Analysis of São Paulo boroughs based on restaurants prices

Rafael Mota Gregorut

November 27th, 2019

1. Introduction

1.1. Background

São Paulo is the biggest city in Brazil, with over 12 million habitants with different backgrounds spread across an area of more than 1500 squared kilometres. The city is organised in 32 administrative boroughs placed in 5 regions (centre, north, south, east and west). In such a diverse environment, prices of services can dramatically change from one borough to another, food services included.

1.2. Business Problem

In this project, we aim to cluster the boroughs of São Paulo, Brazil, based on the price of their restaurants. Generally more expensive restaurants are located in business or richer areas. On the other hand, cheaper restaurants tend to be more concentrated on more poor areas.

1.3. Interest and target

This analysis can be used to understand the purchasing power of the people who live in each borough of the city. It can be useful to regular citizens to understand the costs of meals in the city, and also to restaurateurs who are willing to expand their business and need to be aware of the boroughs that will allow then to keep their average ticket value.

2. Data collection and preparation

2.1. Data source

We can retrieve the boroughs present in São Paulo by scraping this Wikipedia page. However, that page lacks the information of latitude and longitude coordinates for each borough. To overcome that issue, we use Nominatim, from Geopy, to retrieve the latitude and longitude of each borough.

The data related to the restaurants in each borough are obtained by calling the Foursquare API to get venue recommendations documented here. As parameters for this call, for each borough, we pass:

- Credentials
- Version
- Latitude and longitude coordinate
- Radius of 2km
- Section ‘food’ to restrict the results to restaurants

- Category ID of the restaurant category documented here
- A limit to restrict the number of venues returned in the response. This parameter is actually not that important for us since we will be interested in the total results field, which does not consider the limit parameter
- Price parameter with the price level from 1 to 4, where: 1 is the cheapest level and 4 is the most expensive level

2.2. Data cleaning

When scraping the Wikipedia page, due to the structure of the page, some rows and columns were retrieved some empty values. In these cases, the rows and columns with empty values were dropped without damaging the data of the boroughs. In addition, the columns that contained the boroughs' data are splitted in the page, so we needed to concatenate them to have all the boroughs' names in one single column.

There was also the need to treat missing values for latitude and longitude coordinates: we used Nominatim to generate them, but it happened that for some boroughs, Nominatim was not able to return a pair of coordinates. When that happened, we removed the row for the corresponding borough.

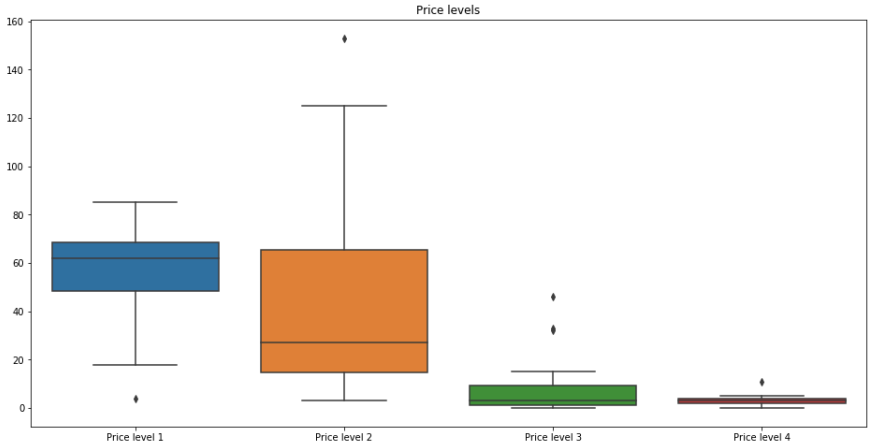
After this clean up, we ended up with 30 boroughs of São Paulo with latitude and longitude coordinates, from an original list of 32 boroughs.

From Foursquare API, we retrieved the data regarding the price for restaurants in each borough. More specifically, we were interested in the total number of restaurants for a given price level in a given borough. The call for the APIs were done for all boroughs for all price levels and the returned json was parsed to obtain the value of interest.

In the end, the count of restaurant for each price level was inserted in the boroughs' table: one column per price level, and one row per borough.

2.3. Exploratory data analysis and Feature selection

If we plot the boxplot of all price levels, we can notice that the price level 4 has fewer data when compared to the other price levels. Checking it with more detail, it seems that the price level 4 is redundant with price level 3.

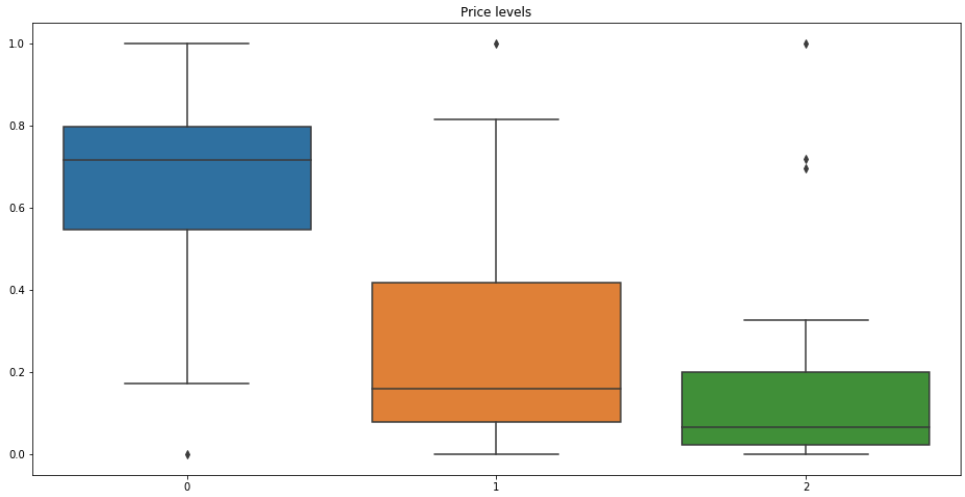


To confirm that theory, during our analysis, we experimented two cases: one in which price level 4 was a feature, and another one in which price level 4 was dropped. The results for both were equal to all boroughs except one. Therefore, we dropped the price level 4 column for our final analysis, and we considered as features the count of restaurants for price levels 1, 2 and 3 only.

3. Modelling methodology

3.1. Normalisation

Before creating the clusters for the boroughs, we normalised the features for the data retrieved from the Foursquare APIs. In order to do that, we used the MinMax scaler making our selected features range from 0 to 1.



3.2. Clustering

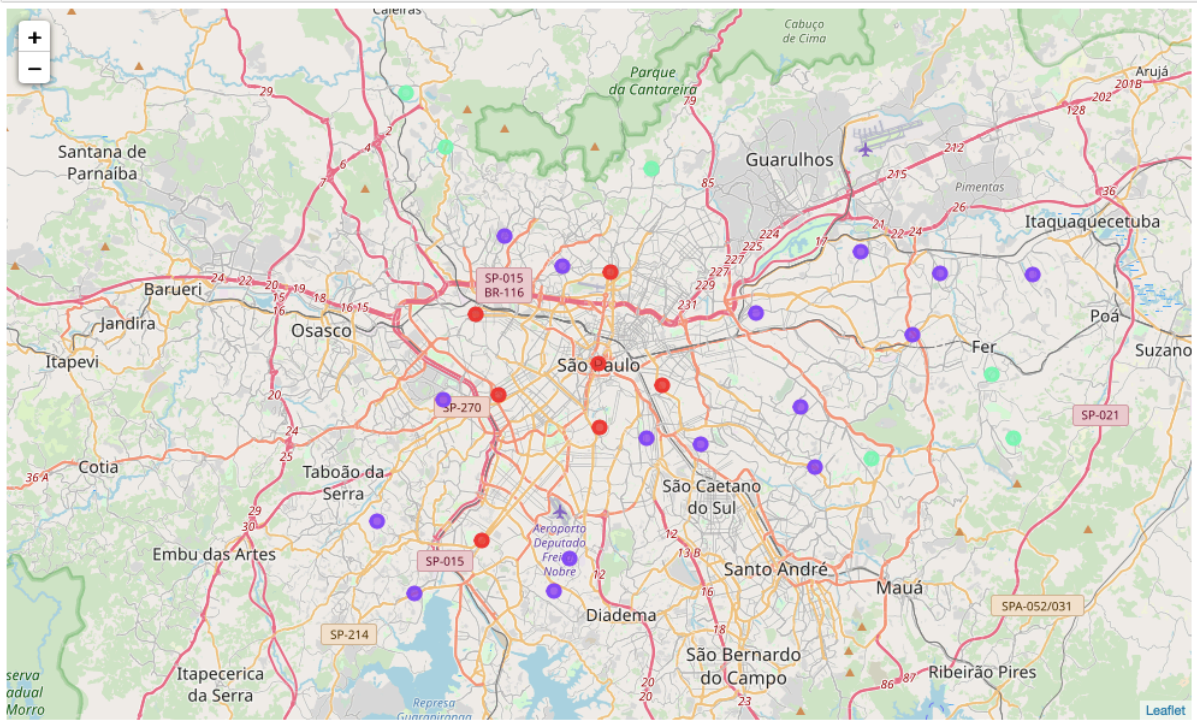
For this project we used k-means as the algorithm to divide the boroughs into clusters. We consider ‘k’ to be equal to 3, meaning that we will have three clusters at the conclusion of the algorithm. Note that value for ‘k’ was chosen for better interpretation of the returned clusters.

To optimise the algorithm, we used two parameters provided by implementation of k-means in Scikit-learn:

- ‘init’ for better initialisation of the centroids of the algorithm
- ’n_init’ to define the number of times the algorithm would run with different centroids

4. Results and discussion

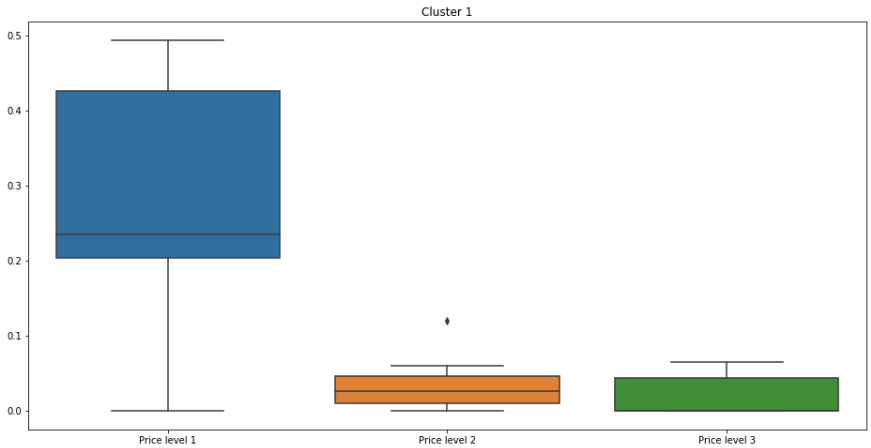
The 3 clusters obtained from k-means were plotted on São Paulo map as you can see below:



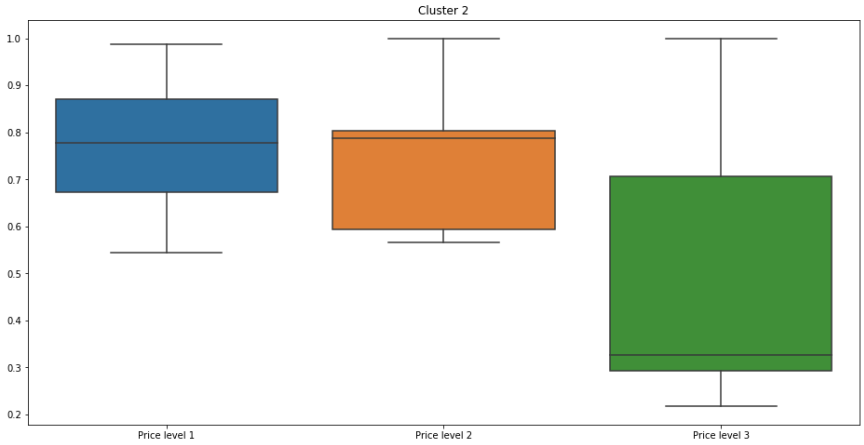
From the map, we clearly noticed that the cluster marked in red is concentrated closer to the city centre, another cluster marked in purple is spread across the city and another cluster marked in light-green is placed on the extremes of the north and east regions mainly.

We can plot the price levels of each cluster in order to look into them with more details:

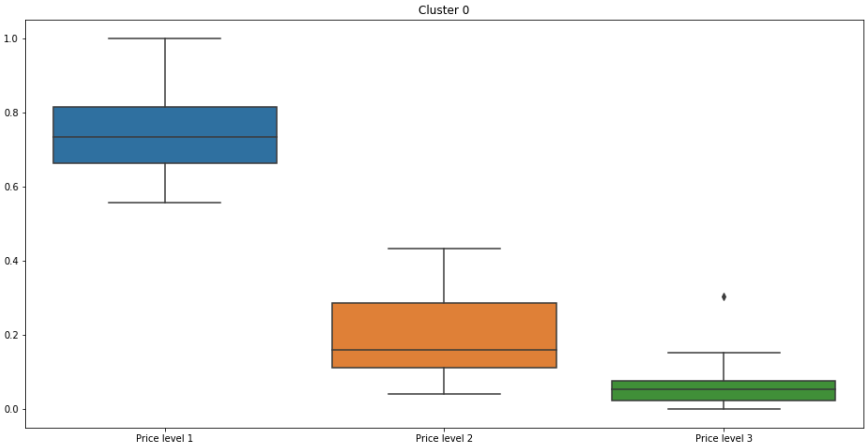
- The cluster labeled as 1, coloured in light-green, contains the boroughs with a big presence of restaurants in the cheapest price level. In general, these boroughs are located in the suburban areas of the city



- The cluster labeled as 2, coloured in red, contains the boroughs with more restaurants in the 3rd and 2nd price level, therefore with higher prices. These boroughs are located closer to city centre and in business areas



- The cluster labeled as 0, coloured in purple, contains boroughs with price levels mainly in the cheapest and medium cost. These boroughs are spread over the city, but are not business areas and neither are located in suburban areas



5. Conclusion

The returned clusters reflect what can be observed empirically in the city of São Paulo: city centre and business areas have more expensive restaurants, while the extremes generally have cheaper locations for meals. For a better clustering, as a next step, we could consider to get the data of each neighbourhood, and not only borough, to have a more granular view of the city. Besides, the price information for the venues could be retrieved from a platform that is more popular among the local population.