Project 2 - Clustering

Kostas botsi

June 2024

1 Introduction I

The purpose of this analysis is to determine how the municipalities in the country are grouped based on their age composition. We aim to identify the number of such groups and to characterize them. Additionally, the analysis assesses the quality of the resulting groupings.

2 Chapter II

The data we will analyze pertains to Portugal. While the dataset includes both NUTS (Nomenclature of Territorial Units for Statistics) and municipalities for geographical grouping, our analysis will focus solely on the municipalities. The variables included in the dataset are as follows:

Geographical group	2 levels: NUTS and Municipaty
Age group	18 levels: 0-04, 05-09, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85+
Territories	349 in total

Table 1: Variables into the data portugal

3 Data processing III

The data is in an Excel file, which we download and load into the statistical package R. Both the data processing and analysis are conducted using R.

Before starting the analysis, we read the data from the Excel file and converted it into a data frame. Since there is no data available for the year 2000, we removed the columns corresponding to the year 2000 for each age group. We then named the remaining columns according to their respective age groups.

Additionally, two numeric variables were mistakenly read as characters, so we converted these variables back to numeric format.

4 Data analysis IV

In this section, we present two methods for clustering analysis to explore how the country's municipalities are grouped based on age composition.The first method is hierarchical clustering. We explore different linkage methods to determine which one best separates the groups and how good the clustering is.

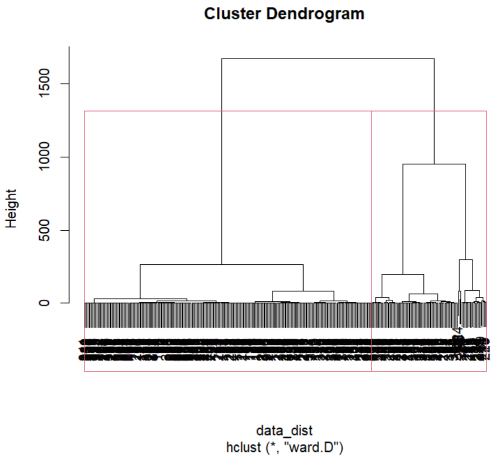

Figure 1: Hierarchical cluster linkage method Ward

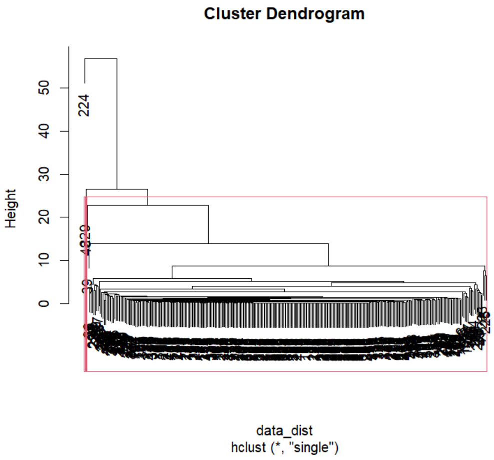

Figure 2: Hierarchical cluster linkage method single

We present two linkage methods for hierarchical clustering. The first method is the Ward method, which clearly separates the two clusters. The second method is the single linkage method, which does not clearly show the number of clusters.

In the following graph we show another two linkage method which is the average and the complete method There is presenting the two graphs,with the average linkage and the complete linkage,which dont show clearly how many are the cluster for this methods. We present the hierarchical clustering analysis to illustrate how the municipalities' territories are grouped based on age composition.

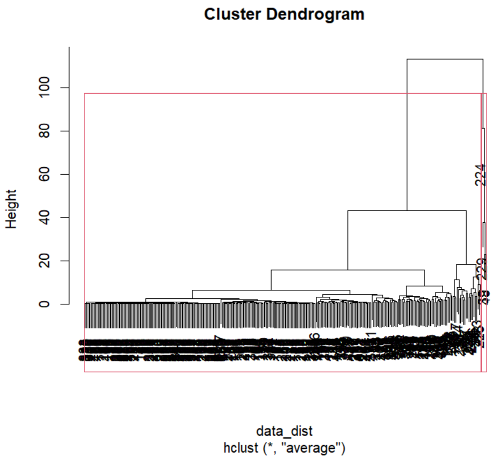

Figure 3: Hierarchical cluster linkage method average

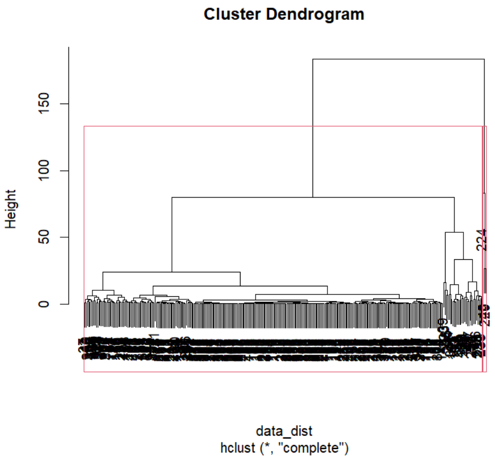

Figure 4: Hierarchical cluster linkage method complete

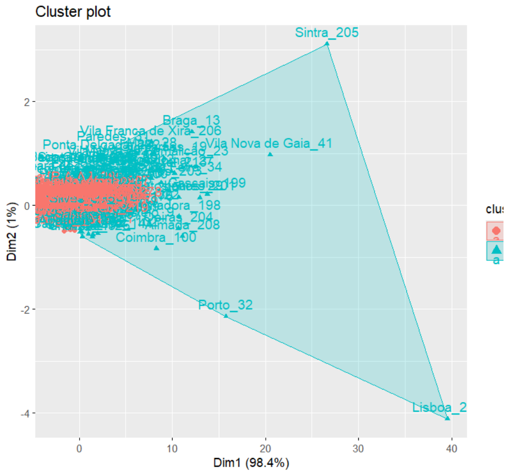

Figure 5: cluster with the wald method

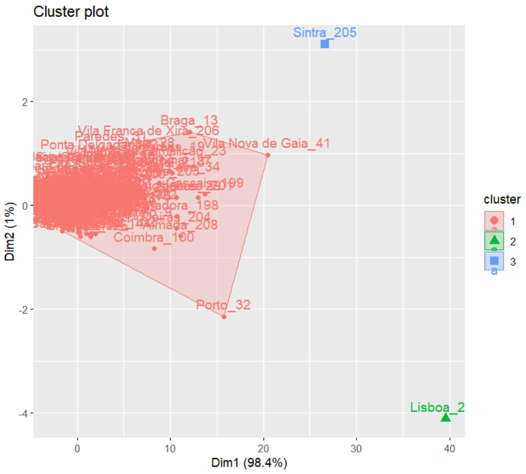

Figure 6: cluster with the single method

From the graphs, we observe that with the Ward linkage method, the two clusters are not well separated because they overlap. With the single linkage method, the three clusters are not clearly separated, as one cluster contains most of the territories and the another two are only one observation in each cluster.
Next we present the another two graphs with the average and the complete linkage method to see if the clusters are good seperate.

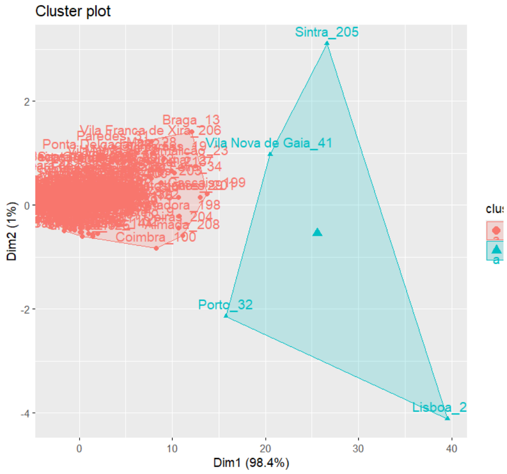

Figure 7: cluster with the average method

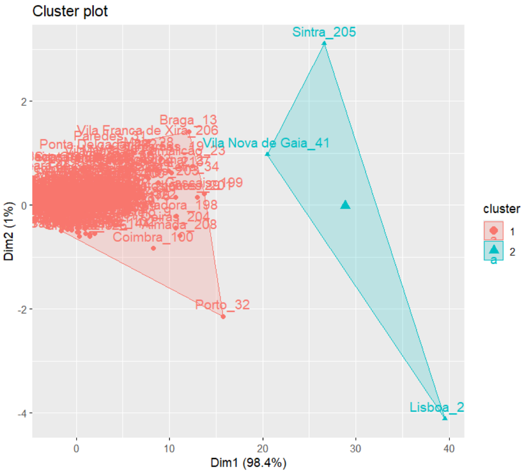

Figure 8: cluster with the complete method

The complete and average linkage methods better separate the two clusters, but one cluster has many more observations compared to the other.

Next, we should evaluate the quality of the clusters obtained with different linkage methods. We do this using silhouette values, which range from 0 to 1. Values close to 1 indicate better clustering.

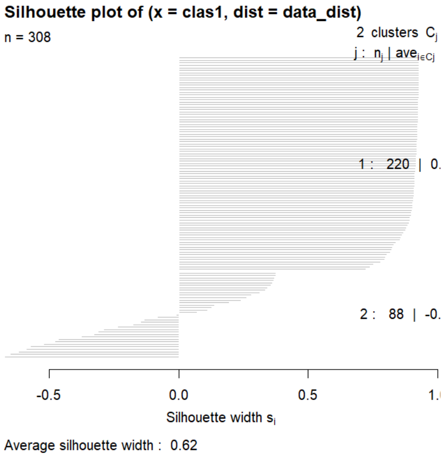

Figure 9: silhouette values for ward method

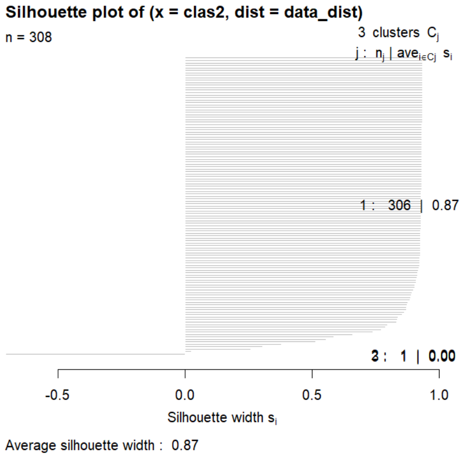

Figure 10: silhouette values for method Single

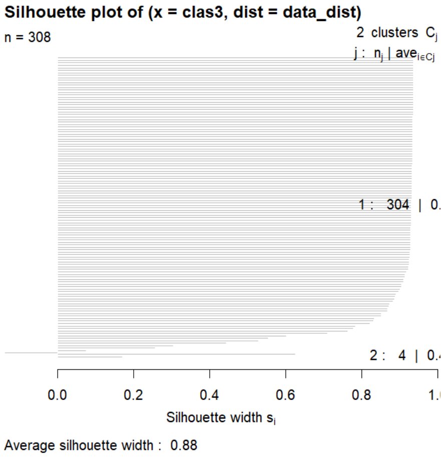

Figure 11: silhouette values for method average

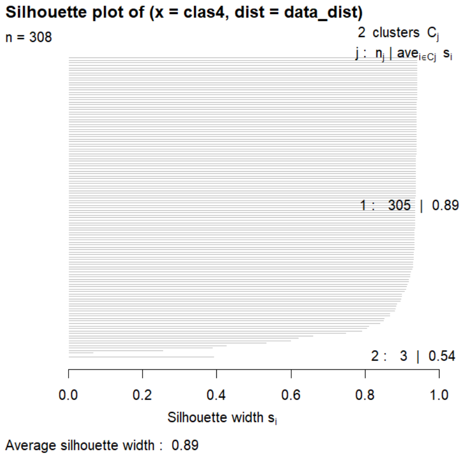

Figure 12: silhouette values for method complete

From the silhouente plot we see that the method average has the highest average.So we conclude for the hierachical method the the clusters are 2 with linkage method average.But is not the optimal because most of the observation are on the one cluster

Next method that we use for clustering is the k-means method. To start we need to see how many cluster should start for the kmeans method.To do this we must do the elbow method

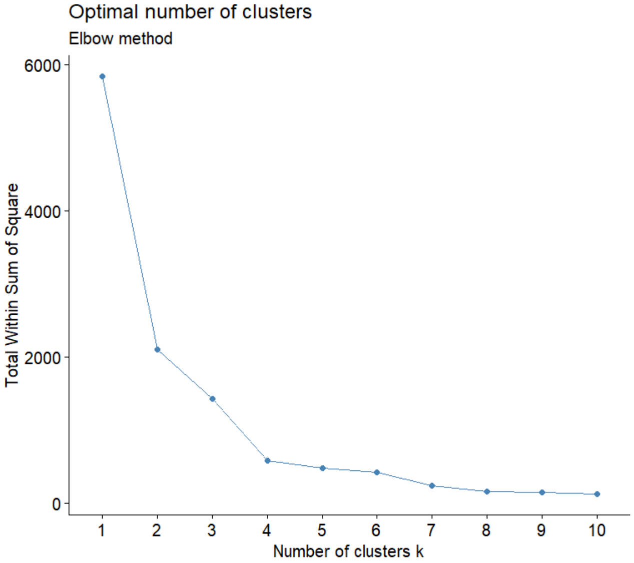

Figure 13: Elbow graph

From this graph we see how many cluster we need for the k-means cluster.We see that the first change is at cluster 2 so for the k-means we need 2 cluster. Then we presend the graph for the two cluster.

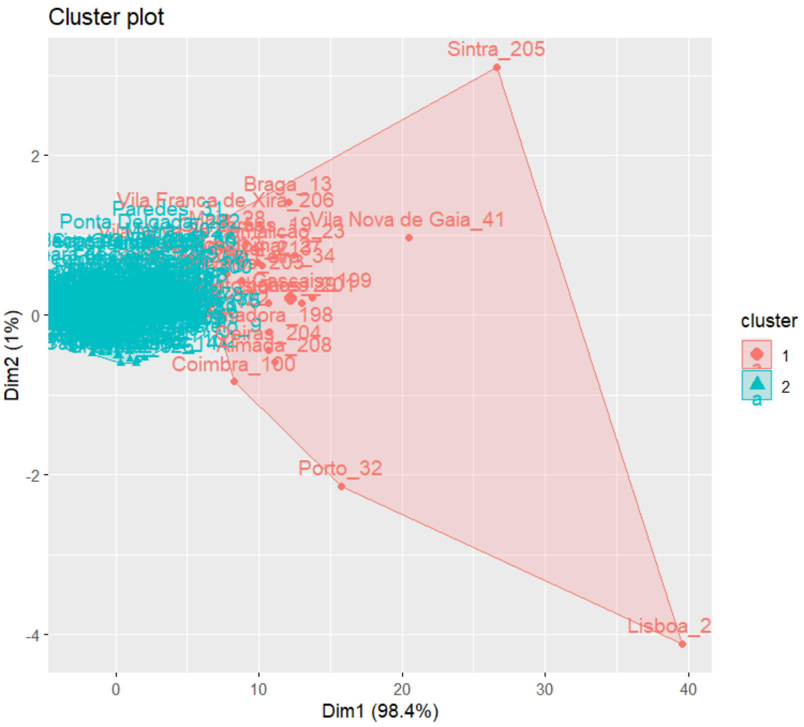

Figure 14: K-means cluster

From this plot we see that the territories are better seperate and dont overlap the clusters,the clusters seem to be good seperated.
Next we should to see if the the k-means cluster are good,with the silhouente average value that we present in the following graph

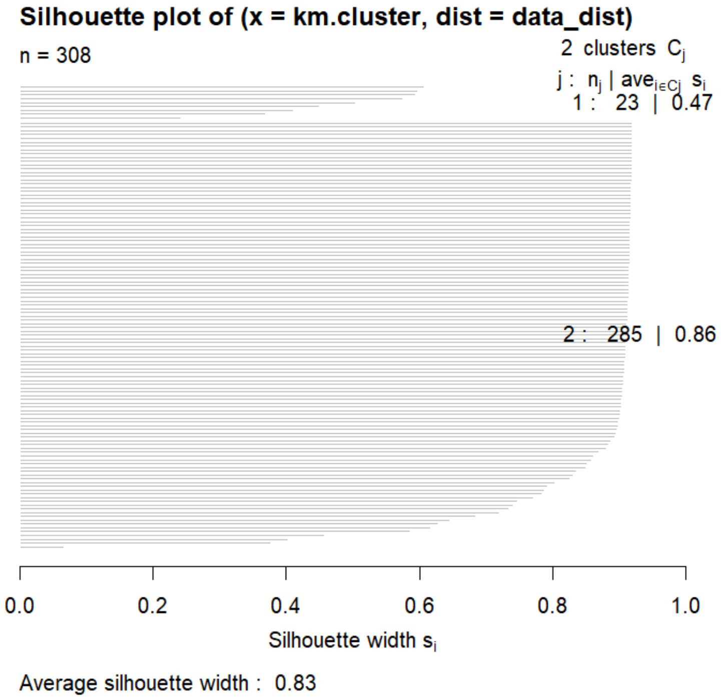

Figure 15: Average silhouente value

We observe a high average silhouette value of approximately 83%, which indicates that the clusters are well separated. This suggests that the clustering method effectively distinguishes between groups based on age.

5 Conclusion VII

Based on different clustering methods, the best approach for clustering the data was the K-means method. According to the graph, we achieved a good separation into two clusters. Additionally, the average silhouette values are high, indicating that the clusters are well-separated.

The main insights from the clustering analysis are as follows:

One cluster contains 285 territories, indicating that these municipalities have a predominantly younger population. The other cluster contains only 28 municipalities, which tend to have an older population. This is primarily because these municipalities are located farther from the city and have fewer young people.