Exploring São Paulo Neighbourhoods Venues

Paulo Cardoso - 2019/06/15

Introduction

The work is a capstone project for the Coursera course IBM Data Science Professional Certificate Specialization. The objective of this first part is to present a description of the problem and also point out with data will be used on the development of de project.

Problem Statement

- It's possible replicate parts of the analysis performed during the course with São Paulo Data?
- How can performed a grouping the neighborhoods and find out top 10 venues corresponding to that neighborhood in Sao Paulo, SP - Brazil?

Problem Description

Grouped by neighborhood exploring the venues around creating a top 10 venues by neighborhood in São Paulo. This kind of analysis can lead to possibility of identification of new opportunities for businesses, identification of the saturation of business segments in the neighborhood and also possibility to better understand the citie with its characteristics based on distribution of the venues.

Target Audience

- Payment industry, this segment could take advantage of an analysis like to that to better position their self on the market, knowing where their product could have a better acceptance.
- Investors thinking on open a new business in town, this kind of person could benefit from this analysis to rule out segment of business in tow due saturation of the kind of business on the neighborhood.
- Social Science researchers curious with the behaviour and distribution of venues in the city.

Sucess Critteria

The success criteria of the project will be a good presentation of which are the top 10 kind of venue by neighborhood and some recommendation regarding which are the worst business segments to be opened on some neighborhoods.

Data

On this project was used data originated from three different sources, the first dataset gathered was the brazilian postal code dataset, the next one was data retrieved from the geocoder ArcGIS APIs/Python Library and the third and last source Foursquare API data. It's important to say that the data gathered from the APIs was used to enrich the original postal code dataset.

- Postal code: http://cep.la/CEP-dados-2018-UTF8.zip
- Geocoder ArcGIS: https://developers.arcgis.com/python/guide/
- Foursquare: https://developer.foursquare.com/docs/api

Data Cleansing

The original postal code data gathered on http://cep.la/CEP-dados-2018-UTF8.zip as several problems as for example:

- Name of the city separated on more than one column
- State abbreviation combined with second column of the city name
- Neighborhood name and Street name concatenated
- Three empty columns

0	1	2	3	4	5	6
1001000	São	Paulo/SP	Sé\tPraça da Sé - lado ímpar	NaN	NaN	NaN
1001001	São	Paulo/SP	Sé\tPraça da Sé - lado par	NaN	NaN	NaN
1001010	São	Paulo/SP	Sé\tRua Filipe de Oliveira	NaN	NaN	NaN
1001900	São	Paulo/SP	Sé\tPraça da Sé, 108 \t UNESP - Universidade E...	NaN	NaN	NaN
1001901	São	Paulo/SP	Sé\tPraça da Sé, 371 \t Edifício Santa Lídia	NaN	NaN	NaN

Data Cleansing

To mitigate this issue and enable analyses continuation a data cleansing must be applied, techniques of string handling, and dataframe entanglements were utilized on this cleansing. In order to demonstrate the result of cleansing process below is presented the cleaned dataset.

	City	State	CEP	Neighborhood	Street
0	São Paulo	SP	1001000	Sé	Praça da Sé - lado ímpar
9	São Paulo	SP	1002020	Centro	Viaduto do Chá
173	São Paulo	SP	1017000	Brás	Avenida Rangel Pestana - até 499/500
337	São Paulo	SP	1035100	República	Avenida São João - de 651 a 1339 - lado ímpar
540	São Paulo	SP	1101000	Luz	Avenida Santos Dumont - até 999/1000
546	São Paulo	SP	1101080	Ponte Pequena	Ponte Santos Dumont
591	São Paulo	SP	1107000	Bom Retiro	Avenida do Estado - até 2599 - lado ímpar
602	São Paulo	SP	1109000	Canindé	Avenida Cruzeiro do Sul - até 1299 - lado ímpar
723	São Paulo	SP	1134901	Barra Funda	Avenida Rudge, 700
734	São Paulo	SP	1136005	Várzea da Barra Funda	Rua Américo Del Veneri

Methodology

Aquire Postal Code data from Brazilian cities, clean the data, perform data explorationg, plot basic statistics to describe the data and present metrics, plot maps to present the data gathered, perform feature engineering, clusterization using K-Means, demonstrate results.

Data Exploration - Part 1

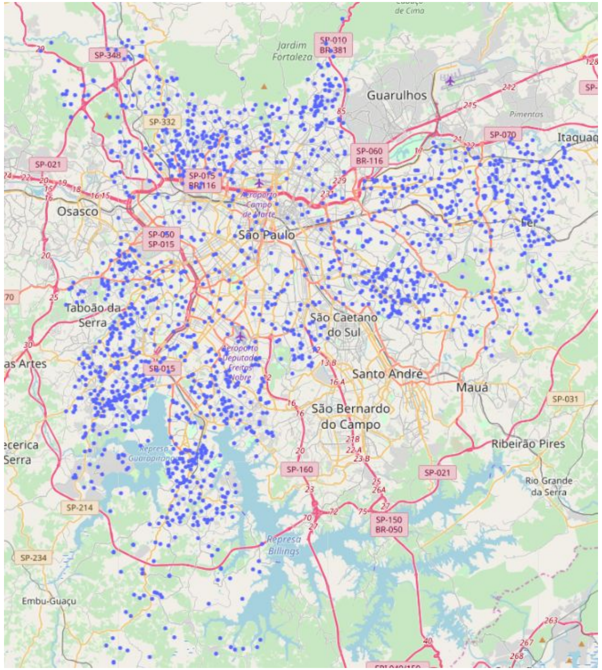
The first step of the data exploration was performed the enrichment of the postal code data with geolocation information, on another words latitude and longitude was added to the original cleaned data in order o enrich it. Below is presented the process of data enrichment by the creation of a function, the execution of the function and result dataset.

```
# Creating a function to get the Lat Long data from the Postal Code
def get_geocoder(postal_code_from_df):
    lat_lng_coords = None
    while(lat_lng_coords is None):
        g = geocoder.arcgis('{}, São Paulo, São Paulo'.format(str(postal_code_from_df).strip()))
        lat_lng_coords = g.latlng
        latitude = lat_lng_coords[0]
        longitude = lat_lng_coords[1]
    return latitude,longitude
```

```
# Adding Latitude and Longitude columns to dataframe
dt1['Latitude'], dt1['Longitude'] = zip(*dt1['CEP'].apply(get_geocoder))
dt1.head()
```

	City	State	CEP	Neighborhood	Street	Latitude	Longitude
0	São Paulo	SP	1001000	Sé	Praça da Sé - lado ímpar	-23.562870	-46.654680
9	São Paulo	SP	1002020	Centro	Viaduto do Chá	-23.546685	-46.637805
173	São Paulo	SP	1017000	Brás	Avenida Rangel Pestana - até 499/500	-23.549709	-46.630488
337	São Paulo	SP	1035100	República	Avenida São João - de 651 a 1339 - lado ímpar	-23.540970	-46.642602
540	São Paulo	SP	1101000	Luz	Avenida Santos Dumont - até 999/1000	-23.562870	-46.654680

Data Exploration - Part 1



Data Exploration - Part 2

The second step of the data exploration is very similar to the first one, it also aimed into enrich the original data. But this turn the data source utilized was the Foursquare API, to do that, it was created a function responsible for based on a input of the neighbourhood name, latitude and longitude performed a API call to retrieve the foursquare data. The result of put all this API response data into a dataframe can be observed below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Sé	-23.56287	-46.65468	Museu de Arte de São Paulo (MASP)	-23.561585	-46.655832	Art Museum
1	Sé	-23.56287	-46.65468	Seen	-23.563984	-46.656098	Restaurant
2	Sé	-23.56287	-46.65468	Teatro Popular do Sesi	-23.563635	-46.654605	Theater
3	Sé	-23.56287	-46.65468	Tivoli Mofarrej - São Paulo	-23.563923	-46.656146	Hotel
4	Sé	-23.56287	-46.65468	Starbucks	-23.562082	-46.655736	Coffee Shop
5	Sé	-23.56287	-46.65468	Centro Cultural FIESP - Ruth Cardoso	-23.563376	-46.654245	Cultural Center
6	Sé	-23.56287	-46.65468	Pello Menos	-23.562758	-46.656545	Health & Beauty Service
7	Sé	-23.56287	-46.65468	Starbucks	-23.563384	-46.653057	Coffee Shop
8	Sé	-23.56287	-46.65468	Granado	-23.563961	-46.652955	Cosmetics Shop
9	Sé	-23.56287	-46.65468	Lindt	-23.564155	-46.653155	Chocolate Shop

Data Presentation

- Top 10 Venues Categories

Venue Category	counts
Brazilian Restaurant	1858
Bakery	1808
Pizza Place	1461
Hotel	1060
Pharmacy	1007
Burger Joint	1003
Gym / Fitness Center	1002
Coffee Shop	908
Restaurant	906
Middle Eastern Restaurant	856

- Top 10 Venues Categories in the neighborhoods

Neighborhood	Venue Category	counts
Vila São Francisco	Brazilian Restaurant	16
Sumarezinho	Bar	15
Vila Guarani (Z Sul)	Brazilian Restaurant	12
Cerqueira César	Bar	12
Itaim Bibi	Restaurant	11
Sumarezinho	Art Gallery	11
Vila São Francisco (Zona Sul)	Brazilian Restaurant	11
Centro	Brazilian Restaurant	10
Paraíso	Coffee Shop	10
Itaim Bibi	Italian Restaurant	10

Data Presentation

- Top 3 Venues Categories by neighborhoods

Neighborhood	Venue Category	counts
Área Rural de São Paulo	Hotel	7
Área Rural de São Paulo	Brazilian Restaurant	6
Área Rural de São Paulo	Coffee Shop	5
Água Rasa	Brazilian Restaurant	3
Água Rasa	Bakery	2
Água Rasa	Farmers Market	2
Água Funda	Bakery	1
Água Funda	Brazilian Restaurant	1
Água Funda	Furniture / Home Store	1
Água Fria	Japanese Restaurant	4
Água Fria	Pharmacy	4
Água Fria	Pizza Place	4
Água Branca	Hotel	7
Água Branca	Brazilian Restaurant	6
Água Branca	Coffee Shop	5
Várzea de Baixo	Farmers Market	2
Várzea de Baixo	Italian Restaurant	2
Várzea de Baixo	Mineiro Restaurant	2

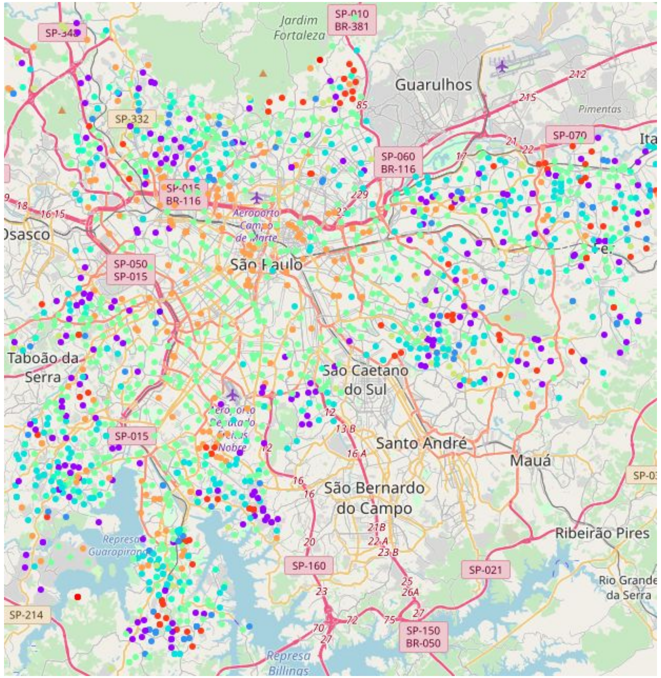
- Top 10 neighborhoods with more venues

	Neighborhood	counts
1871	Área Rural de São Paulo	100
1071	Moema	100
1278	Santa Amélia	100
1266	Roda a Roda Jequiti	100
1258	República	100
1245	Protendit	100
1232	Pinheiros	100
1180	Parque Rodrigues Alves	100
244	Jardim Aladim	100
321	Jardim Bronzato	100

Obs.: Due Foursquare API restriction, the maximum number of venues by neighborhood is 100

Results

The result of this project can be understood as the clustering of the neighborhoods based on the similarity of the top 10 venues. To define the ideal amount of clusters several tests were performed with several different number of clusters as for example 3, 4, 5, 6, 7, 8, 9 and 10 clusters. Below is the plot that represents the distribution of the ten clusters:



Results

Regarding the Cluster size, below is presented the amount of clusters created in this analysis and also the amount of neighborhoods on each of this clusters.

Cluster #	Amount of Neighborhoods
Cluster 0	238
Cluster 1	5
Cluster 2	83
Cluster 3	327
Cluster 4	95
Cluster 5	681
Cluster 6	65
Cluster 7	278
Cluster 8	96
Cluster 9	4

Discussion

As presented on de previous section, the map plot and the frequency of neighborhoods by cluster. Adding to that, when looing at this data and the data of the data exploration it's possible to extrapolate some recommendations having on mind the target audience mentioned at the begining of the work.

Recommendation

- Payment industry: could use this data to improve market research amd target specific segments to getter a better penetration and capilarity in terms os sales. Also, another point it would be knowing the neighborhood distribution on the clusters would be easyer to target campaings to stimulate sales.
- Investors: São Paulo is one of best gastronomical city in the world, several excelent restaurants. I would recommend open a nich restaurant on neighborhoods of cluster 4 because its already a place where the best restaurants are located.
- Social Science researchers: the cluster distribution in the map denote a clear income differance, as rings the clusters are opening letting clear the differences on social class.

Conclusion

To conclude is important to point out that all problem statements and descrtiption were achived and provided insightful response and recommendations about which would be the best actions depending on certain scenarios. So in a whole this analysis was succesful and achived its objectives.