UNIVERSITY AT BUFFALO

CSE601 – Data Mining and Bioinformatics

Project 1- Part 1 Report (PCA)

Aayush Shah (50207564)
Aniruddh Chaturvedi (50206958)
Haril Satra (50208283)

9/30/2017

Implementation of Principal Component Analysis (PCA) for Dimensionality Reduction.

Steps in PCA Implementation:

- Read the input data from text file as a two-dimensional array.
- Calculate centred data from the input data by calculating the mean of each dimension and subtracting the data by the mean.
- Calculate covariance matrix (Cov) from the centred data (D) using the following formula:

  Cov = D * D_Transpose / Number of data points

  We used the following function available in the numpy package in python:

  numpy.cov(D,bias=true)
- Calculate Eigen vectors and Eigen Values form the covariance matrix using the relation:

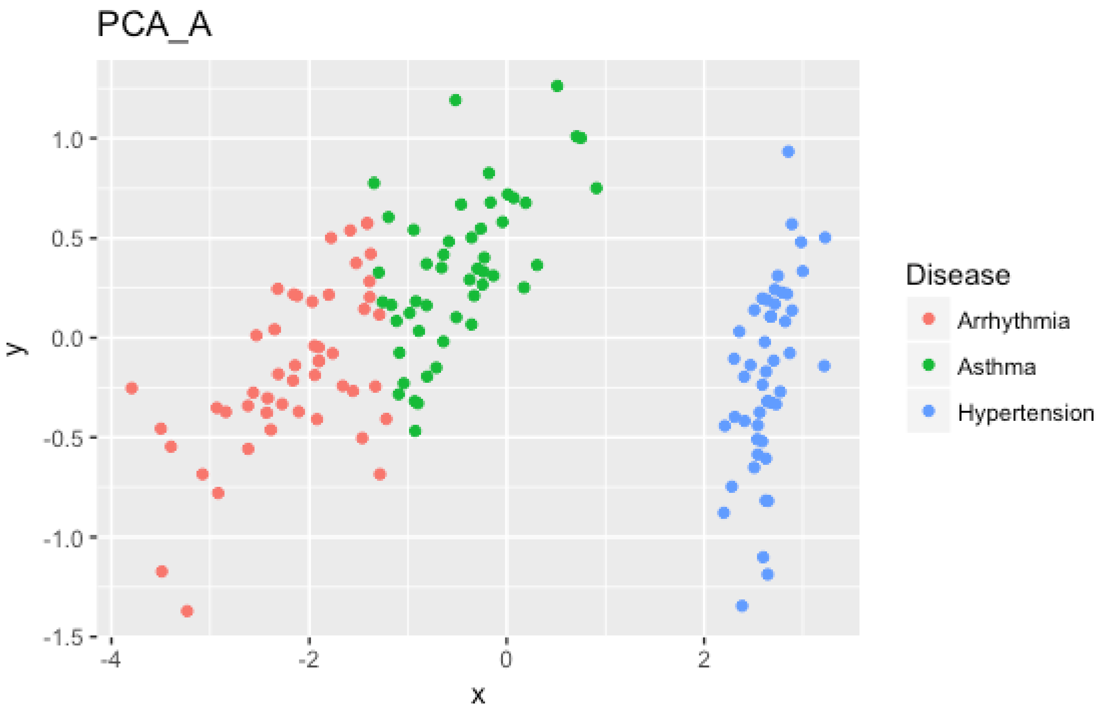
  [Covariance Matrix] * [Eigen Vector] = [Eigen Value] * [Eigen Vector]

  We used the following function available in the numpy.linalg package in python:

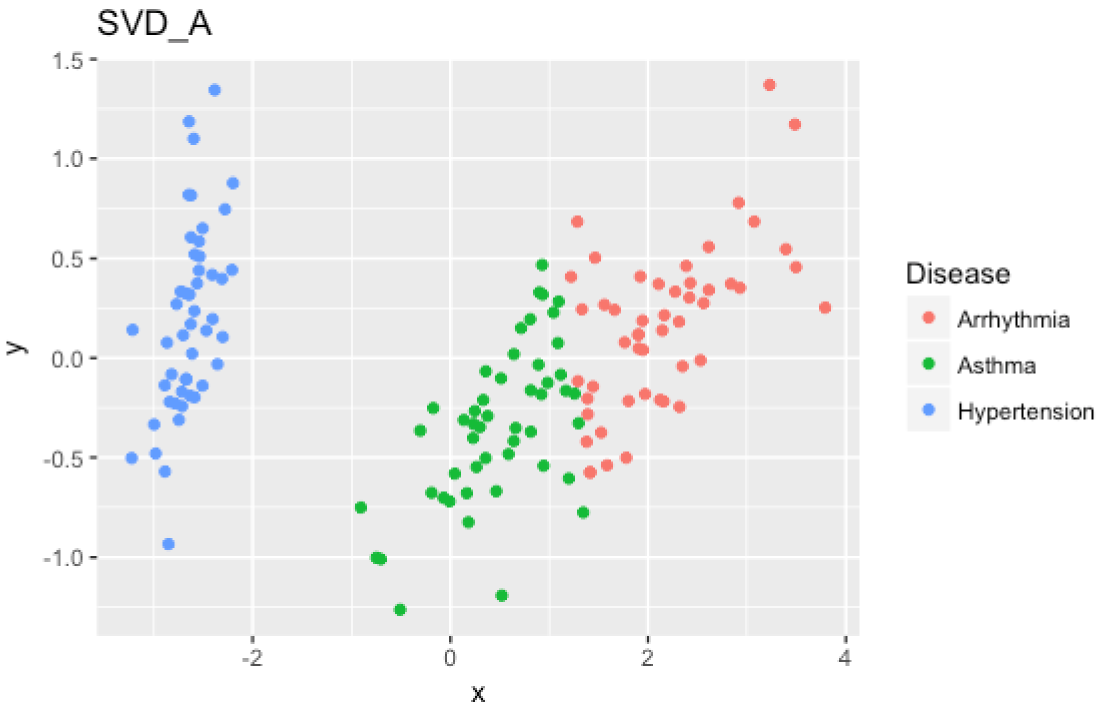
  np.linalg.eig(Cov)
- Sort the Eigen Vectors according to Eigen Values and select the top two Eigen Vectors, since we want to reduce the original data to two dimensions.
- Then we performed dot product of the Centred Data and Eigen Vectors to obtain new data points of the reduced dimension.

Plots obtained from running different Dimensionality reduction algorithms on three different datasets are as follows:

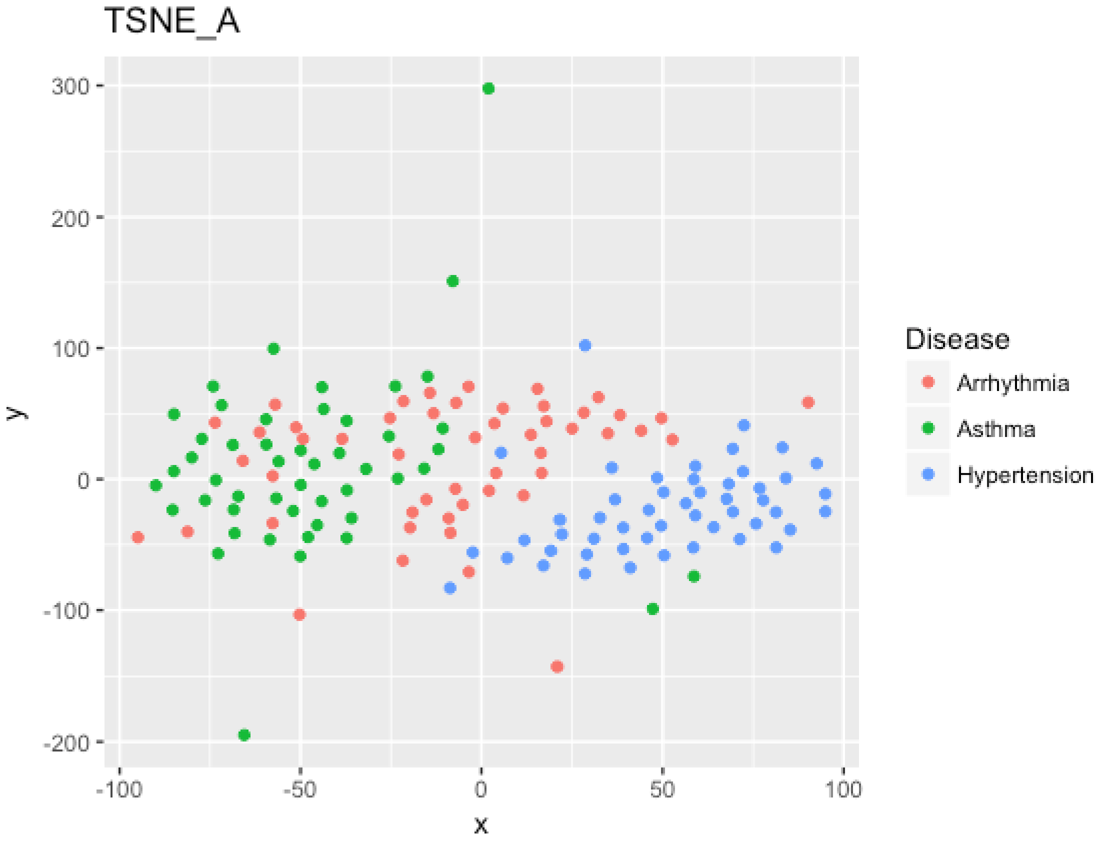
1) Plot obtained from running PCA on dataset A.



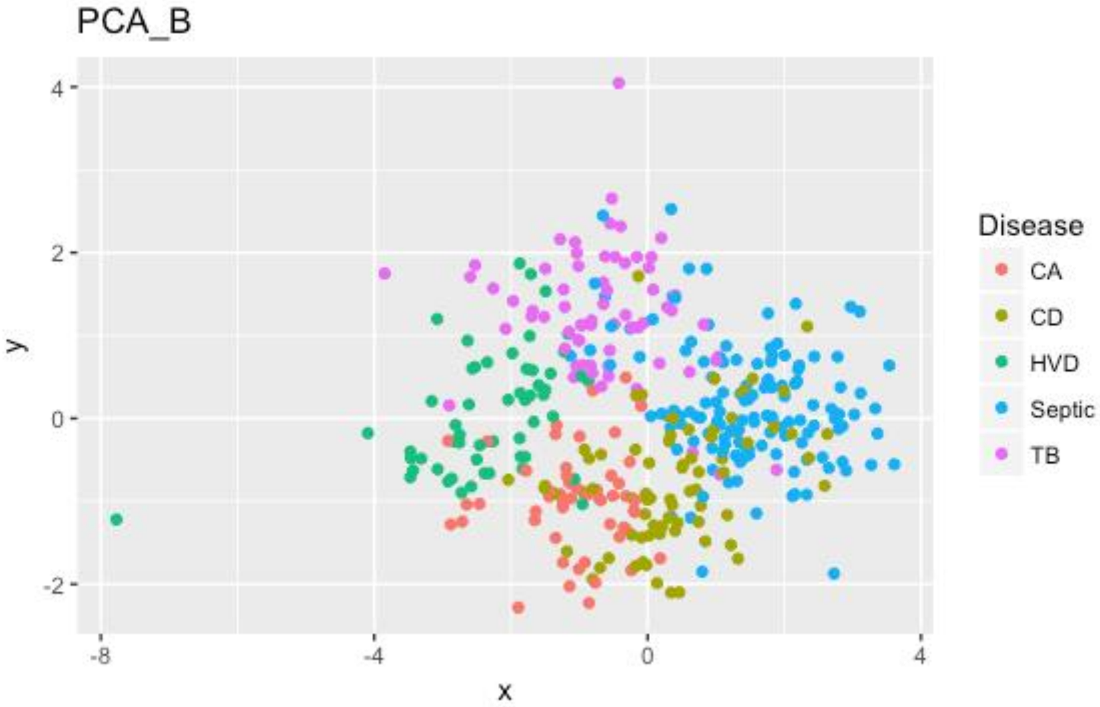
2) Plot obtained from running SVD on dataset A.



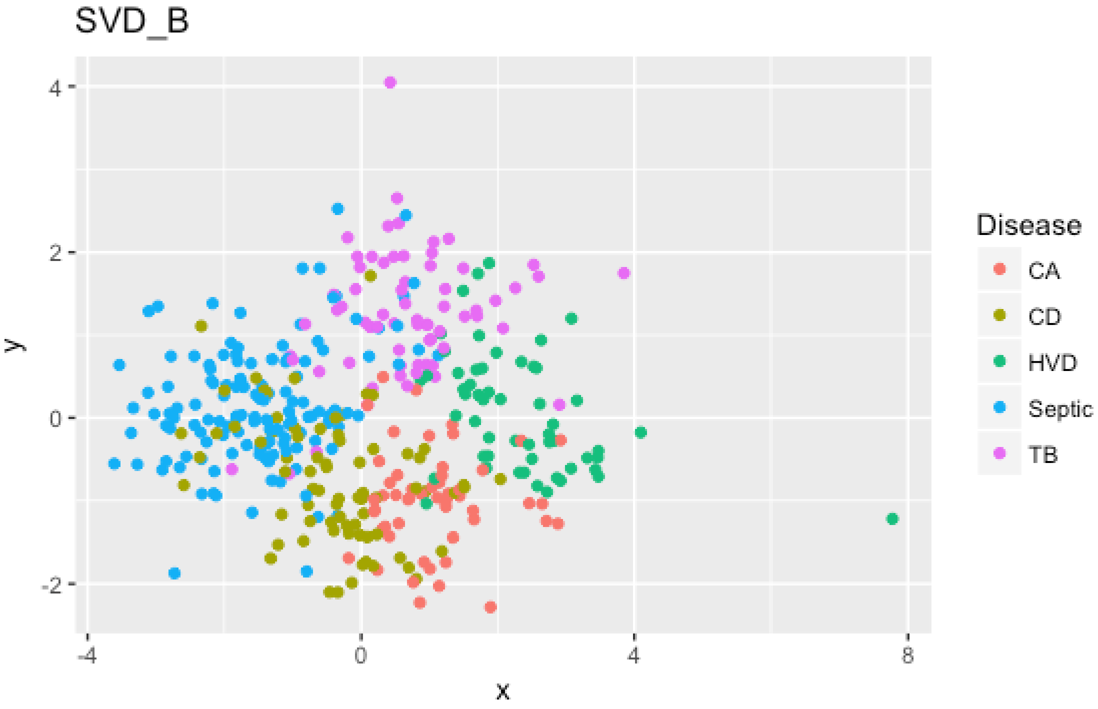
3) Plot obtained from running TSNE on dataset A.



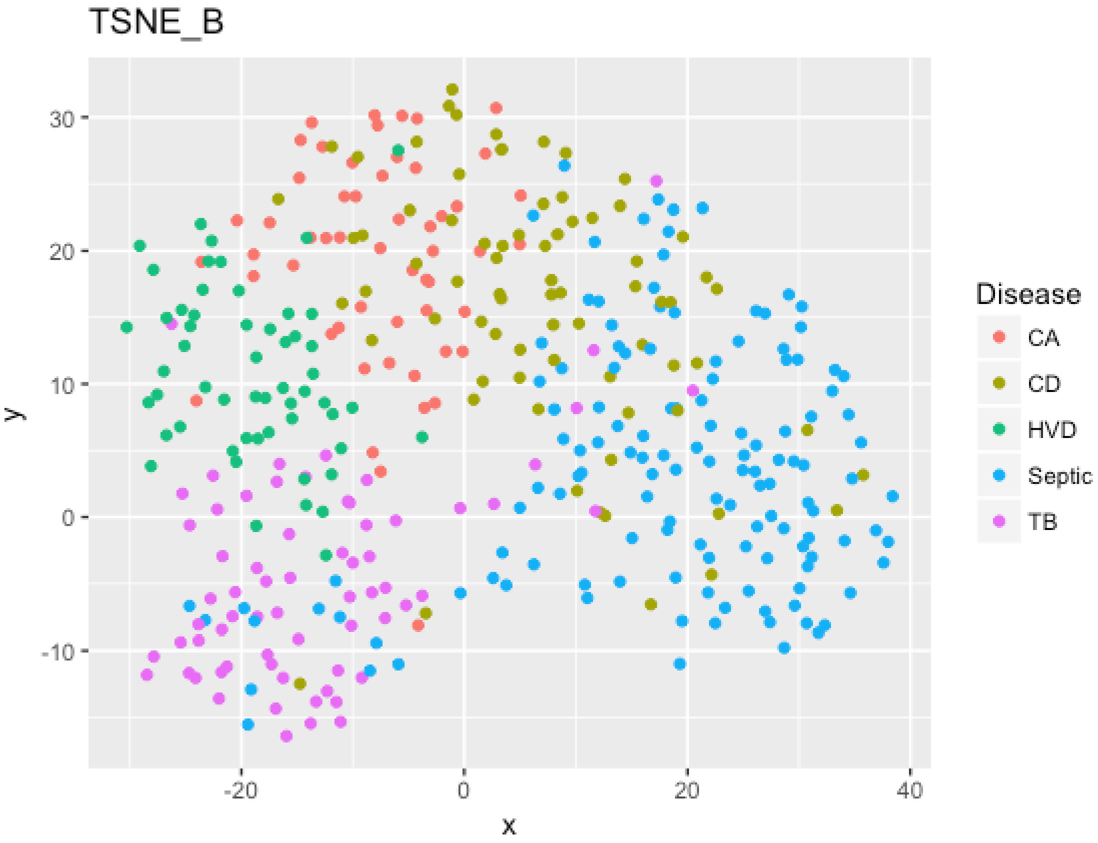
4) Plot obtained from running PCA on dataset B.



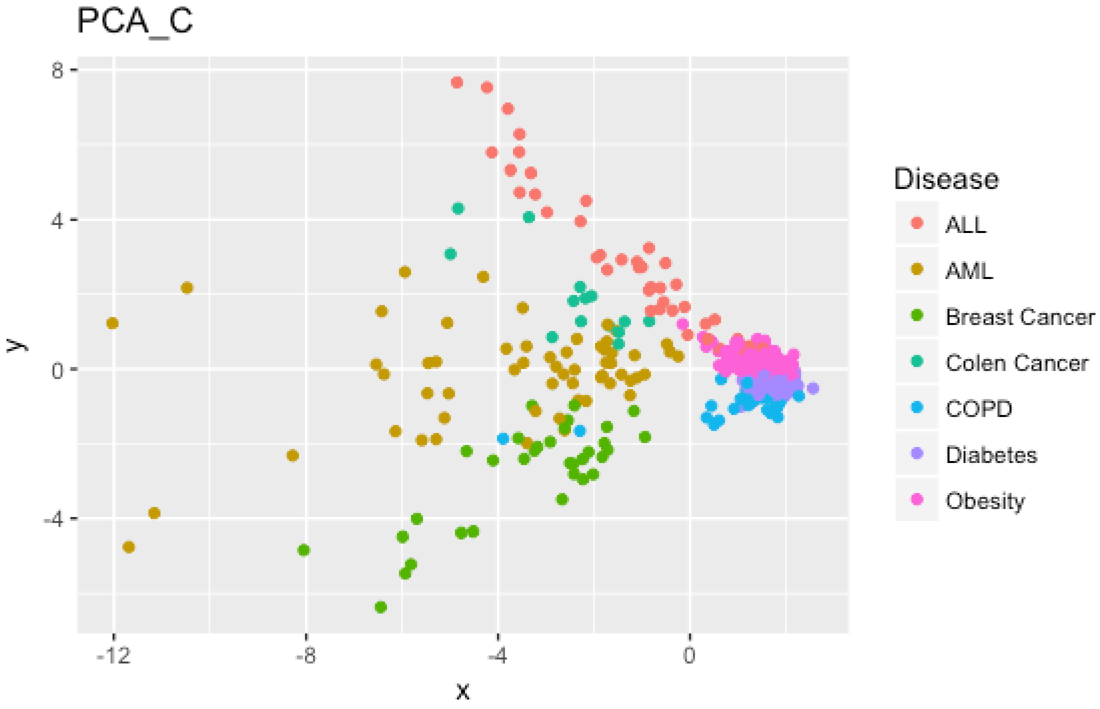
5) Plot obtained from running SVD on dataset B.



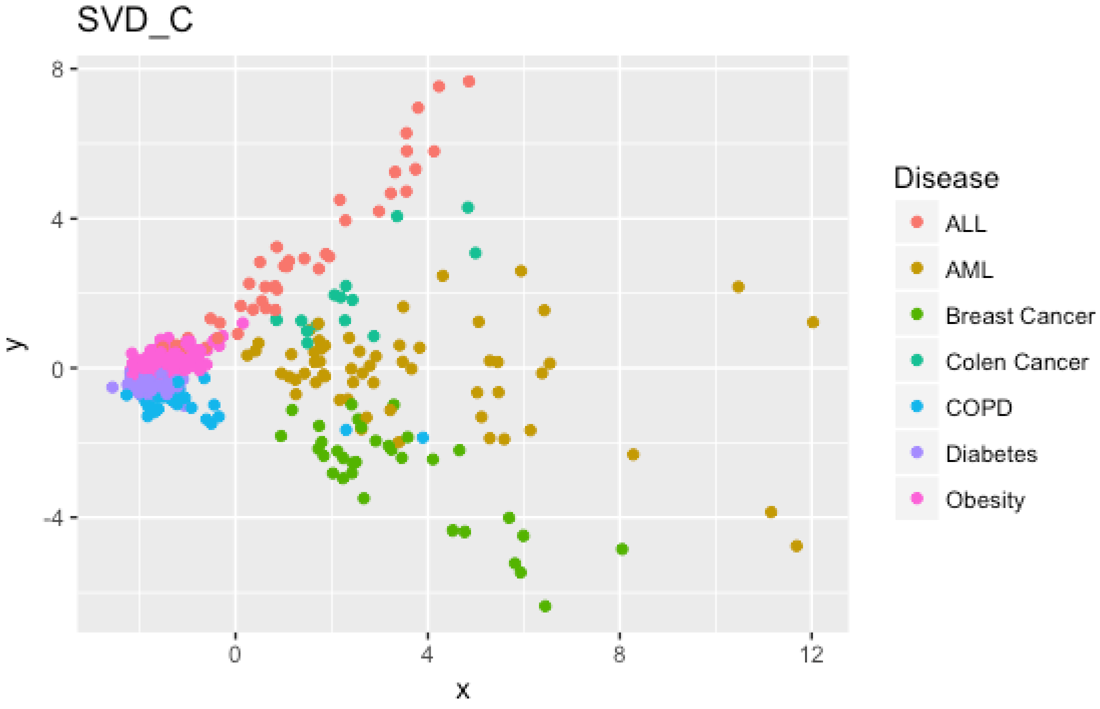
6) Plot obtained from running TSNE on dataset B.



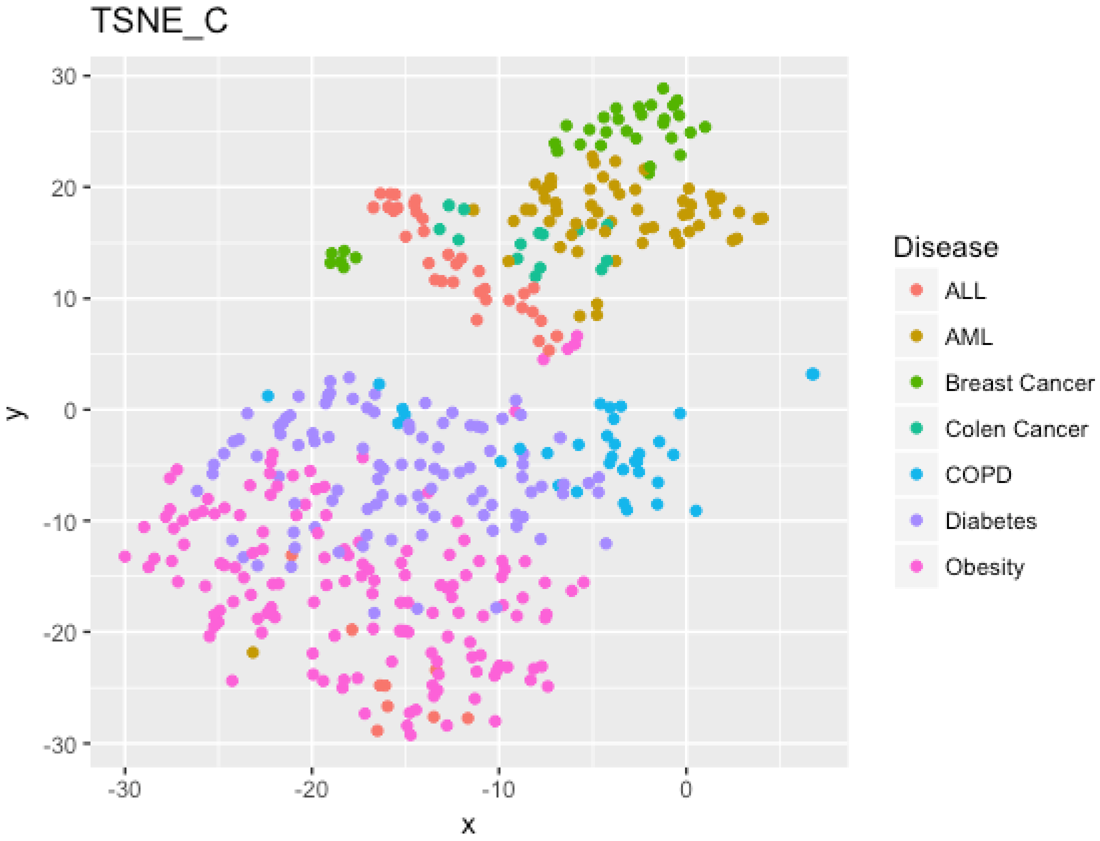
7) Plot obtained from running PCA on dataset C.



8) Plot obtained from running SVD on dataset C.



9) Plot obtained from running TSNE on dataset C.



Observations comparing PCA, SVD and TSNE derived from all the scatter plots:

- Principal Components Analysis (PCA) is a commonly used dimensionality reduction method and often acts as a starting point for data exploration. PCA works best on linear data and is sometimes only able to capture the linear structure of the data.
- SVD is another way to perform PCA. SVD makes much more sense numerically, as in PCA, forming the covariance matrix can cause loss of precision. Also, SVD gives much better results on sparse input data.
- t-SNE is a non-linear method and works better on high dimensional data.

As visible from the scatterplots of the three datasets:

- For dataset with lesser number of dimensions, as in the case of PCA_A.txt, the PCA algorithm works fine and reduces the dataset to clearly distinct data points.
- SVD gives a similar visualization of the data as PCA. If the input data would have been sparse, SVD would have still given a clear visualization of the dataset.
- As the number of dimensions increases, as in the case of PCA_B.txt and PCA_C.txt, the data points obtained by PCA are not as clearly distinct as compared to t-SNE.
- t-SNE gives a much more clearly separated visualization of the data which could then be clearly classified by a clustering algorithm.
- However, different methods perform differently depending on the number of dimensions you are trying to reduce by as well as how dense or sparse your data is. Hence, you need to experiment to find out the method, which works best for you.