Sample Solution

Project Objective:

What is multicollinearity and how it affects the regression model? Multicollinearity occurs when the independent variables of a regression model are correlated and if the degree of collinearity between the independent variables are high, it becomes difficult to estimate the relationship between each independent variable and the dependent variable and the overall precision of the estimate coefficients. Even though the regression models with high multicollinearity can give you a high R squared but hardly any significant variables.

The objective of the project is to use the dataset *Factor-Hair-Revised.csv* to build a regression model to predict *satisfaction*.

Project Approach:

- Data Exploration
- Collinearity of the variables
- Initial Regression analysis
- Factor Analysis
- Labelling and interpreting of the factors
- Regression analysis using the factors as independent variable
- Model performance measures

Data Exploration:

Let's import the data and check the basic descriptive statistics.

```
1. data <- read.csv("Factor-Hair-Revised.csv", header = TRUE, sep = ",")
2. head(data)
```

```
  ID ProdQual Ecom TechSup CompRes Advertising ProdLine SalesFImage
1  1      8.5  3.9     2.5     5.9         4.8      4.9         6.0
2  2      8.2  2.7     5.1     7.2         3.4      7.9         3.1
3  3      9.2  3.4     5.6     5.6         5.4      7.4         5.8
4  4      6.4  3.3     7.0     3.7         4.7      4.7         4.5
5  5      9.0  3.4     5.2     4.6         2.2      6.0         4.5
6  6      6.5  2.8     3.1     4.1         4.0      4.3         3.7
  ComPricing WartyClaim OrdBilling DelSpeed Satisfaction
1        6.8        4.7        5.0      3.7          8.2
2        5.3        5.5        3.9      4.9          5.7
3        4.5        6.2        5.4      4.5          8.9
4        8.8        7.0        4.3      3.0          4.8
5        6.8        6.1        4.5      3.5          7.1
6        8.5        5.1        3.6      3.3          4.7

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

Now let's check the structure and summary statistics of the data set.

```
> str(data)
'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 ...
```

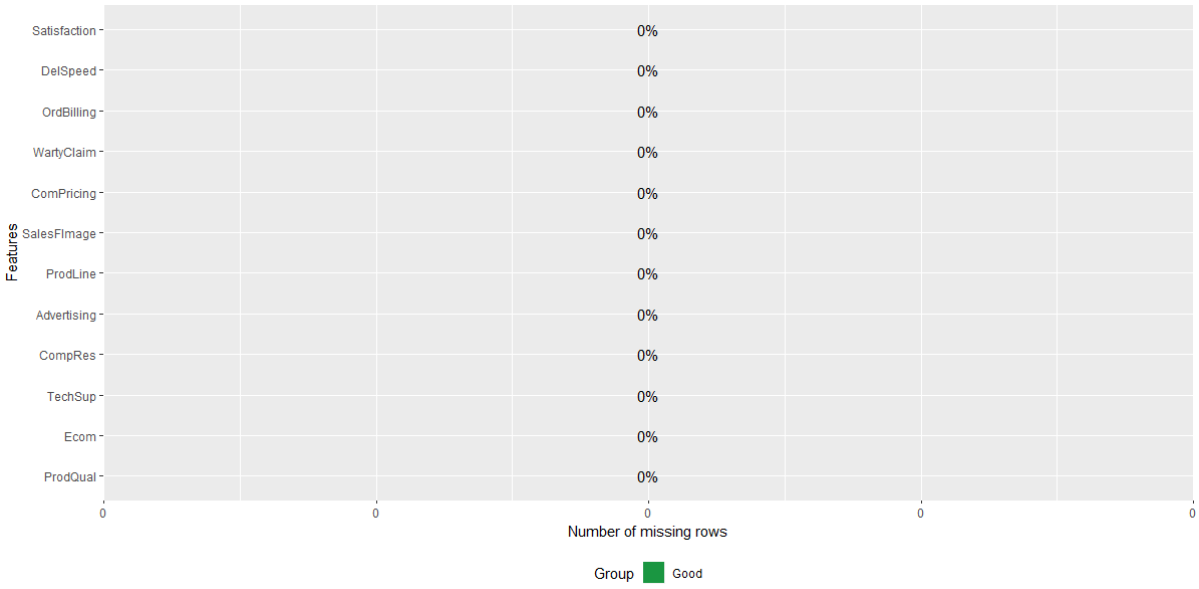
```
> describe(data)
             vars   n  mean    sd median trimmed   mad min   max range  skew kurtosis
ID              1 100 50.50 29.01  50.50   50.50 37.06 1.0 100.0  99.0  0.00    -1.24
ProdQual        2 100  7.81  1.40   8.00    7.85  1.78 5.0  10.0   5.0 -0.24    -1.17
Ecom            3 100  3.67  0.70   3.60    3.63  0.52 2.2   5.7   3.5  0.64     0.57
TechSup         4 100  5.37  1.53   5.40    5.40  1.85 1.3   8.5   7.2 -0.20    -0.63
CompRes         5 100  5.44  1.21   5.45    5.46  1.26 2.6   7.8   5.2 -0.13    -0.66
Advertising     6 100  4.01  1.13   4.00    4.00  1.19 1.9   6.5   4.6  0.04    -0.94
ProdLine        7 100  5.80  1.32   5.75    5.81  1.56 2.3   8.4   6.1 -0.09    -0.60
SalesFImage     8 100  5.12  1.07   4.90    5.09  0.89 2.9   8.2   5.3  0.37     0.26
ComPricing      9 100  6.97  1.55   7.10    7.01  1.93 3.7   9.9   6.2 -0.23    -0.96
WartyClaim     10 100  6.04  0.82   6.10    6.04  0.89 4.1   8.1   4.0  0.01    -0.53
OrdBilling     11 100  4.28  0.93   4.40    4.31  0.74 2.0   6.7   4.7 -0.32     0.11
DelSpeed       12 100  3.89  0.73   3.90    3.92  0.74 1.6   5.5   3.9 -0.45     0.09
Satisfaction   13 100  6.92  1.19   7.05    6.90  1.33 4.7   9.9   5.2  0.08    -0.86
```

The variable ID is unique number given to individuals and as such should not have any explanatory power for explaining *Satisfaction* in the regression equation. So we can safely drop ID from the dataset.

```
1.  data1 <- subset(data, select = -c(1))
```

Now let's check for the missing values in the dataset.

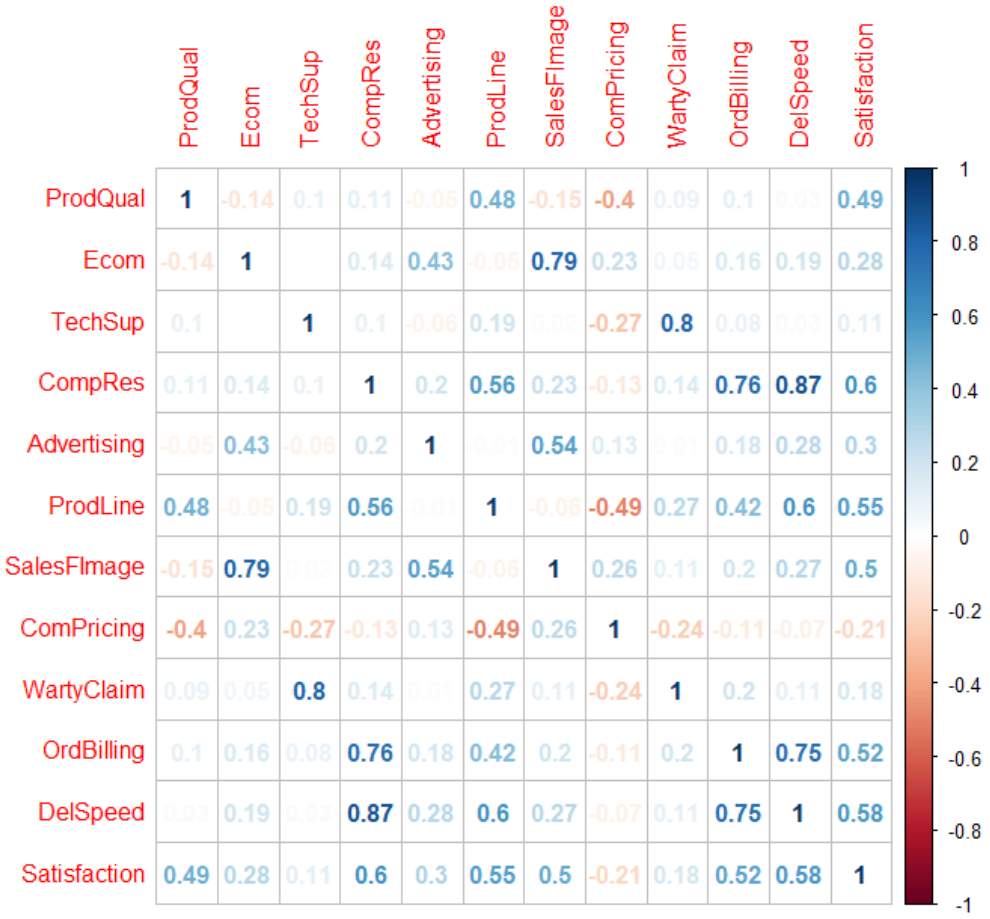
```
1.  library(DataExplorer)
2.  plot_missing(data)
```



There are no missing values in the datasets.

Inter-item Correlation analysis:

Now let's plot the correlation matrix plot of the data.



Observations:

1. CompRes and DelSpeed are highly correlated
2. OrdBilling and CompRes are highly correlated
3. WartyClaim and TechSupport are highly correlated
4. CompRes and OrdBilling are highly correlated
5. OrdBilling and DelSpeed are highly correlated
6. Ecom and SalesFImage are highly correlated

For examining the patterns of multicollinearity, it is required to conduct t-test for correlation coefficient. Let's use the *ppcor* package to compute the partial correlation coefficients along with the t-statistics and corresponding *p* values for the independent variables.

```
1. library(ppcor)
2. pcor(data1, method = "pearson")
```

```
> pcor(data1, method = "pearson")
$estimate
                ProdQual          Ecom      TechSup      CompRes   Advertising
ProdQual      1.000000000  0.1549597037  0.002424222 -0.056354142  0.1123767461
Ecom          0.154959704  1.0000000000  0.082359000 -0.033513777 -0.0002972504
TechSup       0.002424222  0.0823590001  1.000000000  0.143603415 -0.0593002540
CompRes      -0.056354142 -0.0335137768  0.143603415  1.000000000 -0.0648060931
Advertising   0.112376746 -0.0002972504 -0.059300254 -0.064806093  1.0000000000
ProdLine      0.281144724  0.1538660545 -0.125349050  0.020305650 -0.1319231866
SalesFImage  -0.376228551  0.7321011890 -0.093310796 -0.022150933  0.2628793429
ComPricing   -0.014021386  0.0149131857 -0.132972485 -0.004487151 -0.0632980381
WartyClaim   -0.042752416 -0.1232553822  0.787729606 -0.109686211  0.0275909587
OrdBilling    0.054523215  0.1546990521 -0.165914724  0.288344382 -0.0326580430
DelSpeed     -0.335084210 -0.0083917930 -0.021914916  0.528932328  0.2046544064
Satisfaction  0.607438787 -0.3308440834  0.055111867  0.172416347 -0.0449735411
                ProdLine  SalesFImage   ComPricing  WartyClaim  OrdBilling
ProdQual      0.28114472 -0.376228551 -0.014021386 -0.04275242  0.05452322
Ecom          0.15386605  0.732101189  0.014913186 -0.12325538  0.15469905
TechSup      -0.12534905 -0.093310796 -0.132972485  0.78772961 -0.16591472
CompRes       0.02030565 -0.022150933 -0.004487151 -0.10968621  0.28834438
Advertising  -0.13192319  0.262879343 -0.063298038  0.02759096 -0.03265804
ProdLine      1.00000000 -0.230170570 -0.361757904  0.25718360 -0.28098068
SalesFImage  -0.23017057  1.000000000  0.126612489  0.18930027 -0.18235930
ComPricing   -0.36175790  0.126612489  1.000000000  0.02035155 -0.08650087
WartyClaim    0.25718360  0.189300271  0.020351554  1.00000000  0.25984451
OrdBilling   -0.28098068 -0.182359301 -0.086500869  0.25984451  1.00000000
DelSpeed      0.50189505  0.005889925  0.190437993 -0.09164900  0.35053021
Satisfaction  0.18325156  0.660251850 -0.087467711 -0.08868071  0.14880055
                DelSpeed Satisfaction
ProdQual     -0.335084210   0.60743879
Ecom         -0.008391793  -0.33084408
TechSup      -0.021914916   0.05511187
CompRes       0.528932328   0.17241635
Advertising   0.204654406  -0.04497354
ProdLine      0.501895051   0.18325156
SalesFImage   0.005889925   0.66025185
ComPricing    0.190437993  -0.08746771
WartyClaim   -0.091649002  -0.08868071
OrdBilling    0.350530212   0.14880055
DelSpeed      1.000000000   0.08955614
Satisfaction  0.089556143   1.00000000
```

```
$p.value
                ProdQual          Ecom      TechSup      CompRes Advertising
ProdQual     0.000000e+00 1.447435e-01 9.819081e-01 5.977987e-01  0.29163501
Ecom         1.447435e-01 0.000000e+00 4.402823e-01 7.538375e-01  0.99778145
TechSup      9.819081e-01 4.402823e-01 0.000000e+00 1.769171e-01  0.57876015
CompRes      5.977987e-01 7.538375e-01 1.769171e-01 0.000000e+00  0.54395113
Advertising  2.916350e-01 9.977814e-01 5.787601e-01 5.439511e-01  0.00000000
ProdLine     7.269030e-03 1.476353e-01 2.391187e-01 8.493380e-01  0.21517368
SalesFImage  2.576212e-04 2.442012e-16 3.817042e-01 8.358301e-01  0.01230672
ComPricing   8.956443e-01 8.890480e-01 2.115149e-01 9.665194e-01  0.55338281
WartyClaim   6.890839e-01 2.471182e-01 3.262868e-20 3.034147e-01  0.79629803
OrdBilling   6.097697e-01 1.454288e-01 1.180864e-01 5.850710e-03  0.75993047
DelSpeed     1.245192e-03 9.374303e-01 8.375553e-01 8.359069e-08  0.05300070
Satisfaction 2.182238e-10 1.447603e-03 6.059096e-01 1.041593e-01  0.67382405
                ProdLine  SalesFImage   ComPricing   WartyClaim   OrdBilling
ProdQual     7.269030e-03 2.576212e-04 0.8956443272 6.890839e-01 0.6097697424
Ecom         1.476353e-01 2.442012e-16 0.8890479591 2.471182e-01 0.1454287628
TechSup      2.391187e-01 3.817042e-01 0.2115148546 3.262868e-20 0.1180864174
CompRes      8.493380e-01 8.358301e-01 0.9665194173 3.034147e-01 0.0058507099
Advertising  2.151737e-01 1.230672e-02 0.5533828069 7.962980e-01 0.7599304715
ProdLine     0.000000e+00 2.907563e-02 0.0004593443 1.440249e-02 0.0073046065
SalesFImage  2.907563e-02 0.000000e+00 0.2343791374 7.394541e-02 0.0853804743
ComPricing   4.593443e-04 2.343791e-01 0.0000000000 8.490015e-01 0.4175555780
WartyClaim   1.440249e-02 7.394541e-02 0.8490014635 0.000000e+00 0.0133877361
OrdBilling   7.304606e-03 8.538047e-02 0.4175555780 1.338774e-02 0.0000000000
DelSpeed     4.664278e-07 9.560619e-01 0.0721942454 3.902770e-01 0.0007064114
Satisfaction 8.383625e-02 1.449719e-12 0.4123500152 4.058729e-01 0.1615974575
                DelSpeed Satisfaction
ProdQual     1.245192e-03 2.182238e-10
Ecom         9.374303e-01 1.447603e-03
TechSup      8.375553e-01 6.059096e-01
CompRes      8.359069e-08 1.041593e-01
Advertising  5.300070e-02 6.738241e-01
ProdLine     4.664278e-07 8.383625e-02
SalesFImage  9.560619e-01 1.449719e-12
ComPricing   7.219425e-02 4.123500e-01
WartyClaim   3.902770e-01 4.058729e-01
OrdBilling   7.064114e-04 1.615975e-01
DelSpeed     0.000000e+00 4.012356e-01
Satisfaction 4.012356e-01 0.000000e+00

$statistic
                ProdQual          Ecom      TechSup      CompRes  Advertising
ProdQual      0.00000000  1.471424516  0.02274128 -0.52949015  1.060907458
Ecom          1.47142452  0.000000000  0.77522957 -0.31456380 -0.002788456
TechSup       0.02274128  0.775229571  0.00000000  1.36122814 -0.557266374
CompRes      -0.52949015 -0.314563798  1.36122814  0.00000000 -0.609215688
Advertising   1.06090746 -0.002788456 -0.55726637 -0.60921569  0.000000000
ProdLine      2.74821968  1.460787002 -1.18522655  0.19052316 -1.248460807
SalesFImage  -3.80921125 10.081854496 -0.87916864 -0.20784517  2.555921881
ComPricing   -0.13154519  0.139913642 -1.25856891 -0.04209363 -0.594981366
WartyClaim   -0.40142023 -1.165122048 11.99562484 -1.03519395  0.258924708
OrdBilling    0.51223505  1.468888754 -1.57829305  2.82489237 -0.306523103
DelSpeed     -3.33624278 -0.078724768 -0.20562952  5.84663073  1.961341689
Satisfaction  7.17336519 -3.288799958  0.51778207  1.64199901 -0.422316520
                ProdLine SalesFImage  ComPricing WartyClaim OrdBilling    DelSpeed
ProdQual       2.7482197 -3.80921125 -0.13154519 -0.4014202  0.5122350 -3.33624278
Ecom           1.4607870 10.08185450  0.13991364 -1.1651220  1.4688888 -0.07872477
TechSup       -1.1852266 -0.87916864 -1.25856891 11.9956248 -1.5782930 -0.20562952
CompRes        0.1905232 -0.20784517 -0.04209363 -1.0351939  2.8248924  5.84663073
Advertising   -1.2484608  2.55592188 -0.59498137  0.2589247 -0.3065231  1.96134169
ProdLine       0.0000000 -2.21876450 -3.64012829  2.4965744 -2.7464786  5.44344736
SalesFImage   -2.2187645  0.00000000  1.19736653  1.8084929 -1.7398559  0.05525335
ComPricing    -3.6401283  1.19736653  0.00000000  0.1909540 -0.8145030  1.81976992
WartyClaim     2.4965744  1.80849286  0.19095404  0.0000000  2.5242649 -0.86337748
OrdBilling    -2.7464786 -1.73985585 -0.81450302  2.5242649  0.0000000  3.51103503
DelSpeed       5.4434474  0.05525335  1.81976992 -0.8633775  3.5110350  0.00000000
Satisfaction   1.7486638  8.24679932 -0.82367672 -0.8351894  1.4115877  0.84350046
              Satisfaction
```

```
ProdQual         7.1733652
Ecom            -3.2888000
TechSup          0.5177821
CompRes          1.6419990
Advertising     -0.4223165
ProdLine         1.7486638
SalesFImage      8.2467993
ComPricing      -0.8236767
WartyClaim      -0.8351894
OrdBilling       1.4115877
DelSpeed         0.8435005
Satisfaction     0.0000000

$n
[1] 100

$gp
[1] 10

$method
[1] "pearson"
```

As expected the correlation between *sales force image* and *ecommerce* is highly significant; so is the correlation between *delivery speed* and *order billing* with *complaint resolution*. Also, the correlation between *order & billing* and delivery speed. We can safely assume that there is a high degree of collinearity between the independent variables.

Initial Regression Model using the data as it is.

Let's build a model using all the independent variables and check its performance.

```
1. model0 = lm(Satisfaction~., data1)
2. summary(model0)
3. vif(model0)
```

```
Call:
lm(formula = Satisfaction ~ ., data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43005 -0.31165  0.07621  0.37190  0.90120

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.66961    0.81233  -0.824  0.41199
ProdQual     0.37137    0.05177   7.173 2.18e-10 ***
Ecom        -0.44056    0.13396  -3.289  0.00145 **
TechSup      0.03299    0.06372   0.518  0.60591
CompRes      0.16703    0.10173   1.642  0.10416
Advertising -0.02602    0.06161  -0.422  0.67382
ProdLine     0.14034    0.08025   1.749  0.08384 .
SalesFImage  0.80611    0.09775   8.247 1.45e-12 ***
ComPricing  -0.03853    0.04677  -0.824  0.41235
WartyClaim  -0.10298    0.12330  -0.835  0.40587
OrdBilling   0.14635    0.10367   1.412  0.16160
DelSpeed     0.16570    0.19644   0.844  0.40124
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5623 on 88 degrees of freedom
Multiple R-squared:  0.8021,   Adjusted R-squared:  0.7774
```

```
F-statistic: 32.43 on 11 and 88 DF,  p-value: < 2.2e-16

> vif(model0)
   ProdQual        Ecom     TechSup     CompRes Advertising
   1.635797    2.756694    2.976796    4.730448    1.508933
   ProdLine SalesFImage  ComPricing  WartyClaim  OrdBilling
   3.488185    3.439420    1.635000    3.198337    2.902999
   DelSpeed
   6.516014
```

Here, even though independent variables explain 78% of variance of the dependent variable, only 3 variables are significant out of 11 independent variables.

Now let's study the VIF values.

High Variable Inflation Factor (VIF) is a sign of multicollinearity. There is no formal VIF value for determining presence of multicollinearity; however in weaker models VIF value greater than 2.5 may be a cause of concern.

From the VIF values we can infer that variables `DelSpeed and CompRes` are a cause of concern.

Remedial Measures:

Two of the most commonly used methods to deal with multicollinearity in the model is the following.

- Remove some of the highly correlated variables using VIF or stepwise algorithms.
- Perform an analysis design like principal component analysis (PCA)/ Factor Analysis on the correlated variables.

Factor Analysis:

Now let's check the factorability of the variables in the dataset.

First create a new dataset by taking a subset of all the independent variables in the data and perform Kaiser-Meyer-Olkin (KMO) Test.

```
1. data2 <- subset(data1, select = -c(12))
2. datamatrix<-cor(data1)
3. KMO(r=datamatrix)
```

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = datamatrix)
Overall MSA =  0.66
MSA for each item =
    ProdQual         Ecom      TechSup      CompRes  Advertising     ProdLine
        0.49         0.59         0.52         0.83         0.83         0.70
 SalesFImage   ComPricing   WartyClaim   OrdBilling     DelSpeed Satisfaction
        0.52         0.77         0.52         0.79         0.72         0.66
```

Since MSA > 0.5, we can run Factor Analysis on this data.

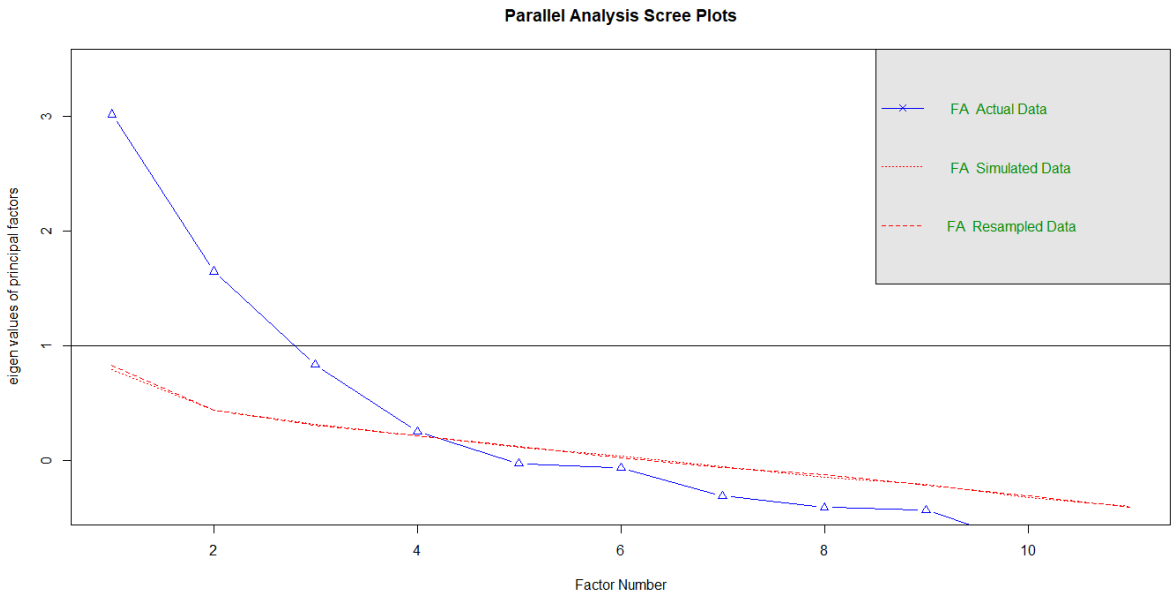
Bartlett's test of sphericity should be significant.

```
1.  cortest.bartlett(datamatrix, n = 50)
```

```
$chisq
[1] 360.9826

$p.value
[1] 3.049209e-42

$df
[1] 66
```

Now let's use *Psych* package's *fa.parallel* function to execute parallel analysis to find acceptable number of factors and generate the scree plot.

```
1.  parallel <- fa.parallel(data2, fm = 'minres', fa = 'fa')
```



The blue line shows eigenvalues of actual data and the two red lines (placed on top of each other) show simulated and resampled data. Here we look at the large drops in the actual data and spot the point where it levels off to the right.

Looking at the plot 3 or 4 factors would be a good choice.

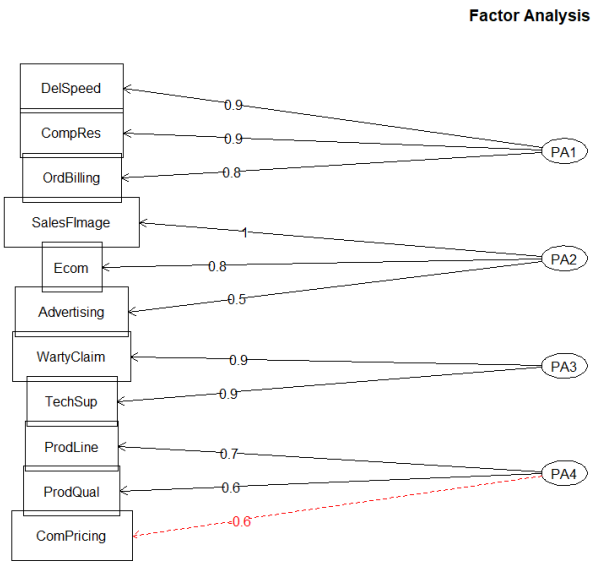
Let's use 4 factors to perform the factor analysis. Also, let's use orthogonal rotation (varimax) because in orthogonal rotation the rotated factors will remain uncorrelated whereas in oblique rotation the resulting factors will be correlated.

```
1. fa1<- fa(r=data2, nfactors = 4, rotate="varimax",fm="pa")
2. print(fa1)
3. fa.diagram(fa1)
```

```
Factor Analysis using method =  pa
Call: fa(r = data2, nfactors = 4, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
              PA1   PA2   PA3   PA4   h2    u2 com
ProdQual     0.02 -0.07  0.02  0.65 0.42 0.576 1.0
Ecom         0.07  0.79  0.03 -0.11 0.64 0.362 1.1
TechSup      0.02 -0.03  0.88  0.12 0.79 0.205 1.0
CompRes      0.90  0.13  0.05  0.13 0.84 0.157 1.1
Advertising  0.17  0.53 -0.04 -0.06 0.31 0.686 1.2
ProdLine     0.53 -0.04  0.13  0.71 0.80 0.200 1.9
SalesFImage  0.12  0.97  0.06 -0.13 0.98 0.021 1.1
ComPricing  -0.08  0.21 -0.21 -0.59 0.44 0.557 1.6
WartyClaim   0.10  0.06  0.89  0.13 0.81 0.186 1.1
OrdBilling   0.77  0.13  0.09  0.09 0.62 0.378 1.1
DelSpeed     0.95  0.19  0.00  0.09 0.94 0.058 1.1

                       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 funct
ion was  6.55 with Chi Square of  619.27
The degrees of freedom for the model are 17  and the objective function wa
s  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 =  3
0.27  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
```

Let's visualize the factors.



Labelling and interpretation of the factors.

#	Factors	Variables	Label	Short Interpretation
1	PA1	DelSpeed, CompRes, OrdBilling	Purchase	All the items are related to purchasing the product; from placing order to billing and getting it delivered.
2	PA2	SalesFImage, Ecom, Advertising	Marketing	In this factors the items are related to marketing processes like the image of sales force and spending on advertising.
3	PA3	WartyClaim, TechSup	Post Purchase	Post purchase activities are included in this factor; like warranty claims and technical support.
4	PA4	ProdLine, ProdQual, CompPricing	Product Position	Product positioning related items are grouped in this factor.

Regression analysis using the factors scores as the independent variable.

Let's combine the dependent variable and the factor scores into a dataset and label them.

```
1. regdata <- cbind(data1[12], fa1$scores)
2. head(regdata)
3. names(regdata) <- c("Satisfaction", "Purchanse", "Marketing",
   "Post_purchase", "Prod_positioning")
```

```
> head(regdata)
  Satisfaction  Purchanse  Marketing Post_purchase Prod_positioning
1          8.2 -0.1338871  0.9175166  -1.719604873      0.09135411
2          5.7  1.6297604 -2.0090053  -0.596361722      0.65808192
3          8.9  0.3637658  0.8361736   0.002979966      1.37548765
4          4.8 -1.2225230 -0.5491336   1.245473305     -0.64421384
5          7.1 -0.4854209 -0.4276223  -0.026980304      0.47360747
6          4.7 -0.5950924 -1.3035333  -1.183019401     -0.95913571
```

Let's the dataset into training and testing dataset (70:30)

```
1. set.seed(100)
2. indices= sample(1:nrow(regdata), 0.7*nrow(regdata))
3. train=regdata[indices,]
4. test = regdata[-indices,]
```

Let's train the regression model.

```
1. model1 = lm(Satisfaction~., train)
2. summary(model1)
```

```
Call:
lm(formula = Satisfaction ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6857 -0.4018  0.1051  0.4027  1.2036

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.92625    0.08263  83.827  < 2e-16 ***
Purchanse         0.62022    0.08408   7.377 3.73e-10 ***
Marketing         0.57735    0.08047   7.175 8.50e-10 ***
Post_purchase     0.09567    0.08667   1.104    0.274
Prod_positioning  0.66562    0.09374   7.101 1.15e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6814 on 65 degrees of freedom
Multiple R-squared:  0.7079,   Adjusted R-squared:   0.69
F-statistic: 39.39 on 4 and 65 DF,  p-value: < 2.2e-16
```

The factors Purchasing, Marketing and Prod_positioning are significant in the model.

Let's check the VIF scores.

```
> vif(model1)
       Purchanse        Marketing    Post_purchase Prod_positioning
        1.012217         1.009683         1.009037         1.012533
```

All the VIF values are reasonably alright; we don't have multicollinearity in the model.

Root mean square error of the model:

```
> rmse1 <- sqrt((mean(model1$residuals^2)))
> rmse1
[1] 0.6566331
```

Now let's check prediction of the model in the test dataset.

```
1.  pred=predict(model1, newdata = test, type = "response")
2.  test$Satisfaction.Predict <- pred
```

```
> head(test[c(1,6)],10)
   Satisfaction Satisfaction.Predict
1           8.2              7.269232
2           5.7              7.158146
3           8.9              8.550469
4           4.8              5.541333
5           7.1              6.690958
7           5.7              4.661277
14          7.6              7.963941
21          5.4              5.570249
23          7.0              7.704405
27          6.3              7.361437
```

Now let's check the R-squared in the test data set.

```
> cor(test$Satisfaction, test$Satisfaction.Predict)^2
[1] 0.6396429
```

Root mean square error of the model in the test dataset:

```
> rmse2 <- sqrt(mean((test$Satisfaction - pred)^2))
> rmse2
[1] 0.6618739
```

#	R-Squared	RMSE
Train	0.69	0.656633
Test	0.6396429	0.661874

The R squared is not varying much for the model both in test and train dataset and so is the RMSE; so we can infer that the model is valid and also not overfit.