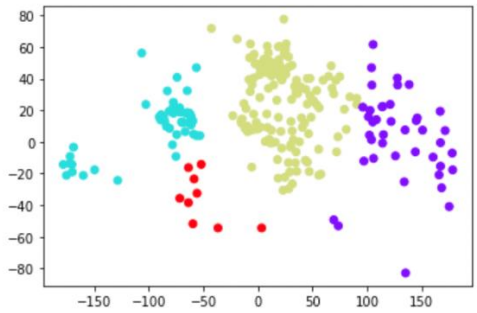
Assignment 2

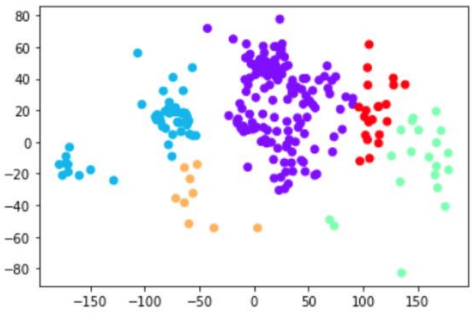
Data Preprocessing:
Unique identification: country_name is removed from the dataset

Scatterplot based on number of clusters:
Part 1: Agglomerative Clustering
Linkage: Complete

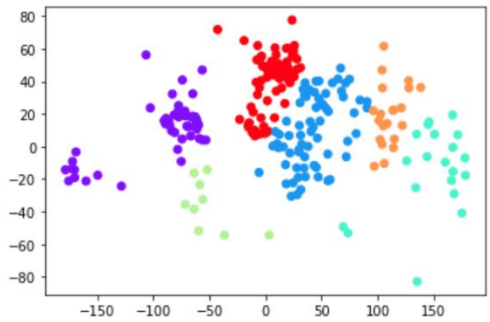
Cluster = 4



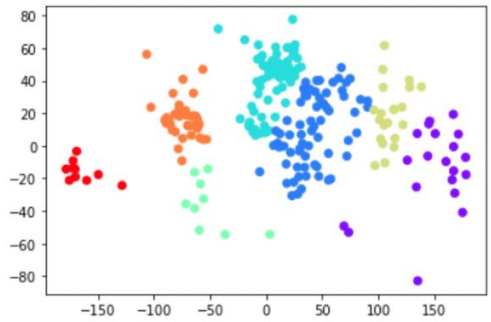
Cluster = 5



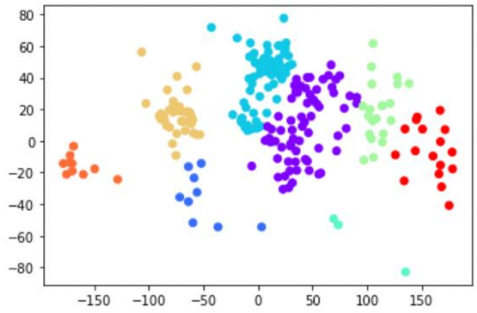
Cluster = 6



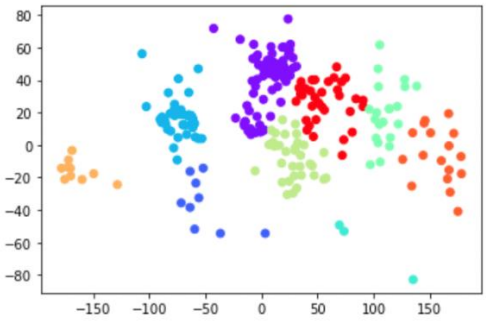
Cluster = 7



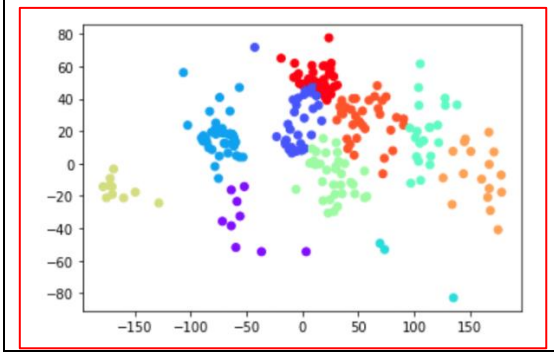
Cluster = 8



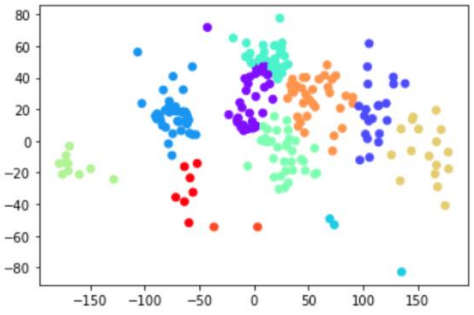
Cluster = 9



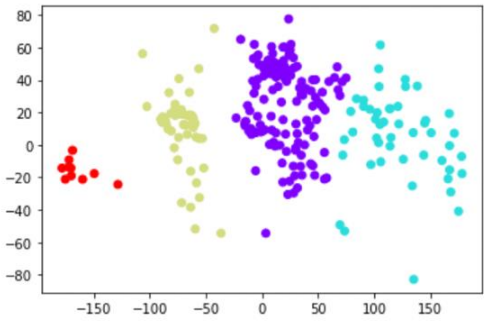
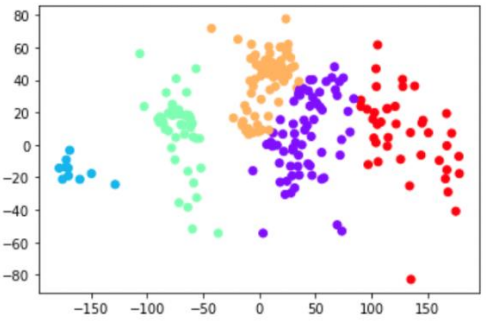
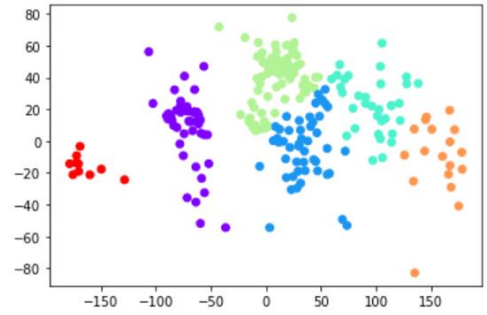
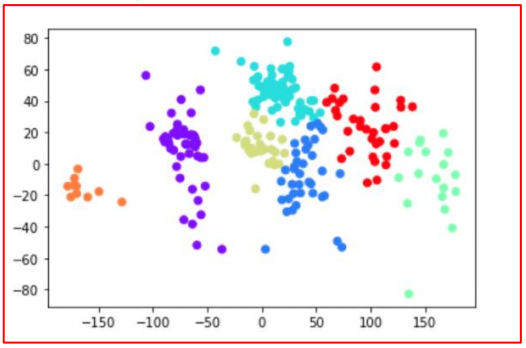
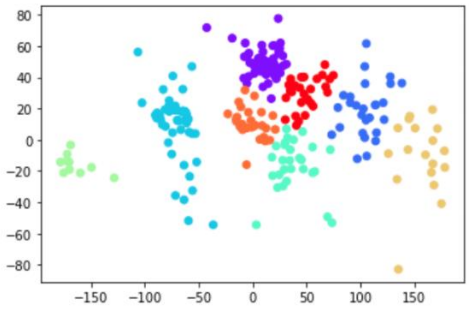
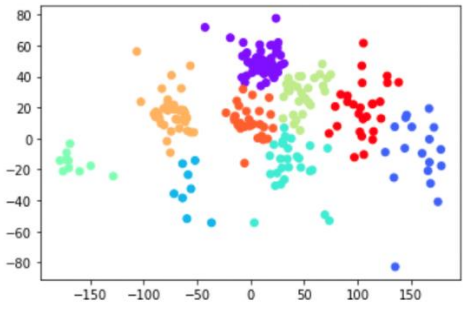
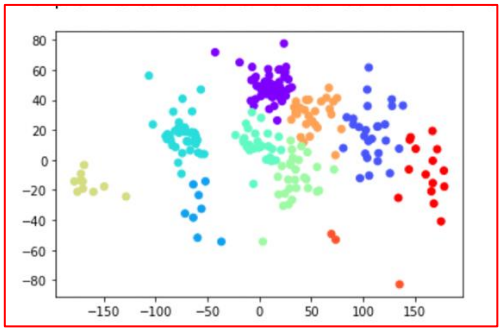
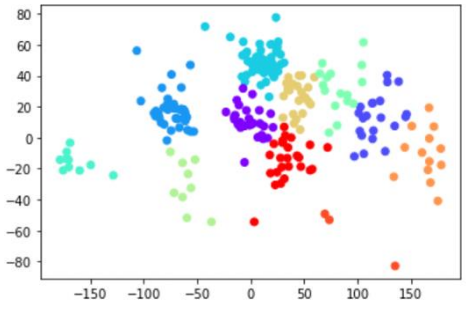
Cluster = 10



Cluster = 11



KMeans Clustering

Cluster = 4 ; SSE = 272505 	Cluster = 5 ; SSE = 200584 
Cluster = 6 ; SSE = 157589 	Cluster = 7 ; SSE = 132052 
Cluster = 8 ; SSE = 107389 	Cluster = 9 ; SSE = 88939 
Cluster = 10 ; SSE = 81485 	Cluster = 11 ; SSE = 74739 

Part2: Result & Discussion

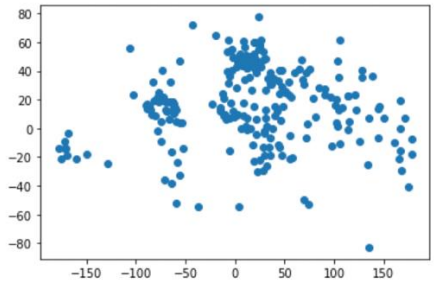
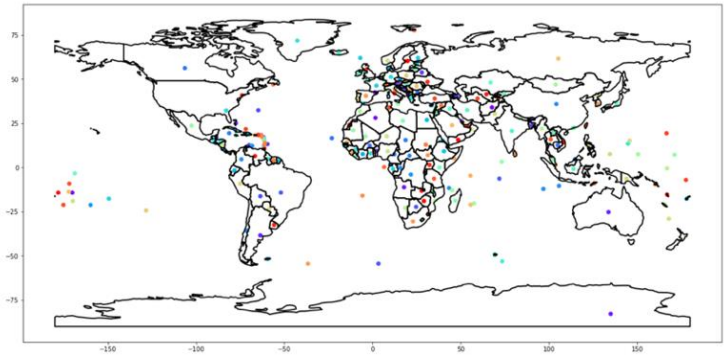


Figure 1.0: Image without clustering

Currently there's a total of 7 contingents in the world namely: Africa (56), Antarctica (1), Asia (49), Europe (46), North & Central America (22), South America (16), Oceania (14)
*Number in the () indicates the number of countries.

Theoretically, k = 7 should give the best clustering outcome. However, due to the close range of geographical location among the continents, it is challenging to separate the countries by continents with 100% accuracy. Based on cluster k =7 overlapping of continents is found in some countries such as Europe & Asia, Africa & Asia, North and Central America & South America. Dataset overlap with actual map can be represented in the figure below:



Part 2a: KMeans Clustering:

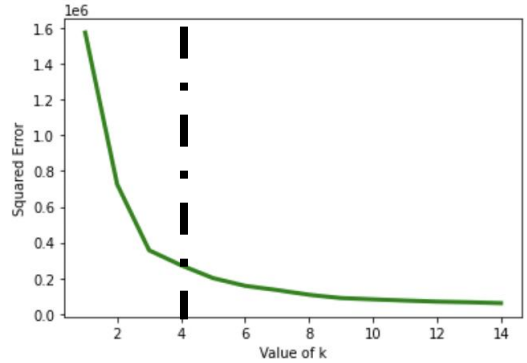
Several options are explored to find the optimum k-value. **Priority will be given to countries placed in the correct continent.**

Option 1: Elbow method

Option 2: Based on existing number of continents where k = 7

Option 3: The manual way of comparing countries in predicted cluster vs actual continent

Option 1: Elbow method



Cluster	Continent
0	Asia, South Africa, Europe
1	Antarctica, Oceania, Asia, South Africa,
2	North & Centra America, South America
3	Oceania

*multiple continents in cluster 0, 1, 2, 3

Value k = 4 is the optimum value

Option 2: Based on existing number of continents, k = 7

*country in bracket denotes outliers, not in the correct continent

Cluster	Continent
0	Combination of North & Central America, South America
1	Africa (Oman, Qatar, Saudi Arabia, UAE, Yemen)
2	Europe (Egypt, Estonia, Iran, Iraq, Israel, Libya, Syria, Tunisia)
3	Oceania (Antarctica)
4	South Africa
5	Oceania
6	Asia

Option 3: The manual way of comparing countries in predicted cluster vs actual continent, k =10

*country in bracket denotes outliers, not in the correct continent

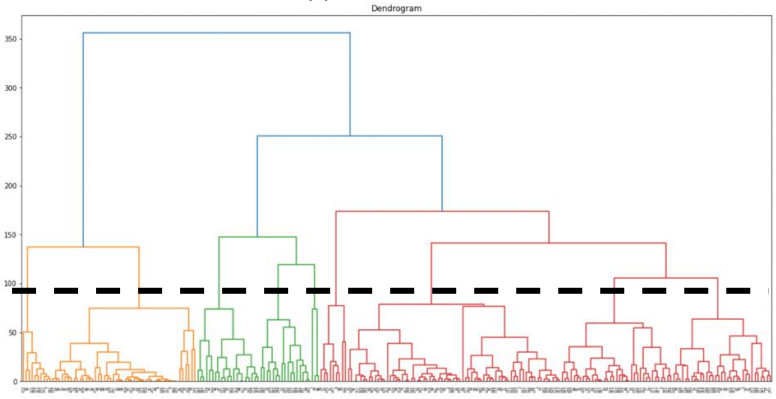
Cluster	Continent	Centroid
0	Europe (Libya, Tunisia)	49.13415273, 11.04527028
1	Asia	18.33127919, 108.62935637
2	South America	-33.3509395, -57.62434537
3	North & Central America (Ecuador, Guyana, Peru, Venezuela)	16.78186493, -71.06985896
4	South Africa	10.75386428, -1.10167986
5	South Africa	-12.06182649, 35.16229227
6	Oceania	-15.84210807, -164.351001
7	Asia (Cyprus, Eritrea, Turkey)	29.9210421, 51.06670653
8	Antarctica	-61.74164267, 92.6175717
9	Oceania	-7.36144245, 160.9867064

Conclusion:

When k = 4 & k =7, there's overlapping of countries in continents vs actual continents. Therefore, higher k value where **K = 10 is selected as the best k value for KMeans** clustering technique as it can cluster countries into right continent with minimum difference against actual continent (except for a few outliers). Further increase k-value to 11 does not reduce the number of outliers in respective continent compared to k = 10, despite having lower SSE value as similar group of countries from the same continents are divided into new clusters.

Part 2b: Agglomerative Clustering ; Linkage = Complete

Best Agglomerative n_cluster = 10 (cluster stop here since increase to n_cluster = 11 does not reduce the number of outliers as recorded in n_cluster =10). Comparison against linkage 'Average' is made. Result is shown in appendix section.



Cluster	Continent	Outliers
0	South America	-
1	South Africa	France, Italy, Lichtenstein, Malta, Monaco, Portugal Spain, Switzerland, Tunisia, Vatican City
2	North & Central America	Ecuador, French Guiana, Peru, Venezuela
3	Antarctica	-
4	Asia	-
5	South Africa	-
6	Oceania	-
7	Oceania	-
8	Asia	Cyprus, Eritrea, Ethiopia, Somalia, Turkey
9	Europe	-

Part 3: Identify the best continent clustering

Best continent clustering selected: KMeans.

Justification: In comparison of KMeans to Agglomerative Clustering, KMeans is better at clustering countries in the correct continent with lesser number of outliers vs actual continents as summarized in both scatterplot & summary table.

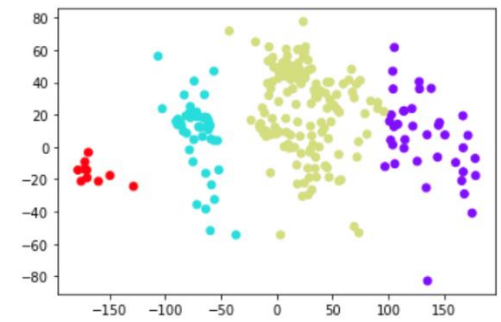
Cluster	Continent
0	Europe
1	Asia
2	South America
3	North & Central America
4	South Africa
5	South Africa
6	Oceania
7	Asia
8	Antarctica
9	Oceania

Reference:

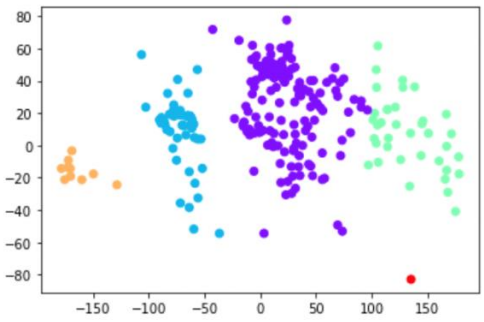
https://blog.jcharistech.com/2020/07/20/clustering-countries-into-continents-using-unsupervised-machine-learning/

Appendix:
Agglomerative Clustering; Linkage = Average

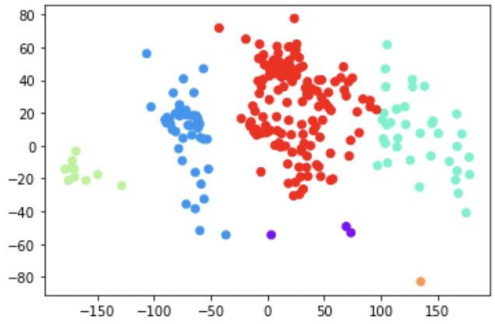
Cluster = 4



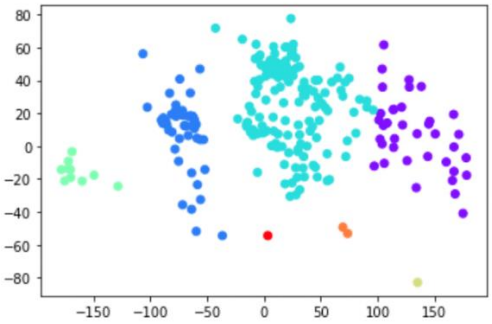
Cluster = 5



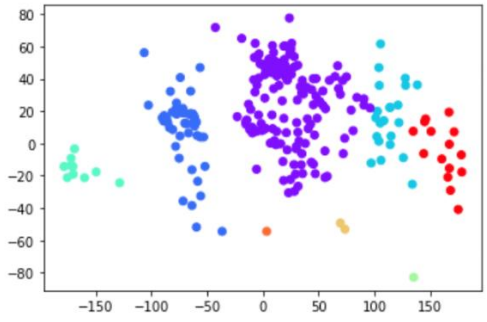
Cluster = 6



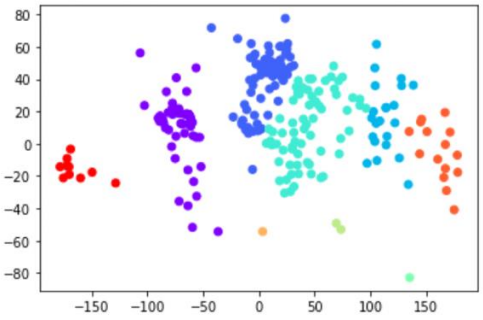
Cluster = 7



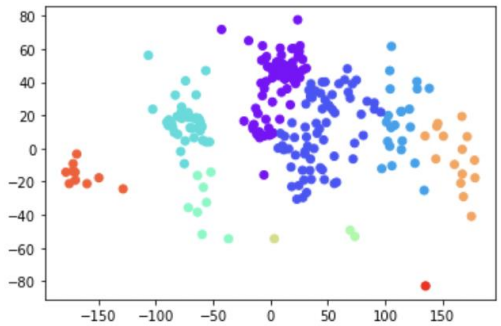
Cluster = 8



Cluster 9



Cluster = 10



Cluster = 11