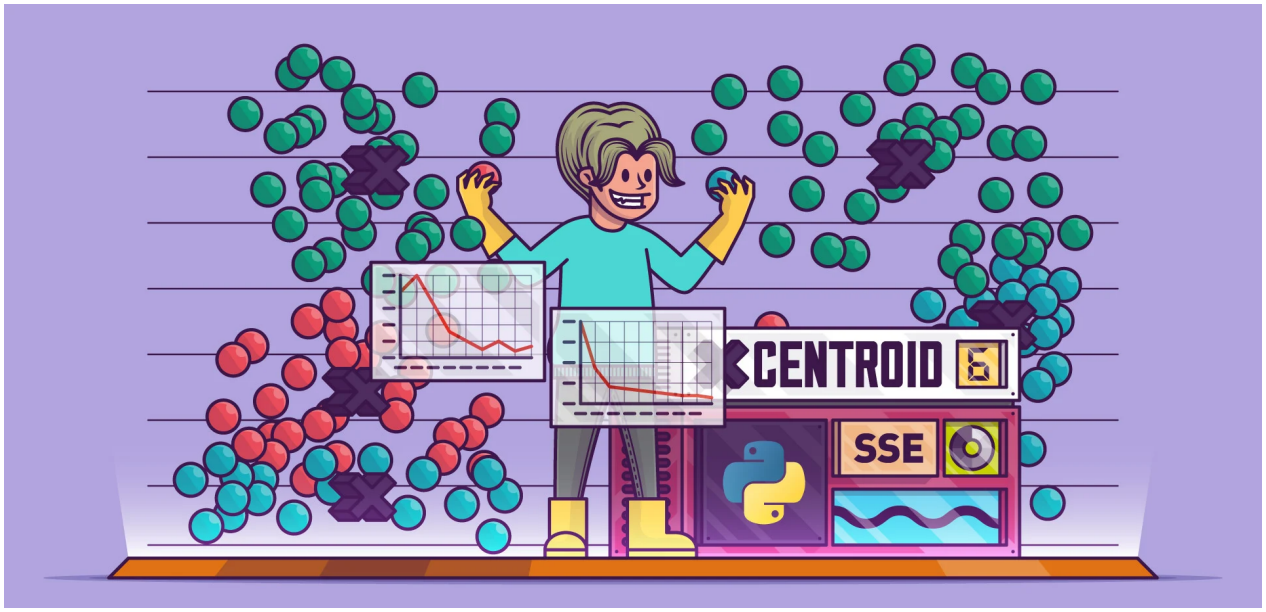
Pattern Recognition and Machine Learning

Lab 9 Report (Kartik Choudhary B20CS025)

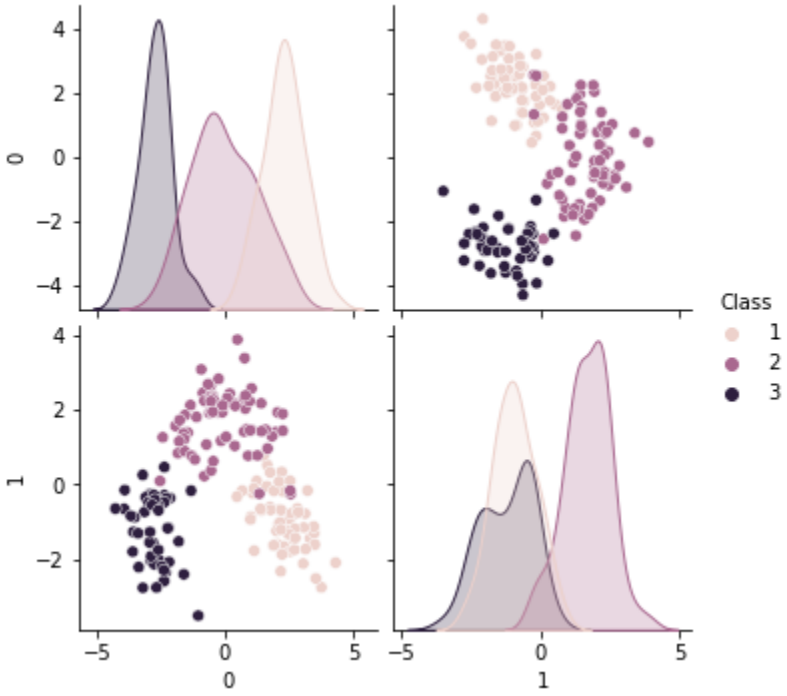
Clustering



Question 1:

a. PCA is used as the dimension reduction technique

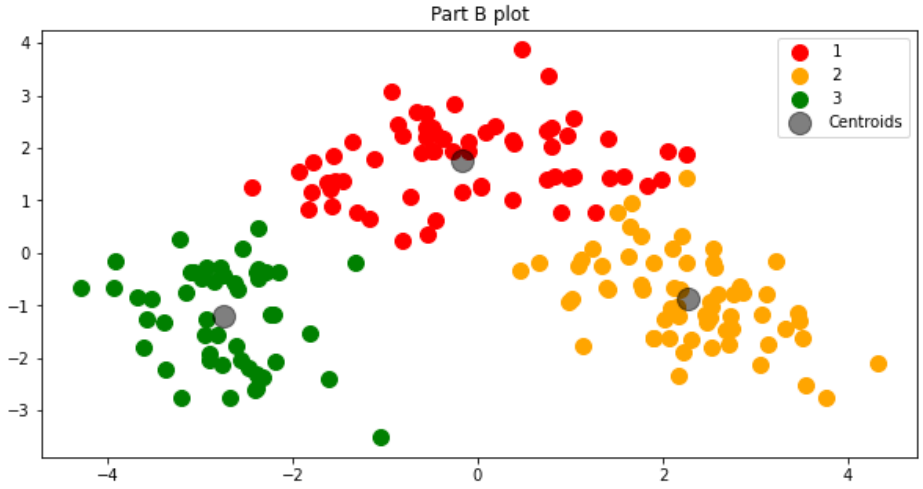
Visualization of the data:



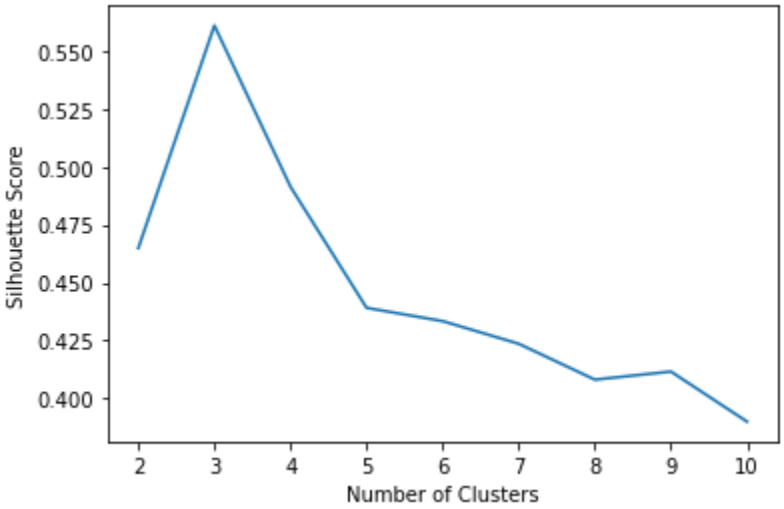
By looking at the data, we can see that the best suited value for k will be 3, because there are only three classes.

b. Building the k means clustering algorithm using sklearn and choosing the value k = 3 from above.

Plotting the centroids :

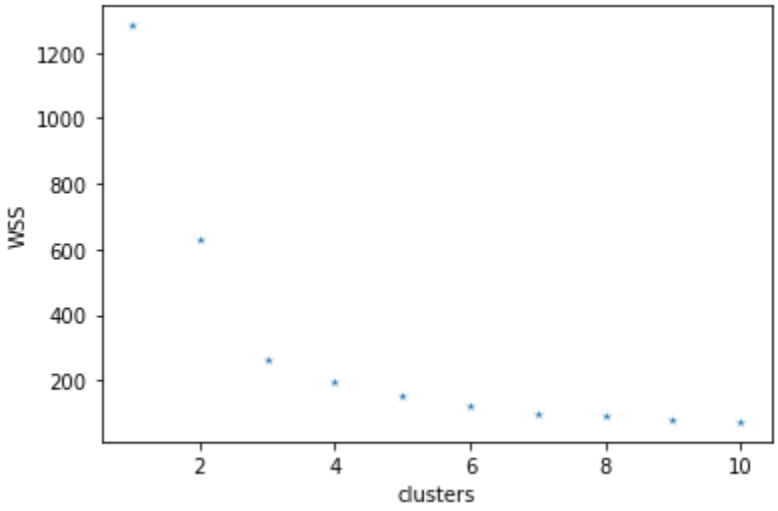


c. Using different values of K and finding the silhouette score and finding the optimal value of k



Clearly the best silhouette score is when the number of clusters is 3.

d. Using the elbow method we can see that the elbow is formed when the number of clusters are 3.



Question 2:

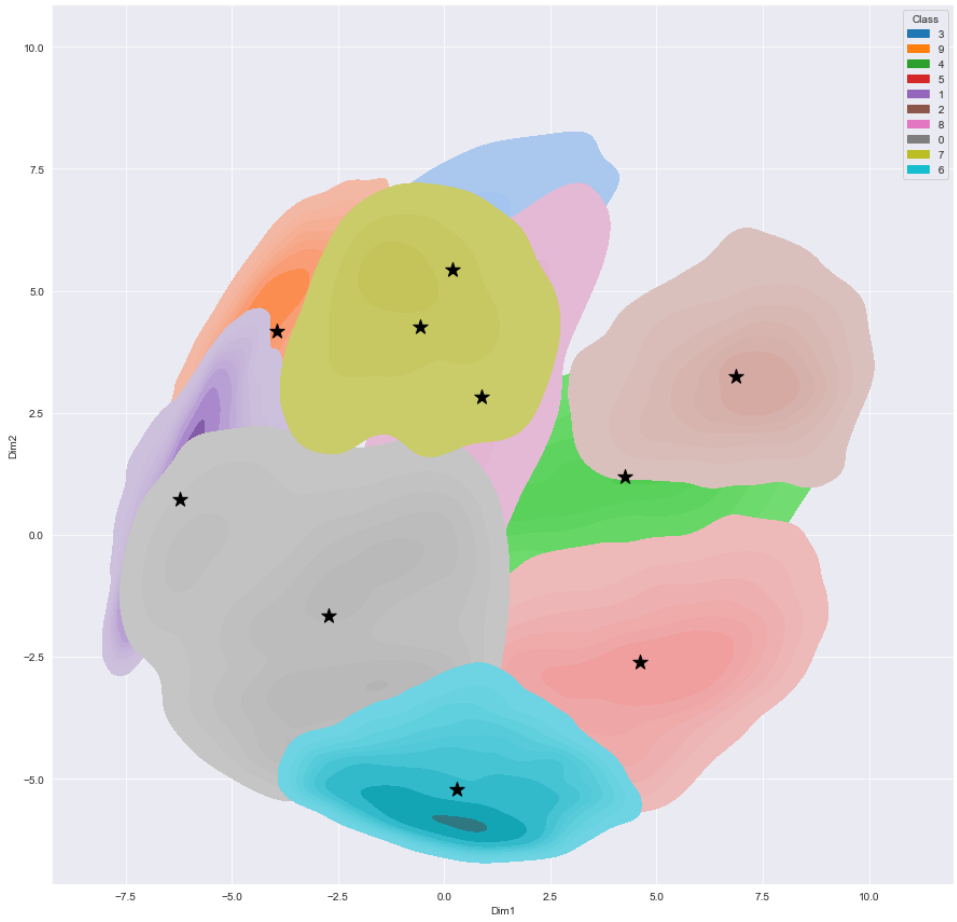
a. Implementing the K Means algorithm from scratch has been shown in the code.

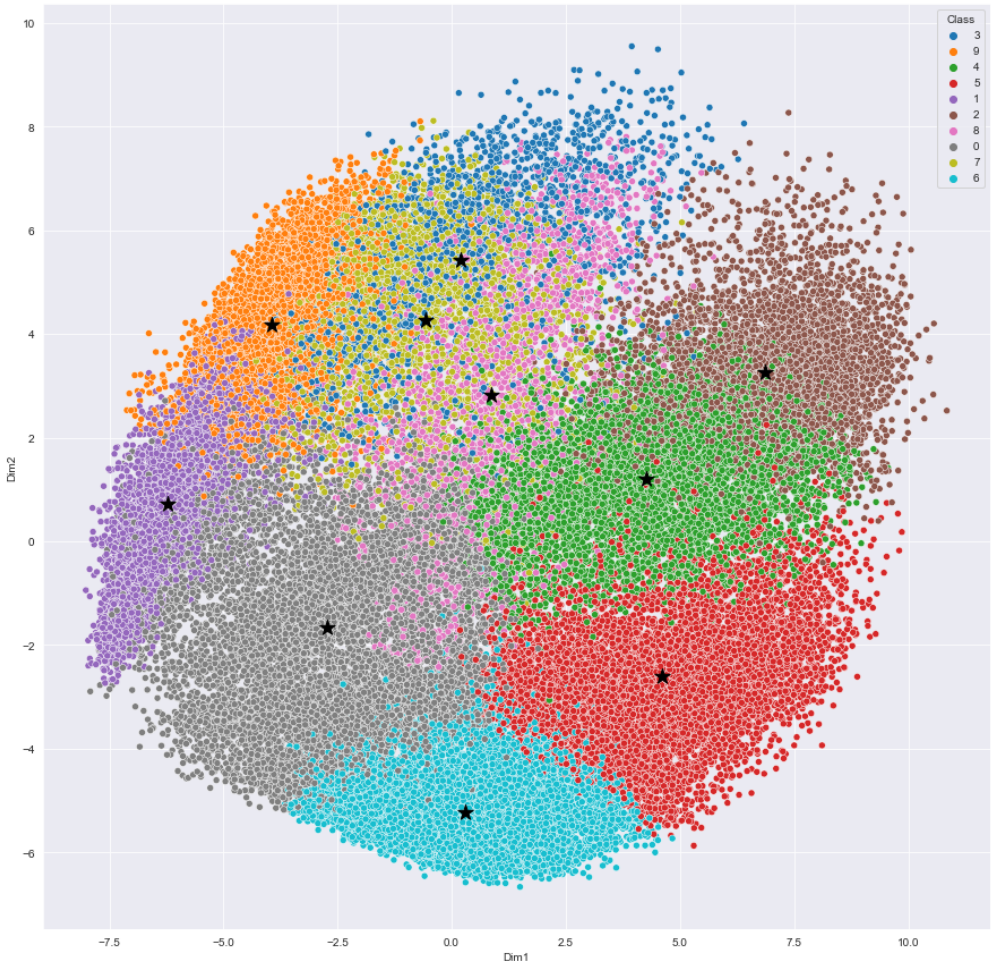
b. All the given conditions are successfully met in the scratch implementation in the code.

c. After training K-Means model on f-MNIST data with k=10 and 10 random 784 dimensional points the reported number of points in each cluster are as follows:

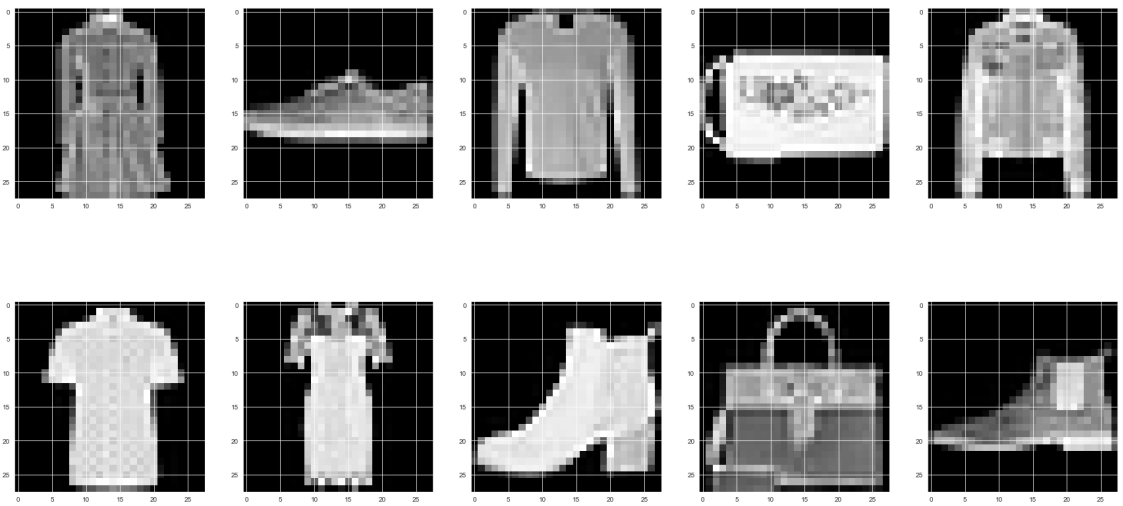
```
Cluster: 3.0 ==> 2337
Cluster: 9.0 ==> 4632
Cluster: 4.0 ==> 6892
Cluster: 5.0 ==> 7701
Cluster: 1.0 ==> 9003
Cluster: 2.0 ==> 4857
Cluster: 8.0 ==> 2577
Cluster: 0.0 ==> 10072
Cluster: 7.0 ==> 3086
Cluster: 6.0 ==> 8843
```

d. Visualizing the cluster centers of each cluster as 2-d images of all clusters





e. Visualization of 10 images corresponding to each cluster is as follows:

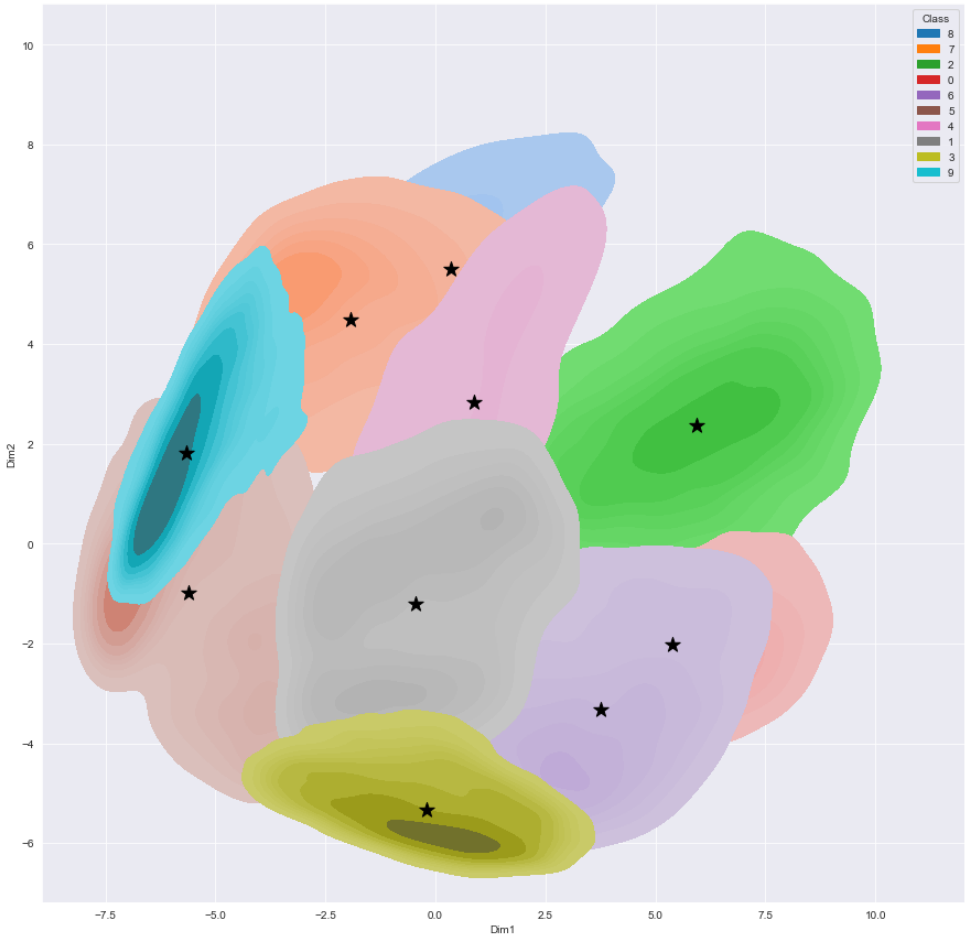


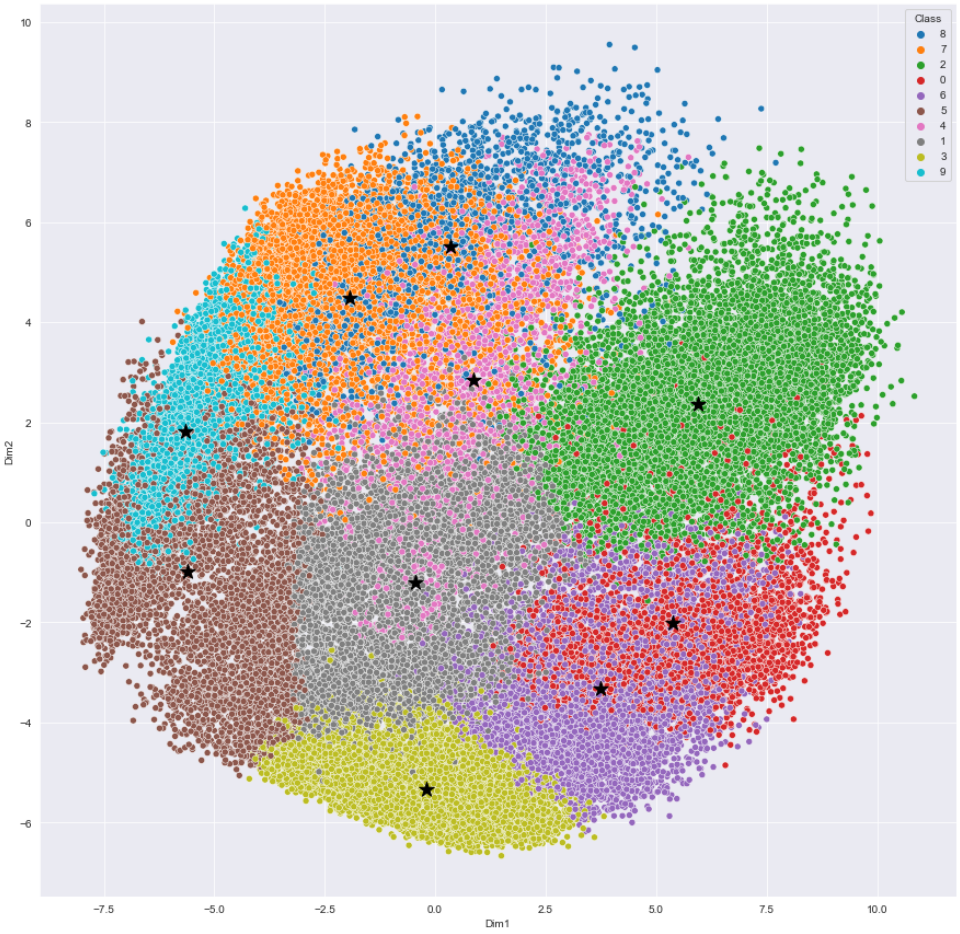
f. Training another k-means model with 10 images from each class as initializations,

Reported number of points in each cluster are:

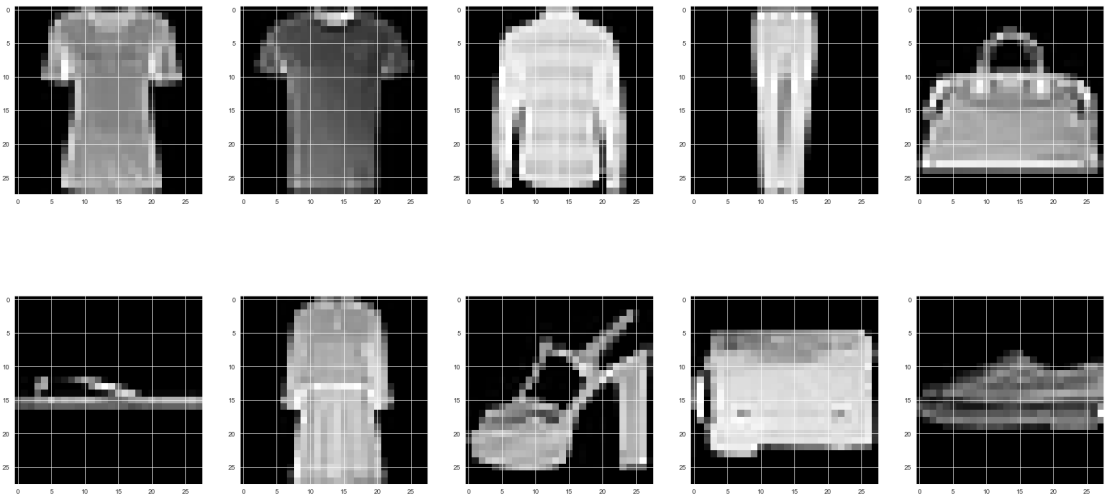
```
Cluster: 8.0 ==> 2349
Cluster: 7.0 ==> 5834
Cluster: 2.0 ==> 9759
Cluster: 0.0 ==> 3810
Cluster: 6.0 ==> 5189
Cluster: 5.0 ==> 7510
Cluster: 4.0 ==> 2558
Cluster: 1.0 ==> 7712
Cluster: 3.0 ==> 7772
Cluster: 9.0 ==> 7507
```

Visualization of cluster centers:

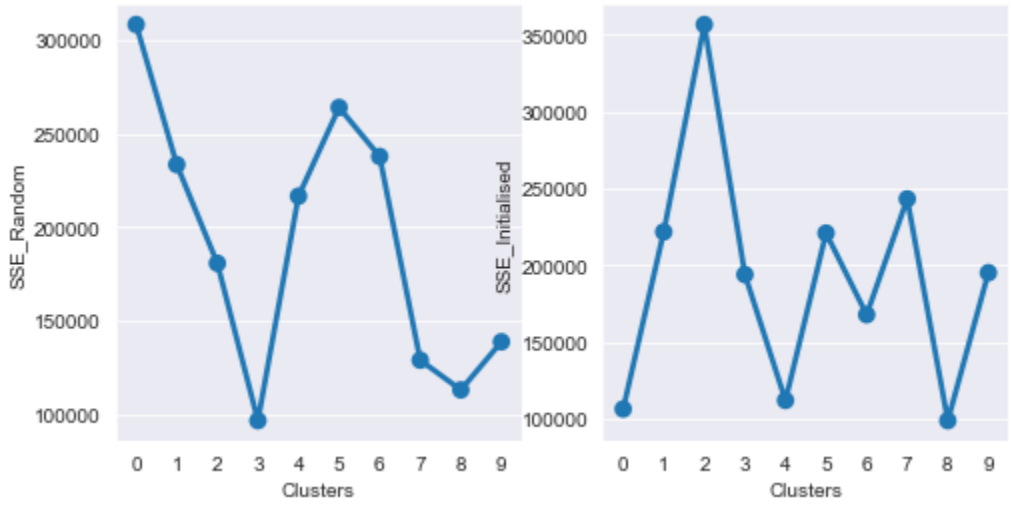




g. Visualization of 10 images corresponding to each cluster:



h. Plotting the SSE's reported:



On calculating the sum of SSE's Results for both the models are almost similar with Initialized centroids approach performing slightly better.

```
SSE --> Random --> 1916427.0093593828
SSE --> Initialised --> 1915335.1297465428
```

Question 3: Hierarchical clustering

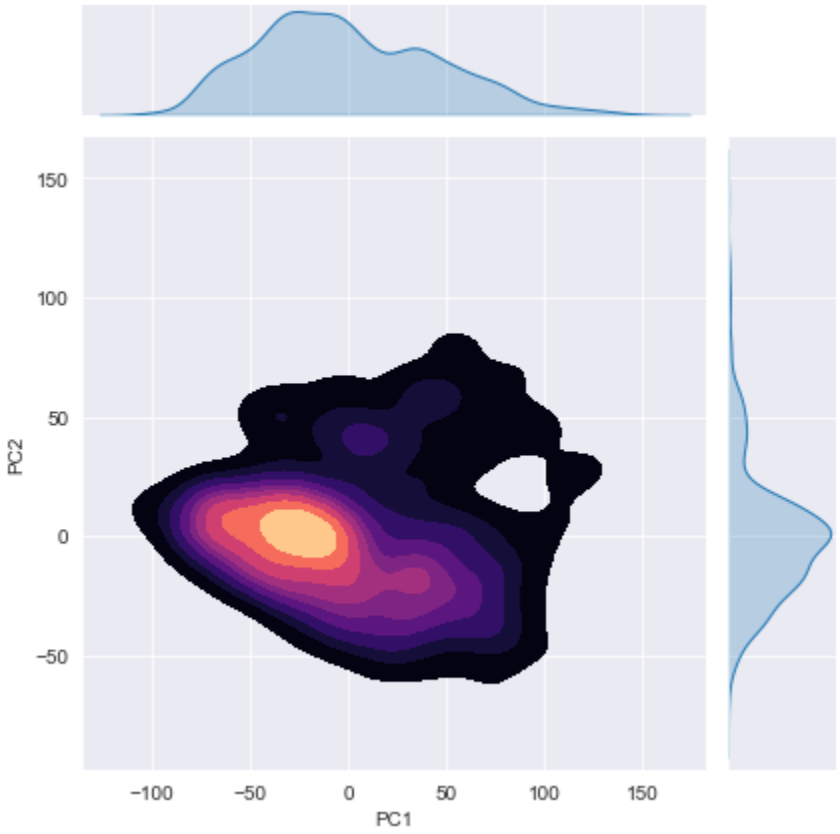
a. A joint dataset containing 1500 images from the given 'Yes' folder, and 1500 images from the given 'No' folder (total of 3000 images) is constructed. The images from the said folder are read in and resized (100 x 100) using the cv2 library, and stored as a list. The list is converted to a NumPy array, resized (3000, 10000), and finally converted to a Pandas Dataframe with appropriate column names and class labels. Statistics of the data are analyzed via '.head()', '.info()', and '.describe()'. Following observations are made:

- There are ten thousand(10000) features and three thousand(3000) data points.
- Only continuous features are available.

● The scale of continuous features is different.

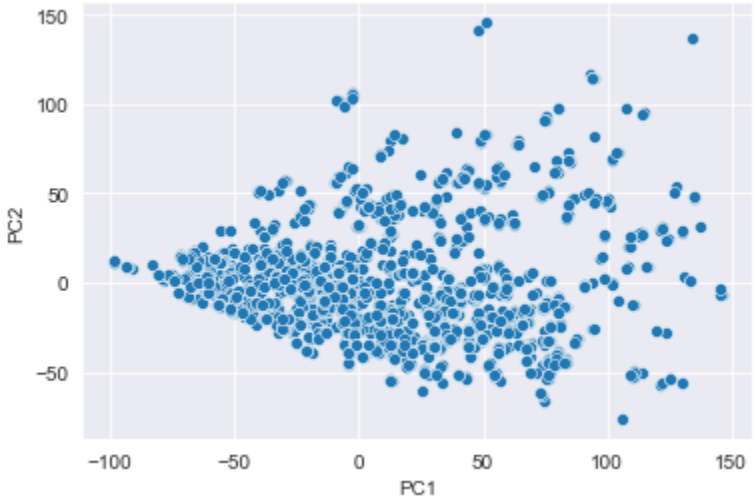
● No missing (NaN) values. The features are normalized via StandardScaler() so that the scale of each variable will be the same.

b. Principal component analysis (PCA) dimension reduction technique with ‘n_components = 2’ (for two-dimensional data) is used. Visualization of the data after implementing the mentioned technique is as follows:

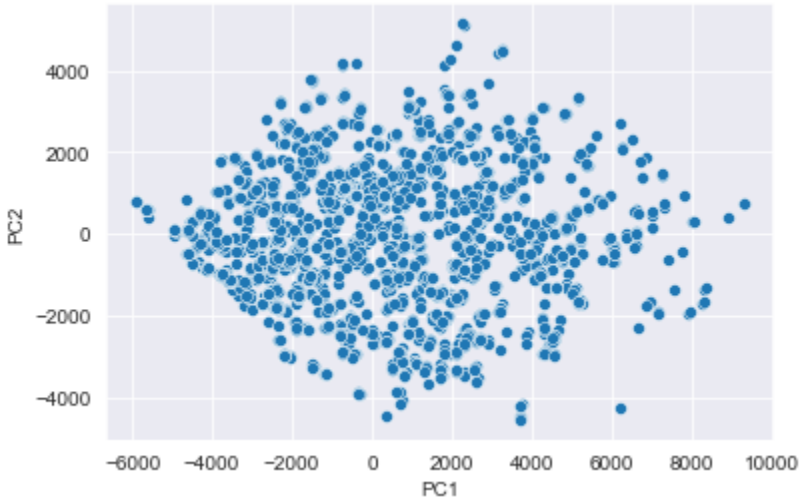


c. Visualization of the communities from ‘Part A’ is as follows:

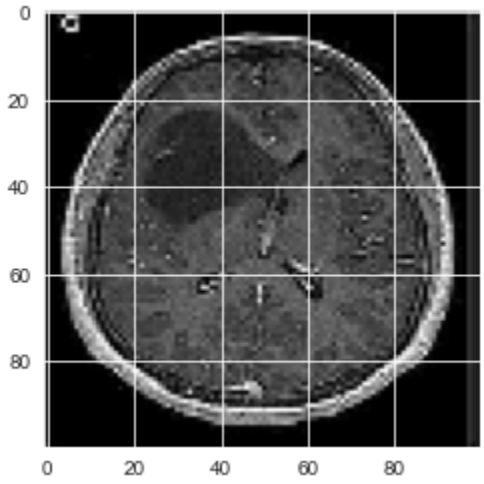
With Scaling:



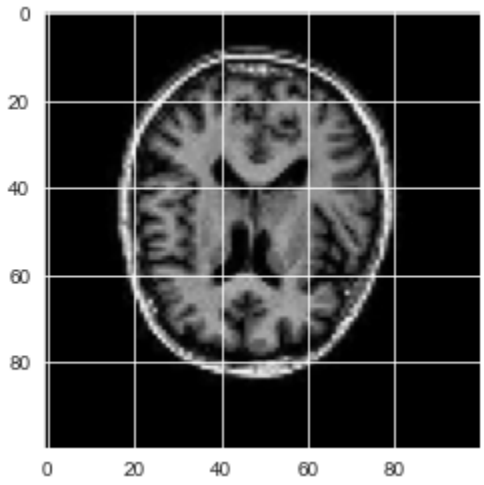
Without Scaling:



Community 1: (YES)

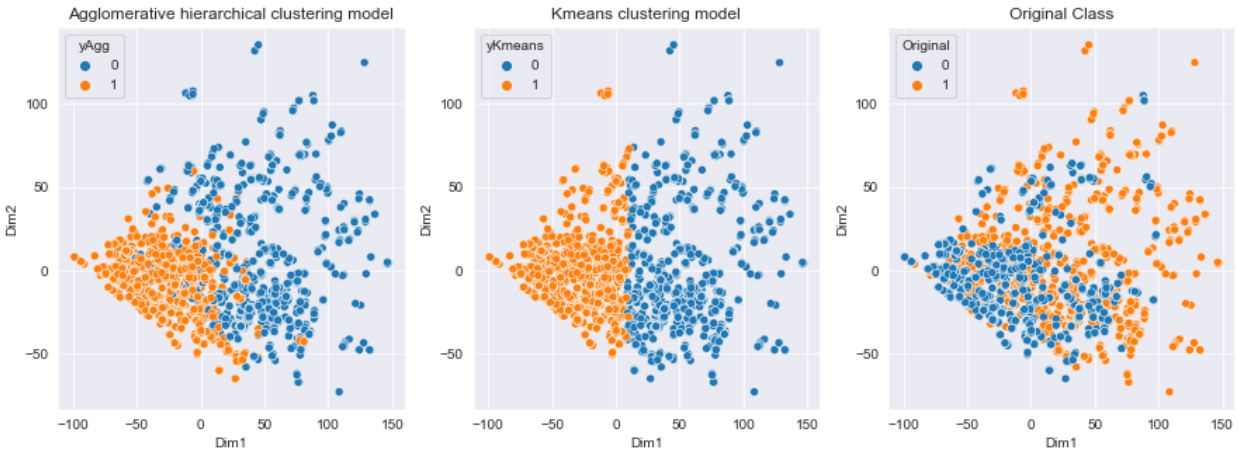


Community 2: (NO)



d. Agglomerative hierarchical clustering (using sklearn) is applied on the mentioned dataset.
e. KMeans clustering (using sklearn) is applied to the mentioned dataset.

Comparison based on clustering is visualized below:



Observations:

Based on the above-mentioned data, the following observations can be made:

- KMeans clustering has a better accuracy score
- KMeans divides the given samples into strictly two halves(left and right) from the middle
- Agglomerative hierarchical clustering divides the given samples along with a diagonal (left upper and right lower).

Conclusions It can be concluded that though KMeans has a better accuracy score and separated unique clusters, clustering implemented by Agglomerative hierarchical clustering is more efficient in the sense that it resembles the original classification. This can be justified by the fact that though KMeans cluster based on distance from the centroid of the cluster sample points, Agglomerative hierarchical clustering follows a bottom-up approach. A structure that is more informative than the unstructured set of clusters returned by flat clustering