Assignment 3 – k-Means Clustering

Elif Erden
Student ID: 041503038
Date: 13.03.2019

COMP 462
Introduction to Machine Learning

1. Algorithm Explanation

In this assignment, k-Means clustering algorithm was implemented and tested for grouping the given data into intended number of clusters. Three different datasets were given for applying the clustering algorithm. The dataset files contain two features as x1 and x2 variables and class labels. Class labels were discarded for this assignment and only features were used since k-Means is an unsupervised algorithm and does not need class labels.

Before implementing the algorithm, the dataset is read, its first and second columns are taken as features and one column is added to the original dataset for filling with the cluster numbers assigned to every data point. When the dataset is read, the scatterplot of these variables are plotted for visually understanding the original data without clusters. Plots of the three different data are shown in Figure 1.

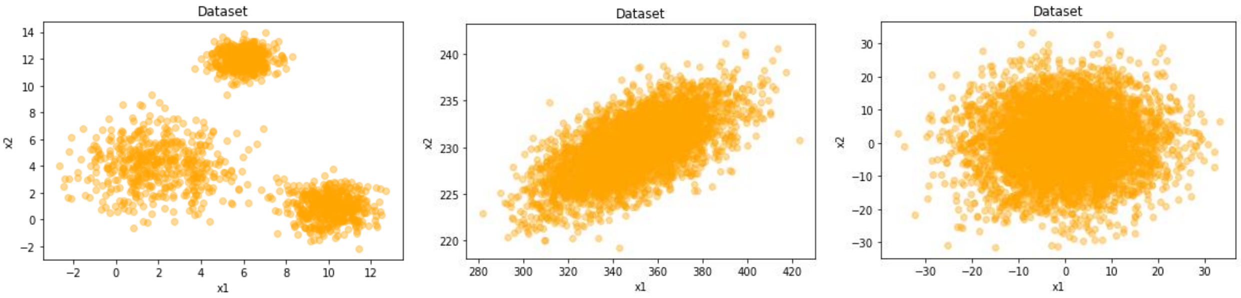

Figure 1. Scatter plots of the three datasets.

For applying the k-Means algorithm, the number of clusters are taken from the user as the k value and according to the number of clusters, random initial cluster centroids are identified. The minimum and maximum values for the two variables of the data are calculated and random numbers for the two variables are generated for every cluster between minimum and maximum values. When the initial random centroid is generated, it is added to the cluster centers array. Cluster centers array consists of k numbers of rows that indicate the clusters and three columns that first two columns for the x1 and x2 coordinates of the clusters. For assigning the observations to each cluster, distances from the point in the dataset to every cluster are calculated based on Euclidean distance metric and added to the third column in the cluster centers array. Then, the index of the cluster with the distance that has the lowest value is considered as the cluster to that point. This process continues until every point in the dataset has the cluster name. At the same time, objective value is calculated by adding the squares of the distances of every point to their clusters.

If no points can be assigned to at least one cluster, then a warning is given to the user for trying again with different initial random centroids and execution is stopped. If all clusters have at least one point, then algorithm continues to run and plots the data with cluster centers. Cluster centers array is copied as previous cluster centers array before calculating new cluster centers to identify the total movement of two consecutive cluster centers. For updating the cluster centers, averages of x1 and x2 variables for every cluster are calculated and cluster centers array is updated with these new centers. While cluster centers are being calculated, total movement of two cluster centers are also calculated with Euclidean distance metric. If the total movement is less than 0.001, then algorithm stops, plots the objective function value according to the number of iterations and prints the initial and final objective function values.

2. Results

The k-Means algorithm was tested for different k values and datasets. Sample outputs of the algorithm are shown in below.

Dataset 1

Dataset 1 was tested with the k values of 3 and 7. The initial and final plots with the cluster centers for dataset 1 with the k value of 3 are shown in Figure 2.

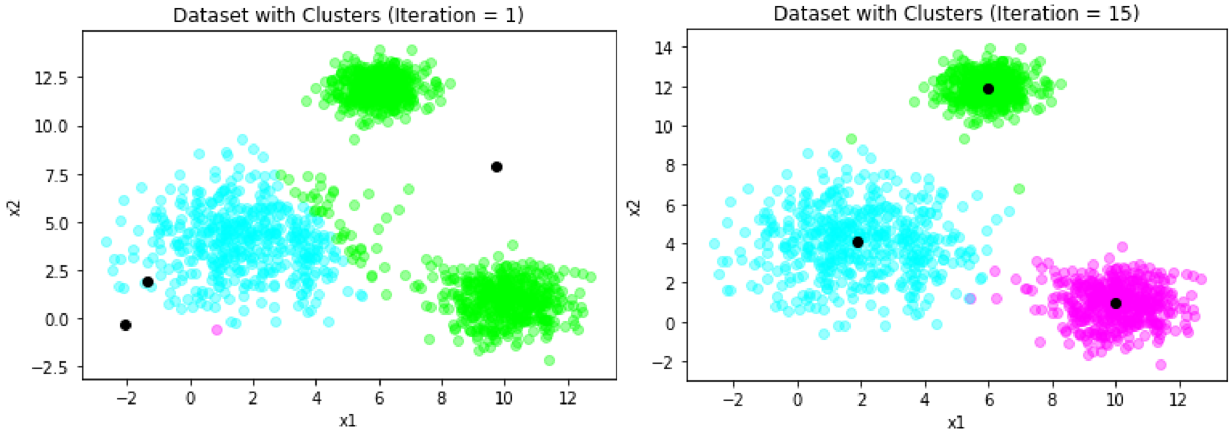

Figure 2. Scatter plots with initial and final cluster centers (Dataset=1, k=3).

The initial objective function value was calculaed as 50817.16 and the final objective function value was calculated as 4489.45 with the total iteration number of 15. The plot of the objective function value according to the number of iterations can be seen in Figure 3.

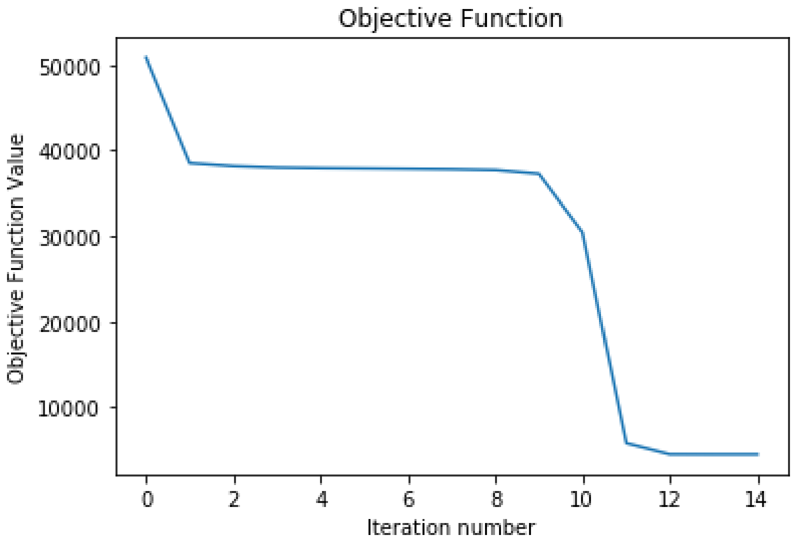

Figure 3. Objective function value vs iteration count. (Dataset=1, k=3).

The initial and final plots with the cluster centers for dataset 1 with the k value of 7 are shown in Figure 4.

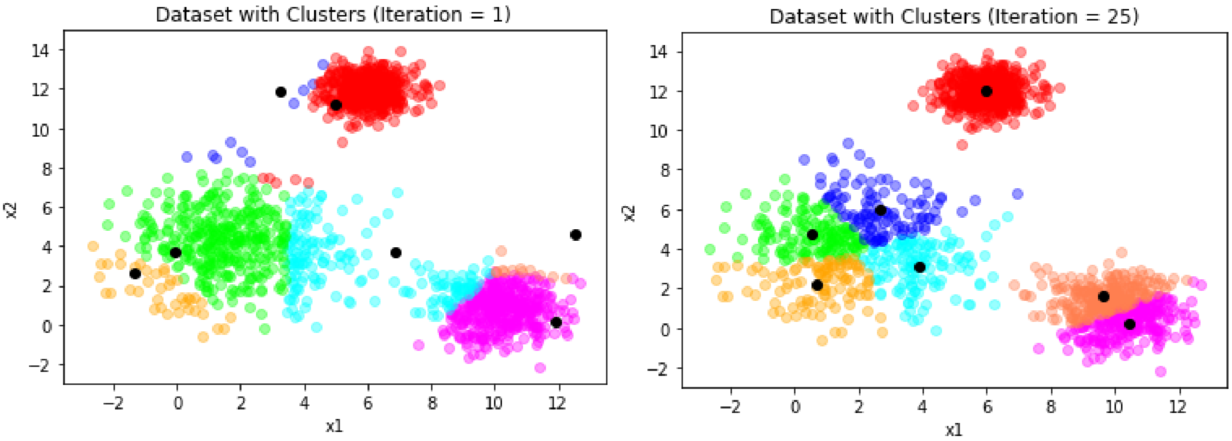

Figure 4. Scatter plots with initial and final cluster centers (Dataset=1, k=7).

The initial objective function value was calculaed as 7228.90 and the final objective function value was calculated as 2172.08 with the total iteration number of 25. The plot of the objective function value according to the number of iterations can be seen in Figure 5.

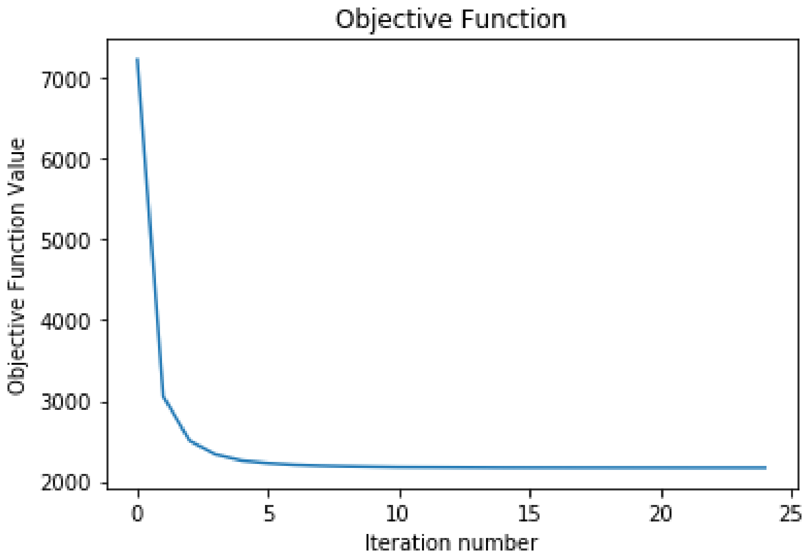

Figure 5. Objective function value vs iteration count. (Dataset=1, k=7).

Dataset 2

Dataset 2 was tested with k values of 2 and 5. The initial and final plots with the cluster centers for dataset 2 with the k value of 2 are shown in Figure 6.

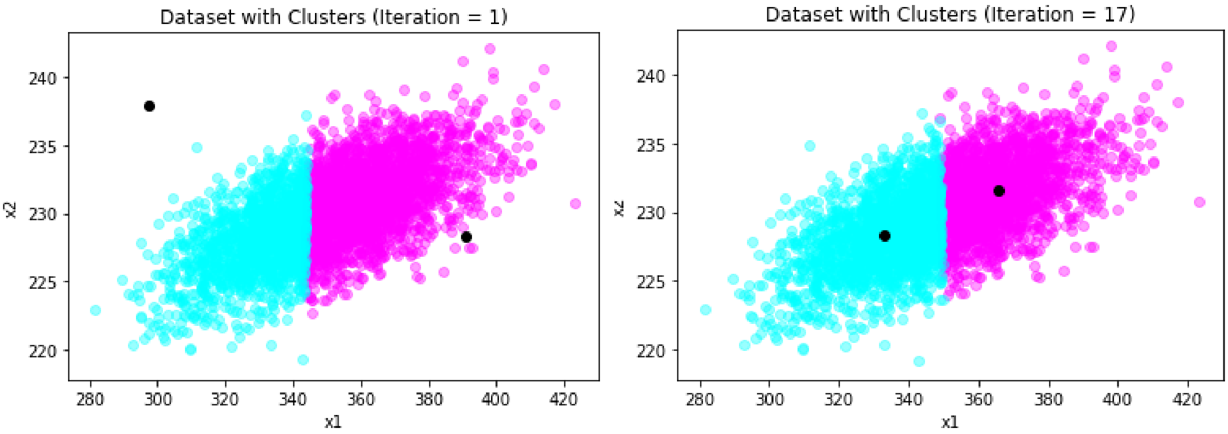

Figure 6. Scatter plots with initial and final cluster centers (Dataset=2, k=2).

The initial objective function value was calculaed as 4363791.65 and the final objective function value was calculated as 632866.76 with the total iteration number of 17. The plot of the objective function value according to the number of iterations can be seen in Figure 7.

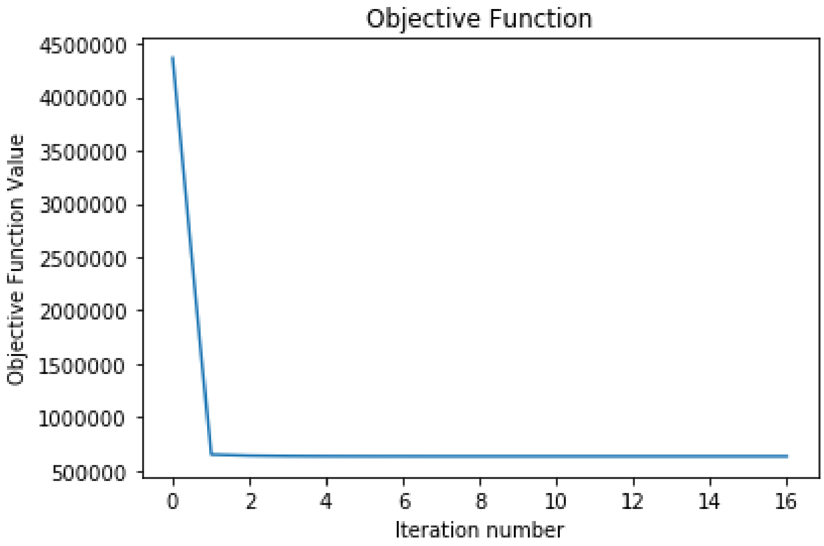

Figure 7. Objective function value vs iteration count. (Dataset=2, k=2).

The initial and final plots with the cluster centers for dataset 2 with the k value of 5 are shown in Figure 8.

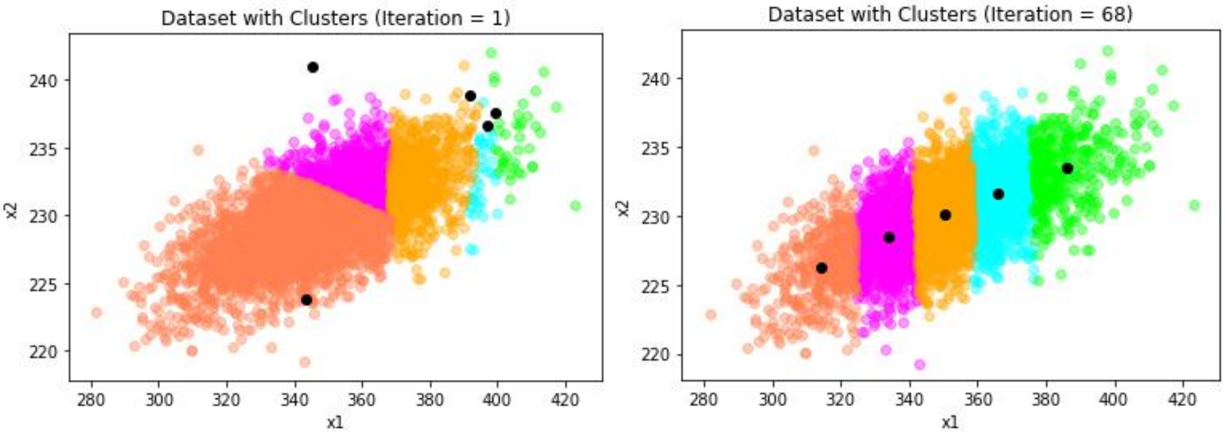

Figure 8. Scatter plots with initial and final cluster centers (Dataset=2, k=5).

The initial objective function value was calculaed as 1115943.16 and the final objective function value was calculated as 154914.29 with the total iteration number of 68. The plot of the objective function value according to the number of iterations can be seen in Figure 9.

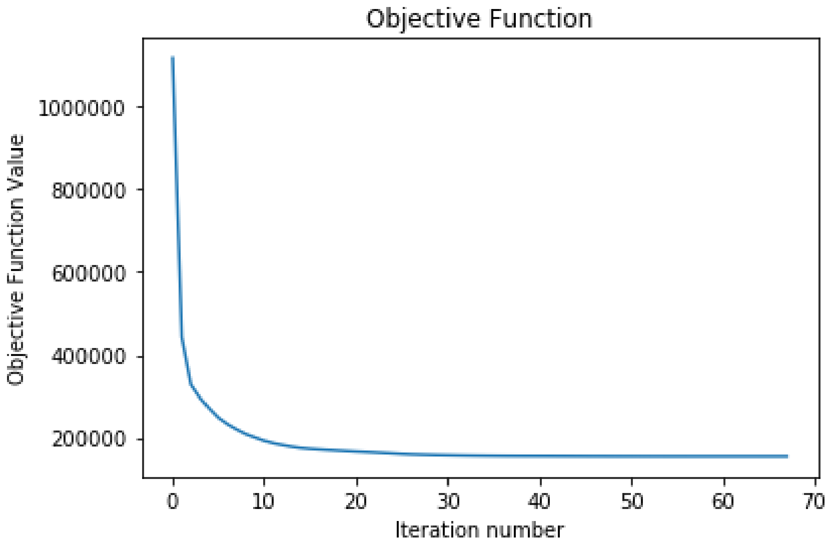

Figure 9. Objective function value vs iteration count. (Dataset=2, k=5).

Dataset 3

Dataset 3 was tested with k values of 3 and 8. The initial and final plots with the cluster centers for dataset 3 with the k value of 2 are shown in Figure 10.

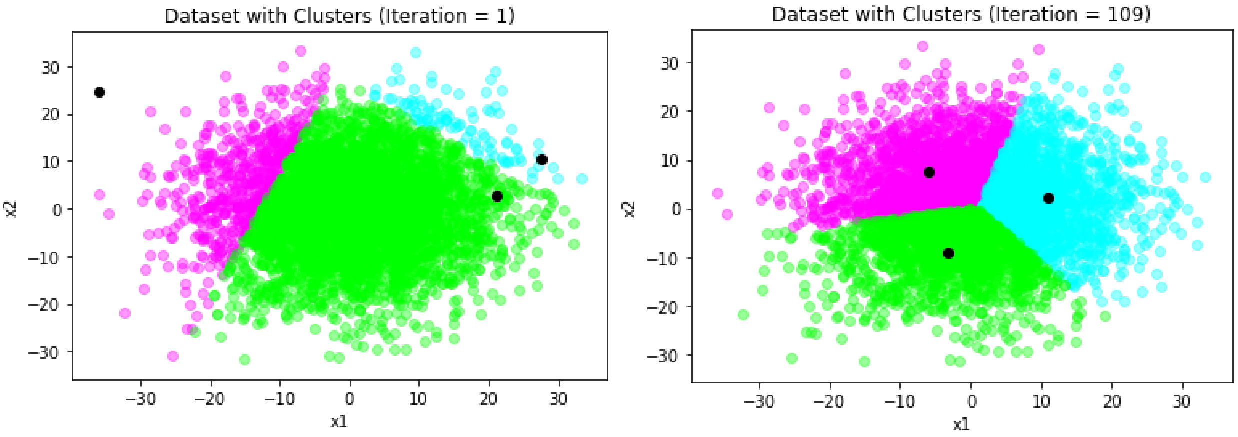

Figure 10. Scatter plots with initial and final cluster centers (Dataset=3, k=3).

The initial objective function value was calculaed as 2844754.95 and the final objective function value was calculated as 444138.81 with the total iteration number of 109. The plot of the objective function value according to the number of iterations can be seen in Figure 11.

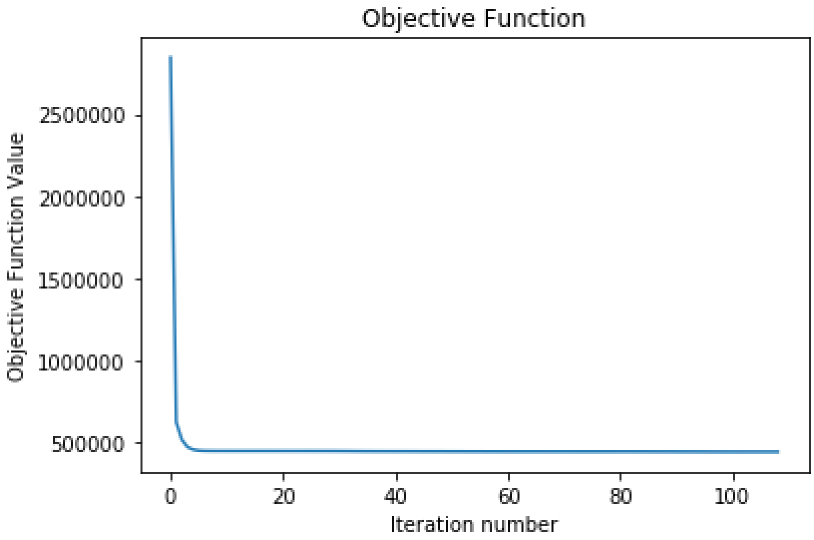

Figure 11. Objective function value vs iteration count. (Dataset=3, k=3).

The initial and final plots with the cluster centers for dataset 3 with the k value of 8 are shown in Figure 12.

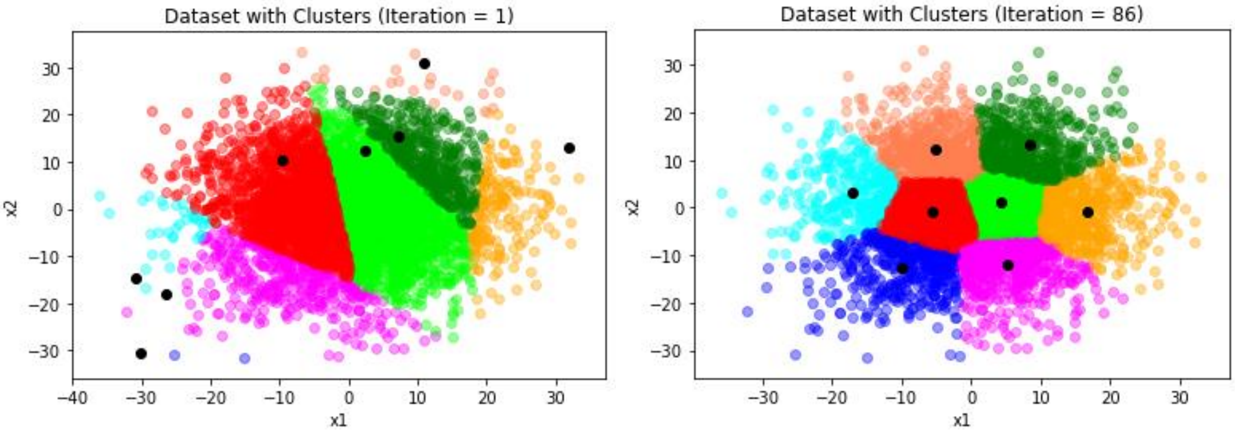

Figure 12. Scatter plots with initial and final cluster centers (Dataset=3, k=8).

The initial objective function value was calculaed as 1103093.13 and the final objective function value was calculated as 189677.56 with the total iteration number of 86. The plot of the objective function value according to the number of iterations can be seen in Figure 13.

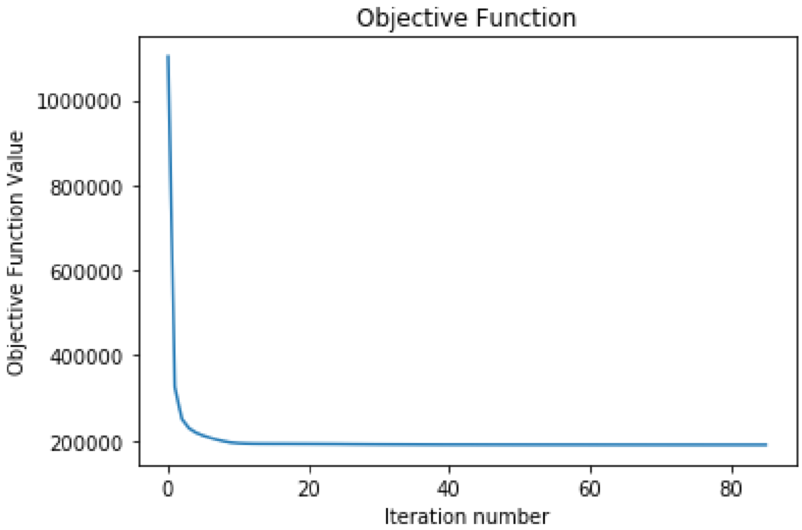

Figure 13. Objective function value vs iteration count. (Dataset=3, k=8).

Pivot table of all results shown above can be seen in Table 1 for identifying easily the differences between the dataset objective function values.

Table 1. Pivot table of results according to different k values and datasets.

	Dataset 1		Dataset 2		Dataset 3	
	k = 3	k = 7	k = 2	k = 5	k = 3	k = 8
Initial Objective Function Value	50,817.16	7,228.90	4,363,791.65	1,115,943.16	2,844,754.95	1,103,093.13
Final Objective Function Value	4,489.45	2,172.08	632,866.76	154,914.29	444,138.81	189,677.56
Iteration Number	15	25	17	68	109	86