CDS503 MACHINE LEARNING

ASSIGNMENT 2

ACADEMIC SESSION 2021/2022
SEMESTER 2

Part 1: K-Means Clustering (25 marks) Use the K-Means cluster algorithm to find clusters representing continents.

a)

Experiment with some different sizes of k and observe the range of the Sum of Squares Error (SSE) (see Appendix for more details on SSE). What k value would you pick to best cluster the countries into continents? Briefly justify why you select the k value.

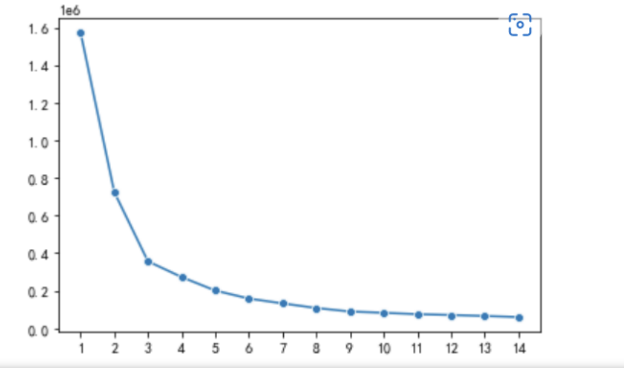


Figure1

I would choose K=5 according to the elbow method, and Figure1 shows that the SSE value goes down when k reaches 4. Even though the SSE is lower when k =6, we know there are five continents in the world

Commented [JLSY1]: There are 7 continents in the world so setting k = 7 is the most intuitive.

b) Report the final parameters you set including the value of k selected to obtain your final clusters. Also, report the coordinates of the centroids, sum of squared differences to centroids and the number of iterations from the best clustering you have found. Explain in one sentence why you think this is your best clustering.

Value Name	value
k	5
coordinates of cluster centers	[6.47298142 129.70481495] [9.62003573 -69.14907131] [2.52989902 42.76142619] [39.02417162 10.32558702] [-15.84210807 -164.35100116]
Sum of squared distances	200747.04206284357
Number of iterations	4

Figure2

This is the best clustering to me, since Figure4 makes the most sense to me, it agree with the real world Situation.

Commented [JLSY2]: Need to provide evidence from data.

C）Name the continent each cluster represents in the table below. Describe each cluster according to the centroid values of each attribute. For each cluster, be sure to report the attribute centroid in terms of the original attribute values. Also, you can visualize the clusters on a scatter plot to help you describe and identify the continent represented by each cluster. You can also concatenate the cluster labels with longitude, latitude, and country names to analyze the countries in each cluster.

Cluster	Centroid	Continent
0	6.47298142 129.70481495	AISA
1	9.62003573 -69.14907131	America
2	2.52989902 42.76142619	Africa
3	39.02417162 10.32558702	Europe
4	-15.84210807 -164.35100116	Oceania

Figure3

Commented [JLSY3]: Missing Antarctica.

Commented [JLSY4]: Asia.

Commented [JLSY5]: North American and South America are two different continents.

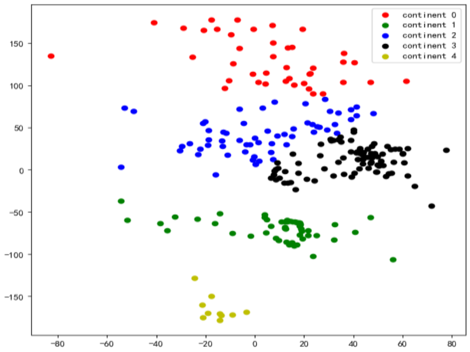


Figure4

Commented [JLSY6]: Word map is flipped. Longitude should be the x-axis and latitude should be the y-axis.

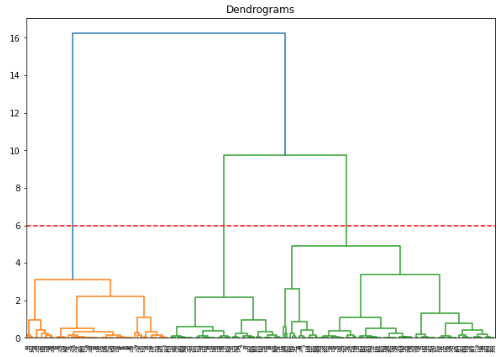
From Figure3, this shows that cluster0 is Asia, cluster1 is America ,cluster2 is Africa, cluster3 is Europe, and cluster4 is Oceania.

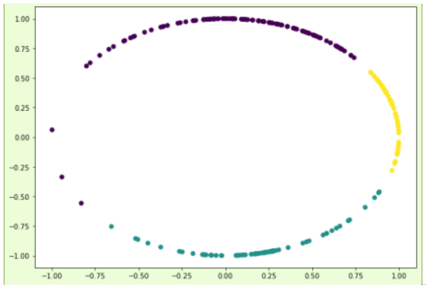
Part 2: Hierarchical Clustering (25 marks)

a) Use the Agglomerative Clustering algorithm. Experiment with different number of clusters (n_clusters) and other parameters (e.g., affinity, linkage, etc.) to get the best clusters to represent continents. Use dendrogram and scatter plot to help you visualize the clusters. Report the best parameters, the number of clusters you have selected. Explain in one sentence why you think this is your best clustering.
This is my best clustering since the dendrogram is balanced and makes the most sense.

n_clusters=3, affinity='euclidean', linkage='ward'

Commented [JLSY7]: Not representative of actual world continents.





Commented [JLSY8]: Not quite sure what are you trying the convey through this plot.

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
6.74638591 -82.962124	Oceania
22.80562742 26.44343502	europe
7.43620576 132.59868938	asia

Commented [JLSY9]: Missing 4 continents.

Part 3: Identify the best continent clustering (50 marks)

Based on the best clusters obtained respectively from K-means and AgglomerativeClustering, choose ONE algorithm that would give you the most accurate grouping of countries into the correct continent (final_cluster). Marks will be based on the number of your continent cluster labels matching the actual continent labels.

a) Based on the algorithm you have selected, provide a final mapping of the cluster number to the continent name in the following table.

I choose K means as my final model.

Cluster	Centroid	Continent
0	6.47298142 129.70481495	AISA
1	9.62003573 -69.14907131	America
2	2.52989902 42.76142619	Africa
3	39.02417162 10.32558702	Europe
4	-15.84210807 -164.35100116	Oceania

Commented [JLSY10]: Asia.

Commented [JLSY11]: Mapped to the larger continent: North America.