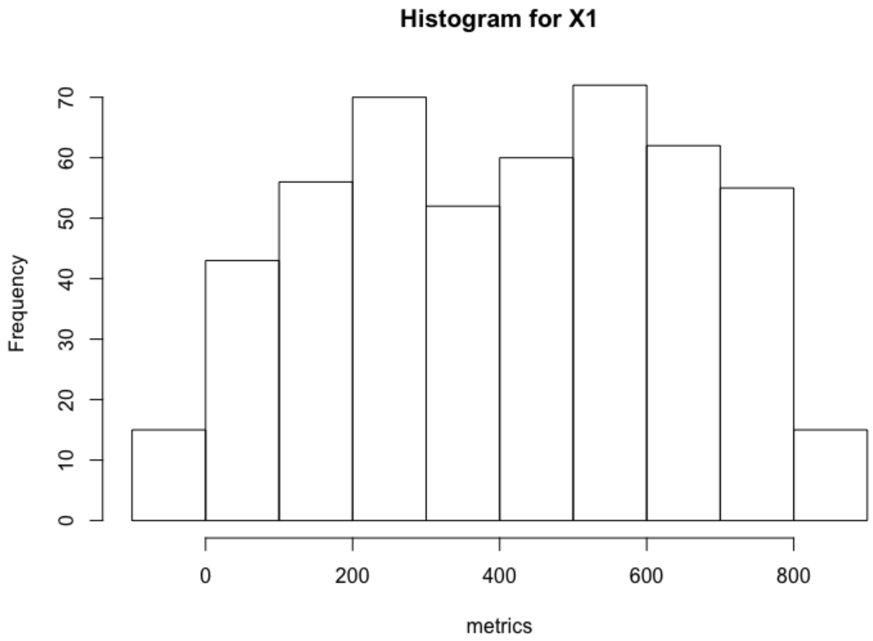
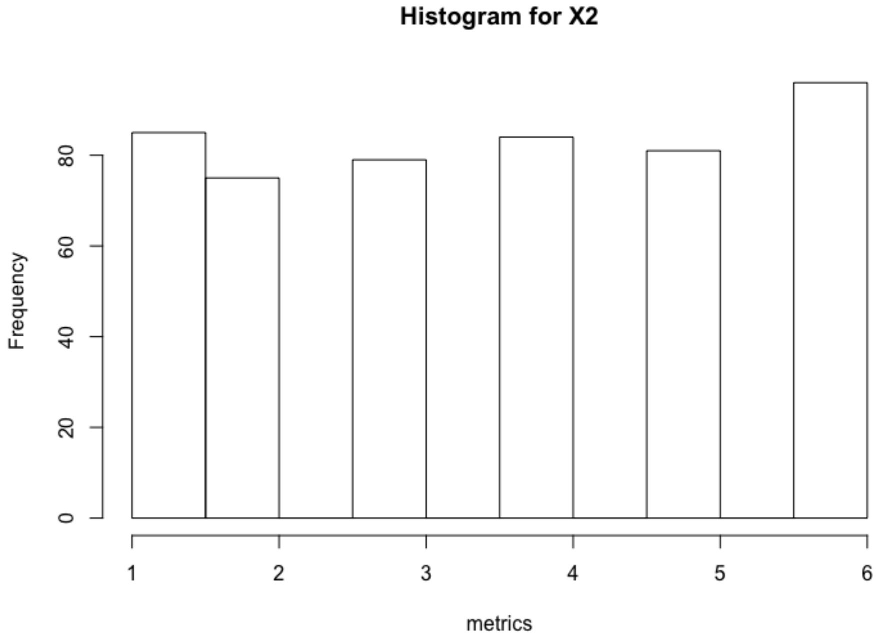
PROJECT – 2 REPORT
ADITYA BHARDWAJ
ABHARDW2

TASK 1: Basic statistics analysis

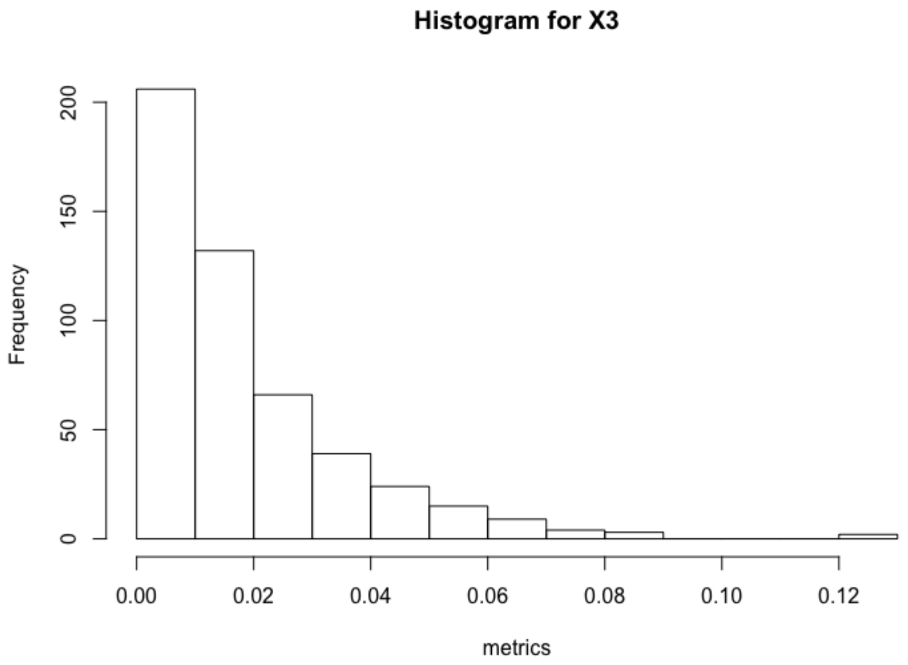
1.1 For each variable Xi, i.e. column in the data set corresponding to Xi, calculate the following:
Histogram, mean, variance.



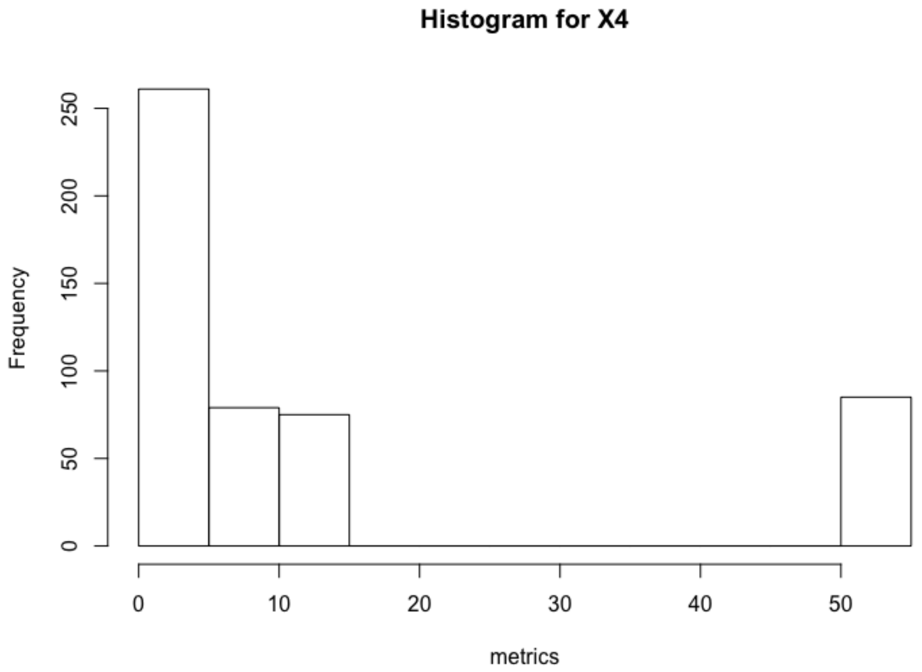
```
[1] "Mean of X1:411.491409903481"
[1] "Variance of X1:58093.3759206524"
```



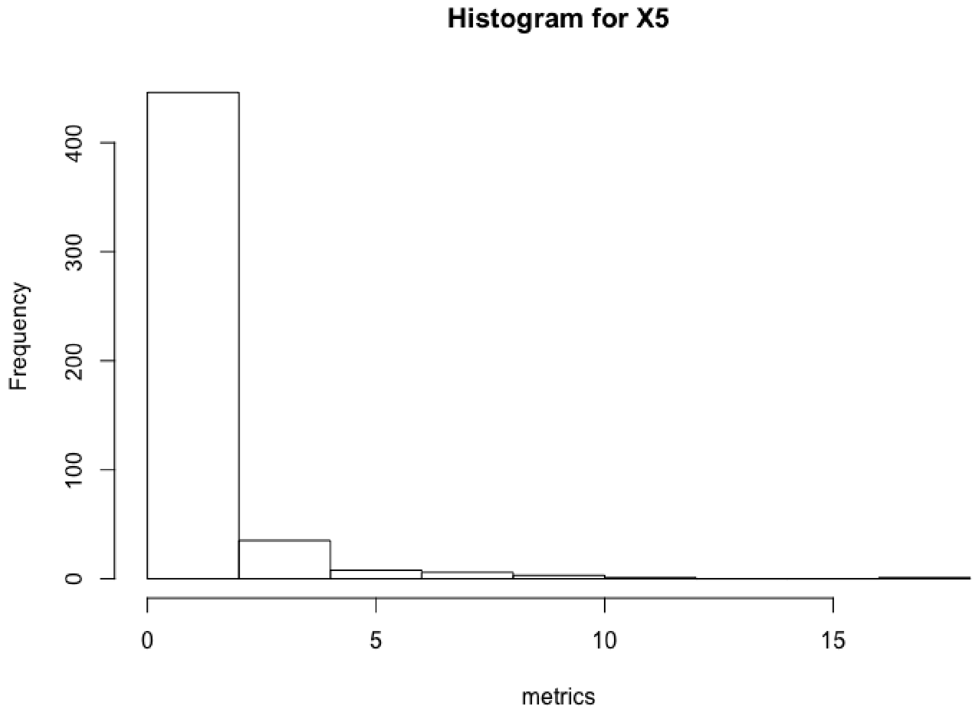
```
[1] "Mean of X2:3.578"
[1] "Variance of X2:3.04600801603206"
```



```
[1] "Mean of X3:0.0183302295289708"
[1] "Variance of X3:0.000325273875963521"
```

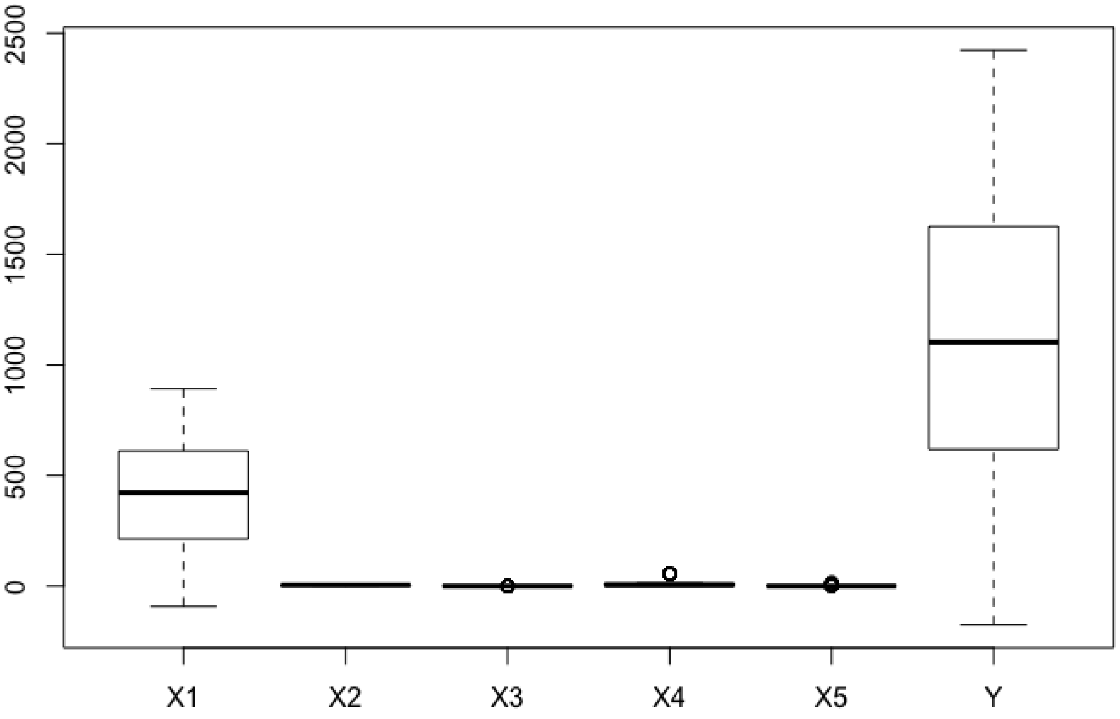


```
[1] "Mean of X4:13.35792"
[1] "Variance of X4:354.121167809218"
```

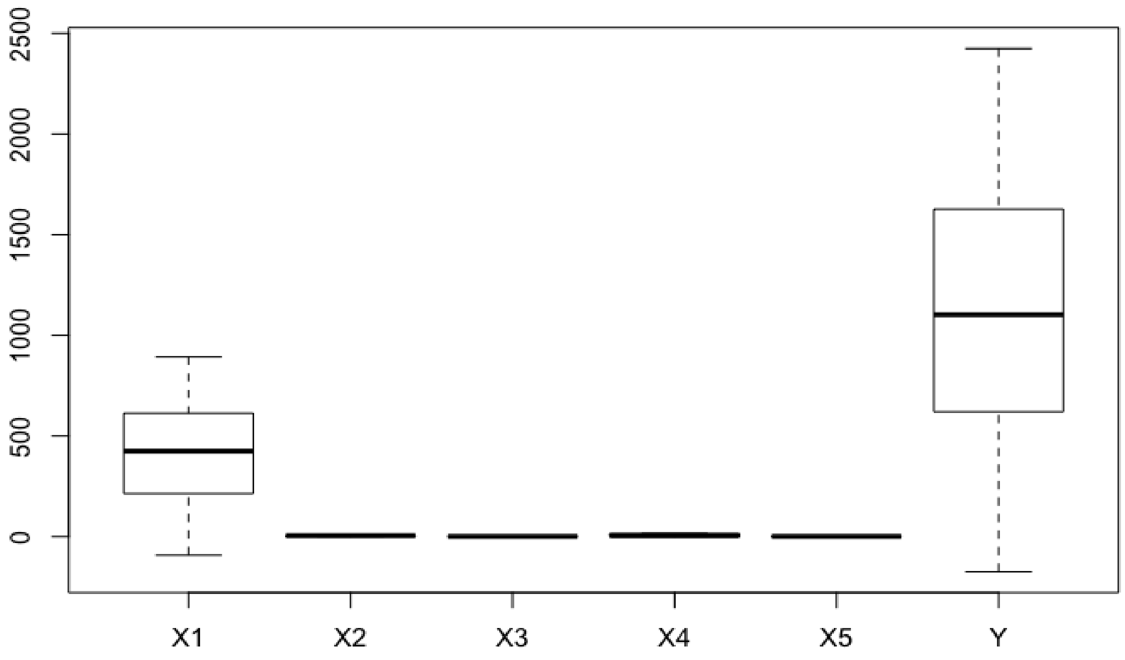


```
[1] "Mean of X5:0.732233670821059"
[1] "Variance of X5:2.50686413261778"
```

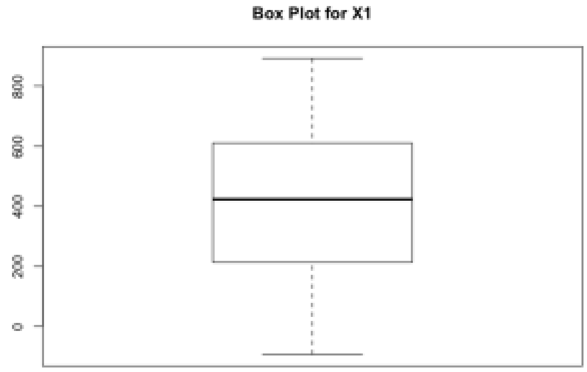
1.2 Use box plot or any other function to remove outliers (do not overdo it!), or you can do that during the model building phase (tasks 2 and 3)

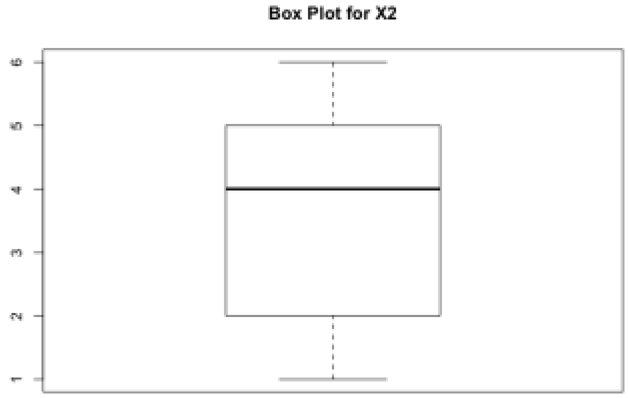


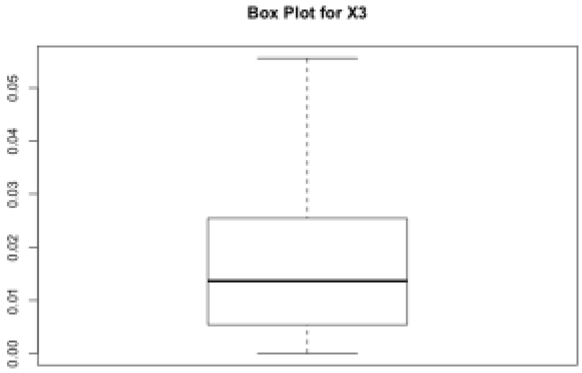
The Box Plot above shows the strong correlation between X1 and Y. Also, the dots at X3, X4 and X5 are the outliers which needs to be removed from further calculations.
After removing outliers, we get box plot as:

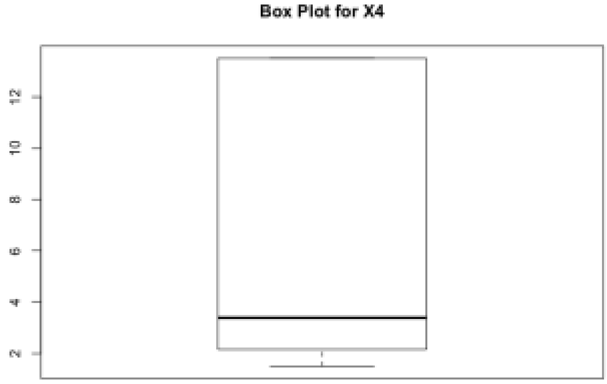


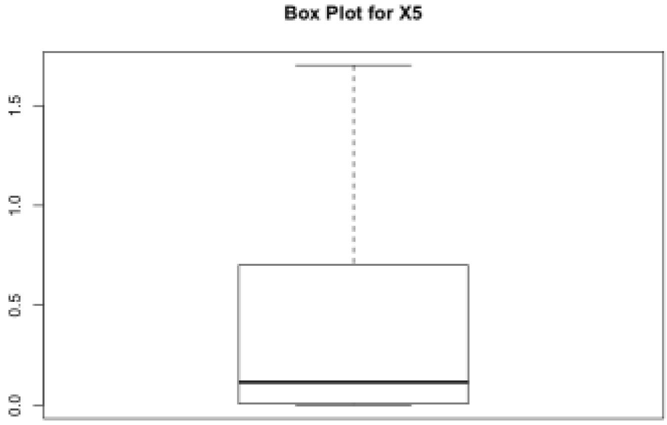
Below are the Box Plots of all the variables without outliers:



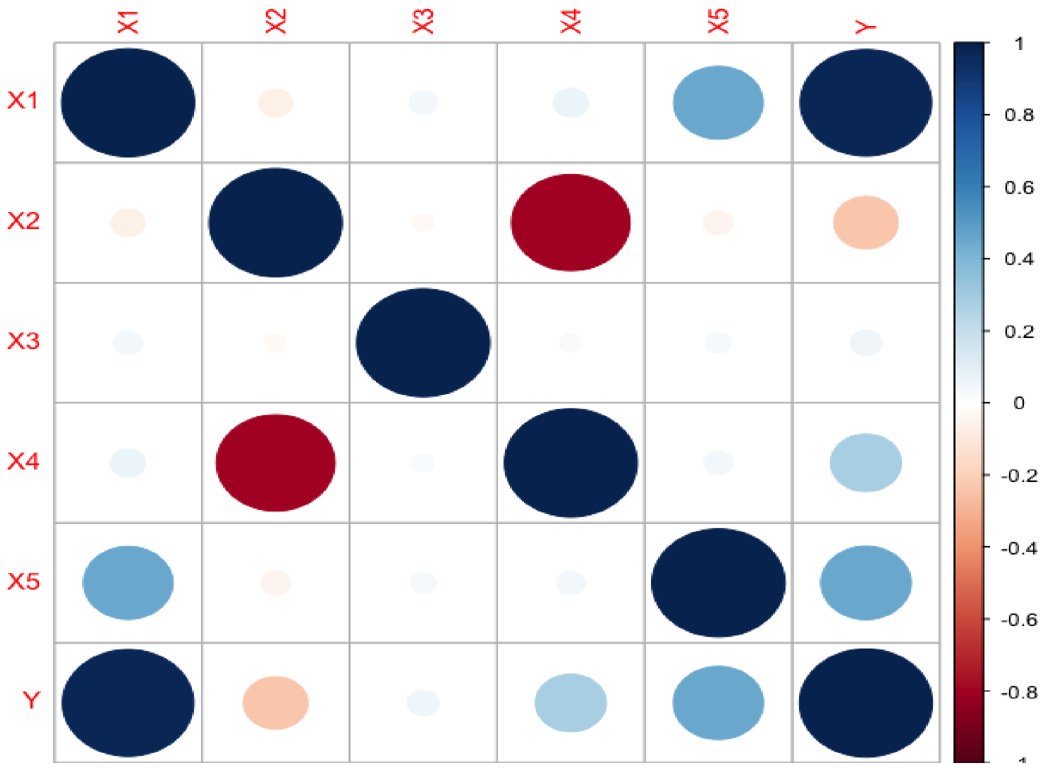








1.2 Calculate the correlation matrix Σ among all variables, i.e., Y, X1, X2, X3, X4 and X5. Draw conclusions related to possible dependencies among these variables.



Conclusions:

From Above Matrix, it is evident that there is strong correlation between X1 and Y.
The least correlation can be observed between X4 and Y which means they are not correlated at all.

1.3 Comment on the results.

With the help of Correlation Matrix, following comments can be made on the results:

X1	1. X1 have very low positive correlation with X3, X4. 2. X1 is not correlated to X2. 3. X1 is moderately correlated to X5. 4. X1 have strong positive correlation with Y.
X2	1. X2 is not correlated to X1, X3, X4, X5 and Y. i.e., All other variables.
X3	1. X3 is slightly correlated to X1, X4, X5 and Y. 2. X3 is not correlated to X2.
X4	1. X4 is slightly correlated to X1 and X5. 2. X4 is low moderately correlated to Y.
X5	1. X5 is slightly correlated to X3 and X4. 2. X5 is not correlated to X2. 3. X5 is low moderately correlated to X5. 4. X5 is moderately correlated to Y.

TASK 2: Linear regression

Function Call in R:

```
Call:
lm(formula = dataset$Y ~ dataset$X1, data = dataset)

Residuals:
   Min      1Q  Median      3Q     Max
-168.18  -83.90  -54.29    1.41  367.48

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.82016   12.11883   6.751 4.09e-11 ***
dataset$X1   2.48065    0.02542  97.590  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 136.9 on 498 degrees of freedom
Multiple R-squared:  0.9503,	Adjusted R-squared:  0.9502
F-statistic:  9524 on 1 and 498 DF,  p-value: < 2.2e-16
```

2.1 Determine the values for a0, a1, and s2.

a_0 = 81.82016

a_1 = 2.48065

s^2 = 136.7215

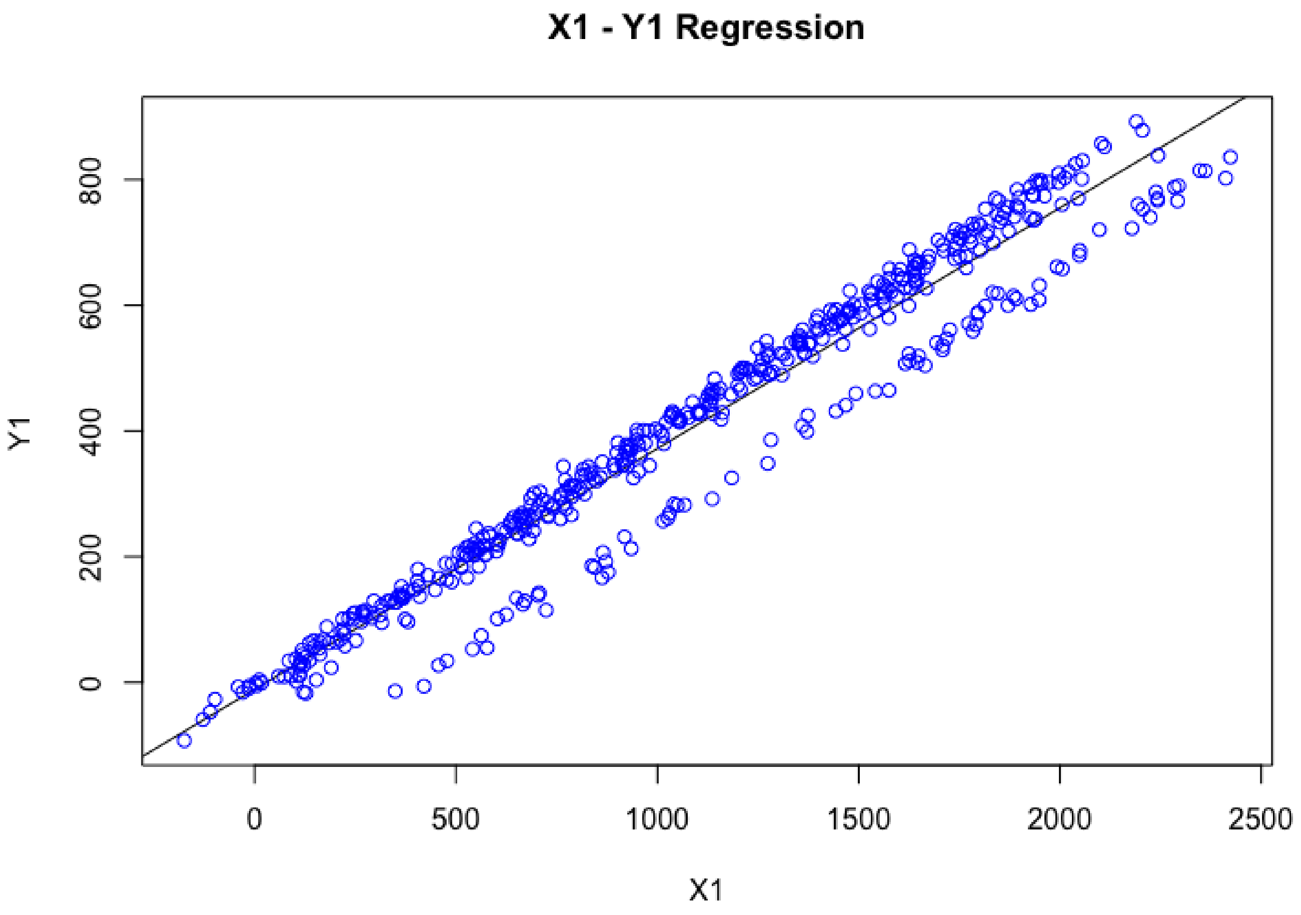
2.2 Check the p-values, R2, F value to determine if the regression coefficients are meaningful.

p-values: Since the p value for both the coefficients is ~ 0, the null hypothesis that the coefficients are zero is rejected. Thus, both the coefficients a_0 and a_1 are meaningful and can be used.

R^2: R^2 is not very high, hence, we can enhance this model further for a better fit.

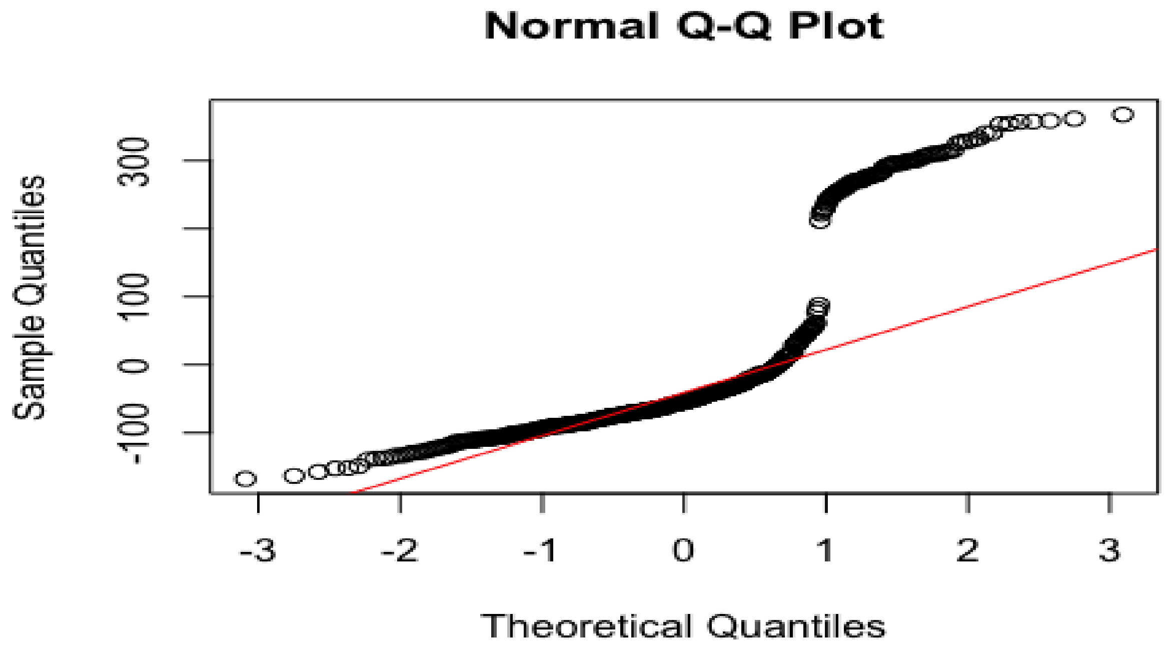
F-Value:
Since the model have only one parameter X1, the degree of freedom is 1. The F-statistic value is good and also the p-value ~ 0, which implies this model is a good fit.

2.3 Plot the regression line against the data



2.4 Do a residuals analysis:

QQ Plot for the regression:

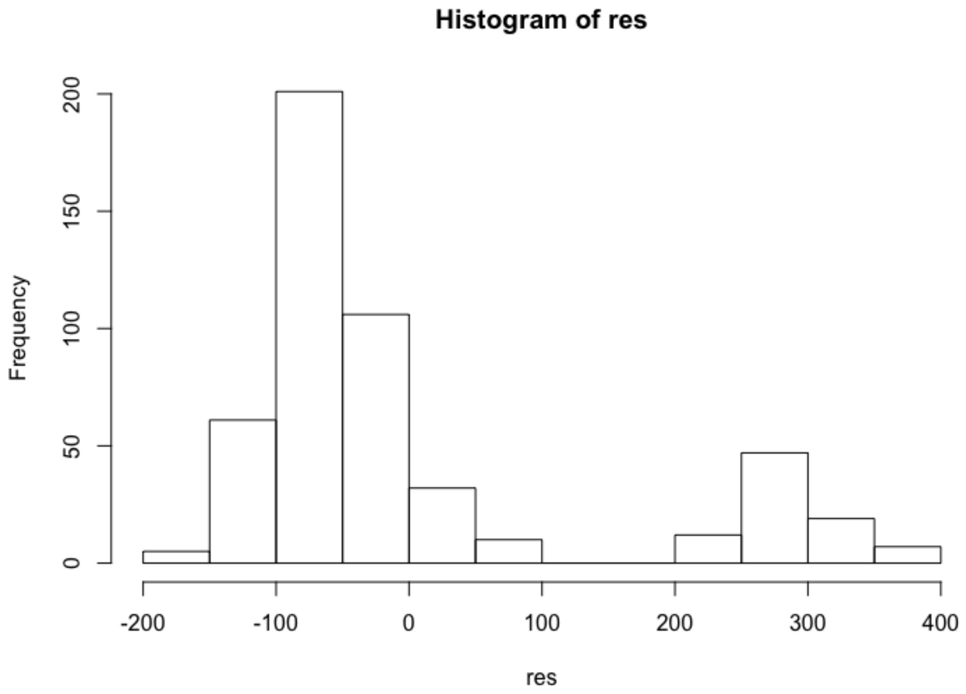


As the QQ Plot is not a straight line, the residuals are not distributed normally which means the data is discrete.

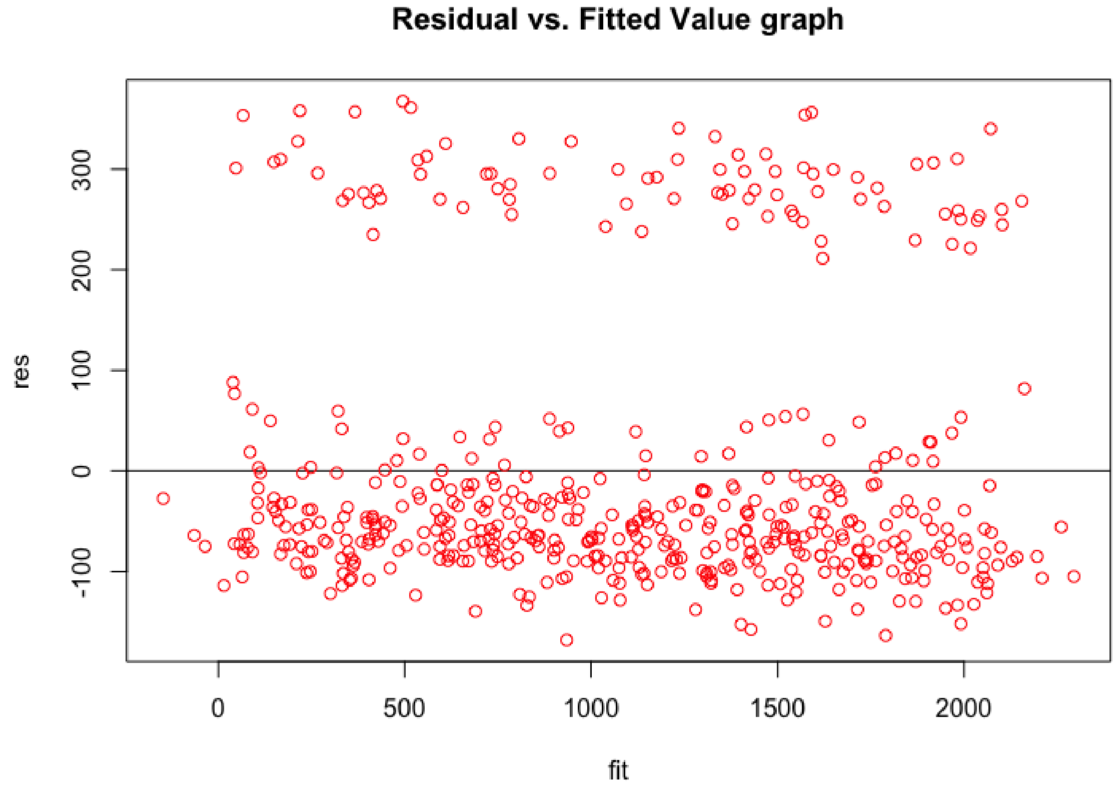
```
        Pearson's Chi-squared test

data:  tbl
X-squared = 249500, df = 249000, p-value = 0.2396
```

Residual Histogram:



Residual Scatter Plot:



2.7 Use a higher-order polynomial regression, i.e., Y = a0 + a1X + a2X2 + ε, to see if it gives better results.

R Function Call:

```
> polyReg.fun(prob)

Call:
lm(formula = dataset$Y ~ dataset$X1 + dataset$X1S, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-165.17  -83.38  -54.05    0.10  367.61

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.896e+01  1.690e+01   5.263 2.11e-07 ***
dataset$X1  2.426e+00  9.416e-02  25.762  < 2e-16 ***
dataset$X1S 6.807e-05  1.123e-04   0.606    0.545
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 136.9 on 497 degrees of freedom
Multiple R-squared:  0.9503,	Adjusted R-squared:  0.9501
F-statistic:  4756 on 2 and 497 DF,  p-value: < 2.2e-16
```

a_0 = 8.896e+01

a_1 = 2.426e+00

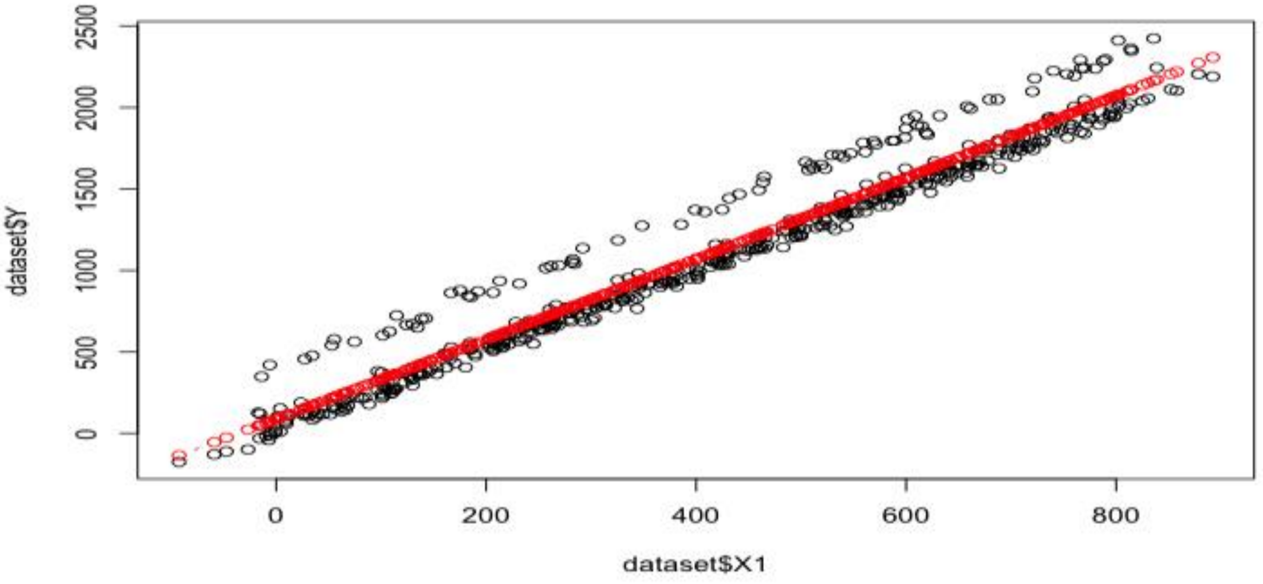
a_2 = 6.807e-05

σ^2 = 136.671

p values: Since the p value for a_0 and a_1 is < 0.05, the null hypothesis that the coefficients are zero is rejected. Thus, both the coefficients a_0 and a_1 are meaningful and can be used. However, for a_2, the value of p is very high and thus a_2 is not helpful in the model.

R^2: Since R^2 is similar to the R^2 for simple linear regression model, this model is not improving significantly. Both the models perform similar.

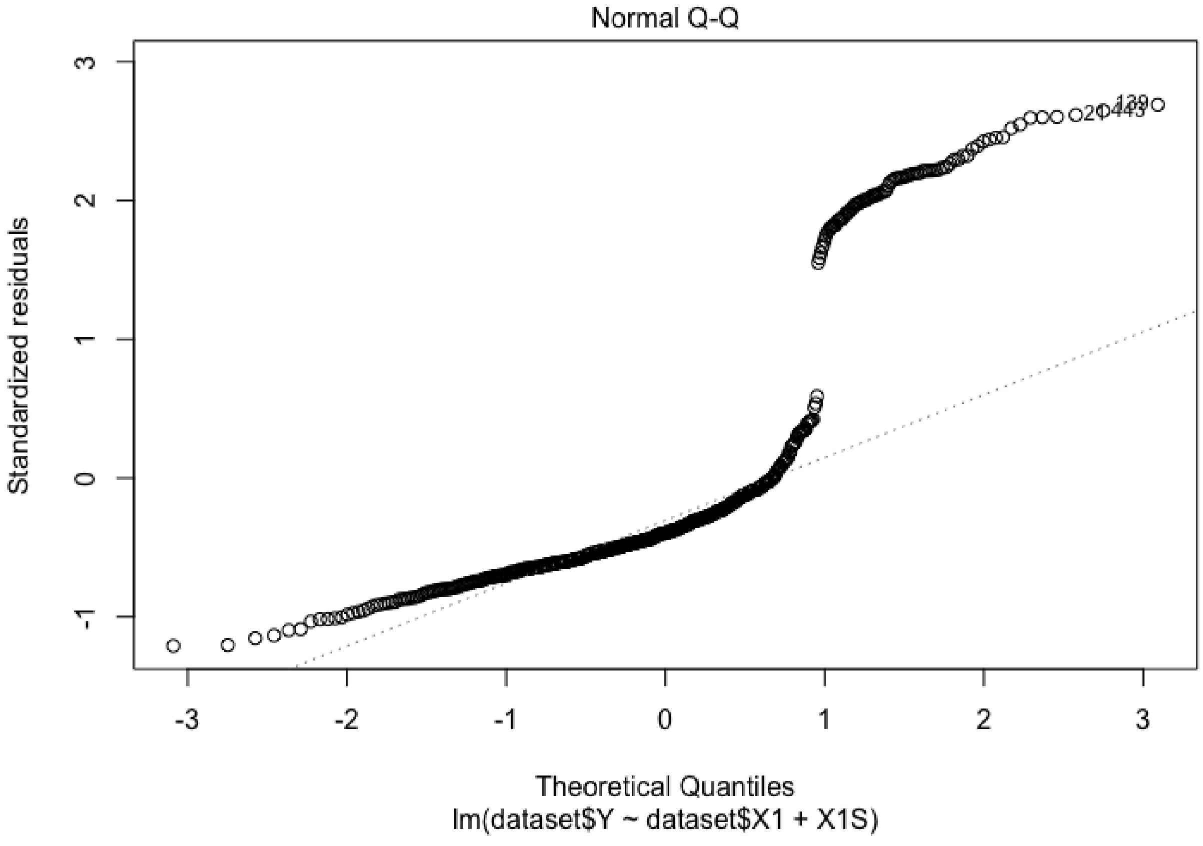
F-Statistic: The F-statistic of polynomial model is similar to the linear regression model, that tells us that the linear regression model is a better fit.
Hence, we can conclude that both the linear regression model and multiple regression model are very similar.



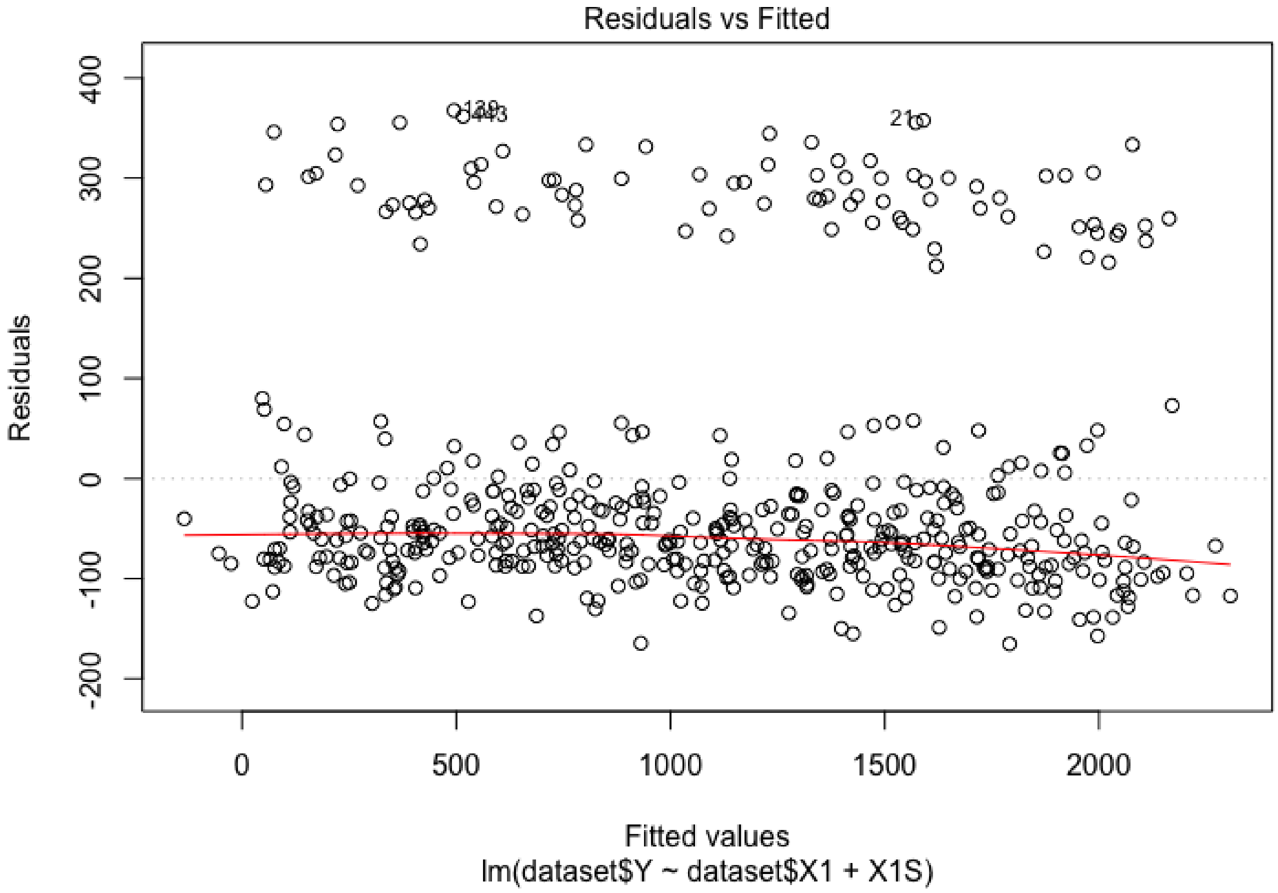
Pearson's Chi-squared test

```
data:  tbl
X-squared = 249500, df = 249000, p-value = 0.2396
```

QQ Plot for the Polynomial Regression Model:



Residual Scatter Plot:
The plot show the regression line closer to the fitted model which shows the Polyregression model is better but other factors show little difference in the regression line. Hence, calculating extra variable doesn't make much sense.



2.8 Comment on your results in a couple of paragraphs.

In the linear regression model, the variable X_1 is significant. The R2 value suggests that our model is good but not good enough. It can be further tweaked to achieve higher R^2 values (closer to 1).

In the polynomial regression model, the variable X_1 is significant but X_1^2 is not significant. So, the X_1^2 component does not contribute much to the model. R^2 values are similar to the linear regression model. This suggests that there is not much of a gain in training a polynomial regression model.

TASK 3: Multivariate regression

Function call in R:

```
> multivariate.fun(prob)

Call:
lm(formula = dataset$Y ~ dataset$X1 + dataset$X2 + dataset$X3 +
    dataset$X4 + dataset$X5, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-86.954 -19.249  -1.436  20.711  79.024

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.492e+00  6.228e+00   0.400   0.6892
dataset$X1  2.419e+00  5.942e-03 407.090   <2e-16 ***
dataset$X2  3.450e-01  1.198e+00   0.288   0.7735
dataset$X3  1.522e+02  7.081e+01   2.149   0.0321 *
dataset$X4  7.111e+00  1.112e-01  63.967   <2e-16 ***
dataset$X5  7.814e+00  9.031e-01   8.652   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.49 on 494 degrees of freedom
Multiple R-squared:  0.9979,	Adjusted R-squared:  0.9978
F-statistic: 4.617e+04 on 5 and 494 DF,  p-value: < 2.2e-16
```

3.1 Carry out a multiple regression on all the independent variables, and determine the values for all the coefficients, and σ2.

a_0 = 2.492e+00

a_1 = 2.419e+00

a_2 = 3.450e-01

a_3 = 1.522e+02

a_4 = 7.111e+00

a_5 = 7.814e+00

σ^2 = 28.34244

3.2 Based on the p-values, R2, F value, and correlation matrix Σ, identify which independent **variables need to be left out (if any) and go back to step 3.1.**

We don't need to train model again by leaving X2 and X3 as they are not much significant in the regression.

After removing X2 and X3, and retraining, we get:

```
> multivariate_new.fun(prob)
```

```
Call:
lm(formula = dataset$Y ~ dataset$X1 + dataset$X4 + dataset$X5,
   data = dataset)

Residuals:
   Min      1Q  Median      3Q     Max
-88.000 -18.554  -1.425  20.480  78.481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.623814   2.679647   2.472   0.0138 *
dataset$X1  2.419315   0.005955 406.269   <2e-16 ***
dataset$X4  7.088840   0.068118 104.067   <2e-16 ***
dataset$X5  7.843066   0.905314   8.663   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.56 on 496 degrees of freedom
Multiple R-squared:  0.9978,	Adjusted R-squared:  0.9978
F-statistic: 7.653e+04 on 3 and 496 DF,  p-value: < 2.2e-16
```

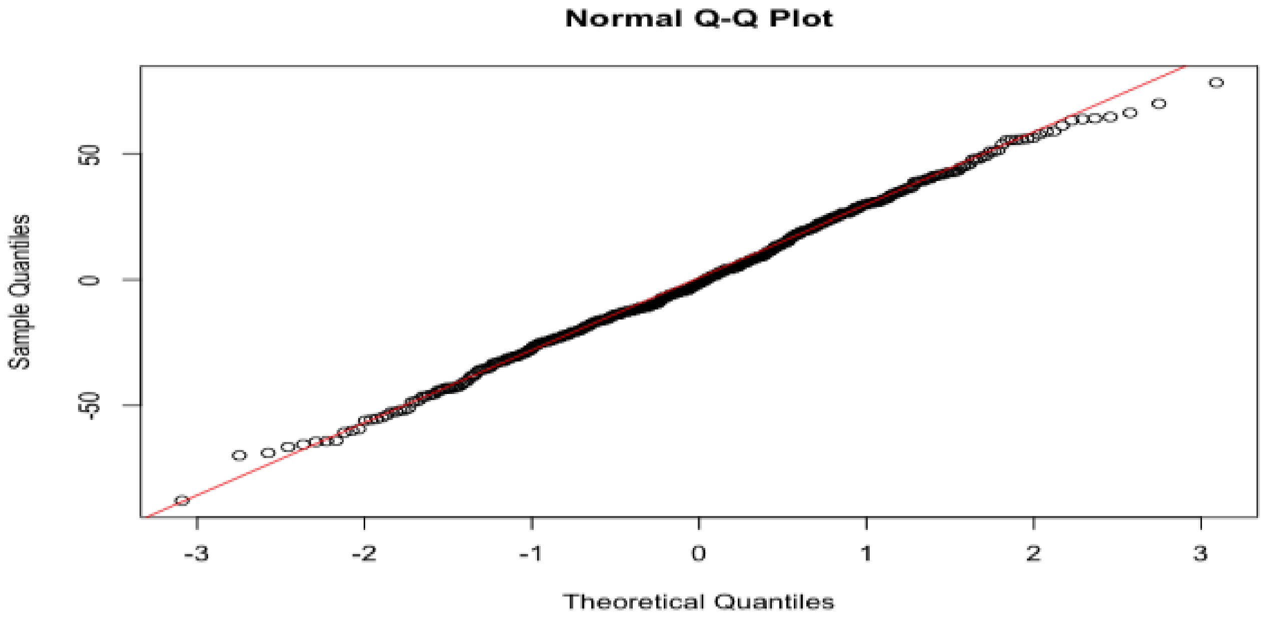
On observation:

- All coefficients are significant when we look at their p-values.
- R^2 is same as the previous model.
- F-statistic have increased in this model.

Model is similar to the regression model obtained above but this model would be preferable over the last one.

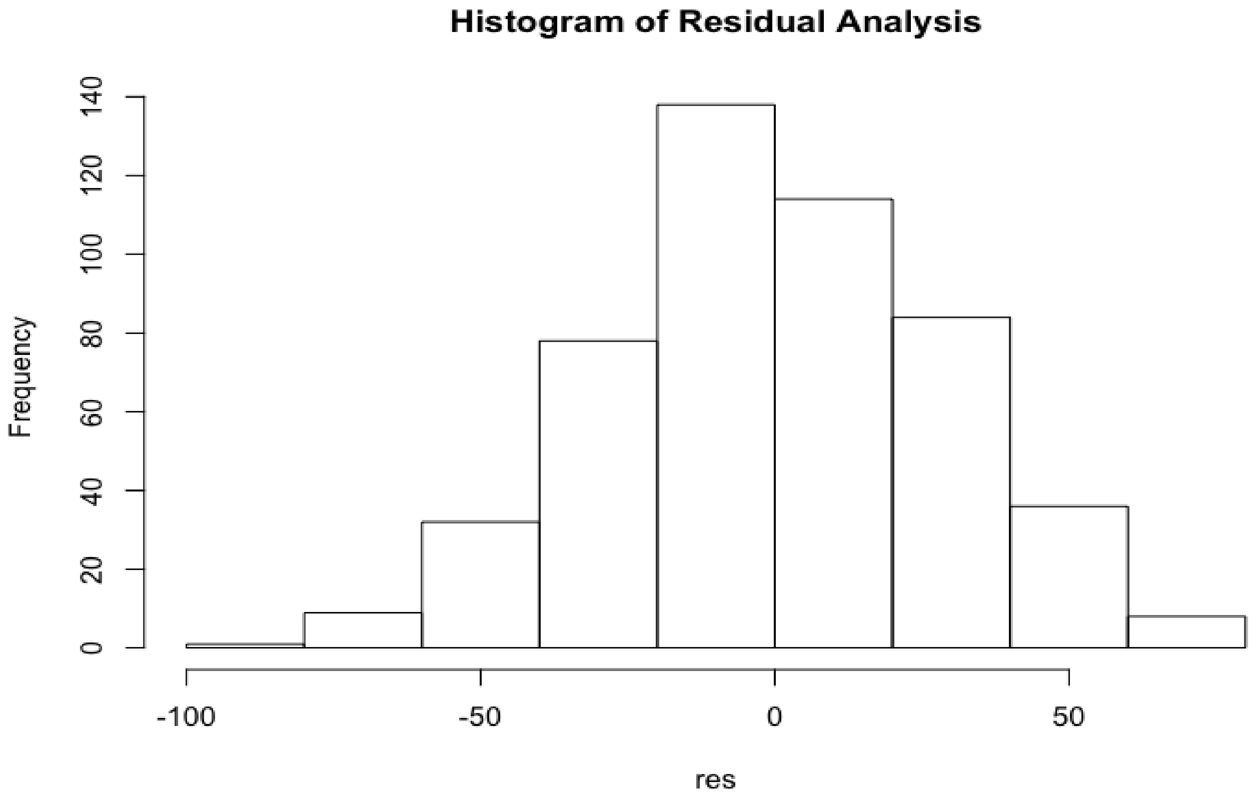
3.3 Residual Analysis for second multiple regression model

QQ Plot:

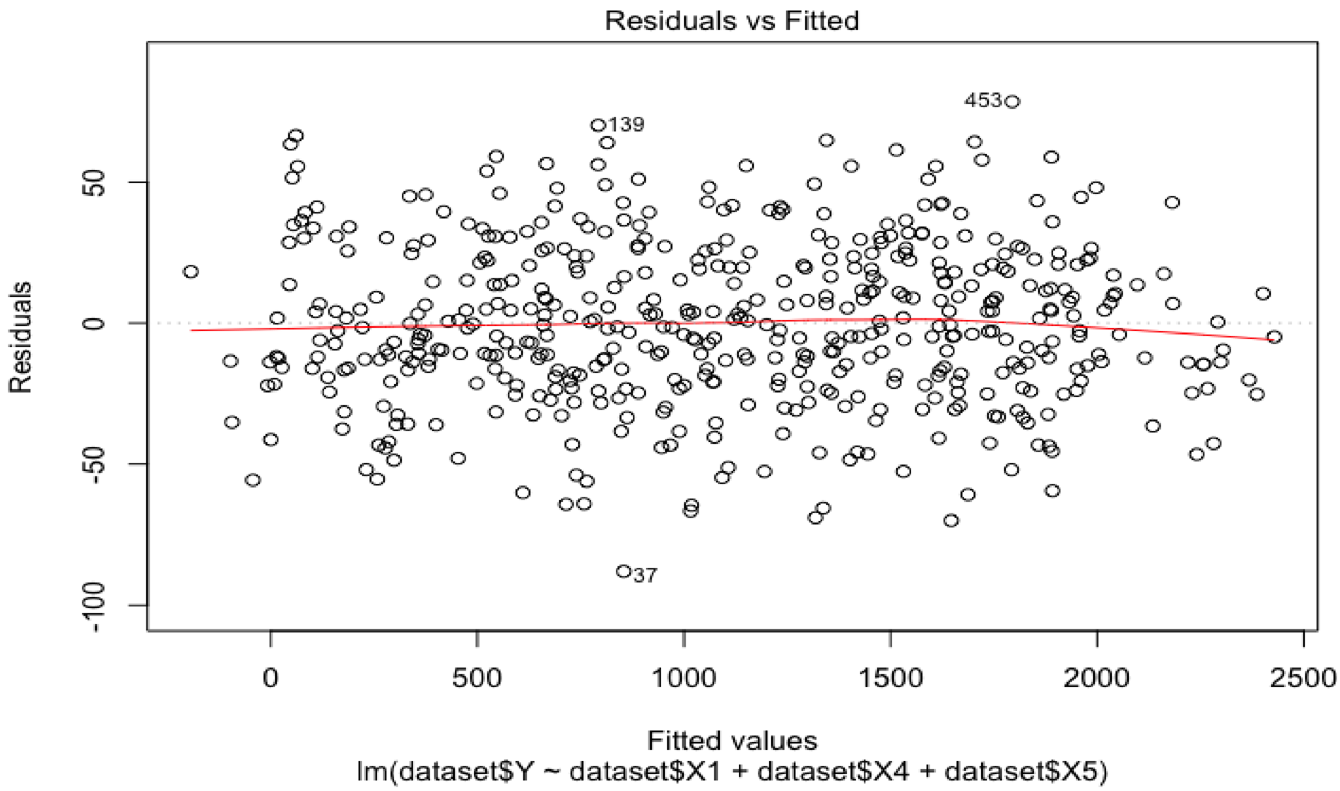


The residuals are normally distributed. The QQ plot for multiple regression model is slightly light tailed which means more data is concentrated near the center of the histogram than at the extremes. This is evident from the histogram of residuals below.

Histogram:



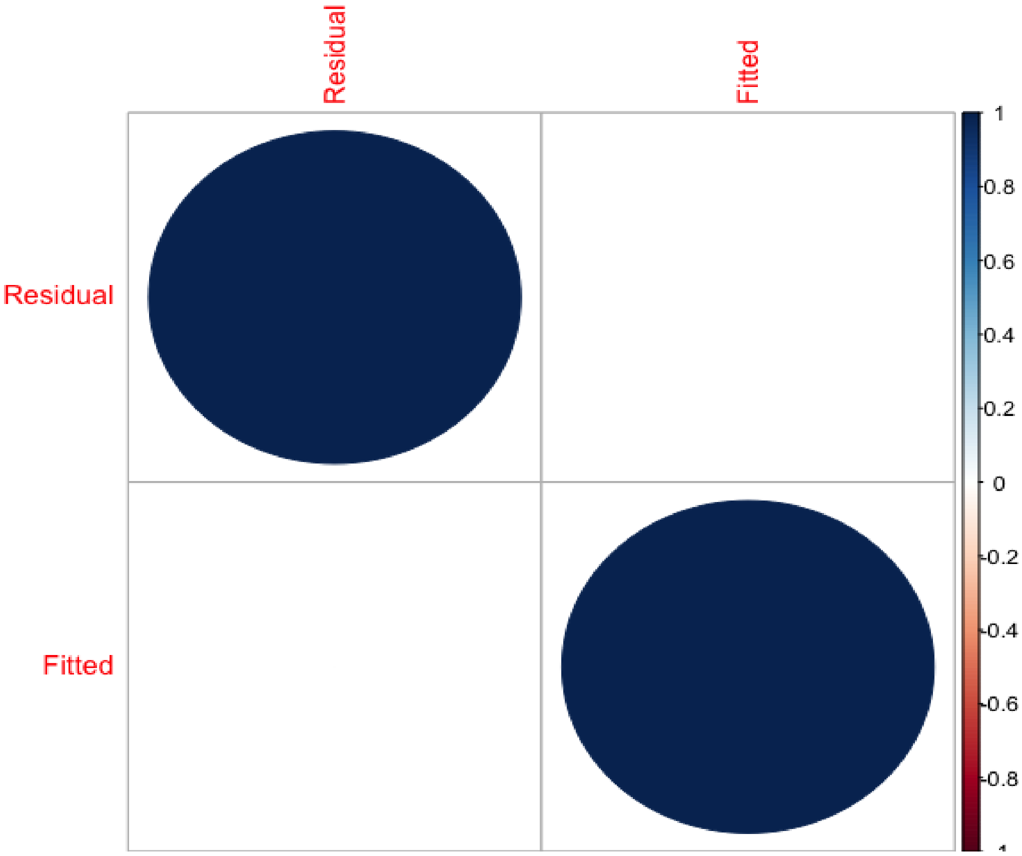
Scatter Plot:



On observing the scatter plot, we can conclude that there is no correlation between both Residual and Fitted data.

The correlation matrix between both also proves the same:

Correlation Matrix:


Residual
Fitted
Residual
Fitted