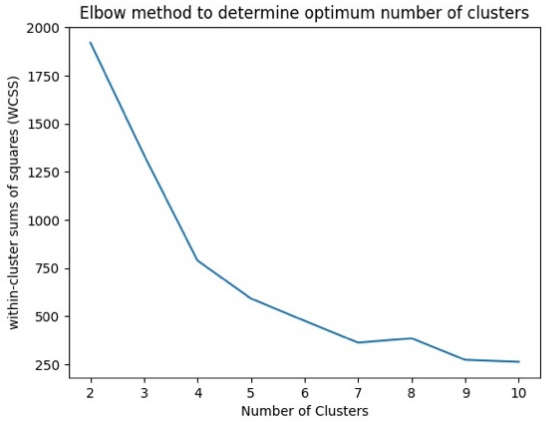
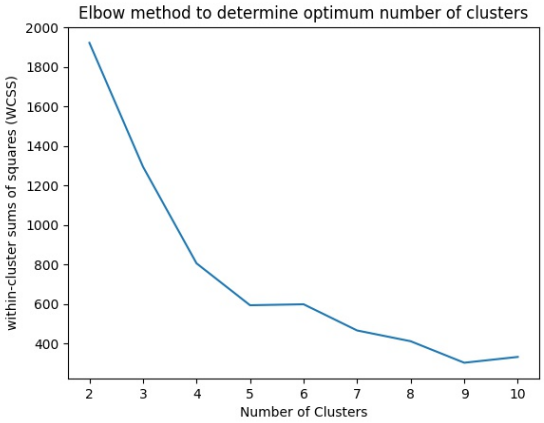
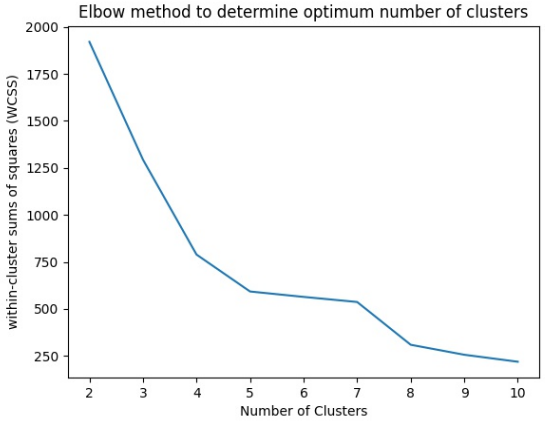
Project-2
K-means Clustering

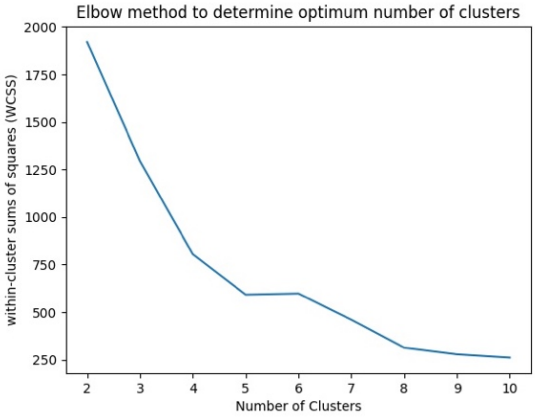
Strategy 1:
randomly pick the initial centers from the given samples. Run 1 and 2.





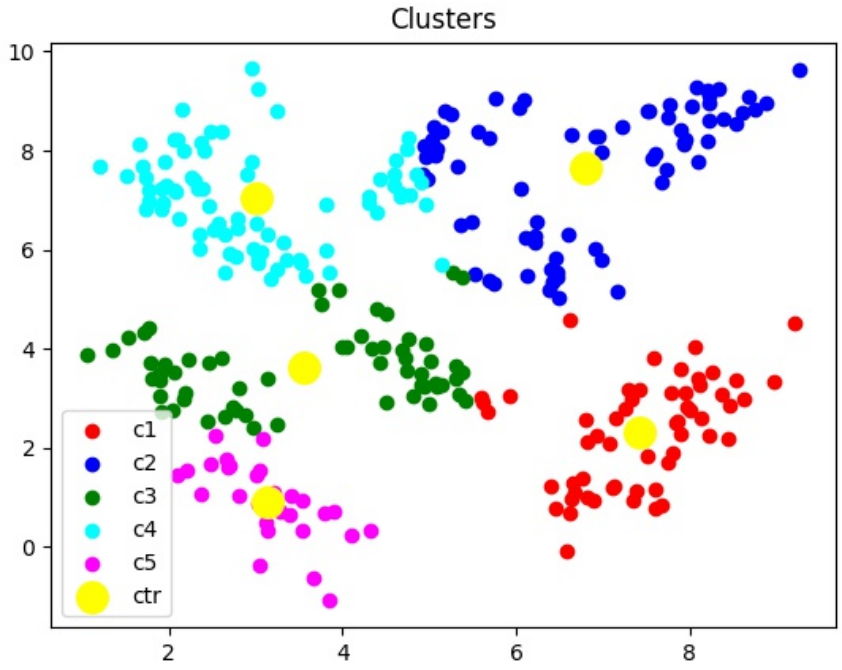
Strategy 2:
pick the first center randomly; for the i^{th} center (i>1), choose a sample (among all possible samples) such that the average distance of this chosen one to all previous (i-1) centers is maximal. Run 1 and 2.



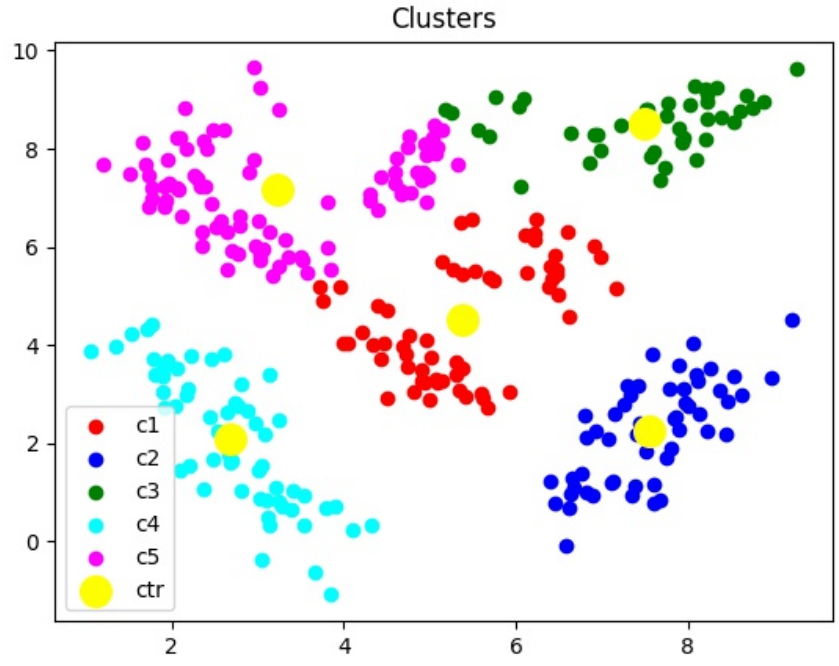


From the graph, I can conclude that **C=5** is the optimal number of clusters for given dataset and the data distribution for **C=5** is in the graph below.

For strategy 1 and C=5:



For strategy 2 and C=5:



Observations:

- At each run, for fixed clusters value K, I am getting different number of inputs classified in the clusters depending upon the initial centroids. For same initialization strategy, I am getting different cluster strength for each cluster in every run.
- For 100 runs for K-means clustering, **64** times the initialization strategy 2 got lower objective function value which suggests that strategy 2 is better for initialization of centroids.