CDS503: Machine Learning
Assignment 2: Clustering

Part 1: K-Means Clustering

a)

The elbow method was used to determine an appropriate k-value. Here the sum of squared errors of the respective k values are considered. Of course, the error becomes smaller the larger the value of k becomes, but there is a point at which the drop becomes significantly smaller. This point is called the elbow point. In figure 1 it can be seen that k = 5 and k = 8 are possible points.

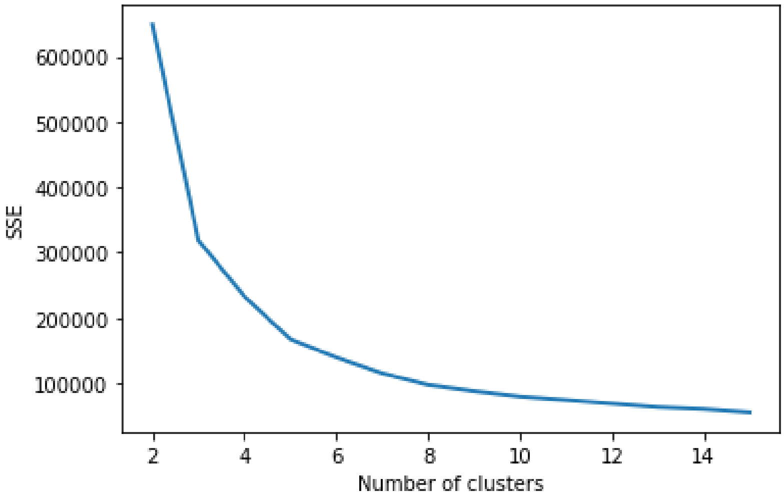

Figure 1: Elbow method

To find a suitable value, the data were also considered in the real-life context. Each cluster should represent a continent, so k should correspond to the number of continents. However, there is room for maneuver here because there are four to seven continents, depending on the consideration.

Considering the two arguments mentioned above, k = 5 was chosen. Which means that the continents Europe, Asia, America, Africa, and Oceania should be represented by the clusters. This is illustrated again by figure 2. The continent of Antarctica is ignored in this case, as it does only contain one country at all.

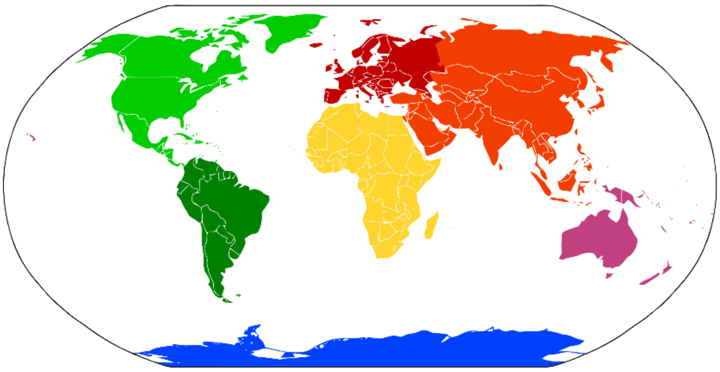

Figure 2: continents [Wikipedia.com]

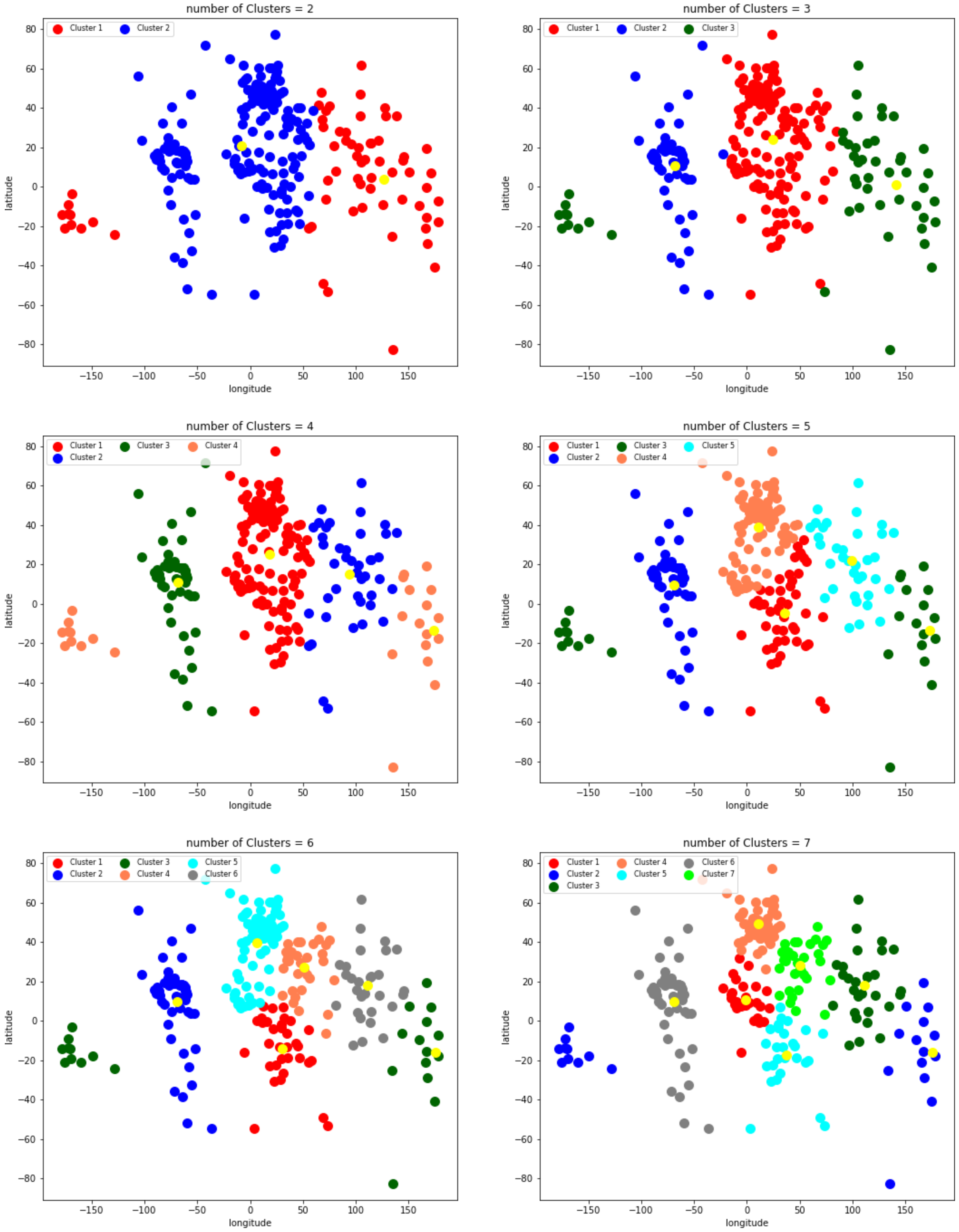

Figure 3: plots of the maps depending on the k-value

As can be observed in Figure 3, it is not useful to set k < 5, otherwise continents will merge with each other, and false results are inevitable. As an example, at k = 4, Europe and Africa are combined into one cluster; at smaller values, this problem is exacerbated. With k > 5, on the other hand, continents are divided. For example, with k = 6, the continent of Asia is divided into different clusters, so that it can be said that k > 5 is not useful.

This reinforces the assumption that k = 5 is the best choice.

b) Final parameters

Several hyperparameters were tried, but they did not improve the result. It would have been possible to set the initial centroids to the different continents, but this would have manipulated the results.

Finally, the following parameters were used: n_clusters = 5, init = random, algorithm = full and random_state = 0.

Outcomes:

Centroids:

```
[ 21.98576744  98.8705185 ]
[ 38.95511233  10.72265493]
[ -4.61213085  35.91495756]
[-13.52713344 173.31886863]
[  9.62003573 -69.14907131]
```

Sum of squared distances to centroids: 167176.92210148106

Number of iterations run: 13

This is the best clustering as the continents are best captured here.

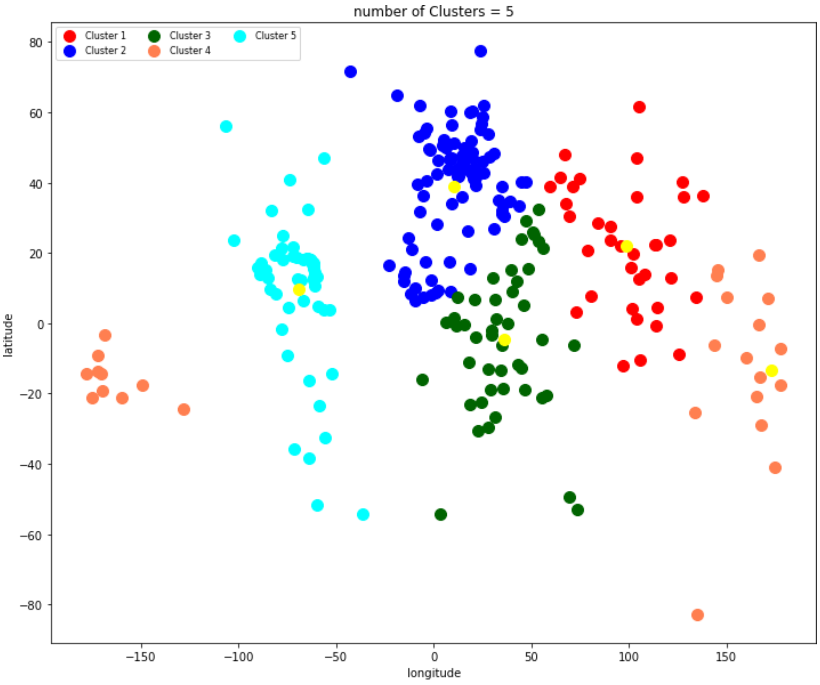

Figure 4: best clustering k-means

c) Clusters

Table 1: K-Means clusters

Cluster	Centroid	Continent
0	21.98576744 98.8705185	Asia
1	38.95511233 10.72265493	Europe
2	-4.61213085 35.91495756	Africa
3	-13.52713344 173.31886863	Ozeania
4	9.62003573 -69.14907131	America

Part 2: Hierarchical Clustering

a)

First, different numbers of clusters were tried out, with 5 clusters fitting best, as with K-Means.

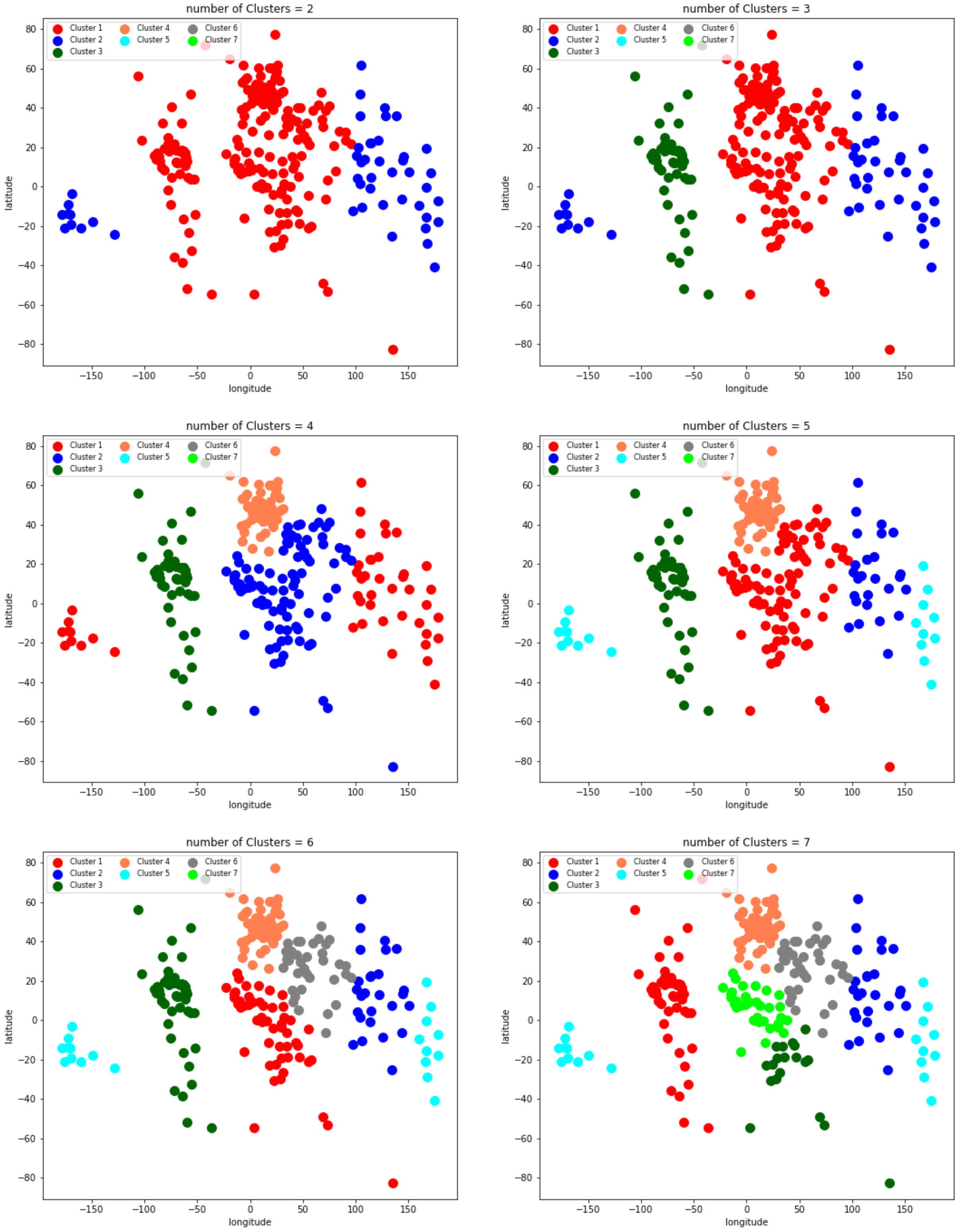

Figure 5: plots of the maps depending on the number of clusters

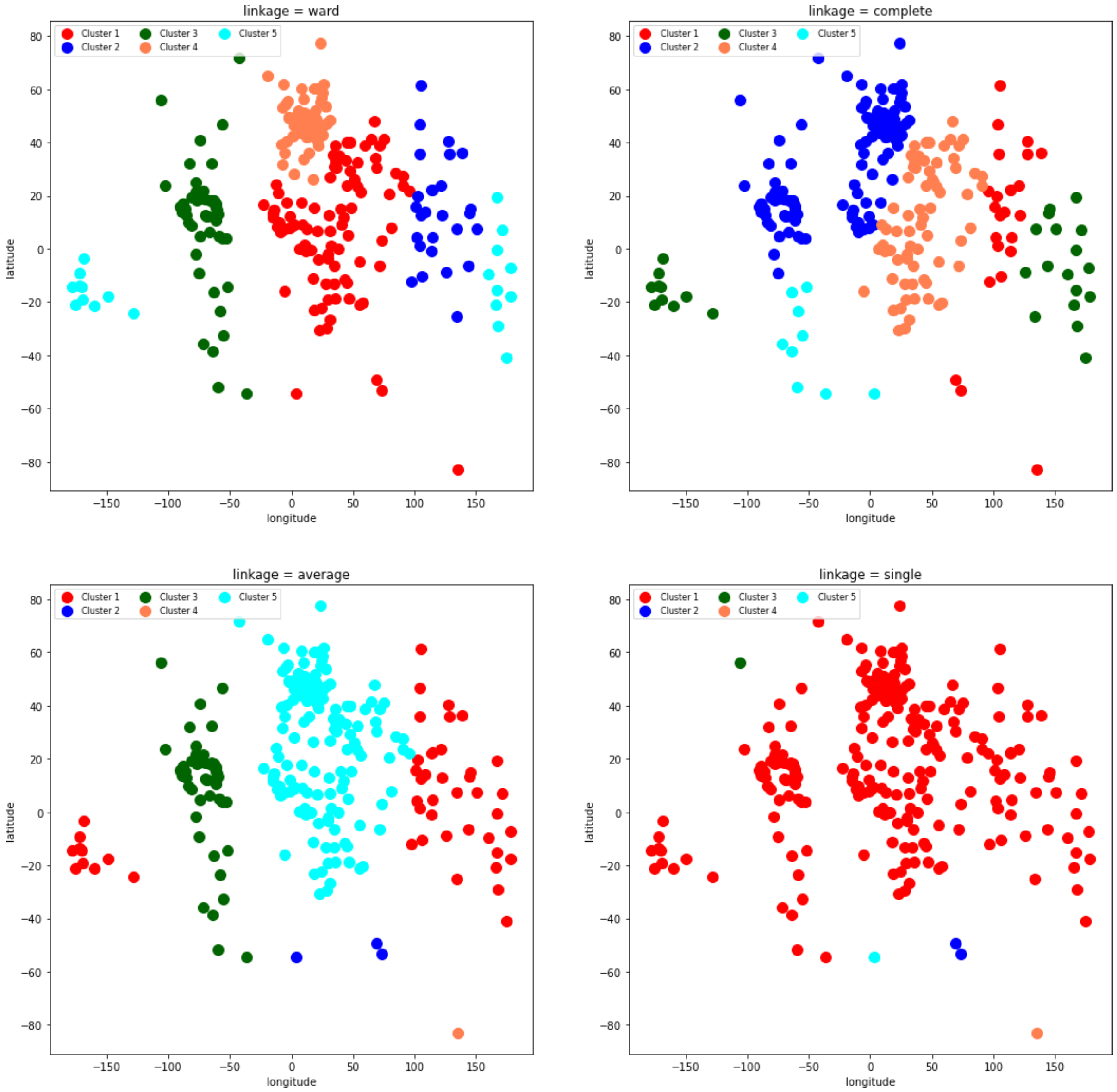

Figure 6: plots of the maps depending on the linkage

Then, for this number of clusters, the other parameters were optimized. For the parameter linkage the option ward is best suited. (See figure 6) Single linkage tends to form one very large cluster, since the shortest distance is considered here in each case. Complete and average linkage do not manage to distinguish Africa and Europe. This is a difficult task, since these continents are very close to each other. Ward linkage, however, manages to overcome this problem to some extent, since it tries to reduce the variance within a cluster and since Europe consists of small countries in terms of area, i.e., many data points close together, a cluster can be formed here. All linkages fail to cleanly separate Africa from Asia or eastern Europe. This is due to the fact that the continents lie directly next to each other here and purely from the point of view of the coordinates no border is to be expected here. For ward linkage there is only euclidean as an option for the affinity. The other parameters have no remarkable influence on the result.

The best result is obtained with the parameters n_clusters = 5, affinity = Euclidean and linkage = ward, because the continents are best separated from each other.

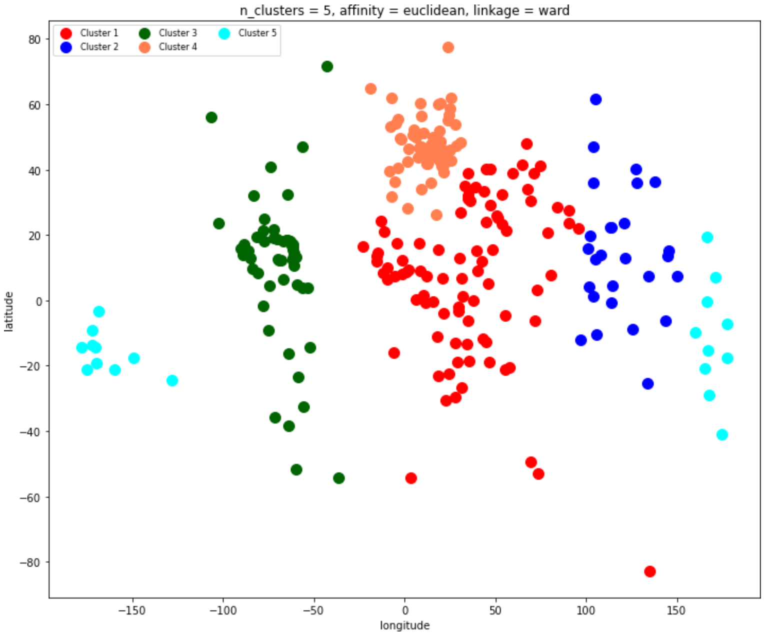

Figure 7: best hierarchical clustering

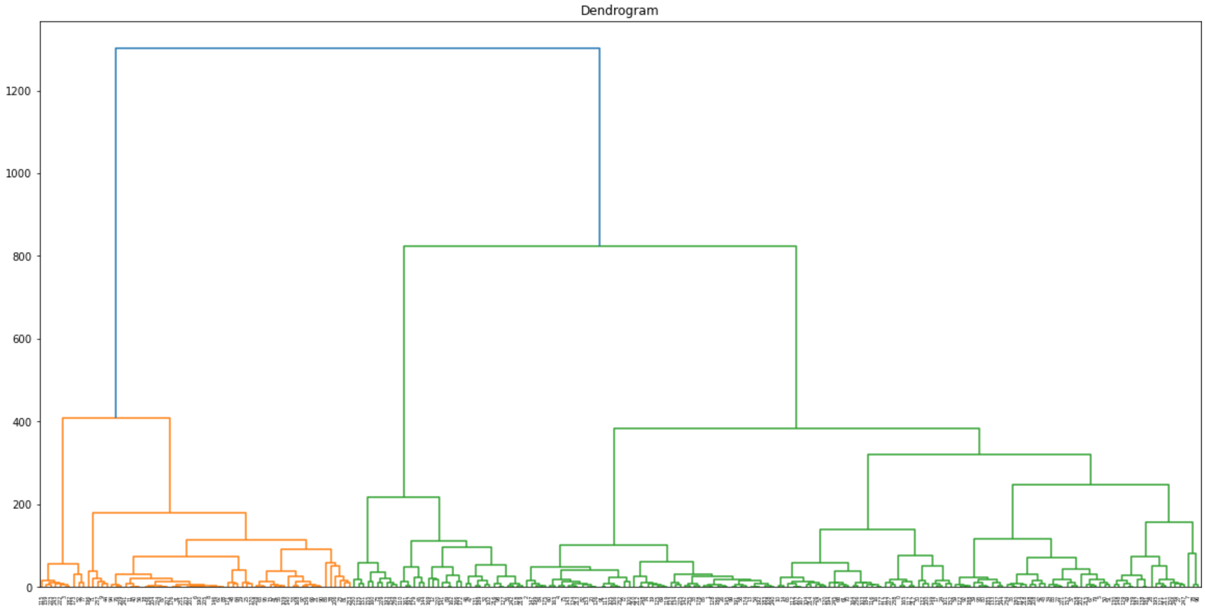

Figure 8: dendrogram

Table 2: Hierarchical clusters

Cluster	Continent
0	Africa
1	Asia
2	America
3	Europe
4	Ozeania

Part 3: best continent clustering

The best clustering is the hierarchical clustering with the following parameters: n_clusters = 5, affinity = Euclidean and linkage = ward.

Table 3: best clustering

Cluster	Continent
0	Africa
1	Asia
2	America
3	Europe
4	Ozeania