Assignment I: Introduction to Data Mining
Submitted By: Harshal Rasal
B-Number: B00816445

Q1.

1. Distributions of the vector component values for both datasets?

Answer: I have found out that all of them are much closer to the Gaussian Distribution.

2. For each dataset, randomly pick up 10 samples and report the distribution parameters for each of the 10 samples?

A.

```
 Dateset 1 :

Sample : 1 Mean : 137.6682985 STD :  88.40874268873087
Sample : 2 Mean : 159.681627 STD :  87.60167155028363
Sample : 3 Mean : 114.9144778 STD :  66.84431867783732
Sample : 4 Mean : 130.5938003 STD :  75.15781252210024
Sample : 5 Mean : 144.09173200000004 STD :  79.95283446624751
Sample : 6 Mean : 136.02236200000002 STD :  71.09430042376081
Sample : 7 Mean : 65.867902 STD :  35.9140882944924
Sample : 8 Mean : 46.7711145 STD :  27.58930989565188
Sample : 9 Mean : 75.71721959999999 STD :  44.16359857398796
Sample : 10 Mean : 173.251065 STD :  90.59840978477203
```

B.

```
 Dateset 2 :

Sample : 1 Mean : 287.8874 STD :  52.84849913895379
Sample : 2 Mean : 280.8196 STD :  53.95392070869364
Sample : 3 Mean : 140.26248999999999 STD :  29.82673408152324
Sample : 4 Mean : 269.8185 STD :  51.372329563199685
Sample : 5 Mean : 183.34685 STD :  37.94016498708855
Sample : 6 Mean : 140.26248999999999 STD :  29.82673408152324
Sample : 7 Mean : 243.67430000000004 STD :  50.96031903854214
Sample : 8 Mean : 254.5552 STD :  50.43111641595891
Sample : 9 Mean : 217.43820000000002 STD :  43.58798390336493
Sample : 10 Mean : 253.66879 STD :  47.629239365602935
```

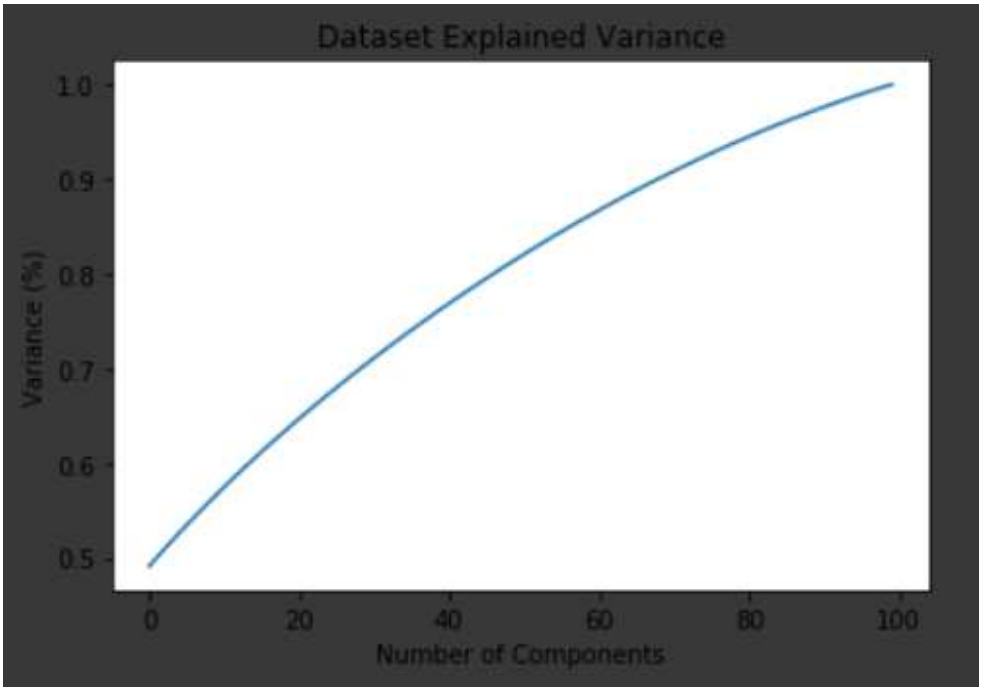
Q2.

1. Implement PCA and DCT methods and apply them for feature extraction to the two datasets, respectively?

Answer: Both are implemented, you may refer to the code.

2. 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?

Answer: First, I have applied the PCA and then calculated the Cumulative Summation of the Explained Variance to check the variance versus number of features. I have found out if we can retain 90% features, we will be able to work with whole data with 90% variance. We use this principle for the PCA, whereas DCT is data-independent and it does the feature extraction process itself as we can remove the values having Zeros from the data.



Q3.

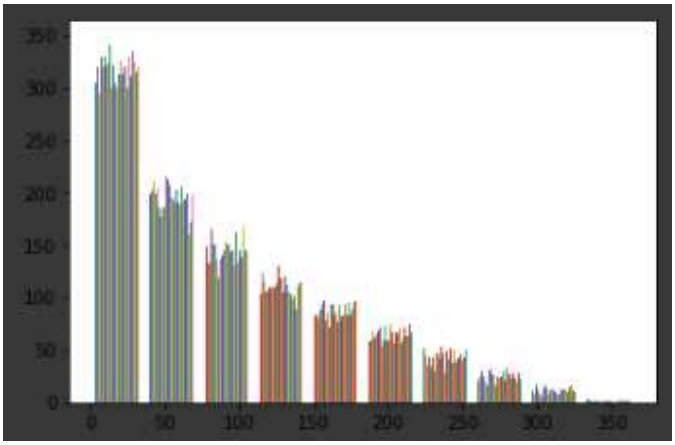
1. Compare the feature extraction results between the two methods for the two datasets, respectively?

Answer: Results are shown in the code as well as below.

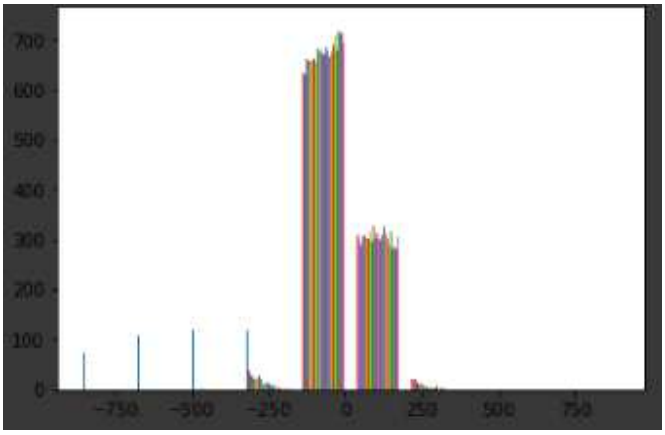
2. Report your comparison conclusion?

Answer: The results on both datasets are very much consistent. The following image of DCT shows that it compresses the values, due to which, most of the values are compressed to ZERO values.

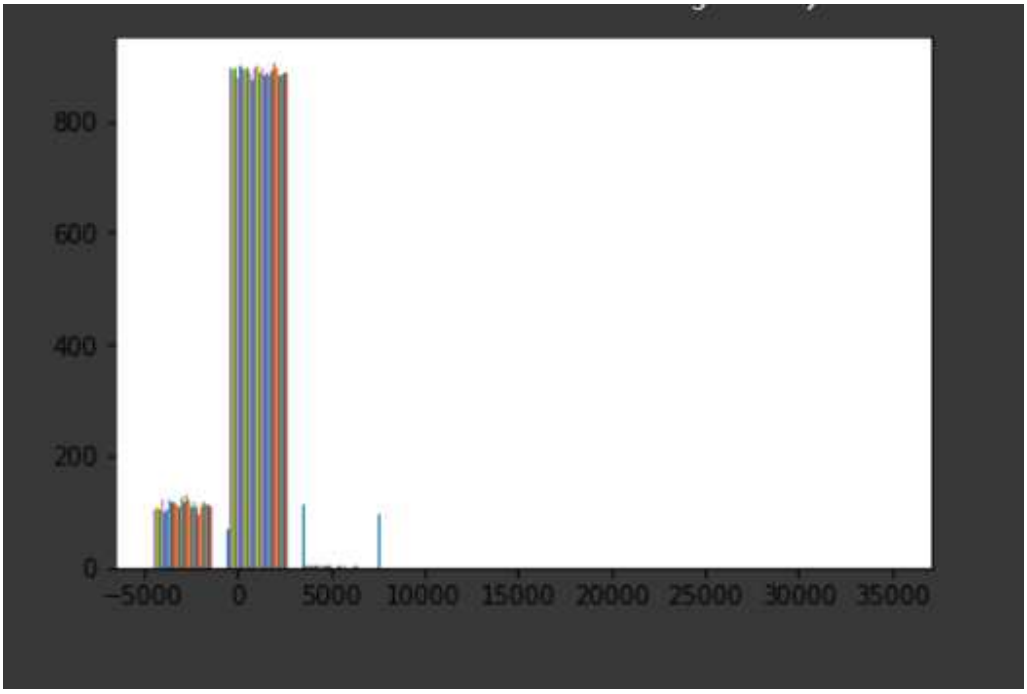
Before PCA, we had following distribution for all the features in the dataset.



On the other side, for PCA, we remove the features ourselves by choosing the dataset features appropriately. PCA result is shown below. These results are of Dataset 1 results.



DCT results are shown below:


800
600
400
200
0
−5000
0
5000
10000
15000
20000
25000
30000
35000

Q4.

1. Read the literature on Independent Component Analysis (ICA) and implement ICA?

Answer: I have implemented the ICA, please refer to the results.

2. Apply ICA to the two datasets, respectively. Report your comparison studies on the two datasets between PCA and ICA on feature extraction?

Answer: ICA behavior is kind of same as PCA, and for the feature extraction step, we follow the PCA principle i.e. we use PCA to find out the number of features needed to be extracted. ICA Results are shown below while PCA results are shown in Q3. One major change, we noticed is that the values are compressed like most of the values are between -0.2 and .2 where PCA has same actual values of dataset. Again, these results are consistent in both the results.