Capstone Project - The Battle of Neighborhoods

DETERMINING POSSIBLE LOCATIONS FOR A NEW ITALIAN RESTAURANT IN PORTO ALEGRE, BRAZIL

Bruno Peixoto Aguilar

January 29, 2021

1. INTRODUCTION AND BUSINESS PROBLEM

Porto Alegre is the largest city in the south of Brazil. Porto Alegre is home to more than 1.5 million people. In addition to its population, there are more than 3 million people who live in its metropolitan area. Economically, Porto Alegre is also a very important city of Brazil as its GDP is the eighth highest among all cities of Brazil.

Determining the location of a business is a crucial decision that business owners must make. In the case of restaurants, this decision is very critical. A bad location might be the reason for a good restaurant to close permanently. In this context, the objective of this project is to study what are the most promising areas to start a new Italian restaurant in the city of Porto Alegre so that the business has better chances of being successful. The present project will attempt to answer this question supported by data.

To determine the location of the restaurant, some basic characteristics are required from the neighbourhoods. First, the neighbourhood must have a significant number of restaurants. Second, the neighbourhood must have a considerable number of potentials clients. Therefore, the ideal neighbourhood must have a minimum population with financial conditions. Lastly, to avoid competition, the focus will be to determine a neighbourhood with no Italian restaurants, if possible. Otherwise, the focus will be to choose areas with few Italian restaurants.

2. DATA ACQUISITION AND PREPROCESSING

Data was gathered from varied sources. The data sources are listed below:

1. Wikipedia list of Porto Alegre boroughs, their population and average income.
2. Foursquare data on restaurants locations.
3. Coordinates for each boroughs of Porto Alegre retrieved geocoder.

After gathering all data, data will be treated and cleaned. As the list of boroughs and their coordinates are in two separate datasets, both datasets will be merged. Additionally, the average income in the Wikipedia dataset is provided in terms of “Minimum monthly salary

per household". Therefore, it will be converted to Brazilian Reais (BRL), utilizing the value of the current minimum monthly salary.

The following criteria will be used to analyze each neighbourhood in Porto Alegre:

1. Average income of the residents of the neighbourhood
2. Total population in the neighbourhood
3. Number of restaurants in the neighbourhood
4. Number of Italian restaurants in the neighbourhood

3. METHODOLOGY

In this section the methodology applied to this project will be explained. The first step in the project was collect all the data necessary for this project and prepare it for the next steps. The first data set to be acquired was the list of neighbourhoods of Porto Alegre. This was retrieved from Wikipedia (link for the original page: https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Porto_Alegre). The dataset was cleaned so it remained only with the following features: neighbourhood name, average household income and the population. The list of neighbourhoods was used to collect the coordinates for each of them using geocoder. The coordinates (latitude and longitude) were merged to the data.

The venues of Porto Alegre were obtained using Foursquare by utilizing the list created in the previous step. All venue categories not related to restaurants was excluded from the dataset as they are not the focus of this project. One hot encoding was used to convert data so it can be used for next steps. K-Means was used to identify clusters based on the existing restaurants (i.e., venue categories). The focus of this analysis was to identify clusters of neighbourhoods that does not fit in the characteristics that are wanted. The "k" constant was determined using the elbow method.

Lastly, after narrowing down the candidate neighbourhoods for the new restaurant, all remaining neighbourhoods were analyzed individually. As abovementioned, the goal was to find neighbourhoods that have restaurants, but that do not have any Italian restaurant ideally. So, the neighbourhoods that have any Italian restaurant were excluded from the final list.

The result of this methodology is a list of neighbourhoods that meet the minimum requirements listed in the beginning of this report.

4. DATA ANALYSIS

Figure 1 shows how the first 5 rows of the dataset after retrieving all required data and cleaning. The complete dataset contains 79 rows.

	Neighbourhood	Population	Average income (BRL)	Latitude	Longitude
0	Agronomia	12222	4,378.00	-30.09	-51.12
1	Anchieta	203	9,251.00	-29.98	-51.17
2	Arquipélago	5061	3,256.00	-29.99	-51.23
3	Auxiliadora	9985	21,527.00	-30.02	-51.19
4	Azenha	13449	11,803.00	-30.05	-51.22

Figure 1 - Neighbourhoods dataset

The map of Porto Alegre with the neighbourhoods was initially generated to assure that coordinates were precise enough for the analysis. Figure 2 shows the map of Porto Alegre.

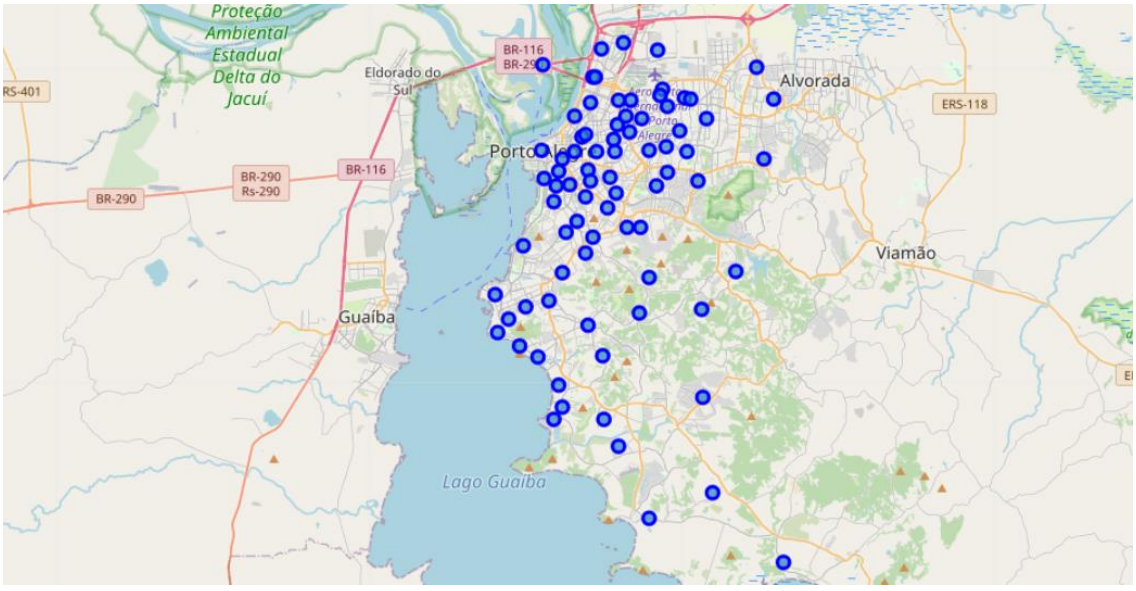

Figure 2 – Map of Porto Alegre

The average income per neighbourhood and the population per neighbourhood are shown in Figure 3 and 4, respectively. In addition, to understand if there is any sort of relation between these two variables, a scatter plot was created (Figure 5). As seen in the scatter plot there is no clear relationship between the two variables. However, the charts allowed some neighbourhoods to be excluded from the list of candidates based on their population (>5000) and average income (>BRL 10,000). These values were determined only to exclude very low values for both variables.

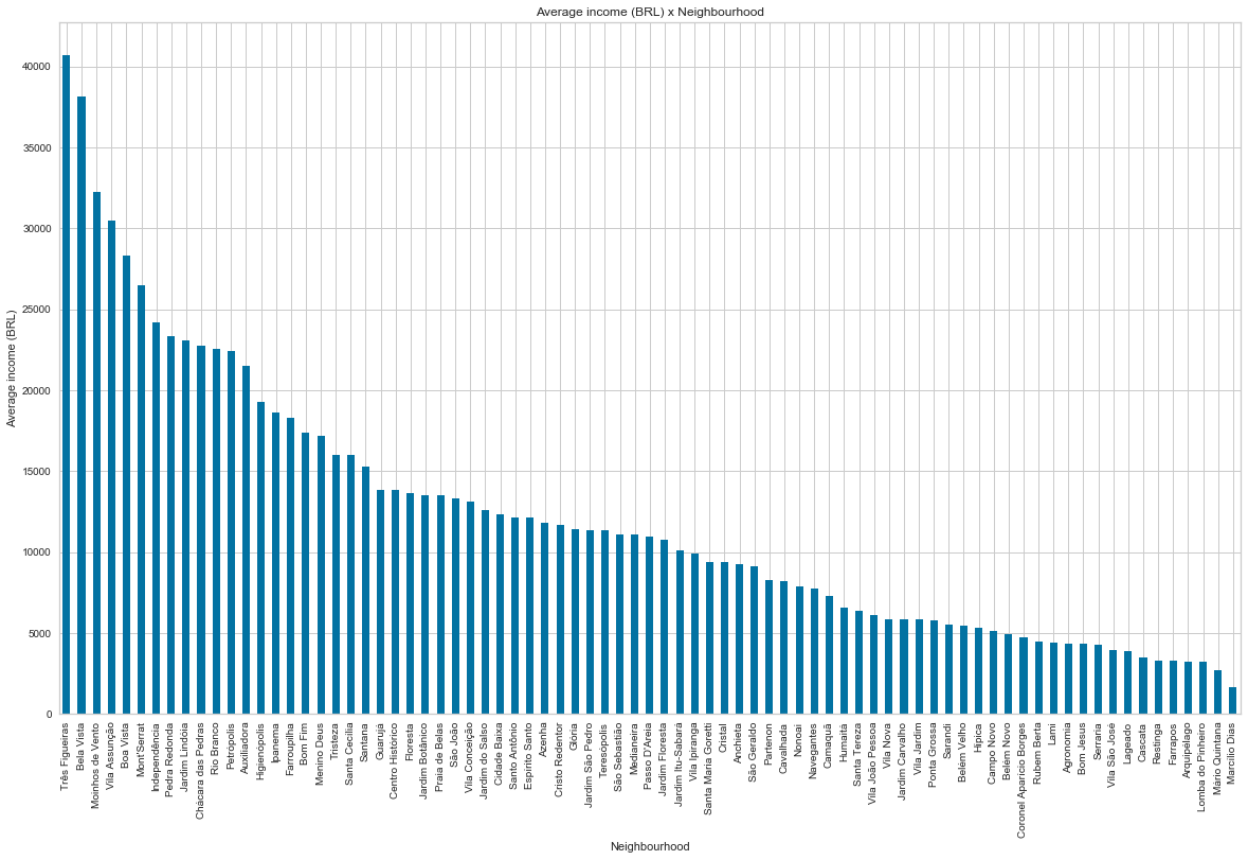

Figure 3 – Average income per neighbourhood

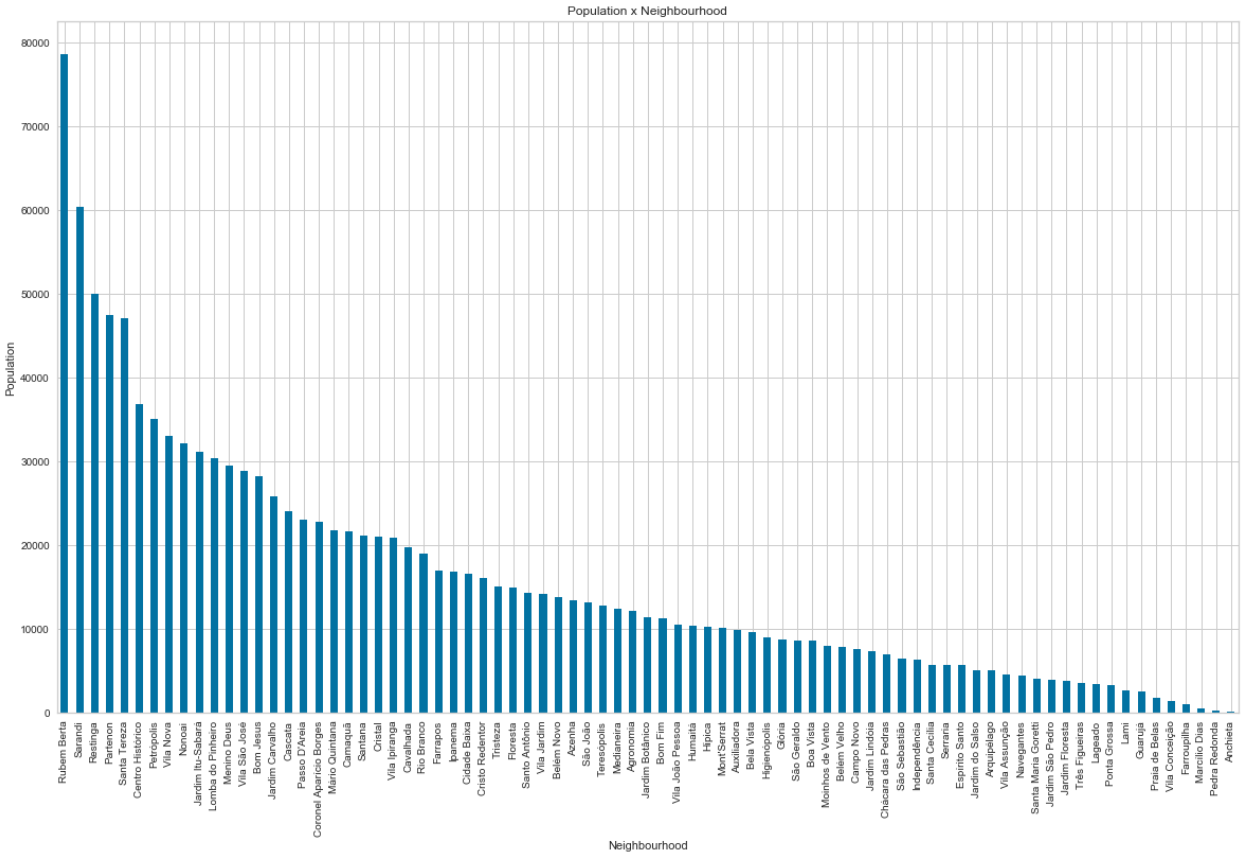

Figure 4 – Average income per neighbourhood

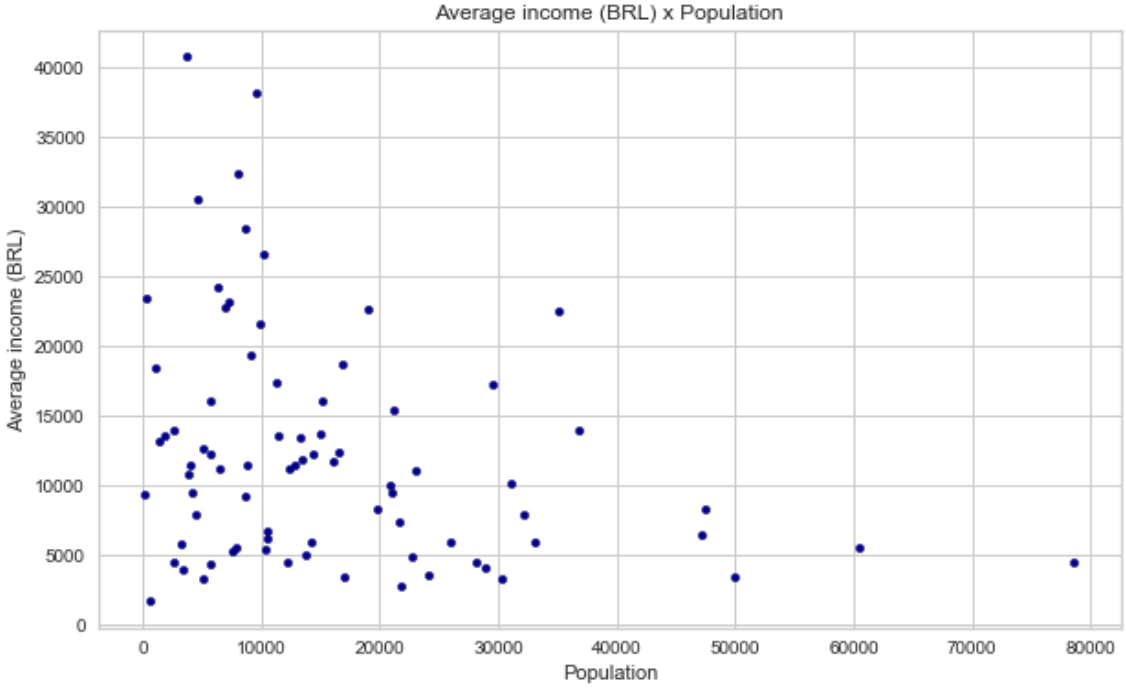

Figure 5 –Population vs average income scatter plot

Figure 6 shows the first five rows of the list of venues from retrieved from Foursquare. The list was treated to group these venues in their neighbourhoods and show the frequency of each venue category for the neighborhoods.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Anchieta	-29.98	-51.17	Locare - Locadora de Materiais para Eventos	-29.99	-51.16	Department Store
1	Arquipélago	-29.99	-51.23	Restaurante da Ilha	-29.99	-51.23	Restaurant
2	Arquipélago	-29.99	-51.23	Motel da Ilha	-29.99	-51.23	Motel
3	Auxiliadora	-30.02	-51.19	Natasul	-30.02	-51.19	Gym Pool
4	Auxiliadora	-30.02	-51.19	Estética Visualite	-30.02	-51.19	Salon / Barbershop

Figure 6 – Porto Alegre venues dataset

K-Means was used to organize neighbourhoods into clusters based on their similarities in terms of venues. The elbow analysis (Figure 7) was used to determine the ideal "k" in this analysis. Values tested for "k" varied from 2 to 9. The algorithm did not suggested any values for "k" as it did not identify any elbow according to its own criteria. However, in a visual analysis there is a small elbow in k=4. Therefore, this value was used for clustering.

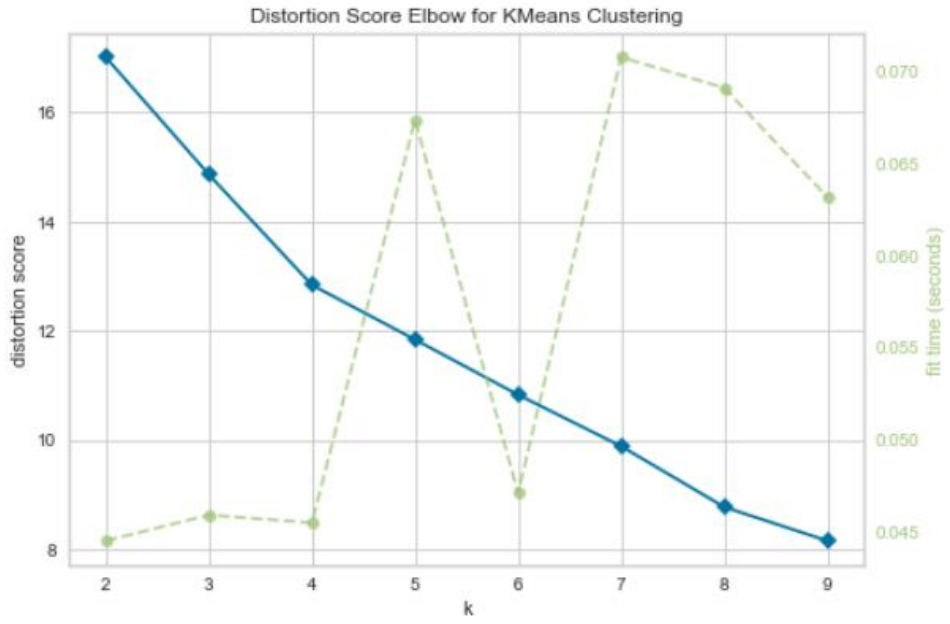

Figure 7 – K-Means elbow analysis

With k=4, K-Means algorithm was used, and Figure 8 shows the resulting clustering map. In red (cluster 1), it is the main cluster of the analysis with many neighbourhoods with similar characteristics. The cyan (cluster 3) and the purple (cluster 2) cluster are very similar to the red cluster in its elements. Despite being classified as three different clusters, the analysis of their elements is not enough to assure that there is a significant difference between them. As the analysis relies of the venue categories used in Foursquare, they might be inaccurate. For this reason, all three clusters will be considered in the next step. However, cluster 4 (in yellow) and cluster 5 (not shown in the map) were used to exclude neighbourhoods from the list of candidates. Cluster 4 is mostly comprised by neighbourhoods in which the most frequent venues are bars and pubs. Therefore, there is no evidence that these neighbourhoods have areas with restaurants. Cluster 5 was created to accommodate all neighbourhoods with food-related venues. Consequently, its elements were also excluded from the list of candidates for the same reasons of cluster 4.

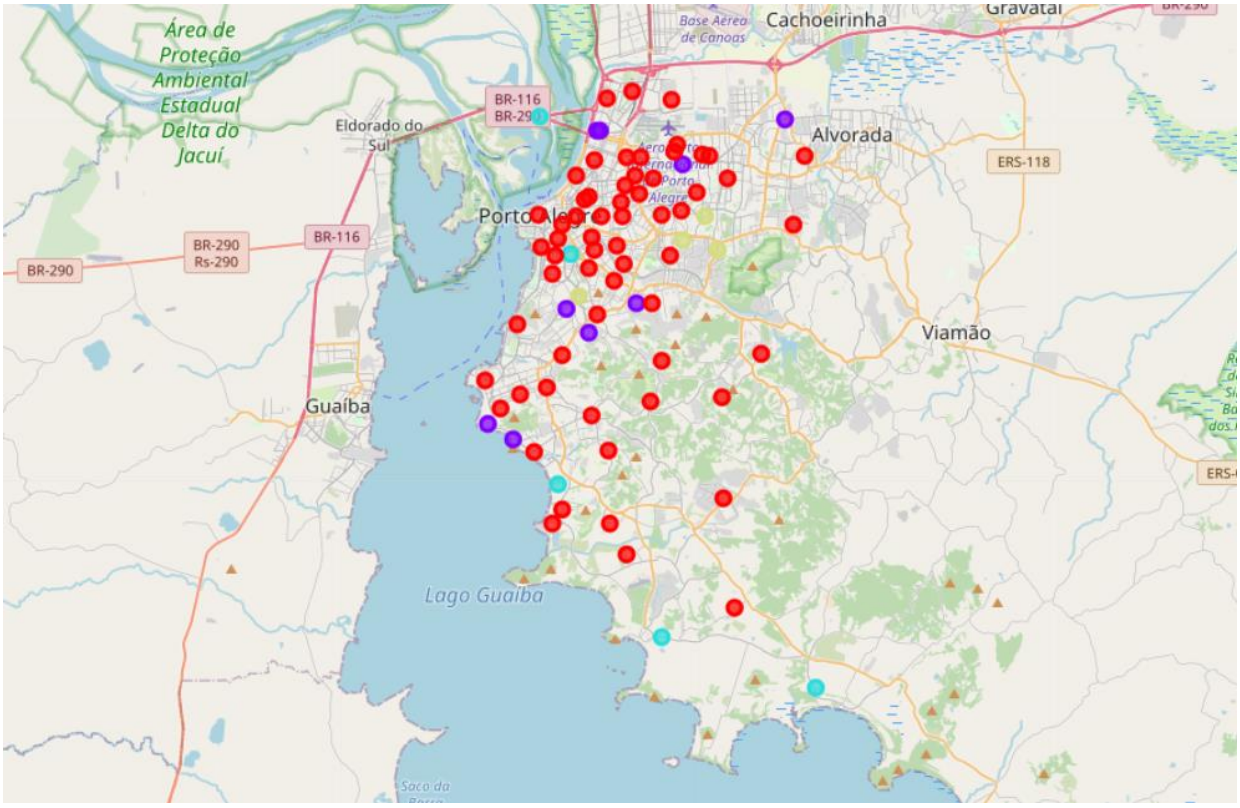

Figure 8 – Clustering analysis of neighbourhoods in Porto Alegre

As the last step of the analysis, the focus was to identify among the list of possible locations which ones still does not have an Italian restaurant. Therefore, the list of venues per neighbourhood was analyzed once more and every neighbourhood with at least one Italian restaurant was excluded from the list of candidate neighbourhoods. Figure 9 shows the table with the neighbourhoods and their respective number of Italian restaurants.

Neighbourhood	Venue Category
Auxiliadora	2
Boa Vista	1
Bom Fim	2
Centro Histórico	1
Chácara das Pedras	2
Cidade Baixa	2
Espírito Santo	1
Humaitá	1
Independência	1
Jardim Botânico	2
Jardim Lindóia	1
Moinhos de Vento	2
Rio Branco	5
Santana	2
São Geraldo	2
São Sebastião	2
Tristeza	1

Figure 9 – Clustering analysis of neighbourhoods in Porto Alegre

Figure 10 shows the final list of possible neighbourhoods to open a new Italian restaurant in Porto Alegre according to the criteria used in this project.

	Neighbourhood	Population	Average income (BRL)	Latitude	Longitude
0	Azenha	13449	11,803.00	-30.05	-51.22
1	Bela Vista	9621	38,148.00	-30.03	-51.19
2	Cristo Redentor	16103	11,671.00	-30.01	-51.16
3	Floresta	14941	13,629.00	-30.02	-51.21
4	Glória	8809	11,407.00	-30.07	-51.20
5	Higienópolis	9096	19,283.00	-30.02	-51.18
6	Ipanema	16877	18,634.00	-30.13	-51.23
7	Jardim do Salso	5143	12,584.00	-30.05	-51.17
8	Menino Deus	29577	17,160.00	-30.06	-51.22
9	Mont'Serrat	10236	26,477.00	-30.03	-51.19
10	Passo D'Areia	23083	10,956.00	-30.02	-51.18
11	Petrópolis	35069	22,407.00	-30.05	-51.19
12	Santa Cecília	5800	15,983.00	-30.04	-51.21
13	Santo Antônio	14392	12,133.00	-30.05	-51.21
14	São João	13238	13,354.00	-30.01	-51.19
15	Teresópolis	12844	11,341.00	-30.08	-51.21

Figure 10 – List of possible neighbourhoods to establish a new Italian restaurant in Porto Alegre

5. RESULTS AND DISCUSSION

The analysis showed that restaurants are pulverized in all Porto Alegre. In other words, there is no specific area in which restaurants are extremely concentrated. Moreover, there are some areas in which Foursquare did not provide any data on their restaurants. In this project, there were no constraints in terms of neighbourhoods in Porto Alegre. Therefore, the entire city was considered.

After collecting all data required for this project, the first criteria to narrow down the possible neighbourhoods for establishing a new restaurant were the population and the average income in each neighbourhood. This process excluded many locations, reducing the candidate neighbourhoods to 33.

K-Means clustering provided some more information to exclude more neighbourhoods based on the characteristic of the restaurants in those clusters. However, this process was limited because of the lack of data. It allowed only two clusters to be excluded. The first excluded cluster was characterized by the pubs and bars. These neighbourhoods were excluded because the areas of interest must be characterized by restaurants specifically. The other cluster excluded from the analysis was characterized by not having any data on its restaurants.

At this point, the candidate areas were characterized by medium to high income and population, and the presence of restaurants. However, the objective was to find an area with restaurants, but with no Italian restaurant. Therefore, the last step was to exclude all other neighbourhoods that already have an Italian restaurant. The result of this process was a list of 16 neighbourhoods with no Italian restaurant and with people who have financial conditions to be a customer.

6. CONCLUSIONS

The main objective of this project was to determine the location of a new Italian restaurant in the city of Porto Alegre, Brazil. By using Foursquare data to determine the distribution of restaurants in the city and using socio-economical data from the neighbourhoods, it was possible to reduce the area where the restaurant could be established.

The process is replicable to other cities depending on the available data. Other sources of evaluation of venues can be used to improve the results of the process. Lastly, to determine the final location for the restaurant, other factors would have to be taken into consideration including other characteristics of the neighbourhoods and the strategy to be implemented for this new restaurant.