Priya Diwakar
200205361

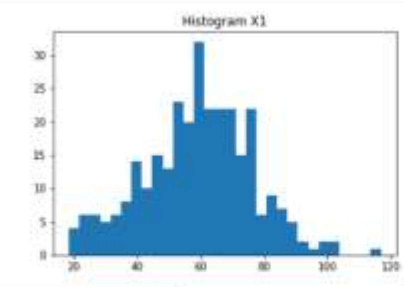
IOT Analytics

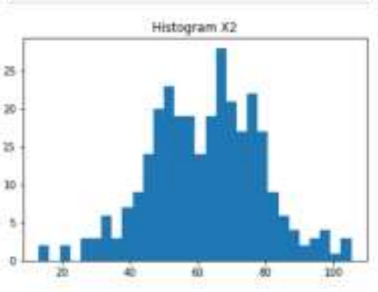
Project 2 Regression

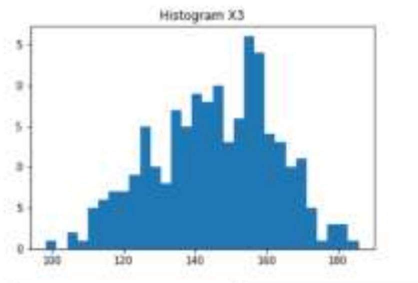
1. Task 1 Analysis of the Variables

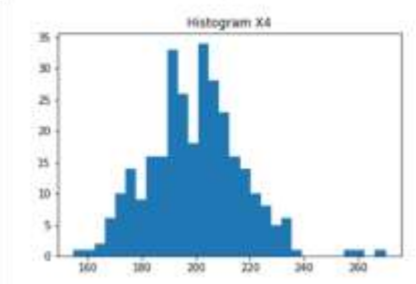
1.1.

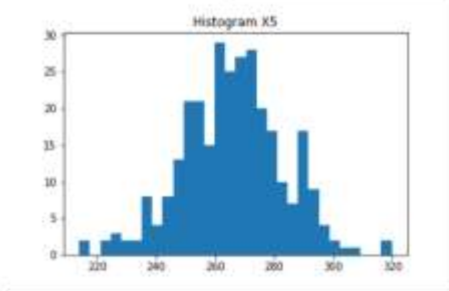
a. Histogram plots of the Variables:











b. Mean of the Variables:

X1 mean: 59.26281
X2 mean: 62.1877166667
X3 mean: 145.5158
X4 mean: 199.822333333
X5 mean: 265.965166667

c. Variance of the Variables:

X1 variance: 291.991919761
X2 variance: 265.69593293
X3 variance: 283.862741693
X4 variance: 295.077073222
X5 variance: 294.332428306

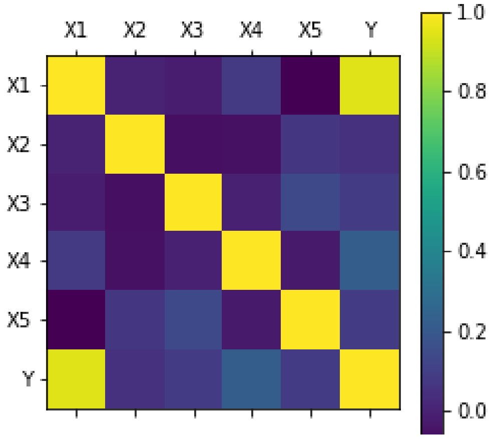
1.2. Outliers:

Z score has been used to remove outliers, resulting in 8 values being found as outliers and removed.

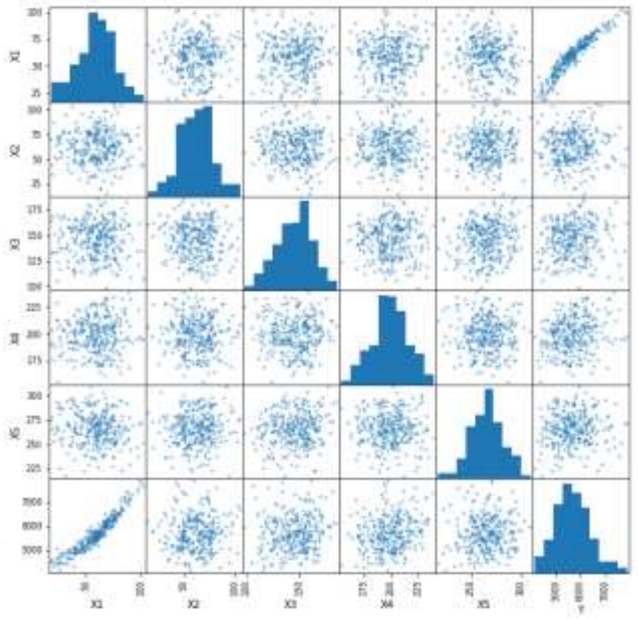
1.3. Correlation Matrix:

From the correlation matrix a strong positive linear relationship is observed between X1 and Y. X4 has weaker relation with Y compared to X1. X2, X3, X5 almost negligible relation to Y. Even among the variables there is almost negligible dependence. The variables could be assumed as independent. This is also evident from the heat map and scatter plot.

	X1	X2	X3	X4	X5	Y
X1	1	0.004753	-0.02079	0.083499	-0.10381	0.948151
X2	0.004753	1	-0.0573	-0.05496	0.069588	0.05167
X3	-0.02079	-0.0573	1	-0.00104	0.139087	0.092948
X4	0.083499	-0.05496	-0.00104	1	-0.02507	0.223321
X5	-0.10381	0.069588	0.139087	-0.02507	1	0.087266
Y	0.948151	0.05167	0.092948	0.223321	0.087266	1



Correlation Heat Map



Correlation Scatter Plot

2. Task 2 Simple Linear Regression and higher order polynomial regression

2.1. Estimate of the coefficients a_0, a_1 and σ^2:

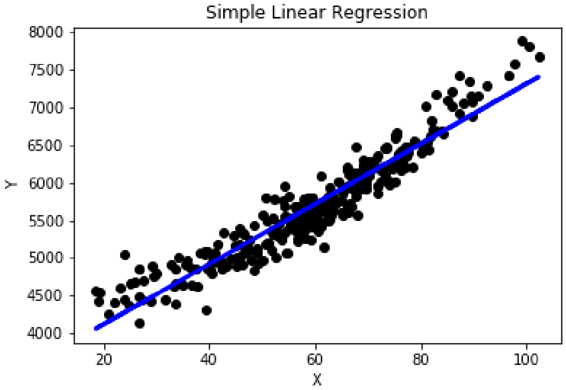
The estimates are a_0 = 3319.3573, a_1 = 39.9718 and σ^2 = 51501.5748

2.2. p values, R^2 and F values:

p values of zero are observed for the coefficients, so null hypothesis that the coefficients are zero is rejected by 90% confidence, and shows there is a statistical dependence of X1 and Y. R^2 value of 0.899 indicates a good fit as it is close to 1. F value of 2581 and p value for F test overall model is also zero indicating null hypothesis that all coefficients are zero can be rejected and there is overall significance of the model.

2.3. Plot regression line:

We can observe a good linear fit through the data.

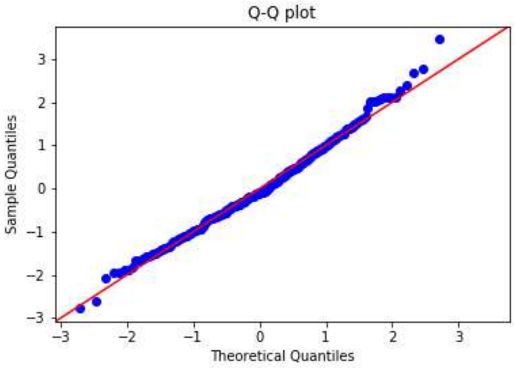


2.4. Residual Analysis:

a. Q-Q plot and Chi-squared test

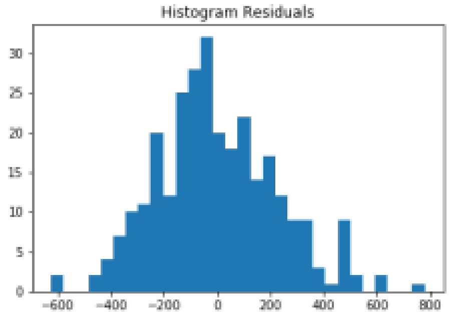
1. Q-Q plot:

The pdf of residuals against the pdf of $N(0,s^2)$ is shown in the Q-Q plot. The two distributions are almost equal though the right tail of residuals is a bit longer than $N(0,s^2)$ as it strays off a bit from the line and the quantile is a slightly larger value.



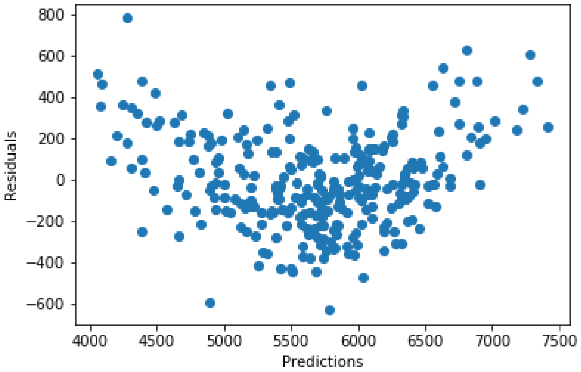
2. Chi-Squared Test:

   The chi-squared test gives a p value of 0.051861 so we say the pdf of the residual is normal distribution at 95% and 99% confidence level and is not a normal distribution at 90% confidence level. The conclusion is a bit ambiguous and questionable.



b. Residual Scatter Plot:

   The scatter plot shows the residuals are not random and there is some shape visible indicating some correlation trend. The variance is not same for the residuals.



2.5. Higher Order Polynomial Regression:

a. Estimate of the coefficients a_0, $a_{1,}$ a_2 and σ^2:

   The estimates are $a_0 = 4273.5435$, $a_1 = 3.7846$, $a_2 = 0.3135$ and $\sigma^2 = 37418.9412$

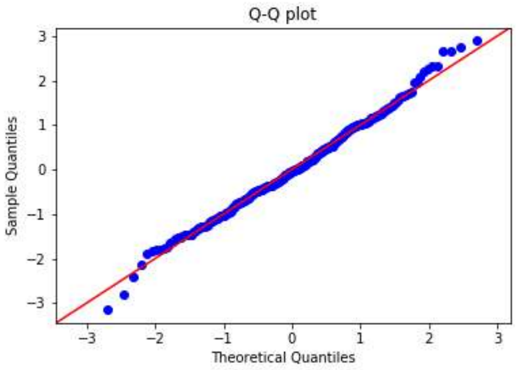
b. p values, R^2 and F values:

   p values of zero are observed for the a_0 and a_2 coefficients, so null hypothesis is rejected by 90% confidence, and shows there is a statistical dependence of X1^2 and Y. But a p value of 0.282 is observed for $a_{1.}$ This indicates we could fail to reject the null hypothesis that a_1 is zero. But the F value of 1831 and p value for F test overall model is zero indicating null hypothesis that all coefficients are zero can be rejected and there is overall significance of the model. R^2 value of 0.927 indicates a good fit and better compared to previous regression.

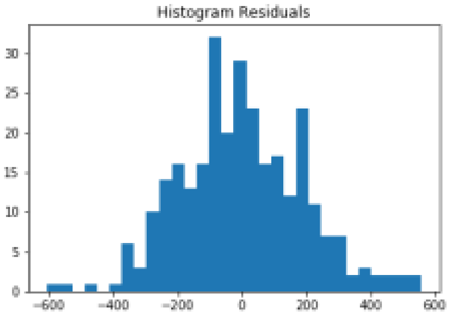
c. Residual Analysis:

   a. Q-Q plot and Chi-squared test

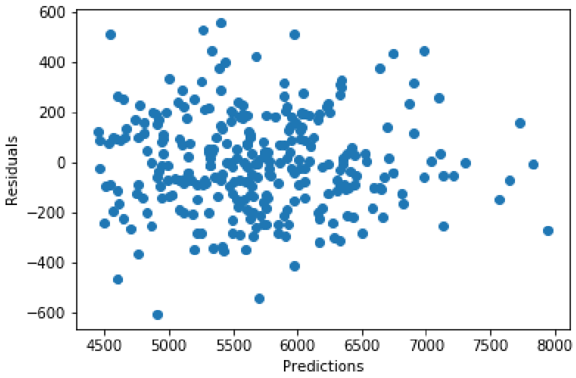
1. Q-Q plot:
   The pdf of residuals against the pdf of N(0,s^2) is shown in the Q-Q plot. The two distributions are almost equal though the left tail of residuals is a bit shorter than N(0,s^2) as it strays off a bit from the line and the quantile is a slightly smaller value.



2. Chi-Squared Test:
   The chi-squared test gives a p value of 0.3795 so we say the pdf of the residual is normal distribution at all confidence levels 90%, 95% and 99%.



d. Residual Scatter Plot:
   The scatter plot shows the residuals are random compared to the previous regression as there isn't a shape much evident. The variance does vary and a few outliers can be seen. It seems a better fit than previous regression.



3. Task 3 Multivariable Linear Regression

3.1. Estimate of the coefficients and σ^2:

The estimates are a_0 = -739.8153, a_1 = 40.2738, a_2 = 2.1324, a_3 = 3.8446, a_4 = 6.6792, a_5 = 7.6036 and σ^2 = 17762.3872

3.2. p values, R^2 and F values:

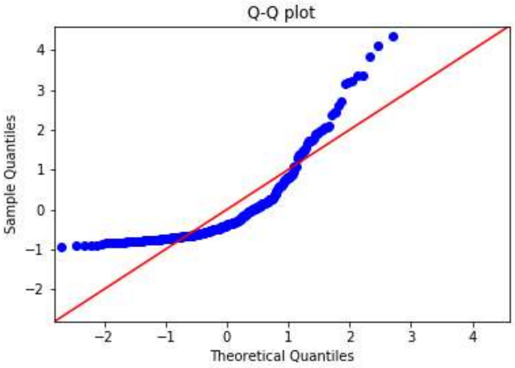
p values of zero are observed for the coefficients, so null hypothesis is rejected by 90% confidence, and shows there is a statistical dependence of the independent variables and Y. R^2 value of 0.966 indicates a good fit as it is close to 1. F value of 1608 and p value for F test overall model is also zero indicating null hypothesis that all coefficients are zero can be rejected and there is overall significance of the model. Considering all these values no independent variables need to be removed.

3.3. Residual Analysis:

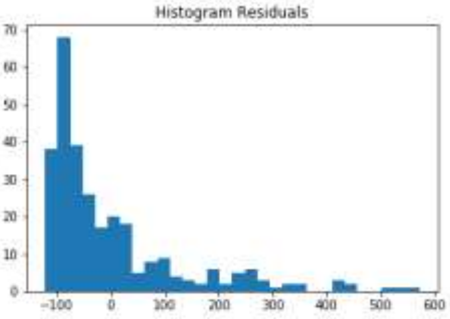
a. Q-Q plot and Chi-squared test

1. Q-Q plot:

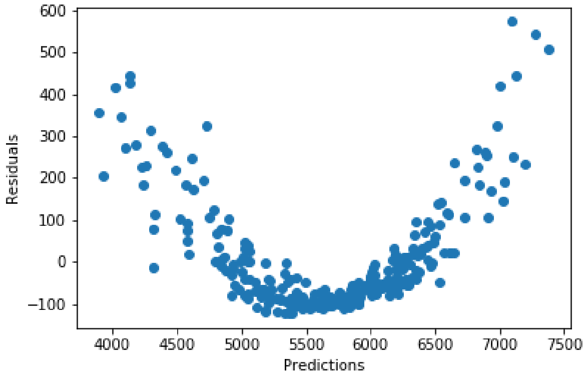
The pdf of residuals against the pdf of $N(0,s^2)$ is shown in the Q-Q plot. The two distributions are not equal. The pdf is not along the line of $N(0,s^2)$ pdf and completely strays off at both ends. We observe it is not a normal distribution.



2. The chi-squared test gives a p value of close to zero, so we say the pdf of the residual is not a normal distribution at all confidence levels 90%, 95% and 99%. It is also observed from the histogram.



b. Residual Scatter Plot:
The scatter plot shows the residuals are not random and there is a clear shape visible indicating a correlation trend. The variance is not same for the residuals. This pattern could indicate a need for a non-linear model.


Residuals
Predictions