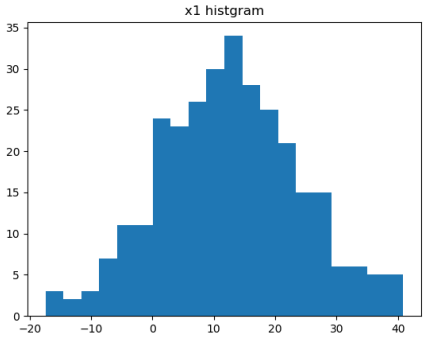
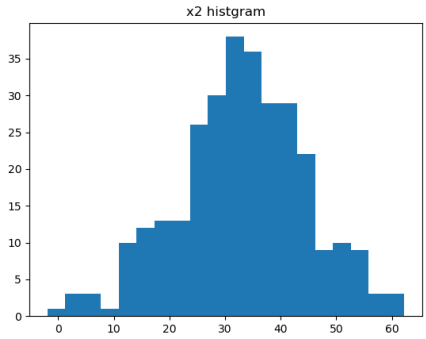
Multivariable Regression Results

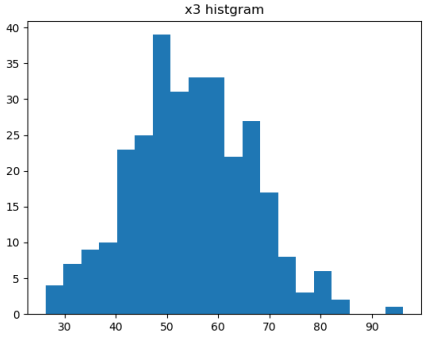
Wen-Han Hu(whu24)

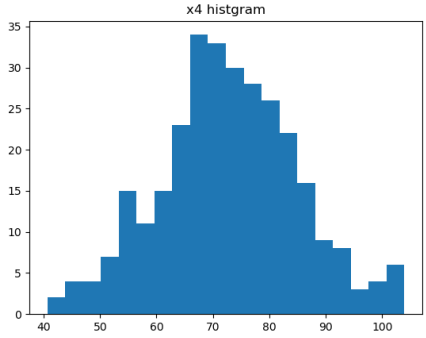
Task .1 Basic Statistics Analysis

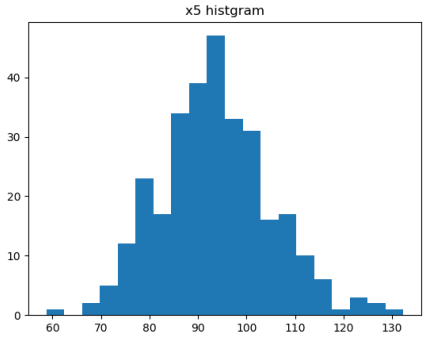
Variables histogram







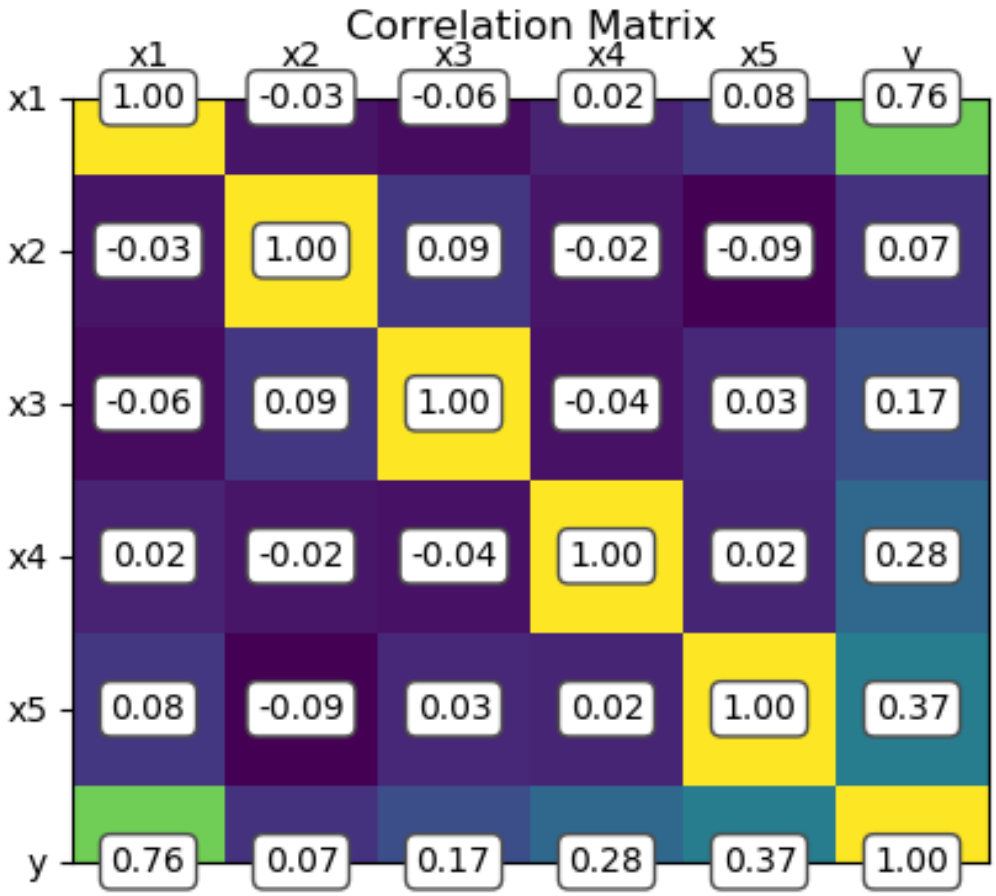




Variables mean and variance

	Mean	Variance
x1	12.706961	128.748759
x2	33.094677	131.084286
x3	54.805887	140.669748
x4	72.772503	151.835022
x5	93.539193	130.266234

Correlation Matrix



Comment

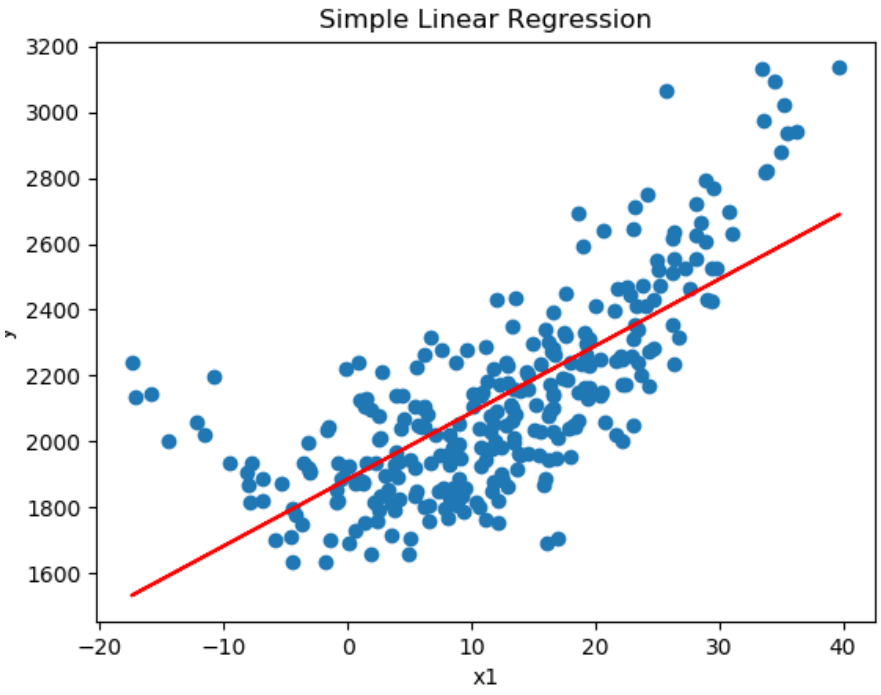
Based on the basic analysis above, we can know that all the variables are normal distributed with different mean and variance. The correlation matrix shows that all the variables are independent, and x1 has higher correlation with y comparing to other variables.

Task 2

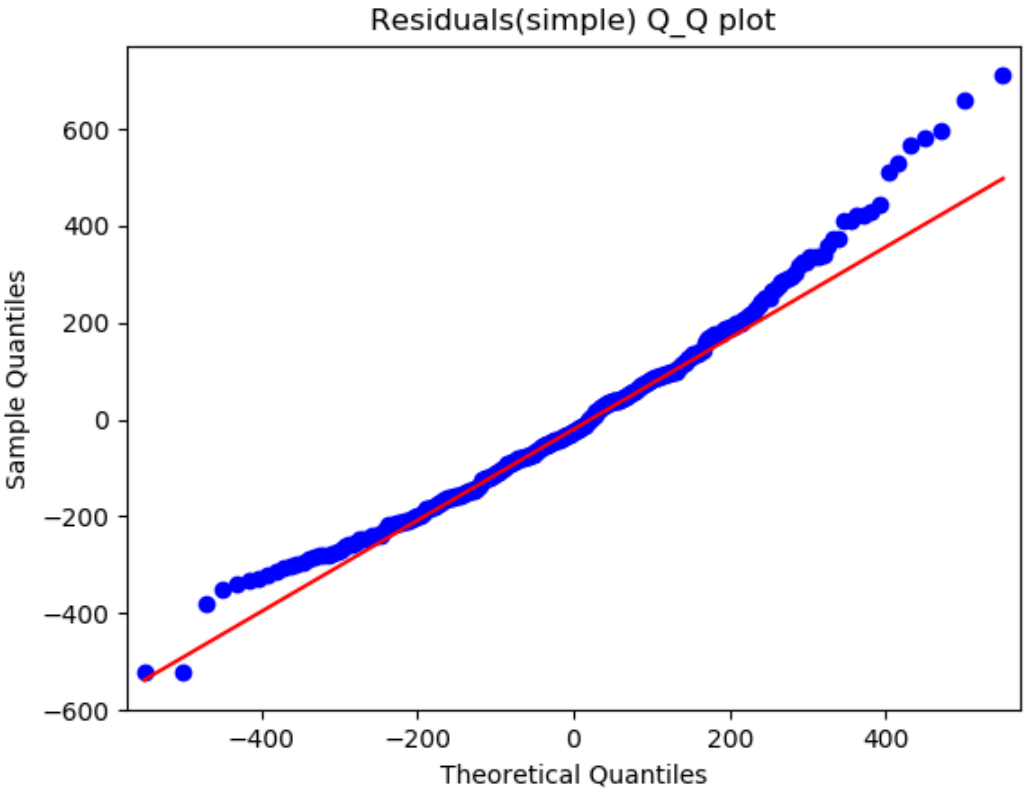
Summary of simple linear regression

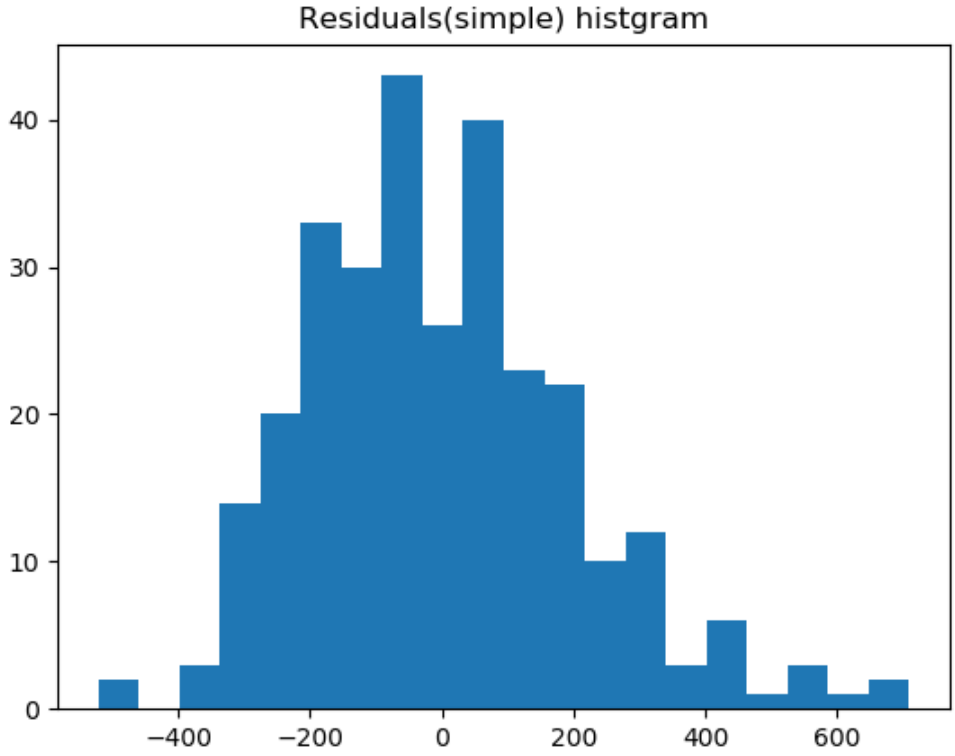
```
Simple linear regression....
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.538
Model:                            OLS   Adj. R-squared:                  0.537
Method:                 Least Squares   F-statistic:                     340.4
Date:                Tue, 22 Oct 2019   Prob (F-statistic):           6.34e-51
Time:                        16:33:21   Log-Likelihood:                -1979.7
No. Observations:                 294   AIC:                             3963.
Df Residuals:                     292   BIC:                             3971.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1884.7648     17.898    105.307      0.000    1849.540    1919.990
x1            20.2598      1.098     18.450      0.000      18.099      22.421
==============================================================================
Omnibus:                       19.994   Durbin-Watson:                   2.029
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               22.427
Skew:                           0.606   Prob(JB):                     1.35e-05
Kurtosis:                       3.601   Cond. No.                         24.6
==============================================================================
```

Plot of simple linear regression:

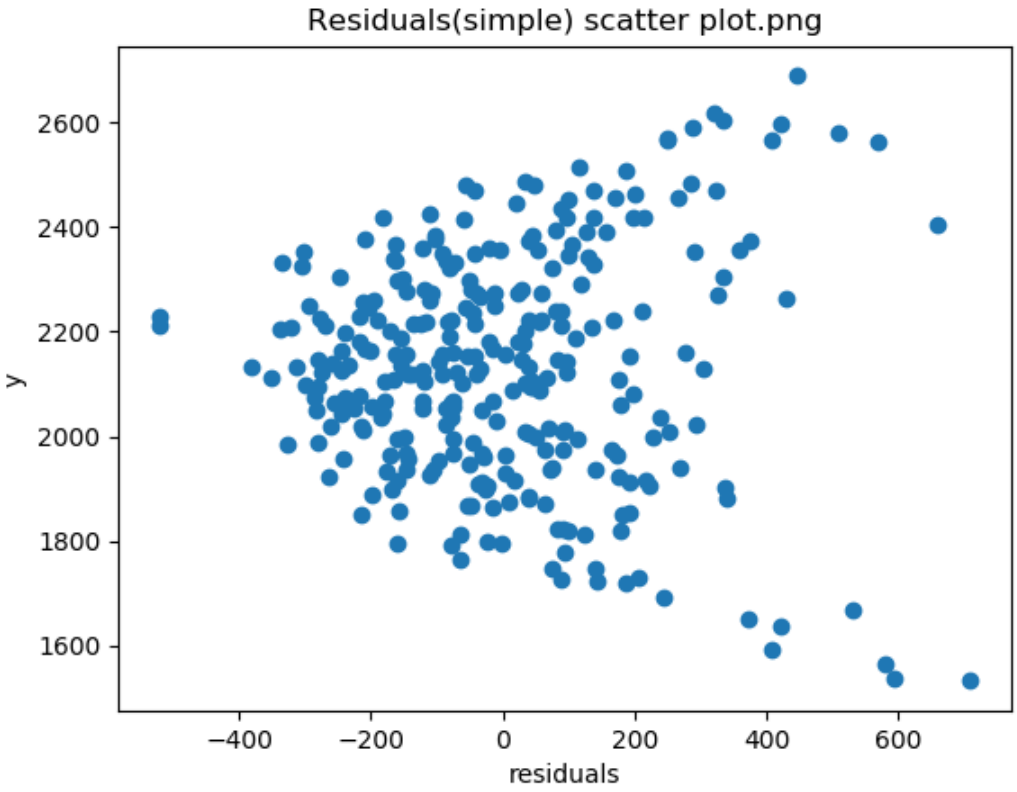


Residual analysis of simple linear regression





```
Running chisquare test.........
We have p-value is 4.554051754991799e-05
Significant Result, the null hypothesis rejected
```



Comment

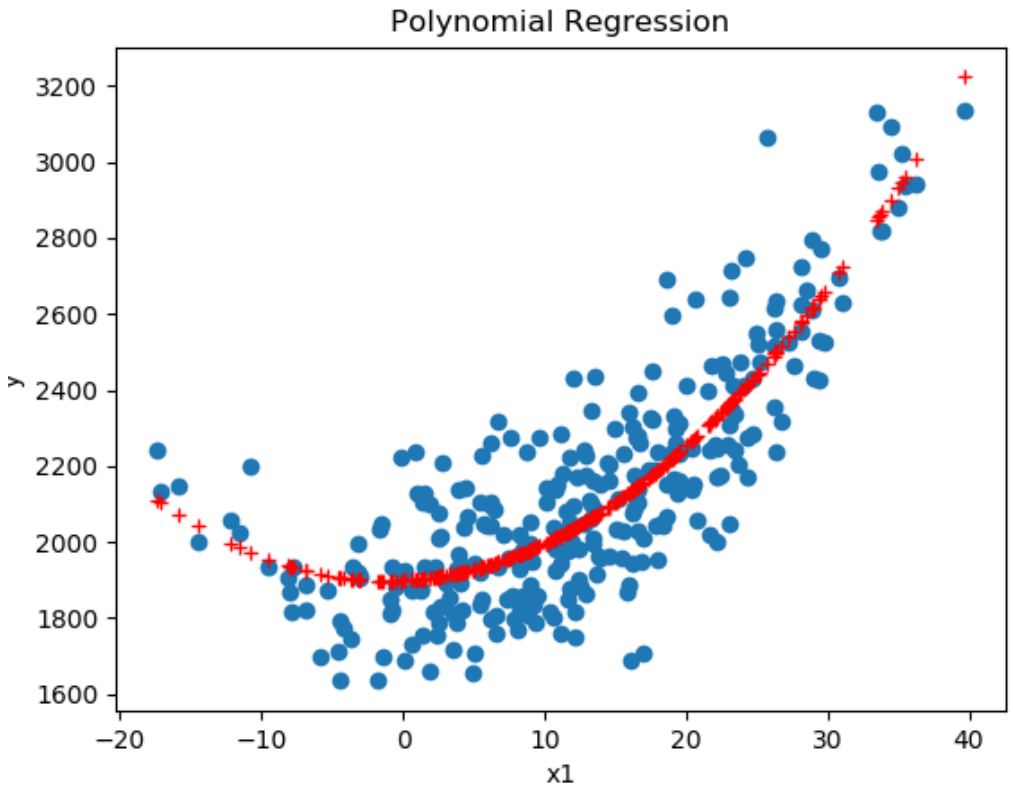
First, we remove the outliers if z-score of data in that independent variable is greater than 3. In other word, we only keep the data which is within 3 standard deviation. Next, based on the results of simple linear regression, we have the p-values of coefficients which is less the 95% confidence. Consequently, we reject the null hypothesis that the coefficients a_0 and a_1 are zero.

For residual analysis, both left tail and right tail of distribution of the residuals does not match that of the normal distribution. We cannot make sure the residuals if normal distributed. After we run the chi-square test, we reject the null hypothesis which is residuals are normal distributed. This makes some sense since the R-squared only 0.538. which is not very good fit. Finally, there is no discernible trend in the residuals scatter plot.

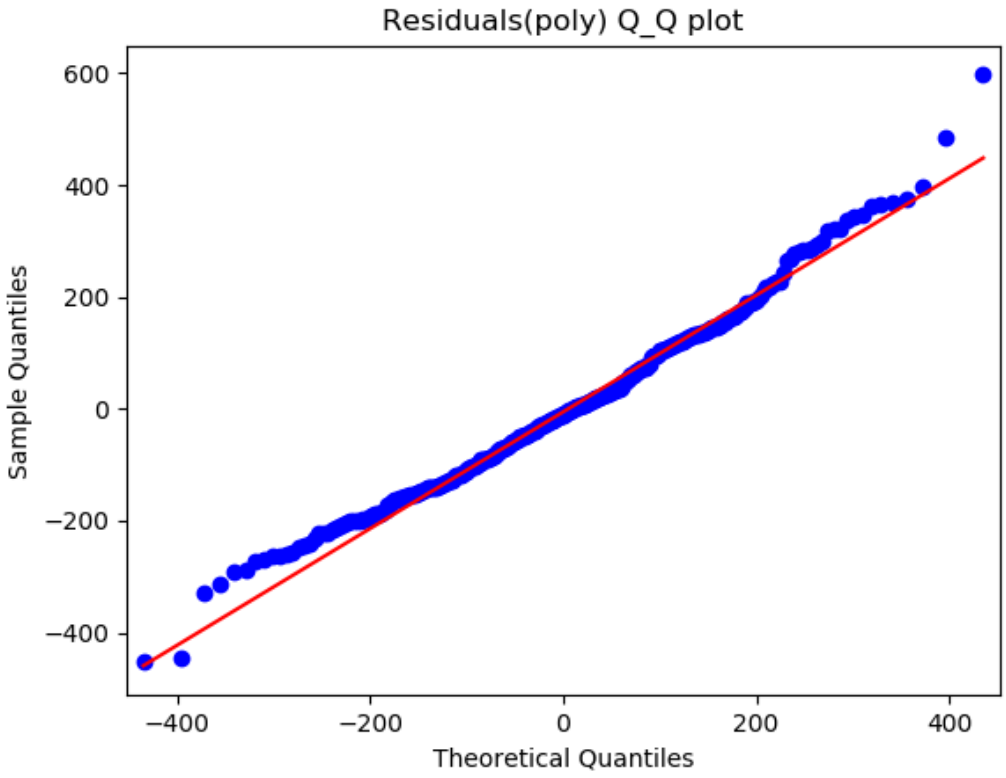
Summary of polynomial regression

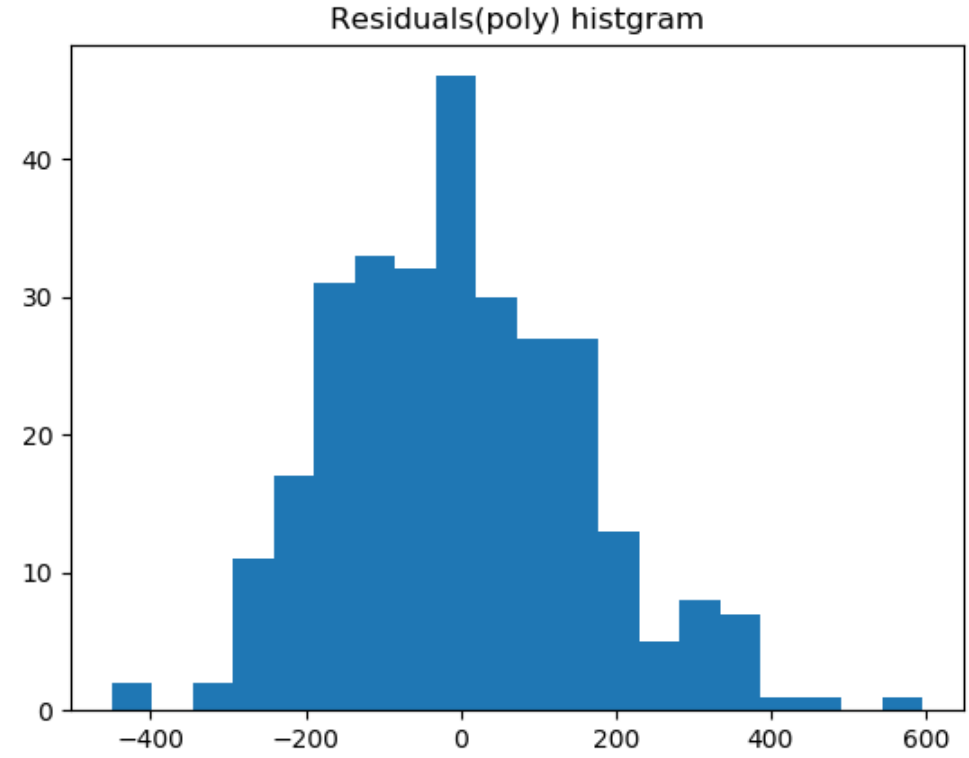
```
Polynomial Regression....
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.711
Model:                            OLS   Adj. R-squared:                  0.709
Method:                 Least Squares   F-statistic:                     358.3
Date:                Tue, 22 Oct 2019   Prob (F-statistic):           3.31e-79
Time:                        16:33:22   Log-Likelihood:                -1910.7
No. Observations:                 294   AIC:                             3827.
Df Residuals:                     291   BIC:                             3838.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1899.2724     14.222    133.543      0.000    1871.281    1927.264
x1             1.7260      1.652      1.045      0.297      -1.525       4.977
x1^2           0.7950      0.060     13.199      0.000       0.676       0.914
==============================================================================
Omnibus:                        9.658   Durbin-Watson:                   2.089
Prob(Omnibus):                  0.008   Jarque-Bera (JB):                9.745
Skew:                           0.398   Prob(JB):                      0.00765
Kurtosis:                       3.402   Cond. No.                         603.
==============================================================================
```

Plot of polynomial regression:

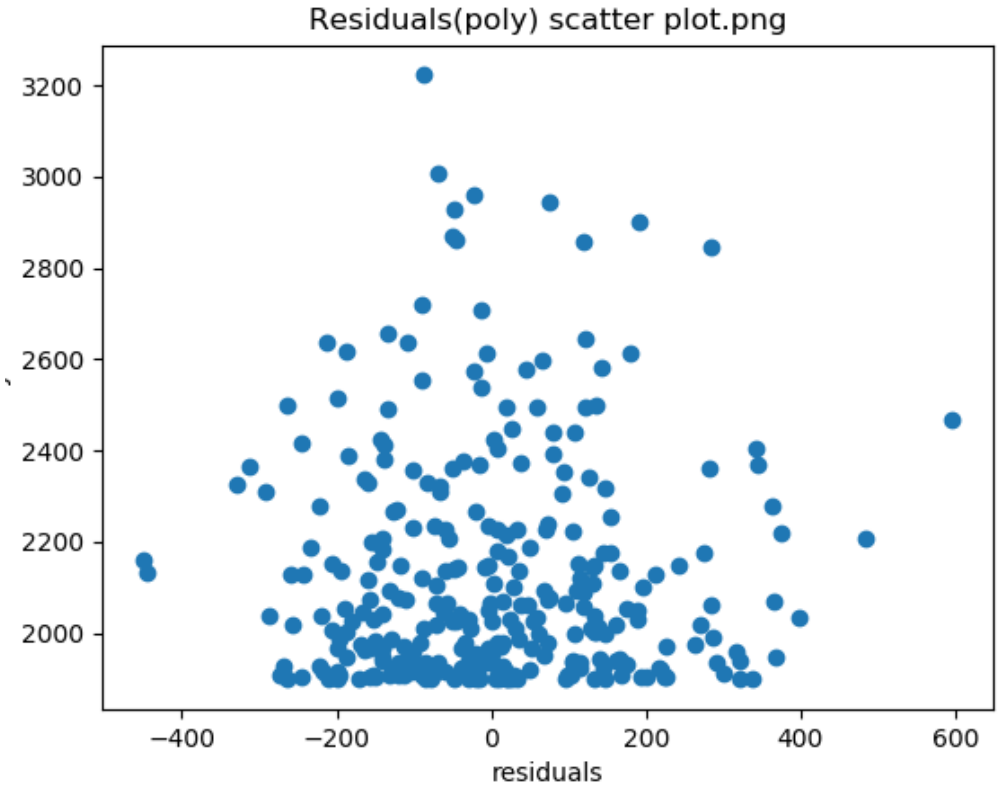


Residual analysis of polynomial regression





```
Running chisquare test.........
We have p-value is 0.007996440656196195
Significant Result, the null hypothesis rejected
```



Comment

The results of polynomial regression show that the model has better fit comparing to simple linear regression. The R-squared, Q-Q plot have better scores than simple linear regression, and there is no discernible trend in the residuals scatter plot as well. However, the a_1 coefficients are not significant in this model and the residuals still not normal distributed even though the overall result is improved.

Task 3

Summary of multivariable regression

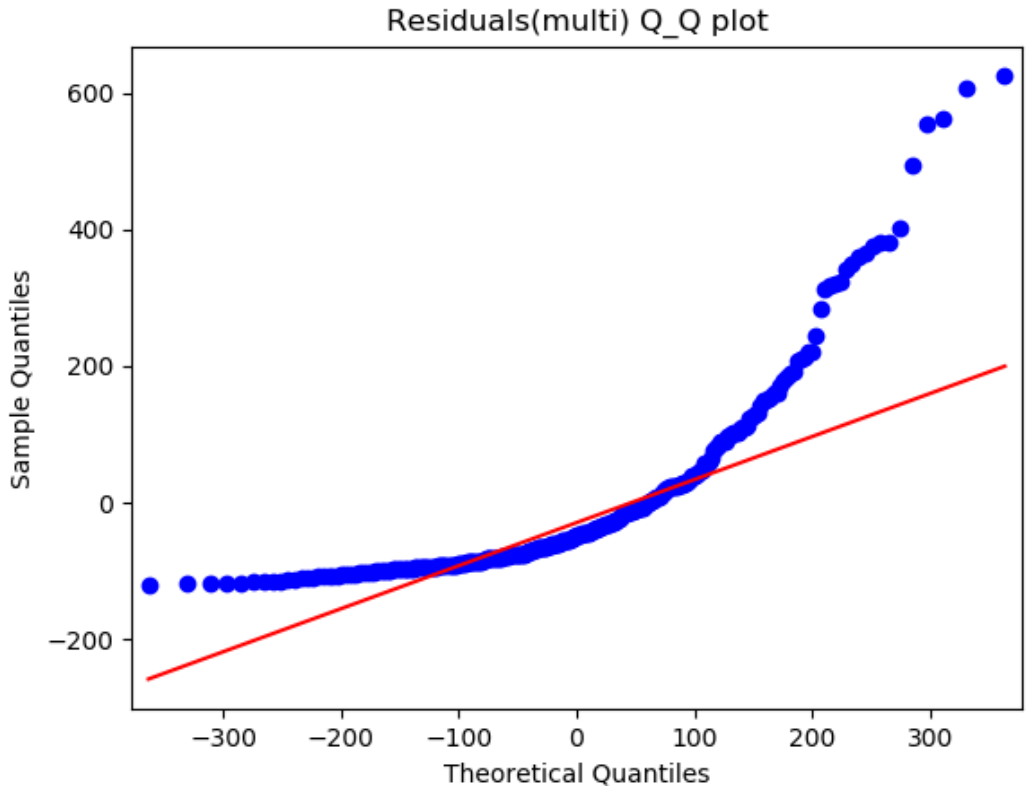
OLS Regression Results

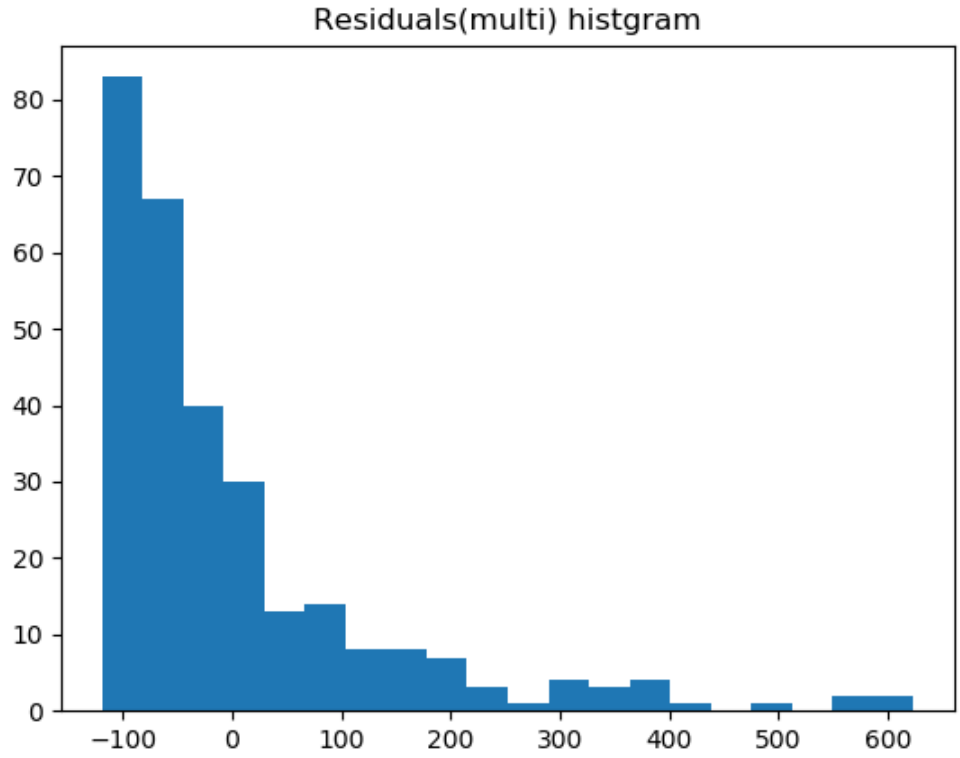
Dep. Variable:	y	R-squared:	0.792
Model:	OLS	Adj. R-squared:	0.788
Method:	Least Squares	F-statistic:	216.7
Date:	Tue, 22 Oct 2019	Prob (F-statistic):	7.23e-95
Time:	16:33:23	Log-Likelihood:	-1838.9
No. Observations:	291	AIC:	3690.
Df Residuals:	285	BIC:	3712.
Df Model:	5		
Covariance Type:	nonrobust		

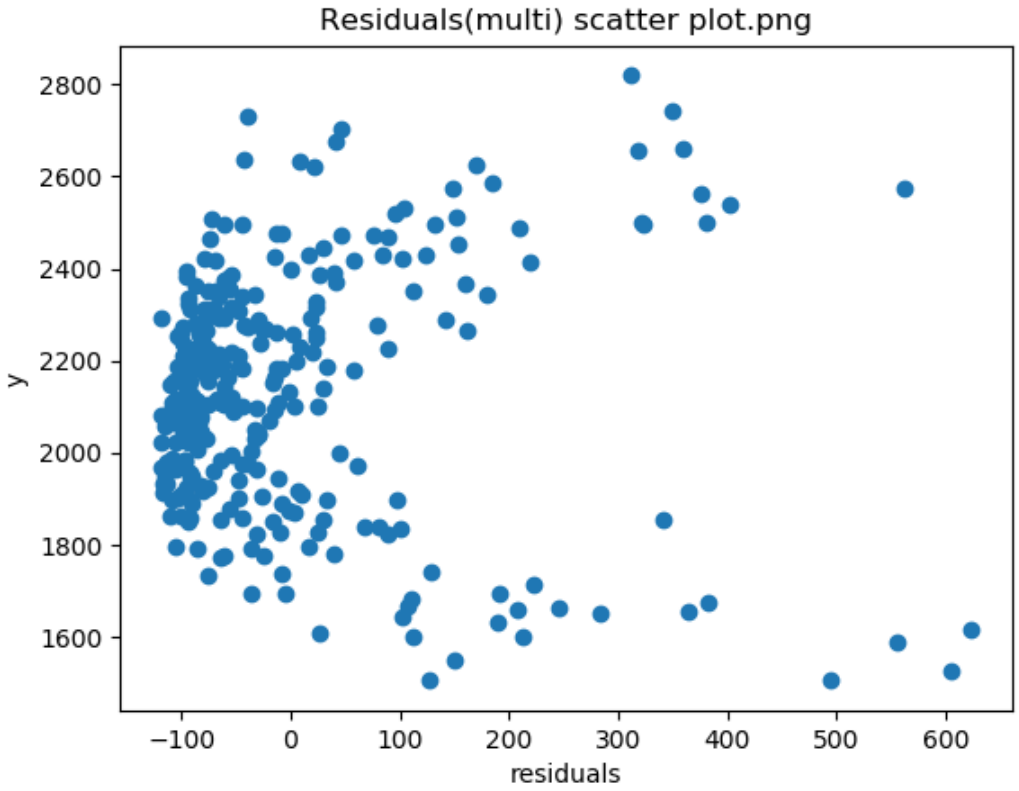
	coef	std err	t	P>|t|	[0.025	0.975]
const	154.3212	94.911	1.626	0.105	-32.494	341.136
x1	20.0759	0.739	27.149	0.000	18.620	21.531
x2	2.7251	0.705	3.866	0.000	1.338	4.113
x3	5.4953	0.686	8.008	0.000	4.145	6.846
x4	7.3586	0.645	11.416	0.000	6.090	8.627
x5	8.6291	0.739	11.675	0.000	7.174	10.084

Omnibus:	143.109	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	555.990
Skew:	2.175	Prob(JB):	1.86e-121
Kurtosis:	8.190	Cond. No.	1.62e+03

Residual analysis of multivariable regression




Residuals(multi) histgram
80
70
60
50
40
30
20
10
0
−100
0
100
200
300
400
500
600


Residuals(multi) scatter plot.png
2800
2600
2400
2200
2000
1800
1600
y
−100
0
100
200
300
400
500
600
residuals

```
Running chisquare test.........
We have p-value is 8.398302150686801e-32
Significant Result, the null hypothesis rejected
```

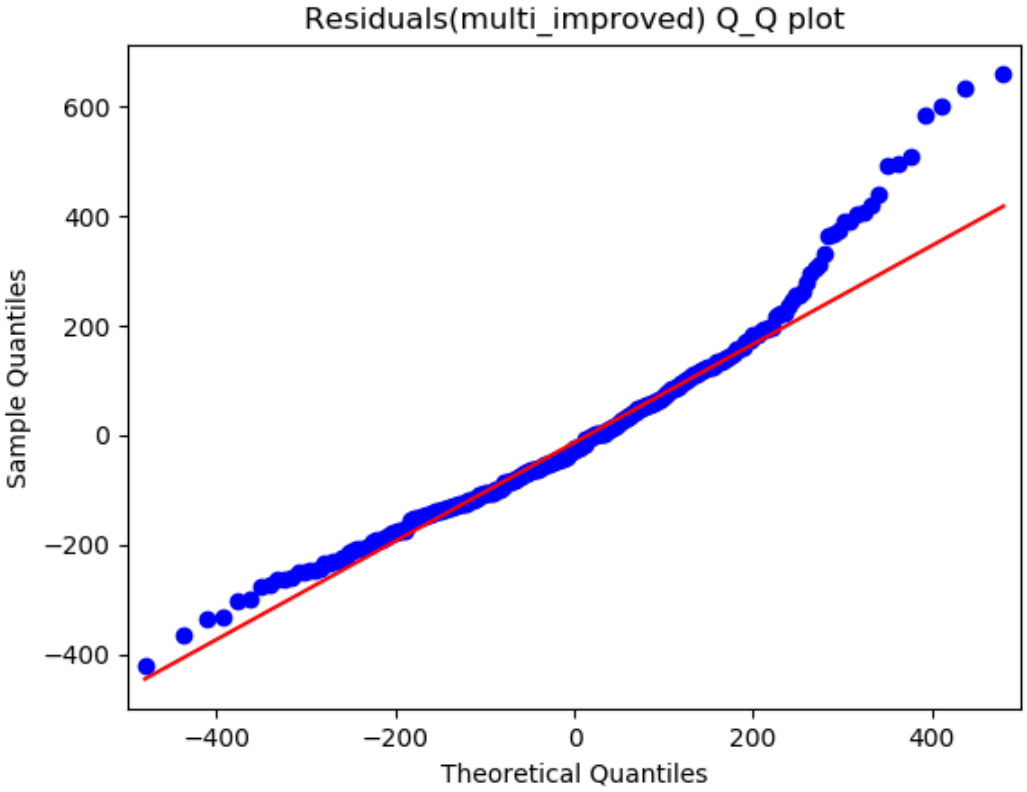
Comment

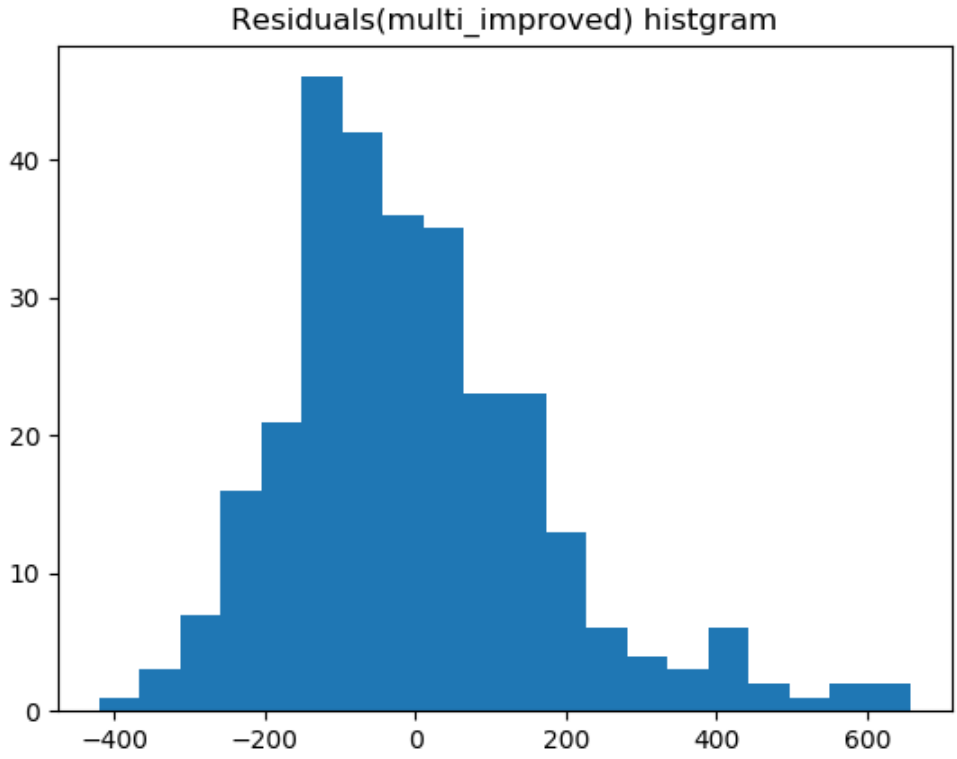
If we simply used all the independent variables, the residuals will strongly non-normal distributed. We can improve it by choose the observed strong correlation between Y and variables. In our case, we can simply only used x_1, which has the highest correlation with Y and the remaining variables are all under 0.5, then the result will be same as simple linear regression. However, in order to make them slightly different, we choose the first and second high correlation, which is x_1 and x_5, and re-do the multivariable regression.

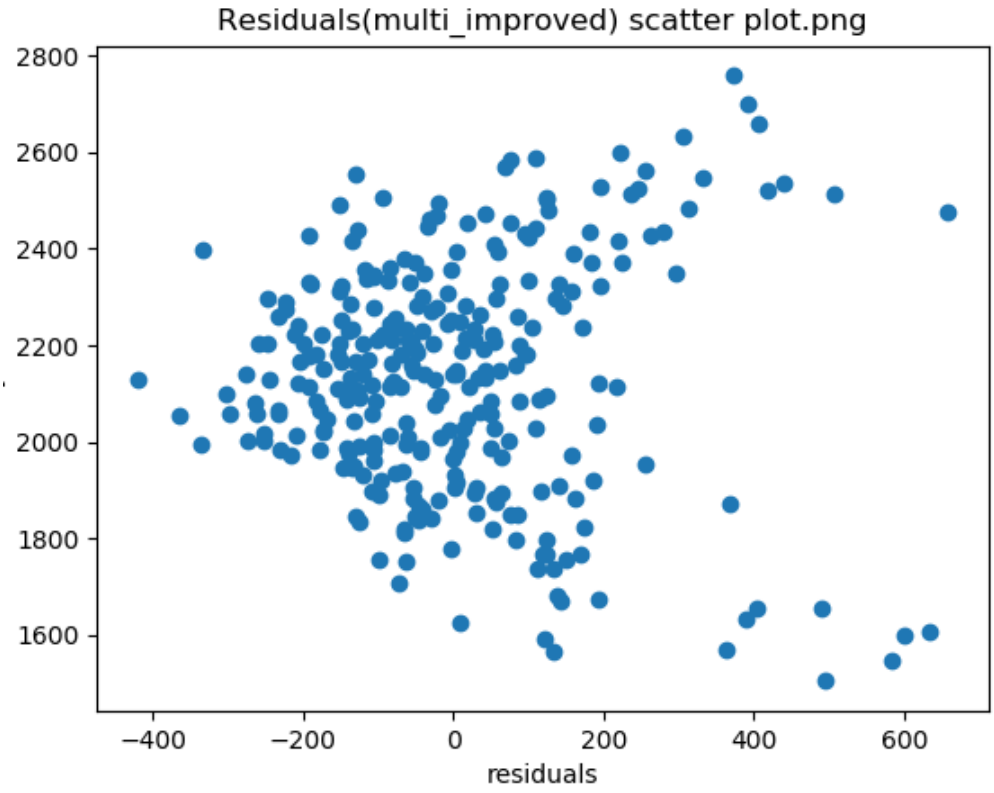
Summary of improved multivariable regression

```
Better attempt multivariate regression...
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.644
Method:                 Least Squares   F-statistic:                     264.8
Date:                Tue, 22 Oct 2019   Prob (F-statistic):           4.62e-66
Time:                        21:06:37   Log-Likelihood:                -1926.7
No. Observations:                 292   AIC:                             3859.
Df Residuals:                     289   BIC:                             3871.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1036.9900     90.287     11.486      0.000     859.287    1214.693
x1            19.7136      0.965     20.423      0.000      17.814      21.613
x5             9.1505      0.958      9.549      0.000       7.264      11.037
==============================================================================
Omnibus:                       45.366   Durbin-Watson:                   1.863
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               68.867
Skew:                           0.940   Prob(JB):                     1.11e-15
Kurtosis:                       4.458   Cond. No.                         820.
==============================================================================
```

Residual analysis of multivariable regression







Comment

The residual results are improved after the better feature selection; however, we lost some data presentation to our model.