The Battle of Neighborhoods

JULY 20

Authored by: Rodrigo Eiras

Final Report

Finding the better neighborhood to start an English School in Rio de Janeiro, Brazil

1) Introduction

The customer owns a franchise of English Schools named “English4You” and they are interested to open a new schools in Brazil, specifically in Rio de Janeiro. Rio de Janeiro is one of the most excited cities in Brazil, culturally diverse, beautiful beaches, have good restaurants and pubs to go out at night and a very receptive people that make it a good option to invest.

This would be the first school from this franchise in Brazil, they do not know very well the places, neighborhoods, and best locations to set up the business. As Rio is a very intense city in tourism, there are many English schools around and choosing a location that minimize the competition is a prior challenge for the project.

2) Business Problem

What neighborhood from Rio de Janeiro has good conditions to establish an English school considering some metrics such less competition, per capita income, life expectancy, education rate and so on.

3) Data Sources

The data to be used for this project comes from two different locations:

Foursquare API

- It is a local search-and-discovery service which provides information on different types of entertainment, drinking and dining venues. Foursquare

has an API that can be used to query their database and find information related to the venues, such as location, overall category, reviews, and tips.

Rio de Janeiro Neighborhood Census Data

- This data is available through Wikipedia and contains the neighborhood names also the main metrics about life quality living in the city.

3.1) Foursquare API

For this project we will use the Foursquare Places API. One of the features of this API is to provide a list of venues within a specific location, based on the Lat/Lon coordinates and a radius. In order to obtain a list of venues within a specified area, we use the "explore" endpoint from the API. By passing the proper parameters via an HTTP request to the explore endpoint, we get a JSON object with the information shown in below:

```
1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue
5. Name of the venue e.g. the name of a store or restaurant
6. Venue Latitude
7. Venue Longitude
8. Venue Category
```

3.2) Rio de Janeiro Neighborhood Census Data

The data is based on the last official census published by the government and now is public through Wikipedia. The Rio de Janeiro City has at least 158 neighborhoods and some of them are aggregated in the same line, so I need to split them to make a good data frame about the correct locations. The URL where is it located in Wikipedia is:
https://pt.wikipedia.org/wiki/Lista_de_bairros_do_Rio_de_Janeiro_por_IDH

Some feature that can be extracted includes life expectancy, education rate, per capita income and some others.

The table in Wikipedia does not have the location values (lat/long) so I need to geocode it through geopy library.

4) Methodology

The capstone project was separated in six parts to be more organized:

a. Part 1: Data Loading and Data Wrangling
b. Part 2: Geocoding the Data
c. Part 3: Data Exploration and Data Visualization
d. Part 4: Machine Learning – Clustering
    i. K-Means and Dendrogram
e. Part 5: Venues Evaluation using Foursquare API
f. Part 6: Conclusion
g. This section will show all the steps done in the analysis: data exploration, charts, and models. The results and discussions will follow in the next sections of the report.

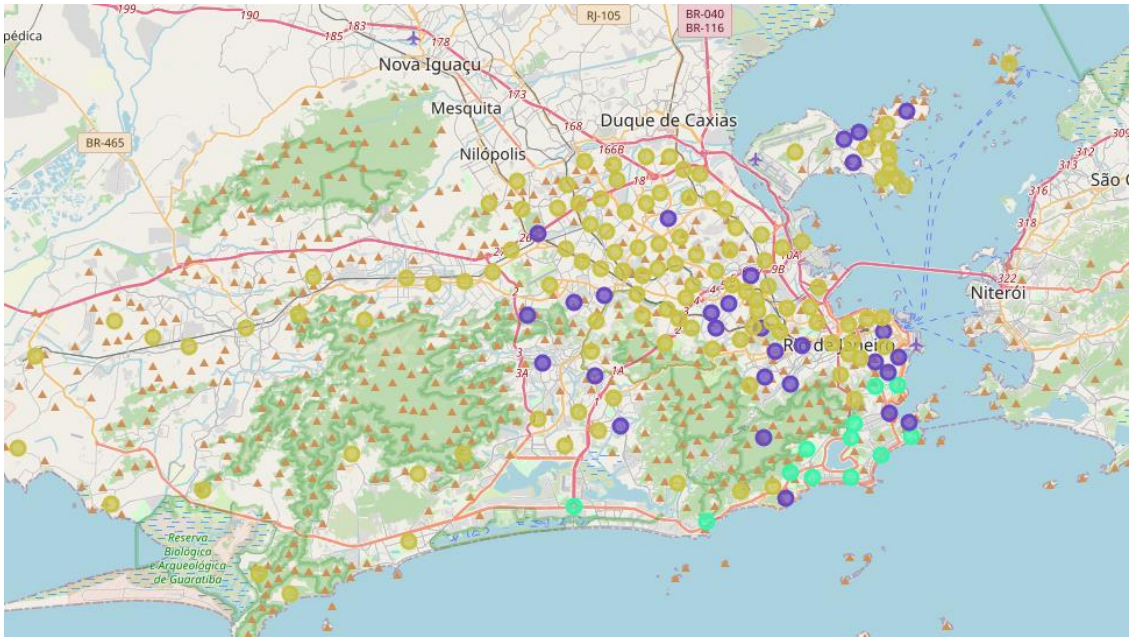

Figure 1 Clusters provided by K-Means

5) Results

On this section, will be presented all the data analysis about the neighborhoods in Rio de Janeiro city focusing in understanding and identify a good neighborhood to start an English School from a new franchise that coms to Brazil.

- 1st Discussion: In Rio de Janeiro, most of the people that has higher per capita income lives in the south area of the city and closer to the beach. That explains a lot for example why Copacabana, Ipanema and Leblon are neighborhoods highly looked for tourists to stay when they are in the city. Off course, crime rates in those areas are lower compared to other areas, like the suburb for example.

- 2nd Discussion: Even with a big difference in per capita income in most of the neighborhoods, the city of Rio de Janeiro has a good Human Development Index and I think some reasons can explain that two of them are:

  1) The city is not too big, that means most of activities are accessible by most of the people living in the city,

  2) As we can see in the other metrics the education rate and life longevity rate are particularly good even in areas where the per capita income is low. That values raises the average of the HDI (that is the calculated rate)

- 3rd Discussion: As said in the 2nd discussion, the HDI rate (IDH) is heavily influenced by the education rate (IDH-E) that raises the IDH rate average making the neighborhoods having not much difference in HDI. Because the city is not that big, the inequality rate is high with slums present in almost every neighborhood of the city. In fact, I can conclude the problems that has in the city are not easily saw only looking to HDI metrics, but the data is enough to see where people with more acquisition power lives.
- To a company that is looking to know where to open an english school in Rio de Janeiro, the data is good enough.

- 4th Discussion: If we look the box plots searching for inequality, we can see the south area is more linear on that. In general, people that lives in the south area have better results in every metrics available and discussed here. The North area has some difference depending on the localization. How far north, worst is the conditions. The Downtown area usually only have commercial buildings, few people lives there compared to others area. The West zone is like the south area only around "Barra da Tijuca". The rest of the west area is more like the north area.

- 5th Discussion: As I suspected before looking the box plots, the distribution of south area is more linear when compared to the other areas. That means the inequality there is concentrated in some small slums around the neighborhoods. The other areas, such West and North has some outliers that comproves they are more unequal. There are some regions (outliers) that maybe have some opportunity there to establish an english school as some neighborhood has the same Per Capita Income and HDI from south area. Off course we will need to set a range for the price of the monthly subscription.

- 6th Discussion: At least, the result of K-Means is remarkably interesting! Look at the markers on the coast area (those are the south area) that are very asked from tourists to have a stay in RJ. Off course there are many english schools around and we can check on this later. But the most interesting cluster is the "0" (zero) that point some area with similar metrics with good per capita income rate.

- 7th Discussion: The cluster 0 provided by K-Means was considered the "middle term" of the neighborhood in Rio de Janeiro. They are not totally in the south area that has many competitions and the cost to start any business is high (ex: rentals) and they are not in the worse area of the city. To help to understand a bit more, I selected the cluster 0 result and applied a dendrogram to cluster the similar neighborhoods inside the cluster. The result was particularly good. The dendrogram creates 3 big clusters, the first

one on the top is a kind of outlier. Urca and Botafogo are in the south area and it is any square is very disputed. Jardim Guanabara is a a good option as it is not in the south area and most of the people that lives there has good living conditions. The second cluster is a good option too, excluding Jardim Sulacap and Vidigal. The first one is too far from downtown and it is a kind of army village and the second one is a slum located in the south area. The other ones: Tijuca, Maracanã, Alto da Boa Vista and Glória can be considered a good option to start the business.

- 8th Discussion: Looking at the data provided by the Foursquare API, it is possible to see, the 10 most frequents venues does not have any language school that can be a competitor against my customer. That is cool! Any of this neighborhood probably is a good location to start their business. To help to choose one location we will consider some human experience about the city and the neighborhoods. But, before that, I will check if there are any language school that it is not fitted in the tens and do some other checks looking for better infrastructure such metros, trams, and restaurants.

- 9th Discussion: Some insights from the Foursquare data:

    No english schools registered in the Foursquare API for those selected neighborhoods. But maybe there are some that is not registered in the Foursquare API (need a better check using other API or data source)

    Maracanã and Glória has some Tram Stations.

        Quick note on this: Tijuca, Glória and Maracanã already have metro and train stations that was not returned by the Foursquare API

    All of the selected neighborhoods have good restaurants to serve our customers during the classes period.

6) Discussion

The data analysis using the census data from the Rio de Janeiro city is good enough to know and understand which neighborhood offer better conditions for some investment, such opening an English school.

Although, will be more interesting enrich the data sources using more data coming from another systems, such crime rates, metro usage, traffic hours and many others.

The Foursquare API is very nice, but some places are not registered, and the results based only on that can be imprecise.

7) Conclusion

Our Customer will be advised that three of the selected neighborhoods are good options to start their first english school in Rio de Janeiro according to the data source evaluated, they are:

- 1st: Glória
  - Glória is the 1st option because it is closer to the downtown and for people that works every day in offices, the proximity can be a good differential compared to the other neighborhoods. Glória is part of the south area of the city but because of the proximity with downtown the rental prices are not too high.

- 2nd: Tijuca
  - Tijuca is the 2nd option because it is the neighborhood in the north area that is more traditional. It is served by metro stations; the per capita income is similar with some neighborhoods of the south area

(including Glória). The only cons are the distance from downtown compared to Glória.

- 3rd: Jardim Guanabara
  - Jardim Guanabara is the 3rd options because is too far from downtown and it is not good on transportations (only taxis and buses in there). The per capita income there is one of the best in the city also the HDI rate. It is a good option but only for people that lives there.