Analysis of São Paulo boroughs based on restaurants prices

Rafael Mota Gregorut
Nov 28th 2019

Introduction

- São Paulo is the biggest city in Brazil, organised in 32 boroughs. In such environment, prices of services can dramatically change from one borough to another

- In this project, we aim to cluster the boroughs of São Paulo based on the price of their restaurants

- This analysis can be used to understand the purchasing power of the people who live in each borough of the city

Data source

- We retrieved the list of boroughs by scraping this page
- We used Nominatim, from Geopy, to obtain the latitude and longitude of each borough
- The data related to the restaurants in each borough were obtained by calling the Foursquare API to get venue recommendations documented here
  - The call to the API limited the venues to restaurants in a radius of 2km from the defined latitude and longitude
  - The price for a venue is divided in price levels from 1 to 4: 1 being the cheapest and 4 being the most expensive

Data cleaning

- When scraping the page, the rows and columns with empty values were dropped without damaging the data of the boroughs

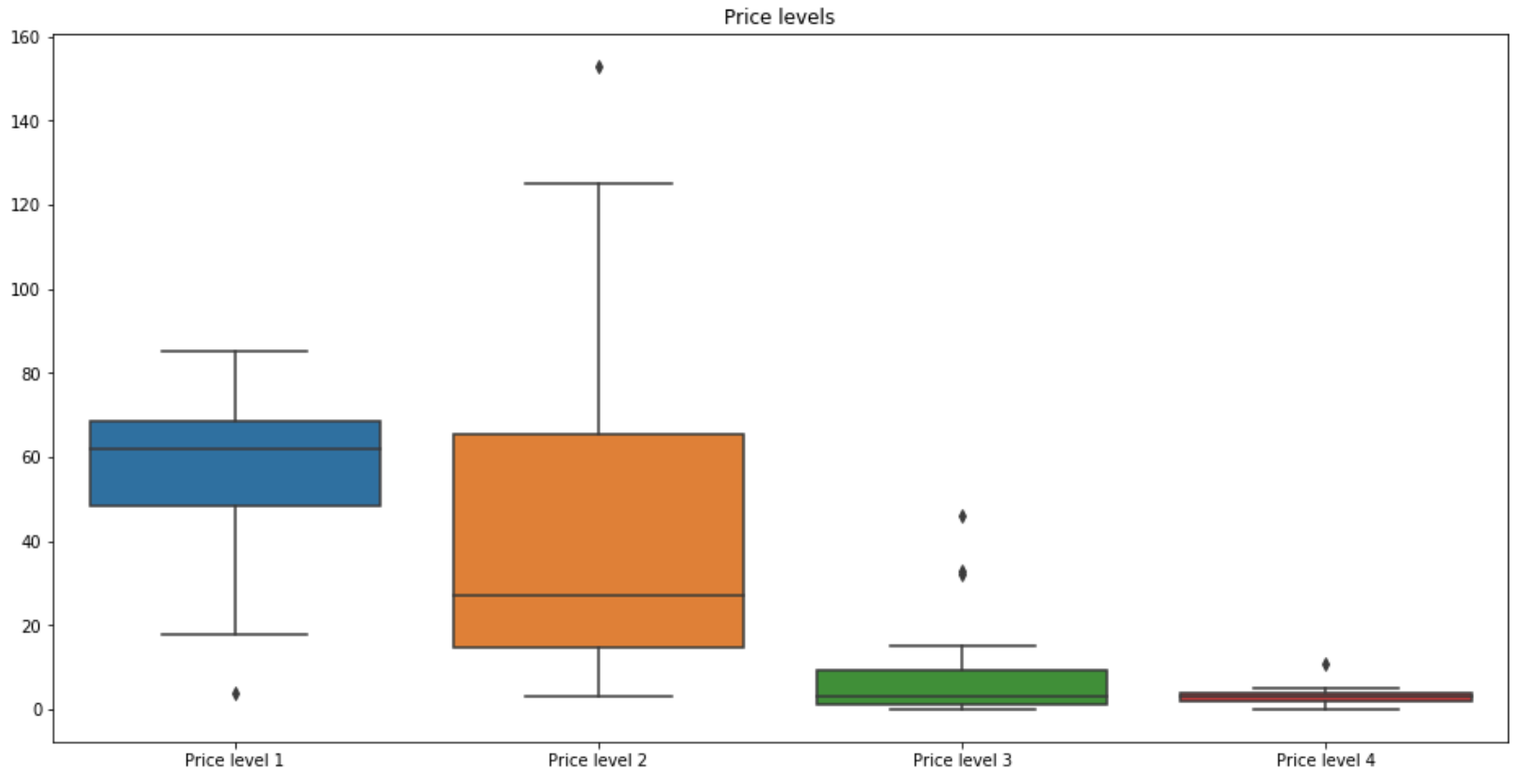
- It happened that for some boroughs, Nominatim was not able to return a pair of coordinates. When that happened, we removed the row for the corresponding borough

- 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, we retrieved the total number of restaurants for a given price level in a given borough

EDA and feature selection

- Plotting the boxplot of all price levels, we can notice that the price level 4 has fewer data



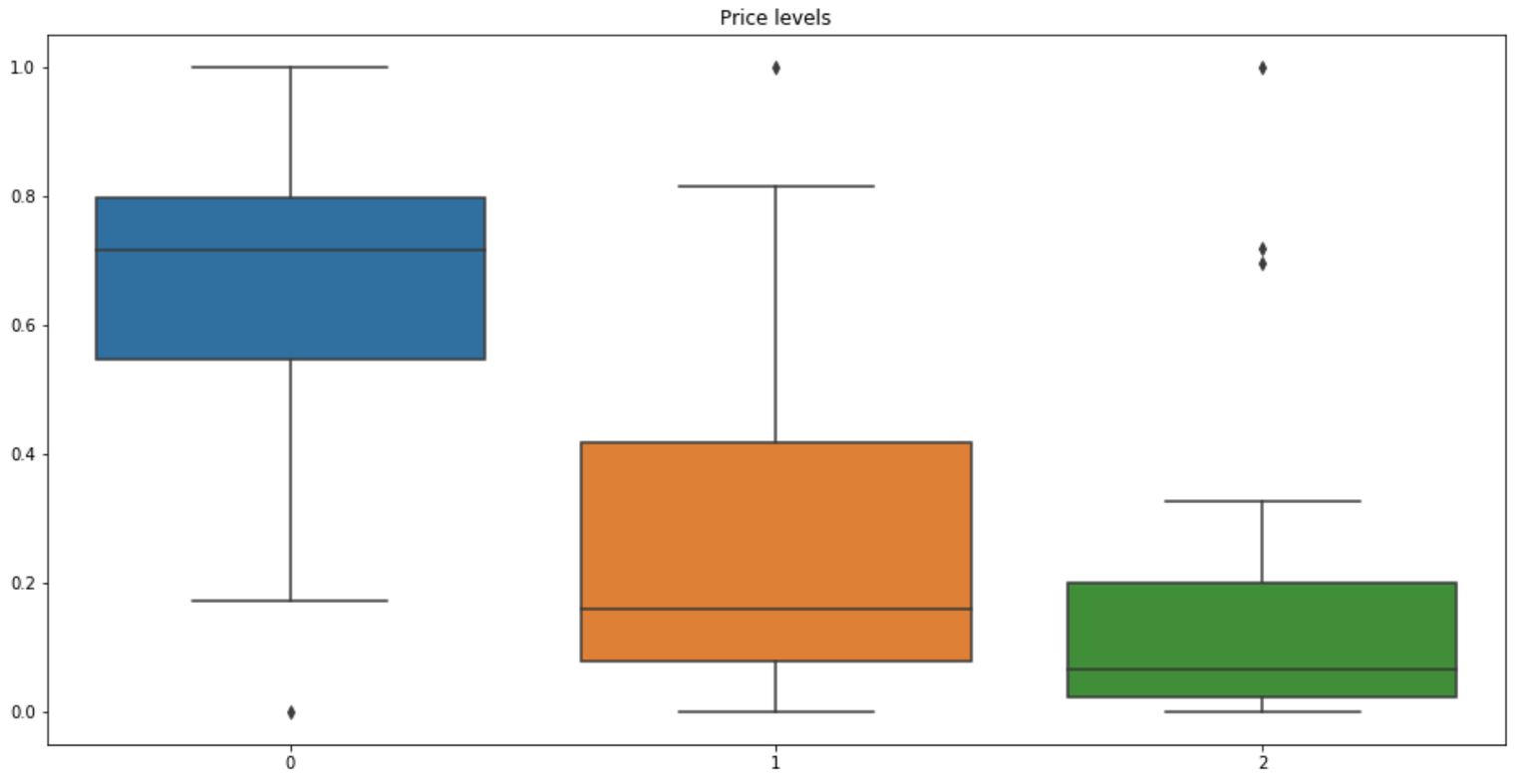
EDA and feature selection

- 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

Normalisation

- Before creating the clusters for the boroughs, we normalised the features, making them range from 0 to 1



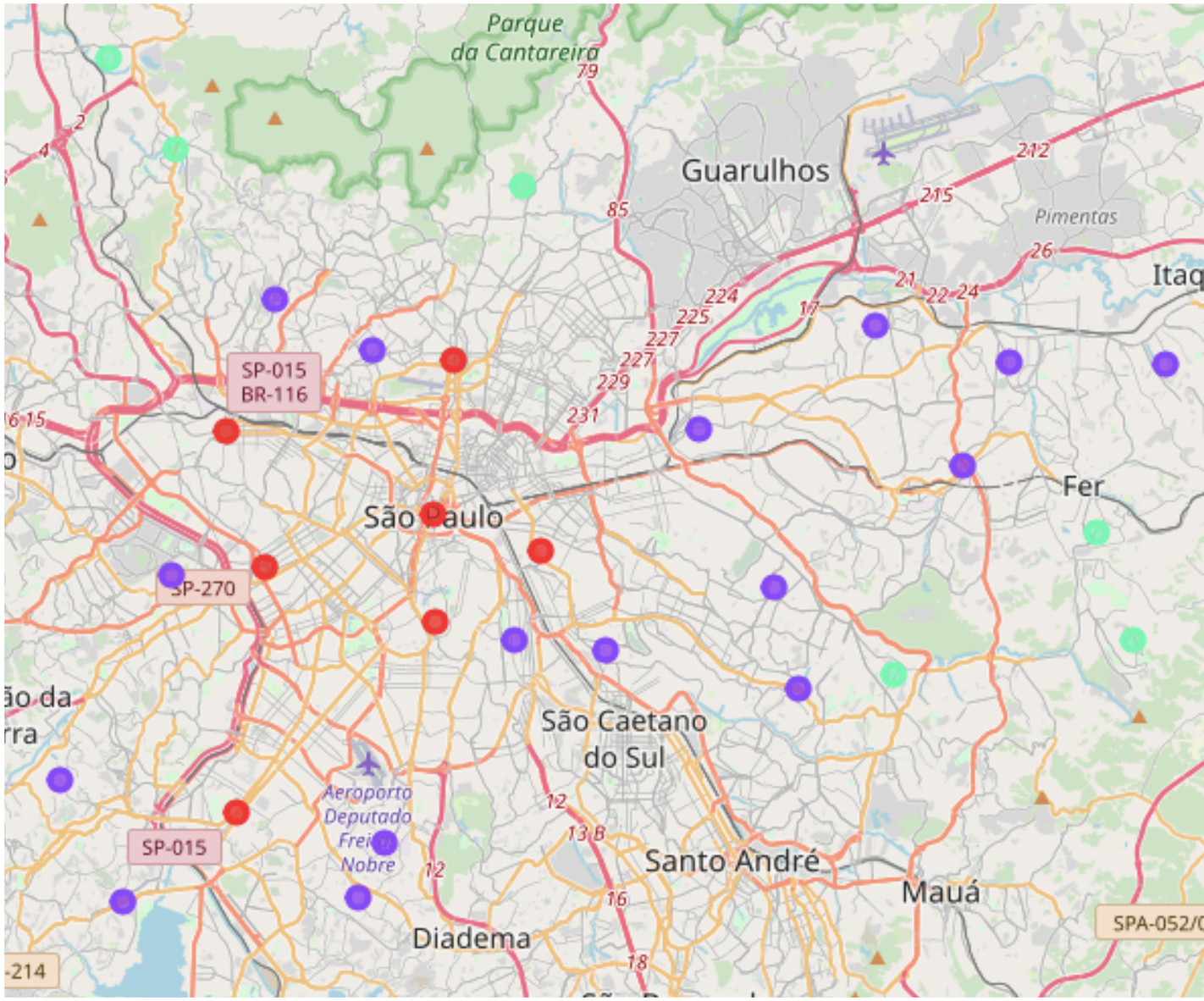
Clustering

- 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.

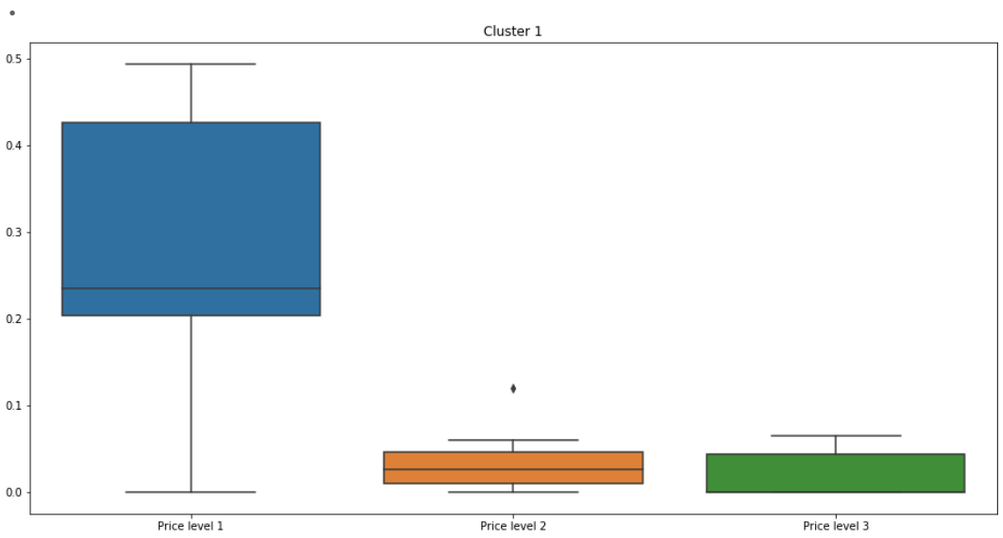
Results and discussion

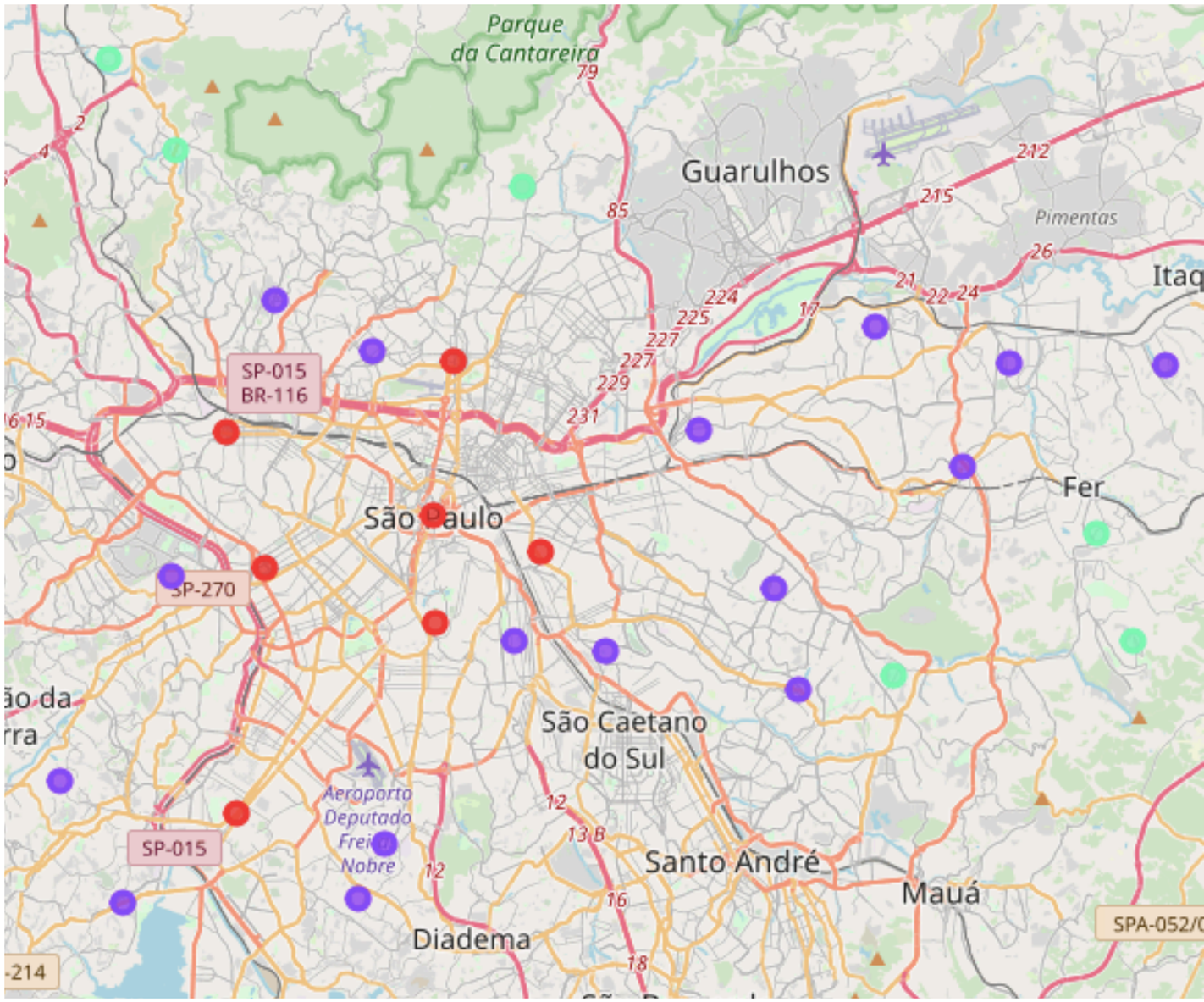
- From the map, we clearly noticed that the cluster marked in red is concentrated closer to the city centre, another cluster marked in purple is spread across the city and another cluster marked in light-green is placed on the extremes of the north and east regions mainly.



Results and discussion

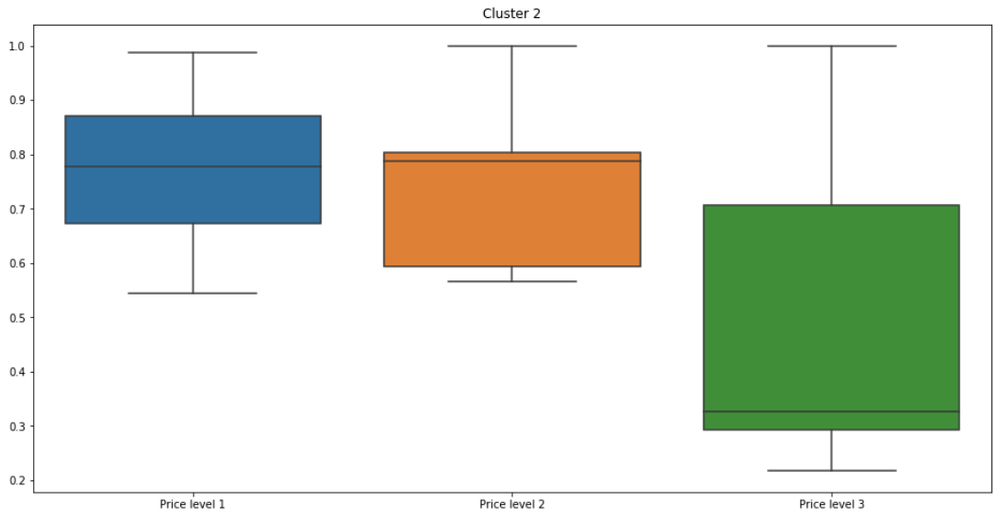
- The cluster coloured in light-green, contains the boroughs with a big presence of restaurants in the cheapest price level. In general, these boroughs are located in the suburban areas of the city

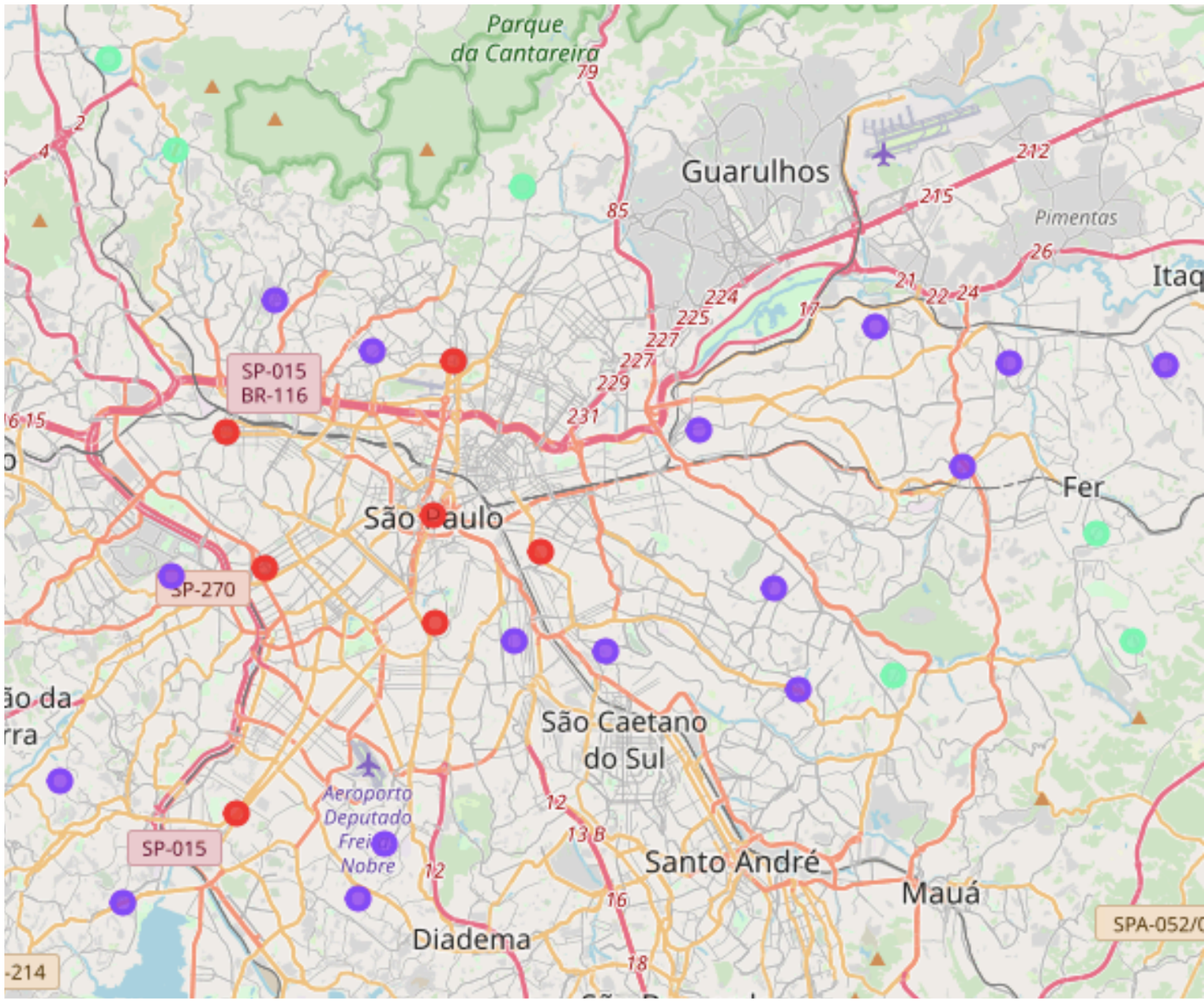




Results and discussion

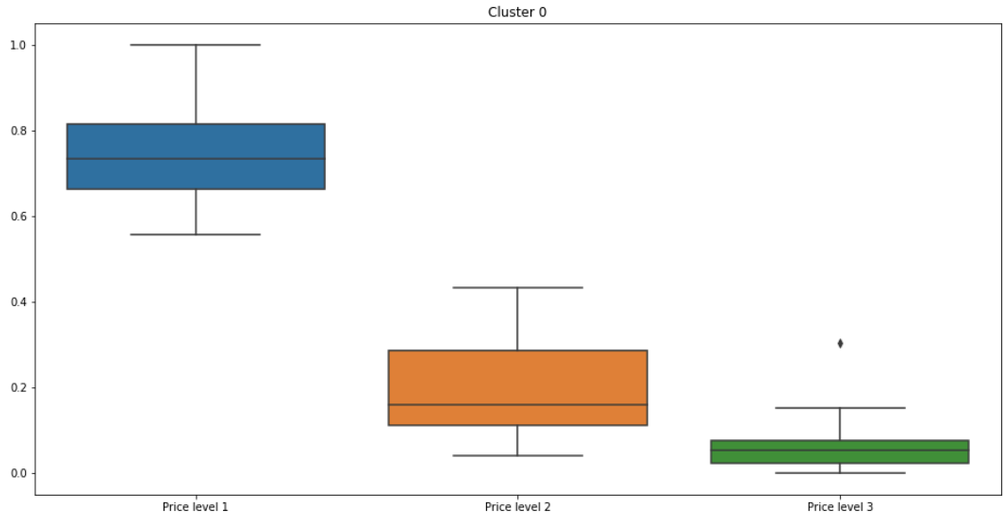
- The cluster coloured in red, contains the boroughs with more restaurants in the 3rd and 2nd price level, therefore with higher prices. These boroughs are located closer to city centre and in business areas

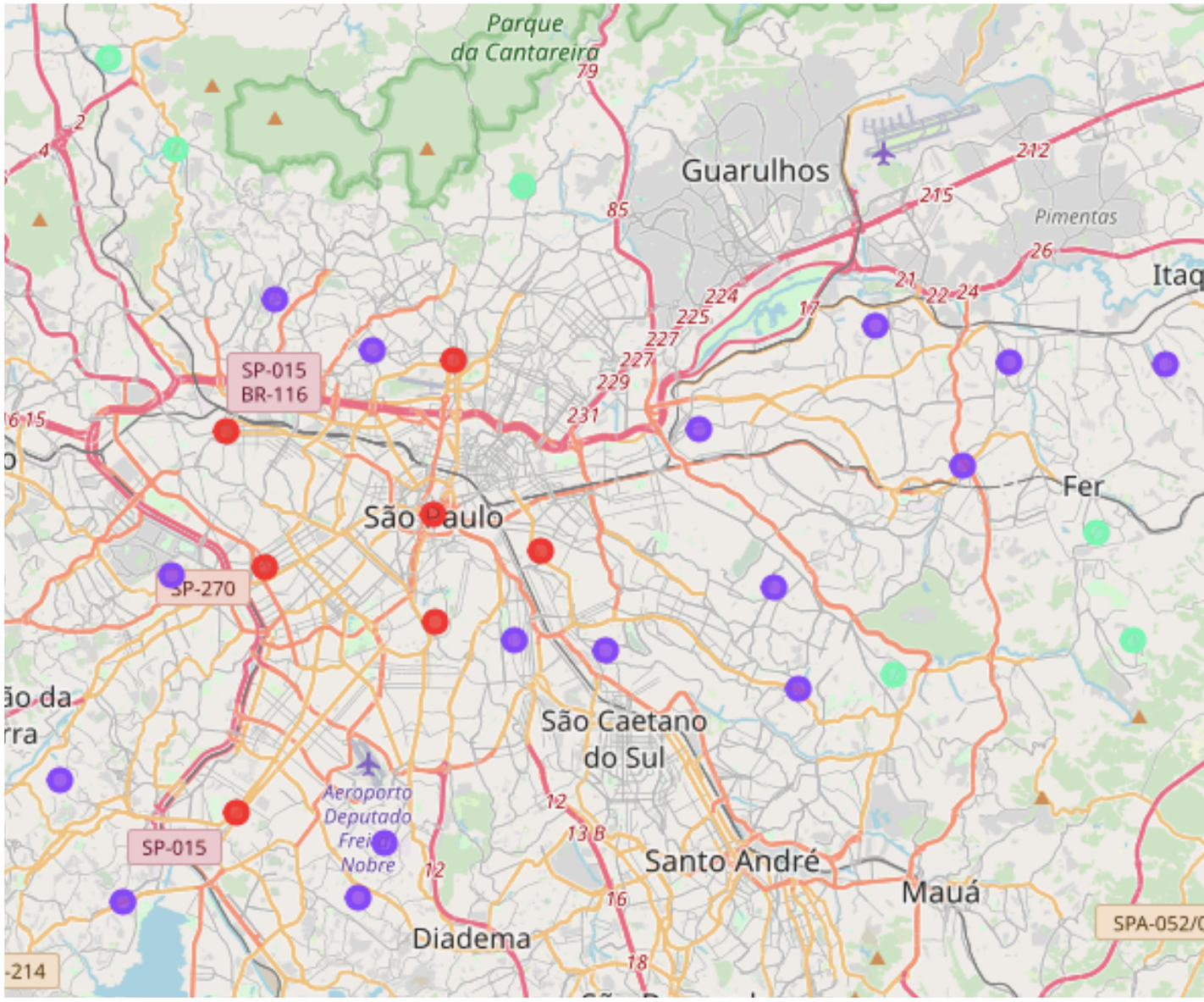




Results and discussion

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





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