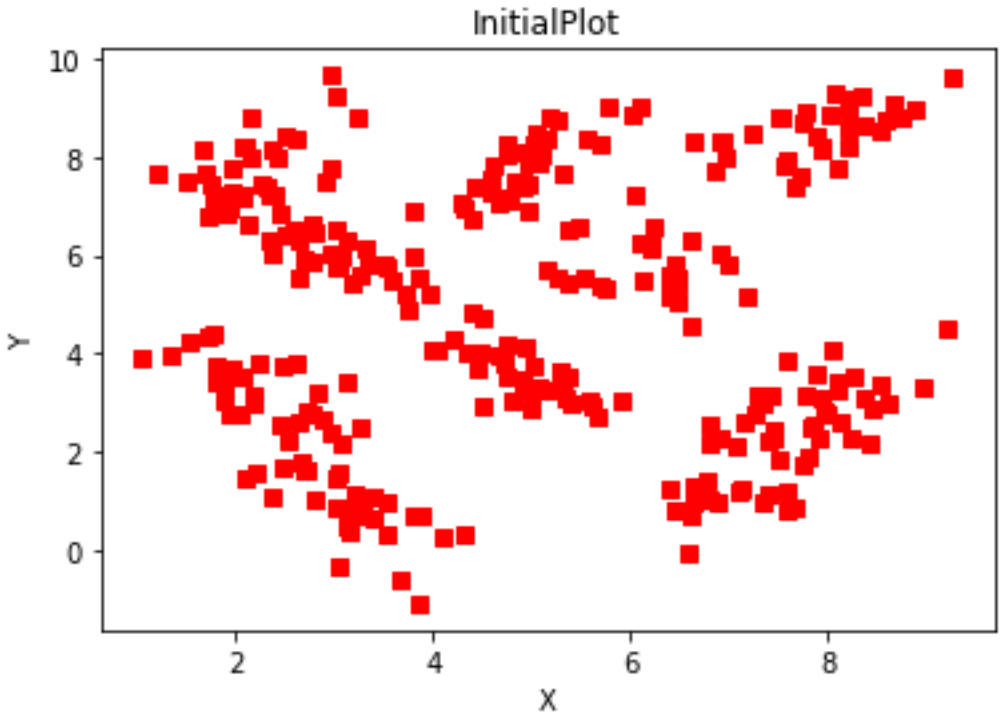
CSE 575 STATISTICAL MACHINE LEARNING

PROJECT 2 – UNSUPERVISED LEARNING (K-MEANS)

NAME: VAMSI KRISHNA KANAGALA **ID: 1218608781**

INTRODUCTION:

In this project, we have performed unsupervised learning using a clustering algorithm, K-Means. The objective of the project is to implement the K-Means algorithm on the given the dataset, which contains a set of 2-D points. The K-Means algorithm is implemented using two different strategies for choosing the initial cluster centers. Also, we have tested the implementation on the given data, with the number K of clusters ranging from 2-10. The dataset contains x and y coordinates of 300 datapoints. Below is the scatter plot of the given dataset.



Below are the plots and implementation details of both the strategies.

Strategy-1:

In this strategy, initially the centroids are chosen randomly before implementing the algorithm. And then, we have calculated the distance of each datapoint to these initial centroids using the below formula

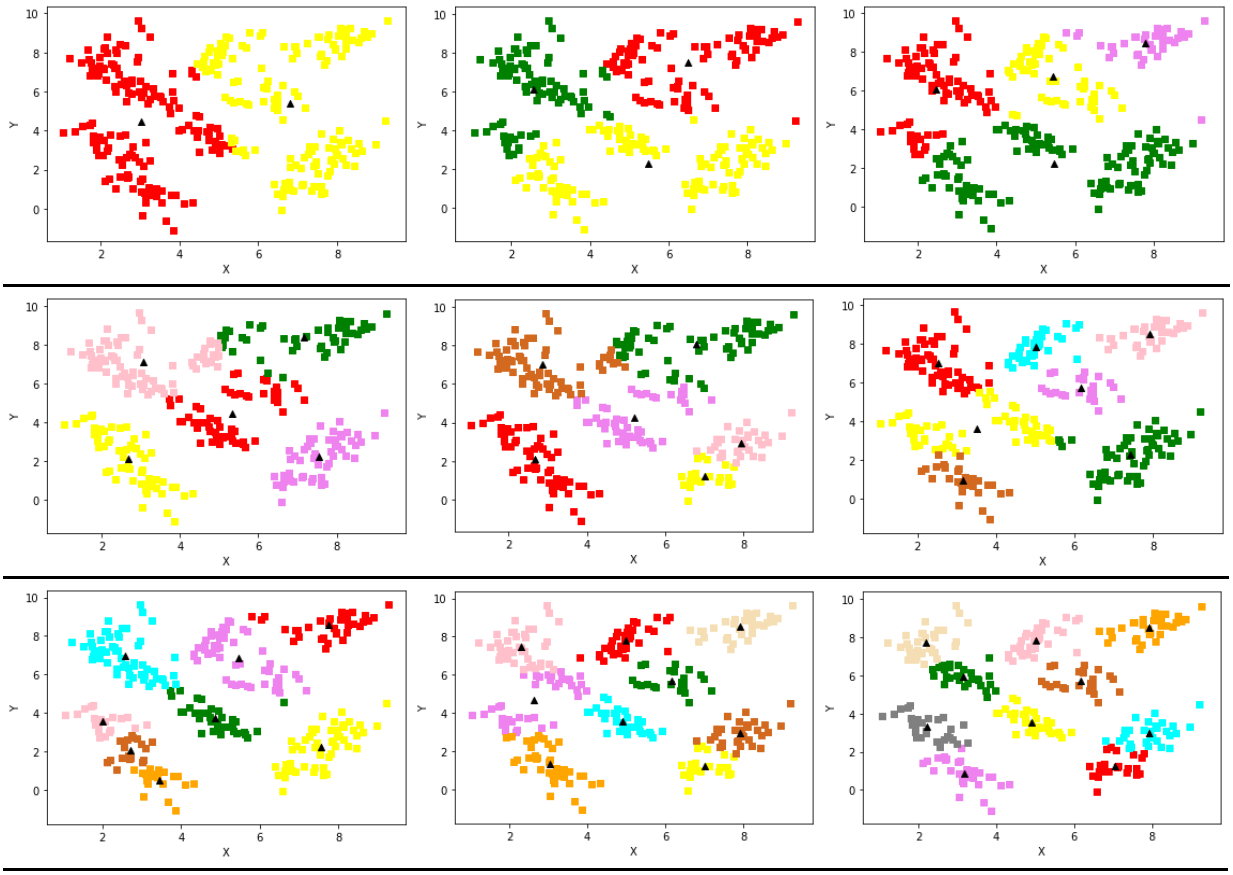
$$For\ a\ centroid\ \mu\ with\ x\ coordinate\ \ \mu_x\ and\ y\ coordinate\ \mu_y$$

$$d_x = |x_i - \mu_{xj}|^2 \; where \; 1 \le i \le n \; and \; 1 \le j \le k$$
$$d_y = |y_i - \mu_{yj}|^2 \; where \; 1 \le i \le n \; and \; 1 \le j \le k$$

From the above formula, distance between the x and y coordinates of each data point to the centroid will be calculated. Each data point is assigned to the nearest cluster by comparing the distances calculated. After assigning all the datapoints to each of the clusters, new centroid of each cluster will be calculated using the below formula.

$$new\ Centroid = \left(\frac{\sum_{i=1}^{c} x_i}{c}, \frac{\sum_{i=1}^{c} y_i}{c} \right)$$

These steps will be repeated by changing the values of K from 2 to 10. Below are the plots of the final clusters after the first initialization with random centroids.



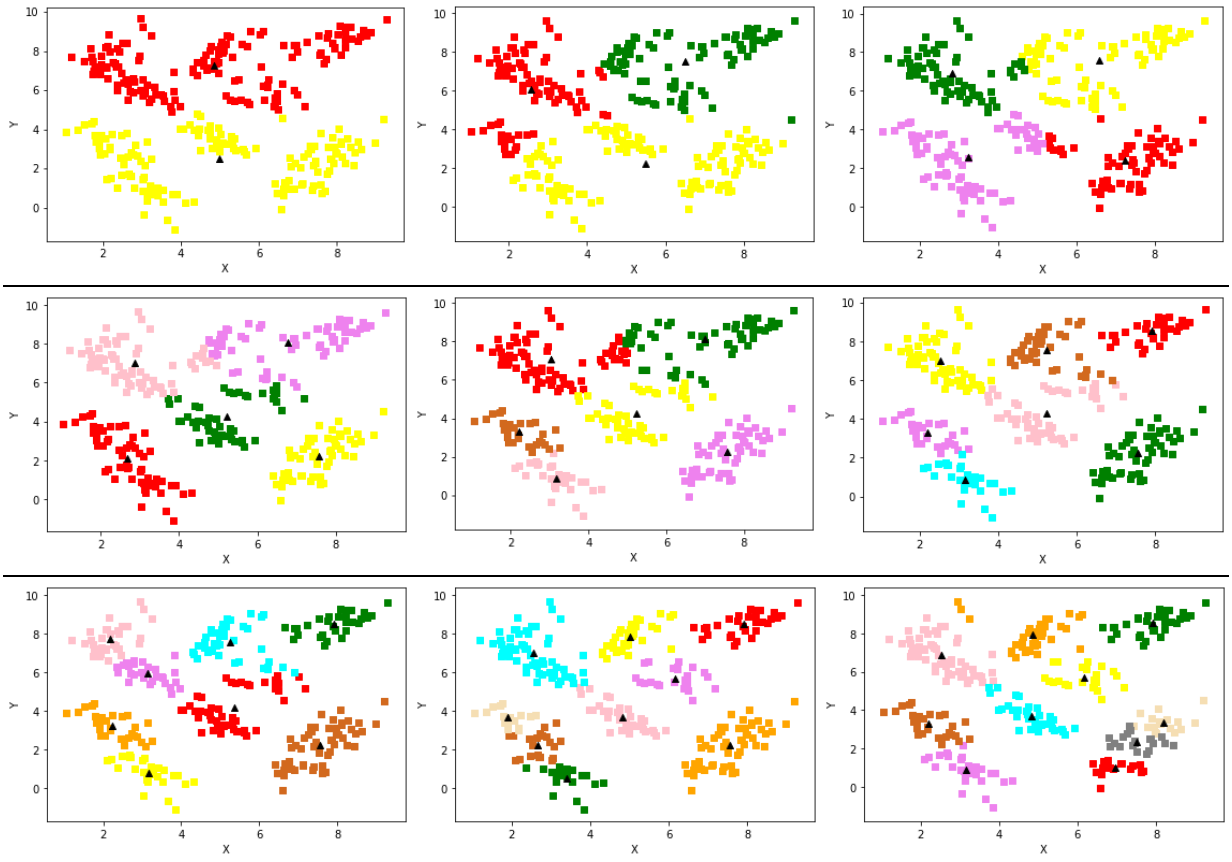
Below is the formula to compute the loss using objective function

$$Loss\ computed\ by\ objective\ function = \sum_{i=1}^{k} \sum_{x \in D_i} |x - \mu_i|^2$$

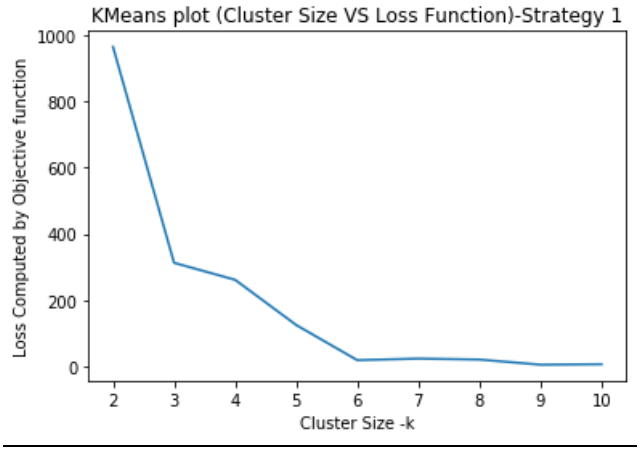
Below is the plot for number of clusters versus the loss computed by the objective function

For k = 4 to 6 and k = 9 there is a slight increase in the loss function whereas from k = 2 to 4 and k = 6 to 9 there is a decrease in the loss function.

Below are the plots for final clusters after second initialization with random centroids.



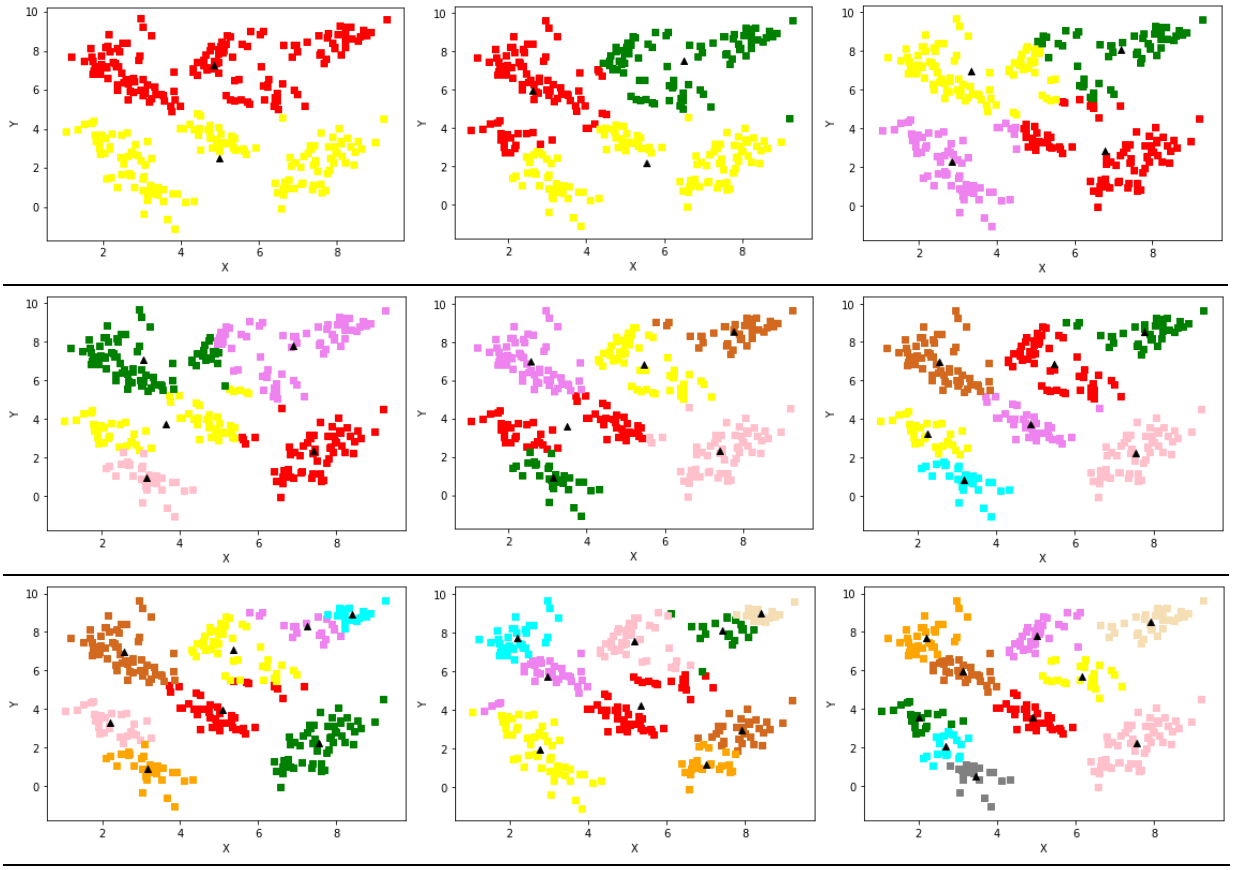
Below is the plot for number of clusters versus the loss computed by the objective function after second initialization



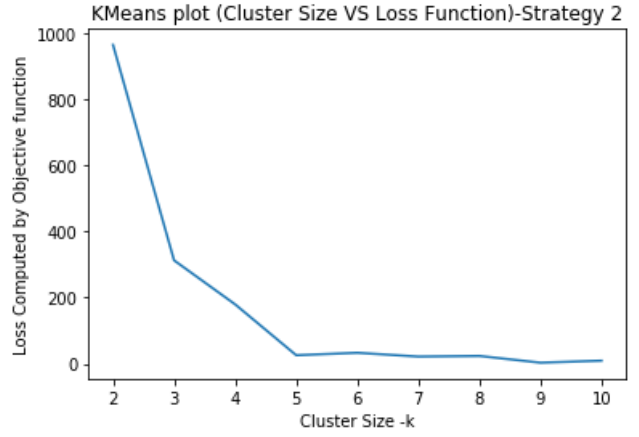
Here there is a slight increase in the graph from k = 6 to 8 and it is decreasing in remaining portions.

Strategy-2 :

In this strategy only the first centroid is taken at random and for the i-th (i>1) center, we have chosen another sample point such that the average distance of this chosen one to all previous (i-1) centers is maximum. We computed all the centroids like this for K clusters (K ranging from 2 to 10). After the first initialization, the plots of the final clusters are as follows.

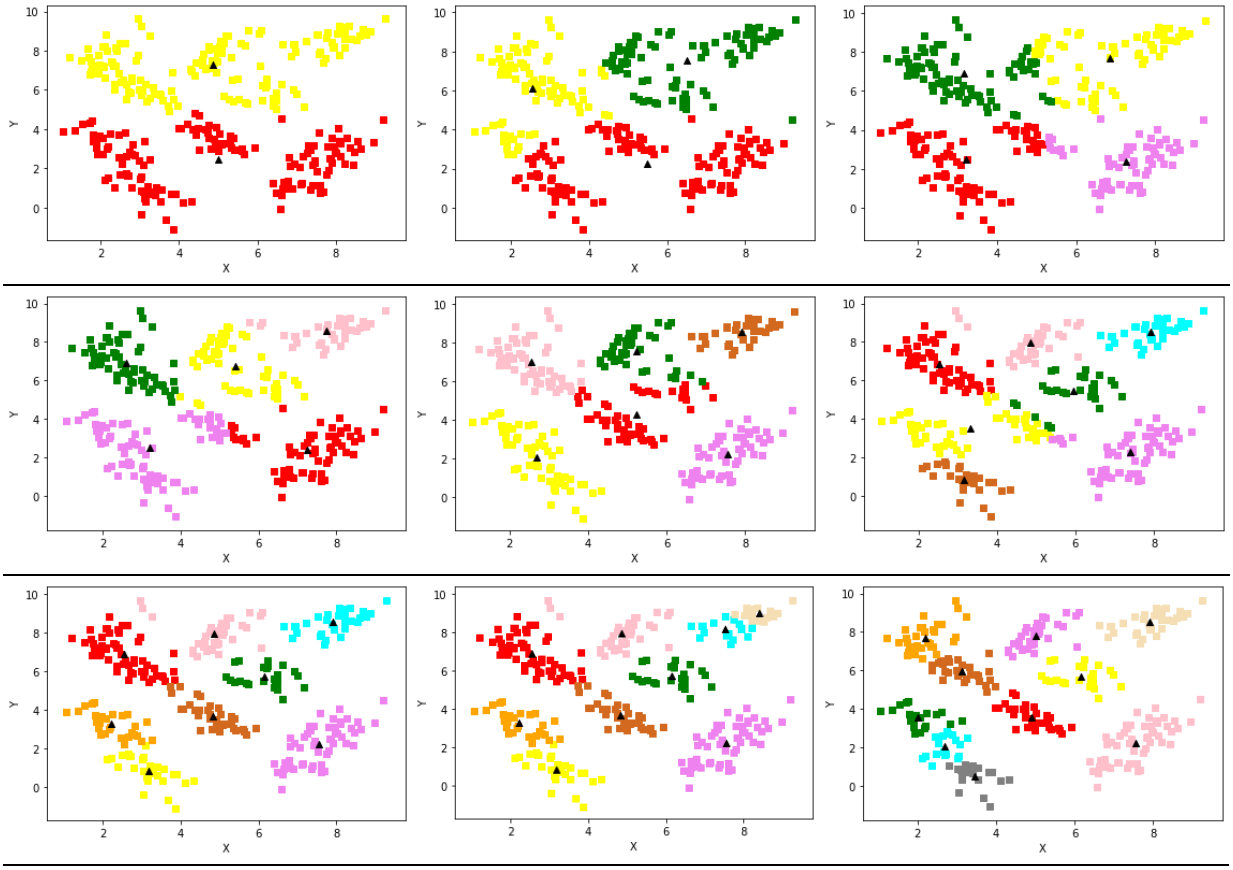


Below is the plot for number of clusters versus the loss computed by the objective function.

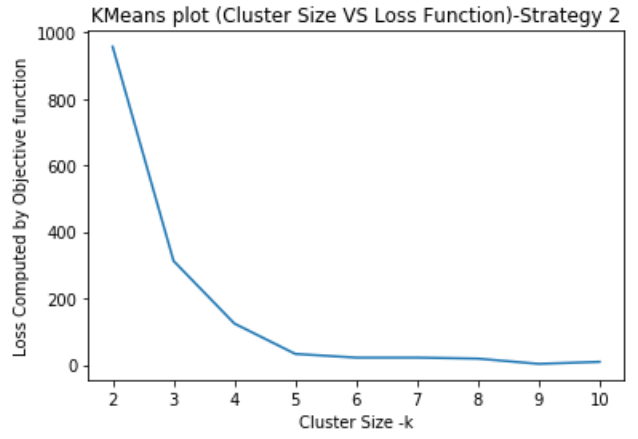


Here apart from k = 5 to 6 the graph is strictly decreasing.

Below are the plots for final clusters after second iteration.



Below is the plot for number of clusters versus the loss computed by the objective function after second iteration.



Apart from k = 6 to 7, the graph is still decreasing.

Conclusion :

Finally, upon comparing both the strategies based on the loss computed by the objective function, we have noticed that Strategy-2 produced better results that Strategy-1. Hence it is advised to initialize only one centroid randomly while implementing K-Means algorithm.