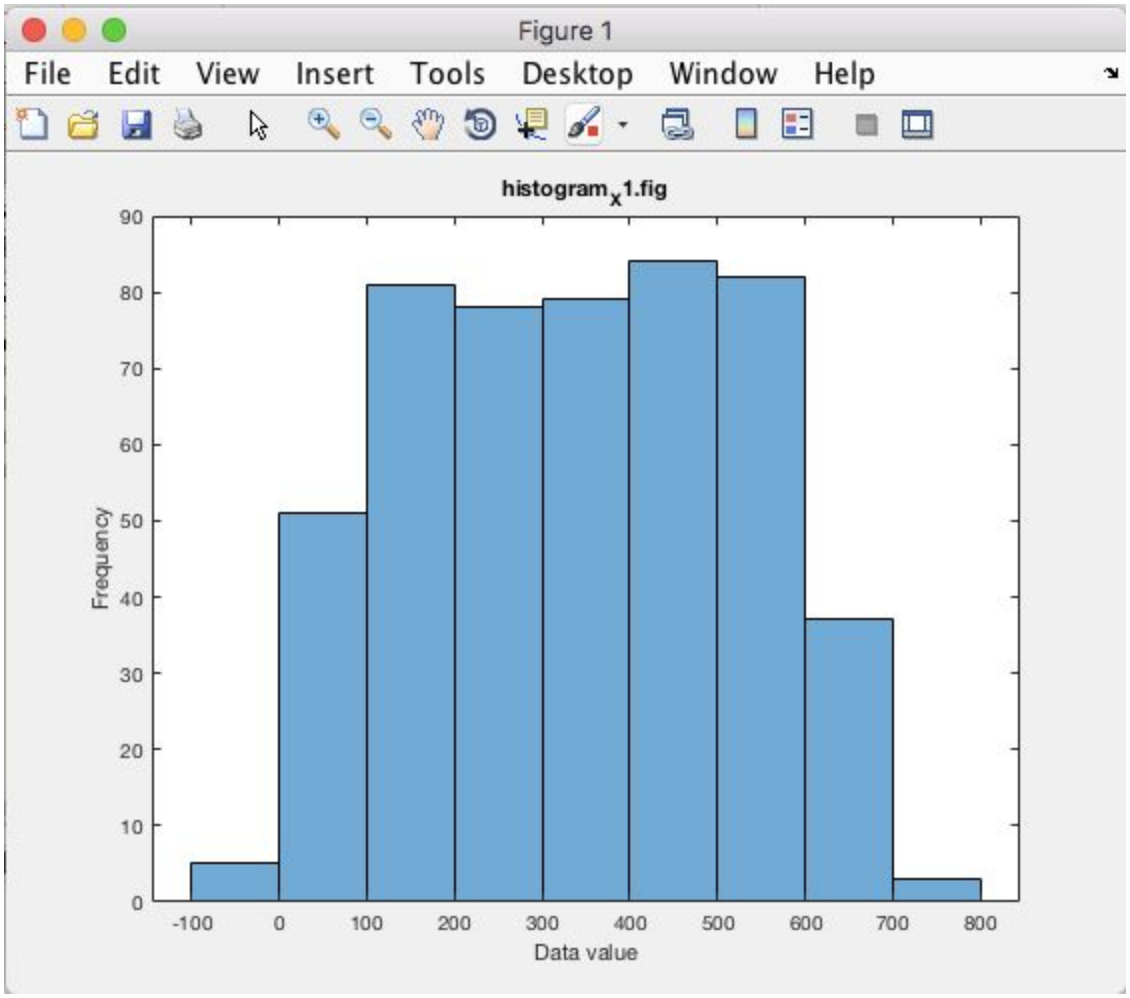
Task 1. Basic statistics analysis

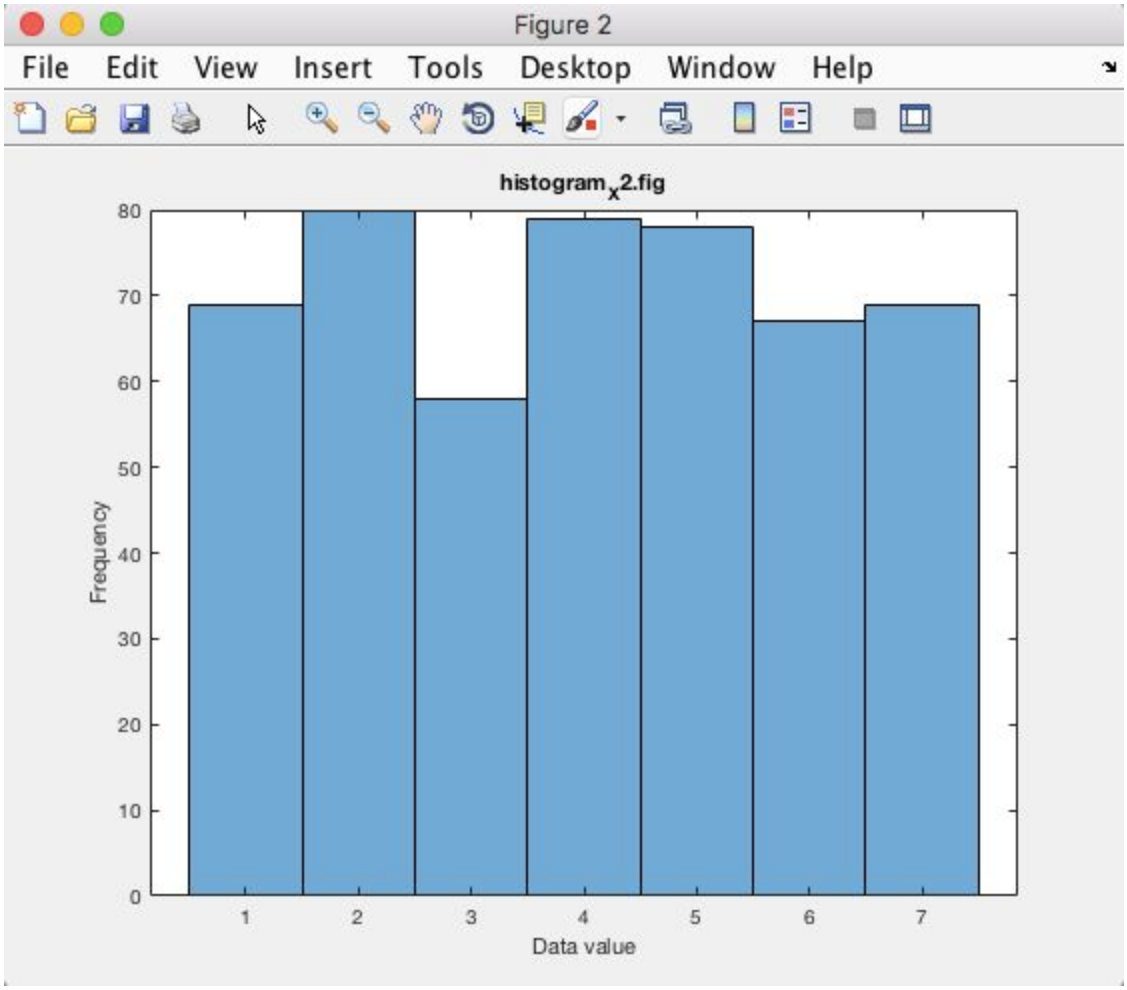
1.1. For each variable Xi, i.e. column in the data set corresponding to Xi, calculate the following: Histogram, mean, variance.

1.2. Use box plot or any other function to remove outliers (do not over do it !), or you can do that during the model building phase (tasks 2 and 3)
1.3 Calculate the correlation matrix Σ among all variables, i.e., Y, X1, X2, X3, X4 and X5. Draw conclusions related to possible dependencies among these variables.
1.4 Comment on the results

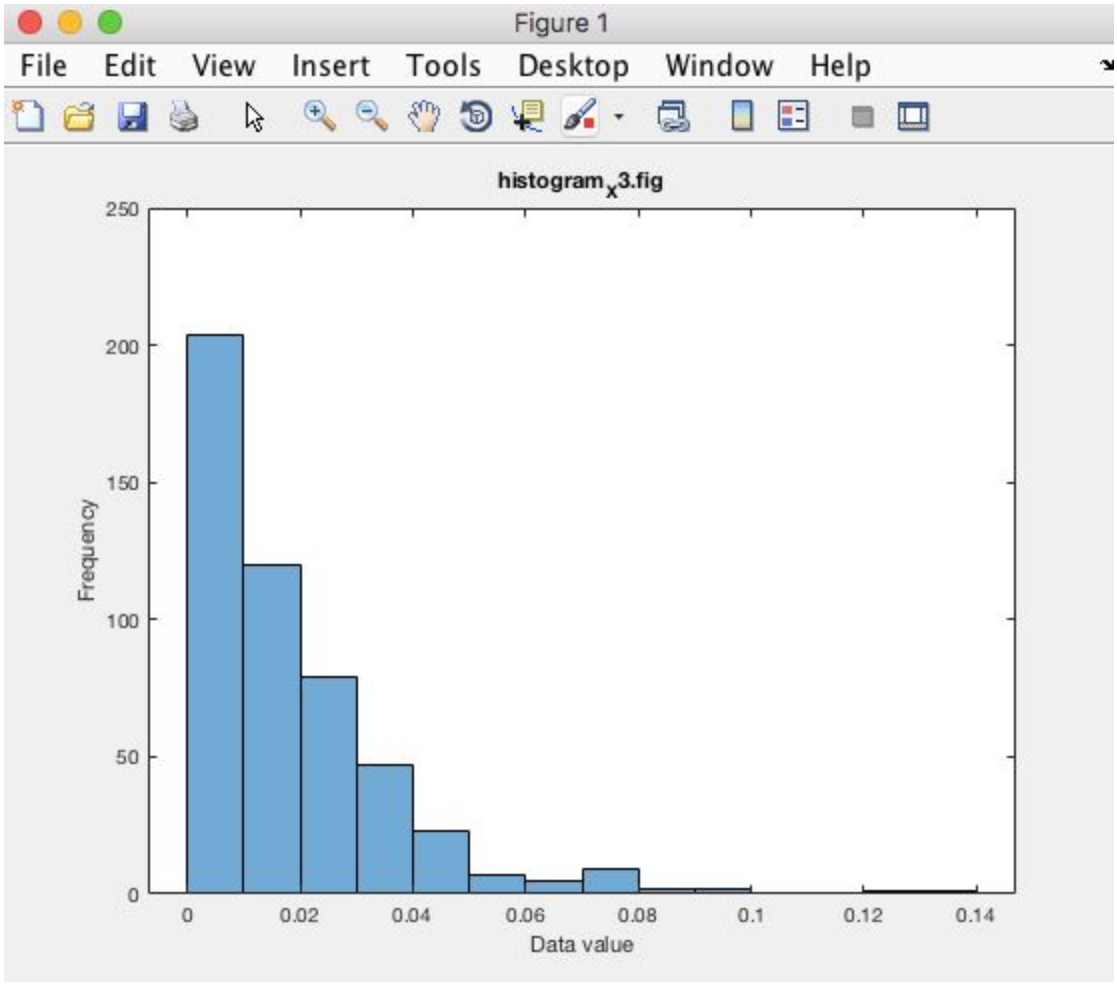
Solutions:
The tasks were completed using matlab and the code is uploaded with this document.
Task 1.1:
X1:
Mean: 341.246168
Variance: 33981.297803
Histogram:



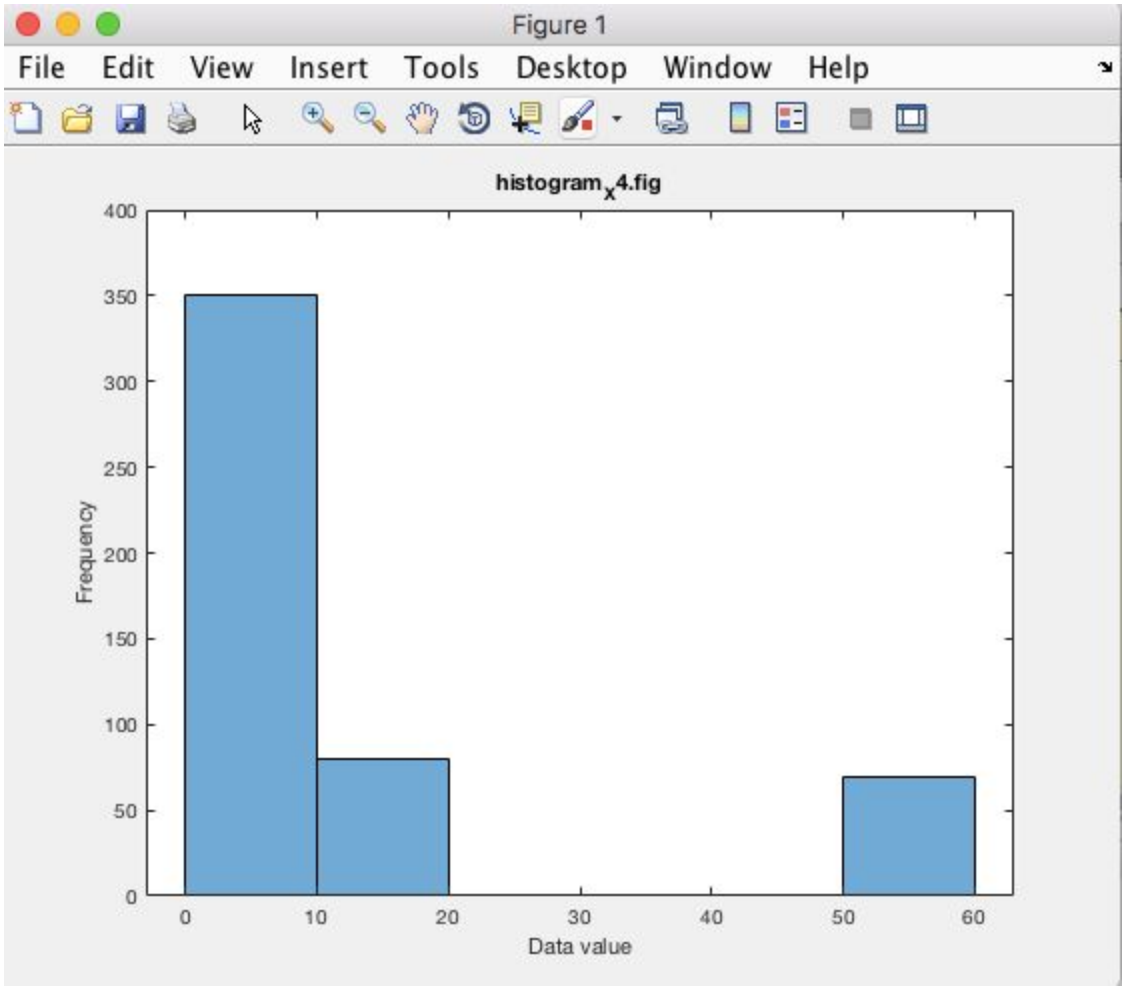
X2:
Mean: 3.988
Variance: 3.931
Histogram:



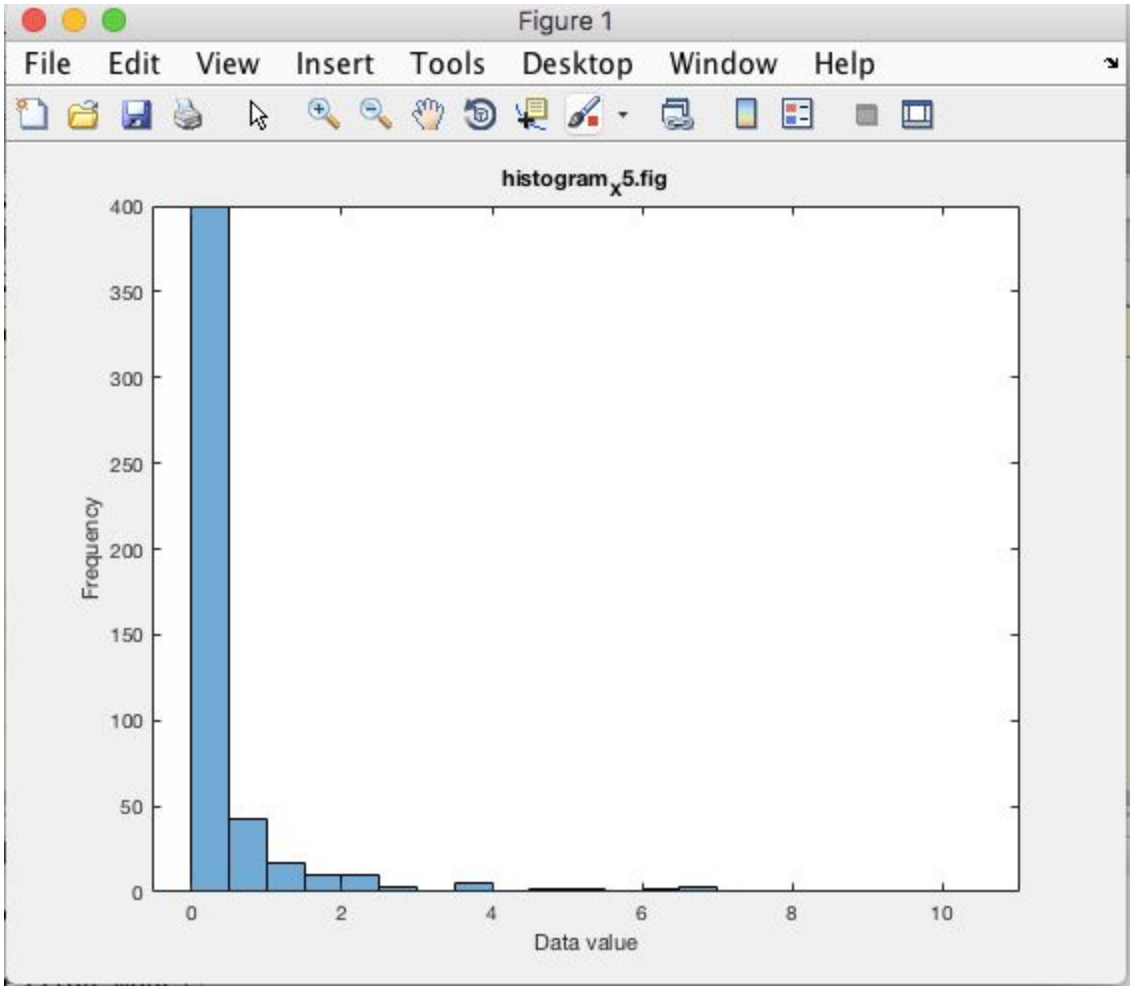
X3:
Mean: 0.018529
Variance: 0.000335
Histogram:



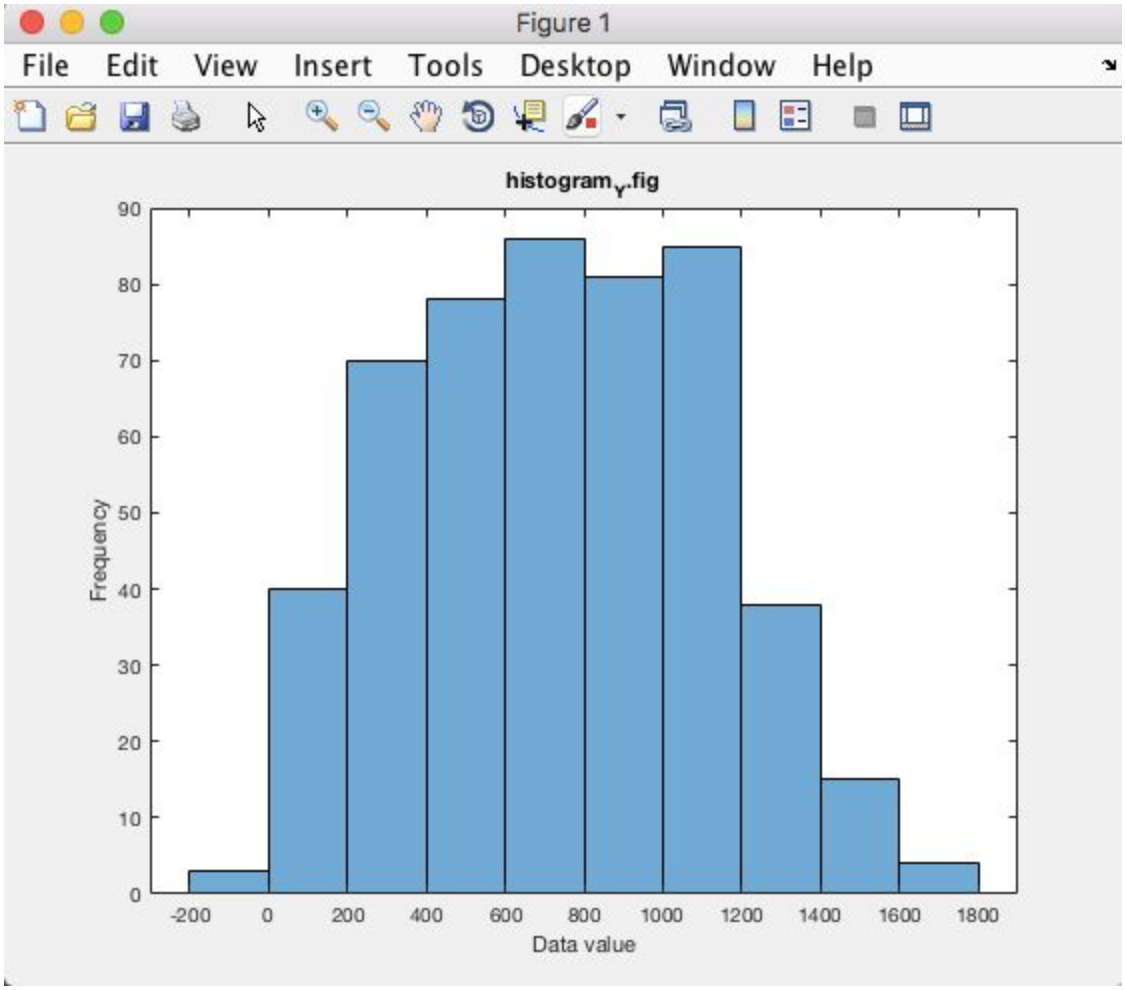
X4:
Mean: 11.531292
Variance: 305.769966
Histogram:



X5:
Mean: 0.484657
Variance: 1.382036
Histogram:



Y:
Mean: 743.637445
Variance: 143981.055313
Histogram:



Task 1.2:

Box plot:

X1:

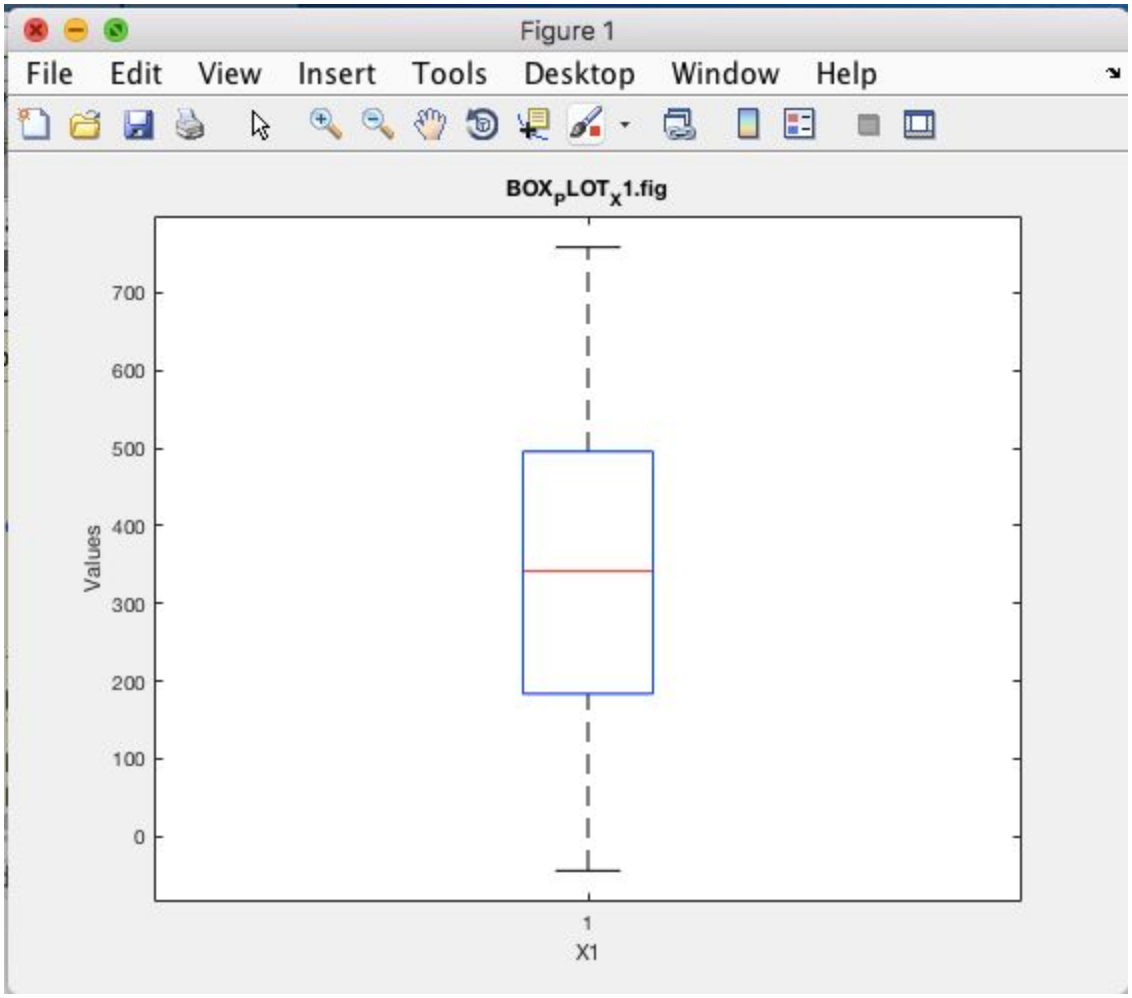


Figure 1
File Edit View Insert Tools Desktop Window Help
BOXPLOTX1.fig
Values
700
600
500
400
300
200
100
0
1
X1

X2:

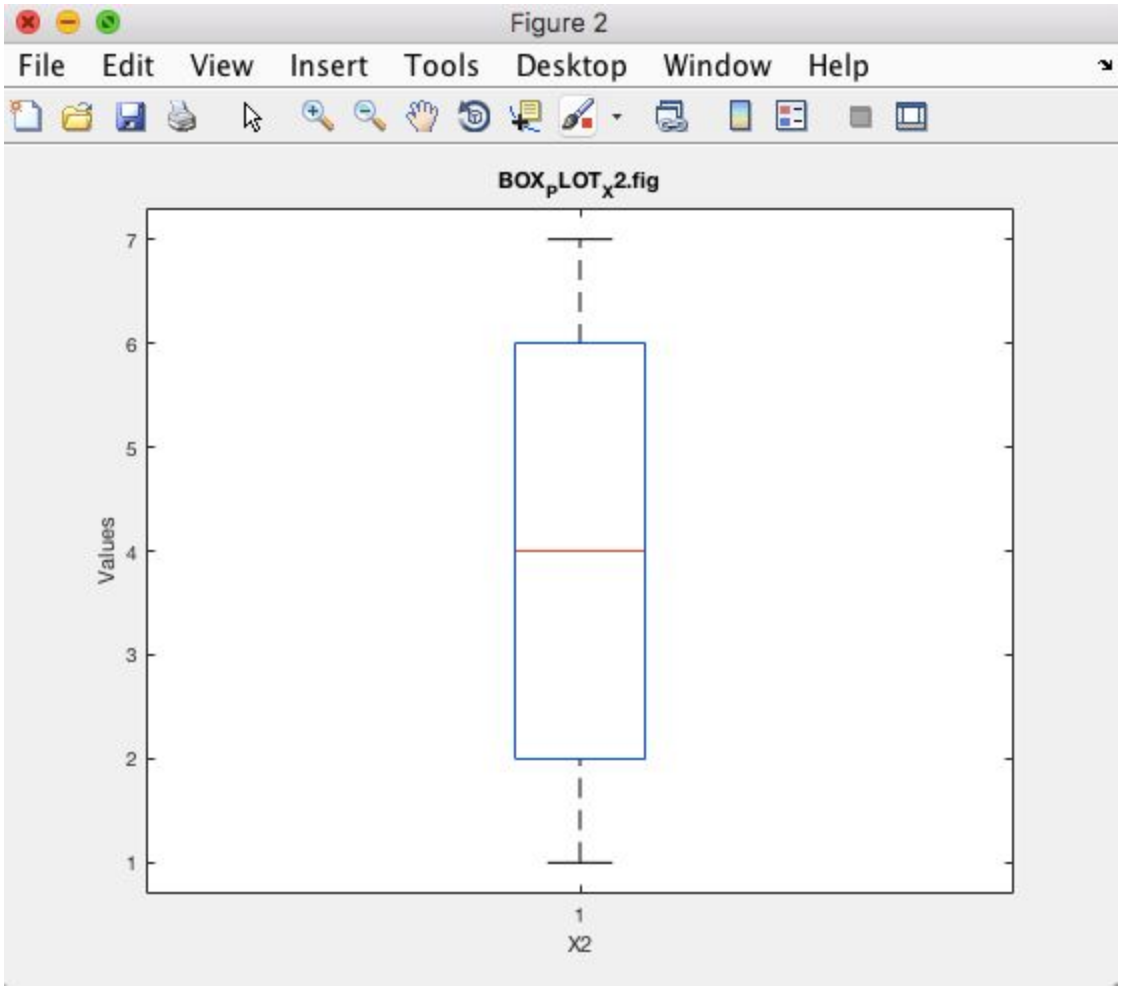


Figure 2
File Edit View Insert Tools Desktop Window Help
BOX_PLOT_X2.fig
Values
7
6
5
4
3
2
1
1
X2

X3:

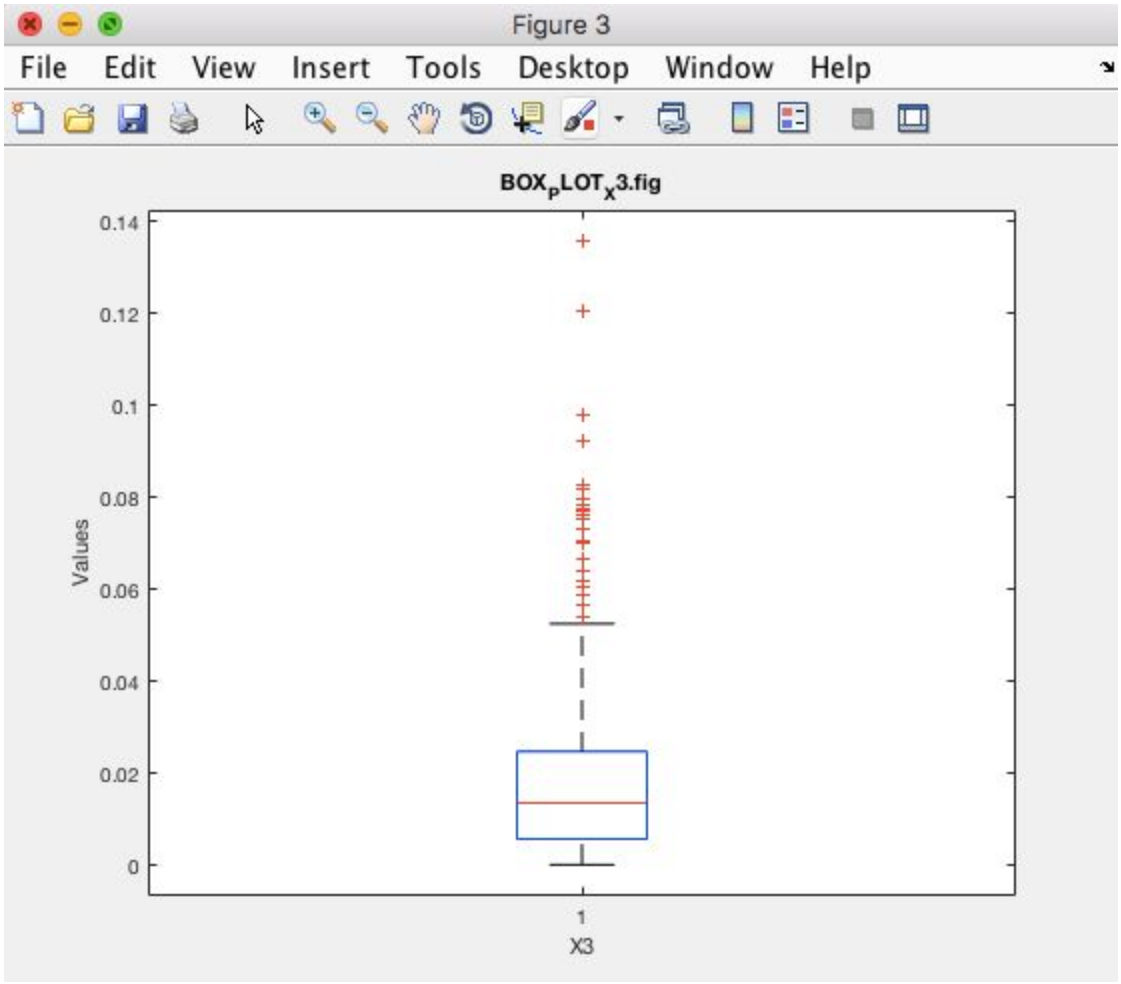


Figure 3
File Edit View Insert Tools Desktop Window Help
BOX_PLOT_X3.fig
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Values
1
X3

X4:

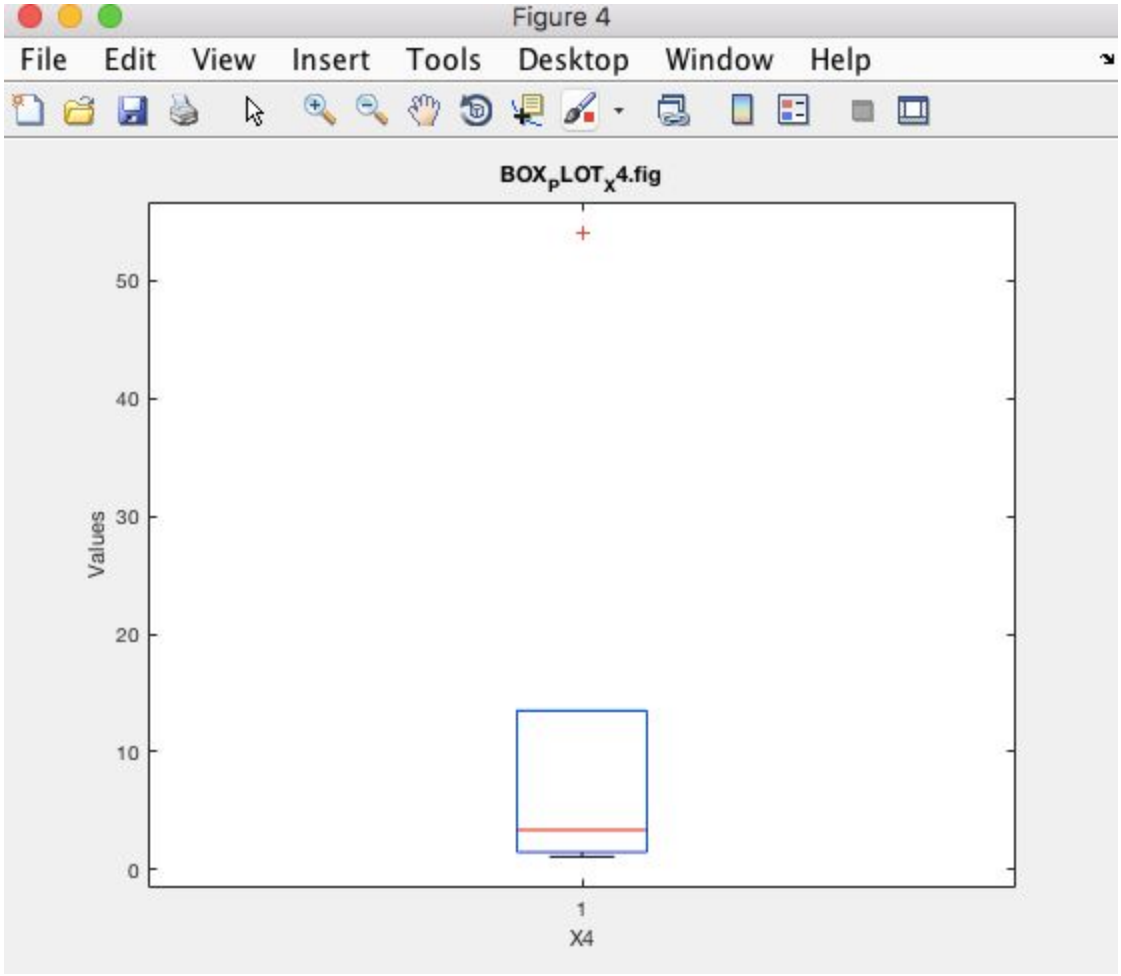


Figure 4
File Edit View Insert Tools Desktop Window Help
BOXpLOTX4.fig
50
40
30
20
10
0
Values
1
X4

X5:

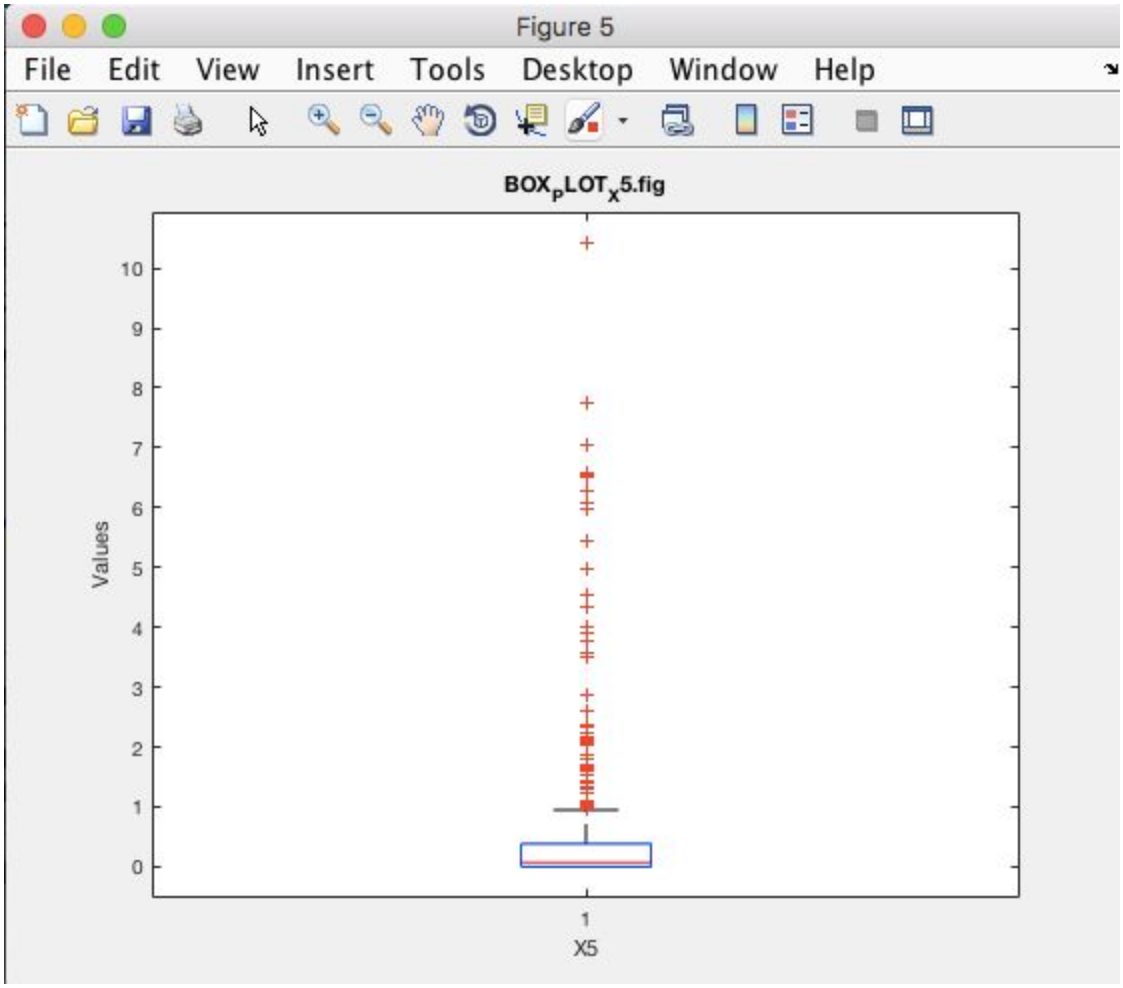
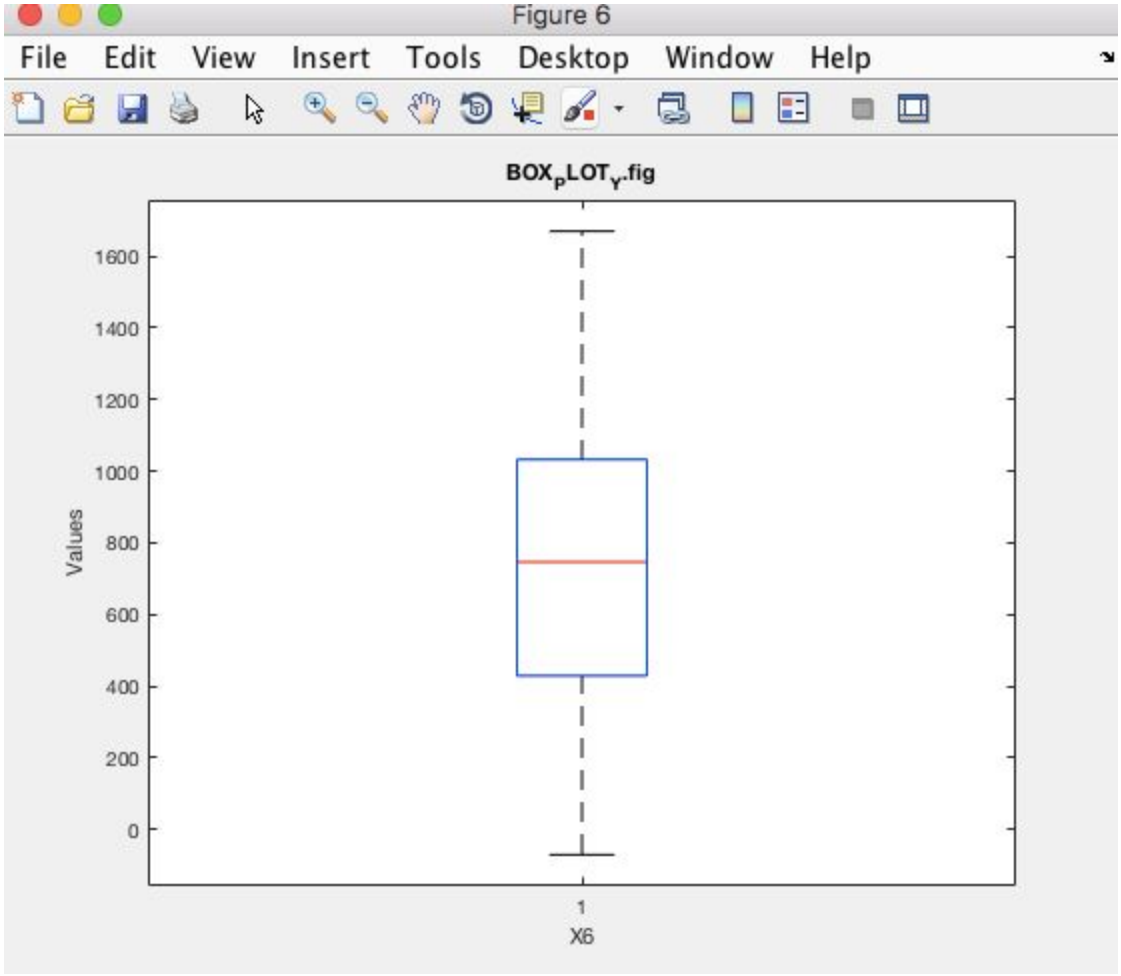


Figure 5
File Edit View Insert Tools Desktop Window Help
BOXPLOTX5.fig
Values
10
9
8
7
6
5
4
3
2
1
0
1
X5

Y:



Task 1.3 Correlation Matrix:

The correlation matrix is as shown below :

	X1	X2	X3	X4	X5	Y
X1	1.0000	-0.0347	0.0109	0.0052	0.3871	0.9162
X2	-0.0347	1.0000	-0.0336	-0.7510	-0.0332	-0.3284
X3	0.0109	-0.0336	1.0000	0.0435	0.0051	0.0267
X4	0.0052	-0.7510	0.0435	1.0000	0.0195	0.3989
X5	0.3871	-0.0332	0.0051	0.0195	1.0000	0.3848
Y	0.9162	-0.3284	0.0267	0.3989	0.3848	1.0000

As we can see from the correlation matrix:
There is a strong positive correlation between X1 and Y. X1 has very less negative linear correlation with X2.There is good amount of correlation between X1 and X5.
If the correlation coefficient is in between -0.1 to 0.1 then the variables are said to have no linear relationship or a very weak linear relationship.

Hence we can say there is no linear correlation between X1 and X3 and X1 and X4.

X2 has negative correlation between all the other variables (X1,X3,X4,X5,Y)
X3 has no linear correlation with any of the variables.
X4 has no correlation with X1,X3 and X5, negative correlation with X2 and less correlation with Y
X5 good amount of correlation with X1, no linear correlation with X2, X3, X4 and a decent correlation with Y.
Y has good correlation with X1, negative correlation with X2, no correlation with X1, and a decent correlation with X4 and X5.

Since X3 is not correlated with any other vairables and there is no correlation with Y, X3 is not a good candidate for the modeling.

Task 1.4
We observe that X3 is not correlated with any other variable and also not with Y. Hence, we should exclude it when predicting the value of Y. Since it has no correlation with other variables (when doing Multiple linear regression). When included, it may result in the case of overfitting.

Also, X3 shows minimal correlation with other independent variables X1, X2, X4 and X5 which makes it a good candidate for zero contribution to multi-collinearity when performing multivariate linear regression. We will analyze it more in the third task discussed below.

We also, see that the independent variables, X1, X2, X4 and X5 are highly correlated with Y. Hence, they are good candidates for predictor variables. But there is a high correlation between X1 and X5 and also X2 and X4 which can result in multicollinearity problem
We will analyze the effect in task 3 below
Comments on Y distribution: Y has a high mean. It is much likely that it has only positive correlation with predictor variables. Also, on plotting Y's histogram, it looks close to a normal distribution

Task 2: Linear regression

Before proceeding with the multiple regressions, you will carry out a simple linear regression to estimate the parameters of the model: $Y = a0 + a1X + \varepsilon$, where $X = X1$.

2.1 Determine the values for a0, a1, and s^2.

2.2 Check the p-values, R^2, F value to determine if the regression coefficients are meaningful.

2.3 Plot the regression line against the data.

2.4 Do residuals analysis:

a. Do a Q-Q plot of the pdf of the residuals against N (0, s2) Alternatively, draw the residuals histogram and carry out a χ 2 test that it follows the N (0, s^2).

b. Do a scatter plot of the residuals to see if there are any correlation trends.

2.7 Use a higher-order polynomial regression, i.e., $Y = a0 + a1X + a2X2 + \varepsilon$, to see if it gives better results.

2.8 Comment on your results in a couple of paragraphs.

Solutions:

The code is written in MATLAB and uploaded with this document

2.1)

The values are:

a0: 100.05

a1: 1.886

s^2 : 120871.1

2.2)

The output of the linear model:

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    ___________

    (Intercept)    100.05        14.333    6.9803     9.4478e-12
    x1              1.886       0.036954    51.036    5.7396e-200


Number of observations: 500, Error degrees of freedom: 498
Root Mean Squared Error: 152
R-squared: 0.839,  Adjusted R-Squared 0.839
F-statistic vs. constant model: 2.6e+03, p-value = 5.74e-200
The value of a0, a1, and s2 is: 1.000478e+02 1.885998e+00 1.208711e+05
The value of Rsquared is:8.391708e-01
The value of p is:5.739566e-200
The value of F is:2.604670e+03
```

The values are:
p : 5.739566 e^{-200} ~ 0
F: 2.604670
R^2 : 0.839

Comments:
R^2 value is around 0.839 which is a good value in order to ensure goodness of the model to fit the data. High R shows that the points are less scattered around the regression line.

P value corresponding to the F-test is 0.000 which is significant enough to state that our model provides a better fit than intercept only model. So, just by checking R^2 and P value we can say X1 might a meaningful addition. Higher F value and low p would mean a meaningful predictor.

If we check the p value of X1 is nearly 0, this implies that, Y variability of Y with changing X1. So, X1 might be meaningful addition to the model. But we need to do other tests(QQ plot) to give the final verdict.

Task 2.3

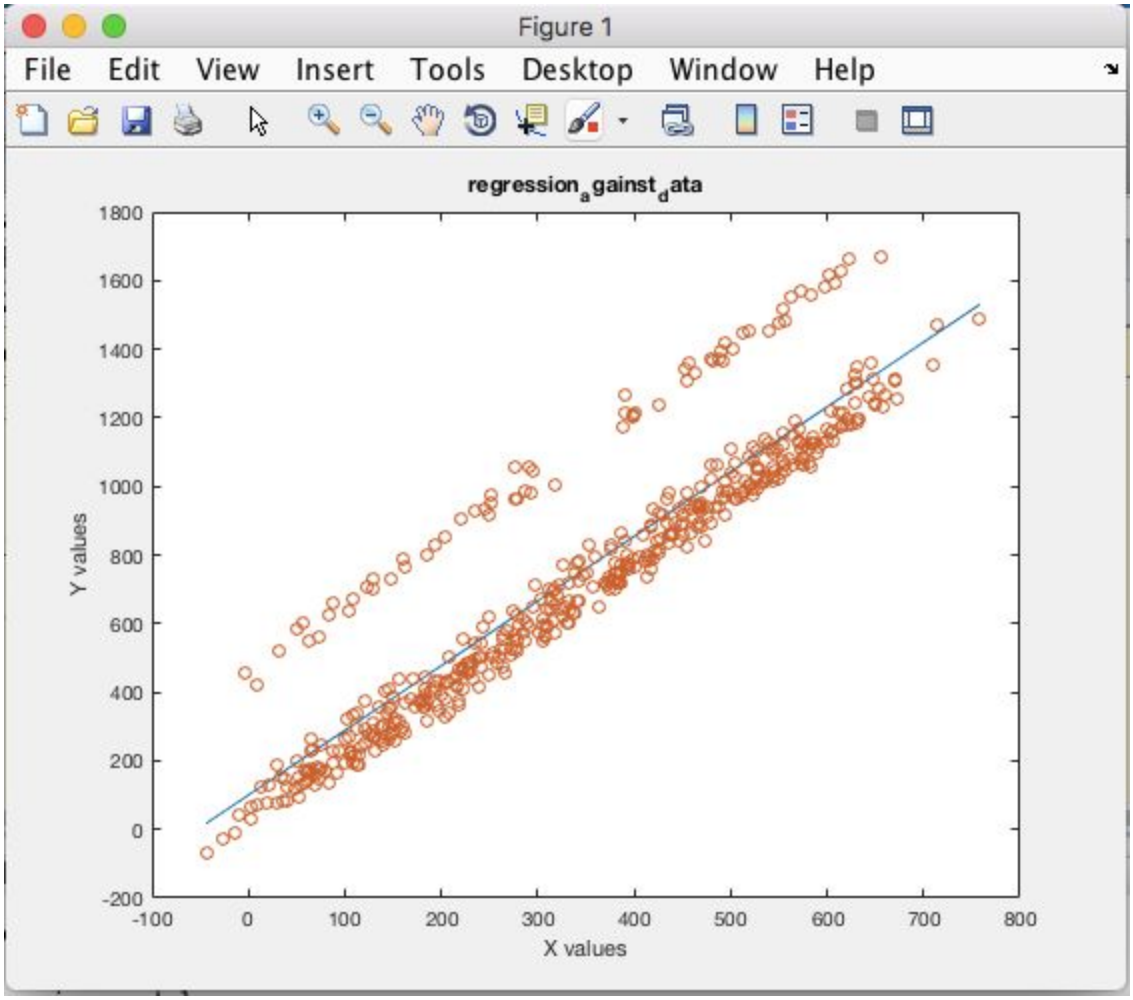
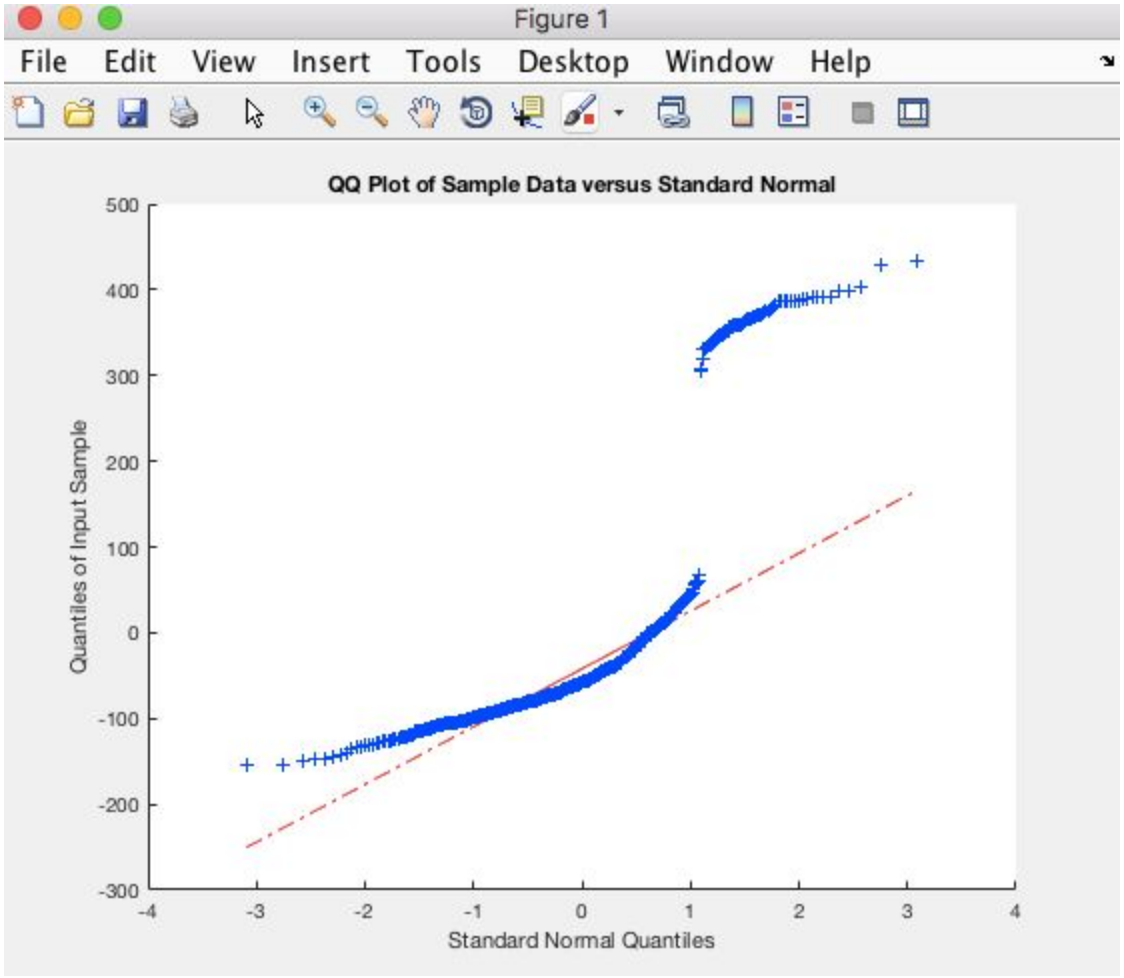


Figure 1
File Edit View Insert Tools Desktop Window Help
regression_against_data
Y values
X values
1800
1600
1400
1200
1000
800
600
400
200
0
-200
-100
0
100
200
300
400
500
600
700
800

Task 2.4

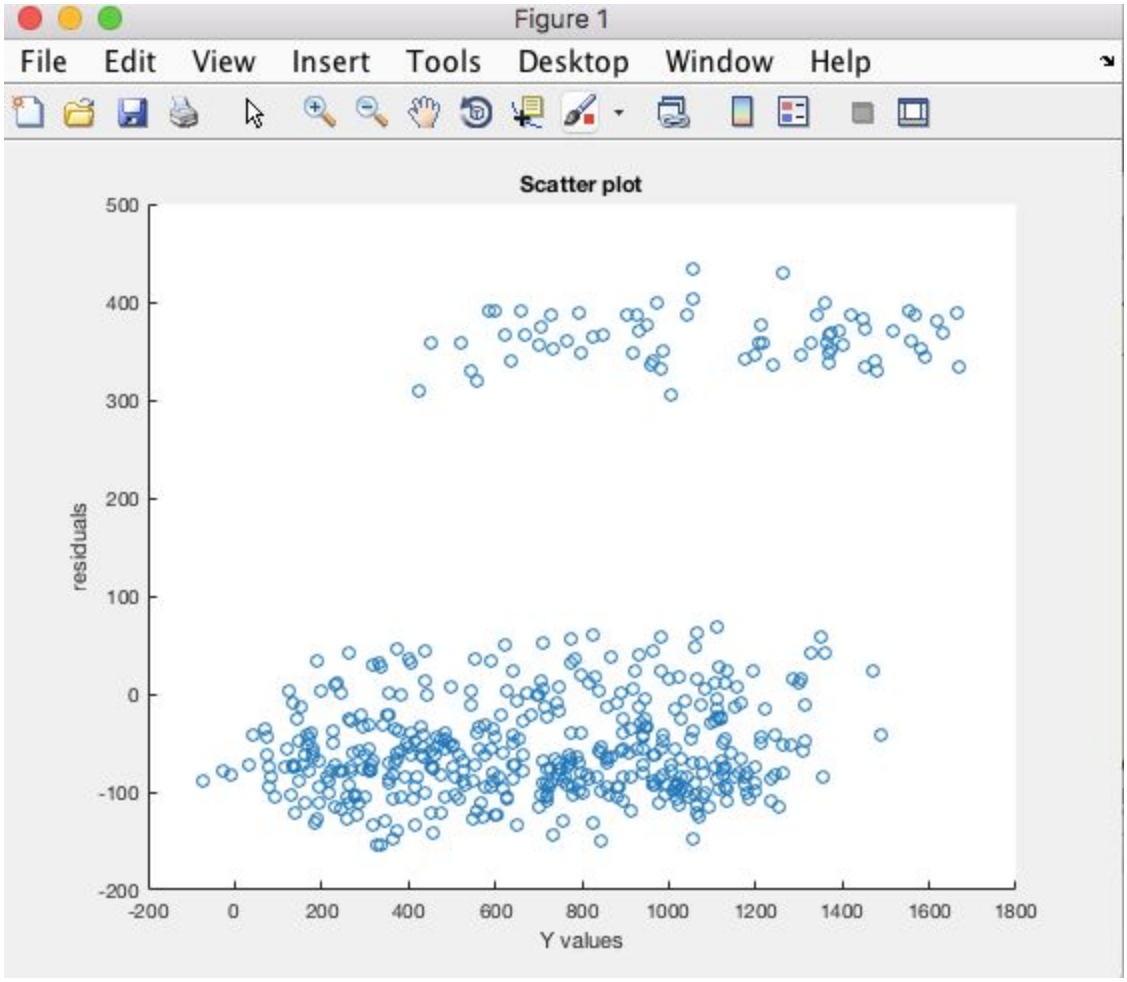
a) Q-Q plot:



As can be seen, the distribution of residuals is very far fetched from the normal distribution line but it has tails and high modality. There are multiple breaks in residual when compared with a normal distribution (shows a multi-modal behavior). So, residuals distribution here doesn't resemble a normal distribution perfectly.

b)

Scatter plot:



Comments:
From the scatter plot, it can be seen that the residuals are scattered all more near the higher range, but average out to 0 in between. X1 is not a good candidate for the model.

Task 2.7:

When I run a higher order polynomial test on the X1, I get the output message:
“1. Removing x1^2, FStat = 0.062839, pValue = 0.80217”
Since the p value is very high it removes on its own and hence the model received is a linear model and hence its the same as the linear model.

Task 2.8
Comments:
For Linear Fit :

X1 alone provides a bad predictor model. Eventhough R_squared value is quite good but the pvalue with is also high to be considered as the best model.

The residuals do not have a normal distribution and show multi-modal behavior. This can be seen using Q-Q plots and histograms.

For Polynomial Fit
$X1^2$ fails to improve the model and cannot provide any better fit or prediction model. In fact, the standard errors for the predictor variables is increased with polynomial fit clearly indicating that the model gets worse with increasing order of X1. Also, the p values start getting out of significance zone with polynomial fits.

Hence, we need to consider other predictor variables to decide the best regression model for the given data.

Task 3: Multivariate regression:

Task3.1

Multiple regression with all the variables(X1, X2, X3, X4, X5):
The output is:

```
the estimated coefficients are as displayed

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
                   Estimate       SE         tStat       pValue
                   ________    _________    ________    __________

    (Intercept)      7.3764       5.2714      1.3993       0.16235
    x1                 1.86    0.0067765      274.47             0
    x2              -0.21589      0.87996    -0.24534       0.80629
    x3               -8.1772       62.934    -0.12993       0.89667
    x4                8.5239      0.09976      85.444    4.4267e-298
    x5                8.8214        1.062      8.3064     9.5562e-16


Number of observations: 500, Error degrees of freedom: 494
Root Mean Squared Error: 25.7
R-squared: 0.995,  Adjusted R-Squared 0.995
F-statistic vs. constant model: 2.16e+04, p-value = 0
The value of variance is: 1.433267e+05
```

R-squared value is very high(0.995) which is indicative of good fit of the model. But it can also be a case of overfitting. As we can see, the p-values for X2 and X3 are very high which clearly indicates that the model is bad with all the variables included. The standard error in these independent variables (X2 and X3) is also very high. Overall, the model looks very promising with following values on residual analysis

Task 3.2:
But, as we can see from the above output the p value for intercept, X2 and X3 is very high. We have to leave out X2, X3 and intercept and generate the model.

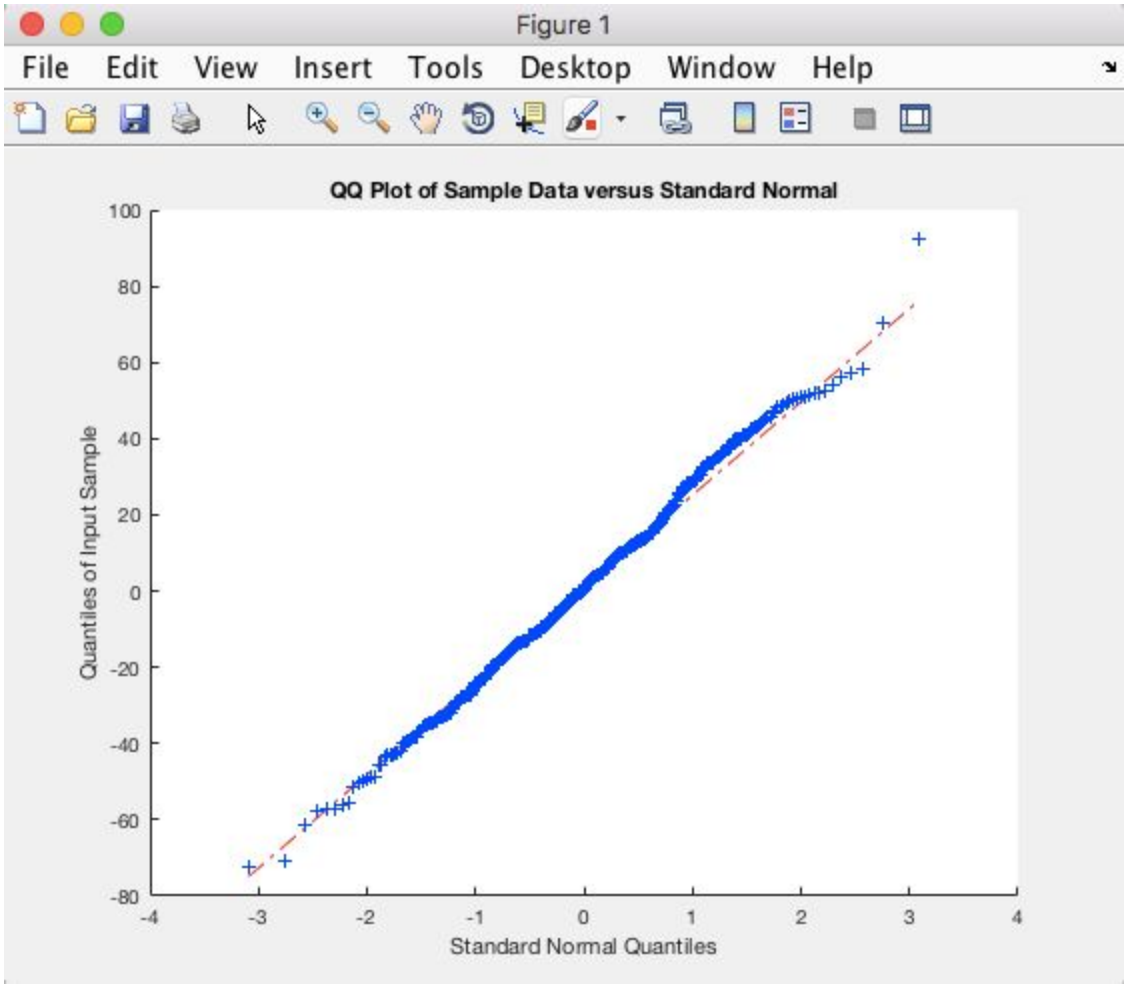
The output for the model is as shown:

```
Linear regression model:
    y ~ x1 + x2 + x3

Estimated Coefficients:
           Estimate       SE        tStat       pValue
           ________    _________    ______    __________

    x1     1.8734      0.0038474    486.92             0
    x2     8.5883       0.063088    136.13             0
    x3     8.4339          1.0523    8.0149    7.9275e-15


Number of observations: 500, Error degrees of freedom: 497
Root Mean Squared Error: 25.8
The value of Rsquared is:9.954094e-01
The value of p is:0
The value of F is:1.742505e+05
The value of variance is: 1.433267e+05
```

As we can see from the above picture we get the values:
p: 0, F: 174250.5 Variance: 143326.7
Coefficients:
x1(a1): 1.8734
x2(a4): 8.5883
x4(a5): 8.4339

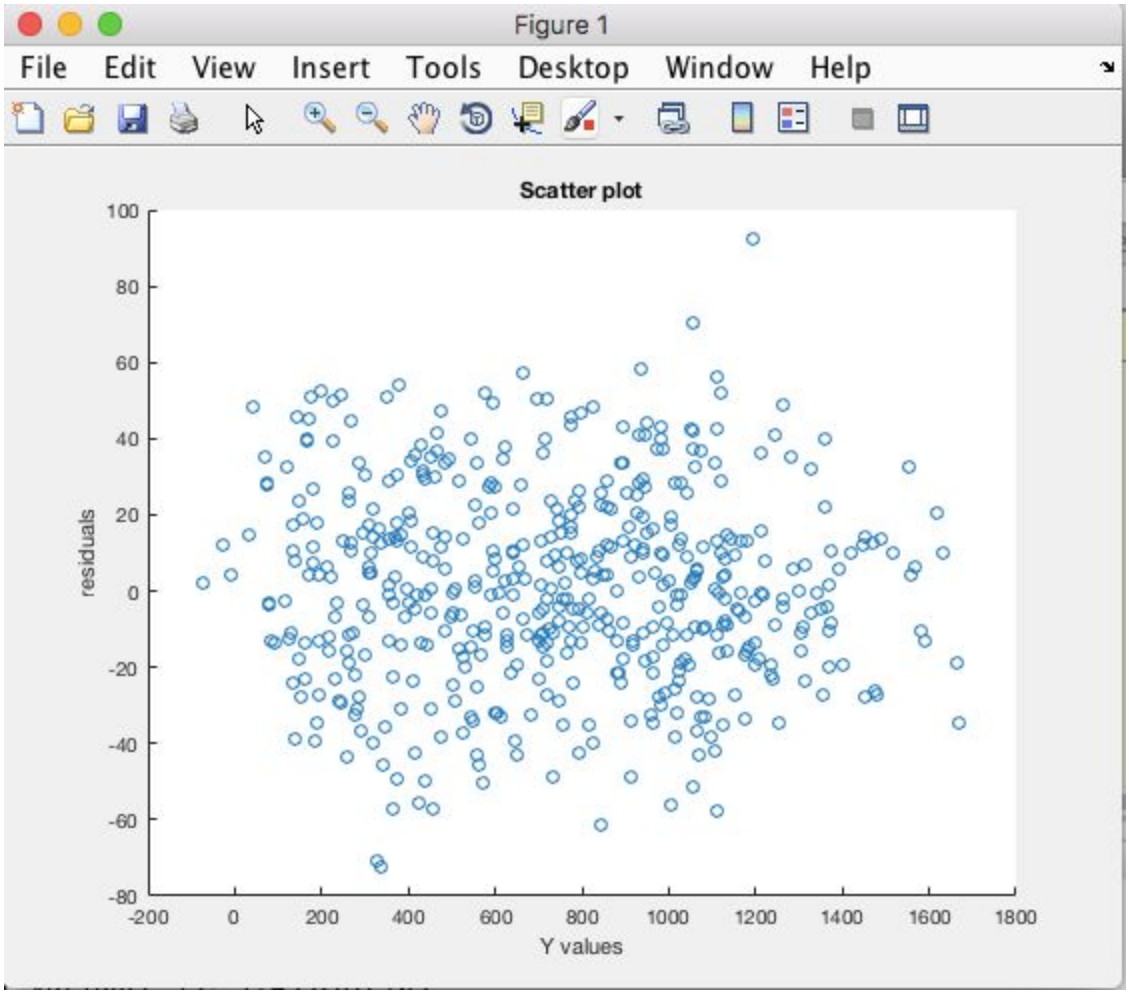
Task 3.3(a):

QQplot:



Looking at the QQ plot the residuals have a normal distribution.

Task 3.3(b):
Scatter plot:



Looking at the scatter plot we can say that the residuals are well scattered, and residuals average out to zero with no correlation trends.

By looking at coefficients(p, r squared, p value, F value, variance), QQ plot and scatter plot we can come to the conclusion that this is the best regression model with the given data with variables X1, X4 and X5. Also the chi square test gives the value of 0, and hence the best regression model.