Title: Assignment 2

Part 1: K-Means Clustering

a.) Start the clustering with 2 attributes (longitude and latitude) with K-Means clustering which the k=3, and continue subsequently by increasing k value. From the results (showed as below table), the best value for k is 9, to perform clustering for the countries into 9 continents. The sum squares error (SSE) is decreasing significantly from k=3 until k=9. The sum squares error (SSE) for k=10 compare to k=9 is having the smallest changes. Hence, number of clustering 9 can be justified as the best size for k.

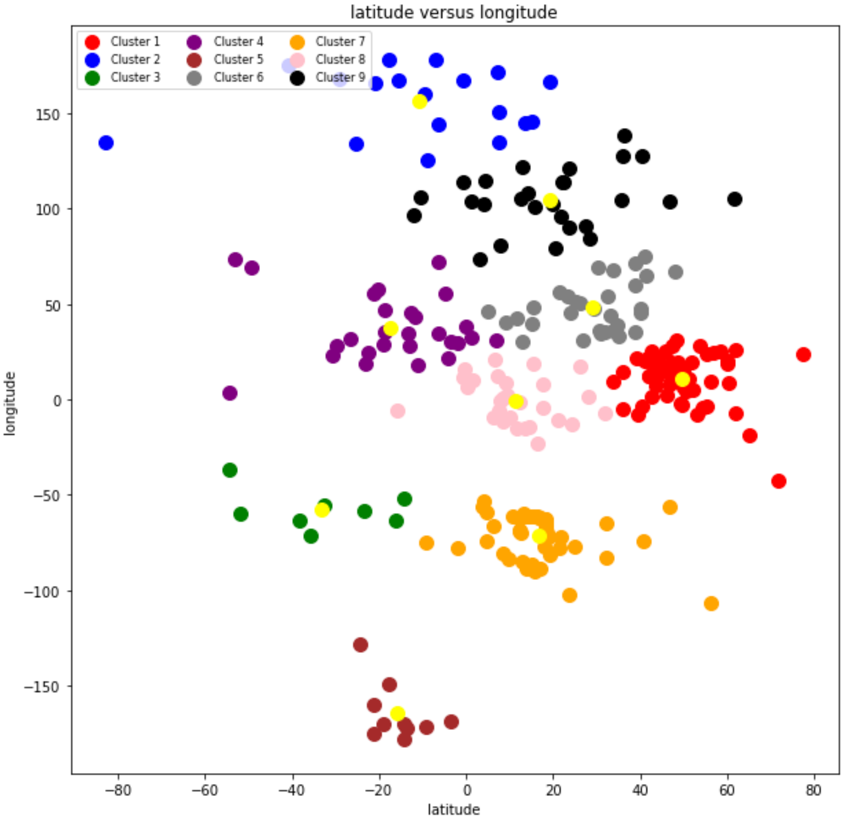
No. of K	No of Iteration	SSE to centroids
K=3	8	356365.87
K=4	8	272505.5
K=5	8	200584.03
K=6	8	157589.89
K=7	8	132052.61
K=8	6	107389.45
K=9	4	88939.85
K=10	9	81485.89

b.) By using n_clusters = 9 and random_state = 0 in K-Means algorithm, reaching the convergence with 4 iterations where the sum squares error is reaching minimal reduction.

c.) The continent and centroids value for all the 9 clusters as below.

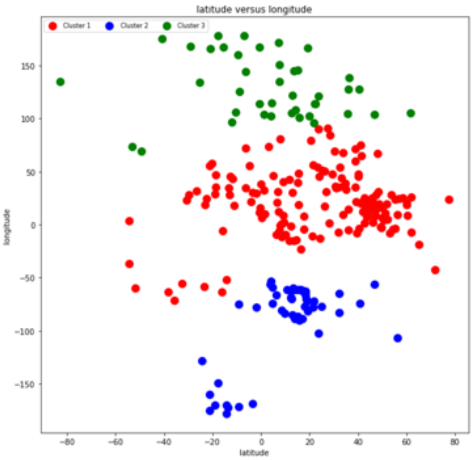
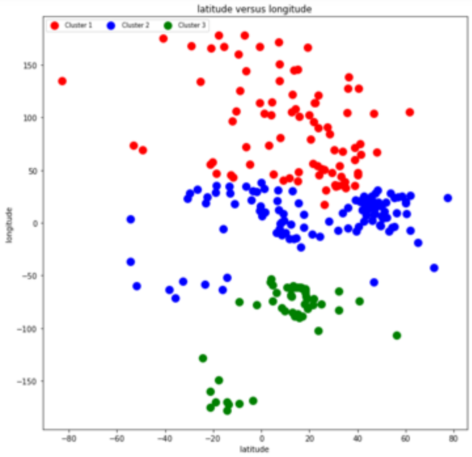
Cluster	Centroid	Continent	Label in scatterplot
0	[49.57259605 10.92636527]	Europe	Cluster 1
1	[-10.81353477 156.11725861]	Antarctic	Cluster 2
2	[-33.3509395 -57.62434537]	South America	Cluster 3
3	[-17.3056768 37.42555458]	East Africa	Cluster 4
4	[-15.84210807 -164.35100116]	Oceania	Cluster 5
5	[29.064347 48.28143745]	Middle East	Cluster 6
6	[16.78186493 -71.06985896]	North America	Cluster 7
7	[11.31033699 -0.44703661]	West Africa	Cluster 8
8	[19.26297174 104.62467122]]	Asia	Cluster 9

From the scatterplot, the continent of Africa has further divided into East Africa (Cluster 4) and West Africa (Cluster 8) because the countries stated are having large differences on longitude and they couldn’t group under same cluster. For Cluster 2 and Cluster 5 are consider under the continent of Australia but the countries are scattered too far from each other. Hence, did not group the countries under the continent of Australia but split them into two optimal clusters by Antarctic and Oceania.



Part 2: Hierarchical Clustering

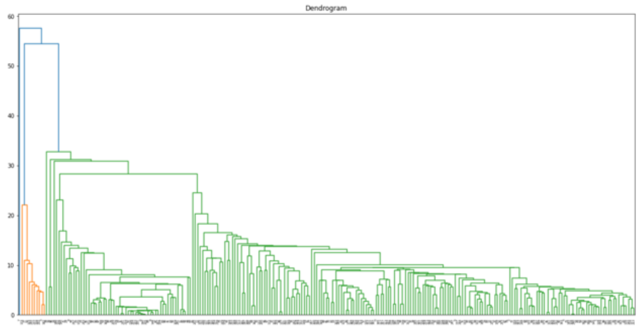
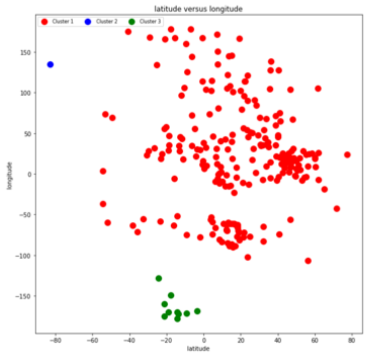
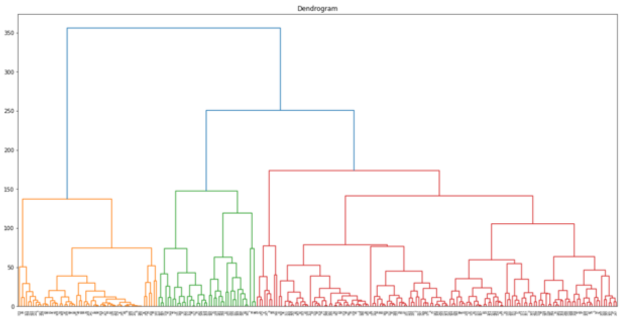
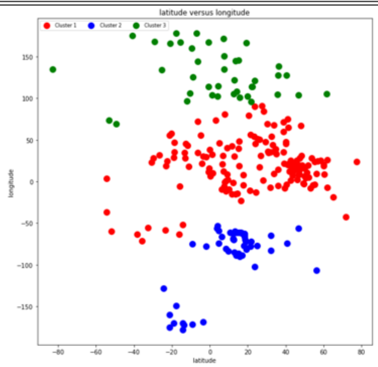
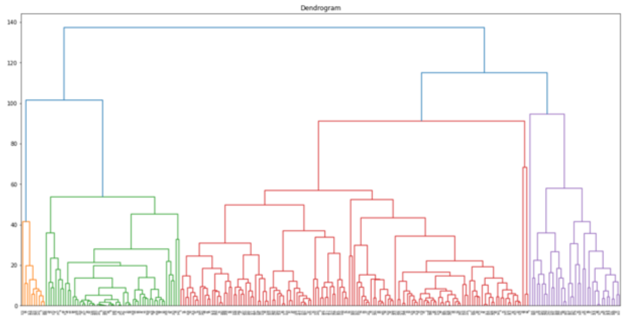
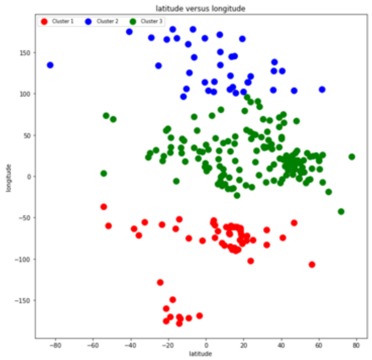
a.) Using the Agglomerative clustering algorithm, firstly compare whether any differences in clustering by using affinity 'Euclidean' and 'Manhattan'. The scatterplot showed that both parameters in affinity are actually producing same clustering outlook.

n_cluster = 3	Affinity = Euclidean	Affinity = Manhattan
linkage = 'complete'		

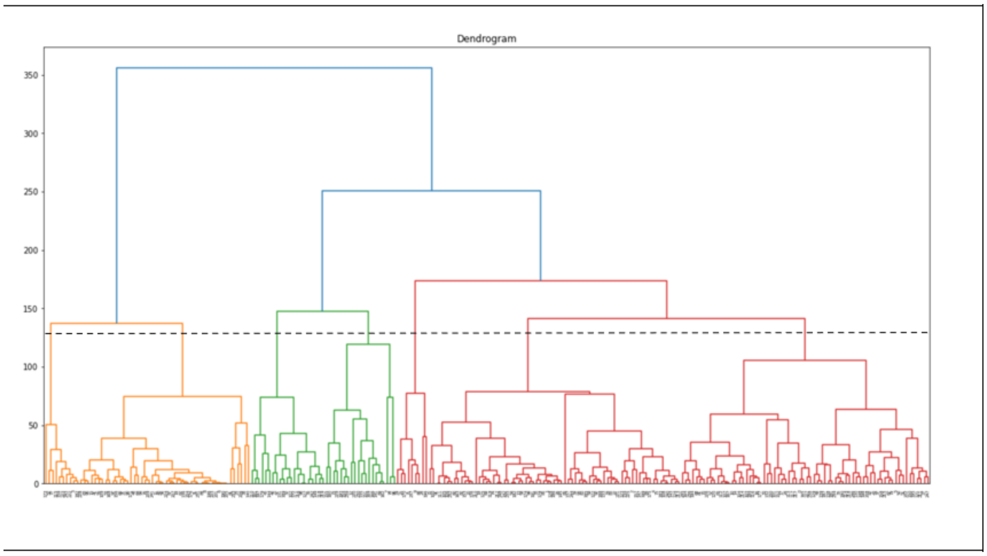
Hence, proceed the dendrogram by using affinity 'Euclidean' and all three types of linkage. From the table below, the better results of clustering is from the parameter linkage set to 'complete'.

Even it is showing slant to one side slightly but compare to the parameter with ‘single’ and ‘average’, it is better.

Dendrogram with parameter linkage ‘ single’ is definitely not a good cluster because it is highly slanted to one side and it is very difficult to split into balance clusters. The dendrogram with parameter linkage ‘ average’ is showing clusters concentrated in the center. It also make the clustering split more difficult.

Linkage	Dendrogram	Scatterplot (n-cluster = 3)
"Single"		
"Complete"		
"Average"		

b.) Referring to the dendrogram with linkage method ‘complete’, form the clustering into 7 clusters (refer below).



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

Part 3: Identify the best continent clustering

a.) K-means clustering which had produced 9 clusters are better and considered as best clusters. If follow Agglomerative clustering results, found that there are countries been grouped into inappropriate continents especially between Asia with Africa, and Oceania with Antarctic.

Cluster	Continent
0	Europe
1	Antarctic
2	South America
3	East Africa
4	Oceania
5	Middle East
6	North America
7	West Africa
8	Asia