IBM Data Science Certificate Final Report

1. Introduction

Imagine you are a person who needs to move to a city you don't know anything about it. A big one, like New York or London. You have to find a place to stay, a place that fit your personal budget, of course, but also a place where you feel confortable and that has some interesting and helpful locations around it, like a grocery store or a park to relax in weekends. When the time comes, wouldn't be nice if you're able to find this kind of information at once, without having to search for it in many different locations? Or, if you're a person who works in house sales/rent field, wouldn't be nice if you're the person who delivers such information to a custumer?

These were the two main questions that leaded me to propose a business problem and execute some data science tasks to try to solve it. All processes and results are discussed in this report.

1.1 Business Problem

There is a fact that some characteristics in a house or apartment has directly influence in their prices of rent or sale. For people who is moving to a new and unknow city - São Paulo, for instance - this is just one of many other concerns they can have at the moment: São Paulo is the financial center of Brazil[1] and is the most populated city in the country - and in south hemisfery at all[2], wich more than 12 million people[3] living there. This can be a quite challenge for some people.

So, to help them, this project aimed to **segment São Paulo apartments avaliable for rent** according to some characteristics – more specifically: size, price of rent and price of condominuim. Among with that, some information about each districts location was provided by Foursquare to give them a better ideia about what kind of places they can expect to find around these apartments.

With this information in hand, they can find the best place to live according to the amount of money they can spend on a rent and to some characteristics of the districts. Companhies and stakeholders working in this field can also use these information to make better decisions about market and client segmentation.

1.2 Data sources

Two data sources were used in this project. One of them was Kaggle plataform, wich provided **data about apartments avaliable in São Paulo**. The complete dataset covered apartments for rent and sale in this city in April 2019 and presented data about size, price, condominium, number of rooms, toilets and suites, if they were already furnished and if it was the first time they were avaliable to rent. It also presented some data about the building, like latitude and longitude locations, district and if they have elevator, parking place and swimming pool.

The other data source was Foursquare, wich provided what they call venues, **places of interest, around each district**. To get these venues, I've used the Foursquare API. In order to maintain the consistence with my primary data, the API version used was April 2019, according to

[1] *PIMENTA, Angela. Esqueça os países. O poder está com as cidades. Revista Exame. Nov 2, 2007.*

[2] *INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Estimativas da população residente nos municípios brasileiros com data referência em 1º de julho de 2018. Aug 29, 2018.*

[3] *INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). São Paulo. Oct 6, 2019.*

documentation. But, first, it was also necessary to get oficial **latitude and longitude locations of each district**, using Geocoder Python library.

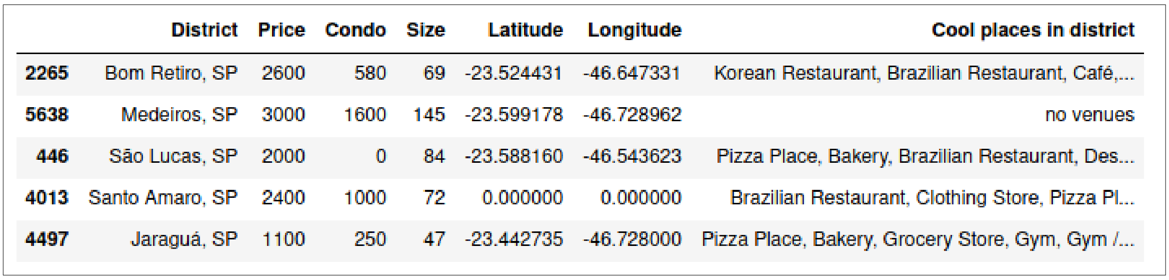
1.3 Data cleaning and feature selection

The **primary data** about apartments avaliable were very consistent and didn't present any null or missing values. After checking them, I've made some small adjusts in columns names and values and selected only the rows and columns relevant for my problem: **apartments avaliable for rent**.

To get **Foursquare venues**, I've used every district name (from primary dataset) and geocoordenates in a API search query, with a radius default value of 1km (1000m) and venues results limited by 200. This **districts geocoordenates** were provided by Geocoder Python Library.

So, my **final dataset** presented information about apartments sizes, prices of rent and condominium, latitude and longitude, district location and the list of venues categories from respective district. Some of them had no venues assigned (because some districts didn't retrieve any venues), so I've filled this blank data with a message for the public, as you can see in Figure 1. This final dataset has 7228 rows and 7 columns.

Figure 1 - Final dataset sample



	District	Price	Condo	Size	Latitude	Longitude	Cool places in district
2265	Bom Retiro, SP	2600	580	69	-23.524431	-46.647331	Korean Restaurant, Brazilian Restaurant, Café,...
5638	Medeiros, SP	3000	1600	145	-23.599178	-46.728962	no venues
446	São Lucas, SP	2000	0	84	-23.588160	-46.543623	Pizza Place, Bakery, Brazilian Restaurant, Des...
4013	Santo Amaro, SP	2400	1000	72	0.000000	0.000000	Brazilian Restaurant, Clothing Store, Pizza Pl...
4497	Jaraguá, SP	1100	250	47	-23.442735	-46.728000	Pizza Place, Bakery, Grocery Store, Gym, Gym /...

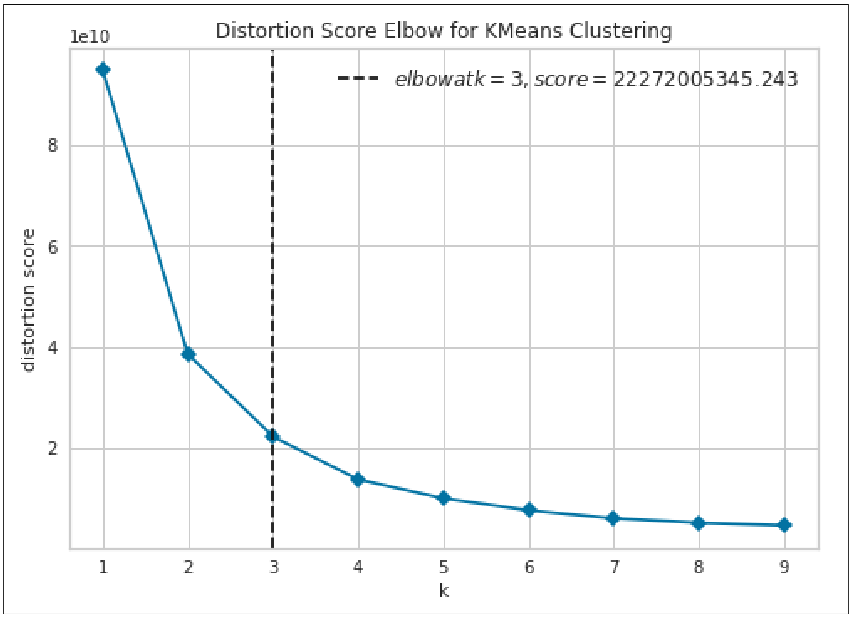
As **features** for machine learning algorithm, I've selected only sizes, price of rent and price of condominium. The other information were important for map visualization and cluster analysis.

2. Segmenting apartments with clustering

Clustering is a non-supervised machine learning algorithm. It's frenquently used for client and market segmentation. Here, I've decided to use it **to segment São Paulo apartments avaliable for rent** according to the features selected above in order to offer future renters and stakeholders in this field better information about the options they can have based on client personal budget.

As a non-supervised task, clustering demands you to find "manually" the best possible number of clusters for your situation. To do that in my case, I've used the Elbow Point method, as shown in Figure 2.

Figure 2 – Elbow Method showing the optimal k



The main idea of Elbow method is understand how your model behaves in terms of distortion every time you increase k. Distortion “computes the sum of squared distances from each point to its assigned center”[4] wich means that we need to choose the k value that shows a good decrease of this distance when compared with the next one. Looking the figure above, **k equals 3** seemed to be the best choice for my problem.

After apply the task again, now with the best number of clusters, I was finally able to se what kind of groups my data could form. The results are discussed bellow.

3. Results

As described above, my clustering features involved three numerical categories about prices and sizes of apartments avaliable for rent. After applying the task, I’ve found some interesting **characteristics about each cluster**. These characteristics involve price of rent and condominium ranges, range and mean apartment sizes and number of apartments avaliable per cluster. To build this table I needed to find the minimum and maximum values for each feature per cluster, as well as the mean size of apartments per cluster and the size of the cluster itself. It’s important to note that I haven’t changed clusters order when enumerating them. Table 1, bellow, presents all this gattered information.

Table 1 – Clusters characteristics

	Prices range (R$)	Condominium range (R$)	Size range (m²)	Mean Size (m²)	Number of apartments
Cluster 1	480 to 4,800	0 to 7,500	30 to 400	72	6108
Cluster 2	13,500 to 50,000	0 to 8,800	58 to 880	285	169
Cluster 3	4,500 to 13,350	0 to 9,500	30 to 852	168	951

[4] ELBOW METHOD. Avaliable at https://www.scikit-yb.org/en/latest/api/cluster/elbow.html. Feb 11, 2020.

Based on this, I've named each cluster and made some abstractions in order to describe them, as follows. A **map visualization** of all apartments avaliable can also be seen bellow, in Figure 3, when each point represents an apartment for rent with the correspondent collor of the cluster it belongs.

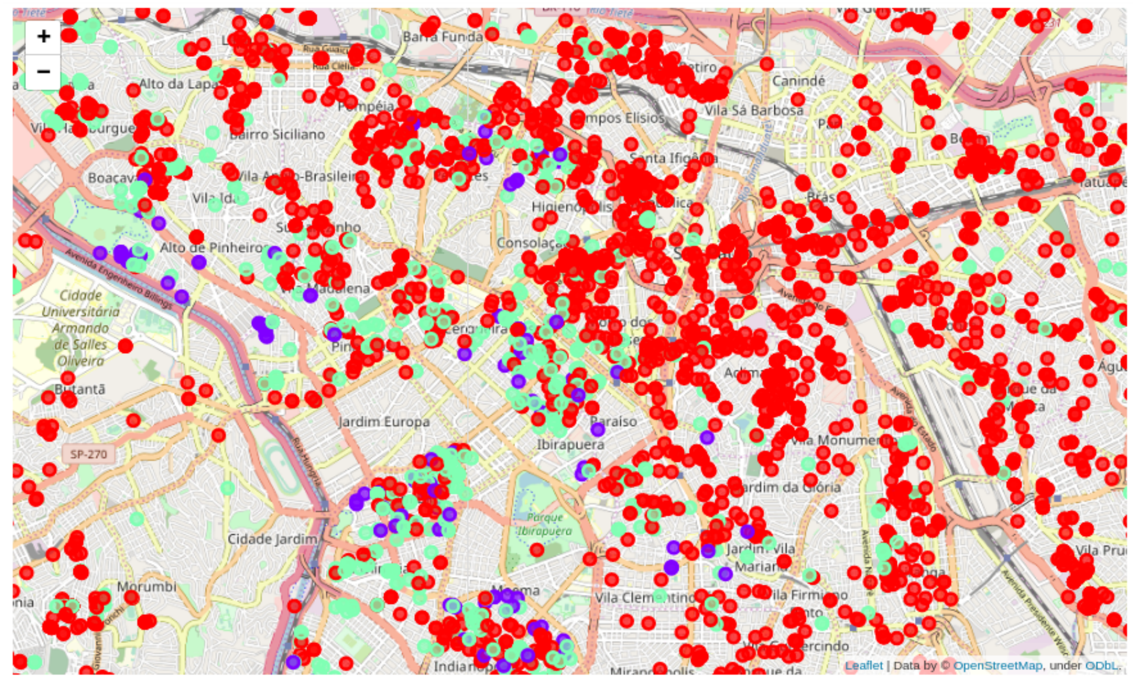
- Cluster 1 - **Small sizes and prices:** this cluster is the bigger one and have the smaller sizes and rent/condominium prices of apartments. It's probably the first choice for most part of future renters, specialy if they pretend to live alone or have small families. It's marked in red collor on the map.

- Cluster 2 – **A+ class:** this is the most expensive and exclusive cluster, with feel huge and really expensive apartments avaliable for rent. It's marked in light green collor on the map.

- Cluster 3 – **Big families:** with high costs and apartments sizes, future renters in this middle area must have a very good personal budget. It can be a good choice for people with many roommates or big families. It's marked in purple collor on the map.

We can see from the map bellow that all places, big or small, cheap or expensive, are distributed into all districts, wich is expected since my clustering task didn't include location as a feature. So, it's important to note and remember that **most part of districts has apartments avaliable from diferent clusters**, wich means that they can suit diferent tipe of personal budgets and necessities.

Figure 3 – Cluster visualization on São Paulo map



As for the **venues**, the main idea was just list them in a good way to present future renters and stakeholders some information about what kind of places they can find in each district. For instance, if a person choose an aparment in Bom Retiro he or she would find Korean and Brazilian restaurants, coffee shops and other kind of venues listed in the final dataset, as shown in Figure 1 above. If you really don’t know anything about the place you’re moving, this kind of information can already help you a lot.

4. Conclusions and future considerations

Moving to a unknow city can be fun, but also an uneasy thing to do. My idea here was segment São Paulo apartments in order to fit personal budgets and provide some other information that can help people to choose and better adapt to your new homes. After applying a clustering task, I’ve found **three main clusters, or groups, of apartments avaliable for rent** that can guide future renters and stakeholders in this field to achieve this. This three clusters consisted in a big one with small places and cheaper prices, an intermediate one with big prices and places and a very small and exclusive one for A+ class. A visualization of these apartments on a São Paulo map was also presented.

As for **future aproaches** to the same kind of problem, it would be very interesting to include venues or other districts characteristics gattered from Foursquare (or other data source), as well as other apartments characteristics – number of rooms, for instance - as features for the clustering tasks. Another ideia for stakeholders specialists in this field would be use a classification model to make a better market analysis based on data about average salaries in São Paulo or other cities in Brazil.