Introduction:

This report is on analysis of K-Means clustering for two different given datasets.
1. Gene Expression Dataset
2. Human Hereditary Disease Data

User can use any number of clusters and method for calculating distance while forming clusters. Here I've used euclidean distance and Modified Pearson Correlation(spearman) for generating report. But code is valid for any type of distance method taken by pdist2 function like correlation,hamming etc.

- Euclidean Distance = sqrt((x1-x2)^2+(y1-y2)^2)
- Modified Pearson Correlation
  Pearson Correlation measures the similarity in shape between two profiles.

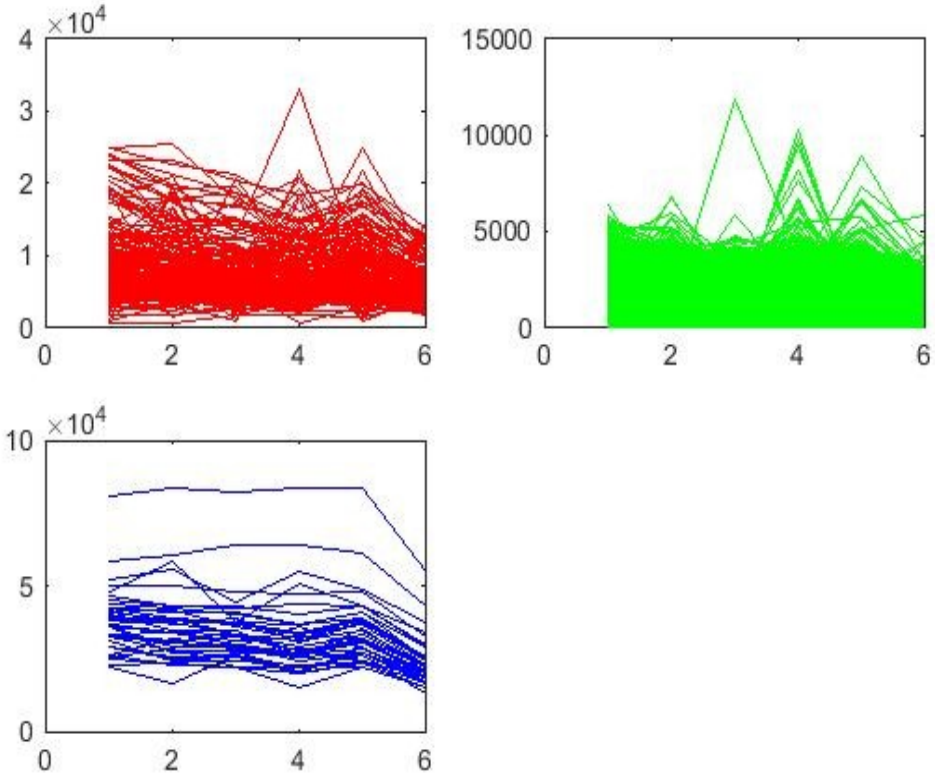
Code Documentation:

Whole code is written in matlab for implementation of K Means clustering on two datasets.
There are five matlab files for it.

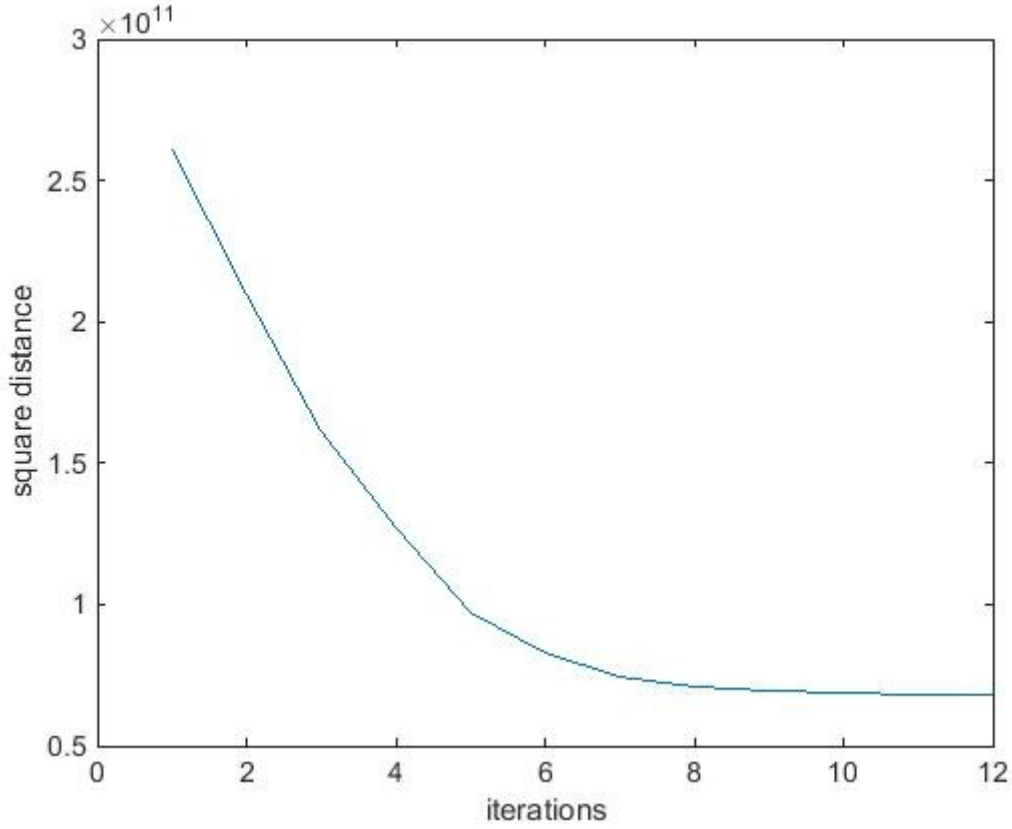
1. **genes.m:** This file import genes dataset for k mean clustering. It preprocesses data ,call knn function and draw plot for different clusters . It prompts user on how many clusters and which method of distance they want to use for k mean cluster .
2. **disease.m:** This file imports disease dataset. It calls knn function and draw 3d plot for different clusters and label data which have been given. It also prompts user on how many clusters and which method of distance they want to use for k mean cluster .
3. **knn.m:** function used for k mean clustering . It uses matrix B as input data . It uses function centroid and distance for clustering . Initially it takes k random points in dataset B for selecting centroid. Then after each loop new centroids are formed till no more centroid changes its position. It also plot square sum of distances with every iteration.
4. **Centroid.m:** This function generates centroid and clusters for the dataset using distance of cluster data points from previous centroid.
5. **Distance.m:** calculate square of sum of distances of data points to their respective cluster.

GENE DATASET:

1. This is for euclidean distance ,k=3. Plot of different clusters w.r.t. column.

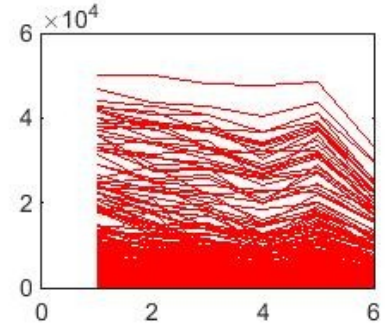


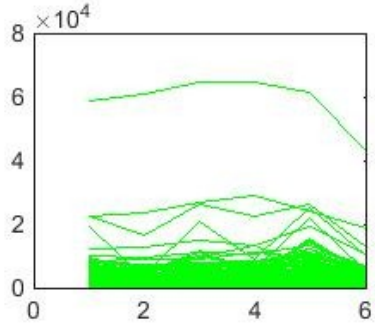
Plot of sum of square distance w.r.t. iterations.

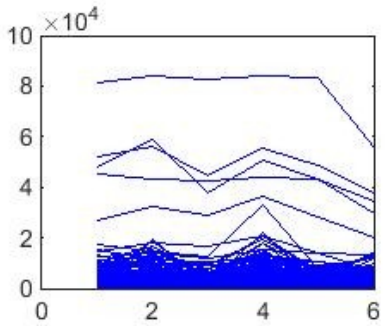


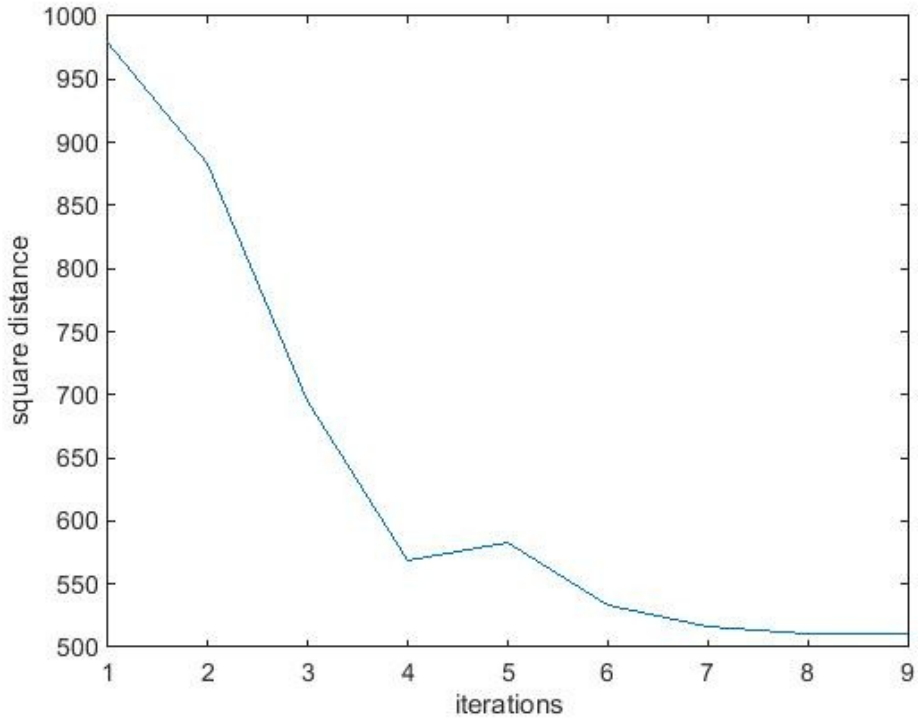
This shows how after each iteration we are reaching towards stable centroid of the cluster as the distance is decreasing.

2) method = Spearman distance , k=3





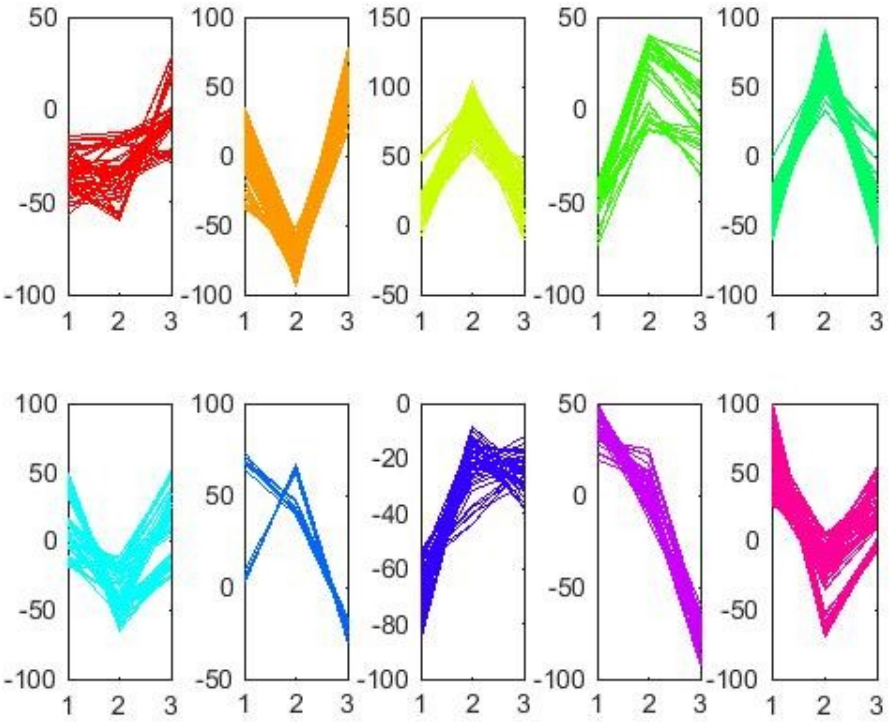




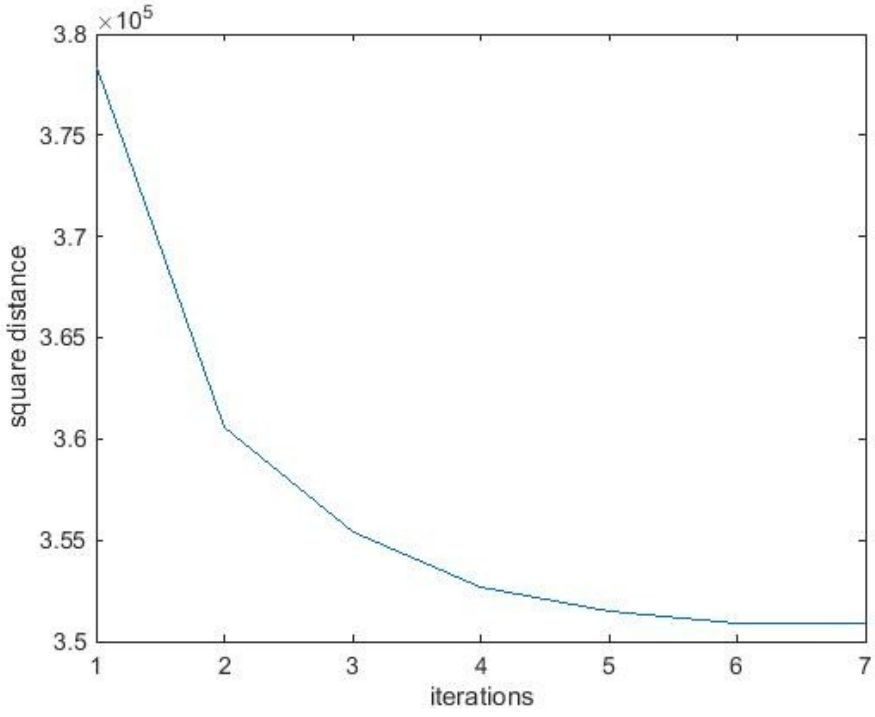
DISEASE DATASET

1) k=10, method = euclidean

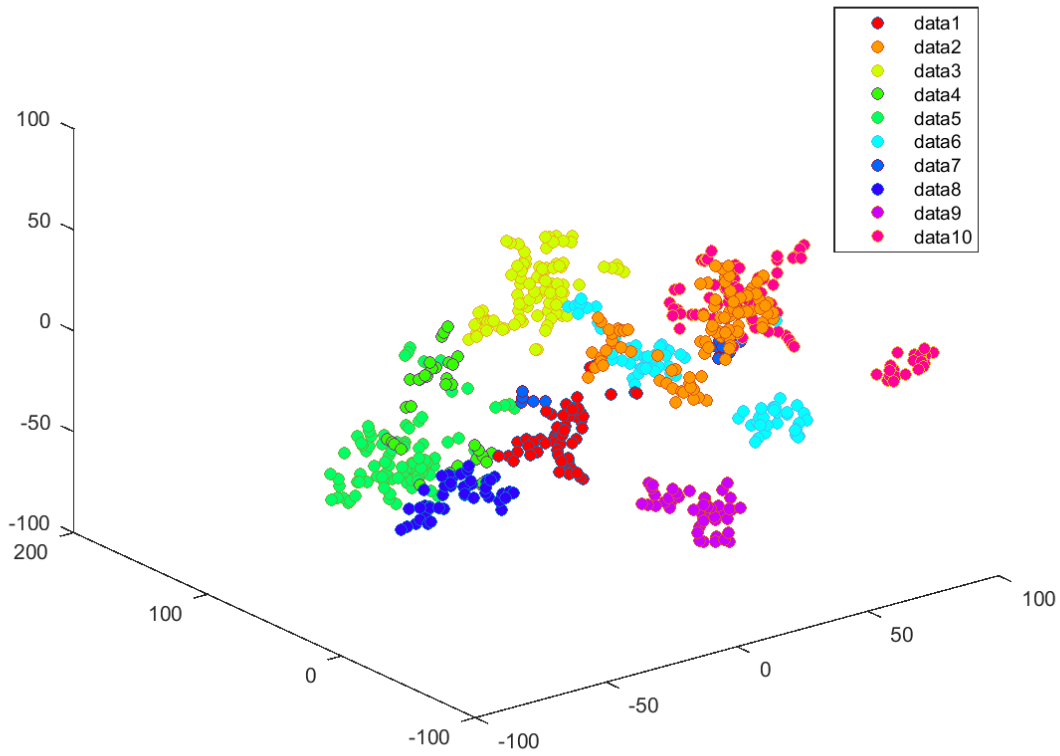
Plot of different clusters w.r.t. columns of the disease dataset



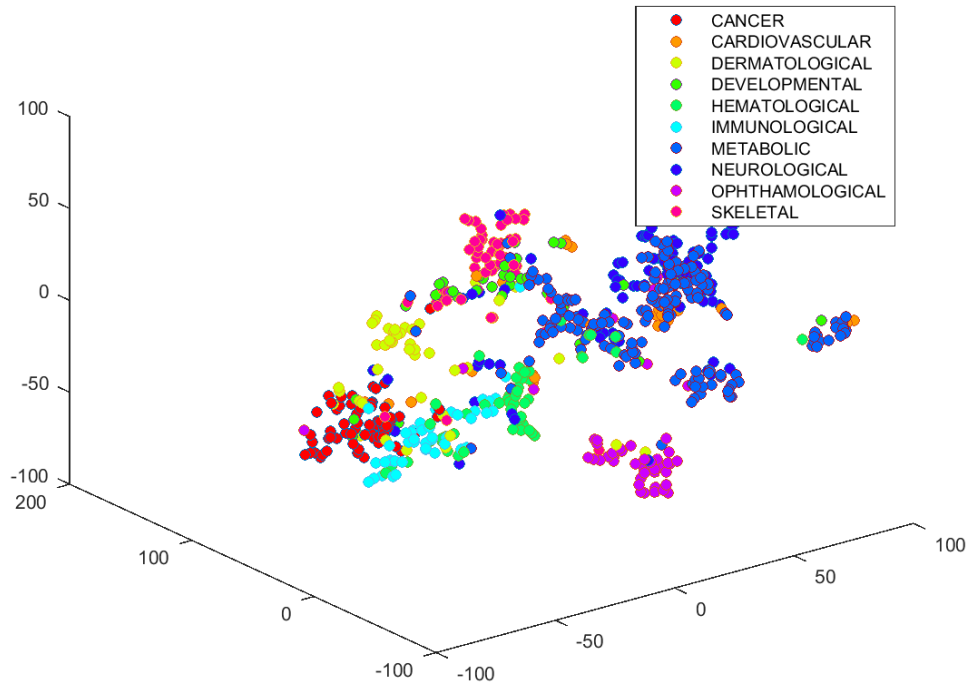
Square of sum of distances reduces in each iteration.



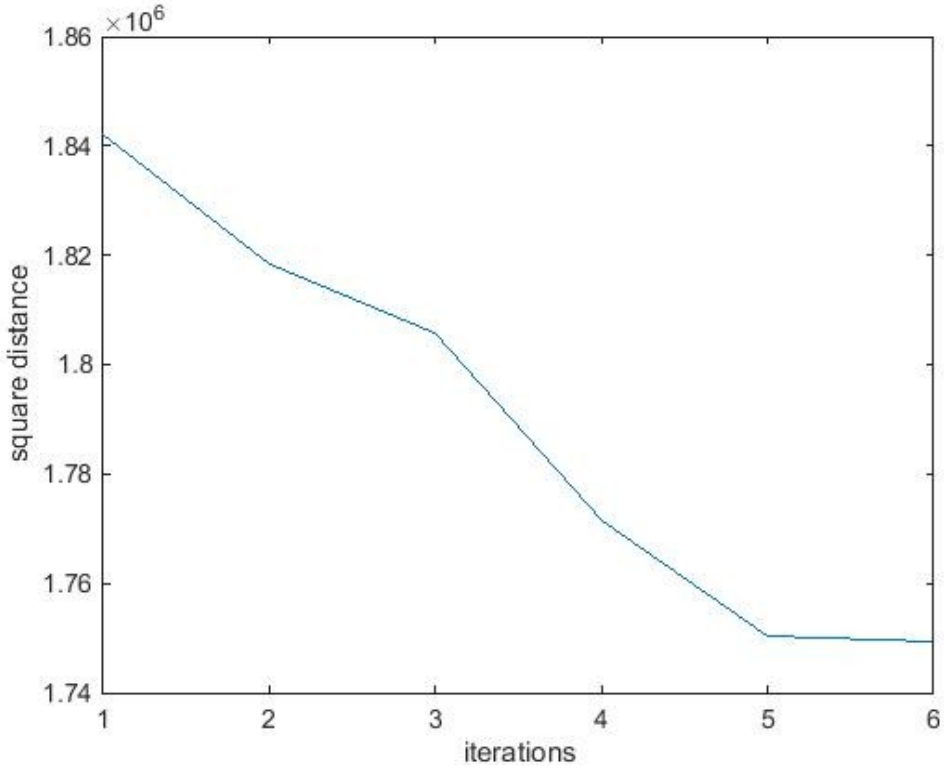
3d plot for different clusters where each axis represents column of the dataset.

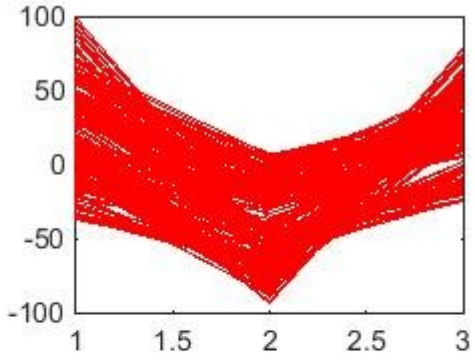


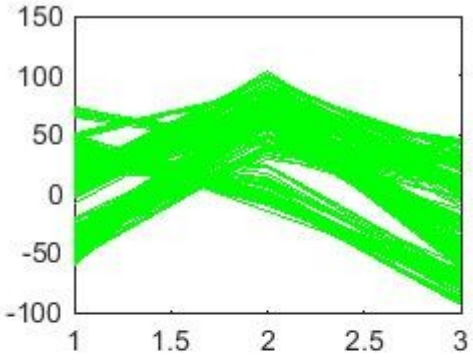
3d plot for various labels given as dataset.

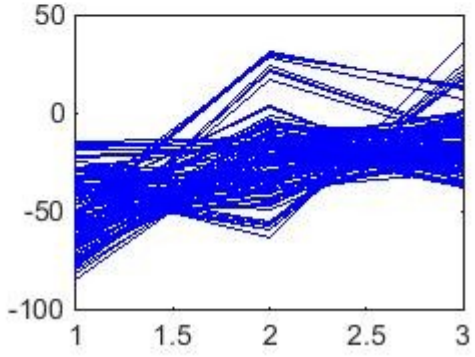


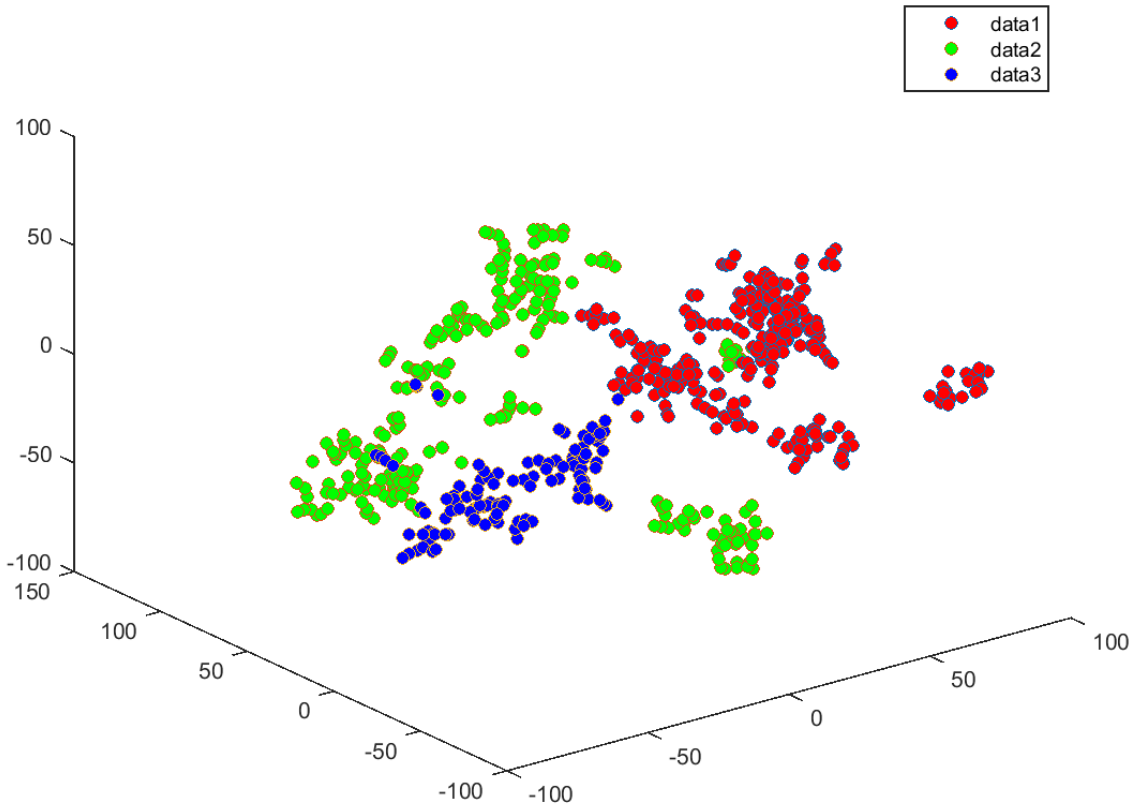
2) method=euclidean, k =3



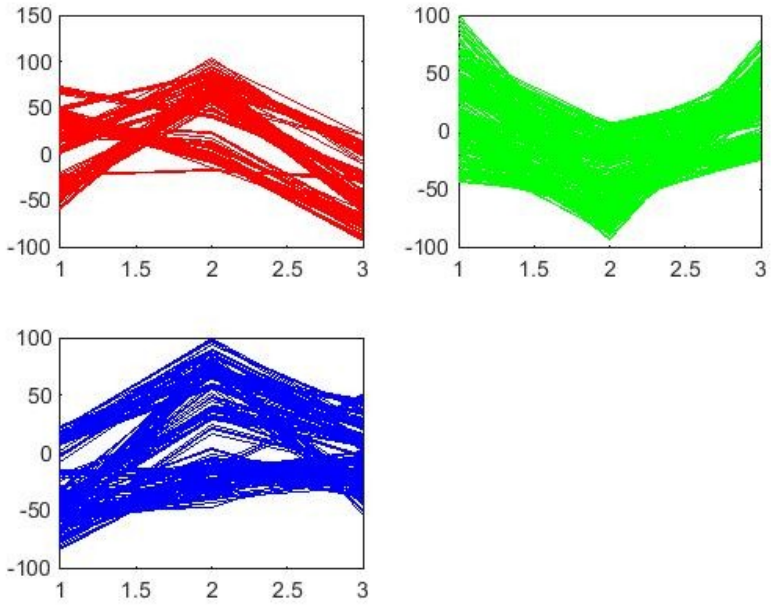


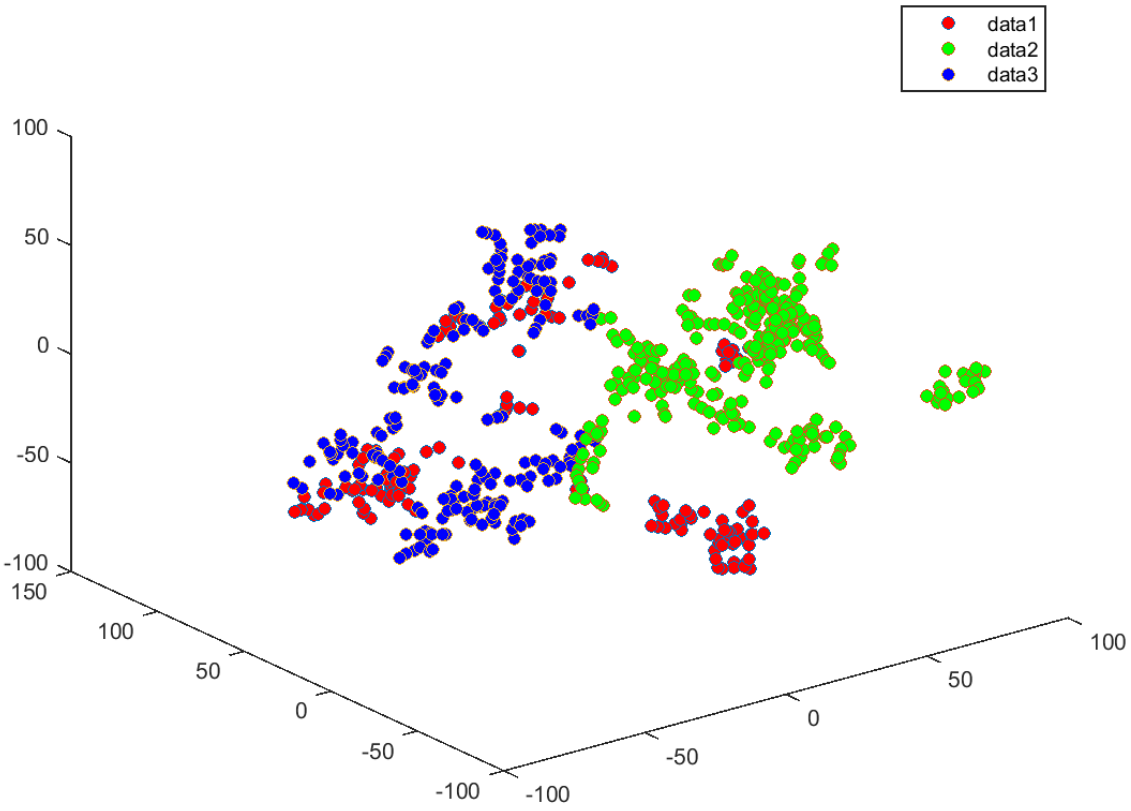




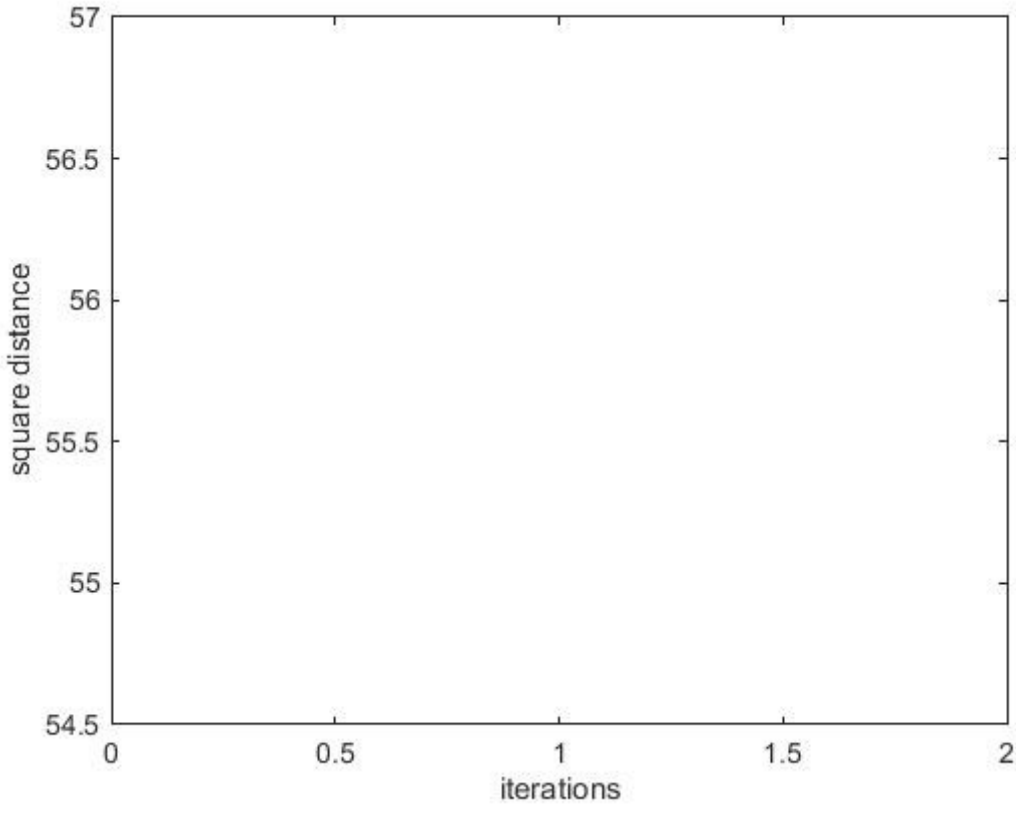


3. method= spearman, k=3





3d plot for clusters.



It goes for only one iteration after randomly choosing centroids as dataset is small and our distance matrix is coming very small for spearman coefficient. Here sum of square distance comes out to be **55.75.**

OBSERVATIONS:

1. Number of iterations are almost similar for various starting points of centroid.
2. Number of iterations needed to find stable centroid becomes more when we increase number of clusters.
3. Distance matrix calculated by euclidean distance method is much larger than calculated by modified pearson coefficient . Hence iterations for euclidean method is larger than modified pearson coefficient method.
4. Plot for square of sum of distance for modified pearson coefficient method doesn't always comes smooth as compare to exponential decreasing in euclidean case.
5.Human Hereditary Disease dataset is very small , better clusters will be formed if we increase data for it.
6. Sometimes for modified pearson coefficient few clusters are not formed and hence it gives error in the plot function.
7. k-means clustering is a heuristic algorithm, there is no guarantee that it will converge to the global optimum, and the result may depend on the initial clusters. By changing initial seeds one can get global optimum.
8. Different clusters are formed using different distance metric. For example euclidean metric tells how close data points are whereas modified pearson coefficient tells how similar two profiles are .