Project 2

Task 1

1.1) For mean, variance, histogram

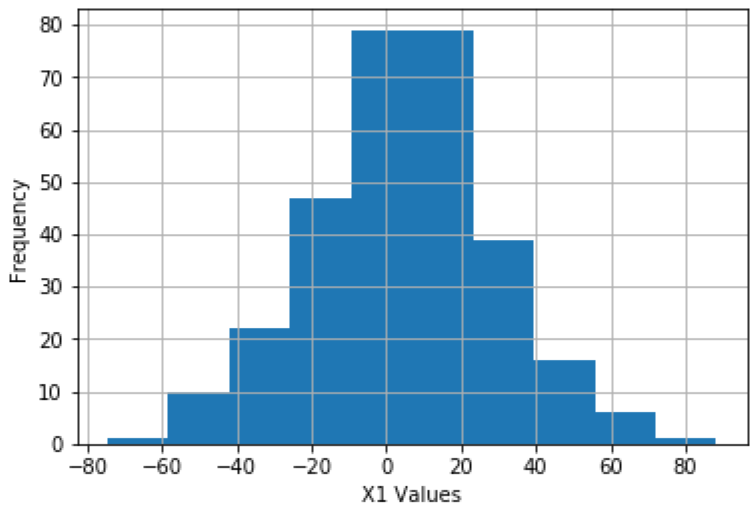
Mean:

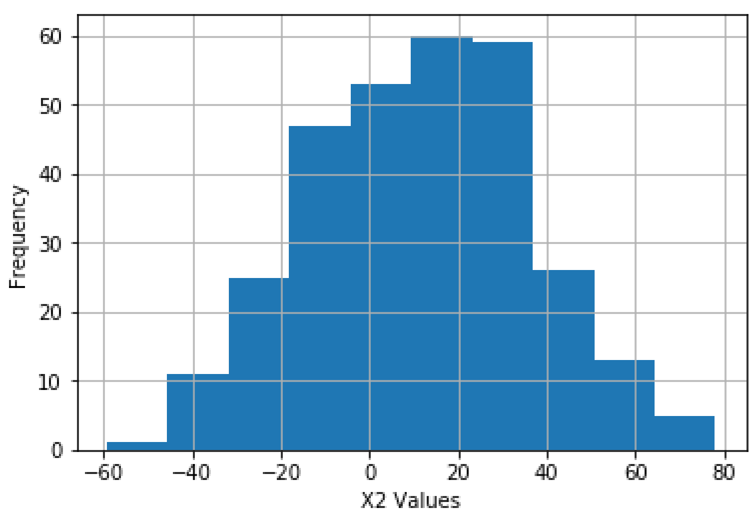
X1	4.048153
X2	11.602439
X3	17.769881
X4	17.831539
X5	29.524337
Y	1391.236605

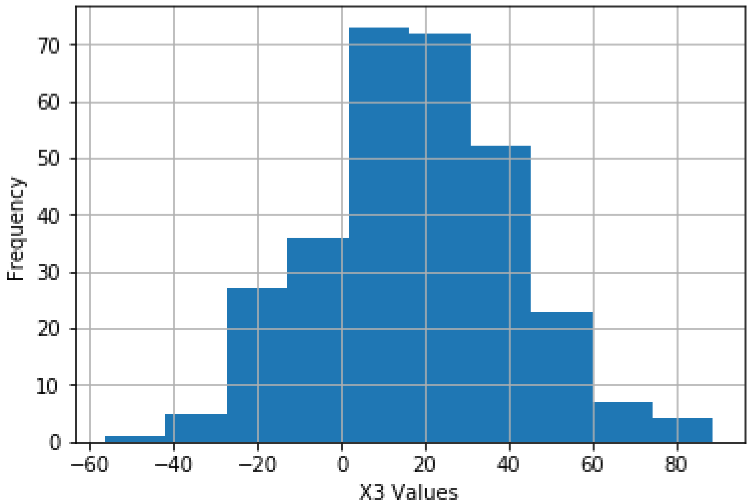
Variance:

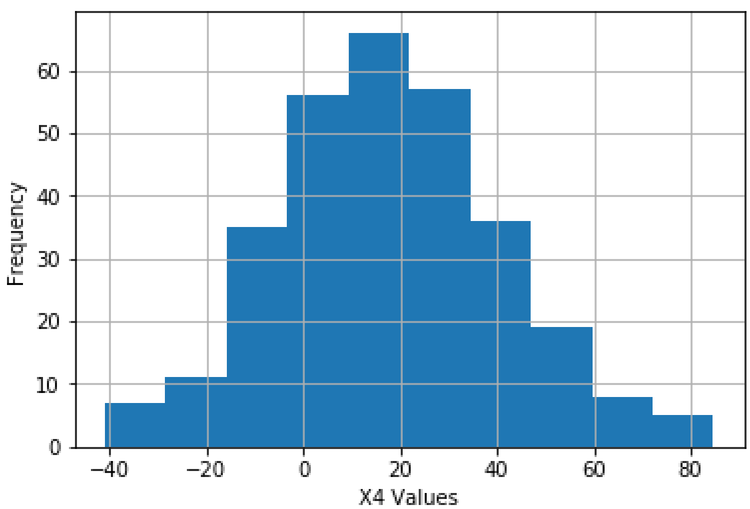
X1	6.222418e+02
X2	6.215748e+02
X3	5.497936e+02
X4	5.360255e+02
X5	5.246722e+02
Y	3.117595e+06

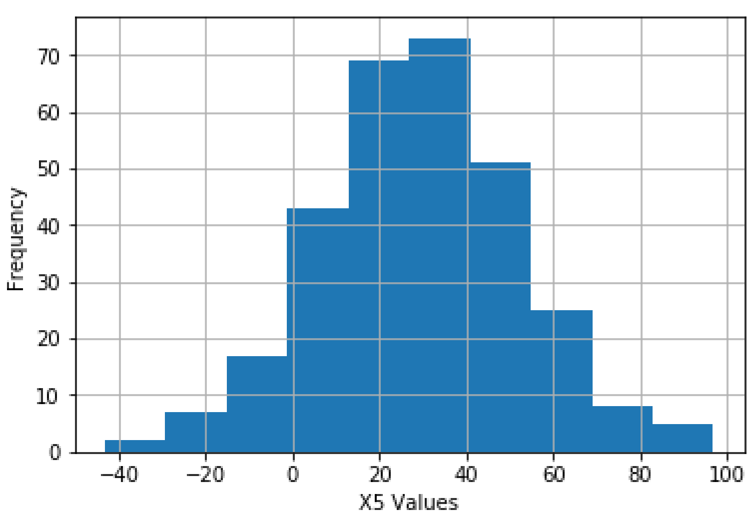
Histogram



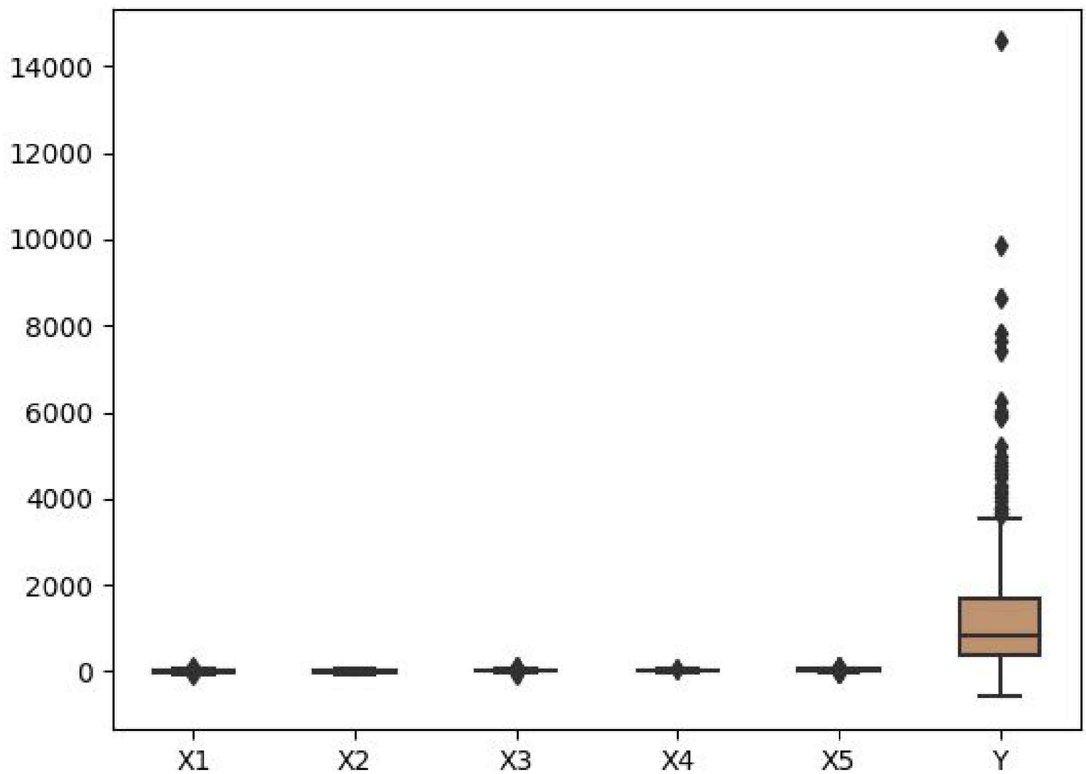



70
60
50
40
30
20
10
0
Frequency
−60
−40
−20
0
20
40
60
80
X3 Values


60
50
40
30
20
10
0
Frequency
−40
−20
0
20
40
60
80
X4 Values


70
60
50
40
30
20
10
0
Frequency
−40
−20
0
20
40
60
80
100
X5 Values

1.2) Box Plot



Although there are outliers in the data, we are going to go ahead and skip removing outliers as they are not many.

1.3) Corr(X, Y) = r_{xy} = Cov(X, Y)/($\sqrt{var(X)\ .\ var(Y)}$)
It can be seen that, -1 <= r_{xy} <= 1
X and Y are positively correlated if $r_{xy} > 0$
and they are strongly positively correlated if $r_{xy} = 1$

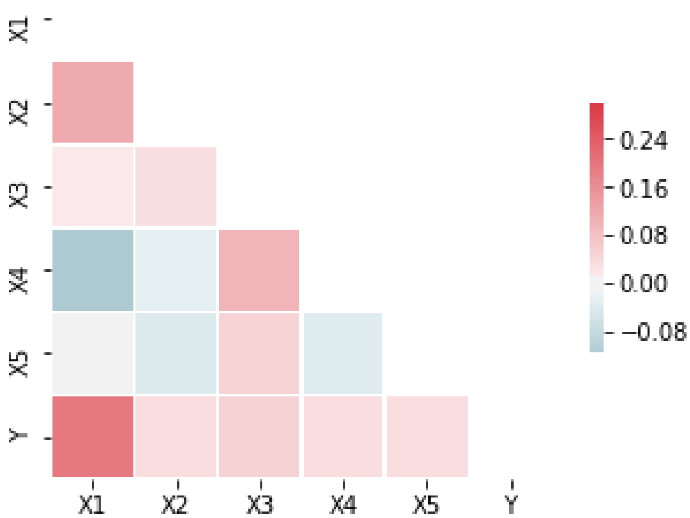
X and Y are negatively correlated if $r_{xy} < 0$
and they are strongly negatively correlated if $r_{xy} = -1$

X and Y are uncorrelated if $r_{xy} = 0$

Considering the above statements,

Correlation Matrix:

	X1	X2	X3	X4	X5	Y
X1	1.000000	0.114921	0.014953	-0.117389	-0.003737	0.198738
X2	0.114921	1.000000	0.032483	-0.023898	-0.038486	0.032932
X3	0.014953	0.032483	1.000000	0.100298	0.051097	0.052723
X4	-0.117389	-0.023898	0.100298	1.000000	-0.034567	0.035399
X5	-0.003737	-0.038486	0.051097	-0.034567	1.000000	0.034853
Y	0.198738	0.032932	0.052723	0.035399	0.034853	1.000000



Result Evaluation:

Assuming values in the range (-0.15, 0.15) ~ 0
X1 is positively correlated to X2, Y
X1 is negatively correlated to X4
X1 is uncorrelated to X3, X5

X2 is positively correlated to X1
X2 is somewhat negatively correlated to X3, X4, X5, Y

X2 is uncorrelated to X5

X3 is somewhat positively correlated to X4, X5
X3 is somewhat negatively correlated to X3, X4, X5, Y
X3 is uncorrelated to X5

X4 is uncorrelated to X5, Y
X5 is uncorrelated to Y

Heatmap:

X1 seems to be correlated to Y much more than other independent variables.
Therefore, Y seems to be dependent on X1 either in linear or polynomial manner. Although, we cannot conclude this by just using the correlation matrix.

Task 2

Simple linear regression where, Y = a0 + a1X1 + e
2.1)

H_0 = Regression coefficients are significant.
H_a = Regression coefficients are not significant

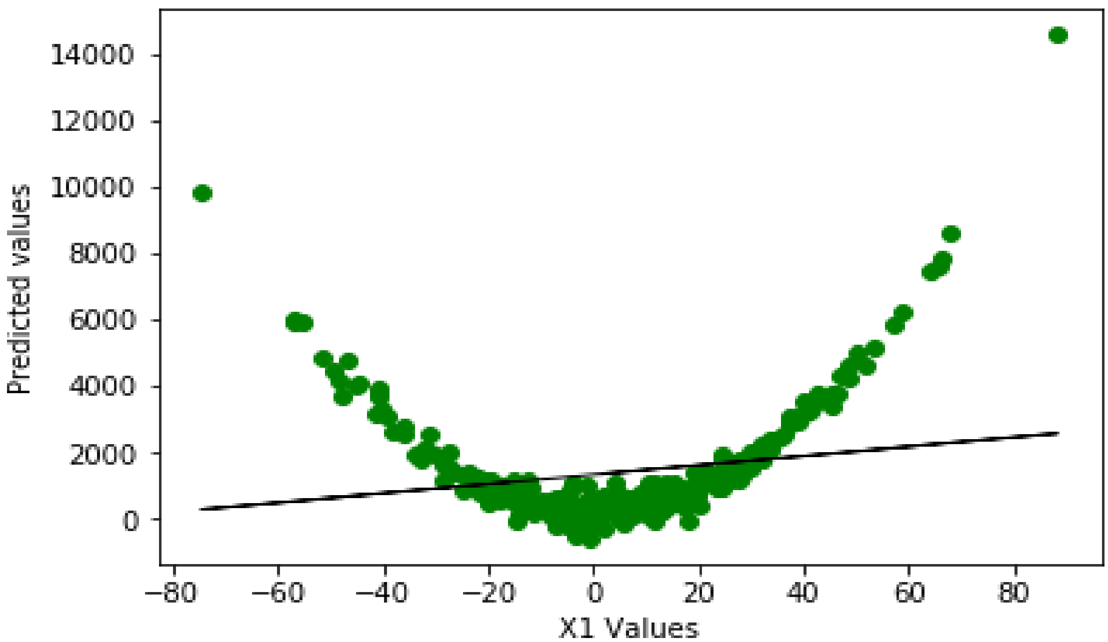
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.039
Model:                            OLS   Adj. R-squared:                  0.036
Method:                 Least Squares   F-statistic:                     12.25
Date:                Sun, 28 Oct 2018   Prob (F-statistic):           0.000535
Time:                        11:35:19   Log-Likelihood:                -2662.0
No. Observations:                 300   AIC:                             5328.
Df Residuals:                     298   BIC:                             5335.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1334.2900    101.389     13.160      0.000    1134.761    1533.819
X1            14.0673      4.019      3.501      0.001       6.159      21.976
==============================================================================
Omnibus:                      209.536   Durbin-Watson:                   2.249
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2058.836
Skew:                           2.847   Prob(JB):                         0.00
Kurtosis:                      14.502   Cond. No.                         25.6
==============================================================================
```

2.2)

p = 0.01 for X1
R^2 = 0.039
Adjusted R^2 = 0.036
F = 12.25
Since p and R^2 are very small and ~ 0, we can say that X1 has a normal distribution and regression coefficients are significant.

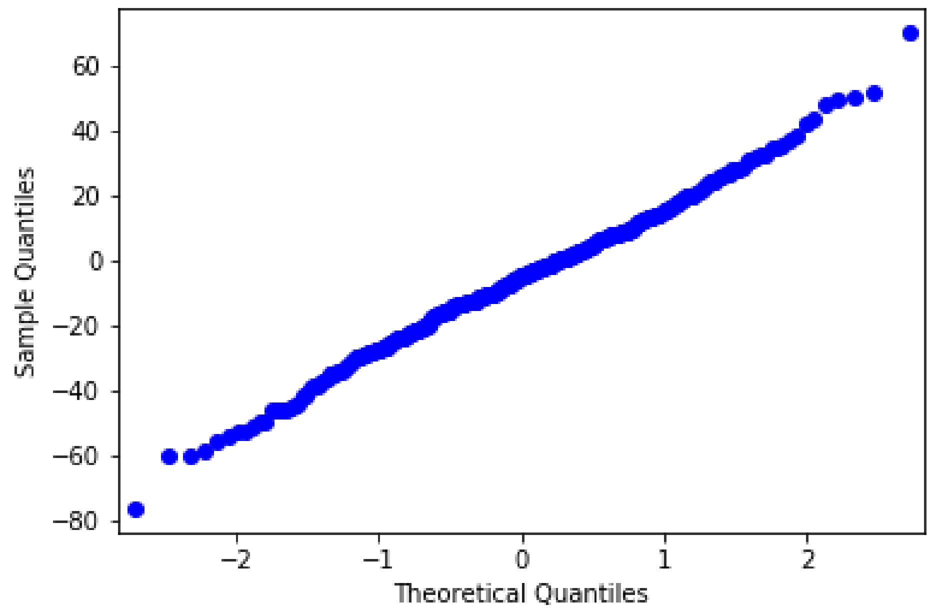
Thus, Null hypothesis accepted

2.3) Linear regression graph



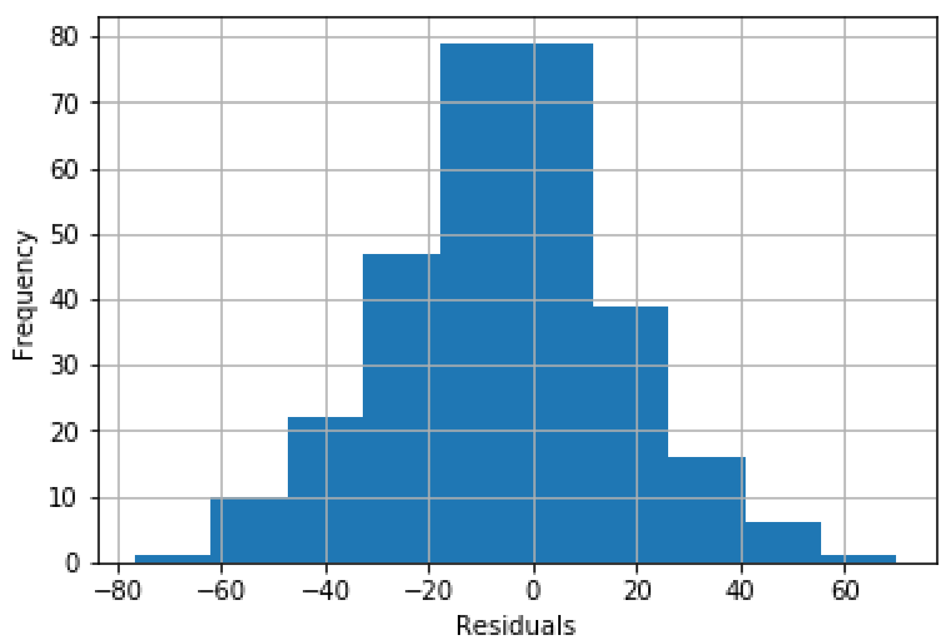
2.4)

a) QQplot between residuals and predicted Y values



We can clearly see from the QQ plot that residuals are following a normal distribution.

b) Residuals histogram



Based on the histogram, residuals are following a normal distribution.

Chi Square Test

H_0 = Residuals follow a normal distribution in N(0, s^2)
H_a = Residuals do not follow a normal distribution in N(0, s^2)

Chi square test is only applicable for categorical data with only positive values. Since we have negative values in our sample data set, we cannot use this test.
With the chi square and the degrees of freedoms (dof), we get the p-value to determine significance and result of our hypothesis.

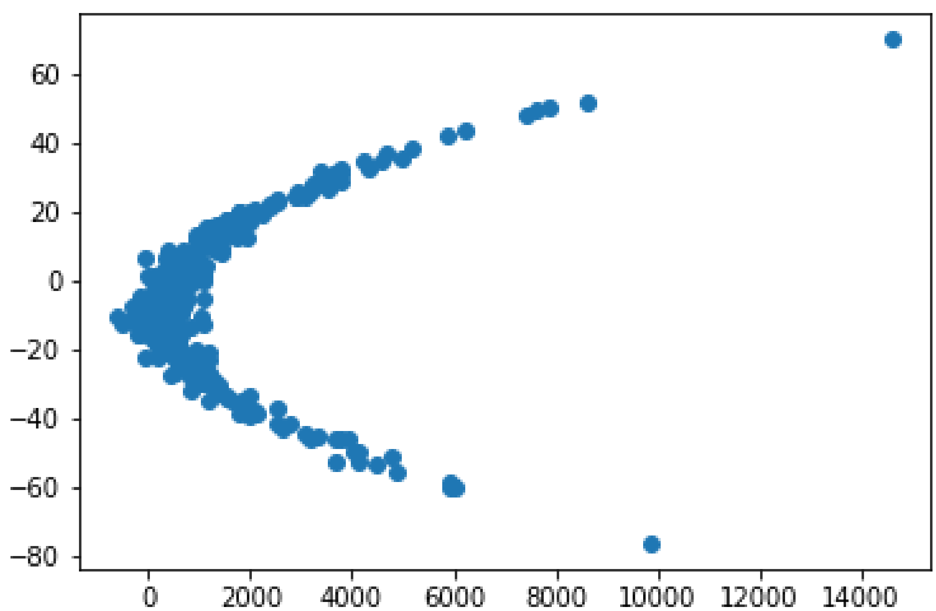
Normal test

H_0 = Residuals follow a normal distribution in N(0, s^2)
H_a = Residuals do not follow a normal distribution in N(0, s^2)

Using normal test, we obtain p = [0.40070655]

Depending on the test, as our p-value is above our threshold (0.05), we accept the null hypothesis or The null hypothesis cannot be rejected.
Thus, Residuals follow a normal distribution in N(0, s^2)

b) Scatter plot:

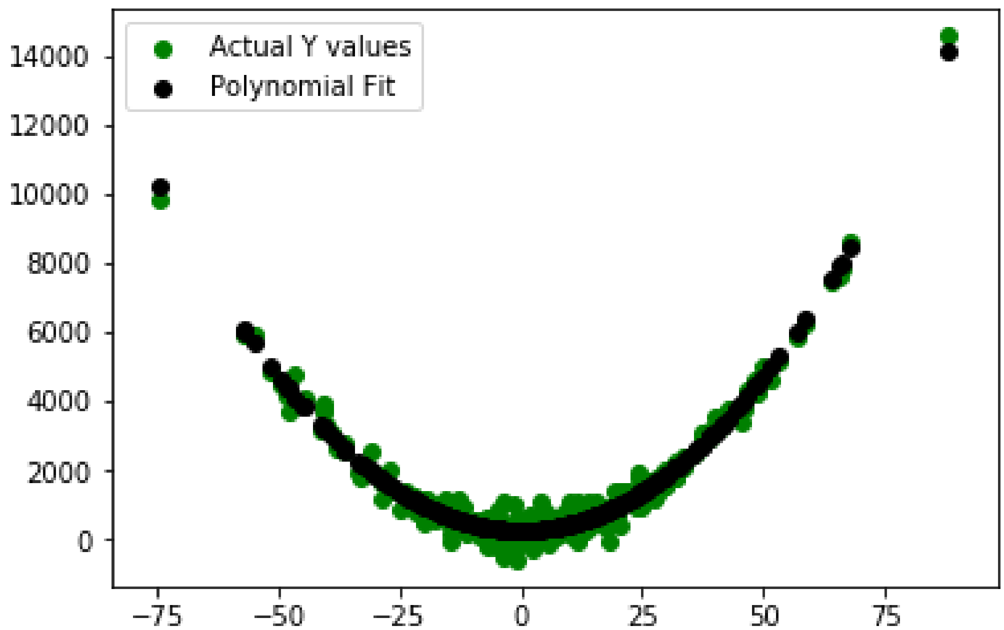


We can see that outliers exist and residuals are correlated. Outliers are influencing our distribution and errors are clearly visible here.

2.7) Using polynomial distribution of the form:

$$Y = a_0 + a_1X_1 + a_2X_1^2$$

We obtain the following scatter plot,



We can see that the polynomial equation is fitting our data set very well. Thus, we can say that our Y is dependent on feature X1.

Task 3

3.1) For values of all coefficients:

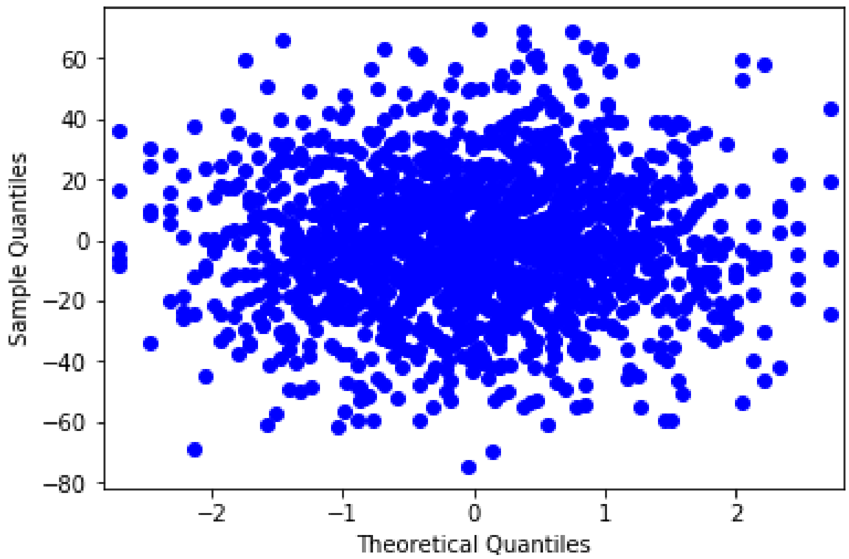
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.046
Model:                            OLS   Adj. R-squared:                  0.030
Method:                 Least Squares   F-statistic:                     2.854
Date:                Sun, 28 Oct 2018   Prob (F-statistic):             0.0156
Time:                        13:03:32   Log-Likelihood:                -2661.0
No. Observations:                 300   AIC:                             5334.
Df Residuals:                     294   BIC:                             5356.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1109.3913    201.097      5.517      0.000     713.620    1505.163
X1            14.4142      4.087      3.527      0.000       6.371      22.457
X2             0.7723      4.066      0.190      0.849      -7.230       8.775
X3             3.1487      4.321      0.729      0.467      -5.356      11.653
X4             4.3173      4.401      0.981      0.327      -4.345      12.979
X5             2.7638      4.403      0.628      0.531      -5.902      11.430
==============================================================================
Omnibus:                      214.174   Durbin-Watson:                   2.235
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2156.031
Skew:                           2.925   Prob(JB):                         0.00
Kurtosis:                      14.758   Cond. No.                         94.3
==============================================================================
```

3.2)

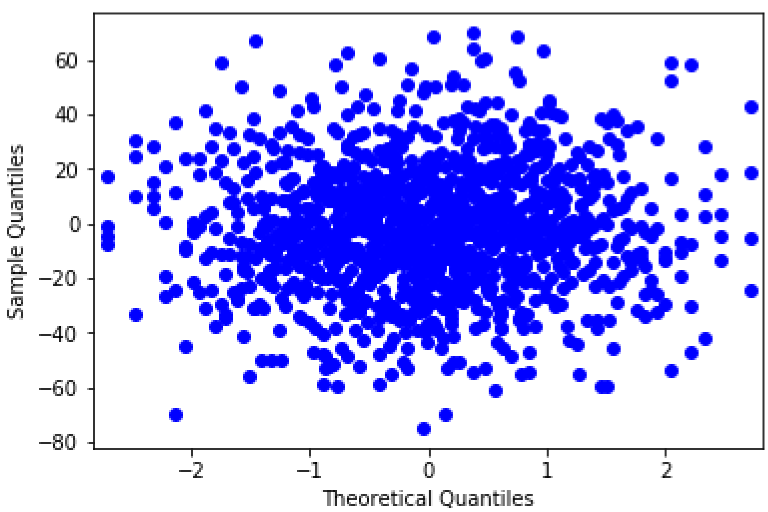
$R^2 = 0.046$
F = 2.854

QQ plot with multivariable regression

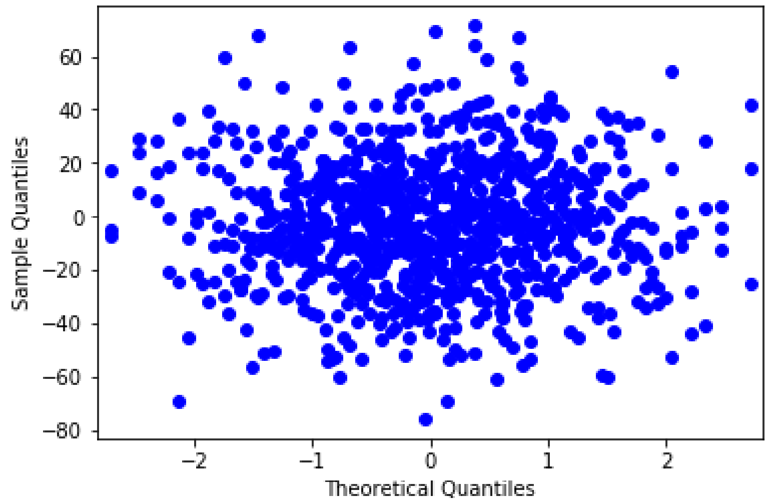
$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5$



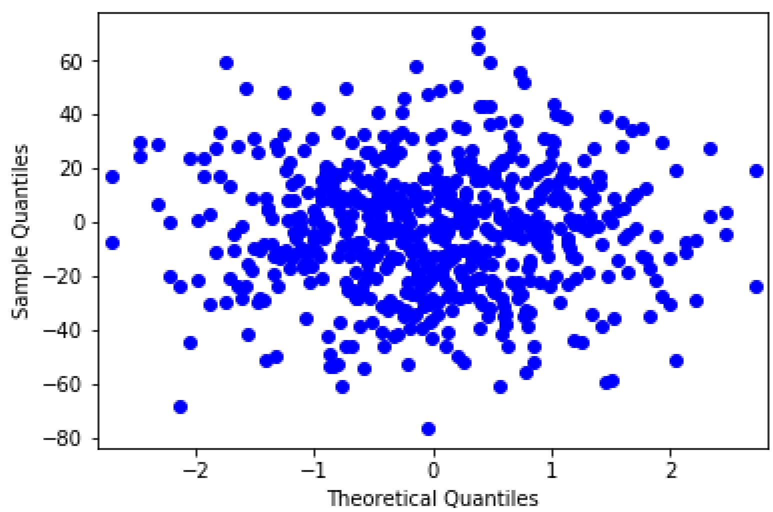
$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4$



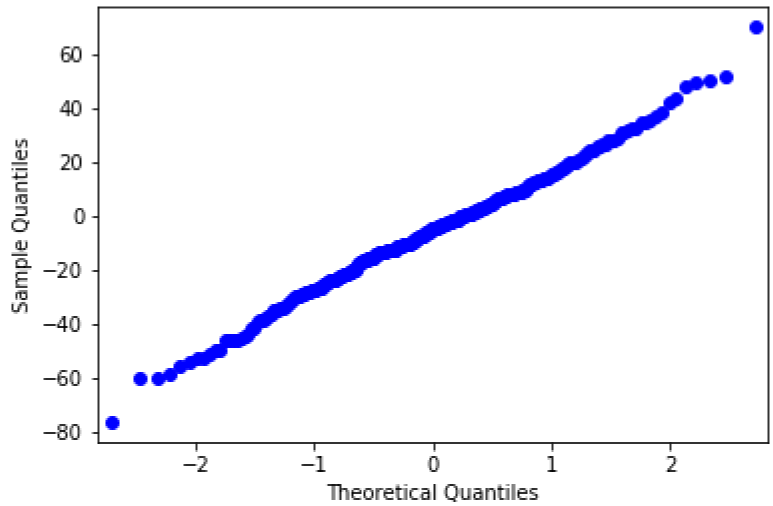
$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3$



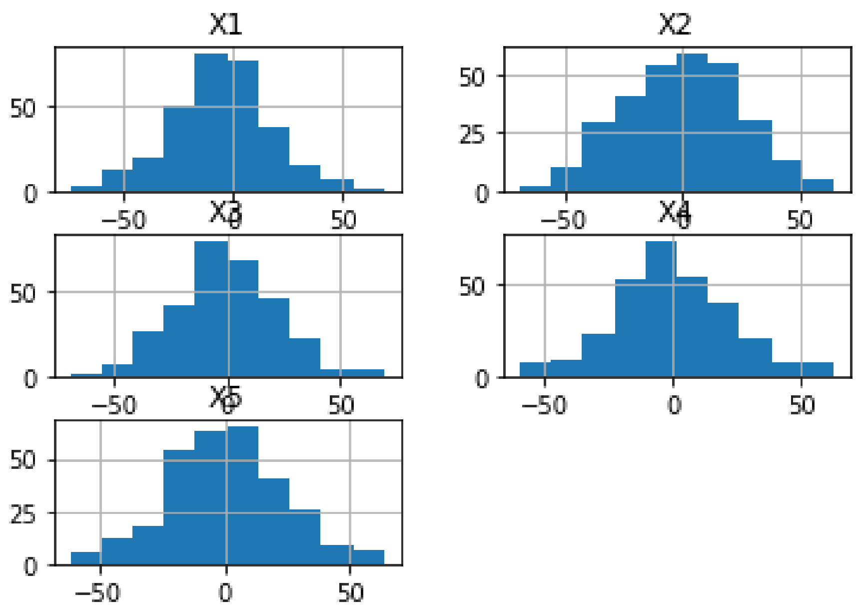
$Y = a_0 + a_1X_1 + a_2X_2$



$Y = a_0 + a_1X_1$



Histogram



We can see that, although the histograms look normally distributed, the scatter plots of residuals against Y, do not indicate a normal distribution. Only for residuals of X1 and Y are normally distributed.

Chi Square Test

H_0 = Residuals follow a normal distribution in N(0, s^2)
H_a = Residuals do not follow a normal distribution in N(0, s^2)

Chi square test is only applicable for categorical data with only positive values. Since we have negative values in our sample data set, we cannot use this test.
With the chi square and the degrees of freedoms (dof), we get the p-value to determine significance and result of our hypothesis.

Normal test

H_0 = Residuals follow a normal distribution in N(0, s^2)
H_a = Residuals do not follow a normal distribution in N(0, s^2)

Using normal test, we obtain,

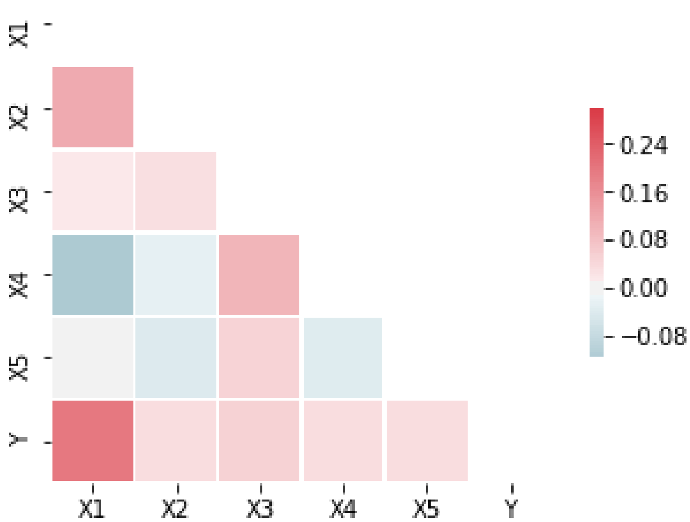
Variable	X1	X2	X3	X4	X5
p	0.40070655	0.59183702	0.53158877	0.4414195	0.4509979

Depending on the test, as our p-value is above our threshold (0.05), we accept the null hypothesis or The null hypothesis cannot be rejected.

Thus, All residuals follow a normal distribution in N(0, s^2)

Correlation Matrix:

	X1	X2	X3	X4	X5	Y
X1	1.000000	0.114921	0.014953	-0.117389	-0.003737	0.198738
X2	0.114921	1.000000	0.032483	-0.023898	-0.038486	0.032932
X3	0.014953	0.032483	1.000000	0.100298	0.051097	0.052723
X4	-0.117389	-0.023898	0.100298	1.000000	-0.034567	0.035399
X5	-0.003737	-0.038486	0.051097	-0.034567	1.000000	0.034853
Y	0.198738	0.032932	0.052723	0.035399	0.034853	1.000000



Based on our correlation matrix, X2, X3, X4, X5 all independent variables can be removed and we can obtain a good fit in X1 as we saw in task 2 for polynomial fit.

Comments

Using multivariable regression obtains R^2 = 0.046 as compared to linear regression using X1 obtains R^2 = 0.039. Thus, we can say that the data Y is dependent on X1.
According to polynomial equation with degree = 2, we obtain a perfect fit for the data.

Although all residuals are normally distributed, we can see that the independent variables are not strongly correlated to Y or each other except for X1 and Y.

Thus, we should not pick multivariable equation for fitting our dataset as it does not give us the best fit.