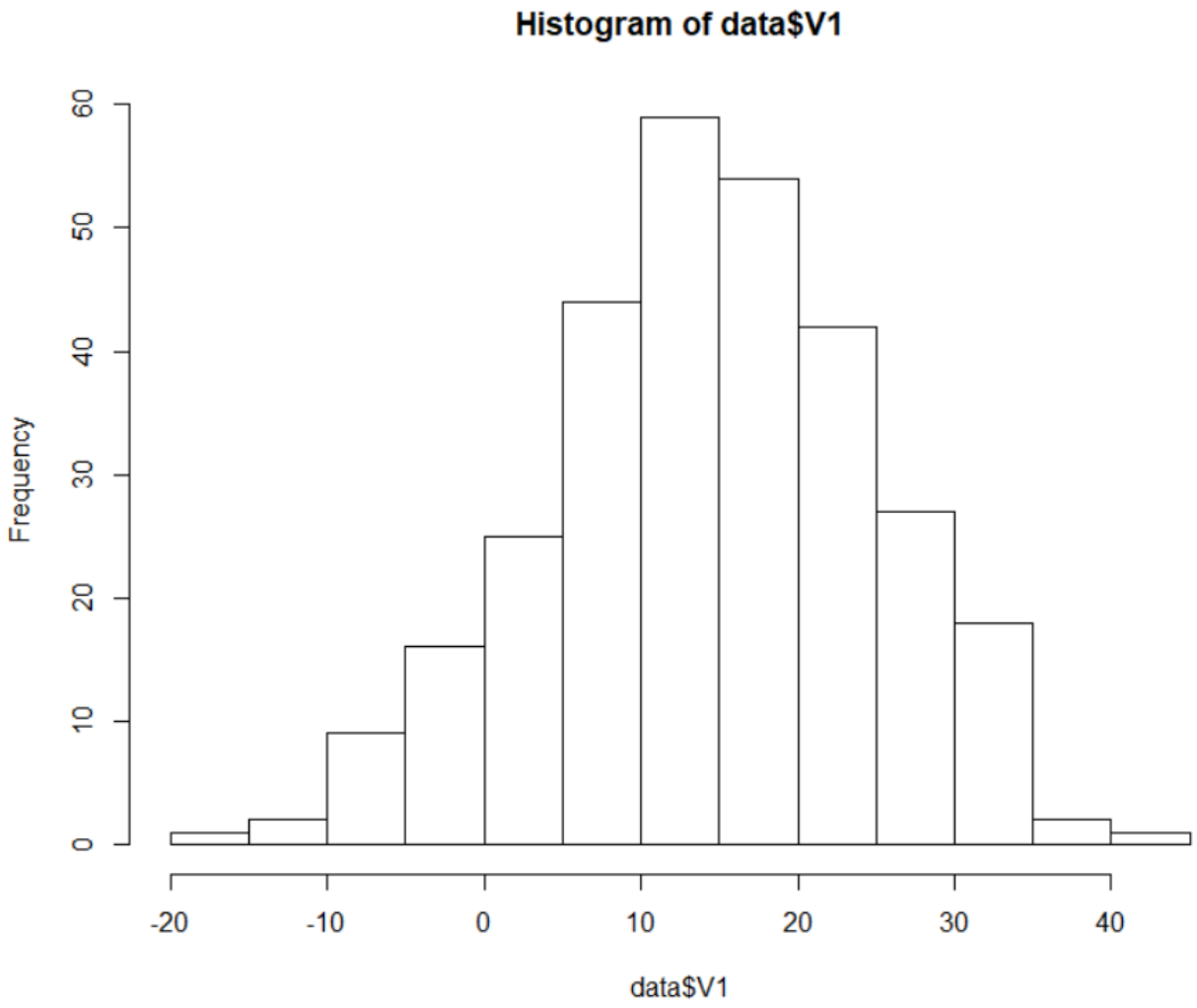
Project 2: Regression Project

Task 1

1)

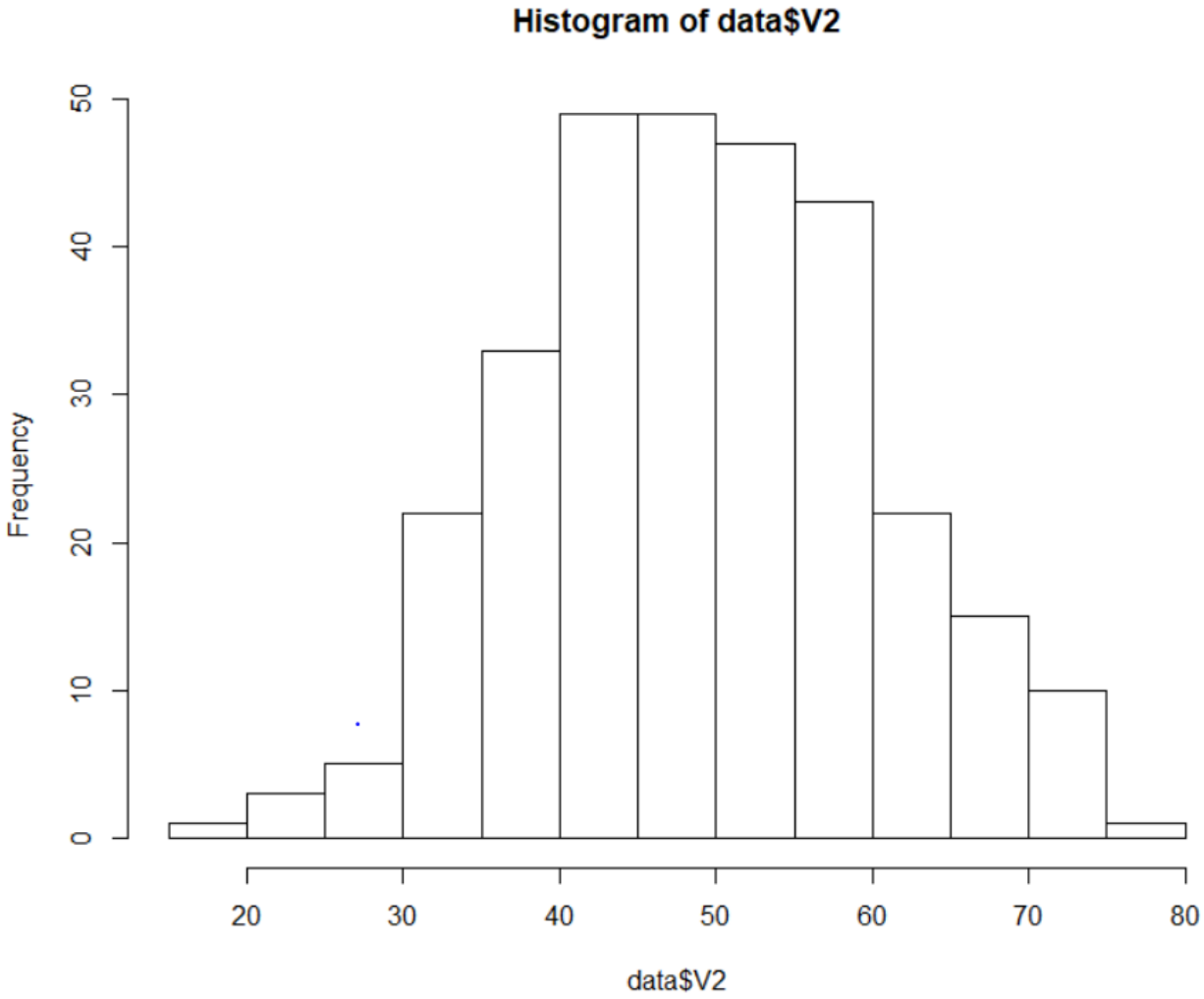
1.1)

(a) V1



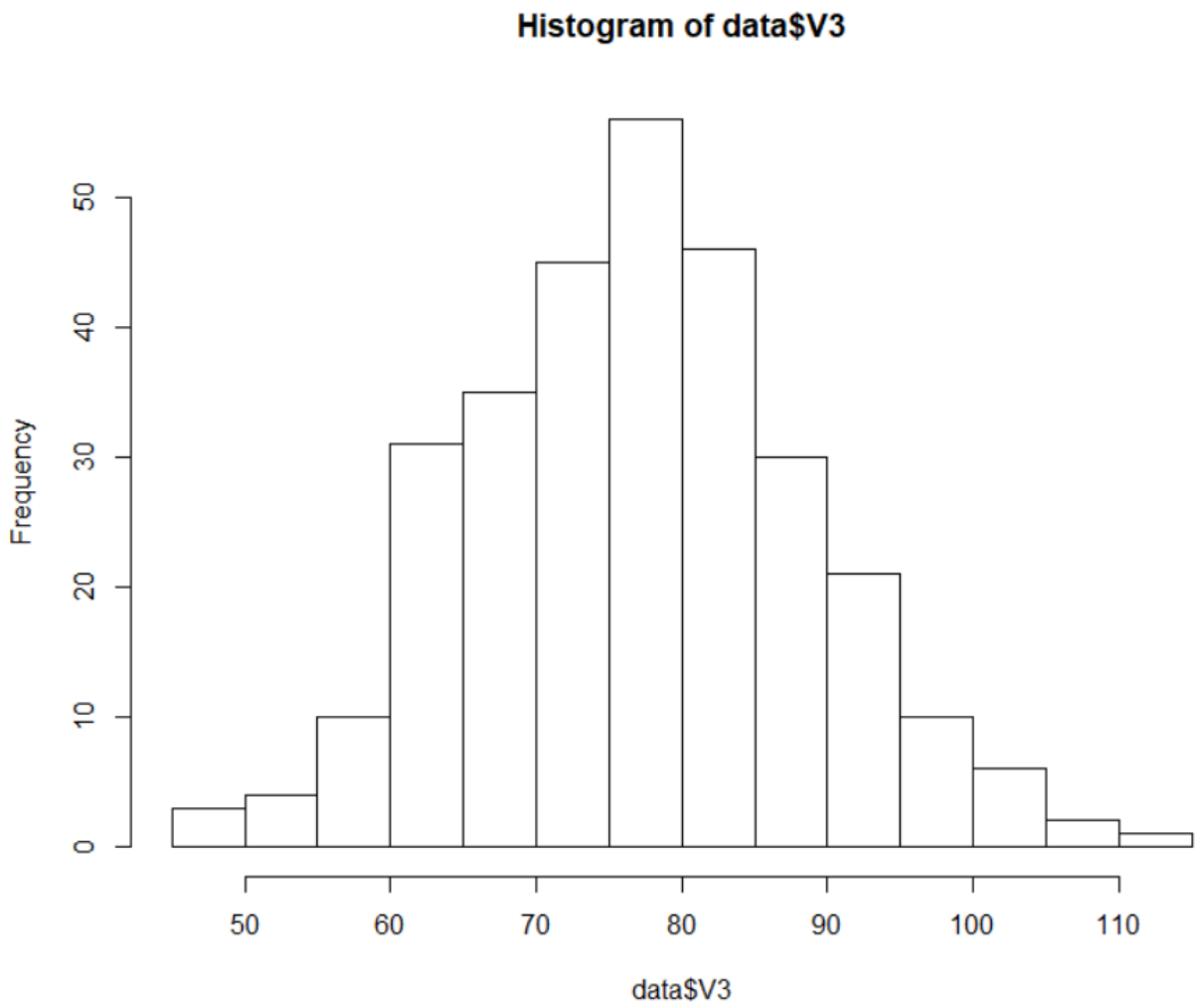
Mean = 14.43793
Variance = 106.492

(b) V2



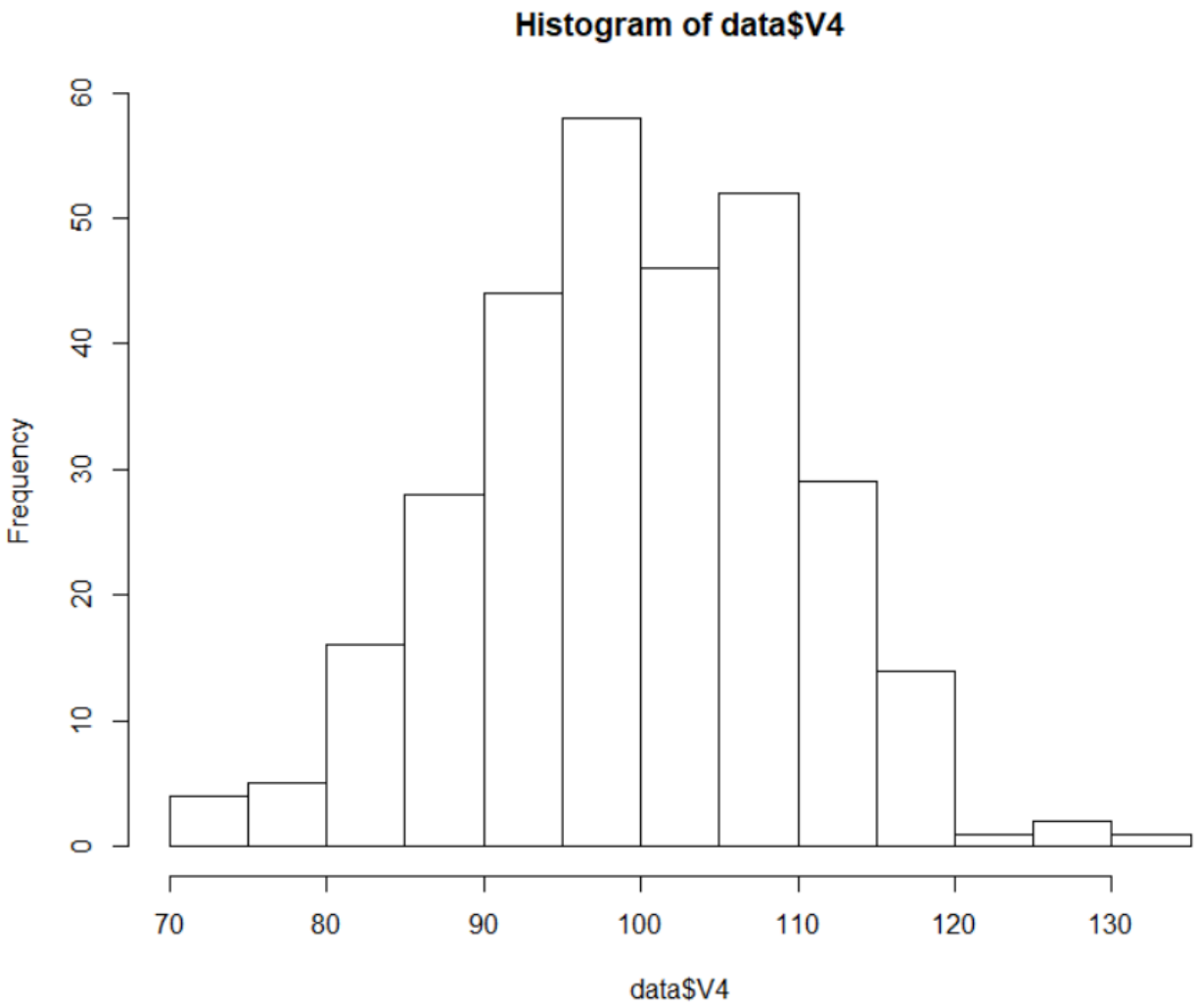
Mean = 48.94855
Variance = 118.4082

(c) V3



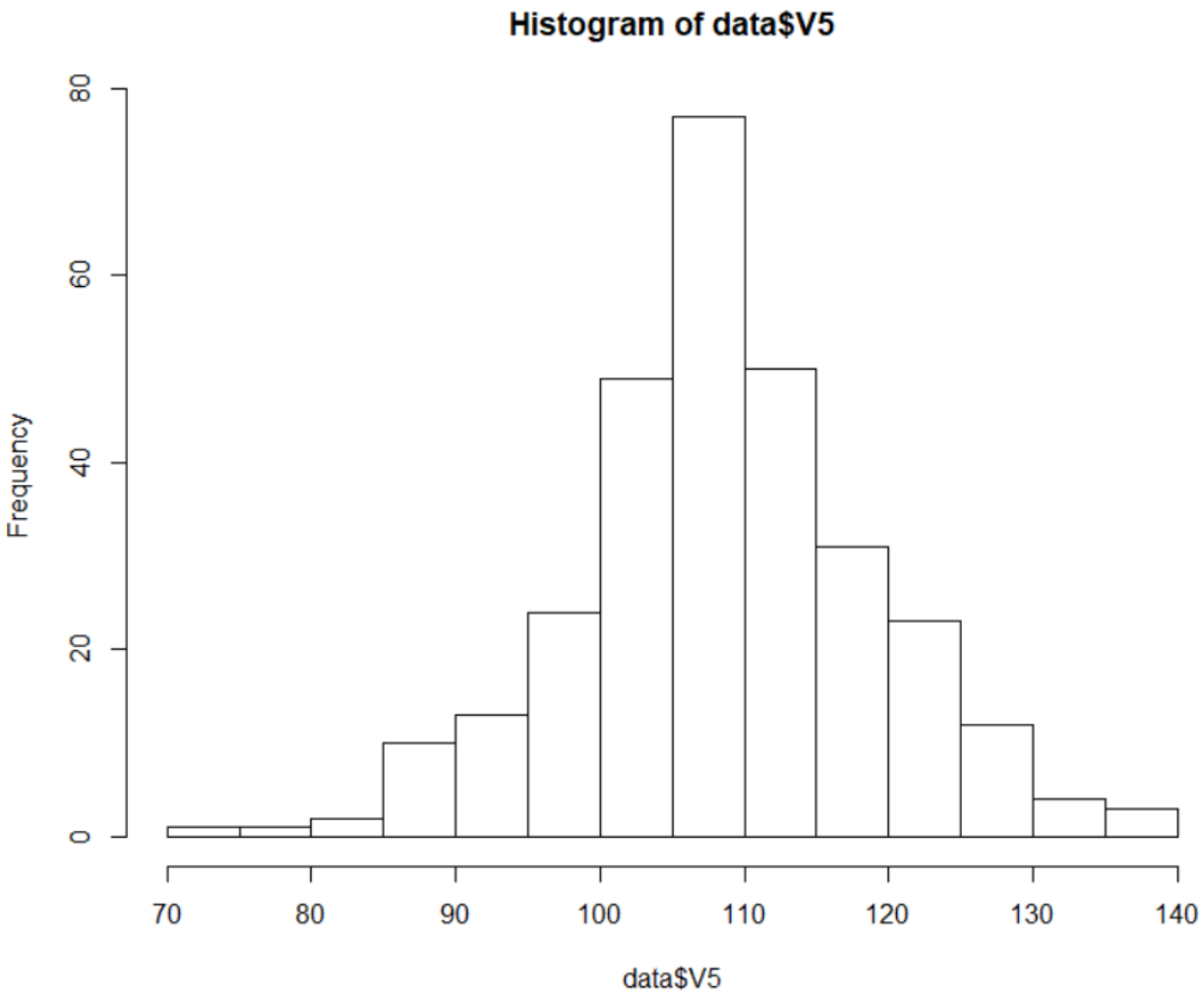
Mean = 77.00445
Variance = 131.3139

(d) V4



Mean = 99.67765
Variance = 107.1422

(e) V5



Mean = 108.7772
Variance = 109.4551

1.2) Using box-plot to check for outliers

For X1 -> boxplot shows to outliers but then checking the two values using the sca tterplot of (X1 and Y), we find that the outliers actually follow the regression line a nd do not fall apart. Hence, keeping the values as it is.

For X2 -> boxplot shows 1 outlier. Checking this value shows it to be away from m ost other points, hence removing this data point.

For X3 -> boxplot shows 3 outliers. Checking the values using the scatterplot show s that the higher end value is away from the other data set and regression line. He nce, removing it. The other lower end outliers are relatively close to the regressio n line, hence keeping them.

For X4 -> boxplot shows 1 outlier. Checking the value using the scatterplot shows that the value is away from the regression line, hence removing it.

For X5 -> boxplot shows 7 outliers. Checking the value using the scatterplot shows that the minimum value of the dataset is away from the regression line and hence removing it.

A total of 4 outliers have been removed.

1.3)

Relationship between dependent and independent variables: X5 and Y show the h ighest positive linear correlation followed by X1 and X4. X2 and X3 don't show go od correlation with Y. None of the independent variables show high correlation wi th each other and hence have very little or no multicollinearity.

Correlation Matrix of Original Data

	X1	X2	X3	X4	X5	Y
X1	1.00000000	-0.029714999	-0.060902129	0.02416089	0.04952514	0.4710583
X2	-0.02971500	1.000000000	0.005895711	-0.06232862	0.14900695	0.2919424
X3	-0.06090213	0.005895711	1.000000000	-0.05960414	0.03531775	0.3413193
X4	0.02416089	-0.062328624	-0.059604143	1.00000000	-0.06560340	0.4156236
X5	0.04952514	0.149006954	0.035317746	-0.06560340	1.00000000	0.5910064
Y	0.47105830	0.291942420	0.341319254	0.41562359	0.59100636	1.0000000

Correlation Matrix of Data after removing Anomalies

	X1	X2	X3	X4	X5	Y
X1	1.00000000	-0.02287457	-0.04315181	0.03049943	0.07283193	0.4913636
X2	-0.02287457	1.00000000	0.00108724	-0.04419480	0.13673755	0.2925367
X3	-0.04315181	0.00108724	1.00000000	-0.04131837	0.03421129	0.3478558
X4	0.03049943	-0.04419480	-0.04131837	1.00000000	-0.05598619	0.4205753
X5	0.07283193	0.13673755	0.03421129	-0.05598619	1.00000000	0.5971843
Y	0.49136363	0.29253669	0.34785576	0.42057533	0.59718426	1.0000000

1.4)

After removing the outliers, a comparison of the correlation matrix shows better li near relationship between the dependent and the independent variables ie. the v alues of correlation for the independent variable with the dependent variable are closer to 1 than before showing better positive correlation.

#Task 2

2.1)

The estimate for:

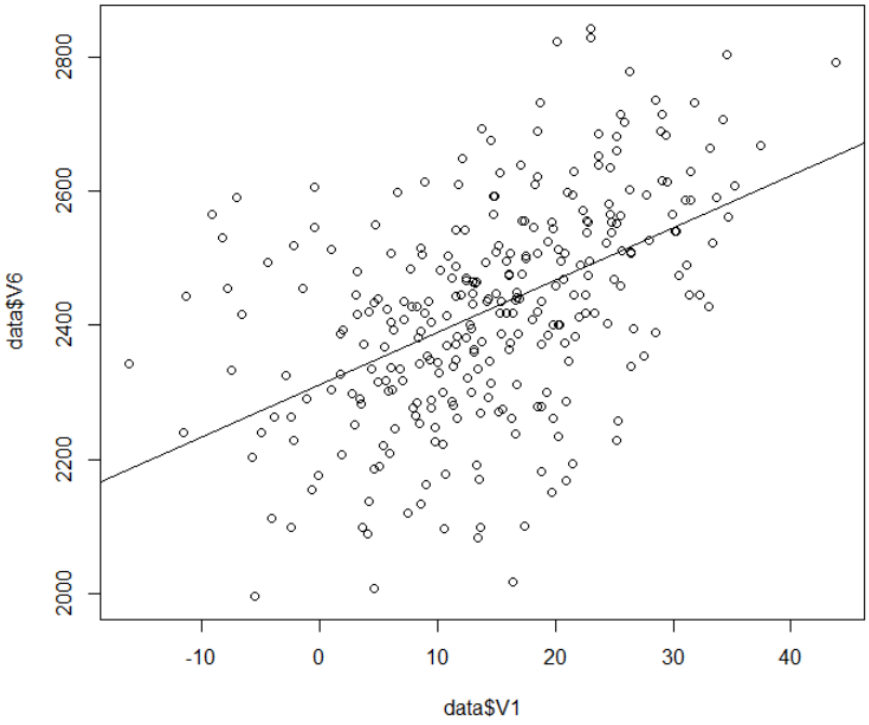
a0 = 2311.8982
a1 = 7.7977
σ = 19909.21

2.2) Individual p-values for both the coefficients as well as the model p-values are <2e-16 (approx. 0) indicating that the model (both regression coefficients) are significant.

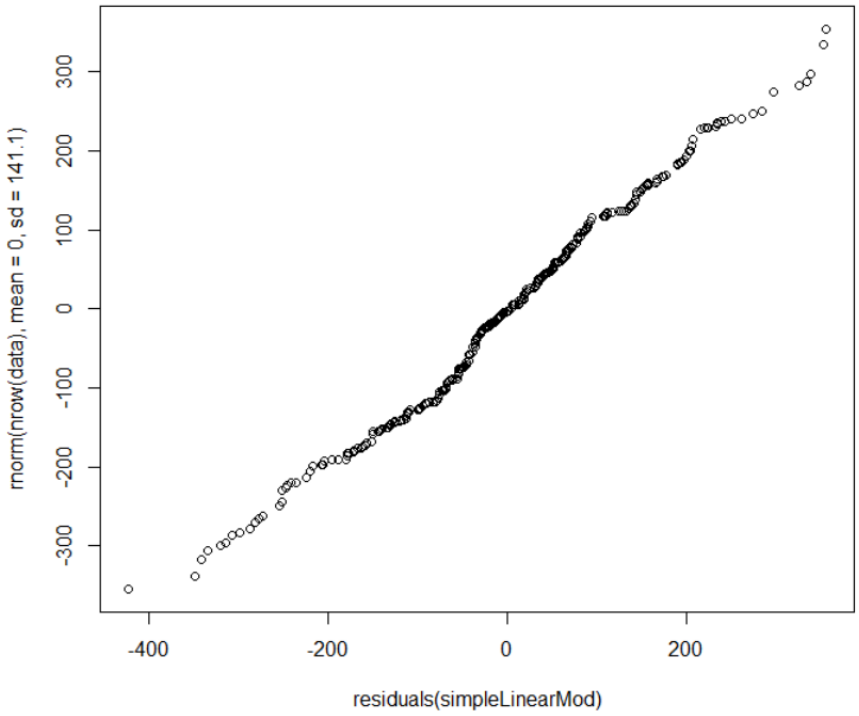
The value of R2 is 0.2414 which is significantly low but we do not reject th e model based on this but agree to the point that a lot of the variation in y accoun ts to the error and is not explained by X1.

F-statistic: 93.58 indicates that there is a higher proportion of the variation in Y th at is explained by X1 compared to the error.

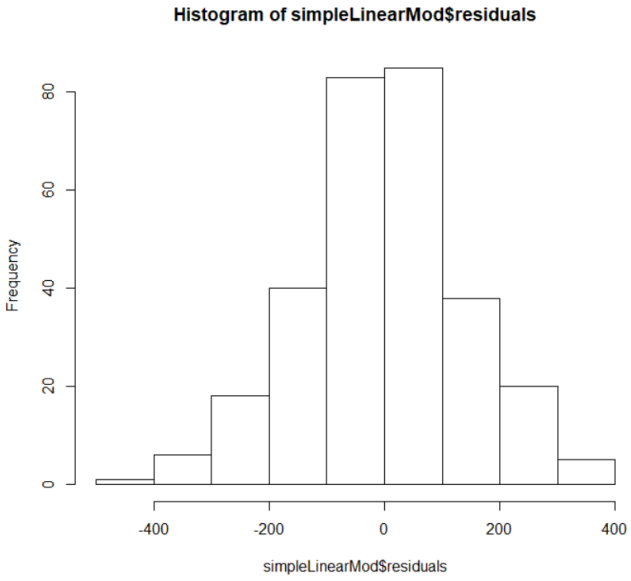
2.3) Regression line against data:



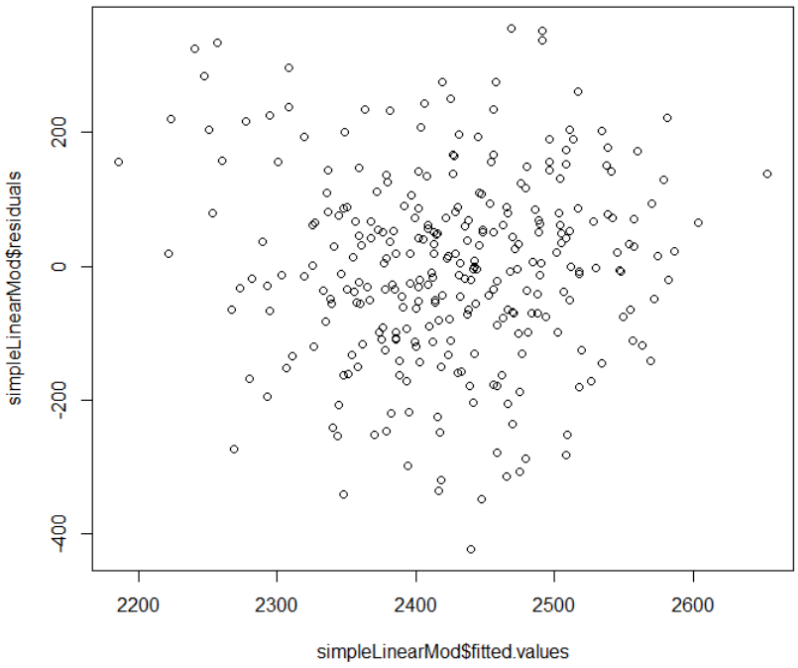
2.4) a) QQ Plot



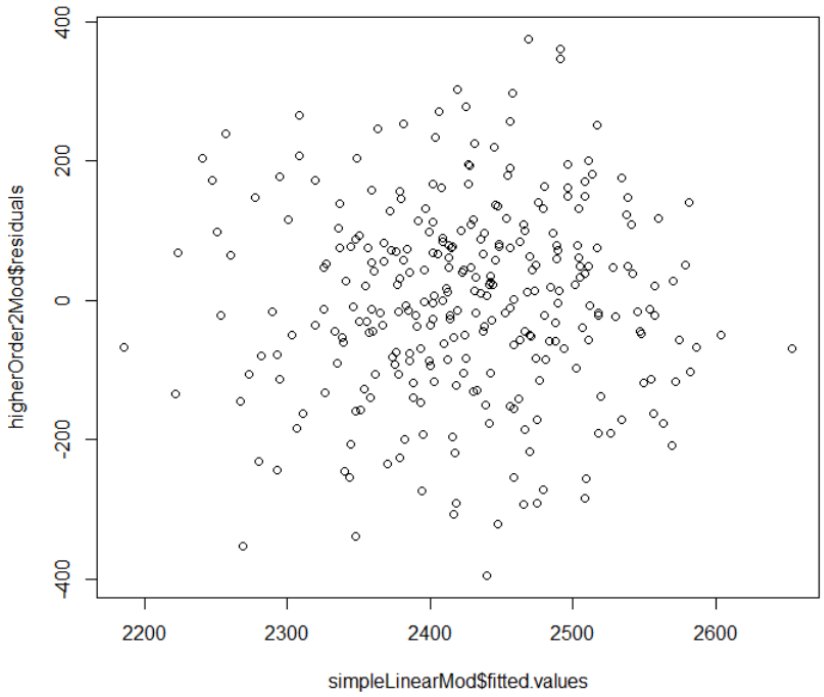
Histogram:



b)



There is a slight top-down coning effect indicating slight heteroscedasticity. This is evident from the fact that adding a quadratic term improves the mo del (fixes the problem to some extent). Below is the scatter plot for the fitte d values with 2nd order polynomial regression model. The residuals are mor e randomly distributed and do not follow any particular trend.



Regarding the Chi-Square Test (written in Python):

The values obtained are:

Chi-square obtained = 9.466666666666667
Chi-square Critical = 14.067140449340169
P-value = 0.2208638586286018 (for alpha = 0.05)

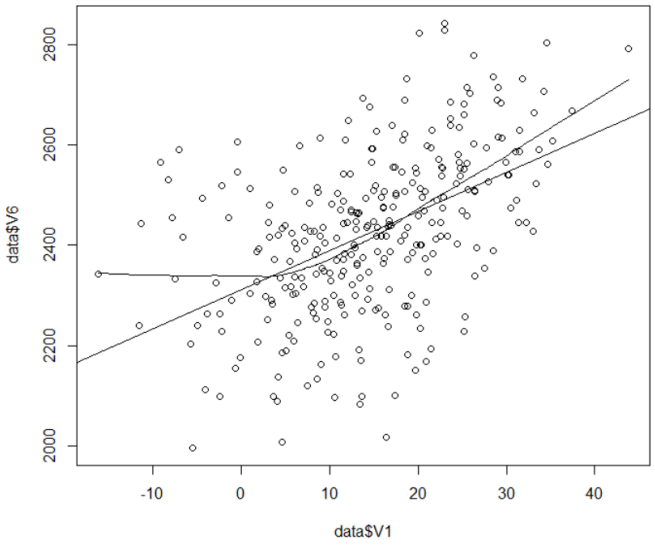
Since the value of Chi-square obtained < Chi-square Critical and the P-value is > 0.05, we conclude that we cannot reject the NULL hypothesis and that the residuals follow Normal Distribution with $N(0, s^2)$.

2.7) and 2.8)

```
> summary(higherOrder2Mod)

Call:
lm(formula = data$V6 ~ I(data$V1^2), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-395.66  -84.64   -0.01   83.66  374.96

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.340e+03  1.133e+01  206.57   <2e-16 ***
I(data$V1^2) 2.710e-01  2.554e-02   10.61   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 137.7 on 294 degrees of freedom
Multiple R-squared:  0.277,     Adjusted R-squared:  0.2745
F-statistic: 112.6 on 1 and 294 DF,  p-value: < 2.2e-16

> summary(simpleLinearMod)

Call:
lm(formula = data$V6 ~ data$V1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-422.56  -84.34    0.59   80.70  355.73

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2311.8982    14.2797 161.901   <2e-16 ***
data$V1        7.7977     0.8061   9.673   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 141.1 on 294 degrees of freedom
Multiple R-squared:  0.2414,    Adjusted R-squared:  0.2389
F-statistic: 93.58 on 1 and 294 DF,  p-value: < 2.2e-16
```

Above is the comparison between first order and 2nd order polynomial regression and it suggests a slight improvement in the predictions from the model obtained by the 2nd order polynomial regression. F-statistic has a higher value for the 2nd order regression with an increase adjusted R-squared. We cannot trust R-squared in this case because the number of coefficients are higher in 2nd order regression which will definitely cause an increase in the value of R-squared due to the effect of the new term. Therefore, considering adjusted R-squared. Also, the residual standard error is reduced slightly.

So, in summary, the models are significant (2nd order model being better than the 1st order). R^2 adjusted of 0.2389 suggests that we can predict only 23.89% of the dependent variable from the given independent variable. The rest of it accounts to actual error or other variables adding those would improve the prediction of the model overall.



Above scatterplot shows the simple regression line and the smoothing line.

3)

3.1) The values of all the coefficients are:

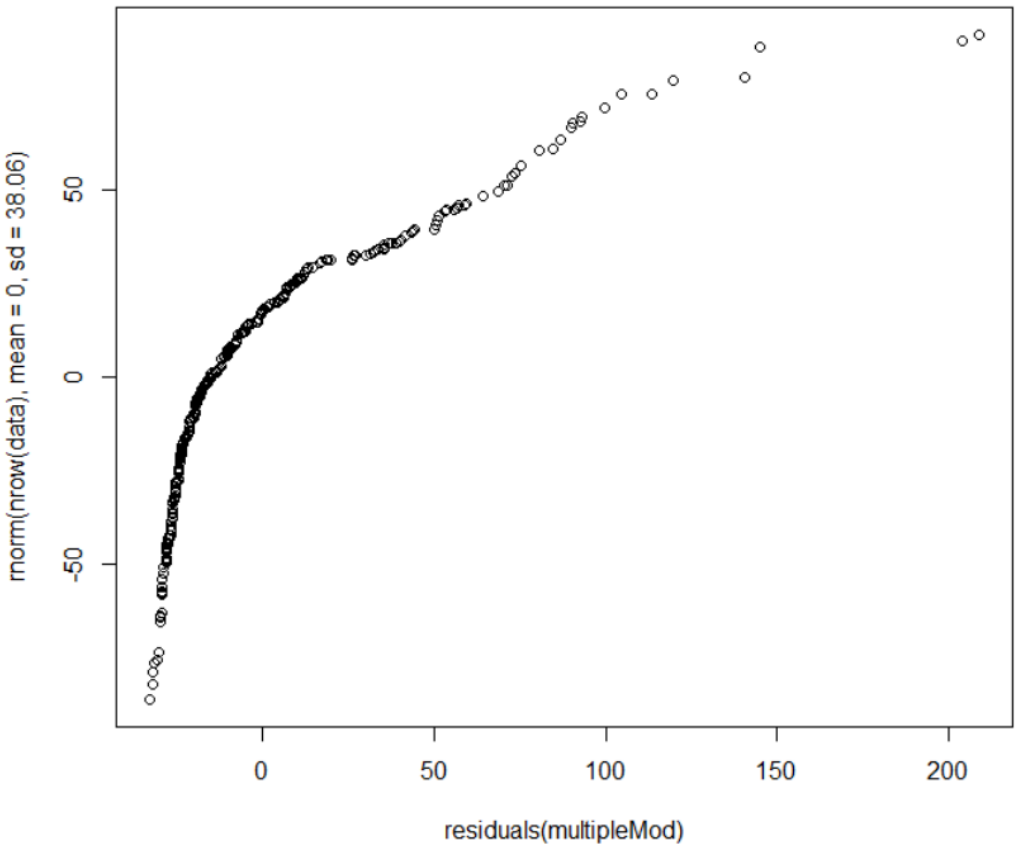
$a_0 = 74.0771$
$a_1 = 7.2881$
$a_2 = 3.7590$
$a_3 = 5.2517$
$a_4 = 7.3451$
$a_5 = 8.4966$

Standard error: 38.06 and hence the value of variance (σ2) = 1448.5636

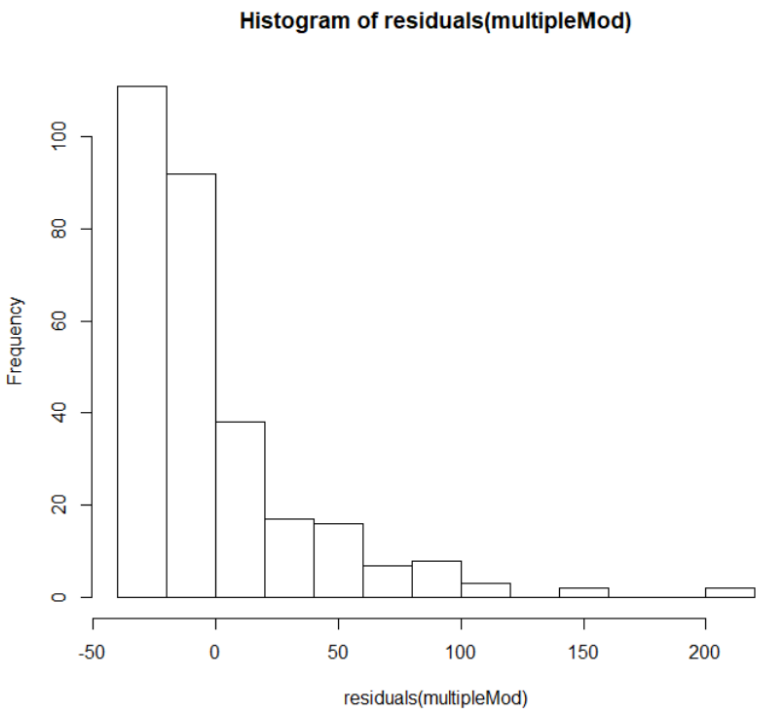
3.2) The p-values for all the coefficients are less than 0.05 indicating th at the coefficients are significant. F-statistic is high, and the models p-value is < 0.0 5 indicating that the model is significant. The value of R-square may not be a good indicator because of a presence of high number of variables. Adding a variable incr eases the effect of the R-square. A better indicator is the value of the adjusted R-s quared that normalizes for the number of the variables. The value of adjusted R-sq uared is high indicating that the model is significant.
Since, the p-values for all the coefficients are less than 0.05, all of them are signific ant and none of them need to be removed.

3.3)

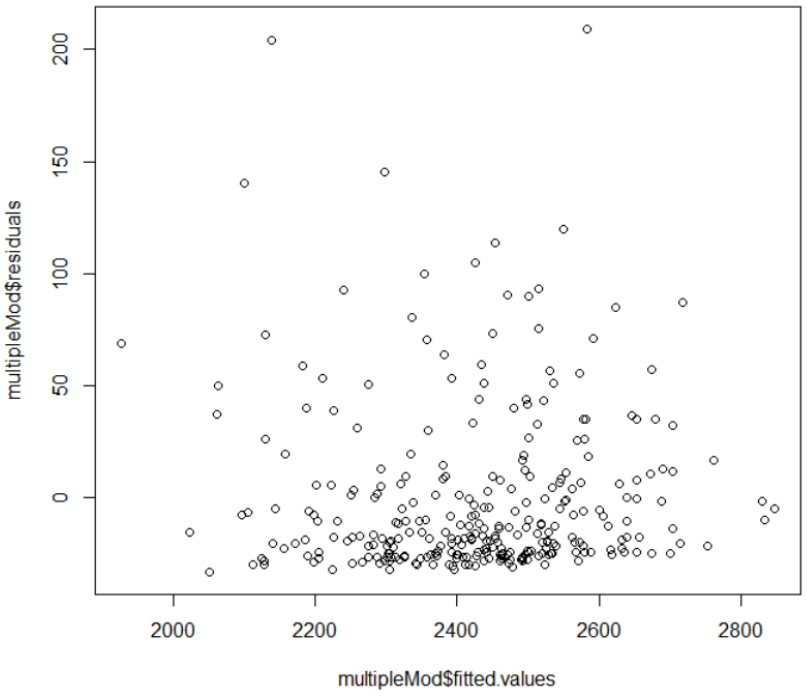
a) QQ Plot:



This suggests that the residuals aren't normally distributed since the 1st quantile o f the residuals is reached in a very short band compared to 1st quantile of a norma l distribution with sd = 38.06. Hence, it shows a upper tilted curve confirmed by th e below histogram:



b)



Regarding the Chi-Square Test (written in Python):

The values obtained are:

Chi-square obtained = 486.66666666666663
Chi-square Critical = 14.067140449340169
P-value = 0.0 (for alpha = 0.05)

Since the value of Chi-square obtained > Chi-square Critical and the P-value is 0 < 0.05, we conclude that we reject the NULL hypothesis and accept the alternate hy pothesis that the residuals don't follow the Normal Distribution.

3.4)

The distribution of the residuals is not normal. The scatter plot, histogram or the q qplot show that most of the residuals are very small in magnitude. There are a few random errors that show high values (fluctuations). This indicates that most of the times the predictions will be accurate except for a few times where the predicted value may vary largely from the observed value. With higher values of n (the numb er of observations), most of the test results are reliable even when a few of the re siduals depart substantially from the normal distribution.