Exploring neighborhoods in Curitiba (Brazil)

José H. K. Larcher

15/01/2020

1 Introduction

1.1 Background

Curitiba is a Brazilian city, capital of the state of Paraná. The city of Curitiba is divided into a total of 75 neighborhoods, grouped into ten administrative regions. The regional are species of subprefectures, treated here as boroughs, whose headquarters are represented by the units of the so-called "Rua da Cidadania" (citizenship street), and have the purpose of decentralizing public agencies and the provision of social, structural and leisure services within the city. According to the Brazilian Institute of Geography and Statistics, the city had, in 2012, 108,474 local units, 103,211 companies and active commercial establishments and 780,390 workers, of which 1,084,369 were employed and 931,971 employees.

1.2 Problem

This project aims to serve people looking for a place to open a business by answering where that business should be opened.

1.3 Interest

Entrepreneurs who are looking for places to open their businesses and wondering in which neighborhoods their business will be most appropriate and prosperous.

2 The Data

2.1 Data sources

The data used in the project will be socioeconomic data obtained through the Curitiba neighborhood page on Wikipedia and data obtained through the Foursquare API.

2.2 Description of the datasets

The dataset obtained from the Wikipedia page has data for each neighborhood and to which borough it belongs. The fields obtained will be:

- borough
- neighborhood
- men: number of men in the neighborhood
- women: number of women in the neighborhood
- total: number of people in the neighborhood
- households: number of private households.
- income_per_head: average monthly income per household head [BRL].
- area: [km^2]

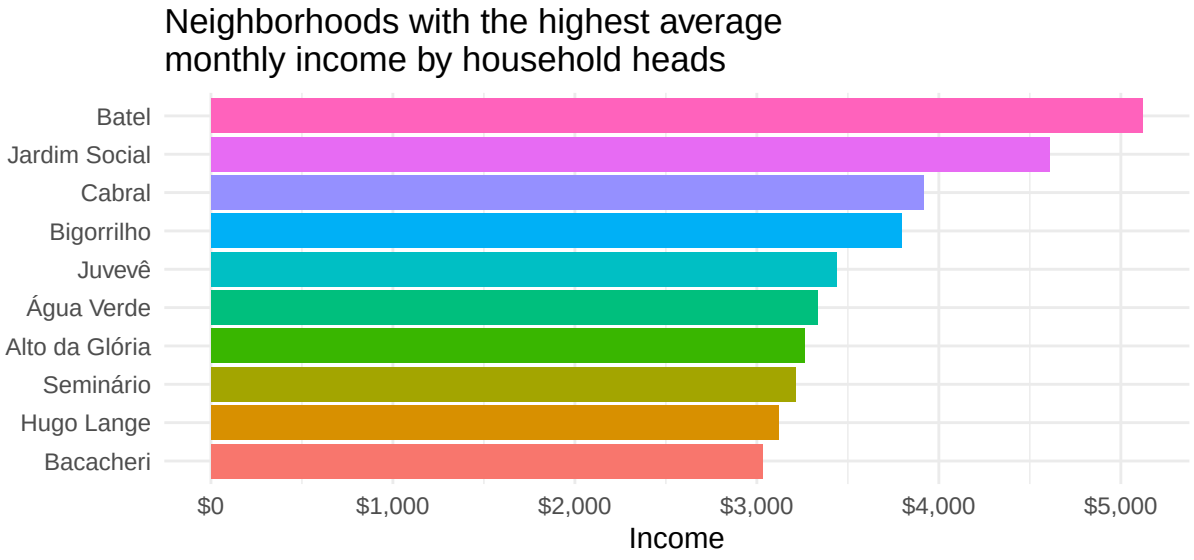
2.3 Data cleaning and Feature Selection

Neighborhood socioeconomic data was scraped from the Wikipedia page using the pandas library's own HTML page reading function. The required columns were found and regular expressions were used to clean some fields. Some fields were normalized using z-score. The neighborhood coordinates were obtained using the geopy library. For the venues and business data of Curitiba, the Foursquare API was used.

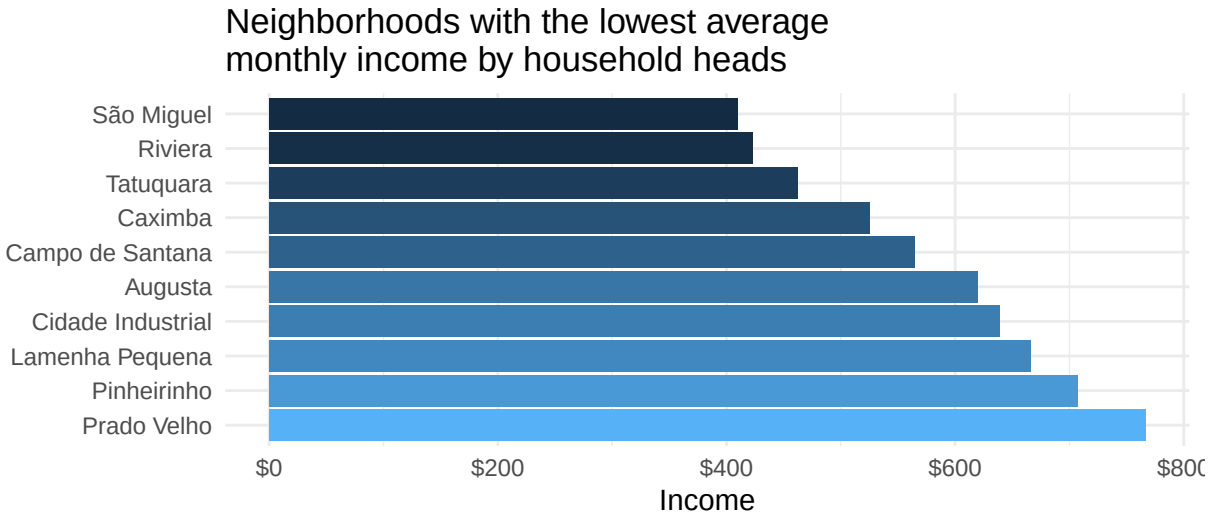
With these data two main dataframes were obtained. The first consisted of all neighborhoods, with their regional, their socioeconomic data, their coordinates, and their top ten business categories. The second consisted of neighborhoods related to the proportion of business types in it.

3 Exploratory Data Analysis (Methodology)

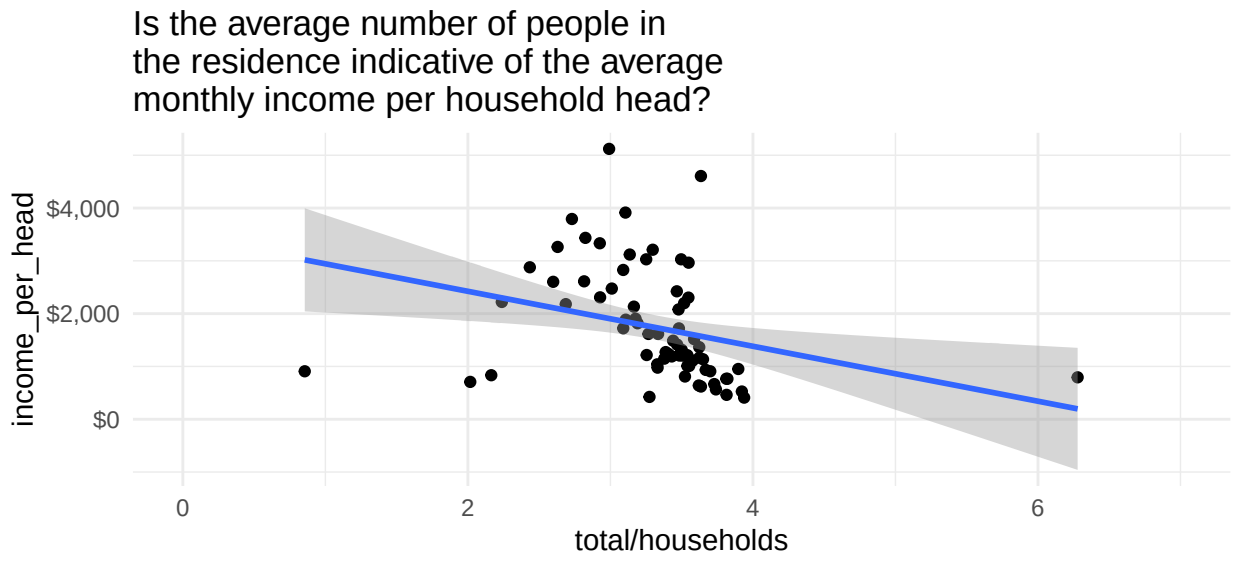
First were explored the neighborhoods with the highest monthly income per head of the house.



And then the ones with the lowest monthly income per head of the house.

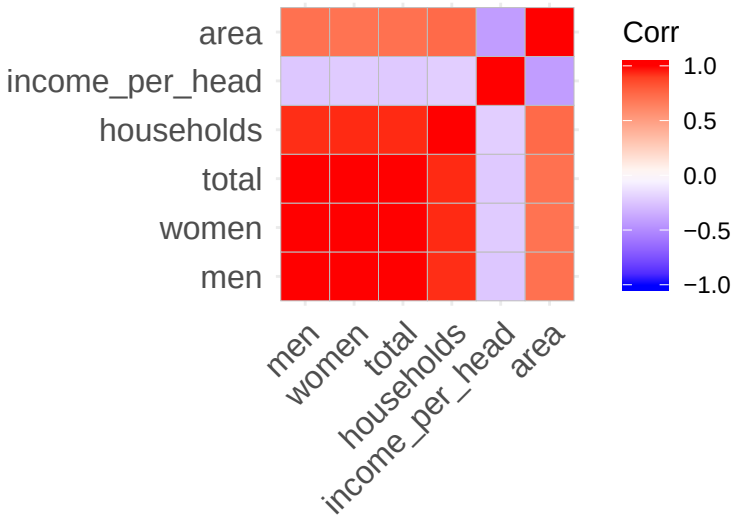


Next, was investigated the average number of people in the household versus the average monthly income by household heads.



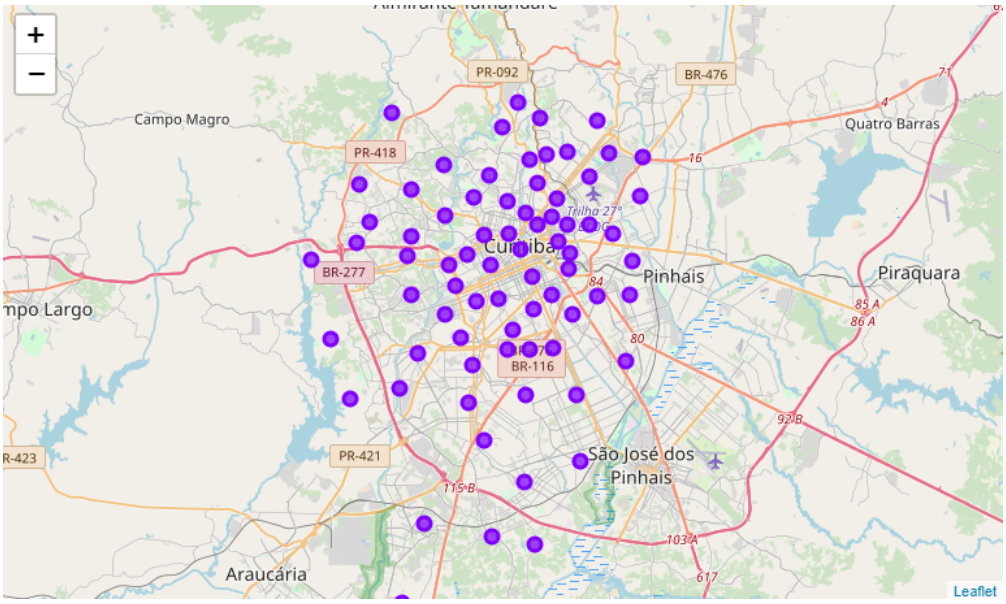
It seems to have a correlation, but not very high, as the R^2 is of just 0.05245.

Looking at the correlogram, we see that the number of households has a high correlation with the number of people (men and women alike).



3.1 Map

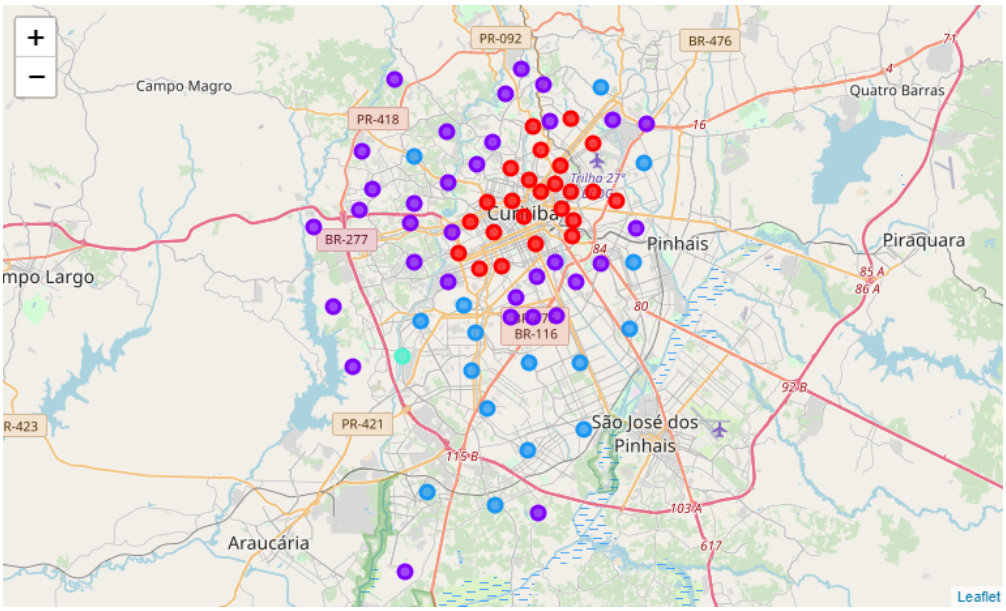
Then we can check out the neighborhoods map.



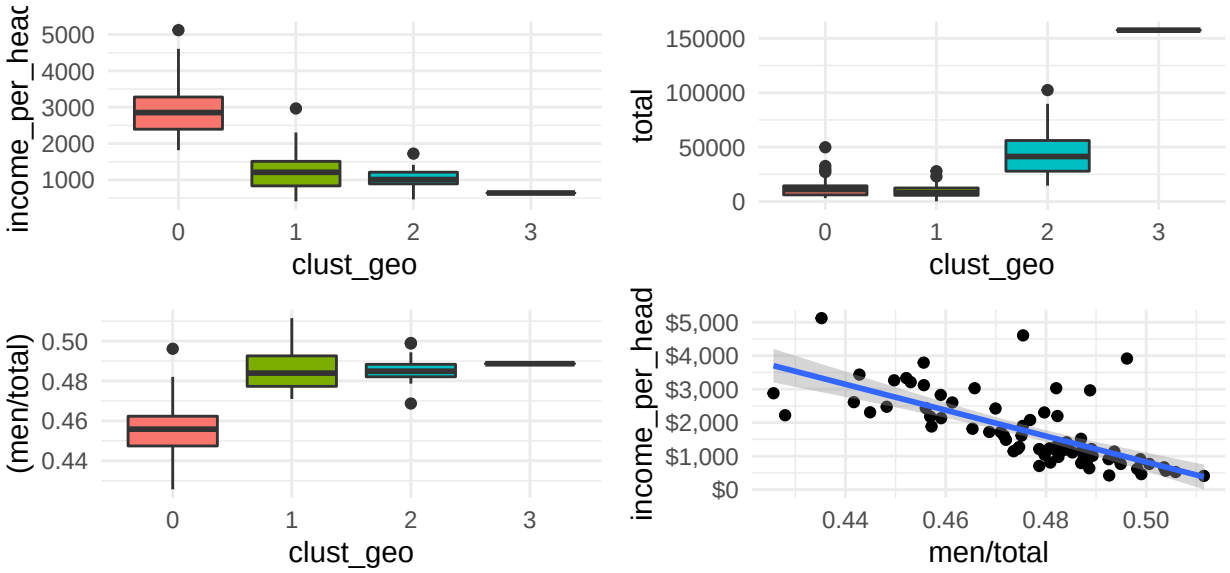
4 Model / Results

4.1 Clustering based on socioeconomic data

The model use for clustering the neighborhoods was a KMeans model. For the firt model, the socioeconomic data was used to group the neighborhoods in 3 different clusters, as seen below.



Analising the cluster by some variables, we can take a look in what influenced the clustering.

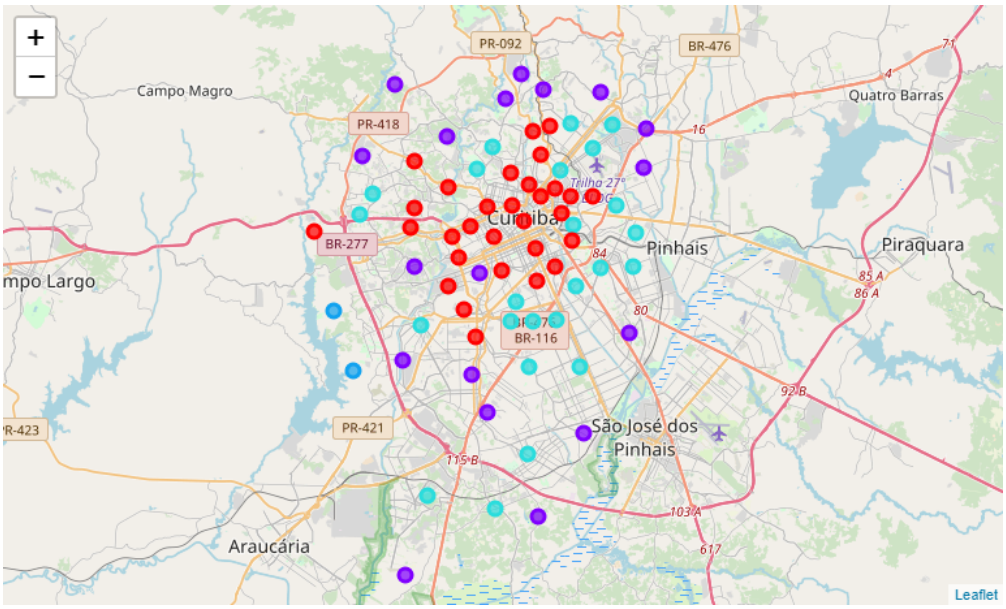


We can observe that the clusters have a clear division on the income per household head. Also, the number of men over the total number of people presents another clear division on the clusters, as the total number of people does.

A interesting thing is to notice that neighborhoods with less men/total have a greater income.

4.2 Clustering based on venue data

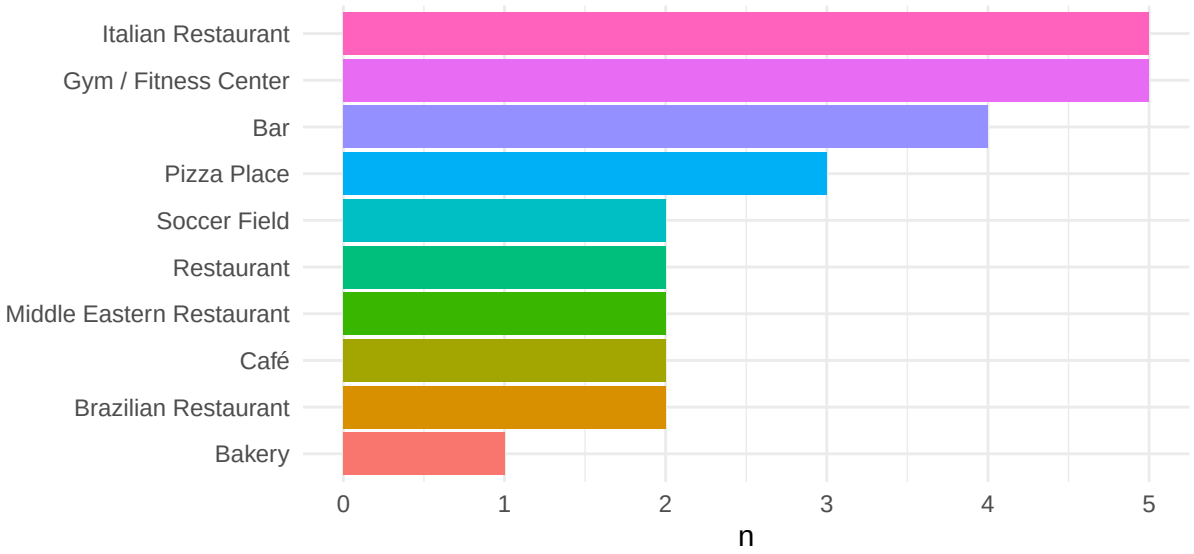
Another KMeans clustering was aplied based on the business categories on the neighborhoods. The result can be seen in the next picture.



5 Discussion

The central cluster of the venue data has a great superposition with the central cluster in the socioeconomic data. As the central cluster also represents one with the major incomes in the neighborhoods, this means it is a good place to open a business.

Grouping the number of first places in the central cluster for venues, the count becomes as follows.



We can see that this places might be good for Italian restaurants and gyms, but, on the other side, these kinds of venues might be saturated in the region. Thinking like this, invest in a type of venue more down the count line might be a good solution.

6 Conclusion

Taking the intersection of the central clusters of both model as good neighborhoods to open a business, these neighborhoods would be:

neighborhood
Batel
Jardim Social
Cabral
Bigorrilho
Juvevê
Água Verde
Alto da Glória
Seminário
Hugo Lange
Bacacheri

7 References

https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Curitiba

https://developer.foursquare.com/docs/api/

https://pypi.org/project/geopy/