# Advanced Statistics project based on Factor Hair Revised file

Code

Setting the working directory

Hide

```
getwd()
```

```
[1] "D:/chandrima/BACP - GreatLearning/Advanced Stats - Project"
```

Importing the dataset within R environment

Hide

```
FactorHairRevised_DF=read.csv("Factor-Hair-Revised.csv", header = TRUE)
```

Checking the dimension of the dataset

Hide

```
dim(FactorHairRevised_DF)
```

```
[1] 100  13
```

The dataset so imported has 100 rows and 13 columns

Checking the first and last few rows

| | ID <int> | ProdQual <dbl> | E… <dbl> | TechS… <dbl> | Comp… <dbl> | Advertising <dbl> | ProdLine <dbl> | SalesFImage <dbl> | ComPricing <dbl> |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 8.5 | 3.9 | 2.5 | 5.9 | 4.8 | 4.9 | 6.0 | 6.8 |
| 2 | 2 | 8.2 | 2.7 | 5.1 | 7.2 | 3.4 | 7.9 | 3.1 | 5.3 |
| 3 | 3 | 9.2 | 3.4 | 5.6 | 5.6 | 5.4 | 7.4 | 5.8 | 4.5 |
| 4 | 4 | 6.4 | 3.3 | 7.0 | 3.7 | 4.7 | 4.7 | 4.5 | 8.8 |
| 5 | 5 | 9.0 | 3.4 | 5.2 | 4.6 | 2.2 | 6.0 | 4.5 | 6.8 |

5 rows | 1-10 of 13 columns

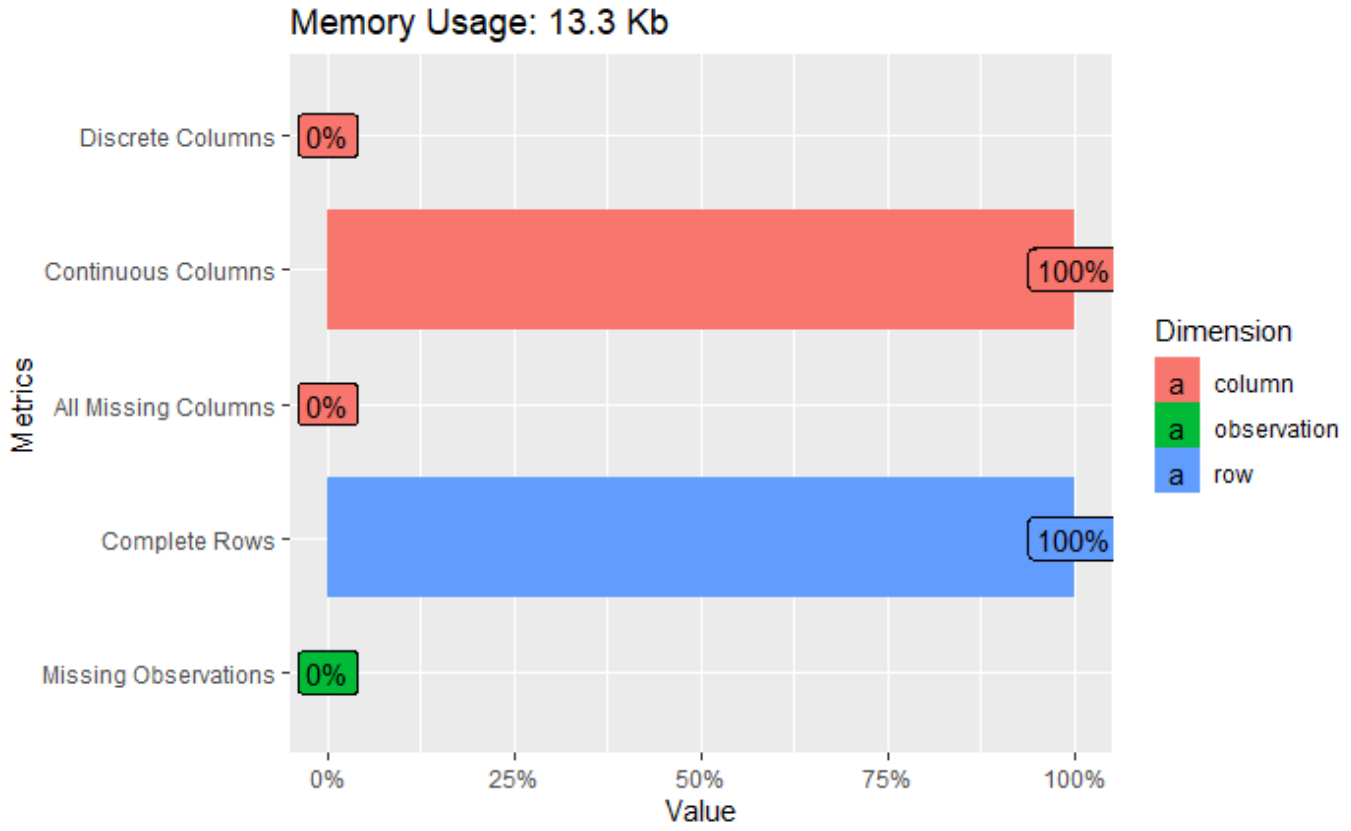
| | ID <int> | ProdQ… <dbl> | E… <dbl> | Tech… <dbl> | Com… <dbl> | Advertising <dbl> | ProdLine <dbl> | SalesFImage <dbl> | ComPricing <dbl> |
|---|---|---|---|---|---|---|---|---|---|
| 96 | 96 | 8.6 | 4.8 | 5.6 | 5.3 | 2.3 | 6.0 | 5.7 | 6.7 |
| 97 | 97 | 7.4 | 3.4 | 2.6 | 5.0 | 4.1 | 4.4 | 4.8 | 7.2 |
| 98 | 98 | 8.7 | 3.2 | 3.3 | 3.2 | 3.1 | 6.1 | 2.9 | 5.6 |
| 99 | 99 | 7.8 | 4.9 | 5.8 | 5.3 | 5.2 | 5.3 | 7.1 | 7.9 |
| 100 | 100 | 7.9 | 3.0 | 4.4 | 5.1 | 5.9 | 4.2 | 4.8 | 9.7 |

5 rows | 1-10 of 13 columns

Understanding the data structure

| r… <int> | colu… <int> | discrete_columns <int> | continuous_columns <int> | all_missing_columns <int> | total_miss |
|---|---|---|---|---|---|
| 100 | 13 | 0 | 13 | 0 | |

1 row | 1-6 of 9 columns




As is clear, there are no missing records in the data and all the variables are of continuous data type

Checking the data types

Hide

```
str(FactorHairRevised_DF)
```

```
'data.frame':   100 obs. of  13 variables:
 $ ID          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ ProdQual    : num  8.5 8.2 9.2 6.4 9 6.5 6.9 6.2 5.8 6.4 ...
 $ Ecom        : num  3.9 2.7 3.4 3.3 3.4 2.8 3.7 3.3 3.6 4.5 ...
 $ TechSup     : num  2.5 5.1 5.6 7 5.2 3.1 5 3.9 5.1 5.1 ...
 $ CompRes     : num  5.9 7.2 5.6 3.7 4.6 4.1 2.6 4.8 6.7 6.1 ...
 $ Advertising : num  4.8 3.4 5.4 4.7 2.2 4 2.1 4.6 3.7 4.7 ...
 $ ProdLine    : num  4.9 7.9 7.4 4.7 6 4.3 2.3 3.6 5.9 5.7 ...
 $ SalesFImage : num  6 3.1 5.8 4.5 4.5 3.7 5.4 5.1 5.8 5.7 ...
 $ ComPricing  : num  6.8 5.3 4.5 8.8 6.8 8.5 8.9 6.9 9.3 8.4 ...
 $ WartyClaim  : num  4.7 5.5 6.2 7 6.1 5.1 4.8 5.4 5.9 5.4 ...
 $ OrdBilling  : num  5 3.9 5.4 4.3 4.5 3.6 2.1 4.3 4.4 4.1 ...
 $ DelSpeed    : num  3.7 4.9 4.5 3 3.5 3.3 2 3.7 4.6 4.4 ...
 $ Satisfaction: num  8.2 5.7 8.9 4.8 7.1 4.7 5.7 6.3 7 5.5 ...
```

Hide

```
summary(FactorHairRevised_DF)
```

```
       ID            ProdQual           Ecom          TechSup         CompRes       Advertisi
ng       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
```

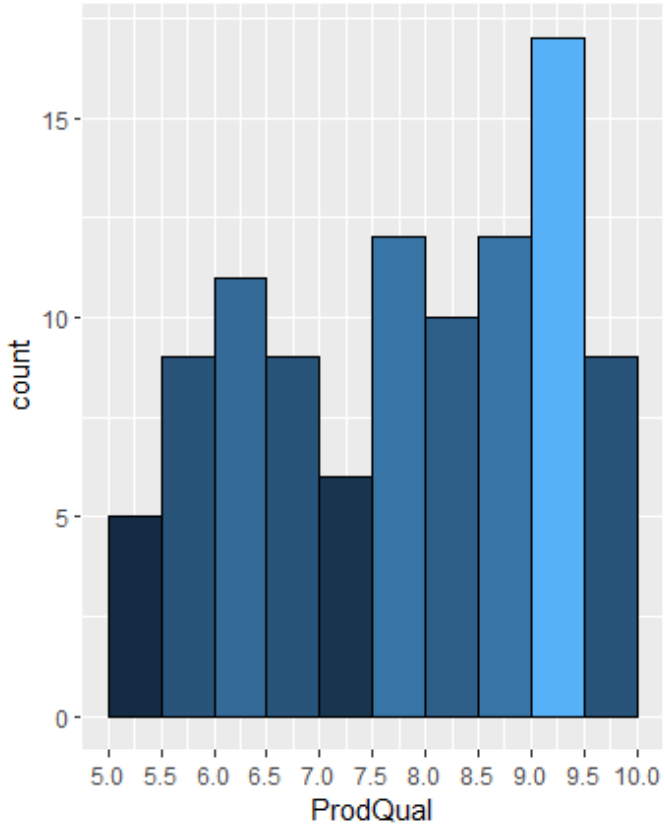
Exploratory data analysis - Since all the variables are continuous in nature we would stick to histogram and boxplots

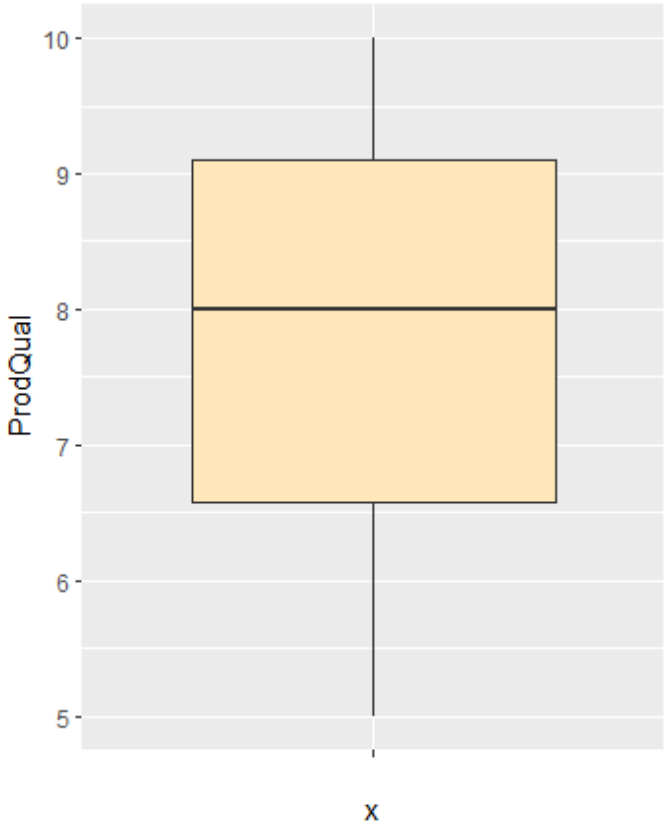
Hide

```
library(ggpubr)
```

```
Loading required package: magrittr
```

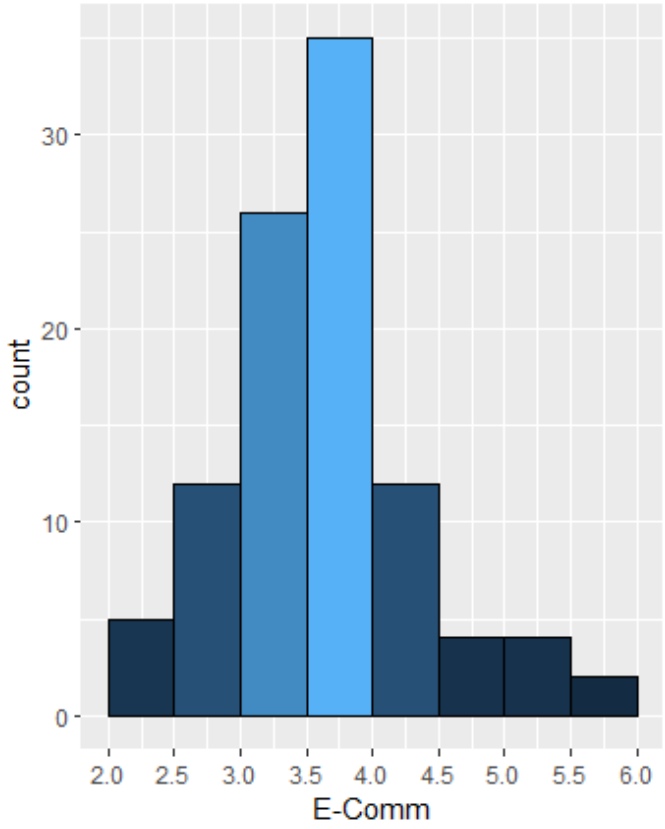
Analyzing Product Quality

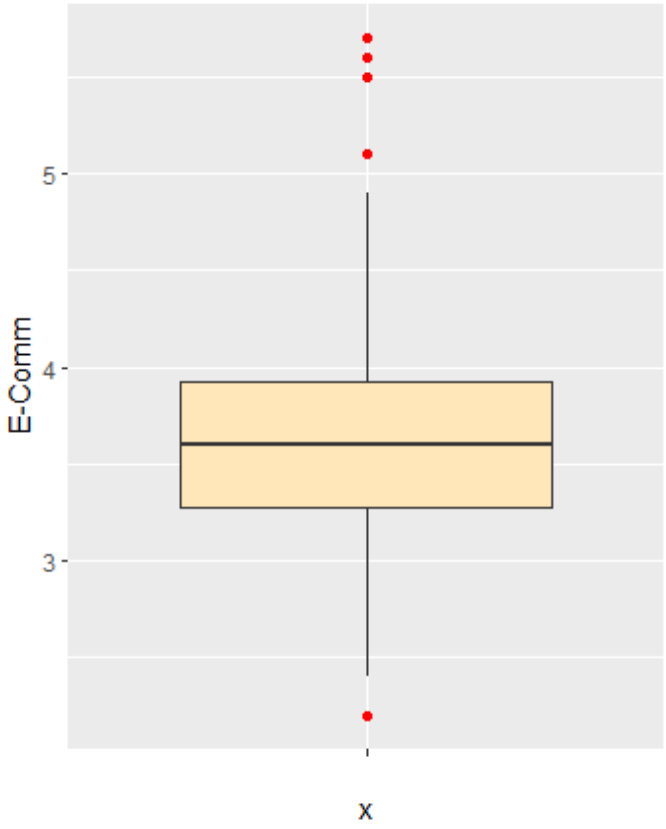



Not a normal distribution and no outliers present for Product Quality; a large number of data points are present between 9 and 9.5.

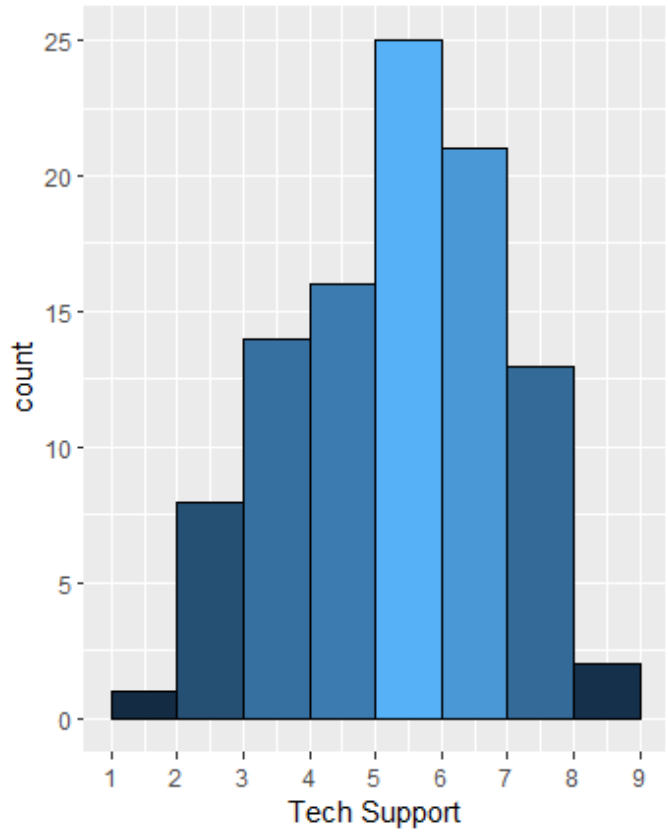
Analyzing E-Commerce

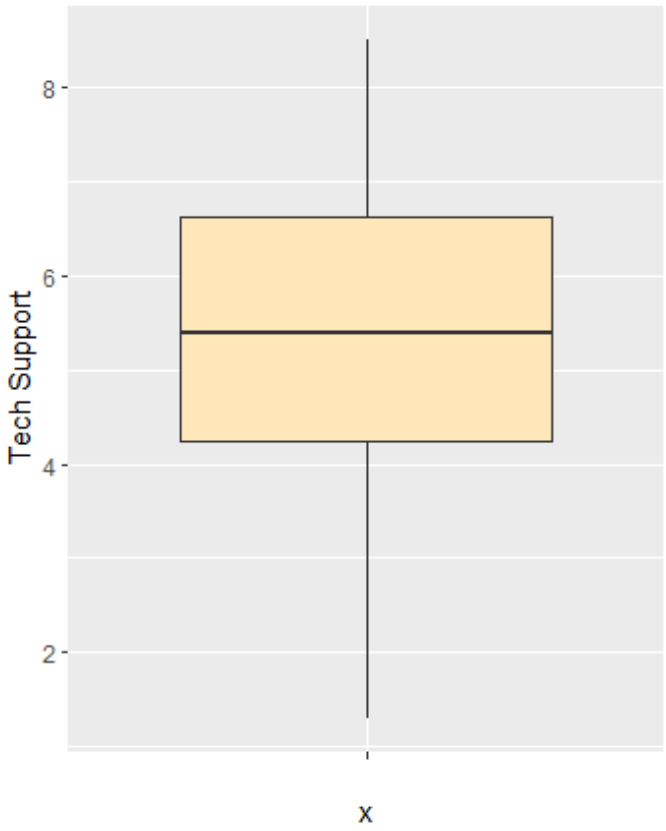



Somewhat normally distributed with a significant number of data points lying between 3.5 and 4. E-Comm has a few outliers present in its data.

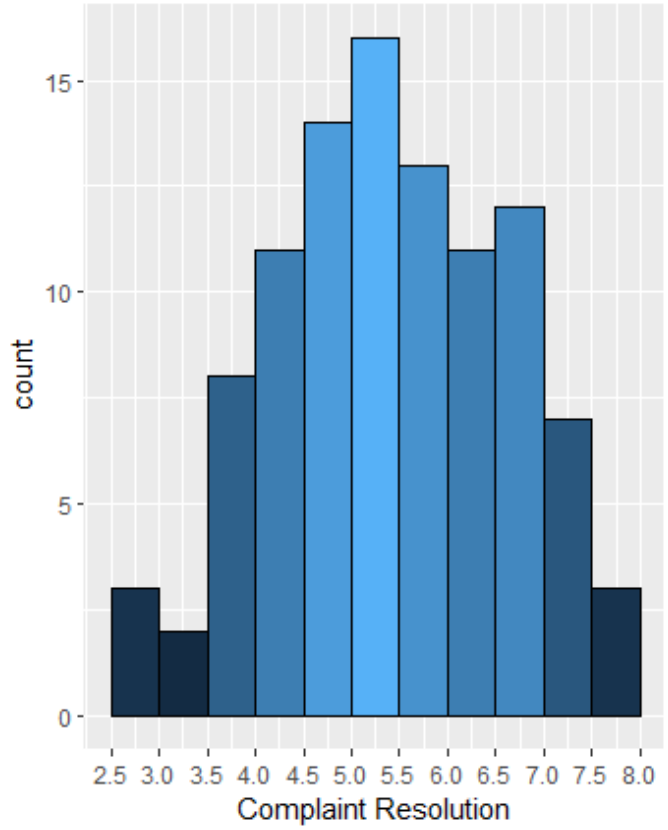
Analyzing Technical Support

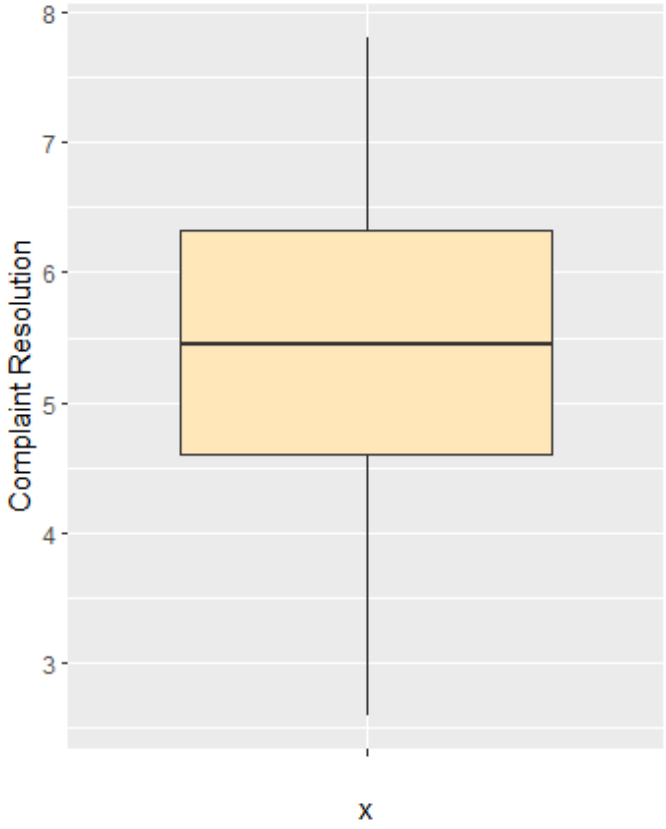



Somewhat normally distributed with ~45% of data points lying between 5 and 7. No outliers present in Tech Support.

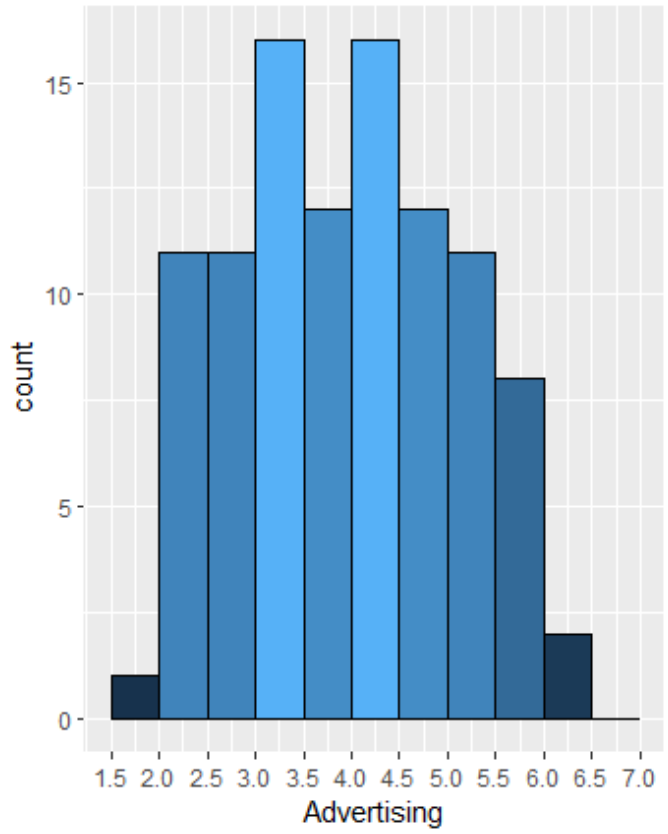
Analyzing Complaint Resolution

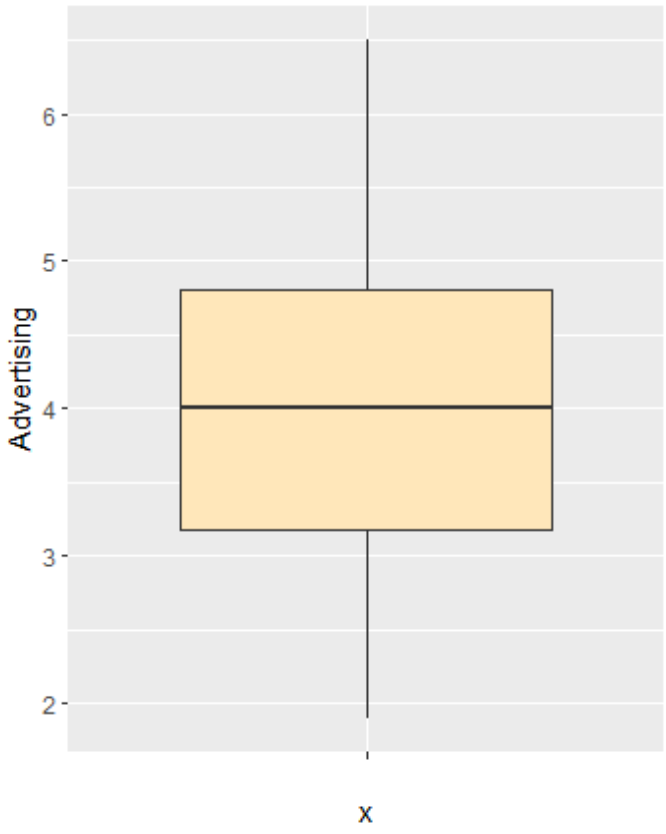



A weak normal distribution and no outliers are present in Complaint Resolution.

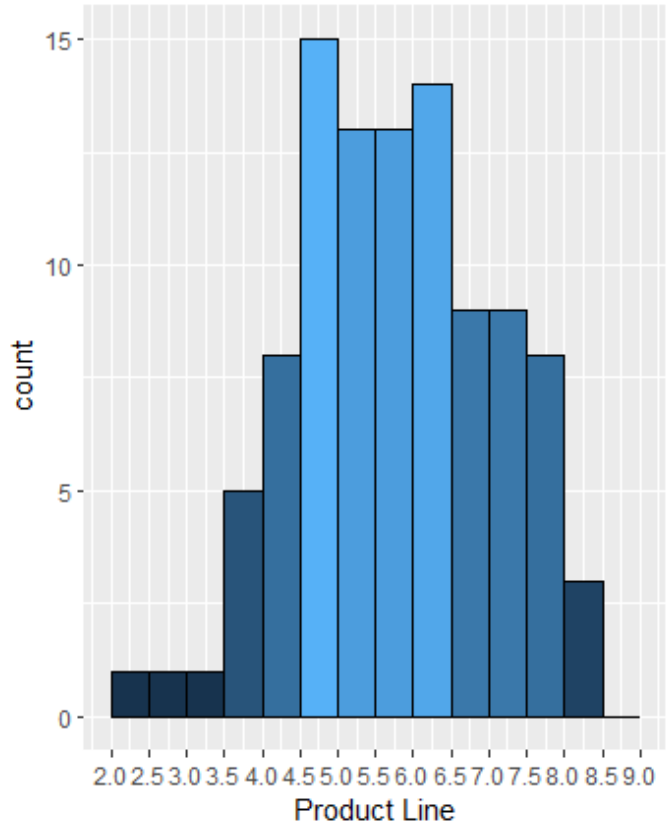
Analyzing Advertising

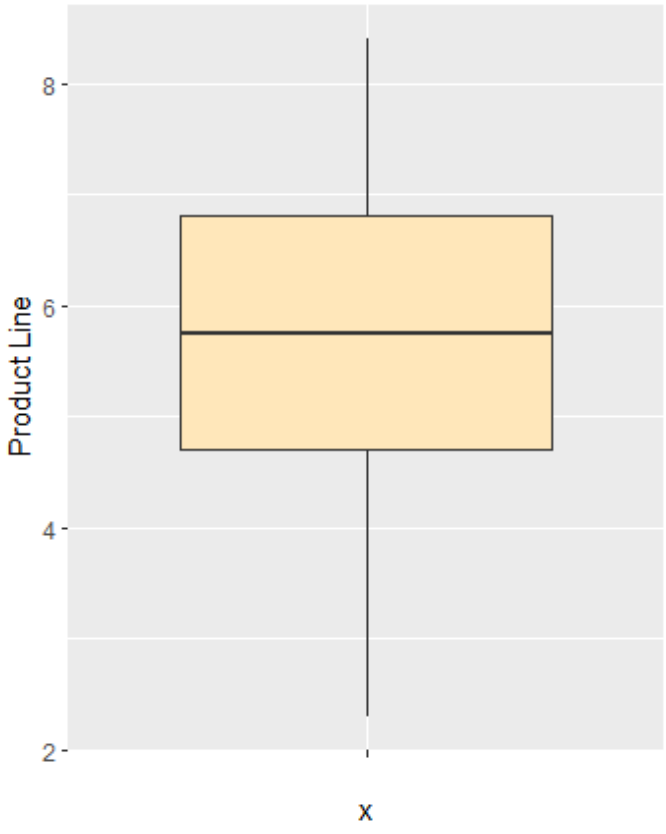



Not a normal distribution and no outliers are present in Advertising.

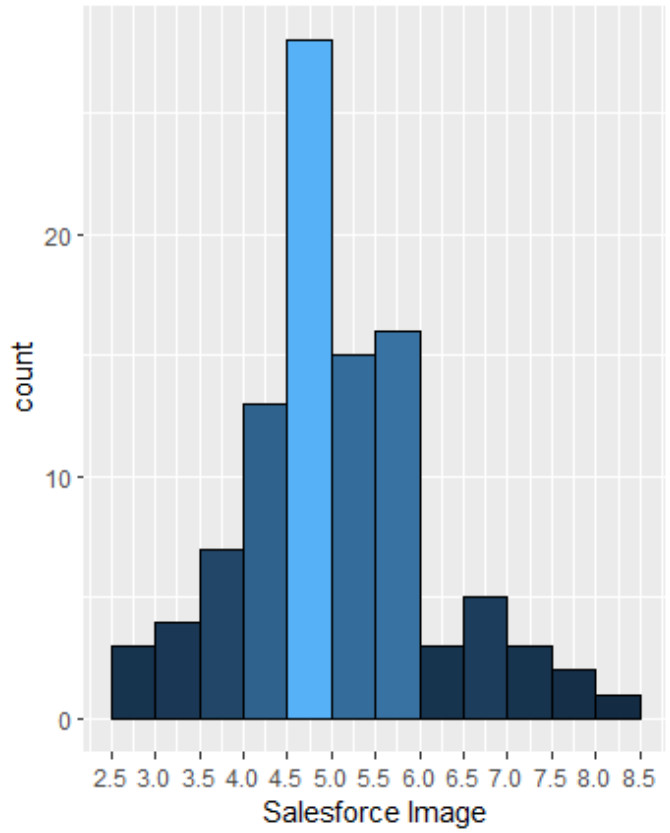
Analyzing Product Line

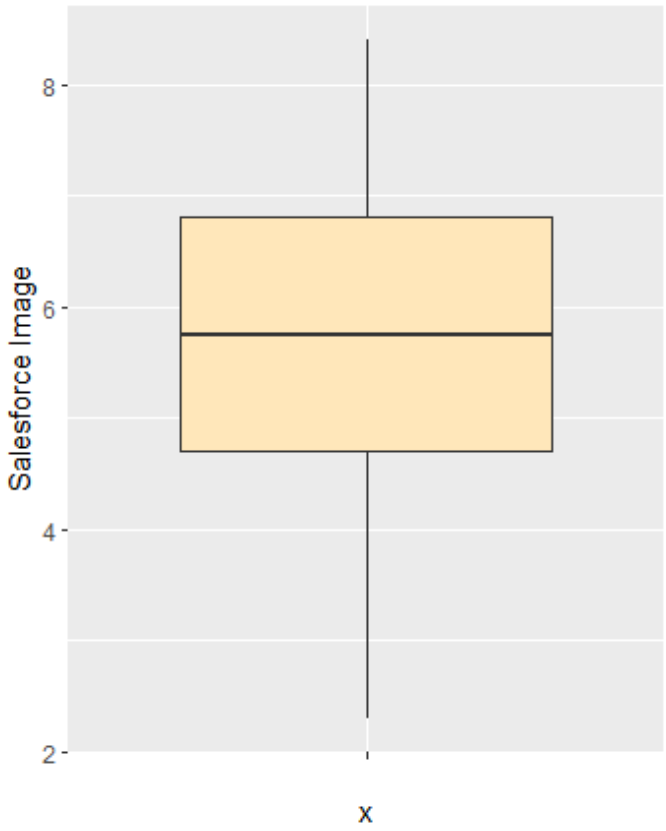



A weak normal distribution and no outliers are present in Product Line.

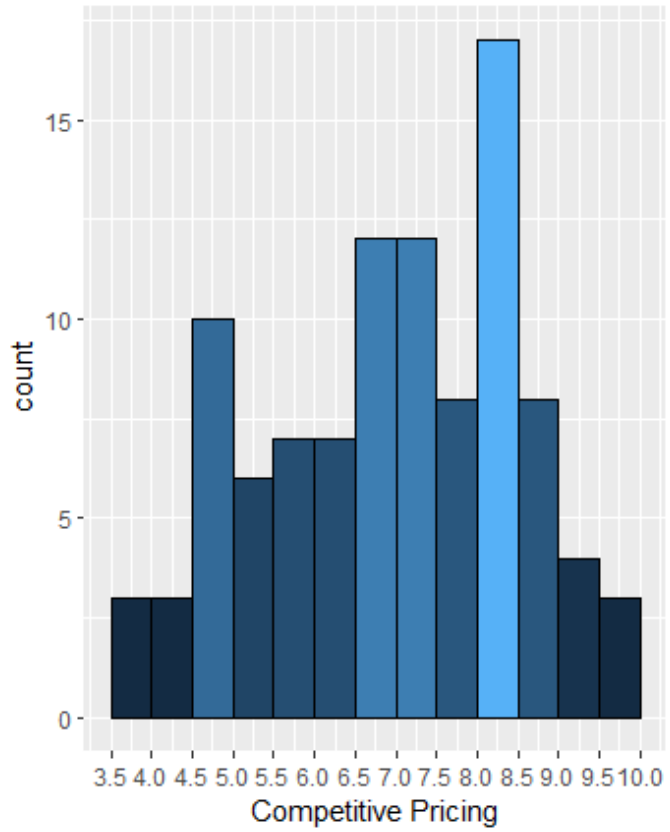
Analyzing Salesforce Image

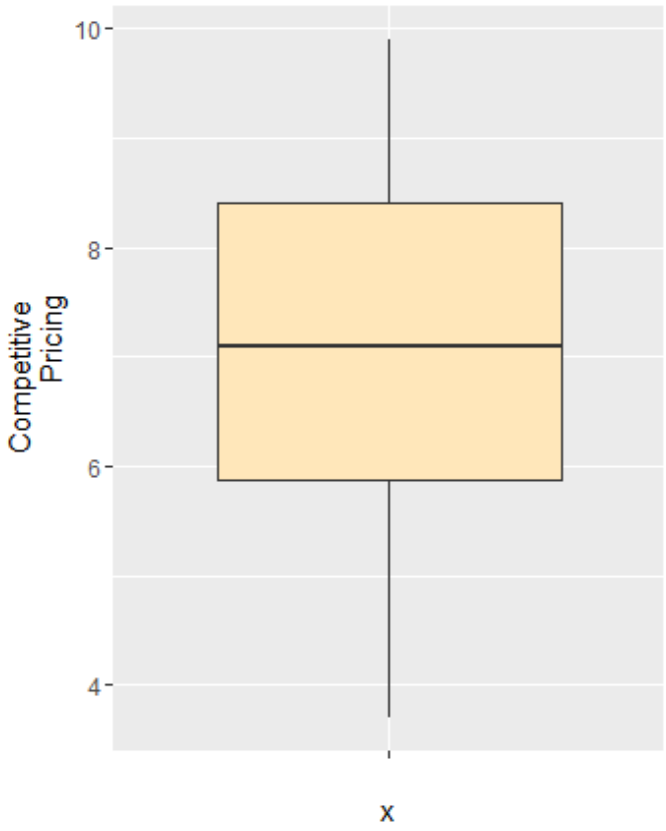



A weak normal distribution and no outliers are present in Salesforce Image.Close to 30% data points lie between 4.5-5.

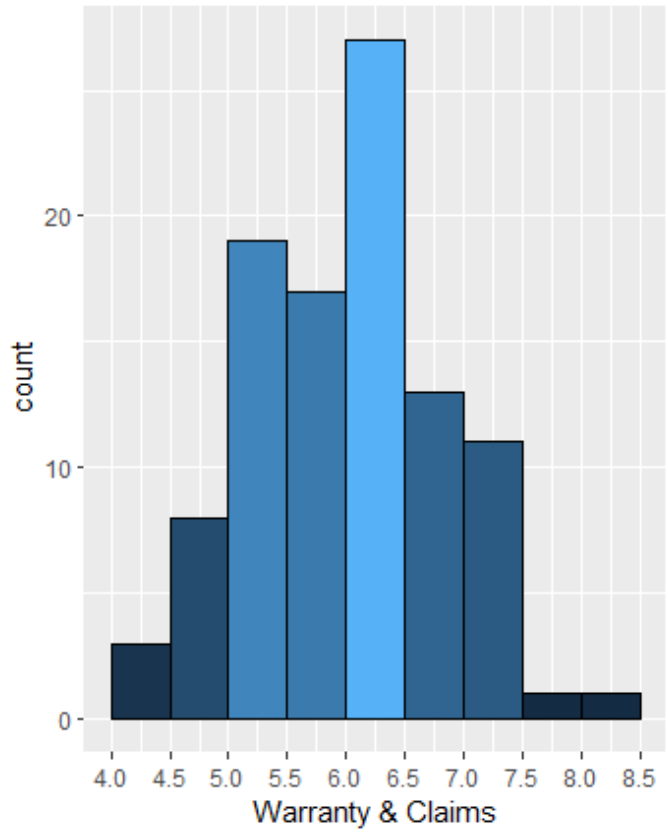
Analyzing Competitive Pricing

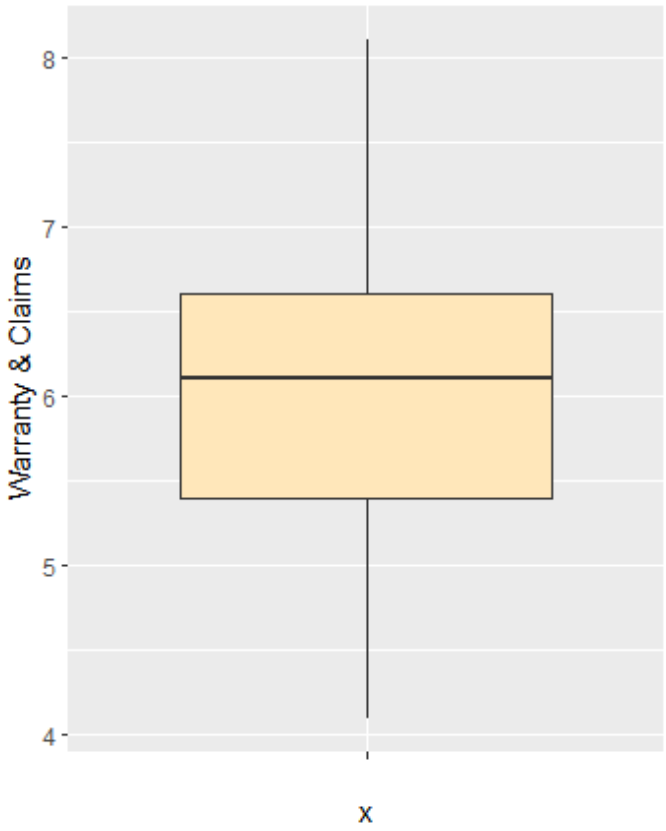



Not a normal distribution and no outliers are present in Competitive Pricing.

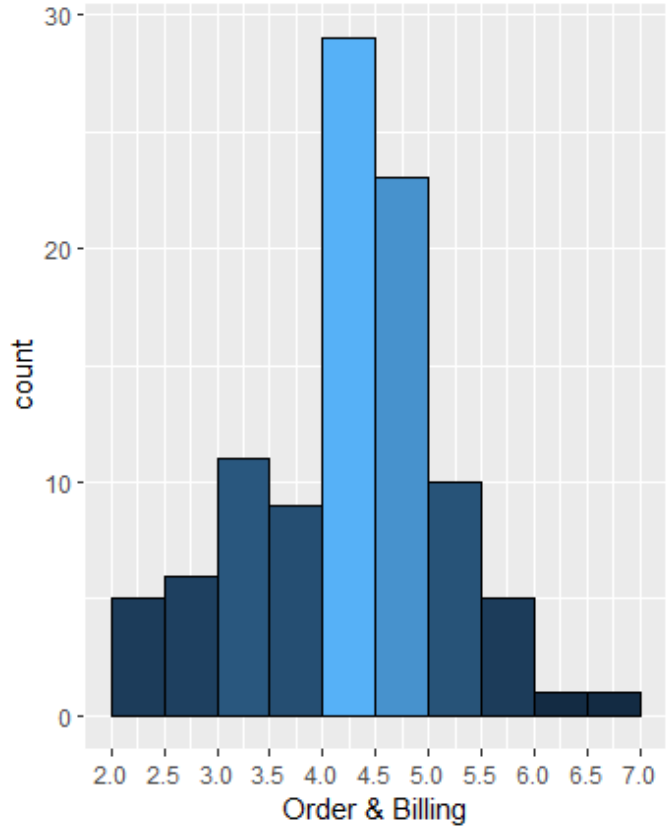
Analyzing Warranty & Claims

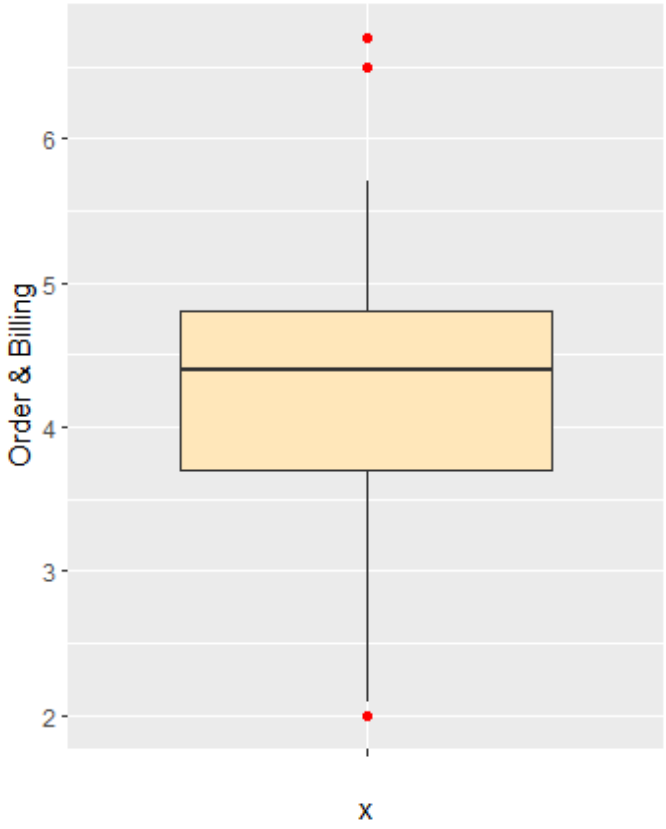



A weak normal distribution and no outliers are present in Warranty & Claims. Over 25% of data points lie between 6 and 6.5.

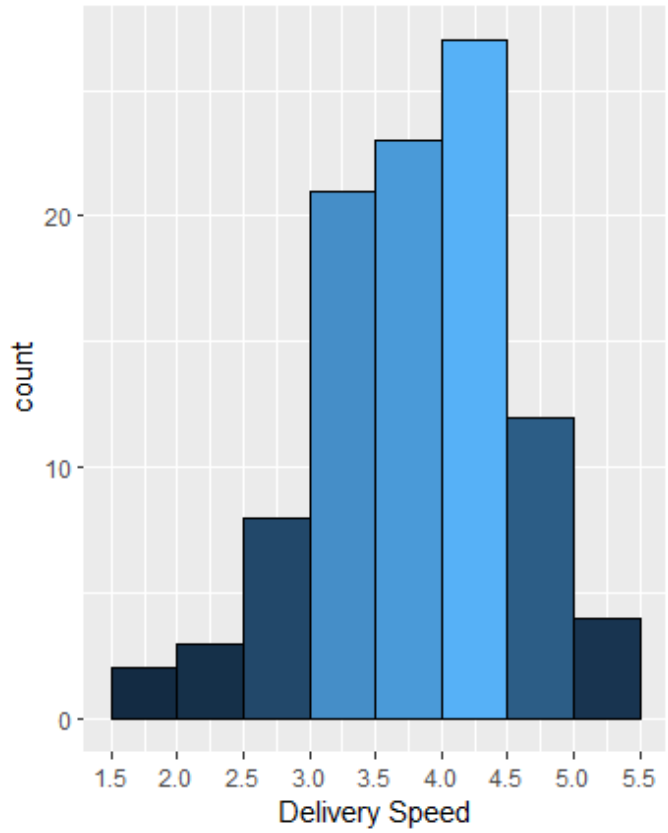
Analyzing Order & Billing

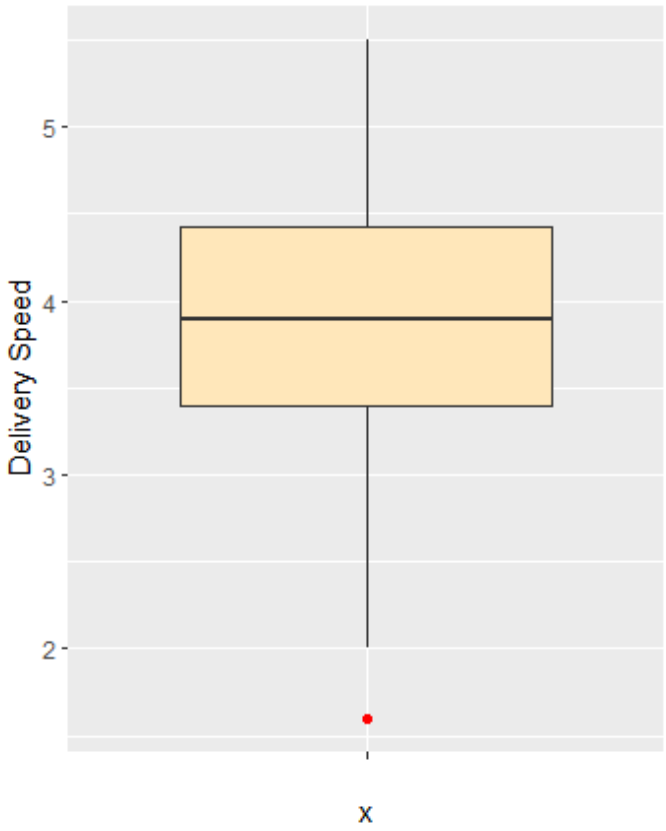



A weak normal distribution with outliers being present in Order & Billing. ~50% of data points lie between 4 and 5.

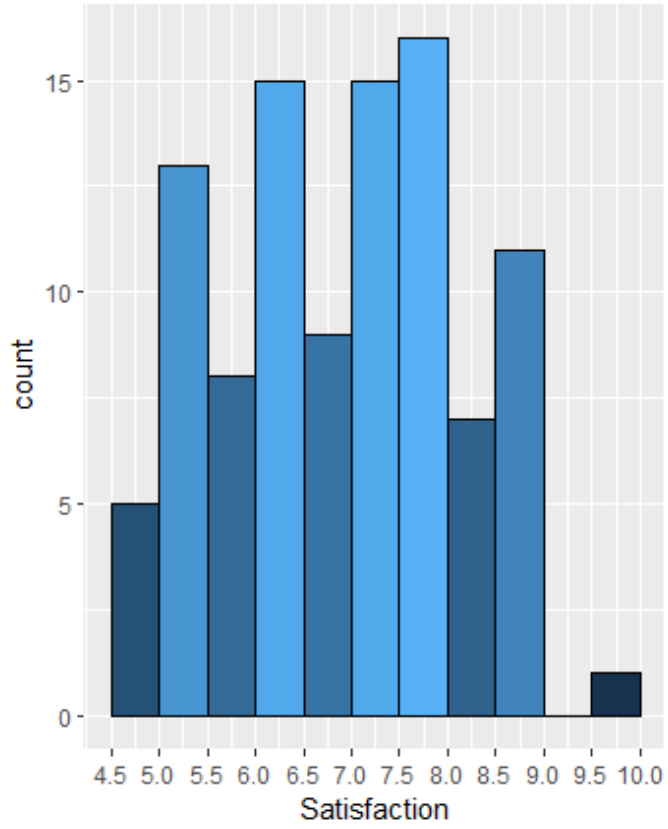
Analyzing Delivery Speed

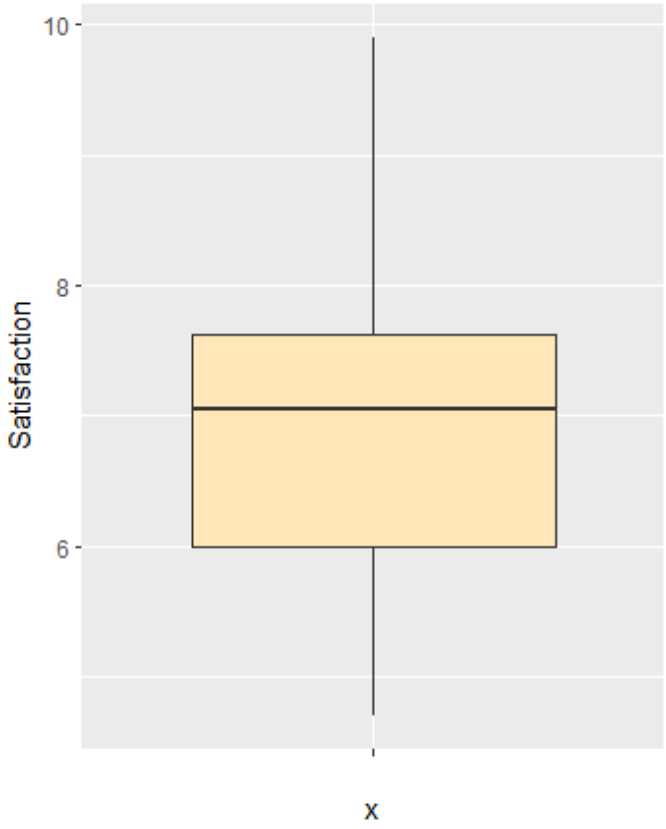



A weak normal distribution with an outlier being present in Delivery Speed. More than 65% of data points lie between 3 and 4.5.

Analyzing Customer Satisfaction



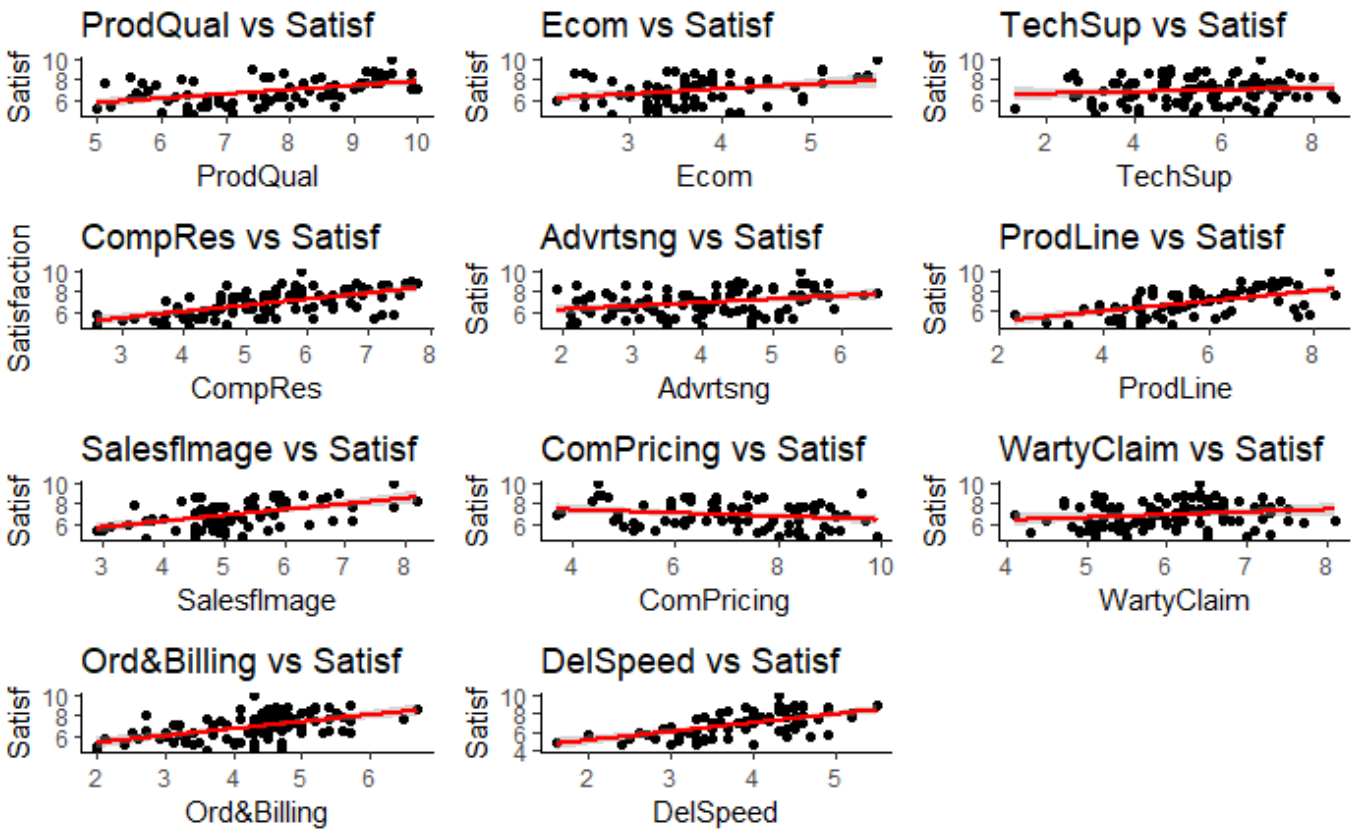

Not a normal distribution and no outliers are present in Customer Satisfaction.

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

Hide

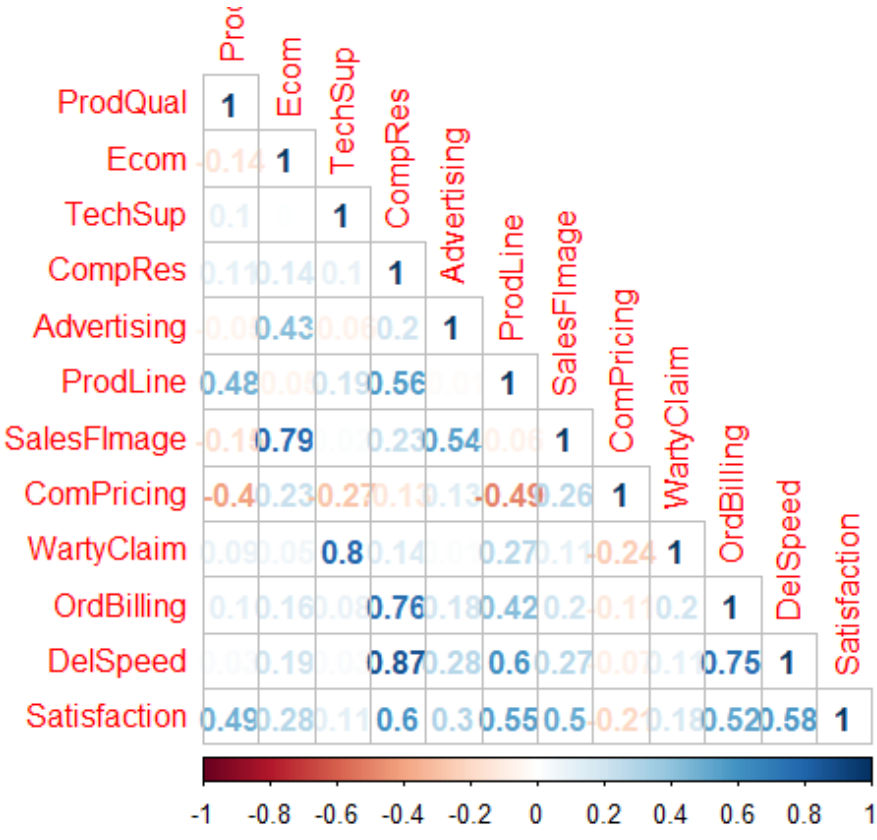
```
ggarrange(

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$ProdQual,y=FactorHairRevised_DF$Sat
isfaction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "ProdQual", y = "Satisf",tit
le = "ProdQual vs Satisf"),

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$Ecom,y=FactorHairRevised_DF$Satisfa
ction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "Ecom", y = "Satisf",title =
"Ecom vs Satisf"),

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$TechSup,y=FactorHairRevised_DF$Sati
sfaction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "TechSup", y = "Satisf",titl
e = "TechSup vs Satisf"),

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$CompRes,y=FactorHairRevised_DF$Sati
sfaction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "CompRes", y = "Satisfactio
n",title = "CompRes vs Satisf"),

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$Advertising,y=FactorHairRevised_DF
$Satisfaction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "Advrtsng", y = "Satisf",tit
le = "Advrtsng vs Satisf"),

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$ProdLine,y=FactorHairRevised_DF$Sat
isfaction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "ProdLine", y = "Satisf",tit
le = "ProdLine vs Satisf"),

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$SalesFImage,y=FactorHairRevised_DF
$Satisfaction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "SalesfImage", y = "Satisf",
title = "SalesfImage vs Satisf"),

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$ComPricing,y=FactorHairRevised_DF$S
atisfaction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "ComPricing", y = "Satisf",t
itle = "ComPricing vs Satisf"),

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$WartyClaim,y=FactorHairRevised_DF$S
atisfaction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "WartyClaim", y = "Satisf",t
itle = "WartyClaim vs Satisf"),

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$OrdBilling,y=FactorHairRevised_DF$S
atisfaction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "Ord&Billing", y = "Satisf",
title = "Ord&Billing vs Satisf"),

ggplot(FactorHairRevised_DF, aes(x = FactorHairRevised_DF$DelSpeed,y=FactorHairRevised_DF$Sat
isfaction))+geom_point()+
       geom_smooth(method=lm,col="red")+theme_classic()+labs(x = "DelSpeed", y = "Satisf",tit
le = "DelSpeed vs Satisf"),
```

```
heights = c(35,35,35,35,35,35,35,35,35,35,35), widths = c(35,35,35,35,35,35,35,35,35,35,35),n
col = 3, nrow = 4)
```




Overall its a weak association. Still satisfaction is somewhat more dependent on Product Quality, Complaint Resolution, Product Line, Salesforce Image, Order & Billing and Delivery Speed as against Ecommerce, Tech Support, Advertising and Warranty&Claims.Satisfaction has a weak negative relation with Competitive Pricing.

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 amogst themselves.For instance between Delivery Speed (DelSpeed) and Complaint Resolution (CompRes) the correlation coefficient is 0.87. Similarly the correlation coefficient is 0.8 between Warranty&Claim (WartyClaim) and Tech Support(TechSup). Competitive Pricing exhibits negative correlation with Product Quality and Product Line.

Hide

```
vif(model1)
```

```
   ProdQual        Ecom     TechSup     CompRes Advertising    ProdLine SalesFImage  ComPrici
ng  WartyClaim  OrdBilling    DelSpeed
   1.635797    2.756694    2.976796    4.730448    1.508933    3.488185    3.439420    1.6350
00    3.198337    2.902999    6.516014
```

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

Normality Tests

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 the a significance test called Shapiro-Wilk's test.

Null Hypothesis H0 : The variable considered is normally distributed Alternative Hypothesis H1: The variable considered is not normally distributed

Hide

```
shapiro.test(FactorHairRevised_DF$ProdQual)
```

```

    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$ProdQual
W = 0.94972, p-value = 0.0007953
```

Hide

```
shapiro.test(FactorHairRevised_DF$Ecom)
```

```

    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$Ecom
W = 0.95852, p-value = 0.003157
```

Hide

```
shapiro.test(FactorHairRevised_DF$TechSup)
```

```
    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$TechSup
W = 0.98626, p-value = 0.39
```

Hide

```
shapiro.test(FactorHairRevised_DF$CompRes)
```

```
    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$CompRes
W = 0.98646, p-value = 0.4023
```

Hide

```
shapiro.test(FactorHairRevised_DF$Advertising)
```

```
    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$Advertising
W = 0.97626, p-value = 0.06769
```

Hide

```
shapiro.test(FactorHairRevised_DF$ProdLine)
```

```
    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$ProdLine
W = 0.98692, p-value = 0.4324
```

Hide

```
shapiro.test(FactorHairRevised_DF$SalesFImage)
```

```
    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$SalesFImage
W = 0.97403, p-value = 0.04534
```

Hide

```
shapiro.test(FactorHairRevised_DF$ComPricing)
```

```
    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$ComPricing
W = 0.96758, p-value = 0.01448
```

Hide

```
shapiro.test(FactorHairRevised_DF$WartyClaim)
```

```
    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$WartyClaim
W = 0.99094, p-value = 0.7404
```

Hide

```
shapiro.test(FactorHairRevised_DF$OrdBilling)
```

```
    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$OrdBilling
W = 0.97405, p-value = 0.04549
```

Hide

```
shapiro.test(FactorHairRevised_DF$DelSpeed)
```

```
    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$DelSpeed
W = 0.98161, p-value = 0.177
```

Hide

```
shapiro.test(FactorHairRevised_DF$Satisfaction)
```

```
    Shapiro-Wilk normality test

data:  FactorHairRevised_DF$Satisfaction
W = 0.97516, p-value = 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 normall distributed.

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

Simple linear regression with every variable

Between Satisfaction and Product Quality

Hide

```
summary(Model_ProdQual)
```

```

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 Satisfaction and Ecommerce

Hide

```
Model_Ecommerce=lm(Satisfaction~Ecom)
summary(Model_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 Satisfaction and Tech Support

Hide

```
Model_TechSup=lm(Satisfaction~TechSup)
summary(Model_TechSup)
```

```

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 Satisfaction and Complaint Resolution

Hide

```
Model_CompRes=lm(Satisfaction~CompRes)
summary(Model_CompRes)
```

```

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 Satisfaction and Advertising

Hide

```
Model_Advertising=lm(Satisfaction~Advertising)
summary(Model_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 Satisfaction and Product Line

Hide

```
Model_ProdLine=lm(Satisfaction~ProdLine)
summary(Model_ProdLine)
```

```
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 Satisfaction and Salesforce Image

Hide

```
Model_SalesFImage=lm(Satisfaction~SalesFImage)
summary(Model_SalesFImage)
```

```
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 Satisfaction and Competitive Pricing

Hide

```
Model_ComPricing=lm(Satisfaction~ComPricing)
summary(Model_ComPricing)
```

```
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 Satisfaction and Warranty & Claims

Hide

```
Model_WartyClaim=lm(Satisfaction~WartyClaim)
summary(Model_WartyClaim)
```

```
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 Satisfaction and Order & Billing

Hide

```
Model_OrdBilling=lm(Satisfaction~OrdBilling)
summary(Model_OrdBilling)
```

```
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 Satisfaction and Delivery Speed

Hide

```
Model_DelSpeed=lm(Satisfaction~DelSpeed)
summary(Model_DelSpeed)
```

```
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.

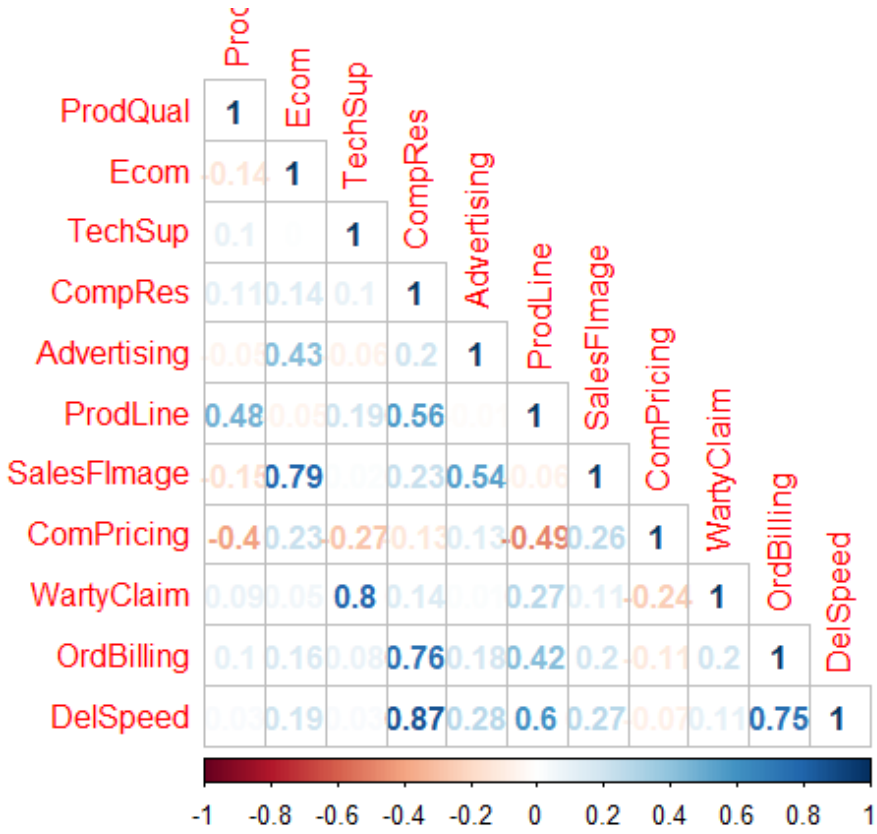
Due to the presence of multicollinearity in the dataset we now proceed to perform PCA/FA and do dimensionality reduction, thereby reducing the number of predictors to a smaller set of uncorrelated components

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

Before begining 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.

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.

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.




Hide

```
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

Hide

```
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  ComPrici
ng  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.

Proceeding with the factor analysis

Evaluating the eigen values for the correlation matrix

Hide

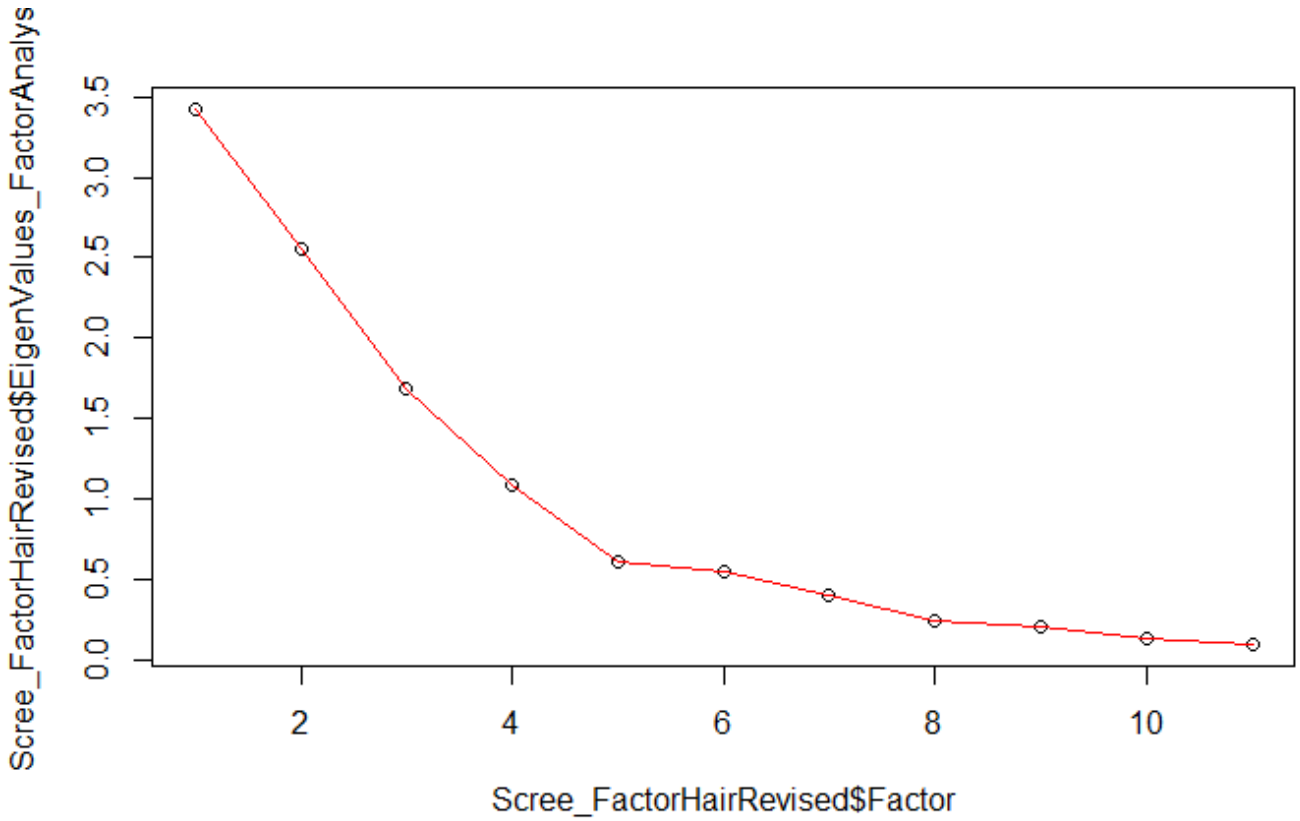
```
EigenValues_FactorAnalysis
```

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

Hide

```
plot(x=Scree_FactorHairRevised$Factor, y=Scree_FactorHairRevised$EigenValues_FactorAnalysis,
data=Scree_FactorHairRevised)
lines(x=Scree_FactorHairRevised$Factor, y=Scree_FactorHairRevised$EigenValues_FactorAnalysis,
data=Scree_FactorHairRevised, col="red")
```

```
"data" is not a graphical parameter
```




According to the Kaizer rule, number of factors to extract depend on the Eigen Value - Extract all the factors whoose eigen value is greater than or equal to 1.

Using the Kaizer Rule, we decide to reduce the dataset to 4 factors

Hide

```
Unrotate_FactorHairRevised
```

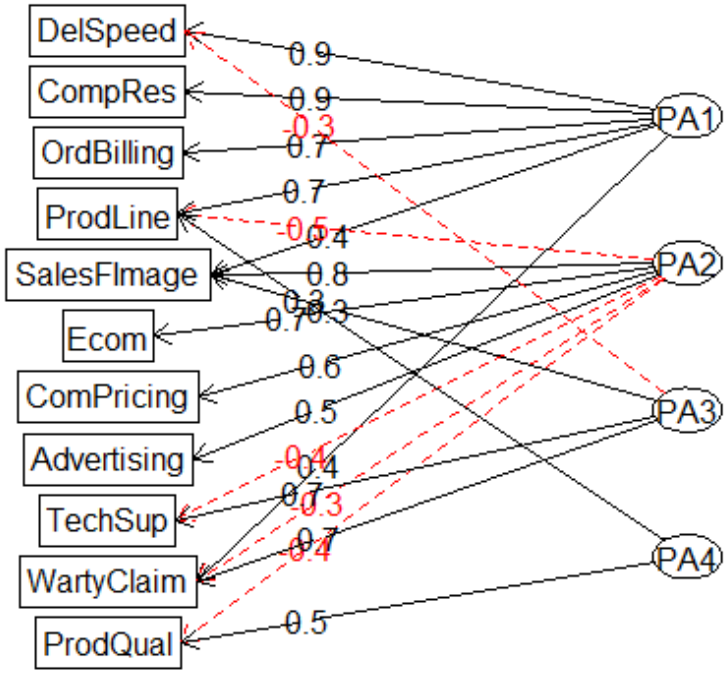
```
Factor Analysis using method =  pa
Call: fa(r = FactorHairRevised_DF[, 2:12], nfactors = 4, rotate = "none",
    fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
```

| | PA1 <S3: AsIs> | PA2 <S3: AsIs> | PA3 <S3: AsIs> | PA4 <S3: AsIs> | h2 <dbl> | u2 <dbl> | com <dbl> |
|---|---|---|---|---|---|---|---|
| ProdQual | 0.20 | -0.41 | -0.06 | 0.46 | 0.4242958 | 0.57570420 | 2.395874 |
| Ecom | 0.29 | 0.66 | 0.27 | 0.22 | 0.6381735 | 0.36182647 | 2.002805 |
| TechSup | 0.28 | -0.38 | 0.74 | -0.17 | 0.7946147 | 0.20538530 | 1.945189 |
| CompRes | 0.86 | 0.01 | -0.26 | -0.18 | 0.8428100 | 0.15718999 | 1.272078 |
| Advertising | 0.29 | 0.46 | 0.08 | 0.13 | 0.3142090 | 0.68579095 | 1.947435 |
| ProdLine | 0.69 | -0.45 | -0.14 | 0.31 | 0.8002906 | 0.19970935 | 2.300098 |
| SalesFImage | 0.39 | 0.80 | 0.35 | 0.25 | 0.9792432 | 0.02075678 | 2.114574 |
| ComPricing | -0.23 | 0.55 | -0.04 | -0.29 | 0.4432708 | 0.55672916 | 1.905757 |
| WartyClaim | 0.38 | -0.32 | 0.74 | -0.15 | 0.8135338 | 0.18646624 | 2.036647 |
| OrdBilling | 0.75 | 0.02 | -0.18 | -0.18 | 0.6218211 | 0.37817894 | 1.233989 |

1-10 of 11 rows Previous **1** 2 Next

```
                       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

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

The degrees of freedom for the null model are  55  and the objective function was  6.55 with
Chi Square of  619.27
The degrees of freedom for the model are 17  and the objective function was  0.33

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  0.03

The harmonic number of observations is  100 with the empirical chi square  3.19  with prob <
1
The total number of observations was  100  with Likelihood Chi Square =  30.27  with prob <
0.024

Tucker Lewis Index of factoring reliability =  0.921
RMSEA index =  0.096  and the 90 % confidence intervals are  0.032 0.139
BIC =  -48.01
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                PA1  PA2  PA3  PA4
Correlation of (regression) scores with factors   0.98 0.97 0.95 0.88
Multiple R square of scores with factors          0.96 0.95 0.91 0.78
Minimum correlation of possible factor scores     0.92 0.90 0.82 0.56
```

# Factor Analysis




Hide

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

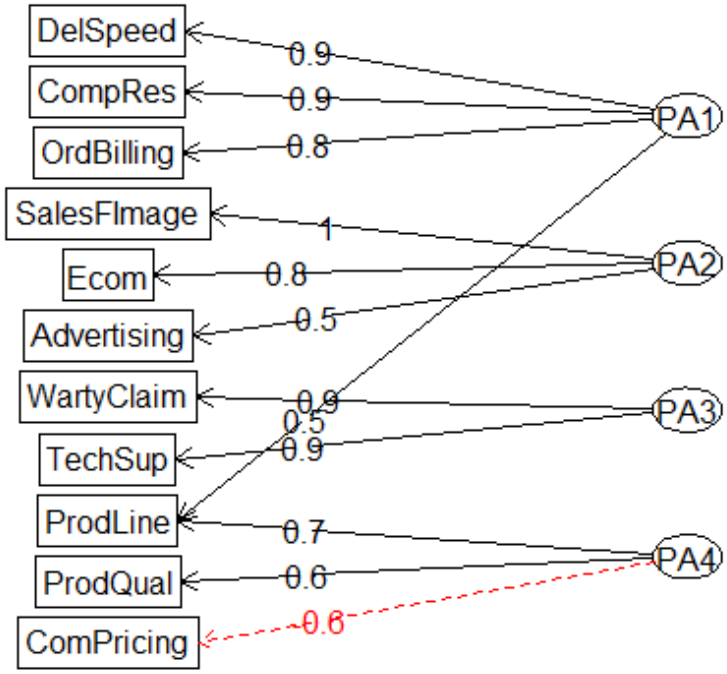
```
Factor Analysis using method =  pa
Call: fa(r = FactorHairRevised_DF[, 2:12], nfactors = 4, rotate = "varimax",
    fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
```

| | PA1 <S3: AsIs> | PA2 <S3: AsIs> | PA3 <S3: AsIs> | PA4 <S3: AsIs> | h2 <dbl> | u2 <dbl> | com <dbl> |
|---|---|---|---|---|---|---|---|
| ProdQual | 0.02 | -0.07 | 0.02 | 0.65 | 0.4242958 | 0.57570420 | 1.027410 |
| Ecom | 0.07 | 0.79 | 0.03 | -0.11 | 0.6381735 | 0.36182647 | 1.058914 |
| TechSup | 0.02 | -0.03 | 0.88 | 0.12 | 0.7946147 | 0.20538530 | 1.037430 |
| CompRes | 0.90 | 0.13 | 0.05 | 0.13 | 0.8428100 | 0.15718999 | 1.092936 |
| Advertising | 0.17 | 0.53 | -0.04 | -0.06 | 0.3142090 | 0.68579095 | 1.239268 |
| ProdLine | 0.53 | -0.04 | 0.13 | 0.71 | 0.8002906 | 0.19970935 | 1.921786 |
| SalesFImage | 0.12 | 0.97 | 0.06 | -0.13 | 0.9792432 | 0.02075678 | 1.075909 |
| ComPricing | -0.08 | 0.21 | -0.21 | -0.59 | 0.4432708 | 0.55672916 | 1.565831 |
| WartyClaim | 0.10 | 0.06 | 0.89 | 0.13 | 0.8135338 | 0.18646624 | 1.077658 |

| | PA1 <S3: AsIs> | PA2 <S3: AsIs> | PA3 <S3: AsIs> | PA4 <S3: AsIs> | h2 <dbl> | u2 <dbl> | com <dbl> |
|---|---|---|---|---|---|---|---|
| OrdBilling | 0.77 | 0.13 | 0.09 | 0.09 | 0.6218211 | 0.37817894 | 1.109102 |

1-10 of 11 rows Previous **1** 2 Next

```
                       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

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

The degrees of freedom for the null model are  55  and the objective function was  6.55 with
Chi Square of  619.27
The degrees of freedom for the model are 17  and the objective function was  0.33

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  0.03

The harmonic number of observations is  100 with the empirical chi square  3.19  with prob <
1
The total number of observations was  100  with Likelihood Chi Square =  30.27  with prob <
0.024

Tucker Lewis Index of factoring reliability =  0.921
RMSEA index =  0.096  and the 90 % confidence intervals are  0.032 0.139
BIC =  -48.01
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                 PA1  PA2  PA3  PA4
Correlation of (regression) scores with factors   0.98 0.99 0.94 0.88
Multiple R square of scores with factors          0.96 0.97 0.88 0.78
Minimum correlation of possible factor scores     0.93 0.94 0.77 0.55
```




Hide

```
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
```

| | Cust_Satisfaction <dbl> | Order_Processing <dbl> | Marketing <dbl> | PostSales_Service <dbl> | Product_Manager < |
|---|---|---|---|---|---|
| 1 | 8.2 | -0.1338871 | 0.9175166 | -1.719604873 | 0.0913 |
| 2 | 5.7 | 1.6297604 | -2.0090053 | -0.596361722 | 0.6580 |
| 3 | 8.9 | 0.3637658 | 0.8361736 | 0.002979966 | 1.3754 |
| 4 | 4.8 | -1.2225230 | -0.5491336 | 1.245473305 | -0.6442 |
| 5 | 7.1 | -0.4854209 | -0.4276223 | -0.026980304 | 0.4736 |

5 rows

Hide

```
summary(FactorHairRevised_FinalDF)
```

```
 Cust_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
```

Hide

```
summary(FactorHairRevised_Model)
```

```

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
```

It can be seen that p-value of the F-statistic is < 2.2e-16, which 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-statitic p-values:

Hide

```
summary(FactorHairRevised_Model)$coefficient
```

```
                    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 while changes in PostSales_Service is not significantly associated with Customer Satisfaction.

The confidence interval of the model coefficient can be extracted as follow:

Hide

```
confint(FactorHairRevised_Model)
```

```
                       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
```

Hide

```
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

Model accuracy and assessment - calculating Residual Sandard Error

Hide

```
sigma(FactorHairRevised_Model)/mean(FactorHairRevised_FinalDF$Cust_Satisfaction)
```

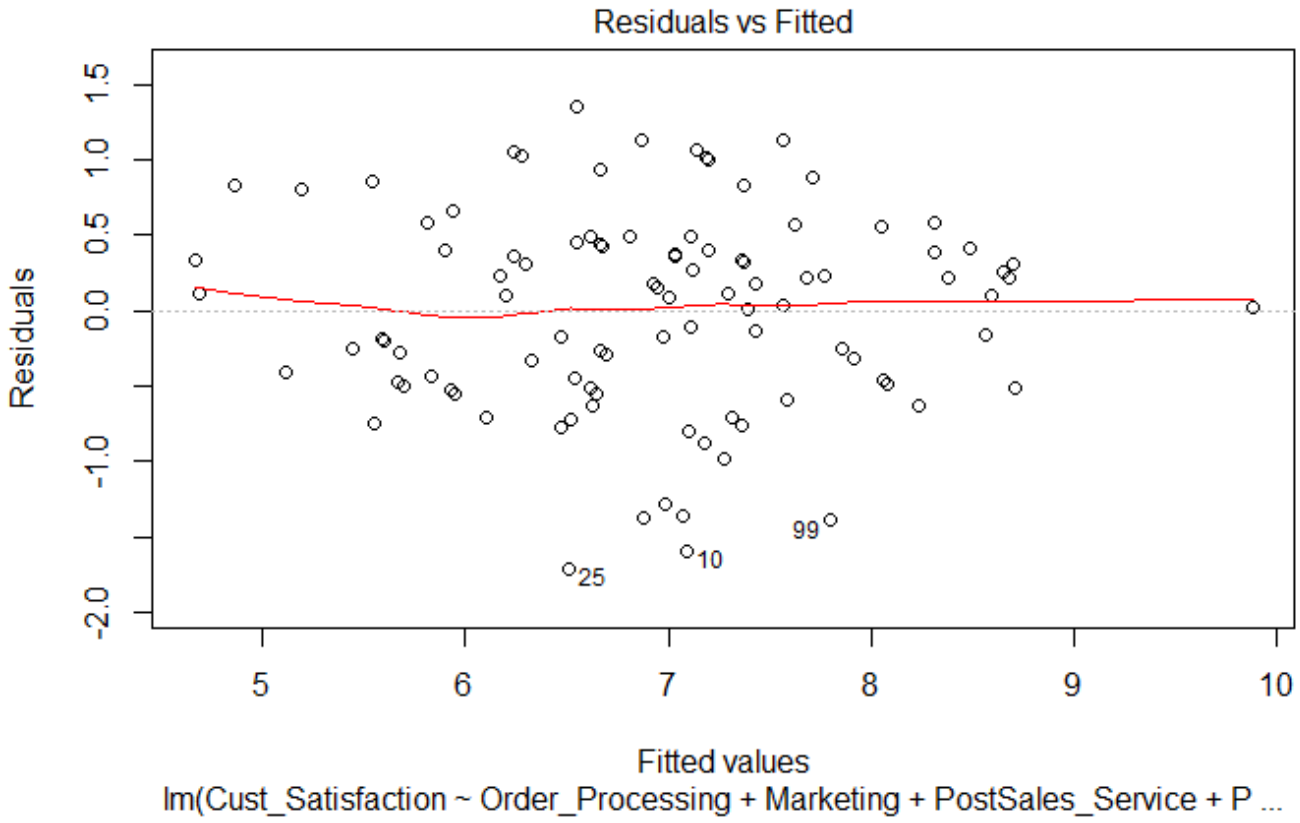
```
[1] 0.09678969
```

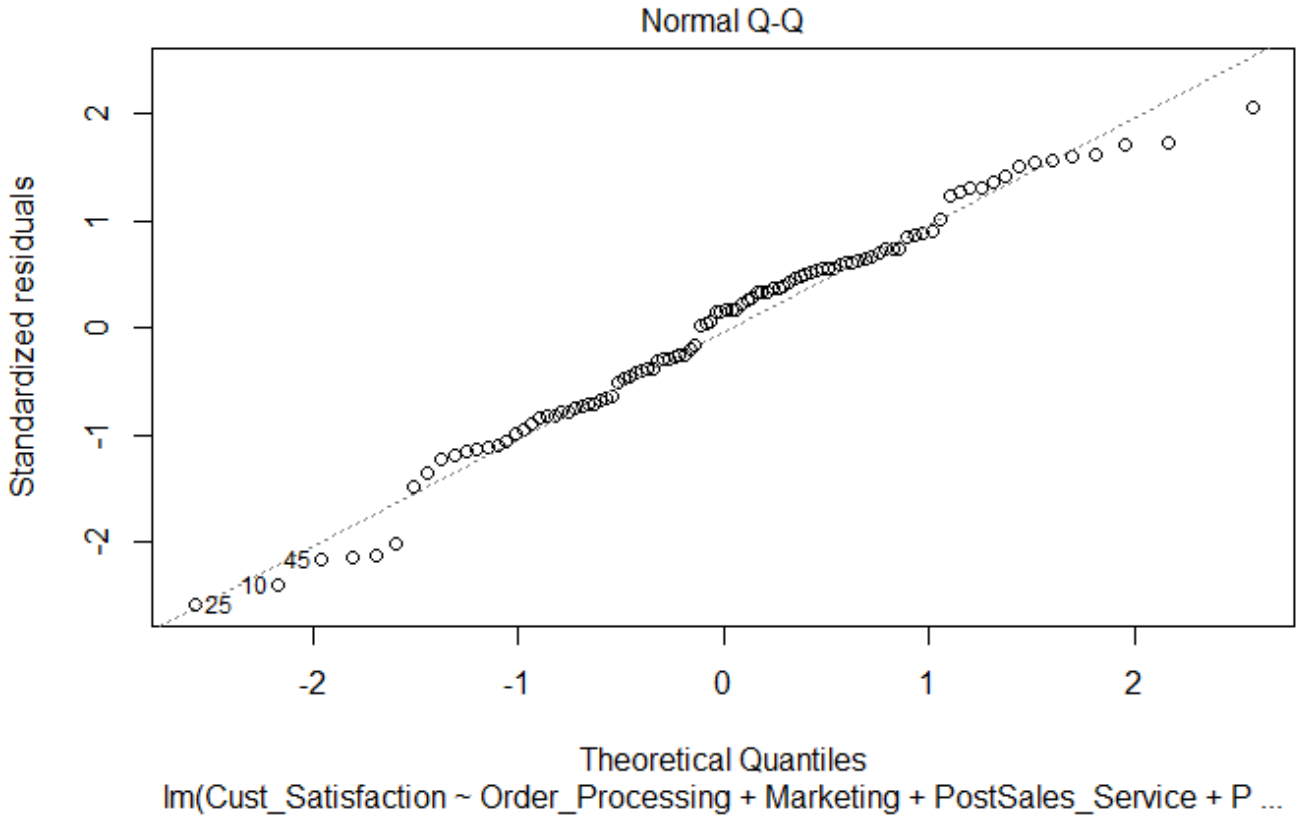
Therefore the model has only 9.6% error rate.

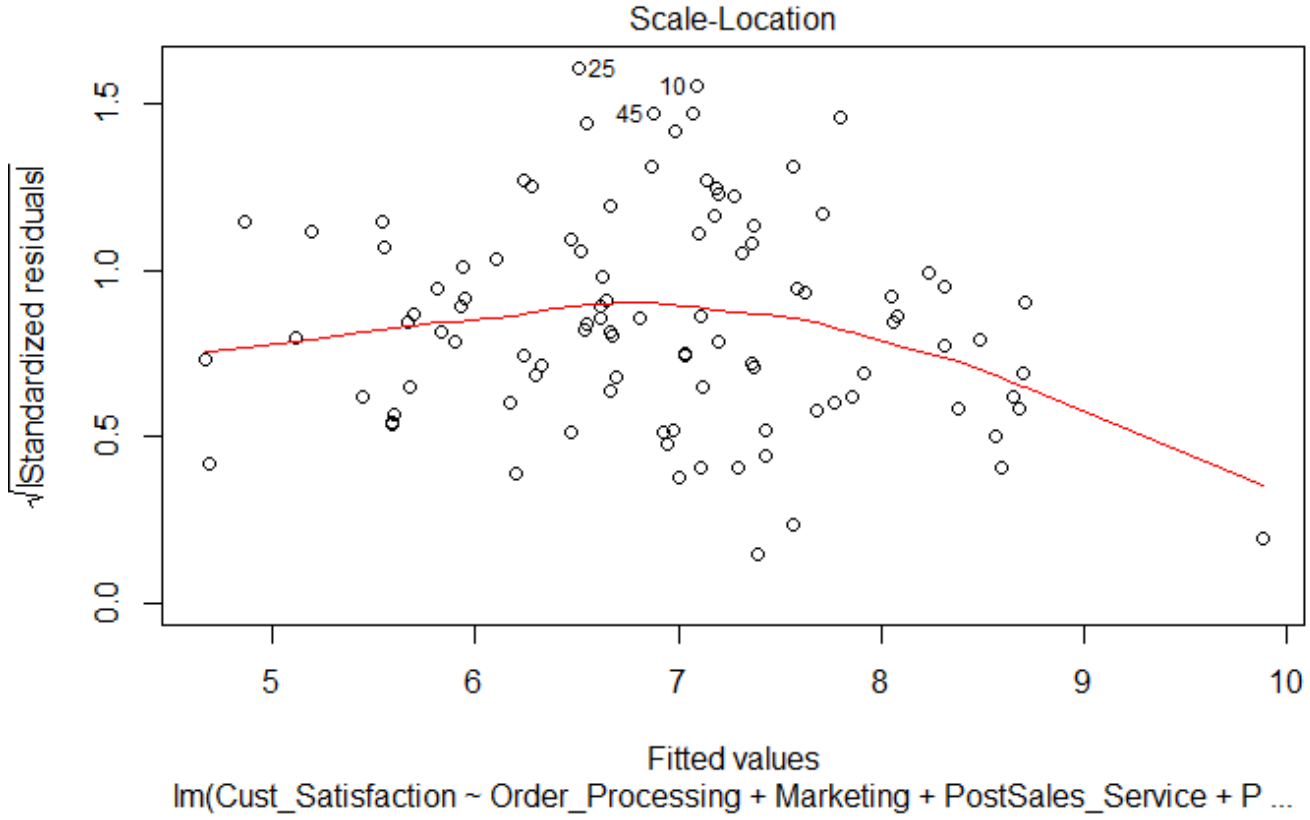
Residual Plot Analysis

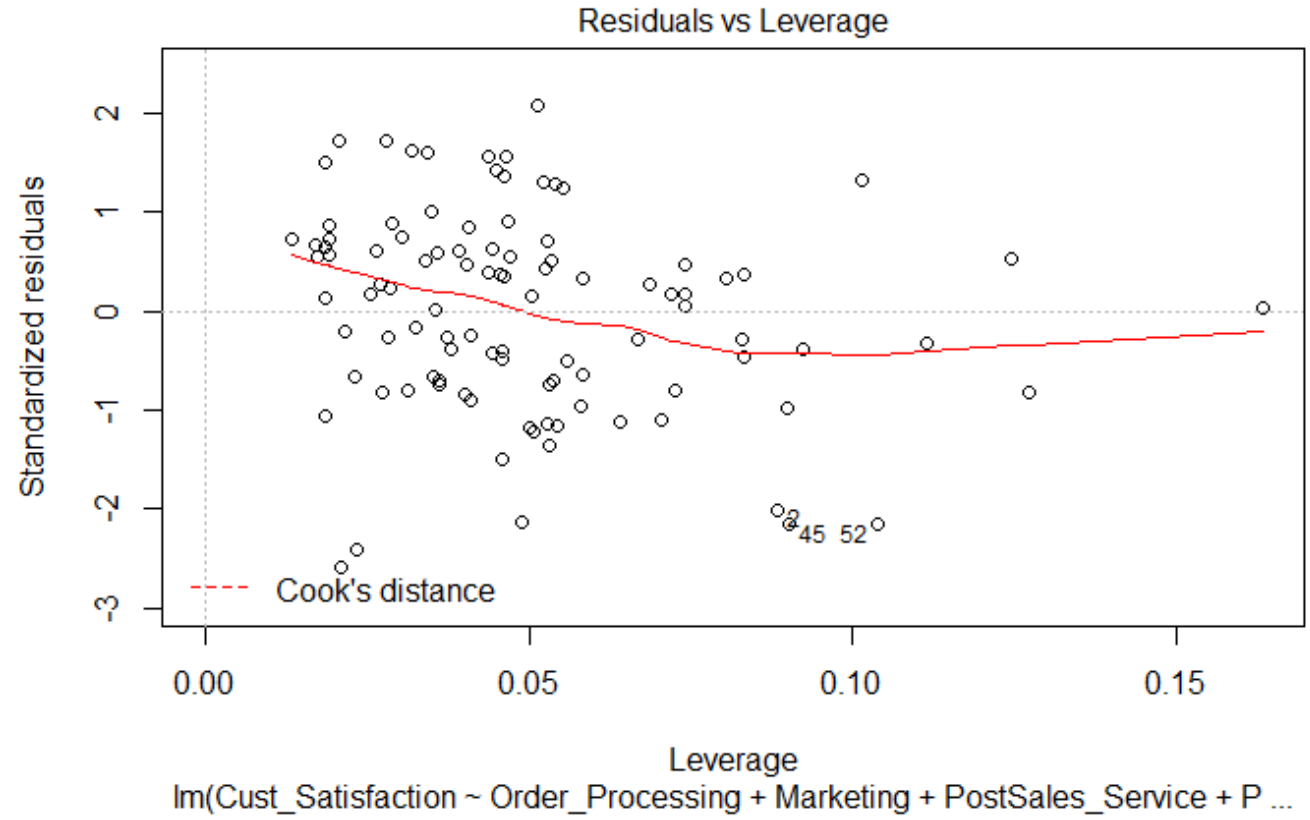
Hide

```
plot(FactorHairRevised_Model)
```


Residuals vs Fitted
Residuals
1.5
1.0
0.5
0.0
-1.0
-2.0
5
6
7
8
9
10
99
10
25
Fitted values
lm(Cust_Satisfaction ~ Order_Processing + Marketing + PostSales_Service + P ...


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-2
-1
0
1
2
45
10
25
Theoretical Quantiles
lm(Cust_Satisfaction ~ Order_Processing + Marketing + PostSales_Service + P ...


Scale-Location
25
10
45
√|Standardized residuals|
0.0
0.5
1.0
1.5
5
6
7
8
9
10
Fitted values
lm(Cust_Satisfaction ~ Order_Processing + Marketing + PostSales_Service + P ...


Residuals vs Leverage
Standardized residuals
2
1
0
-1
-2
-3
2
45
52
Cook's distance
0.00
0.05
0.10
0.15
Leverage
lm(Cust_Satisfaction ~ Order_Processing + Marketing + PostSales_Service + P ...