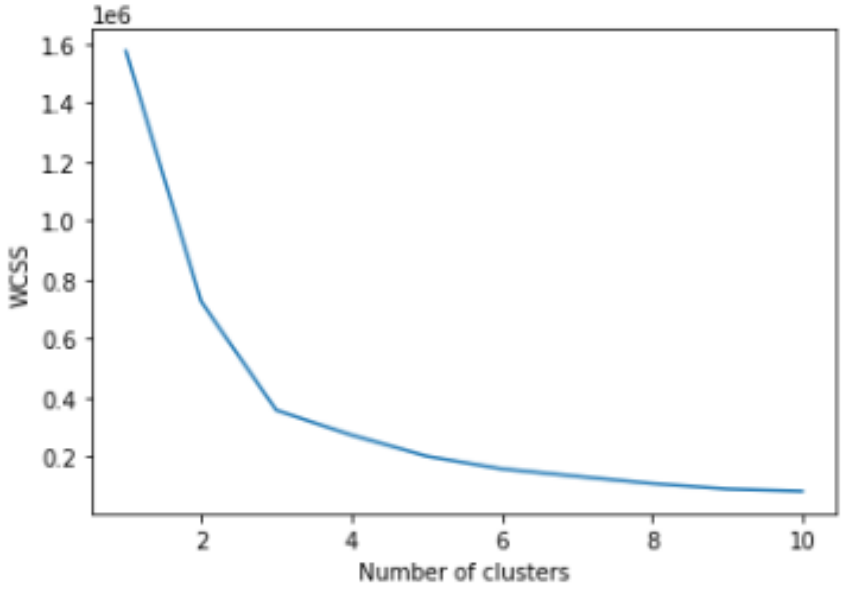
Part 1 Kmeans Clustering

(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.



The above graph shows the relationship between cluster and sum of squared error. Based on the graph, k value of 3 is elbow point where adding the k value would not give much better modelling of the data. Hence, k value of 3 is best clustering by analysing sum of square error.

(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.

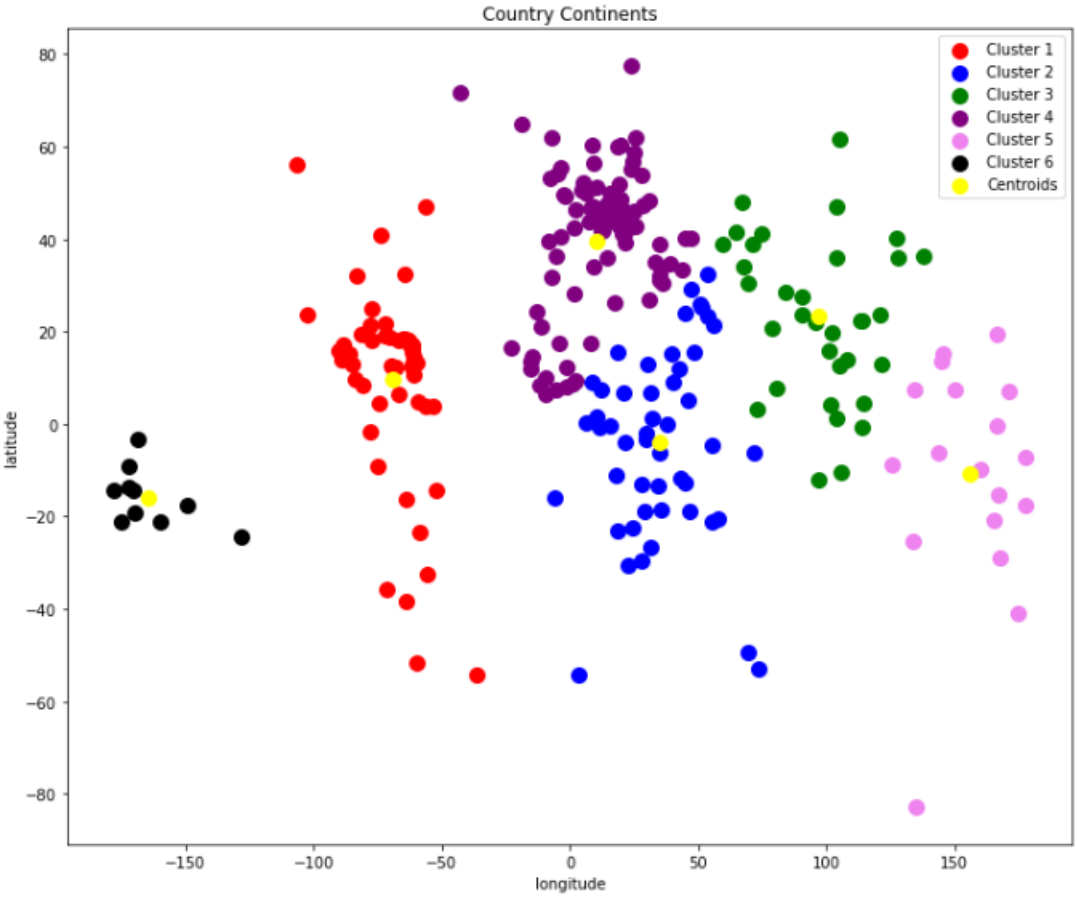
Items	Description	
Parameter	n_cluster = 6 random state = 0	
Coordinate of cluster centers	Longitude	Latitude
	-69.1491	9.620036
	35.00838	-3.92314
	97.03025	23.31903
	10.64995	39.60602
	156.1173	-10.8135
	-164.351	-15.8421
Number of iterations	8	
Sum of squared differences to centroids	157589.89379630276	

This is the best clustering because the countries listed are separated into 6 continents, hence it is best to cluster it into 6 components.

(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 analyse the countries in each cluster.

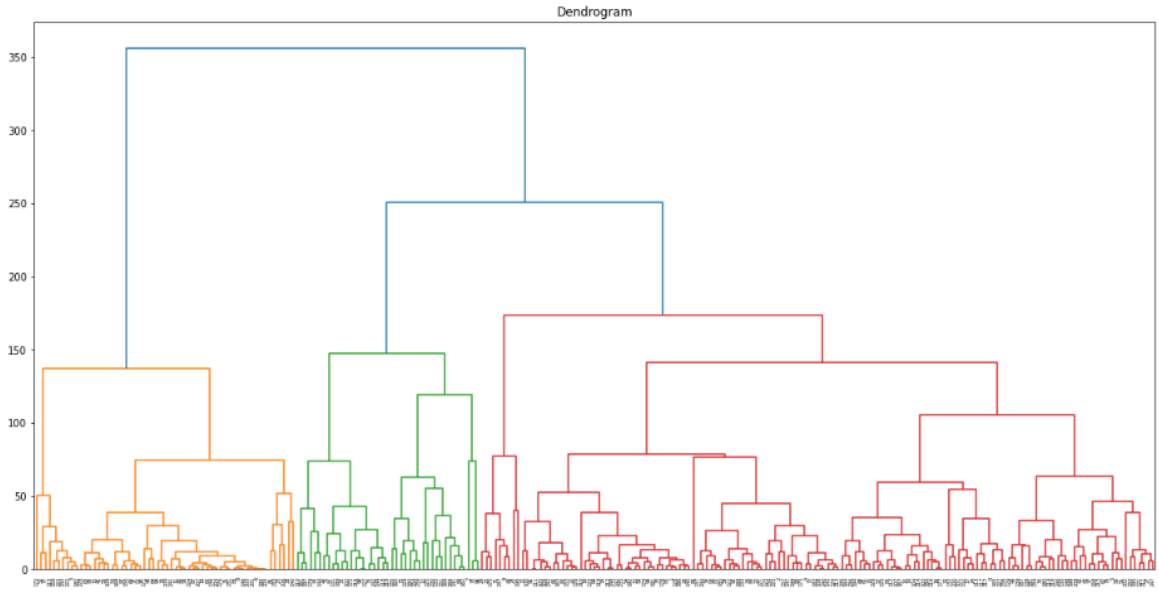
Cluster	Longitude	Latitude	Continents
0	-69.1491	9.620036	Americas
1	35.00838	-3.92314	Africa
2	97.03025	23.31903	Asia
3	10.64995	39.60602	Europe
4	156.1173	-10.8135	Australia / Oceania
5	-164.351	-15.8421	Australia / Oceania



Part 2: Hierarchical Clustering (25 marks)
Use the hierarchical clustering algorithm to find clusters representing continents.

a) Use the Agglomerative Clustering algorithm. Experiment with different number of clusters (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.

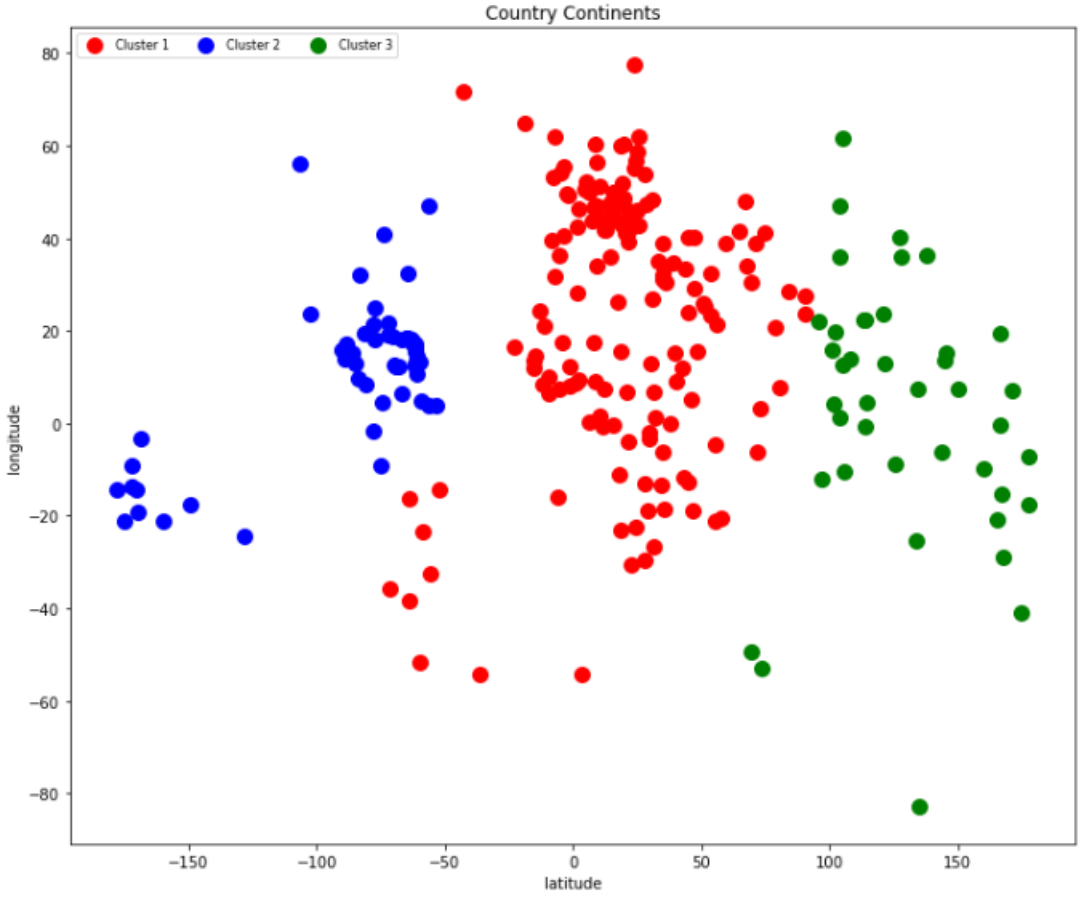
Parameters	Description
Number of clusters chosen	3
Affinity	Euclidean
Linkage	Complete



This is the best clustering using agglomerative because it produces a more balanced result where the outcomes are distributed more fairly compared to the other parameters.

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	Continents
0	North America and Oceania
1	South America, Europe and Africa
2	Asia and Australia



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.

Cluster	Continents
0	Americas
1	Africa
2	Asia
3	Europe
4	Australia / Oceania
5	Australia / Oceania