The Battle of Neighborhoods - Curitiba

Using data science to find the best location for a new business

Fernanda Sonego Bonaldo

April 2021

1. Introduction

1.1 Background

Curitiba is one of the biggest and most multicultural cities of Brazil. It is regarded the best in which to invest in Brazil and because of its diverse economy and cultural background it is also considered a great place to launch new business ideas, services and even products. The city is divided into 9 boroughs covering the 75 neighborhoods, each with its own set of characteristics.

1.2 Problem

The same features that make Curitiba the great city that it is, also turn the task of choosing a location for a new business into a real challenge. The goal of this project is to figure out the best place to open a new special diets and healthy foods store, taking in consideration the owner's preference for a region and the profile of the neighborhoods being analysed based on local venues and the local population's demographics.

2. Data acquisition

2.1 Web Scraping

The first step was using web scraping to read the wikipedia page containing a table for every borough in Curitiba and data collected in 2010 about each neighborhood covered by the borough. The data can be accessed in https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Curitiba and is displayed in the web page as follows:

Bairro Novo

Bairros oficiais de Curitiba - Regional Bairro Novo (IBGE-IPPUC/2010)[3][1][5]						
Bairro	Área (km²)	Habitantes			Domicilios particulares	Rendimento mensal médio por responsáveis dos domicílios (R$)
		Homens	Mulheres	Total		
Ganchinho	11,20	3 667	3 658	7 325	1 921	767,35
Sitio Cercado	11,12	50 631	51 779	102 410	27 914	934,95
Umbará	22,47	7 280	7 315	14 595	17 064	908,70

where "Bairro Novo" is the borough, the first column is the name of the neighborhoods in said borough, the second column is the area of the neighborhood in square kilometers, the third column is the number of citizens separated into men, women and total, the fourth

column is the number of private homes and the fifth column is the average monthly income per citizen responsible for the home in Reais.

2.2 Geographical coordinates

The second step was adding the geographical coordinates of each neighborhood to the dataframe using the geopy library and, using the collected data, narrowing down the search to one specific borough.

2.3 Foursquare API

Next, using the Foursquare location data, a second dataframe is created with nearby venues of every neighborhood in the chosen borough. This data will be leveraged for clustering the neighborhoods and determining the neighborhoods' profiles.

3. Methodology

3.1 Data pre-processing and analysis

3.1.1 Data cleaning

Before the data can be used and visualized it needs to be cleaned and transformed into the appropriate formats.

After scraping the wikipedia page with the pandas library we end up with an array of dataframes, each containing the data of a borough that were then appended together into one dataframe.

We know that the store is aimed toward the middle class so average income data is very important for future comparison. On the other hand the store is not aimed at a specific gender so the gender data can be dropped, along with the second and fourth columns shown in the image of the wikipedia page, keeping only the relevant data on the dataframe. New column names were set, with the final dataframe having 76 rows, and 5 columns, including the index column.

	index	neighbourhood	total_population	monthly_income	Borough
0	3	Ganchinho	7 325	76735	Bairro Novo
1	4	Sitio Cercado	102 410	93495	Bairro Novo
2	5	Umbará	14 595	90870	Bairro Novo
3	3	Abranches	11 165	1 009,67	Boa Vista
4	4	Atuba	12 632	1 211,60	Boa Vista
...	...	...	...	...	...
71	11	Santo Inácio	6 037	1 518,26	Santa Felicidade
72	12	São Braz	23 119	1 206,50	Santa Felicidade
73	13	São João	2 950	1 166,03	Santa Felicidade
74	14	Seminário	7 395	3 210,65	Santa Felicidade
75	15	Vista Alegre	9 930	2 079,83	Santa Felicidade

76 rows × 5 columns

After analysing the dataframe above it was noticed that the numerical data of the column “monthly_income” was missing the comma in a few rows. Since all of the values were supposed to have a comma indicating the cents, in the cases where the comma was not present the last two digits were removed. This loss of precision was not relevant when comparing values in a macro scale.
Next, the commas were exchanged for periods and spaces were removed in order to cast the values of the columns “total_population” and “monthly_income” from object to integer and float, respectively. A new dataframe called df_income containing the average income per borough was also created to facilitate future visualization of the data.
In order to visualize the location of the neighborhoods in a map of Curitiba, the geopy library was used to add the geographical coordinates of each neighborhood to the dataframe, ending up with the following dataframe:

	index	neighbourhood	total_population	monthly_income	Borough	latitude	longitude
0	3	Ganchinho	7325	767.00	Bairro Novo	-25.572076	-49.263667
1	4	Sitio Cercado	102410	934.00	Bairro Novo	-25.542701	-49.269106
2	5	Umbará	14595	908.00	Bairro Novo	-25.568169	-49.285699
3	3	Abranches	11165	1009.67	Boa Vista	-25.361474	-49.272054
4	4	Atuba	12632	1211.60	Boa Vista	-25.387500	-49.206606
...	...	...	...	...	...	...	...
71	11	Santo Inácio	6037	1518.26	Santa Felicidade	-25.425206	-49.328578
72	12	São Braz	23119	1206.50	Santa Felicidade	-25.418226	-49.350834
73	13	São João	2950	1166.03	Santa Felicidade	-25.391453	-49.311479
74	14	Seminário	7395	3210.65	Santa Felicidade	-25.448910	-49.305147
75	15	Vista Alegre	9930	2079.83	Santa Felicidade	-25.406766	-49.295578

76 rows × 7 columns

Using data visualization techniques described in the next section of this report, the search was narrowed down to one borough containing 10 neighborhoods, and the next step, as mentioned in the introduction, was to get Foursquare location data of all the nearby venues of the selected borough. A new dataframe was created containing all of the Foursquare API data. In total there were 173 venues of which there were 72 unique categories, with “bakery” being the most prominent one, with 16 acurrences.
One hot encoding was applied to the dataframe and the resulting dataset was grouped by neighborhood, resulting in a dataframe with 11 rows, one for each neighborhood and one for column names, and 73 columns, one for each unique venue category and one for index.
Using the dataframe previously created, a new dataframe with the 10 most common venues of each neighborhood was created as shown in the next image:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Campo Comprido	Pizza Place	Bakery	Middle Eastern Restaurant	Gym	Women's Store	Dessert Shop	Convenience Store	Cosmetics Shop	Dance Studio	Deli / Bodega
1	Fanny	Bakery	Gym	Hot Dog Joint	Soccer Field	Liquor Store	Pet Store	Pharmacy	Gym / Fitness Center	Market	Electronics Store
2	Guaíra	Bakery	Electronics Store	Market	Café	Steakhouse	Bookstore	Japanese Restaurant	Diner	Dessert Shop	Cosmetics Shop
3	Lindoia	Buffet	Park	Bakery	Market	Convenience Store	Pizza Place	Recording Studio	Ice Cream Shop	Bus Station	Deli / Bodega
4	Novo Mundo	Soccer Field	Pet Store	Supermarket	Market	Bus Station	Gym Pool	Electronics Store	Brazilian Restaurant	Gym	Gas Station
5	Parolin	Tennis Court	Gym / Fitness Center	Café	Gastropub	Brazilian Restaurant	Gym	Event Service	Electronics Store	Cosmetics Shop	Clothing Store
6	Portão	Restaurant	Bakery	Bar	Dance Studio	Pizza Place	Paper / Office Supplies Store	Candy Store	Italian Restaurant	Coffee Shop	Convenience Store
7	Santa Quitéria	Bakery	Brazilian Restaurant	Soccer Field	Churrascaria	Pizza Place	Business Service	Recording Studio	Farmers Market	Furniture / Home Store	Wine Shop
8	Vila Izabel	Pizza Place	Churrascaria	Bakery	Sushi Restaurant	Brazilian Restaurant	Coffee Shop	Paper / Office Supplies Store	Food Truck	Deli / Bodega	Plaza
9	Água Verde	Brazilian Restaurant	Pizza Place	Beer Store	Argentinian Restaurant	Burger Joint	Diner	Dessert Shop	Dance Studio	Gym	Gym / Fitness Center

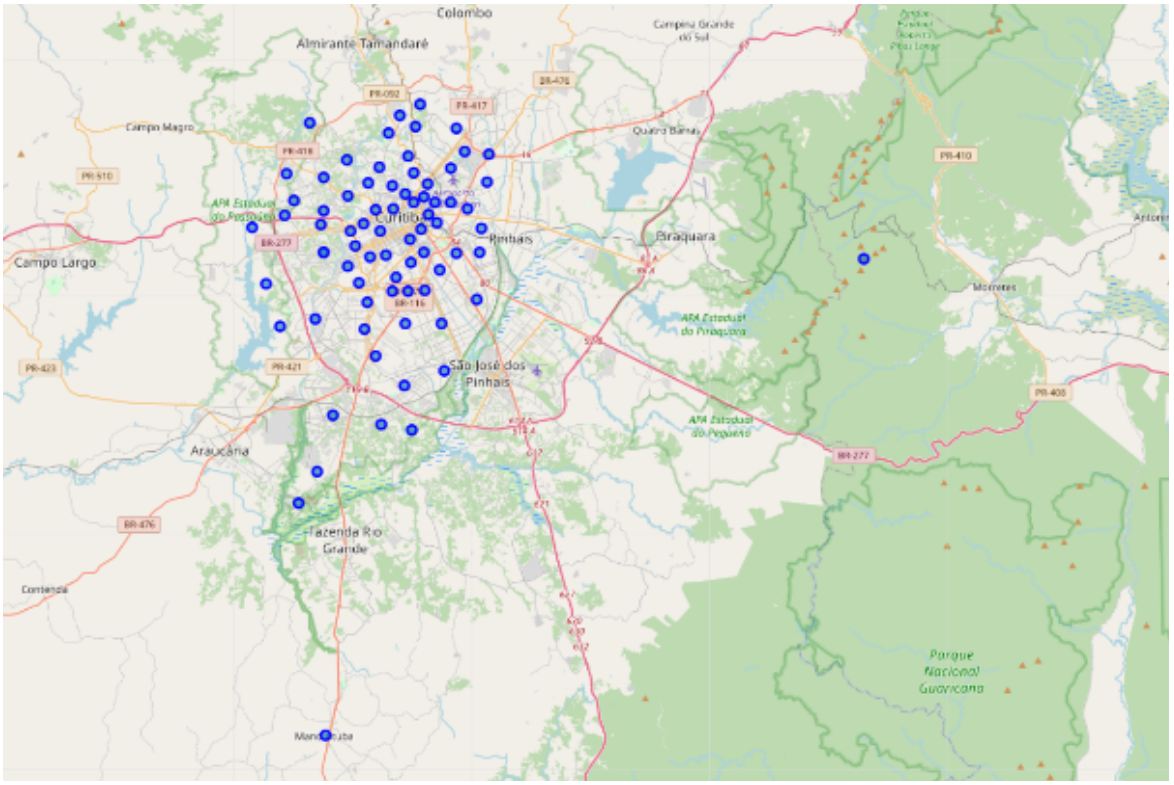
3.2 Data clustering

This new dataframe was then used to partition the neighborhoods into 3 clusters labeled as 0, 1 and 2. The clustering was done using k-means clustering from the sklearn library. The initial borough dataframe was merged with the dataframe containing the 10 most common venues per neighborhood and the cluster labels were added creating a final dataframe called "fp_merged":

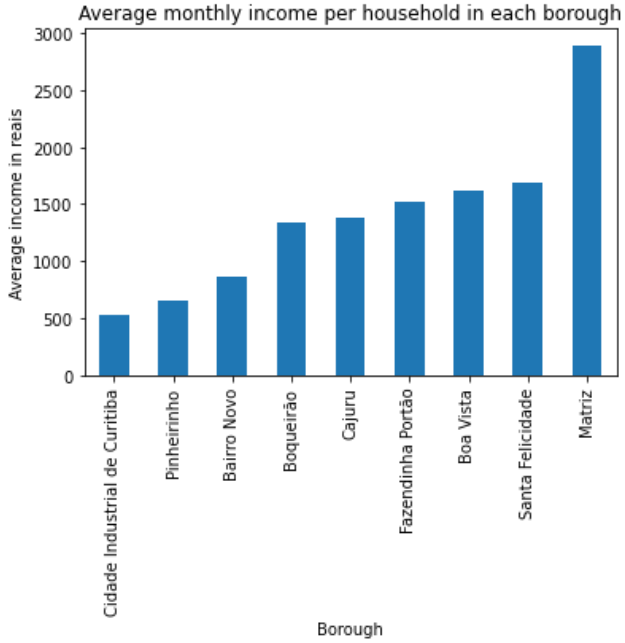
	level_0	index	Neighborhood	total_population	monthly_income	Borough	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	0	3	Água Verde	49866	3332.57	Fazendinha Portão	-25.455263	-49.282808	2.0	Brazilian Restaurant	Pizza Place	Beer Store	Argentinian Restaurant	Burger Joint	Diner	Dessert Shop	Dance Studio	Gym	Gym / Fitness Center
1	1	4	Campo Comprido	21638	1216.71	Fazendinha Portão	-25.453340	-49.328432	0.0	Pizza Place	Bakery	Middle Eastern Restaurant	Gym	Women's Store	Dessert Shop	Convenience Store	Cosmetics Shop	Dance Studio	Deli / Bodega
2	2	5	Fanny	7866	1189.54	Fazendinha Portão	-25.479200	-49.266138	1.0	Bakery	Gym	Hot Dog Joint	Soccer Field	Liquor Store	Pet Store	Pharmacy	Gym / Fitness Center	Market	Electronics Store
3	4	7	Guaíra	14268	1235.61	Fazendinha Portão	-25.470043	-49.275242	2.0	Bakery	Electronics Store	Market	Café	Steakhouse	Bookstore	Japanese Restaurant	Diner	Dessert Shop	Cosmetics Shop
4	5	8	Lindoia	8343	809.00	Fazendinha Portão	-25.479004	-49.277692	2.0	Buffet	Park	Bakery	Market	Convenience Store	Pizza Place	Recording Studio	Ice Cream Shop	Bus Station	Deli / Bodega
5	6	9	Novo Mundo	42999	1040.40	Fazendinha Portão	-25.486966	-49.296063	1.0	Soccer Field	Pet Store	Supermarket	Market	Bus Station	Gym Pool	Electronics Store	Brazilian Restaurant	Gym	Gas Station
6	7	10	Parolin	11982	1365.48	Fazendinha Portão	-25.459976	-49.263767	1.0	Tennis Court	Gym / Fitness Center	Café	Gastropub	Brazilian Restaurant	Gym	Event Service	Electronics Store	Cosmetics Shop	Clothing Store
7	8	11	Portão	40735	1722.89	Fazendinha Portão	-25.473700	-49.302414	2.0	Restaurant	Bakery	Bar	Dance Studio	Pizza Place	Paper / Office Supplies Store	Candy Store	Italian Restaurant	Coffee Shop	Convenience Store
8	9	12	Santa Quitéria	11720	1487.95	Fazendinha Portão	-25.462602	-49.310944	2.0	Bakery	Brazilian Restaurant	Soccer Field	Churrascaria	Pizza Place	Business Service	Recording Studio	Farmers Market	Furniture / Home Store	Wine Shop
9	10	13	Vila Izabel	10949	2438.13	Fazendinha Portão	-25.456326	-49.293979	2.0	Pizza Place	Churrascaria	Bakery	Sushi Restaurant	Brazilian Restaurant	Coffee Shop	Paper / Office Supplies Store	Food Truck	Deli / Bodega	Plaza

3.3 Data visualization and exploratory analysis

The first step in visualizing the data was to use folium to create an interactive map of the city with its neighborhoods superimposed on top:



The second step was creating a bar chart with the dataframe “df_income”, previously created containing the average income per borough, to better visualize the data as shown below:



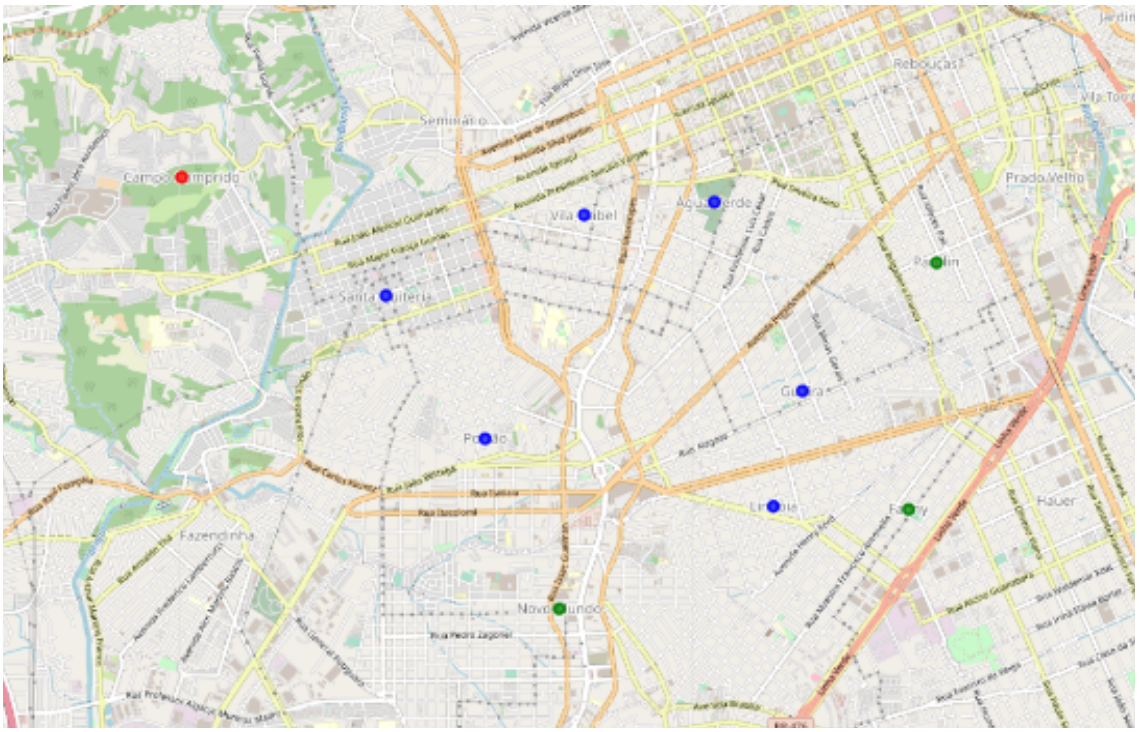
Based on the chart and the fact that the store is aimed at the middle class population, it was possible to narrow down the search from 9 to 5 boroughs: Boqueirão, Cajuru, Fazendinha Portão, Boa Vista and Santa Felicidade. Furthermore, based on the map of the city and the store owner’s personal preference for a city region, it was determined that the best borough is Fazendinha Portão and its 10 neighborhoods.

After getting the venue data from Foursquare API, as described in the data pre-processing and analysis session, each neighborhood was printed along with its 5 most common venues

for better visualization purposes. The following image is an example of one of the neighborhoods:

```
----Fanny----
           venue  freq
0         Bakery  0.19
1            Gym  0.12
2  Hot Dog Joint  0.12
3      BBQ Joint  0.06
4      Pet Store  0.06
```

With neighborhoods already clustered into 3 groups, a new map was created of Fazendinha Portão where each dot represents a neighborhood and its color represents it's cluster: red for 0, green for 1 and blue for 2.



Displaying the "fp_merged" dataframe separated by cluster we get the following:
Cluster 0 (red): Campo Comprido

Borough	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Fazendinha Portão	-25.45334	-49.328432	0.0	Pizza Place	Bakery	Middle Eastern Restaurant	Gym	Women's Store	Dessert Shop	Convenience Store	Cosmetics Shop	Dance Studio	Deli / Bodega

This cluster is made up of only one neighborhood that is located farther away from the other neighborhoods in the borough. As can be seen in the image above, the neighborhood doesn't have many health related venues, with "gym" being the only category present in the list of most common venues.
Cluster 1 (green): Fanny, Novo Mundo, Parolin

Borough	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Fazendinha Portão	-25.479200	-49.266138	1.0	Bakery	Gym	Hot Dog Joint	Soccer Field	Liquor Store	Pet Store	Pharmacy	Gym / Fitness Center	Market	Electronics Store
Fazendinha Portão	-25.486966	-49.296063	1.0	Soccer Field	Pet Store	Supermarket	Market	Bus Station	Gym Pool	Electronics Store	Brazilian Restaurant	Gym	Gas Station
Fazendinha Portão	-25.459976	-49.263767	1.0	Tennis Court	Gym / Fitness Center	Café	Gastropub	Brazilian Restaurant	Gym	Event Service	Electronics Store	Cosmetics Shop	Clothing Store

This cluster is made up of 3 neighborhoods where health related venues are very prominent, with at least 3 categories of exercise and health related venues being mentioned in each neighborhood.

Cluster 2 (blue): Água Verde, Guaíra, Lindoia, Portão, Santa Quitéria, Vila Izabel

Borough	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mc Comm Venue
Fazendinha Portão	-25.455263	-49.282808	2.0	Brazilian Restaurant	Pizza Place	Beer Store	Argentinian Restaurant	Burger Joint	Diner	Dessert Shop	Dance Studio	Gym	Gym / Fitness Center
Fazendinha Portão	-25.470043	-49.275242	2.0	Bakery	Electronics Store	Market	Café	Steakhouse	Bookstore	Japanese Restaurant	Diner	Dessert Shop	Cosmet Shop
Fazendinha Portão	-25.479004	-49.277692	2.0	Buffet	Park	Bakery	Market	Convenience Store	Pizza Place	Recording Studio	Ice Cream Shop	Bus Station	Deli / Bodega
Fazendinha Portão	-25.473700	-49.302414	2.0	Restaurant	Bakery	Bar	Dance Studio	Pizza Place	Paper / Office Supplies Store	Candy Store	Italian Restaurant	Coffee Shop	Conven Store
Fazendinha Portão	-25.462602	-49.310944	2.0	Bakery	Brazilian Restaurant	Soccer Field	Churrascaria	Pizza Place	Business Service	Recording Studio	Farmers Market	Furniture / Home Store	Wine St
Fazendinha Portão	-25.456326	-49.293979	2.0	Pizza Place	Churrascaria	Bakery	Sushi Restaurant	Brazilian Restaurant	Coffee Shop	Paper / Office Supplies Store	Food Truck	Deli / Bodega	Plaza

This is the biggest cluster, with 6 neighborhoods. As can be seen in the dataframe, food related venues are the most common ones, as opposed to health related venues, with the category "gym" being mentioned in only one neighborhood.

After analysing the clusters and concluding that the cluster 1 was the ideal one for the store, further data about the neighborhoods was displayed:

index	Neighborhood	total_population	monthly_income	Borough	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	Fanny	7866	1189.54	Fazendinha Portão	-25.479200	-49.266138	1.0	Bakery	Gym	Hot Dog Joint	Soccer Field	Liquor Store	Pet Store	Pharmacy	Gym / Fitness Center	Market	Electronics Store
9	Novo Mundo	42999	1040.40	Fazendinha Portão	-25.486966	-49.296063	1.0	Soccer Field	Pet Store	Supermarket	Market	Bus Station	Gym Pool	Electronics Store	Brazilian Restaurant	Gym	Gas Station
10	Parolin	11982	1365.48	Fazendinha Portão	-25.459976	-49.263767	1.0	Tennis Court	Gym / Fitness Center	Café	Gastropub	Brazilian Restauran	Gym	Event Service	Electronics Store	Cosmetics Shop	Clothing Store

4. Results and discussion

Considering all of the previously explained data analysis and visualization, and especially the 3 neighborhood clusters' profiles based on the venue categories, it is possible to conclude that the best place to open the new store is in the second cluster of neighborhoods, labeled 1, that is comprised of Fanny, Novo Mundo and Parolin. Further location search is out of the scope of this project and would be carried out by the owner's preferred real estate agency, where

available locations and prices would have to be analysed along with the business goals and financial projections.

4. Conclusion

The aim of this project was to solve the business problem of figuring out the best location to open a new special diets and healthy foods store. To achieve this goal broadly available data, along with information provided by the store owner, was processed, visualized and analysed using multiple data science techniques resulting in the location search being narrowed down from 75 to 3 neighborhoods.

This study demonstrates the immense, and mostly untapped, potential of data science in the resolution of business problems and how location data can be leveraged to achieve a multitude of goals from a micro scale, for example a person trying to decide where to go for lunch, to a macro scale, such as comparing countries and their cultural influence on different parts of the world.