Project 2: Unsupervised Learning (K-means)

1. Introduction:

The main objective of this project is to perform clustering on a set of data points provided which contains a set of 2-D points using k means clustering algorithm. There are two different implementations using two different strategies for choosing the initial cluster centers.

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. The main objective of the k means algorithm is to group similar data points together and find the underlying patterns.

2. Dataset:

In this project, we are using the “All Samples” dataset which is a collection of 2-dimensional dataset. There are 300 rows and 2 columns. Here is the scatter plot which shows the distribution of the given dataset.

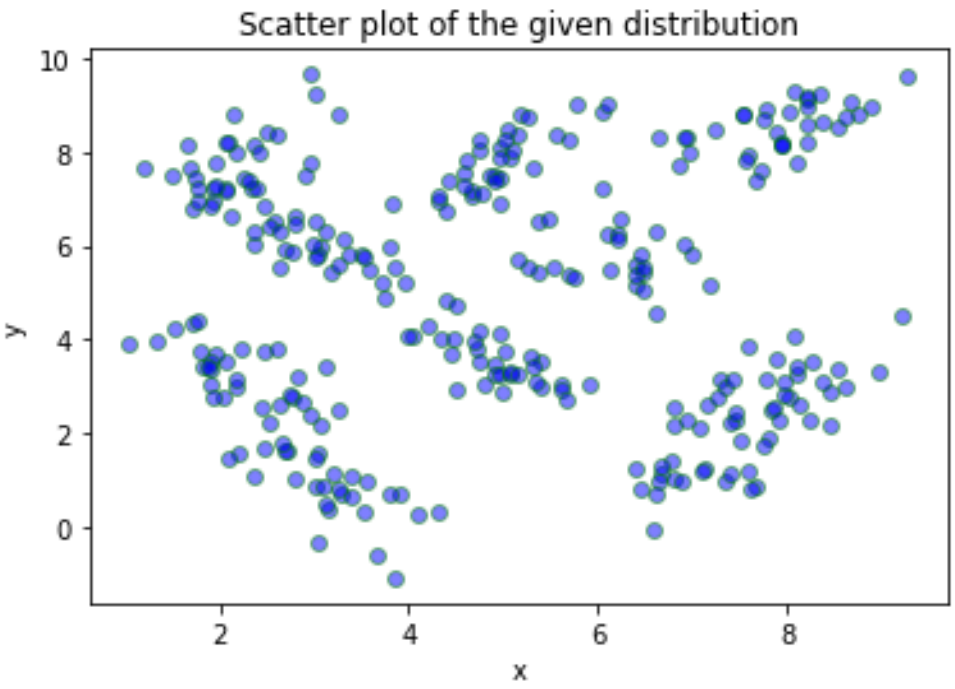


Fig 1. Scatter plot for the entire

3. Implementation of K-means algorithm:

K means is one of the most popular clustering algorithms which is used widely by most of the Machine Learning Engineers. In order to cluster or learn the data without any labels, k means initially randomly picks k different values and assigns them as the initial centroids where each centroid signifies each cluster. Once the initialization is done, it assigns each of the remaining points to the closest clusters which is represented by the centroids selected.

Once each of the points in the 2D dataset is assigned to one of the clusters, the new centroids are calculated by calculating the mean of each clusters. Once the new centroids are calculated, the same

assignment step as defined in the above steps is carried on and the same sequence of steps carry on until the centroids converge or in other words when the centroids don't change their position. In the given project the above steps are repeated for different values of k from 2 to 10. Once the final centroids are calculated the Objective function value vs number of clusters k graph is plotted to check the optimal value for k.

As suggested in the project, there are two strategy to implement the k means.

Implementation 1:

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

```
centroids = {
    i+1 : [x1[0][np.random.randint(0,299)],
x1[1][np.random.randint(0,299)]]
    for i in range(k)
   }
```

```
the initial centroids are
{1: [8.12343077727999, 3.0514253903709436], 2: [6.4842301079969005, 4.9007
01139147971]}
```

Results for the Strategy 1:

Iteration1:

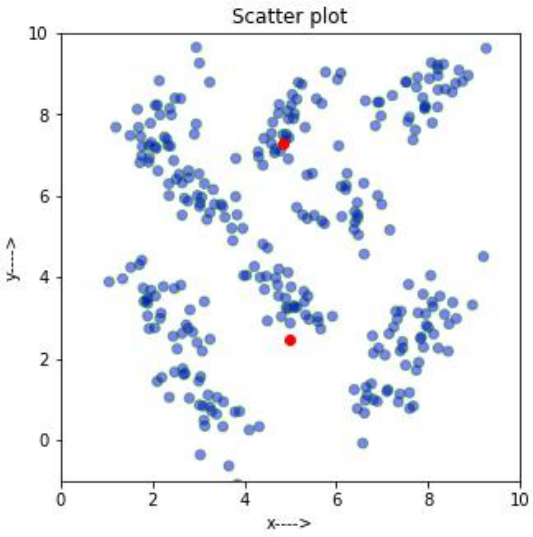


Fig 2. Scatter plot for k=2

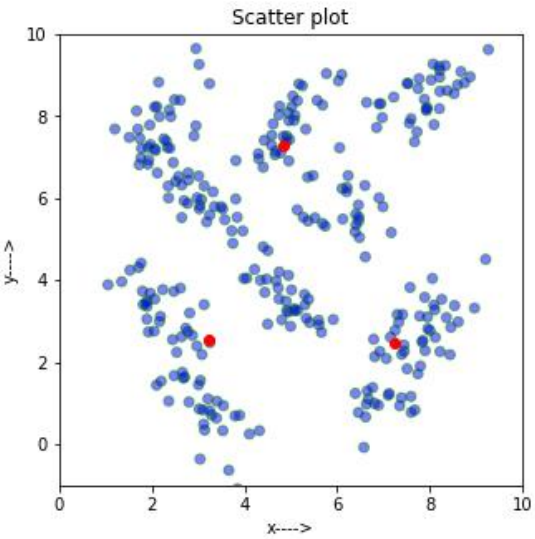


Fig 3. Scatter plot for k=3

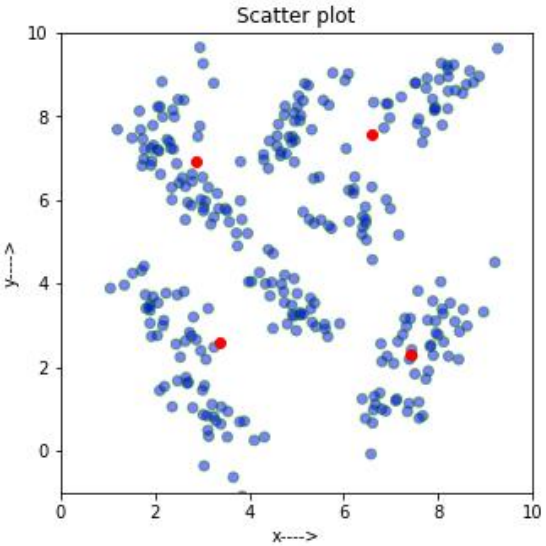


Fig 4. Scatter plot for k=4

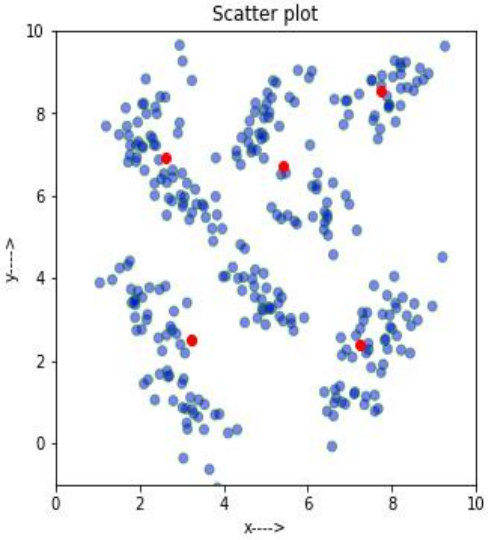


Fig 5. Scatter plot for k=5

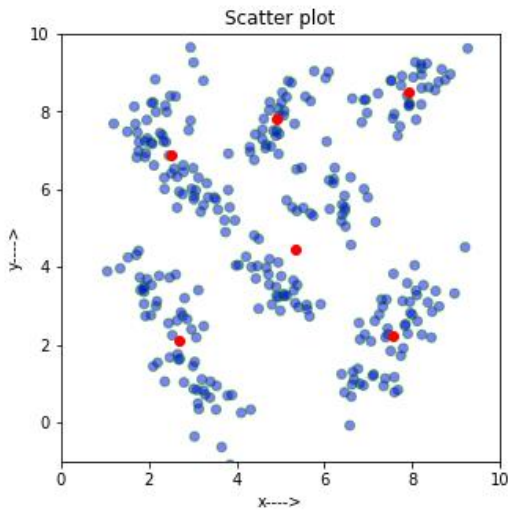


Fig 6. Scatter plot for k=6

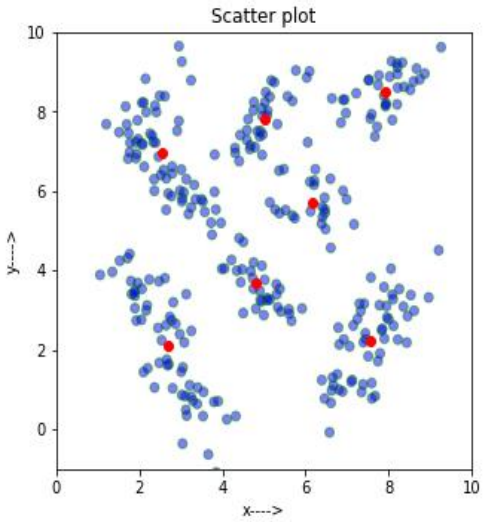


Fig 7. Scatter plot for k=7

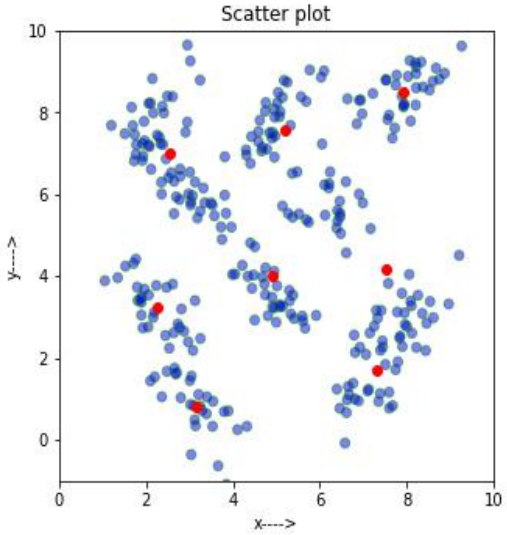


Fig 8. Scatter plot for k=8

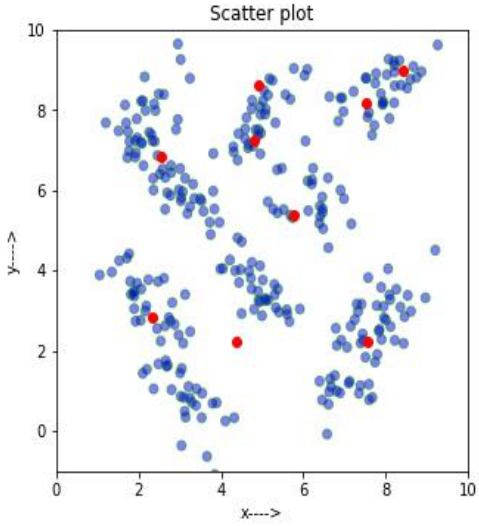


Fig 9. Scatter plot for k=9

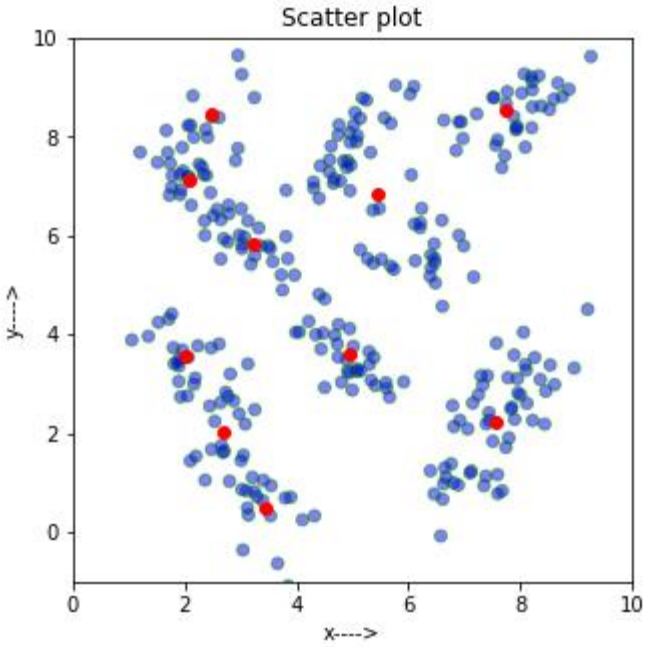


Fig 10. Scatter plot for k=10

Iteration2:

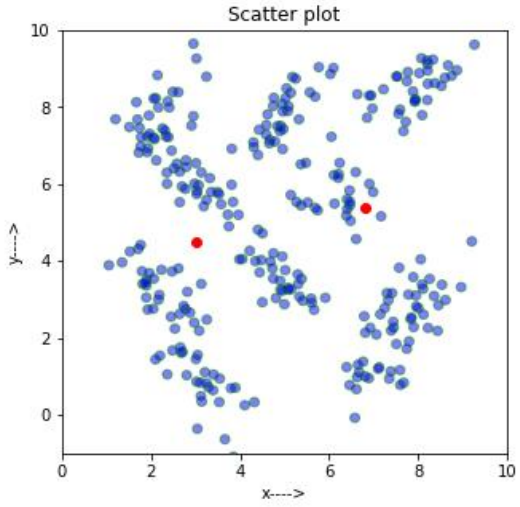


Fig 11. Scatter plot for k=2

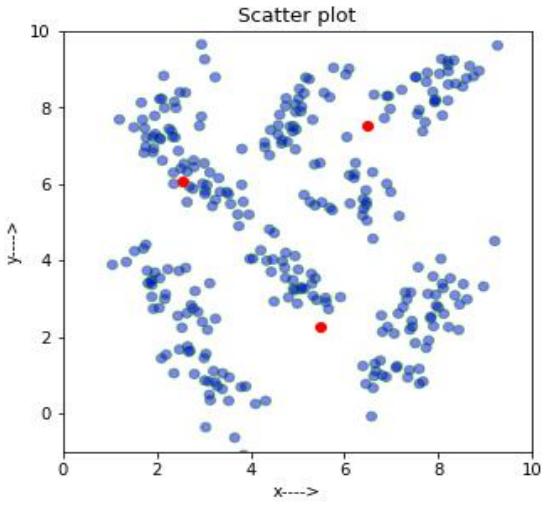


Fig 12. Scatter plot for k=3

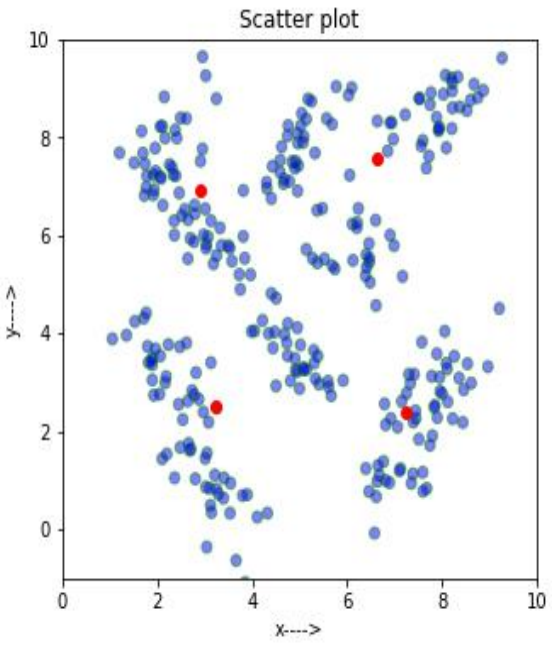


Fig 13. Scatter plot for k=4

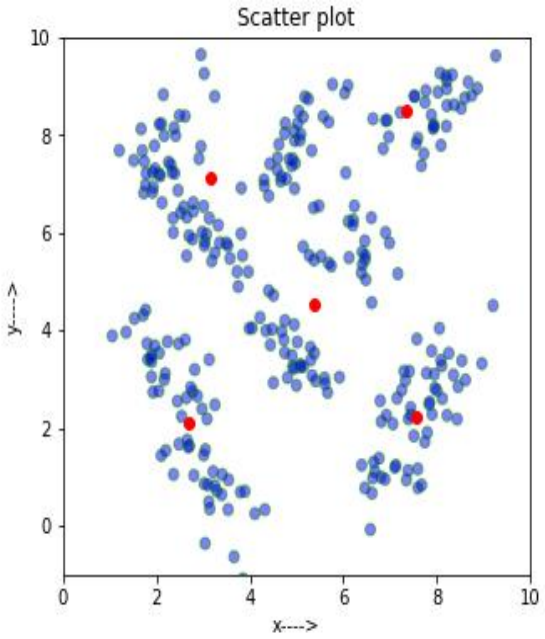


Fig 14. Scatter plot for k=5

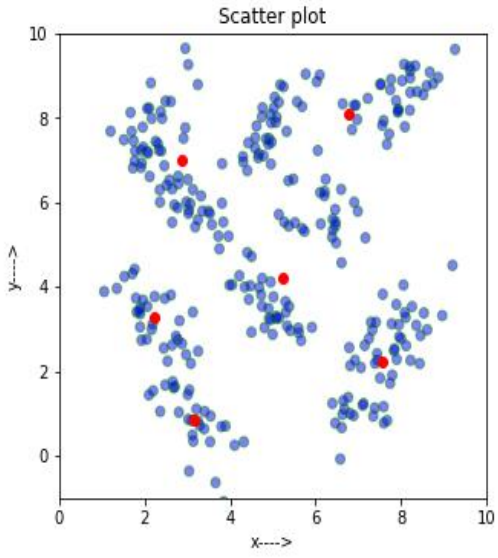


Fig 15. Scatter plot for k=6

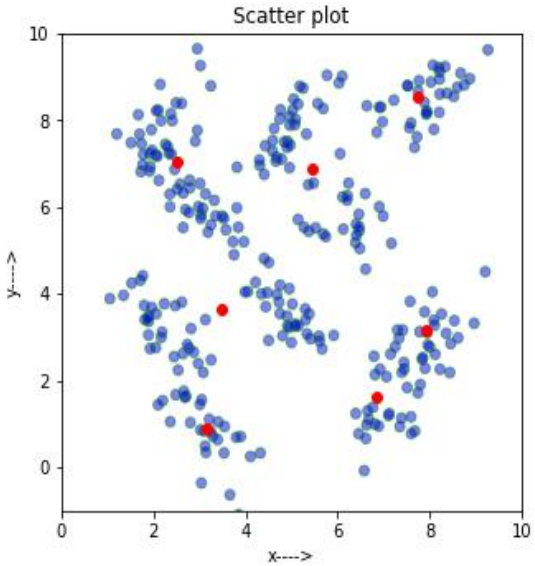


Fig 16. Scatter plot for k=7

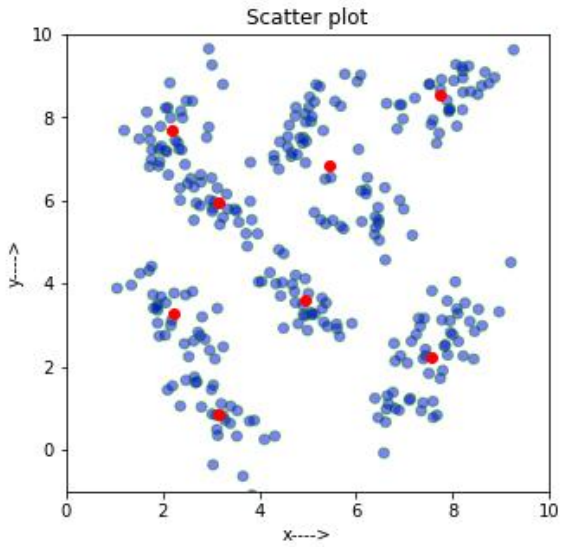


Fig 17. Scatter plot for k=8

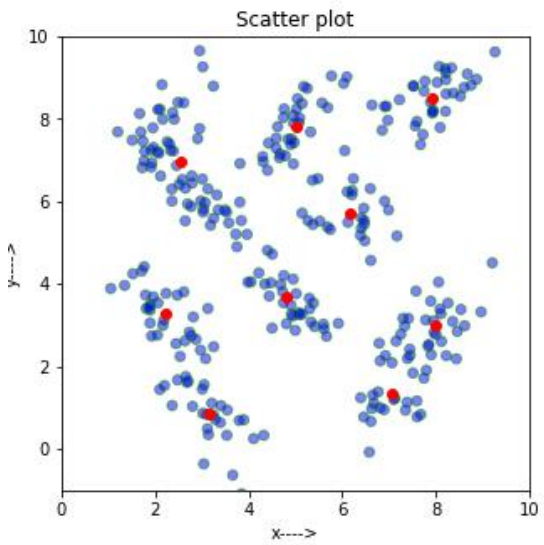


Fig 18. Scatter plot for k=9

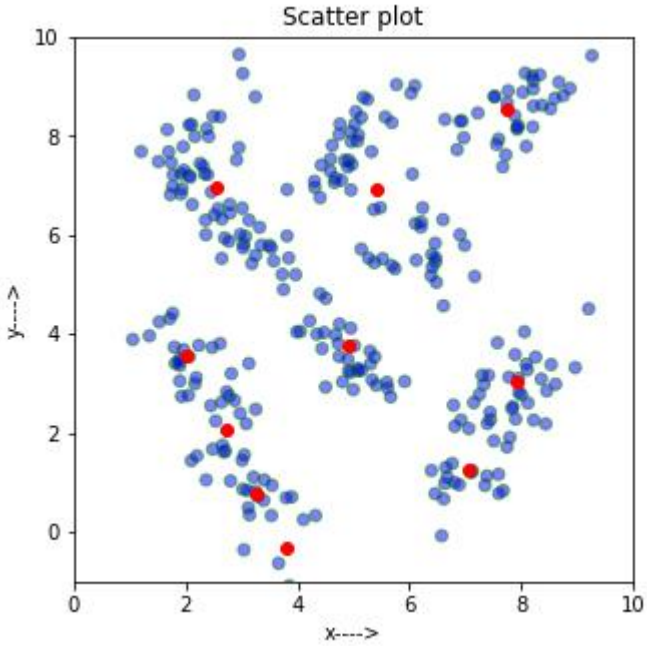
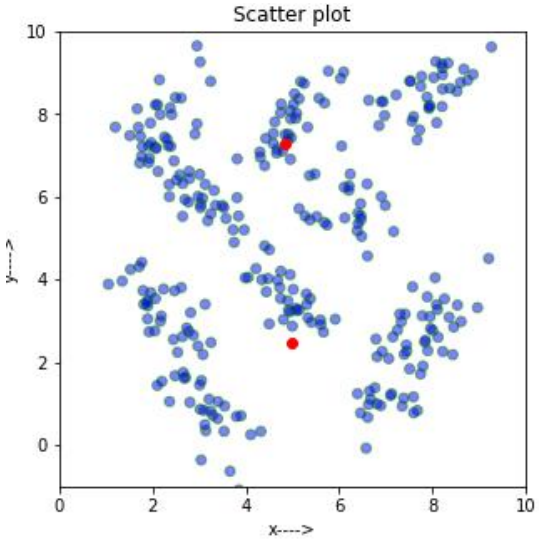


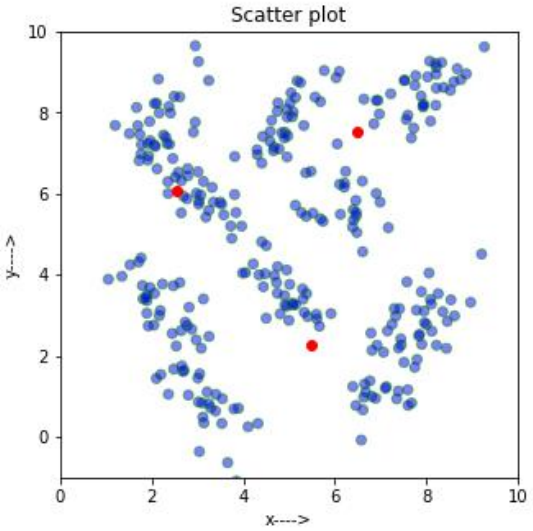
Fig 19. Scatter plot for k=10

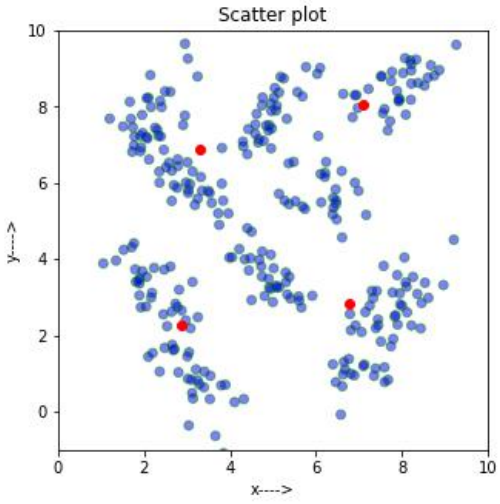
Implementation 2:

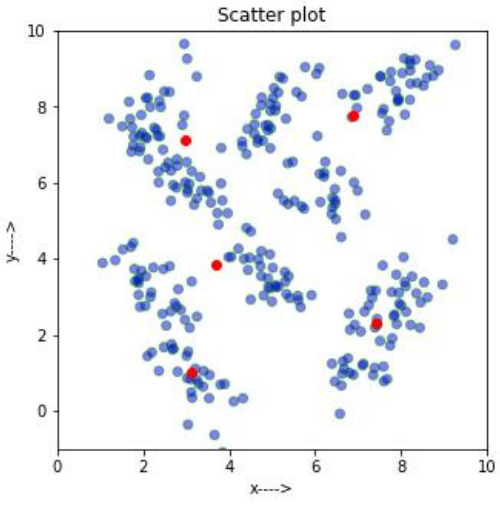
Strategy 2: Pick the first center randomly; for the i-th center (i>1), choose a sample (among all possible samples) such that the average distance of this chosen one to all previous (i-1) centers is maximal.

Iteration 1:









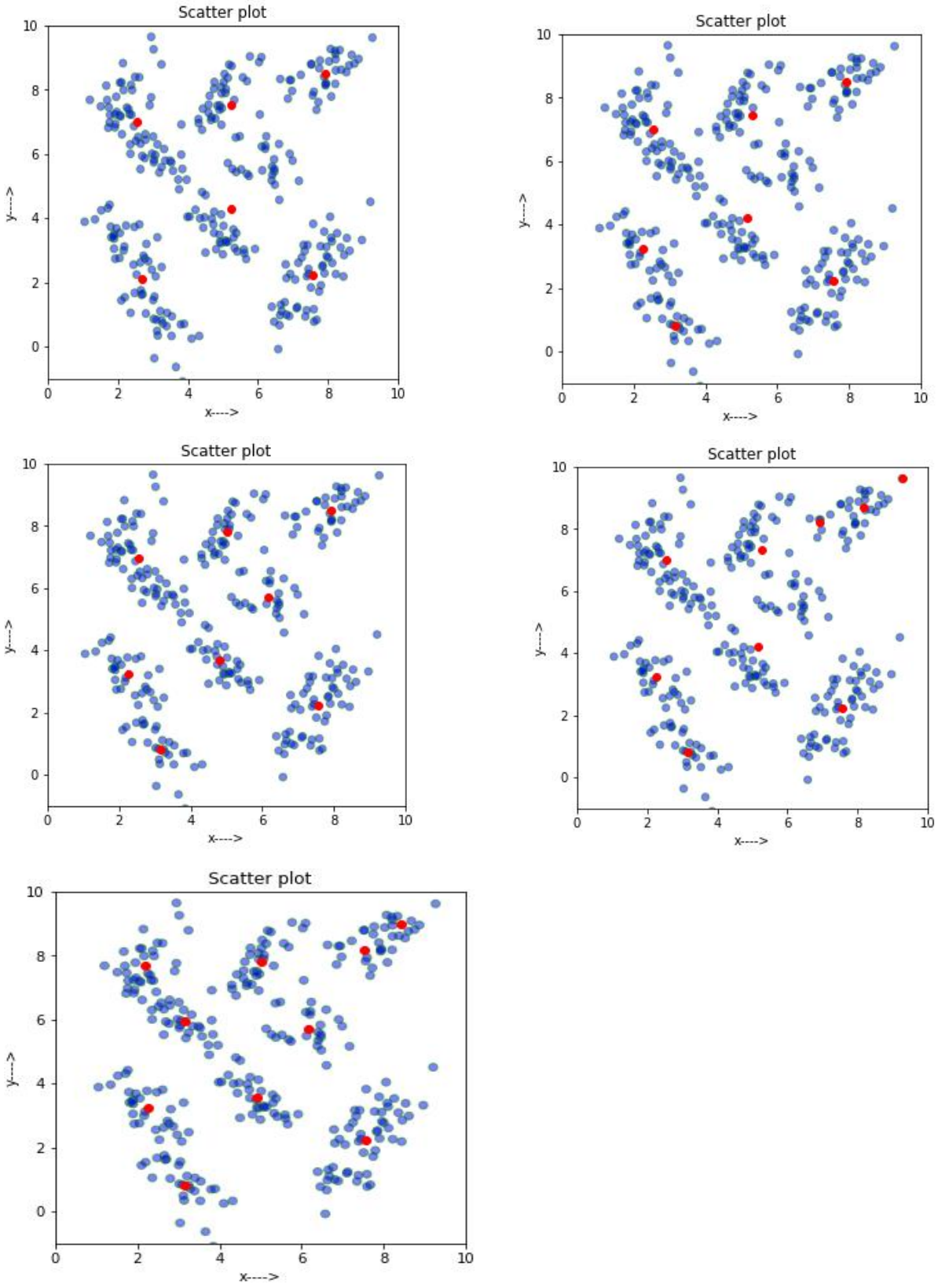
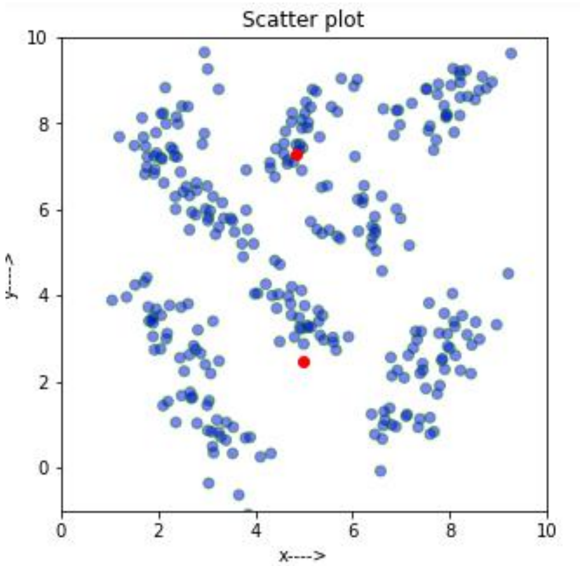
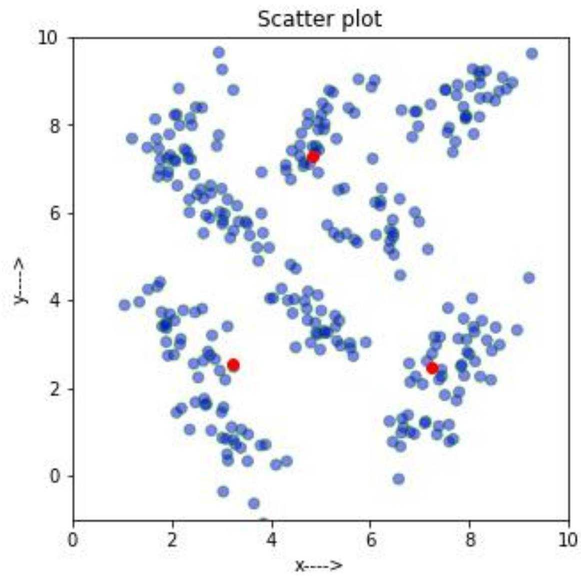


Fig 20. Scatter plots for k value from (2-10) – Strategy 2

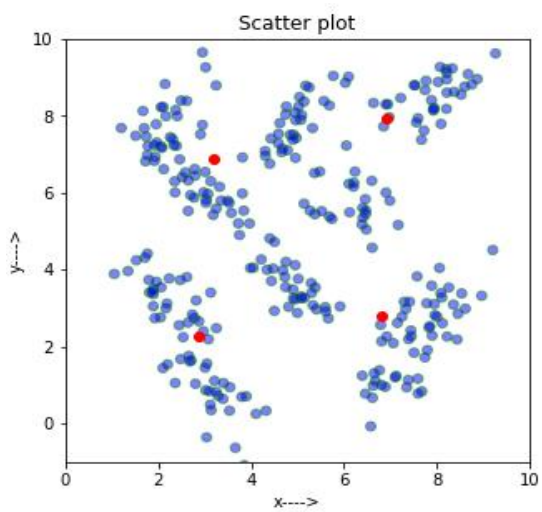
Iteration 2:



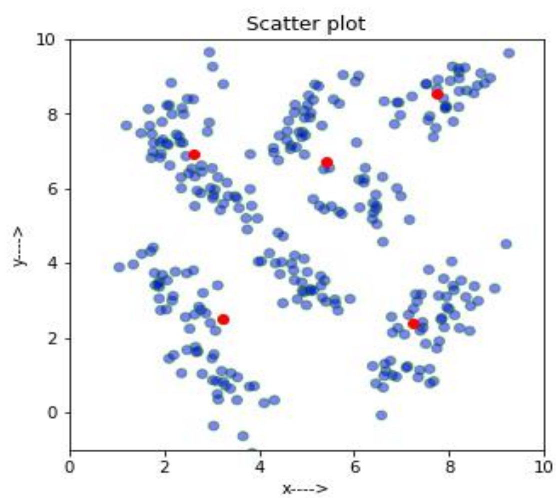
Scatter plot
y---->
x---->



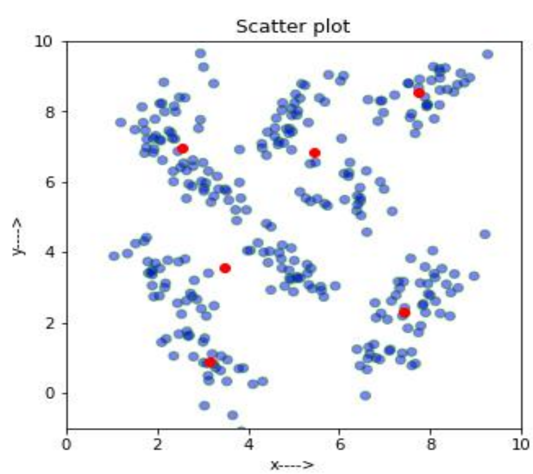
Scatter plot
y---->
x---->



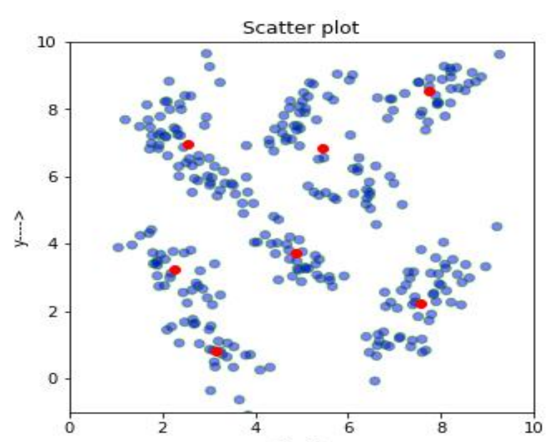
Scatter plot
y---->
x---->



Scatter plot
y---->
x---->



Scatter plot
y---->
x---->



Scatter plot
y---->
x---->

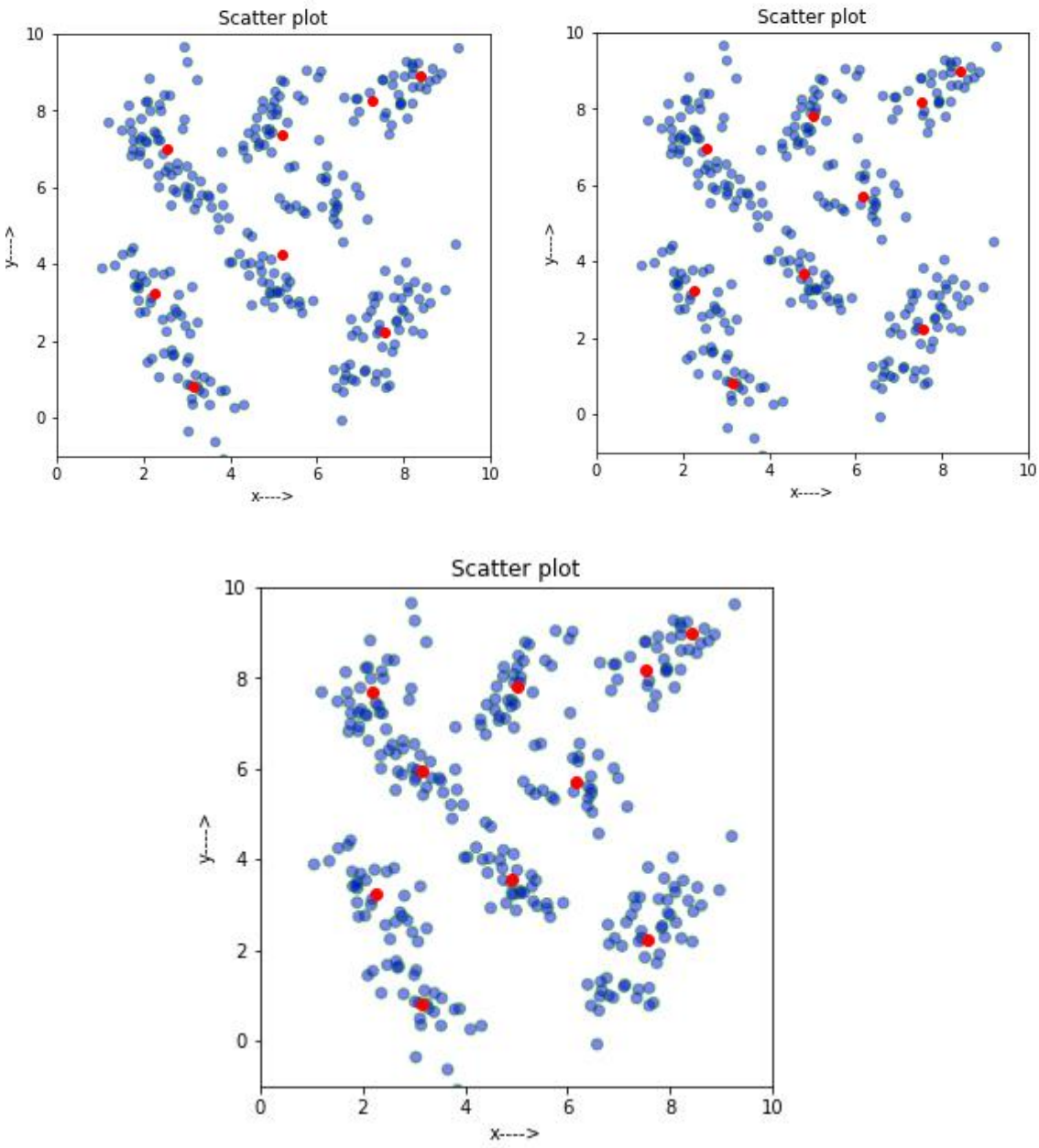


Fig 20. Scatter plots for k value from (2-10) – Strategy 2

Objective Function:

The main objective of the objective function is to find the optimal value of k in the k means algorithm. Find the variance of the data points with regard to the final centroids calculated and plot a curve of Objective fun ction vs number of clusters k. As the value of k increases the variance

decreases and one point the variance doesn't decrease to a major exte nt.
(Referring to the course notes: When clustering the samples into k clusters/sets Di, with respective center/mean vectors $\mu_1, \mu_2, \ldots \mu_k$, the objective function is defined as)

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in D_i} ||\mathbf{x} - \boldsymbol{\mu}_i||^2 \ )$$

Strategy 1:

Iteration 1:

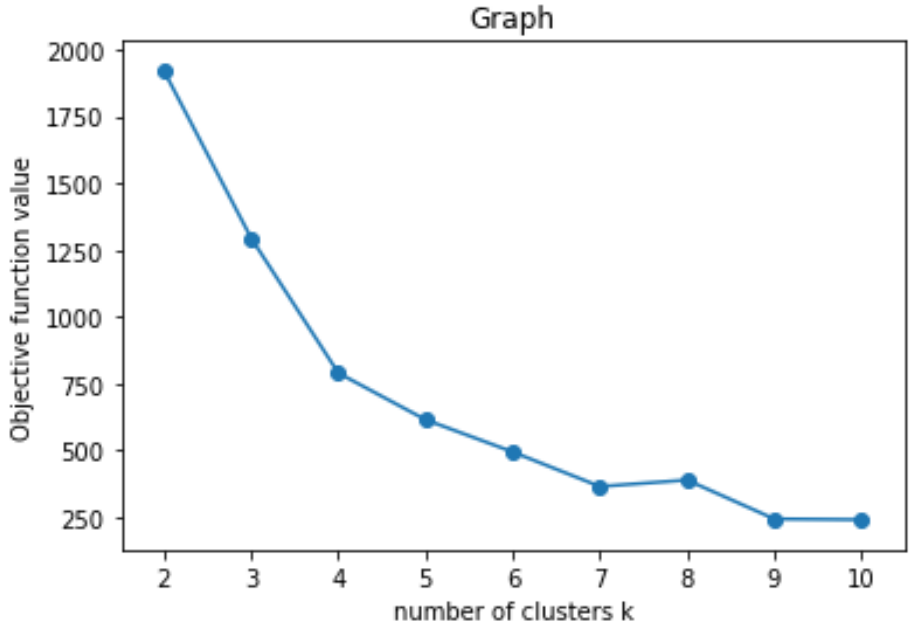


Fig 21. Objective Funtion vs K value Graph (Strategy 1)

Iteration 2:

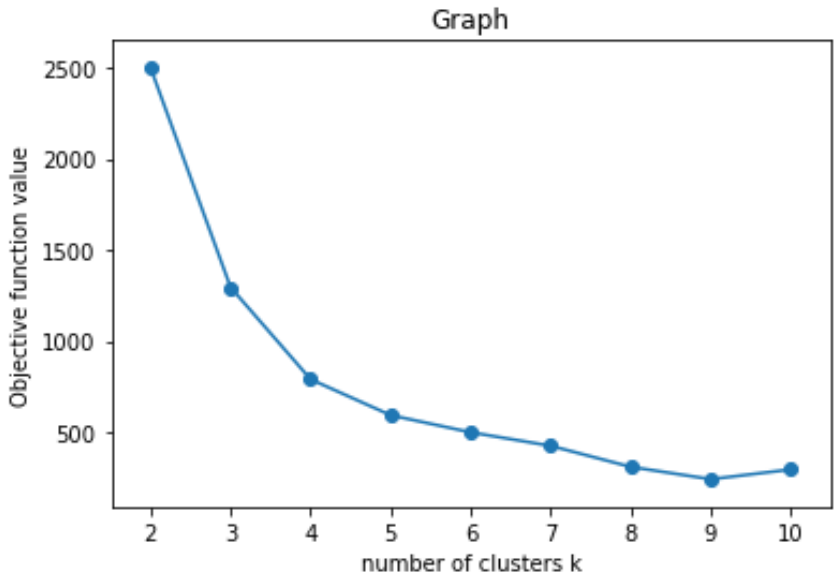


Fig 22. Objective Funtion vs K value Graph (Strategy 1)

Strategy 2:

Iteration 1:

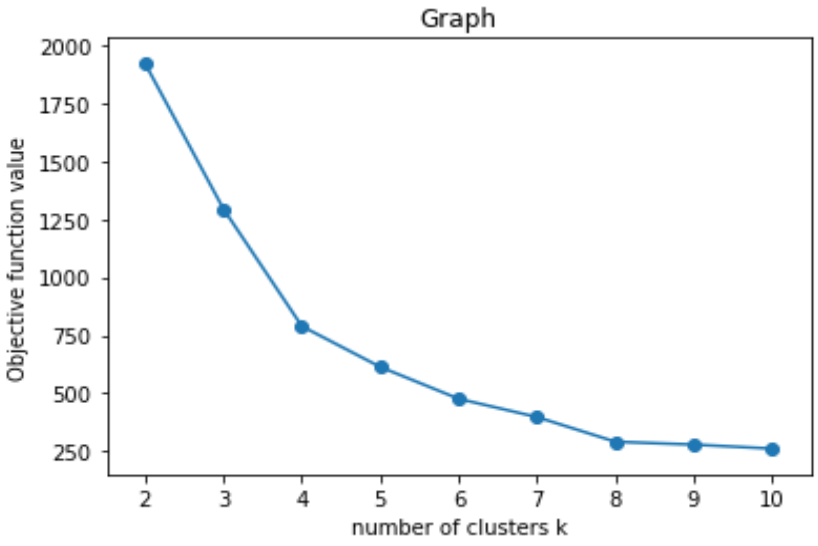


Fig 23. Objective Funtion vs K value Graph (Strategy 2)

Iteration 2:

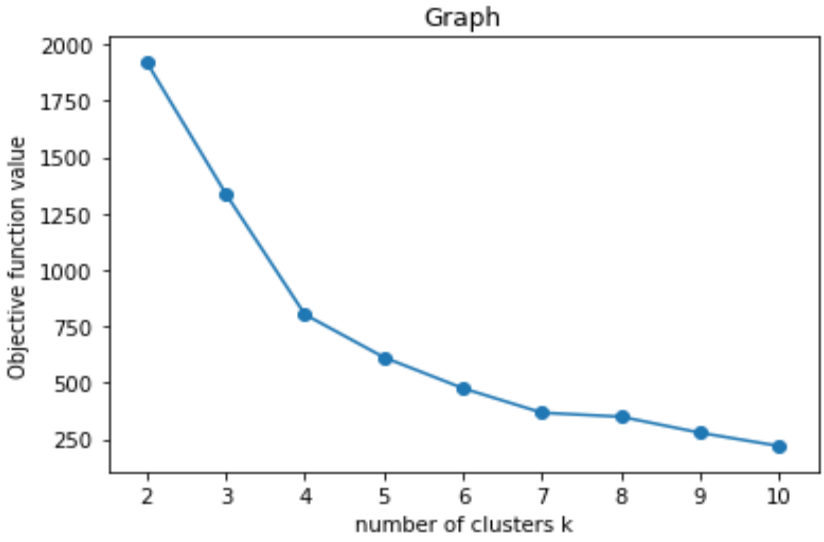


Fig 21. Objective Funtion vs K value Graph (Strategy 2)

Conclusion:

K - means seems to be one of the best and popular clustering algorithm s. Looking at the Elbow plot it's very evident that the number of optimal clusters is 4 in both the stratagies as the elbow is formed at k=4.