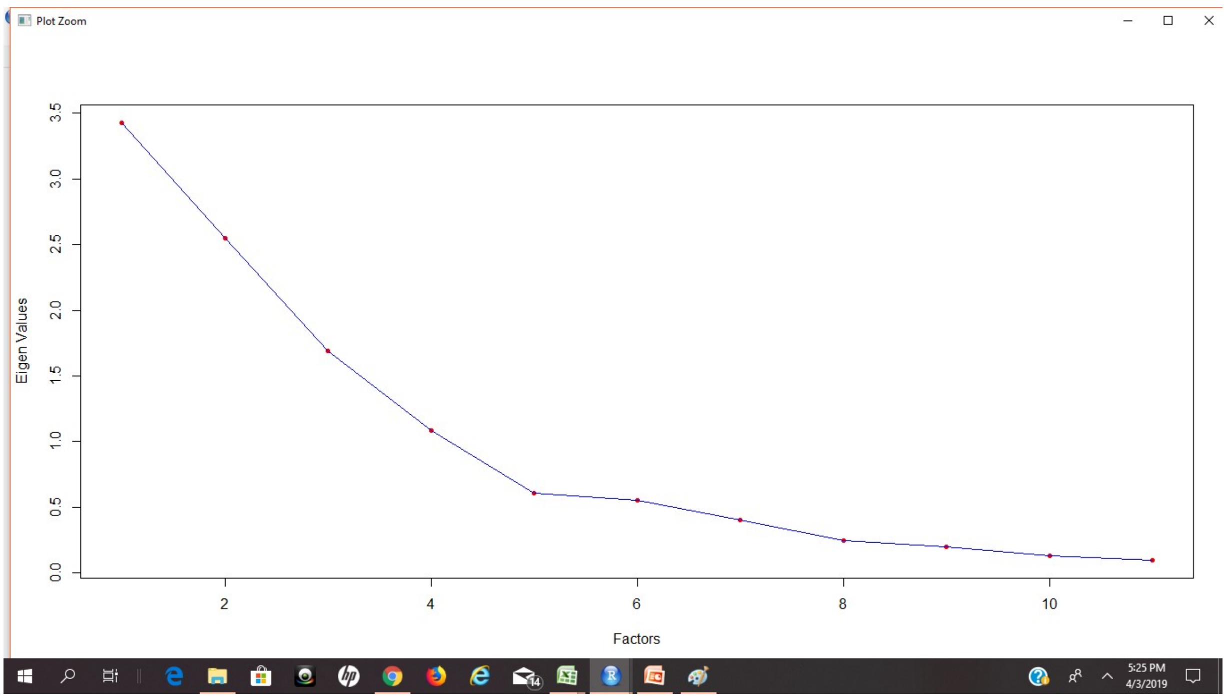
3.6. Perform Factor Analysis

3.6.1. Calculate Eigen Values

Eigen values of independent variables are used to perform Factor Analysis and the Eigen values of the independent variables are calculated as shown below.

```
> eigenValues
 [1] 3.42697133 2.55089671 1.69097648 1.08655606 0.60942409 0.55188378 0.40151815 0.24695154
 [9] 0.20355327 0.13284158 0.09842702
```

As per the Kaiser rule, the values which are > 1 are considered as the Principal Component Factors.
In the above result, the first 4 values are > 1 and these are considered as Principal Component Factors.
A Scree plot is used to visualize and select the Principal Component Factors.
Scree plot uses Eigen values on the Y-Axis and Factors on the X-Axis.



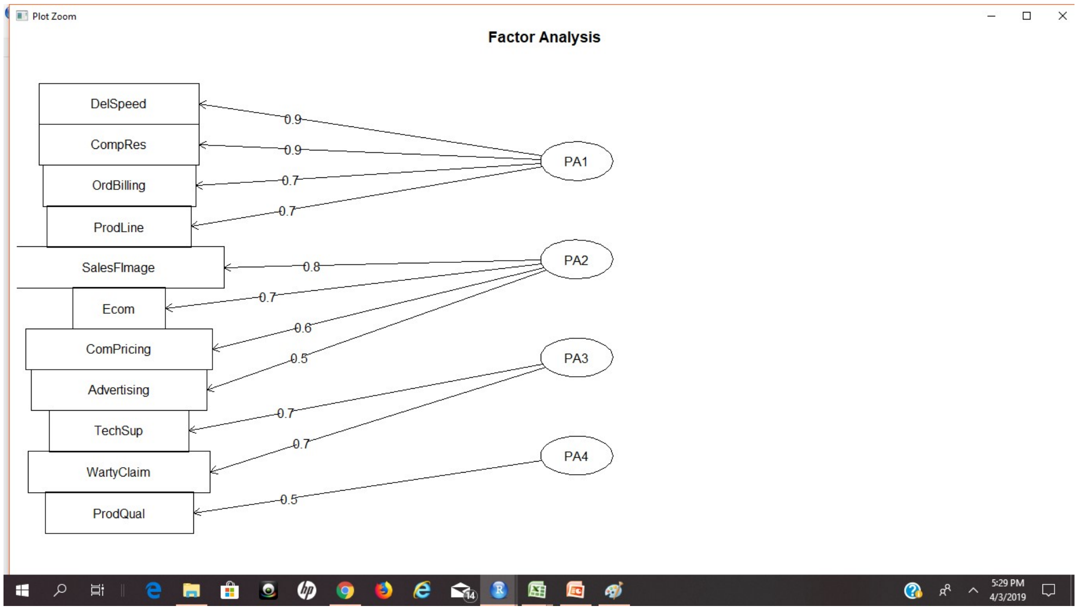
Please refer Appendix A for Source Code.

3.6.2. Perform Factor Analysis

Perform Factor Analysis with the 4 principal component factors. The factor scores associated with each Principal component factor as shown below:

```
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
>
```

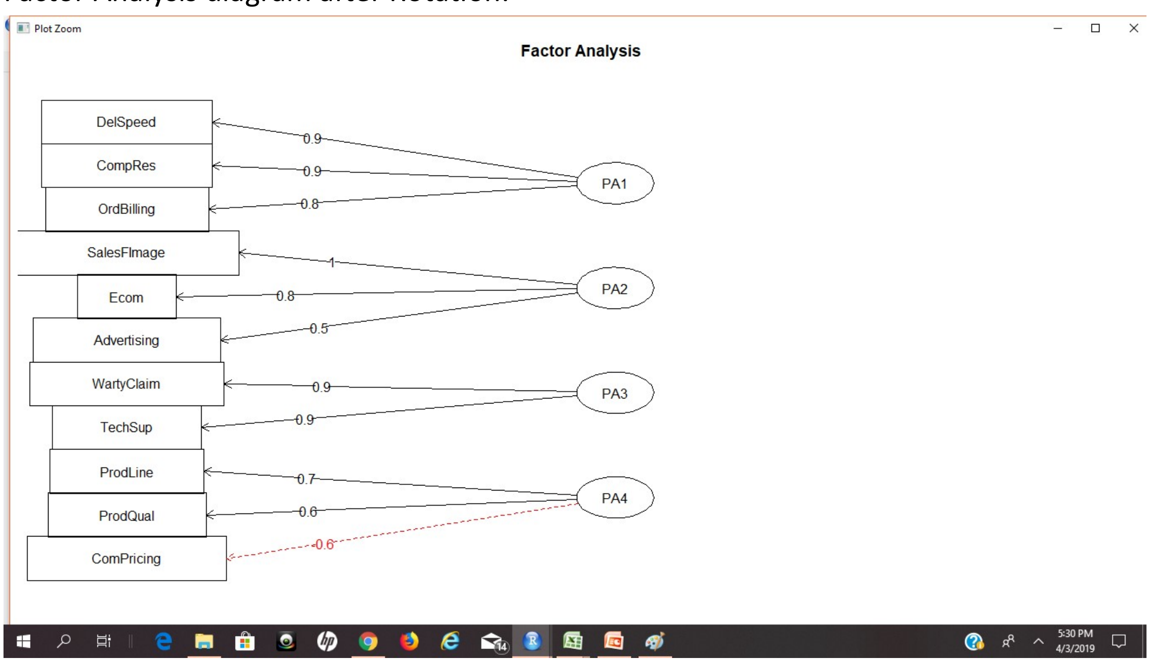
A Factor analysis diagram also shows the Independent variables and the factor scores associated with each principal component.



Rotate the plot using 'varimax' to see if there is any change in the factor mapping. The below diagram shows Factor scores and factor diagram after the rotation.

```
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 diagram after Rotation:



There is a change in the mapping of factors into Principal Component Factors and their factor scores. This mapping seems to be more relevant than the previous mapping. Hence, the Principal Component Factors along with their correlated independent variables have been identified as

PA1	DelSpeed, CompRes, OrdBilling
PA2	SalesFImage, Ecom, Advertising
PA3	WartyClaim, TechSup
PA4	ProdLine, ProdQuality, ComPricing

Please refer Appendix A for Source Code.

3.7. Name the Factors

Naming the Principal component factors is subjective and in this example, the names have been identified as follows:

Dependant Variable	Customer Satisfaction
PA1	Customer Service
PA2	Product Marketing
PA3	Product Support
PA4	Product Line Pricing

Now, bind the dependant variable 'Customer Satisfaction' With the newly derived Principal component factors and their scores are as shown below.

```
> head(finalData)
  Customer Satisfaction Customer Service Product Marketing Product Support
1                   8.2       -0.1338871         0.9175166    -1.719604873
2                   5.7        1.6297604        -2.0090053    -0.596361722
3                   8.9        0.3637658         0.8361736     0.002979966
4                   4.8       -1.2225230        -0.5491336     1.245473305
5                   7.1       -0.4854209        -0.4276223    -0.026980304
6                   4.7       -0.5950924        -1.3035333    -1.183019401
  Product Line Pricing
1           0.09135411
2           0.65808192
3           1.37548765
4          -0.64421384
5           0.47360747
6          -0.95913571
```

Please refer Appendix A for Source Code.

3.8. Perform Multiple Linear Regression

From the above Factor Analysis, Customer Satisfaction is dependant on 4 independent variables, we will use Multiple Linear Regression to build a model to predict the future analytics of the Market Segmentation in the context of Product Service Management.

Following are the assumptions of Multiple Linear Regression.

i) **Linear Relationship:** There must be a linear relation between the dependant and Independent variables. Scatter plots can show whether there is linear or curvilinear relationship.

ii) **Multivariate Normality:** Multiple regression assumes that the residuals are Normally distributed.

iii) **No Multicollinearity:** Multiple Regression assumes that the independent Variables are not highly correlated. This is tested with Variance Inflation Factor (VIF) values.

iv) **Homoscedasticity**: This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

First step is to build a liner regression model using the dependat variable and the derived 4 principal component factors as independent variables.

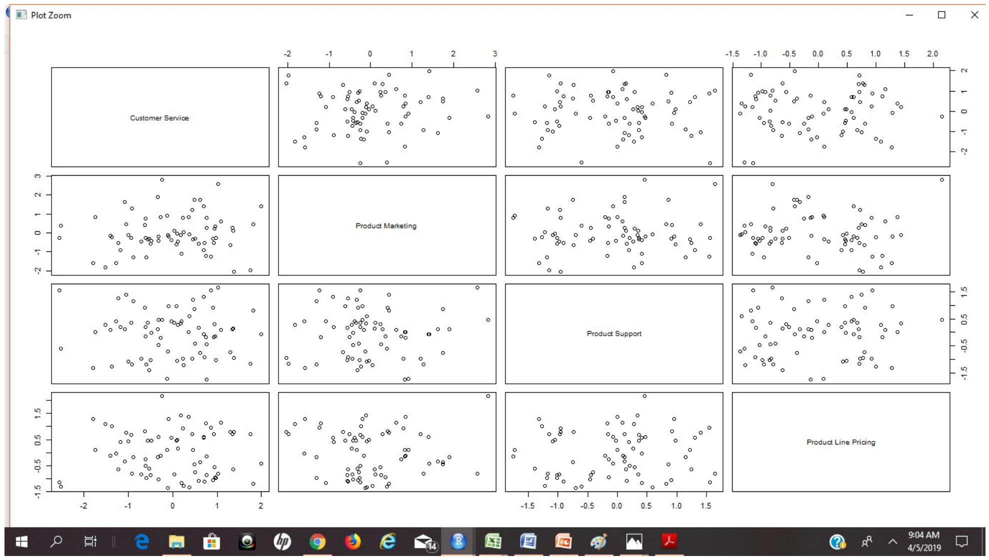
Set the seed value to 300 and split the data into 70:30 as Training and Test Data and build a linear regression model with Training Data and use Test Data to predict the model.

Please refer Appendix A for Source Code to split the data into Training and Test data and to build a liner model with Training data.

Once the model is built, test for assumptions of Multiple Liner Regression.

i) Linear Relationship:

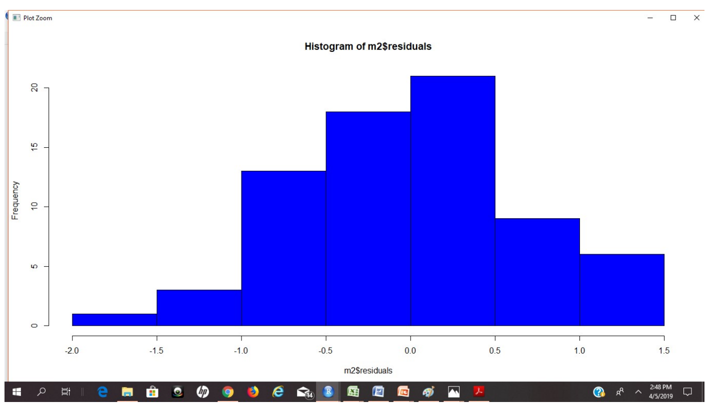
Following scatter plots explain the linear relationship between the independent variables.

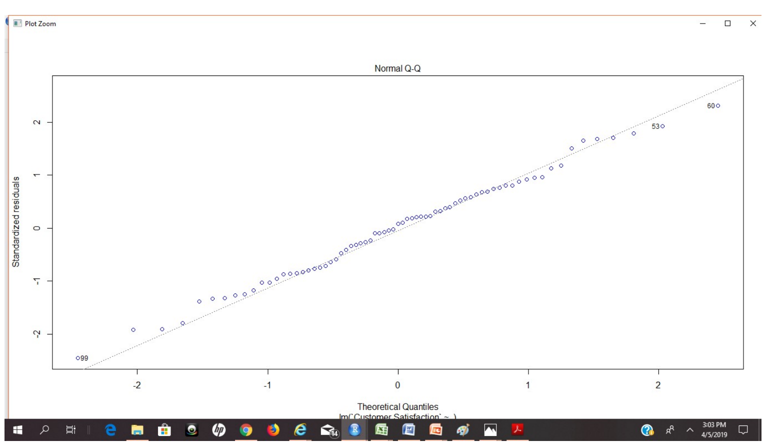


From the above plot, it appears that the independent variables are not correlated.

ii) Multivariate Normality

The following Histogram of residuals and the Normal Q-Q plot of the regression model explains the normality of the residuals.





As per the histogram and Normal Q-Q plot, the residuals are normally distributed.

iii) No Multicollinearity:

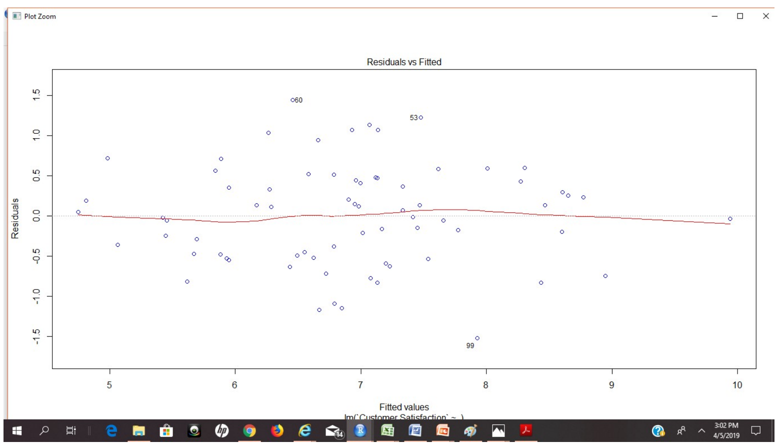
Multicollinearity between regression coefficients can also be tested with Variance Inflation factors as shown below.

```
vif(m2)
  `Customer Service`    `Product Marketing`      `Product Support` `Product Line Pricing`
            1.010464               1.013471               1.008309               1.008043
```

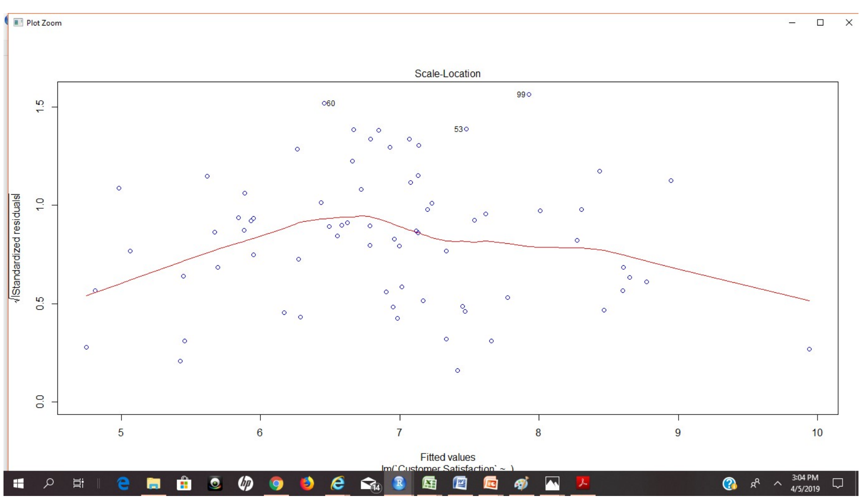
In the above result, as the values are < 5, the coefficients are not multicollinear.

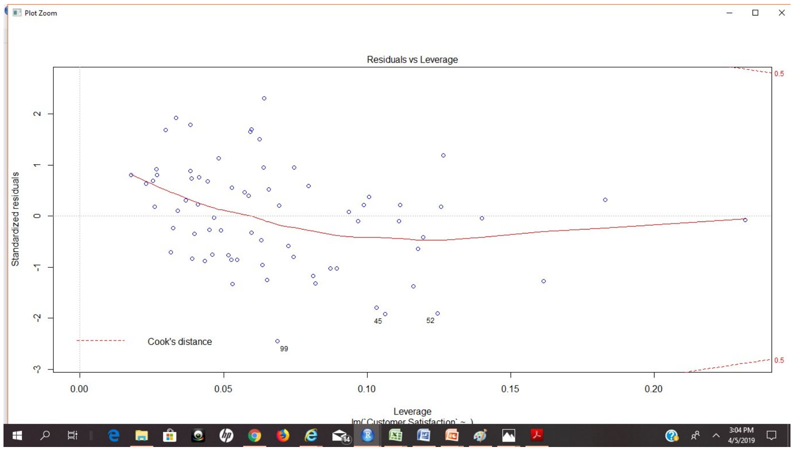
iv) Homoscedasticity:

In the following 'Residuals vs fitted values' plot it can be seen that residuals are linearly distributed and hence variance is uniform.



The Scale-Location plot and Residual vs Leverage plots of the regression model are shown below.





3.9. Interpreting the result and Model Validity

When we apply the regression on Train Data, the result is:

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.92249    0.07699  89.917  < 2e-16 ***
`Customer Service`        0.56394    0.07452   7.567 1.58e-10 ***
`Product Marketing`       0.67726    0.07606   8.904 6.44e-13 ***
`Product Support`         0.09038    0.08404   1.075    0.286
`Product Line Pricing`    0.55484    0.08576   6.470 1.40e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6454 on 66 degrees of freedom
Multiple R-squared:  0.7366,    Adjusted R-squared:  0.7206
F-statistic: 46.13 on 4 and 66 DF,  p-value: < 2.2e-16
```

As per the above result, the adjusted R squared value is 0.7206, which means,
The four independent variables/factors cause 72.06% variation in the ‘Customer Satisfaction’(Dependant Variable). Also the p-values of 3 independent variables/coefficients is much < 0.05, meaning that these three coefficients(Customer Service, Product Marketing, and Product Line Pricing) are highly significant. Hence, the model is robust and statistically significant.

In order to say that the model is valid, we need to predict the model with Test data and calculate the R Square value.

Below mentioned is the Calculation for R-Squared value.

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

Where

$$SSR = \sum(\hat{y} - \overline{y})^2 \text{ (measure of explained variation)}$$

$$SSE = \sum(y - \hat{y}) \quad \text{(measure of unexplained variation)}$$

$$SST = SSR + SSE = \sum(y - \overline{y})^2 \text{ (measure of total variation in y)}$$

As per the above formula, the R-Squared value calculated by R is

```
> pred = predict(m2, newdata = Test)
> SST = sum((Test$`Customer Satisfaction` - mean(Train$`Customer Satisfaction`))^2)
> SST
[1] 36.27184
> #SSR
> SSE = sum((pred - Test$`Customer Satisfaction`)^2)
> SSE
[1] 15.80878
> #SSR
> SSR = sum((pred - mean(Train$`Customer Satisfaction`))^2)
> SSR
[1] 22.04063
> SSR/SST
[1] 0.6076513
```

In the above Result, the R-Squared value is 0.6077, which is in the range of +/-15 of R-Squared value of Training Data Set. Hence, the model is statistically significant.

3.10. Conclusion

- From the preliminary Data Analysis, we confirmed that there are 11 continuous independent variables , 1 continuous dependant variable(Customer Satisfaction) and there exists multicollinearity between the Independent variables.
- With the help of Factor Analysis, we have reduced the dimensions to 4 Principal Component Factors which are majorly contributing to the variance in Customer Satisfaction variable.
- Further, we have split the data into two data sets in 70:30 ratio, Training and Test data sets. With the help of Multiple Linear Regression, we have built Model with Training data set which shows 72% variance in the dependant variable that is Customer Satisfaction.
- We have also tested all the assumptions of Multiple Linear Regression successfully with the help of Scatter plots, Histogram, Normality Q-Q Plot and Variance Inflation Factor(VIF).
- Finally, we have predicted the model with the Test data by calculating R-Squared value which shows 60% of variation in the dependant variable by the independent variables(4 Principal Component Factor Coefficients).

➢ As both the R-Squared values of Training data set and Test data set seems to be high, we conclude that the model is robust and statistically significant and hence can be used in the Predictor Analysis of Market Segmentation in the context of Product Service Management.

4. Appendix A – Source Code

```
#==============================================================================
# Data Analysis - Factor-Hair
#==============================================================================
#Environment Set up and Data Import
#Set up working Directory
setwd("C:/Users/Radhika/Desktop/R Programming/Project_AdvanceStatistics_Module2")
getwd()
#
#Read the input file
hair = read.csv("Factor-Hair-Revised.csv")
attach(hair)
#Find the internal structure of the data
str(hair)
#Find the descriptive statistics of the data
summary(hair)
#Find the column names of the data
names(hair)
#Column 1 is an ID column, coulumns 2 to 12 are the independant variables
#and column 13 is the dependant variable. Since we are going to find the
#correlation between independant variables, Discard column 1 and 13.
hair1=hair[,2:12]
#Get the correlation matrix of all independant variables to test
#the multicollinearity between the independant variables.
matrix=cor(hair1)
#Display the correlation matrix
matrix
# Install packages "car" and "corrplot" to visualize the correlation between
# the variables diagramatically.
install.packages("car")
library(car)
#vif(m1)
install.packages("corrplot")
library(corrplot)
#Display the correlation diagram
corrplot(matrix, type="upper", method="number")
```

```r
#
#Load "psych" package to use bartlett test to check if there is a need
# for dimension reduction in the given data set.
library(psych)
# Bartlett test gives a p-value and if the p-value is < 0.05, it confirms the
# scope for dimension reduction in the data set.
cortest.bartlett(matrix, 100)
# The above test results in p-value of 1.79337e-96 which is < 0.05.
# Hence, we need to perform Factor analysis for reducing the factors to dimensions.
#
#Perform KMO Test to measure the Sampling adequacy.
# KMO Test Specifies if overall MSA > 0.5, the sample size is good enough
# for performing FActor Analysis.
KMO(matrix)
# KMO Test results in overall MSA = 0.65, hence, the given sample size is
# good enough for performing Factor Analsis.
#
# Find out the eigen values to identify the Principal factors which are
# contributing the most.
eigenvector = eigen(matrix)
# Extract the eigen values and display them.
eigenValues = eigenvector$values
eigenValues
# The first four values of the eigen values are > 1 which are contributing the
# most to derive the customer satisfaction. As per the Kaiser rule, select these
# variables as the principal factors.
# Visualize the principal factors in the scree plot.
plot(eigenValues, xlab="Factors", ylab = "Eigen Values", col="red", pch=20)
lines(eigenValues, col="blue")
#
#Apply the Factor Analysis with the selected 4 principal factors and
#check the factor scores associated with each principal factor
fa1=fa(r=hair1, nfactors = 4, rotate="none", fm="pa")
print(fa1)
#Print the Factor Analysis diagram
fa.diagram(fa1)
#
#As per the diagram, the factors doesn't seem to be mapped properly.
#Hence, rotate using varimax and check the factor analysis scores and diagram.
fa2=fa(r=hair1, nfactors = 4, rotate="varimax", fm="pa")
print(fa2)
#After rotation, the factors seem to be mapped to the principal factors
#properly with commonality between the factors.
fa.diagram(fa2)
# Get the attributes of the derived factors and print the scores associated
# with the new derived factors.
attributes(fa2)
fa2$scores
# Now bind the dependant variable customer satisfaction with the newly derived
# independant variables using their scores.
finalData = cbind(hair[,13], fa2$scores)
#Print the above matrix with headers. As the headers doesn't seem to be meaningful,
#assign some meaningful names to the derived factors.
head(finalData)
#Map the column names with meaningful names.
colnames(finalData) = c("Customer Satisfaction", "Customer Service", "Product Marketing",
                        "Product Support", "Product Line Pricing")
head(finalData)
class(finalData)
#Since the final data is in matrix form, convert it into a data frame.
finalData = as.data.frame(finalData)
finalData
#Now, split the data into two sets, Train and Test to predict the model with new data.
#In this case, we build the model with Training data and predict the model with Test data.
install.packages("caTools")
library(caTools)
set.seed(300)
#Split the data into 70:30
spl = sample.split(finalData$`Customer Satisfaction`, SplitRatio = 0.7)
Train = subset(finalData, spl==T)
```

```
Test = subset(finalData, spl==F)
dim(Train)
dim(Test)
plot(Train[c(2,3,4,5)])
#Apply regression to the train data and check the R Square value.
m2 = lm(`Customer Satisfaction` ~., data = Train)
#
# Plot regression model
plot(m2, col="blue")
# Plot Histogram of residuals to check for normality
hist(m2$residuals, col="blue")
summary(m2)
#Residuals:
#Min       1Q   Median       3Q      Max
#-1.52566 -0.48763  0.04769  0.43246  1.44382
#
#Coefficients:
#Estimate Std. Error t value Pr(>|t|)
#(Intercept)             6.92249    0.07699  89.917  < 2e-16 ***
#  `Customer Service`      0.56394    0.07452   7.567 1.58e-10 ***
#  `Product Marketing`     0.67726    0.07606   8.904 6.44e-13 ***
#  `Product Support`       0.09038    0.08404   1.075    0.286
#`Product Line Pricing`  0.55484    0.08576   6.470 1.40e-08 ***
 # ---
#  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#Residual standard error: 0.6454 on 66 degrees of freedom
#Multiple R-squared:  0.7366, Adjusted R-squared:  0.7206
#F-statistic: 46.13 on 4 and 66 DF,  p-value: < 2.2e-16
#
#The above result shows that the adjusted R squared value is 0.7206
#Which says, there is 72.06% variation present in Customer Satisfation dependant variable
#by the four independant variables. Also, the p-values of 3 independant variables
#is much less than 0.05 meaning that the three independant variables are highly significant.
#Hence, the regression model is robust and statistically significant.
#
#variance inflation factor (or VIF), which measures how much the variance of a
#regression coefficient is inflated due to multicollinearity in the model.
vif(m2)
#`Customer Service`     `Product Marketing`   `Product Maintenance` `Product Line Pricing`
#1.016756               1.025960              1.033948              1.028339
#In the above result, the VIF values of regression coefficients are less than 5 which
#shows the coefficients are not multicollinear.
#
#Predict the model with test data by calculating R-squared value
pred = predict(m2, newdata = Test)
#
#SST
SST = sum((Test$`Customer Satisfaction` - mean(Train$`Customer Satisfaction`))^2)
SST
#SSR
SSE = sum((pred - Test$`Customer Satisfaction`)^2)
SSE
#SSR
SSR = sum((pred - mean(Train$`Customer Satisfaction`))^2)
SSR
SSR/SST
#The above result gives an R Square value of 0.6080142, which is in the range of +/-15 of
#Training R-Squared value.
#Hence, this model is statistically significant.
```