Specialty Coffee Shop & Venues Data Analysis of São Paulo City

A. Introduction

A.1. Description & Discussion of the Background

Is it possible to determine the best place to start a new Specialty Coffee Shop business?
According to the Specialty Coffee Association of America (SCAA), coffee which scores 80 points or above on a 100-point scale is graded "specialty."
Brazil is the top coffee producer nation and one of the top 15 consumer in the world.
As Specialty Coffee usually costs higher them regular coffee, using information of regions human development index (HDI) and Foursquare venues data, I will check if it is possible to determine the better location to start a new business.

A.2. Data Description

To consider the problem we can list the data as below:

- A list of some of the TOP Specialty Coffee Shops in São Paulo. Provided by a group of coffee lovers;
- A list of São Paulo´s coffee shops got from Foursquare API;
- Human Development Index by neighborhood scrapped from Wikipedia;
- A list containing the center coordinates of each neighborhood of São Paulo.

B. Methodology

As a database, I used GitHub repository in my study.
Bellow the data frame containing the Specialty Coffee Shops selected for this study:

	Name	Address	Neighborhood	Lat	Long
0	COFFEE LAB	Rua Fradique Coutinho, 1340	VILA MADALENA	-23.556041	-46.691383
1	UM COFFEE CO.	Rua Julio Conceicao, 553	BOM RETIRO	-23.527525	-46.641472
2	UM COFFEE CO.	Rua Iaia, 62	ITAIM BIBI	-5.806105	-35.235661
3	POR UM PUNHADO DE DOLARES	Rua Nestor Pestana, 115	CONSOLACAO	-23.548556	-46.645346
4	URBE CAFE	Rua Antonio Carlos, 404	CONSOLACAO	-16.578470	-43.936821
5	SOFA CAFE	Rua Artur de Azevedo, 514	PINHEIROS	-23.559819	-46.677229
6	SOFA CAFE	Rua Bianchi Bertoldi, 130	PINHEIROS	-23.568654	-46.690946
7	FREAK CAFE	Avenida Jurema, 359	MOEMA	-23.611181	-46.655530
8	KING OF THE FORK	Rua Artur de Azevedo, 1317	PINHEIROS	-23.563975	-46.683830
9	CUPPING CAFE	Rua Wisard, 171	VILA MADALENA	-23.555726	-46.690367
10	TORRA CLARA	Rua Oscar Freire, 2.286	PINHEIROS	-23.567231	-46.664316

I used python **folium** library to visualize geographic details of São Paulo. I created a map with Specialty Coffee Shops superimposed on top. I used latitude and longitude values to get the visual as below:
In Wikipedia I found a list of São Paulo´s HDI.
I´ve scrapped it and the result is a data frame with 96 rows. A head below:

	Distrito	IDH
0	MOEMA	0.981
1	PINHEIROS	0.980
2	PERDIZES	0.977
3	JARDIM PAULISTA	0.975
4	ALTO DE PINHEIROS	0.972

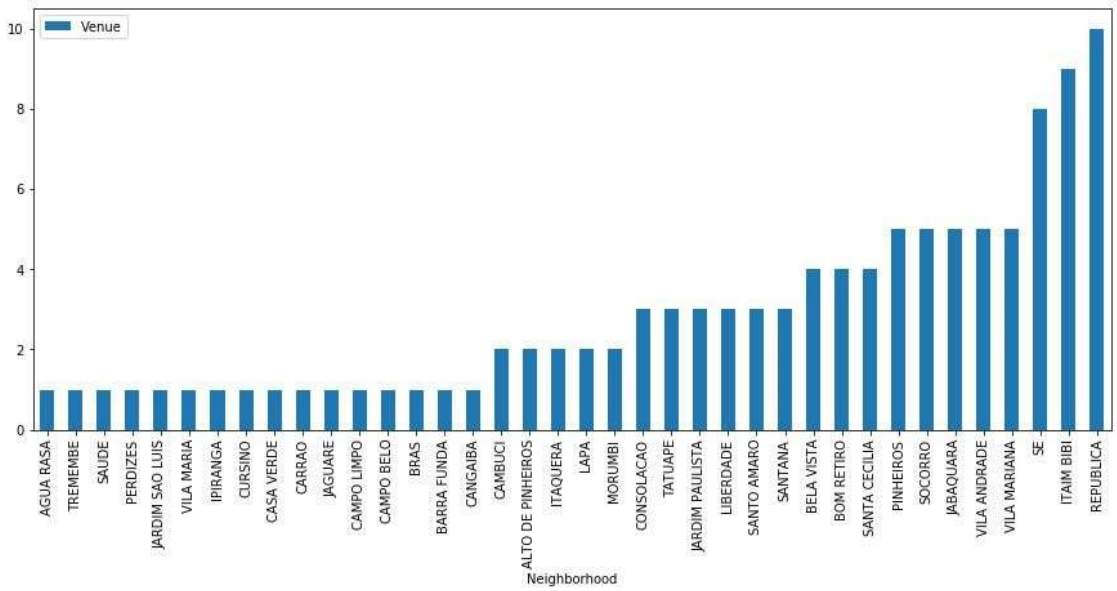
I utilized the Foursquare API to explore the São Paulo and find all Coffee Shops listed there. I designed the limit as **100 venue** per neighborhood and the radius of **250 meter** for each from their given latitude and longitude information. Here is a head of the list of neighborhoods:

	uf	municipio	bairro	longitude	latitude
0	SP	Sao Paulo	ACLIMACAO	-46.630972	-23.571487
1	SP	Sao Paulo	ALTO DA BOA VISTA	-46.692066	-23.635254
2	SP	Sao Paulo	ALTO DA LAPA	-46.717820	-23.535388
3	SP	Sao Paulo	ALTO DA MOOCA	-46.587312	-23.564019
4	SP	Sao Paulo	ALTO DA RIVIERA	-46.763657	-23.699916

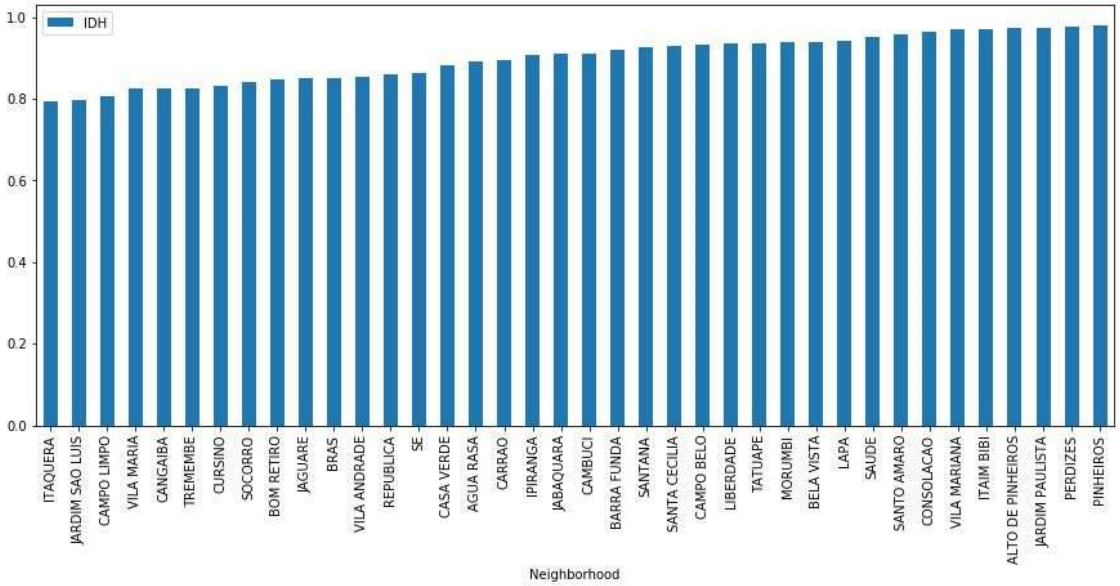
And a head of the list of Venues name, latitude and longitude informations from Forsquare API.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
8	ALTO DE PINHEIROS	-23.553709	-46.708835	Leo Dolci	-23.553818	-46.707798	Café
16	ALTO DE PINHEIROS	-23.553709	-46.708835	McCafé	-23.553481	-46.707820	Coffee Shop
44	BARRA FUNDA	-23.530376	-46.657432	Maná Doçaria e Café	-23.529816	-46.656263	Coffee Shop
101	BELA VISTA	-23.562831	-46.646259	Moscatel Doceria e Bar de Açúcar	-23.558924	-46.646283	Café
137	BELA VISTA	-23.562831	-46.646259	Café Família	-23.559764	-46.649551	Café

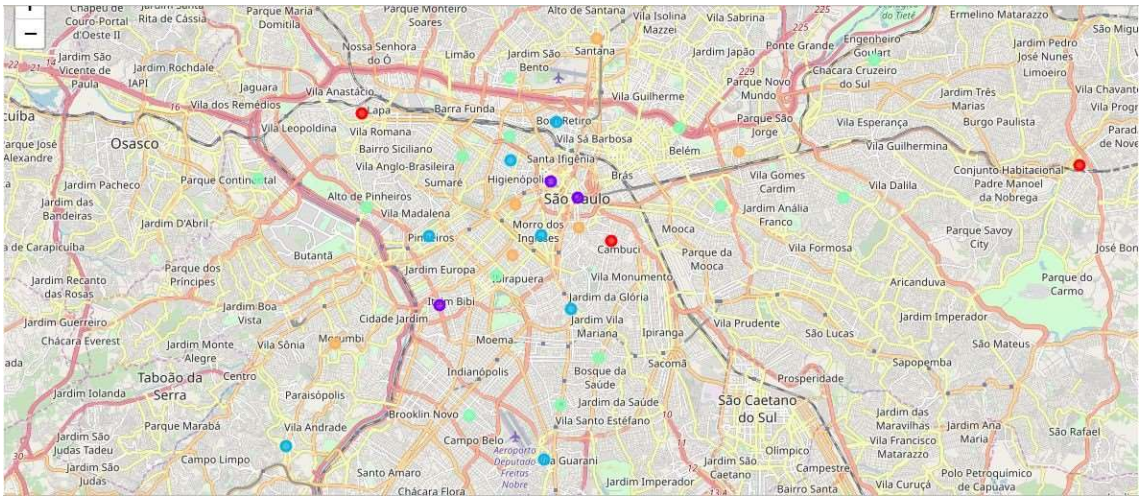
After cleaning some NaN. In summary 108 venues returned by Foursquare. The result doesn't mean that inquiry run all the possible results in all neighborhoods. Actually, it depends on given Latitude and Longitude information and here is we just run single Latitude and Longitude pair for each location. We can increase the possibilities with Neighborhood information with more Latitude and Longitude information. And certainly, this number will increase.



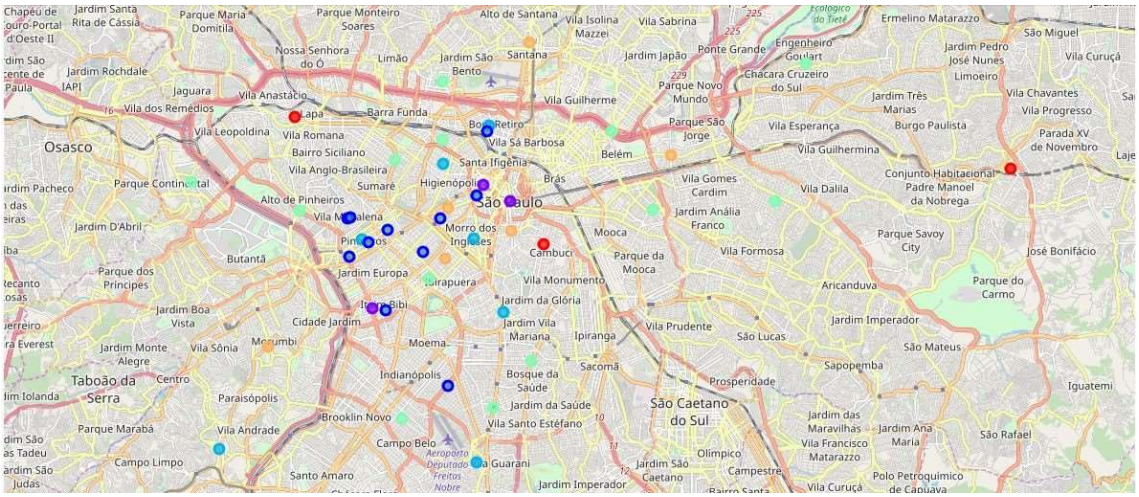
Now we can take a look at HDI graph. Does it ring any bells?



Not at all for me. Let´s make some further analysis.
We can see a clustered map of São Paulo´s neighborhoods below.



I used unsupervised learning **K-means algorithm** to cluster the neighborhoods. K-Means algorithm is one of the most common cluster methods of unsupervised learning.
First, I will run K-Means to cluster the neighborhoods into 5 clusters because when I analyze the K-Means with elbow method it ensured me the 5 degree for optimum k of the K-Means.
Below the map with clustering the neighborhoods based on the number of venues.



And if we create 4 groups considering the HDI range from low to high. Will this model fit better?
Based on the describe function from our data frame, we can obtain the values of the intervals of each group.

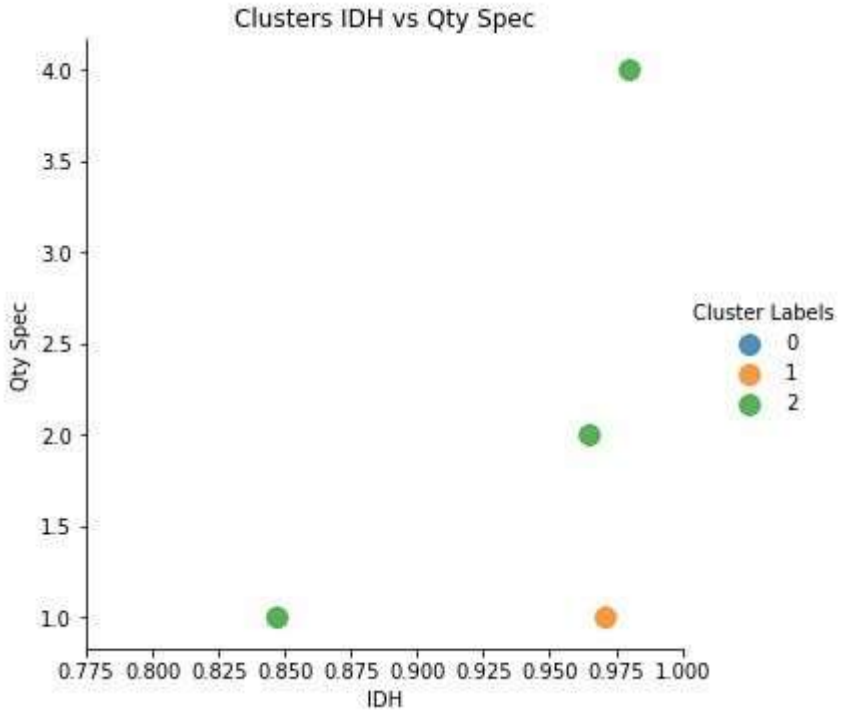
	Cluster Labels	Venue	longitude	latitude	IDH
count	35.000000	35.000000	35.000000	35.000000	35.000000
mean	2.742857	2.971429	-46.609469	-23.517163	0.903200
std	1.313792	2.357538	0.194813	0.194567	0.057183
min	0.000000	1.000000	-46.747535	-23.653663	0.795000
25%	2.000000	1.000000	-46.673168	-23.580847	0.852000
50%	3.000000	3.000000	-46.645191	-23.552568	0.920000
75%	4.000000	4.000000	-46.623326	-23.532811	0.945500
max	4.000000	10.000000	-45.547962	-22.590591	0.980000

We will consider the following:

- High HDI - values between 0.98 and 0.9455;
- Medium HDI - values between 0.9454 and 0.9200;
- Medium Low HDI - values between 0.9199 and 0.85200;
- Low HDI - values between 0.85199 and below.

C. Results

The plot below shows a cluster of IDH and the quantity of Specialty Coffee Shops by neighborhood.



As we can see. Specialty Coffee Shops are usually installed in neighborhood with High HDI.

D. Discussion

São Paulo is a big city, one of the largest in the world. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high-quality results for this.

I used the Kmeans algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 5. However, only 88 district coordinates were used. For more detailed and accurate guidance, the data set can be expanded, and the details of the neighborhood or street can also be drilled.

I was couldn´t find the correct geojson of the city. For this reason, I was unable to plot a choropleth map with HDI index and a visualization of the venues on it.

I ended the study by visualizing the data and clustering information. In future studies, I will try to include some additional social information and to find additional data so I can use more neighborhoods.

E. Conclusion

As a result, Specialty Coffee Shop business tends to be in neighborhoods with high HDI. One investor should look closer to these locations when thinking of starting his new business.

Above and beyond,

Fábio Saito