CSE – 575 STATISTICAL MACHINE LEARNING

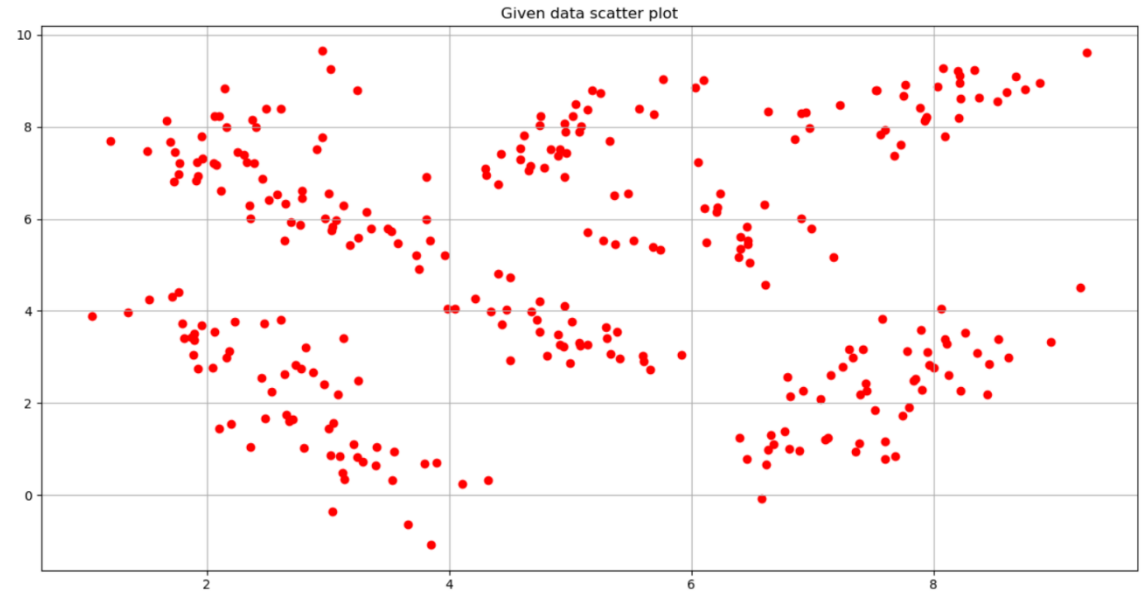
PROJECT 2

NAME: KADIRI GURU PREETAM

ASURITE ID: gkaidri

GIVEN DATA:

The data points given to us consists of 300 points on a X-Y plane.



2 to 10 cluster groups are to be used with two different strategies.

STRATEGY 1:

Randomly pick the initial centres from the given samples.

- k – points (centroids) are picked from the given 300 points randomly.
- Distance between centroids and all the other points is calculated. This calculated distance result in a matrix of shape.

$$distance = \left|\left|x_i - x_j\right|\right|^2 \ where\ 1 \le i \le n; 1 \le j \le n$$

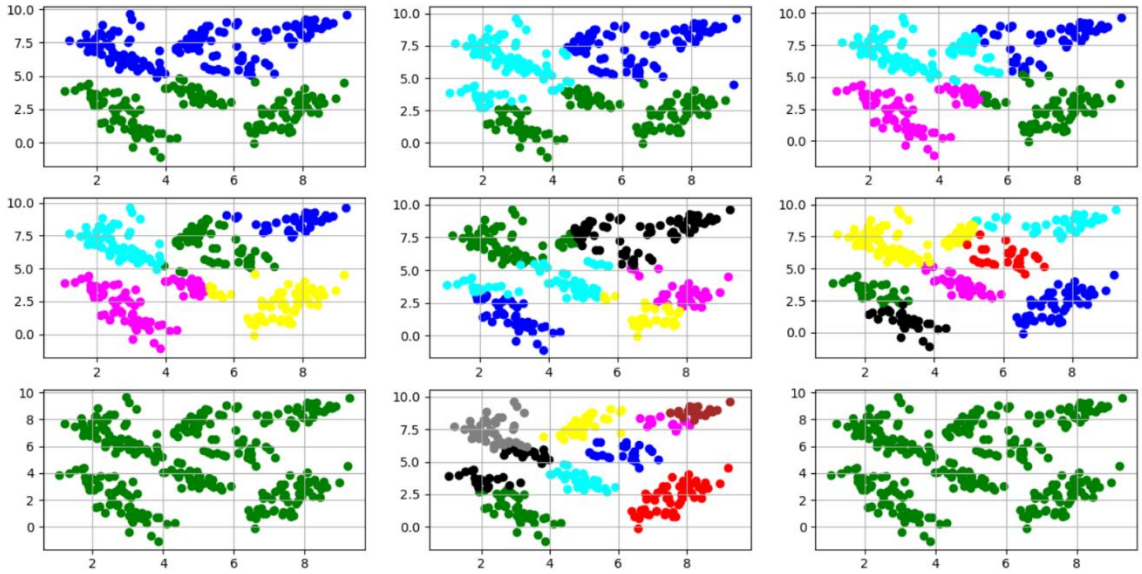
- From each row of the matrix, the minima cluster for that centroid and assign it to the cluster.
- New clusters are calculated, by calculating the mean of the points.

$$centroid = (\frac{\sum x_i}{c}, \sum \frac{y_i}{c})$$

- Next we calculate the objective function.
- ALL OF THE ABOVE STEPS ARE REPEATED 50 TIMES.
- REPEAT THE PROCESS FOR k = 2 to 10.

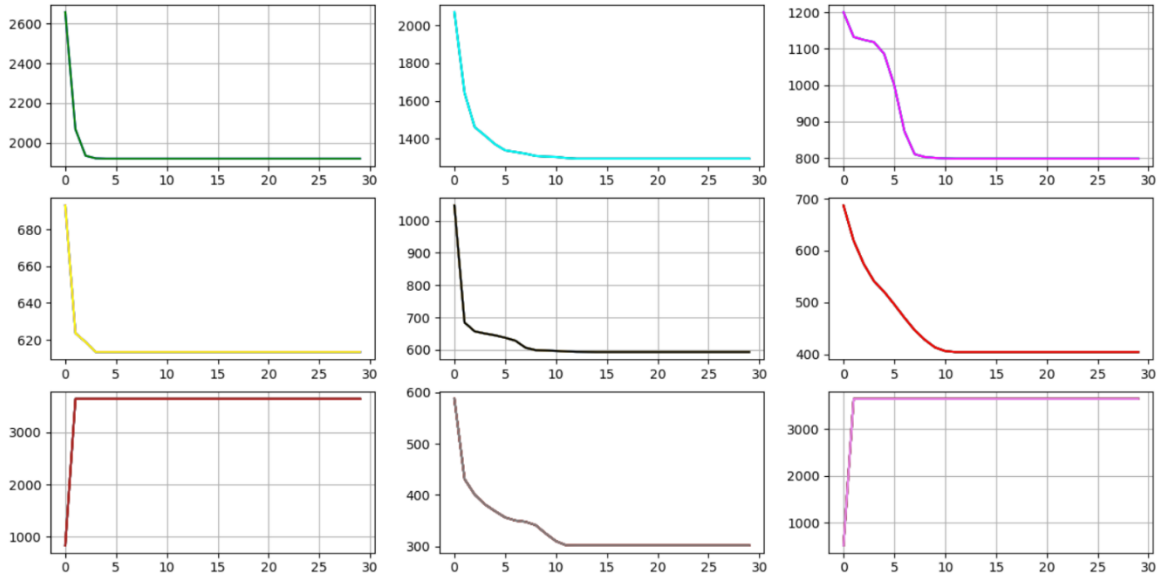
FIRST INITIALIZATION

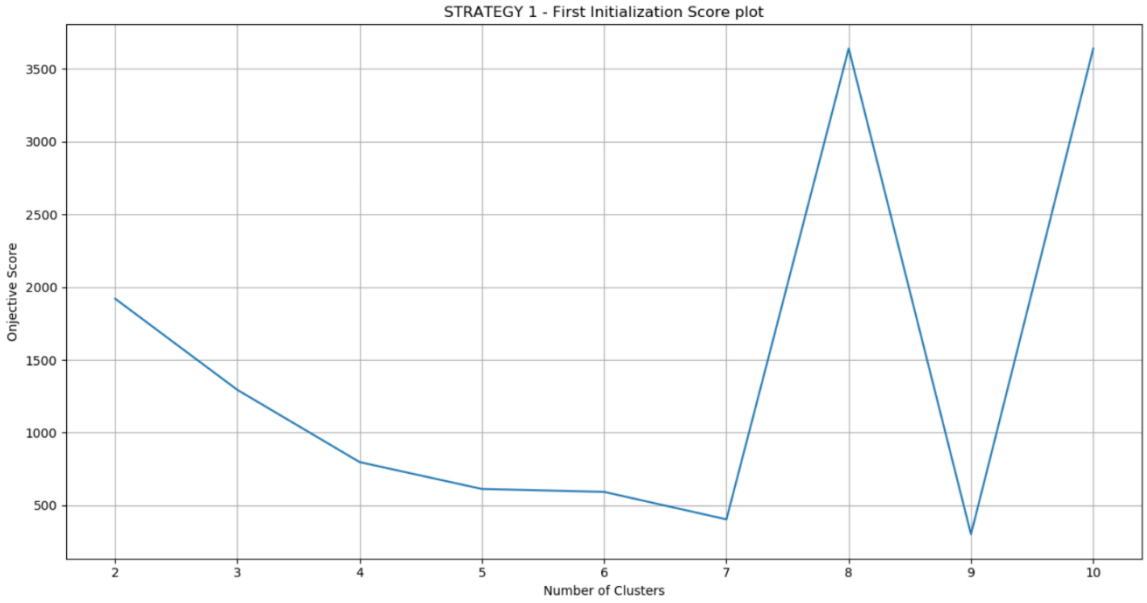
Clustering of random centres.



Objective functions are the clusters converge

This graph clearly depicts the elbow point which avoids overfitting.

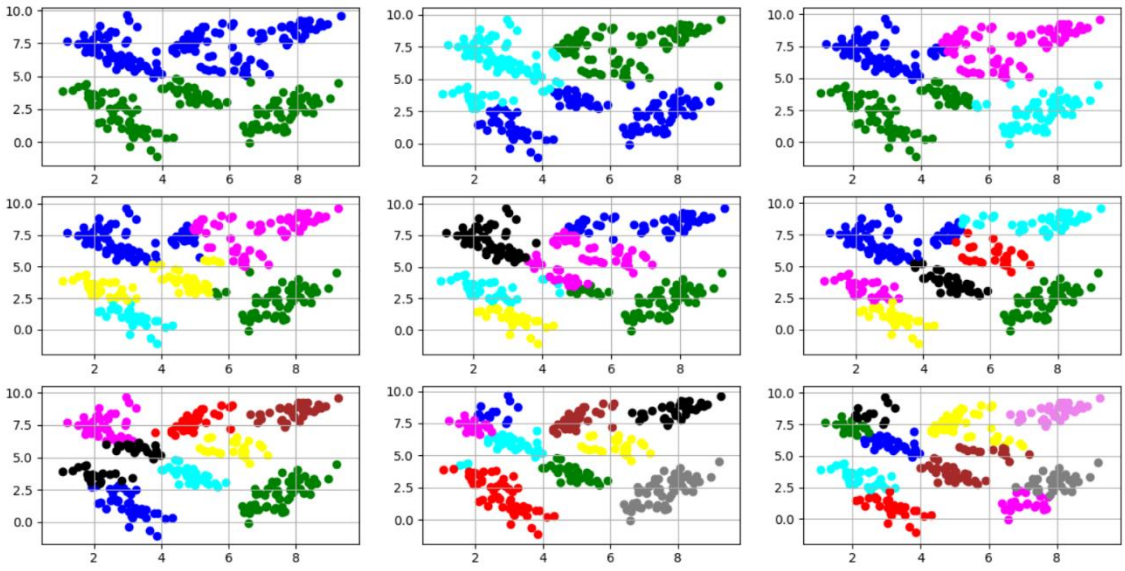


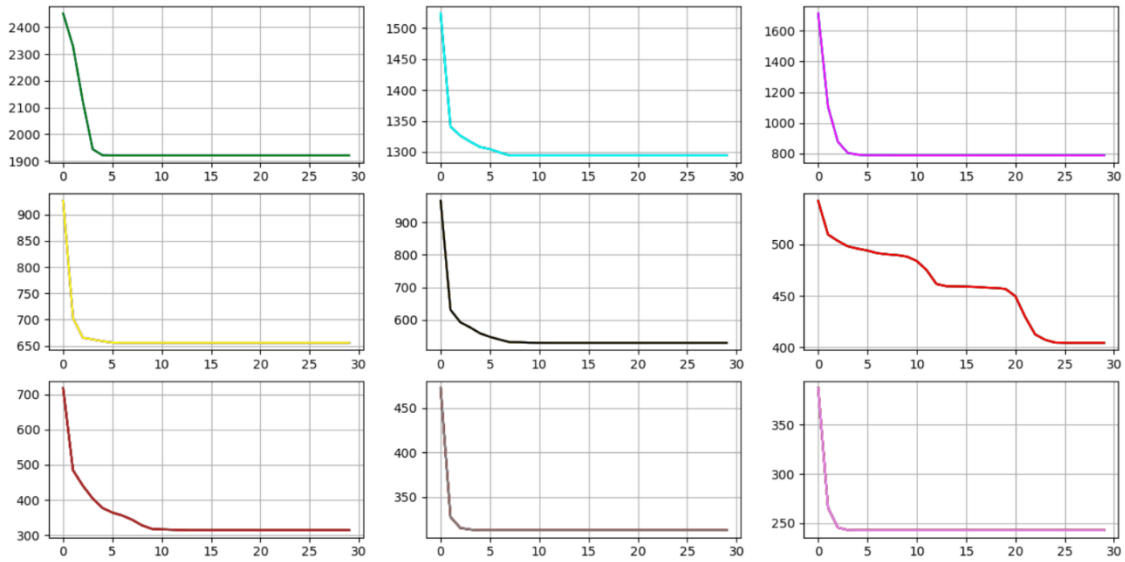


The ABOVE graph depicts how the objective function score behaves based on the number of clusters.

SECOND INITIALIZATION

Clustering of random centres.

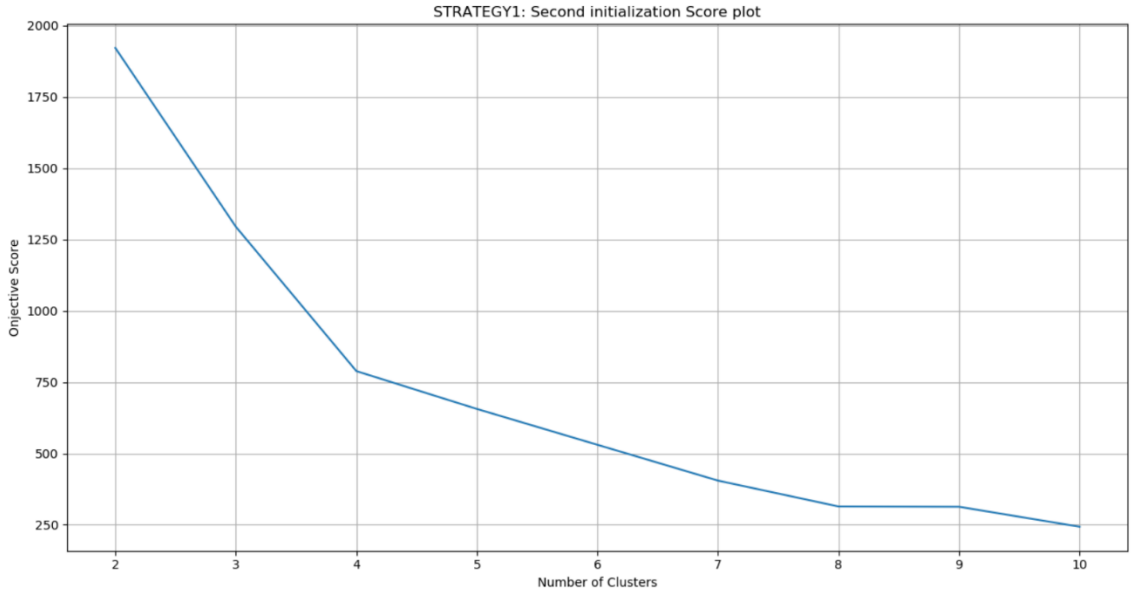


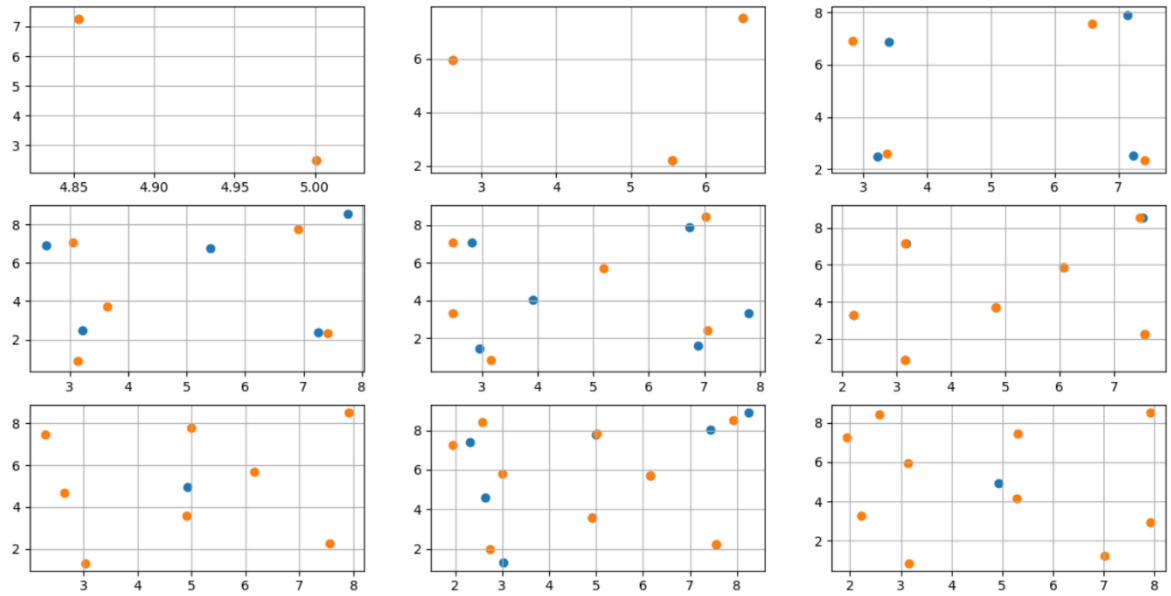


The ABOVE graph show objective functions are the clusters converge.

This graph clearly depicts the elbow point which avoids overfitting.

The graph BELOW depicts how the objective function score behaves based on the number of clusters.





The graph ABOVE shows combined cluster centres for INITIALZATION 1 and 2 for STRATEGY 1.

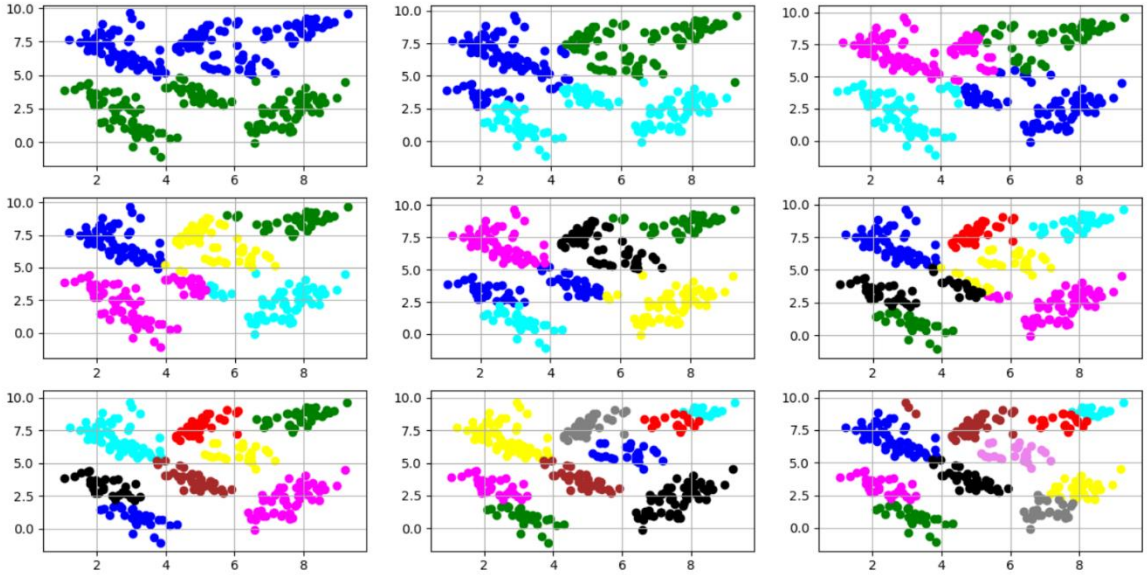
STRATEGY – 2:

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.

- The first point alone is picked up randomly.
- The subsequent points are calculated by taking the point furthest away. This process is repeated for cases with multiple cluster centers, average distance to all previous points is the farthest.
- The algorithm for STRATEGY – 1 is used to calculate the objective function.
- THE ABOVE STEPS ARE REPEATED 50 TIMES.
- REPEAT THE PROCESS FOR k = 2 to 10.

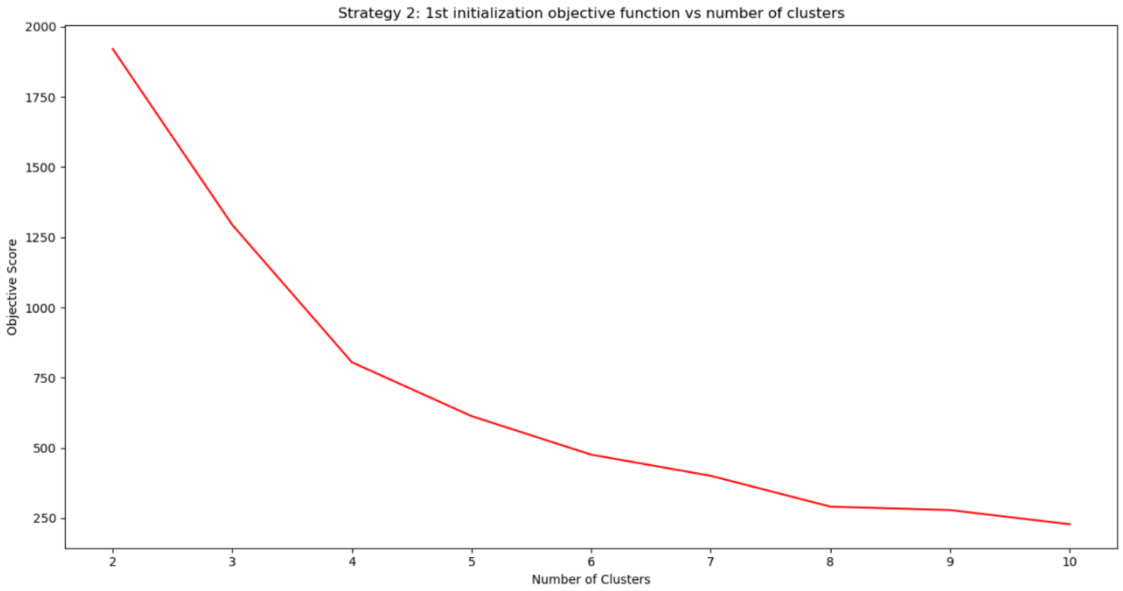
FIRST INTIALIZATION

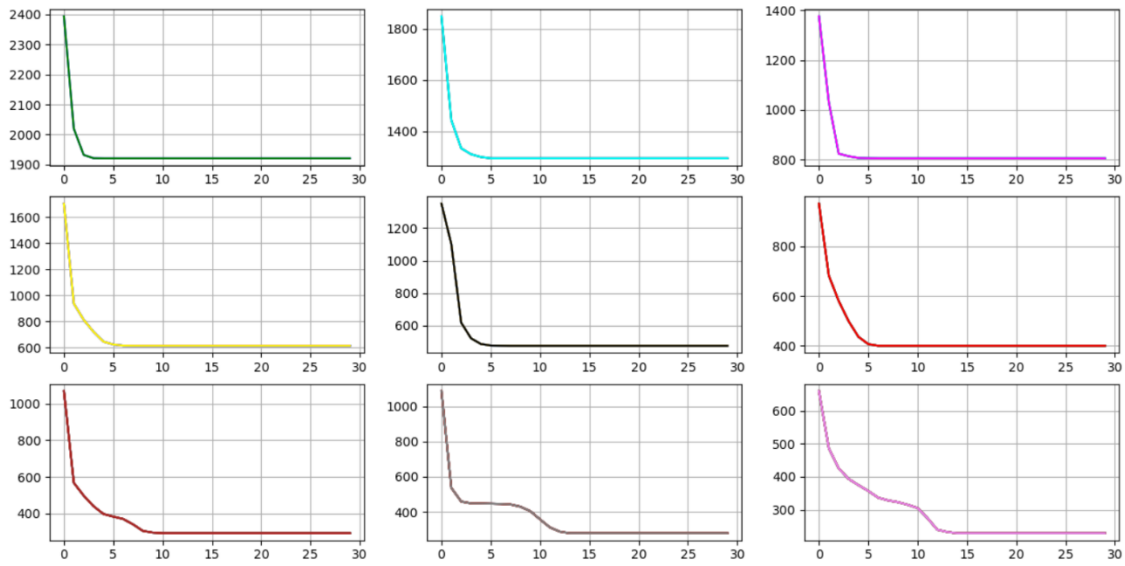
Clustering of random centres.



Objective functions are the clusters converge

This graph clearly depicts the elbow point which avoids overfitting.

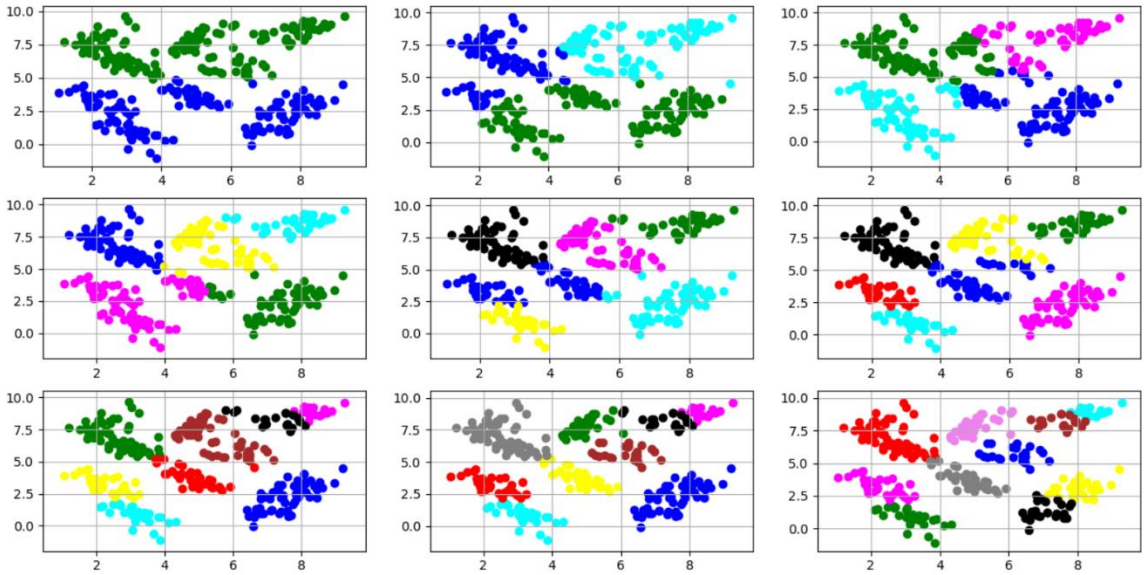


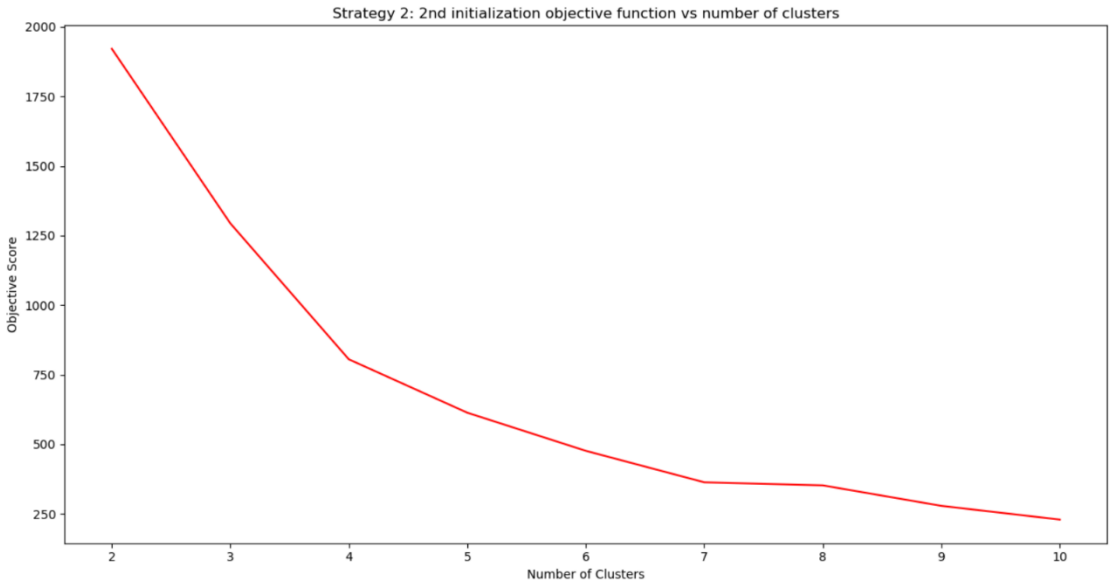


The ABOVE graph depicts how the objective function score behaves based on the number of clusters.

SECOND INITIALIZATION

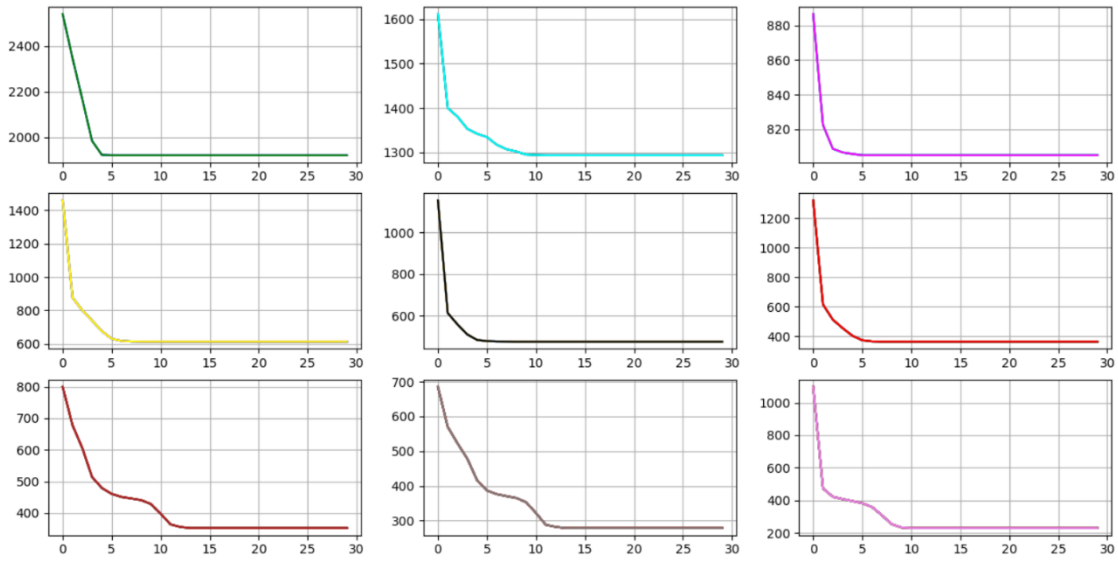
Clustering of random centres.



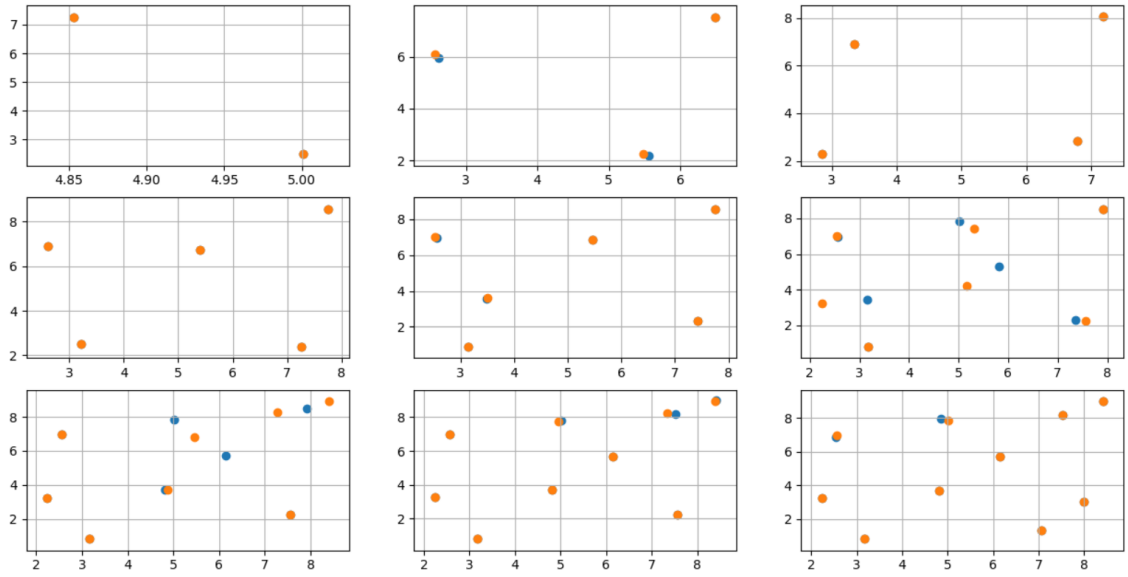


The ABOVE graph show the Objective functions are the clusters converge

This graph clearly depicts the elbow point which avoids overfitting.



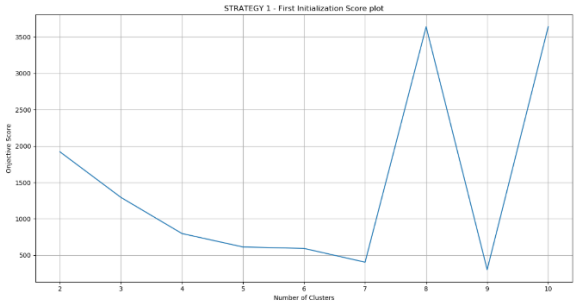
The ABOVE graph depicts how the objective function score behaves based on the number of clusters.



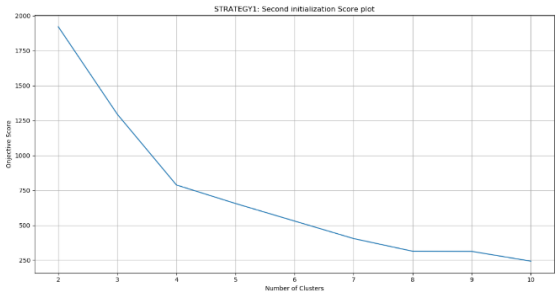
The graph ABOVE shows combined cluster centres for INITIALZATION 1 and 2 of STRATEGY 2.

STRATEGY 1

INITILIZATION 1

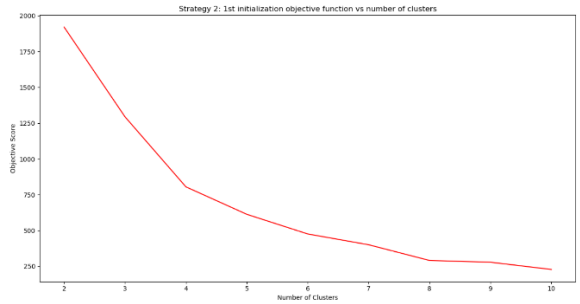


INITLIZATION 2

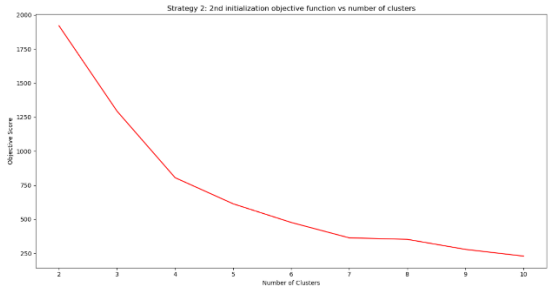


STRATEGY 2

INITLIZATION 1



INITLIZATION 2



The spikes for Initiation 1 in Strategy 1 suggest that the initial chosen random points were too far away from the data points and the 50 iterations weren't enough to bring it with proper range of the data points.

We can see that that for STRATEGY – 1, we get elbow points at

K = 4 for FIRST INITLIZATION.

K = 4 for SECOND INITLIZATION.

Which shows the optimal number of clusters to have without overfitting.

We can see that that for STRATEGY – 2, we get elbow points at

K = 4 for FIRST INITLIZATION.

K = 4 for SECOND INITLIZATION.

Which shows the optimal number of clusters to have without overfitting.