Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Data Transformation and Predictive Modeling in Patient Satisfaction Analysis

Yogeshwaran R , Naresk K , Yuvaraj Pandian K
Panimalar Engineering college, Chennai

CHAPTER 5
IMPLEMENTATION

5.1 SAMPLE CODE

Data Transformation

Importing data set

```
data_NK <- read.table("PROG8435-24W-Final_train.txt", header = TRUE, sep = ",")
data_NK <- as.data.frame(data_NK)
```

Display the structure of the R object

```
str(data_NK)
```

```
## 'data.frame':    912 obs. of  9 variables:
##  $ Index           : int  1 2 4 5 ...
##  $ Age             : int  74 6 77 3 43 ...
##  $ Serverity       : int  52 4 50 5 ...
##  $ Surgical.Medical: int  1 1 0 1 ...
##  $ Anxiety         : num  7 4.6 4.2 6 6.6 4.3 5.3 5.6 6.6 2.2 ...
##  $ Type            : chr  "Neuro" "Plastic Surgery" "No Surgery" "No Surgery" ...
##  $ Marital         : chr  "M" "M" "M" "S" ...
##  $ FSA             : chr  "V6Y" "E3H" "S4L" "O7U" ...
##  $ Satisfaction    : int  55 4 53 7 ...
```

```
dim(data_NK)
```

```
## [1] 912   9
```

```
head(data_NK,5)
```

```
##   Index Age Serverity Surgical.Medical Anxiety              Type Marital FSA
## 1     1  74        52                1     7.0            Neuro       M V6Y
## 2     2  67        43                1     4.6 Plastic Surgery       M E3H
## 3     3  68        56                0     4.2       No Surgery       M S4L
## 4     4  44        45                0     6.0       No Surgery       S O7U
## 5     5  63        58                1     6.6       Orthopedic       M T5Y
##   Satisfaction
## 1           55
## 2           46
## 3           44
## 4           71
## 5           79
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Transform character variables (Type and Martial) to factor variable

```
data_NK$Type <- as.factor(data_NK$Type)
data_NK$Marital <-as.factor(data_NK$Marital)
```

Removing Index and FSA column from the dataset as it does not add any value

```
data_NK$Index <- NULL
data_NK$FSA <- NULL
```

Summary and Descriptive Statistics of the dataset after Transformation

```
#options(width = 90)
summary(data_NK)
```

```
##       Age            Serverity      Surgical.Medical     Anxiety
##  Min.   :-2.00   Min.   :19.00   Min.   :0.0000   Min.   :1.900
##  1st Qu.:37.00   1st Qu.:40.00   1st Qu.:0.0000   1st Qu.:3.700
##  Median :52.00   Median :46.00   Median :1.0000   Median :4.700
##  Mean   :51.36   Mean   :46.75   Mean   :0.5702   Mean   :4.796
##  3rd Qu.:65.00   3rd Qu.:54.00   3rd Qu.:1.0000   3rd Qu.:5.900
##  Max.   :79.00   Max.   :81.00   Max.   :1.0000   Max.   :7.800
##                Type       Marital   Satisfaction
##  Abdominal        :135   M:653   Min.   : 26.00
##  Cardiovascular : 97     S:259   1st Qu.: 59.00
##  Neuro            : 79           Median : 69.00
##  No Surgery       : 392          Mean   : 68.08
##  Orthopedic       : 80           3rd Qu.: 78.00
##  Plastic Surgery:129             Max.   :102.00
```

```
stat.desc(data_NK)
```

```
##                      Age      Serverity Surgical.Medical       Anxiety Type
## nbr.val      912.0000000   912.0000000    912.00000000 912.00000000   NA
## nbr.null       0.0000000     0.0000000    392.00000000   0.00000000   NA
## nbr.na         0.0000000     0.0000000      0.00000000   0.00000000   NA
## min           -2.0000000    19.0000000      0.00000000   1.90000000   NA
## max           79.0000000    81.0000000      1.00000000   7.80000000   NA
## range         81.0000000    62.0000000      1.00000000   5.90000000   NA
## sum        46838.0000000 42635.0000000    520.00000000 4374.20000000   NA
## median        52.0000000    46.0000000      1.00000000   4.70000000   NA
## mean          51.3574561    46.7489035      0.57017544   4.79627193   NA
## SE.mean        0.5355493     0.3266344      0.01640177   0.04534840   NA
## CI.mean.0.95   1.0510538     0.6410434      0.03218964   0.08899948   NA
## var          261.5735119    97.3013160      0.24534443   1.87550749   NA
## std.dev       16.1732344     9.8641429      0.49532255   1.36949169   NA
## coef.var       0.3149150     0.2110027      0.86871954   0.28553254   NA
##                Marital   Satisfaction
## nbr.val             NA 912.0000000
## nbr.null            NA   0.0000000
## nbr.na              NA   0.0000000
## min                 NA  26.0000000
```

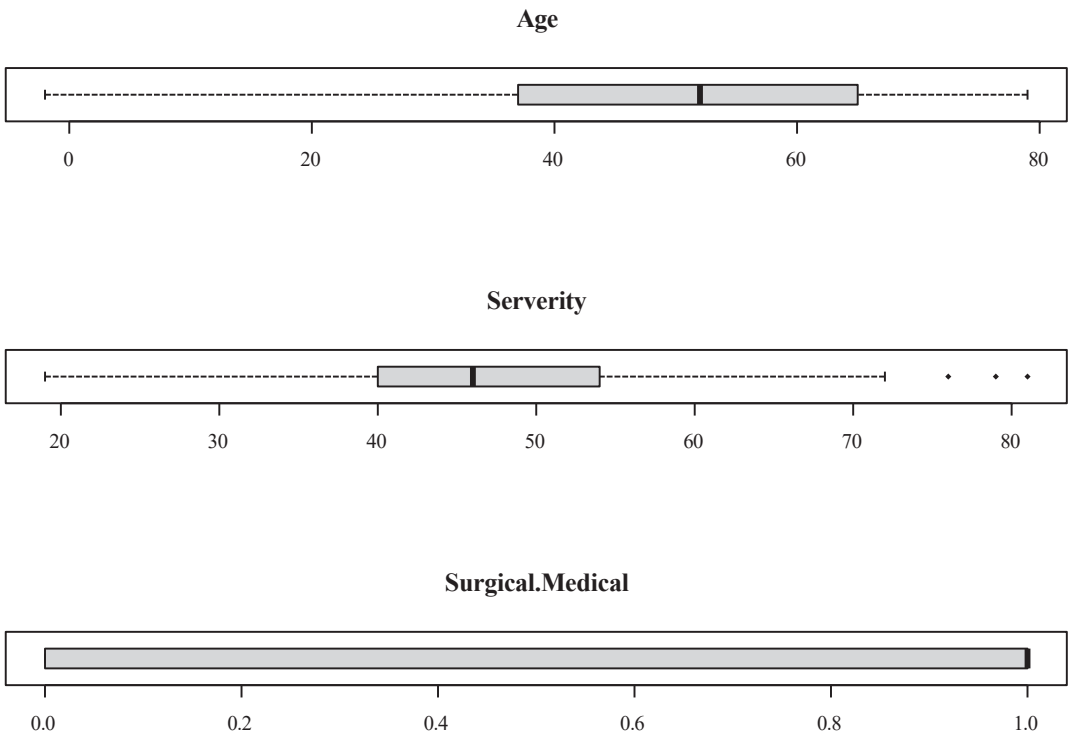
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

```
##  max             NA    102.0000000
##  range           NA     76.0000000
##  sum             NA  62085.0000000
##  median          NA     69.0000000
##  mean            NA     68.0756579
##  SE.mean         NA      0.4512237
##  CI.mean.0.95    NA      0.8855588
##  var             NA    185.6858173
##  std.dev         NA     13.6266583
##  coef.var        NA      0.2001693
```

Preliminary Analysis

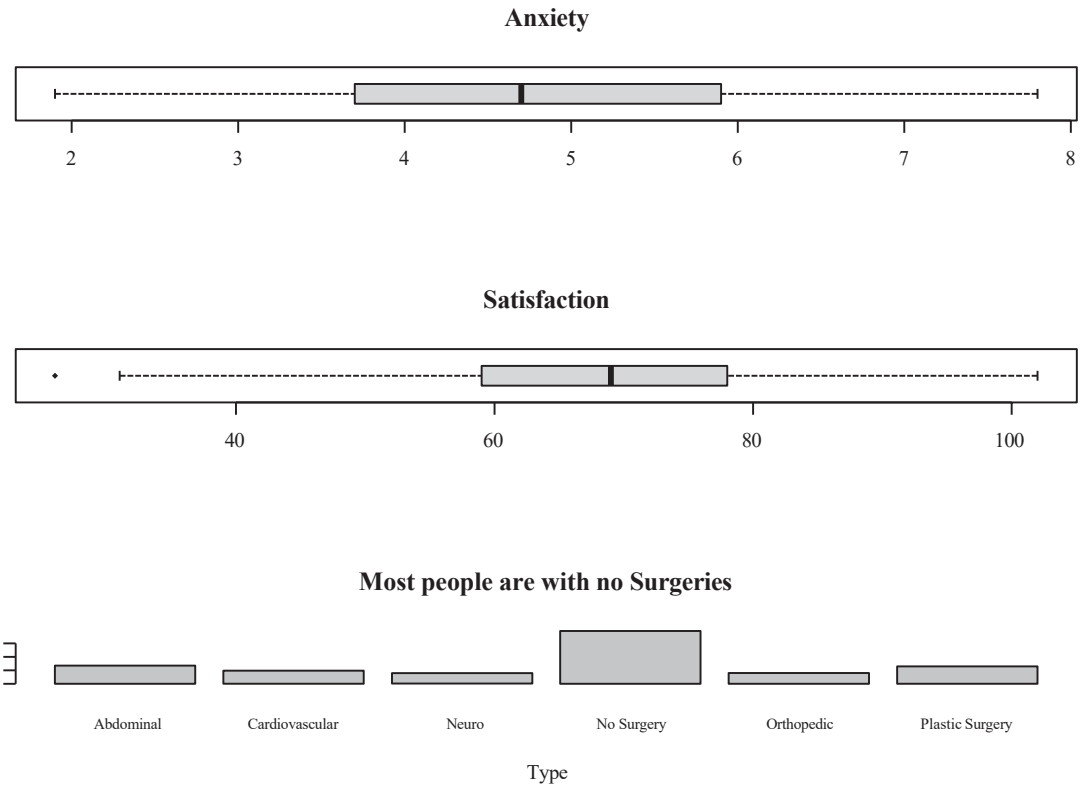
Creating boxplots to the dataframe to identidy outliers and gain insights

```
par(mfrow=c(3,1))
for (i in 1:ncol(data_NK))
{
  if (is.numeric(data_NK[,i])) {
       boxplot(data_NK[i], main=names(data_NK)[i],
                horizontal=TRUE, pch=18)
  }
}
```



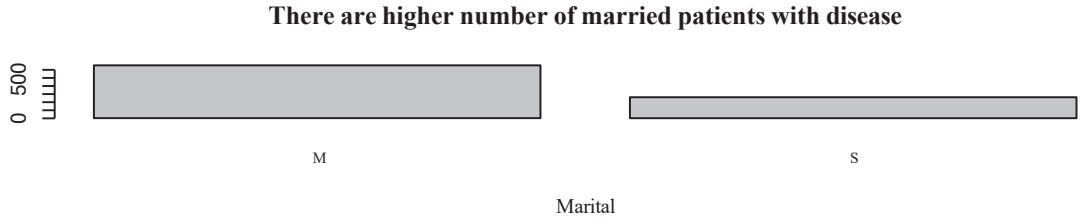
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

```
barplot(table(data_NK$Type), cex.names=.75,main='Most people are with no Surgeries',xlab='Type')
```




```
barplot(table(data_NK$Marital), cex.names=.75,main='There are higher number of married patients with di
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026



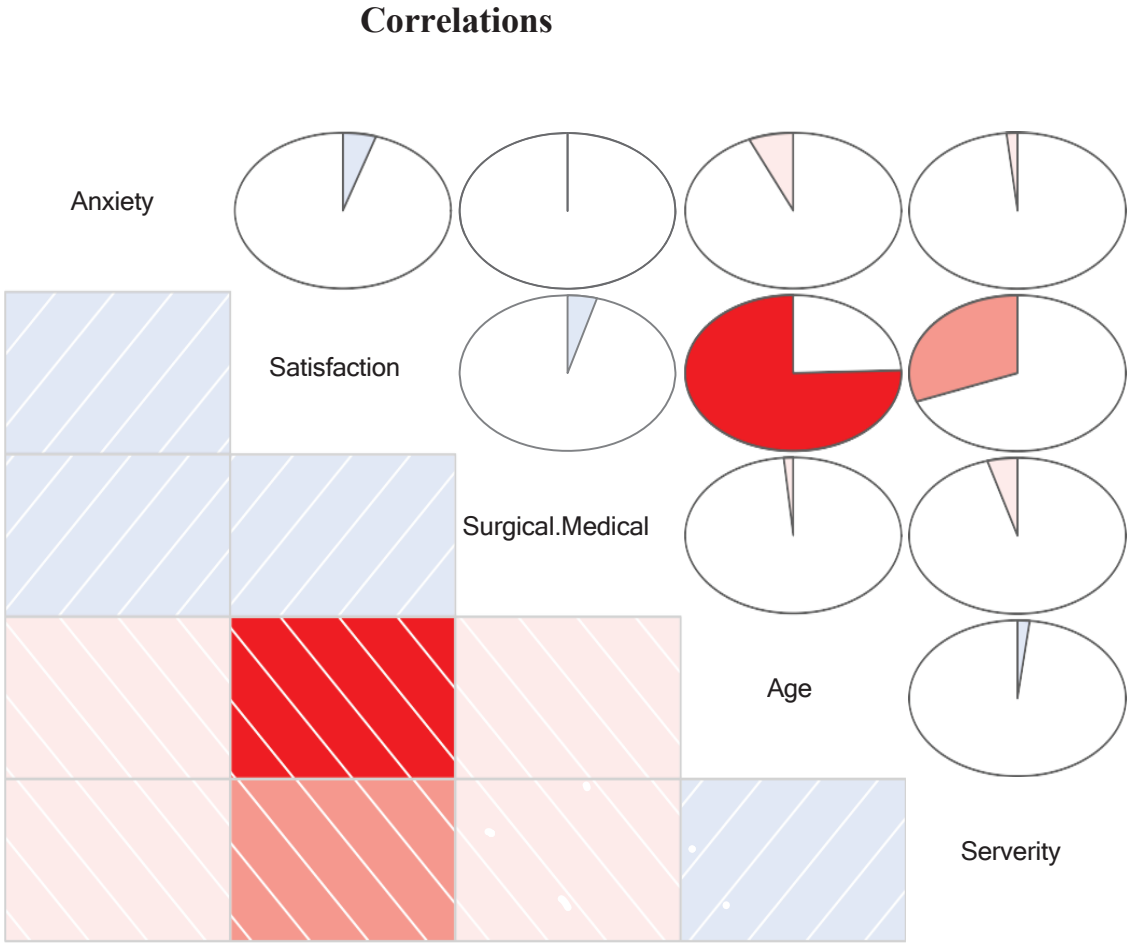
Removing records in dataset where Age is less than 0

There is a patient with age -2. It is not technically not possible

```
data_NK <- data_NK[!data_NK$Age < 0,]
```

Correlations on the dataset

```
corrgram(data_NK, order=TRUE, lower.panel=panel.shade,
         upper.panel=panel.pie, text.panel=panel.txt,
         main="Correlations")
```



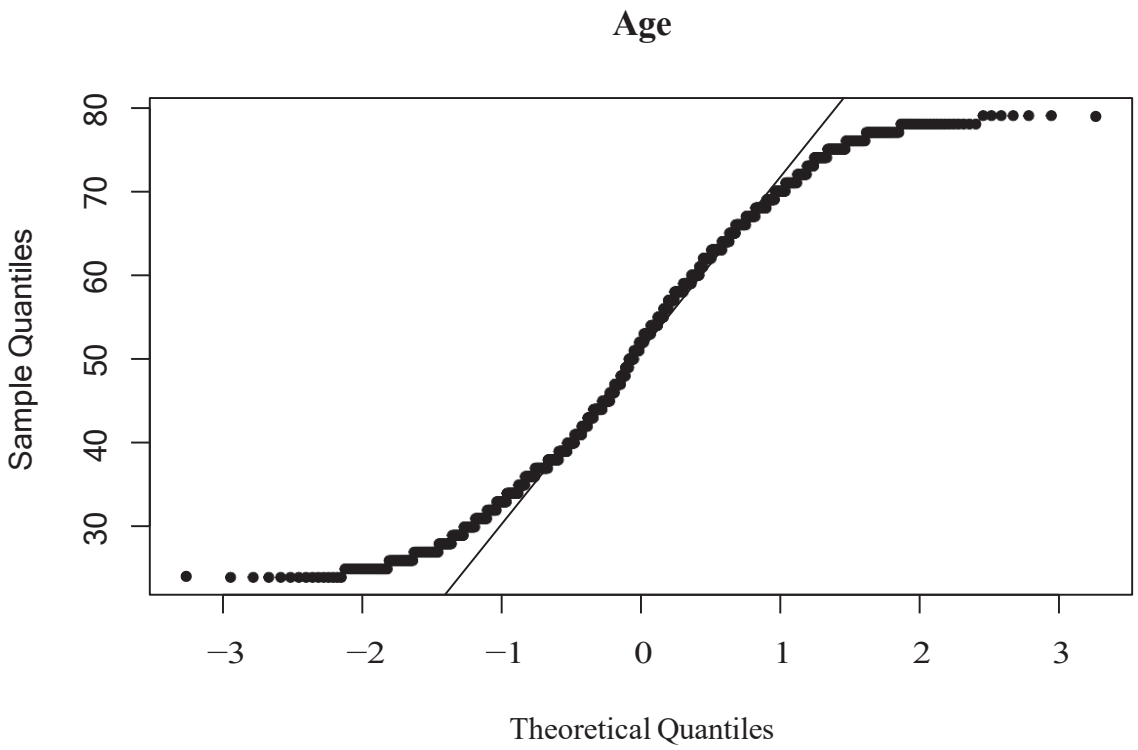
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Some observations on correlation

Satisfaction has a negative correlation with Age, shows that Satisfaction might decrease for older age people

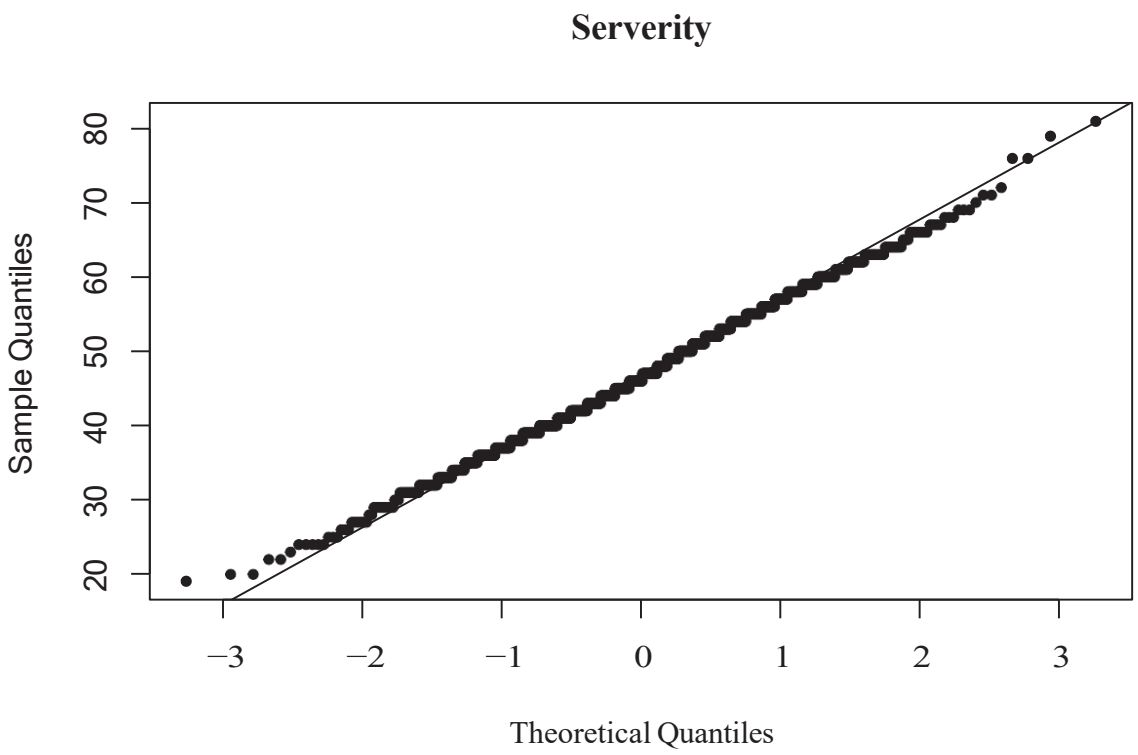
Distribution

```
qqnorm(data_NK$Age , main="Age", pch=20)
qqline(data_NK$Age)
```

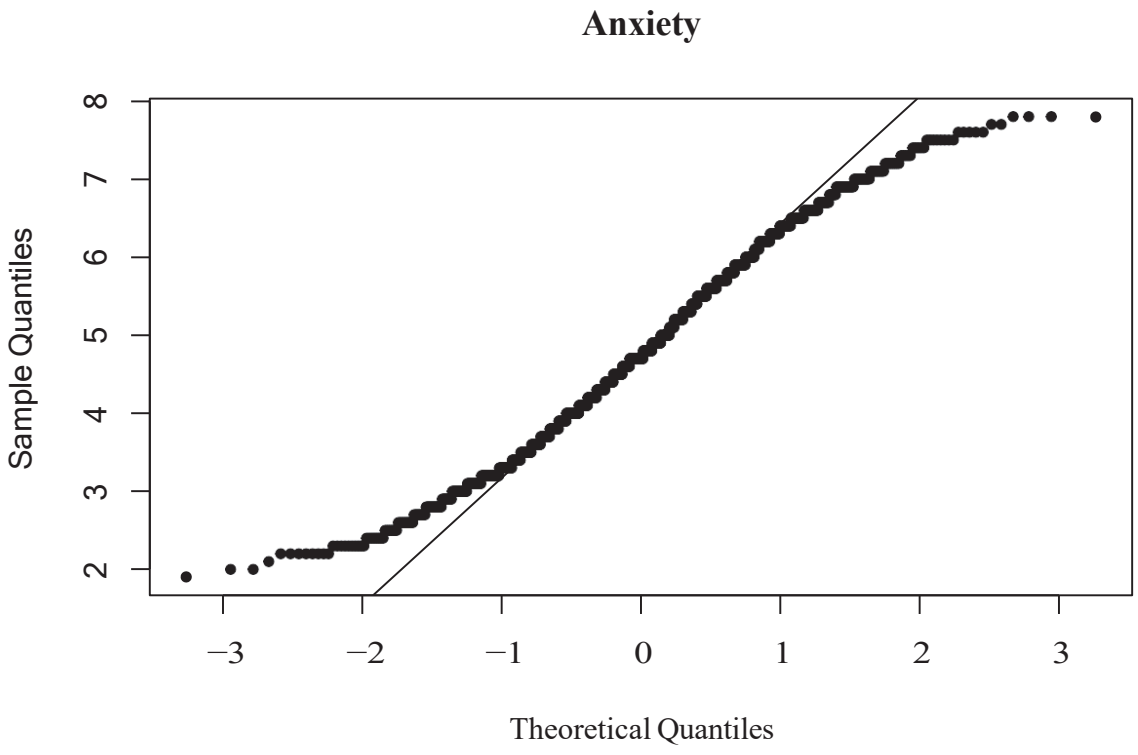



```
qqnorm(data_NK$Serverity , main="Serverity", pch=20)
qqline(data_NK$Serverity)
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

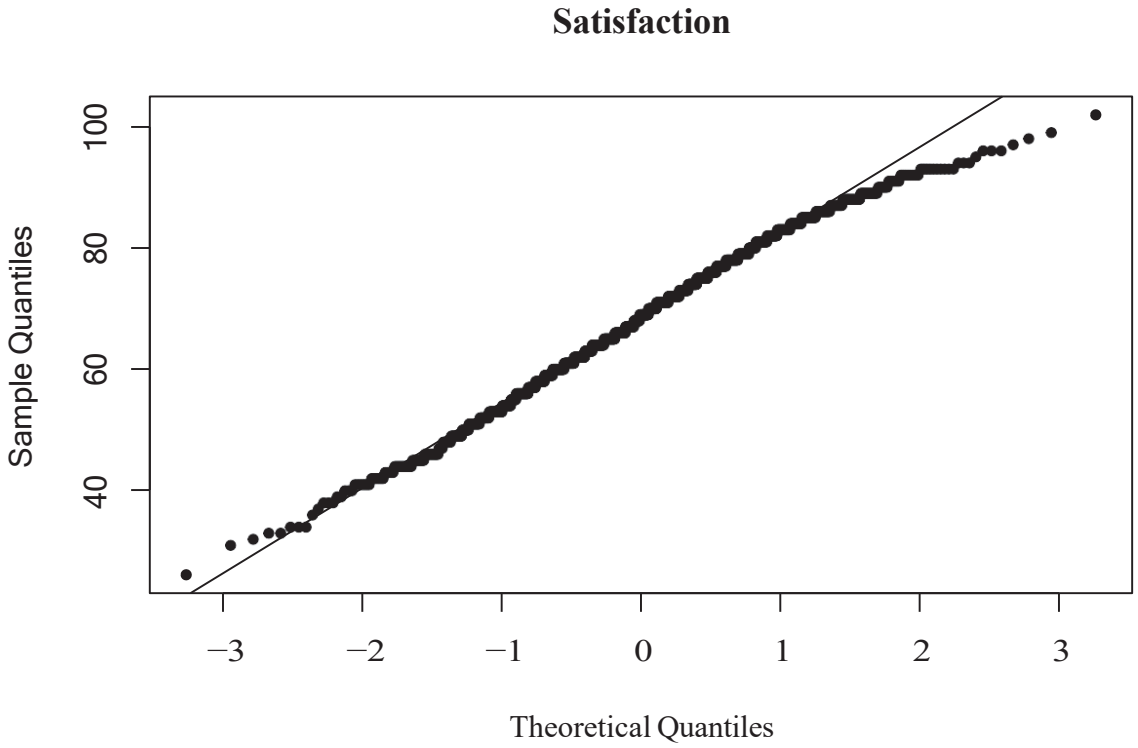



```
qqnorm(data_NK$Anxiety , main="Anxiety", pch=20)
qqline(data_NK$Anxiety )
```




```
qqnorm(data_NK$Satisfaction , main="Satisfaction", pch=20)
qqline(data_NK$Satisfaction)
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026



Some observations on distribution
Age is not evenly distributed Serverity is moderately distributed Anxiety is not evenly distributed
Satisfaction is evenly distributed, but there is drop for higher values

Model Development

Splitting the dataframe to Train and Test

```
# Setting the seed based on student number
set.seed(1464)

# Choosing sampling rate
tr <- 0.80
ts <- 0.20

# Finding the number of rows of data
n_row <- nrow(data_NK)

# Choose the rows for the training sample
training_rows <- sample(1:n_row, tr * n_row, replace = FALSE)

# Assign to the training sample
train <- subset(data_NK[training_rows,])
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

```
# Assign the balance to the test sample
test <- subset(data_NK[-c(training_rows),])
```

Summary of train dataset

```
summary(train)
```

```
##       Age            Serverity      Surgical.Medical    Anxiety
##  Min.   :24.0   Min.   :19.00   Min.   :0.0000   Min.   :2.000
##  1st Qu.:37.0   1st Qu.:40.00   1st Qu.:0.0000   1st Qu.:3.700
##  Median :52.0   Median :46.00   Median :1.0000   Median :4.700
##  Mean   :51.4   Mean   :46.71   Mean   :0.5673   Mean   :4.765
##  3rd Qu.:66.0   3rd Qu.:54.00   3rd Qu.:1.0000   3rd Qu.:5.825
##  Max.   :79.0   Max.   :81.00   Max.   :1.0000   Max.   :7.800
##                    Type      Marital   Satisfaction
##  Abdominal         :103   M:518    Min.   : 26.00
##  Cardiovascular    : 85   S:210    1st Qu.: 59.00
##  Neuro             : 64            Median : 68.00
##  No Surgery        :315            Mean   : 68.04
##  Orthopedic        : 63            3rd Qu.: 78.00
##  Plastic Surgery   : 98            Max.   :102.00
```

Correlations on train dataset

```
options(width = 100)
ht <- hetcor(train)      #from polycor library
round(ht$correlations,2)
```

```
##                    Age Serverity Surgical.Medical Anxiety  Type Marital Satisfaction
## Age               1.00      0.03            -0.01   -0.05  0.00   -0.09        -0.75
## Serverity         0.03      1.00            -0.03   -0.03 -0.01   -0.04        -0.32
## Surgical.Medical -0.01     -0.03             1.00   -0.01 -0.16   -0.07         0.04
## Anxiety          -0.05     -0.03            -0.01    1.00  0.02   -0.03         0.04
## Type              0.00     -0.01            -0.16    0.02  1.00    0.10        -0.11
## Marital          -0.09     -0.04            -0.07   -0.03  0.10    1.00         0.00
## Satisfaction     -0.75     -0.32             0.04    0.04 -0.11    0.00         1.00
```

Creating Full model

```
# Creating full model
full_model <- lm(Satisfaction ~ ., data = train)
# Summary of full model
summary(full_model)
```

```
##
## Call:
## lm(formula = Satisfaction ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.5785  -5.1477  -0.0803   4.9491  25.7570
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

```
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value   Pr(>|t|)
## (Intercept)            120.16332    2.06181  58.280    < 2e-16 ***
## Age                     -0.62968    0.01812 -34.758    < 2e-16 ***
## Serverity               -0.41132    0.02946 -13.960    < 2e-16 ***
## Surgical.Medical         2.78991    0.89549   3.116   0.001909 **
## Anxiety                 -0.11049    0.21469  -0.515   0.606943
## TypeCardiovascular      -1.40456    1.15306  -1.218   0.223579
## TypeNeuro                0.08295    1.25798   0.066   0.947446
## TypeNo Surgery               NA         NA      NA         NA
## TypeOrthopedic          -4.41496    1.26137  -3.500   0.000494 ***
## TypePlastic Surgery     -4.55395    1.11181  -4.096  0.0000468 ***
## MaritalS                -1.54240    0.64804  -2.380   0.017569 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.864 on 718 degrees of freedom
## Multiple R-squared:  0.6711, Adjusted R-squared:  0.667
## F-statistic: 162.8 on 9 and 718 DF,  p-value: < 2.2e-16
```

```
# Predicting Satisfaction for training dataset
pred_train_full <- predict(full_model, newdata = train)
# Calculating RMSE for training dataset
RMSE_train_full <- sqrt(mean((train$Satisfaction - pred_train_full)^2))
# Predicting Satisfaction for test dataset
pred_test_full <- predict(full_model, newdata = test)
# Calculating RMSE for test dataset
RMSE_test_full <- sqrt(mean((test$Satisfaction - pred_test_full)^2))
round(RMSE_train_full, 2)
```

```
## [1] 7.81
```

```
round(RMSE_test_full, 2)
```

```
## [1] 7.47
```

Creating Stepwise model

```
# Creating Stepwise model
stepwise_model <- step(full_model,trace=0)
# Summary Stepwise selection model
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = Satisfaction ~ Age + Serverity + Type +
Marital, ##data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.465   -5.168   -0.028    4.950   25.485
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

```
##
## Coefficients:
##                              Estimate  Std. Error  t value    Pr(>|t|)
## (Intercept)                122.39530     1.83431   66.725     < 2e-16 ***
## Age                         -0.62916     0.01808  -34.802     < 2e-16 ***
## Serverity                   -0.41089     0.02944  -13.958     < 2e-16 ***
## TypeCardiovascular          -1.42509     1.15178   -1.237    0.216381
## TypeNeuro                    0.04069     1.25466    0.032    0.974136
## TypeNo Surgery              -2.81047     0.89414   -3.143    0.001740 **
## TypeOrthopedic              -4.43181     1.26030   -3.516    0.000465 ***
## TypePlastic Surgery         -4.57105     1.11074   -4.115   0.0000431 ***
## MaritalS                    -1.53289     0.64745   -2.368    0.018168 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.86 on 719 degrees of freedom
## Multiple R-squared:  0.671,  Adjusted R-squared:  0.6673
## F-statistic: 183.3 on 8 and 719 DF,  p-value: < 2.2e-16
```

```
# Predicting Satisfaction for training dataset
pred_train_stepwise <- predict(stepwise_model, newdata = train)
# Calculating RMSE for training dataset
RMSE_train_stepwise <- sqrt(mean((train$Satisfaction - pred_train_stepwise)^2))
# Predicting Satisfaction for test dataset
pred_test_stepwise <- predict(stepwise_model, newdata = test)
# Calculating RMSE for test dataset
RMSE_test_stepwise <- sqrt(mean((test$Satisfaction - pred_test_stepwise)^2))
round(RMSE_train_stepwise, 2)
```

```
## [1] 7.81
```

```
round(RMSE_test_stepwise, 2)
```

```
## [1] 7.46
```

Creating backward selection model

```
# Creating backward model
backward_model <- step(full_model, direction = "backward",trace=0)
# Summary backward selection model
summary(backward_model)
```

```
##
## Call:
## lm(formula = Satisfaction ~ Age + Serverity + Type +
Marital, ##data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.465   -5.168   -0.028    4.950   25.485
##
## Coefficients:
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

```
##                          Estimate  Std. Error  t value    Pr(>|t|)
## (Intercept)             122.39530     1.83431   66.725    < 2e-16 ***
## Age                      -0.62916     0.01808  -34.802    < 2e-16 ***
## Serverity                -0.41089     0.02944  -13.958    < 2e-16 ***
## TypeCardiovascular       -1.42509     1.15178   -1.237    0.216381
## TypeNeuro                 0.04069     1.25466    0.032    0.974136
## TypeNo Surgery           -2.81047     0.89414   -3.143    0.001740 **
## TypeOrthopedic           -4.43181     1.26030   -3.516    0.000465 ***
## TypePlastic Surgery      -4.57105     1.11074   -4.115   0.0000431 ***
## MaritalS                 -1.53289     0.64745   -2.368    0.018168 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.86 on 719 degrees of freedom
## Multiple R-squared:  0.671,  Adjusted R-squared:  0.6673
## F-statistic: 183.3 on 8 and 719 DF,  p-value: < 2.2e-16
```

```r
# Predicting Satisfaction for training dataset
pred_train_backward <- predict(backward_model, newdata = train)
# Calculating RMSE for training dataset
RMSE_train_backward <- sqrt(mean((train$Satisfaction - pred_train_backward)^2))
# Predicting Satisfaction for test dataset
pred_test_backward <- predict(backward_model, newdata = test)
# Calculating RMSE for test dataset
RMSE_test_backward <- sqrt(mean((test$Satisfaction - pred_test_backward)^2))
round(RMSE_train_backward, 2)
```

```
## [1] 7.81
```

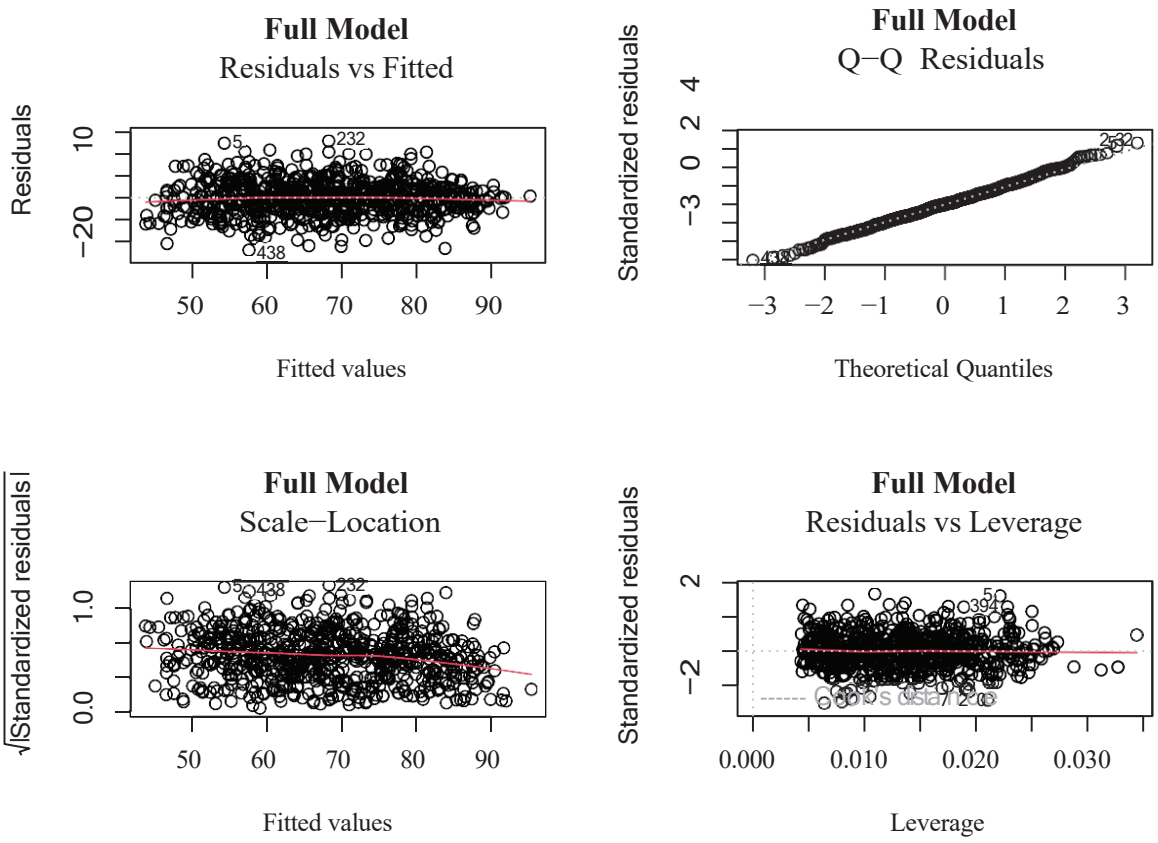
```r
round(RMSE_test_backward, 2)
```

```
## [1] 7.46
```

Graphical observation of models

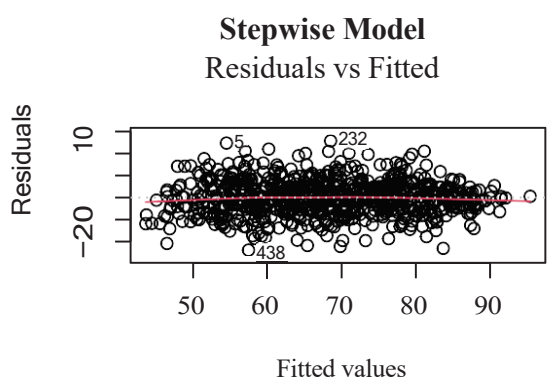
```r
par(mfrow = c(2, 2))
plot(full_model,main="Full Model")
```

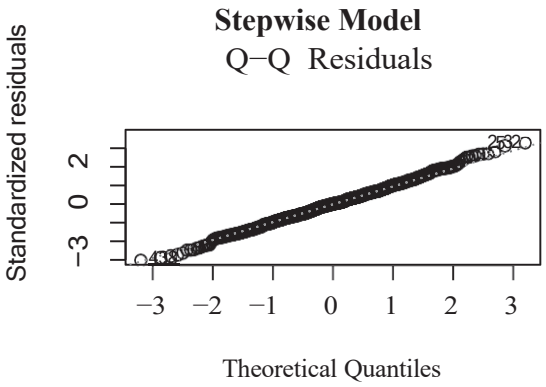
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

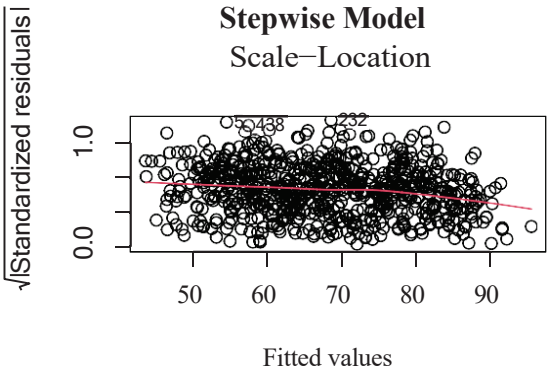


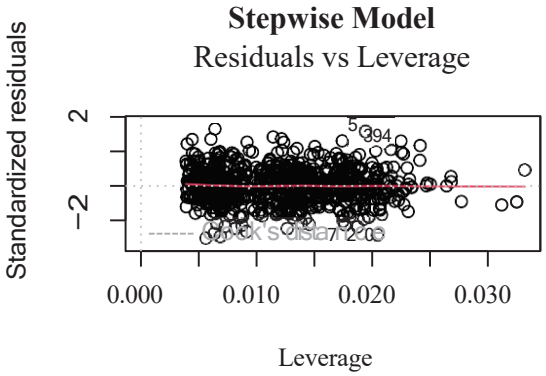

```
par(mfrow = c(1, 1))

par(mfrow = c(2, 2))
plot(stepwise_model,main="Stepwise   Model")
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026



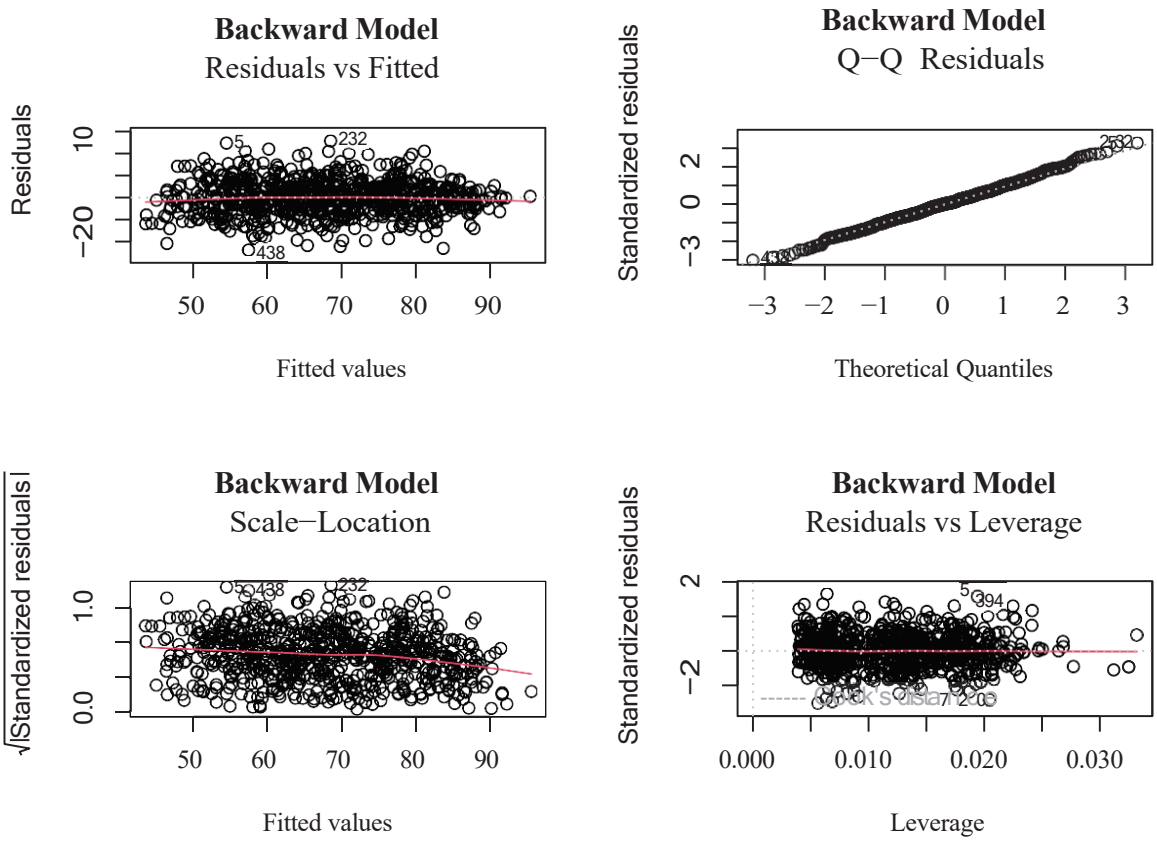







```r
par(mfrow = c(1, 1))

par(mfrow = c(2, 2))
plot(backward_model,main="Backward   Model")
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026




```
par(mfrow = c(1, 1))
```

The residuals are normally distributed in all three models
The cooks distance tells that leverage and influence is satisfiable

Comparison between full model, Stepwise model and Backward model

Tests	Full	Stepwise	Backward
F-Stat	Pass	Pass	Pass
R^2	0.671	0.671	0.671
Adj R^2	0.667	0.6673	0.6673
Res Er	7.864	7.86	7.86
T-P.val	Pass	Pass	Pass
RMSE Tr	7.81	7.81	7.81
RMSE Ts	7.47	7.46	7.46

From the comparison we can say all three model perform good
In Full model RMSE test is 7.47 which is .01 close to the training, So I choose Full model for my prediction

Model Prediction - Full Model

Preparing Test data for prediction

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

```
Final_data <- read.table("PROG8435-24W-Final_test.txt", header = TRUE, sep = ",")
```

Transform character variables (Type and Martial) to factor variable

```
Final_data$Type <- as.factor(Final_data$Type)
Final_data$Marital <-as.factor(Final_data$Marital)
```

Removing Index and FSA column from the dataset

```
Final_data$Index <- NULL
Final_data$FSA <- NULL
```

Summary of prediction file

```
str(Final_data)
```

```
## 'data.frame':    203 obs. of  6 variables:
##  $ Age             : int  72 59 76 58 64 68 73 59 52 71 ...
##  $ Serverity       : int  31 43 31 49 31 53 47 49 49 31 ...
##  $ Surgical.Medical: int  1 0 0 1 1 0 0 1 1 1 ...
##  $ Anxiety         : num  4.9 6.3 6.4 2.9 2.8 5.2 2.7 4.5 6.9 6.4 ...
##  $ Type            : Factor w/ 6 levels "Abdominal","Cardiovascular",..: 1 4 4 2 3
4 4 1 5 6 ...
##  $ Marital         : Factor w/ 2 levels "M","S": 2 2 1 2 2 1 1 1 1 2 ...
```

Using full model and predicting the Satisfaction

```
Predicted <- round(predict(full_model, newdata=Final_data),0)
head(Predicted)
```

```
##  1  2  3  4  5  6
## 63 63 59 63 68 55
```

Appending the Predicated value to the final dataset

```
test_final <- cbind(Final_data,Predicted)
head(test_final)
```

```
##   Age Serverity Surgical.Medical Anxiety          Type Marital Predicted
## 1  72        31                1     4.9     Abdominal       S        63
## 2  59        43                0     6.3    No Surgery       S        63
## 3  76        31                0     6.4    No Surgery       M        59
## 4  58        49                1     2.9 Cardiovascular       S        63
## 5  64        31                1     2.8         Neuro       S        68
## 6  68        53                0     5.2    No Surgery       M        55
```

Creating Final Dataset

```
write.csv(test_final,"PROG8435-24W-Final-NSK.txt")
```