Mini Project 2 – Advanced Statistics

Submitted by

Sita K

BACP Batch (Dec 19-May 20)

Great Learning

February 22, 2020

Table of Contents

Table of Figures

1. Project Objective

The objective of this report is to explore the factor hair dataset (Factor-Hair-Revised.csv) in R and to build an optimum regression model to predict satisfaction. This report will contain:

- Exploratory data analysis of the dataset
- Multicollinearity – evidence and explanation
- Linear regression of dependent variable on all remaining 12 independent variables
- PCA/Factor Analysis to extract the relevant factors and explanation
- Multiple linear regression on dependent variable and independent variables(factors from PCA) and final explanation

2. Assumptions

- The data provided is conclusive and contains the required data

3. Data Analysis – Approach

1. Environment data setup and data import
2. Calculating the required values using inbuilt functions
3. Plot various graphs to understand and explore the data
4. Perform correlation between various variables and check for multicollinearity
5. Form the conclusion
6. Perform linear regression between dependent and independent variables
7. Perform PCA to extract the final list of 4 factors and name them.
8. Perform multiple linear regression on the final list of factors and dependent variable Satisfaction.

   For environment data setup, R's inbuilt packages were used. Also for setting up working directory '**setwd**()' function was used. The given dataset is in .csv format, so we can use **read.csv** function to import the data. All the R commands are in Appendix A.

4. Problem Responses

1. For Basic data summary we can use the function **summary** in R, which provides us with mean, median etc of each column. **Cor** function provides us with the correlation of each variable with each other. **Dim** function provides us with the dimensions of the data. **Str** function provides us with the structure of the data.
   To check if any blank values are there we use **is.na** function. Since the first column is ID values to which the dependent variable **Satisfaction** does not have much correlation we remove it from the dataset.

To do Univariate analysis and Bivariate analysis and for graphs we can use the library **DataExplorer.** We can use **corrplot** function to view the correlation diagram between various variables.

Plot_histogram and **plot_density** show various behaviour of all the variables in graphical format.

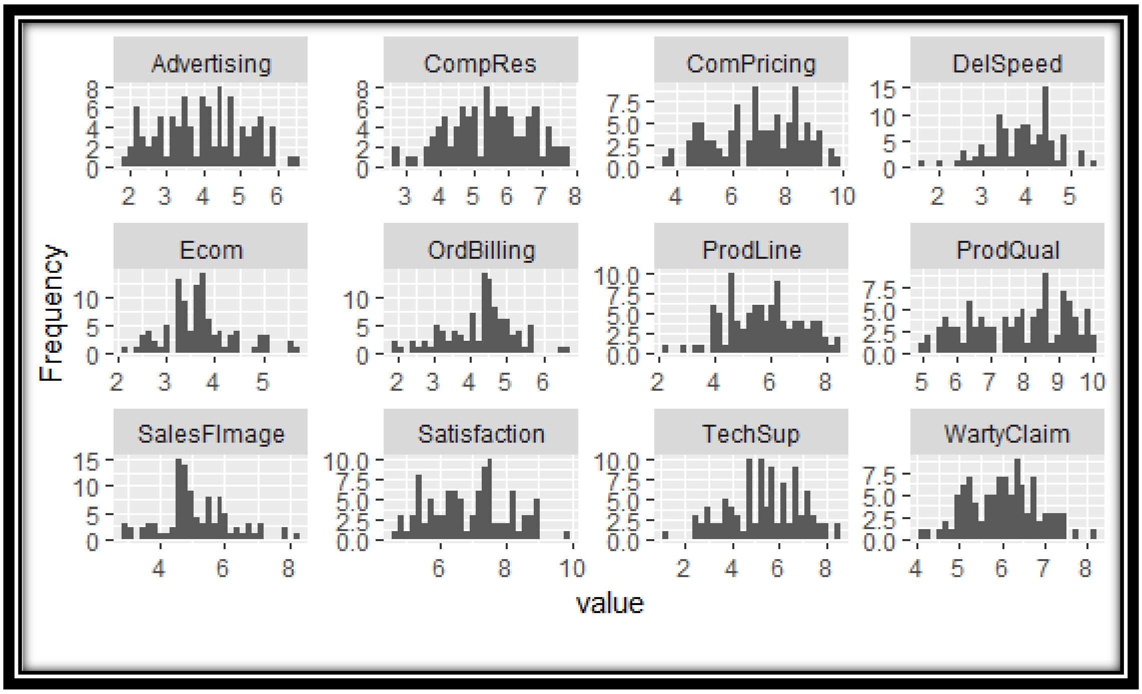

Figure 1: Histogram Plots

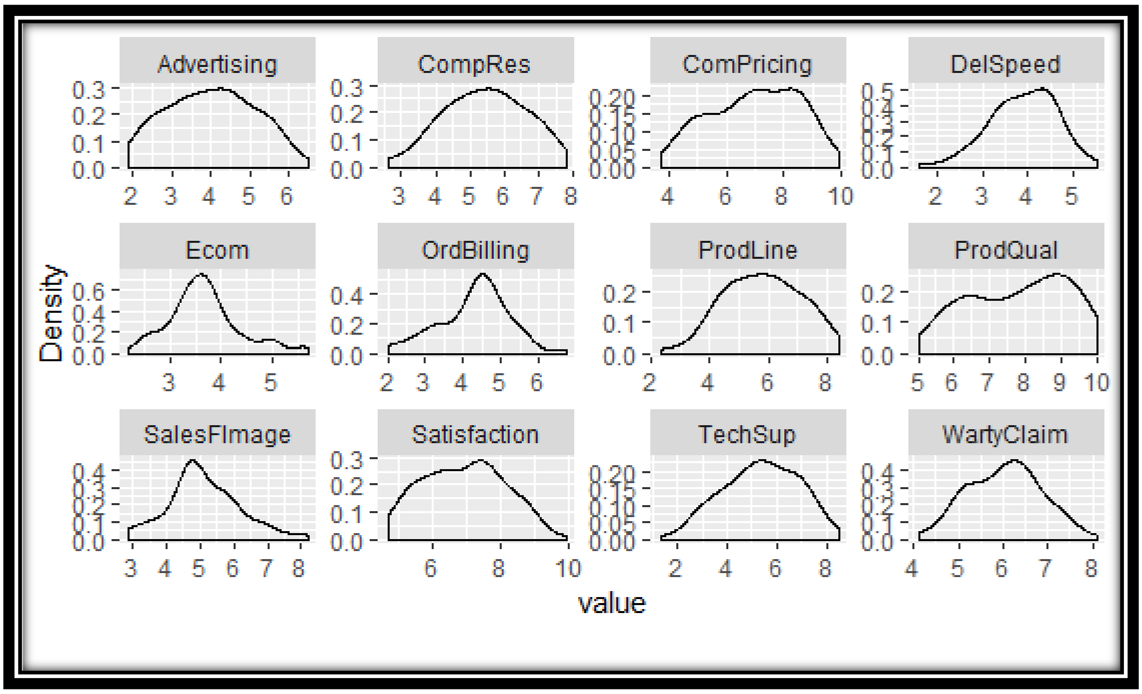

Figure 2: Density Plots

From the diagrams we can understand some variables like Delivery speed, Tech support are left skewed. Others like Sales Force Image is right skewed. Some are bimodal like Product quality and Warranty claims. Most resemble normal distribution like Ecommerce, Complaint Resolution.

We can use the **boxplot** function to identify outliers.

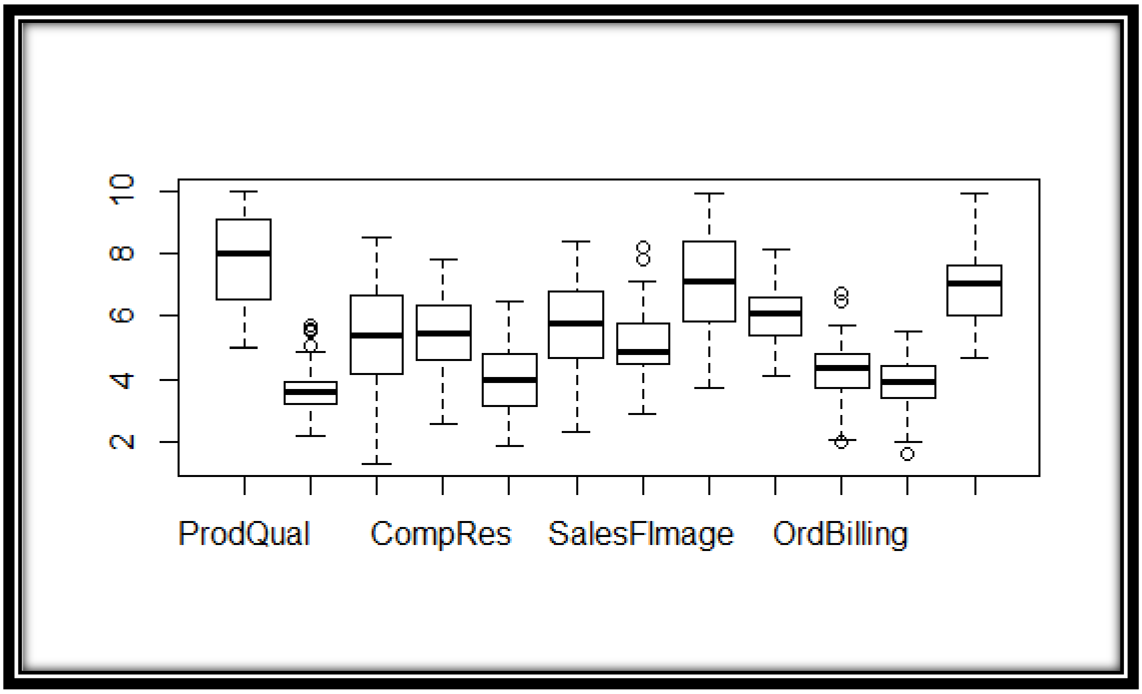

Figure 3: Boxplot

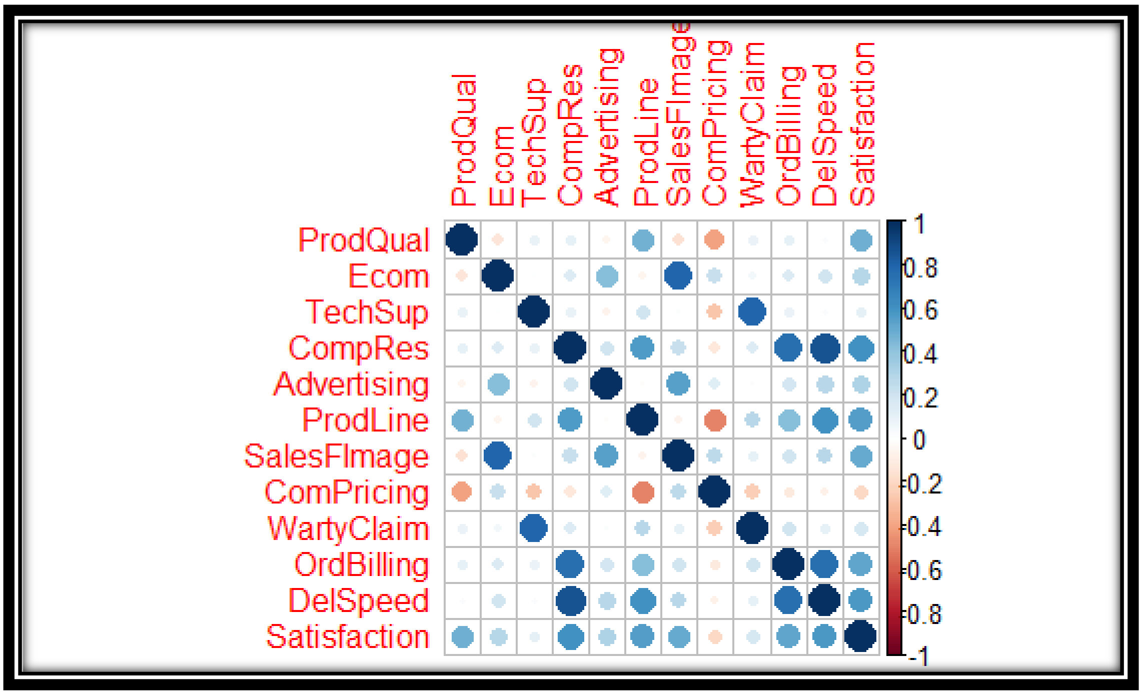

Figure 4: Correlation plot

2. To explain and check whether multicolinearity is there we can use the function **vif** which will indicate which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model.
   The smallest possible value of VIF is one (absence of multicollinearity). As a rule of thumb, a VIF value that exceeds 2.5 indicates a problematic amount of collinearity.
   Once we run the function, we can see that the vif value for delivery speed is greater than 5, complaint resolution is almost 5 and other variables like Product Line, Warranty claim etc are greater than 2.5 which indicates the subtle presence of multicolinearity.

```
> vif(lm(Satisfaction~.,data=mydata))
   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
```

Figure 5: VIF Values

3. We will do a simple linear regression with every variable. For that we will use **lm** function for calculating the linear model of dependent variable as a function of each independent variable. Here the dependent variable is Satisfaction and other 11 variables are the independent variables
4. To perform PCA/Factor Analysis, we first check whether the dataset can be subjected to PCA using **Bartlett test** with **cortest.bartlett** function. Since p value is significantly low we reject the null hypothesis that PCA cannot be conducted.

```
> cortest.bartlett(cor(mydata[1:11]),100)
$chisq
[1] 619.2726

$p.value
[1] 1.79337e-96

$df
[1] 55
```

Figure 6: Bartlett test

We can find the eigen values using the **eigen** function and we can see that four of the factors have eigen values more than 1 from the scree plot.

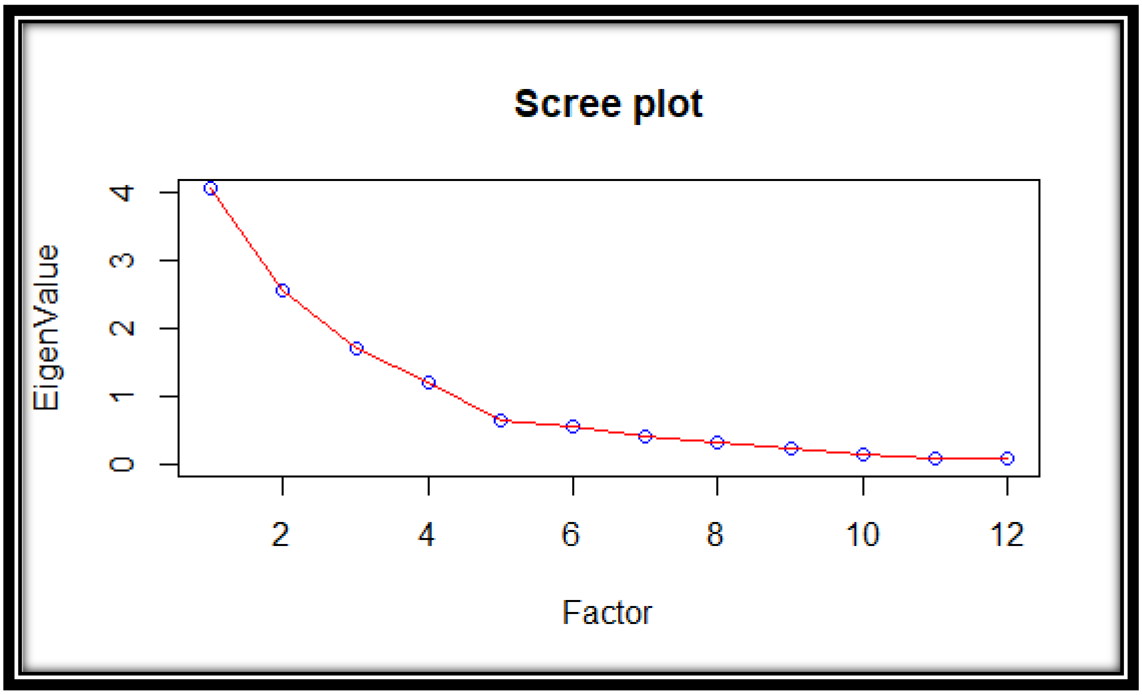

Figure 7: Scree Plot

Then we can perform the unrotate and rotate PCA Analysis using the **principal** function and we can see that the 4 factors significantly contribute to the variation. The most accurate result is obtained as output of rotation type of varimax.

```
Principal Components Analysis
Call: principal(r = mydata[1:11], nfactors = 4, rotate = "v
Standardized loadings (pattern matrix) based upon correlati
              RC1    RC2    RC3    RC4    h2     u2  com
ProdQual     0.002 -0.013 -0.033  0.876 0.768 0.2320 1.00
Ecom         0.057  0.871  0.047 -0.117 0.777 0.2229 1.05
TechSup      0.018 -0.024  0.939  0.101 0.893 0.1069 1.03
CompRes      0.926  0.116  0.049  0.091 0.881 0.1187 1.06
Advertising  0.139  0.742 -0.082  0.015 0.576 0.4240 1.10
ProdLine     0.591 -0.064  0.146  0.642 0.787 0.2129 2.12
SalesFImage  0.133  0.900  0.076 -0.159 0.859 0.1406 1.12
ComPricing  -0.085  0.226 -0.246 -0.723 0.641 0.3594 1.47
WartyClaim   0.110  0.055  0.931  0.102 0.892 0.1078 1.06
OrdBilling   0.864  0.107  0.084  0.039 0.766 0.2339 1.05
DelSpeed     0.938  0.177 -0.005  0.052 0.914 0.0856 1.08
```

Figure 8: PCA output

From the output we can understand that Factor 1 consists of Complaint Resolution, Order & Billing and Delivery Speed. Factor 2 consists of E-commerce, Salesforce Image and Advertising. Factor 4 consists of Product Quality, Product Line and Competitive Pricing. Factor 3 consists of Technical Support, Warranty & Claims.

We will name the four factors as follows.

Component names	Meaningful Names for factors	Names shortened
RC1	Purchasing Experience	PrchExp
RC2	Brand Recognition	BndRecog

RC3	After Sales service	AftSaSrvc
RC4	Product Features	ProdFtr

a. RC1 - Purchasing Experience explains about variables affecting Complaint resolution, Order and Billing and delivery speed to customers
b. RC2 - Brand recognition handles E-commerce, image of Sales force and Advertising which is face of the product
c. RC3 - After Sales Service gives information about Technical support, and Warranty and claims if there is any problem to customer after he has bought the item
d. RC4 – Product Feature talks about the qualities of product like its varieties and types, prices and its quality i.e all tangible aspects about the very existence of company

5. We have to create a data frame with 4 of the different factors and the dependent variable Satisfaction to perform multiple linear regression. We can use **as.data.frame** and **cbind** function for the same.
   To do multiple linear regression of the same, we have to use **lm** function. If you take the summary after doing the multiple regression we can see the output as below.

```
    Min      1Q Median      3Q     Max
-1.631 -0.500  0.137  0.462  1.523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.9180     0.0709   97.59  < 2e-16 ***
PrchExp       0.6180     0.0712    8.67  1.1e-13 ***
BndRecog      0.5097     0.0712    7.15  1.7e-10 ***
AftSaSrvc     0.0671     0.0712    0.94     0.35
ProdFtr       0.5403     0.0712    7.58  2.2e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.709 on 95 degrees of freedom
Multiple R-squared:  0.661,     Adjusted R-squared:  0.646
F-statistic: 46.2 on 4 and 95 DF,  p-value: <2e-16
```

Figure 9: Multiple Regression Output

From the summary we can understand the Pr(>|t|) value is significant for Purchase Experience, Brand Recognition and Product Features are statistically significant. After Sales service has a higher p value and is not statistically significant. Adjusted R squared value is around 65% which explains predicts the variability of the data set upto a level which is good. Degrees of freedom is n-k-1(100-4-1=95)

Overall p-value is much lower than 0.05 which indicates the model is significantly valid. The t-statistic values are relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists. Residual standard error is very low. F-statistic is relatively larger than 1 given the size of our data. So it indicates a good relationship between dependent and independent variables.

We can predict the dependent variable values using the new Model with the function **predict**. Then we can plot the graph against predicted and actual values using **plot** function. We can see the actual and predicted values are not varying much.

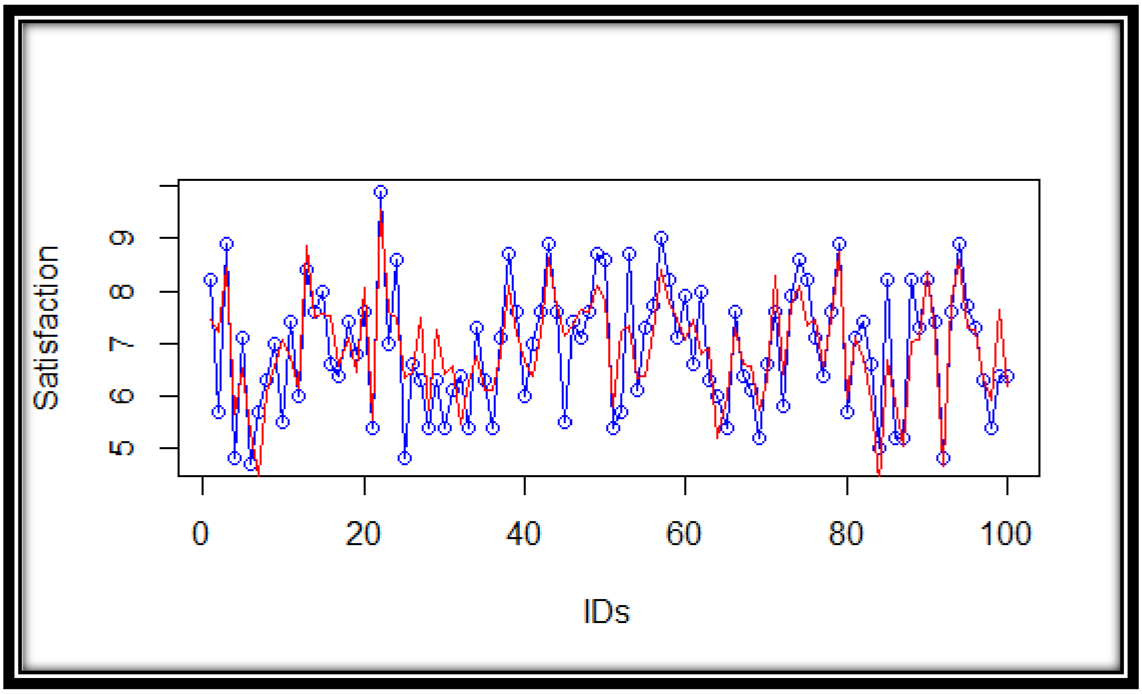

Figure 10: Plot of Actual(Blue) vs Predicted(Red)

5. Conclusion

We can see that the Satisfaction ratings of hair product depends highly on Purchase Experience, Brand Recognition and Product Features. After Sales Service comes after all this while considering the hair product.

6. References

https://cran.r-project.org/web/packages/dlookr/vignettes/EDA.html

https://rpubs.com/

https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Appendix A

Answer_AS.R contains the responses. Sample code is given below.

```
##Read the data
getwd()
setwd("E:/Sita/BACP/R Data")
mydata=read.csv("Factor-Hair-Revised.csv",header=T)
mydata
attach(mydata)
names(mydata)
##Summary of the data
summary(mydata)
## To check the correlation
COR=cor(mydata)
#dimesnsions
dim(mydata)
#structure of the data
str(mydata)
##To check if any blank data is present
sum(is.na(mydata))
##Remove the first row since its ID values
mydata=mydata[2:13]
# library DataExplorer for EDA and plots
library("DataExplorer")
library("corrplot")
corrplot(COR)
plot_histogram(mydata)
plot_density(mydata)
boxplot(mydata)
##Multicollinearity
library("car")
vif(lm(Satisfaction~.,data=mydata))
##Plot on multicolinearity
plot(mydata)
#Simple linear Regression
Model_ProdQ = lm(Satisfaction~ProdQual)
summary(Model_ProdQ)
Model_Ecom= lm(Satisfaction~Ecom)
summary(Model_Ecom)
Model_TechSup= lm(Satisfaction~TechSup)
```

```r
summary(Model_TechSup)
Model_CR = lm(Satisfaction~CompRes)
summary(Model_CR)
Model_Adv = lm(Satisfaction~Advertising)
summary(Model_Adv)
Model_PL = lm(Satisfaction~ProdLine)
summary(Model_PL)
Model_SalesF = lm(Satisfaction~SalesFImage)
summary(Model_SalesF)
Model_ComP = lm(Satisfaction~ComPricing)
summary(Model_ComP)
Model_WC = lm(Satisfaction~WartyClaim)
summary(Model_WC)
Model_OB = lm(Satisfaction~OrdBilling)
summary(Model_OB)
Model_DS = lm(Satisfaction~DelSpeed)
summary(Model_DS)
#PCA/Factor Analysis
library(psych)
cortest.bartlett(cor(mydata[1:11]),100)
ev=eigen(cor(mydata[1:11]))
ev
EigenValue = ev$values
EigenValue
Factor=c(1,2,3,4,5,6,7,8,9,10,11)
Scree = data.frame(Factor,EigenValue)
plot(Scree,main="Scree plot",col="Blue",ylim=c(0,4))
lines(Scree,col="Red")
Unrotate = principal(mydata[1:11], nfactors=4, rotate="none")
print(Unrotate,digits=3)
UnrotatedProfile=plot(Unrotate,row.names(Unrotate$loadings))
Rotate=principal(mydata[1:11],nfactors=4,rotate="varimax")
print(Rotate,digits=3)
RotatedProfile=plot(Rotate,row.names(Rotate$loadings),cex=1.0)
```

```r
## data frame with four factors and dependent variable
as.data.frame(Rotate$scores)
mydata2=cbind(mydata[,12],Rotate$scores)
colnames(mydata2)=c("Satisfaction","PrchExp","BndRecog","AftSaSrvc","ProdFtr")
mydata2=as.data.frame(mydata2)
attach(mydata2)
#Multiple linear regression
Model1=lm(Satisfaction~PrchExp+BndRecog+AftSaSrvc+ProdFtr,mydata2)
print(summary(Model1),digits=3)
#Predict using new model
mydata3=predict(Model1)
mydata3 = as.data.frame(mydata3)
colnames(mydata3)=c("Pred_Satis")
mydata3=cbind(mydata2,mydata3)
mydata3$Pred_Satis=round(mydata3$Pred_Satis,2)
plot(mydata3$Satisfaction,col="Blue",xlab="IDs",ylab="Satisfaction")
lines(mydata3$Satisfaction,col="Blue")
plot(mydata3$Pred_Satis,col="Red")
lines(mydata3$Pred_Satis,col="Red")
```

Figure 11: R Code