Capstone Project - The Battle of Neighborhoods

IBM Professional Data Science Specialization

Clustering Pet Stores in Sao Paulo using Machine Learning

Danilo Ribeiro de Lima

February, 21 2021

1 - Introduction

Sao Paulo is the largest city in Brazil, presenting a city population of about 12.25 million and almost 22 million in its metropolitan region (2019). It is the Southeastern state of São Paulo and one of the richest cities in the southern hemisphere. Historically attractive to immigrants and (somewhat later) Brazilians from other states, it’s one of the world's most diverse cities.

Sao Paulo is Brazil's technological and economic hub. It has the largest economy by Gross Domestic Product (GDP) in Latin America and the Southern Hemisphere, representing 10.7% of all Brazilian GDP and being home to 63% of established multinational companies in the country.

According to the Brazilian Association of the Pet Products (Abinpet), Brazil has the second largest population of dogs, cats, and domestic songbirds worldwide and is the third-largest country in pet total population only. According to Euromonitor, there are about 150 million pets in the country — a larger number than Brazilian children, for example. According to Euromonitor, the Brazilian market for pet products is the fourth largest globally in terms of sales volume and presented a 10.8% Compound Annual Growth Rate (CAGR) between 2014 and 2019. The increase in the pet population mainly drove the evolution of this market in the country. In the next five years, this growth rate is expected to reach a 16.6% CAGR.

The number of pet stores (or pet shops) in Sao Paulo city almost doubled between 2006 and 2016, reaching 3072 stores. According to Instituto Pet Brazil, this business generates R$ 23 billion (USD 4 billion) each year, including veterinary products, pet food, accessories, and services.

Considering this growing market in Sao Paulo, what should a business person consider before deciding to open a pet store? What is the best location in Sao Paulo and why?

2 - Business Problem

The goal is to find a proper location to open a pet store by clustering Sao Paulo's districts. I will explore, segment and cluster neighborhoods in Sao Paulo and find the main features related to the business. This report will find the pet stores based on the total number of stores and their ratings. By clustering data using Foursquare API, I can provide information about the best location to open a pet store in the city.

1. What is/are the best location(s) for a pet store in Sao Paulo city?
2. In what district should the investor open a pet store to have the best chance of being successful?

3 - Data Analysis Methodology

I used the BeautifulSoap library for web scraping data frames in the Wikipedia website. The code will be provided in Jupyter Notebook.

In the next step, I searched for Sao Paulo districts' geographical coordinates. At first, I tried to use Google services, but it asks the user for a credit card number. Then geocoder class from Geopy client worked just fine to extract latitude and longitude coordinates for each Sao Paulo district. This data is uploaded to Google Drive and Geodown library is used to download data.

For venues in each Sao Paulo district, I used Foursquare API tools. It collects venues available along with their categories, ratings and counts for likes and tips.

For data preparation, I checked the file for any empty cell concerning latitude and longitude coordinates for the districts. There are 96 districts on the CSV data frame and dropped the original ‘Population’ column for this project. Later I used the Foursquare API tool to extract a maximum of 120 venues located within a 600-meter radius based on latitude and longitude coordinates from city districts.

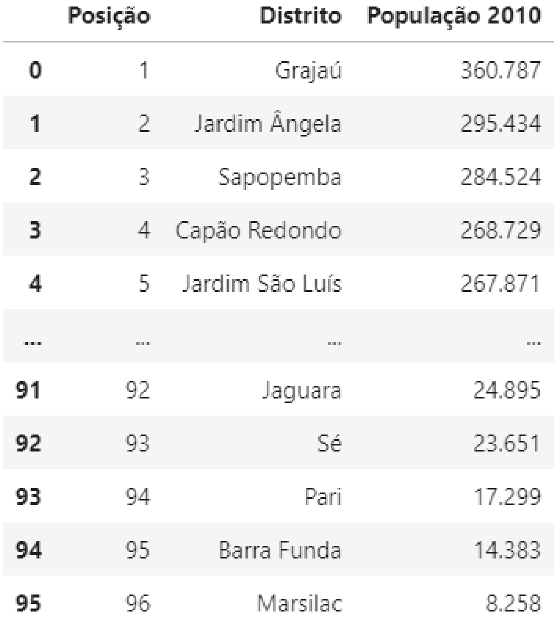
Finally, Foursquare API is used to collect ratings, likes and tips for Pet Stores in Sao Paulo districts. Then filtered Pet Stores based on minimum ratings and plotted a bar chart for decision making. I will select districts where I can find at least one Pet Store, along with their respective average ratings and merged these columns with geographical coordinates.

4 - Results and Discussion

4.1 - Web scraping

After web scraping using BeautifulSoap, Sao Paulo districts is as follow:

Table 1 : Data scraped from Wikipedia



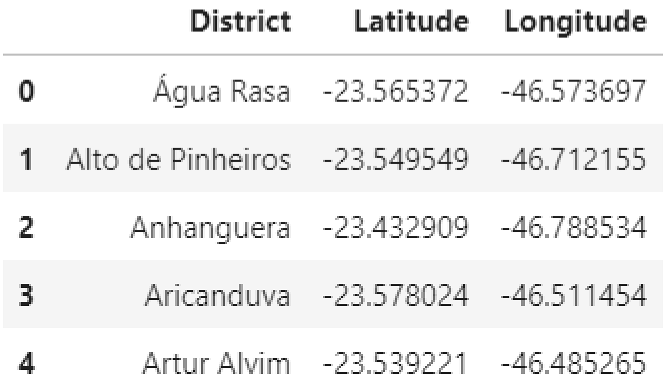
	Posição	Distrito	População 2010
0	1	Grajaú	360.787
1	2	Jardim Ângela	295.434
2	3	Sapopemba	284.524
3	4	Capão Redondo	268.729
4	5	Jardim São Luís	267.871
...	...	...	...
91	92	Jaguara	24.895
92	93	Sé	23.651
93	94	Pari	17.299
94	95	Barra Funda	14.383
95	96	Marsilac	8.258

Source: https://pt.wikipedia.org/wiki/Lista_dos_distritos_de_S%C3%A3o_Paulo_por_popula%C3%A7%C3%A3o

4-2 Adding geographical data

Then geocoder class from Geopy client worked just fine to extract latitude and longitude coordinates for each Sao Paulo district. This data is downloaded from Google Drive.

Table 2: Sao Paulo districts and their geographical coordinates.



	District	Latitude	Longitude
0	Água Rasa	-23.565372	-46.573697
1	Alto de Pinheiros	-23.549549	-46.712155
2	Anhanguera	-23.432909	-46.788534
3	Aricanduva	-23.578024	-46.511454
4	Artur Alvim	-23.539221	-46.485265

Source: https://drive.google.com/uc?id=1qa_jjyP9mwXkxc03A2jqkMnYMq8QlU6E

4-3 Exploratory data analysis

After collecting all venues and filtering the category column for Pet Stores, I grouped data by district and plotted a bar chart.

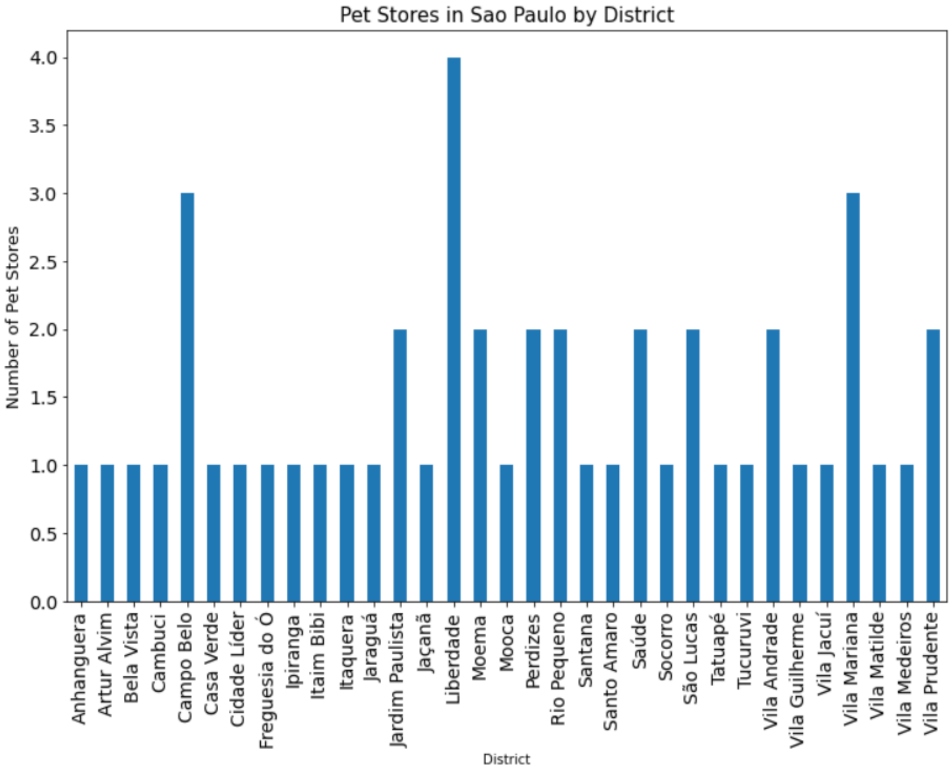


Figure1: Pet Store counting by district.

There are 47 pet stores in Sao Paulo city. *Campo Belo, Liberdade and Vila Mariana districts present a higher number of Pet Stores.*

Then need to have a look at the information collected from Foursquare. Not all districts have available ratings, for example.

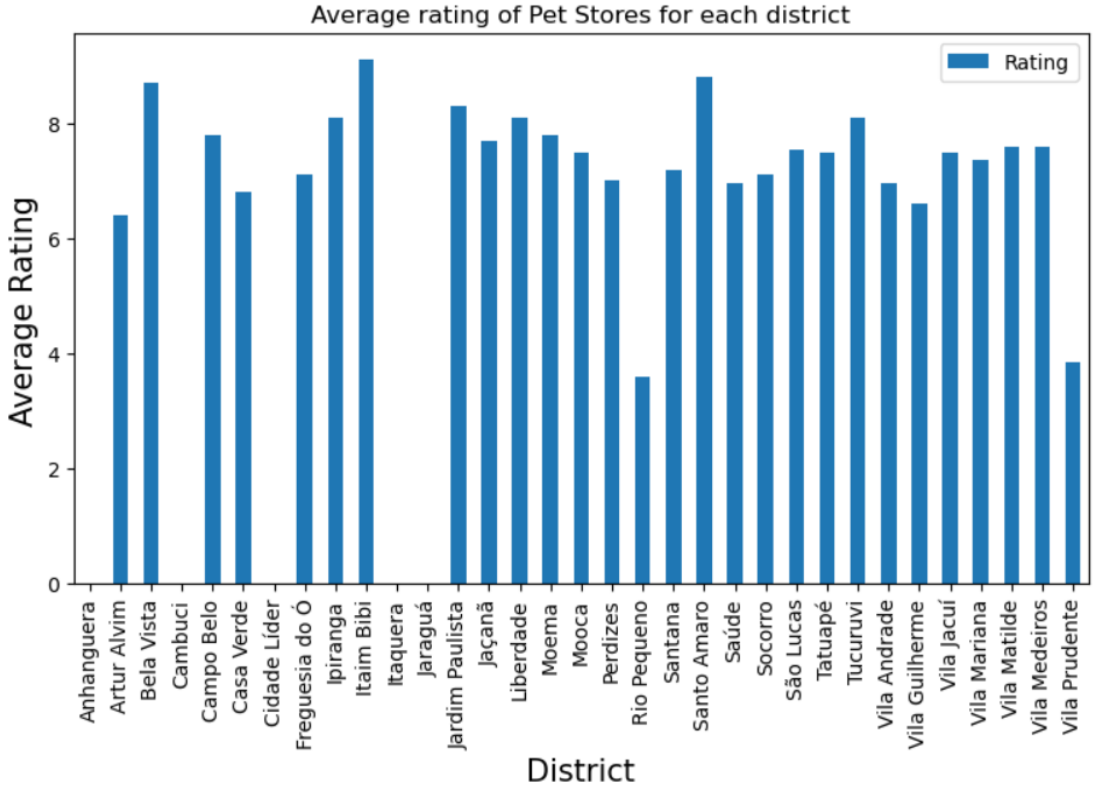


Figure 2: Pet stores in Anhanguera, Cambuci, Cidade Lider, Itaquera and Jaragua districts have no rating.

Instead of dropping districts due to missing data, in the next step, all Pet Stores with ratings higher than 7.0 are filtered; thus, districts rating lower than 7.0 and null data are removed from the data frame.

Best Pet Stores by average ratings are located in Bela Vista, Itaim Bibi and Santo Amaro. And most Pet Stores present average ratings between 7.0 and 8.0.

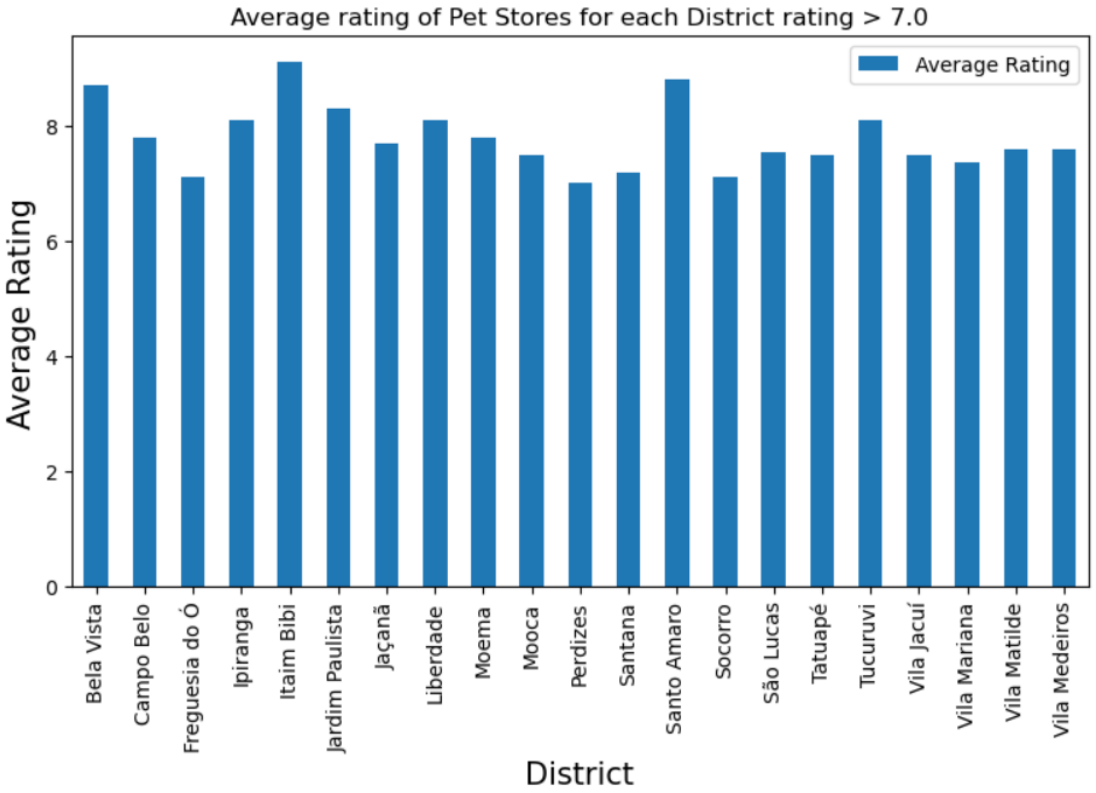


Figure 3: Districts with average rating higher than 7.0

At this point, it is important to check districts by population size. Calling the data frame just after the Wikipedia web scraping is possible to plot each Sao Paulo district population in 2010.

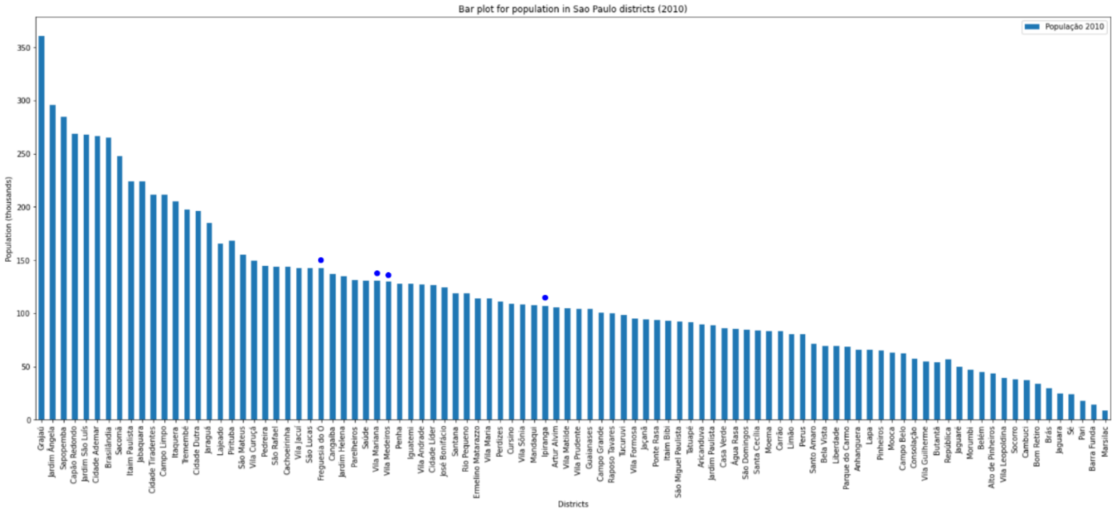


Figure 4: Sao Paulo districts by population (2010)

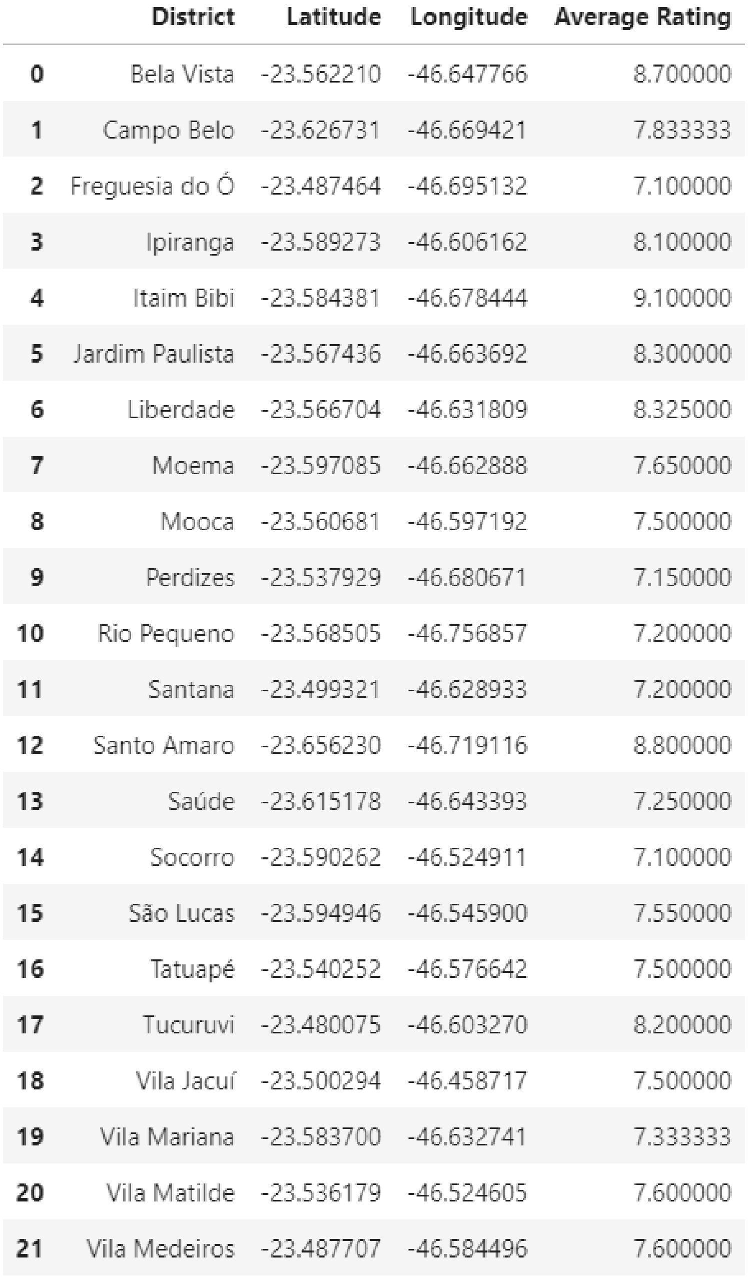
Compared to total population by district, **Freguesia do O, Vila Mariana, Vila Medeiros and Ipiranga** are ahead in the list and they also present average ratings higher than 7.0. As mentioned before, Vila Mariana has a higher number of Pet Stores by district and may point to market saturation in the area.

Freguesia do O, Vila Mariana and Vila Medeiros have one Pet Store each and a great potential for business in the respective areas.

4-4 Feature Selection and One Hot Encoding

Now that we have the average ratings for every district, we will merge this column with our first data frame containing the districts' geographical coordinates. We will call this new data frame pet_merged. The latter 3 columns will be used as features for clustering.

Table 3: Merged feature data frame for clustering



	District	Latitude	Longitude	Average Rating
0	Bela Vista	-23.562210	-46.647766	8.700000
1	Campo Belo	-23.626731	-46.669421	7.833333
2	Freguesia do Ó	-23.487464	-46.695132	7.100000
3	Ipiranga	-23.589273	-46.606162	8.100000
4	Itaim Bibi	-23.584381	-46.678444	9.100000
5	Jardim Paulista	-23.567436	-46.663692	8.300000
6	Liberdade	-23.566704	-46.631809	8.325000
7	Moema	-23.597085	-46.662888	7.650000
8	Mooca	-23.560681	-46.597192	7.500000
9	Perdizes	-23.537929	-46.680671	7.150000
10	Rio Pequeno	-23.568505	-46.756857	7.200000
11	Santana	-23.499321	-46.628933	7.200000
12	Santo Amaro	-23.656230	-46.719116	8.800000
13	Saúde	-23.615178	-46.643393	7.250000
14	Socorro	-23.590262	-46.524911	7.100000
15	São Lucas	-23.594946	-46.545900	7.550000
16	Tatuapé	-23.540252	-46.576642	7.500000
17	Tucuruvi	-23.480075	-46.603270	8.200000
18	Vila Jacuí	-23.500294	-46.458717	7.500000
19	Vila Mariana	-23.583700	-46.632741	7.333333
20	Vila Matilde	-23.536179	-46.524605	7.600000
21	Vila Medeiros	-23.487707	-46.584496	7.600000

To evaluate other Pet Stores in the region pandas one hot encoding tool is used to find the 10 most common venues in each of the 22 districts in the table above.

The get_dummies function is used to create one column for each category followed by grouping venues by district and calculating proportions for each category. A loop is created for the 1st to 10th most common venue categories. This data frame is merged with the merged feature data frame (above) and cluster labeled to examine common patterns in the data set.

4-5 Clustering districts

Description of all features necessary to run the k-means algorithm:

1 — Drop District column from the pet_merged data frame (k-means does not handle categorical variables);

2 — Run StandardScaler function from sklearn.preprocessing to normalize our features;

3 — Run KMeans algorithm to cluster data. Use the elbow method to select the optimal number of clusters.

The **elbow method** is a prevalent technique and the idea is to run k-means clustering for a range of clusters k (let’s say from 1 to 10). For each value, we are calculating the sum of squared distances from each point to its assigned center (distortions). When the distortions are plotted and the plot looks like an arm then the “elbow” (the point of inflection on the curve) is the best value of k. The optimal k is 4.

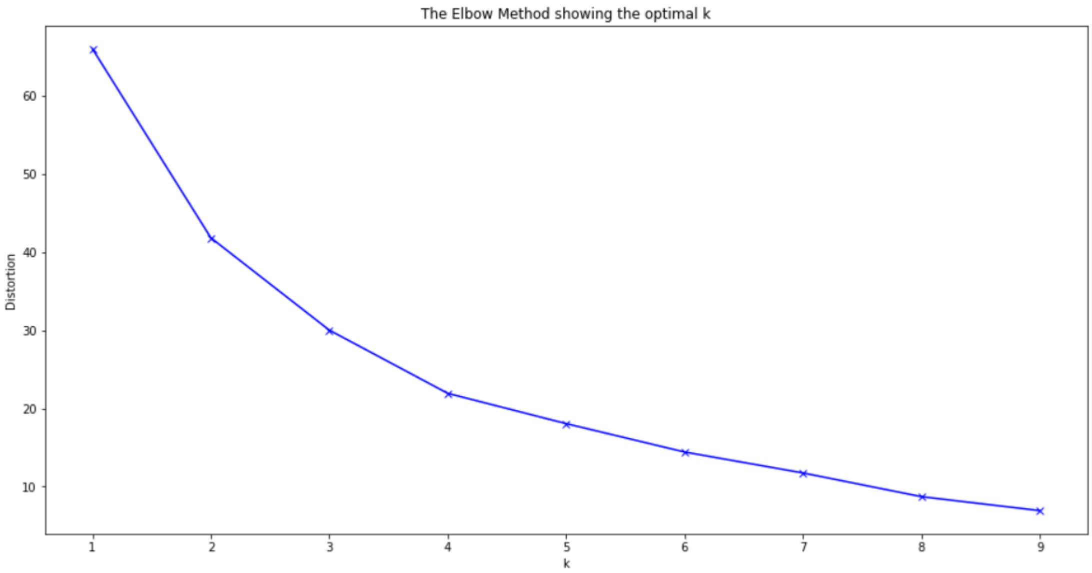


Figure 5: Elbow method for optimal k mean

Clusters labels (0, 1, 2 and 3) are saved in the features data-frame columns and used to create a map (folium) centered in Sao Paulo city. Markers are showed for each district and colored by cluster label.

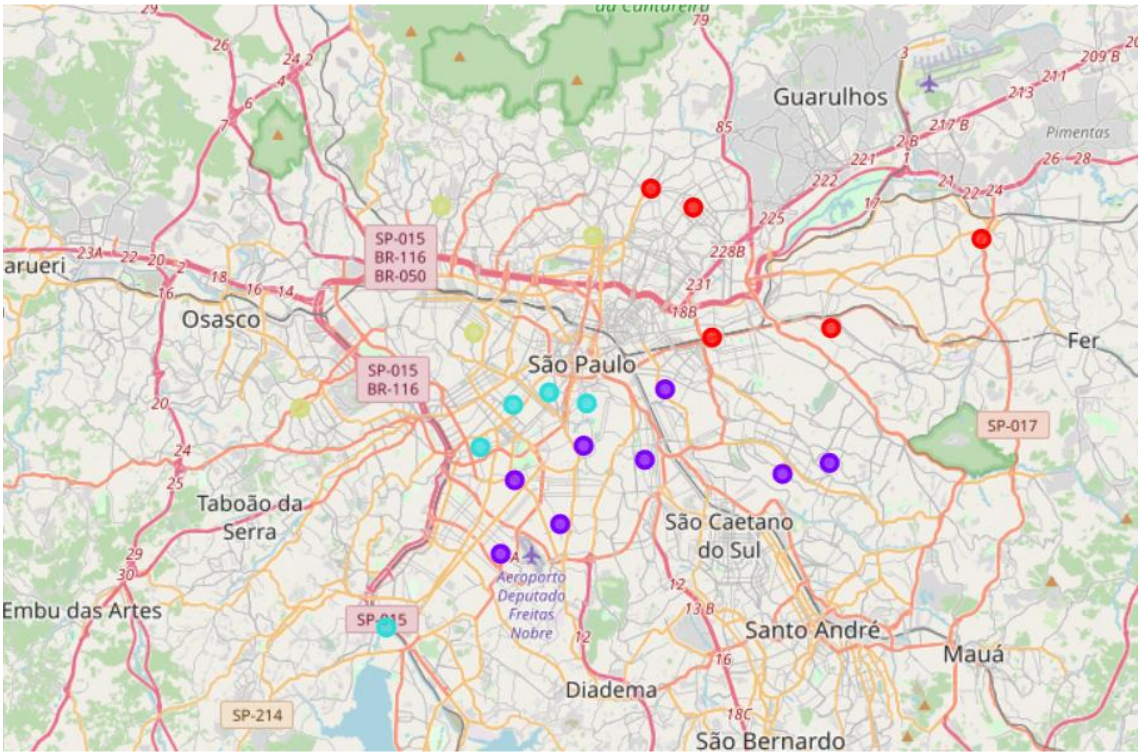
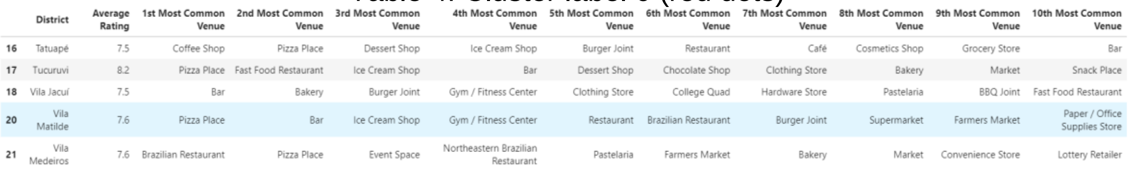


Figure 6: Four clusters for Pet Stores in Sao Paulo city

Districts assigned to cluster label 0 are red-colored in the map. One potential place for Pet Store is in this cluster: **Vila Medeiros**. By evaluating the table below, Vila Medeiros has many grocery stores, bakeries and bars. Pet Store is still no widely spread in this region. **Considering the population and scarcity of Pet Stores in this area, a business person should seriously consider a store in this region.**

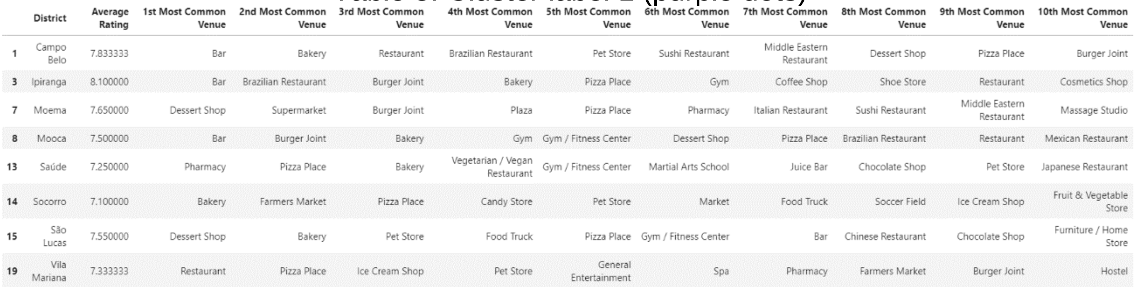
Table 4: Cluster label 0 (red dots)



	District	Average Rating	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
16	Tatuapé	7.5	Coffee Shop	Pizza Place	Dessert Shop	Ice Cream Shop	Burger Joint	Restaurant	Café	Cosmetics Shop	Grocery Store	Bar
17	Tucuruvi	8.2	Pizza Place	Fast Food Restaurant	Ice Cream Shop	Bar	Dessert Shop	Chocolate Shop	Clothing Store	Bakery	Market	Snack Place
18	Vila Jacuí	7.5	Bar	Bakery	Burger Joint	Gym / Fitness Center	Clothing Store	College Quad	Hardware Store	Pastelaria	BBQ Joint	Fast Food Restaurant
20	Vila Matilde	7.6	Pizza Place	Bar	Ice Cream Shop	Gym / Fitness Center	Restaurant	Brazilian Restaurant	Burger Joint	Supermarket	Farmers Market	Paper / Office Supplies Store
21	Vila Medeiros	7.6	Brazilian Restaurant	Pizza Place	Event Space	Northeastern Brazilian Restaurant	Pastelaria	Farmers Market	Bakery	Market	Convenience Store	Lottery Retailer

On the other hand, cluster label 1 (purple dots) was assigned to districts whose average ratings are about average and Pet Store can be found more often. These districts concentrate on the south side of Sao Paulo, which hosts the city's upper-middle class. Additionally, Campo Belo and Vila Mariana, places with the highest number of Pet Stores are located in this region. **Finally, considering the fragmented nature of this region (which should create different strategies for customers) I would avoid this region.**

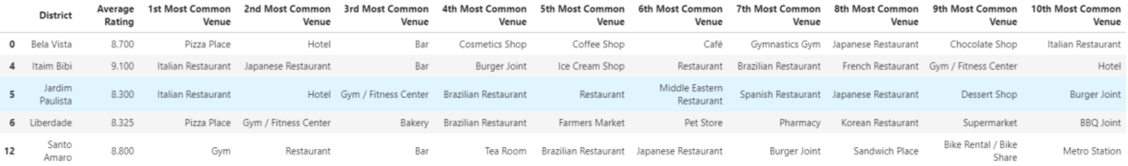
Table 5: Cluster label 1 (purple dots)



	District	Average Rating	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
1	Campo Belo	7.833333	Bar	Bakery	Restaurant	Brazilian Restaurant	Pet Store	Sushi Restaurant	Middle Eastern Restaurant	Dessert Shop	Pizza Place	Burger Joint
3	Ipiranga	8.100000	Bar	Brazilian Restaurant	Burger Joint	Bakery	Pizza Place	Gym	Coffee Shop	Shoe Store	Restaurant	Cosmetics Shop
7	Moema	7.650000	Dessert Shop	Supermarket	Burger Joint	Plaza	Pizza Place	Pharmacy	Italian Restaurant	Sushi Restaurant	Middle Eastern Restaurant	Massage Studio
8	Mooca	7.500000	Bar	Burger Joint	Bakery	Gym	Gym / Fitness Center	Dessert Shop	Pizza Place	Brazilian Restaurant	Restaurant	Mexican Restaurant
13	Saúde	7.250000	Pharmacy	Pizza Place	Bakery	Vegetarian / Vegan Restaurant	Gym / Fitness Center	Martial Arts School	Juice Bar	Chocolate Shop	Pet Store	Japanese Restaurant
14	Socorro	7.100000	Bakery	Farmers Market	Pizza Place	Candy Store	Pet Store	Market	Food Truck	Soccer Field	Ice Cream Shop	Fruit & Vegetable Store
15	São Lucas	7.550000	Dessert Shop	Bakery	Pet Store	Food Truck	Pizza Place	Gym / Fitness Center	Bar	Chinese Restaurant	Chocolate Shop	Furniture / Home Store
19	Vila Mariana	7.333333	Restaurant	Pizza Place	Ice Cream Shop	Pet Store	General Entertainment	Spa	Pharmacy	Farmers Market	Burger Joint	Hostel

Cluster label 2 (blue dots) was assigned to districts whose average ratings are the highest in the clustering. The values range from 8.3 to 9.1, with a mean value of 8.64. Unlike previous clusters, Pet Stores are top rated. This region is located in the southwest part of Sao Paulo, which hosts the city’s wealthiest neighborhoods. **Most venues are restaurants and Pet Stores are scarce in this region, for this reason, it should be considered a potential place for this study.** A more detailed study based on demographics and purchase power could indicate whether Cluster label 0 or label 2 is the most indicated place for Pet Stores, for example.

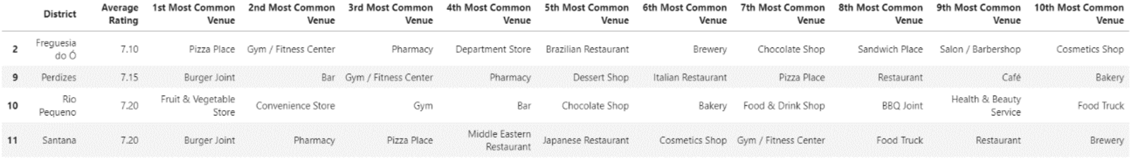
Table 6: Cluster label 2 (blue dots)



	District	Average Rating	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	Bela Vista	8.700	Pizza Place	Hotel	Bar	Cosmetics Shop	Coffee Shop	Café	Gymnastics Gym	Japanese Restaurant	Chocolate Shop	Italian Restaurant
4	Itaim Bibi	9.100	Italian Restaurant	Japanese Restaurant	Bar	Burger Joint	Ice Cream Shop	Restaurant	Brazilian Restaurant	French Restaurant	Gym / Fitness Center	Hotel
5	Jardim Paulista	8.300	Italian Restaurant	Hotel	Gym / Fitness Center	Brazilian Restaurant	Restaurant	Middle Eastern Restaurant	Spanish Restaurant	Japanese Restaurant	Dessert Shop	Burger Joint
6	Liberdade	8.325	Pizza Place	Gym / Fitness Center	Bakery	Brazilian Restaurant	Farmers Market	Pet Store	Pharmacy	Korean Restaurant	Supermarket	BBQ Joint
12	Santo Amaro	8.800	Gym	Restaurant	Bar	Tea Room	Brazilian Restaurant	Japanese Restaurant	Burger Joint	Sandwich Place	Bike Rental / Bike Share	Metro Station

Cluster label 3 (yellow dots) is present on the map. One potential place for Pet Store is in this cluster: **Freguesia do O**. Also, most venues are restaurants and Pet Stores are scarce in this region; however, as mentioned before, a more detailed study based not only on demographics but also purchase power could indicate the most indicated place for Pet Stores.

Table 7: Cluster label 3 (yellow dots)



	District	Average Rating	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
2	Freguesia do Ó	7.10	Pizza Place	Gym / Fitness Center	Pharmacy	Department Store	Brazilian Restaurant	Brewery	Chocolate Shop	Sandwich Place	Salon / Barbershop	Cosmetics Shop
9	Perdizes	7.15	Burger Joint	Bar	Gym / Fitness Center	Pharmacy	Dessert Shop	Italian Restaurant	Pizza Place	Restaurant	Café	Bakery
10	Rio Pequeno	7.20	Fruit & Vegetable Store	Convenience Store	Gym	Bar	Chocolate Shop	Bakery	Food & Drink Shop	BBQ Joint	Health & Beauty Service	Food Truck
11	Santana	7.20	Burger Joint	Pharmacy	Pizza Place	Middle Eastern Restaurant	Japanese Restaurant	Cosmetics Shop	Gym / Fitness Center	Food Truck	Restaurant	Brewery

5 - Conclusion

Sao Paulo is a large city and one of the places in the southern hemisphere. The number of Pet Stores found in each district is not uniform with most of them hosting only one. This type of business is concentrated in three districts (Campo Belo, Liberdade and Vila Mariana).

This project considered the population in each district as indicative of the potential need for Pet Stores (more people, more pet animals). However, I ignored other factors that may affect a business's success, such as pet store size and market share, price range across stores, purchase power by region, etc, due to a lack of available data. Thus this analysis gives only a broad view about the matter and does not aim to exhaustive or detailed.

This is an example of how Data Science is a valuable tool for decision-making in daily life.

Code:

https://github.com/sommersut/Coursera_Capstone/blob/main/Final_Capstone/Pet%20Store%20Battle_of_Neighborhoods-Final.ipynb