Summary

Context

São Paulo is a metropolis of many faces. At the same time it is the most important economic center of Brazil, it is the capital of culture in Latin America; with leisure, knowledge and entertainment offers that match to no other. A typically urban metropolis covered in a vast green area.

Diversified and extremely rich gastronomy; the city gathers some of the best restaurants in Latin America and the world – amongst its over 15 thousand restaurants and 20 thousand bars. There are national and international options, which reach every customer. Besides the culinary of 52 countries, São Paulo is famous for its food trucks and "gastronomic little fairs", which nowadays take over the streets of the city.

São Paulo is a city of numbers. There are over 12 million inhabitants that share space with 14,9 million of tourists per year. It is 1,521 km² wide, offering several interesting places to meet, restaurants to taste different gastronomies, malls and many shopping centers for all kinds of goods. And even though buildings surround it, there are lots of parks and green areas.

Speaking of it, here are some numbers regarding what "Paulistanos" and tourists can find on the city:

320 hotels (44,000 rooms)
63 hostels
314 movie theaters
138 theaters
115 cultural centers
150 libraries
158 museums
333 sport and leisure centers
11 soccer stadiums
109 parks and green areas
20 thousand restaurants
30 thousand bars
53 malls
33 thousand taxis

Audience

Explore, monitor and understand the market is a concern that suits very well many situations: business owners who want to expand or defend, entrepreneurs who want to make a market entry, even employees in order to select the best places to work.To do

so, is essencial to identify the best neighborhoods to start a new restaurant and help this public.¶

The business problem

In this São Paulos's cenario, we see a very consolidated and diverse restaurant's market that attracks a large public. Thus, to start a new restaurant in the capital it is essencial a well planned market entry. To do so, we can analyse the current market with machine learning techniques and deliver the best spots and types of cusine to start a new business.

Methodology Overview

The methodoly will be:

1)Data extraction from Foursquare API and Geoenconding from Google Cloud API

2)Data profiling and preparation: understand the data distribution, missing values, restaurants descriptions. Prepare the data to run the models.

3)Modeling: stablish a clustering model and by regression with other businesses occurrences find the best neighborhood to start a new restaurant

4)Discussion and success criteriar: for each cluster, we run a linear regression taking out a neighborhood of our interest. We input the neighborhood dependent variables and run the trained model to understand the projection. The difference between the projection and the real value reflects the lack of restaurants in that neighborhood.

5)Conclusion: recommend the best place to start a restaurant

Success Criteria

Identify the neighborhood with the highest lack of restaurants inside the cluster of restaurants(the one that restaurant is the predominant business and consequently it has the best correlation and sinergy with betewen restaurants and other businesses).¶

Data extraction

In this section, we will extract, clean and prepare the data to be analysed. The main features or informations are:

1)The neighborhood

2)The geospacial cordinates

2)The business located in each venues of these neighborhoods

3)The frequency of occurrences of each business in each neighborhood

To get this variables we will need to

1)Collect the neighborhoods from São Paulo and the geospatial coordinates. Source: https://www.estadosecapitaisdobrasil.com/listas/lista-dos-bairros-de-sao-paulo/

For this, I manually downloaded and organized in a excel file.

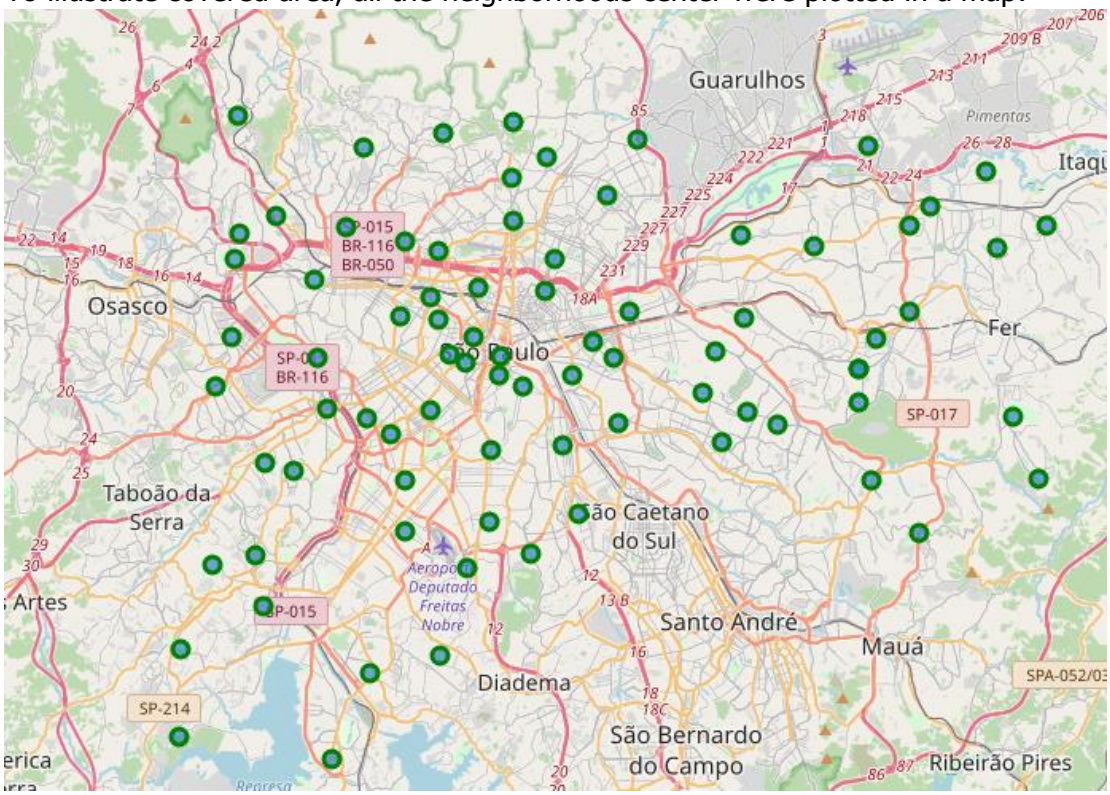
The total number of official neighborhoods is 90.

2)The googlemaps api allow us to extract the geospacial coordinates from the neighborhood name. In order to extract, I created a imported the googlemaps library and I run a loop:

	FullAddress	long	lat
0	Água Rasa, São Paulo	-46.5819	-23.5532
1	Alto de Pinheiros, São Paulo	-46.7096	-23.5533
2	Anhanguera, São Paulo	-46.7279	-23.4976
3	Aricanduva, São Paulo	-46.511	-23.5795
4	Artur Alvim, São Paulo	-46.469	-23.546

Data Profiling

To illustrate covered area, all the neighborhoods center were plotted in a map:



3)With these coordinates, we extract all the businesses presented in a 1000 meters radius from the foursquare API. The result is the venue and the venue's category:

Água Rasa, São Paulo	-23.553209	-46.581890	Padaria Santa Branca	-23.553953	-46.583706	Bakery
Água Rasa, São Paulo	-23.553209	-46.581890	Bona's Carnes	-23.552434	-46.583091	Restaurant
Água Rasa, São Paulo	-23.553209	-46.581890	Temaki Station	-23.553987	-46.583660	Restaurant
Água Rasa, São Paulo	-23.553209	-46.581890	Padaria Carillo	-23.553214	-46.578554	Bakery

Here is the category rank sorted by occurrences:

Restaurant	1399
Bakery	296
Bar	188
Gym / Fitness Center	175
Burger Joint	151
Gym	136
Pharmacy	124
Dessert Shop	119

Restaurant is the top 1 occurrence. This reflects the market focus of Foursquare and also expands our data sample.

We can see some prominent neighborhoods in terms of restaurants:

Itaim Bibi, São Paulo	Restaurant	47
Jardim Paulista, São Paulo	Restaurant	41
Santa Cecília, São Paulo	Restaurant	39
Iguatemi, São Paulo	Restaurant	36
Liberdade, São Paulo	Restaurant	33
Pari, São Paulo	Restaurant	32
Santana, São Paulo	Restaurant	32
República, São Paulo	Restaurant	32
Sé, São Paulo	Restaurant	30
Saúde, São Paulo	Restaurant	30
Campo Belo, São Paulo	Restaurant	30
Consolação, São Paulo	Restaurant	30

Modeling

In São Paulo we have common venues. To understand and perform better correlations, I created the following method:

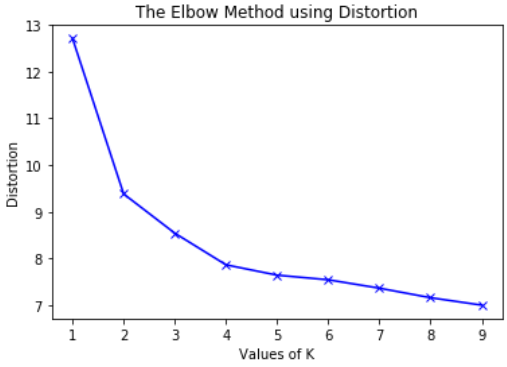
1)Run a clustering model

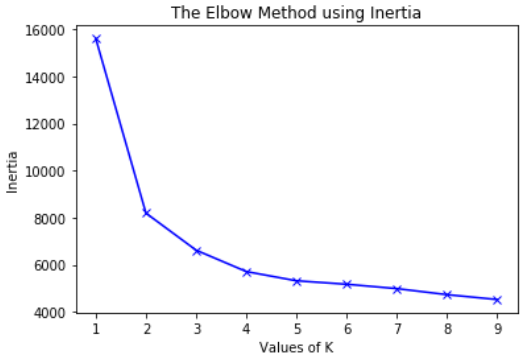
2)Select the clusters with more restaurants

3)Train and test a linear regression for restaurant's ocurrances x other places for each cluster alone and together. Select the best method by mean error.

4)With the winner method, run a train the linear regression without a target neighborhood and project the number of restaurants of that neighborhood. The difference between the projection and the real number represents the lack of restaurants in that specific neighborhood.

In order to cluster the model, I used the K-means method for 10 Ks. The best number os clusters was 4, according to the Elbow and Inertia Method:





The dataset clustered enable, at least, the 2 method metioned earlier:

a)Train and test the restaurant projection for each cluster

Cluster	Mean Error
0	2.46%
1	1.94%
2	2.56%
3	4.28%
Total mean	2.81%

b)Train and test the restaurant projection for all cluster together

Cluster	Mean Error
0,1,2,3	5.26%

The best method is a) with smallest error.

With the method a) was run a traning and test loop, performing the following instruction:

1)Remove the target neighborhood from the cluster sample

2)Train the linear regression with restaurants(dependent variable) and other places(independent variable)

3)Predict the number of restaurants for the target cluster with its depend variable

4)The difference between the predicton and the real number of restaurants is the lack of restaurants in the target neighborhood.

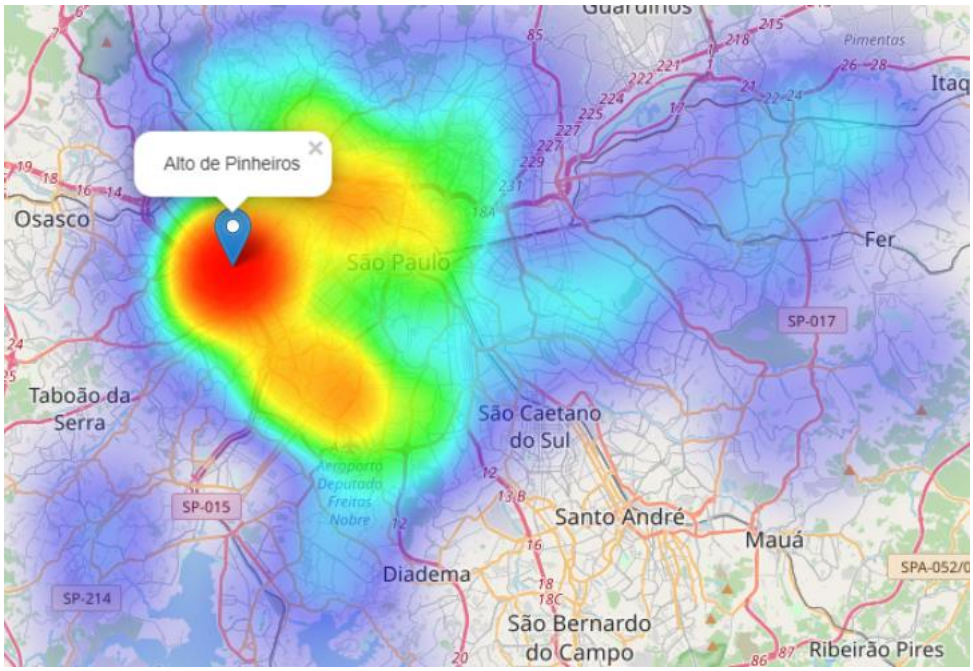
5)This results is stored and the algorithm moves on to the next neighborhood.

Discussion

This is the lack of restaurants rank:

Neighborhood	Actual of Number Restaurants	Predicted Number of Restaurants	Lack of Restaurants
Alto de Pinheiros, São Paulo	3.0	14.483317	11.483317
Moema, São Paulo	17.0	24.306804	7.306804
Consolação, São Paulo	30.0	37.252818	7.252818
Casa Verde, São Paulo	18.0	23.175517	5.175517
Brasilândia, São Paulo	4.0	8.894292	4.894292
Saúde, São Paulo	30.0	34.626458	4.626458
Perdizes, São Paulo	20.0	24.390057	4.390057
Cidade Ademar, São Paulo	9.0	13.053998	4.053998
República, São Paulo	32.0	35.483911	3.483911
Capão Redondo, São Paulo	10.0	13.033215	3.033215

We can see there is tendecy of lack of restauraunt toward the west from south and north, where Alto dos Pinheiros is located.



Conclusion

The recomendation is that the best place to start or maintain a restaurant is Alto de Pinheiros. This neigborhood has others businesses correlated to restaurants that indicates by linear regression a lack of 11.4 restaurants.