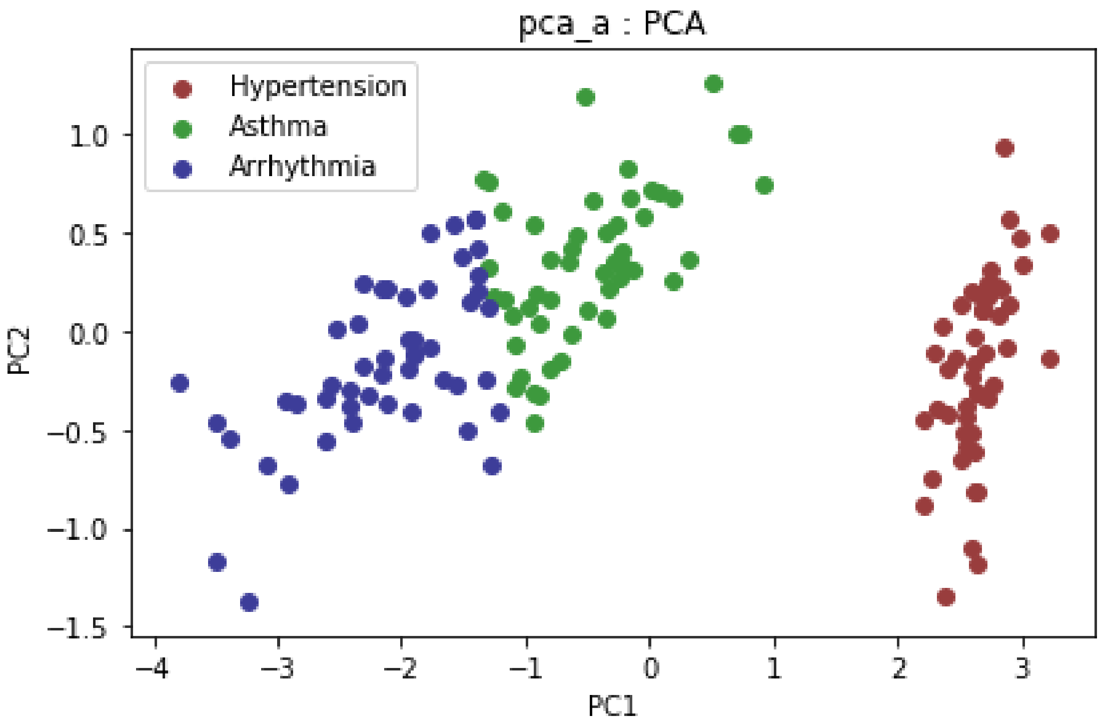
## Principal Component Analysis (PCA) :

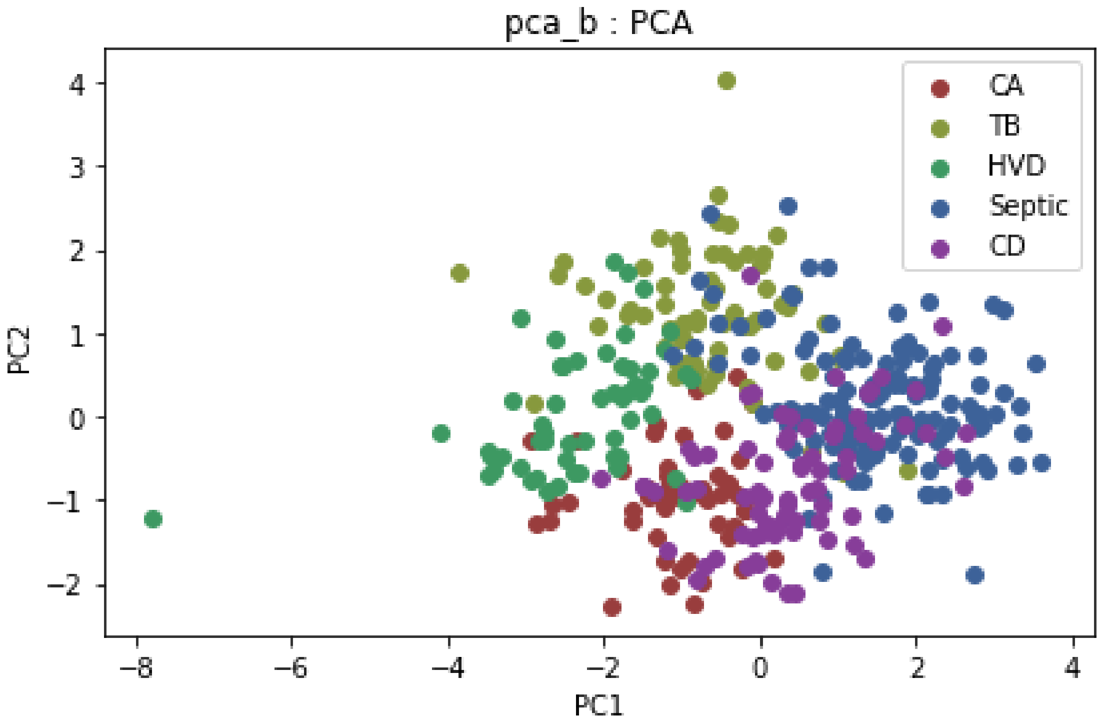
PCA is one of the dimensionality reduction techniques that converts the set of possibly correlated features to set of linearly uncorrelated variables. These new uncorrelated variables are called "Principal Components". The principal components are chosen in such a way that first principal component has maximum variance in it and second principal component has second maximum variance and so on.

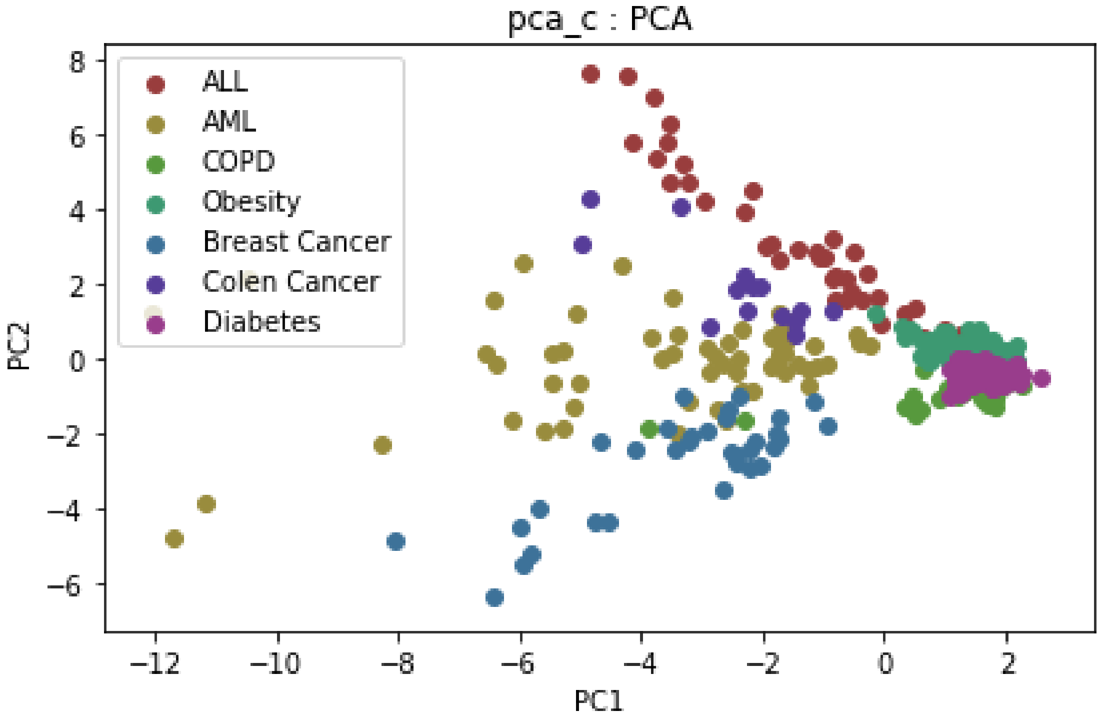
**STEPS TO IMPLEMENT PCA** :

1. Compute the mean of each column in the input data set
2. Compute the mean adjusted matrix
3. Find the Co-variance matrix of the mean adjusted matrix
4. Find the eigen vectors and eigen values of the covariance matrix
5. Arrange the eigen vectors in the descending order.
6. Select top K eigen values that cover the maximum variance and select their corresponding eigen vectors

We have three datasets pca_a.txt, pca_b.txt and pca_c.txt files. PCA algorithm is implemented on these three data sets and top 2 eigen vectors are selected. Plots are given below :

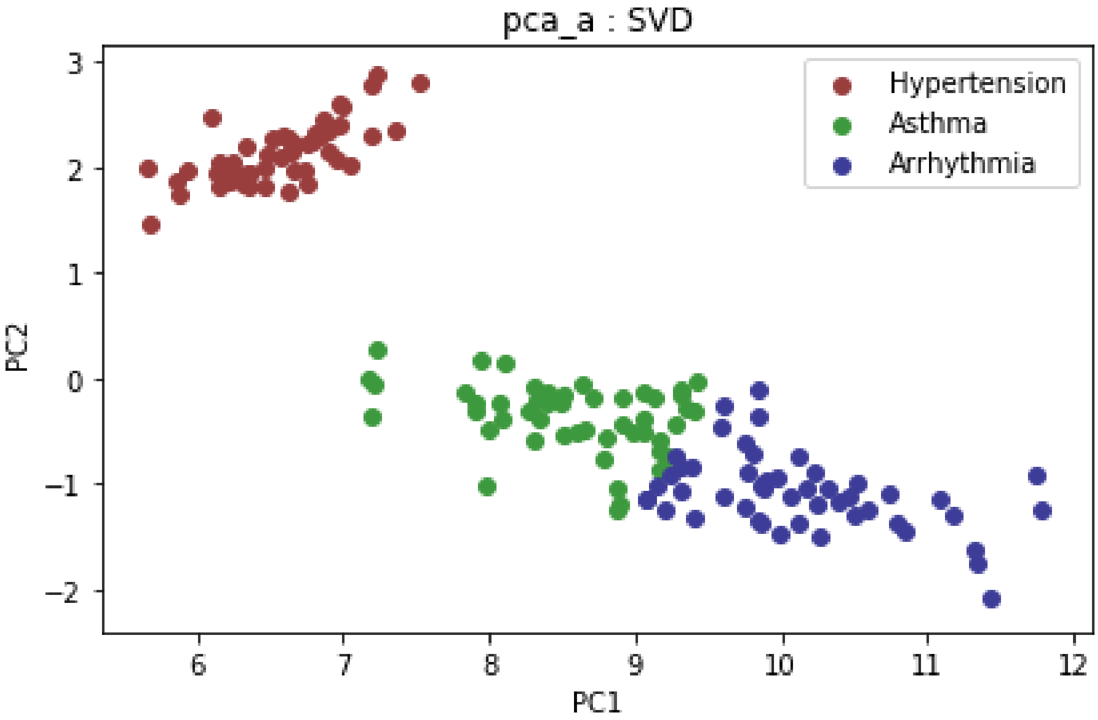


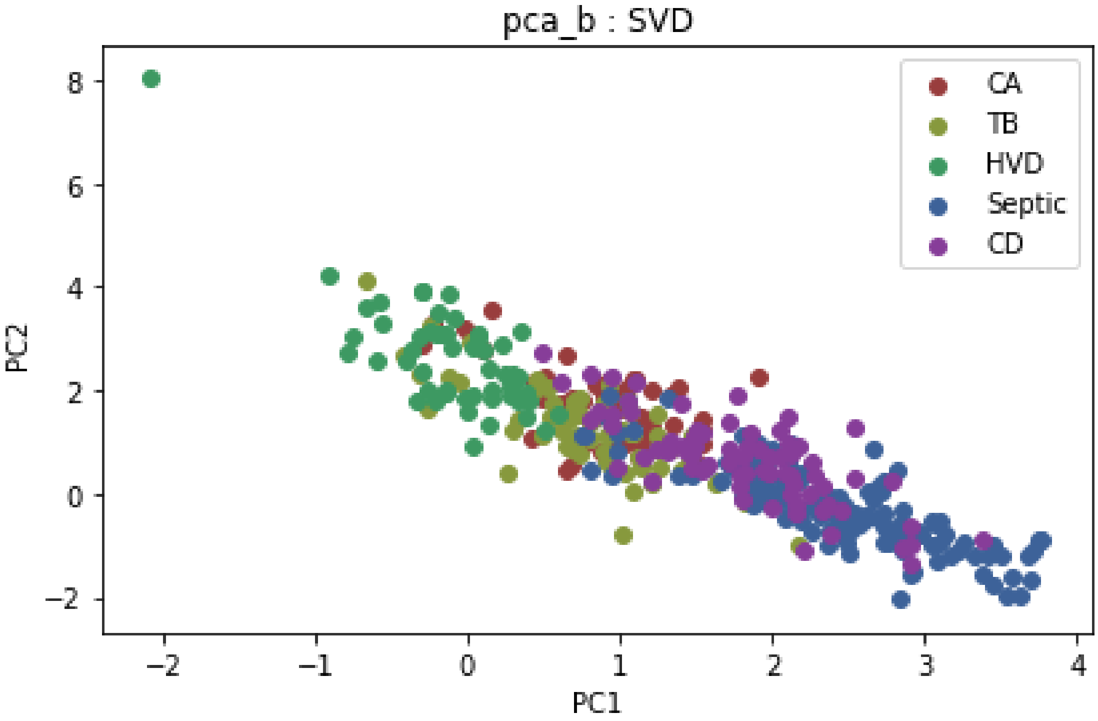

pca_b : PCA
CA
TB
HVD
Septic
CD
PC2
4
3
2
1
0
−1
−2
PC1
−8
−6
−4
−2
0
2
4

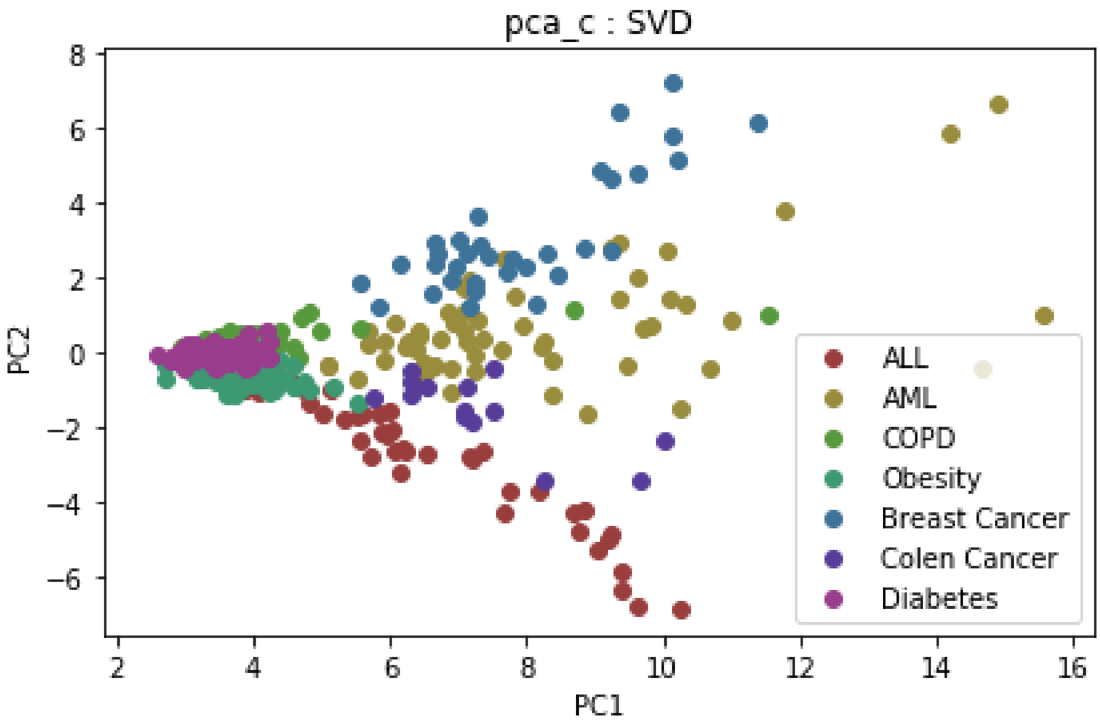

pca_c : PCA
ALL
AML
COPD
Obesity
Breast Cancer
Colen Cancer
Diabetes
PC2
8
6
4
2
0
−2
−4
−6
PC1
−12
−10
−8
−6
−4
−2
0
2

## Singular Value Decomposition :

SVD is another technique to implement dimensionality reduction. We do not use Mean adjusted matrix in SVD as we did in PCA. Plots for 3 different data sets when SVD is applied are given below :



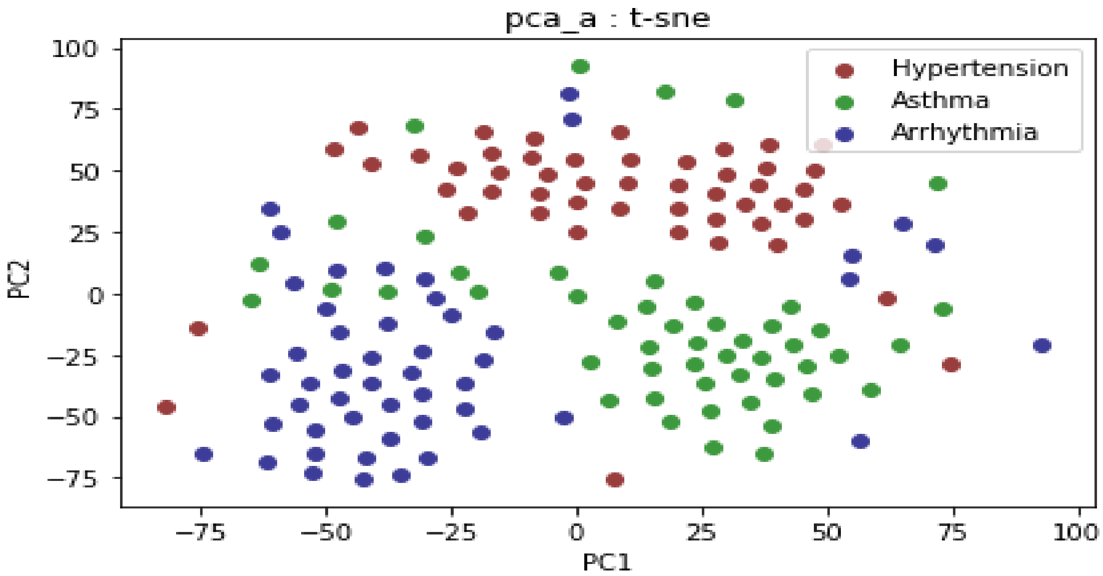


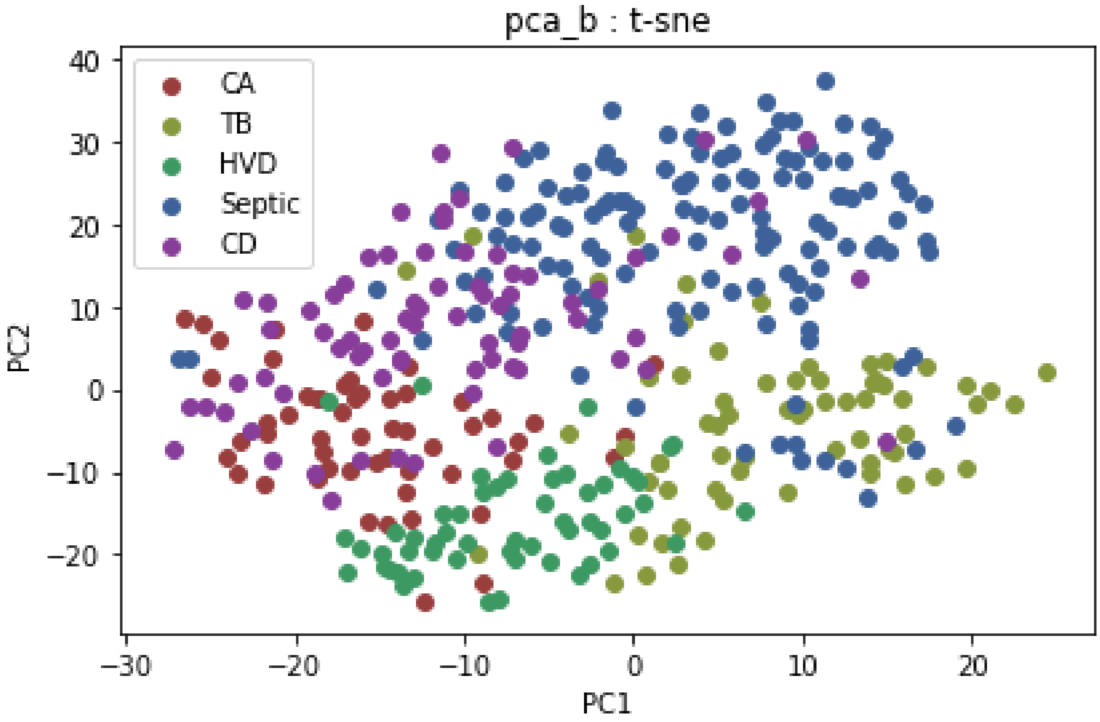


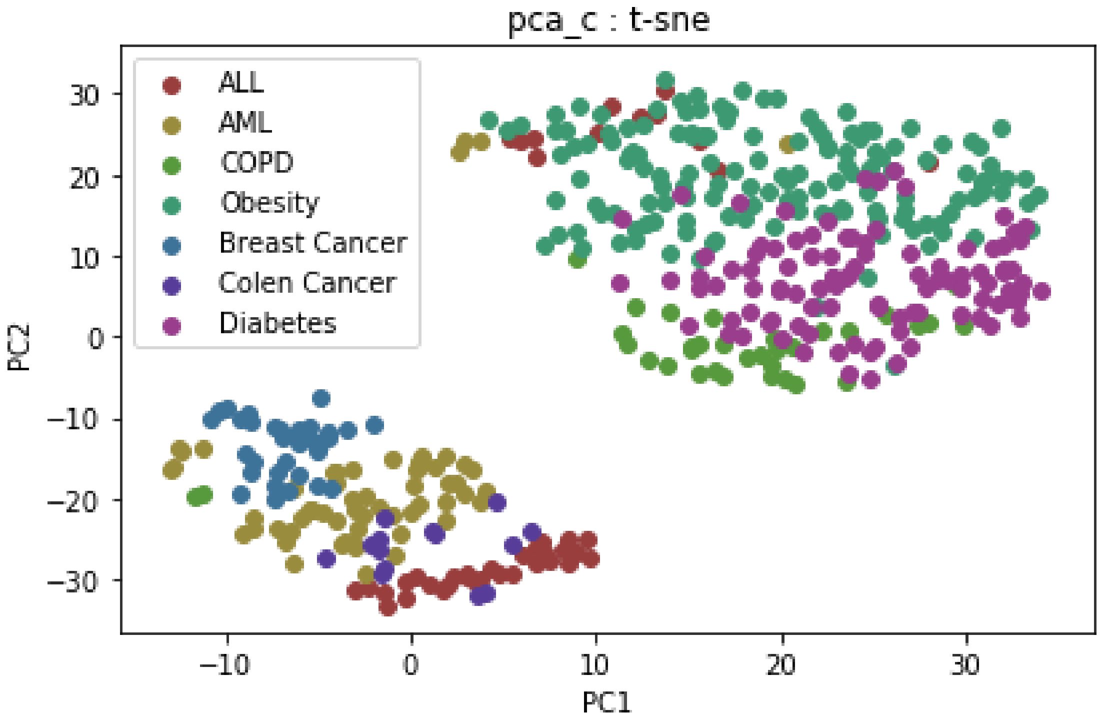
We used sci-learn function to compute the SVD**. if we use mean adjusted matrix as an input to the function, we get the results same as PCA**.

## T-Distributed Stochastic Neighbor Embedding (t-SNE)

T-SNE is another technique for dimensionality reduction that is well suited for high dimensional data sets. It is a non-linear dimensionality reduction technique. Here are the results obtained by applying T-SNE on 3 different data sets.




pca_b : t-sne
● CA
● TB
● HVD
● Septic
● CD
PC2
40
30
20
10
0
−10
−20
−30
−20
−10
0
10
20
PC1


pca_c : t-sne
● ALL
● AML
● COPD
● Obesity
● Breast Cancer
● Colen Cancer
● Diabetes
PC2
30
20
10
0
−10
−20
−30
−10
0
10
20
30
PC1

## Analysis of plots :

- T-SNE is a non-linear and probabilistic model. So the plot generated by each run of the algorithm will vary with each run.
- SVD suits better for the data sets which are sparse as there is no need to center the data around the mean.
- PCA and SVD for the data sets pca_a and pca_c are almost similar. It indicates that data is already centered around mean
- PCA and SVD for the data sets pca_a and pca_b are completely different. It indicates that data is highly distributed.