PROJECT PART 2:

Unsupervised Learning (K- Means Clustering)

Setup:

Define four functions which perform the following tasks:

1. Find the Euclidean distance between two data points
2. Implement k- means using strategy 1
3. Implement k- means using strategy 2
4. Finding the Objective Function

Implementation:

1. **Finding Euclidean Distance between two data points:**
   Given two points X and Y in d- dimensional space such that X = [x1, x2,x3……..xd] and Y = [y1, y2, y3, ……..yd], the euclidean distance between X and Y is defined as:

   $$d(x, y) = \sqrt{(x_1 - y_1)^2}$$

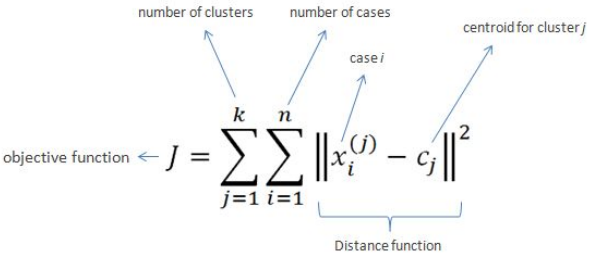
   The function returns the value of d(X, Y).

2. **K- means using Strategy 1:**
   Randomly pick the initial centers from the given samples.
   Steps:
   - Initialize the centroids at random in k clusters
   - Run the loop till the time the next calculated centroids are equal to the value of the previous centroids.
   - Inside the loop:
     - Assign the data point to k clusters
     - Store the euclidean distance of each point to the randomly chosen centroids and find their minimum distance.
     - Create the clusters based on the distance and assign the data point to their respective centroid (with which it has the minimum distance)
     - Store the clusters numbers of each data point
     - Update the clusters with new values which will now serve as the new centroids and repeat the procedure until the convergence criteria are met

3. **K- means using Strategy 2:**
   Pick the first center randomly; for the i-th center (i>1), choose a sample (among all possible samples) such that the average distance of this chosen one to all previous (i-1) centers is maximal.
   Steps:
   - Initialize the first set of centroids with random value

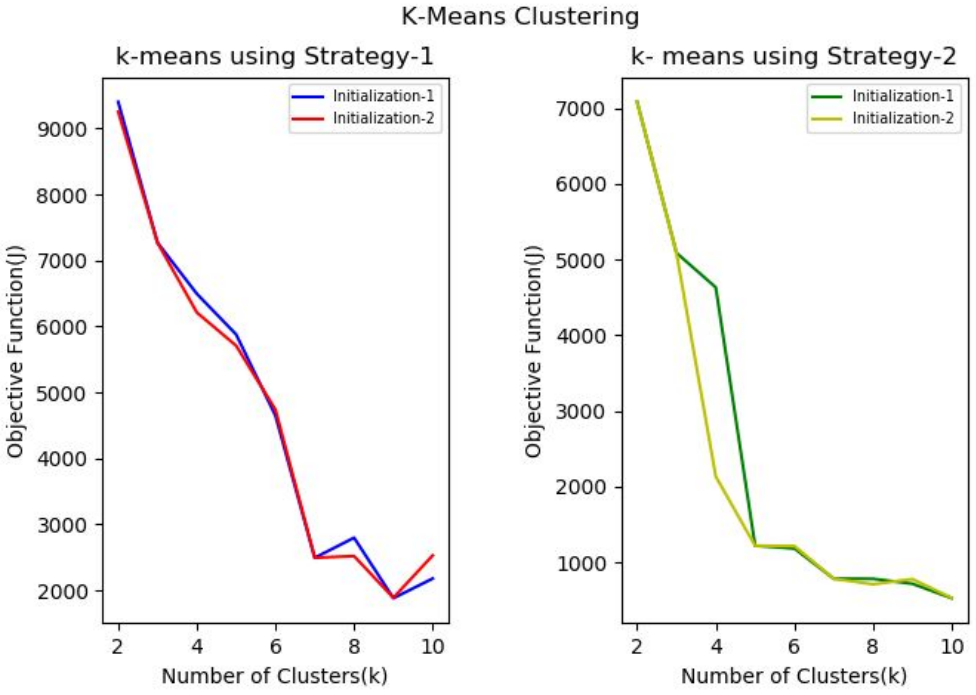
- Store the euclidean distance of each data point with the centroids chosen in the previous step
- Until the convergence criteria are met, keep assigning the data points into the cluster (based on the minimum distance to the centroid)
- The new set of centroid values is calculated by calculating the mean values of all the data points present in the corresponding clusters
- Update the k centroids and repeat the procedure until the convergence criteria have been met.

4. **Finding the Objective Function:**
   The corresponding cost/objective function J that is minimized when the data points are assigned to clusters using the euclidean distance metric is given by:



Calculate the objective function or the cost of the clustering for the number of clusters (k) varying from 2 to 10. Then plot the function values for Strategy 1 and 2. Under each strategy, plot the objective function twice, each start from a different initialization.

Graph Representation:



Caveat:
The graph shown above is just one scenario. The centroids are chosen randomly for strategy 1. Also, for strategy 2, the first set of centroids chosen is random. Thus running the code again can produce different results/graphs.

Results (as per the above graph):

Strategy-1:
Number of clusters required for convergence is between 3- 4 (with J = 7200 (approx))
Strategy-2:
Number of clusters required for convergence is 5 (with J = 1200 (approx))