Introduction:

In recent years, urban centers worldwide have experienced rapid transformations, necessitating the reassessment of administrative boundaries and structures to accommodate changing demographics and urban dynamics. The city of Lisbon is no exception, and in response to these challenges, the Portuguese government implemented administrative reforms in July 2012. The Parliament approved a proposal, including Chapter Two regarding the reconfiguration of Lisbon's parish map, which stipulated that the reconfiguration was based on the principles of territorial organization rationalization and adjustment, with the purpose of creating larger and more balanced parishes. The original 53 parishes were merged, modified, or maintained, and these changes took effect in January 2013. The implementation of Lisbon's administrative reform was a complex and challenging process, but it resulted in a more modern and efficient administrative structure for the city. Notably, the most modern and trendsetting parish, Parque das Nações(Expo), was born from this reform.

During the same period, the Portuguese National Institute of Statistics (INE) conducted the 2011 Portuguese National Census from March 21 to May 2011. Census enumerators collected data from households and individuals across the country. The INE was responsible for organizing and implementing the census, ensuring the accuracy of data collection and analysis. The comprehensive results and data from the census were gradually released starting on July 29, 2011. According to the census, Lisbon's population in 2011 was 547,733, a decrease from 564,477 in 2001, with a population density of approximately 6,378 people per square kilometer. Lisbon's population was aging, with a higher proportion of elderly people and a median age of around 42.2 years. Compared to other regions in Portugal, Lisbon had smaller average household sizes, with 2.4 people per household. Lisbon had a relatively high number of vacant housing units, accounting for about 12.2% of the total housing stock. The educational level of Lisbon's population was higher than that of other regions in Portugal, with a higher proportion of individuals completing higher education. The unemployment rate was higher than the national average, but the city also had a higher proportion of individuals engaged in professional, scientific, and technical activities. Lisbon had a diverse population, with a considerable proportion of foreign-born residents. The city attracted immigrants from various countries, particularly from African Portuguese-speaking countries, Brazil, and other European nations.

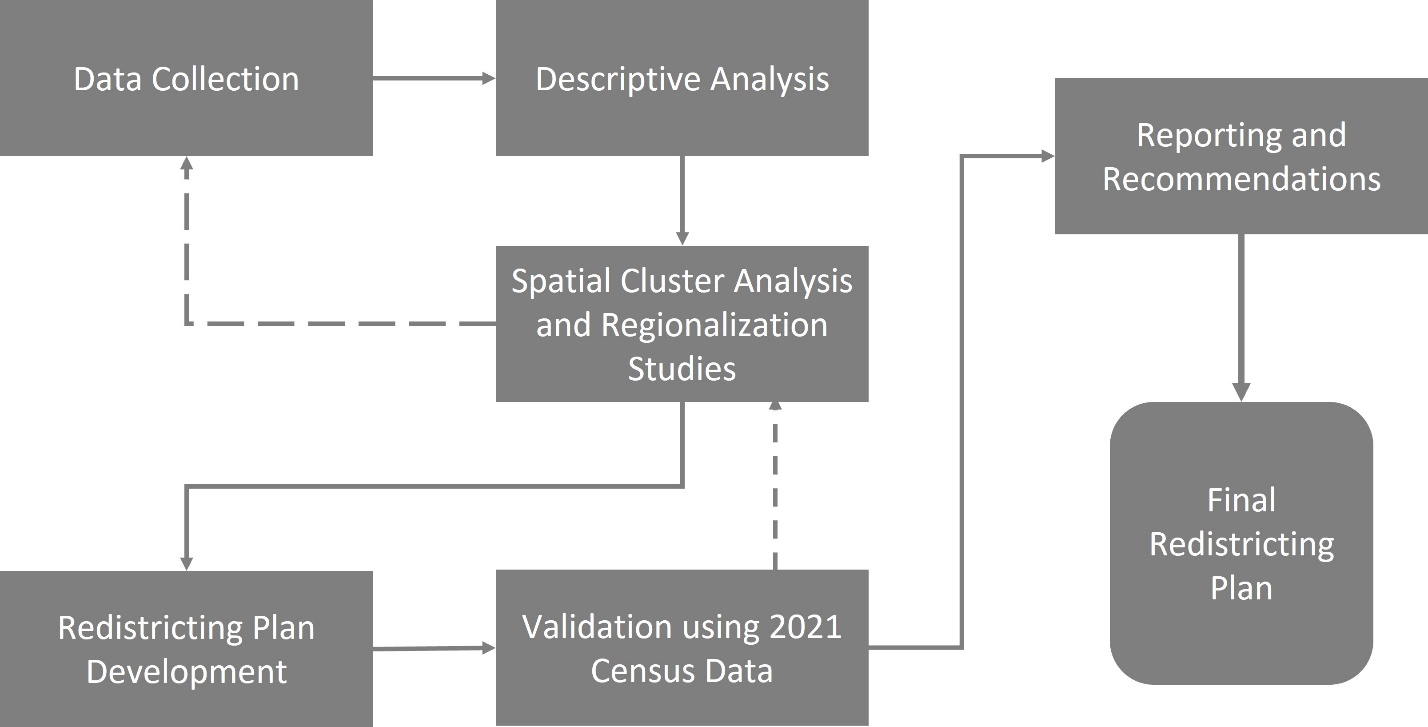
This study aims to propose a redistricting plan for the Lisbon administrative area based on a data-driven approach, utilizing the 2011 census data to inform the decision-making process. The objective of this research is to provide a comprehensive and well-informed redistricting proposal that takes into account the current and future needs of Lisbon's diverse population. By analyzing key demographic variables, such as population density, age distribution, education, employment, housing, and migration patterns, we seek to identify trends, disparities, and opportunities that can inform the redistricting process. In doing so, we aim to ensure that the proposed administrative boundaries are equitable, functional, and reflective of the city's changing demographics.

The study will commence with a meticulous examination of Lisbon's current administrative structure, followed by an analysis of the 2011 census data to underscore the patterns and trends that will inform the proposed redistricting plan. Subsequently, we will utilize advanced methods such as spatial cluster analysis from data science disciplines, regionalization studies, and other relevant tools to pinpoint Lisbon's specific challenges and opportunities at the granular level of the smallest geographical units. This will allow us to outline the goals and objectives of the redistricting process more effectively. Based on these findings, a comprehensive redistricting plan will be developed, with recommendations for redefining administrative boundaries, reallocating resources, and implementing new policies and plans to address the identified needs.

Once the initial redistricting plan is formulated using the 2011 census data, the latest 2021 census data will be employed to validate the results. By comparing the proposed plan with the most recent demographic data, we can ensure that the redistricting plan remains relevant and responsive to Lisbon's evolving needs. This validation process will allow us to refine the proposed plan, make adjustments where necessary, and ultimately create a more robust and effective administrative framework for the city.

In conclusion, the proposed redistricting plan for Lisbon's administrative area, based on a census data-driven approach, seeks to improve the city's governance and adapt it to the shifting needs of its population. By incorporating both the 2011 and 2021 census data and focusing on key demographic variables, we aim to develop a comprehensive, evidence-based proposal that addresses the challenges and opportunities faced by each administrative area. Ultimately, this research aims to contribute to the creation of a more equitable, functional, and responsive administrative framework that can better serve the diverse needs of Lisbon's residents, both now and in the future.

Conceptual model:



**Methodology:**

This study employs a mixed-methods approach, combining quantitative and qualitative data analysis techniques to develop a comprehensive redistricting plan for the Lisbon administrative area. The methodology is organized into several stages, as detailed below:

1. Data Preparation:

The primary data source for this research is the 2011 Portuguese National Census conducted by the Portuguese National Institute of Statistics (INE). The dataset will include variables such as population density, age distribution, education, employment, housing, and migration patterns. Before conducting any analysis, the raw census data will be preprocessed, cleaned, and transformed as needed to ensure its usability for the study. This process may include handling missing values, converting data types, and standardizing units of measurement. Additionally, the latest 2021 census data will be prepared to validate and update the initial redistricting plan.

1. Descriptive Analysis:

The descriptive analysis will involve calculating summary statistics for each demographic variable of interest, including means, medians, standard deviations, and ranges. Visualizations such as bar charts, histograms, and scatterplots will be created to illustrate the distribution and relationships between variables. Maps will be used to display the spatial distribution of key demographic variables across Lisbon's current administrative areas, highlighting areas of high or low concentration, as well as potential anomalies or outliers.

1. Spatial Cluster Analysis and Regionalization Studies:

Spatial cluster analysis techniques, such as k-means clustering and hierarchical clustering, will be employed to identify areas within Lisbon that share similar demographic characteristics. Geographic Information Systems (GIS) software will be used to visualize and analyze the spatial patterns of demographic variables and identify potential clusters. Regionalization studies, which involve aggregating smaller geographical units into larger, homogeneous regions, will be conducted using methods such as the Ward's method or the SKATER algorithm. This will enable the identification of potential new administrative boundaries that align with the observed demographic patterns.

1. Redistricting Plan Development:

The redistricting plan development will involve integrating the findings from the descriptive analysis and spatial cluster analysis to propose new administrative boundaries, resource allocations, and policies. Key considerations during this stage will include ensuring equitable representation, maintaining the integrity of communities, and facilitating efficient governance. Scenario analysis will be conducted to explore the potential implications of different redistricting options, and the proposed plan will be refined iteratively based on stakeholder input, feasibility assessments, and alignment with research objectives.

1. Validation using 2021 Census Data:

The validation process will involve comparing the initial redistricting plan, based on the 2011 census data, with the most recent 2021 census data. This comparison will help determine if the proposed administrative boundaries, resource allocations, and policies remain relevant and responsive to Lisbon's evolving needs. Discrepancies between the two datasets will be analyzed, and the redistricting plan will be updated accordingly. The validation process will also serve as a robustness check, ensuring that the proposed plan is adaptable to changing demographic trends.

1. Reporting and Recommendations:

The final stage of the methodology involves synthesizing the research findings, redistricting plan, and validation results into a comprehensive report. This report will provide a detailed account of the methodology, data sources, and analytical techniques used, as well as a summary of the key findings and proposed redistricting plan. Recommendations for implementing the proposed plan will be provided, including steps for engaging stakeholders, conducting public consultations, and integrating the new administrative boundaries into existing governance structures. The report will also offer suggestions for future research, such as exploring the impact of the redistricting plan on service provision, community cohesion, and political representation, as well as potential policy interventions to address the identified challenges and opportunities in each administrative area.

By following this expanded methodology, the study aims to develop a comprehensive, data-driven redistricting plan for the Lisbon administrative area that is responsive to the current and future needs of the city's diverse population, ultimately contributing to a more equitable, functional, and responsive administrative framework. This in-depth approach will help ensure that the redistricting plan is well-informed, evidence-based, and adaptable to changing circumstances.

Furthermore, the proposed methodology is designed to promote transparency and stakeholder engagement throughout the redistricting process. By actively involving key stakeholders, such as community leaders, policymakers, and residents, this research aims to foster a collaborative and inclusive approach to redistricting that takes into account diverse perspectives and needs.

In addition to the primary research objectives, this study can also serve as a model for other cities facing similar challenges related to urban growth, demographic shifts, and administrative restructuring. The methodology outlined here can be adapted and applied to different contexts, helping other urban centers develop data-driven redistricting plans that promote effective governance and equitable representation.

Overall, the expanded methodology outlined above will enable this study to develop a robust, evidence-based redistricting plan for Lisbon's administrative area. By considering a wide range of demographic variables, utilizing advanced data analysis techniques, and engaging stakeholders in the decision-making process, this research aims to create an administrative framework that can better serve the diverse needs of Lisbon's residents, both now and in the future.

Data Preparation:

Initially, we procured the original BRGI data containing geographic information from the National Statistic Institute's official website. The 2011 census data can be accessed at https://mapas.ine.pt/download/index2011.phtml, while the 2021 census data is available at https://mapas.ine.pt/download/index2021.phtml. Utilizing geopandas, we processed the 2011 census data, which comprises 3623 entries and 134 columns. The data encompasses various geographic levels, from the most granular (sub-section) to the broadest (Lisbon city), and includes 53 parishes.

The dataset incorporates 28 columns related to building characteristics, including the number of buildings, structure types, building functions, floor quantities, construction years, and materials used. Furthermore, it contains 23 columns regarding accommodations, describing different aspects such as accommodation types, vacancy statuses, area sizes, number of rooms, available parking spaces, and ownership types.

In addition, the dataset features 15 columns on family data, highlighting family counts, family members, the presence of elder individuals or children, unemployed family members, and couples with children across various age ranges. It also consists of 33 columns on individual demographic information, presenting the number of residents segmented by gender and age groups.

Lastly, there are 23 columns on education and employment, detailing individuals' educational levels, those actively studying, individuals cohabiting with others of different educational levels, and employment statuses across diverse industries. The final column, labeled 'geometry,' comprises MULTIPOLYGON data, signifying non-overlapping polygons for intricate geographic features with disconnected areas.

We commenced the data preparation phase by comprehending the data structure and generating supplementary features. Based on the 'geometry' column, we calculated the area size and the latitude and longitude of each entry's centroid. Additionally, we aggregated the number of buildings erected before 1990 to establish a new feature. Then we computed some proportions from our data. The primary reason for computing these proportions is to enable a more nuanced and comprehensive understanding of the data. By normalizing the data, we can make meaningful comparisons between different geographical units, accounting for variations in size, population, and other factors. Furthermore, these proportions facilitate the identification of trends, disparities, and relationships within the data, which can inform the redistricting process, ensuring that the proposed administrative boundaries are equitable, functional, and reflective of Lisbon's changing demographics. Moreover, we computed numerous ratios to facilitate a deeper understanding of the data:

1. The proportion of exclusively residential buildings within the total building count.
2. The proportion of buildings with five or more floors within the total building count.
3. The proportion of buildings constructed before 1995 within the total building count.
4. The proportion of conventional dwellings for habitual residence within the total accommodations.
5. The proportion of conventional dwellings with an area exceeding 200 m2 within all conventional dwellings for habitual residence.
6. The proportion of owner-occupied conventional dwellings within all conventional dwellings for habitual residence.
7. The proportion of traditional families with individuals aged below 15 years in the aggregate number of private households.
8. The proportion of resident individuals aged between 20 and 64 years within the overall population.
9. The proportion of female residents within the total number of resident individuals

Following this, we calculated the aggregate number of residents employed across all industries to derive one more proportions which is the proportion of resident individuals working within their municipality of residence in the total number of resident individuals employed across all sectors.

Following the feature engineering process, we implemented the StandardScaler from the sklearn.preprocessing package to normalize all the variables in our dataset. This step is essential to ensure that the differing scales and ranges of the variables do not disproportionately impact the subsequent analysis. After standardization, we employed Principal Component Analysis (PCA) as a dimensionality reduction technique to streamline our dataset for further examination.

We ascertained the number of PCA components possessing eigenvalues exceeding 1, signifying their capacity to comprehensively explain the features in our dataset. We identified 17 such components, and consequently, our final dataset comprised all 17 PCA components. This transformed dataset, with reduced dimensionality, will facilitate more efficient and accurate analysis in the subsequent stages of our research.

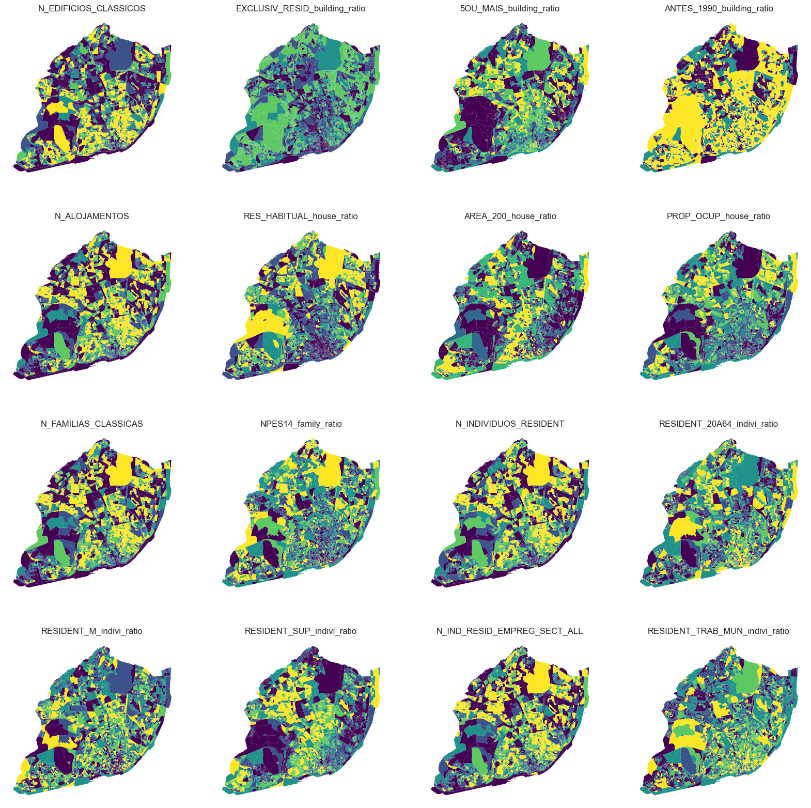
Descriptive Analysis:

Initially, we constructed a list designated as 'ratio\_variables,' which encompassed all the proportion variables derived in the previous steps. Additionally, we generated a list titled 'analysis\_variables,' containing all the proportion variables alongside the primary variables of distinct categories, including the total number of classic buildings, total number of accommodations, total number of classical families, and total number of resident individuals. These lists serve as a systematic approach to organizing the variables for subsequent analysis and ensuring a comprehensive examination of the relevant factors.

The spatial distribution maps

Afterwards, I generated visualizations to display the spatial distribution of each variable within the 'analysis\_variables' list, using the processed GeoDataFrame 'df11', which contains the Census 2011 data after all necessary feature engineering and data transformations. To achieve this, I create a 4x4 grid of subplots using the plt.subplots() function, flatten the axes array for easier indexing, and then iterate through each variable in the 'analysis\_variables' list.

For each variable, I plot a choropleth map using the GeoDataFrame 'df11' and the plot() function, specifying the 'column' parameter as the current variable, and using a "Quantiles" classification scheme along with the "viridis" color map. The linewidth is set to 0 to remove lines between polygons. Afterward, I remove axis clutter by calling the set\_axis\_off() method and set the title of the subplot to the variable name with set\_title(col). Lastly, I display the final figure containing the spatial distribution maps for each variable using plt.show().



From the figure, it can be observed that there is a higher concentration of buildings in the center of Lisbon. However, the proportion of exclusively residential buildings is low in the center, while the proportion of tall buildings is higher in the north of the center. Additionally, the proportion of old buildings is quite high in most areas, except for some regions in the north.

The number of accommodations is also high in the city center, but the proportion of conventional dwellings for habitual residence is notably low in the center. In the western part of the city center and Expo, the proportion of dwellings larger than 200 m2 is high. Furthermore, the proportion of owner-occupied conventional dwellings is higher in some northern areas and Expo.

Excluding the Avenida Liberdade area, the number of families is high in most of the central areas of Lisbon. The proportion of families with children under 15 years old is higher around the city center. The visualization of the number of resident individuals is similar to that of families. The proportion of residents aged 20 to 64 years old is dispersed throughout the city, with noticeable concentrations in the center, Expo, and Benfica. The proportion of female individuals is evenly distributed across the city, as illustrated in the chart. The proportion of residents who have completed their university studies is high in areas ranging from the west of the center to Benfica, as well as in Belem and Expo.

The number of employees is high in most areas, with the exception of the city's fringes and certain border regions. The proportion of resident individuals working within their municipality of residence increases as one moves closer to the city center, but there are also a few highlighted fringe areas.

Moran's I score

First, I updated the 'analysis\_variables' list by appending the 'Area' variable. This action allows for the inclusion of the 'Area' variable in subsequent analyses and visualizations that utilize the 'analysis\_variables' list. By adding the 'Area' variable to the list, I can examine its relationship with other variables and evaluate its potential influence on the clustering results. This added variable can provide additional insights into the spatial distribution and characteristics of the study area.

In our study, we introduced Moran's I score as a means of quantifying spatial autocorrelation for each variable in the 'analysis\_variables' list. Moran's I score gauges the extent to which the value of a given variable at one location correlates with the values in its neighboring locations. Ranging from -1 to 1, values close to -1 signify negative spatial autocorrelation (neighboring values tend to be dissimilar), while values close to 1 indicate positive spatial autocorrelation (neighboring values tend to be similar). A value near 0 implies no significant spatial autocorrelation, suggesting a random distribution of values.

Moran's I score is instrumental in spatial data analysis as it facilitates the identification of patterns or clusters in the data, thereby providing insights into the spatial structure of the observed variables. A high positive Moran's I score may point to spatial clustering, whereas a low negative score can imply spatial dispersion or dissimilarity among neighboring values. In academic research, Moran's I score is valuable for discerning underlying spatial relationships between variables, thus informing policy or intervention strategies.

To investigate the spatial autocorrelation of the variables, I utilized the Queen contiguity method to derive the weights from the processed GeoDataFrame 'df11'. By setting a seed for reproducibility using NumPy's random.seed() function, I ensured that the results could be replicated in future analyses. Subsequently, I computed Moran's I for each variable in the 'variables' list by employing a list comprehension and the Moran() function. I then structured the results as a list of tuples, with each tuple encompassing the variable name, the corresponding Moran's I value, and the p-value associated with the spatial autocorrelation test. This approach allowed me to systematically assess the spatial autocorrelation patterns within the variables and gain insights into their spatial relationships.

|  |  |  |
| --- | --- | --- |
| Variable | Moran's I | P-value |
| 5OU\_MAIS\_building\_ratio | 0.443371722 | 0.001 |
| RESIDENT\_SUP\_indivi\_ratio | 0.425766199 | 0.001 |
| ANTES\_1990\_building\_ratio | 0.413361392 | 0.001 |
| PROP\_OCUP\_house\_ratio | 0.408845195 | 0.001 |
| AREA\_200\_house\_ratio | 0.373674246 | 0.001 |
| EXCLUSIV\_RESID\_building\_ratio | 0.368130602 | 0.001 |
| N\_FAMILIAS\_CLASSICAS | 0.35148788 | 0.001 |
| N\_ALOJAMENTOS | 0.339248687 | 0.001 |
| N\_INDIVIDUOS\_RESIDENT | 0.336289604 | 0.001 |
| N\_IND\_RESID\_EMPREG\_SECT\_ALL | 0.332473839 | 0.001 |
| N\_EDIFICIOS\_CLASSICOS | 0.297378502 | 0.001 |
| RES\_HABITUAL\_house\_ratio | 0.247221393 | 0.001 |
| NPES14\_family\_ratio | 0.183663563 | 0.001 |
| Area | 0.17902436 | 0.001 |
| RESIDENT\_M\_indivi\_ratio | 0.154660659 | 0.001 |
| RESIDENT\_TRAB\_MUN\_indivi\_ratio | 0.153725148 | 0.001 |
| RESIDENT\_20A64\_indivi\_ratio | 0.117590909 | 0.001 |

The table above presents the Moran's I scores and associated p-values for each variable in the 'analysis\_variables' list. A higher Moran's I score indicates a stronger positive spatial autocorrelation, while a lower score suggests weaker or negative spatial autocorrelation. The p-values provide insight into the statistical significance of the Moran's I scores. In our analysis, all the variables have a p-value of 0.001, indicating that the spatial autocorrelation observed for each variable is statistically significant.

Some variables with the highest Moran's I scores include 5OU\_MAIS\_building\_ratio, RESIDENT\_SUP\_indivi\_ratio, ANTES\_1990\_building\_ratio, PROP\_OCUP\_house\_ratio and AREA\_200\_house\_ratio. These scores signify that there is a notable positive spatial autocorrelation for these variables, suggesting that the data points with similar values tend to cluster together spatially. In contrast, variables with lower Moran's I scores, such as RESIDENT\_20A64\_indivi\_ratio (0.117591) and RESIDENT\_TRAB\_MUN\_indivi\_ratio (0.153725), exhibit weaker spatial autocorrelation.

Spatial Cluster Analysis and Regionalization Studies:

Kmeans clustering analysis

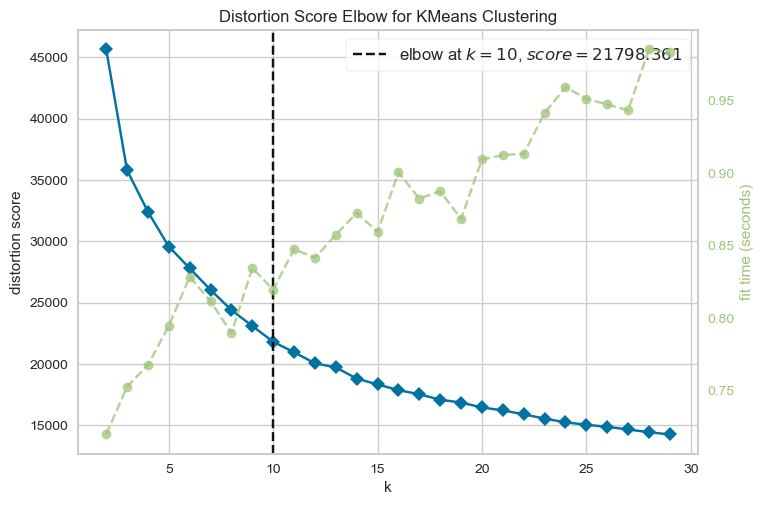
In order to further explore the relationships between the variables and identify potential patterns or clusters, I employed the k-means clustering algorithm. K-means clustering is a widely used unsupervised machine learning technique that aims to partition a dataset into k distinct, non-overlapping clusters based on the similarity of data points. The algorithm works by iteratively updating the centroids (mean values) of the clusters until convergence is reached or a predetermined number of iterations has been completed.

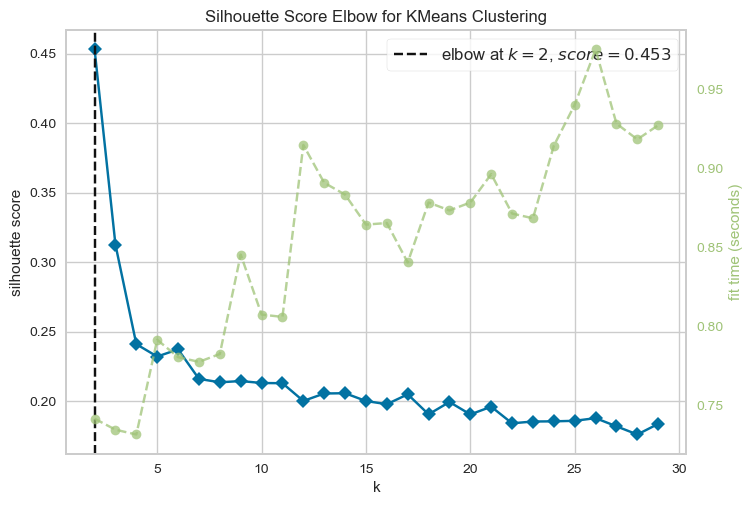
Determine the optimal number of clusters (k):

To identify the appropriate number of clusters for the dataset, I used the elbow method or silhouette analysis. These techniques involve plotting the sum of squared errors (SSE) or silhouette scores for various values of k and selecting the k value where the SSE starts to decrease less rapidly (elbow point) or the silhouette score is maximized. Considering there are currently 24 parishes in Lisbon, which can be viewed as 24 clusters from a data science perspective, I also set an additional k value to be 24 to explore the possibility of achieving a clustering solution that aligns with the existing administrative divisions.

In this analysis, I employed the elbow method and silhouette analysis to determine the optimal number of clusters for the k-means algorithm. Initially, I instantiated a KMeans model and created a KElbowVisualizer (Yellowbrick library) instance to evaluate a range of cluster numbers (k) between 2 and 30. By setting the timings parameter to True, the visualizer also displayed the computation time for each k value. Subsequently, I executed the silhouette analysis with the same k range by specifying the 'metric' parameter as 'silhouette' to use the silhouette score for evaluating the clustering performance.

Upon completion, I obtained two distinct figures: the elbow plot and the silhouette score plot. These visualizations facilitate the selection of the optimal number of clusters by identifying the point at which the sum of squared errors (SSE) starts to decrease less rapidly (elbow point) or the silhouette score is maximized. By comparing the results of these two techniques, I can confidently determine the most appropriate number of clusters for the k-means algorithm, which will provide valuable insights into the spatial relationships and patterns within the dataset.





KElbowVisualizer is a visualization tool provided by the Yellowbrick library in Python, designed to facilitate the selection of the optimal number of clusters for KMeans clustering. Based on the elbow method, this tool recommends choosing the optimal number of clusters at the "elbow" point in a plot of the sum of squared distances between each point and its closest centroid versus the number of clusters. The KElbowVisualizer creates a KMeans model for a range of cluster numbers and fits it to the data. It then computes the sum of squared distances for each cluster and generates a plot of the within-cluster sum of squared distances versus the number of clusters. The optimal number of clusters is selected at the "elbow" point, which is the point of inflection in the plot where the rate of decrease in the sum of squared distances starts to slow down. In this analysis, the elbow plot revealed that 10 clusters were the optimal choice based on the program's selection.

Utilizing the KElbowVisualizer with the 'silhouette' metric, a KMeans model is created for a range of cluster numbers and fitted to the data. The silhouette scores for each cluster are computed, and a plot of the silhouette scores versus the number of clusters is generated. The optimal number of clusters is determined at the point where the silhouette score is maximized, indicating an ideal balance between intra-cluster similarity and inter-cluster dissimilarity. This method offers an alternative to the elbow method for selecting the optimal number of clusters and ensures that the chosen clustering solution is both meaningful and robust, thereby facilitating a deeper understanding of the underlying spatial relationships between different variables. From the silhouette score plot, it is evident that the highest score is achieved with an initial 2 clusters, followed by a steep drop to a lower score. After 4 clusters, the score gradually decreases until the end of the evaluated range.

In an academic context, it is important to acknowledge that data may not always exhibit clear clustering patterns, especially in cases involving high-dimensional data, such as the variables under consideration in this study. Consequently, relying solely on the silhouette score method may not always yield an optimal clustering solution. In such scenarios, the elbow method serves as a valuable alternative, as it provides an automated result that can help inform the selection of an appropriate number of clusters for further analysis.

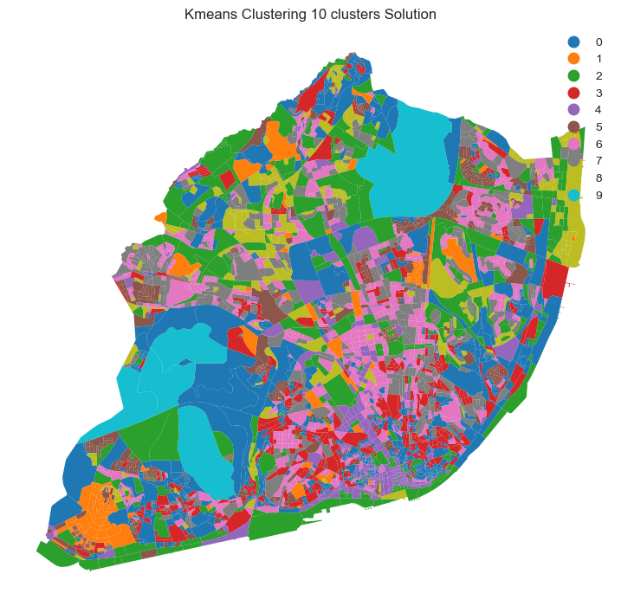
Perform k-means clustering

In the study, I initialized a KMeans instance with the chosen number of 10 clusters based on the elbow method. To ensure reproducibility, I set a seed using NumPy's random.seed() function. Then, I ran the k-means algorithm on the 'analysis\_variables' subset of the 'df11' DataFrame using the fit() method.

After fitting the model, I accessed the generated cluster labels with the 'labels\_' attribute and printed the first five labels for a quick inspection. Subsequently, I added the cluster labels to the 'df11' DataFrame as a new column named 'kmcls'. This addition allows for the visualization and examination of the clusters in relation to the spatial distribution of the study area.

Next, I employed the same approach, but this time using 24 clusters, which corresponds to the original number of parishes in Lisbon. I created another k-means model with 24 clusters and added the resulting labels to the 'df11' DataFrame as a new column named 'kmcls24'. This step enables a comparison between the clusters identified using data-driven techniques (10 clusters) and the clusters based on real-life divisions (24 parishes). Such a comparison can provide valuable insights into the relationships between the data-driven clusters and the existing administrative boundaries.

In this research, I employed a 'plot\_choropleth' function to produce spatial distribution maps for KMeans clustering analysis. The function accepts a GeoDataFrame, a column name, and an optional title, and subsequently creates a choropleth map that displays the unique values of the specified column. It sets up a figure and axis, plots the data, removes axis lines, and adds a title. The map is then displayed using the plt.show() function. This function facilitates the visualization and interpretation of clustering solutions by presenting them within a geospatial context. The next two figures showcase the visualization results of the two clustering solutions.



In the k-means clustering solution with 10 clusters, we observe that the clusters are evenly distributed across the entire Lisbon map, and most of them exhibit a geospatially dispersed visual effect. Comparing this to the results of the 24 clusters solution, we notice a substantial variation in color. However, upon closer examination, the distribution patterns of the 24 clusters solution appear strikingly similar to those of the 10 clusters solution. This observation suggests that when the number of clusters is set to 24, the resulting clusters are likely generated by further subdividing the 10 clusters solution and recombining the subdivisions into 24 distinct clusters.

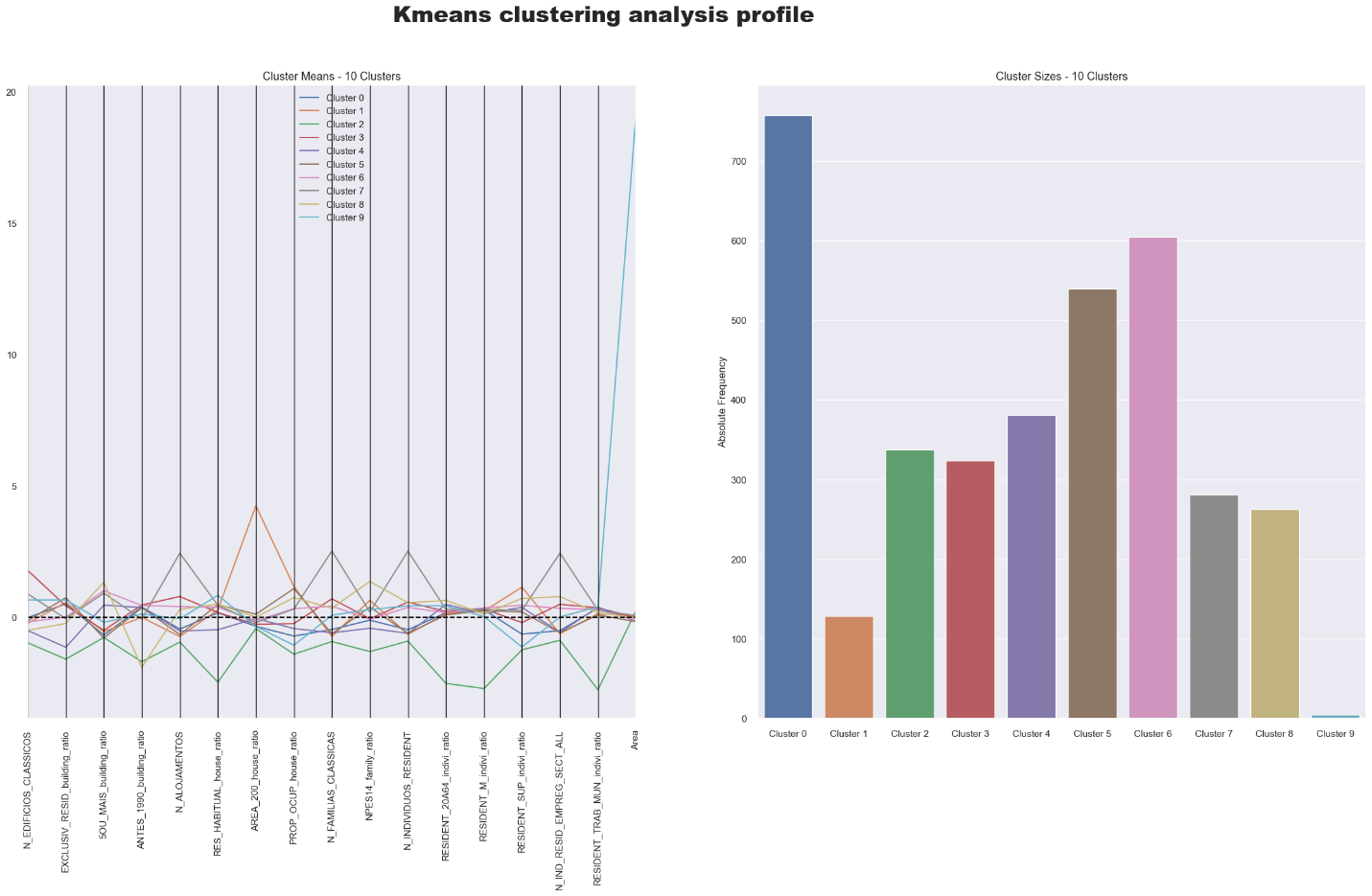
Certainly, the observed results are closely related to the mechanics of the k-means algorithm. The specified number of clusters only constrains the final outcome, but it does not influence the algorithm's process of iteratively updating the centroids (means) of the clusters. The algorithm's primary aim is to achieve convergence or complete a predetermined number of iterations, regardless of the number of clusters specified. Consequently, the similarities observed between the 10 clusters and 24 clusters solutions can be attributed to the intrinsic properties of the k-means clustering technique.

The 10 clusters solution presents a valuable insight into the clustering results, allowing for a visual examination of the degree to which Tobler's first law of geography is evident in multivariate clusters. According to this law, regions in closer proximity should display greater resemblance than those that are more distant. The clustering diagram demonstrates this law, as unit blocks that are adjacent to one another often share the same color, creating similar clusters. While the colors are dispersed throughout the diagram, this distribution does not prevent the formation of similar clusters among many neighboring units. Nevertheless, the complexity of the underlying spatial units limits the effectiveness of visual inspection. As the polygons differ in size and shape, relying solely on visual observation is not adequate to accurately interpret the spatial distribution of these clusters.

Analyze the results:

Then we created a function names ‘cluster\_profiles’ which is designed to generate visualizations of cluster profiles for one or more clustering label columns in a DataFrame. By creating parallel coordinates plots for cluster centroids and bar plots for cluster sizes, this function provides insights into the means and sizes of the clusters, allowing for a deeper understanding of the underlying data patterns and facilitating comparison of different clustering solutions.

The figure below displays the analysis results for the 10 clusters k-means clustering solution generated by the ‘cluster\_profiles’ function. The visualization provides a clear representation of the cluster means and sizes, allowing for an in-depth examination of the data patterns within each cluster. The parallel coordinates plots for cluster centroids highlight the differences and similarities among clusters, while the bar plots for cluster sizes reveal the distribution of data points across the clusters.



In the figure above, the distribution of the number of clusters reflects the relative frequency of each cluster in the data. The cluster sizes range from the smallest, Cluster 9, with only 5 units, to the largest, Cluster 0, with 757 units. This wide range in cluster sizes indicates that some areas share more common attributes, while others are more distinct. But if we properly adjust the display order of different clusters, we can also get a graph of the number of clusters that presents a normal distribution, which can prove the rationality of the cluster's own existence.

Examining the degree of change of each feature across the clusters, we notice that some features show considerable variability, while others are more consistent. For example, the 'Area' attribute demonstrates a significant change, with Cluster 9 having an exceptionally high value compared to the rest. This suggests that the spatial units in Cluster 9 may cover larger geographical areas and could be fundamentally different from other clusters in terms of their spatial characteristics. Furthermore, the number of classic buildings ('N\_EDIFICIOS\_CLASSICOS') also shows considerable variation across clusters.

In contrast, some features, like the proportion of occupied properties ('PROP\_OCUP\_house\_ratio'), display more consistent values across most clusters, with a few exceptions like Cluster 1, which has a notably high value. Other features, such as the number of classic families ('N\_FAMILIAS\_CLASSICAS'), show more moderate variation across clusters.

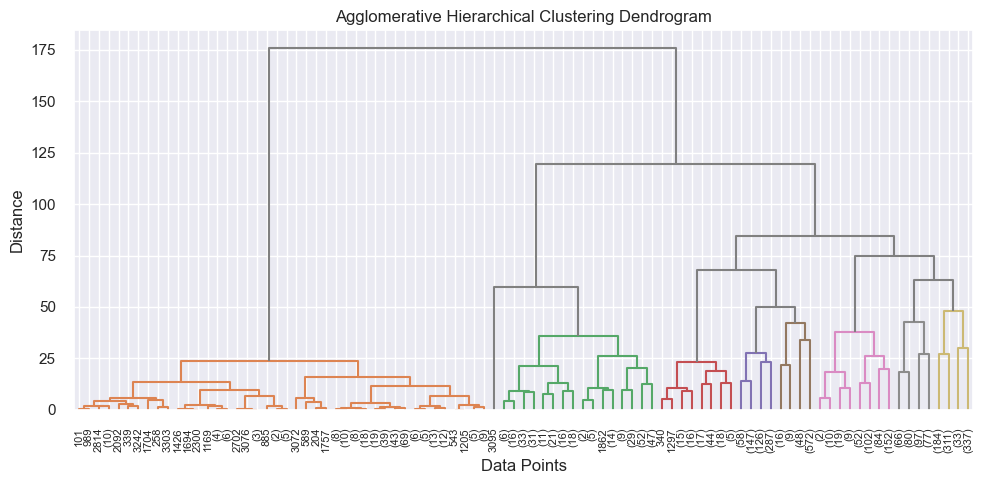
By analyzing the distribution of clusters, the degree of change in each feature, and special values like the high 'Area' value in Cluster 9, we can gain a better understanding of the spatial patterns and relationships within the data. This integrated analysis reveals unique characteristics or outlier regions that may warrant further investigation and helps identify the underlying trends in the study area.

Hierarchical clustering

As previously discussed, k-means is merely one clustering algorithm among many others. In this section, we will examine the 2011 Lisbon Census dataset employing an alternative prominent clustering technique: Agglomerative Hierarchical Clustering (AHC). This method offers another perspective on the data and allows for comparative analysis.

Hierarchical clustering is a clustering method that builds a hierarchy of groups based on the similarity of data points. There are two main approaches: agglomerative (bottom-up) and divisive (top-down). Agglomerative clustering, which is often considered superior due to its more natural and interpretable approach, starts with each point as a separate cluster and gradually merges the closest clusters. Divisive clustering, on the other hand, starts with all points in one cluster and splits it into smaller ones. A tree-like structure called a dendrogram is used to visualize the hierarchy and determine the optimal number of clusters. Hierarchical clustering provides an interpretable representation of the data structure but can be computationally complex for large datasets. The choice of distance metric and linkage method can also impact the results.

To generate a dendrogram plot for hierarchical clustering analysis, I utilized the linkage() function from the scipy.cluster.hierarchy module with the 'ward' method. The resulting linkage matrix was then passed to the dendrogram() function for visualization. The plot was customized using various parameters such as rotating x-axis labels, truncating branches, setting a color threshold, and adjusting the plot layout. This plot is useful in visualizing the hierarchical clustering results and identifying distinct clusters within the data. The resulting dendrogram plot is shown in the figure below.

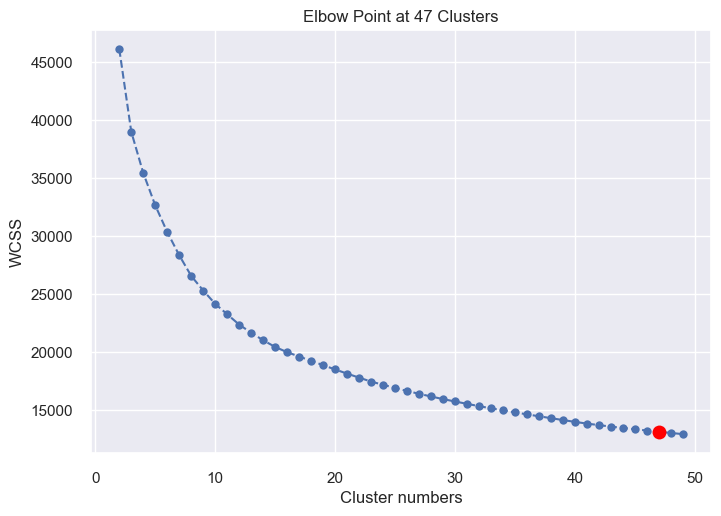
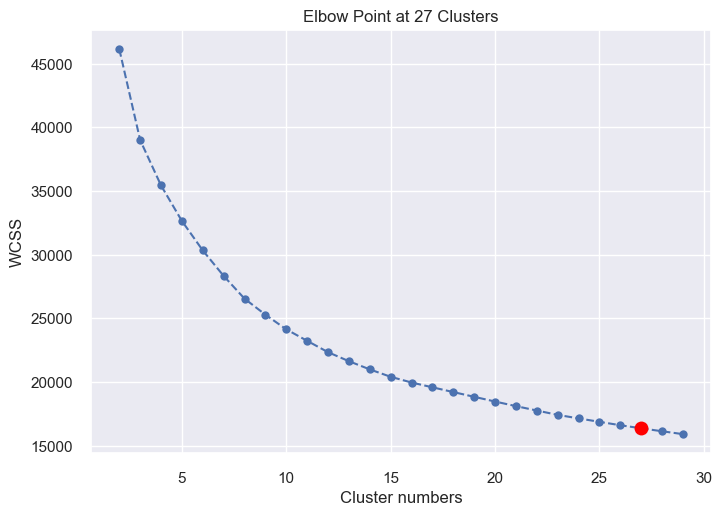


Determine the optimal number of clusters (k):

The dendrogram visualization produced inconclusive results for determining the optimal number of clusters. To address this issue, the within-cluster sum of squares (WCSS) was imported as a measure to determine the ideal number of clusters for a given dataset. The WCSS calculates the sum of the squared distances between each data point and the centroid of its assigned cluster. The metric is calculated for each cluster, and the sum of all clusters' WCSS is utilized to identify the ideal number of clusters. While the WCSS generally decreases as the number of clusters increases, the decrease becomes less pronounced beyond a certain point. The optimal number of clusters is often identified at the "elbow point," where the rate of WCSS decrease slows down significantly, indicating that additional clusters would not substantially improve clustering. The WCSS is a vital metric in cluster analysis since it assists in determining the optimal number of clusters, which is a critical parameter for interpreting and utilizing the clustering results.

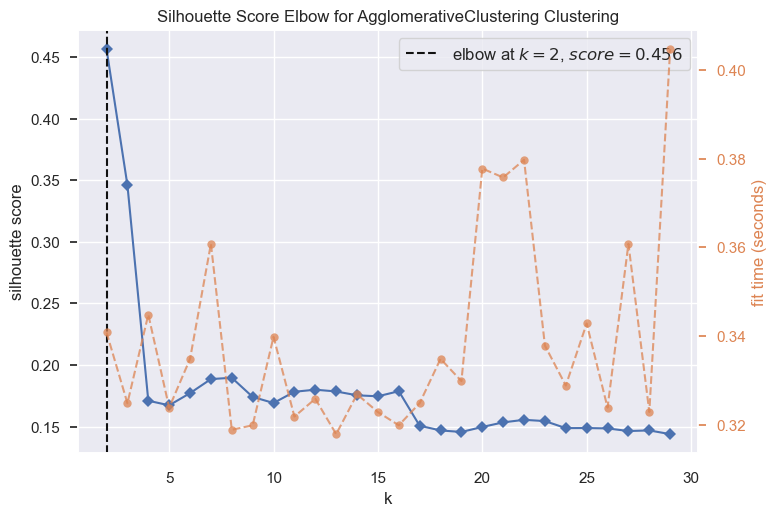
A customized code has been developed to determine and visualize the optimal number of clusters for the given dataset. The code includes a function called "find\_elbow\_point" which computes the WCSS for various cluster numbers and returns the index of the elbow point. The identified elbow point is used to plot a red dot on the visualization result to signify the optimal number of clusters, while the elbow\_point variable stores the number of clusters at the elbow point, which is displayed in the plot title.

The find\_elbow\_point function adopts Min-Max normalization to normalize the WCSS values between 0 and 1, and then calculates the percentage change in WCSS values by finding the absolute difference between each pair of normalized WCSS values divided by the first normalized WCSS value. The elbow point is identified by locating the highest percentage change in WCSS values in a loop that commences from the second cluster number and continues until the second to last cluster number. The computed percentage change values are stored in a list, and the elbow point is defined as the cluster number with the maximum percentage change. Finally, the visualization result is annotated with a vertical line and a text annotation to display the elbow point. Two visualization results are shown, where the upper limit of the number of clusters is set to 30 and 50, respectively.



An issue arises as the elbow point consistently appears at the third-to-last point, even when the clustering range is extended to less practical values, such as 75 or 100. Despite these adjustments, the elbow point remains fixed, indicating that the optimal number of clusters for the data is relatively small. This observation suggests that WCSS may not be an effective tool in this particular research context.

An alternative method for selecting the optimal number of clusters in AHC is the Silhouette score, which has been previously explained. In this approach, we utilize the KElbowVisualizer library to calculate the Silhouette score for a range of cluster numbers from 2 to 30. The resulting visualization is shown in the following figure.

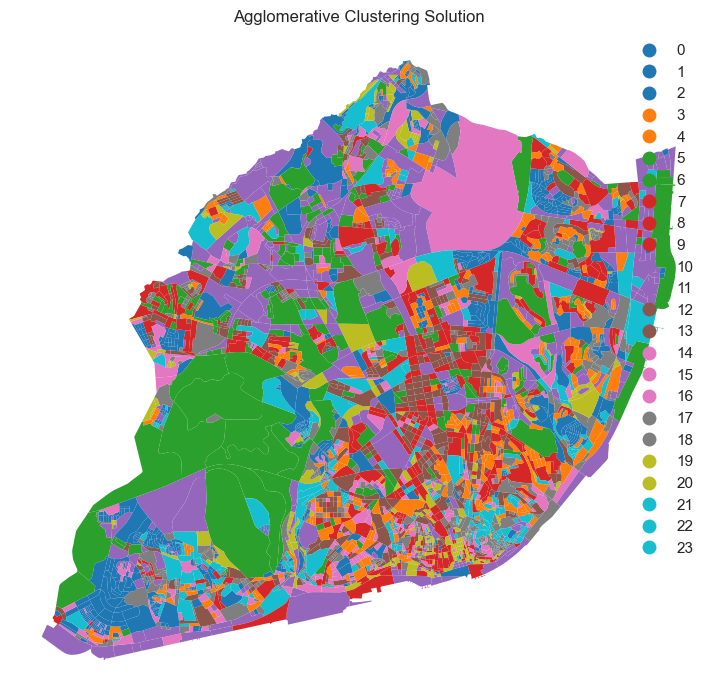


A similar plot to the silhouette score plot in K-means was obtained, and the results showed a similar trend, rendering the obtained evidence inconclusive.

In the current study, domain knowledge plays a crucial role in determining the optimal number of clusters. It is important to take into consideration any prior information or knowledge related to the underlying structure of the data or the system being studied. Even if clustering algorithms suggest a different number of clusters, the predetermined number based on domain knowledge may still be the best choice. Therefore, in our study, the initial number of 24 parishes, based on prior knowledge, is considered the optimal starting number of clusters.

Perform AHC clustering:

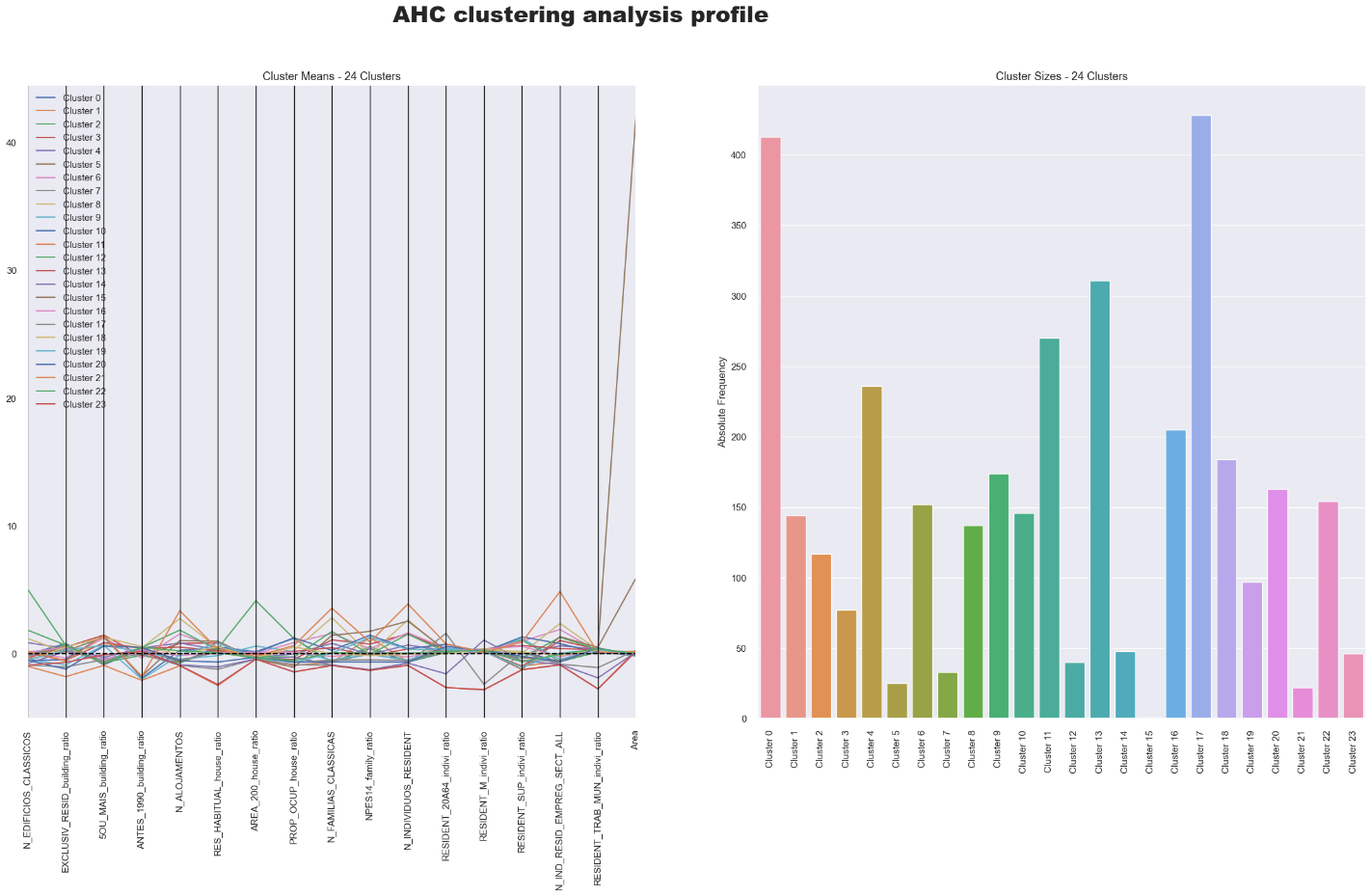
We can use the previously developed function to examine the AHC solution with 24 clusters. Firstly, we define the AHC model with a ward linkage and 24 clusters. Then, we fit the model to the dataset and assign the resulting labels to a new column in the DataFrame. Finally, we use the plot\_choropleth function to generate a choropleth map of the resulting clusters, with the title "Agglomerative Clustering Solution".



Analyze the results:

Upon examining the geographically dispersed cluster, it becomes clear that it is not inherently invalid, especially when the focus is on analyzing the overall structure and geographic context of the multivariate data. Visually, there may not be a striking difference since both the initial and k-means solutions would display similarly color-saturated maps to an observer unfamiliar with the data. Nevertheless, a rigorous and careful observer would discern a superior solution, as more regions are interconnected and colors are distributed across the map in larger segments. This results in a marked improvement of Tobler's first law of geography's visibility in the context of multivariate clustering.

We will proceed with the utilization of the "cluster\_profiles" function. The subsequent figure illustrates the analysis outcome for the AHC clustering solution, which generates 24 clusters.



Upon examining the AHC cluster\_profile plot, an integrated analysis can be presented that encompasses the distribution of clusters, variations in feature values, distinctive values, and the Area value outlier.

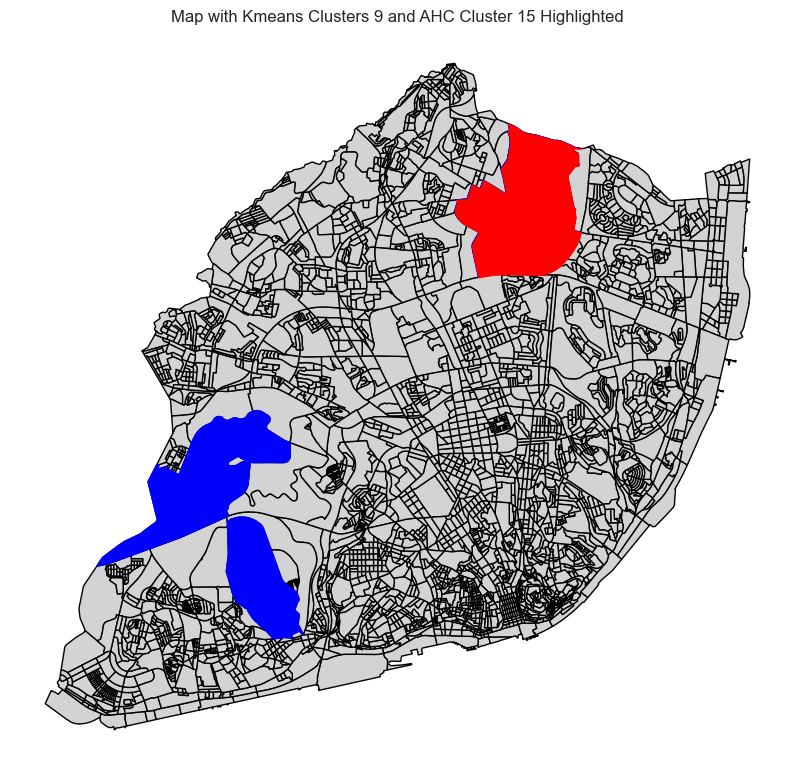
Cluster distribution: The dataset has been segmented into 24 distinct clusters, each comprising a varying number of units. Some clusters encompass over 400 units, while others contain fewer than 50. This disparity in distribution might suggest differing levels of similarity among the units within each cluster.

Variability in feature values: The features within the table reveal a spectrum of values across distinct clusters. Certain features, such as N\_EDIFICIOS\_CLASSICOS and N\_ALOJAMENTOS, demonstrate incremental changes among clusters. Conversely, other features like PROP\_OCUP\_house\_ratio and RESIDENT\_SUP\_indivi\_ratio exhibit more pronounced variations between clusters. These discrepancies in feature values might contribute to each cluster's distinctiveness.

Unique values: Specific clusters are distinguished by their exceptional feature values. For instance, cluster 11 possesses the lowest values for EXCLUSIV\_RESID\_building\_ratio and ANTES\_1990\_building\_ratio, while cluster 12 exhibits the highest value for N\_EDIFICIOS\_CLASSICOS. These particular values might signify distinct attributes within these clusters.

Outlier in Area value: It has been observed that in the AHC clustering method, cluster 15 exhibits an anomalously high Area value (42.231064) in comparison to other clusters. A similar phenomenon has been identified in cluster 9 during the K-means cluster analysis. This outlier may imply that the units within this cluster span a considerably larger geographical area than those in the remaining clusters. It is crucial to conduct a more in-depth investigation into the cause of this elevated Area value, as it could potentially impact the comprehensive analysis and interpretation of the clustering outcomes. Gaining a deeper understanding of the rationale behind the unusually high Area value will offer invaluable context for interpreting the clustering results and deriving insightful conclusions from the data.

In summary, this cohesive analysis of the centroid table underscores the diverse distribution of clusters, the variability in feature values, and the presence of unique values and outliers. Grasping these elements is vital for interpreting the clustering outcomes and gleaning valuable insights from the data.



Upon examining and comparing K-means cluster 9 and AHC cluster 15, we observed that they correspond to the most extensive sub-sections in Lisbon. The map above illustrates the largest sub-section, represented by the red region, which is encompassed by K-means cluster 9 and also forms part of AHC's cluster 15. The blue regions depict the other four sizeable sub-sections belonging to K-means cluster 9.

Further analysis revealed that the largest sub-section corresponds to Lisbon Airport, accounting for its unusual size. Due to its distinct nature and the fact that it is neither a residential area nor under the direct jurisdiction of the Lisbon Municipal Government, the airport was excluded from the analysis data. However, the Municipal Government is responsible for providing essential services such as security, fire protection, and medical assistance to the airport.

To ensure a more accurate and meaningful analysis, we adhere to the principle of proximity. After obtaining the final result, we integrate the outlier into the nearest group, thereby providing a comprehensive understanding of the clustering results in the context of Lisbon's sub-sections.