**ICP-7:**

**Name: Amarnadha Reddy Ankireddypalli**

**ID: 16314889**

**Email: aa9f3@umsystem.edu**

**Description:**

Use a different dataset (than the one we used in class) and use the model provided in ICP7 to perform clustering. You must try 5 different number of clusters (for example, n­\_clusters= 5 or n\_clusters=6,7,8,or 9 etc.) based on elbow curve and for each cluster visualize the clustering results and report your findings in detail. Also provide the elbow curve screen shots in your report.

**Objective:**

1. Successfully executing the code and making clusters with different value of K.
2. Using a new and good dataset
3. Providing the logical explanation of the changes that you made to model and over all code quality
4. Wiki Report quality, video explanation

**Implementation (with screenshots):**

1. Import the libraries required for clustering and visualization.

Graphical user interface, text

Description automatically generated

1. This is a California Housing dataset taken from Kaggle. Dataset has following attributes for each block of group in California: Population, Median Income, Median Housing Price and more.

A picture containing graphical user interface

Description automatically generated

1. Here, I’m only including the important features of the dataset to perform clustering.

Table

Description automatically generated

1. Here, we see that there is a lot of variation in the magnitude of the data. Variables like medium\_income and households have low magnitude whereas variables like total\_rooms, population, median\_house\_value, have a higher magnitude.
2. Since K-Means is a distance-based algorithm, this difference of magnitude can create a problem. So, let’s first bring all the variables to the same magnitude:

Table

Description automatically generated with medium confidence

1. The magnitude looks similar now. Next, let’s create a k-means function and fit it on the data:

Text

Description automatically generated

1. To evaluate how well the formed clusters are, we will calculate the inertia of the clusters:

**Inertia**: It is the sum of squared distances of samples to their closest cluster center.

Graphical user interface, application

Description automatically generated

1. We got an inertia value of almost 34832. Now, let’s see how we can use the elbow curve to determine the optimum number of clusters.

Chart, line chart

Description automatically generated

1. Looking at the above elbow curve, we can choose any number of clusters between 4 to 8. Let’s set the number of clusters as 4 and fit the model:

Graphical user interface, text, application

Description automatically generated

1. Let's now plot these clusters on a 3D scatter plot.

4 CLUSTERS:

Chart, scatter chart

Description automatically generated

5 CLUSTERS:

Chart, scatter chart

Description automatically generated

6 CLUSTERS:

Chart, scatter chart

Description automatically generated

7 CLUSTERS:

Chart, scatter chart

Description automatically generated

8 CLUSTERS:

Chart, scatter chart

Description automatically generated

In conclusion,

* The scatter plot above shows the geographic distribution of the clusters. It appears that the algorithm created separate segments for median\_house\_value depending on total\_rooms and population.
* This is how we can use Python to create the K-means clustering algorithm. Before executing the K-means, it's critical to scale the input features; otherwise, the clusters would become excessively stretched, and the algorithm will perform poorly.
* Scaling the features, on the other hand, does not guarantee that the clusters will become spherical, but it usually improves a lot.

**Difficulties faced:** In visualizing the clusters in 3D scatterplot**.**

**Video Link:** <https://youtu.be/zODVwr3Go1M>

**Conclusion:**

From this ICP7, I’ve learnt:

* How to scale the important features of the data in case of lot of variation in the magnitude.
* What is inertia and how to form elbow curve based on inertia.
* How to analyze elbow curve to decide on no.of clusters and their visualization.