**IS 733-HW3**

1.a) <https://github.com/Sumanth457/is7332025/blob/main/data-mining-project-repo/HW3/DC93457_IS733_HW3.ipynb>

1.b) As we can see in the plot there is a clear gap between two groups hence it clearly shows that there are two distinct clusters of data points. One cluster has shorter eruption times around 2-3 minutes and shorter waiting times around 50-70 minutes. Whereas the second cluster has longer eruption times around 4-5 minutes and longer waiting times around 75-95 minutes.

1.c) There are two types of hierarchical clustering –

a) Divisive Hierarchical Clustering Steps

1. Start with all the data points in a single cluster.
2. Using methods like K means(k=2), spatial clustering and other methods, identify the least coherent cluster and split the cluster into two sub clusters. The split should have high intra-cluster similarity and low intra-cluster similarity.
3. By using measures like distance metrics, variance or silhouette scores, evaluate the split the clusters and recursively consider whether it is worth splitting each cluster further.
4. Visualize the splits as a binary tree which is called dendrogram.

b)Agglomerative Hierarchical Clustering Steps

1. Start considering each data point as its own cluster.
2. Calculate distance between all pairs of clusters.
3. Find clusters with smallest distance and merge them into a single cluster.
4. Repeat steps 2-3 until all points are merged into a single cluster.
5. The hierarchy of merges is visualized as a tree called dendrogram.

Hierarchical clustering method is the appropriate method for this dataset. The scatter plot clearly shows two clusters groups and hierarchical clustering works perfectly for discovering such groupings without needing to specify the number of clusters. This method also provides a dendrogram which helps to visualize the data and see how many clusters exists.

2.a,b,c) <https://github.com/Sumanth457/is7332025/blob/main/data-mining-project-repo/HW3/DC93457_IS733_HW3.ipynb>

2.d) Yes, K means successfully found the two clusters we identified in problem 1. When we plotted the scatter plot, we observed two cluster groups, one with short eruption and short waiting times and the other with long eruption and long waiting times.

After running k means clustering, the final clustering cleanly separated the data into the same two regions. One cluster has shorter eruptions and shorter waiting times and the other has long eruptions and waiting times. Thus, it is clearly visible that K means correctly spotted the separation which we observed in scatter plot.