IBM Data Science Capstone Project- New Apartment

Andre Saito Guerreiro

I. Introduction

A. Background

The pandemic has caused several changes to our society. One of the effects in the city of São Paulo was a sharp increase in the demand for rental properties. According to [1] during the third trimester of 2020, there was an 18% increase in the number of searches in a popular real state website compared to the same period in 2019. Article [2] states that the number of officialized real state transactions was 336,968, a 37% increase compared to 10 years prior.

One of the many people that moved during that period is my girlfriend. The sudden change to remote work and of her daily habits, along with a loss of income in her family influenced her choice of moving back to her parent household. However, now that the pandemic is (hopefully) nearing its end and the prospect of her job requiring her to return to her office, she is looking for a new place to live.

B. Problem

In this project, we intend to use the skills acquired in data scrapping, data wrangling, and machine learning to aid us in the search for a new apartment.

II. Data sources

In this project, we will use data from three different sources.

A. Quinto andar

Quinto Andar is one of the most famous real state websites in São Paulo. To obtain the real state information, we access Quinto Andar's API and perform inquiries. We search for real state in a square area with

the two of the edges diagonal to each other located in the following coordinates: (-23.478476683596345,-46.76044092347318)(-23.692276596008664,-46.59753427723765). This square covers most of the city of São Paulo. The data from 4317 apartments were obtained. The result was then filtered considered the following criteria. The data presented in this report were obtained on July 24, 2021.

- Total rental price (Rent + condominium fee + taxes) larger than 1000,00 BRL and lower than 2800,00 BRL.
- Total number of bedrooms larger than 2.
- Number of parking spots larger than 1.
- Close to train or metro stations.
- Accepts pets.
- Currently available.

After filtering the total number of apartments was 281. The criteria were chosen by the client (i.e. Girlfriend).

B. Foursquare

Foursquare is an app, which is purpose is to clients discover and share information about businesses and attractions around designated locations. From Foursquare we obtained data on the venues surrounding each of the remaining apartments after filtering. For each inquiry, it was considered an area with a radius of 800 meters. Although a myriad of information for each venue is provided, only the venue's category was stored. The frequency of the venue's categories is used to aid us in clustering similar locations into clusters. We also add data concerning venues that surround two places that the client has lived before and liked. We call these locations **test locations**. They are being used to choose one cluster after clustering is performed.

C. Google Maps

Similar to other large cities in the world, São Paulo has problems with traffic. Therefore transit times from and to work greatly impact the quality of life of their citizens. Using Google Maps "Distance Matrix" API, we calculate the transit time using public transportation from each of the apartments to the client's workplace. This transit time is calculated considering the arrival time at 9 am of the next Monday, counting from the time the code was run. For the data presented in this report, it was considered transit times estimated for July 26, 2021.

III. Methodology

In this project, we intend to evaluate apartments in São Paulo according to several characteristics. The goal is to have a shortlist of 5 to 20 apartments that would fit our client's needs. It is important to not excessively filter candidate apartments, leading to a small number of choices. There are important properties of the apartment that cannot be easily evaluated through the use of algorithms, such as personal taste. We also cannot be too loose on our filtering, leading to too many choices the client would have to go through manually.

The first step of the process has already been performed when the data was obtained, by filtering considering several characteristics of the apartment.

The second step consists of finding apartments in locations that are similar to places the client has lived before and approved (the test locations). This is done by a K-means clustering algorithm. We use the data obtained from Foursquare, detailed in II-B. The K-means algorithm separates the apartments into mutually exclusive clusters. We then choose the cluster that contains both test locations. Although the problem could be modeled as a classification problem, due to the small number of test locations, we decided to use clustering instead.

Our first attempt to determine the number of clusters is to analyze the mean squared error between the centroids of the clusters to the elements in its cluster for a different number of clusters. Called the "elbow method", the inflection point of the curve is chosen as the number of clusters. Figure 1 presents the results. It is possible to observe from 1 that the curve does not have a clear inflection point. Therefore we use a different method to determine the number of clusters. Based on the client's preferences, we know that the two test locations should be on the same cluster. We also want the clustering to be selective, as the number of apartments at this point is still high. Therefore, after iterating on the number of clusters, we chose the highest number of clusters that abides by the rule stated above, which was a total of 7 clusters.

The final step is filtering the apartments in which the transit time to the client's workplace is acceptable (60 minutes).

IV. Results and Discussions

A. Filtering

In Figure 2 we show the location of the apartments after filtering. It is possible to observe that there

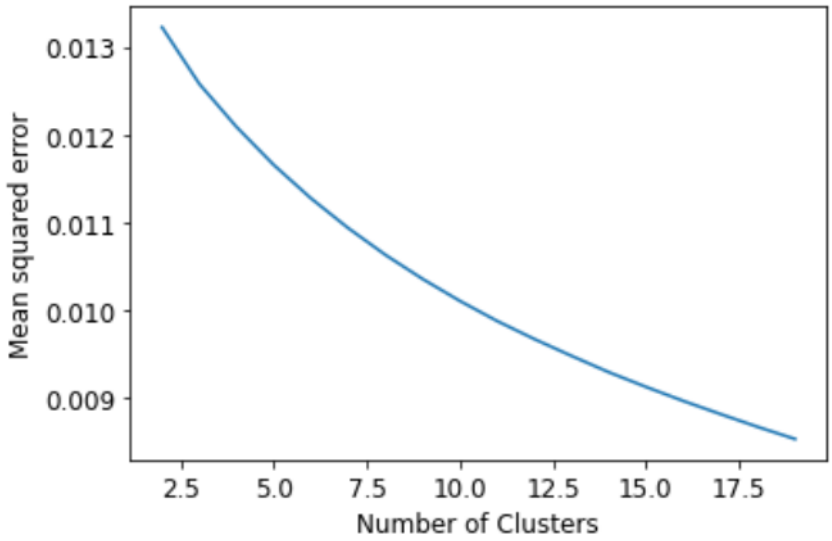


Fig. 1. Men squared error vs Number of clusters. In this case, no clear inflection point is observed. Therefore we cannot use the elbow method to determine the number of clusters to be used in the K-means algorithm.

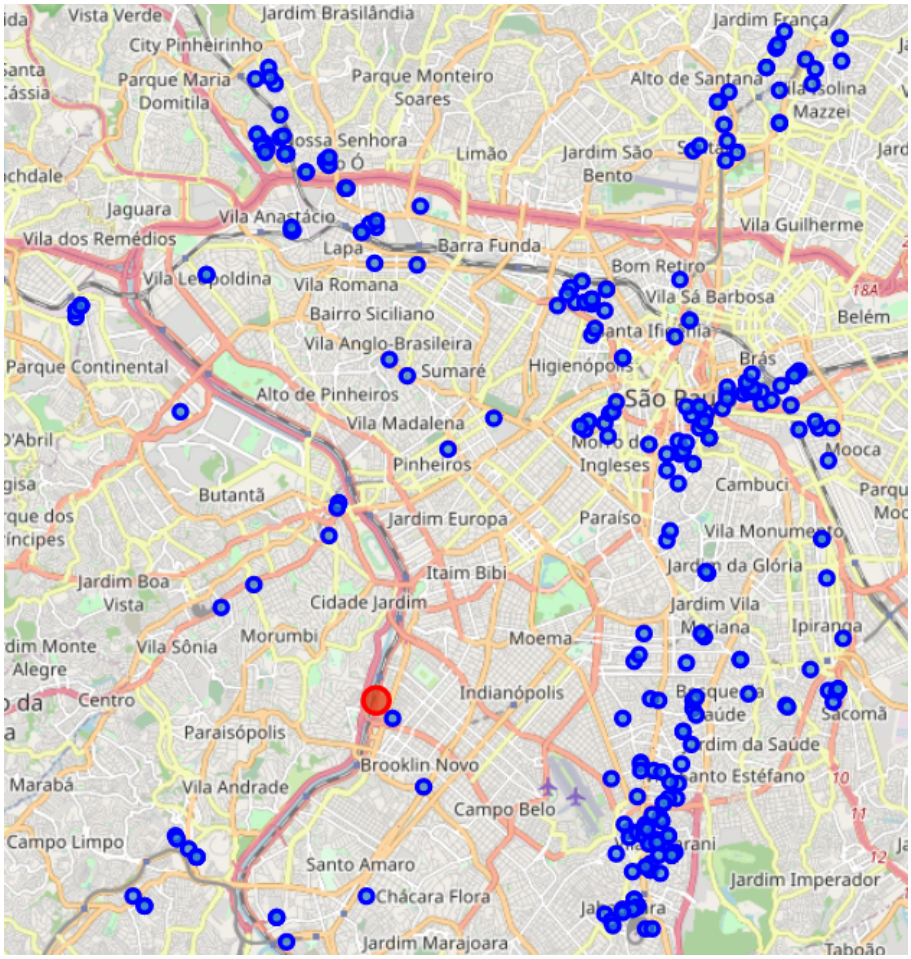


Fig. 2. Map of São Paulo with the location of the apartments after the first filter. The blue dot denotes the apartments and the red dots the approximate location of the client's workplace.

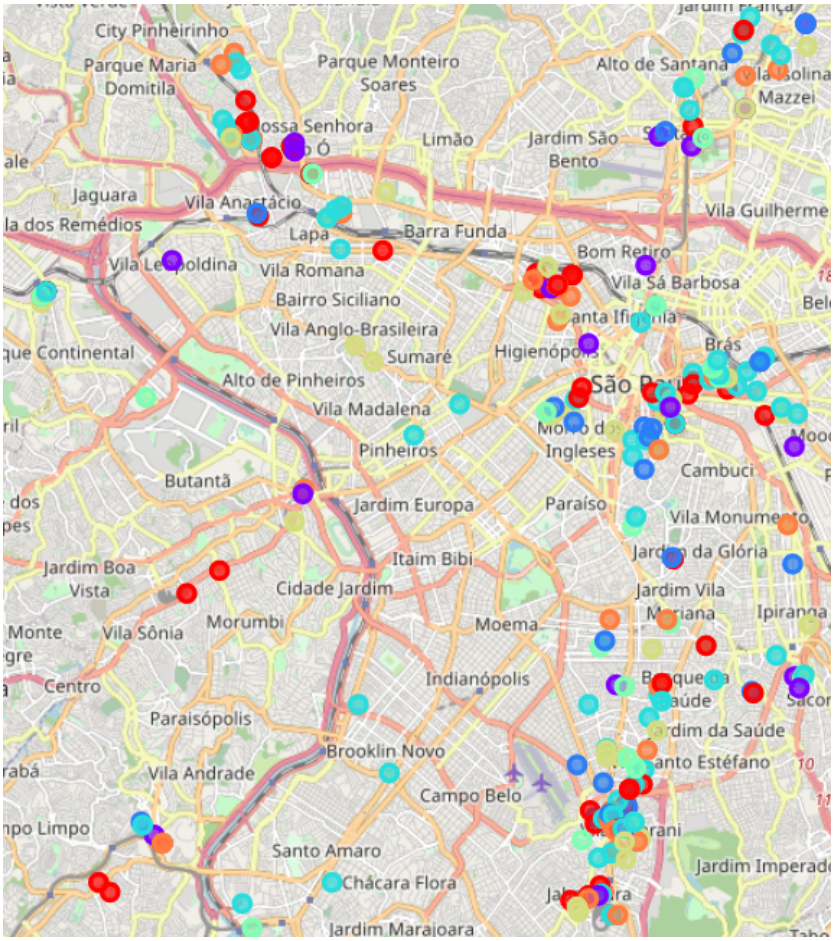


Fig. 3. Map of São Paulo with the location of the apartments after clustering. Each color represents one of the 7 clusters.

is a small number of apartments near the client's workplace. This is due to the high price of real estate in that region, which in most cases goes above the 2800 BRL limit.

B. Clustering

Figure 3 presents the results of the clustering with K-means. The colors denote the cluster of each of the apartments. An interesting insight from figure 3 is that there is a low correlation between the apartment location and its cluster. Apartments closed to each other may belong to different clusters, which reflects the diverse nature of the city.

C. Transit data

Figure 4 presents a histogram of the transit time. It shows that there are very few apartments that satisfies the client's criteria closer than 40 minutes to her workplace, which was expected given the high value of properties near it. The transit time threshold chosen by the client is close to the median of the

```
The transit time median is 64.5 minutes.
```

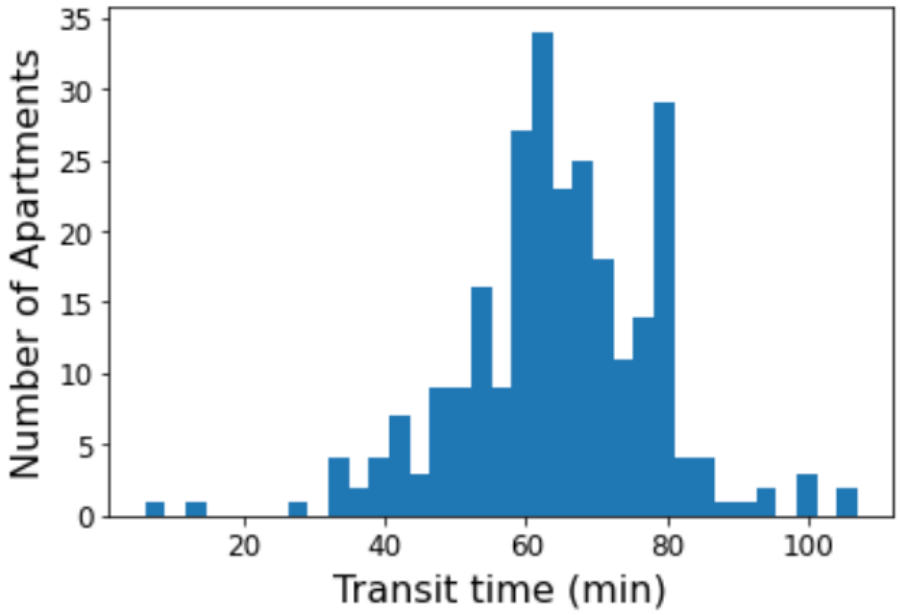
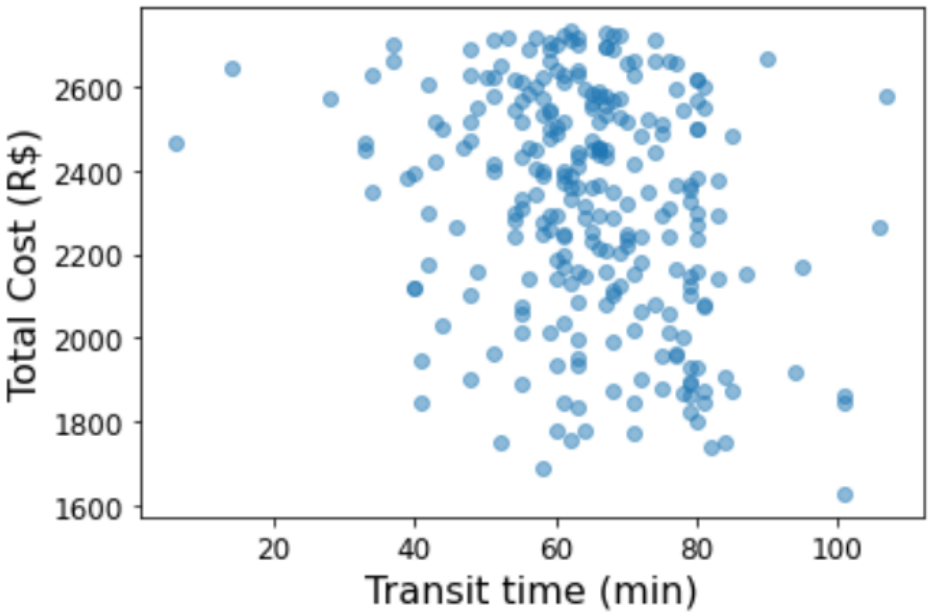


Fig. 4. Histogram of the transit time.




```
R square is 0.06804674150324908
```

Fig. 5. Histogram of the transit time.

transit time, showing that the client already had a good intuitive idea of the apartments she could afford. Finally, Figure 5 presents a scatter plot of the transit time vs the total cost. It shows that from 40 minutes of transit time and on, there is almost no correlation between the transit time and the total cost.

Website_Id	Lat	Lng	Bedrooms	Parking_Slots	Total_Value	Total_Area	Floor	Neighborhood	Status	Traffic_time(min)
101409	-23.608980	-46.630734	2	1	2717.0	57.0	6	Bosque da Saúde	publicado	57
119491	-23.524734	-46.697685	2	1	2690.0	90.0	8	Vila Romana	publicado	56
257592	-23.586649	-46.723164	2	1	2448.0	60.0	10	Butantã	publicado	33
291088	-23.601331	-46.643220	2	1	2575.0	72.0	13	Vila Clementino	publicado	51
335416	-23.607796	-46.619220	2	1	2448.0	70.0	7	Vila Gumercindo	publicado	57
518304	-23.608234	-46.630049	2	1	2454.0	47.0	3	Bosque da Saúde	publicado	56
615232	-23.622575	-46.641706	2	1	2538.0	65.0	8	Saúde	publicado	59
625436	-23.571916	-46.705593	2	1	2628.0	68.0	7	Butantã	publicado	34
627613	-23.600038	-46.642174	2	1	2689.0	70.0	1	Vila Clementino	publicado	48
632196	-23.526911	-46.733039	2	1	2653.0	75.0	18	Vila Leopoldina	publicado	52
632693	-23.603011	-46.606042	2	1	2400.0	67.0	7	Vila Nair	publicado	58
635542	-23.596033	-46.629236	2	1	2416.0	60.0	4	Vila Mariana	publicado	51
643807	-23.596178	-46.641170	2	1	2609.0	60.0	3	Vila Mariana	publicado	55
662260	-23.612393	-46.645429	2	1	2710.0	72.0	3	Planalto Paulista	publicado	51
671551	-23.591110	-46.729930	2	1	2262.0	60.0	14	Vila Sônia	publicado	46
83787	-23.625629	-46.687598	2	1	2646.0	85.0	3	Jardim das Acácias	publicado	14

Fig. 6. Characteristics of the apartments chosen by the algorithm.

D. Final results

After all of the filters, the algorithm selected 16 apartments out of the initial 4317. Figure 6 presents the characteristics of these apartments. Figure 7 presents the top 5 most frequent venue categories surrounding each of the chosen apartments. We can clearly see a trend. All of the locations are mostly surrounded by restaurants, cafes and Gyms/Fitness centers. These characteristics are very similar to the characteristics of our client's test locations (GF1 and GF2).

V. Conclusion

The purpose of this project was to identify São Paulo's apartments in order to aid the client in narrowing down the search for the optimal apartment for her new home. This was done in 3 steps. The first was to filter the apartments based on their characteristics based on the client's wishes. Secondly, a clustering algorithm was used to select apartments in locations that are similar to places that the client has lived

	Website_Id	1	2	3	4	5	labels
0	101409	Fruit & Vegetable Store	Gym / Fitness Center	Pizza Place	Burger Joint	Bar	3
2	119491	Bakery	Gym / Fitness Center	Pharmacy	Restaurant	BBQ Joint	3
28	257592	Pharmacy	Pizza Place	Gym / Fitness Center	Coffee Shop	Brazilian Restaurant	3
35	291088	Gym / Fitness Center	Brazilian Restaurant	Pet Store	Burger Joint	Dessert Shop	3
45	335416	Bakery	Gym	Gym / Fitness Center	Pharmacy	Dessert Shop	3
83	518304	Gym / Fitness Center	Japanese Restaurant	Fruit & Vegetable Store	Burger Joint	Dance Studio	3
131	615232	Bakery	Pizza Place	Brazilian Restaurant	Pharmacy	Gym / Fitness Center	3
146	625436	Bar	Café	Brazilian Restaurant	Chocolate Shop	Ice Cream Shop	3
152	627613	Bar	Coffee Shop	Brazilian Restaurant	Café	Pet Store	3
160	632196	Dessert Shop	Brazilian Restaurant	Japanese Restaurant	Gym / Fitness Center	Burger Joint	3
161	632693	Pizza Place	Gym / Fitness Center	Optical Shop	Bar	Pet Store	3
165	635542	Burger Joint	Pet Store	Pizza Place	Pharmacy	BBQ Joint	3
179	643807	Bar	Coffee Shop	Brazilian Restaurant	Dessert Shop	Pizza Place	3
223	662260	Cosmetics Shop	Bakery	Pizza Place	Dessert Shop	Pharmacy	3
245	671551	Pharmacy	Pizza Place	Bakery	Japanese Restaurant	Grocery Store	3
257	83787	Gym	Salon / Barbershop	Mineiro Restaurant	Supermarket	Italian Restaurant	3
264	GF0	Dessert Shop	Cocktail Bar	Restaurant	Hostel	Spa	3
265	GF1	Cosmetics Shop	Dessert Shop	Gym / Fitness Center	Pizza Place	Bar	3

Fig. 7. Top 5 most frequent venue categories surrounding each of the candidate apartments.

before. Finally, it was analyzed the transit time from each candidate apartment to the client's workplace. The final result was a shortlist of 16 candidates out of over 4200 apartments.

The algorithm was successful in generating a list of candidates that have the desired requirements defined by the client are in locations with similar characteristics as other places the client has lived before and have acceptable transit times to her workplace. The final decision on the optimal apartment will be made by the client's preferences.

REFERENCES

[1] M. Schlindwein, "Efeito covid: aumenta a procura por aluguel de imóveis no país: Radar," Nov 2020. [Online]. Available: https://veja.abril.com.br/blog/radar/demanda-por-alugueis-em-alta-no-terceiro-trimestre/

[2] E. Veiga, "'da cidade para praia': os paulistanos que aproveitaram pandemia de covid para mudar de casa," Jun 2021. [Online]. Available: https://www.bbc.com/portuguese/brasil-57590323