Name: Prathamesh N. Lonkar

B-Number: B00811727

1. (20 pts.) Determine the distributions of the vector component values for both datasets. For each dataset, randomly pick up 10 samples and report the distribution parameters for each of the 10 samples.

Ans.) After taking 10 samples when we plotted the vectors on a histogram where we came to the conclusion that the first dataset is of uniform distribution and the distribution for the second dataset is of Gaussian distribution.

The distribution parameters for each dataset are as follows:

min and max for dataset1:

[[2.8249, 264.47],

[2.5124, 91.95200000000001],

[2.8202, 260.5],

[0.69226, 210.53],

[2.0823, 183.19],

[0.31656, 200.65],

[1.4891, 148.72],

[5.3418, 212.48],

[5.1806, 172.08],

[1.1581, 350.23]]

mean and standard Deviation for dataset1 is: Where 1st value in each array is the mean and second is the standard deviation

[[130.99282899999997, 72.98327697792735],

[43.45846699999999, 25.19970462730488],

[143.75709200000006, 76.81609302414395],

[103.4966876, 65.13418699985425],

[86.15304300000004, 55.698602909561835],

[103.57654009999996, 58.01378757948532],

[80.45554299999999, 43.314700692982406],

[107.33760099999998, 62.08759952990533],

[82.35331299999999, 49.78512927748738],

[170.32301099999992, 98.54138517901997]]

min and max for dataset2:

[[124.6, 337.38],

[0.17076, 95.32],

[115.6, 284.13],

[105.27, 413.66],

[142.62, 391.86],

[125.31, 389.48],

[182.16, 552.8],

[118.91, 267.14],

[0.9051299999999999, 239.14],

[177.6, 423.95]

mean and standard Deviation for dataset2 is:

[225.49199999999993, 46.00005560866204],

[47.73649259999999, 27.573681276630822],

[200.11899999999997, 33.86352815345737],

[273.6412999999999, 57.75984847028255],

[271.24430000000007, 51.99496487651473],

[275.0513, 56.46244005805984],

[358.2119999999998, 76.36980075920061],

[190.11180000000007, 34.25862015259809],

[119.30088330000001, 71.39630573308924],

[291.866, 52.48498072782345]

When we plot the dataset using the qqplot we get a plot indicating all the datapoints as a straight line hence we conclude that the distribution is Gaussian for the second dataset.

Note that the min, max, mean, standard deviation as mentioned above will keep on changing when we run our program multiple times as we are takin 10 random samples from the whole dataset.

For this question i have used the standard libraries as follows:

numpy: For Linear Algebra

pandas: data processing, CSV file I/O

matplotlib: data plotting

sklearn.preprocessing: used for standard scalling of data.

Q2.) (50 pts.) Implement PCA and DCT methods and apply them for feature extraction to the two datasets, respectively. Report the principle you have proposed to truncate the dimensionality and the reduced dimensionalities for the two datasets after the feature extraction for PCA and DCT, respectively.

Ans.) After implementation of PCA on the two datasets the PCA will be responsible for maximizing the variance by selecting a particular number of data points. Here we manually specify the truncation parameters to the PCA here I have given n\_components a 2 which means that the PCA will reduce the dimensionality to nx2 and will do so by maximizing the variance. If we increase the the truncation parameters then the variance will change accordingly. There will be a time when the variance does not change to a great extent, we can remove those datapoints or truncate before those data points.

For the PCA &DCT feature Extraction we can use 2 method which are, Either we can manually give the maximum number components to be considered or we can apply the power law to truncate the dataset.

For these data samples i have manually given the number of components to be considered which is 2 which obtained a variance of approximately 74%, we can increase the variance by increase the number of components to be considered which will in turn give a higher variance but only till a certain point.

For DCT we can use another method in which we can remove the outliers in the program using Boxplot which contains interquartile range Q1 to Q3 the if most of the box plot lies between the first 2 quartile range and the rest of the data is lies far from the box plot then they are removed and considered as outliers.

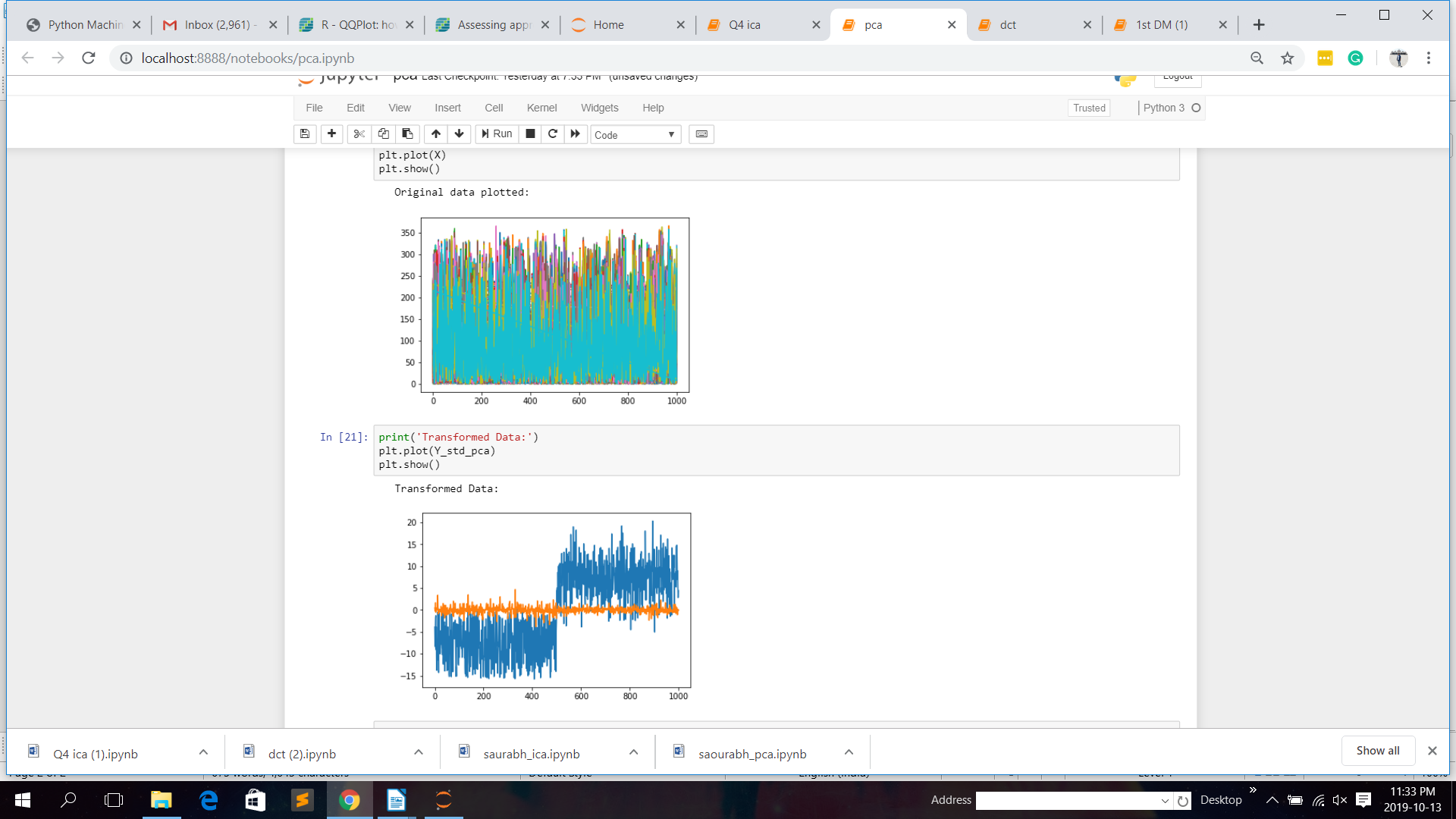
Q3.) (30 pts.) Compare the feature extraction results between the two methods for the two datasets, respectively, and report your comparison conclusion.

Ans.)

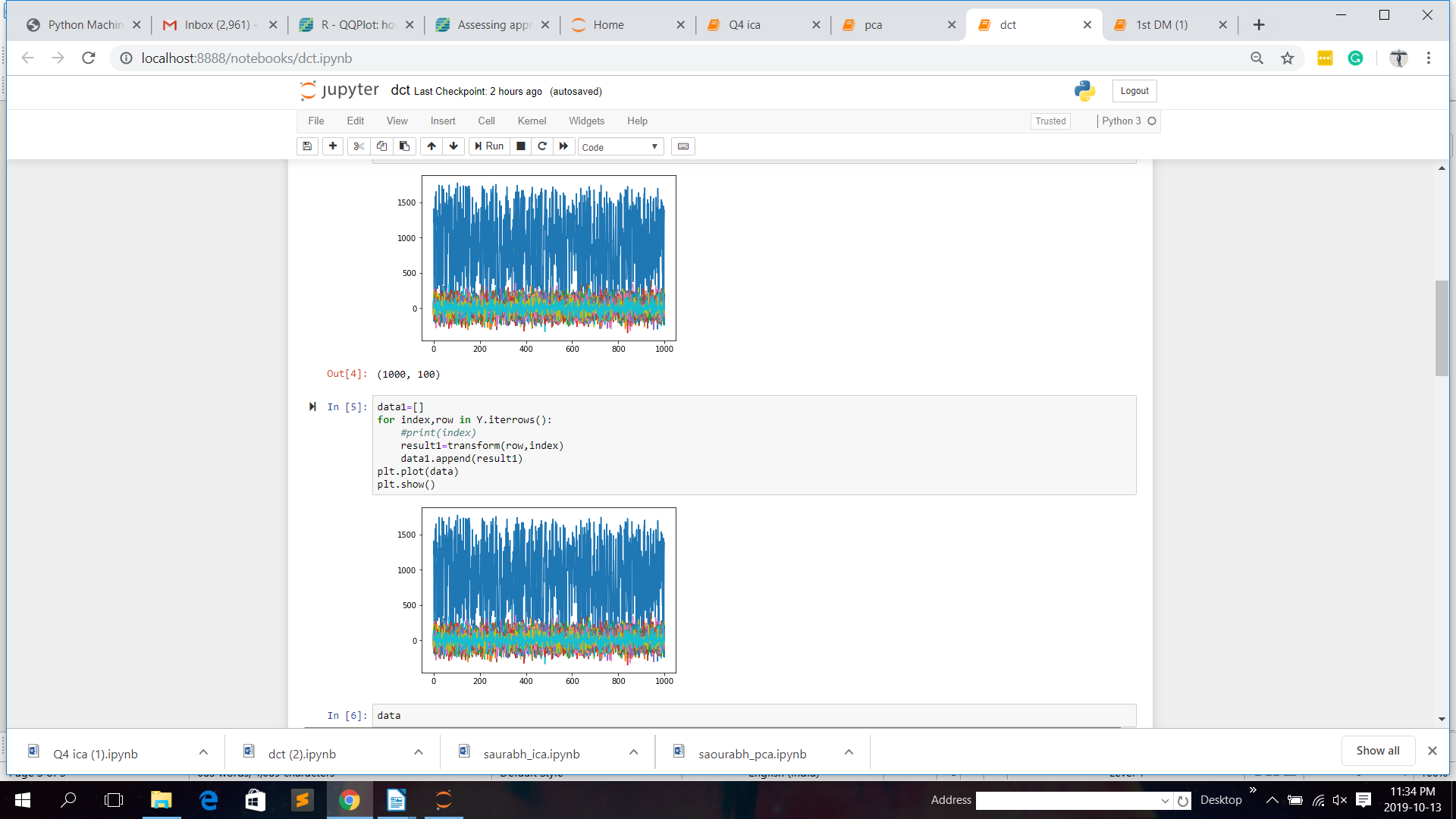
* For the PCA selects the data points to maximize the variance hence even though we have truncated it to the lowest possible dimensions it will still try to give the maximum variance.

While in DCT we will transform the vector components and then truncate the data or reduce dimensionality

* For PCA we will get the following graphs:



* For DCT we will get the following results:



Q4.)(50 pts.) Read the literature on Independent Component Analysis (ICA) and implement ICA. Then apply ICA to the two datasets, respectively. Report your comparison studies on the two datasets between PCA and ICA on feature extraction.

Ans.)

* Firstly in PCA we will plot the graph according to the variance. Whereas for ICA we take independent components to calculate the result.
* In PCA input and output for all the datapoints are independent of each other where as ICA

Input and output are mapped to each other

* In PCA mutual information is shared within the datapoints such as the mean and std deviation etc.

The Output for the ICA is as follows:

