Choosing a neighborhood to live in Porto Alegre

Marcus Vinicius Pereira

August 2020

1. INTRODUCTION

1.1 Background

Porto Alegre is the capital of the southernmost state of Brazil, Rio Grande do Sul. In Porto Alegre, approximately 1.5 million people live in an area of 496,682 km². The city was formed from the arrival of Azorean couples in the middle of the 18th century. Around the 19th century, the city received many German and Italian immigrants, as well as Spaniards, Africans, Poles and Lebanese.

In Porto Alegre, approximately 37% of people live in apartments. The city suffers from problems such as sub habitation, high cost of living and traffic jams. On the other hand, the city also stands out for being one of the best cities in Brazil to live, work, study and have fun.

1.2 The Problem

If someone were planning to move to Porto Alegre, it would be very useful to know which neighborhoods to consider when choosing a home. Those looking for housing do this type of neighborhood categorization intuitively. It would be much easier just by using population density, average income per housing and facilities in the neighborhoods, and then grouping them. Grouping the neighborhoods would greatly reduce the effort to find housing for a newcomer who does not know the city.

2. DATA

2.1 Data acquisition

Three data sources will be used, as follows:

- Wikipedia for data on population density and average income per housing in each neighborhood (here)
- Foursquare API to get the most common venues of given neighborhood of Porto Alegre
- File downloaded from the Porto Alegre Observa POA website with the Porto Alegre neighborhoods area, in shp format (here)

It is intended in this work to group the neighborhoods of Porto Alegre by similarity, identifying each group by its characteristics.

2.2 Data cleaning

Firstly, it is possible to see that the data on population density and average income per household taken from Wikipedia does not contain data from many neighborhoods in Porto Alegre. Such neighborhoods were disregarded from the analysis.

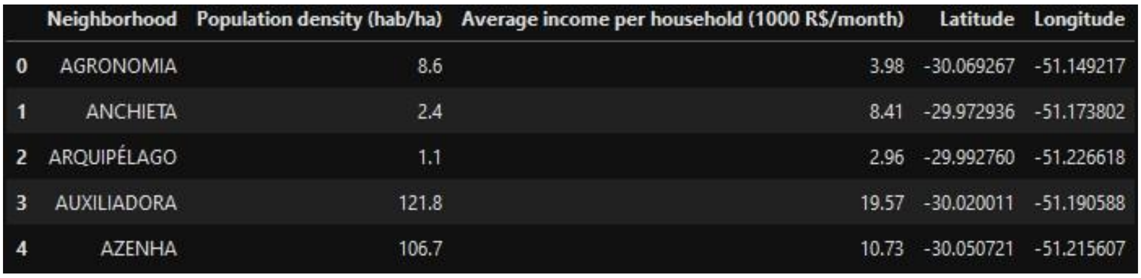
It was found that the longitude and latitude data for the São José neighborhood were wrong. An adjustment was made to adapt the position of the neighborhood in the city of Porto Alegre.

3. METHODOLOGY

The methodology applied in this work followed the following steps:

I. Data acquisition: Wikipedia, Porto Alegre neighborhoods geolocation, Foursquare Venues data. For getting venues, it is considered a radius of 500 meters around of the centroid of each neighborhood
II. Preparing data for Analysis: data cleaning e neighborhoods geolocation adjustments
III. Clustering neighborhoods: clustering neighborhoods by venues, population density and average income per household using Kmeans from scikit-learn library
IV. Analyze clusters, identifying the main characteristics of each cluster.

Data acquisition and data cleaning are described in section 2 of this document. After those, the master data is as follows:

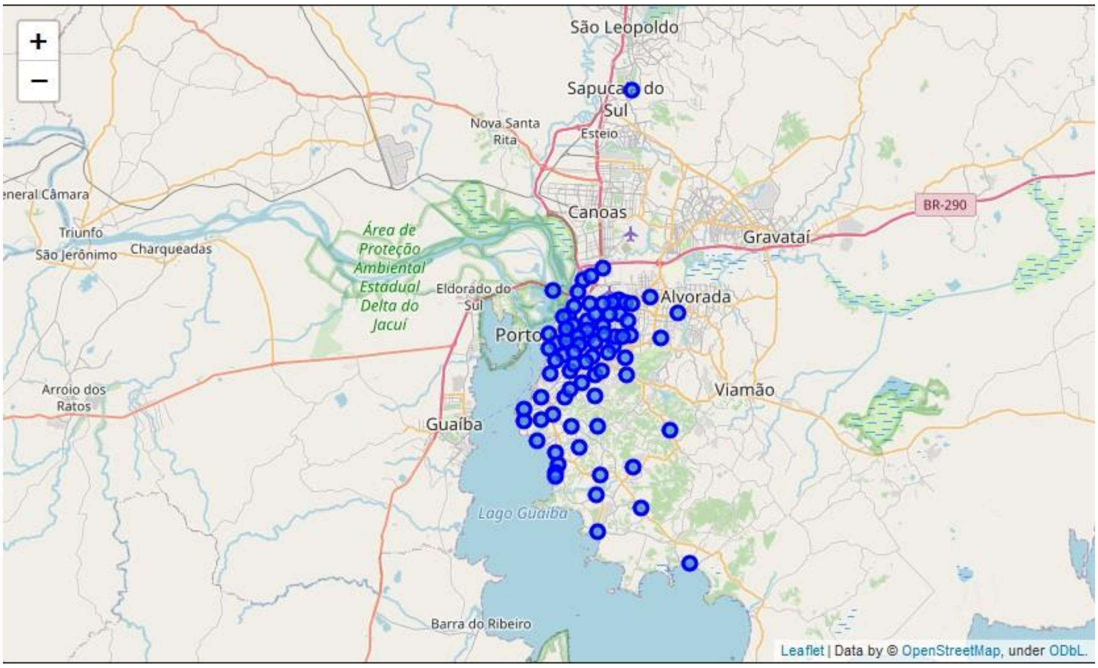


	Neighborhood	Population density (hab/ha)	Average income per household (1000 R$/month)	Latitude	Longitude
0	AGRONOMIA	8.6	3.98	-30.069267	-51.149217
1	ANCHIETA	2.4	8.41	-29.972936	-51.173802
2	ARQUIPÉLAGO	1.1	2.96	-29.992760	-51.226618
3	AUXILIADORA	121.8	19.57	-30.020011	-51.190588
4	AZENHA	106.7	10.73	-30.050721	-51.215607

Table 1 – Master data (5 first rows)

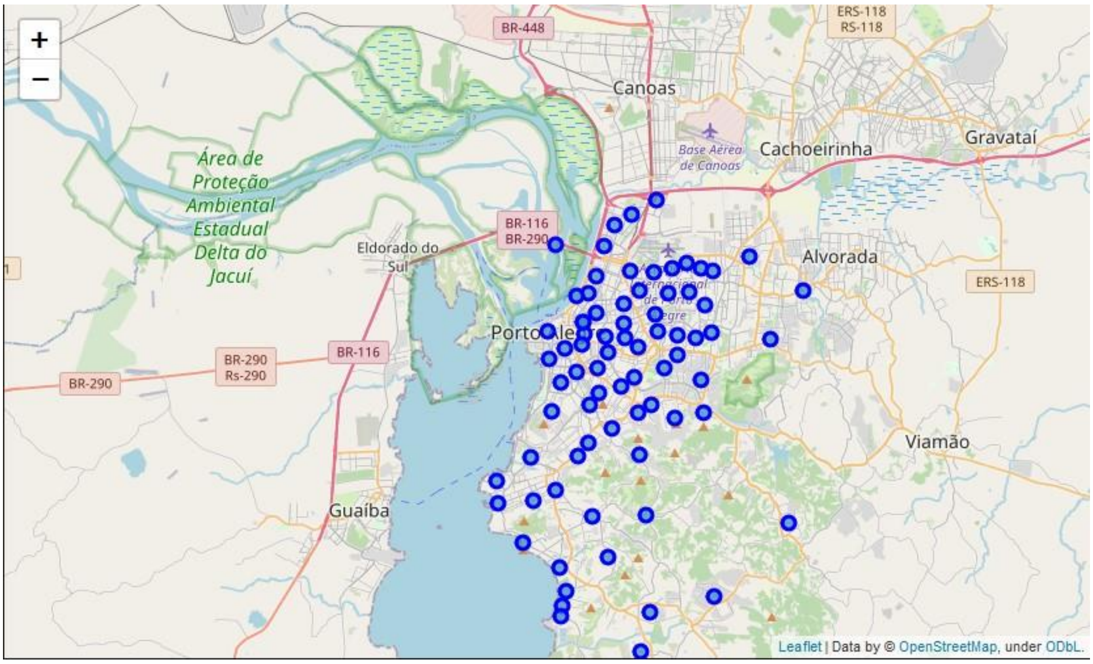
I was used folium library to visualize where the center of each neighborhoods are, generating figure 1. Wikipedia provided data of 79 neighborhoods of Porto Alegre.

Figure 1 – Porto Alegre neighborhoods considered in this analysis, São José misplaced



As we can see in the figure 1, São José neighborhood is misplaced. After an adjustment, the map becomes as is in figure 2.

Figure 2 - Porto Alegre neighborhoods considered in this analysis



After merging master data (from Wikipedia) with data of boundaries of each neighborhood (geolocation of each neighborhood), it was created figures 3 and 4. The figures show how neighborhoods are represented in terms of population density and average income per household, respectively.

Figure 3 - Porto Alegre neighborhoods – population density (hab/ha)

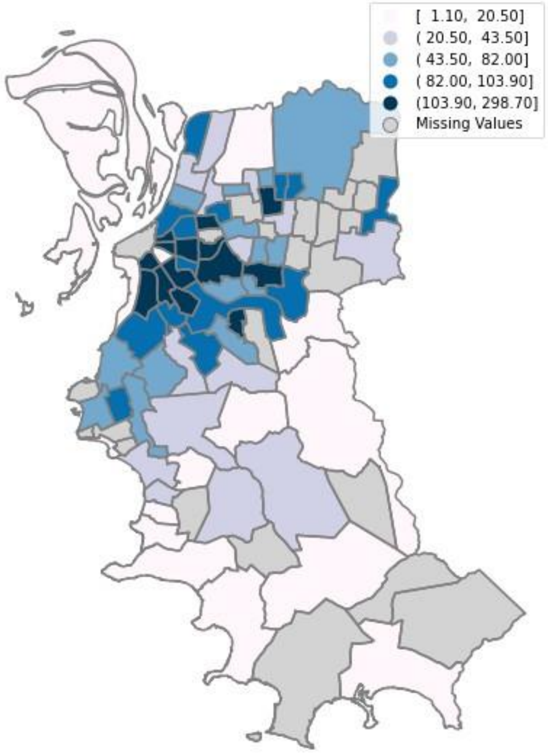
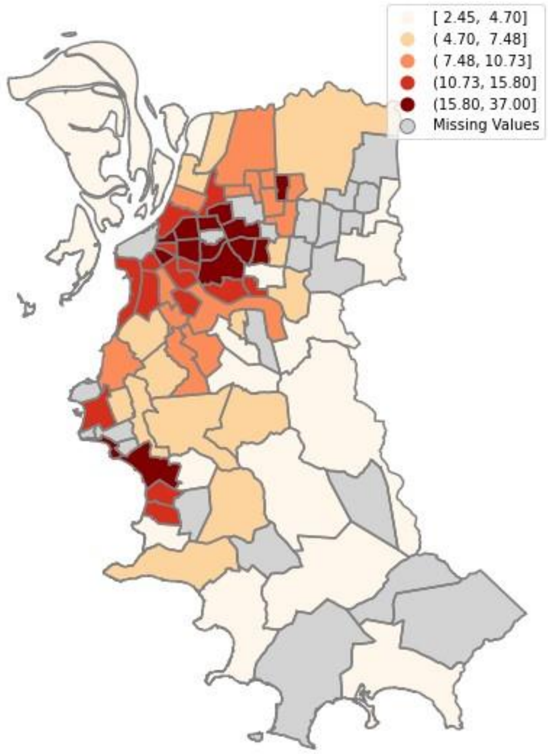
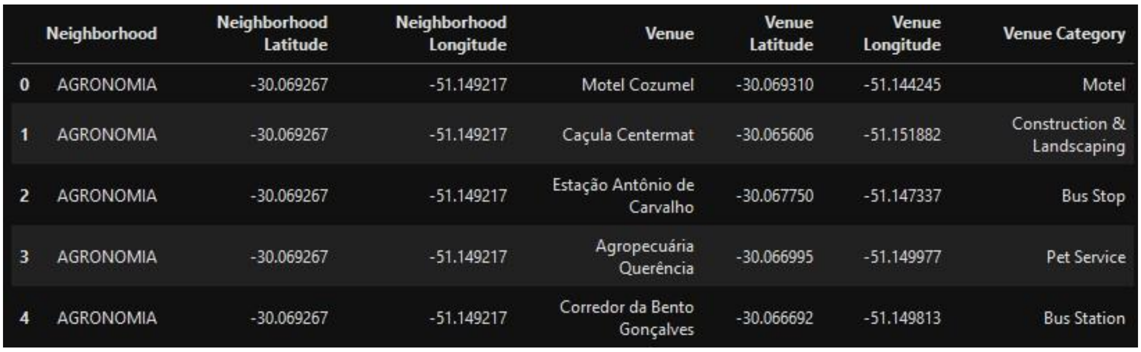


Figure 4 - Porto Alegre neighborhoods – average income per household (R$ 1000/house)



Using Foursquare API, it was collected data of the venues for each neighborhood. The venues are inside a circle that have 500 meters of radius, whose center is the centroid coordinates of each neighborhood. Table 2 shows the first 5 rows of the data. There are 278 unique venues in the data collected.



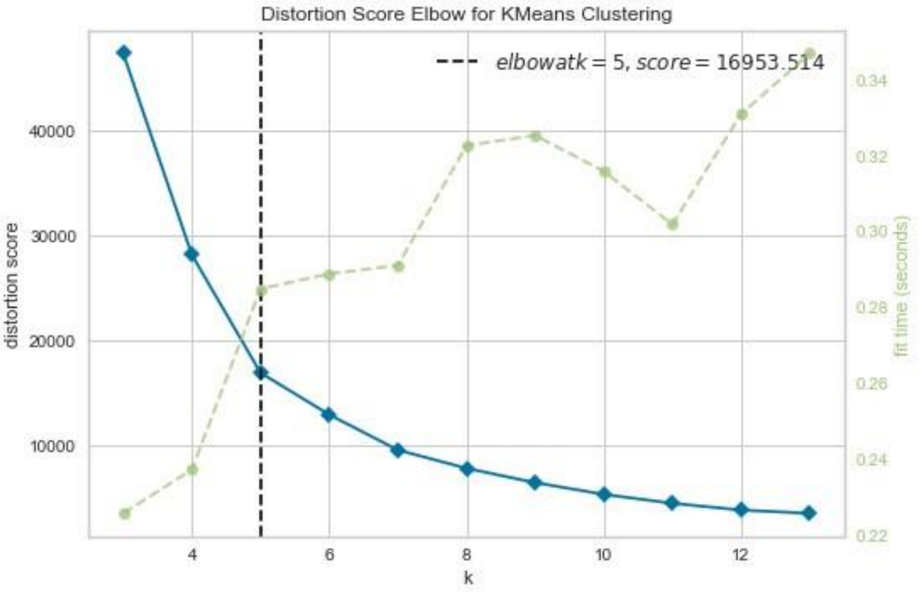
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	AGRONOMIA	-30.069267	-51.149217	Motel Cozumel	-30.069310	-51.144245	Motel
1	AGRONOMIA	-30.069267	-51.149217	Caçula Centermat	-30.065606	-51.151882	Construction & Landscaping
2	AGRONOMIA	-30.069267	-51.149217	Estação Antônio de Carvalho	-30.067750	-51.147337	Bus Stop
3	AGRONOMIA	-30.069267	-51.149217	Agropecuária Querência	-30.066995	-51.149977	Pet Service
4	AGRONOMIA	-30.069267	-51.149217	Corredor da Bento Gonçalves	-30.066692	-51.149813	Bus Station

Table 2 – Venues data of each neighborhood

After that, the neighborhoods were grouped by venues, but also by population density and average income per household. It was used K-means algorithm (from scikit-learn library) to cluster the neighborhoods.

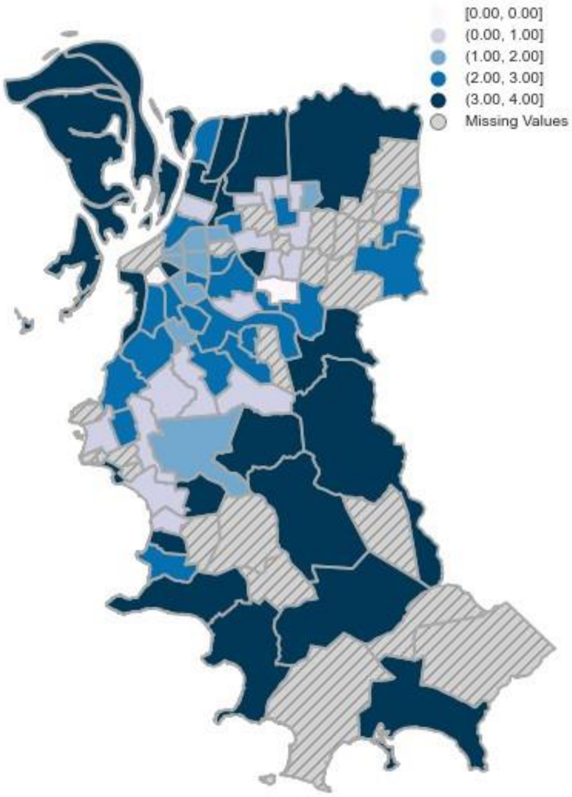
To select the optimum number of clusters, I used the elbow method aided by yelowbrick library. Figure 5 shows that the optimum number of clusters is 5 (0 to 4).

Figure 5 – Elbow method

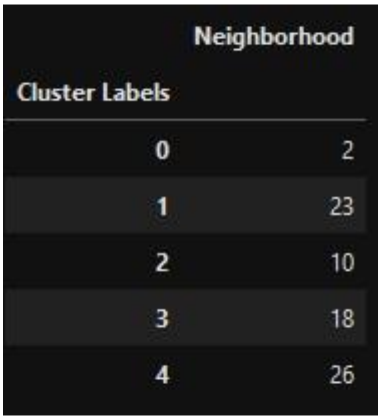


After proceeding with clustering, Porto Alegre’s map with clustered neighborhoods are show in figure 6.

Figure 6 – Porto Alegre’s map with clustered neighborhoods



The number of neighborhoods that each cluster contains is shown in table 3.



	Neighborhood
Cluster Labels	
0	2
1	23
2	10
3	18
4	26

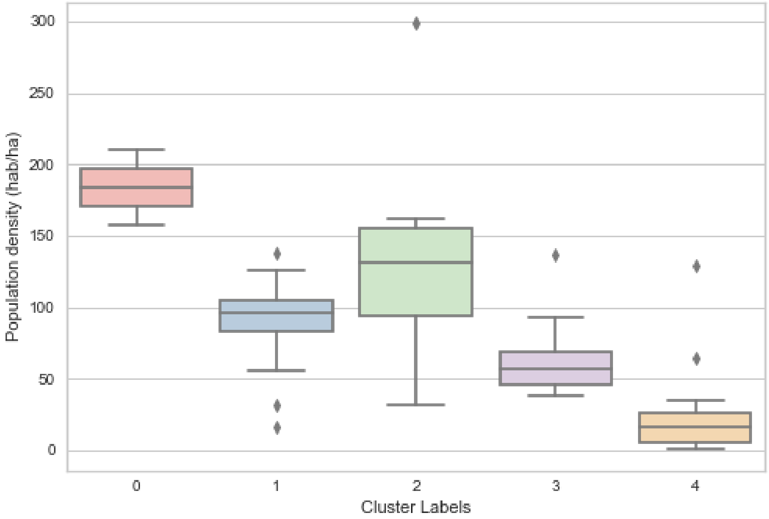
Table 3 – Number of neighborhoods in each cluster

4. ANALYSIS

It seems evident from the figure 7 that Cluster 0, which has only two neighborhoods, is characterized by high population density, which leads us to state that people in these neighborhoods live mostly in apartments. On the other hand, Cluster 4, which has 26 neighborhoods, has the characteristic of being the low densely populated. This means that this cluster has mostly suburban neighborhoods.

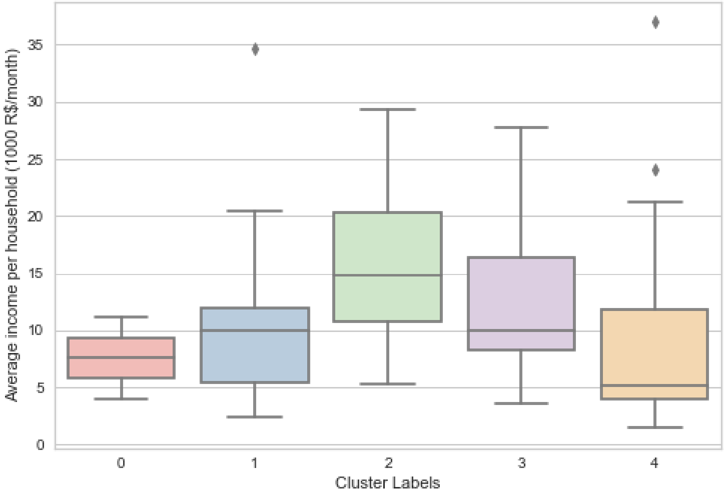
The other clusters contain neighborhoods with a population density mostly between 50 and 150 hab / ha. It cannot be said of clusters 1 and 2 if they are mostly composed of houses or apartment buildings, since the discrepancy within them is relatively large. Cluster 3 has a population density that indicates that most houses are houses.

Figure 7 – Clusters and population density



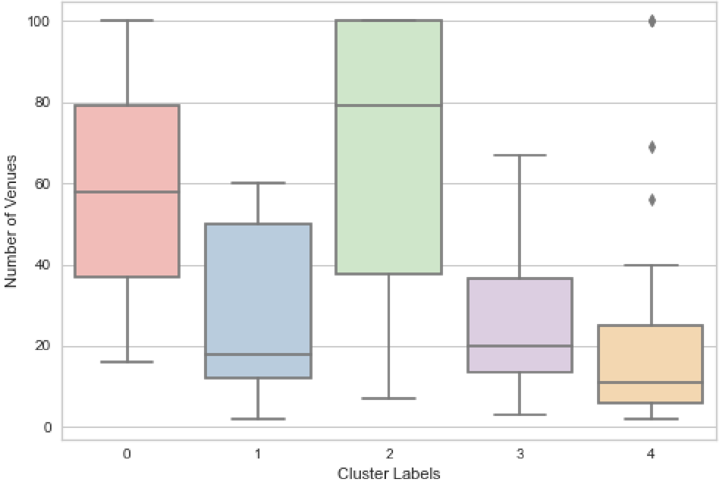
In terms of average income per housing unit, it is clear that there is not a big discrepancy between the clusters regarding the minimum income. However, the median and maximum values are quite different. Clusters 1, 3 and 4 show the greatest discrepancy, with the median values close to the IQR limits. Figure 8 also shows that income inequality, a strong characteristic of Brazil, is present in all clusters. That is, it is not possible to distinguish the clusters as high, medium or low income.

Figure 8 – Clusters and average income per household

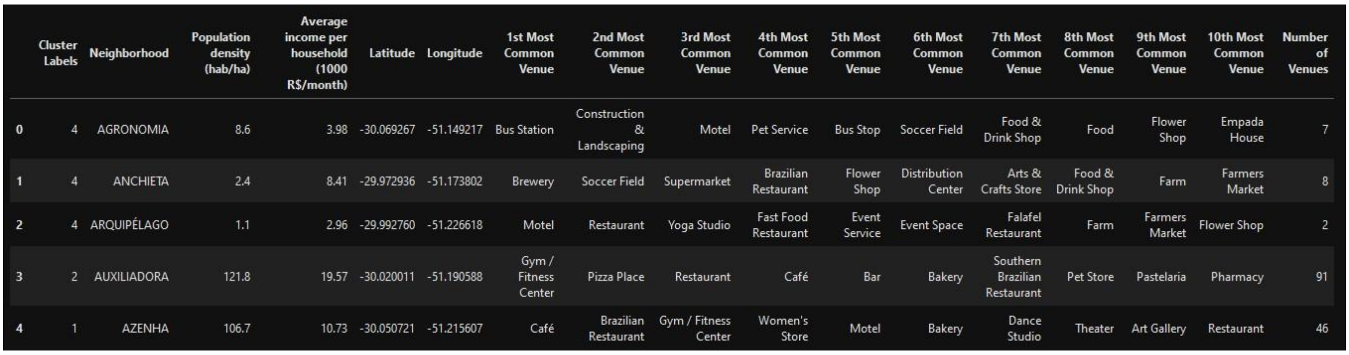


From the number of venues of each cluster, 75% of the neighborhoods contained in cluster 4 have less than 25 venues. In cluster 2, 75% of the neighborhoods have more than 35 venues.

Figure 9 – Clusters and number of venues



After a more peripheral analysis of the clusters, table 3 shows the main categories of venues contained in each cluster (from first to tenth).

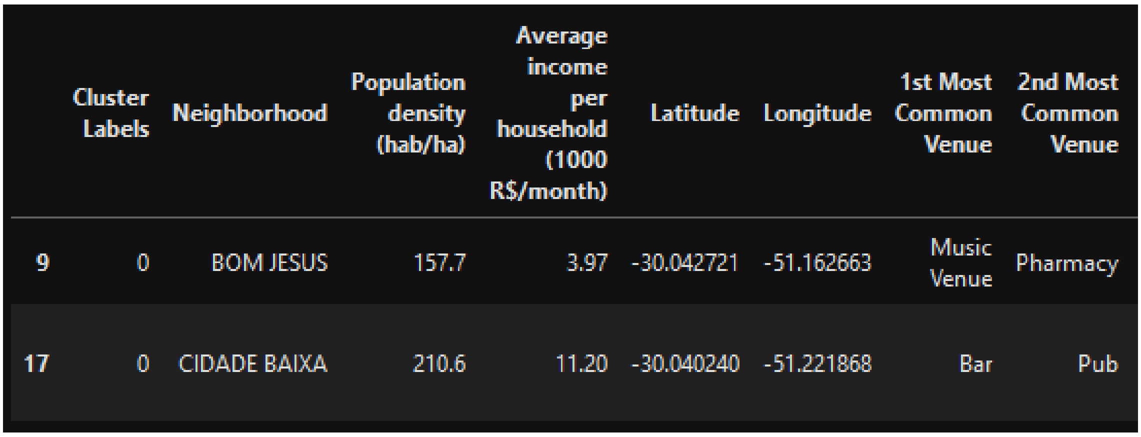


	Cluster Labels	Neighborhood	Population density (hab/ha)	Average income per household (1000 R$/month)	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Number of Venues
0	4	AGRONOMIA	8.6	3.98	-30.069267	-51.149217	Bus Station	Construction & Landscaping	Motel	Pet Service	Bus Stop	Soccer Field	Food & Drink Shop	Food	Flower Shop	Empada House	7
1	4	ANCHIETA	2.4	8.41	-29.972936	-51.173802	Brewery	Soccer Field	Supermarket	Brazilian Restaurant	Flower Shop	Distribution Center	Arts & Crafts Store	Food & Drink Shop	Farm	Farmers Market	8
2	4	ARQUIPÉLAGO	1.1	2.96	-29.992760	-51.226618	Motel	Restaurant	Yoga Studio	Fast Food Restaurant	Event Service	Event Space	Falafel Restaurant	Farm	Farmers Market	Flower Shop	2
3	2	AUXILIADORA	121.8	19.57	-30.020011	-51.190588	Gym / Fitness Center	Pizza Place	Restaurant	Café	Bar	Bakery	Southern Brazilian Restaurant	Pet Store	Pastelaria	Pharmacy	91
4	1	AZENHA	106.7	10.73	-30.050721	-51.215607	Café	Brazilian Restaurant	Gym / Fitness Center	Women's Store	Motel	Bakery	Dance Studio	Theater	Art Gallery	Restaurant	46

Table 3 – Venues data of each neighborhood

4.1 Cluster 0

Cluster 1 consists of 23 neighborhoods. Figures 10 and 11 show, respectively, the first most common venues and the second most common venues in the neighborhoods that make up the cluster.



	Cluster Labels	Neighborhood	Population density (hab/ha)	Average income per household (1000 R$/month)	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue
9	0	BOM JESUS	157.7	3.97	-30.042721	-51.162663	Music Venue	Pharmacy
17	0	CIDADE BAIXA	210.6	11.20	-30.040240	-51.221868	Bar	Pub

Table 4 – Cluster 0 neighborhoods

4.2 Cluster 1

As we can see, the cluster is made up of neighborhoods that contain squares, bus stops, soccer fields, gyms and restaurants as the most frequent venues. However, it is not possible to give a unique characteristic to the cluster, since for 23 neighborhoods there are 14 different first most common venues.

Figure 10 – 1st Most Common Venue in Cluster 1

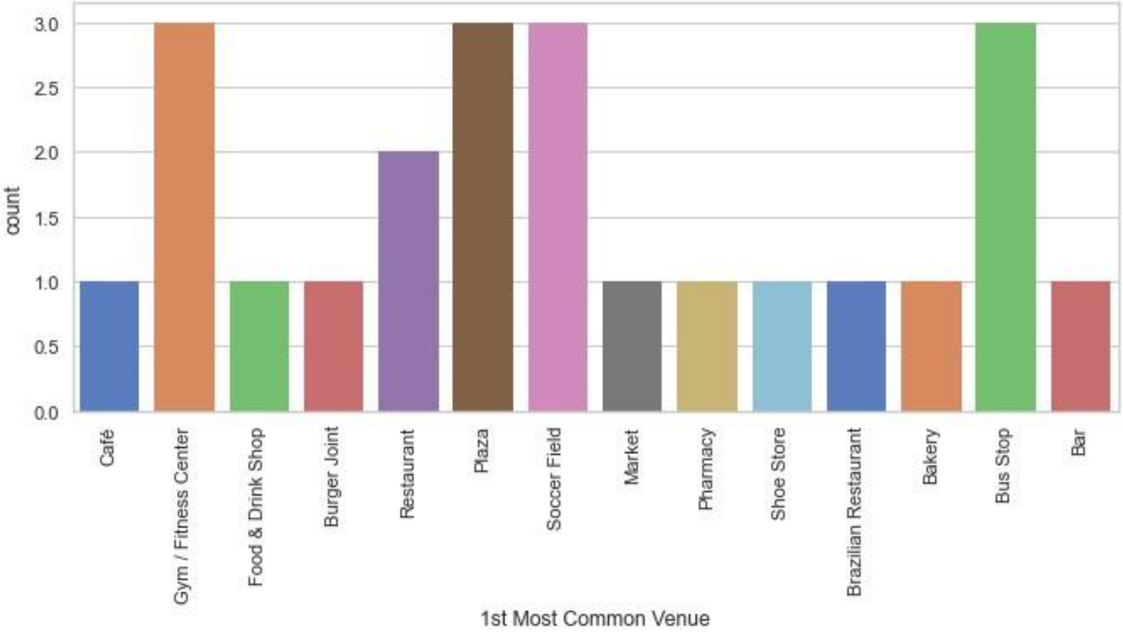
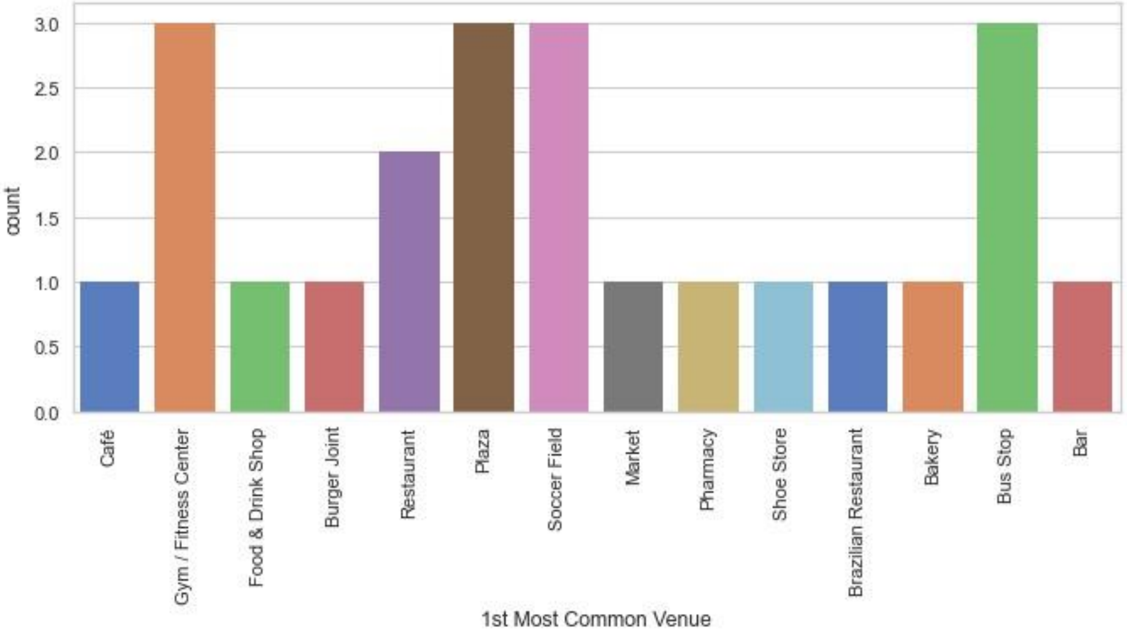


Figure 11 – 2nd Most Common Venue in Cluster 1



4.3 Cluster 2

Cluster 2 has 10 neighborhoods. Gyms, cafes, bakeries and pizza places are the most common types of venues in the neighborhoods. Thus, it is possible to say that it is a residential but urban neighborhood, with no apparent nightlife, but with facilities such as bakeries and cafes.

Figure 12 – 1st Most Common Venue in Cluster 2

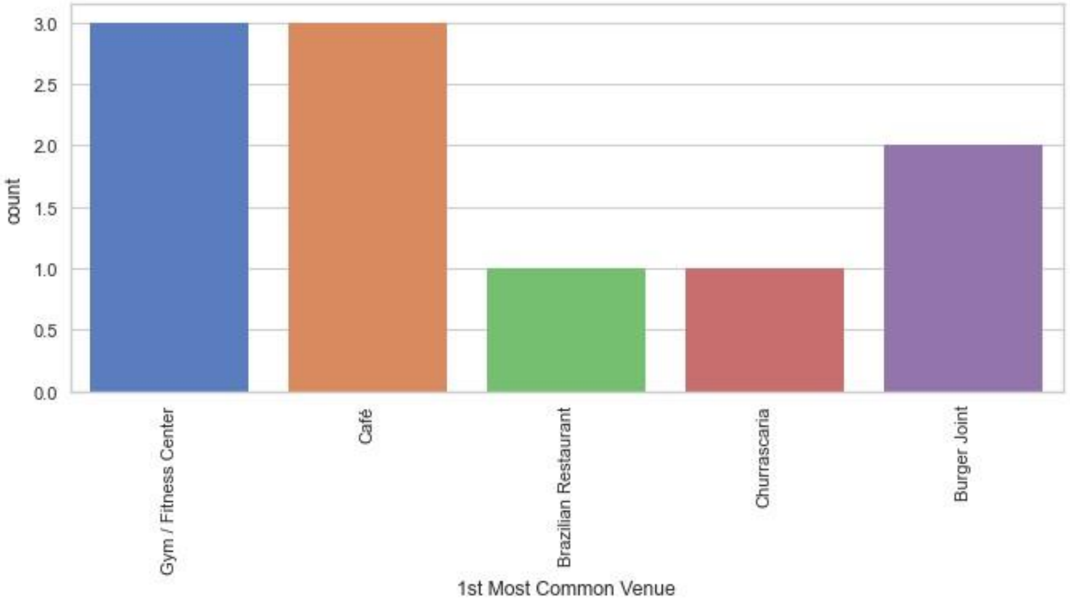
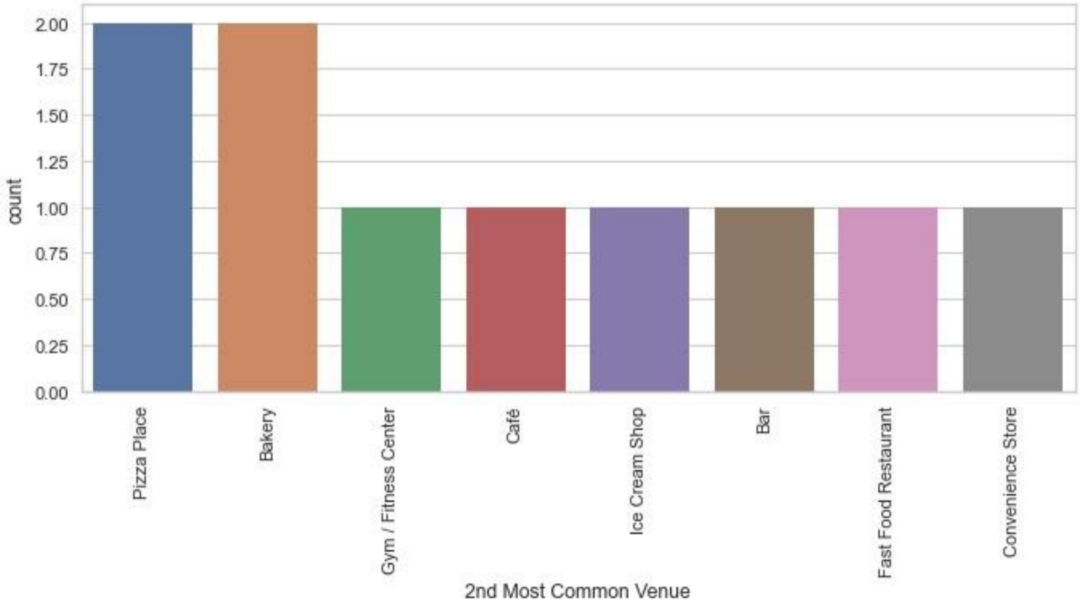


Figure 13 – 2nd Most Common Venue in Cluster 2



4.4 Cluster 3

Cluster 3 has 18 neighborhoods. The first most common venues are squares, pizza places, Brazilian restaurants, soccer fields and gyms. On this side, we can infer that the cluster contains residential neighborhoods, houses (due to the low population density). On the other hand, it is interesting to note that the most common types of second venues are very varied.

Figure 14 – 1st Most Common Venue in Cluster 3

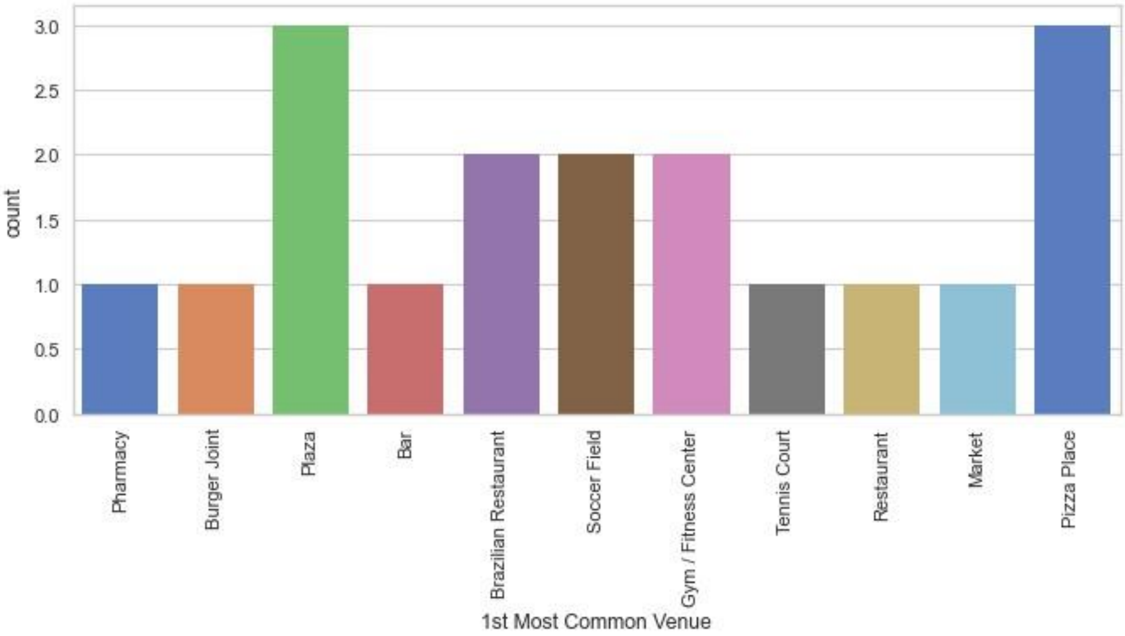
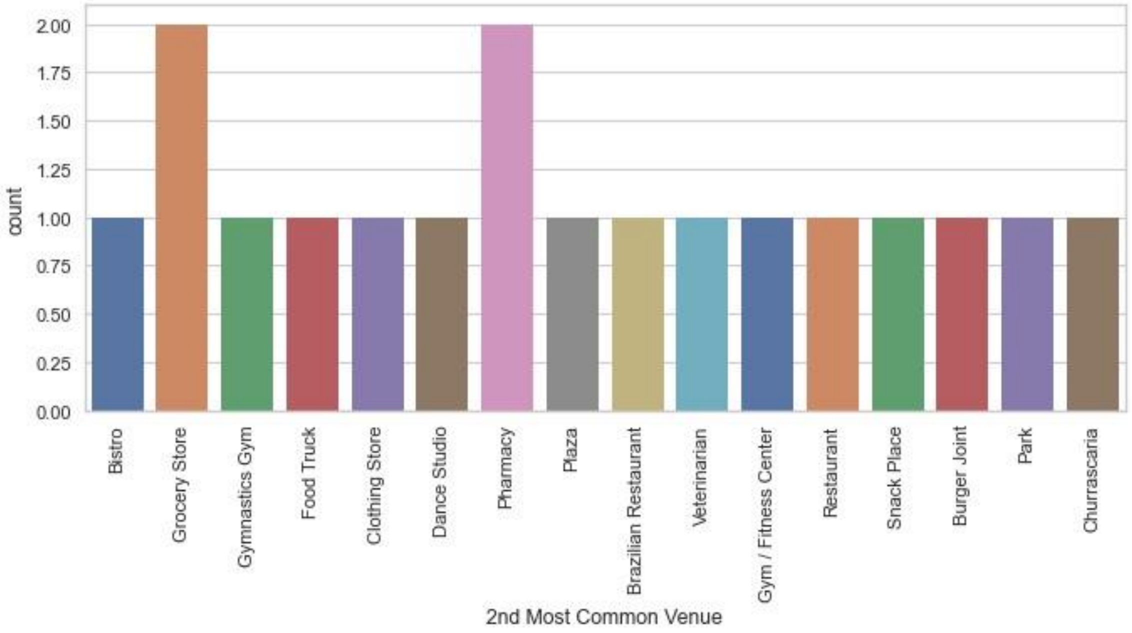


Figure 15 – 2nd Most Common Venue in Cluster 3



4.5 Cluster 4

Cluster 3 has 26 neighborhoods. There are 18 different types of the most common first venues. The second most common venues are also quite diverse. Thus, based on the venues, it is not possible to give the cluster a characteristic that identifies it.

Figure 16 – 1st Most Common Venue in Cluster 4

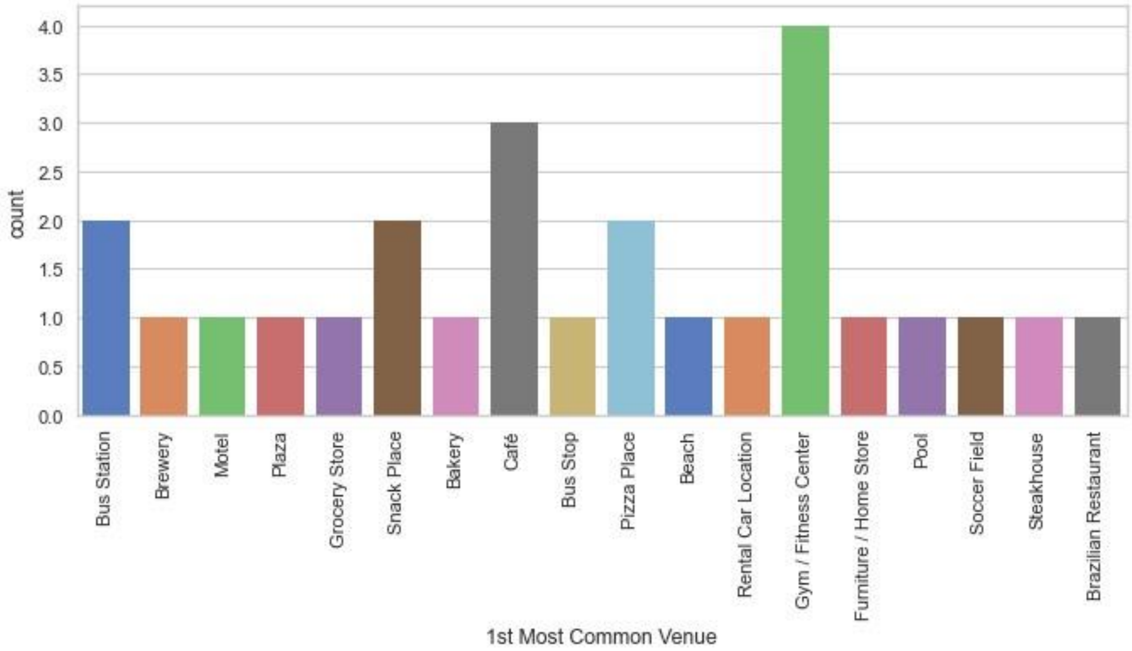
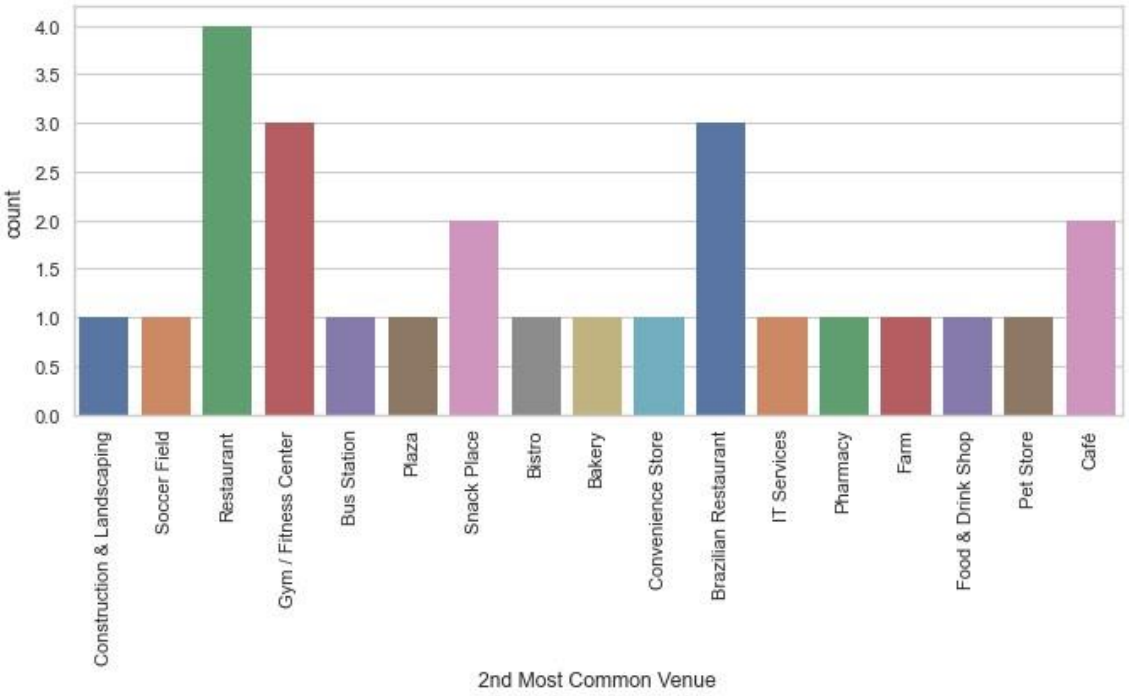


Figure 17 – 2nd Most Common Venue in Cluster 4



5. CONCLUSION

It was possible to characterize the clusters based on the population density factor. In the other hand, the average income per housing information does not characterize them, since Brazil is one of the most unequal countries in the world.

The data collected from Foursquare about the venues did not contribute to characterize the neighborhoods, and perhaps even worsened the clustering method used. This happens mainly because foursquare is in disuse in Brazil, especially among young people.

There were also problems with the lack of data on important neighborhoods in Porto Alegre, such as the Bomfim neighborhood.

Thus, due to the data used, especially those collected from Foursquare, the clustering done in this work would not help someone choose a neighborhood in Porto Alegre to live. Crime rate and property rental price would be more useful information for this analysis. However, this data is not readily available, especially residential rental value data.