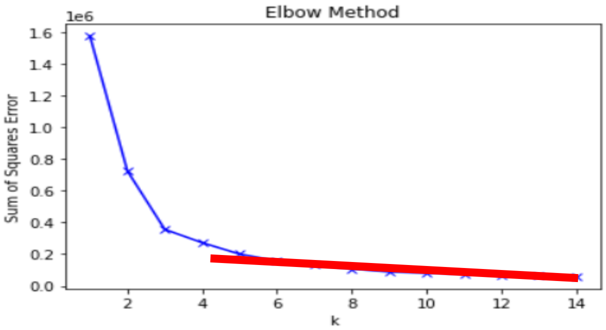
Assignment Report

Part 1: K-Means Clustering

(a)



The number of clusters, k = 6 was selected to best cluster the countries into continents. This is because by observing the SSE for different sizes of k, we can see that at k = 6, the SSE starts to decrease in an approximately linear fashion, which means that adding another cluster doesn't give much better modelling of the data.

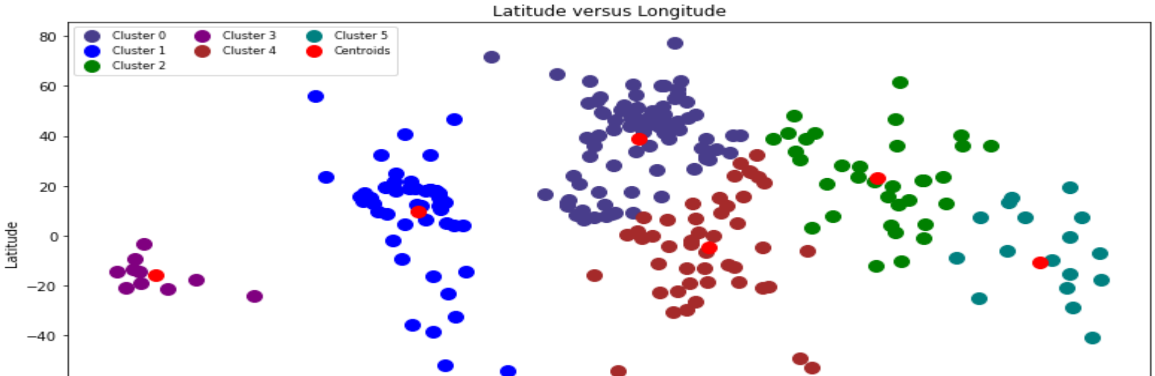
(b) Final parameters for KMeans: n_cluster = 6, n_init = 100, random_state = 0, and the rest of the parameters following the default.
The coordinates of the centroids are as below:

Cluster	Latitude	Longitude
0	38.95511233	10.72265493
1	9.62003573	-69.14907131
2	23.31902544	97.03025315
3	-15.84210807	-164.35100116
4	-4.61213085	35.91495756
5	-10.81353477	156.11725861

Sum of squared distances to centroids: 157574.21116675233
Number of iterations: 7

I think this is the best clustering because the number of iterations until it reaches convergence is low, the sum of squared distances to centroids is low, the size of each cluster is more or less similar (did not exist the case where a majority of points fall under one cluster), the clusters are nicely represented in the scatterplot with the centroids in their centres as shown in (c), which is something similar to the location of each continent in the real-world map.

(c)



Cluster	Centroid	Continent
0	(38.95511233, 10.72265493)	Europe
1	(9.62003573, -69.14907131)	America
2	(23.31902544, 97.03025315)	Asia
3	(-15.84210807, -164.35100116)	Oceania
4	(-4.61213085, 35.91495756)	Africa
5	(-10.81353477, 156.11725861)	Oceania

The centroid values represents the coordinates in (latitude, and longitude). The cluster labels are concatenated with longitude, latitude and country names to analyse the countries in each cluster. Even though are there 6 clusters, there are only 5 continents observed. The following explains what each cluster is.

- The centroid for Cluster 0 is (38.95511233, 10.72265493), which is located in the Europe continent. For instance, Denmark, Finland, Germany, Iceland, and so on are clustered together in this cluster.

- The centroid for Cluster 1 is (9.62003573, -69.14907131), which is located in the America continent. It includes both North America and South America. For instance, Brazil, Canada, Dominica, Peru, Uruguay, and so on are clustered together in this cluster.

- The centroid for Cluster 2 is (23.31902544, 97.03025315), which is located in the Asia continent. For instance, Brunei, China, Japan, India, and so on are clustered together in this cluster.

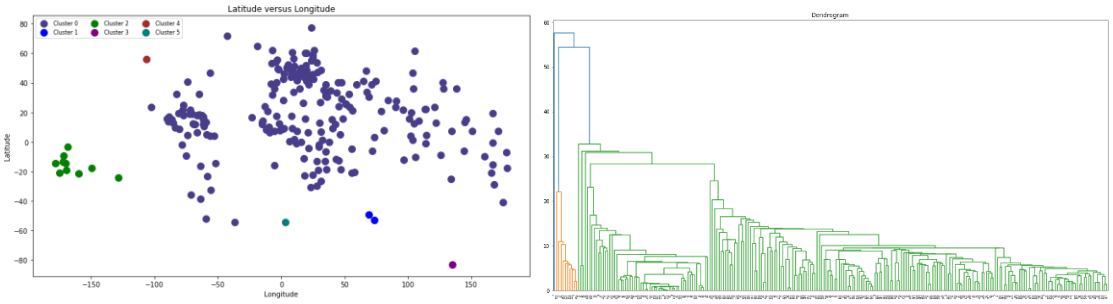
- The centroid for Cluster 3 is (-15.84210807, -164.35100116), which is located in the Oceania continent. It is the only cluster that shared the same continent as Cluster 5, containing territories of Oceania. Cluster 3 is actually right beside Cluster 5. This is because the Earth is spherical in shape, but the scatterplot is only showing the countries in a 2D plot. For instance, American Samoa, Cook Islands, French Polynesia, and so on are clustered together in this cluster.

- The centroid for Cluster 4 is (-4.61213085, 35.91495756), which is located in the Africa continent. For instance, Angola, Cameroon, Ethiopia, Madagascar, and so on are clustered together in this cluster.

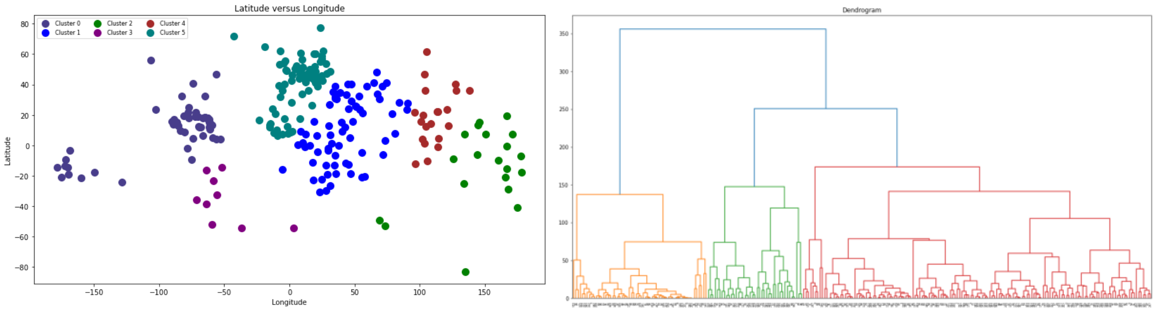
- The centroid for Cluster 5 is (-10.81353477, 156.11725861), which is located in the Oceania continent. It shares the same continent as Cluster 3, containing countries and territories of Oceania. For instance, Australia, Fiji, New Zealand, Papua New Guinea, and so on, are clustered together in this cluster.

Part 2: Hierarchical Clustering

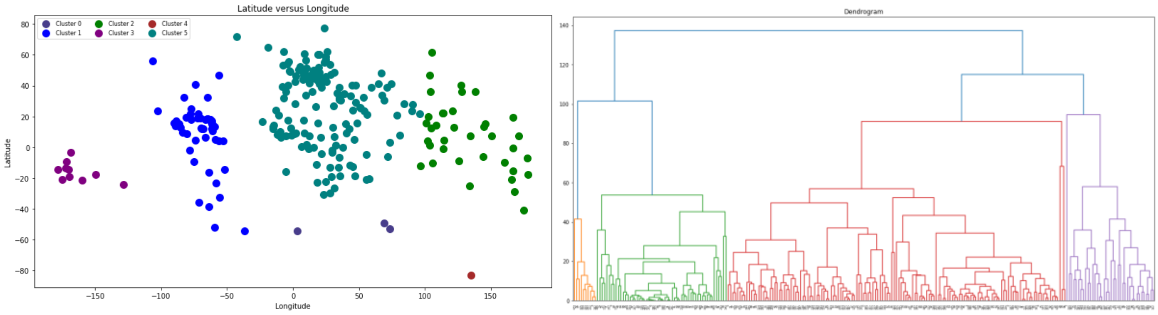
(a) n_clusters = 6, affinity = ‘euclidean’, linkage = ‘single’



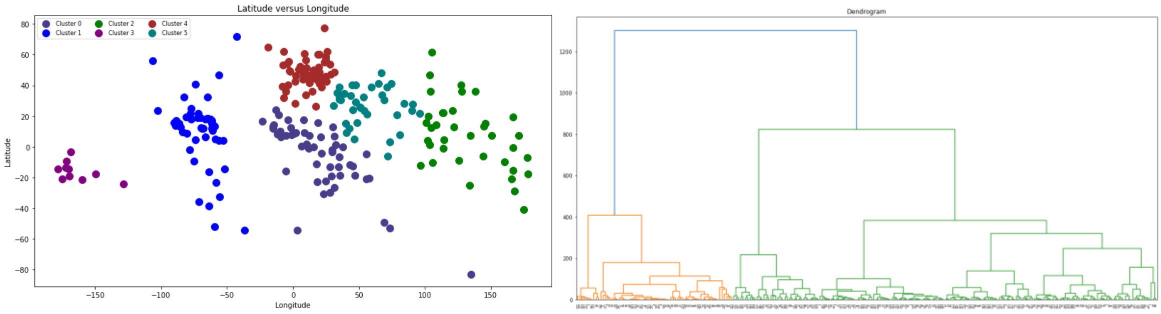
n_clusters = 6, affinity = ‘euclidean’, linkage = ‘complete’



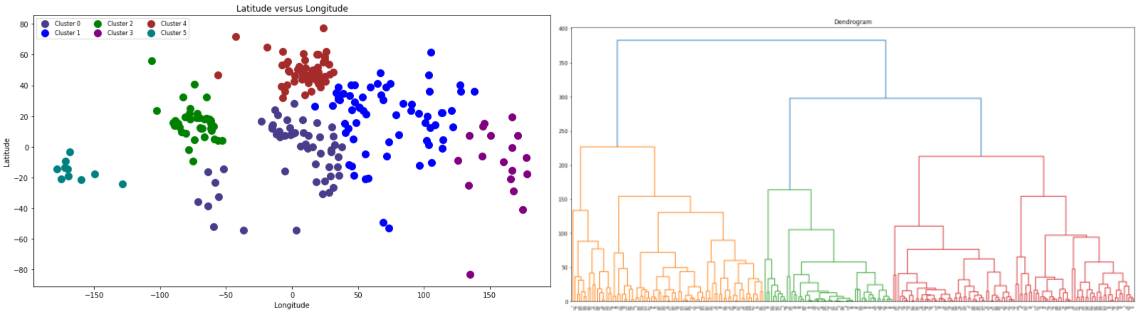
n_clusters = 6, affinity = ‘euclidean’, linkage = ‘average’



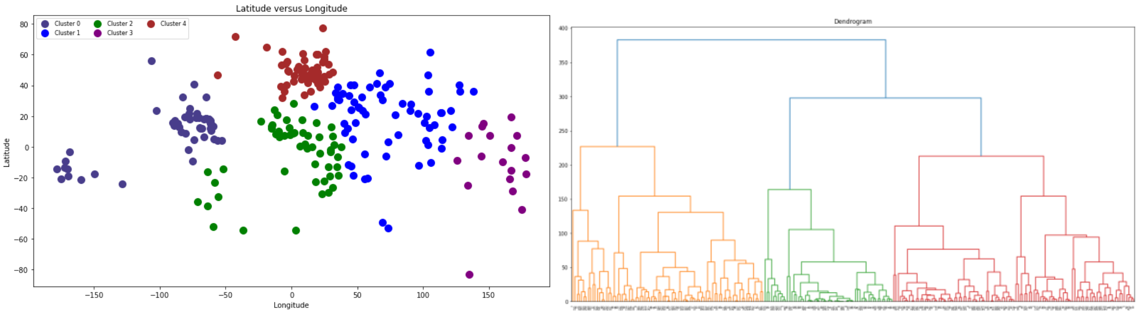
n_clusters = 6, affinity = ‘euclidean’, linkage = ‘ward’



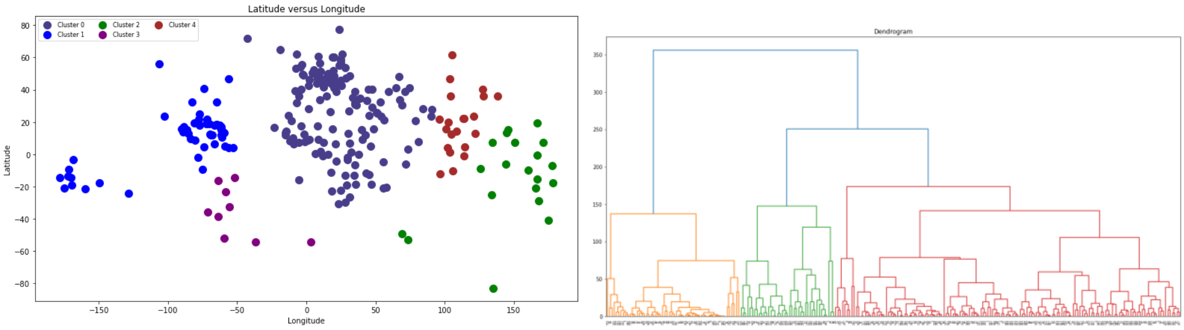
n_clusters = 6, affinity = ‘manhattan’, linkage = ‘complete’



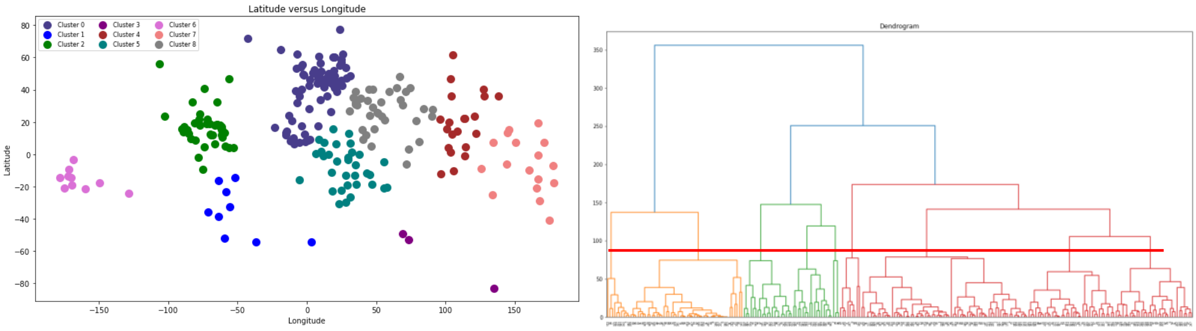
n_clusters = 5, affinity = ‘manhattan’, linkage = ‘complete’



n_clusters = 5, affinity = ‘euclidean’, linkage = ‘complete’



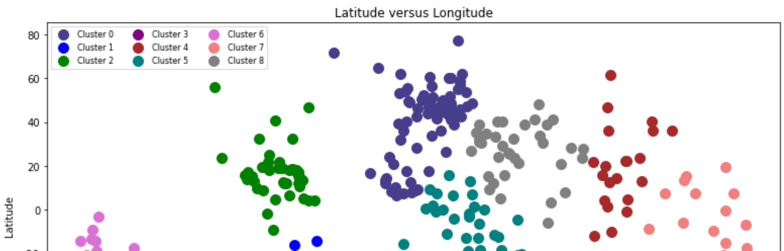
n_clusters = 9, affinity = ‘euclidean’, linkage = ‘complete’



The worst combination of parameters is n_clusters = 6, affinity = ‘euclidean’, linkage = ‘single’, as most of the data points are clustered together under one cluster as shown in the scatterplot and dendrogram.

Therefore, according to the scatterplots and dendrograms, the best parameters are n_clusters = 9, affinity = ‘euclidean’, linkage = ‘complete’. This is the best clustering because the clusters shown in its scatterplot and dendrogram are more or less equal in size and look more similar to what we would expect to see in the continent on a real-world map.

(b)



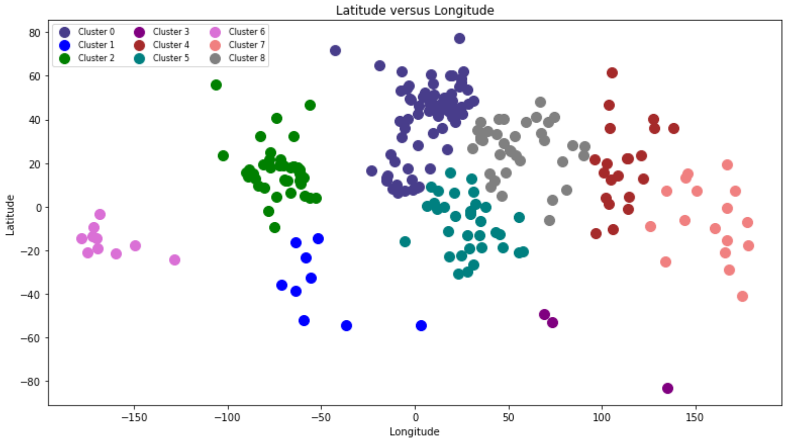
Cluster	Continent
0	Europe
1	South America
2	North America
3	Antarctica
4	Asia
5	Africa
6	Oceania
7	Oceania
8	Asia

Cluster 4 and Cluster 8 belong to the same continent, Asia. Cluster 6 and Cluster 7 belong to the same continent, Oceania.

Part 3: Identify the best continent clustering

(a) When the best clusters from KMeans is compared to the best clusters from AgglomerativeClustering, it seems that the best and most accurate clusters is the one using AgglomerativeClustering because it further separates America into South and North America and could identify the Antarctica continent.

Scatterplot of AgglomerativeClustering



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