PROJECT 2: REGRESSION PROJECT

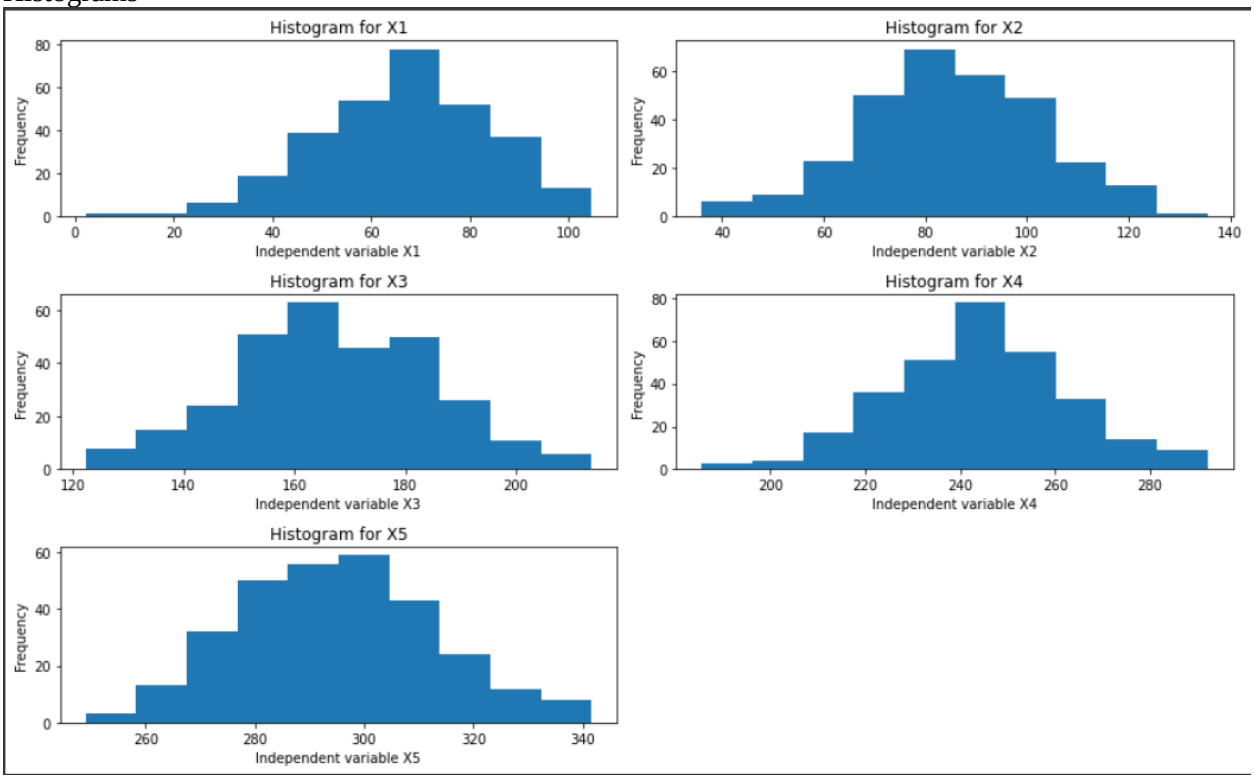
File number: 2

Task 1: Basic Statistics Analysis

Mean and variance

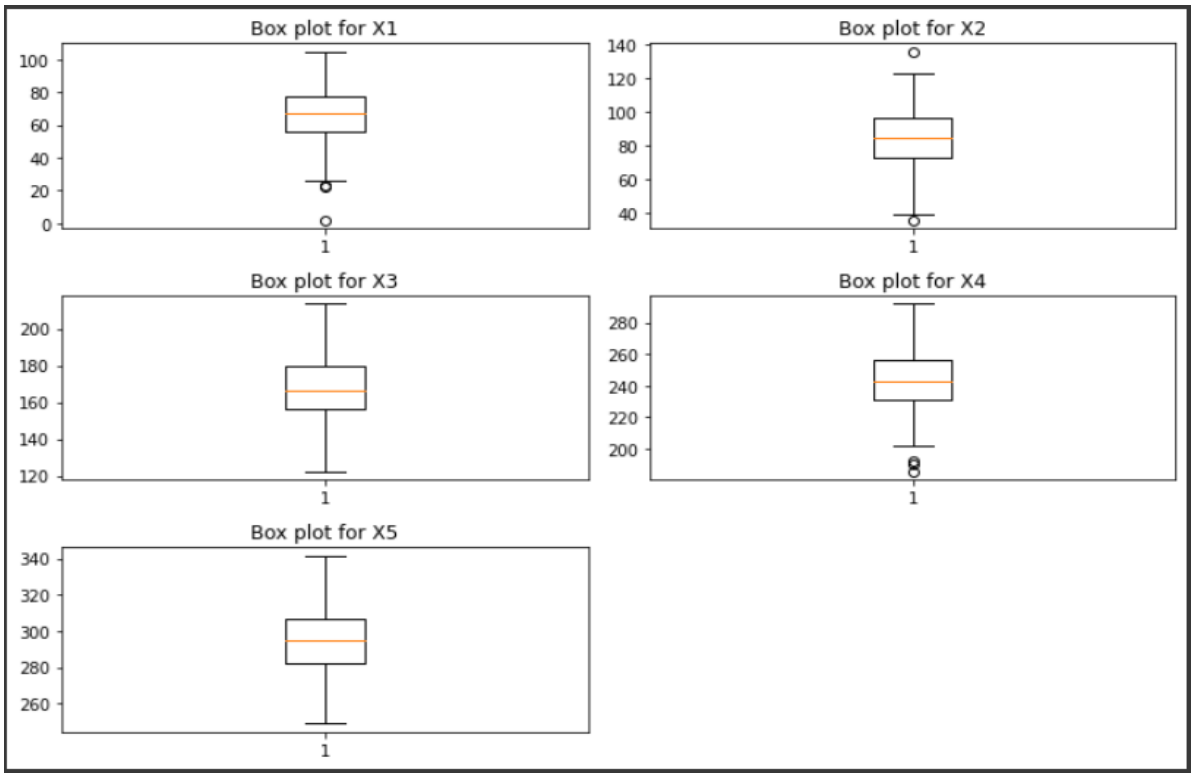
	variable	mean	variance
0	X1	66.596186	282.386731
1	X2	85.085927	306.372422
2	X3	167.362167	325.784891
3	X4	243.571400	344.207283
4	X5	294.752233	320.202993

Histograms

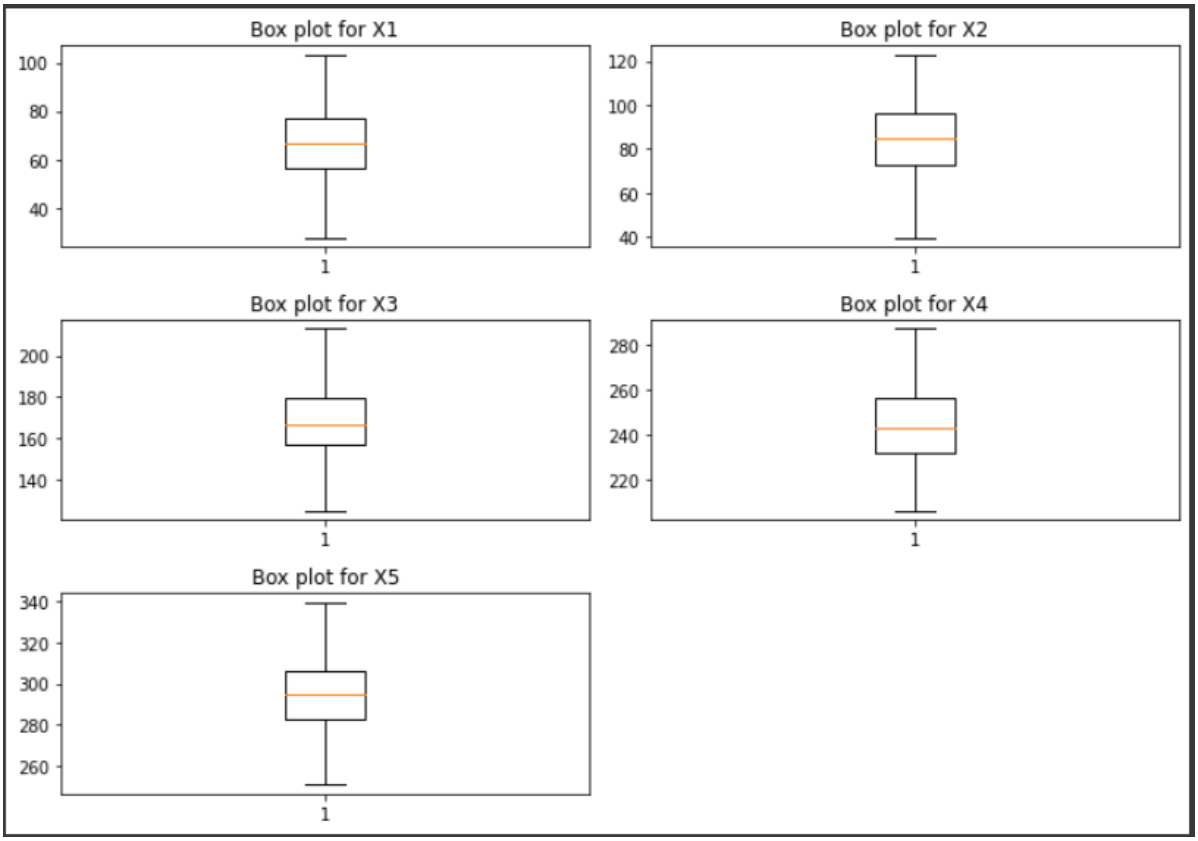


The histograms for all the independent variable show distributions similar to a normal distribution.

Box Plot



Box plot after removing the outliers (a total of 17 outliers were removed)



Correlation matrix:

	X1	X2	X3	X4	X5	Y
X1	1.000000	0.047110	-0.023789	-0.082386	-0.061232	0.950138
X2	0.047110	1.000000	0.014037	0.046273	0.043521	0.118592
X3	-0.023789	0.014037	1.000000	0.036237	0.055953	0.089594
X4	-0.082386	0.046273	0.036237	1.000000	0.128780	0.084423
X5	-0.061232	0.043521	0.055953	0.128780	1.000000	0.155129
Y	0.950138	0.118592	0.089594	0.084423	0.155129	1.000000

From the heatmap, we see the correlation between the independent variables and Y, and the correlation between the independent variable themselves.
X1 and Y are highly correlated with a high positive correlation coefficient of 0.95.
X2 and X5 are also positively correlated to Y, but the correlation is not as high as X1 and Y. This can be seen from the coefficients of the X2-Y (0.12) and X5-Y (0.16).
X3 and X4 are even less correlated with Y with a coefficient of 0.09 and 0.084, respectively.

There is no strong correlation between the independent variables themselves. Thus, we can conclude that this dataset is free of multicollinearity.

Task 2: Simple Linear Regression

Simple linear regression summary:

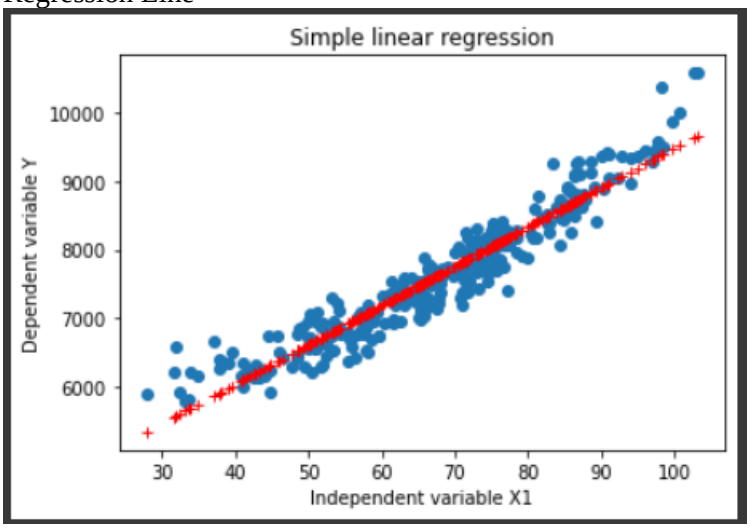
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.903
Model:                            OLS   Adj. R-squared:                  0.902
Method:                 Least Squares   F-statistic:                     2609.
Date:                Wed, 07 Oct 2020   Prob (F-statistic):          3.09e-144
Time:                        17:30:31   Log-Likelihood:                -2011.6
No. Observations:                 283   AIC:                             4027.
Df Residuals:                     281   BIC:                             4035.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3756.9803     77.211     48.658      0.000    3604.994    3908.966
X1            57.1759      1.119     51.076      0.000      54.972      59.379
==============================================================================
Omnibus:                       13.401   Durbin-Watson:                   2.170
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               14.509
Skew:                           0.466   Prob(JB):                     0.000707
Kurtosis:                       3.601   Cond. No.                         302.
==============================================================================
```

a0 (constant) = 3756.908
a1 (X1 coeff) = 57.1759
variance = 87429.318
p-values for the constant and X1 are 0. This indicates that the null hypothesis is rejected for both the coefficients and they should be non-zero.
The R value indicates the correlation between observed values and the predicted value and should be as close to 1 as possible.
The R2 values indicates how close the data is to the adjusted regression line.
R2 = 0 Indicates that the model does not explain anything about the variability of response data around its mean.
R2 = 1 Indicates that the model explains all the variability of the response data around its average.
R2 = 0.903, this indicates a that the linear model explains 90.3% of the variance of the dependent variable from the regressors (independent variables).

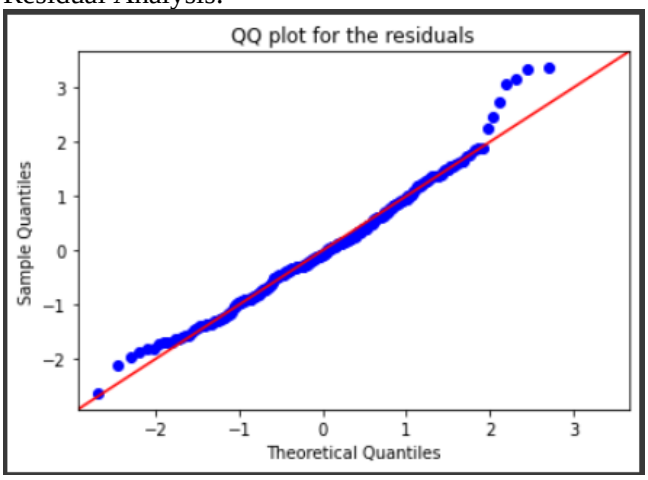
The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.
Adj R2 = 0.902. The value is again close to 1 which is a good indication for the model.

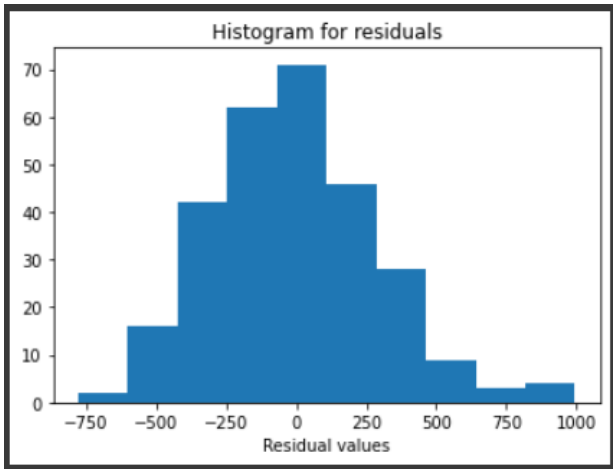
The F value can be used to determine whether the test is statistically significant.
The null hypothesis states that the model with no independent variables fits the data as well as your model. If it is true, you expect F to have a value close to 1.0 most of the time. A large F ratio means that the variation among group means is more than you'd expect to see by chance.
F = 2609 which indicates that the null hypothesis is wrong.

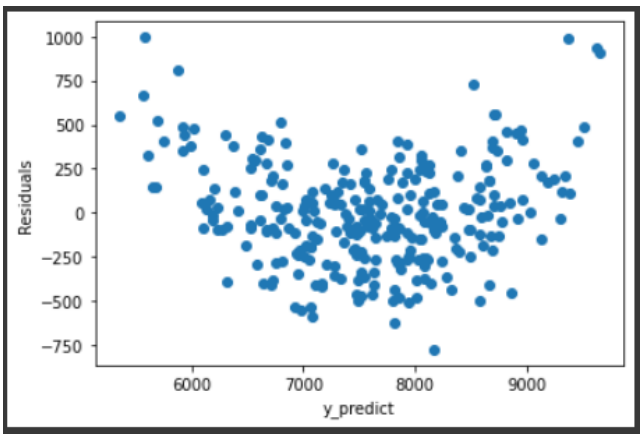
Regression Line



Residual Analysis:








```
Chi-square test
pvalue: 0.0012300380780962553
Null hypothesis rejected.
Residuals do not follow a normal distribution
```

The QQ plot shows that the residuals don't follow the 45-degree always (deviates for left and right tail). The histogram looks like that of a normal distribution, but this is not conclusive. Chi-square test rejects the null hypothesis that the residuals follow a normal distribution. (I am considering 95th-percentile i.e. if $p<0.05$, null hypothesis is rejected, otherwise accepted).

The scatter plot of the residuals does not show any significant trend. The variance in the residuals however seems to vary for left and right tail.

Higher order polynomial regression summary:

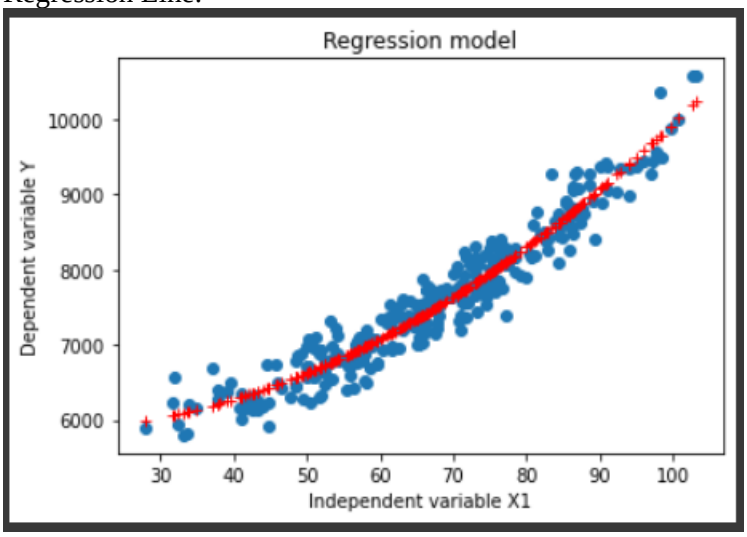
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.931
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                     1895.
Date:                Wed, 07 Oct 2020   Prob (F-statistic):          1.79e-163
Time:                        20:23:36   Log-Likelihood:                -1962.7
No. Observations:                 283   AIC:                             3931.
Df Residuals:                     280   BIC:                             3942.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       5922.8964    211.526     28.001      0.000    5506.513    6339.279
X1           -12.0490      6.502     -1.853      0.065     -24.847       0.749
X1^2           0.5218      0.048     10.761      0.000       0.426       0.617
==============================================================================
Omnibus:                        1.982   Durbin-Watson:                   2.218
Prob(Omnibus):                  0.371   Jarque-Bera (JB):                1.699
Skew:                           0.054   Prob(JB):                        0.428
Kurtosis:                       2.636   Cond. No.                     7.41e+04
==============================================================================
```

a0 (constant) = 5922.8964
a1 (X1 coeff) = -12.0490
a2 (X1^2 coeff) = 0.5218
variance = 61849.786

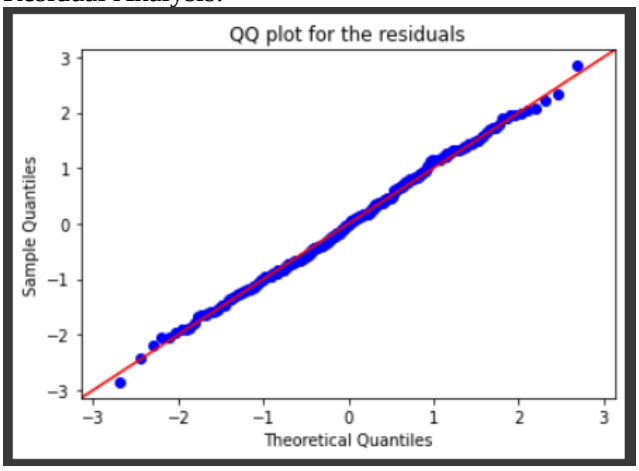
p-value for a0 and a2 is 0, thus we reject the null hypothesis. p-value for a1is 0.065 thus we accept the null hypothesis (considering 95th percentile i.e. p-value < 0.05 => reject, else accept). Thus, a1 is not significant for this model which makes sense because we have X1^2 which is highly correlated with X1.

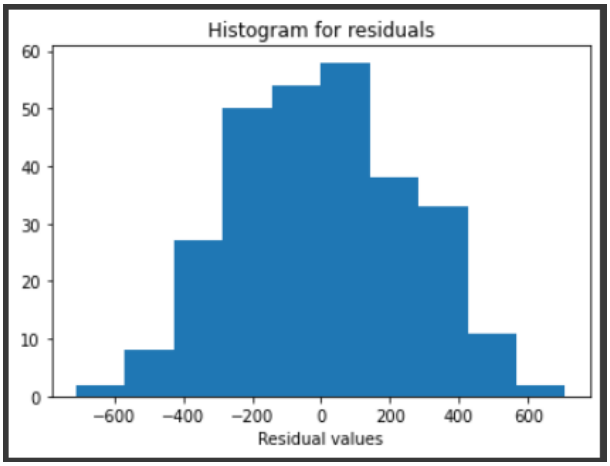
R2 = 0.931. The R2 value has improved further when compared with R2 value of the simple linear regression model.
F = 1895. Again, the F value is much greater than zero, thus the null hypothesis that the model with no independent variables fits the data as well as your mode, is rejected.

Regression Line:

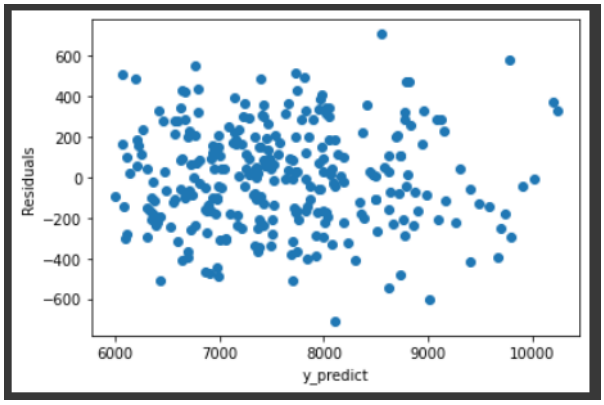


Residual Analysis:






```
Chi-square test
pvalue: 0.37123993006834366
Null hypothesis accepted.
Residuals follow a normal distribution
```



The QQ plot shot shows that the residuals follow the 45-degree (high probability that the distribution is normal). Histogram also shows a normal like distribution. This is confirmed by the chi-square test. The p-value is greater than 0.05 and the null hypothesis is accepted.
The scatter plot does not a follow significant trend.
The polynomial regression model, produces much better results as compared to the simple linear model.

Task 3: Linear Multivariable Regression

Multivariable regression model summary:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.980
Model:	OLS	Adj. R-squared:	0.980
Method:	Least Squares	F-statistic:	2718.
Date:	Wed, 07 Oct 2020	Prob (F-statistic):	5.11e-233
Time:	21:07:24	Log-Likelihood:	-1787.7
No. Observations:	283	AIC:	3587.
Df Residuals:	277	BIC:	3609.
Df Model:	5		
Covariance Type:	nonrobust		

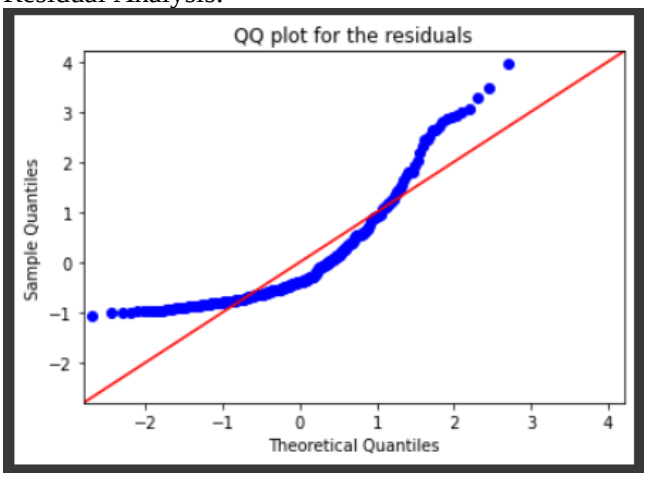
	coef	std err	t	P>|t|	[0.025	0.975]
const	-2303.1764	188.943	-12.190	0.000	-2675.122	-1931.230
X1	58.5151	0.514	113.778	0.000	57.503	59.528
X2	3.2219	0.482	6.687	0.000	2.273	4.170
X3	5.2458	0.463	11.333	0.000	4.335	6.157
X4	7.2332	0.464	15.580	0.000	6.319	8.147
X5	10.3605	0.470	22.057	0.000	9.436	11.285

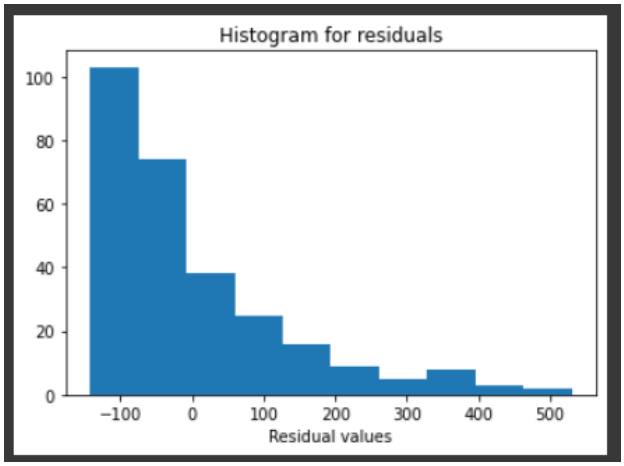
Omnibus:	86.291	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	172.659
Skew:	1.591	Prob(JB):	3.22e-38
Kurtosis:	5.124	Cond. No.	1.01e+04

a0 = -2303.1764
a1 = 58.515
a2 = 3.22
a3 = 5.24
a4 = 7.23
a5 = 10.36
variance = 17962.92

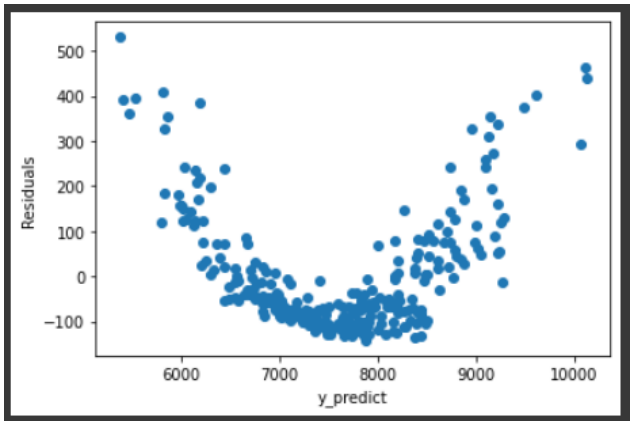
The p-values for all the coefficients is equal to 0. The null hypothesis is rejected for all the coefficients.
R2 = 0.98 has further increased from polynomial regression model.
F = 2718 indicates that the null hypothesis is rejected.

Residual Analysis:






```
Chi-square test
pvalue: 1.8286160456403715e-19
Null hypothesis rejected.
Residuals do not follow a normal distribution
```



The QQ plot shows the deviation of the residuals from the 45-degree line. The Chi-square test further proves that the residuals do not follow a normal distribution. The scatter plot also shows that the variance of the residuals varies a lot at the left and right ends.

Although the null hypothesis is rejected for all the coefficients and the R value is very close to one, this model is still not very good as seen by the residual analysis. Looking at the correlation matrix again.

I will try to improve the model by selecting independent variables with high correlation with Y. Selecting three independent variables with 3 highest correlation coefficients. Selecting X1, X2 and X5.

Improved Multivariable regression model summary:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.953
Model:                            OLS   Adj. R-squared:                  0.952
Method:                 Least Squares   F-statistic:                     1866.
Date:                Wed, 07 Oct 2020   Prob (F-statistic):          2.97e-184
Time:                        21:41:29   Log-Likelihood:                -1910.2
No. Observations:                 283   AIC:                             3828.
Df Residuals:                     279   BIC:                             3843.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.8384    226.229      0.030      0.976    -438.493     452.170
X1            57.7719      0.787     73.373      0.000      56.222      59.322
X2             3.6288      0.739      4.909      0.000       2.174       5.084
X5            11.5403      0.715     16.138      0.000      10.133      12.948
==============================================================================
Omnibus:                       34.869   Durbin-Watson:                   2.028
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               45.315
Skew:                           0.857   Prob(JB):                     1.45e-10
Kurtosis:                       3.953   Cond. No.                     5.75e+03
==============================================================================
```

a0 = 6.8384
a1 = 57.77
a2 = 3.62
a3 = 11.54
variance = 42678.57

The p-values for coefficients of X1, X2 and X5 is equal to 0. The null hypothesis is rejected for these coefficients. The p-value for the constant, 0.976 is > 0.05 and therefore, the null hypothesis is accepted. Thus, the constant is not significant for this model.
R2 = 0.953. Which is less than the previous model.
F = 1866 indicates that the null hypothesis is rejected.

Next, lets remove the constant from this model.

No constant multivariable regression model summary:

OLS Regression Results

Dep. Variable:	Y	R-squared (uncentered):	0.999
Model:	OLS	Adj. R-squared (uncentered):	0.999
Method:	Least Squares	F-statistic:	1.281e+05
Date:	Wed, 07 Oct 2020	Prob (F-statistic):	0.00
Time:	21:47:15	Log-Likelihood:	-1910.2
No. Observations:	283	AIC:	3826.
Df Residuals:	280	BIC:	3837.
Df Model:	3		
Covariance Type:	nonrobust		

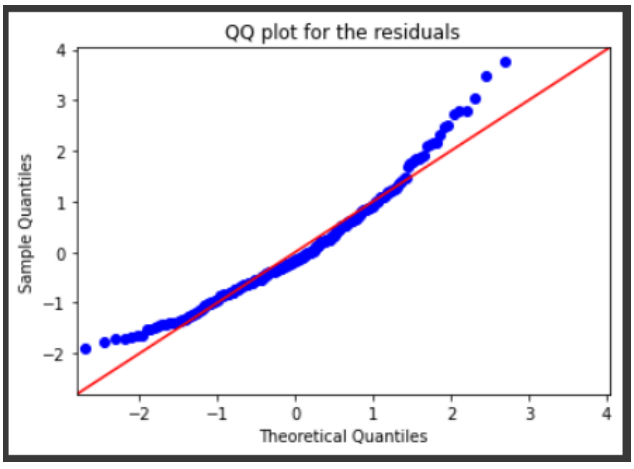
	coef	std err	t	P>|t|	[0.025	0.975]
X1	57.7785	0.755	76.552	0.000	56.293	59.264
X2	3.6338	0.719	5.051	0.000	2.218	5.050
X5	11.5605	0.255	45.275	0.000	11.058	12.063

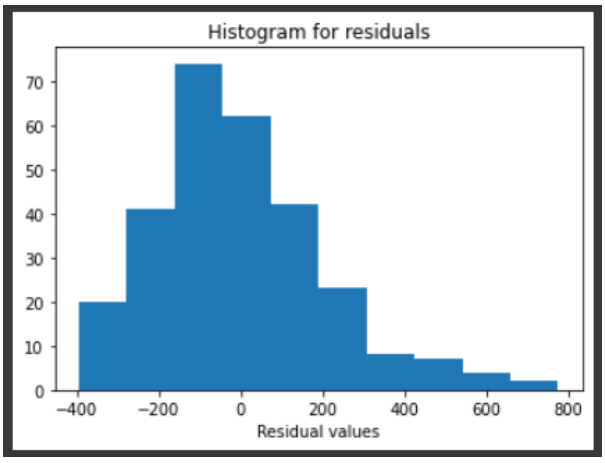
Omnibus:	34.814	Durbin-Watson:	2.027
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45.207
Skew:	0.856	Prob(JB):	1.53e-10
Kurtosis:	3.950	Cond. No.	20.0

a1 = 57.7785
a2 = 3.633
a3 = 11.56
variance = 42678.57

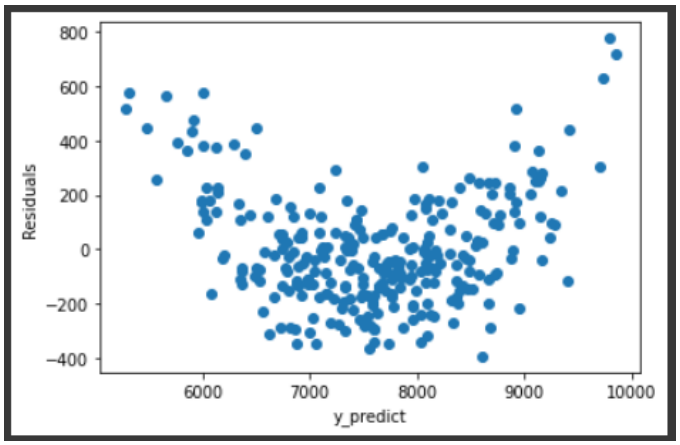
The p-value of all the coefficients is 0. Thus, the null hypothesis is rejected for all the coefficients and they all should be non-zero.
R2 = 0.999. The R2 value for this model is greater than any of the previous model and is nearly equal to 1.
F = 1.281e+05. The null hypothesis is again rejected.

Residual Analysis:





```
Chi-square test
pvalue: 2.6812920785170595e-08
Null hypothesis rejected.
Residuals do not follow a normal distribution
```



The QQ plot for the residuals follows the 45-degree line much better than the original multivariable model (one with all the independent variable). The histogram again looks similar to a normal distribution. However, the chi-square test rejects the null hypothesis and the residuals still don't follow a normal distribution. The scatter plot again doesn't show a discernable trend however, the variance again varies at the left and right end tail.

The multivariable model with just X1, X2, X5 and without an intercept is much better than the multivariable model containing all the independent variable.