IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Applied Data Science Capstone

PREDICTING SÃO PAULO'S FINANCIAL DISTRICTS LOCATION USING COFFEE SHOPS CONCENTRATION THROUGH CLUSTERING TECHNIQUES

Rafael Yonezawa de Mello
linkedin.com/rafaelyonezawa
São Paulo, April 15 of 2019

INTRODUCTION / BUSINESS CASE

According to the International Coffee Association (ICO), Brazil is the biggest Arabica coffee producing country on earth exporting around 58 million bags of grains per year [1]. With around 200 million citizens, it's also the 5th biggest country in the world in terms of population [2] and a developing economy according to the UN standards [3].

Besides all the contribution that coffee does to Brazilian's GDP, it's value for the average brazilian stands way beyond money. For the most of us, a good cup of coffee is an everyday-must-have.

However, as surprising as it may seem, having breakfast at a coffee shop is not part of the Brazilian's routine. That's because we usually have coffee at home and affording a good cup of steaming caffeine at a local coffee shop is kind of a luxury. Coffee shops like Fran's Café, Starbucks and Sterna Café are often associated with business meetings or once in a while purchases.

These observations around the average brazilian's habits raises the question around how improbable it is that coffee shops appearance in brazil tend to represent economic prosperity since they are not used on an everyday basis and are related to businesses. One way of proving this thesis is by testing our capability of correlating coffee shops distributions with financial districts in the city.

This analysis is relevant for both business and economy enthusiasts as well as for government officials who look out for possible economic indicators of development.

Is it possible to predict the location of the biggest business/financial districts of São Paulo just staring at coffee shops?

DATA

For this analysis we'll use Foursquare's API to find all coffee shops in São Paulo's territory as well as their latitude and longitude. With this data we'll be able to identify clusters as agglomerations of venues and their proximities to each other. Having this dataset and being aware of which are the main financial districts in São Paulo will be sufficient to test our main thesis.

According to Wikipedia, the 3 main financial zones in the city of São Paulo are:

1. Avenida Brigadeiro Faria Lima
2. Avenida Paulista
3. The historical Downtown

METHODOLOGY

The main question we want to answer can be seen as an unsupervised learning problem as we expect to identify agglomerations of Coffee Shops. As for the clustering algorithm, our features are going to be the latitude and longitude of the venues.

For collecting the data, I used the Foursquare's API looking out for the words *Coffee*, *Café* (which is coffee in Portuguese) and *Cafeteria* (which means Coffee Shop). Since the results of each

query is limited by 50, I performed multiple queries over the extension of São Paulo gathering the 50 nearest coffee shop and grouped them in a single dataframe. By eliminating duplicates, we have a full dataset of São Paulo's coffee shops. The head of this dataset looks like this:

```
In [6]: Coffee_Shops.head()
```

Out[6]:

id	location.cc	location.city	location.country	location.distance	location.formattedAddress	location.labeledLatLngs	location.lat	location.lng
2be1cd1498efd3ee4c8c521	BR	NaN	Brasil	1495	[Brasil]	[{'label': 'display', 'lat': -23.4261519313639...	-23.426152	-46.430490
4fccf703e4b09949e4fc7ce4	BR	NaN	Brasil	1298	[Brasil]	[{'label': 'display', 'lat': -23.4183807832593...	-23.418381	-46.435666
51f57da8498eeefdff8be716	BR	Guarulhos	Brasil	1307	[Guarulhos, SP, Brasil]	[{'label': 'display', 'lat': -23.4179611828491...	-23.417961	-46.435929
)72817d364d972852d1236b	BR	Guarulhos	Brasil	988	[Avenida Mulungú, 262, Guarulhos, SP, 07151-38...	[{'label': 'display', 'lat': -23.413462, 'lng'...	-23.413462	-46.458443
4fccf703e4b09949e4fc7ce4	BR	NaN	Brasil	1420	[Brasil]	[{'label': 'display', 'lat': -23.4183807832593...	-23.418381	-46.435666

After gathering the data, we must cluster the venues.

Since this clustering is going to be dependent on 2 spacial variables (latitude and longitude), the best way to identify similarity or difference between two venues is by using their distance to the centroids, thus, by calculating it's the Euclidian distance to the points of agglomeration.

The Euclidian distance of two points (p and q), with n dimensions, is expressed as:

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

In our case, we have only 2 dimensions (latitude and longitude).

For the clustering process, I used the k-means algorithm and used n=48. According to Wikipedia, São Paulo has around 96 districts [4], however, n=96 gives us too much granularity, as some clusters become too small (less than 5 venues), so I cut this value in half.

The results were compiled in another dataset, and its head look like this:

```
In [9]: Cluster_Dataframe.head()
```

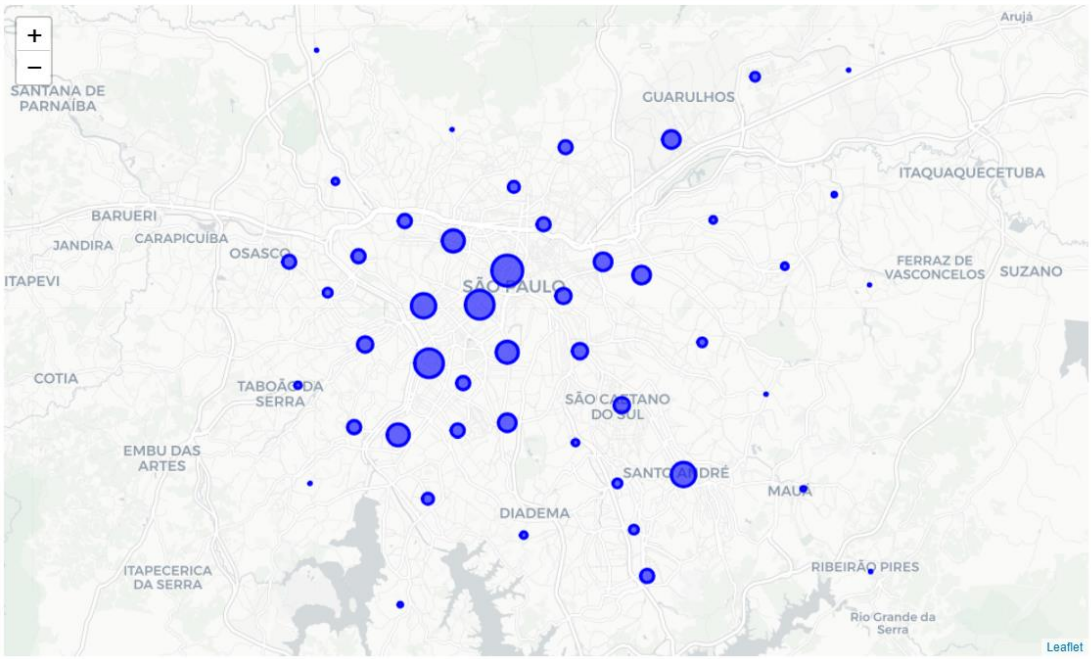
Out[9]:

Cluster	Venues	Latitude_center	Longitude_center
0	32	-23.692481	-46.626891
1	104	-23.547329	-46.600895
2	142	-23.596515	-46.686092
3	110	-23.658285	-46.527461
4	71	-23.583679	-46.727140

Plotting those in a map gives us the following:

```
In [19]: clusters_map
```

Out[19]:



Once we have our clusters and their sizes as well as their centroids, we're going to look if the top 3 clusters (in terms number of venues) are next to São Paulo's main financial districts. The discussion about this process can be seen in the Results section of this report.

RESULTS

The top 3 clusters (in terms of number of venues) are:

```
In [20]: Cluster_Dataframe.sort_values(by='Venues', ascending=False).head(3)
```

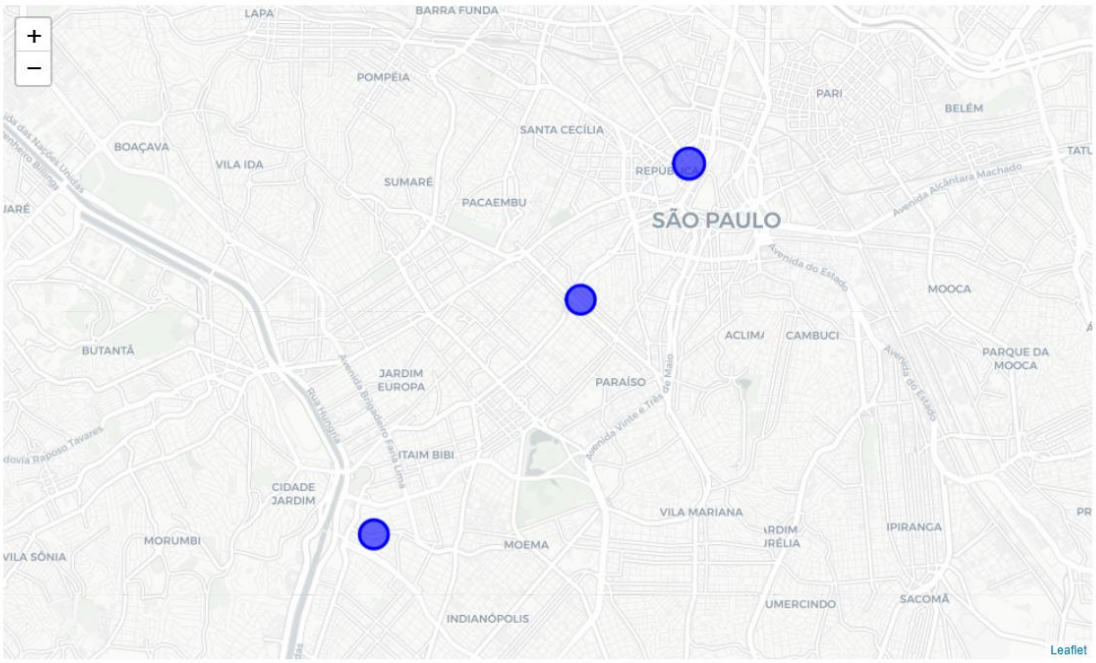
Out[20]:

Cluster	Venues	Latitude_center	Longitude_center
19	141	-23.542867	-46.637640
22	130	-23.595629	-46.686629
27	129	-23.562322	-46.654495

Plotting them in a map gives us:

```
In [22]: top_clusters_map
```

Out[22]:

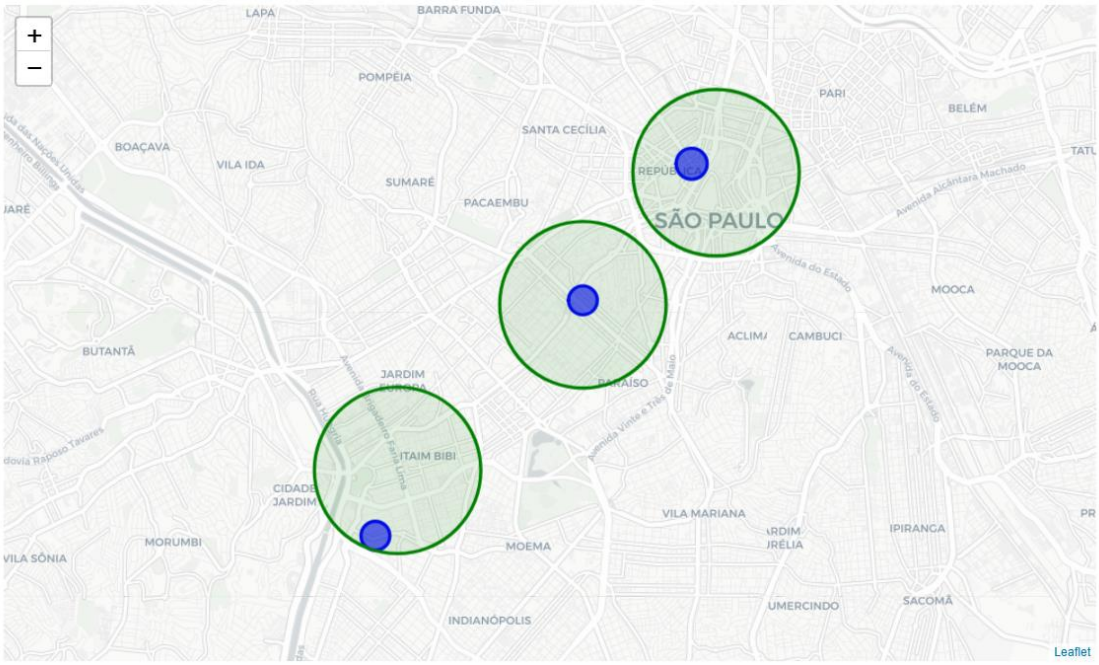


According to Wikipedia, there are 7 financial districts in São Paulo [5]. The biggest 3 are:

1. **Avenida Paulista:** mainly known as one of the biggest tourist attractions in the city for it's cultural value, it holds head offices of some banks and multinationals companies. Located at [-23.5629, -46.6544].
2. **Historical Midtown:** although this part of the city seem like it's been forgotten over time, it's still the district that holds the head office of the brazilian's stock exchange (B3), several companies and hotels. It's located at [-23.5442, -46.6339].
3. **Avenida Brigadeiro Faria Lima:** modern and luxuous, Faria Lima is by far the district with the highest economic development over the last years in São Paulo. It's the favorite spot for companies and hotels that are looking for new opportunities in the city. Located at [-23.5863,-46.6831].

If we plot them with a radius of 75 and the 3 main clusters, we can see that the results are satisfactory:

Out[23]:



DISCUSSIONS AND CONCLUSION

As expected, the 3 biggest agglomerations of coffee shops in São Paulo are near the biggest financial districts of the city. That support our thesis that those are good indicators of economic growth as their appearance are often associated with business development and not just populational density.

In conclusion, we can see that the results are satisfactory and the answer to our main question is yes.

[1] http://www.ico.org/prices/po-production.pdf

[2] http://www.worldometers.info/world-population/brazil-population/

[3] https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf

[4] https://pt.wikipedia.org/wiki/Divis%C3%A3o_territorial_e_administrativa_do_munic%C3%ADpio_de_S%C3%A3o_Paulo

[5] https://pt.wikipedia.org/wiki/Centro_financeiro