Objective and Data

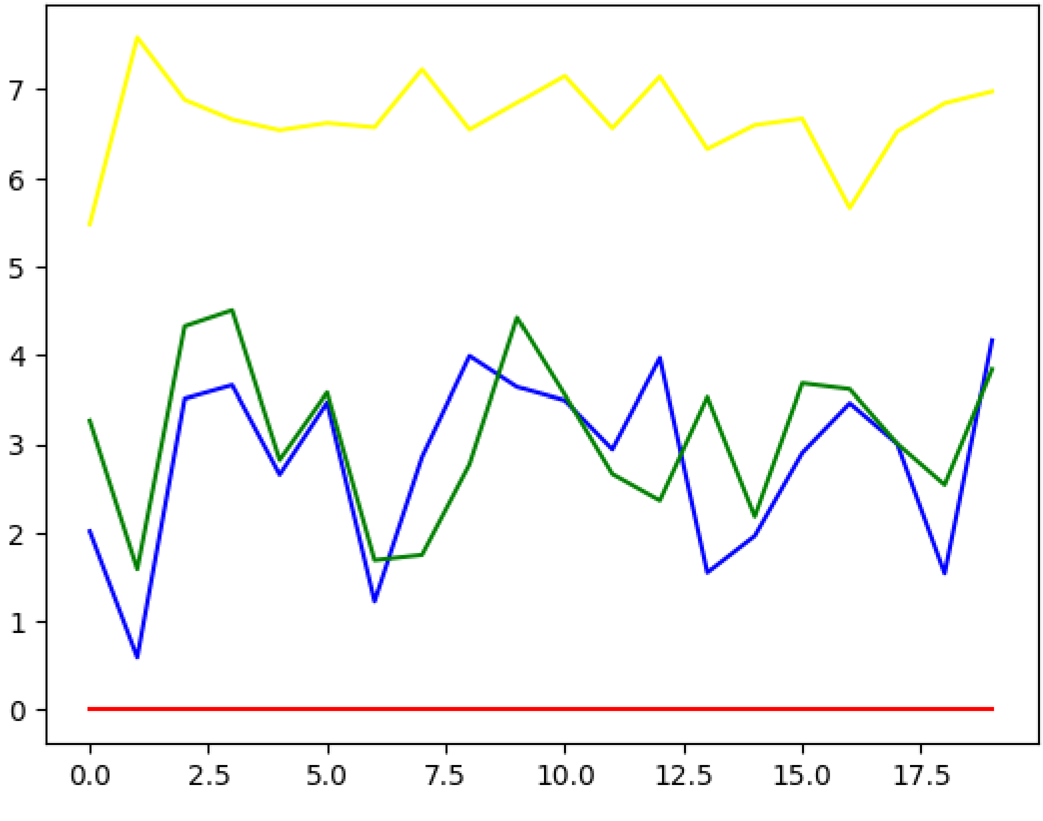
ICMR gene data was used to analyse the gene expression in several cancer samples. The main objective was to find how the data was clustered and whether patterns of clustering could be isolated from the data alone.

The data was first downloaded from kraggle into google drive. Then google drive was connected with google colab.

Data Exploration

An initial exploration showed that the csv file contained data from eight hundred samples. 20530 genes were mapped.

Plotting the first five genes over the first twenty samples showed that some but not all genes are highly correlated.



Hence, clustering is a good candidate for data analysis.

We ran a isnull on the data and observed that the data was already clean.

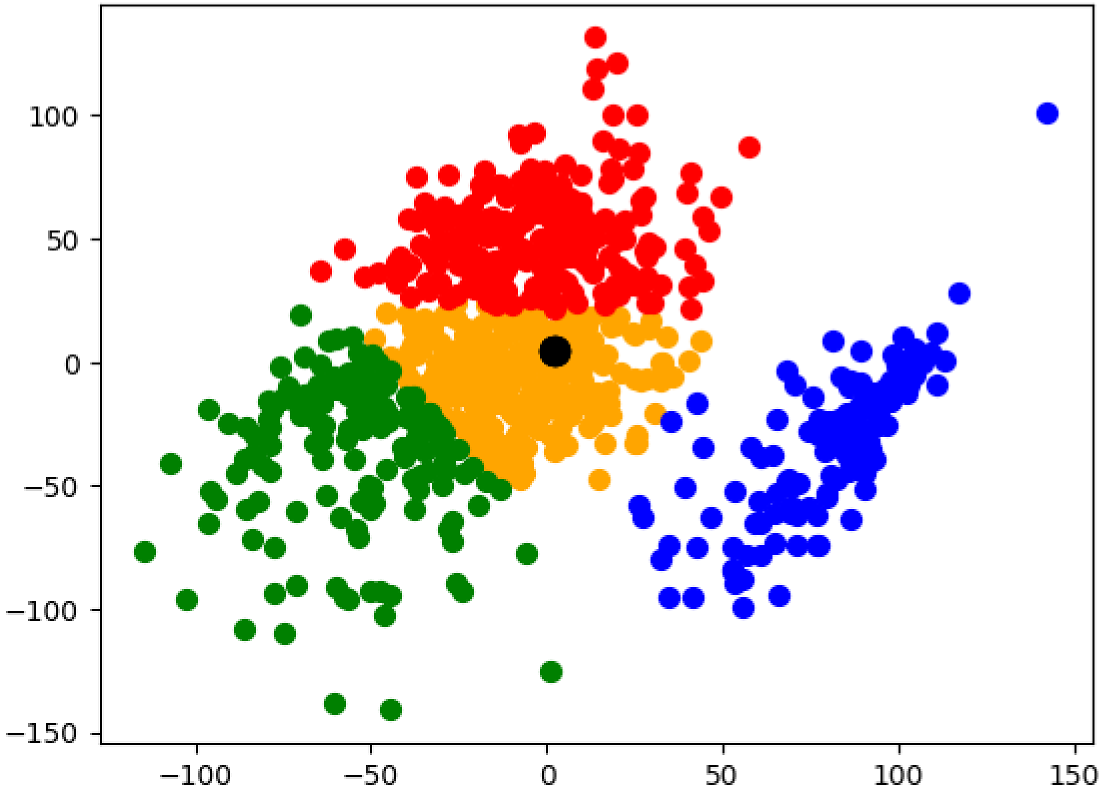
PCA analysis

We first ran a PCA with feature number 4 and also ran a PCA with two features. We also run a PCA for 100 features.

We then run K means with features ranging from 1 to 11 to determine the inertia and hence the optimal cluster number using the elbow method. For different PCA values, our optimal cluster numbers ranged from 4 to 8.

We then ran K means with the optimal cluster number and made scatterplots for clusters using the predicted cluster.

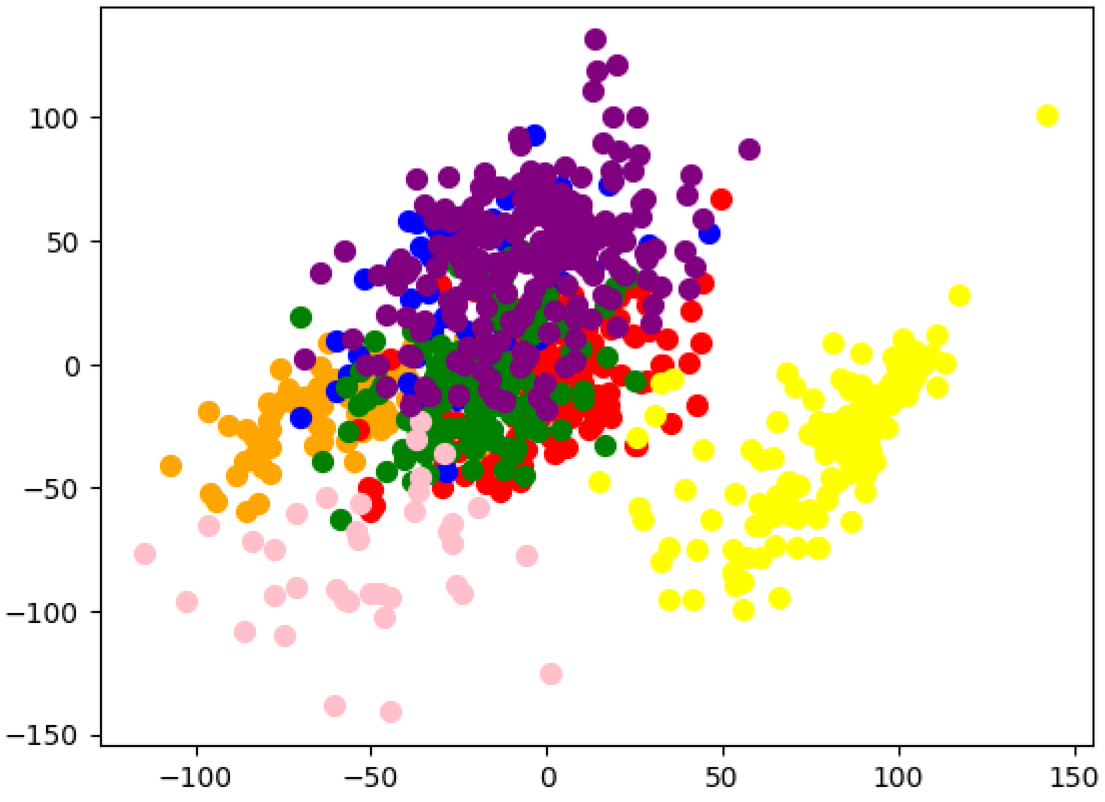
We see that when feature chosen is 2, and cluster chosen is 4, we get very clean cluster. Making variations in either feature number or cluster number did not yield cleaner clustering.



Feature =2, n_cluster =4

This is possibly because we need to plot to higher dimensional spaces to find clean clusters as we increase the number of features. However, the four clean clusters with PCA feature = 2 offers us a glimpse of the first set of clean clusters.

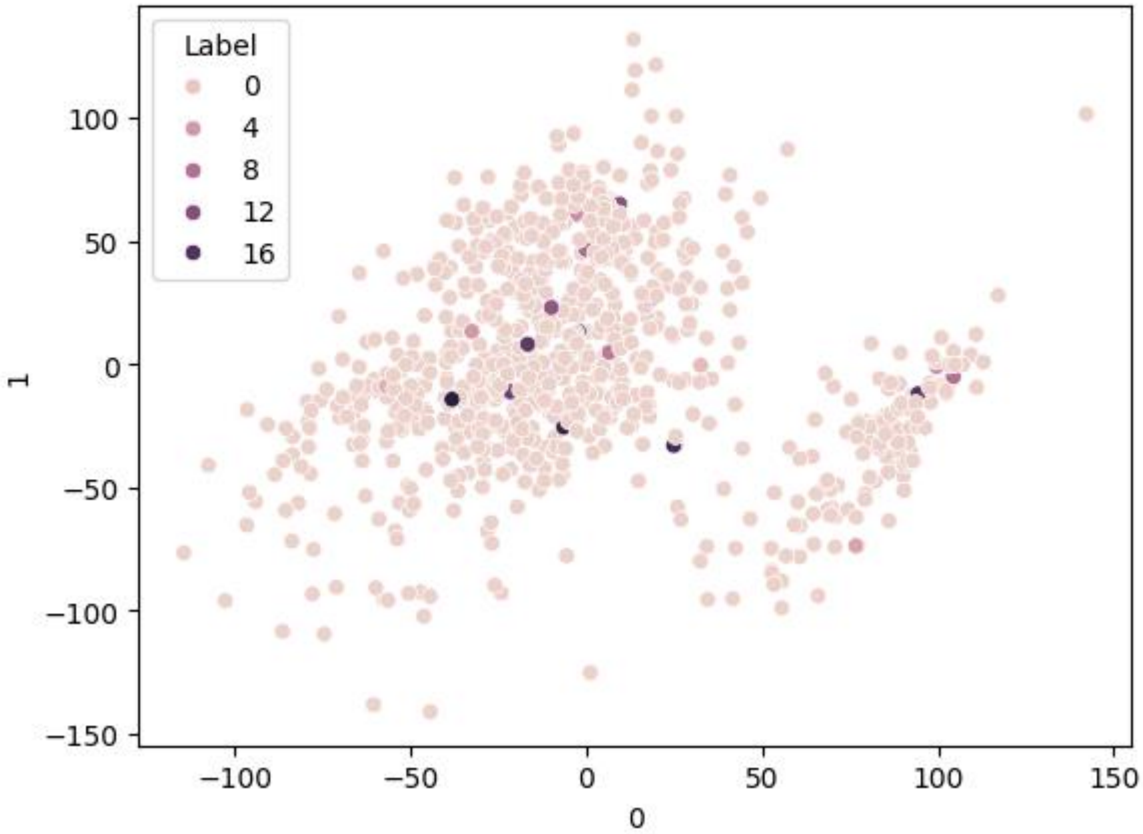
It is to be noticed that even if feature number is increased to 100, the basic idea of four clusters remain visible in the 2d scatterplot



PCA feature 100 and c_cluster = 8

Trying DBSCAN

We also try using DCSCAN for clustering. We use epsilon from .01 to .8 and n_samples = 1, 2, 5 and 8. However, the best plot only revealed small clusters within a larger cluster, hence indicating that DBSCAN is not the best method for this data-set. This is possibly because of the uneven density of data points.



Problems and Future

There are many features in this dataset. PCA involving several features are diffficult to plot.

It would be interesting to be able to plot the higher dimensional clusters for k means. For that purpose, using multi dimensional scaling might be of interest.