README FILE

Shirsha Chowdhury

18EE10068

Imports Used:

Pandas to read the dataset

Numpy for mathematical operations on vectors

Math for mathematical tasks used for formulas

Random for selecting random values

Functions used:

`Zscores(arr):` To find the Z scores of quantitative data.

`distance(A,B):` To find the distance between two data points (i.e Two rows of the dataset)

`Mean(arr):` For quantitative data mean was taken and for categorical data mode will be taken

`randomInit(K):` Randomly Initializes K data points as centroid

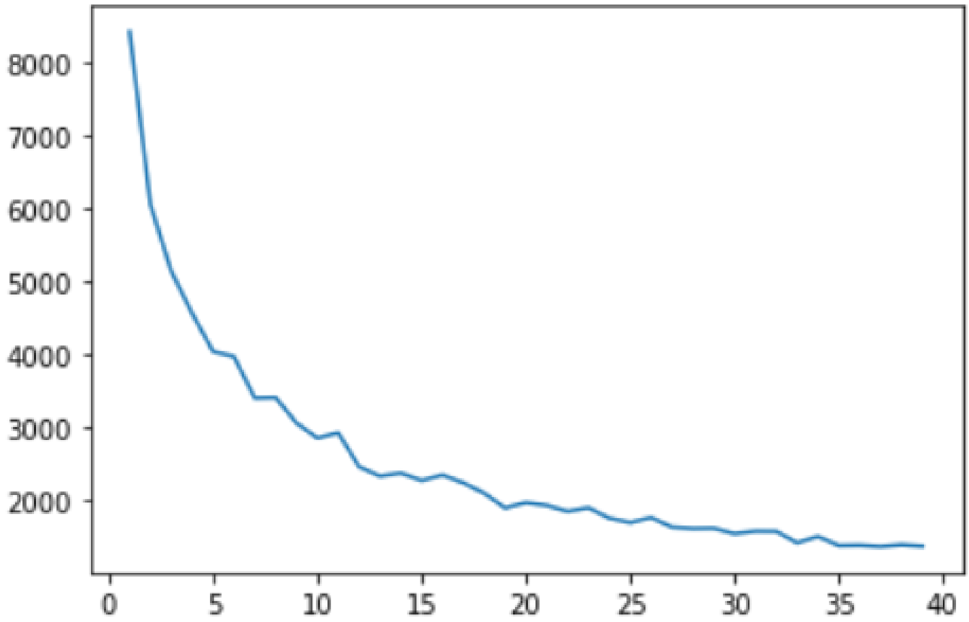
`KMeans(n_iter,K):` n_iter and K was passed in this function and it returns the final cluster which can after n_iter iterations. It was seen that the Cluster were getting converged withing 10 iterations

`MSE(Cluster):`This function was used to see the Error vs K plot from which the optimized K was taken

`dataPrint(Cluster):` Prints the output in a file

`printCluster(Cluster):` Prints the distribution of points in different Clusters

Idea: To use K-Means algorithm to assign clusters to a unsupervised data. The output was stored in a dictionary in which the keys are the cluster number and values are the list of data point belonging to that cluster. To choose the optimized value of K MSE error was taken. The plot obtained was as follows.



In the above plot of MSE vs K (cluster size) we observe that around K=15 the curve gets flatten out in the x-axis. So I am taking K=15 as the optimized value of K. Normally MSE decreases as K increases so we are taking the elbow point of the MSE vs K graph. The clusters were observed to be converging after 10 iterations in most of the cases.

The final distribution for K=15 was as follows after 20 iterations.

137 | 107 | 219 | 70 | 162 | 53 | 157 | 154 | 168 | 109 | 158 | 94 | 84 | 231 | 97 |