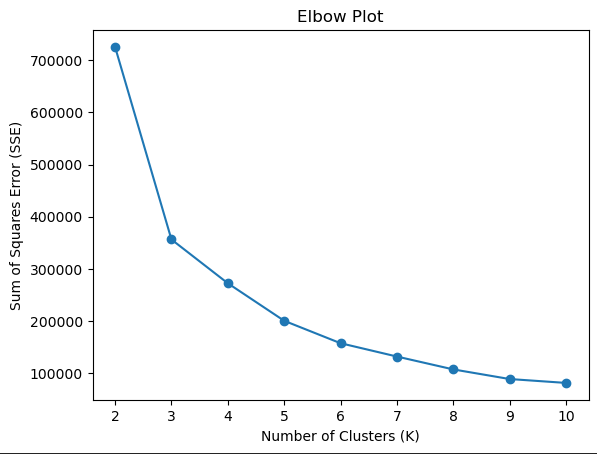
1 a)



Based on the Elbow Plot, shows the Sum Squares Error (SSE) against Number of Clusters(K) plot in K-Means Clustering. Though the optimum k-value is 5, as it is said to be optimum as it falls on the elbow point. This point signifies a notable decrease in SSE as the number of clusters (k) increases and provides the balance between minimizing SSE and avoiding excessive complexity model that overfits. Most notably, it also represents the stage at which including more clusters ceases to yield a significant reduction in SSE.

Though it does not seem to be the case in this dataset, it is notably that a one-unit increase in the k-value from the elbow point can still result in a substantial difference of approximately 40,000 units in SSE. This speculation implies that even a slight change in latitude and longitude, even by a decimal point, holds considerable importance in determining a specific geographical location on Earth. Therefore, the elbow point as k-value is not sufficient in minimizing SSE.

A more optimum k-value would be 8 in this case, as the SSE is ceased to yield significant reduction with each increase in k-value from that point (≈25000 down to ≈18500 from k-value 8 to 9). Furthermore, it is more practical to use k-value 8 to determine each continent as there are many variations in the latitude and longitude in this dataset to represent every continent.

b)

A screenshot of a computer

Description automatically generated

In my best clustering, I selected the k-value as 8 as the final clusters with random\_state set to 0. The coordinates of the centroids for each cluster are as shown in 1(C) along with their respective continents. The sum of squared differences to centroids is 107389.45158. The clustering algorithm took 6 iterations to converge. I believe this is the best clustering because it achieved a significant reduction in SSE, indicating compact clustering around the centroids, and the chosen value of k provided a practical representation of each continent based on the latitude and longitude variations in the dataset.

c)

A screen shot of a computer

Description automatically generated with medium confidence

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Centroid | Continent | Country/Ocean |
| 0 | 49.13415273, 11.04527028 | Europe | Deutschland |
| 1 | 19.26297174, 104.62467122 | Asia | Việt Nam |
| 2 | 9.62003573, -69.14907131 | South America | Venezuela |
| 3 | -16.50361159, 37.73891391 | Africa | Moçambique |
| 4 | -15.84210807, -164.35100116 | Oceania | South Pacific Ocean |
| 5 | -10.81353477, 156.11725861 | Oceania | Solomon Sea |
| 6 | 10.75386428, -1.10167986 | Africa | Ghana |
| 7 | 29.86142026, 48.3508315 | Asia | Iraq |

A picture containing screenshot, colorfulness, text, diagram

Description automatically generated

2 a)

A picture containing diagram, text, plan, line

Description automatically generated

The best parameter for Agglomerative Clustering would be value n\_cluster of 9 with Manhattan distance metric and complete linkage. Considering these factors, the selected parameters aim to achieve a clustering solution that captures the distinct geographic regions represented by the continents. The clusters number 9 is more accurate in this method, as it can separate the continent between Oceania and Antarctica. The Manhattan distance metric is appropriate for calculating distances based on latitude and longitude coordinates, providing an accurate measure of dissimilarity. Finally, complete linkage promotes the identification of well-separated and cohesive clusters, ensuring that each continent is represented distinctly.

b) Name the continent each cluster represent in the table below. To analyze each cluster, observe the data points in each cluster on the scatter plot or look at what country names are in the clusters.

|  |  |
| --- | --- |
| Cluster | Continent |
| 0 | Asia |
| 1 | Europe |
| 2 | North America |
| 3 | Africa |
| 4 | Oceania |
| 5 | Oceania |
| 6 | South America |
| 7 | Asia |
| 8 | Antarctica |

A picture containing screenshot, text, colorfulness, diagram

Description automatically generated

3 a) Based on the algorithm you have selected, provide a final mapping of the cluster number to the continent name in the following table.A map of the world with different colored dots

Description automatically generated with medium confidence

|  |  |
| --- | --- |
| Cluster | Continent |
| 0 | Asia |
| 1 | Europe |
| 2 | North America |
| 3 | Africa |
| 4 | Oceania |
| 5 | Oceania |
| 6 | South America |
| 7 | Asia |
| 8 | Antarctica |