ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Project II

Clustering of Municipal Districts Based on Age Composition

Christina Kataki

Instructor:
Dimitrios Karlis

June 2024

Contents

1 Introduction

Understanding the age composition of a population is crucial for gaining insights into the demographic characteristics and social structure of a country. This report focuses on analyzing the age composition of Portugal's population using data from the 2022 census, specifically at the level of municipal districts. The objective is to identify clusters of municipalities based on their age composition, determine the optimal number of clusters, and provide a detailed characterization of each cluster.

What is Clustering?

The task of grouping data points based on their similarity with each other is called Clustering or Cluster Analysis. This method falls under the branch of Unsupervised Learning, which aims at gaining insights from unlabelled data points. Unlike supervised learning, clustering does not have a target variable. Instead, it seeks to identify inherent structures in the data.

Clustering aims to form groups of homogeneous data points from a heterogeneous dataset. It evaluates the similarity between data points using metrics such as Euclidean distance, Manhattan distance, etc., and then groups the points with the highest similarity scores together.

Two distinct clustering techniques are employed and compared:

- K-means Clustering
- Hierarchical Clustering

The analysis will be presented in a manner that is accessible to readers with limited background in statistics, focusing on clear explanations and essential visualizations. The report is structured as follows: it begins with an Exploratory Data Analysis (EDA) to understand the dataset. This is followed by separate chapters detailing each clustering methodology—K-means, Hierarchical Clustering—and their results. A comparison of the two approaches is then provided. The report concludes with recommendations based on the analysis.

2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is crucial for developing effective machine learning models. It focuses on investigating datasets to identify patterns, detect anomalies, and understand the data's structure. This section uses the 2022 census data of Portugal's municipal districts to explore the age composition of the population. Using visual and statistical methods, the goal is to gain insights and ensure a comprehensive understanding of the data before proceeding to more advanced clustering techniques.

The original dataset was in XLS format, which was converted to CSV for ease of analysis. The initial CSV conversion encountered issues, so it was further altered and cleaned to ensure it is accessible and suitable for analysis.

In the final dataset, all columns containing data from the year 2000 were removed due to a lack of records for this period. Additionally, only rows corresponding to municipalities were retained. The reason that rows that are not municipalities (like NUTS I, NUTS II, NUTS III) were excluded was that they can provide additional context but might interfere with clustering focused on municipalities. Furthermore, no missing values were found in the final dataset.

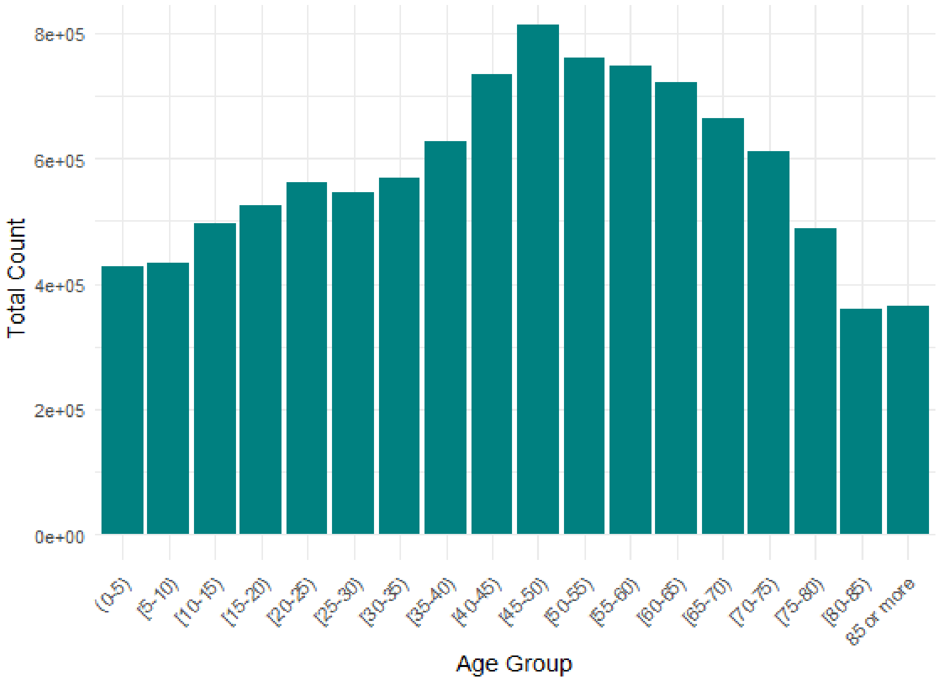


Figure 1: Total Population Distribution by Age Group

The bar plot (Figure 1) visualizes the total population distribution across different age groups in Portugal's municipal districts. The bars represent the count for each age group. The plot indicates that the middle-aged groups (40-60 years) have the highest total population, peaking around 40-45 years. The population gradually decreases in the older age groups, with the lowest counts observed in the 80-85 and 85 or more age groups. This visual representation shows a concentration of the population in middle age and a decline in the elderly population.

Continuously, the map (Figure 2) visualizes the most common age group in each municipality of Portugal, based on the census data. Each municipality is color-coded according to the predominant age group, with the legend indicating the age ranges and corresponding colors. This map reveals regional patterns in age distribution, such as areas with higher prevalence of older populations (e.g., municipalities with more individuals in the 75-80 and 85 or more age groups) and areas with younger or middle-aged populations.

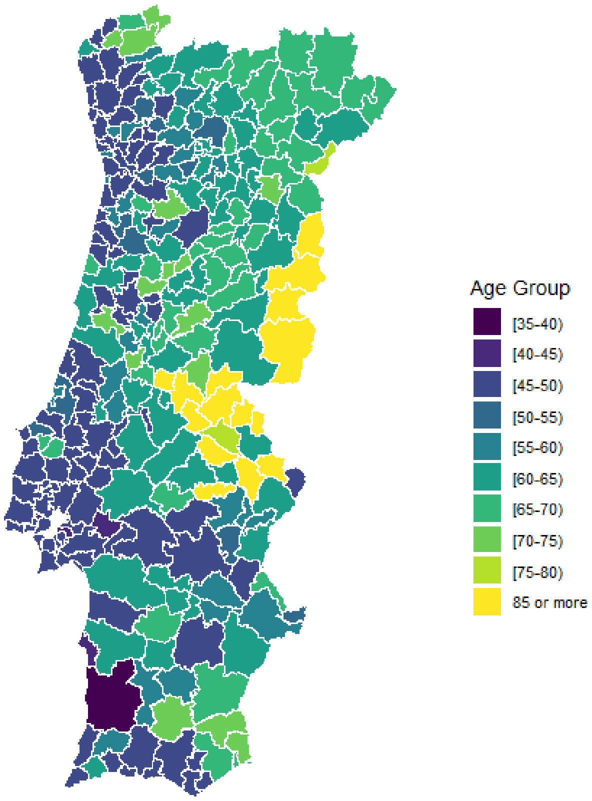


Figure 2: Most Common Age Group in Each Municipality of Portugal

Finally, the top 10 municipalities by total population were identified from the dataset. The population for each age group within these municipalities was normalized by dividing the age group population by the total population of the respective municipality. This normalization allows for a comparative view of age distribution across different municipalities. In the heatmap (Figure 3), the x-axis represents the age groups, the y-axis lists the municipalities, and the color intensity indicates the proportion of the population within each age group. Darker shades represent higher proportions, while lighter shades indicate lower proportions. This visualization provides a clear overview of the age group distribution across the top 10 municipalities.

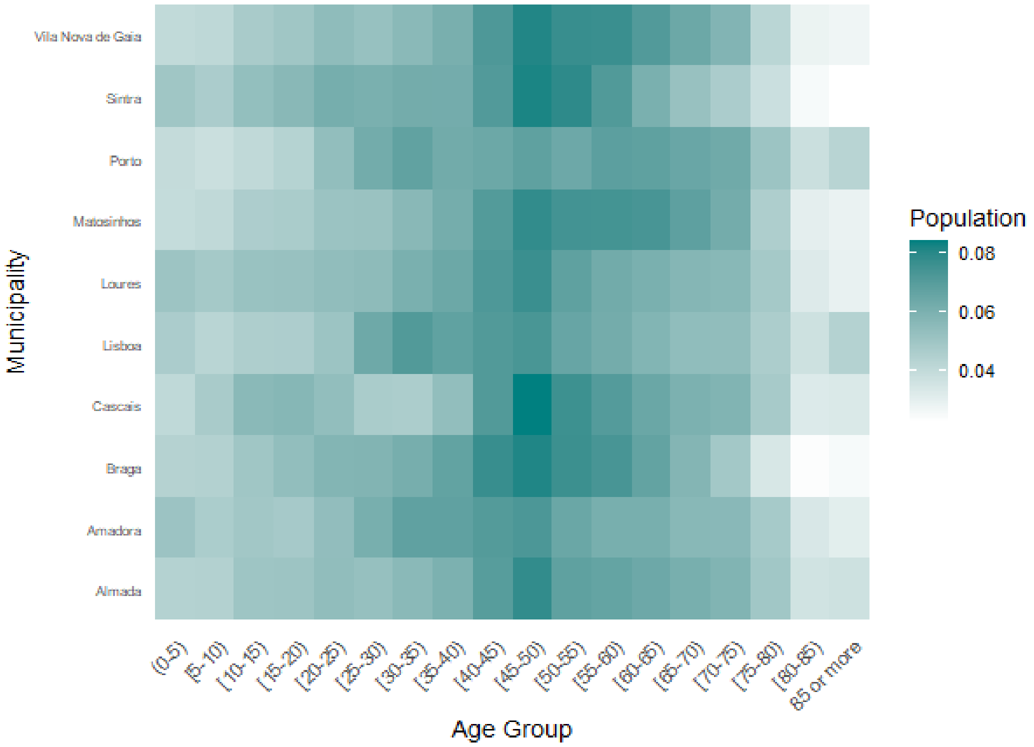


Figure 3: Normalized Heatmap of Age Group Distribution for Top 10 Municipalities

3 Clustering Methodologies and Results

3.1 K-means Clustering

K-means clustering is a popular algorithm used to partition a dataset into a specified number of clusters, denoted by `K`. The algorithm operates iteratively to allocate each observation to the cluster with the nearest mean, which serves as the cluster center. These cluster centers are updated by computing the mean of the observations assigned to each cluster. The iterative process continues until the cluster assignments no longer change significantly, indicating convergence.

The fundamental principle behind K-means clustering is to minimize the within-cluster sum of squares (WCSS), effectively ensuring that observations within a cluster are as similar as possible. One of the critical factors influencing the performance of K-means is the initial selection of cluster centers. Different initializations can lead to different final clusters, necessitating multiple runs with different starting points to achieve a stable and optimal solution.

K-means clustering is efficient in terms of memory usage and can be parallelized, making it suitable for large datasets. However, it is sensitive to the scale of the data, and non-linear transformations can significantly alter the clustering results. Additionally, there is no guarantee that K-means will find the global optimal solution; it may converge to a local minimum instead.

In practical applications, it is essential to run K-means with multiple initializations to increase the likelihood of finding a better clustering solution. K-means can also be complemented with other clustering algorithms to enhance results, and various variants such as K-medoids and K-centroids can be used to handle different types of data and clustering requirements.

In this assignment, the data was scaled to ensure that each numeric variable contributed equally to the clustering process. This step was crucial to avoid any single variable dominating the clustering results due to differences in scale. To determine the optimal number of clusters, the Elbow Method was used.

Elbow Method Analysis

The elbow method helps identify the optimal number of clusters in K-means clustering. By plotting the within-cluster sum of squares (WCSS) against the number of clusters, an ”elbow” point is observed where the rate of decrease in

WCSS slows significantly.

In the Figure 4, the WCSS decreases sharply from 1 to 3 clusters, indicating that the clusters become more compact and well-defined. After 3 clusters, the rate of decrease slows, forming an elbow shape. This suggests that 3 clusters are optimal, as adding more clusters yields diminishing returns in reducing WCSS.

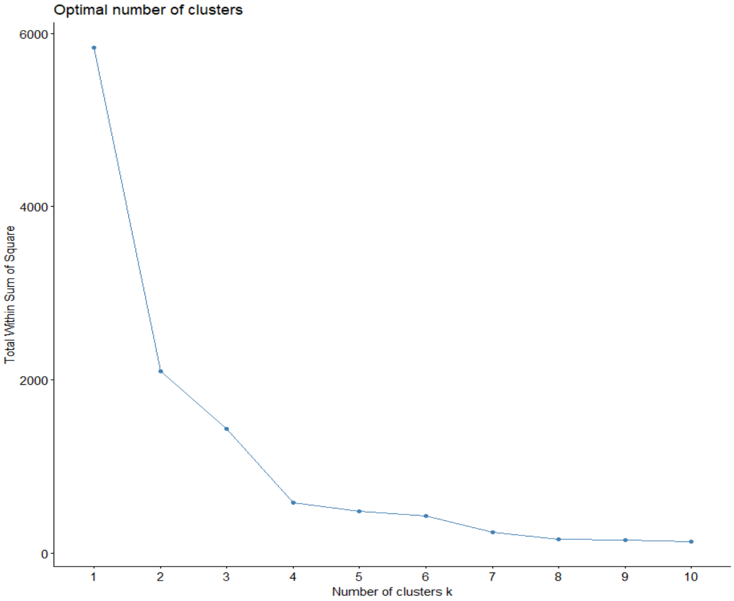


Figure 4: Optimal Cluster Count via Elbow Method

Based on the above, K-means clustering was performed on the scaled dataset with 3 clusters. The algorithm was run with multiple initializations (n=25) to ensure stability and accuracy in the clustering results. The algorithm iteratively assigned each municipality to one of the three clusters, updating the cluster centers until the assignments no longer changed significantly, indicating convergence.

The cluster plot (Figure 5) visualizes the clustering structure in a two-dimensional space, showing the three clusters identified by the K-means algorithm, each represented in different colors: red for Cluster 1, green for Cluster 2, and blue for Cluster 3. The points represent municipalities, positioned based on two principal components (Dim1 and Dim2) that explain 98.4% and 1% of the variance, respectively. Cluster 1 is the most dispersed, indicating diverse age compositions, while Cluster 2 and Cluster 3 are more compact, suggesting homogeneity within these groups. The convex hulls (boundary polygons) highlight the extent of each cluster, with Cluster 1 having the largest boundary due to its variability.

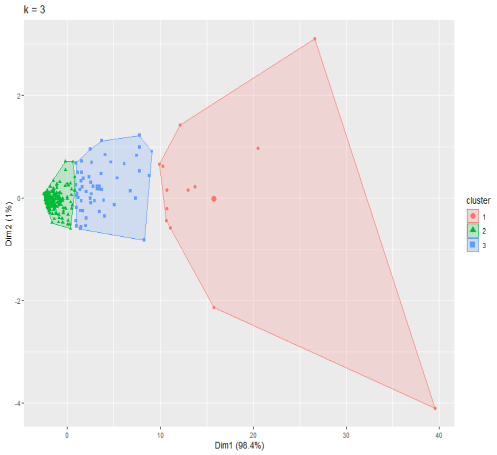


Figure 5: Cluster Plot for K-means Clustering

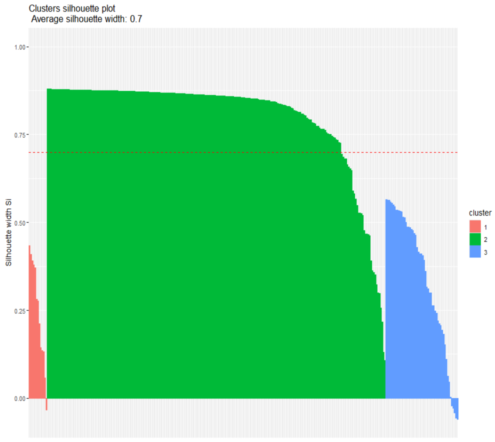


Figure 6: Silhouette Plot for K-means Clustering

To evaluate the clustering quality, the silhouette score was calculated. The silhouette score measures how similar each municipality is to its own cluster compared to other clusters. The average silhouette width of 0.7 indicated a good clustering structure, with most municipalities being well matched to their assigned clusters.

Most municipalities have high silhouette widths (Figure 6), especially those in Cluster 2, which shows clear separation. Cluster 3 has generally high silhouette widths with some variability, while Cluster 1 displays a wider range of widths, suggesting some overlap and less distinct separation. Together, these plots confirm that most municipalities are well-matched to their clusters, demonstrating the effectiveness of K-means clustering based on age composition.

The map (Figure 7) shows the geographical distribution of clusters across Portugal. Cluster 2 covers most of the country, suggesting that many municipalities share similar age composition profiles and population sizes.

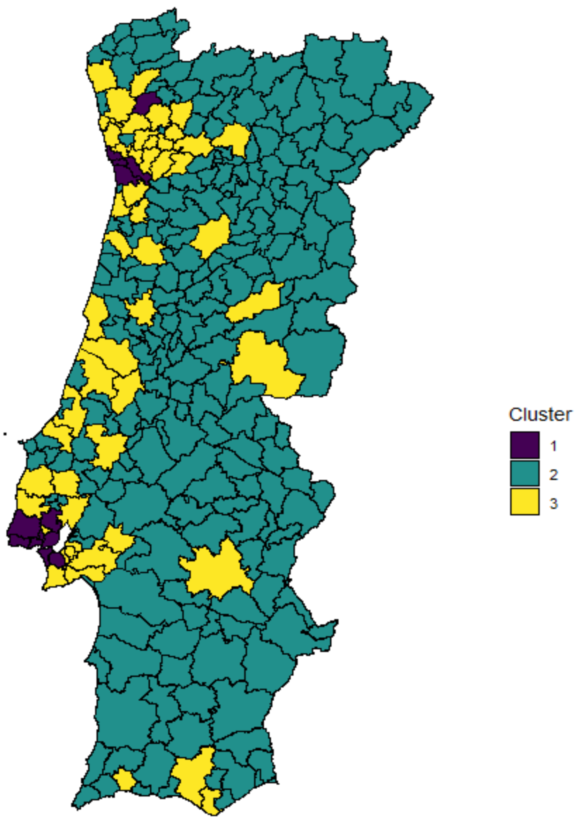


Figure 7: Clusters in Municipalities of Portugal Based on K-means Clustering

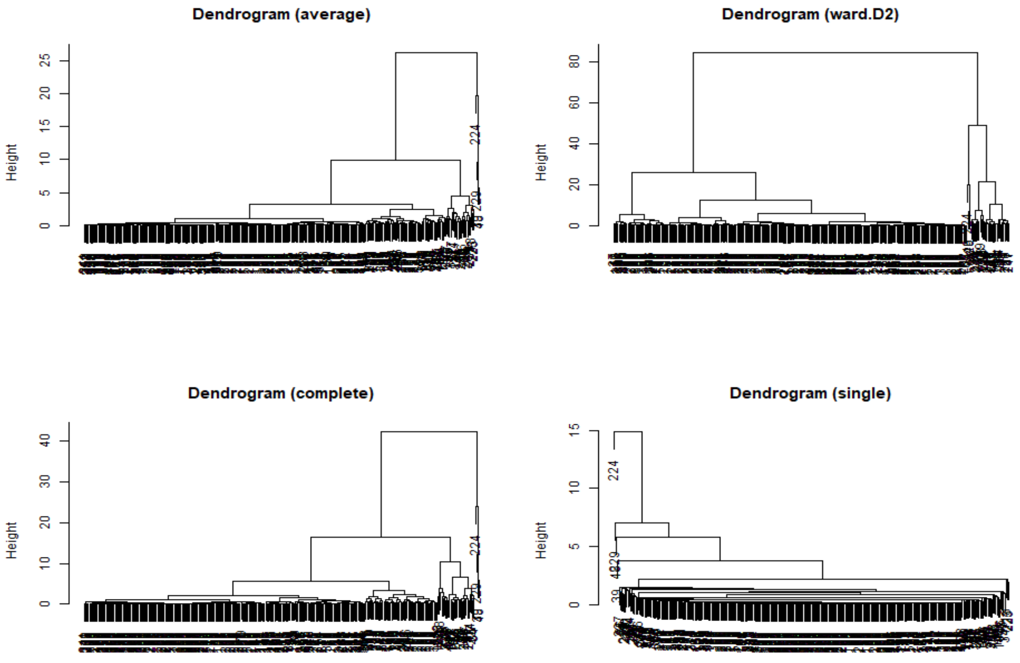
3.2 Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It can be performed in two ways: agglomerative and divisive. Agglomerative hierarchical clustering, which is the more common approach, begins with each observation in its own cluster and iteratively merges the closest pairs of clusters. Conversely, divisive hierarchical clustering starts with all observations in a single cluster and iteratively splits them into smaller clusters.

The process of hierarchical clustering involves calculating a distance matrix, which quantifies how similar or different the observations are from each other. Various linkage criteria can be used to determine the distance between clusters, including single-linkage (minimum distance), complete-linkage (maximum distance), and average-linkage (mean distance). Ward's method, which minimizes the total within-cluster variance, is also a popular choice for hierarchical clustering.

One of the advantages of hierarchical clustering is that it does not require specifying the number of clusters in advance. Instead, the resulting dendrogram, a tree-like diagram, allows for the exploration of the data at various levels of granularity. By cutting the dendrogram at different heights, different numbers of clusters can be identified and analyzed. However, hierarchical clustering can be computationally intensive, especially with large datasets, and the choice of linkage method can significantly affect the results.

The municipalities were grouped based on age demographics using agglomerative hierarchical clustering techniques with Euclidean distance. Several methods were considered: `average`, `ward.D2`, `complete`, and `single`, with their effectiveness assessed through silhouette analysis. The mean silhouette widths for these methods were 0.8745697 (`average`), 0.7997337 (`ward.D2`), 0.8745697 (`complete`), and 0.910053 (`single`), indicating that the highest quality clusters were achieved with the `single` linkage method (values close to 1 signify well-separated clusters). Although the single linkage method provided the highest silhouette width, all methods were used in the analysis to ensure comprehensive evaluation.



Dendrogram (average)

Dendrogram (ward.D2)

Dendrogram (complete)

Dendrogram (single)

Figure 8: Hierarchical Clustering Dendrograms

It is worth mentioning that `Ward.D` and `Ward.D2` are variations of Ward's hierarchical clustering method. `Ward.D` uses the increase in total within-cluster sum of squares, while `Ward.D2` uses squared Euclidean distance, making it more sensitive to cluster separations.

The dendrograms (Figure 8) illustrates how municipalities are grouped based on their similarities, with each leaf representing a municipality. Branches merge at various heights, indicating different levels of similarity. The vertical axis shows dissimilarity, with higher heights indicating less similarity between clusters. The significant vertical separations in each dendrogram point towards forming 2 clusters. It's important to note that only the vertical heights should be used to assess similarity, not the horizontal positions.

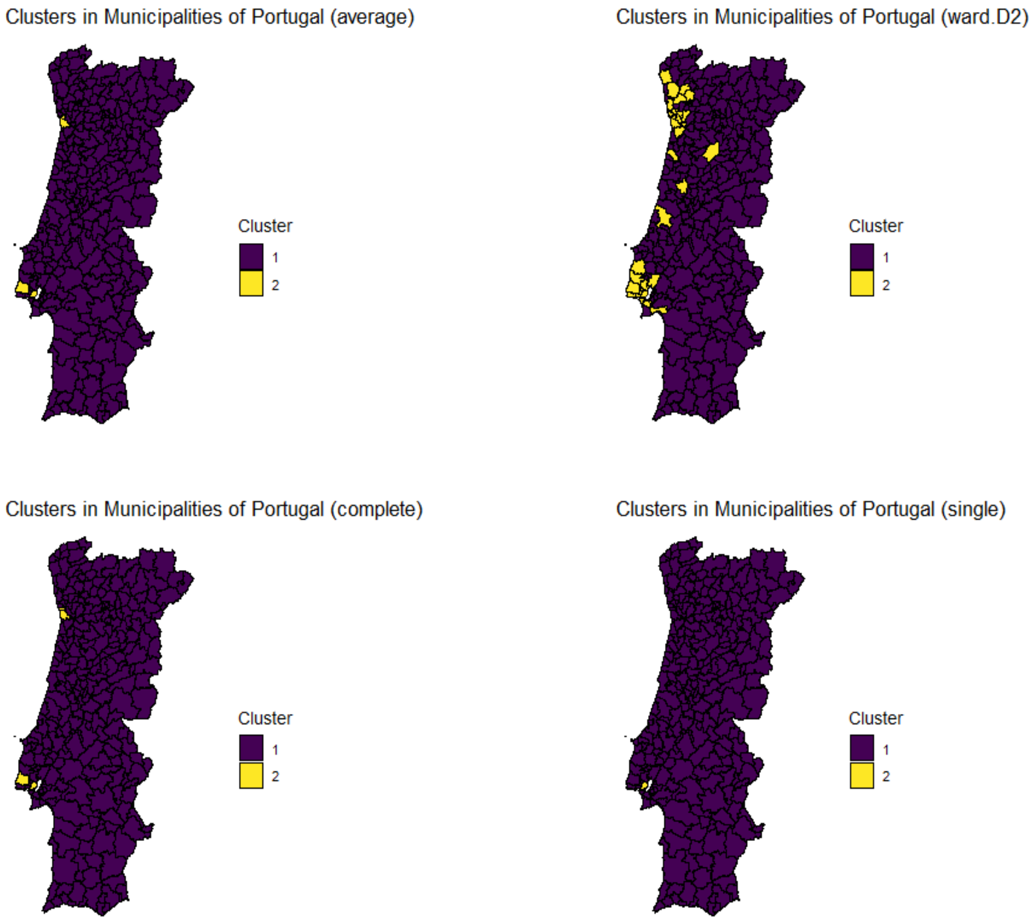

Clusters in Municipalities of Portugal (average)
Clusters in Municipalities of Portugal (ward.D2)
Clusters in Municipalities of Portugal (complete)
Clusters in Municipalities of Portugal (single)
Cluster
1
2

Figure 9: Clusters in Municipalities of Portugal - Different Methods

The Figure 9 presents the clustering using four different hierarchical clustering methods. In the `average` and `complete` linkage methods, the majority of the municipalities fall into a single cluster (purple), with a few exceptions highlighted in yellow. The `ward.D2` method shows a greater number of municipalities in the yellow cluster, indicating a more dispersed clustering pattern. In contrast, the `single` linkage method results in the most distinct clustering with a higher silhouette width, suggesting well-separated clusters. The clear delineation in the `single` linkage method supports the earlier conclusion that it provided the highest quality clustering among the methods evaluated.

4 Comparison of Clustering Techniques: K-means vs. Hierarchical Clustering

In this analysis, two clustering techniques were employed: K-means and hierarchical clustering. Both methods aim to identify groups of municipalities with similar age compositions, but they operate differently and have distinct advantages and limitations.

K-means clustering was used to partition the dataset into a specified number of clusters by minimizing the WCSS. The data was scaled to ensure each variable contributed equally. The Elbow Method helped determine the optimal number of clusters, identifying three as the ideal number. The algorithm was run multiple times with different initializations. The results showed distinct clusters with varying degrees of compactness. The average silhouette score was 0.7, indicating good clustering quality, with most municipalities falling into Cluster 2, covering a significant portion of the country.

Hierarchical clustering, on the other hand, builds a hierarchy of clusters without requiring a predefined number of clusters.. Dendrograms were used to visualize the merging or splitting of clusters at different levels of similarity. The `single` linkage method achieved the highest silhouette width of 0.910053, indicating well-separated clusters. The dendrograms displayed significant vertical separations suggesting the formation of two clusters.

In terms of performance, K-means is more efficient and suitable for large datasets. It provides a straightforward interpretation of clusters but requires careful initialization and scaling. Hierarchical clustering offers flexibility, as it does not need the number of clusters to be predefined, and provides a comprehensive hierarchical view. However, it is more computationally demanding and sensitive to outliers and noise.

When comparing cluster quality, K-means achieved an average silhouette score of 0.7, indicating a good clustering structure. Hierarchical clustering with the `single` linkage method, achieved the highest silhouette width of 0.910053, suggesting superior cluster separation.

5 Conclusion

In summary, this analysis effectively utilized both K-means and hierarchical clustering to uncover patterns in the age distribution of municipalities in Portugal. K-means clustering proved efficient, while hierarchical clustering with the `single` linkage method, provided highly distinct clusters. Both techniques highlighted significant demographic groupings. However, exploring additional techniques, such as model-based clustering, could yield further insights. Model-based clustering, which assumes data generation from a mixture of probability distributions, may provide a deeper understanding of demographic structures across Portugal's municipalities.

Note: Details for the map can be found in Open Net Zero, *Municipalities of Portugal Dataset*, Available at:
https://opennetzero.org/dataset/municipalities-portugal