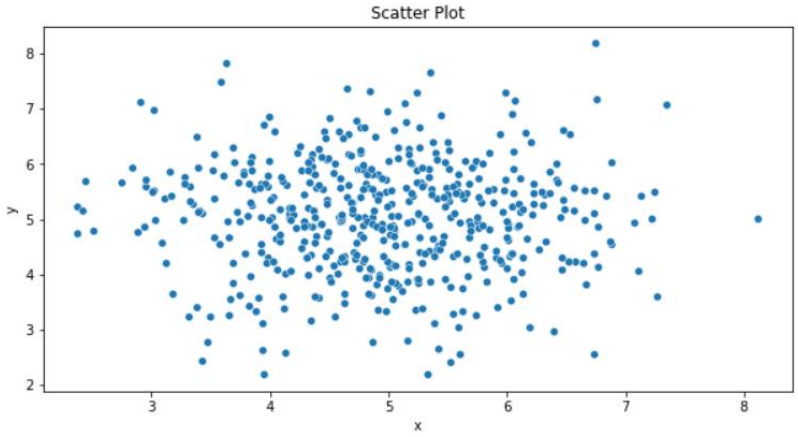
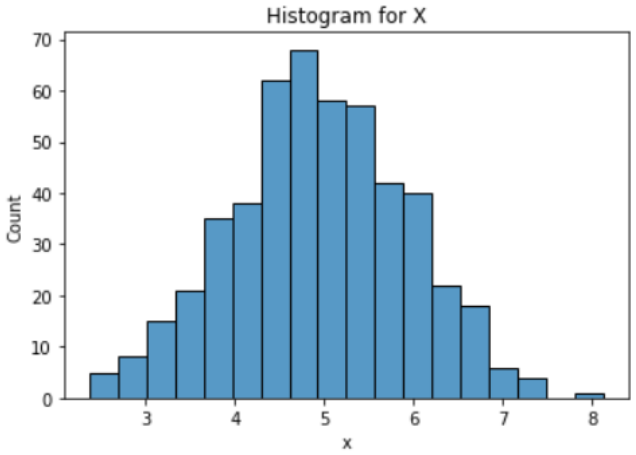
CLL788 – Process Control Dynamics, Report – Assignment 1

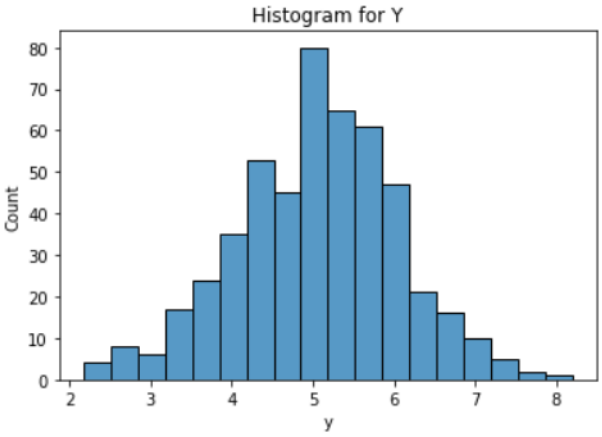
Ansh Lodhi, 2019CH70161

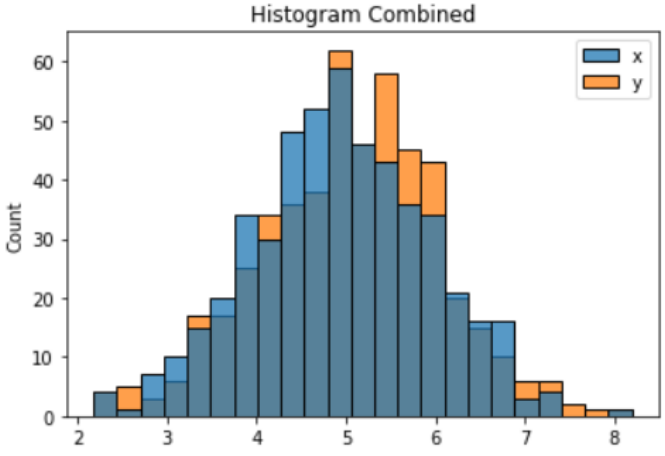
---

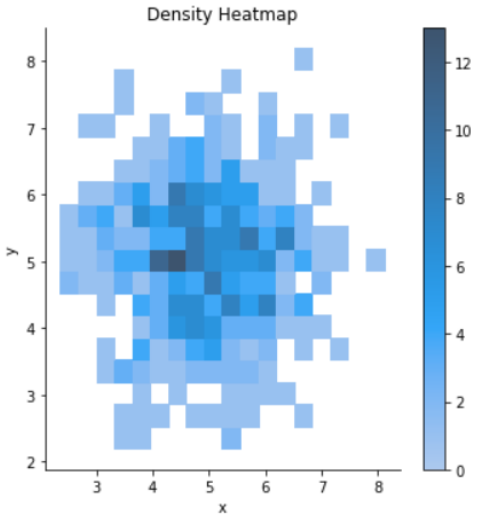
1. Do data visualization using the data given
    a. Show scatter, histogram, heatmap, box plots (data_1).

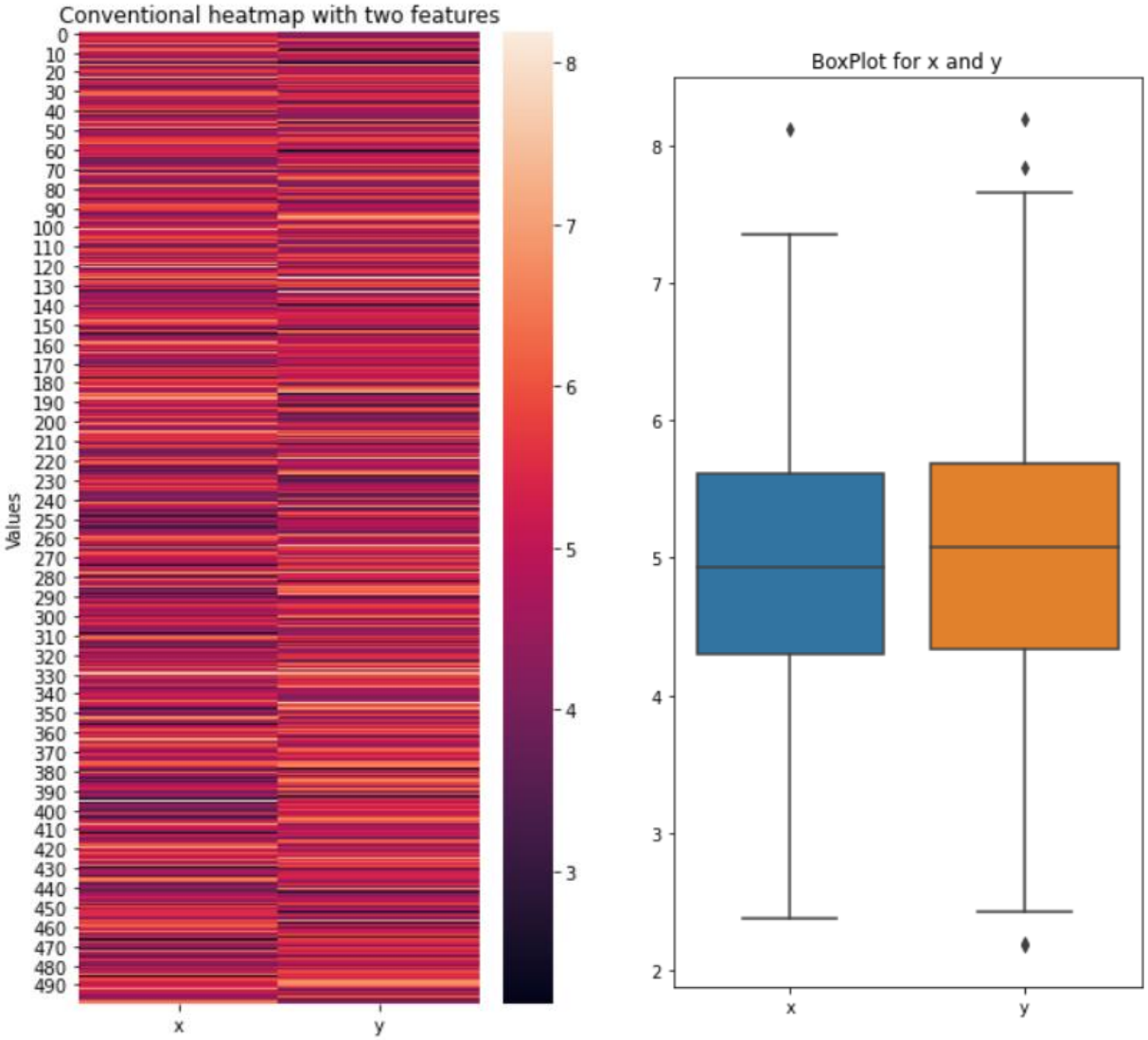




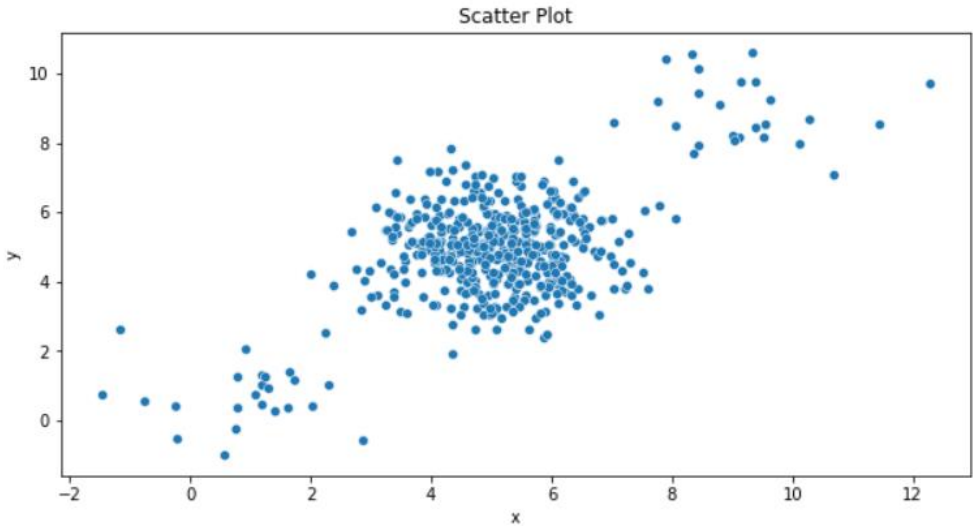


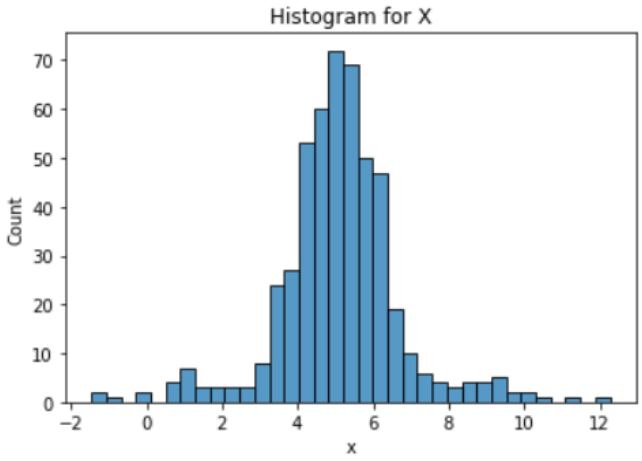


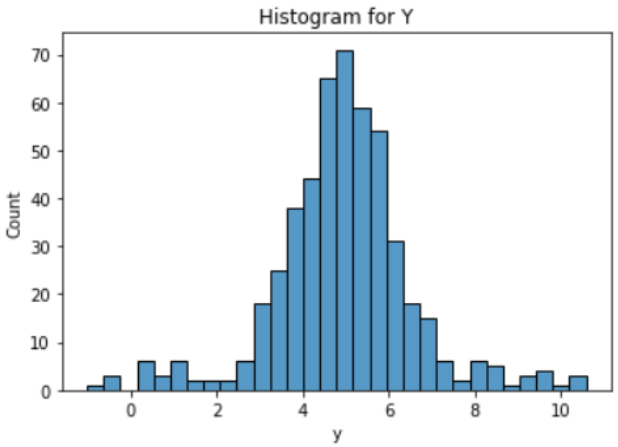


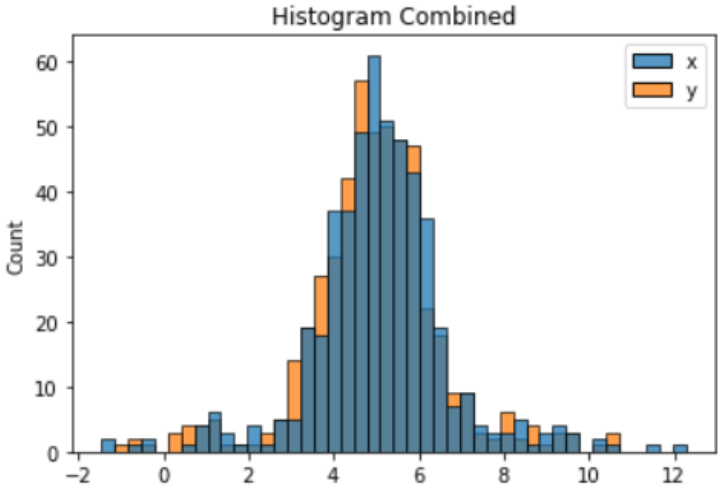


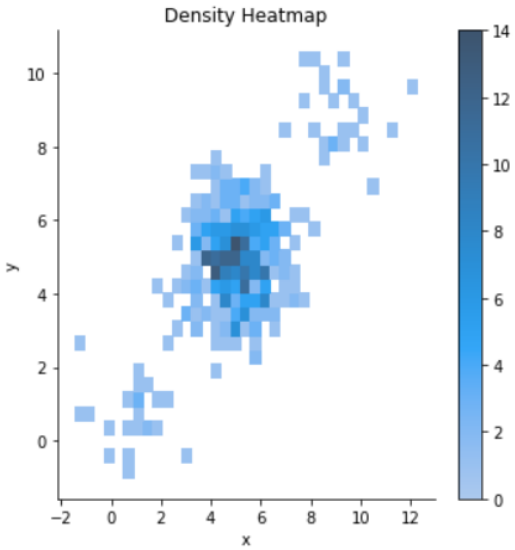
b. Now perform the same for (data_3).

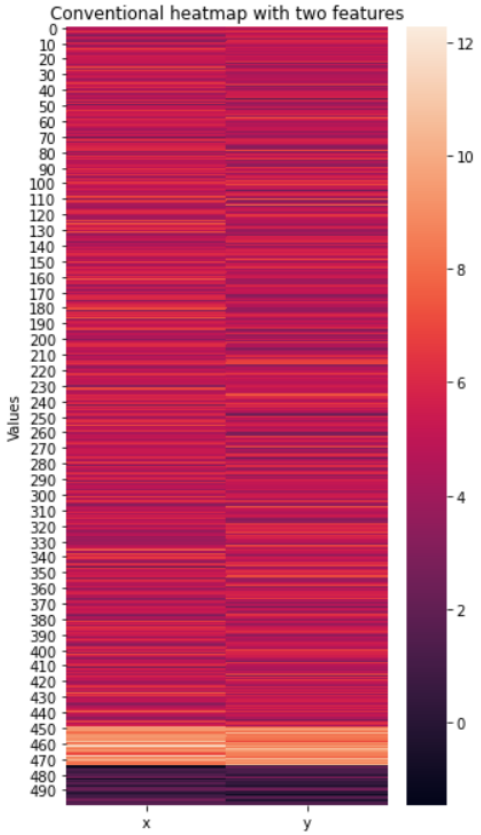


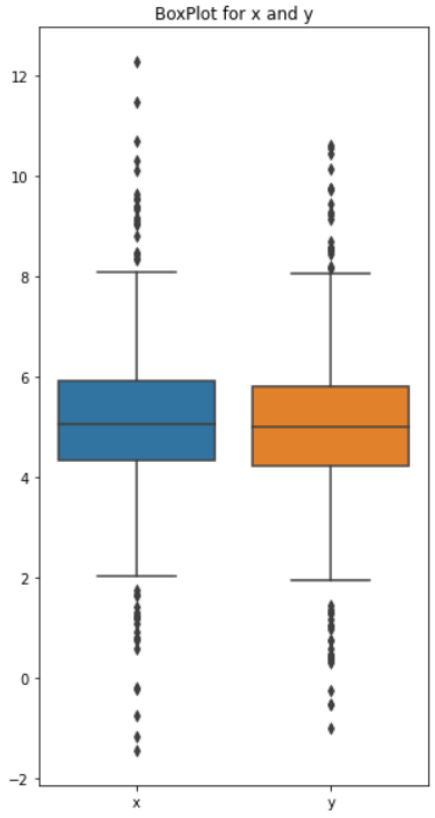

Histogram for X
Count
x


Histogram for Y
Count
y


Histogram Combined
Count
x
y


Density Heatmap
y
x


Conventional heatmap with two features
Values
x
y


BoxPlot for x and y
x
y

c. Calculate the statistics for both data sets.

data_1:

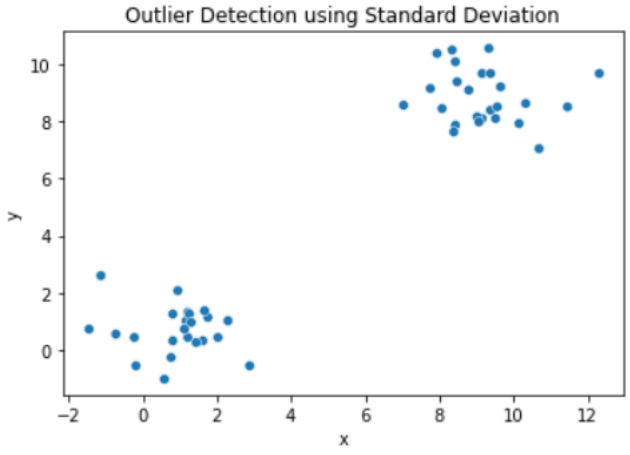
	x	y
count	500.000000	500.000000
mean	4.939743	5.042984
std	0.986803	1.008197
min	2.373638	2.181180
25%	4.303987	4.331464
50%	4.924278	5.074768
75%	5.607214	5.682380
max	8.117045	8.190109

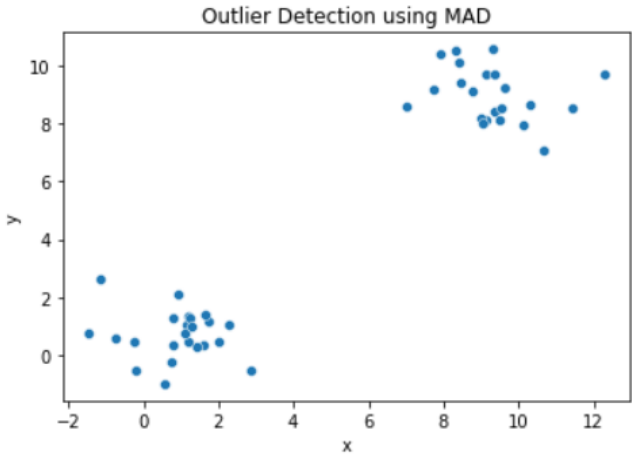
data_2

	x	y
count	500.000000	500.000000
mean	4.939743	5.042984
std	0.986803	1.008197
min	2.373638	2.181180
25%	4.303987	4.331464
50%	4.924278	5.074768
75%	5.607214	5.682380
max	8.117045	8.190109

d. Detect the outliers in data_3 using standard deviation approach and MAD approach. (You can use in-built packages if needed)

-



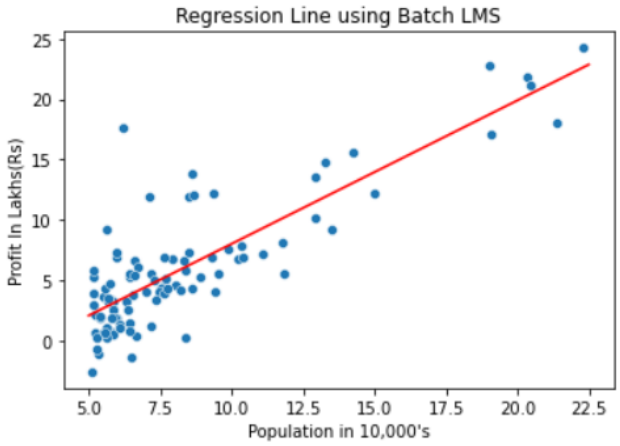


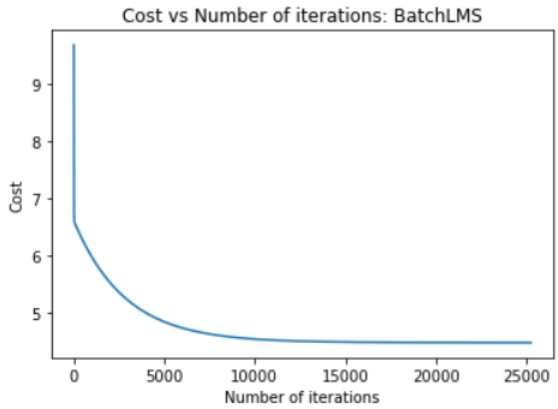

```
------------------------------------------------
Number of Outliers Detected
Standard Deviation Approach:  48
Standard MAD Approach:  45
For the values, refer "q1/Answer1_data_3.ipynb"
------------------------------------------------
```

Please note the values have not been mentioned here due to outliers being in larger quantitites.

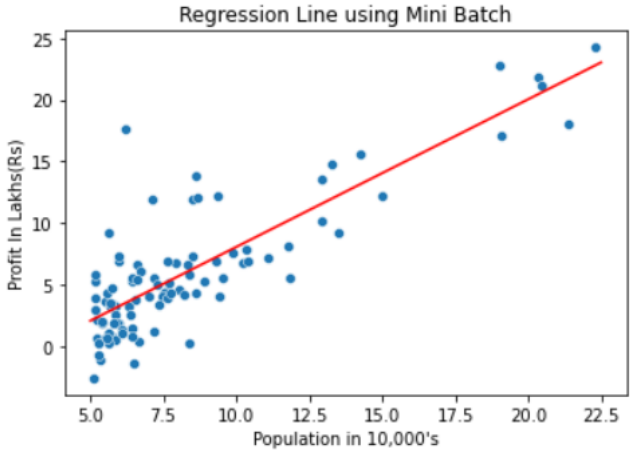
2. You are CEO of a clothing company with outlets in many cities. You have decided to open an outlet in a new city. To help with the decision of selecting a city, you decide to look at population vs profit data and apply linear regression to see if any relation exists between population & profit with population being the independent variable.
    a. Apply Batch LMS, Mini batch, Stochastic LMS and Least Square closed form solution and compare the results. Plot the graphs of the obtained results and training data. Use the learning rate of 0.001. Analyse the results (Convergence time, accuracy etc.) (Don't use in-built packages.)

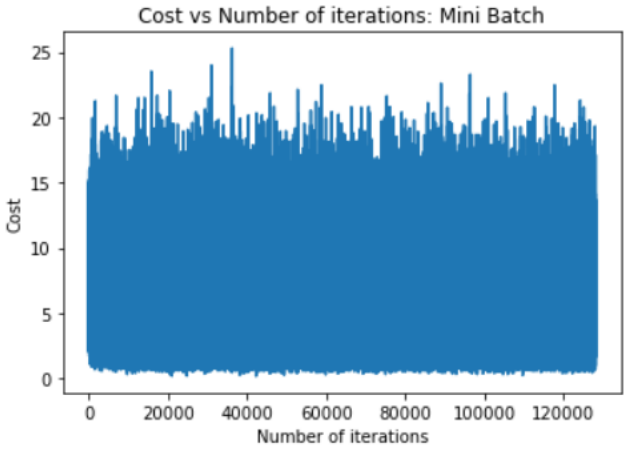
```
------------------------------------------------------------
Batch LMS
Number of Epoches and Iterations:  25255
The Values of theta for Batch LMS Method [[-3.85879218]
 [ 1.18742286]]
The Final cost is:  4.477254157145168
Time Taken:  0.6852922439575195 seconds
------------------------------------------------------------
```



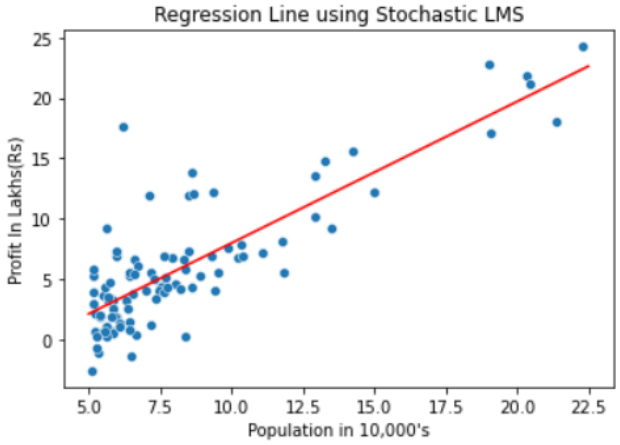


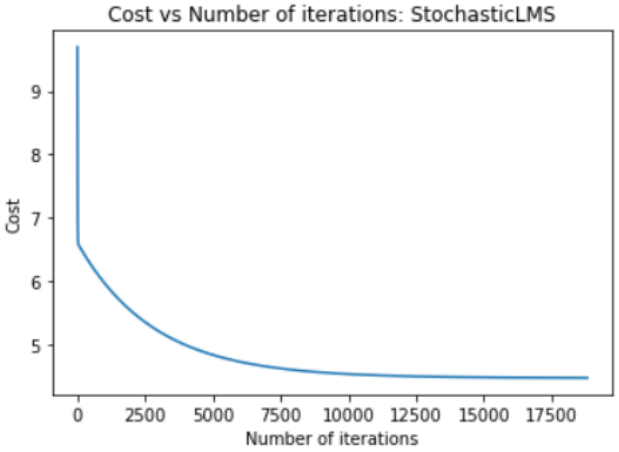
```
------------------------------------------------------------
Mini Batch LMS
Number of Epoches:  12802
Number of iterations:  128020
The Values of theta for Batch LMS Method [[-3.91426366]
 [ 1.19724299]]
The Final cost is:  4.912035542489459
Time Taken:  5.734387159347534 seconds
------------------------------------------------------------
```



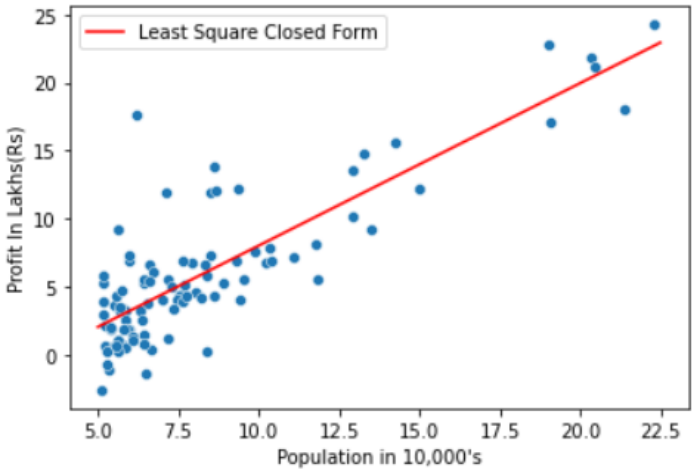


```
------------------------------------------------------------
Stochastic LMS
Number of Epoches and iterations:  18815
The Values of theta for stochastic LMS Method [[-3.73455932]
 [ 1.17092425]]
The Final cost is:  4.480596156620541
Time Taken:  26.55198621749878 seconds
------------------------------------------------------------
```

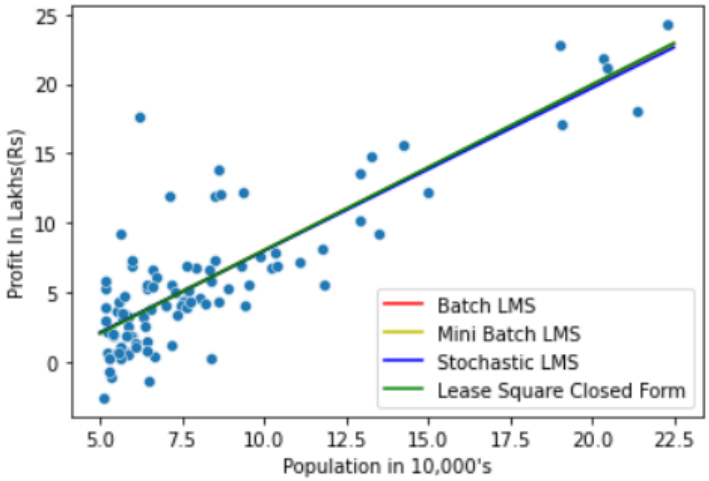




```
------------------------------------------------------------
Least Square Closed Form
The Values of theta for Least Square Closed Form:  [[-3.91508424]
 [ 1.19303364]]
The Final cost is:  4.476971375975179
Time Taken:  0.0 seconds
------------------------------------------------------------
```



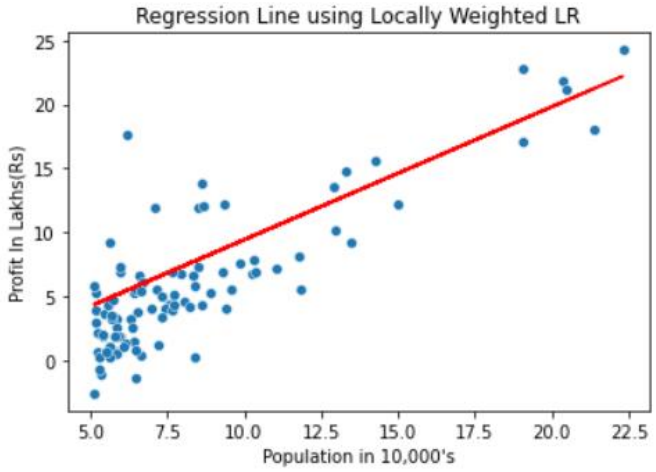
Comparision between above methods

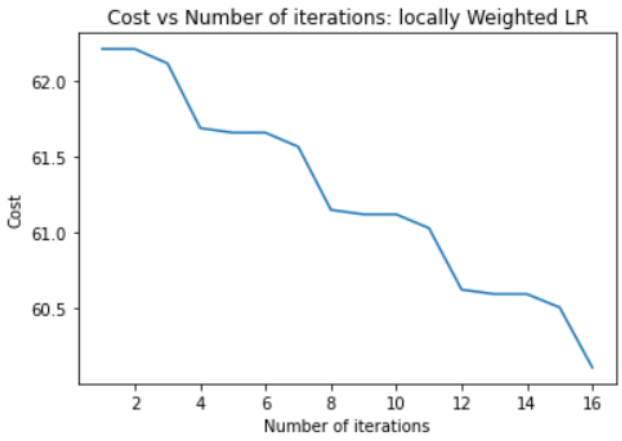


B. Manually perform the locally weighted least linear regression using the first four data points given in excel sheet. Query point is 7.576 and bandwidth parameter is 0.5. Perform four iterations by using stochastic LMS.

Calculation using Python for 16 iterations covering 4 length epoch four times:

```
------------------------------------------------------------
Locally weighted Linear regression
Number of Epoches:  4
Number of Iterations:  16
The Values of theta for Locally weighted LR [-0.96170356  1.03829644]
The final cost is:  60.11176501219228
Time Taken:  0.0039958953857421875 seconds
------------------------------------------------------------
```





Manual Calculation by hand:

Locally weighted Linear Regression:

b) ~~Linearly~~

Population in 10,000's	Profit in Lakhs.
6.2101	17.6920
5.6277	9.2302
8.6186	13.7620
7.1032	11.9540

$$w^{(i)} = \exp\left(-\frac{(x^i - x)^2}{2\tau^2}\right) \qquad x_0 = 7.576, \quad \tau = 0.5$$

$$x = \begin{bmatrix} 17.6920 \\ 9.2302 \\ 13.7620 \\ 11.9540 \end{bmatrix} \qquad w = \begin{bmatrix} 2.3\times10^{-2} \\ 5.04\times e^{-4} \\ 1.13\times e^{-1} \\ 8.3949\times e^{-1} \end{bmatrix}$$

$$J(\theta) = \frac{1}{2}\sum w^{(i)}(\theta^T(x^{(i)} - y^{(i)})^2$$

Iteration-1 $\quad y_{pred} = \begin{bmatrix} 6.2101 \\ 5.6277 \\ 8.6186 \\ 7.1032 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$

$y_{pred} \Rightarrow [0, 0, 0, 0]$

$\theta_0 = \theta_0 + \tau\alpha w^{(i)}(y_{pred}^{(i)} - \theta_0 - \theta_1 x^{(i)}) = 0.042$

$\theta_1 = \theta_1 + \tau\alpha w (y_{pred}^{(i)} - \theta_0 - \theta_1 x^{(i)}) x^{(i)} = 0.262$

$\theta = [0.042, 0.262]$

Locally weighted Linear Regression:

b) ~~Linearity~~

Population in 10,000's	Profit in Lakhs.
6.2101	17.6920
5.6277	9.2302
8.6186	13.7620
7.1032	11.9540

$$w^{(i)} = \exp\left(-\frac{(x^i - x)^2}{2\tau^2}\right) \qquad x_0 = 7.576, \quad \tau = 0.5$$

$$\therefore \quad x = \begin{bmatrix} 17.6920 \\ 9.2302 \\ 13.7620 \\ 11.9540 \end{bmatrix} \qquad w = \begin{bmatrix} 2.3 \times 10^{-2} \\ 5.04 \times e^{-4} \\ 1.13 \times e^{-1} \\ 8.3949 \times e^{-1} \end{bmatrix}$$

$$J(\theta) = \frac{1}{2} \sum w^{(i)} (\theta^T (x^{(i)} - y^{(i)}))^2$$

Iteration-1 $\quad y_{pred} = \begin{bmatrix} 6.2101 \\ 5.6277 \\ 8.6186 \\ 7.1032 \end{bmatrix} * \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$

$$y_{pred} \Rightarrow [0, 0, 0, 0]$$

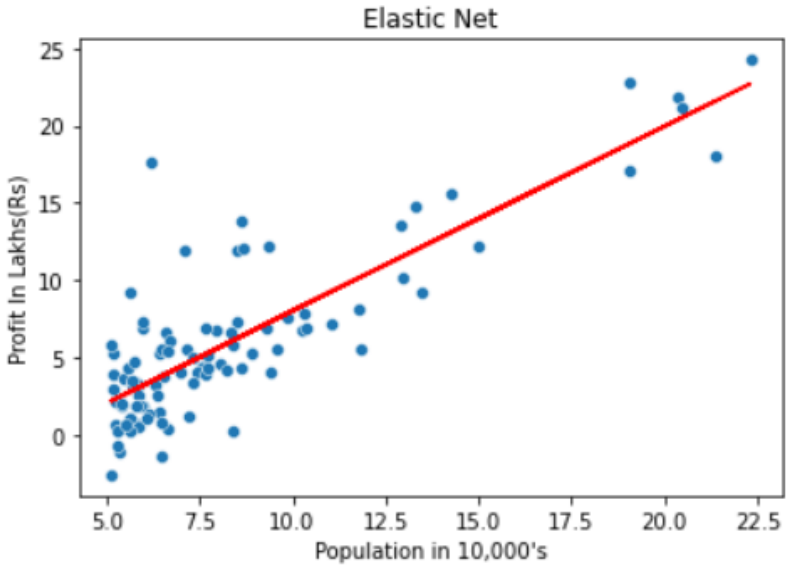
$$\theta_0 = \theta_0 + \tau \alpha w^{(i)} (y_{pred}^{(i)} - \theta_0 - \theta_1 x^{(i)}) = 0.042$$

$$\theta_1 = \theta_1 + \sum \alpha w (y_{pred}^{(i)} - \theta_0 - \theta_1 x^{(i)}) x^{(i)} = 0.262$$

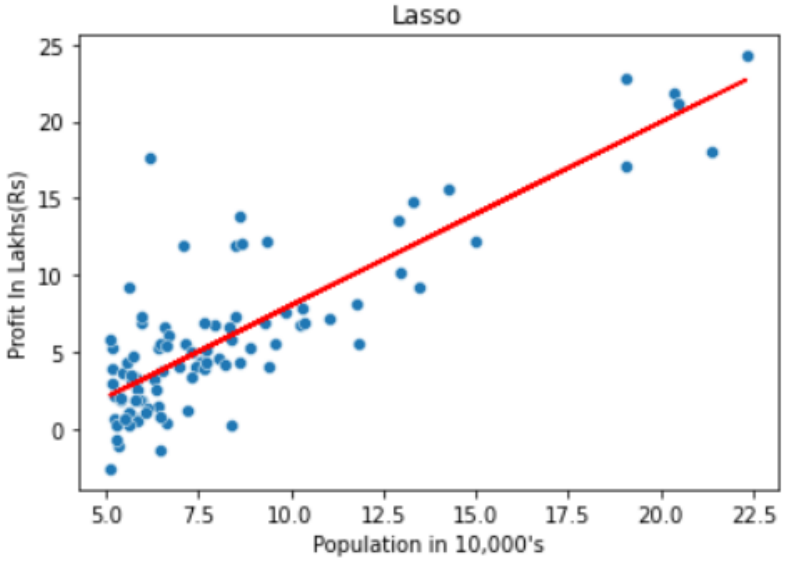
$$\theta = [0.042, 0.262]$$

c. Compare the results of Elastic net, Lasso and Ridge regression. (Use in-built packages)

```
------------------------------------------------------------
Elastic Net
The Values of theta for Elastic Net [-3.9144732] [1.19295967]
------------------------------------------------------------
[<matplotlib.lines.Line2D at 0x2b062b23880>]
```

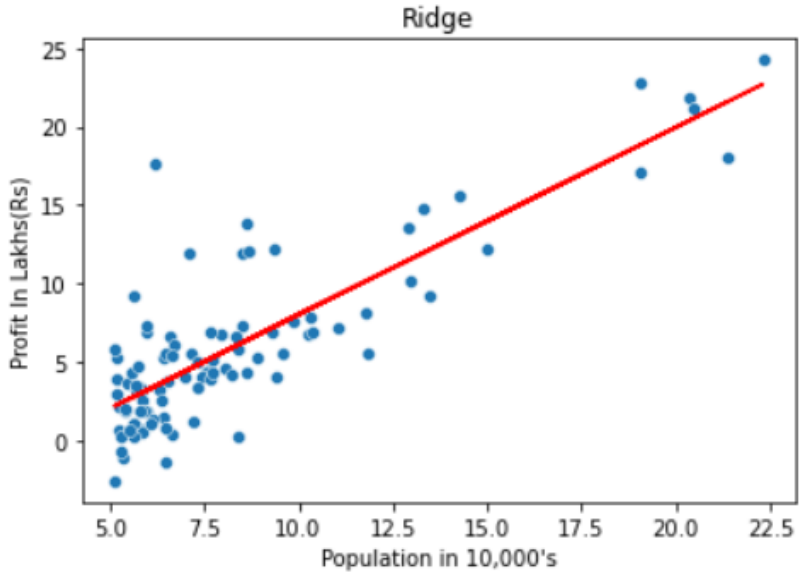


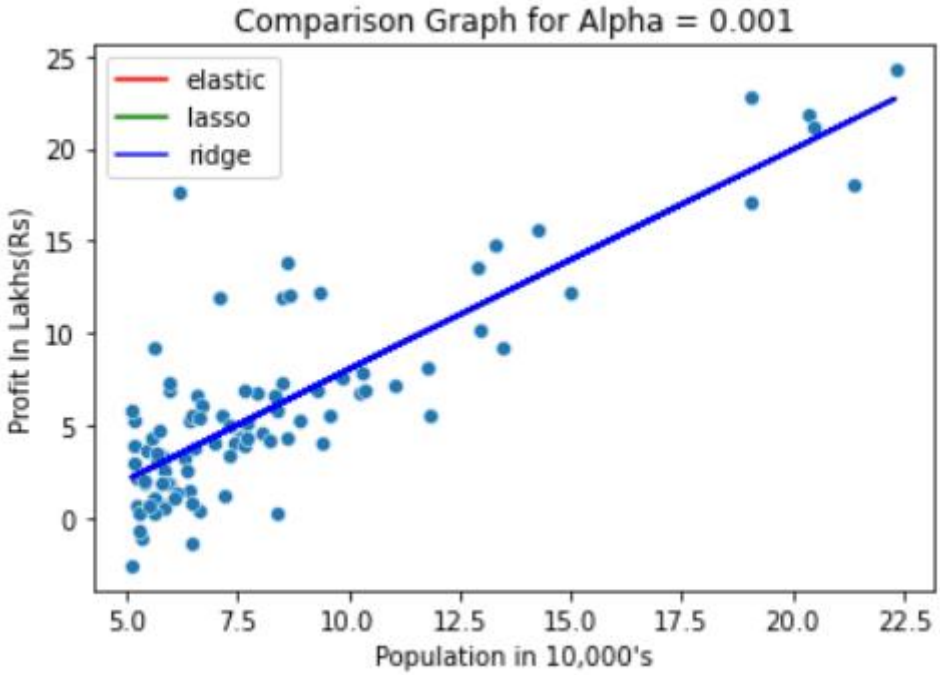
```
------------------------------------------------------------
Lasso
The Values of theta for Lasso [-3.91452696] [1.19296618]
------------------------------------------------------------
[<matplotlib.lines.Line2D at 0x2b06e9a76a0>]
```



```
------------------------------------------------------------
Ridge
The Values of theta for Ridge [-3.91507739] [[1.19303281]]
------------------------------------------------------------

[<matplotlib.lines.Line2D at 0x2b062e5d880>]
```



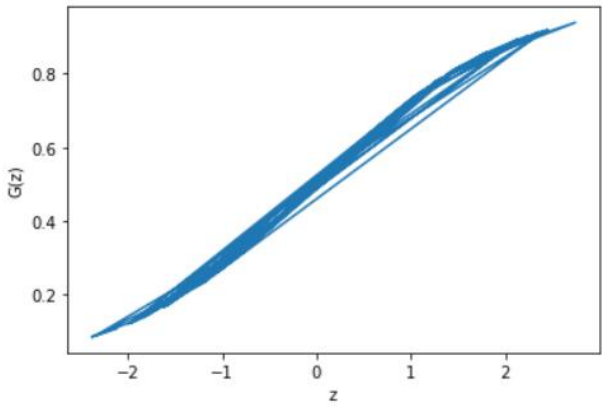


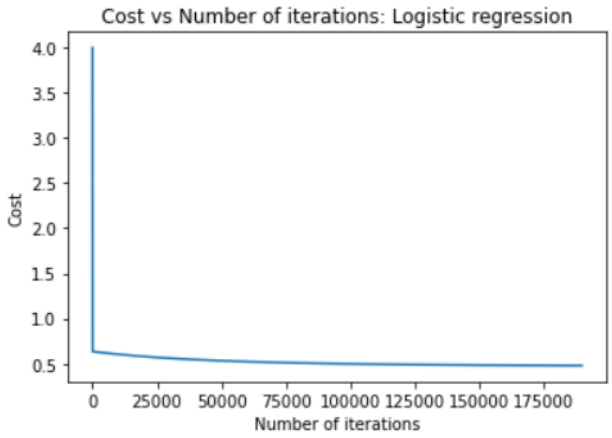
3. A university conducts 2 exams – Aptitude & Verbal as its entrance test to a 2-year program. Based on the scores of these 2 papers, admission is given to students. University has not mentioned the exact criteria of selection. Based on historical data, you need to predict whether a student will get admission based on his/her scores in the 2 exams. Data is provided

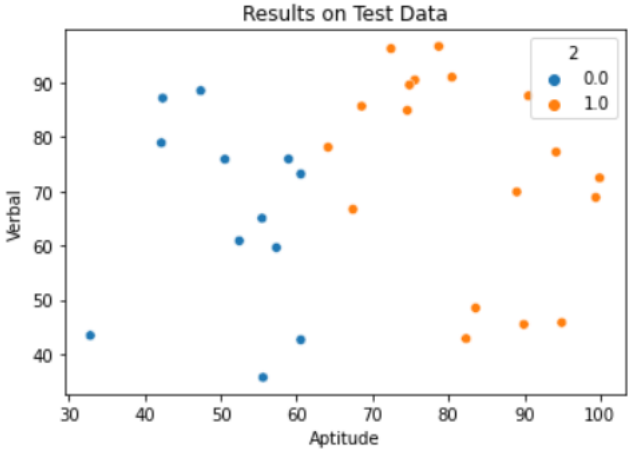
in q2train.csv & q2test.csv. Train.csv contains training data. First column contains the score of Aptitude exam, 2nd column contains the score of verbal exam and 3rd column indicates whether that student got admission or not. 0 indicates not selected whereas 1 means selected. q2test.csv contains test data.

a. Apply logistic regression on training data with the first 2 columns as input data and the third column as output. Use any suitable learning rate. Now predict admission results on test data (q2test.csv) and print the result in output1.txt with every line of the text file containing either 0 or 1. Plot the results. (Don't use in built packages.)

```
------------------------------------------------------------
Logistic Regression
Number of Iterations:  189561
Time Taken:  7.350202322006226
The Values of theta for Logistic regression [[-4.71972859]
 [ 0.06900916]
 [ 0.00665459]]
The Final cost is:  0.480046705755547
------------------------------------------------------------
```







5. Visualize a decision tree that predicts whether tennis will be played on a particular day. Find out the root node of that tree. Use (play_tennis.xslx) dataset for creating the classification tree. Solve this problem manually using the ID3 algorithm.

Calculation has been done using python since manual calculations were too large in numbers to be done.

Code can be found in "./q1/Answer5.ipynb"

```
------------------------------------------------------------
Decision Tree Classifier
For the Following input:  {'Outlook': 'Rainy', 'Temperature': 'Mild', 'Humidity':
'High', 'Windy': 'Weak'}
Result is:  Yes
------------------------------------------------------------
Tree Decision Visualization: {'Outlook': {'Sunny': {'Humidity': {'High':
{'Temperature': {'Hot': {'Windy': {'Weak': 'No', 'Strong': 'No'}}, 'Mild': 'No'}},
'Normal': {'Temperature': {'Cool': 'Yes', 'Mild': 'Yes'}}}}, 'Overcast':
{'Temperature': {'Hot': {'Humidity': {'High': 'Yes', 'Normal': 'Yes'}}, 'Cool': 'Yes',
'Mild': 'Yes'}}, 'Rainy': {'Windy': {'Weak': {'Temperature': {'Mild': {'Humidity':
{'High': 'Yes', 'Normal': 'Yes'}}, 'Cool': 'Yes'}}, 'Strong': {'Temperature': {'Cool':
'No', 'Mild': 'No'}}}}}}
------------------------------------------------------------
```