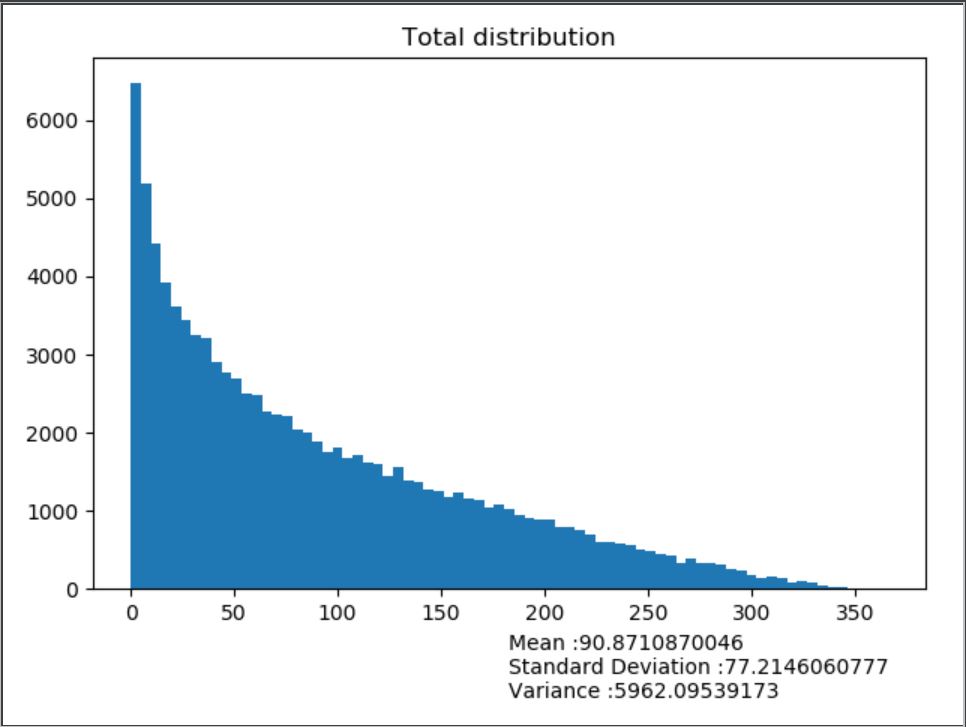
Assignment – I Name: Sushma H

Mail: [shanuma1@binghamton.edu](mailto:shanuma1@binghamton.edu)

ID: B00817656

1. a) Dataset1:

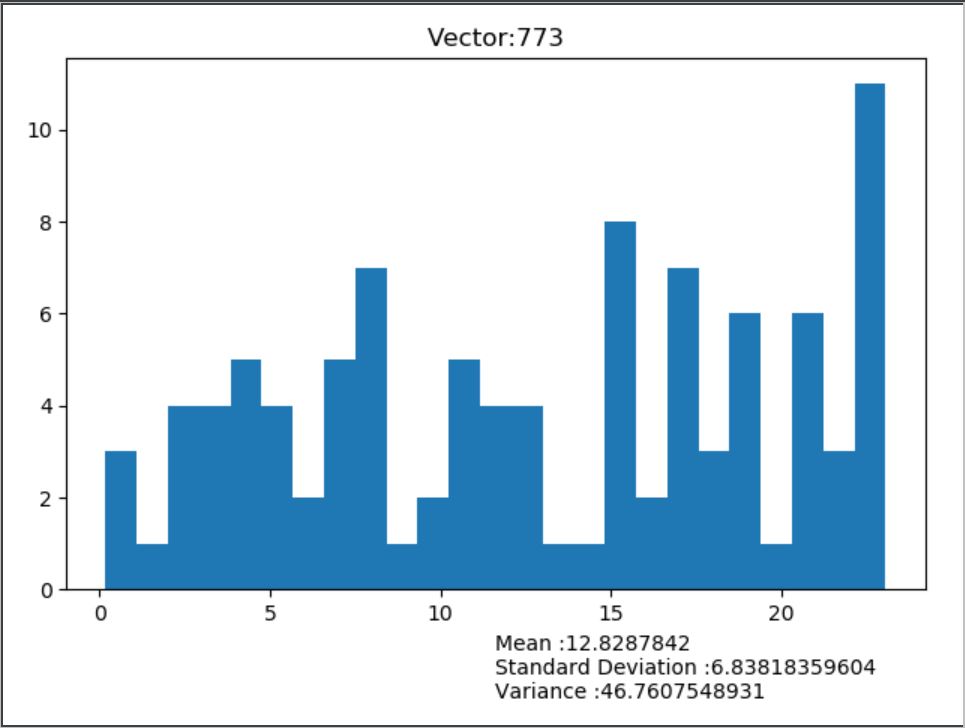
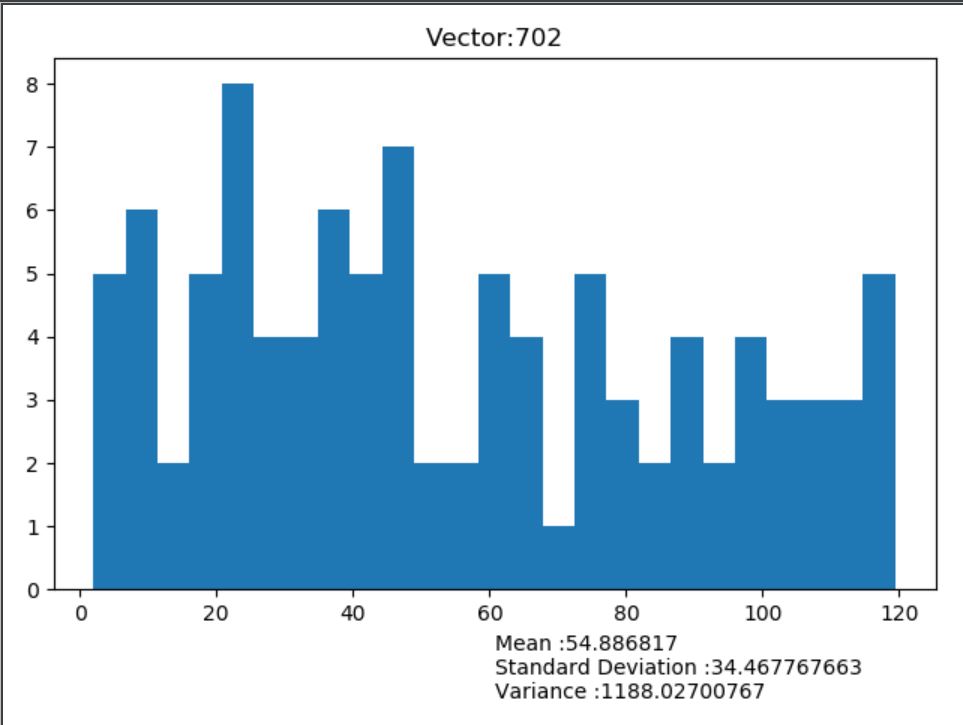
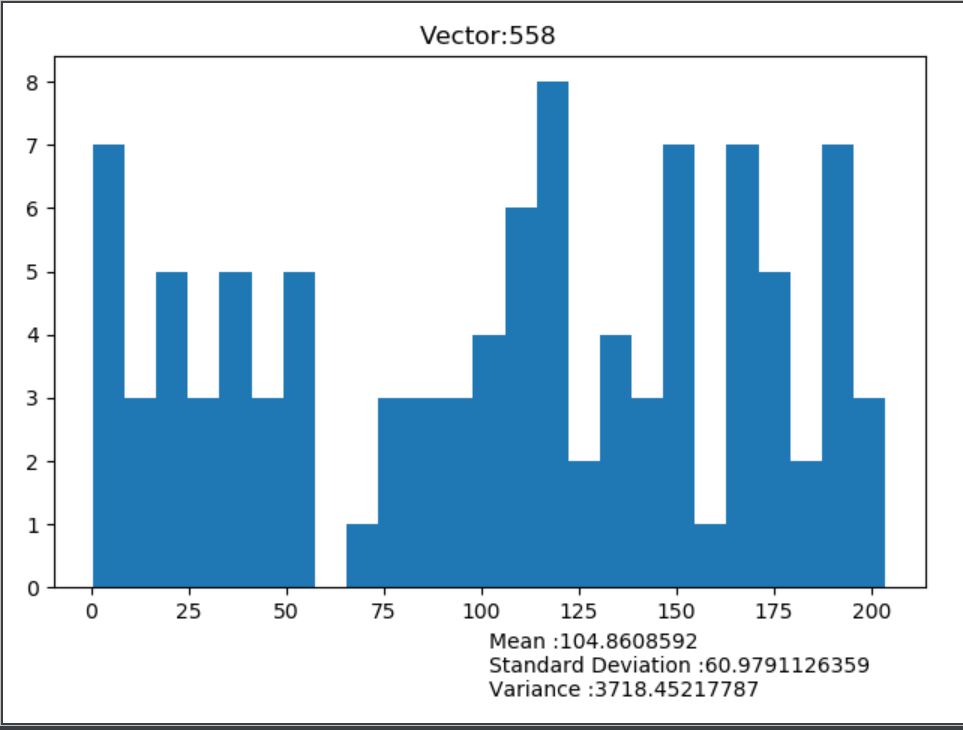
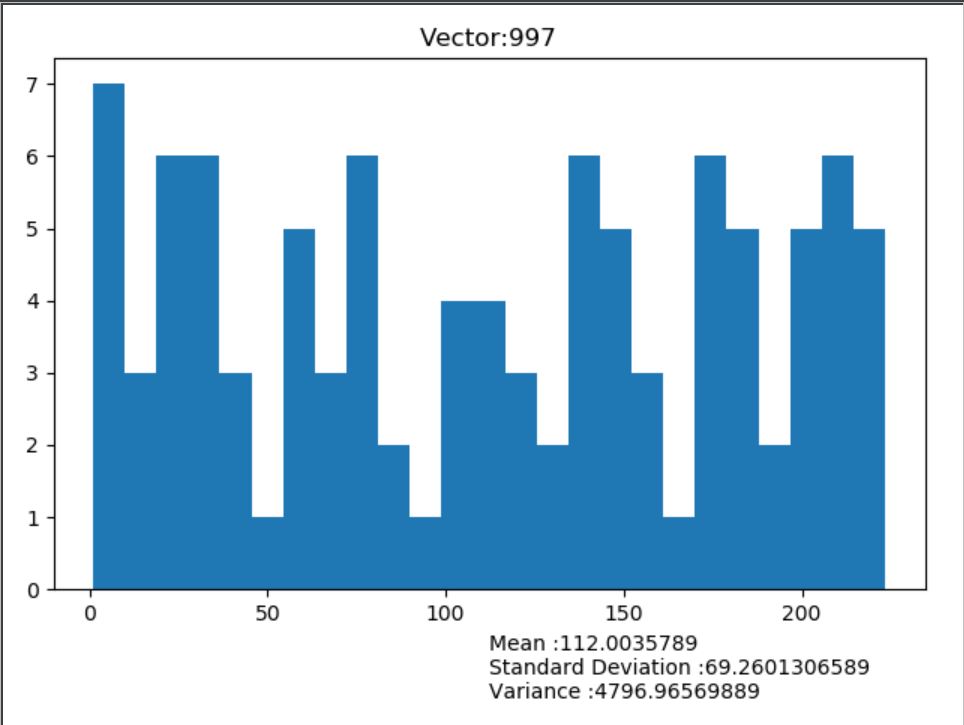
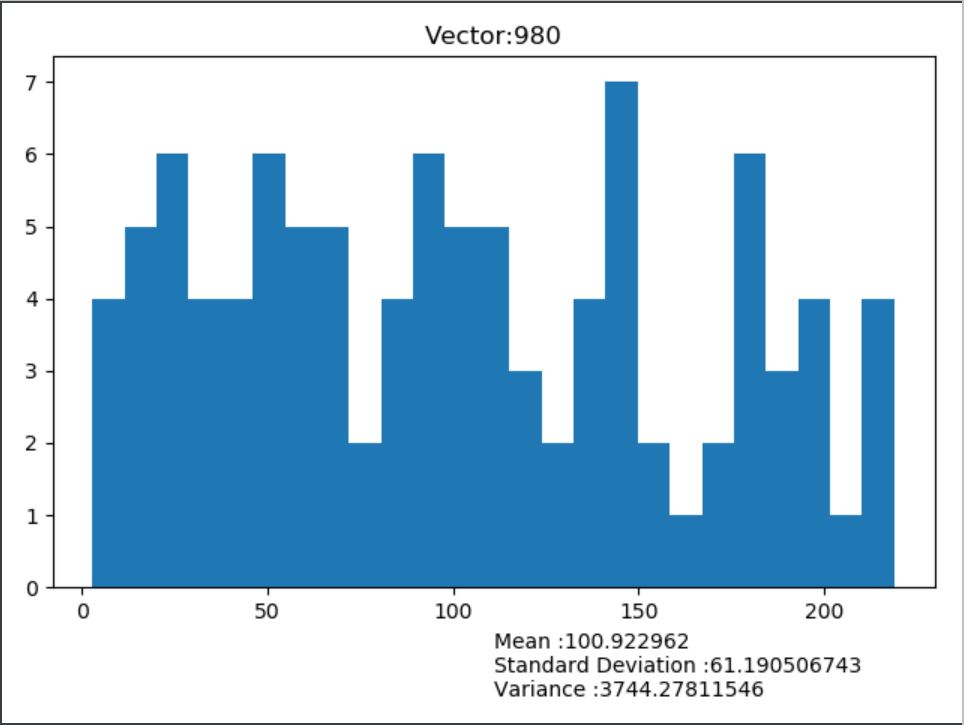
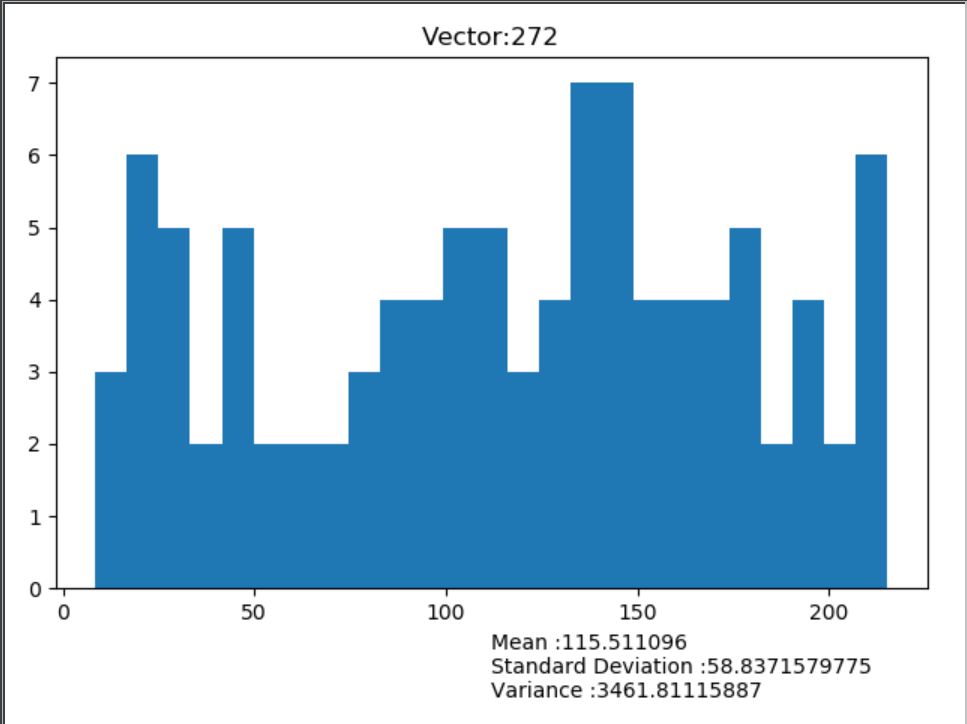
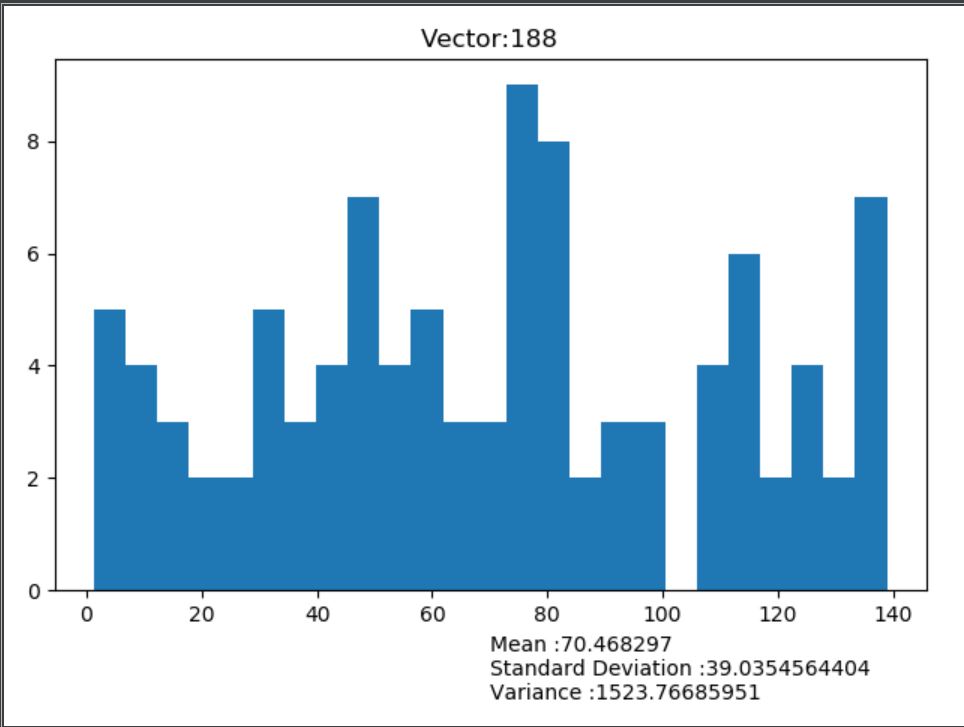
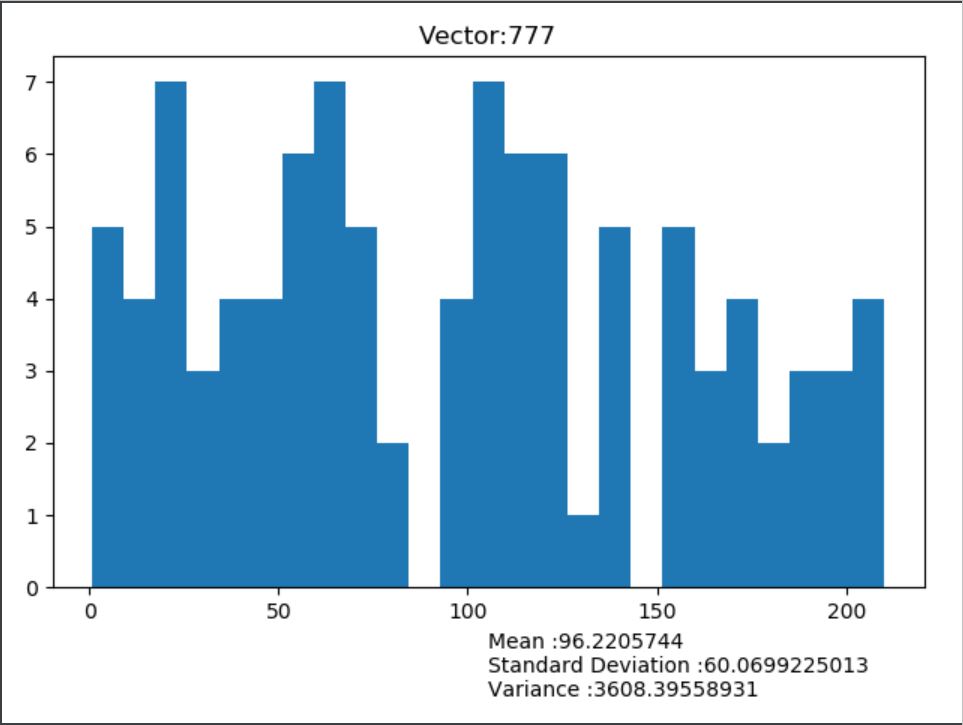
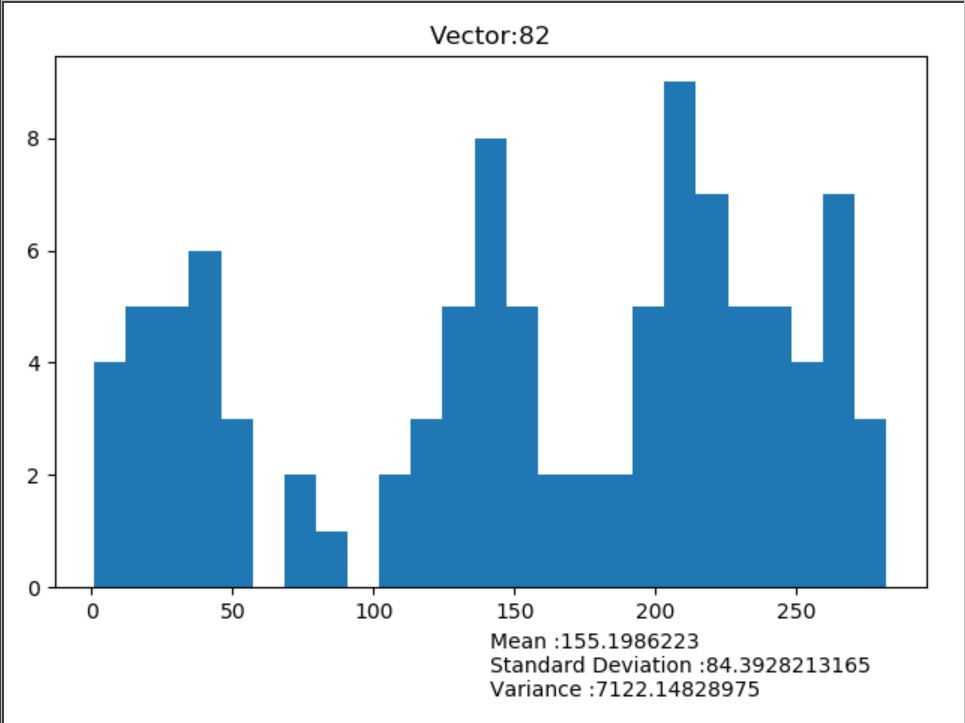
Distribution for the total dataset is plotted using a histogram as shown below,

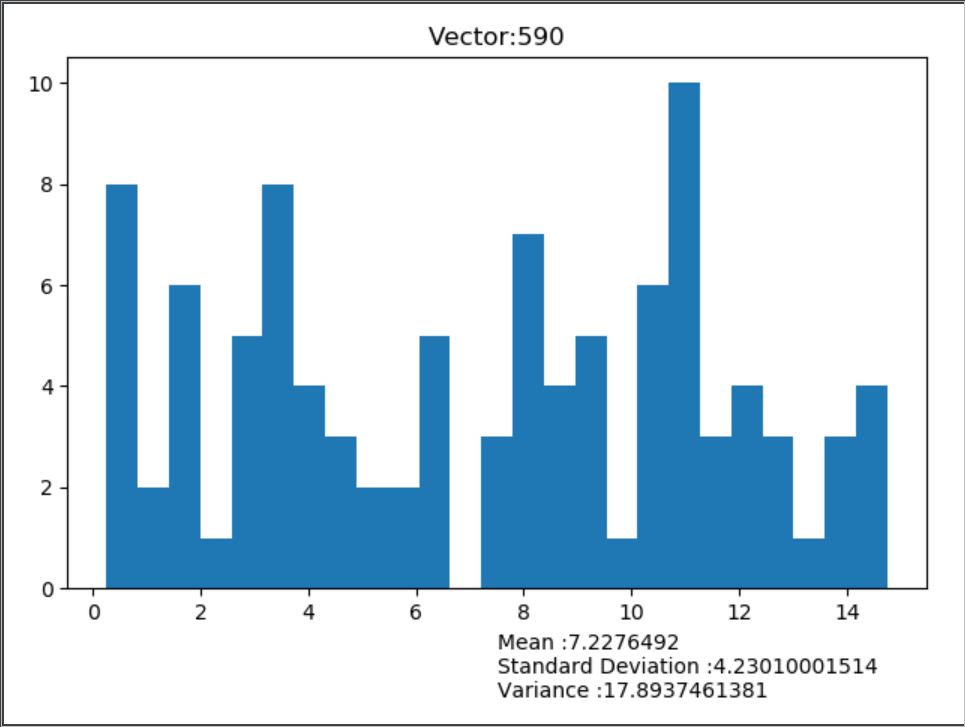


A random of 10 vector components were taken using a random generator and plotted on a histogram. The distribution parameters Mean, standard deviation and variance were calculated respectively.

Note:

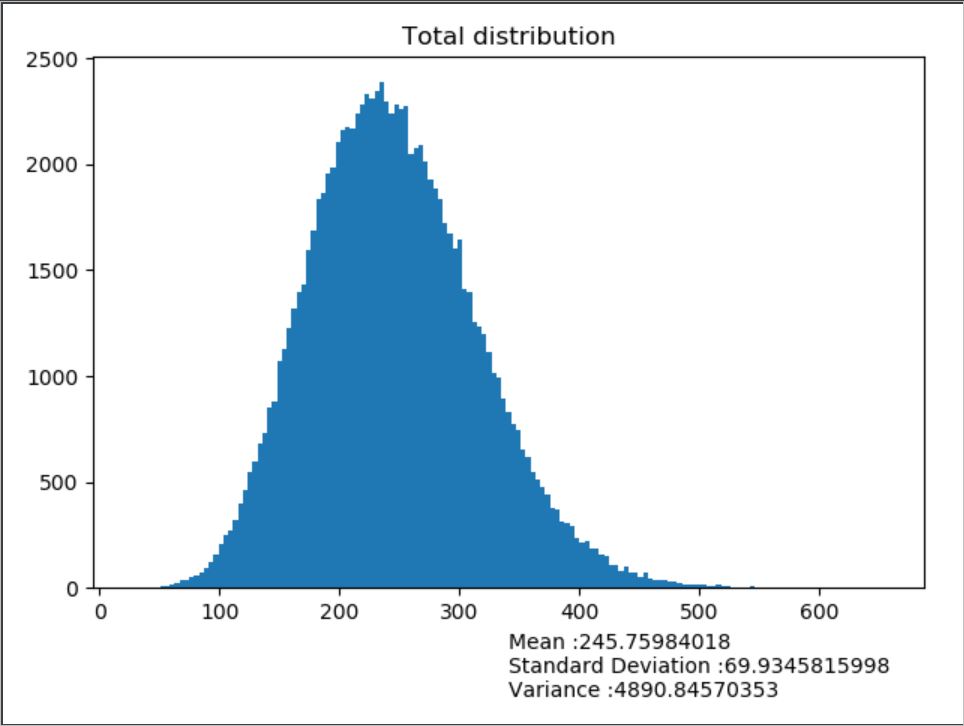
* The vector component values shown in the graph title correspond to the row number in the dataset provided.
* The histograms were plotted using the bin size as 25





b) Dataset 2:

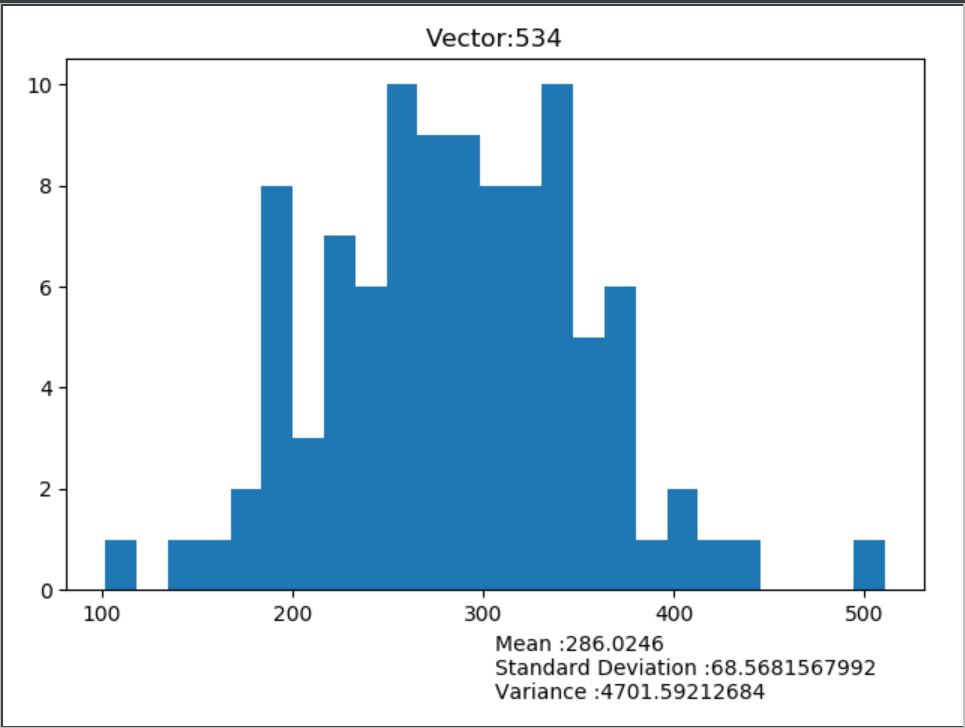
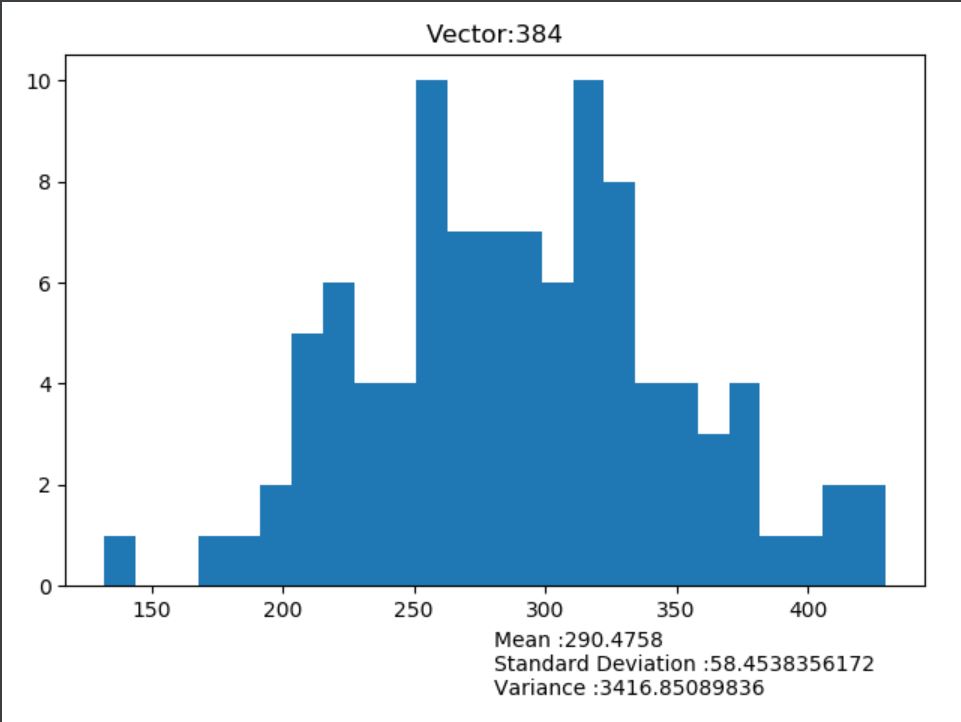
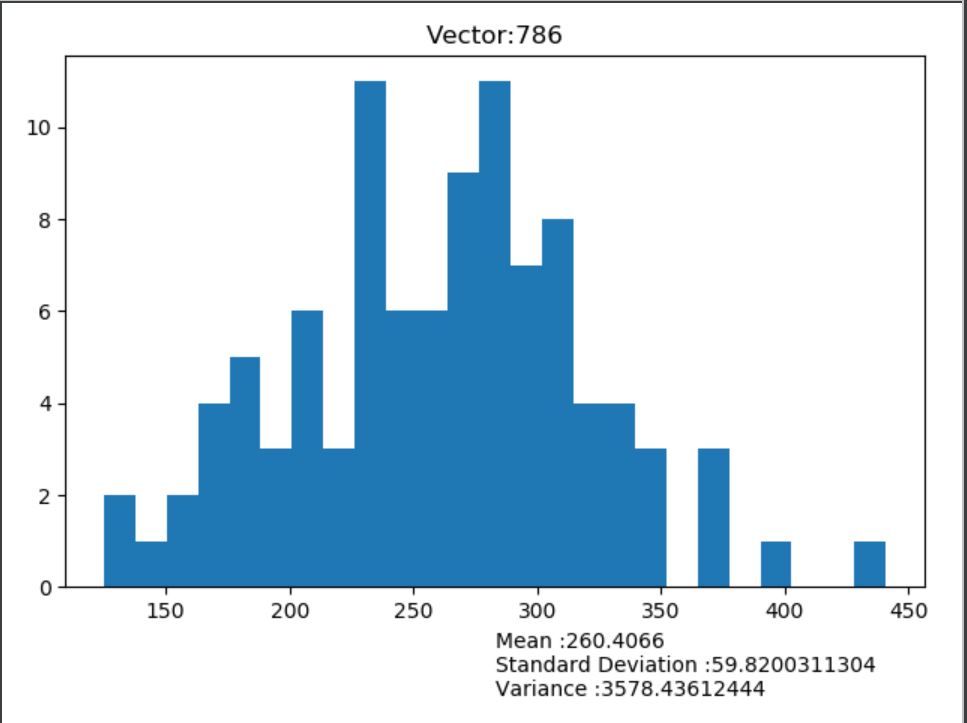
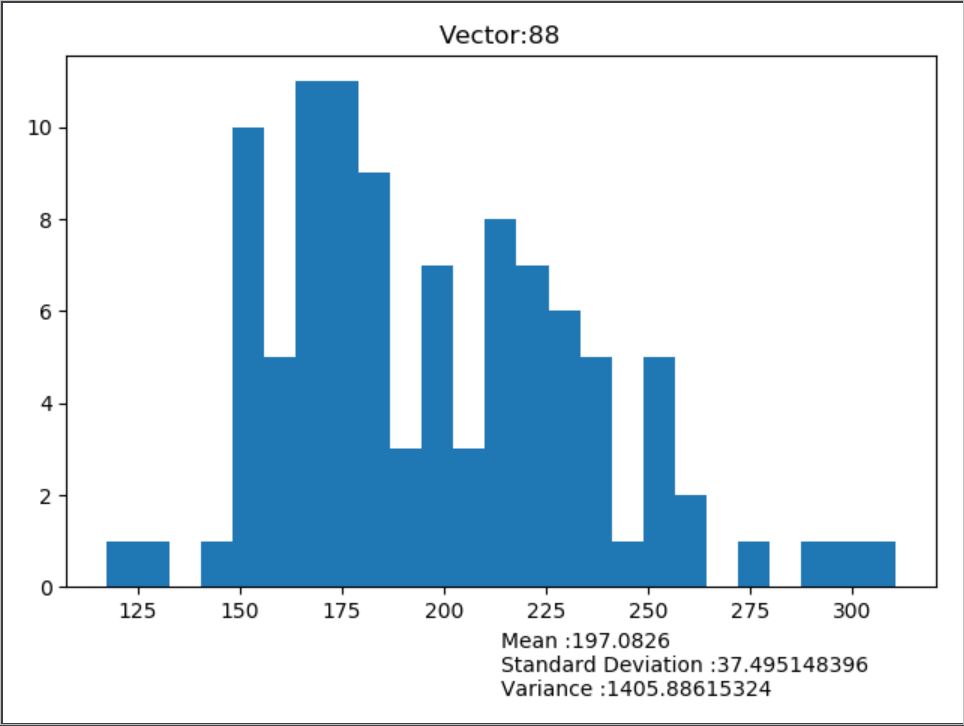
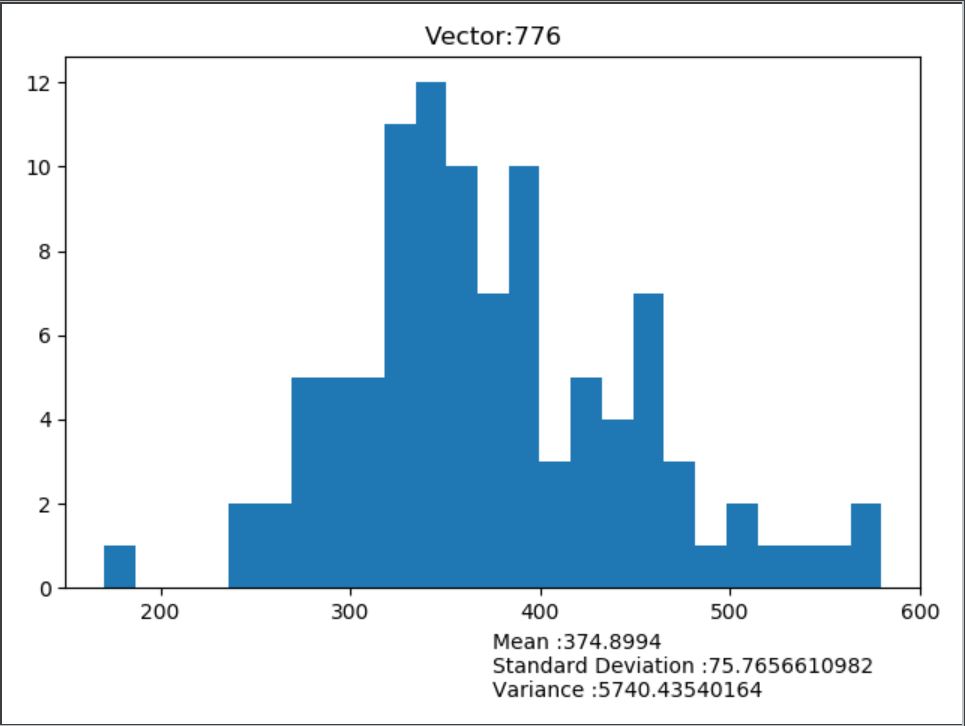
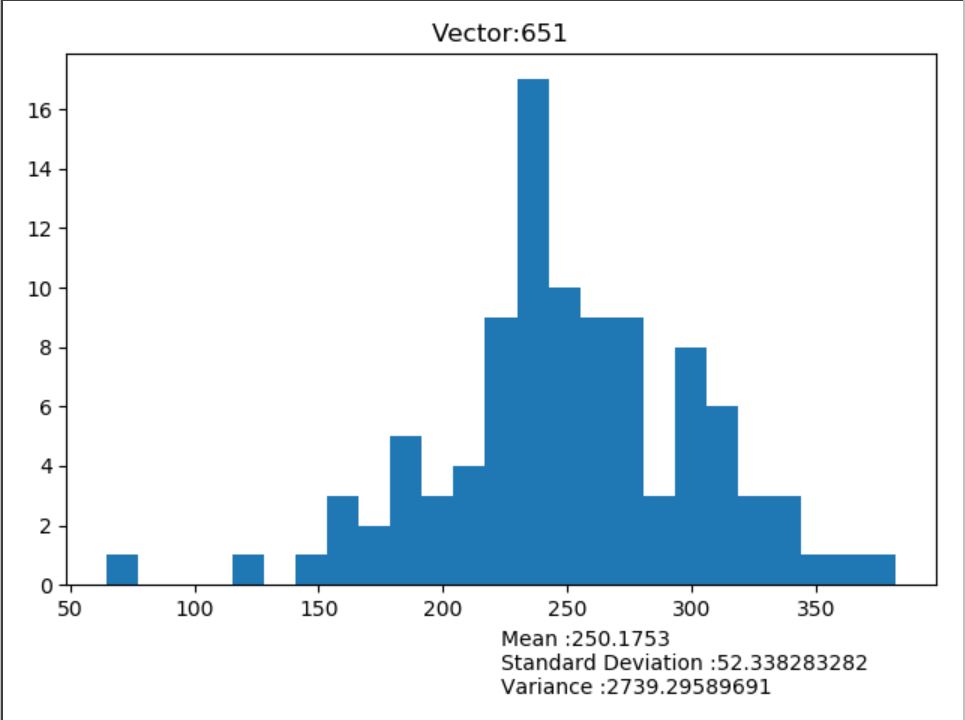
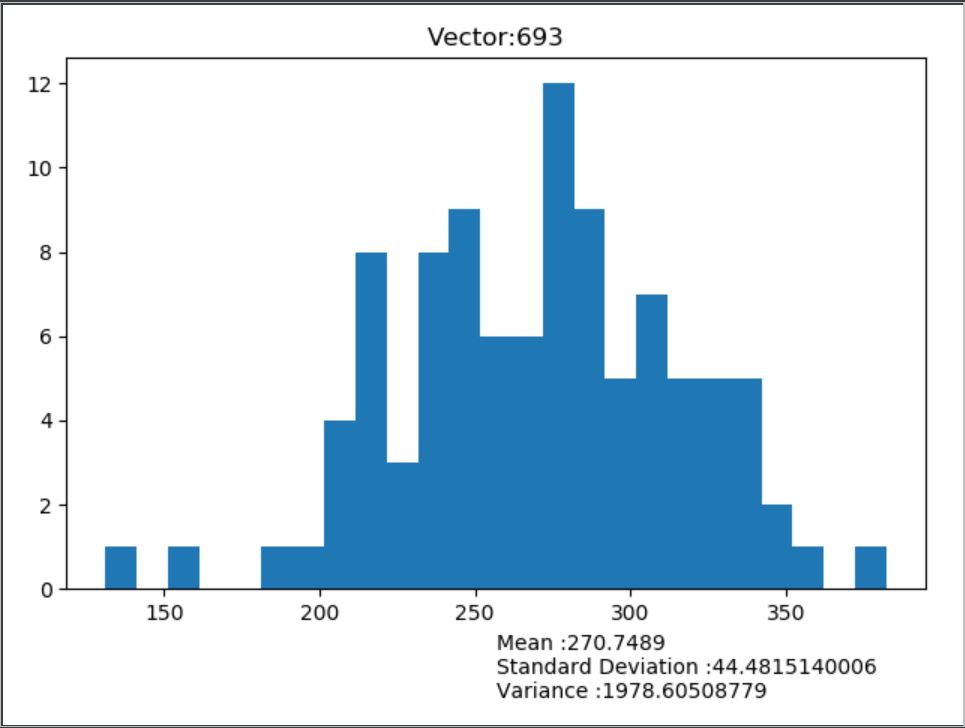
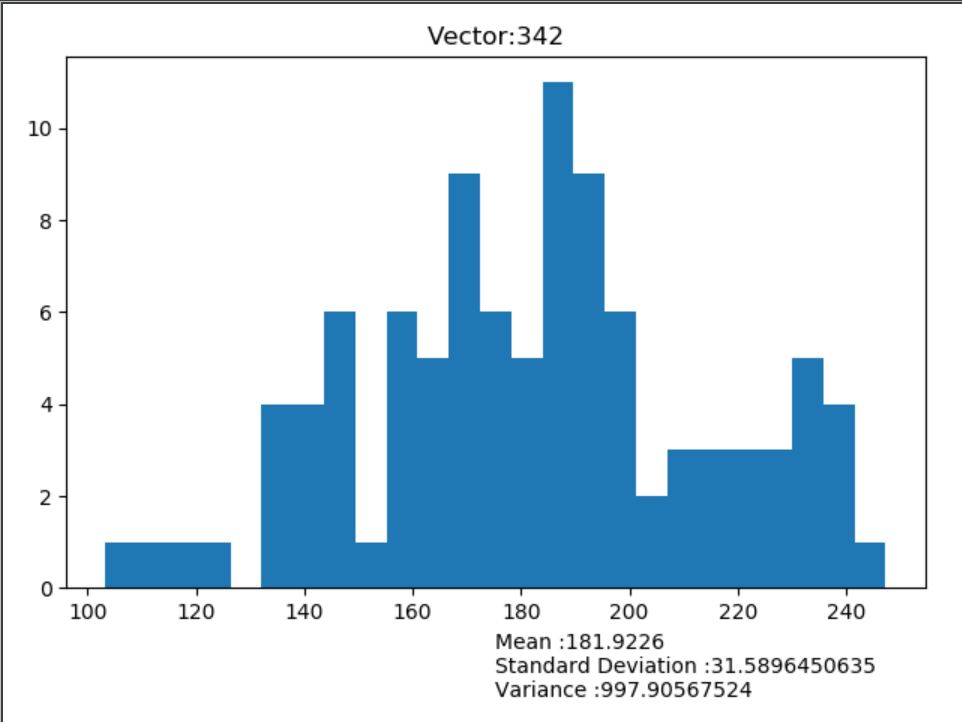
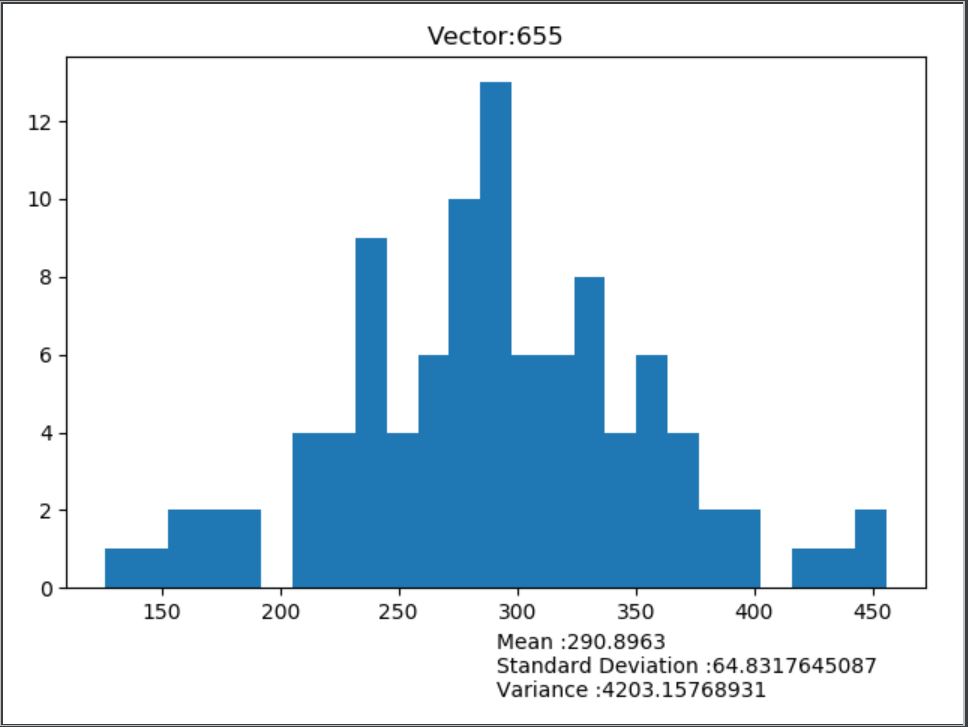
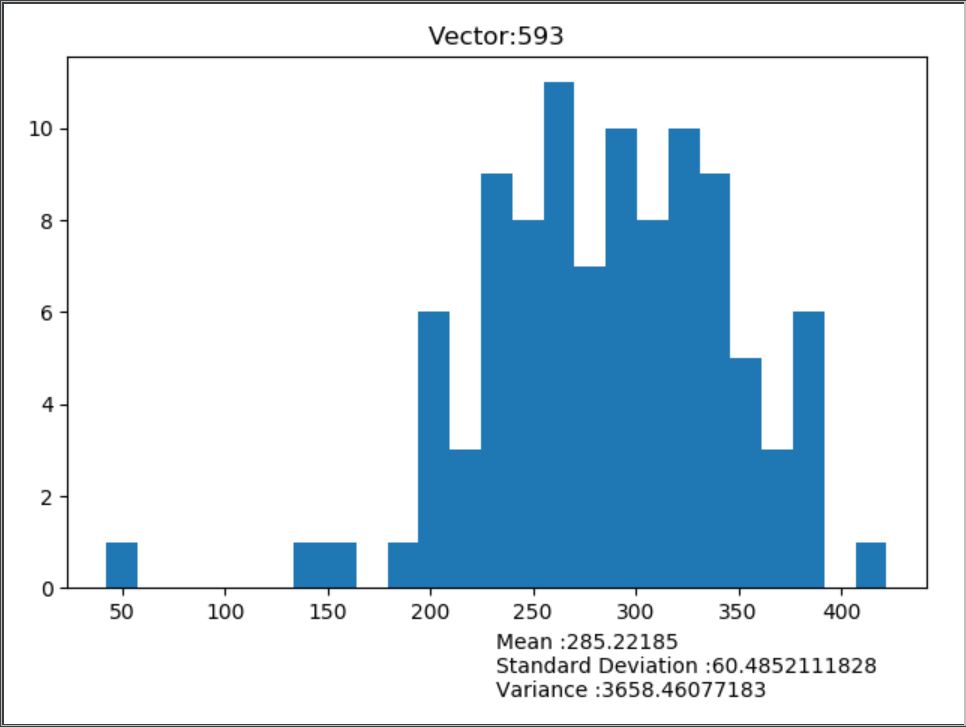
Distribution for the total dataset is plotted using a histogram as shown below,



Similarly, a random of 10 vector components were taken using a random generator and plotted on a histogram. The distribution parameters Mean, standard deviation and variance were calculated respectively.

Note:

* The vector component values shown in the graph title correspond to the row number in the dataset provided.
* The histograms were plotted using the bin size as 25

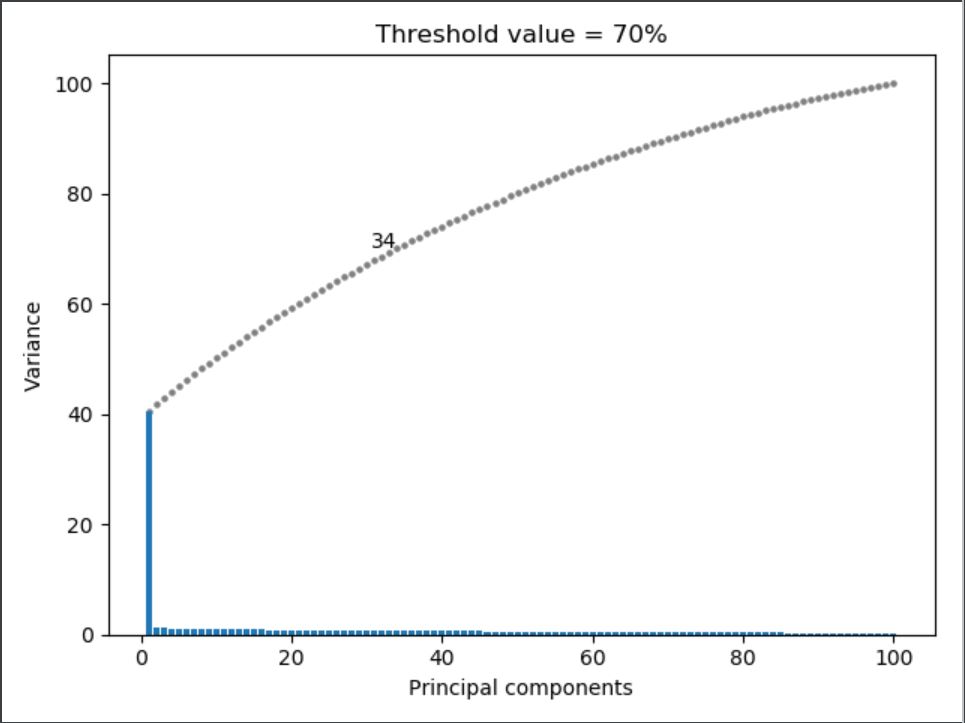


2.

a) The PCA procedure was applied on both datasets in order perform dimensional reduction i.e. reduce the original dataset with 100 dimensions to fewer dimensional data by retaining as much as information as possible. Considering the eigen value and vector pairs (PC’s), the variance was calculated for each principal component.

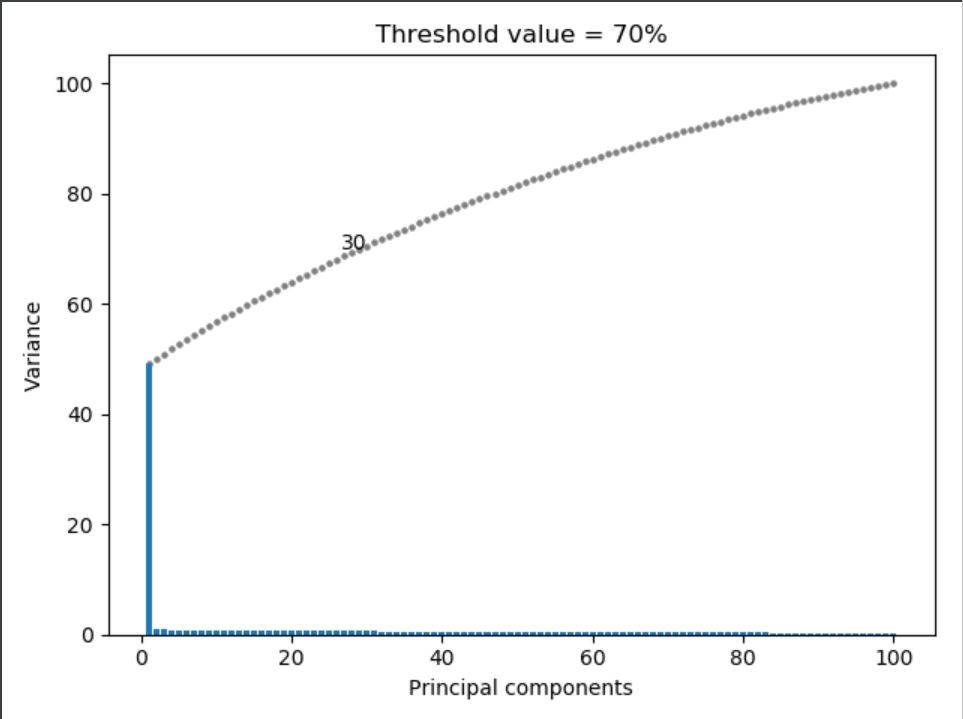
Here, the feature reduction was done on the basis of variance threshold. After observation of the PCs using a box plot to capture variance percent of each PC and a scree graph to represent the cumulative sum of the variance percentage (to note the amount of percent increase), the threshold value for the feature reduction was taken as 70%.

Observation of Dataset 1:



As seen in the above diagram, to retain 70% of the data 34 principle components can be considered.

Observation of Dataset 2:

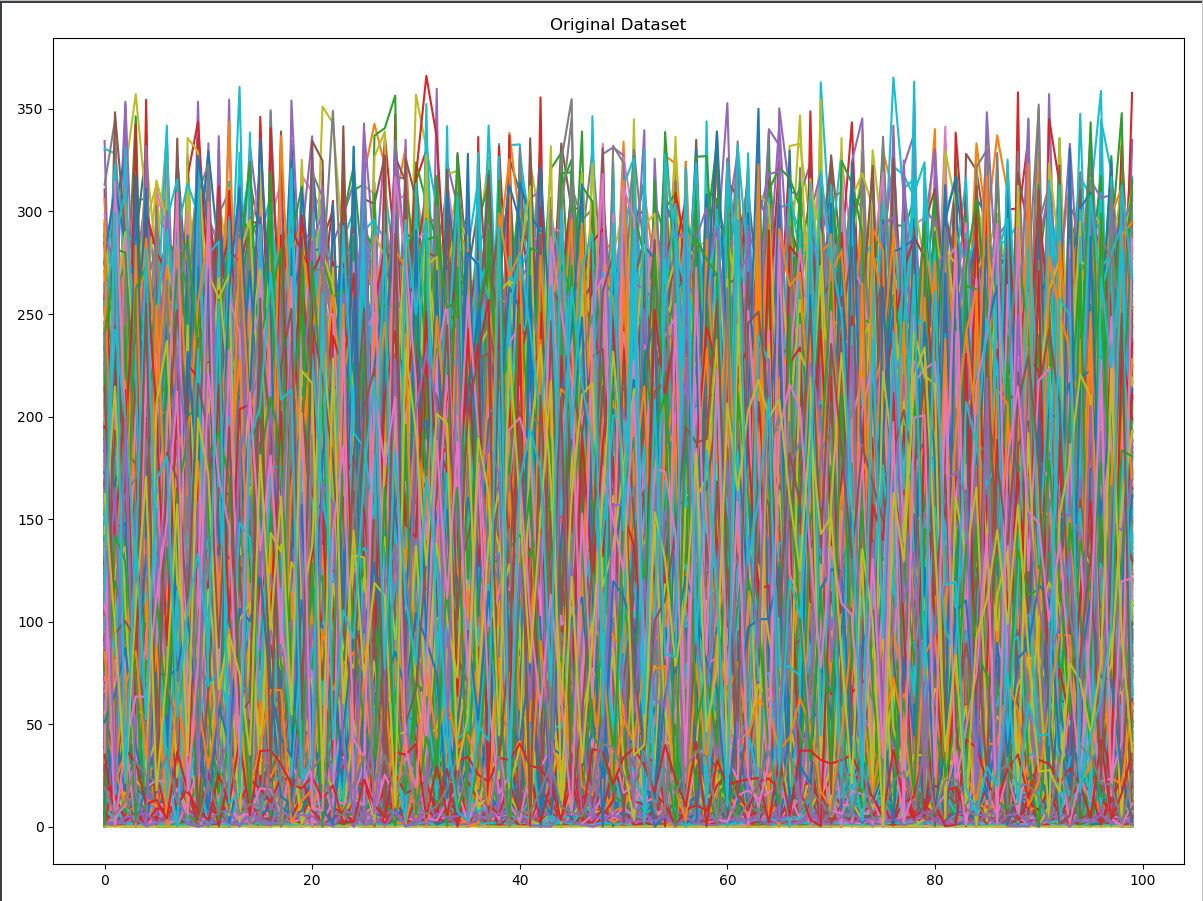


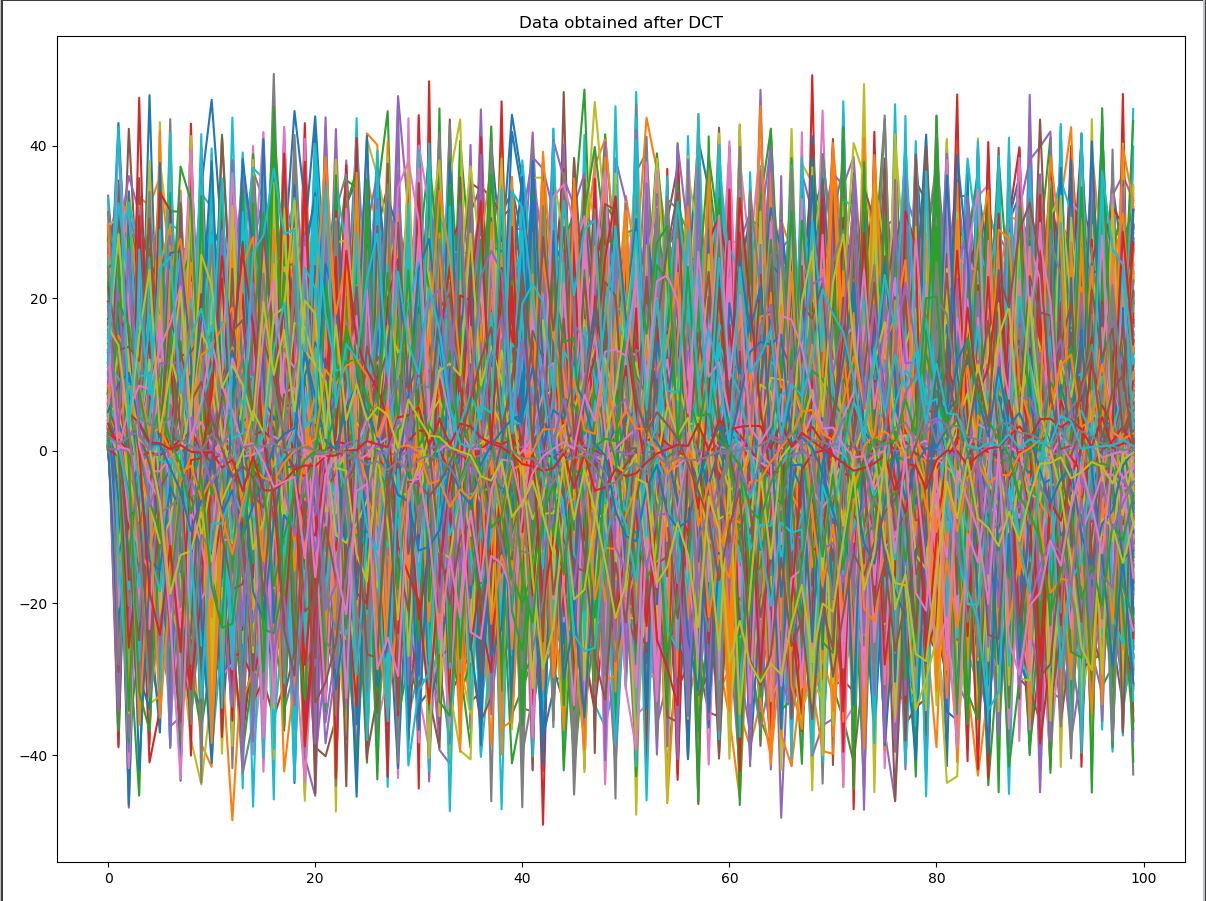
The above diagram shows that 30 principle components can be considered to retain 70% of the data. The remaining principle components gives 30% of information and hence can be discarded in this scenario.

b) DCT algorithm

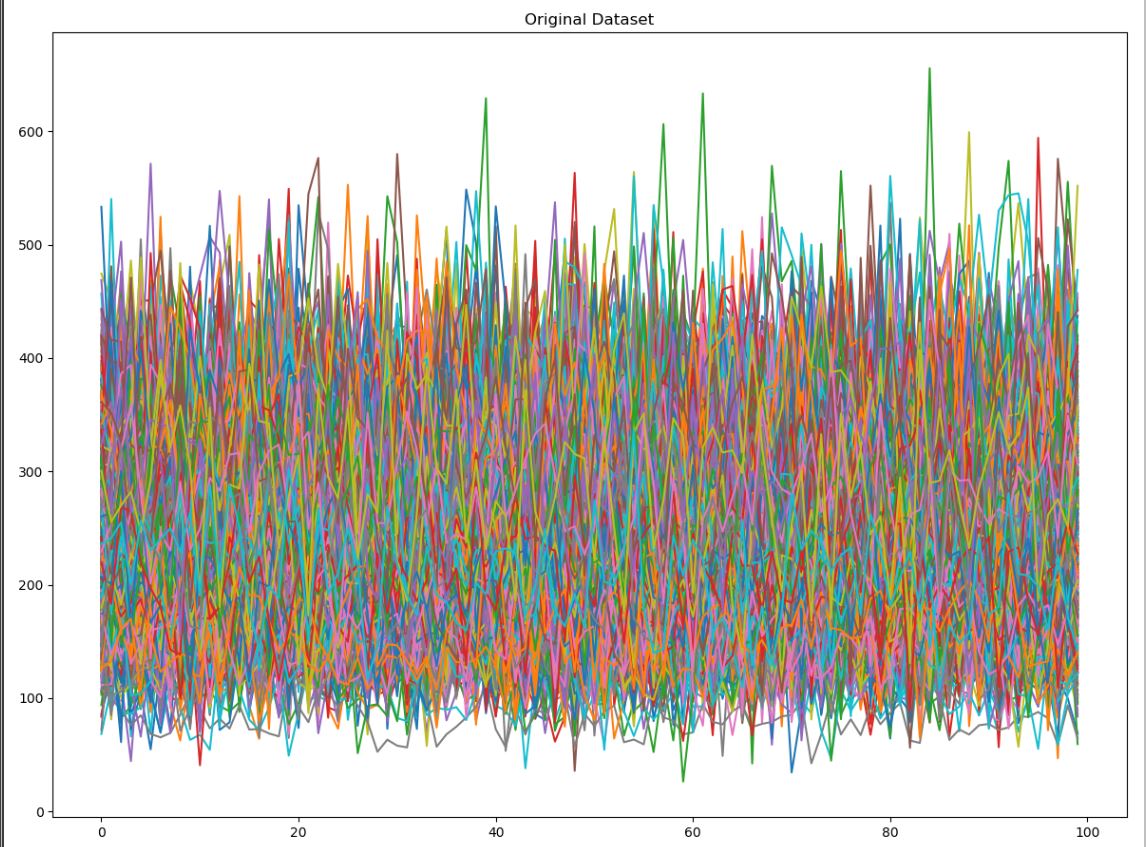
Dataset1:

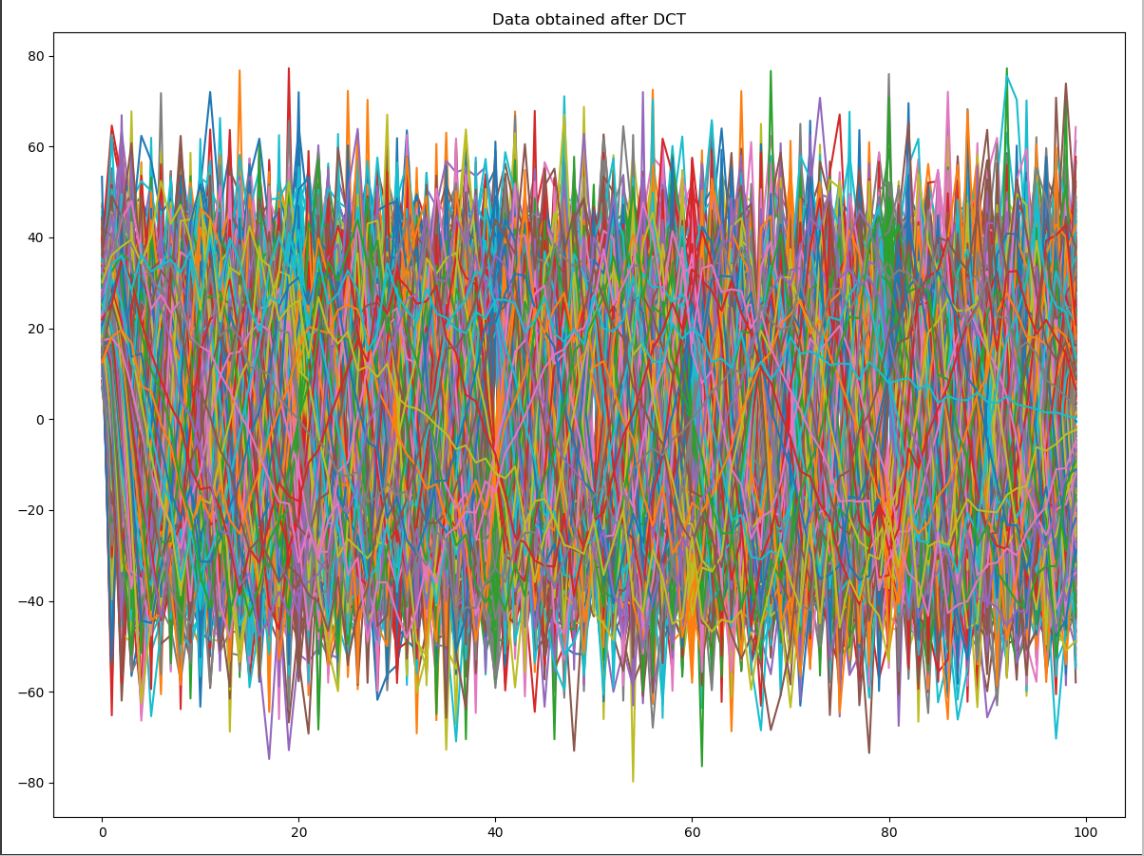
The dataset was compressed using the DCT equation. The result obtained is as shown below,





Dataset 2:

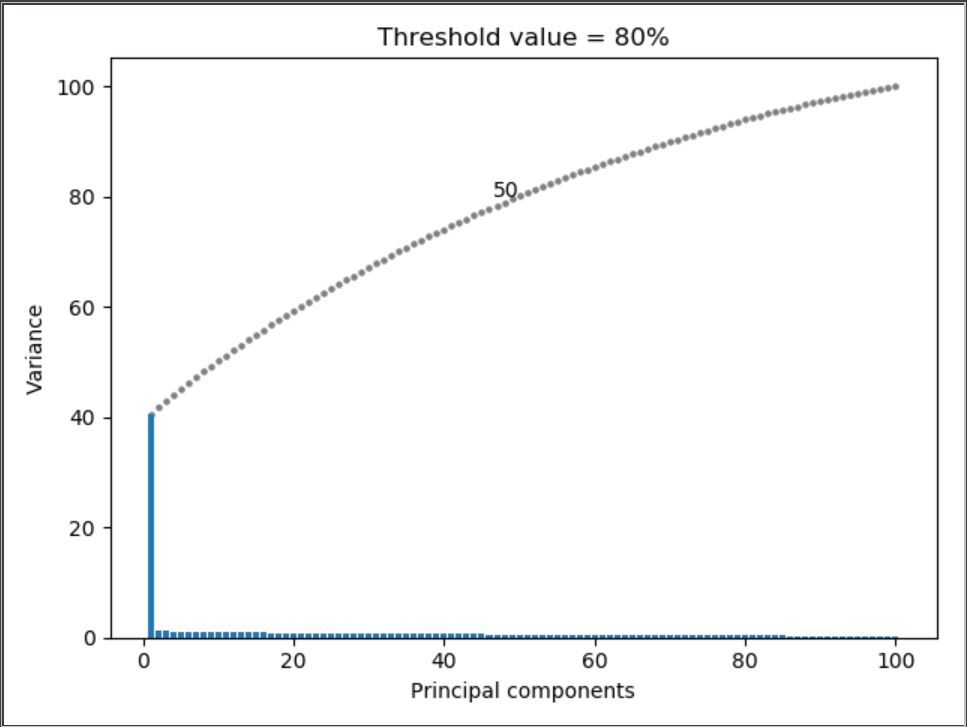




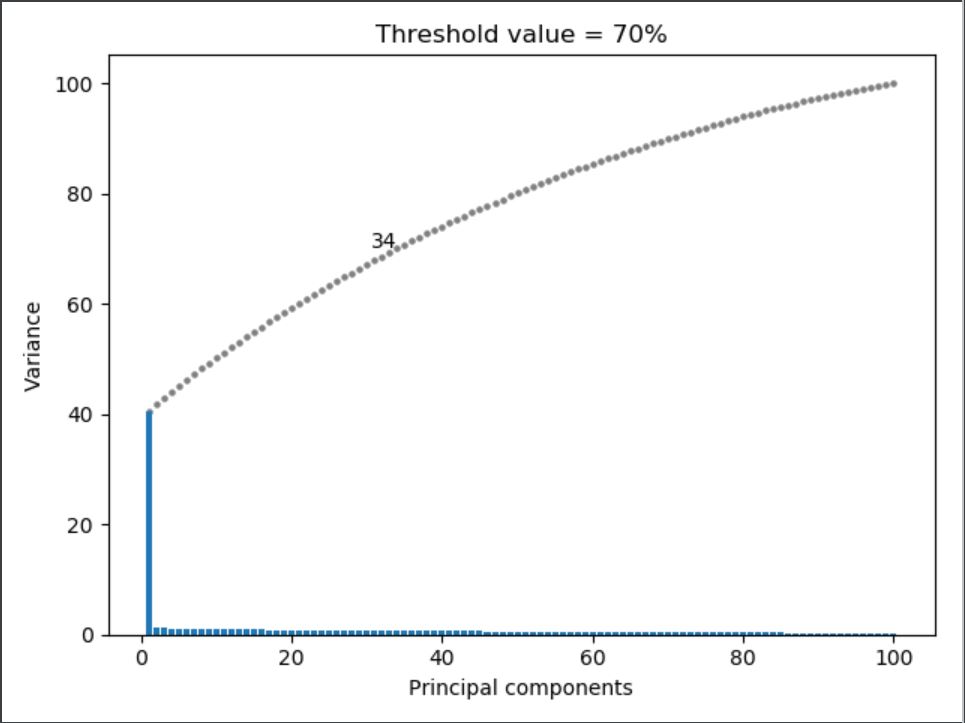
3. a) Dataset 1

Feature extraction using PCA:

Initially, the variance threshold was set to 80%. Below diagram shows the variance-Principal component graph respectively,



50 Principle components are required to retain 80% of information. Now, the threshold was set to 70% and the result obtained is shown below



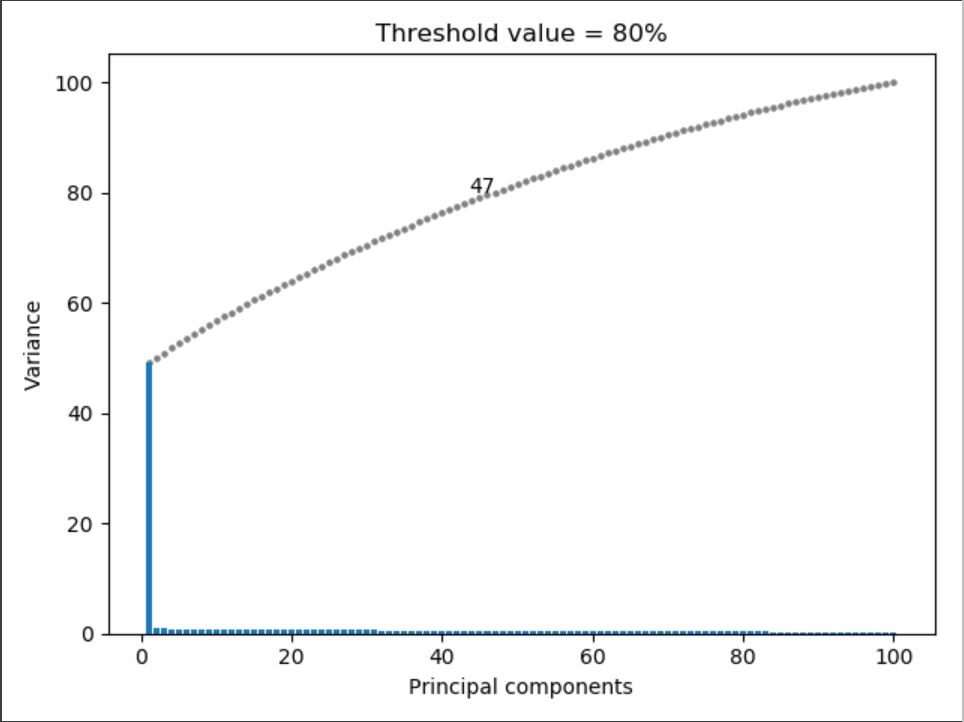
It can be observed that to obtain additional 10% of data (80% from 70%) 16 Principal components are needed to be considered which is not efficient as the aim is to retain maximum information as possible and simultaneously reduce the features to as minimum as possible.

Feature extraction using DCT:

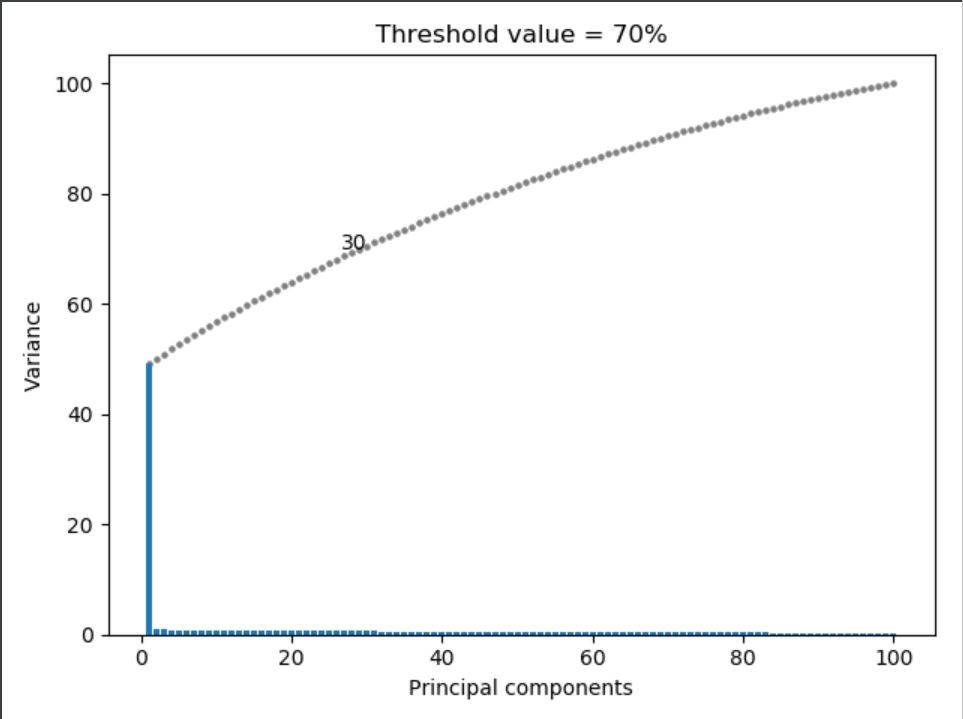
b) Dataset 2:

Feature extraction using PCA:

Similar to the dataset 1, the variance threshold was initially set to 80%. Below diagram shows the Principal component- variance graph respectively,



To retain 80% of data 47 PC’s have to be considered. In comparison the threshold was set to 70% and the result observed is as below



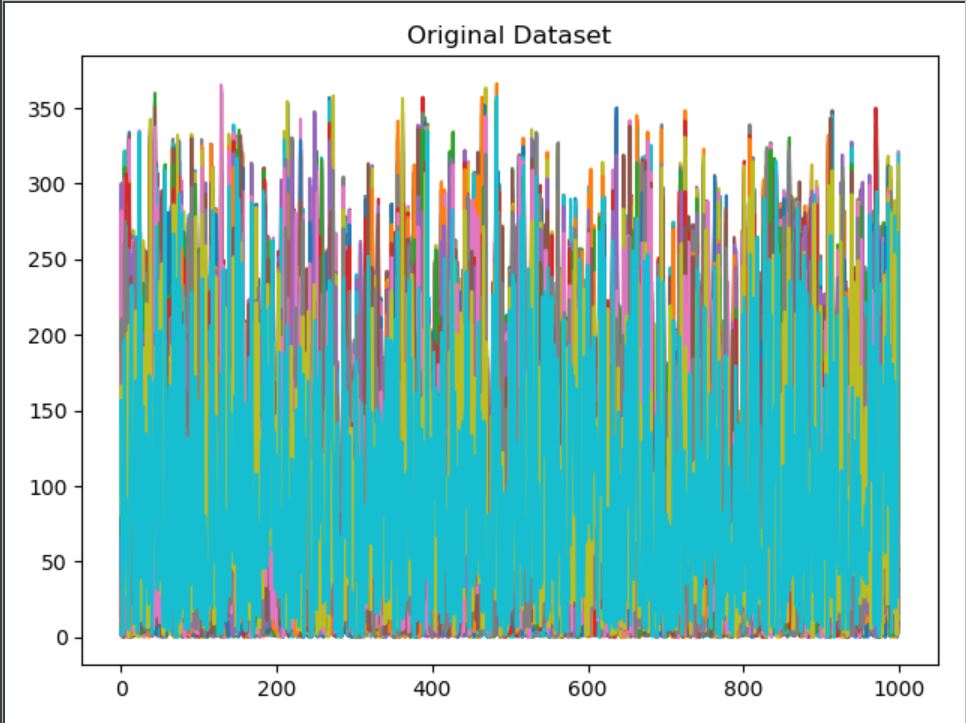
30 PC’s can be taken into account to obtain 70% of data which is more efficient result when compared to the 47 PC’s to obtain 80% of data. Hence, it is ideal to take the variance threshold as 70% for both the dataset.

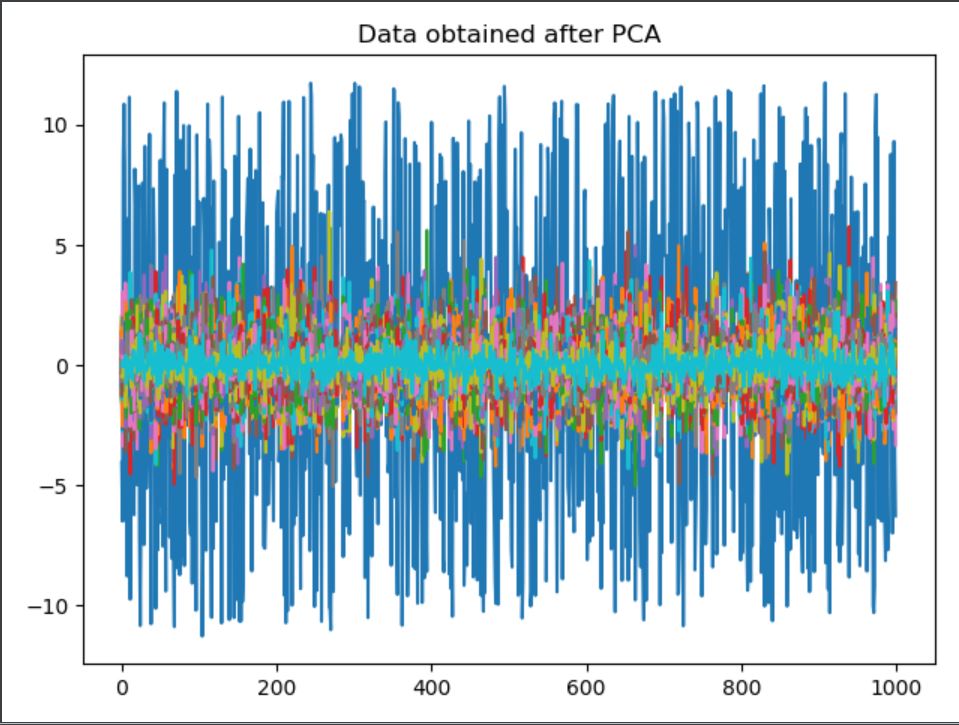
Feature extraction using DCT:

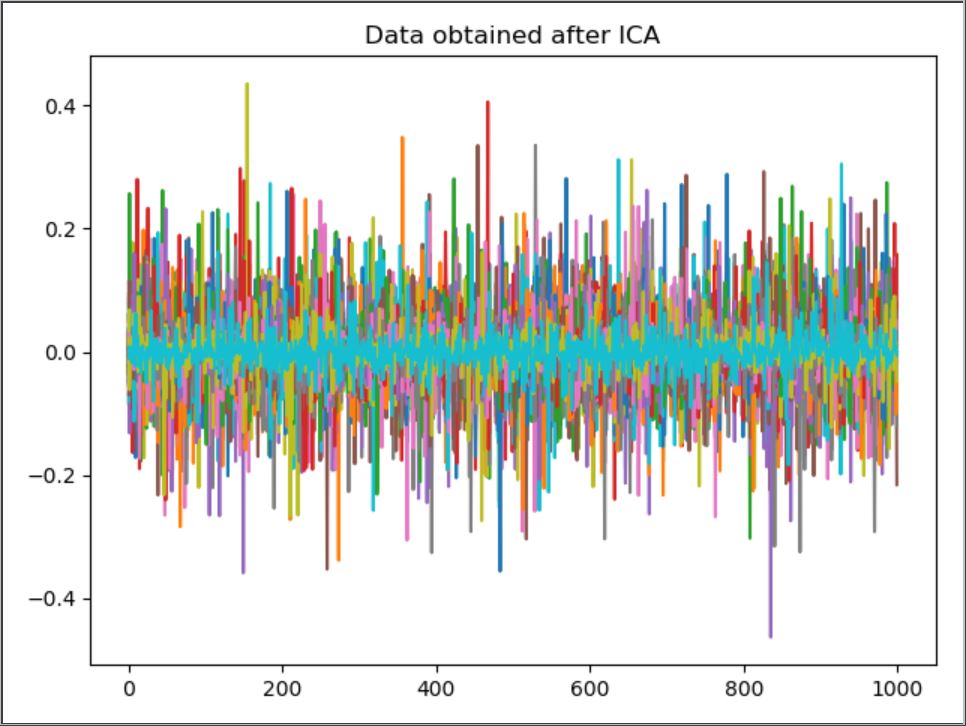
The dataset was compressed using the dct algorithm. The non-zero values obtained (mostly towards the top left of the matrix) represent the lesser frequency values which need to be retained to reconstruct the data and gradually zero values are observed in the resultant matrix i.e. higher frequency values which can be discarded.

4. Dataset 1:

ICA and PCA were applied on the dataset and the results were plotted as shown below,







PCA result:

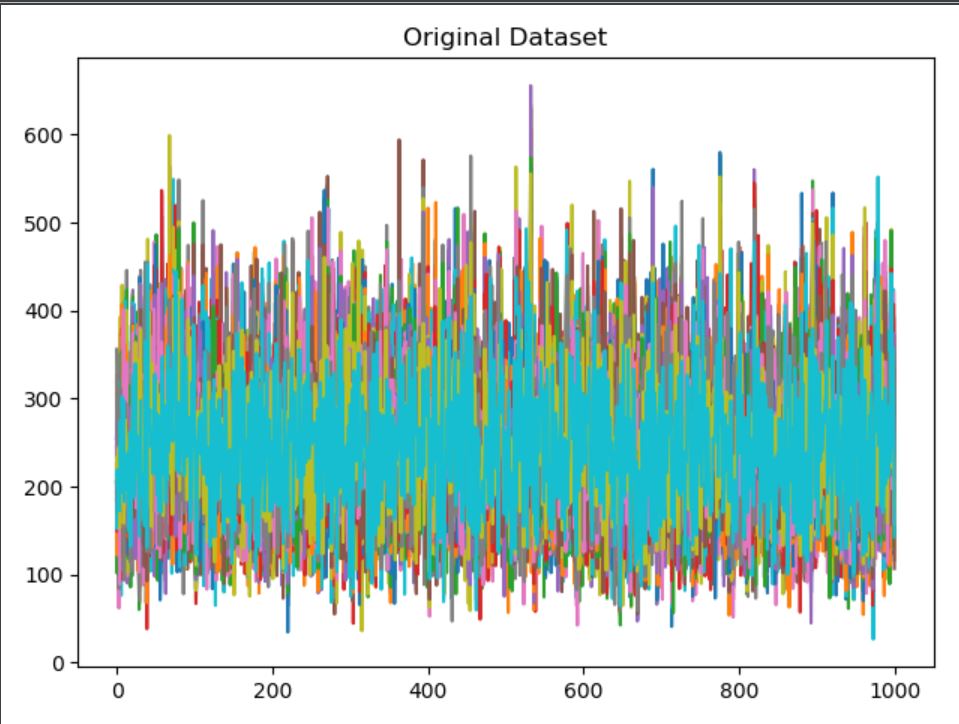
PCA results in obtaining the principal components (PC’s) i.e. the orthogonal vectors which show maximum variance. Hence, the first PC is larger and orthogonal the second and this pattern continues up to all the PC’s. It can be observed that with each PC plotted the portion in the graph reduces because of the decrease in the variance with each PC.

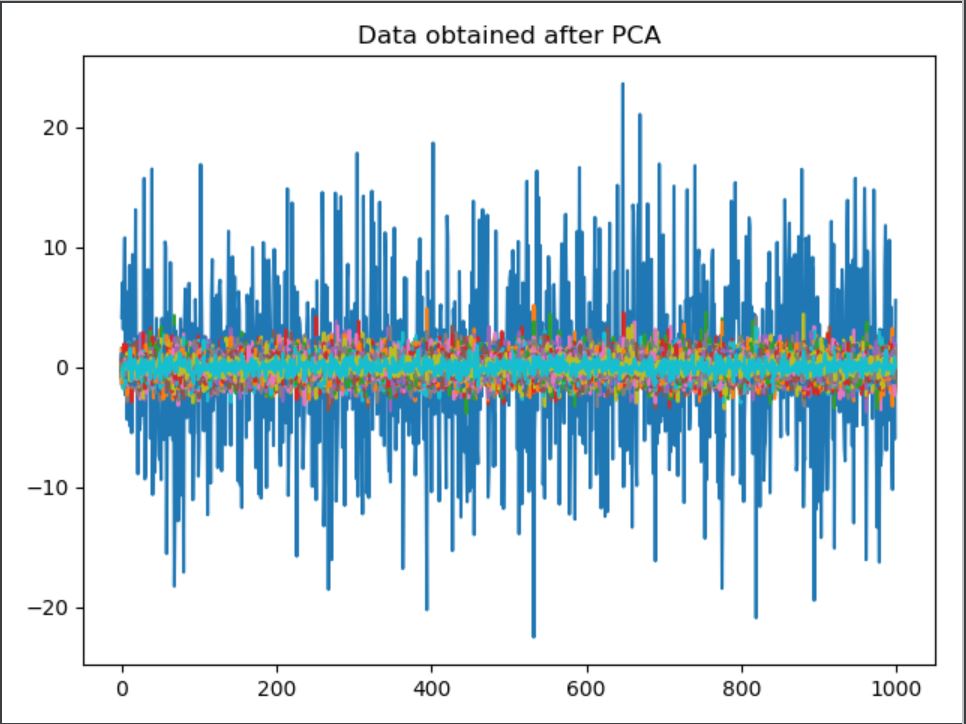
ICA result:

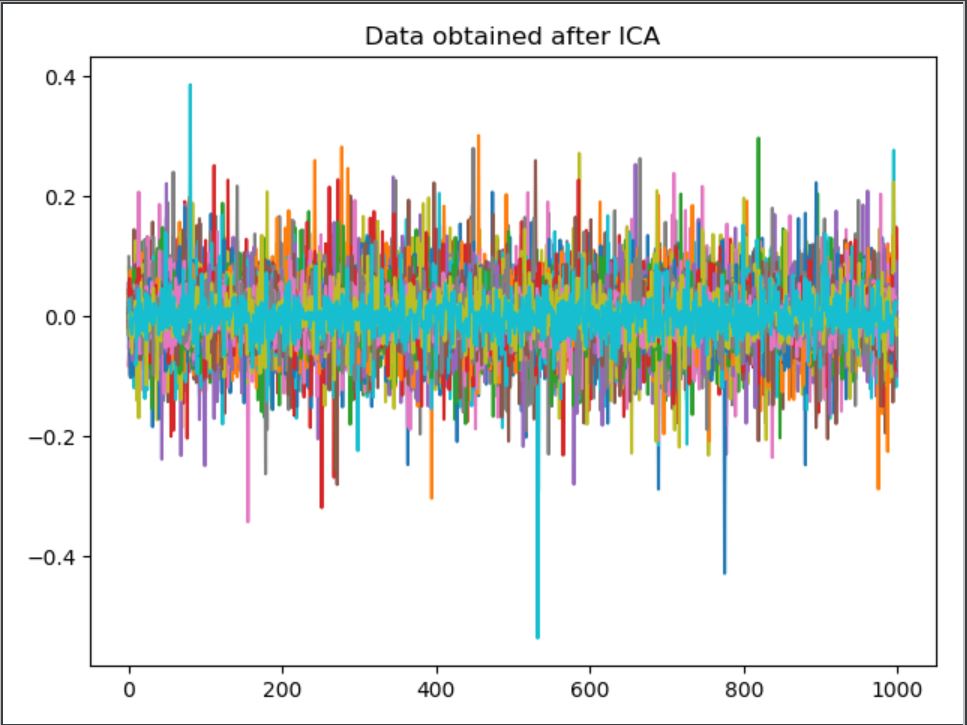
ICA gives the linear transformations which are maximally independent to each other. Therefore, each of every signal can be obtained. As seen in the result, approximately all the signals are equally recovered from the original dataset. Unlike PCA, the resultant vectors in ICA need not be orthogonal to each other.

Dataset 2:

Similarly, the results after PCA and ICA were plotted.







Conclusion:

* It can be concluded that PCA gives the principal components with maximum variance whereas ICA generates all the source signals.
* Unlike PCA whose result summarizes on the dataset as a whole, ICA generates the uncorrelated data thus obtaining the individual components or sources from the original data.