Finding Similar Cities in São Paulo State of Brazil

Luiz Alberto Hiroshi Horita

July 28, 2020

1. Introduction

In the world of entrepreneurship, it is very important to understand your business area and map where the opportunities are. And, in fact, there are countless opportunities of business, but most of time it is not explicitly visible.

Since the rural exodus, which intensified after the great industrial revolution, there is this illusion the best opportunities are at the biggest towns, the capital, with hundreds of thousands of habitants. However, if we take a closer look, we will see there are many products and services that still do not exist, or are scarce, as far as the cities are from these big towns.

It is common, at least in Brazil, to have every kind of establishments in big cities, but miss some categories of establishments in smaller cities, so, they could be great opportunities to endeavor.

Moreover, differently from the past, currently the flow of people is not so intensive to the capital and big towns. Instead, there also people moving from the biggest cities to smaller cities, looking for better quality of life.

In this context, this project's goal is to find similar cities, more specifically in the state of São Paulo (SP) in Brazil, in terms of their population and their most common venue categories.

This kind of analysis could be useful to stakeholders own an establishment and wants to expand it to other similar cities that still have few or no venues of the same category, or to some entrepreneurs who wants to start a new business, or even to someone else who needs to move to another city, but keeping accessibility to venues available at his current city.

2. Data Acquisition and Cleaning

2.1. Data sources

Based on the context of this project, factors that will influence the cities clustering are the list of cities in SP state, the population of each city, the geographical position (latitude and longitude) of each city, and the list of existing venue categories and their quantities in each city.

To acquire these data, we scraped some pages of Wikipedia to collect the list of cities and their respective population, used the GeoPy module from Python to obtain the geographical position of each city, and used the Foursquare API to retrieve the list of venues available in each city, once we know their geographical location.

2.2. Data cleaning

The data acquired from all sources resulted in three dataframes: a dataframe containing all cities of SP state with their respective region of state, registration code, localization; another dataframe containing all cities of Brazil with their respective registration code, ranking in terms of population, state, and population; and the last dataframe containing all venues and their respective latitude, longitude, name, category, and city.

The first dataframe, before mentioned, has data from 643 cities, in which we removed the features region of state, registration code, and localization for being irrelevant to our case. After that, we appended two columns with the geographical position in terms of latitude and longitude.

In the second dataframe, we filtered all cities were not from SP state and removed them, and we removed the features registration code, and ranking in terms of population for being irrelevant to our case. After that, we merged this dataframe with the first one in terms of the city name (see Figure 1).

City	Population	Longitude	Latitude
Aparecida d'Oeste	4196	-50.880871	-20.449811
Aspásia	1822	-50.728046	-20.160028
Dirce Reis	1793	-50.606276	-20.466407

Figure 1 - First three samples of dataframe with list of cities and their respective population and geographical position.

We also removed the row with São Paulo, the capital city, due to its high population value, which characterize it as an outlier (see Figure 2). If we considered São Paulo, the result of clustering would be strongly biased by this high value of population, since this is a feature we considered.

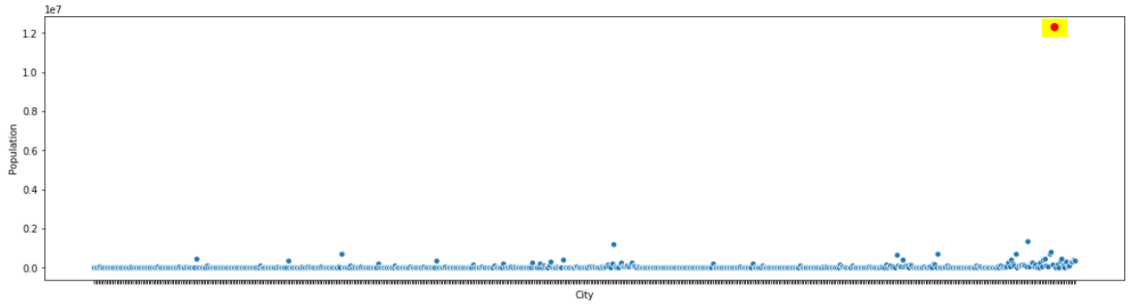

Figure 2 - Plot of all cities' population. São Paulo is an outlier.

In the last dataframe, with 14,757 venues listed, we removed the features name of venue, latitude and longitude for being irrelevant to our case (see Figure 3).

City	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Aparecida d'Oeste	-20.449811	-50.880871	Farmácia do Pedro	-20.451159	-50.881847	Pharmacy
Aparecida d'Oeste	-20.449811	-50.880871	Bar do Fabião	-20.450060	-50.886469	African Restaurant
Dirce Reis	-20.466407	-50.606276	Padaria Doce Pao	-20.464830	-50.606962	Bakery

Figure 3 - First three samples of dataframe with list of venues and their relevant features.

2.3. Feature selection

Since we want to cluster the cities, in this case using K-Means, we need to calculate distances based on some features. However, there are some categorical features are not possible to be processed, so we transformed each of these categories into a binary feature through the one-hot-encoding method.

From this point, we have 392 features. However, there are many redundant categories/features, such as “Art Gallery” and “Art Museum”, or “Bar” and “Beer Bar”. So, based on keywords within each category, we joined redundant features, obtaining in this process a dataframe with 229 features.

When it comes to clustering, we must find the features that best distinguish each cluster. In this context, we saw there were many features with too few occurrences, which would not differentiate cities well, and features with too many occurrences, which also would not differentiate cities, because probably most of them have this kind of establishment.

Following this reason, we removed some features by their number of occurrences. In this case, the number used as threshold for selecting the features were set empirically. With this, in the end, we obtained a dataframe with 23 relevant features that best differentiate the clusters, and 14,757 samples.

3. Exploratory Analysis

3.1. Building cities profile

From the dataframe obtained before, with 23 features (24 columns if counting the cities’ name), we can trace each cities’ profile in terms of mean occurences of each venue categories, by joining samples by cities’name and calculating the average of all features. In this process, we obtained a dataframe with the 24 columns and 623 samples. Note that at the beginning we had 643 cities listed, and now we have 623. This is because in 20 cities there were no venues retrieved from Foursquare API, so it is not possible to compute their profile.

At this point, we have two dataframes, one with population and geographical position of each city, and the other with mean occurences of venue categories in each city. We could consider only the second dataframe to cluster, however, we want to know not only if a city has or not the venue categories, but also the notion of demand versus offer of each of them. For easier understanding, imagine a city has 2 japanese restaurants for a population of 20,000, a second city has 5 japanese restaurants also for

a population of 20,000. If we divide the population for the number of Japanese restaurants we obtain 10,000 for the first city and 4,000 for the second city. From this, we can clearly say the ratio of demand versus offer for this category is higher on the first city than on the second city, and thus, the opportunity of investing on a Japanese food establishment seems better on the first city.

With this in mind, we merged the dataframes of cities' population and geographical position with the dataframe of cities' profile in terms of venue categories occurences, and computed, for each venue category, the reason before explained, dividing the population column by each venue category column.

As result, we obtained the new profile dataframe, but with several cells with infinite values. This outcome was expected, because in almost every city there are venue categories with zero occurrence, so if divide the population by zero the result is infinite.

To treat this, we replaced the infinite values by the double of the highest, non infinite, result. After doing that, we normalized all values in dataframe by column. This process is important to eliminate bias, and reduce the computational power demand.

City	Population	BBQ Joint	Bakery	Bar	Bed & Breakfast	Brazilian Restaurant	Burger Joint	Café	Coffee Shop	...
Adamantina	0.024603	0.02	0.002050	0.001230	1.0	0.003075	0.003075	0.006151	1.0	...
Adolfo	0.001739	0.00	1.000000	1.000000	1.0	1.000000	1.000000	1.000000	1.0	...
Aguaí	0.025500	0.00	0.001211	0.002423	1.0	1.000000	1.000000	1.000000	1.0	...

Figure 4 - First three rows of the new cities' profile dataframe with normalized features.

3.2. Finding best K value

Since the goal of this project is segmenting and clustering of cities, we are dealing with unsupervised machine learning method, so we must explore the best number of clusters for this case. We used the K-Means method, so we explored the results for some different K values, which consists of the number of clusters to be formed.

For this analysis, we clustered the cities varying the K value from 3 to 8, and plotted the distribution curves of each cluster for each K-Means models created (see Figure 5).

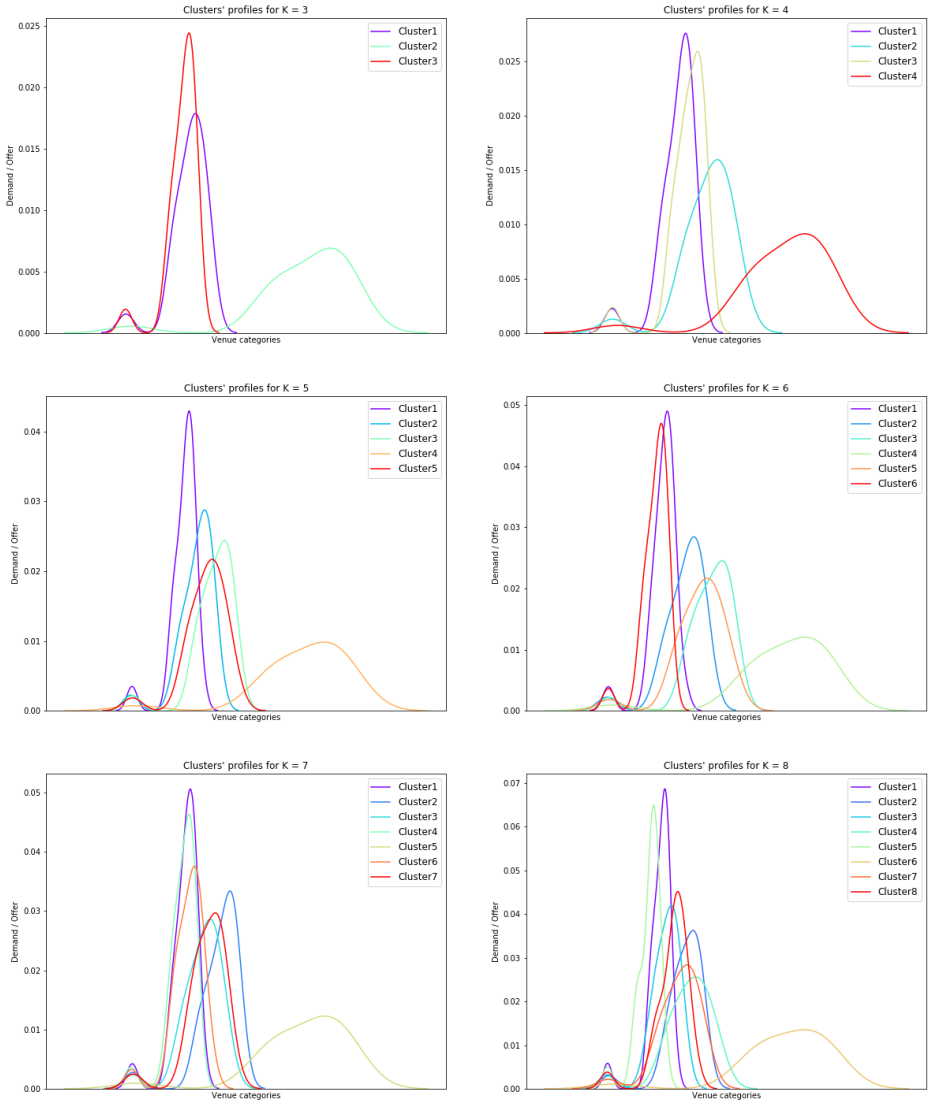

Figure 5 - Distributoin curves of each K-Means models tested.

From this analysis, we can see all the results has intersecting distribution curves, but the result that seems better separate each cluster is for K = 4, i.e., the model with 4 clusters. The other results has almost totally overlapped curves, and, moreover, the more clusters the more agglomerated are the curves.

4. Cluster Modeling

Finally, with the cities' profile dataframe, and the best K value, we modeled the K-Means model to cluster the cities of SP state and plotted the result on the SP state map and the distribution curves of each cluster obtained (see Figures 6 and 7 respectively).

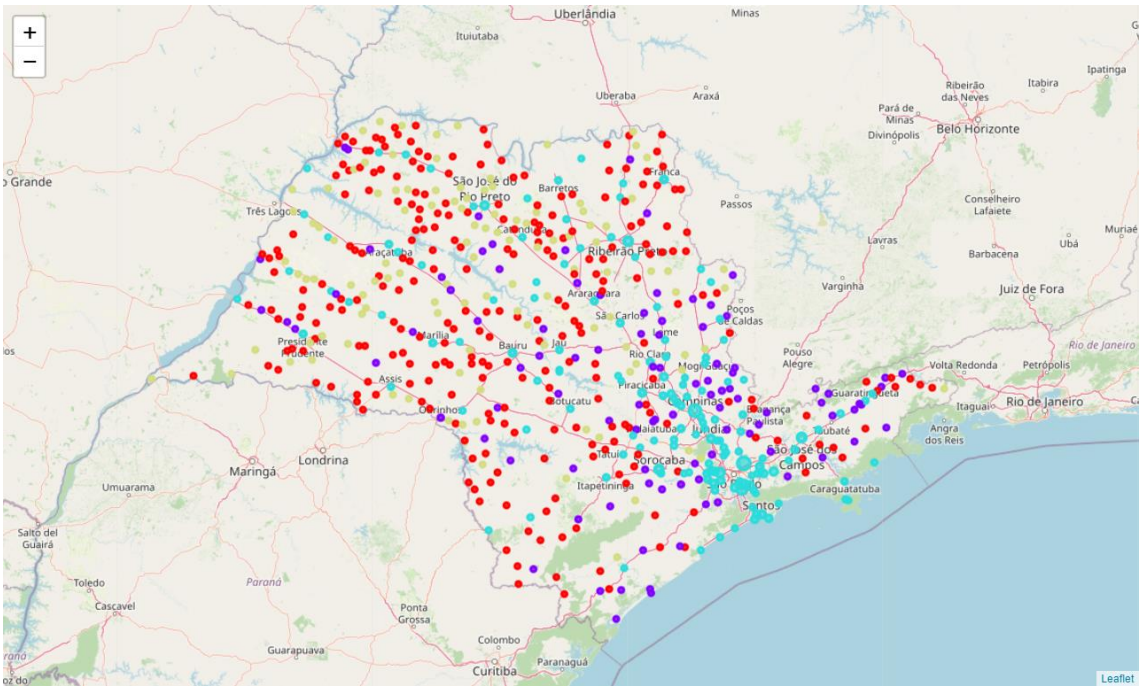

Figure 6 - SP cities clusters.

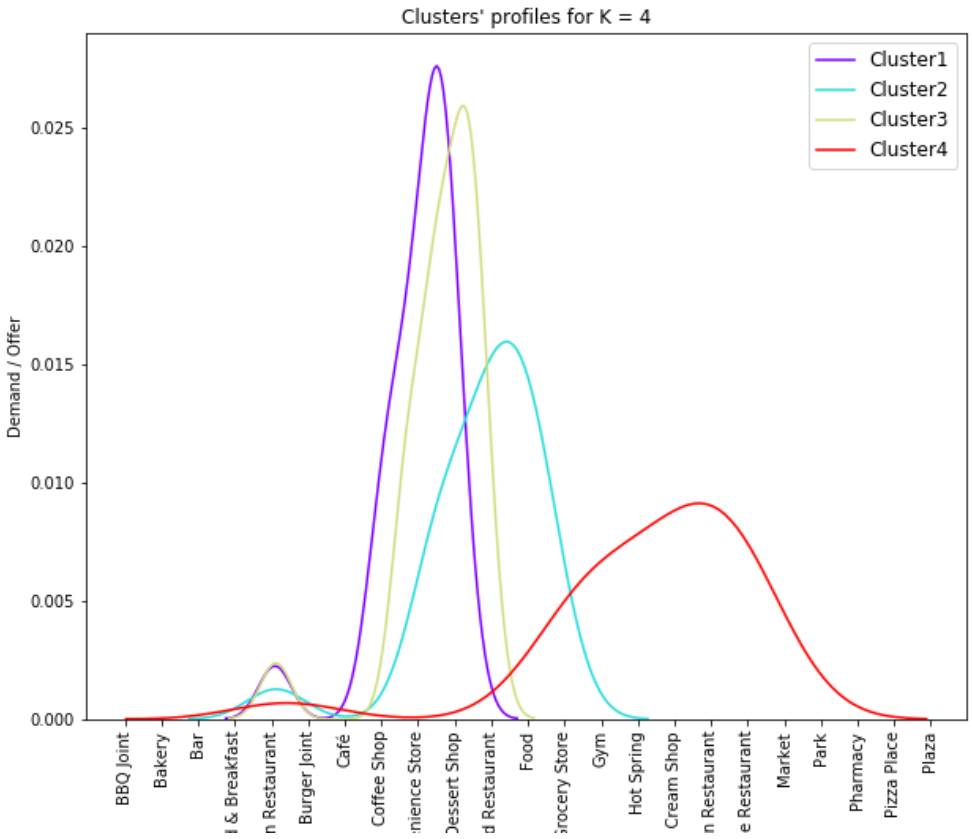

Figure 7 - Distribution curves of each cluster.

5. Results and Discussion

Our analysis shows there is a high concentration of cluster 2 cities (green dots) closer to the coast and around the São Paulo city (the capital of SP state), and sparsed cities of the other 3 clusters. This is expected, because the cluster 2 is composed of bigger and more developed cities of SP state, which are concentrated closer to the capital. Also, if we observe the distribution graph, the cluster 2 shows lower relation between demand and offer of each kind of establishment, because although these cities have huge number of habitants, they also have almost every kind of products and services in high quantities.

However, the cluster 4 also shows low values of demand vs offer on distribution graph, and the cities of this cluster (red dots) are smaller and sparser on the map. These cities seem to have good diversity of venue categories and with enough availability to their population.

The remaining clusters, 1 and 2, consist of cities with some specific venue categories missing. These seems to be great candidate cities for entrepreneurship. The cities in cluster 1, for example, lack in convenience stores, while the cities in cluster 2 lacks in dessert shops.

6. Conclusions

The purpose of this project was to identify similar cities in terms of availability of venue categories in order to support stakeholders to find cities as best opportunities to endeavor their business. By counting the number of each venue category in each city, we can find similar cities in terms of existence of these venues, but to understand better the opportunities we also considered the population of the cities. Taking the population into account, we could calculate the probable demand versus offer of each venue category in each city, and perform a better clustering of cities in terms of opportunities.

Furthermore, this analysis could also help people who wants to move to another city, keeping the lifestyle and access to specific facilities.

7. Future Directions

In this project, we clustered the cities based on venue categories availability and population, computing the demand versus offer ratio. This analysis gives a good notion of possible cities for entrepreneurship, once it is known in which each city lacks in terms of products and services.

To be more assertive in this clustering analysis, another feature could be considered is the feeling of clients in each venue category available in their cities, through feedbacks evaluation. With this information, even if a venue category exists in a certain city, it could be a good opportunity if they do not provide a good product / service.