REGRESSION MODEL TO PREDICT CUSTOMER SATISFACTION

An advanced statistics project using Factor Analysis and Multiple Linear Regression.

Chandrima Nandy

February 2020

Contents

Objective

The objective of the project is to use the dataset 'Factor-Hair-Revised.csv' to build an optimum regression model to predict customer satisfaction

1.1 EDA - Basic data summary, Univariate, Bivariate analysis, graphs

1.1.1 EDA - Check for Outliers and missing values and check the summary of the dataset

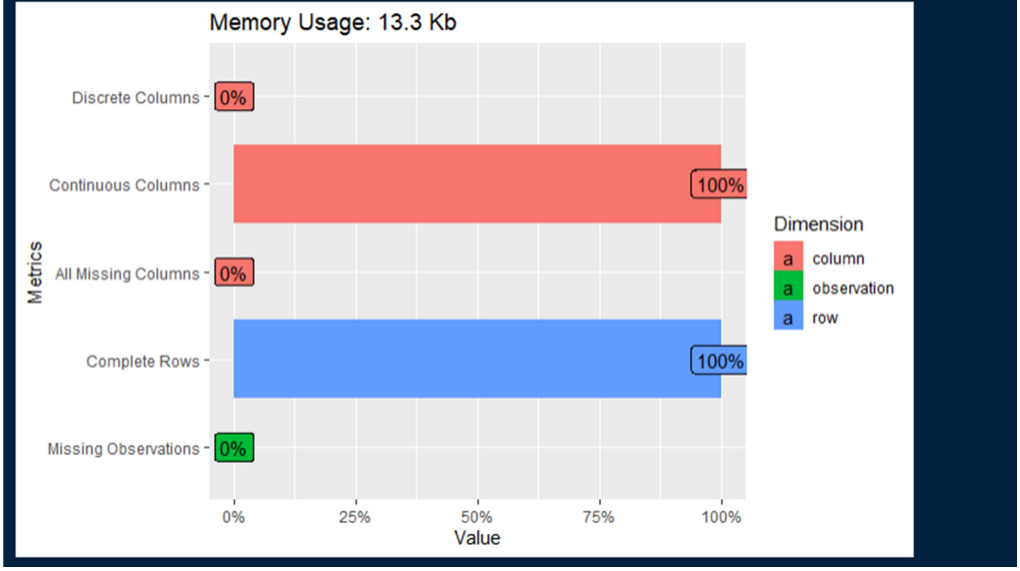
The answers to both of the above questions are as follows:
The given dataset is comprised of 100 rows and 13 columns.

```
Checking the dimension of the dataset
```{r}
dim(FactorHairRevised_DF)
```

 [1] 100 13
```

Further analysis to the structure of the dataset reveals that all the 13 columns have continuous variables with no missing observations.

```
```{r}
plot_intro(FactorHairRevised_DF)
```
```



Provided below is a summary of each of the continuous variables – ID, ProdQual, Ecom, TechSup, CompRes, Advertising, ProdLine, SalesFImage, ComPricing, WartyClaim, OrdBilling, DelSpeed and Satisfaction. (For our analysis purpose, we would ignore the column ID).

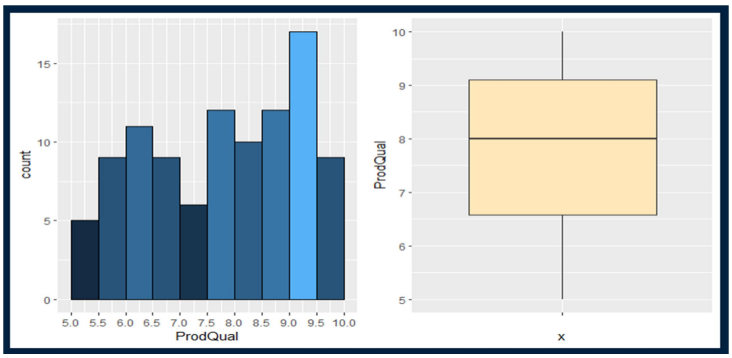
```r
summary(FactorHairRevised_DF)
```

```
       ID            ProdQual          Ecom          TechSup         CompRes       Advertising      ProdLine       SalesFImage      ComPricing   
 Min.   :  1.00   Min.   : 5.000   Min.   :2.200   Min.   :1.300   Min.   :2.600   Min.   :1.900   Min.   :2.300   Min.   :2.900   Min.   :3.700  
 1st Qu.: 25.75   1st Qu.: 6.575   1st Qu.:3.275   1st Qu.:4.250   1st Qu.:4.600   1st Qu.:3.175   1st Qu.:4.700   1st Qu.:4.500   1st Qu.:5.875  
 Median : 50.50   Median : 8.000   Median :3.600   Median :5.400   Median :5.450   Median :4.000   Median :5.750   Median :4.900   Median :7.100  
 Mean   : 50.50   Mean   : 7.810   Mean   :3.672   Mean   :5.365   Mean   :5.442   Mean   :4.010   Mean   :5.805   Mean   :5.123   Mean   :6.974  
 3rd Qu.: 75.25   3rd Qu.: 9.100   3rd Qu.:3.925   3rd Qu.:6.625   3rd Qu.:6.325   3rd Qu.:4.800   3rd Qu.:6.800   3rd Qu.:5.800   3rd Qu.:8.400  
 Max.   :100.00   Max.   :10.000   Max.   :5.700   Max.   :8.500   Max.   :7.800   Max.   :6.500   Max.   :8.400   Max.   :8.200   Max.   :9.900  
   WartyClaim      OrdBilling       DelSpeed       Satisfaction   
 Min.   :4.100   Min.   :2.000   Min.   :1.600   Min.   :4.700  
 1st Qu.:5.400   1st Qu.:3.700   1st Qu.:3.400   1st Qu.:6.000  
 Median :6.100   Median :4.400   Median :3.900   Median :7.050  
 Mean   :6.043   Mean   :4.278   Mean   :3.886   Mean   :6.918  
 3rd Qu.:6.600   3rd Qu.:4.800   3rd Qu.:4.425   3rd Qu.:7.625  
 Max.   :8.100   Max.   :6.700   Max.   :5.500   Max.   :9.900  
```

Now we proceed with the univariate exploratory data analysis. Since all the variables are continuous in nature we would stick to histogram and boxplots.

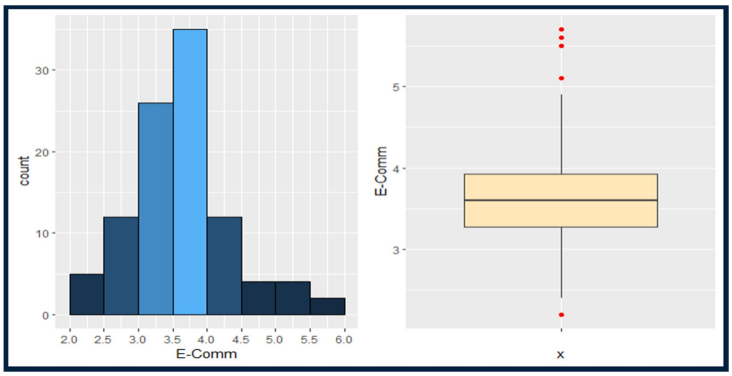
Analysing Product Quality :

Not a normal distribution and no outliers present for ProdQual; a large number of data points are present between 9 and 9.5.



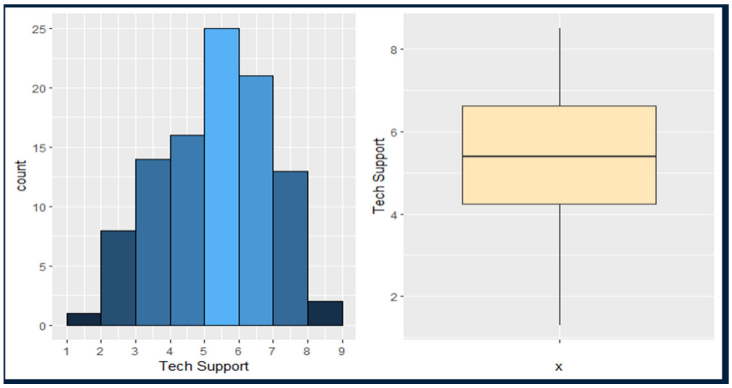
Analysing E-Commerce :

Some resemblance with a normal distribution where a significant number of data points are lying between 3.5 and 4. E-Comm has a few outliers present in its data.



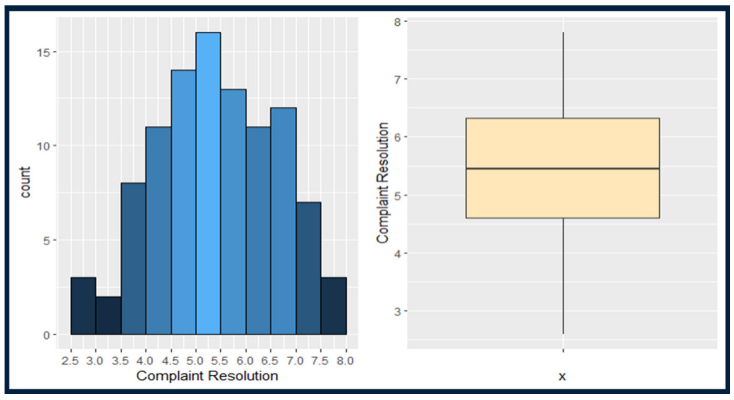
Analysing Technical Support:

Close to normally distributed with ~45% of data points lying between 5and 7. No outliers present in Tech Support.



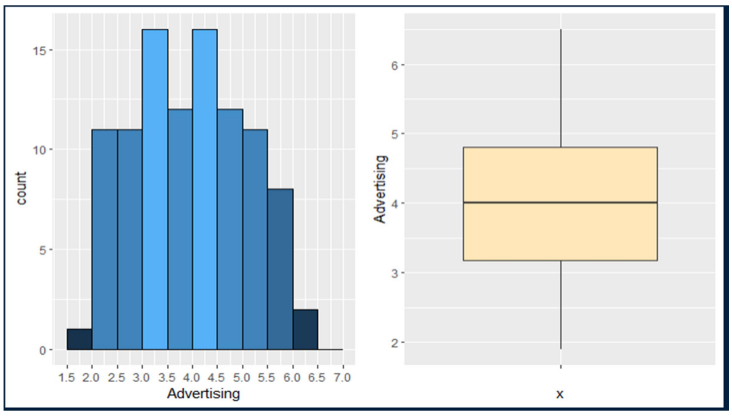
Analysing Complaint Resolution:

A normal distribution and no outliers are present in Complaint Resolution.



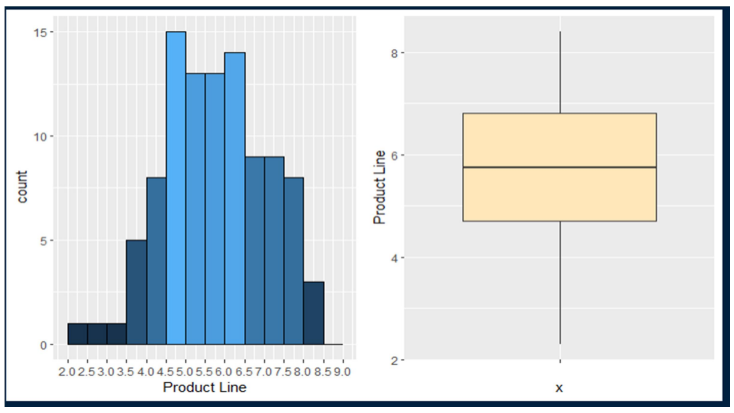
Analysing Advertising:

Not a normal distribution and no outliers are present in Advertising.



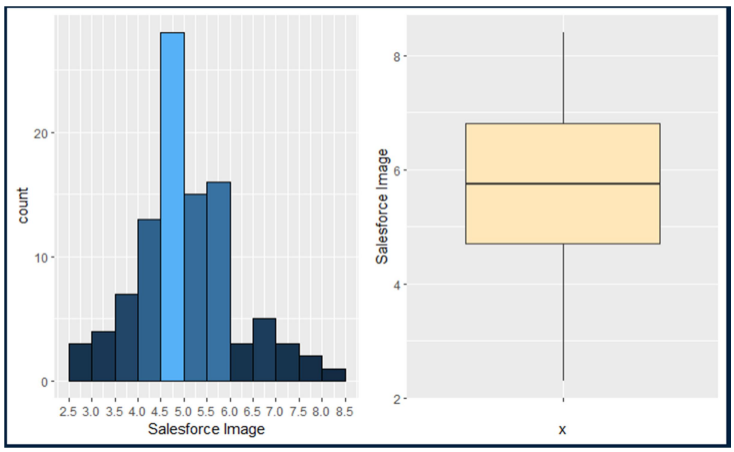
Analysing Product Line :

Some resemblance with a normal distribution and no outliers are present in Product Line.



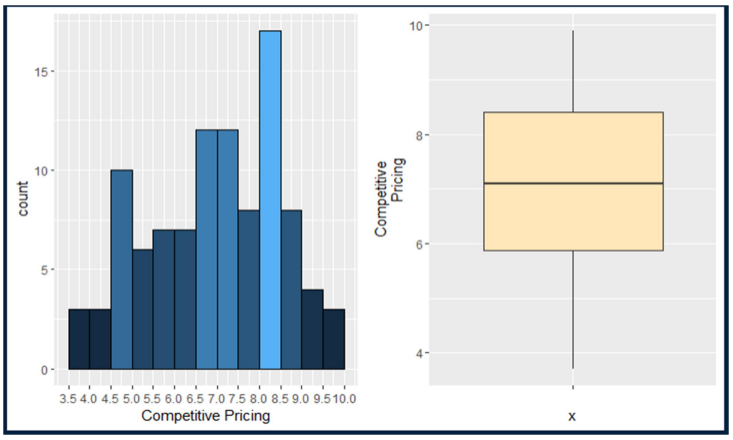
Analysing SalesForce Image :

Some resemblance with a normal distribution and no outliers are present in Salesforce Image. Close to 30% data points lie between 4.5-5.



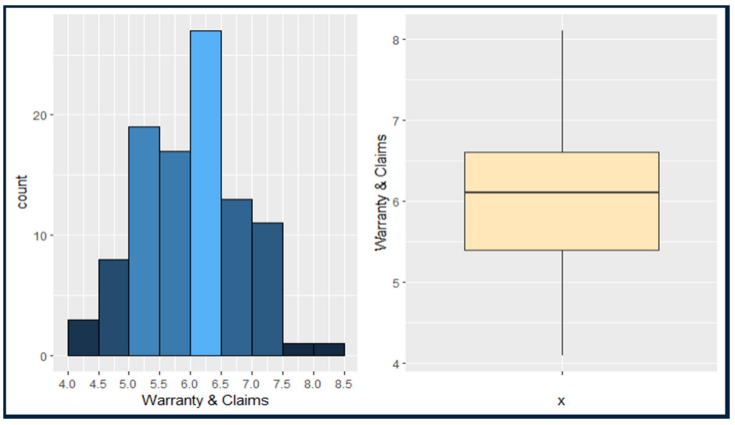
Analysing Competitive Pricing :

Not a normal distribution and no outliers are present in Competitive Pricing.



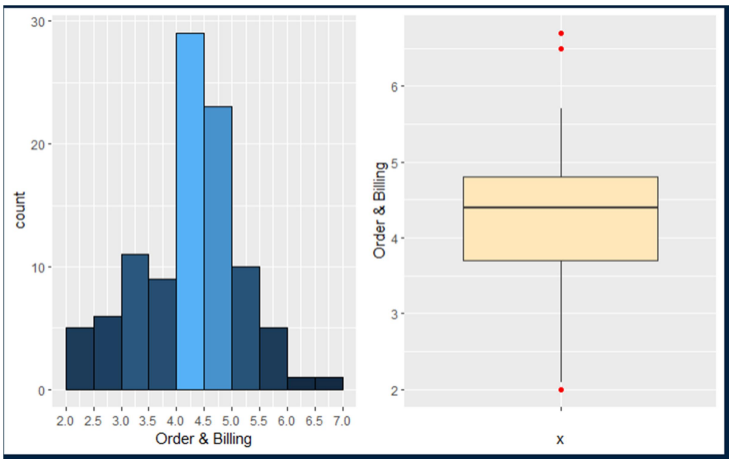
Analysing Warranty & Claims :

Some resemblance with a normal distribution and no outliers are present in Warranty & Claims. Over 25% of data points lie between 6 and 6.5.



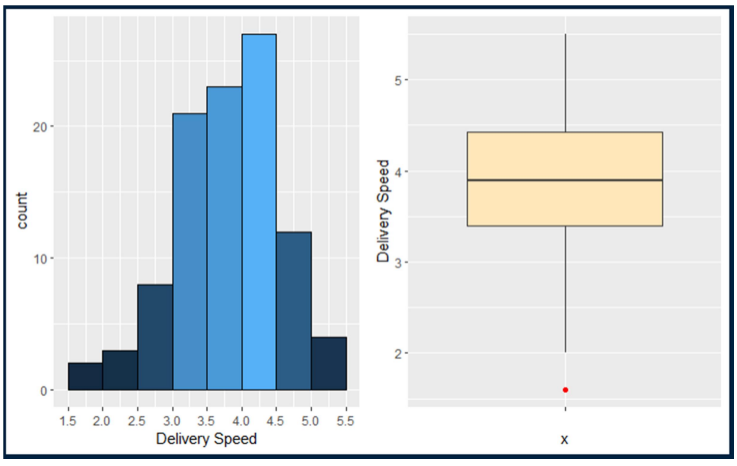
Analysing Order & Billing :

Some resemblance with a normal distribution with outliers being present in Order & Billing. ~50% of data points lie between 4 and 5.



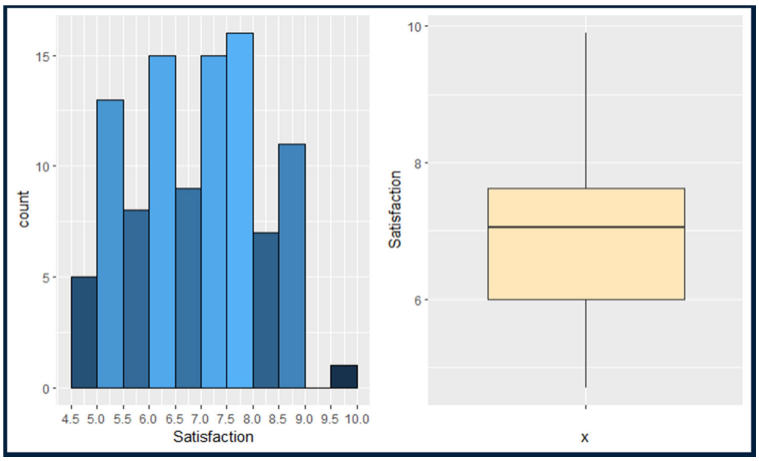
Analysing Delivery Speed :

Some resemblance with a normal distribution with an outlier being present in DeliverySpeed. More than 65% of data points lie between 3 and 4.5.

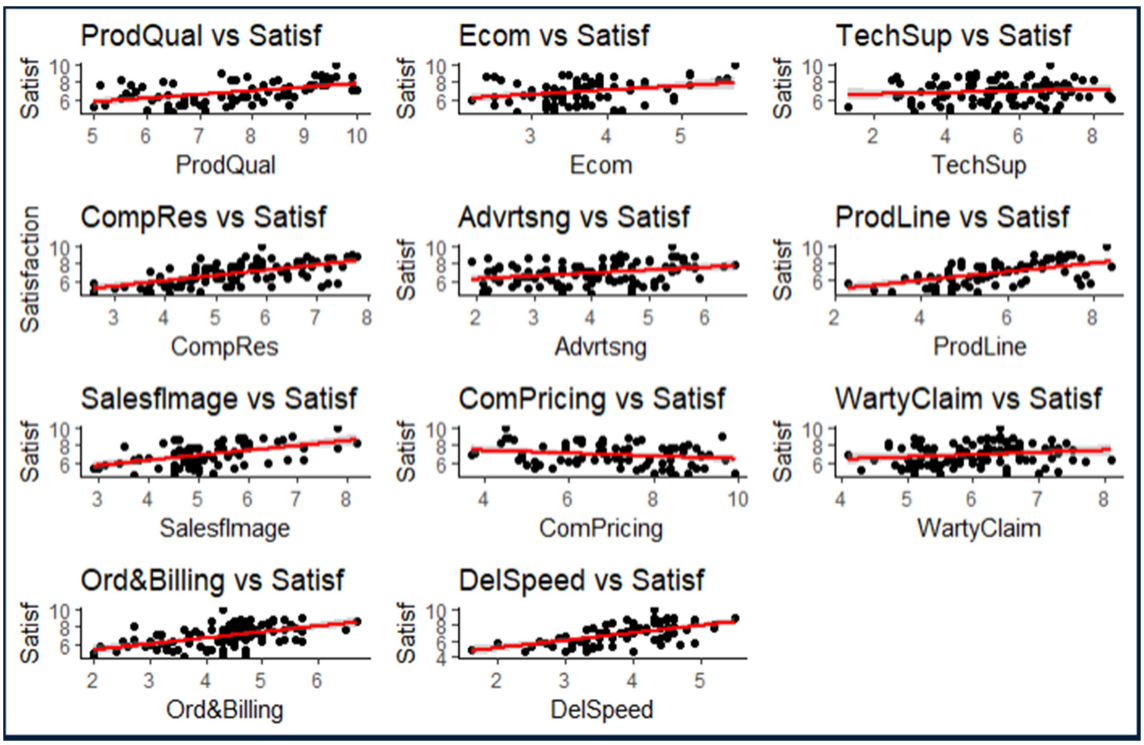


Analysing Customer Satisfaction :

Not a normal distribution and no outliers are present in Customer Satisfaction.

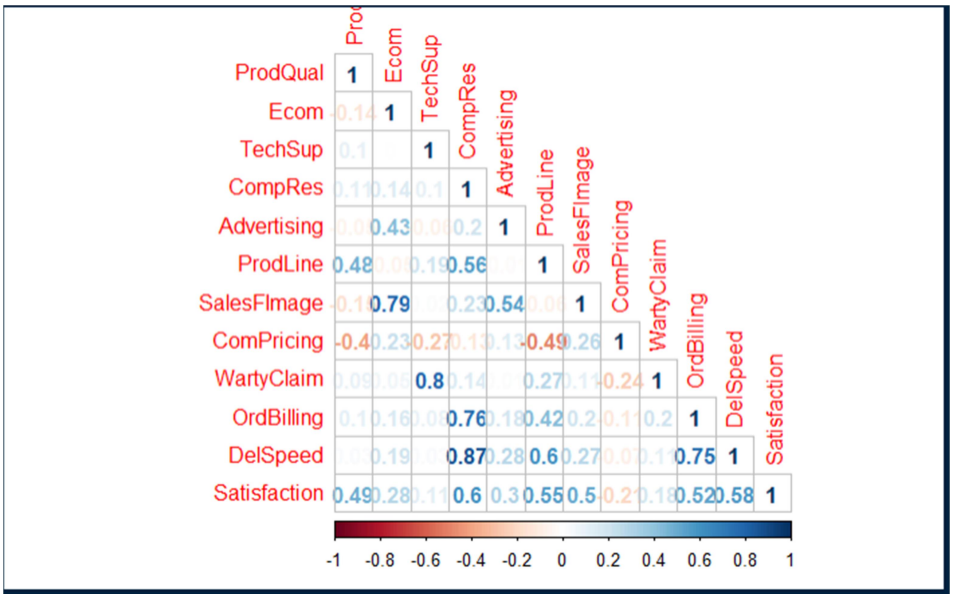


Bi-variate analysis - checking the dependency of Customer Satisfaction on the other 11 variables.



Overall we find weak association between Satisfaction and the other variables. Still, Customer Satisfaction is somewhat more dependent on Product Quality, Complaint Resolution, Product Line, Salesforce Image, Order & Billing and Delivery Speed as against Ecommerce, Technical Support, Advertising and Warranty & Claims. Satisfaction has a weak negative relation with Competitive Pricing.

2. Check for multicollinearity



We haven't taken the 1st column in the correlation matrix as it is an ID column.

We find the presence of multicollinearity in the data as is evident from some of the high correlation coefficients given by the matrix. It is acceptable for Satisfaction to have correlation with the rest of the independent variables but some of the independent variables too exhibit high correlation amongst themselves. For instance, between Delivery Speed and Complaint Resolution the correlation coefficient is 0.87. Similarly, the correlation coefficient is 0.8 between Warranty&Claim and Technical Support. Competitive Pricing exhibits negative correlation with Product Quality and Product Line.

A linear regression was also run, regressing Satisfaction on all the other 11 independent variables and the VIF values were checked which are as follows:

```r
attach(FactorHairRevised_DF)
model1 <- lm( Satisfaction~., data = FactorHairRevised_DF[,2:12])
summary(model1)


library(car)
vif(model1)
```

```
ProdQual      Ecom   TechSup   CompRes Advertising  ProdLine SalesFImage ComPricing WartyClaim OrdBilling  DelSpeed
1.635797  2.756694  2.976796  4.730448  1.508933  3.488185  3.439420  1.635000  3.198337  2.902999  6.516014
```

Since VIF > 2.5 for most of the variables, this confirms the presence of multicollinearity in the data.

3. Simple Linear Regression (with every variable)

Before performing the simple linear regression models, we check for normality of variables. Since, visual exploratory data analysis (as described before) is usually unreliable, we apply a significance test called **Shapiro-Wilk's normality test** whereby the hypothesis are as follows:

Null Hypothesis H0 : The variable considered is normally distributed

Alternative Hypothesis H1: The variable considered is not normally distributed.

```
```{r}
shapiro.test(FactorHairRevised_DF$ProdQual)
shapiro.test(FactorHairRevised_DF$Ecom)
shapiro.test(FactorHairRevised_DF$TechSup)
shapiro.test(FactorHairRevised_DF$CompRes)
shapiro.test(FactorHairRevised_DF$Advertising)
shapiro.test(FactorHairRevised_DF$ProdLine)
shapiro.test(FactorHairRevised_DF$SalesFImage)
shapiro.test(FactorHairRevised_DF$ComPricing)
shapiro.test(FactorHairRevised_DF$WartyClaim)
shapiro.test(FactorHairRevised_DF$OrdBilling)
shapiro.test(FactorHairRevised_DF$DelSpeed)
shapiro.test(FactorHairRevised_DF$Satisfaction)
```
```

The results are as follows:

Variable Name	w- statistic value	p-value
ProdQual	0.94972	0.0007953
Ecom	0.95852	0.003157
TechSup	0.98626	0.39
CompRes	0.98646	0.4023
Advertising	0.97626	0.06769
ProdLine	0.98692	0.4324
SalesFImage	0.97403	0.04534
CompPricing	0.96758	0.01448
WartyClaim	0.99094	0.7404
OrdBilling	0.97405	0.04549
DelSpeed	0.98161	0.177
Satisfaction	0.97516	0.05556

Assessing the p-values at the 5% level of significance, we find that Technical Support, Complaint Resolution, Advertising, Product Line, Warranty&Claims, Delivery Speed and Customer Satisfaction all have p values >= 0.05, hence we accept the null hypothesis that they are all normally distributed.

For the other variables, since the sample size > 30, applying the Central Limit Theorem we assume they are also normally distributed.

The simple linear regression results are as follows:

Between Customer Satisfaction and Product Quality:

```
Call:
lm(formula = Satisfaction ~ ProdQual)

Residuals:
     Min       1Q   Median       3Q      Max
-1.88746 -0.72711 -0.01577  0.85641  2.25220

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.67593    0.59765   6.151 1.68e-08 ***
ProdQual     0.41512    0.07534   5.510 2.90e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.047 on 98 degrees of freedom
Multiple R-squared:  0.2365,    Adjusted R-squared:  0.2287
F-statistic: 30.36 on 1 and 98 DF,  p-value: 2.901e-07
```

Between Customer Satisfaction and Ecommerce:

```
Call:
lm(formula = Satisfaction ~ Ecom)

Residuals:
     Min       1Q   Median       3Q      Max
-2.37200 -0.78971  0.04959  0.68085  2.34580

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.1516     0.6161   8.361 4.28e-13 ***
Ecom          0.4811     0.1649   2.918  0.00437 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.149 on 98 degrees of freedom
Multiple R-squared:  0.07994,   Adjusted R-squared:  0.07056
F-statistic: 8.515 on 1 and 98 DF,  p-value: 0.004368
```

Between Customer Satisfaction and Tech Support

```
Call:
lm(formula = Satisfaction ~ TechSup)

Residuals:
     Min       1Q   Median       3Q      Max
-2.26136 -0.93297  0.04302  0.82501  2.85617

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.44757    0.43592  14.791   <2e-16 ***
TechSup      0.08768    0.07817   1.122    0.265
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.19 on 98 degrees of freedom
Multiple R-squared:  0.01268,   Adjusted R-squared:  0.002603
F-statistic: 1.258 on 1 and 98 DF,  p-value: 0.2647
```

Between Customer Satisfaction and Complaint Resolution

```
Call:
lm(formula = Satisfaction ~ CompRes)

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
```

Between Customer Satisfaction and Advertising

```
Call:
lm(formula = Satisfaction ~ Advertising)

Residuals:
     Min       1Q   Median       3Q      Max
-2.34033 -0.92755  0.05577  0.79773  2.53412

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6259     0.4237  13.279  < 2e-16 ***
Advertising   0.3222     0.1018   3.167  0.00206 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.141 on 98 degrees of freedom
Multiple R-squared:  0.09282,   Adjusted R-squared:  0.08357
F-statistic: 10.03 on 1 and 98 DF,  p-value: 0.002056
```

Between Customer Satisfaction and Product Line

```
Call:
lm(formula = Satisfaction ~ ProdLine)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3634 -0.7795  0.1097  0.7604  1.7373

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.02203    0.45471   8.845 3.87e-14 ***
ProdLine     0.49887    0.07641   6.529 2.95e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1 on 98 degrees of freedom
Multiple R-squared:  0.3031,    Adjusted R-squared:  0.296
F-statistic: 42.62 on 1 and 98 DF,  p-value: 2.953e-09
```

Between Customer Satisfaction and SalesForce Image

```
Call:
lm(formula = Satisfaction ~ SalesFImage)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2164 -0.5884  0.1838  0.6922  2.0728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.06983    0.50874   8.000 2.54e-12 ***
SalesFImage  0.55596    0.09722   5.719 1.16e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.037 on 98 degrees of freedom
Multiple R-squared:  0.2502,    Adjusted R-squared:  0.2426
F-statistic:  32.7 on 1 and 98 DF,  p-value: 1.164e-07
```

Between Customer Satisfaction and Competitive Pricing

```
Call:
lm(formula = Satisfaction ~ ComPricing)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9728 -0.9915 -0.1156  0.9111  2.5845

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.03856    0.54427  14.769   <2e-16 ***
ComPricing  -0.16068    0.07621  -2.108   0.0376 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.172 on 98 degrees of freedom
Multiple R-squared:  0.04339,   Adjusted R-squared:  0.03363
F-statistic: 4.445 on 1 and 98 DF,  p-value: 0.03756
```

Between Customer Satisfaction and Warranty&Claims

```
Call:
lm(formula = Satisfaction ~ WartyClaim)

Residuals:
     Min       1Q   Median       3Q      Max
-2.36504 -0.90202  0.03019  0.90763  2.88985

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3581     0.8813   6.079 2.32e-08 ***
WartyClaim    0.2581     0.1445   1.786   0.0772 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.179 on 98 degrees of freedom
Multiple R-squared:  0.03152,   Adjusted R-squared:  0.02164
F-statistic:  3.19 on 1 and 98 DF,  p-value: 0.0772
```

Between Customer Satisfaction and Order&Billing

```
Call:
lm(formula = Satisfaction ~ OrdBilling)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4005 -0.7071 -0.0344  0.7340  2.9673

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0541     0.4840   8.377 3.96e-13 ***
OrdBilling    0.6695     0.1106   6.054 2.60e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.022 on 98 degrees of freedom
Multiple R-squared:  0.2722,    Adjusted R-squared:  0.2648
F-statistic: 36.65 on 1 and 98 DF,  p-value: 2.602e-08
```

Between Customer Satisfaction and Delivery Speed

```
Call:
lm(formula = Satisfaction ~ DelSpeed)

Residuals:
     Min       1Q   Median       3Q      Max
-2.22475 -0.54846  0.08796  0.54462  2.59432

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2791     0.5294   6.194 1.38e-08 ***
DelSpeed      0.9364     0.1339   6.994 3.30e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9783 on 98 degrees of freedom
Multiple R-squared:  0.333,     Adjusted R-squared:  0.3262
F-statistic: 48.92 on 1 and 98 DF,  p-value: 3.3e-10
```

For all of the models we found the adjusted R score to be very low, ranging between 0.0026 to 0.3574 i.e. the independent variables so considered can help to explain only around 0.26% to 35.74% of the variation in the dependent variable Customer Satisfaction.

Due to the presence of multicollinearity in the dataset we now proceed to perform PCA/FA and do dimensionality reduction, thereby reducing the number of predictors to a smaller set of uncorrelated components.

4.1 Perform PCA/FA and Interpret the Eigen Values (apply Kaiser Normalization Rule)

The given data looks to be an interval survey data i.e. the values for each variable vary between a particular range, therefore we apply Factor Analysis here.

Before beginning we first need to analyze if the given dataset is suitable for factor analysis. So we perform **Bartlett's test of sphericity** and **KMO test**.

Test1 - Bartlett's test checks the null hypothesis that our correlation matrix is an identity matrix, which would indicate that our variables are unrelated and therefore unsuitable for structure detection. Small p values (less than 0.05, considering 5% level of significance) indicate that the null hypothesis is rejected and a factor analysis may be useful with our data.

```
cortest.bartlett(FactorAnalysis_corrplot, n=100)
```

```
$chisq
[1] 619.2726

$p.value
[1] 1.79337e-96

$df
[1] 55
```

Since the p-value is less than 0.05, therefore it indicates that we can perform factor analysis with our data.

Test2 - The Kaiser-Meyer-Olkin Measure of Sampling Adequacy is a statistic that indicates the proportion of variance in your variables that might be caused by underlying factors and returns values between 0 and 1. Values greater than 0.6 generally indicate that a factor analysis may be useful with our data.

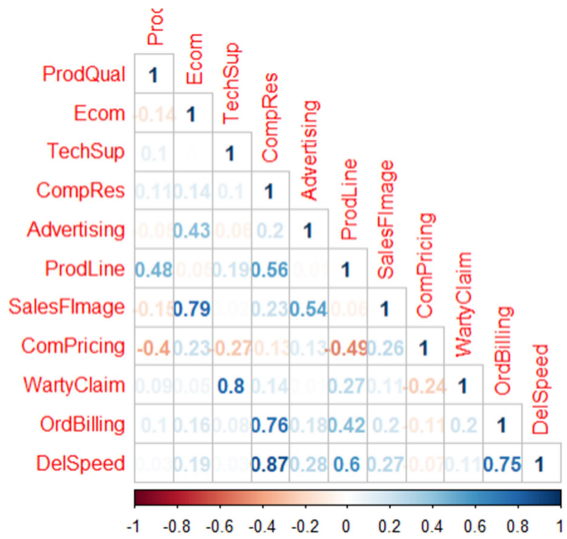
```
```{r}
KMO(FactorAnalysis_corrplot)
```

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

Since the overall Measure of Sampling Adequacy (MSA) is 0.65, the dataset is suitable for factor analysis.

Now, proceeding with factor analysis.

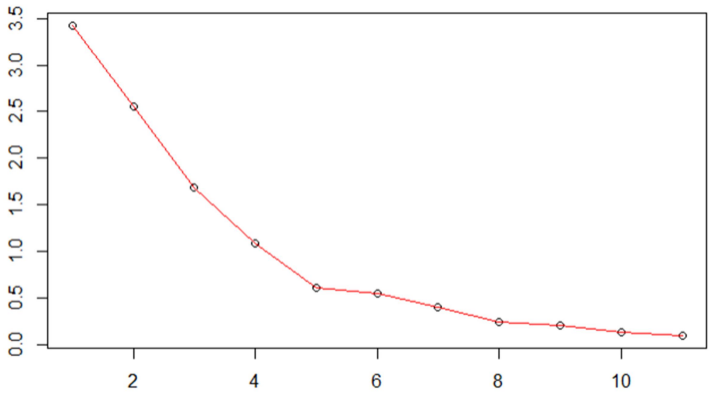
First a correlation matrix has been created taking all the 11 variables other than ID and Satisfaction and the Eigen values were computed as:




```
3.42697133 2.55089671 1.69097648 1.08655606 0.60942409 0.55188378 0.40151815 0.24695154 0.20355327 0.13284158 0.09842702
```

In every factor analysis, there are same number of factors as there are variables. Since we have 11 variables here, there will be 11 factors to begin with. Each of the 11 factors captures a certain amount of the overall variance in Customer Satisfaction which is given by the Eigen values as captured above and the factors are always listed in order of how much variation they explain.

Using the Eigen values the Scree Plot was created as shown below :



According to the Kaizer rule, number of factors to extract depend on the Eigen value –to extract all the factors whose Eigen value is greater than or equal to 1.

4.2 Output Interpretation Tell why only 4 factors are being asked in the questions and tell whether it is correct in choosing 4 factors. Name the factors with correct explanations.

Any factor with an eigenvalue ≥1 explains more variance than a single observed variable. Therefore using the Kaizer Rule, we decide to reduce the dataset to 4 factors.

Next we did an unrotated factor loading:

```
```{r}
Unrotate_FactorHairRevised=fa(FactorHairRevised_DF[,2:12],nfactors = 4, rotate = "none", fm="pa")
Unrotate_FactorHairRevised
```
```

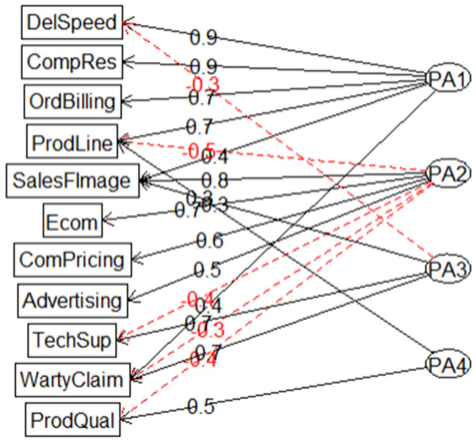
```
Factor Analysis using method =  pa
Call: fa(r = FactorHairRevised_DF[, 2:12], nfactors = 4, rotate = "none",
    fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix

                       PA1  PA2  PA3  PA4
SS loadings           3.21 2.22 1.50 0.68
Proportion Var        0.29 0.20 0.14 0.06
Cumulative Var        0.29 0.49 0.63 0.69
Proportion Explained  0.42 0.29 0.20 0.09
Cumulative Proportion 0.42 0.71 0.91 1.00
```

	PA1 <S3: AsIs>	PA2 <S3: AsIs>	PA3 <S3: AsIs>	PA4 <S3: AsIs>	h2 <dbl>
ProdQual	0.20	-0.41	-0.06	0.46	0.4242958
Ecom	0.29	0.66	0.27	0.22	0.6381735
TechSup	0.28	-0.38	0.74	-0.17	0.7946147
CompRes	0.86	0.01	-0.26	-0.18	0.8428100
Advertising	0.29	0.46	0.08	0.13	0.3142090
ProdLine	0.69	-0.45	-0.14	0.31	0.8002906
SalesFImage	0.39	0.80	0.35	0.25	0.9792432
ComPricing	-0.23	0.55	-0.04	-0.29	0.4432708
WartyClaim	0.38	-0.32	0.74	-0.15	0.8135338
OrdBilling	0.75	0.02	-0.18	-0.18	0.6218211
DelSpeed	0.90	0.10	-0.30	-0.20	0.9420396

Factor loading shows the relationship of each variable to the underlying factor. For instance, the variable ProdQual has a correlation of 0.20 with the factor PA1, negatively correlated with factors PA2 (-0.41) and PA3 (-0.06) and 0.46 with factor PA4.

To see, how the variables relate to the underlying 4 factors, the following graph is plotted:



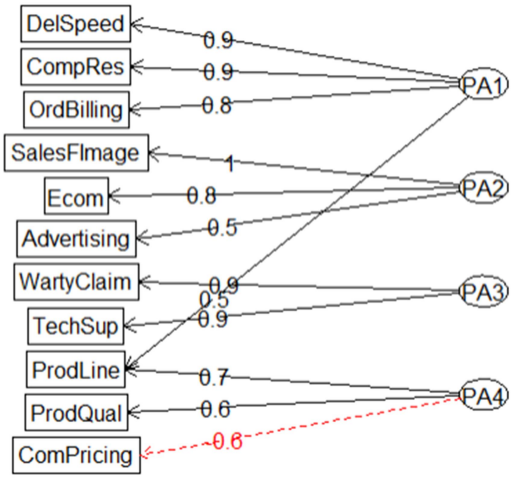
However, to clarify the structure of the loadings matrix and clearly demonstrate which variables are aligned to which factor, we do a Varimax rotation as given:

```{r}
library(psych)
Rotate_FactorHairRevised= fa(FactorHairRevised_DF[,2:12],nfactors = 4, rotate = "varimax", fm="pa")
Rotate_FactorHairRevised
```

```
Factor Analysis using method =  pa
Call: fa(r = FactorHairRevised_DF[, 2:12], nfactors = 4, rotate = "varimax",
    fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix

                       PA1  PA2  PA3  PA4
SS loadings           2.63 1.97 1.64 1.37
Proportion Var        0.24 0.18 0.15 0.12
Cumulative Var        0.24 0.42 0.57 0.69
Proportion Explained  0.35 0.26 0.22 0.18
Cumulative Proportion 0.35 0.60 0.82 1.00
```

	PA1 <S3: AsIs>	PA2 <S3: AsIs>	PA3 <S3: AsIs>	PA4 <S3: AsIs>	h2 <dbl>
ProdQual	0.02	-0.07	0.02	0.65	0.4242958
Ecom	0.07	0.79	0.03	-0.11	0.6381735
TechSup	0.02	-0.03	0.88	0.12	0.7946147
CompRes	0.90	0.13	0.05	0.13	0.8428100
Advertising	0.17	0.53	-0.04	-0.06	0.3142090
ProdLine	0.53	-0.04	0.13	0.71	0.8002906
SalesFImage	0.12	0.97	0.06	-0.13	0.9792432
ComPricing	-0.08	0.21	-0.21	-0.59	0.4432708
WartyClaim	0.10	0.06	0.89	0.13	0.8135338
OrdBilling	0.77	0.13	0.09	0.09	0.6218211
DelSpeed	0.95	0.19	0.00	0.09	0.9420396



Now, examining the variables that are aligned to each factor, we can rename each factor based on some underlying feature that each represent :

DelSpeed, CompRes, OrdBilling are aligned to PA1, we rename PA1 as Order Processing;

SalesFImage, Ecom, Advertising are aligned to PA2, we rename PA2 as Marketing;

WartyClaim, TechSup are aligned to PA3, we rename PA3 as Post Sales Service;

ProdLine, ProdQual, ComPricing are aligned to PA4, we rename PA4 as Product Management

5.1 Create a data frame with a minimum of 5 columns, 4 of which are different factors and the 5th column is Customer Satisfaction

Before creating the dataframe, first the factor scores were produced corresponding to the rotated factor loadings of the 4 factors. Factor scores are composites of the variables that are used to make the latent factor into an observed variable i.e. to give it a scale. Factor scores are derived when the factors are to be used as a predictor or outcome in a regression analysis.

```
FactorHairRevised_Scores=Rotate_FactorHairRevised$scores

head(FactorHairRevised_Scores,5)
```

```
            PA1        PA2         PA3         PA4
[1,] -0.1338871  0.9175166 -1.719604873  0.09135411
[2,]  1.6297604 -2.0090053 -0.596361722  0.65808192
[3,]  0.3637658  0.8361736  0.002979966  1.37548765
[4,] -1.2225230 -0.5491336  1.245473305 -0.64421384
[5,] -0.4854209 -0.4276223 -0.026980304  0.47360747
```

```
cbind(FactorHairRevised_DF$Satisfaction,FactorHairRevised_Scores)
FactorHairRevised_FinalDF=data.frame("Cust_Satisfaction"=FactorHairRevised_DF[,13], "Order_Processing"= FactorHairRevised_Scores[,1],
          "Marketing"=FactorHairRevised_Scores[,2], "PostSales_Service"=FactorHairRevised_Scores[,3],
          "Product_Management"=FactorHairRevised_Scores[,4])
```

Provided below is a summary of the newly created dataframe :

```
  Satisfaction   Order_Processing     Marketing       PostSales_Service  Product_Management
 Min.   :4.700   Min.   :-2.55956   Min.   :-2.0373   Min.   :-2.20200   Min.   :-1.42620
 1st Qu.:6.000   1st Qu.:-0.61566   1st Qu.:-0.4663   1st Qu.:-0.73427   1st Qu.:-0.83402
 Median :7.050   Median : 0.07914   Median :-0.2038   Median : 0.09067   Median : 0.03373
 Mean   :6.918   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
 3rd Qu.:7.625   3rd Qu.: 0.74181   3rd Qu.: 0.5719   3rd Qu.: 0.56502   3rd Qu.: 0.70675
 Max.   :9.900   Max.   : 1.99193   Max.   : 2.8326   Max.   : 2.08285   Max.   : 2.15737
```

5.2 Perform Multiple Linear Regression with Customer Satisfaction as the Dependent Variable and the four factors as Independent Variables

The multiple linear regression result is as follows:

```
call:
lm(formula = Cust_Satisfaction ~ Order_Processing + Marketing +
    PostSales_Service + Product_Management, data = FactorHairRevised_FinalDF)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7125 -0.4708  0.1024  0.4158  1.3483

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.91800    0.06696 103.317  < 2e-16 ***
Order_Processing   0.57963    0.06857   8.453 3.32e-13 ***
Marketing          0.61978    0.06834   9.070 1.61e-14 ***
PostSales_Service  0.05692    0.07173   0.794    0.429
Product_Management 0.61168    0.07656   7.990 3.16e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6696 on 95 degrees of freedom
Multiple R-squared:  0.6971,    Adjusted R-squared:  0.6844
F-statistic: 54.66 on 4 and 95 DF,  p-value: < 2.2e-16
```

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

In multiple linear regresssion, R2 represents the proportion of variance, in the dependent variable that may be predicted by knowing the value of the independent variables. An R2 value close to 1 indicates that the model explains a large portion of the variance in the dependent variable. A problem with the R2, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response. A solution is to adjust the R2 by taking into account the number of independent variables. The adjustment in the “Adjusted R Square” value in the summary output is a correction for the number of independent variables included in the prediction model.

In our model, the adjusted R2 is 0.6844 meaning that 68.44% of the variation in Customer Satisfaction can be predicted by Order_Processing, Marketing, PostSales_Service and Product_Management.of freedom are calculated

Degrees of freedom for RSS= No. of predictor variables = 4;

Degrees of freedom for ESS=Sample Size – No. of predictor variables -1 = 100-4-1 = 95

The F-test of overall significance indicates whether our linear regression model provides a better fit to the data than a model that contains no independent variables.

If the p-value corresponding to the F-statistic is less than the significance level, it means that our sample data provides sufficient evidence to conclude that the independent variables included in our regression model improve the fit.

Since in our model p-value of the F-statistic is < 2.2e-16, it is highly significant. This means that, at least, one of the predictor variables is significantly related to the outcome variable.

Now, to see which predictor variables are significant, we examine the coefficients table, which shows the estimate of regression beta coefficients and the associated t-statistic p-values:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.91800000	0.06695910	103.316794	2.121636e-99
Order_Processing	0.57962798	0.06857416	8.452571	3.320539e-13
Marketing	0.61978029	0.06833625	9.069569	1.609989e-14
PostSales_Service	0.05692291	0.07172935	0.793579	4.294183e-01
Product_Management	0.61167972	0.07655687	7.989873	3.162057e-12

For a given predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the beta coefficient of the predictor is significantly different from zero.

It can be seen that, changes in Order_Processing, Marketing and Product_Management are significantly associated to changes in Customer Satisfaction.

The confidence interval of the model coefficients are as follows:

	2.5 %	97.5 %
(Intercept)	6.78506937	7.0509306
Order_Processing	0.44349106	0.7157649
Marketing	0.48411569	0.7554449
PostSales_Service	-0.08547787	0.1993237
Product_Management	0.45969511	0.7636643

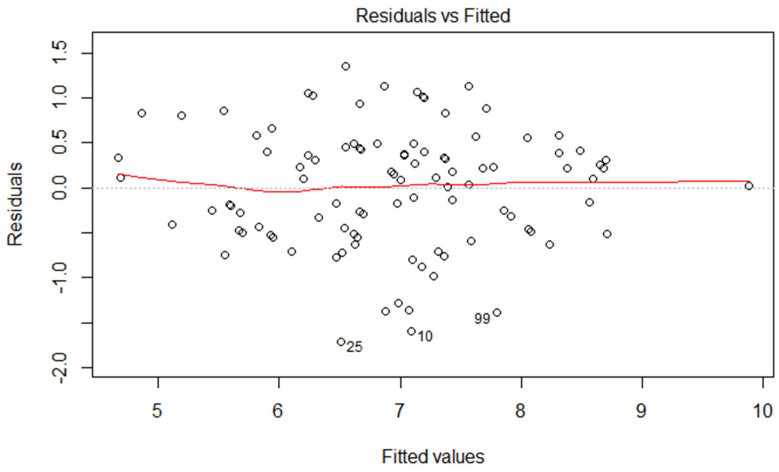
Checking VIF values:

```
```{r}
library(car)
vif(FactorHairRevised_Model)
```
```

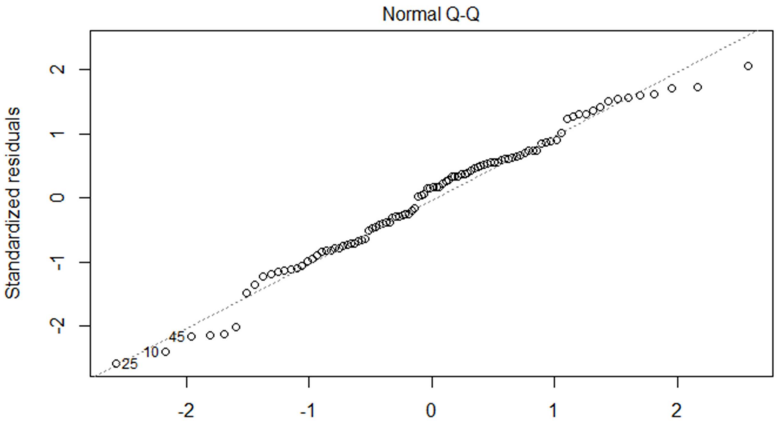
```
   Order_Processing           Marketing   PostSales_Service  Product_Management
           1.001021            1.002683            1.002981            1.005848
```

Since all the dependent variables have VIF around 1, this shows there is no further multicollinearity in the data.

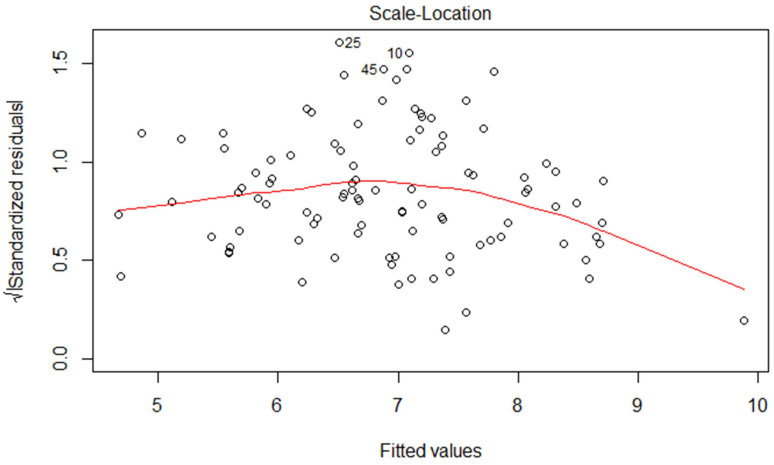
Following is an analysis of residual plots which are a useful to assess veracity of a linear regression model on a particular dataset:



Residual plots are used to look for underlying patterns in the residuals that may mean that the model has a problem. From the Residuals vs Fitted plot, we see that there is no obvious pattern and they're pretty symmetrically distributed, tending to cluster towards the middle of the plot.

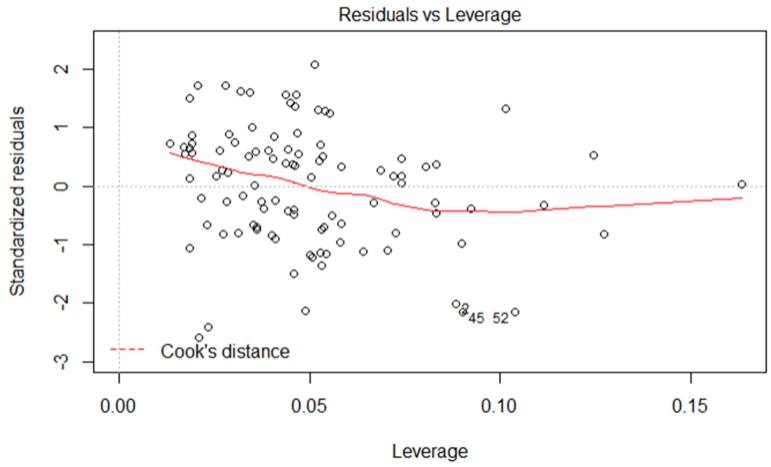


The Q-Q Plot shows if residuals are normally distributed or not. Here since the residuals follow close to a straight line on this plot, it is a good indication they are normally distributed.



Scale Location plot - This plot test the linear regression assumption of equal variance (homoscedasticity) i.e. that the residuals have equal variance along the regression line. It is also

called the Spread-Location plot. The residuals here have equal variance(occupy equal space) above and below the line and along the length of the line.



Residuals vs Leverage plot can be used to find influential cases in the dataset. An influential case is one that, if removed, will affect the model so its inclusion or exclusion should be considered. For instance, outliers will tend to exert leverage and therefore influence on the model. There are no big influencers in this model.

5.4 Output Interpretation <making it meaningful for everybody>

Model accuracy is also judged by looking at the Residual Standard Error (RSE) which is a measure of error of prediction. The lower the RSE, the more accurate the model is.

```{r}
sigma(FactorHairRevised_Model)/mean(FactorHairRevised_FinalDF$Cust_Satisfaction)
```

```
[1] 0.09678969
```

Here we can say that the percentage error is (any prediction would still be off by) 9.6%.

In terms of interpreting our overall model, we believe that is a reasonably fit model (judging by the adjusted R-square value) and that the factors all have moderate influence over Customer Satisfaction. Still, we found Order_Processing, Marketing and Product_Management to have significant influence over Customer Satisfaction as against PostSales_Service.