Restaurant location problem

Igor Sant'Ana Gallo

1. Introduction

This report is part of the final assignment for the Applied Data Science Capstone project. In this report will be presented a business problem concerning the best location of a restaurant in the city of São Paulo – Brazil. In the next subsections, the context of the problem will be described, and the research question will be proposed. After the introduction, it will be presented the process to acquire and clean the data (Section 2), the research method to answer the research question (Section 3), the main results (Section 4), the discussion of the results (Section 5) and the conclusion (Section 6).

1.1 Background

One of the major natural sites of São Paulo – Brazil is the Ibirapuera park. The park has over 100 hectares and it is in Moema neighborhood, in the center-south of São Paulo. The park is one of the most famous parks in Latin America, attracting millions of visitors a year.

Due the large number of visitors the park receives every day, an investor would like to know if the surrounding of the park would be a good place to build his restaurant. To answer this question, the investor will open his restaurant if there is a region, next to the park, with not too many competitors. If such area is available, he also wants to know if it is better to open an Italian, a vegetarian or a Japanese restaurant, the types of food he already invests in.

1.2 Research question

This project aims to answer two questions, where is the best location to start a new restaurant next to Ibirapuera park to avoid a high concentration of competitors? The second question is, which type of restaurant should be open considering that the investor already has experience with Japanese, Italian and vegetarian restaurants?

2. Data acquisition and cleaning

In this section, it will be presented the data sources for the problem and how the data was pre-processed.

2.1 Data sources

To answer the question proposed in the previous section, some location information will be required. First, it is necessary the latitude and longitude of the park. Due to its size, the center of the park will be used as reference for its location. The latitude and longitude were identified using Google maps and is -23.588375, -46.658258. With the latitude and longitude of the park, the closest food venues will be identified through Foursquare API,

with a radius of 2500 meters. This large radius was chosen due the size of the park. With a 2500 meter-radius, we will assure that at least a five-block radius surrounding the park will be considered when looking for food venues.

2.2 Data cleaning

After retrieving the data from Foursquare, it is necessary to choose the columns that have any information related to the venue location and use the IDs retrieved to format the columns. This formatting process is necessary because Foursquare returns codes instead of the names of the venue and categories, which makes the analysis more complex.

After formatting the data frame, it is necessary to check the entries for possible venues that doesn't sell food but were labeled as food venues. To clean the data set, the columns "categories" was checked for the words "café", "restaurant" and "food" because some venues had food in their names but their categories didn't match the characteristics of a restaurant. For example, one of the entries, although had food in their name, was a startup that develops food, therefore, it shouldn't be included in the analysis. After cleaning the data set, the information is ready to be processed.

3. Research method

After collecting the data and cleaning the data, a cluster analysis was performed using the latitude and longitude of the venues to identify regions with lower density of food venues. The cluster analysis was chosen because, when plotted in a folium's map, can provide a visual representation of regions with higher and lower venue density.

After forming the clusters, the region that doesn't present a cluster, or presents the cluster with the lowest density, will be chosen to be the potential location. With the potential location in mind, the surrounding venues of the location will be evaluated. If there is a region with no venues, the closest cluster will be used to choose the appropriate restaurant. Since our investor wants to open a restaurant with fewer competitors, the categories of the venue will be reviewed to see how many Japanese, Italian and vegetarian restaurants are in the region, the type of restaurant that has fewer direct competitors will be chosen.

4. Results

In this section, the main results of the cluster analysis will be presented using the folium module as a representation tool. First, all the venues identified were plotted in a map for a first analysis.
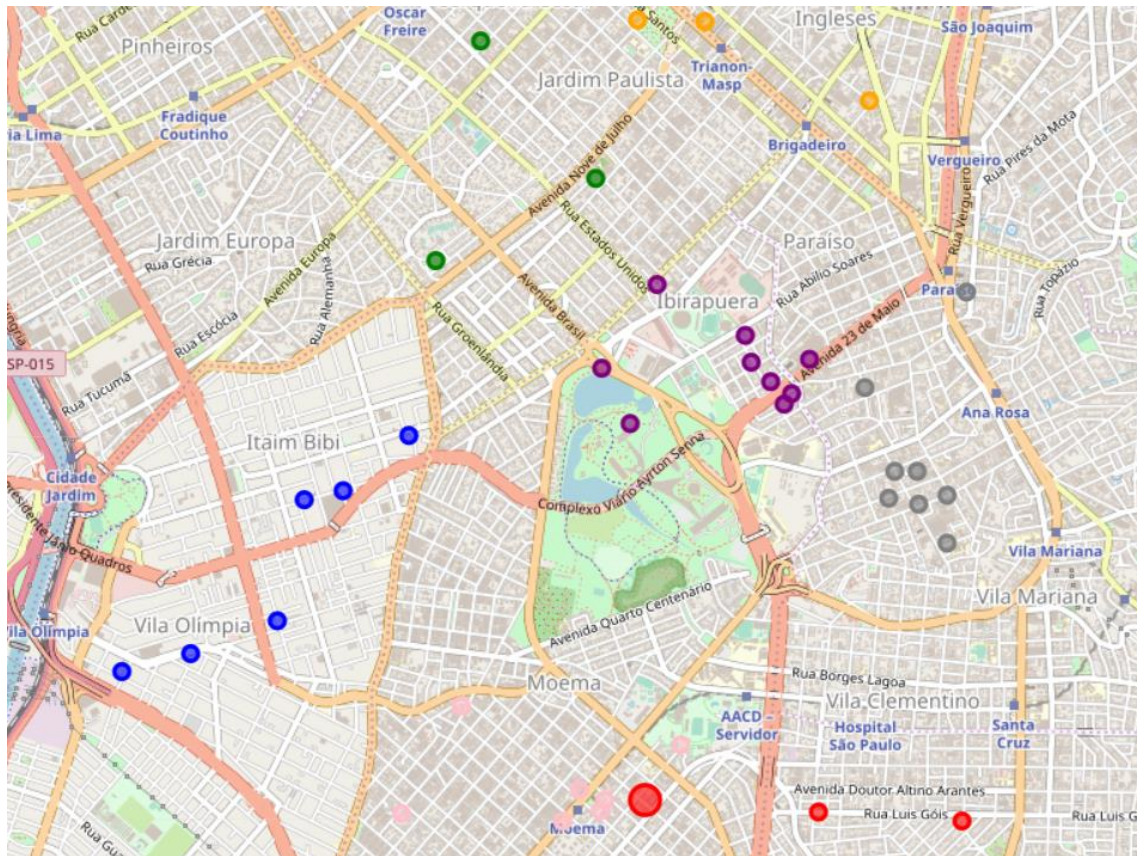
Figure 1: Location of the food venues.
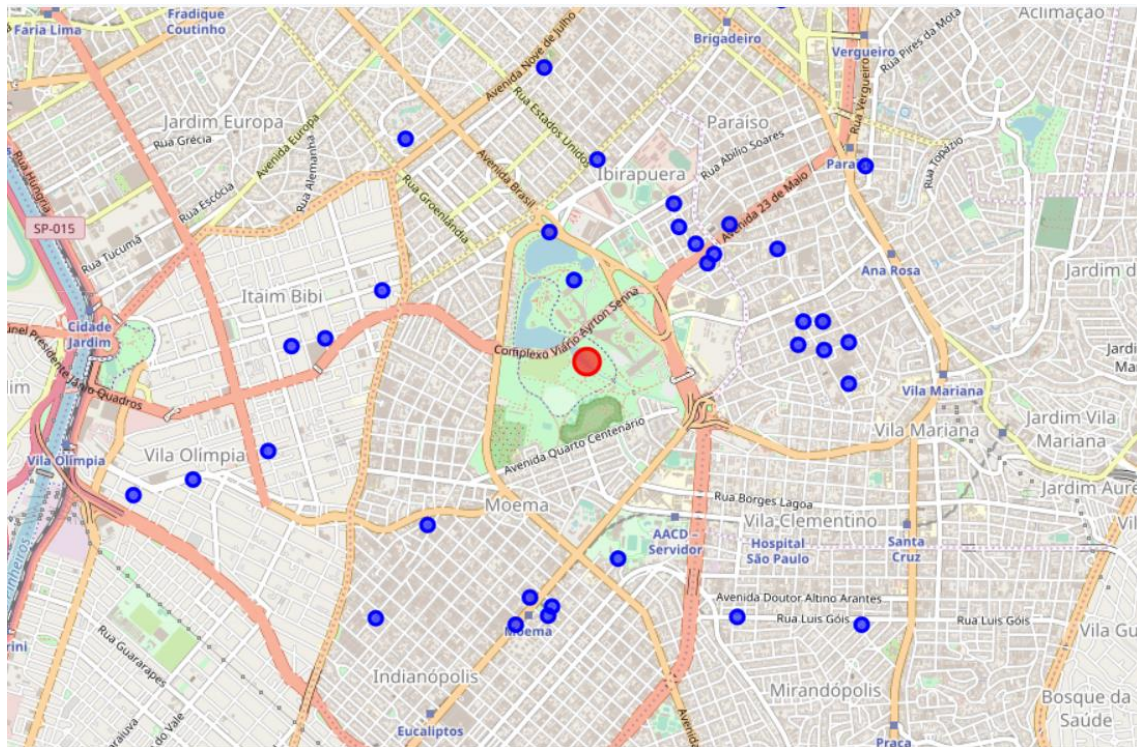
Even without the clusters, it is possible to see that the East region of the map has most of the venues, and the venues in the South and West regions are further from the park, when compared to the ones in the East.

To further explore the regions, the cluster analysis was performed considering seven clusters. The results are presented in the Figure 2.

Figure 2: Location of the food venues considering their clusters.

With Figure 2 is possible to confirm the observation from Figure 1. The East region of the park, next to the purple and gray clusters, has a higher density of food venues, while the East and South regions almost doesn't have any venues and the ones that are in those regions are far from the park. Both South and West regions would be good locations for the restaurant because they have a low density of venues and both can be easily accessed by Quarto Centenário Avenue or República do Líbano Avenue, therefore, both regions will be further explored.

In the West region, the blue cluster is the closest one to the park. In the blue cluster, there is one Japanese restaurant and two restaurants named Fit Food, that might provide some vegetarian options. Therefore, if our investor chooses to stay West of the park, the best option would be opening an Italian restaurant because he will not have any direct competitors.

In the South region, the pink cluster is the closest to the park. In the pink cluster, there are several food trucks and a Japanese restaurant. Since our investor have experience with Japanese, Italian and vegetarian restaurants, he is able to choose between Italian and vegetarian to avoid any direct competitors.

5. Discussion

Considering the results from the cluster analysis, the West and South regions of the park were identified as the best locations for the new restaurant. Both regions not only have a

low density of food venues, but the venues in those regions are further to the park when compared to the ones in the East, for example. Both regions were further explored and the closest clusters in each region, blue in the West and pink in the South, were individually searched for vegetarian, Italian and Japanese restaurants. If the investor chooses to stay in the West, the best option for him would be opening an Italian restaurant, because there isn't any Italian restaurant in the blue cluster. If he chooses to stay in the South, he could choose between Italian or vegetarian restaurants, because there aren't any vegetarian and Italian restaurants in the pink cluster.

6. Conclusion

In this project, an investor wanted to open a restaurant next to the Ibirapuera park in São Paulo - Brazil, to take advantage of the number of tourists that visit the parks every year. A cluster analysis was used to identify the regions next to the park that had the lowest density of food venues, and both South and West regions were identified as good options. Both regions had a lower density of food venues, they can be easily accessed using Quarto Centenário or República do Líbano avenues and the closest food venues were several blocks from the park.

Since both regions were good options, they were both further explored. In the West region the best option would be opening an Italian restaurant, since there are Japanese and vegetarian options available. In the South, the investor can choose between Italian and vegetarian, because there aren't vegetarian and Italian restaurants available nearby, but there is a Japanese option that would make a new Japanese restaurant not so interesting.