
# Mini Project - Customer Satisfaction Survey

## Using PCA & Multiple Linear Regression Method in R

Contents

# 1 Project Objective:

The Project objective is to find the answer for the following questions.

In the given data set of Hair, is there any evidence of Multicollinearity?

Perform Factor Analysis by extracting four factors

Name the factors

Perform Multiple Linear Regression with the customer satisfaction as the dependent variable and the four factors as the independent variables. Comment on Model validity.

## The Dataset:

| 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 |
|  | 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 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **33** | 6.7 | 3.7 | 6.5 | 5.3 | 5.3 | 5.1 | 4.9 | 9.2 | 5.7 | 3.5 | 3.4 | 5.4 |
| **34** | 8 | 3.3 | 6.1 | 5.7 | 5.5 | 4.6 | 4.7 | 8.7 | 5.9 | 4.7 | 4.2 | 7.3 |
| **35** | 6.7 | 4 | 5.2 | 3.9 | 3 | 5.4 | 6.8 | 8.4 | 6.2 | 2.5 | 3.5 | 6.3 |
| **36** | 8.7 | 3.2 | 6.1 | 4.3 | 3.5 | 6.1 | 2.9 | 5.6 | 6.1 | 3.1 | 2.5 | 5.4 |
| **37** | 9 | 3.4 | 5.9 | 4.6 | 3.9 | 6 | 4.5 | 6.8 | 6.4 | 3.9 | 3.5 | 7.1 |
| **38** | 9.6 | 4.1 | 6.2 | 7.3 | 2.9 | 7.7 | 5.5 | 7.7 | 6.1 | 5.2 | 4.9 | 8.7 |
| **39** | 8.2 | 3.6 | 3.9 | 6.2 | 5.8 | 4.9 | 5 | 9 | 5.2 | 4.7 | 4.5 | 7.6 |
| **40** | 6.1 | 4.9 | 3 | 4.8 | 5.1 | 3.9 | 6.4 | 8.2 | 5.1 | 4.5 | 3.2 | 6 |
| **41** | 8.3 | 3.4 | 3.3 | 5.5 | 3.1 | 4.6 | 5.2 | 9.1 | 4.1 | 4.6 | 3.9 | 7 |
| **42** | 9.4 | 3.8 | 4.7 | 5.4 | 3.8 | 6.5 | 4.9 | 8.5 | 4.9 | 4.1 | 4.1 | 7.6 |
| **43** | 9.3 | 5.1 | 4.6 | 6.8 | 5.8 | 6.6 | 6.3 | 7.4 | 5.1 | 4.6 | 4.3 | 8.9 |
| **44** | 5.1 | 5.1 | 6.6 | 6.9 | 4.4 | 5.4 | 7.8 | 5.9 | 7.2 | 4.9 | 4.5 | 7.6 |
| **45** | 8 | 2.5 | 4.7 | 7.1 | 3.6 | 7.7 | 3 | 5.2 | 5.1 | 4.3 | 4.7 | 5.5 |
| **46** | 5.9 | 4.1 | 5.7 | 5.9 | 5.8 | 6.4 | 5.5 | 8.4 | 6.4 | 5.2 | 4.8 | 7.4 |
| **47** | 10 | 4.3 | 7.1 | 6.3 | 2.9 | 5.4 | 4.5 | 3.8 | 6.7 | 5 | 3.5 | 7.1 |
| **48** | 5.7 | 3.8 | 6.8 | 7.5 | 5.7 | 5.7 | 6 | 8.2 | 6.6 | 6.5 | 5.2 | 7.6 |
| **49** | 9.9 | 3.7 | 3.7 | 6.1 | 4.2 | 7 | 6.7 | 6.8 | 5.9 | 4.5 | 3.9 | 8.7 |
| **50** | 7.9 | 3.9 | 4.3 | 5.8 | 4.4 | 6.9 | 5.8 | 4.7 | 5.2 | 4.1 | 4.3 | 8.6 |
| **51** | 6.7 | 3.6 | 5.9 | 4.2 | 3.4 | 4.7 | 4.8 | 7.2 | 5.7 | 4 | 2.8 | 5.4 |
| **52** | 8.2 | 2.7 | 3.7 | 7.4 | 2.7 | 7.9 | 3.1 | 5.3 | 5.3 | 4.5 | 4.9 | 5.7 |
| **53** | 9.4 | 2.5 | 4.8 | 6.1 | 3.2 | 7.3 | 4.6 | 6.3 | 6.3 | 4.7 | 4.6 | 8.7 |
| **54** | 6.9 | 3.4 | 5.7 | 4.4 | 3.3 | 6.4 | 4.7 | 5.2 | 6.4 | 3.2 | 3.3 | 6.1 |
| **55** | 8 | 3.3 | 3.8 | 5.8 | 3.2 | 4.6 | 4.7 | 8.7 | 5.3 | 4.9 | 4.2 | 7.3 |
| **56** | 9.3 | 3.8 | 7.3 | 5.7 | 3.7 | 6.4 | 5.5 | 7.4 | 6.6 | 4.1 | 3.4 | 7.7 |
| **57** | 7.4 | 5.1 | 4.8 | 7.7 | 4.5 | 7.2 | 6.9 | 9.6 | 6.4 | 5.7 | 5.5 | 9 |
| **58** | 7.6 | 3.6 | 5.2 | 5.8 | 5.6 | 6.6 | 5.4 | 4.4 | 6.7 | 4.6 | 4 | 8.2 |
| **59** | 10 | 4.3 | 5.3 | 3.7 | 4.2 | 5.4 | 4.5 | 3.8 | 6.7 | 3.7 | 3.5 | 7.1 |
| **60** | 9.9 | 2.8 | 7.2 | 6.9 | 2.6 | 5.8 | 3.5 | 5.4 | 6.2 | 5.6 | 4 | 7.9 |
| **61** | 8.7 | 3.2 | 8.4 | 6.1 | 2.8 | 7.8 | 3.8 | 4.9 | 7.2 | 5.4 | 4.5 | 6.6 |
| **62** | 8.4 | 3.8 | 6.7 | 5 | 4.5 | 4.7 | 5.9 | 6.7 | 5.1 | 2.7 | 3.6 | 8 |
| **63** | 8.8 | 3.9 | 3.8 | 5.1 | 4.3 | 4.7 | 4.8 | 5.8 | 5 | 4.4 | 2.9 | 6.3 |
| **64** | 7.7 | 2.2 | 6.3 | 4.5 | 2.4 | 4.7 | 3.4 | 6.2 | 6 | 3.3 | 2.6 | 6 |
| **65** | 6.6 | 3.6 | 5.8 | 4.1 | 4.9 | 4.7 | 4.8 | 7.2 | 6.5 | 3.5 | 2.8 | 5.4 |
| **66** | 5.7 | 3.8 | 3.5 | 6.7 | 5.4 | 5.7 | 6 | 8.2 | 5.4 | 4.7 | 5.2 | 7.6 |
| **67** | 5.7 | 4 | 7.9 | 6.4 | 2.7 | 5.5 | 5.1 | 6.2 | 7.5 | 5 | 4.5 | 6.4 |
| **68** | 5.5 | 3.7 | 4.7 | 5.4 | 4.3 | 5.3 | 4.9 | 6 | 5.6 | 4.5 | 4.3 | 6.1 |
| **69** | 7.5 | 3.5 | 3.8 | 3.5 | 2.9 | 4.1 | 4.5 | 7.6 | 5.1 | 4 | 3.4 | 5.2 |
| **70** | 6.4 | 3.6 | 2.7 | 5.3 | 3.9 | 3.9 | 5.3 | 7.1 | 5.2 | 4.7 | 3.9 | 6.6 |
| **71** | 9.1 | 4.5 | 6.1 | 5.9 | 6.3 | 5.3 | 7.1 | 8.4 | 7.1 | 5.4 | 4.4 | 7.6 |
| **72** | 6.7 | 3.2 | 3 | 3.7 | 4.8 | 6.3 | 4.5 | 5 | 5.2 | 2.9 | 3.1 | 5.8 |
| **73** | 6.5 | 4.3 | 2.7 | 6.6 | 6.5 | 6.3 | 6 | 8.7 | 4.7 | 4.6 | 4.6 | 7.9 |
| **74** | 9.9 | 3.7 | 7.5 | 4.7 | 5.6 | 7 | 6.7 | 6.8 | 7.2 | 4.1 | 3.9 | 8.6 |
| **75** | 8.5 | 3.9 | 5.3 | 5.5 | 5 | 4.9 | 6 | 6.8 | 5.7 | 4.4 | 3.7 | 8.2 |
| **76** | 9.9 | 3 | 6.8 | 5 | 5.4 | 5.9 | 4.8 | 4.9 | 7.3 | 3.1 | 3.8 | 7.1 |
| **77** | 7.6 | 3.6 | 7.6 | 4.6 | 4.7 | 4.6 | 5 | 7.4 | 8.1 | 4.5 | 3.9 | 6.4 |
| **78** | 9.4 | 3.8 | 7 | 6.2 | 4.7 | 6.5 | 4.9 | 8.5 | 7.3 | 4.3 | 4.1 | 7.6 |
| **79** | 9.3 | 3.5 | 6.3 | 7.6 | 5.5 | 7.5 | 5.9 | 4.6 | 6.6 | 5.2 | 4.6 | 8.9 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **80** | 7.1 | 3.4 | 4.9 | 4.1 | 4 | 5 | 5.9 | 7.8 | 6.1 | 2.6 | 2.7 | 5.7 |
| **81** | 9.9 | 3 | 7.4 | 4.8 | 4 | 5.9 | 4.8 | 4.9 | 5.9 | 3.2 | 3.8 | 7.1 |
| **82** | 8.7 | 3.2 | 6.4 | 4.9 | 2.4 | 6.8 | 4.6 | 6.8 | 6.3 | 4.3 | 4 | 7.4 |
| **83** | 8.6 | 2.9 | 5.8 | 3.9 | 2.9 | 5.6 | 4 | 6.3 | 6.1 | 2.7 | 3 | 6.6 |
| **84** | 6.4 | 3.2 | 6.7 | 3.6 | 2.2 | 2.9 | 5 | 8.4 | 7.3 | 2 | 1.6 | 5 |
| **85** | 7.7 | 2.6 | 6.7 | 6.6 | 1.9 | 7.2 | 4.3 | 5.9 | 6.5 | 4.7 | 4.3 | 8.2 |
| **86** | 7.5 | 3.5 | 4.1 | 4.5 | 3.5 | 4.1 | 4.5 | 7.6 | 4.9 | 3.4 | 3.4 | 5.2 |
| **87** | 5 | 3.6 | 1.3 | 3 | 3.5 | 4.2 | 4.9 | 8.2 | 4.3 | 2.4 | 3.1 | 5.2 |
| **88** | 7.7 | 2.6 | 8 | 6.7 | 3.5 | 7.2 | 4.3 | 5.9 | 6.9 | 5.1 | 4.3 | 8.2 |
| **89** | 9.1 | 3.6 | 5.5 | 5.4 | 4.2 | 6.2 | 4.6 | 8.3 | 6.5 | 4.6 | 3.9 | 7.3 |
| **90** | 5.5 | 5.5 | 7.7 | 7 | 5.6 | 5.7 | 8.2 | 6.3 | 7.4 | 5.5 | 4.9 | 8.2 |
| **91** | 9.1 | 3.7 | 7 | 4.1 | 4.4 | 6.3 | 5.4 | 7.3 | 7.5 | 4.4 | 3.3 | 7.4 |
| **92** | 7.1 | 4.2 | 4.1 | 2.6 | 2.1 | 3.3 | 4.5 | 9.9 | 5.5 | 2 | 2.4 | 4.8 |
| **93** | 9.2 | 3.9 | 4.6 | 5.3 | 4.2 | 8.4 | 4.8 | 7.1 | 6.2 | 4.4 | 4.2 | 7.6 |
| **94** | 9.3 | 3.5 | 5.4 | 7.8 | 4.6 | 7.5 | 5.9 | 4.6 | 6.4 | 4.8 | 4.6 | 8.9 |
| **95** | 9.3 | 3.8 | 4 | 4.6 | 4.7 | 6.4 | 5.5 | 7.4 | 5.3 | 3.6 | 3.4 | 7.7 |
| **96** | 8.6 | 4.8 | 5.6 | 5.3 | 2.3 | 6 | 5.7 | 6.7 | 5.8 | 4.9 | 3.6 | 7.3 |
| **97** | 7.4 | 3.4 | 2.6 | 5 | 4.1 | 4.4 | 4.8 | 7.2 | 4.5 | 4.2 | 3.7 | 6.3 |
| **98** | 8.7 | 3.2 | 3.3 | 3.2 | 3.1 | 6.1 | 2.9 | 5.6 | 5 | 3.1 | 2.5 | 5.4 |
| **99** | 7.8 | 4.9 | 5.8 | 5.3 | 5.2 | 5.3 | 7.1 | 7.9 | 6 | 4.3 | 3.9 | 6.4 |
| **100** | 7.9 | 3 | 4.4 | 5.1 | 5.9 | 4.2 | 4.8 | 9.7 | 5.7 | 3.4 | 3.5 | 6.4 |

# 2 Assumptions

The sample size of the data set is 100 observations in each variable, and we have 12 variables. Out of 12, 11 Variables are independent variable and 1 variable is a dependent variable (Satisfaction) of remaining 11 variables. We assume all these 11 independent variables are having a high correlation with each other. Hence there will be a high chance of Multicollinearity in the given data sets. So, we decide to find the correlation by using the R with Cor function and further to perform to PCA (Principal Factor Analysis), to reduce the dimension of the entire data set. Finally, to perform Regression by using the newly derived variable to measure the evidence of variation on the Satisfaction Score.

# 3 Step by step approach

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

1. Exploratory Data Analysis & Installing Packages.
2. Correlation Analysis & Bartlett Test on the dependent Variables.
3. Perform in PCA, Start with Eigen Values.
4. Using Scree Plot & Kaiser Criterion to reduce the dimension.
5. Getting the Loadings & Communality with None & Varimax rotation.
6. Visualization of Principal Component Analysis.
7. Generating the scores of RCA for all the observations/ customers.
8. Loading the data of RCA Factors along with the Satisfaction Scores into R.
9. Perform Multiple Linear Regression and confident level of each slopes.
10. Conclusion on the Model Validity.

## 3.4 Exploratory Data Analysis

### 3.4.1 Setting up working Directory, Reading File in R & Installing Packages:

```
setwd("C:/Users/Desktop/Project 2")
FH=read.csv("Factor-Hair-Revised.csv", header = TRUE)
attach(FH)
FH
```

```
install.packages(c("psych","car","foreign","Mass","lattice","nortest","corrplot"))
library(psych)
library(car)
library(foreign)
library(MASS)
library(lattice)
library(nortest)
library(corrplot)
```

### 3.4.2 Number of Rows and Columns:

- The number of rows in the dataset is 100
- The number of columns (Features) in the dataset is 13

```
> dim(FH)
[1] 100  13
```

### 3.4.3 Features and their Types:

All the features are numbers continuous variables.

| Feature Code | Type | Continuous/ Categorical |
|---|---|---|
| ProdQual | num | Continuous |
| Ecom | num | Continuous |
| TechSup | num | Continuous |
| CompRes | num | Continuous |
| Advertising | num | Continuous |
| ProdLine | num | Continuous |
| SalesFImage | num | Continuous |
| ComPricing | num | Continuous |
| WartyClaim | num | Continuous |
| OrdBilling | num | Continuous |
| DelSpeed | num | Continuous |
| Satisfaction | num | Continuous |

### 3.4.4 Check Names/ Column Header:

- To check whether the data has a column header.

- `names(FH)`
- `[1] "ID" "ProdQual" "Ecom" "TechSup" "CompRes" "Advertising"`
- `[7] "ProdLine" "SalesFImage" "ComPricing" "WartyClaim" "OrdBilling"`
- `[12] "DelSpeed" "Satisfaction"`

### 3.4.5 Dataset Summary:

- Summaries the entire datasets to find the Min. Max. Mean, Median 1st & 3rd Quartile of each Variable/Factor. Also to check if there is any Outlier.
- `> summary(FH)`
- Please check the Appendix to code output.

```
> summary(FH)
       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
```

### 3.4.6 Check for Missing Values:

- The data was checked for Missing Values using R function columns(is.na) and found **no missing values**.

### 3.4.7 Ignoring Unwanted Datasets:

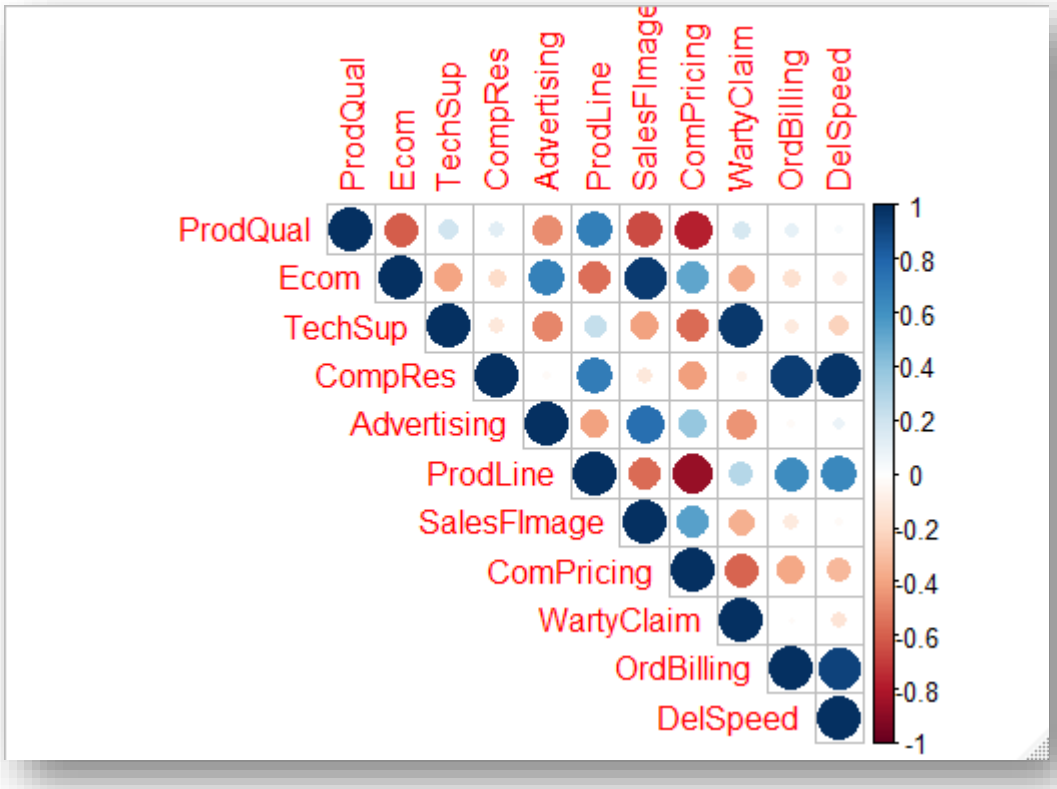
- Ignoring the ID No & Satisfaction Score (Dependent Variable) from the Dataset, which is not required for the correlation analysis
- `FHnew=FH[-c(1,13)]`

## 3.5 Correlation Analysis & Bartlett Test:

### 3.5.1 Correlation:

- Finding the correlation within the variables to find the Multicollinearity.
- `> FHCorr=(cor(FHnew))`
- `> FHCorr`
- Plot the correlation within all the variables to find the Multicollinearity visually.
- `corrplot(cor(FHCorr),method = "circle",type = "upper")`

**Correlation Plot**




- This plot indicates, there was clear demarcation of multicollinearity across all the Variables/ factor. Hence, required to perform PCA for reducing the dimensions.
- Also wrote the correlation matrix in a data frame to export in CSV file to check the correlation with nos.
- `> corrdf=data.frame(FHCorr)`
- `> write.csv(corrdf,"correlation1.csv")`

| Correlation | ProdQual | Ecom | TechSup | CompRes | Advertising | ProdLine | SalesFImage | ComPricing | WartyClaim | OrdBilling | DelSpeed |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ProdQual** | 1.000 | -0.137 | 0.096 | 0.106 | -0.053 | 0.477 | -0.152 | -0.401 | 0.088 | 0.104 | 0.028 |
| **Ecom** | -0.137 | 1.000 | 0.001 | 0.140 | 0.430 | -0.053 | 0.792 | 0.229 | 0.052 | 0.156 | 0.192 |
| **TechSup** | 0.096 | 0.001 | 1.000 | 0.097 | -0.063 | 0.193 | 0.017 | -0.271 | 0.797 | 0.080 | 0.025 |
| **CompRes** | 0.106 | 0.140 | 0.097 | 1.000 | 0.197 | 0.561 | 0.230 | -0.128 | 0.140 | 0.757 | 0.865 |
| **Advertising** | -0.053 | 0.430 | -0.063 | 0.197 | 1.000 | -0.012 | 0.542 | 0.134 | 0.011 | 0.184 | 0.276 |
| **ProdLine** | 0.477 | -0.053 | 0.193 | 0.561 | -0.012 | 1.000 | -0.061 | -0.495 | 0.273 | 0.424 | 0.602 |
| **SalesFImage** | -0.152 | 0.792 | 0.017 | 0.230 | 0.542 | -0.061 | 1.000 | 0.265 | 0.107 | 0.195 | 0.272 |
| **ComPricing** | -0.401 | 0.229 | -0.271 | -0.128 | 0.134 | -0.495 | 0.265 | 1.000 | -0.245 | -0.115 | -0.073 |
| **WartyClaim** | 0.088 | 0.052 | 0.797 | 0.140 | 0.011 | 0.273 | 0.107 | -0.245 | 1.000 | 0.197 | 0.109 |
| **OrdBilling** | 0.104 | 0.156 | 0.080 | 0.757 | 0.184 | 0.424 | 0.195 | -0.115 | 0.197 | 1.000 | 0.751 |
| **DelSpeed** | 0.028 | 0.192 | 0.025 | 0.865 | 0.276 | 0.602 | 0.272 | -0.073 | 0.109 | 0.751 | 1.000 |

- However, also performed the Bartlett Test to proceed on Dimension Reduction using PCA.

```
> cortest.bartlett(FHCorr,nrow(FHnew))
$chisq
[1] 619.2726

$p.value
[1] 1.79337e-96

$df
[1] 55
```

- The Outcome of Bartlett Test, has a high significance pValue, hence it proves the high possibility of Multicollinearity and required to proceed on performing PCA.

## 3.6 Performing Principal Component Analysis

- To start with PCA Analysis, must generate the Eigen Value.

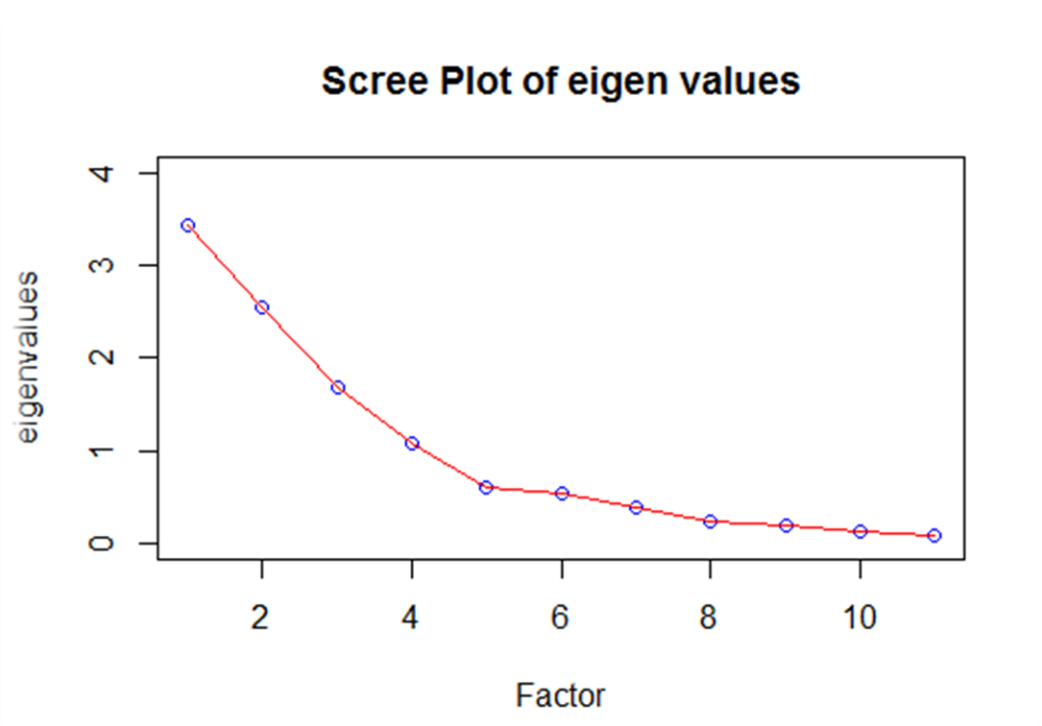
- **Finding out the Eigen Value & Eigen Vector:**
- We require to write the following code in R Function to get the Eigen Values.

```
ev=eigen(FHCorr)
eigenvalues=ev$values
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
```

- Based on Kaiser Criterion, Eigen Values of factors anything greater than 1 can be considered, hence the following 4 we are considering the PCA.
- [1] 3.42697133 2.55089671 1.69097648 1.08655606
- However, there is also another way to finalize the no. of factors which required. Using Scree Plot will help us to identify the same.

- **Scree Plot:**
- Before, creating a plot, require to create a data frame with the Eigen Values generated.
- Creating a Factors for 11 Variables.
- `Factor=c(1:11)`

- Creating a data frame with Factor & Eigen Values and ploting the Scree Plot
- `> scree=data.frame(Factor,eigenvalues)`
- `> plot(scree,main="Scree Plot of eigen values",col="Blue",ylim=c(0,4))`
- `> lines(scree,col="Red")`




- Based on Elbow rule, we could see that there is clear evidence in the plot upto 1, the line doesn't have bend and after 1, the line is getting bend and giving a visual of elbow. So, to perform PCA and reduce a dimension, only 4 factors are required.

- **Loading of 4 Factors & Communality:**
- To get the loading for the four factors with communality, had executed the Principal component function in R with the **Rotation="None"** and the output furnished below.

```
> pc=principal(FHnew,nfactors =4,rotate="none")
> pc
Principal Components Analysis
Call: principal(r = FHnew, nfactors = 4, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
               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

Mean item complexity =  1.9
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.06
 with the empirical chi square  39.02  with prob <  0.0018

Fit based upon off diagonal values = 0.97
```

- In the above output, the communality (h2) clearly shows the % Variation across each variable within these 4 Factors. Cumulative Variation was at 80%, which means 80% of variation being explained by all these 4 Factors. And the next step, had copied this output in a excel file to check the Multicollinearity within this 4 Individual Factors and the observations below. Also understand that, if we have multiple factors correlating, absolute values of the PCA or factor variable are considered, not the signs.

| Rotation None | PC1 | PC2 | PC3 | PC4 | h2 |
|---|---|---|---|---|---|
| ProdQual | 0.25 | -0.5 | -0.08 | 0.67 | 0.77 |
| Ecom | 0.31 | 0.71 | 0.31 | 0.28 | 0.78 |
| TechSup | 0.29 | -0.37 | 0.79 | -0.2 | 0.89 |
| CompRes | 0.87 | 0.03 | -0.27 | -0.22 | 0.88 |
| Advertising | 0.34 | 0.58 | 0.11 | 0.33 | 0.58 |
| ProdLine | 0.72 | -0.45 | -0.15 | 0.21 | 0.79 |
| SalesFImage | 0.38 | 0.75 | 0.31 | 0.23 | 0.86 |
| ComPricing | -0.28 | 0.66 | -0.07 | -0.35 | 0.64 |
| WartyClaim | 0.39 | -0.31 | 0.78 | -0.19 | 0.89 |
| OrdBilling | 0.81 | 0.04 | -0.22 | -0.25 | 0.77 |
| DelSpeed | 0.88 | 0.12 | -0.3 | -0.21 | 0.91 |

Highlighted in Green fills are highest Value across 4 Factors & highlighted in Red Font are Multicollinearity.

This clearly shows that there is a need for redistribution of Factor Values.

So, this time we run the Principal Component Function using the Varimax Rotation.

```
> pc1=principal(FHnew,nfactors =4,rotate="varimax")
> pc1
Principal Components Analysis
Call: principal(r = FHnew, nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
               RC1   RC2   RC3   RC4   h2    u2 com
ProdQual      0.00 -0.01 -0.03  0.88 0.77 0.232 1.0
Ecom          0.06  0.87  0.05 -0.12 0.78 0.223 1.1
TechSup       0.02 -0.02  0.94  0.10 0.89 0.107 1.0
CompRes       0.93  0.12  0.05  0.09 0.88 0.119 1.1
Advertising   0.14  0.74 -0.08  0.01 0.58 0.424 1.1
ProdLine      0.59 -0.06  0.15  0.64 0.79 0.213 2.1
SalesFImage   0.13  0.90  0.08 -0.16 0.86 0.141 1.1
ComPricing   -0.09  0.23 -0.25 -0.72 0.64 0.359 1.5
WartyClaim    0.11  0.05  0.93  0.10 0.89 0.108 1.1
OrdBilling    0.86  0.11  0.08  0.04 0.77 0.234 1.1
DelSpeed      0.94  0.18  0.00  0.05 0.91 0.086 1.1

                       RC1  RC2  RC3  RC4
SS loadings           2.89 2.23 1.86 1.77
Proportion Var        0.26 0.20 0.17 0.16
Cumulative Var        0.26 0.47 0.63 0.80
Proportion Explained  0.33 0.26 0.21 0.20
Cumulative Proportion 0.33 0.59 0.80 1.00

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

The root mean square of the residuals (RMSR) is  0.06
 with the empirical chi square  39.02  with prob <  0.0018

Fit based upon off diagonal values = 0.97
```

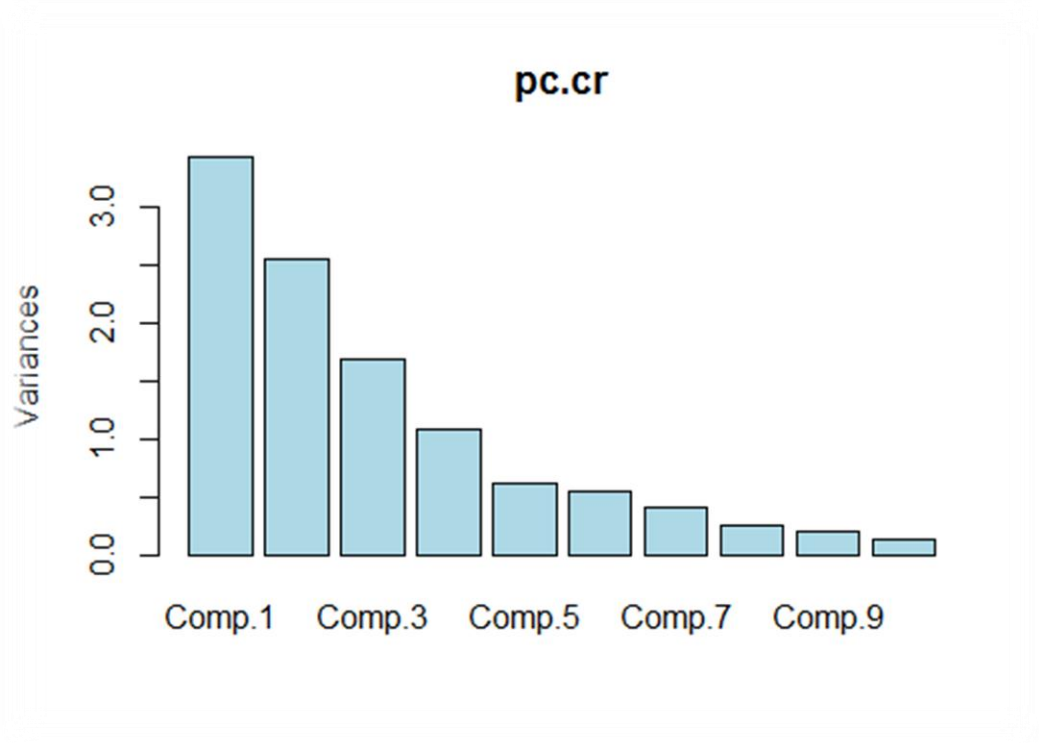
- Copied this Redistributed output Values using Varimax rotation in a excel file to check the Multicollinearity within this 4 Individual Factors and found there is no evidence of multicollinearity and it is perfect data sets to work on a Regression.
- One observation on the RCA Factor 4, is Competition Pricing has a negative loading compare to Product Line and Product Quality, which Product quality and Pricing vary together but travelling in a straight opposite direction.

| Rotation Varimax | RC1 | RC2 | RC3 | RC4 | h2 |
|---|---|---|---|---|---|
| ProdQual | 0 | -0.01 | -0.03 | 0.88 | 0.77 |
| Ecom | 0.06 | 0.87 | 0.05 | -0.12 | 0.78 |
| TechSup | 0.02 | -0.02 | 0.94 | 0.1 | 0.89 |
| CompRes | 0.93 | 0.12 | 0.05 | 0.09 | 0.88 |
| Advertising | 0.14 | 0.74 | -0.08 | 0.01 | 0.58 |
| ProdLine | 0.59 | -0.06 | 0.15 | 0.64 | 0.79 |
| SalesFImage | 0.13 | 0.9 | 0.08 | -0.16 | 0.86 |
| ComPricing | -0.09 | 0.23 | -0.25 | -0.72 | 0.64 |
| WartyClaim | 0.11 | 0.05 | 0.93 | 0.1 | 0.89 |

| | | | | | |
|---|---|---|---|---|---|
| OrdBilling | 0.86 | 0.11 | 0.08 | 0.04 | 0.77 |
| DelSpeed | 0.94 | 0.18 | 0 | 0.05 | 0.91 |

- **Principal Component Visualization:**

- Using the following function had performed the Visualization of Principal Component Analysis. This clearly communicates, most of the variations being explained by the first 4 Factors and al so gives a impression that Kaiser rule was rightly implemented.
- `> FHsc=scale(FHnew)`
- `> pc.cr=princomp(FHsc,cor=TRUE)`
- `> plot(pc.cr, col="lightblue")`




- **Naming/ Labeling the Factors:**

- Factor 1/ RC1 = Complaint Resolution, Order Billing & Delivery Speed
  All these Variables are somewhere connected customer experience, hence naming it as Customer Service Factor.

- Factor 2/ RC2 = ECommerce, Advertising & Sales Force Image.
  All these Variables are connected to Branding, Selling, Marketing & Communication, hence naming it as Marketing Factor.

- Factor 3/ RC3 = Technical Support & Warranty Claims.
  Both the Variables are connected to each other, hence naming it as Technical Support.

- Factor 4/ RC4 = Product Quality, Product Line & Competition Pricing.
  All the Variables are connected to Product Pricing & Quality, hence naming it as Product Price and Quality.

- At the last stage of PCA, generating the PCA Scores of RCA Scores of 4 Factors for all the 100 observations using the following function `> pc1$scores.`

- Copied the output loading of all the 4 factors into csv file along with Satisfaction Score (Dependent Variable) to perform regression.

## 4. Multiple Linear Regression Using 4 Factors & Satisfaction Score:

- Reading the file in R, which contains the 4 Factors & Satisfaction Scores. The file name is FHR.csv, available in the same working directory.

```
FHR=read.csv("Factor-Hair-Regression.csv", header = TRUE)
attach(FHR)
FHR
```

- Performed basic descriptive Stats.

```
str(FHR)
names(FHR)
summary(FHR)
```

- Performed Multiple Linear Regression.

- Created the Linear Regression Model, keeping Satisfaction as a dependent Variable and other 4 factors as Independent Variable by using lm function in R.

```
Model=lm(Satisfaction~CustomerService+Marketing+TechnicalSupport+ProductPricea
ndQuality)
```

- And Summarized the Model, to infer with the outcome.

### R - Output on Multiple Linear Regression:

```
> Model=lm(Satisfaction~CustomerService+Marketing+TechnicalSupport+ProductPriceandQuality)
> summary(Model)

Call:
lm(formula = Satisfaction ~ CustomerService + Marketing + TechnicalSupport +
    ProductPriceandQuality)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6327 -0.5123  0.1392  0.4487  1.5150

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.92065    0.07084  97.699  < 2e-16 ***
CustomerService       0.61896    0.07129   8.682 1.08e-13 ***
Marketing             0.50461    0.07074   7.133 1.93e-10 ***
TechnicalSupport      0.06712    0.07075   0.949    0.345
ProductPriceandQuality 0.54688   0.07142   7.657 1.58e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7083 on 95 degrees of freedom
Multiple R-squared:  0.6611,    Adjusted R-squared:  0.6468
F-statistic: 46.32 on 4 and 95 DF,  p-value: < 2.2e-16
```

- In the above output of lm model, Customer Service, Marketing & Product Price and Quality are highly significance and impacting the overall Satisfaction Score.

- However, Technical Support variable was not much significance.
- So, performed the interaction effect for Technical Support with other variables *(Customer Service, Marketing & Product Price and Quality)* one by one.
- After trying the multiple combination, could not find any interaction effect with Technical support as significance.
- This shows, Technical Support Variable alone not significant variable.

```
> Model1=lm(Satisfaction~CustomerService+Marketing+TechnicalSupport+ProductPriceandQuality+Technic
alSupport*Marketing)
> summary(Model1)

Call:
lm(formula = Satisfaction ~ CustomerService + Marketing + TechnicalSupport +
    ProductPriceandQuality + TechnicalSupport * Marketing)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6102 -0.5014  0.1402  0.4566  1.3554

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 6.91986    0.06994  98.935  < 2e-16 ***
CustomerService             0.62731    0.07053   8.894 4.13e-14 ***
Marketing                   0.51659    0.07015   7.364 6.73e-11 ***
TechnicalSupport            0.06782    0.06986   0.971   0.3342
ProductPriceandQuality      0.54734    0.07052   7.762 1.01e-11 ***
Marketing:TechnicalSupport -0.13175    0.07097  -1.856   0.0665 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6994 on 94 degrees of freedom
Multiple R-squared:  0.673,     Adjusted R-squared:  0.6556
F-statistic:  38.7 on 5 and 94 DF,  p-value: < 2.2e-16
```

- Also checked the confidence level of the model by using Confint Function in R with 99% Confidence level.
  confint(Model,level = 0.99)

```
> confint(Model,level = 0.99)
                                0.5 %    99.5 %
(Intercept)                 6.7344531 7.1068516
CustomerService             0.4315735 0.8063521
Marketing                   0.3186511 0.6905677
TechnicalSupport           -0.1188615 0.2530956
ProductPriceandQuality      0.3591476 0.7346119
```

## 5. Conclusion on the Model Validity

- Basis the Model, all the variables are highly significance and have an impact on the Satisfaction Score except Technical Support.
- Also, scores have increased across all variables, comparing to the 1st Model (without interaction) than in 2nd Model (With interaction).
- Yhat = 6.91986(intercept) +0.62731(Customer service)+0.51659(Marketing) +0.06782(Technical Support)+0.54734(Product Price and Quality).
- Intercept: 6.91986 is the average score been given on the scale of 10 by all the customers.
- Multiple R Squared: implies 67.3% of variations in the Satisfaction Score.
- Adjusted R Squared: implies only 65.56% of variations in the Satisfaction Score.
- The F Statistics is 38.7 and the p-Value is <2.2e-16 which is highly significance and very much lesser than alpha 5%.
- Hence, reject the null hypothesis and all betas are zero.
- Also checked the confidence intervals of Slopes - all 4 factors.
- So, i conclude that regression Model is Valid.

## 6 Appendix A – Source Code:

Here is the Source code of the HAIR Case study:

```
# Project - Advance Statistics - Factor Hair Revised

# Setting up working Directory and reading data file in R
setwd("C:/Users/Dinesh/Desktop/Project 2")
FH=read.csv("Factor-Hair-Revised.csv", header = TRUE)
attach(FH)
FH

# Packages required for Correlation/ Visualization/ Personal Component/ Regression.

install.packages(c("psych","car","foreign","Mass","lattice","nortest","corrplot"))
library(psych)
library(car)
library(foreign)
library(MASS)
library(lattice)
library(nortest)
library(corrplot)

# Basic Descriptive Analytics.
dim(FH)
str(FH)
names(FH)
summary(FH)
is.na(FH)
FHnew=FH[-c(1,13)]
names(FHnew)

# Correlation.

FHCorr=(cor(FHnew))
FHCorr
corrplot(cor(FHCorr),method = "circle",type = "upper")

#Write the correlation matrix in a dataframe

corrdf=data.frame(FHCorr)
write.csv(corrdf,"correlation1.csv")

#Bartlett Test

cortest.bartlett(FHCorr,nrow(FHnew))
print(cortest.bartlett(FHCorr,nrow(FHnew)))

# ************************************************************
```

```
# Finding out the Eigen Values and Eigen Vectors.

ev=eigen(FHCorr)
eigenvalues=ev$values
eigenvalues
Factor=c(1:11)
scree=data.frame(Factor,eigenvalues)
plot(scree,main="Scree Plot of eigen values",col="Blue",ylim=c(0,4))
lines(scree,col="Red")

eigenvectors=ev$vectors
eigenvectors

# Getting the loadings and Communality

pc=principal(FHnew,nfactors =4,rotate="none")
pc

pc1=principal(FHnew,nfactors =4,rotate="varimax")
pc1

#Visualization of PC
FHsc=scale(FHnew)
pc.cr=princomp(FHsc,cor=TRUE)
plot(pc.cr, col="lightblue")

#scores
pc$scores
pc1$scores
dim(pc1$scores)
dim(FHnew)
prcomp(FHnew)

#*********************************************************************
# Multiple Linear Regression

FHR=read.csv("Factor-Hair-Regression.csv", header = TRUE)
attach(FHR)
FHR
str(FHR)
names(FHR)
summary(FHR)
Model=lm(Satisfaction~CustomerService+Marketing+TechnicalSupport+ProductPriceand
Quality)
summary(Model)

#Interaction effect
```

```
Model1=lm(Satisfaction~CustomerService+Marketing+TechnicalSupport+ProductPriceand
Quality+TechnicalSupport*Marketing)
summary(Model1)

anova(Model)
#CI
confint(Model,level = 0.99)
```