Classificação Geodemográfica de Lisboa

Sara Pinto

20230915@novaims.unl.pt

1. Objective

Development of a geodemographic classification for the city of Lisbon, based on a set of data collected in the 2001 Census, and containing metrics on the following areas:

- Level of Education;
- Year and type of construction of the Buildings;
- Age range of the population according to their gender;
- Professional Status;
- Constitution, Families;
- Types of accommodations.

To carry out this exercise, it is necessary to employ unsupervised classification methodologies, which allow the creation of groups, i.e., *clustering*.

2. Methods

2.1 Data Pre-Processing

In summary, the pre-processing of the data consists of the identification of missing values and *outliers*, i.e., extreme values that influence the mean of the variables. Although the transformation of outliers is not mandatory in *Data Mining*, there are some methods that are sensitive to its presence, such as *k-means,* since this algorithm is based on the calculation of averages to create groups.

After consulting the database, two rows with missing values were identified, corresponding to the statistical units of Marvila and Lumiar. These values were imputed by the nearest neighbor method: based on the location of each statistical unit, data from adjacent units were collected and the mean of the variables was calculated. With the imputed values, the ratios were calculated, thus completing the database.

Given the fact that the database already presented standardised data in the form of ratios, it was decided to use these values as an alternative to the variables themselves.

Through the Python libraries Pandas, NumPy and Matplotlib, *the outliers* were identified and dealt with (code attached).

As part of the pre-processing, a transformation of the data relating to the age groups was also carried out. To simplify the analysis, the age groups were reorganized into new variables, adding the values of each gender within each age group.

2.2 Variable Selection

In Knime, the variables were organized by the areas identified in the introduction and were subjected to a Principal Component Analysis (PCA), with the aim of determining which contributed most to the variability of the data.

2.3 Clustering with *K-means*

Or *Clustering* The optimum depends on the number of groups (number of k) that will be formed. In this sense, before proceeding to the *Clustering*, we tried to determine the best value of k. Knime contemplates the possibility of using the *Optimized k-means/Silhouette Coefficient*, which, as its name implies, determines the silhouette coefficient. The calculation of this coefficient includes the mean intra-cluster distance and the mean inter-cluster distance, providing a measure of the compactness of the clusters and the distance separating them (Han et al., 2012). This coefficient varies between -1 and 1, and the closer to 1, the more cohesive and distant the groups will be from each other. Knowing the ideal number of K, the next step consisted of *Clustering* Using the *K-Means*.

3. Results and Discussion

3.1 Exploratory Spatial Analysis of the data

The spatial representation of the percentage of individuals aged 65 and over (Figure 1) reveals that a large part of the population in this age group lives in the most central area of the city and by the river. On the other hand, the Chelas area, and the more peripheral areas (Ameixoeira, Lumiar, Carnide and Parque das Nações) are the ones with the lowest percentage of senior-age residents.

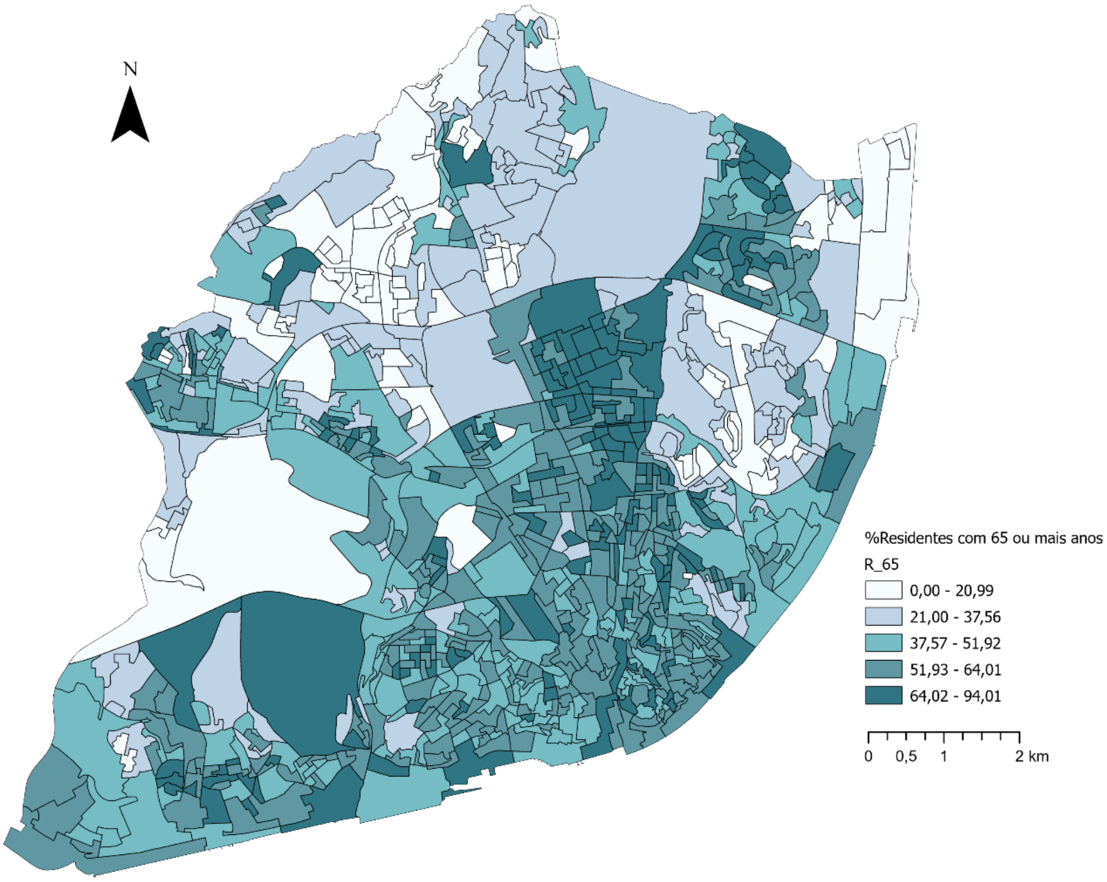


Figure 1 - Spatial representation of residents aged 65 and over.

3.1 Data Pre-Processing

As part of the pre-processing of the data, 978 *outliers (*Figure 2.*)*

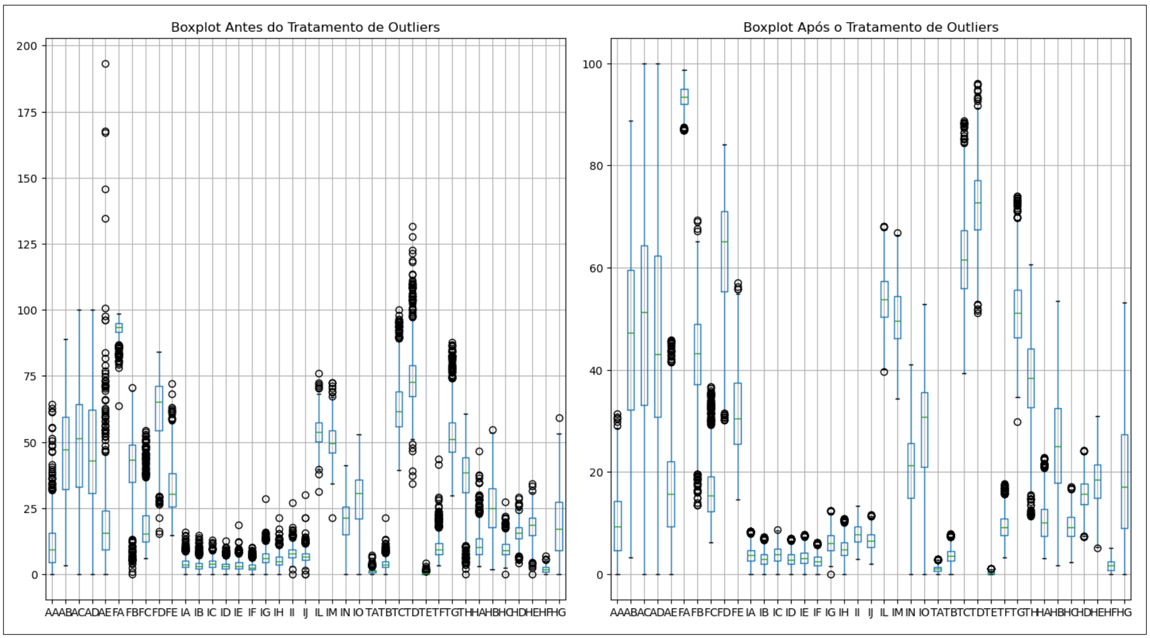


Figure 2 - Graphical representation of interquartile ranges: before treatment of outliers (left) and after treatment (right)

3.2 Variable Selection

According to the results in Appendix 3, the variables that most contributed to the variability of the data were selected for each thematic area (Table 1)

Table 1 - Variables selected for the Geodemographic Classification of Lisbon.

Thematic Area	Code	Assignment
Accommodation	AB	With 3 or 4 rooms
	AC	Leased
	AD	Own Homes
Buildings	E919	Built before 1919
	E1945	Built before 1945
	E1960	Built before 1960
	E2001	Built before 2001
	EARG	Mortared masonry walls
	EBAR	Resistant concrete elements
	ER	Exclusively residential
Schooling	HA	He can't read or write
	HB	1st Cycle of Complete Basic Education
	HE	Complete Secondary Education
	HG	Complete Higher Education
Age group	R5_9	Children (ages 5-9)
	R14_19	Young teens (ages 14-19)
	R25_64	Adults (ages 25-64)
	R65	Seniors (65 years and older)
Classic Families	FB	Made up of persons under the age of 15
	FC	Made up of people over 65 years of age
	FD	Consisting of 1 or 2 people
	FE	Consisting of 1 or 2 people
Professional Status	TC	Employees
	TD	No economic activity
	TH	Pensioners or Retirees

3.3 Clustering

The silhouette coefficient was determined for all thematic areas, and on all occasions, the coefficient was higher for k=2. Although this result was always consistent, in some cases the difference with k=3 was small. For these cases, clustering with k=3 was tested, and the use of an alternative method, the *Self-Organizing-Map,* but the spatial representation of the results did not improve. In this sense, the analysis included in the present study was fully performed with *k-means*, for k=2. For reasons of time, it was not possible to calculate and compare the errors produced by each method, but future improvements of this work should ensure that this step is carried out.

It should be added that the area of the Airport and the Monsanto Forest Park are somewhat atypical, so the results obtained in their statistical units should be interpreted with caution.

3.3.1. Accommodation

The Figure 3 reveals that cluster 0 has a higher average for rented homes, while cluster 1 has a higher average for owned homes. Through the spatial representation, it is observed that, in general, in the quadrant of the North and Southwest periphery of the city, own houses prevail (Olivais, Parque das Nações, Lumiar, Carnide, Benfica, Alvalade, Restelo. In the area by the river, in the central area, most of the residents live in rented houses.

3.3.2. Buildings

Through the Figure 4, it is observed that most of the buildings in the city of Lisbon have an old construction. In fact, the highest averages correspond to buildings constructed before 1919 and 1945 in cluster 0. Logically, it is also in this group that the construction of the walls is done almost exclusively by mortared masonry. Most of these buildings are exclusively residential.

On the other hand, the buildings grouped in cluster 1 correspond to more recent constructions (the highest averages are reached in constructions prior to 2001 and 1960), and in addition to residences, probably also include commercial spaces, services or hotels (and similar accommodations), since the average for the RE variable is low. Logically, the construction of the buildings in this cluster used more concrete than masonry.

In general, it can be inferred that most of the buildings in cluster 0 correspond to the oldest and exclusively residential buildings, and those in cluster 1 to the most recent buildings and that they will be spaces of mixed nature.

Spatially, it is observed that cluster 1 corresponds to the outskirts of the city, and cluster 0 prevails more in the areas close to the Tagus River, excluding the area of Restelo and Parque das Nações.

3.3.3 Schooling

With regard to schooling, the Figure 5 reveals that the population with more education (higher education or complete secondary education) resides in the central area, in the Northwest and at both ends of the city (cluster 1), namely, Avenidas Novas, Areeiro, Alvalade, Carnide, Lumiar, Benfica, Campolide, Sto. António, Campo de Ourique, Estrela, Parque das Nações and Restelo. Cluster 0 corresponds to residents with the lowest levels of schooling (mostly residents with the 1st cycle of basic education), and its spatial distribution coincides with the area of Ameixoeira, Santa Clara, Chelas, Marvila, Alcântara, Ajuda, Castelo, Graça.

3.3.4 Age range

The radar chart of the Figure 6 shows that a large part of the population of the city of Lisbon is adult or elderly, and that only a small proportion is made up of children and young adolescents, which reflects the phenomenon of low birth rate. Although the composition of cluster 0 and 1 is quite similar, cluster 0 is the one with the highest mean for residents over 65 years of age, and cluster 1 is the one with the highest mean for adults aged between 25 and 64 years. Spatially, the oldest population resides in the most central area of the city, including Restelo, while the most peripheral area, including Parque das Nações, tends to be more inhabited by the adult population. These results are in agreement with the spatial representation of residents aged 65 years or older (Figure 1).

3.3.5 Classic Families

The analysis of the Figure 7, points to a pattern that is apparently contradictory to the one identified in the Figure 6 and the Figure 1, which reveal that the younger population lives in the periphery, and that the older population lives in the more central part of the city. In fact, cluster 1 is the one with the highest average for families made up of people over 65 years of age. However, it is important to emphasise that this thematic area concerns families, not isolated individuals. In this sense, it is plausible that in the peripheries there may be more families sharing their home with grandparents, which is corroborated by the higher average of these same clusters for the FE variable (families consisting of 3 or 4 people).

On the other hand, families consisting of only 1 or 2 people are the ones that simultaneously reach the highest average for the inclusion of people under 15 years of age. This cluster may represent the presence of single-parent families or divorced parents, who have young people and children in their care.

3.3.6 Professional status

In this thematic area (Figure 8), both clusters have a high average for the presence of residents without economic activity (TD). What distinguishes the two clusters is the fact that cluster 0 presents a lower mean for the variable TC (employees) and a higher mean for TH (pensioners or retirees), compared to cluster 1. Although the spatial pattern of this thematic area is a little more erratic than the previous ones, there is a tendency for the working population to live closer to the periphery and in the area of Chelas and Olivais. On the other hand, the areas of Alvalade, Areeiro, Penha de França, Marvila, Ajuda, Alcântara are more inhabited by pensioners or retirees. It should be noted, however, that the Parque das Nações area is also included in this last cluster, which contradicts the findings of the Figure 6, i.e., this area is composed of both the population between 25 and 60 years of age and retired persons or pensioners. After verifying the values that the variable TC (Employees) reaches in this geographical area, it was found that it is much higher than that of retirees. In this sense, a clustering error was detected.

Given the low discriminatory power of this thematic area, one of the suggestions to overcome the error would be to eliminate the TD variable (no economic activity), since it may be hindering the process of allocating clusters unnecessarily. It should be noted that this error was easily detected because it is a statistical unit located at one end, but that there are possibly more errors of this type in this thematic category.

ALOJAMENTOS

AC: Arrendados
AB: Com 3 ou 4 divisões
AD: Casas Próprias

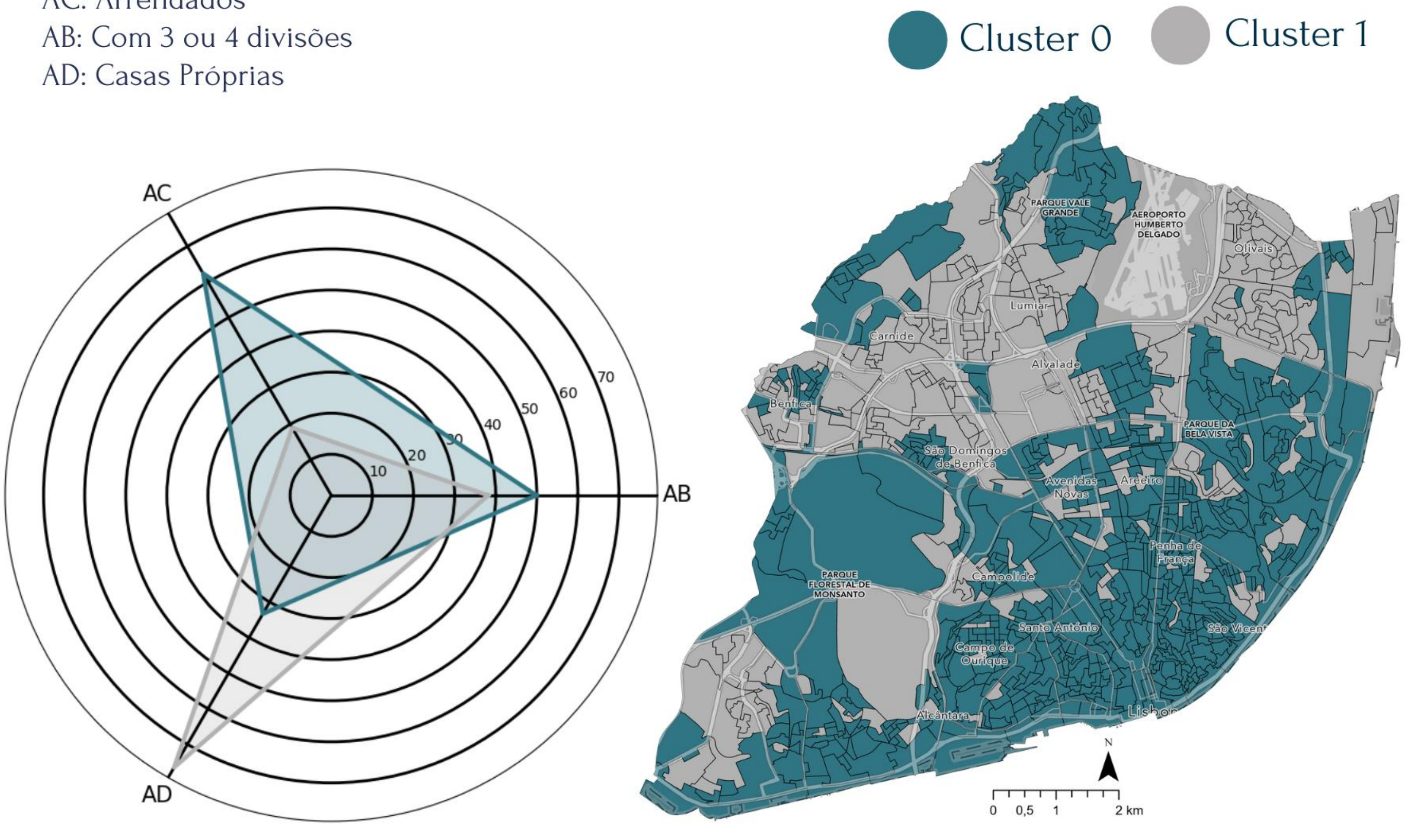


Figure 3 - Graphical and spatial representation of the clustering performed for the variables related to Accommodations.

EDIFÍCIOS

E1919: Construídos antes de 1919
E1945: construídos antes de 1945
E1960: construídos antes de 1960
E2001: construídos antes de 2001
EARG: Paredes de alvenaria argamassada
EBAR: Elementos resistentes de betão
ER: Exclusivamente residenciais

● Cluster 0 ● Cluster 1

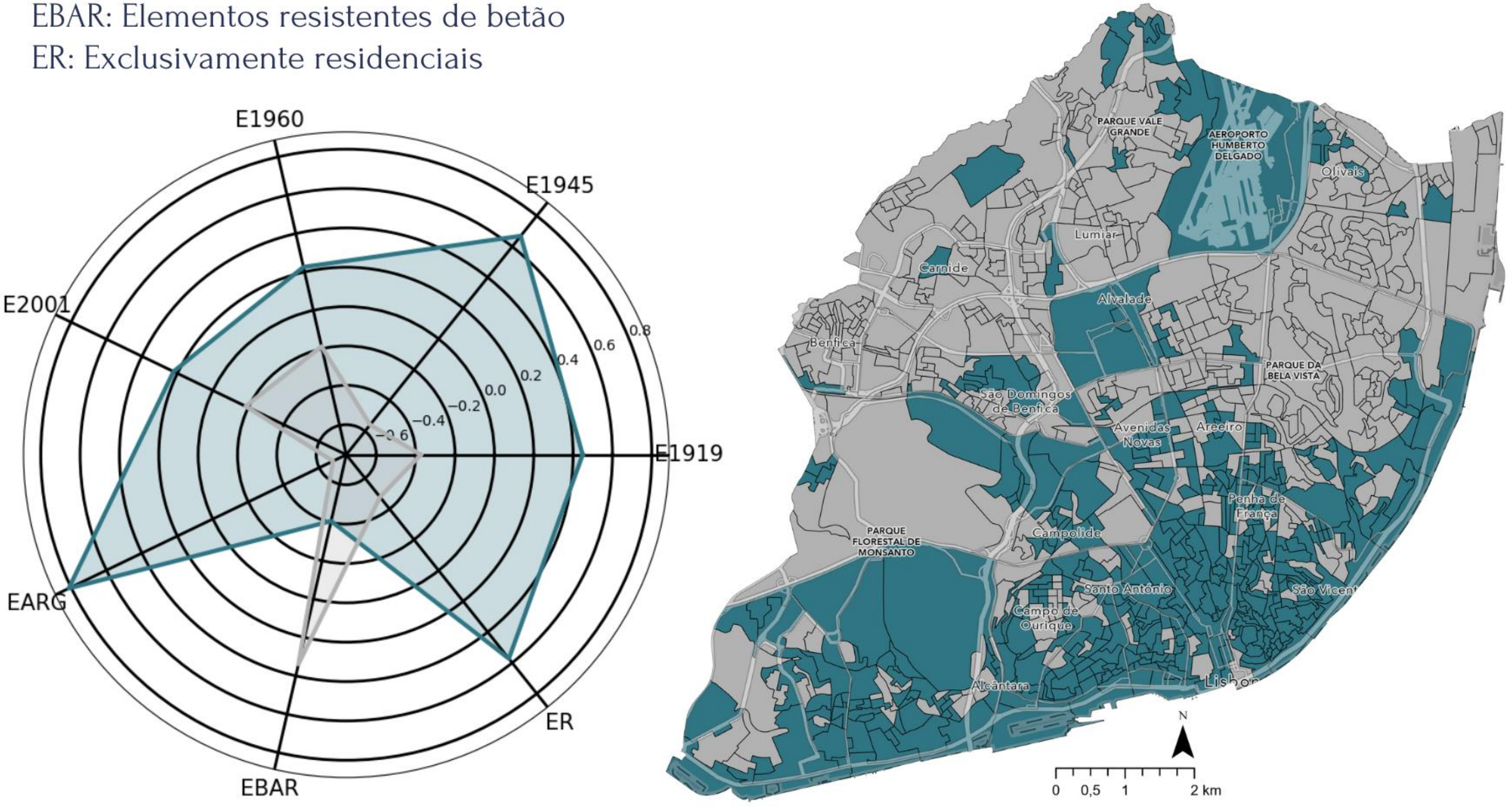
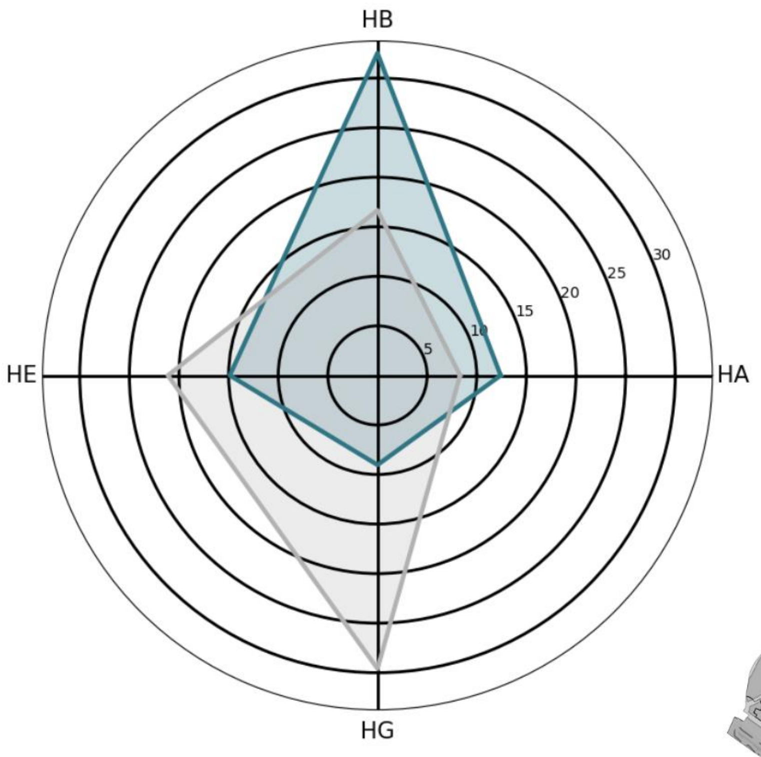


Figure 4 - Graphical and spatial representation of the clustering carried out for the variables related to the Buildings.

ESCOLARIDADE

HB: 1º Ciclo do Ensino Básico Completo
HA: Não sabe ler, nem escrever
HG: Ensino Superior Completo
HE: Ensino Secundário Completo

● Cluster 0 ● Cluster 1



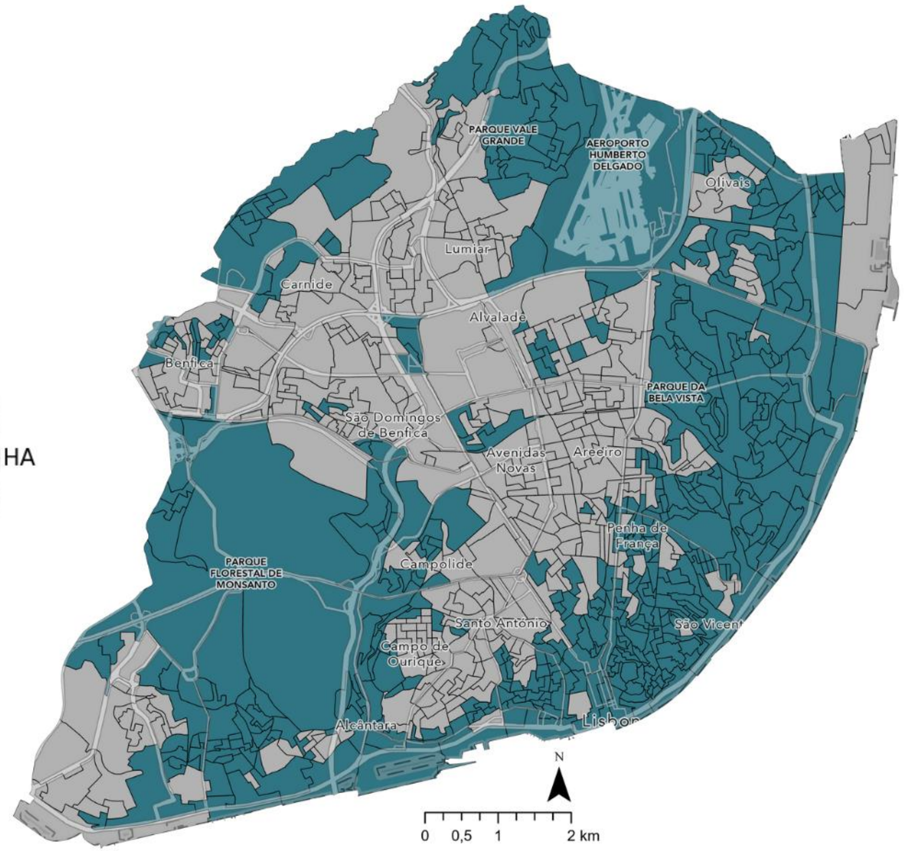
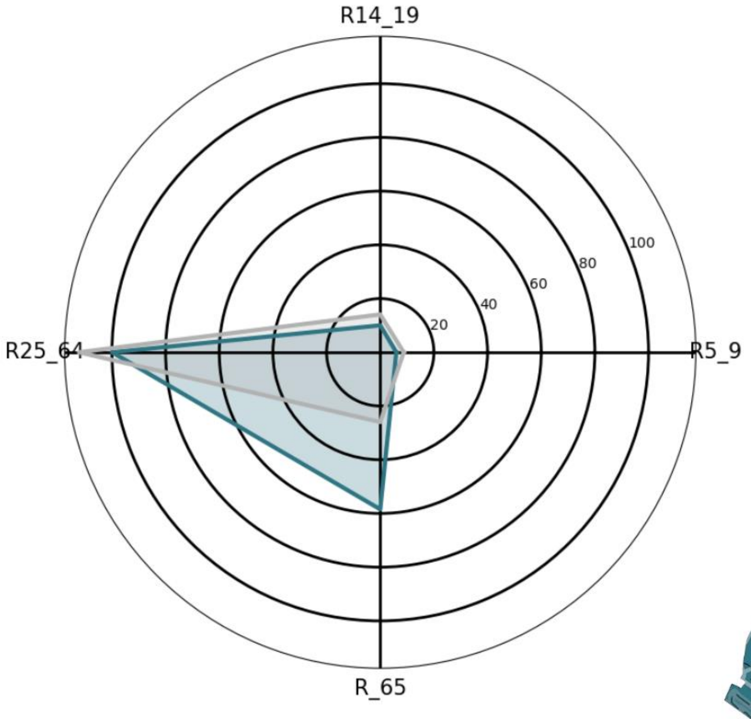


Figure 5 - Graphical and spatial representation of the clustering performed for the variables related to Schooling.

FAIXA ETÁRIA

R5_9: Crianças (idades entre 5 e 9 anos)
R14_19: Jovens adolescentes (idades entre 14 e 19 anos)
R25_64: Adultos (idades entre 25 e 64 anos)
R_65: Séniores (65 ou mais anos)

● Cluster 0 ● Cluster 1



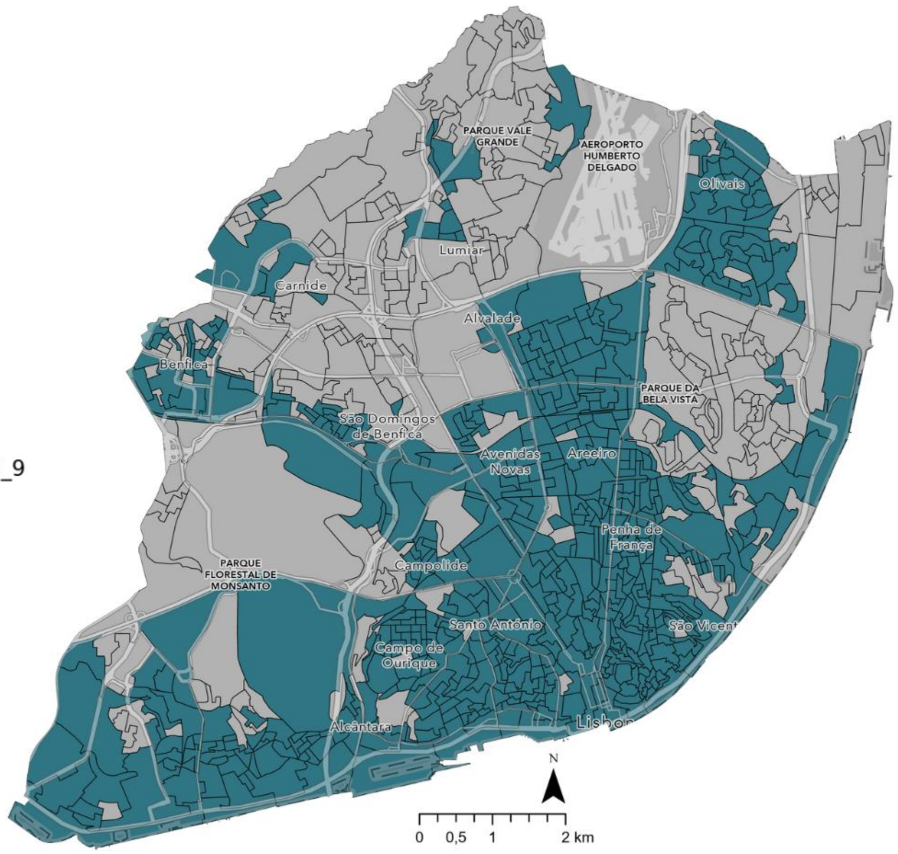
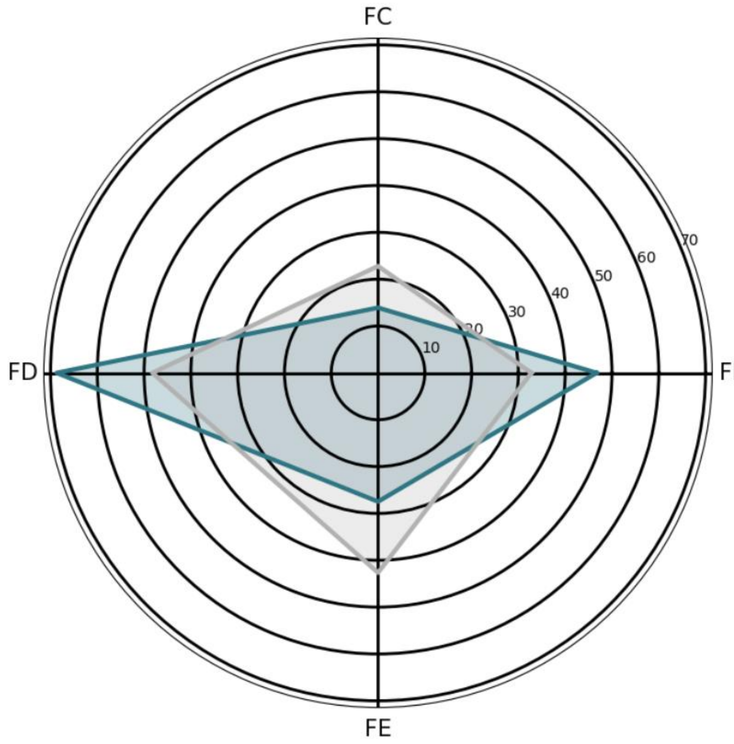


Figure 6 - Graphical and spatial representation of the clustering performed for the variables related to the Age Groups

FAMÍLIAS CLÁSSICAS

FC: Constituídas por pessoas com mais de 65 anos
FB: Constituídas por pessoas com menos de 15 anos
FE: Constituídas por 3 ou 4 pessoas
FD: Constituídas por 1 ou 2 pessoas



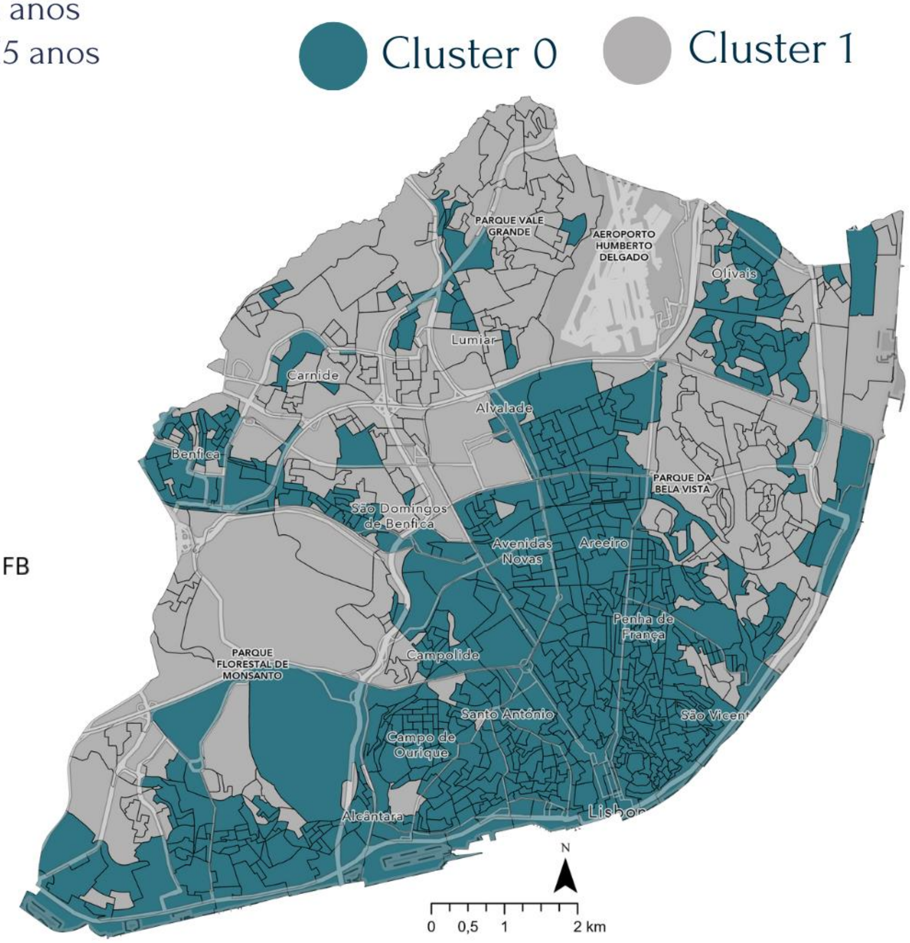
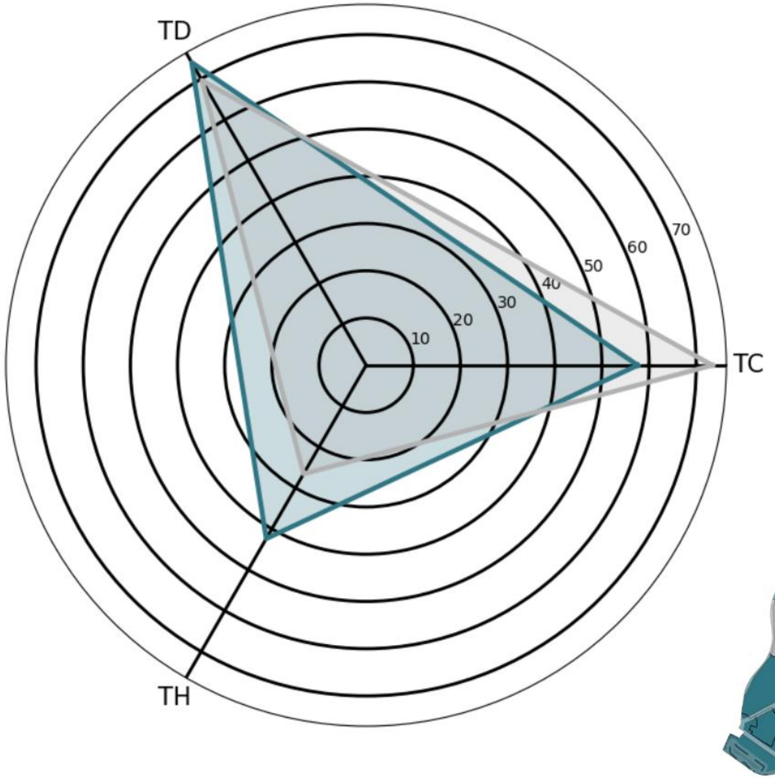


Figure 7 - Graphical and spatial representation of the clustering performed for the variables related to the Classic Families.

SITUAÇÃO PROFISSIONAL

TC: Empregados
TD: Sem Actividade Económica
TH: Pensionistas ou Reformados

● Cluster 0 ● Cluster 1



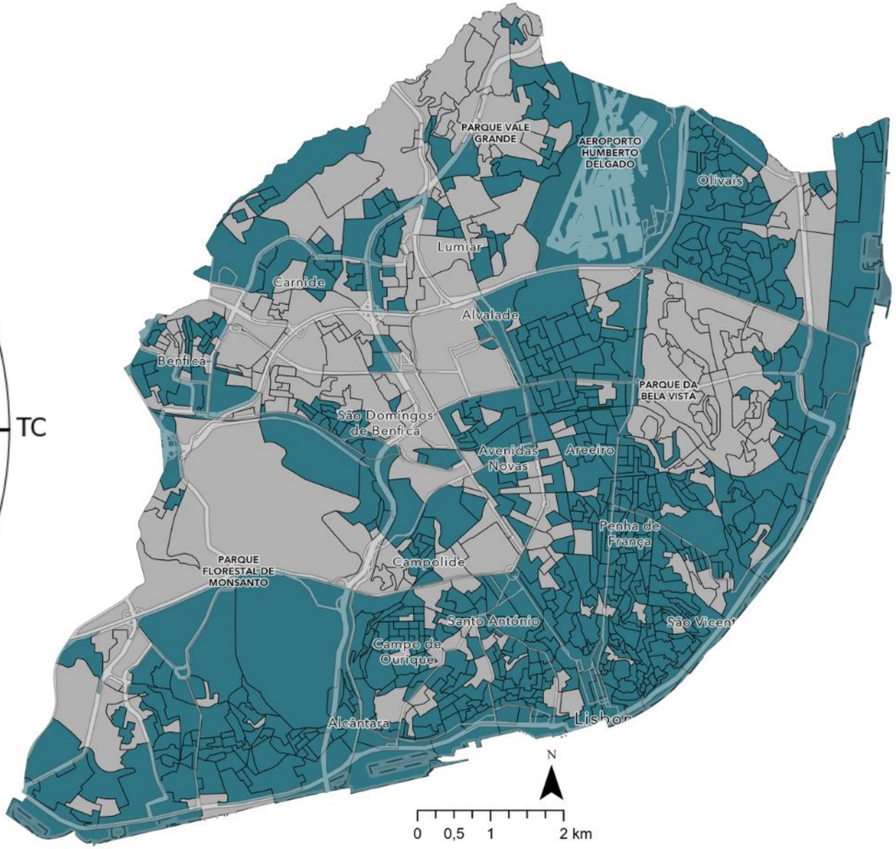


Figure 8 - Graphic and spatial representation of the clustering performed for the variables related to the Professional Situation.

4. Conclusions

Through this work it was possible to develop a geodemographic classification of Lisbon. Despite the errors detected, it is considered that in the vast majority of the results obtained are coherent and that they offer a reliable portrait of the demographic and social dynamics of the city. In future work, it would be important not only to compare the errors obtained between different clustering methods, but also to mix variables in order to realize other clusters. Due to constraints associated with lack of time, it was not possible to carry out all the initially intended clusters, but the overlapping of the information obtained for each of the thematic areas allows us to state that, tendentially and in a general way (there are, of course, many exceptions):

- The older population resides in rented houses in the city centre;
- The adult and working-age population lives mostly in the peripheral area, in their own homes;
- Families living in the most peripheral areas are larger and include members aged 65 and over; On the other hand, families residing in the city center are smaller and include a child or an adolescent;
- The population with lower levels of education lives in the eastern part of the city, with the exception of Parque das Nações, and by the river;
- The most recent buildings are located on the outskirts and the oldest in the central area of the city.

5. References

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining concepts and techniques* (3rd Edition). Morgan Kaufmann.

Appendix 1 – Python code for detection, outlier handling, and boxplot representation

```
Import Pandas as PD
import numpy as np
import matplotlib.pyplot as plt

# Function to handle outliers through the median in all numeric columns except the
'DTCCFRSEC0' column which only identifies the UEs.
def treat_outliers_with_median(df):
 outlier_counts = {}
    for column_name in df.select_dtypes(include='number').columns:
        if column_name == 'DTCCFRSEC0':
            Go on
 col = df[column_name]
 median = col.median()
 q1 = col.quantile(0.25)
 q3 = col.quantile(0.75)
 IQR = Q3 - Q1

        # Setting the thresholds for detecting outliers
 lower_bound = Q1 - 1.5 * IQR
 upper_bound = Q3 + 1.5 * IQR

        print(f"Column: {column_name}")
        print(f"Lower bound: {lower_bound}, Upper bound: {upper_bound}")

        # Count of outliers before replacement
 outliers = col[(col < lower_bound) | (col > upper_bound)]
 outlier_counts[column_name] = len(outliers)
        print(f"Number of outliers detected: {len(outliers)}")

        # Replacement of outliers with median
 df[column_name] = col.apply(lambda x: median if x < lower_bound or x > upper_bound
else x)

        # Verification of the amounts treated
        print(f"Values after: {df[column_name].describe()}\n")
    return df, outlier_counts


# Data import
input_table_1 = pd.read_excel('C:/Lx/Lx_dados_imputados.xlsx')

# Statistics before handling outliers
print("Statistics before handling outliers:")
print(input_table_1.describe())
```

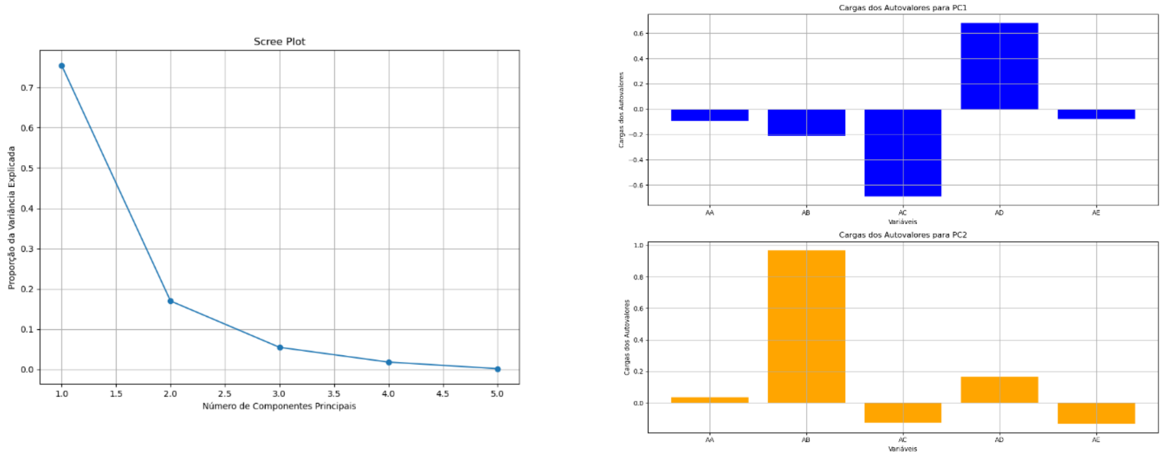
```python
# Applying the function to the input DataFrame
output_table_1, outlier_counts = treat_outliers_with_median(input_table_1.copy())

# Statistics after outliers treatment
print("\nStatistics after handling outliers:")
print(output_table_1.describe())

# Count of outliers handled
print("\nNumber of outliers handled per column:")
for column, count in outlier_counts.items():
    print(f"{column}: {count}")

# Creation of boxplots to compare before and after treatment, excluding the
'DTCCFRSEC0' column
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(14, 7))

# Boxplot before treatment of outliers
input_table_1.drop(columns=['DTCCFRSEC0'], inplace=True)
output_table_1.drop(columns=['DTCCFRSEC0'], inplace=True)

input_table_1.boxplot(ax=axes[0])
axes[0].set_title('Boxplot Before Outliers Treatment')

# Boxplot after treatment of outliers
output_table_1.boxplot(ax=axes[1])
axes[1].set_title('Boxplot After Outliers Treatment')

plt.tight_layout()
plt.show()
```

Appendix 2 – Python code used on a knime node to perform PCA, determine the screeplot and load of eigenvalues

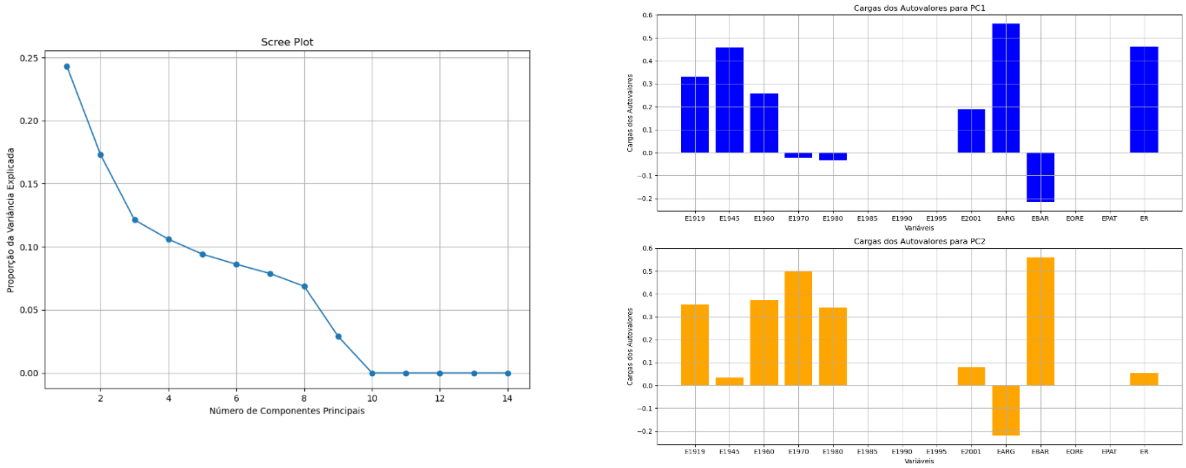
```
# Upload data
data = input_table_1.copy()

# Delete the "STOCFRESCO" column
data = data.drop(columns=['DTCCFRSEC0'])

# Run PCA
PCA = PCA()
pca.fit(data)

# Get the loadings
Loadings = PD. DataFrame(pca.components_. T, columns=[f'PC{i+1}' for i in
range(len(data.columns))], index=data.columns)

# Plot loads of eigenvalues with bar chart for PC1 and PC2
fig, axs = plt.subplots(2, 1, figsize=(12, 10))

axs[0].bar(loadings.index, loadings['PC1'], color='b')
axs[0].set_title('Payloads of Eigenvalues for PC1')
axs[0].set_xlabel('Variables')
axs[0].set_ylabel('Payloads of Eigenvalues')
axs[0].grid(True)

axs[1].bar(loadings.index, loadings['PC2'], color='orange')
axs[1].set_title('Payloads of Eigenvalues for PC2')
axs[1].set_xlabel('Variables')
axs[1].set_ylabel('Payloads of Autovalues')
axs[1].grid(True)

plt.tight_layout()
plt.show()

# Scree Plot
plt.figure(figsize=(10, 7))
plt.plot(range(1, len(pca.explained_variance_ratio_) + 1),
pca.explained_variance_ratio_, marker='o')
plt.title('Scree Plot')
plt.xlabel('Number of Core Components')
plt.ylabel('Ratio of Variance Explained')
plt.grid(True)
plt.show()

# Prepare the output table
output_table_1 = input_table_1.copy()
```

Appendix 3 – Results obtained through Appendix 1: screeplot and eigenvalue load of the analyzed variables

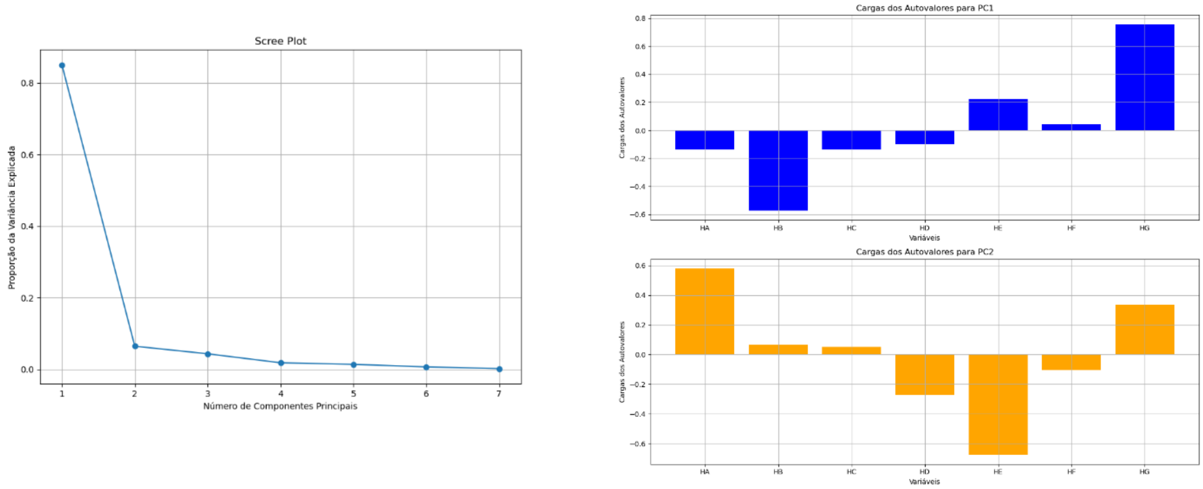
PCA - Accommodations



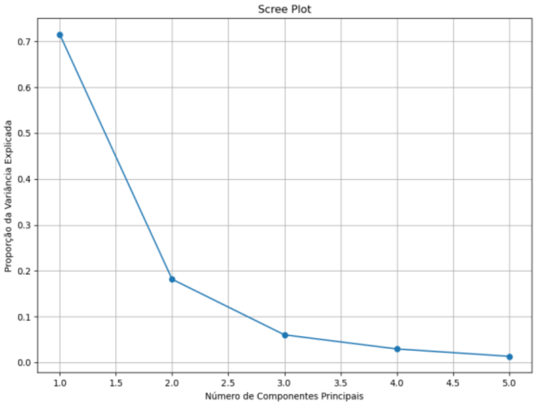
PCA - Buildings

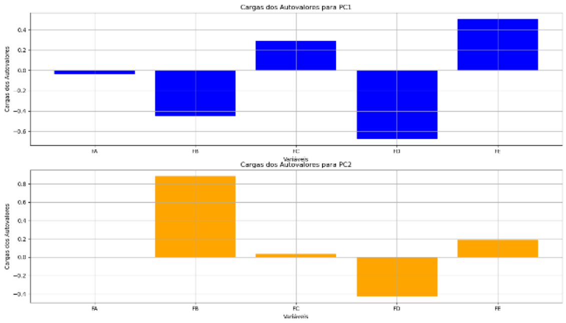


PCA - Schooling

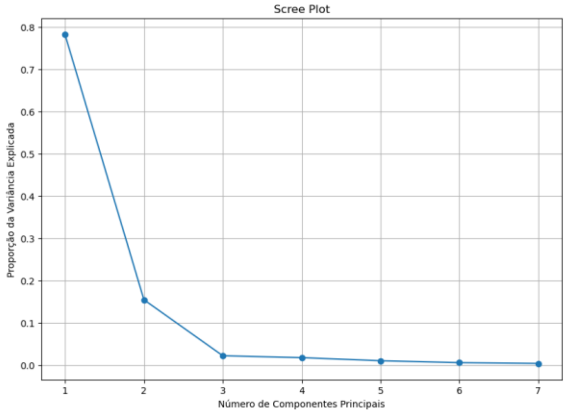


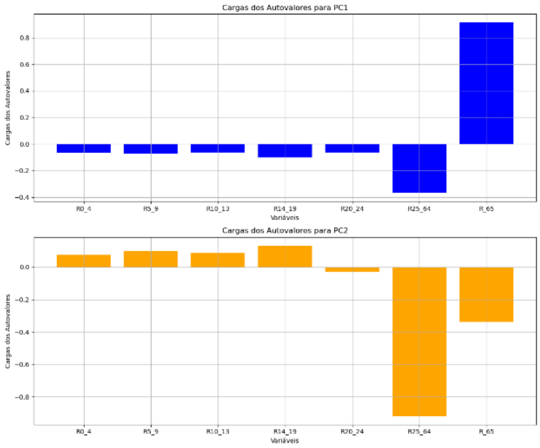
PCA – Classic Families



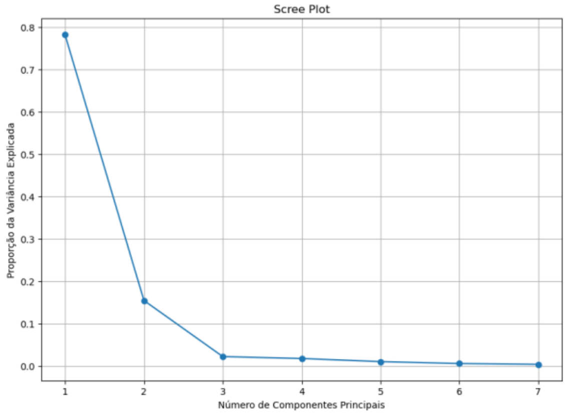


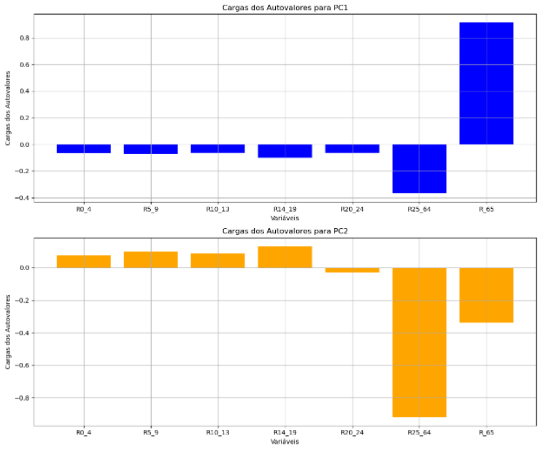
PCA – Age Groups





PCA – Professional Status


Scree Plot
Proporção da Variância Explicada
Número de Componentes Principais


Cargas dos Autovalores para PC1
Cargas dos Autovalores
R0_4
R5_9
R10_13
R14_19
R20_24
R25_64
R_65
Variáveis
Cargas dos Autovalores para PC2
Cargas dos Autovalores
R0_4
R5_9
R10_13
R14_19
R20_24
R25_64
R_65
Variáveis