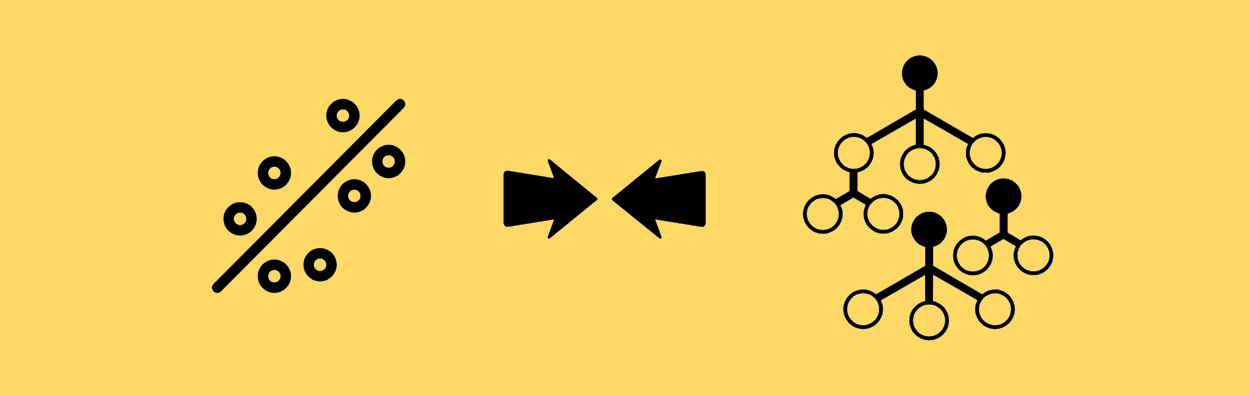
FACTOR AND MULTIPLE LINEAR REGRESSION ANALYSIS

ADVANCED STATISTICS PROJECT

BY SHWETA GUPTA

(PGPBABI)



**INDEX**

* PROJECT OBJECTIVES
* ASSUMPTIONS
* EXPLORATORY DATA ANALYSIS
  + INFERENCES FROM EDA
* INVESTIGATING MULTICOLLINEARITY IN DATA
* SIMPLE LINEAR REGRESSION USING ALL VARIABLES
* MULTIPLE LINEAR REGRESSION USING ALL VARIABLES
* PRINCIPAL COMPONENT/FACTOR ANALYSIS
  + FEASABILITY CHECK
  + IDENTIFYING THE NUMBER OF FACTORS
  + SELECT THE BEST MODEL
  + NAME THE FACTORS
* REGRESSION ANALYSIS USING THE 4 FACTORS
  + MODEL VALIDITY
* CONCLUSION
* APPENDIX – SOURCE CODE

**PROJECT OBJECTIVES**

* Perform **exploratory data analysis** on the dataset
* Identify **outliers and missing values**
* Check for **multicollinearity**
* Perform  **simple linear regression** for the dependent variable with every independent variable
* Perform **PCA/Factor analysis** by extracting 4 factors
* Interpret the **PCA** output and **name the Factors**
* Perform **Multiple linear regression** with customer satisfaction as dependent variables and the four factors as independent variables.
* Comment on the **Model output and validity**

**ASSUMPTIONS**

**Linear Regression**

* A **linear relationship** exists between the independent variable (X) and dependent variable (y)
* Little or **no multicollinearity** between the different features
* Residuals should be normally distributed (**multi-variate normality**)
* Little or **no autocorrelation** among residues
* **Homoscedasticity** of the errors

**Factor Analysis**

* **Scale input variables**
* **Reasonably high correlations are present in data**
* **Correlations are linear**

**EXPLORATORY DATA ANALYSIS**

* The data has **100 observations for 13 variables** as below:

**VARIABLE** **TYPE**  **EXPLANANTION**

$ ID int SERIAL NO

$ ProdQual numeric PRODUCT QUALITY

$ Ecom numeric E-COMMERCE

$ TechSup numeric TECHNICAL SUPPORT

$ CompRes numeric COMPLAINT RESOLUTION

$ Advertising numeric ADVERTISING

$ ProdLine numeric PRODUCT LINE

$ SalesFImage numeric SALES FORCE IMAGE

$ ComPricing numeric COMPETITIVE PRICING

$ WartyClaim numeric WARRANTY AND CLAIMS

$ OrdBilling numeric ORDER & BILLING

$ DelSpeed numeric DELIVERY SPEED

**$ Satisfaction numeric CUSTOMER SATISFACTION**

* **EXPLORING THE FIRST FEW ROWS OF DATA**

> head(mydata[,2:13],5)

ProdQual Ecom TechSup CompRes Advertising ProdLine SalesFImage ComPricing WartyClaim OrdBilling

1 8.5 3.9 2.5 5.9 4.8 4.9 6.0 6.8 4.7 5.0

2 8.2 2.7 5.1 7.2 3.4 7.9 3.1 5.3 5.5 3.9

3 9.2 3.4 5.6 5.6 5.4 7.4 5.8 4.5 6.2 5.4

4 6.4 3.3 7.0 3.7 4.7 4.7 4.5 8.8 7.0 4.3

5 9.0 3.4 5.2 4.6 2.2 6.0 4.5 6.8 6.1 4.5

DelSpeed Satisfaction

1 3.7 8.2

2 4.9 5.7

3 4.5 8.9

4 3.0 4.8

5 3.5 7.1 4

* **FIVE POINT SUMMARY**

> summary(mydata[,2:13])

**ProdQual Ecom TechSup CompRes Advertising ProdLine**

Min. : 5.000 Min. :2.200 Min. :1.300 Min. :2.600 Min. :1.900 Min. :2.300

1st Qu.: 6.575 1st Qu.:3.275 1st Qu.:4.250 1st Qu.:4.600 1st Qu.:3.175 1st Qu.:4.700

**Median : 8.000 Median :3.600 Median :5.400 Median :5.450 Median :4.000 Median :5.750**

**Mean : 7.810 Mean :3.672 Mean :5.365 Mean :5.442 Mean :4.010 Mean :5.805**

3rd Qu.: 9.100 3rd Qu.:3.925 3rd Qu.:6.625 3rd Qu.:6.325 3rd Qu.:4.800 3rd Qu.:6.800

Max. :10.000 Max. :5.700 Max. :8.500 Max. :7.800 Max. :6.500 Max. :8.400

**SalesFImage ComPricing WartyClaim OrdBilling DelSpeed Satisfaction**

Min. :2.900 Min. :3.700 Min. :4.100 Min. :2.000 Min. :1.600 Min. :4.700

1st Qu.:4.500 1st Qu.:5.875 1st Qu.:5.400 1st Qu.:3.700 1st Qu.:3.400 1st Qu.:6.000

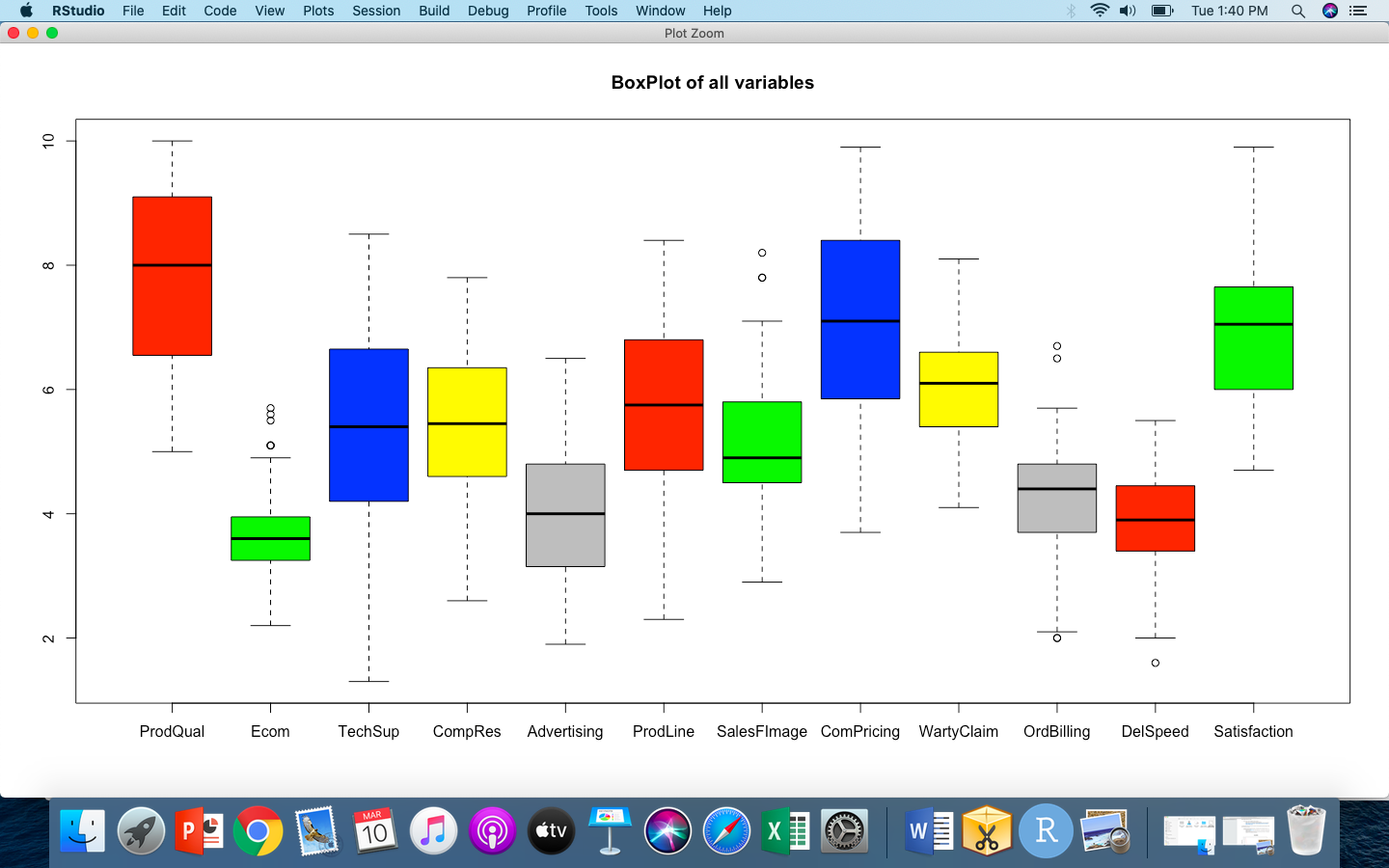
**Median :4.900 Median :7.100 Median :6.100 Median :4.400 Median :3.900 Median :7.050**

**Mean :5.123 Mean :6.974 Mean :6.043 Mean :4.278 Mean :3.886 Mean :6.918**

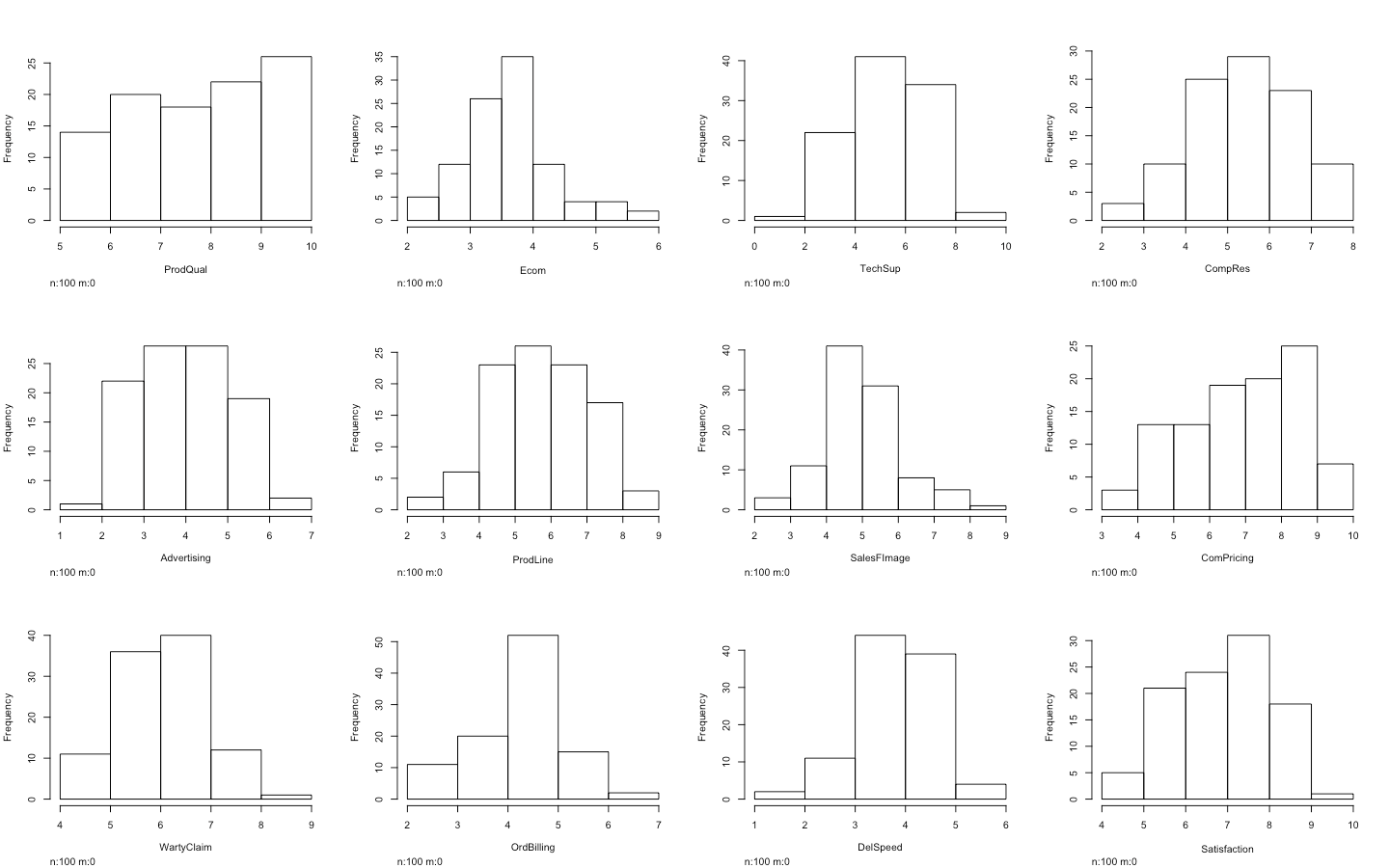
3rd Qu.:5.800 3rd Qu.:8.400 3rd Qu.:6.600 3rd Qu.:4.800 3rd Qu.:4.425 3rd Qu.:7.625

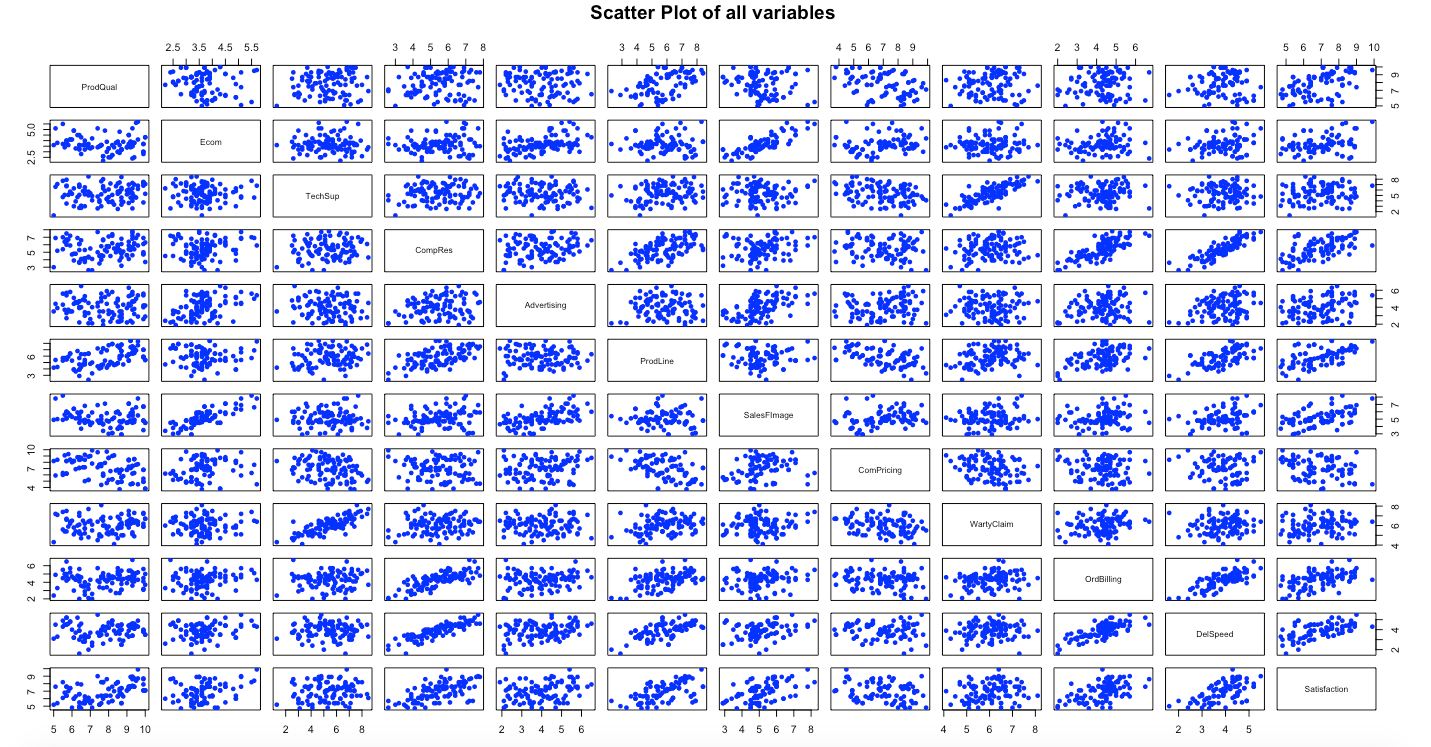
Max. :8.200 Max. :9.900 Max. :8.100 Max. :6.700 Max. :5.500 Max. :9.900

* **No missing values in the data**
* **Mean and Median do not show much deviation from each other**
* **BOXPLOT**



* **Few outliers are present in the data**
* **HISTOGRAM**



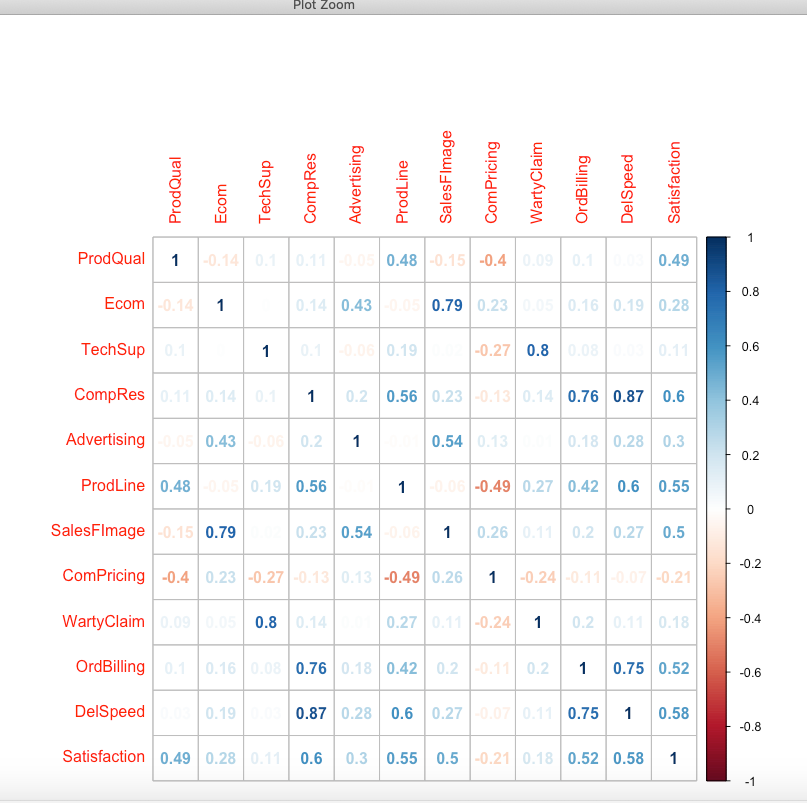
* **Variables seem to follow near normal distribution**
* **SCATTER PLOT**
* **There seems to be multicollinearity present in the data**

**INFERENCES FROM EDA**

* There are no missing values in the data
* There are few minor outliers in the data in the variables Ecom, OrdBilling, Salesfimage and DelSpeed
* Variables seem to follow near normal distribution
* Scatter plot indicates Presence of multi collinearity
* Since the Mean and Median values for all variables do not display high deviation, indicating no major impact of outliers on the data,we will proceed with outliers as is.

**INVESTIGATE MULTICOLLINEARITY IN THE DATA**

corrplot(mydatacor, method = "number")



* Many Independent variables are highly correlated such as :
  + ecom with salesfimage, techsup withwartyclaim, compres with ordbilling and delspeed, ordbilling with delspeed
* Where as Satisfaction , which is our response variable is mildly correlated with some independent variables such as compres, prodline, salesfimage, ordbilling and delspeed

**SIMPLE LINEAR REGRESSION**

* Perform simple linear regression for **Satisfaction with every independent variable**

**Summary statistics for simple linear regression results with 11 independent variables – 11 MODELS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model no | Independent Variable | Mulitple R squared | Adjusted R squared | P Value |
| 1 | ProdQual | 0.2365 | 0.2287 | 2.90E-07 |
| 2 | Ecom | 0.07994 | 0.07056 | 0.004368 |
| 3 | TechSup | 0.01268 | 0.002603 | 0.2647 |
| 4 | CompRes | 0.3639 | 0.3574 | 3.09E-11 |
| 5 | Advertising | 0.09282 | 0.08357 | 0.002056 |
| 6 | ProdLine | 0.3031 | 0.296 | 2.95E-09 |
| 7 | SalesFImage | 0.2502 | 0.2426 | 1.16E-07 |
| 8 | ComPricing | 0.04339 | 0.03363 | 0.03756 |
| 9 | WartyClaim | 0.03152 | 0.02164 | 0.0772 |
| 10 | OrdBilling | 0.2722 | 0.2648 | 2.60E-08 |
| 11 | DelSpeed | 0.333 | 0.3262 | 3.30E-10 |
|  |  |  |  |  |

* N**one of simple linear regression models are successful in explaining significant variance** in our dependent variable
* In Simple Linear regression model - **Model no 4 (var - compres) and 11 (var - delspeed) have the highest variance** explained and are also statistically significant with a lower p value.

LET US DEEP DIVE IN THE RESULTS FROM **MODEL 4** WHICH IS A SLR TO REGRESS **SATISFACTION WITH INDEPENDENT VARIABLE COMPRES** AS BELOW:

Residuals:

Min 1Q Median 3Q Max

-2.40450 -0.66164 0.04499 0.63037 2.70949

Coefficients:

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

(Intercept) 3.68005 0.44285 8.310 5.51e-13 \*\*\*

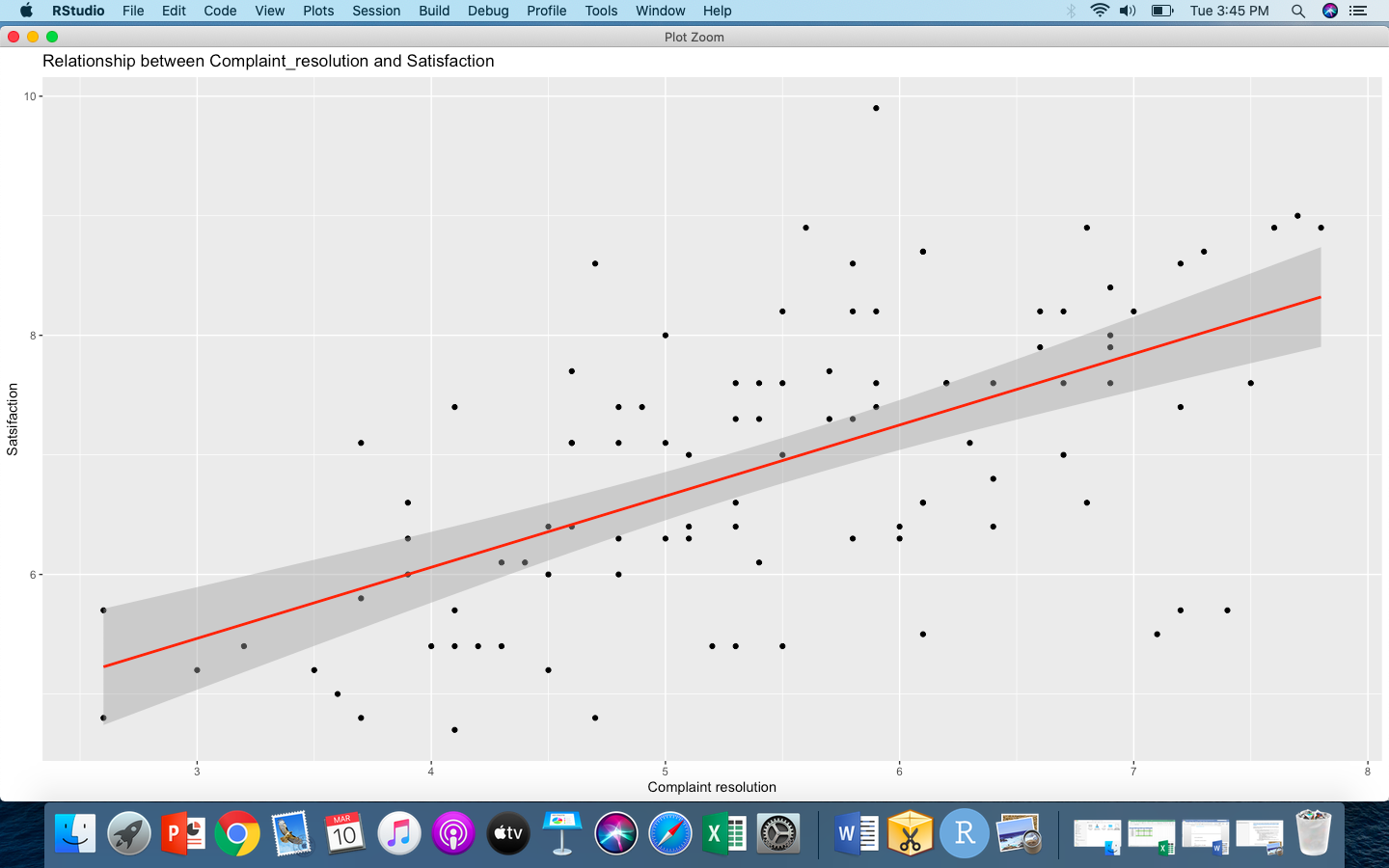
CompRes 0.59499 0.07946 7.488 3.09e-11 \*\*\*

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

Residual standard error: 0.9554 on 98 degrees of freedom

Multiple R-squared:  **0.3639,** Adjusted R-squared**: 0.3574**

F-statistic: 56.07 on 1 and 98 DF, p-value: 3.085e-11

Visualizing the model 4 results

* Therefore from the above model (Model no 4) we can conclude that if we improve the average scores on compress(complain resolution) by 1 point ,it will lead to an improvement of 0.59 points in the average satisfaction score

**MULTIPLE LINEAR REGRESSION**

* Let us also study the results from multiple linear regression taking into account all independent variables together to regress our dependent variable -Satisfaction

lm(formula = Satisfaction ~ ., data = mydata)

Residuals:

Min 1Q Median 3Q Max

-1.52926 -0.30940 0.07764 0.41223 0.88593

Coefficients:

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

(Intercept) -0.751368 0.815650 -0.921 0.35950

ID 0.002145 0.002051 1.046 0.29846

ProdQual 0.363401 0.052300 6.948 6.41e-10 \*\*\*

Ecom -0.428813 0.134355 -3.192 0.00197 \*\*

TechSup 0.036657 0.063777 0.575 0.56693

CompRes 0.164749 0.101694 1.620 0.10884

Advertising -0.032713 0.061905 -0.528 0.59854

ProdLine 0.143919 0.080284 1.793 0.07651 .

SalesFImage 0.798778 0.097946 8.155 2.39e-12 \*\*\*

ComPricing -0.036226 0.046801 -0.774 0.44100

WartyClaim -0.114652 0.123742 -0.927 0.35673

OrdBilling 0.158780 0.104299 1.522 0.13155

DelSpeed 0.173458 0.196473 0.883 0.37975

---

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

Residual standard error: 0.562 on 87 degrees of freedom

**Multiple R-squared: 0.8046, Adjusted R-squared: 0.7776**

F-statistic: 29.85 on 12 and 87 DF, **p-value: < 2.2e-16**

**CHECK VARIANCE INFLATION FACTOR FOR ABOVE MULITPLE LINEAR REGRESSION MODEL to IDENTIFY IMPACT OF MULTICOLLINEARTIY**

vif(model\_12\_multi\_linear\_regression)

ID ProdQual Ecom TechSup CompRes Advertising ProdLine SalesFImage ComPricing WartyClaim

1.109457 1.671241 2.776079 2.985817 4.732628 1.525244 3.494539 3.457106 1.638619 3.224552

OrdBilling DelSpeed

2.941208 6.525324

* Here we can see that our new model which includes all variables is able to explain 78% of variance in our dependent variable
* Ecom, SalesFimage and ProdQual are the most significant variables in the model
* However some variables such compres and Delspeed have a VIF of 4 and above indicating multicollinearity presence in model
* Therefore we will proceed with conducting Factor analysis to reduce the number of variables in form of meaningful factors and achieve a better regression model.

**PRINCIPAL COMPONENT/FACTOR ANALYSIS**

**FEASABILITY ANALYSIS**

**CHECK FEASIBILITY FOR CONDUCTING PCA/FACTOR ANALYSIS USING THE BELOW MENTIONED TESTS:**

**BARTLET TEST FOR MULTICOLLINEARITY IN DATA**

* H0: Data has no multicollinearity
* H1:Data has multicollinearity

**Result**

$chisq

[1] 769.6422

$p.value

[1] 1.65971e-120

$df

[1] 66

* Since the pvalue is less than alpha = 0.05, we reject the null hypothesis and **conclude that there is multicollinearity in the data**

**KMO TEST TO CHECK FOR DATA SAMPLE SUFFICIENCY**

**Result**

Kaiser-Meyer-Olkin factor adequacy

Call: KMO(r = mydata)

Overall MSA = 0.66

MSA for each item =

ID ProdQual Ecom TechSup CompRes Advertising ProdLine SalesFImage ComPricing

0.65 0.50 0.59 0.52 0.84 0.82 0.70 0.52 0.77

WartyClaim OrdBilling DelSpeed Satisfaction

0.51 0.78 0.72 0.66

* Since the overall MSA value is more than 0.5 the **sample size is adequate to perform factor/PCA analysis**

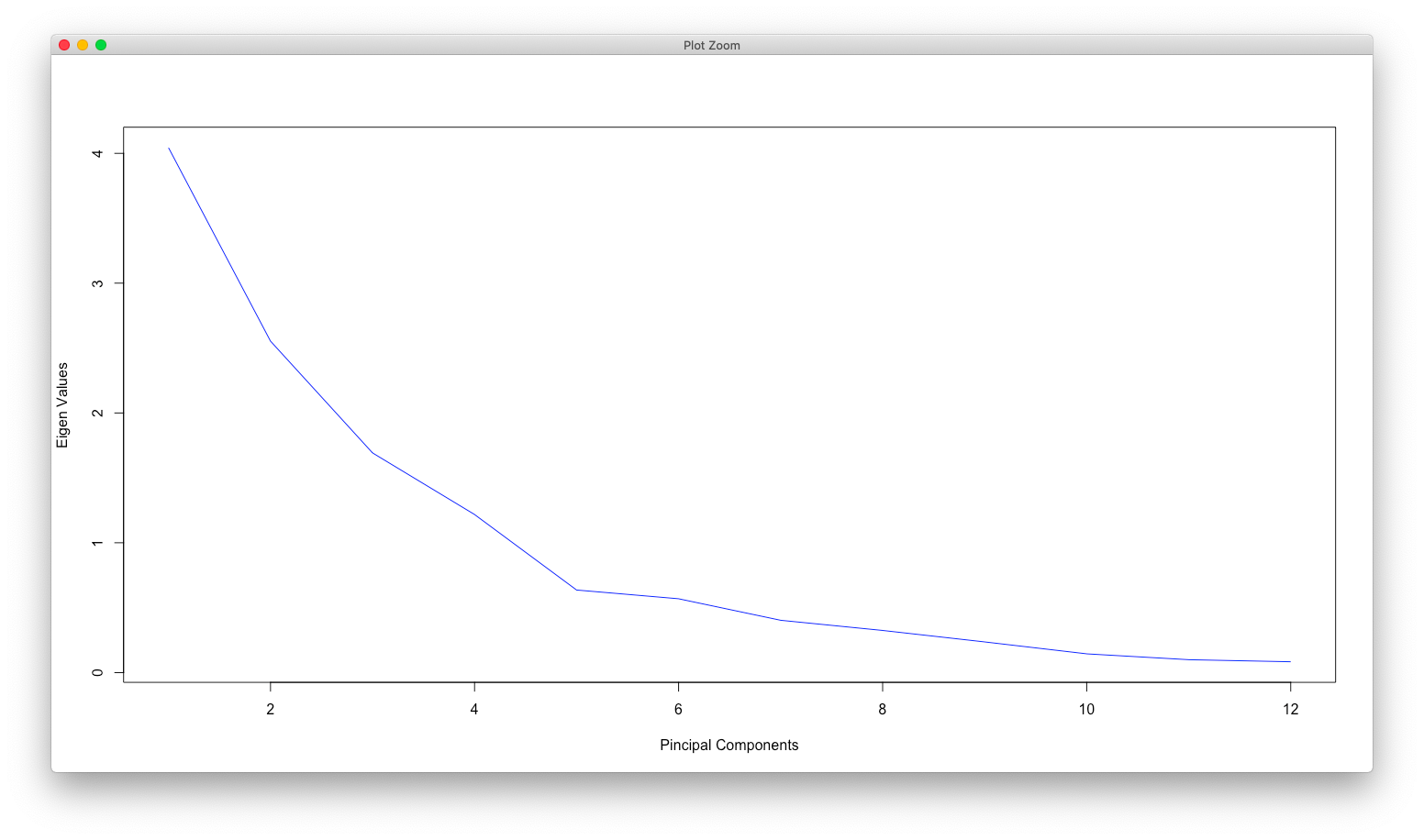
**IDENTIFYING THE NUMBER OF FACTORS**

**EIGEN VALUES**

[1] 4.04285997 2.55292440 1.69222417 1.21754639 0.63596293 0.56853132 0.40282774 0.32448016 0.23613948 0.14422355 0.09913845

[12] 0.08314143

**SCREE PLOT**



* Using Kaiser normalization rule which indicates considering the no of factors until eigen value is more than 1 , we can see in above graph and from eigen values that up to **four factor the eigen value is more than one**
* Hence we select 4 factor solution for our analysis

**IDENTIFYING THE BEST SOLUTION**

**POST IDENTIFYING THE OPTIMAL NO OF FACTORS WE RAN VARIOUS TYPES OF PCA/FACTOR ANALYSIS TO IDENTIFY BEST MODEL AND BELOW MENTIONED ARE THE SUMMARY STATISTIC FROM ALL FACTOR ANALYSIS MODELS**

**SUMMARY STATISTIC – PCA/FACTOR ANALYSES**

|  |  |  |  |
| --- | --- | --- | --- |
| Factor Type | Rotation | Factor Loadings | Cumulative Variance |
| PCA | None | 3.43 2.55 1.69 1.09 | 0.80 |
| PCA | Varimax | 2.89 2.23 1.86 1.77 | 0.80 |
| Factor Analysis | None | 3.21 2.22 1.50 0.68 | 0.69 |
| Factor Analysis | Varimax | 2.63 1.97 1.64 1.37 | 0.69 |

* Since the factor loading are more balanced as well as the cumulative variance explained is highest for **PCA – Varimax method,** we shortlist PCA using Varimax rotation as our factor analysis method

**CLASSIFYING VARIABLES AS PART OF FACTORS BASIS STANDARDISED LOADINGS:**

**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

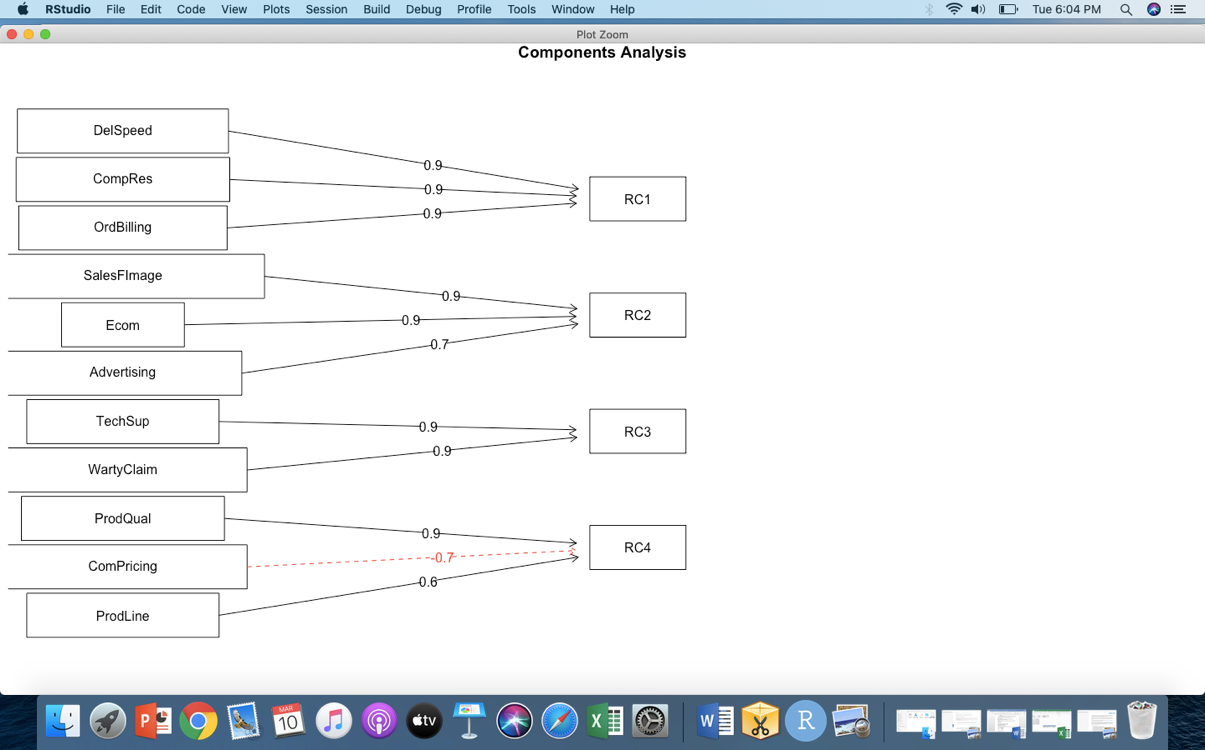
ComPricing -0.09 0.23 -0.25 -0.72 0.64 0.359 1.5

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

**FA DIAGRAM FOR CLASSIFYING VARIABLES INTO FACTORS:**



**HENCE THE 11 VARIABLES ARE CLASSIFIED IN FOUR FACTORS AND NAMED AS BELOW:**

|  |  |  |  |
| --- | --- | --- | --- |
| **RC1** | **RC2** | **RC3** | **RC4** |
| **Order Servicing effectiveness** | **Brand Image** | **After Sales Service** | **Product & Pricing** |
| Complaint resolution | Ecommerce | Tech support | Product Quality |
| Order Billing | Advertising | Warranty &Claims | Product Line |
| Delivery Speed | Sales Force Image |  | Competition pricing |

**REGRESSION ANALYSIS USING THE 4 FACTORS**

**DATA PREPARATION AND KEY STEPS**

* Create a new data frame using the 4 factors and Satisfaction variable from original dataset
* Split the new dataset created in above step in train and test subsets
* Run Multiple Linear regression using satisfaction as dependent variable and the four factors as independent variables
* Predict the response variable using the above model and test data set
* Understand the model accuracy and validity

**Summary statistic for our MLR Model from Train Data set**

lm(formula = V5 ~ ., data = train)

Residuals:

Min 1Q Median 3Q Max

-1.76126 -0.46288 0.08743 0.50066 1.26112

Coefficients:

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

(Intercept) 6.95784 0.08317 83.653 < 2e-16 \*\*\*

RC1 0.63792 0.08035 7.940 3.40e-11 \*\*\*

RC2 0.49173 0.08242 5.966 1.06e-07 \*\*\*

RC3 0.03855 0.08295 0.465 0.644

RC4 0.54751 0.07923 6.910 2.34e-09 \*\*\*

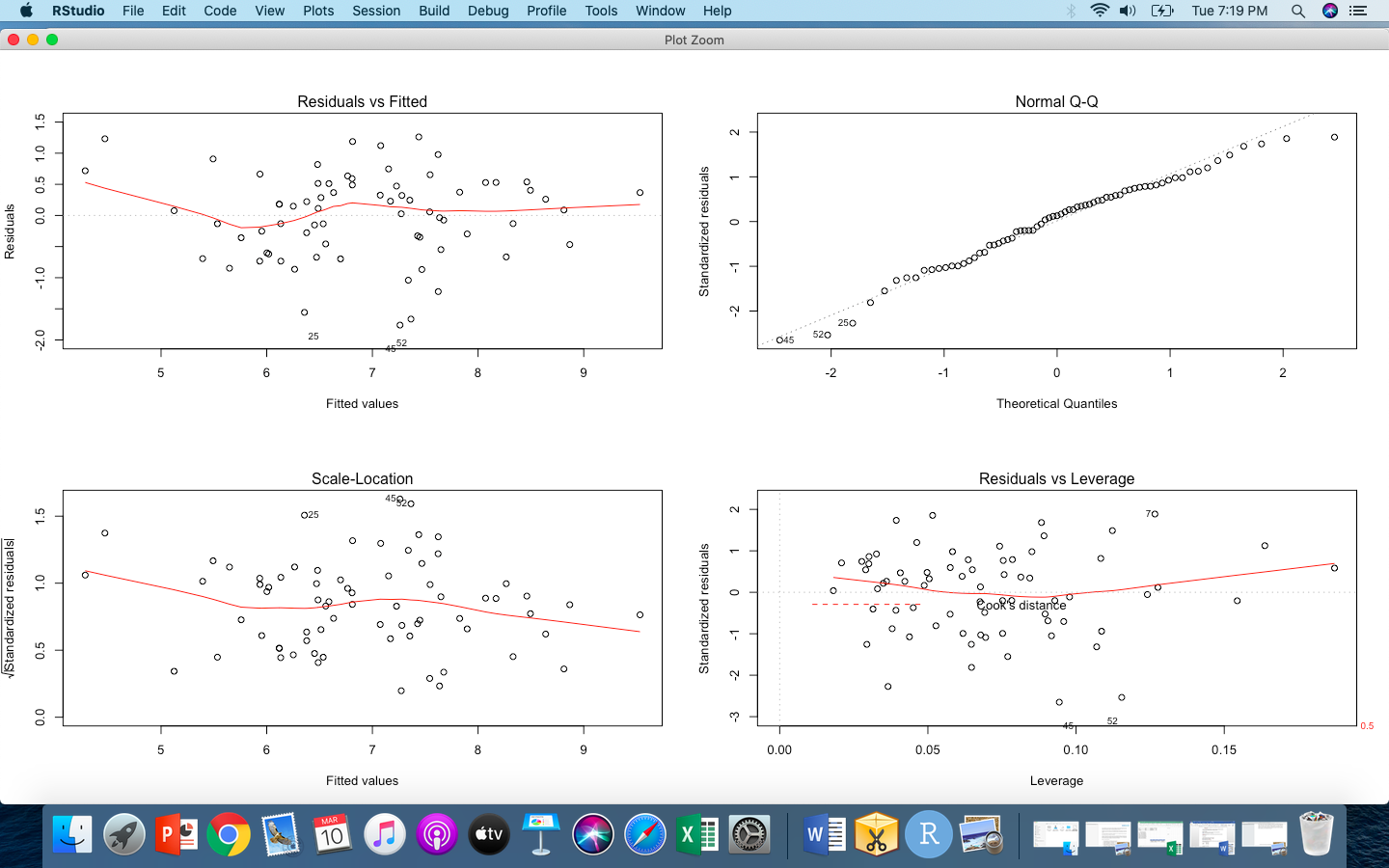
---

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

Residual standard error: 0.6982 on 66 degrees of freedom

**Multiple R-squared: 0.6917, Adjusted R-squared: 0.673**

F-statistic: 37.02 on 4 and 66 DF, **p-value: 3.251e-16**



**REGRESSION EQUATION**

**SATISFACTION = 6.95784 + 0.63792 (ORDER PLACEMENT EFFECTIVENESS) + 0.49173(BRAND IMAGE) + 0.03855(AFTER SALES SERVICE) +0.54751(PRODUCT & PRICING**

**SUMMARY STATISTIC FOR TEST DATA SET**

**Rsquare for Test data**

rsqre

[1] 0.5623987

**Mean absolute percentage error between Test and Train Data**

> mape

[1] 0.02173992

**MODEL VALIDITY**

* We will consider Adjusted R square to comment on model validity instead of Multiple R square since it is adjusted for multiple variables in our model and increases only when the model is improved due to an additional variable in the model and hence is not inflated
* The model indicates that keeping all other factors constant, if we improve RC1 (Order servicing effectiveness ) by one point then customer Satisfaction will improve by 0.639 points on average
* The adjusted R square for our Train data set is 67% where as for our test data set is 56% indicating that model is mildly underfit (difference b/w the Rsquare of train and test dataset is >10%)
* Since p value for model is lower than alpha = 0.05, the model is statistically significant in explaining the variance in our response variable
* RC1,RC2 and RC4 factors namely – Order servicing effectiveness, Brand Image and Product and Pricing are significant variables in the model where as After sales service is not a significant variable in the model and the model may be rerun by removing this factor to improve the model accuracy
* Residuals versus fitted values in the plot – best fit line is not linear indicating scope for improvement in model
* Normal Q-Q plot indicates the residuals are not following normal distribution (some residuals are not falling on the line)
* Scale location plot – indicates that variance of residuals is not constant
* Residual versus leverage plot indicates presence of few extreme values in the data

**TO SUMMARISE**

* The dataset contains 100 observations from 11 Independent variable and 1 dependent variable - Satisfaction
* There **are no missing values** in the data
* There are **few minor outliers** in the data in the variables Ecom, OrdBilling, Salesfimage and DelSpeed
* All **Variables seem to follow near normal distribution**
* Since the Mean and Median values for all variables do not display high deviation, indicating no major impact of outliers on the data,we will proceed with outliers as is
* Many **Independent variables are highly correlated** such as :
  + ecom with salesfimage, techsup withwartyclaim, compres with ordbilling and delspeed, ordbilling with delspeed
* Where as Satisfaction , which is our response variable is mildly correlated with some independent variables such as compres, prodline, salesfimage, ordbilling and delspeed
* N**one of simple linear regression models are successful in explaining significant variance** in our dependent variable
  + In Simple Linear regression model - **Model no 4 (var - compres) and 11 (var - delspeed) have the highest variance** explained and are also statistically significant with a lower p value.
* When we conduct multiple linear regression with all independent varaiables and Satisfaction as dependent variable, our model is able to explain 78% of variance in our dependent variable
  + However some variables such compres and Delspeed have a VIF of 4 and above indicating multicollinearity presence in model
  + Therefore we proceeded with conducting Factor analysis to reduce the number of variables in form of meaningful factors and achieve a better regression model.
* Using Kaiser normalization rule which indicates considering the no of factors until eigen value is more than 1 , **we selected 4 factor solution for our analysis**
  + Since the factor loading are more balanced as well as the cumulative variance explained is highest for **PCA – Varimax method,** we shortlist this as our factor analysis method
  + We formed and named 4 factors as below

|  |  |  |  |
| --- | --- | --- | --- |
| **RC1** | **RC2** | **RC3** | **RC4** |
| **Order Servicing effectiveness** | **Brand Image** | **After Sales Service** | **Product & Pricing** |

* Post this we ran a multiple linear regression using 4 factors as independent variable and satisfaction as dependent variable
  + **The adjusted R square for our Train data set is 67% where as for our test data set is 56% indicating that model is mildly underfit** (difference b/w Rsquare of train and test dataset is >10%)
  + Since p value is lower than alpha = 0.05 **the model is statistically significant in explaining the variance in our response variable**
  + **RC1,RC2 and RC4 factors namely – order servicing effectiveness, Brand Image and Product and Pricing are significant variables in the model where as After sales service is not a significant variable in the model** and the model may be rerun by removing this factor to improve the model accuracy

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**APPENDIX (SOURCE CODE)**

**#Load all required libraries**

**library(MASS)**

**library(car)**

**library(ggplot2)**

**install.packages("Hmisc")**

**library(Hmisc)**

**library(corrplot)**

**library(dplyr)**

**library(rpivotTable)**

**library(psych)**

**library(nfactors)**

**library(xlsx)**

**rpivotTable(mydata)**

**#Load the dataset**

**setwd("/Users/shweta/Desktop/PGPBABI")**

**mydata = read.csv("/Users/shweta/Desktop/PGPBABI/ADVANCED STATISTICS/Project/Factor-Hair-Revised.csv")**

**# EDA**

**dim(mydata)**

**names(mydata)**

**head(mydata[,2:13],5)**

**tail(mydata,5)**

**str(mydata)**

**summary(mydata[,2:13])**

**shapiro.test(mydata$Satisfaction)**

**sum(anyNA(mydata))**

**boxplot(mydata[,2:13], col = c("red","green","blue","yellow","grey") , main = "BoxPlot of all variables",number.cex = 0.7)**

**hist(mydata)**

**hist.data.frame(mydata[,2:13])**

**plot(mydata[,2:13], pch=16, col="blue", main="Scatter Plot of all variables")**

**mydatacor = cor(mydata[2:13])**

**# check for multicollinearity**

**corrplot(mydatacor, method = "number",type = "upper")**

**#Perform Simple Linear regression with every independent variable**

**model\_1 = lm(Satisfaction~ProdQual,data=mydata)**

**summary(model\_1,digits =5)**

**model\_2 = lm(Satisfaction~Ecom,data=mydata)**

**summary(model\_2)**

**model\_3 = lm(Satisfaction~TechSup,data=mydata)**

**summary(model\_3)**

**model\_4 = lm(Satisfaction~CompRes,data=mydata)**

**summary(model\_4)**

**model\_5 = lm(Satisfaction~Advertising,data=mydata)**

**summary(model\_5)**

**model\_6 = lm(Satisfaction~ProdLine,data=mydata)**

**summary(model\_6)**

**model\_7 = lm(Satisfaction~SalesFImage,data=mydata)**

**summary(model\_7)**

**model\_8 = lm(Satisfaction~ComPricing,data=mydata)**

**summary(model\_8)**

**model\_9 = lm(Satisfaction~WartyClaim,data=mydata)**

**summary(model\_9)**

**model\_10 = lm(Satisfaction~OrdBilling,data=mydata)**

**summary(model\_10)**

**model\_11 = lm(Satisfaction~DelSpeed,data=mydata)**

**summary(model\_11)**

**#Visulaise model 4 results which is the best SLR model**

**qplot(mydata$CompRes, mydata$Satisfaction, data = mydata, main = "Relationship between Complaint\_resolution and Satisfaction") +**

**stat\_smooth(method="lm", col="red") + xlab("Complaint resolution") +ylab("Satsifaction")**

**par(mfrow~c(2,2))**

**plot(model\_4)**

**#perform mulitple linear regression with all variables**

**model\_12\_multi\_linear\_regression = lm(Satisfaction~., data=mydata)**

**summary(model\_12\_multi\_linear\_regression)**

**vif(model\_12\_multi\_linear\_regression)**

**# Barlett Sphericity Test for checking the possibility of data dimension reduction**

**print(cortest.bartlett(mydatacor,nrow(mydata)))**

**#KMO test to check for sample adequacy for factor analysis**

**KMO(mydata)**

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

**factors<-eigen(mydatacor)**

**eigenvalues<-factors$values**

**eigenvectors<-factors$vectors**

**eigenvalues**

**#Plotting SCREE Graphs**

**plot(eigenvalues,type="lines",**

**xlab="Pincipal Components",ylab="Eigen Values",col = "blue")**

**# Conduct PCA - rotated and unrotated**

**pc\_unrotate<-principal(mydata[,2:12],nfactors = 4,rotate="none")**

**pc\_rotate<-principal(mydata[,2:12],nfactors = 4,rotate="varimax")**

**pc\_unrotate**

**pc\_rotate**

**RotatedProfile=plot(pc\_rotate,row.names(pc\_rotate$loadings),cex=1.0)**

**fa.diagram(pc\_rotate)**

**# Conduct FA - rotated and unrotated**

**mynewdata = mydata[,2:12]**

**mynewdata**

**summary(mynewdata)**

**newcor = cor(mynewdata)**

**factor\_analysis\_unrotate = fa(r=newcor,nfactors=4,rotate ="none",fm = "pa")**

**factor\_analysis\_rotate = fa(r=newcor,nfactors=4,rotate ="varimax",fm = "pa")**

**factor\_analysis\_unrotate**

**factor\_analysis\_rotate**

**fa.diagram(factor\_analysis)**

**fa.diagram(pc\_rotate)**

**pc\_rotate$scores**

**#create a new dataframe using new factors and satisfaction variable**

**new\_data = as.data.frame(cbind(pc\_rotate$scores, mydata$Satisfaction))**

**summary(new\_data)**

**# multiple linear regression analysis using factor scores**

**install.packages("caTools")**

**library(caTools)**

**# Split data in test and train**

**set.seed(2)**

**index = sample.split(new\_data$V5, SplitRatio = 0.7)**

**train = subset(new\_data,index== TRUE)**

**test = subset(new\_data,index== FALSE)**

**dim(train)**

**dim(test)**

**MLR\_FACTORS = lm(V5~.,data = train)**

**summary(MLR\_FACTORS)**

**pred = predict(MLR\_FACTORS,newdata = test)**

**pred**

**summary(pred)**

**mean(pred)**

**mean(new\_data$V5)**

**# mean absolute percentage error between train and test data**

**mape = ((mean(pred) - mean(new\_data$V5)) /mean(new\_data$V5))**

**mape**

**# r square for test data**

**SSE = sum(pred - test$V5)^2**

**SST = sum((mean(test$V5)- test$V5)^2)**

**rsqre = 1- SSE/SST**

**rsqre**

**# plot our regression model**

**par(mfrow = c(2,2))**

**plot(MLR\_FACTORS)**

**# save the file as rdata**

**save.image()**

**save.image(file = "regression and factor analysis project.RData")**