



# The Battle of Neighborhoods

---

IBM DATA SCIENCE CAPSTONE PROJECT

RODRIGO EIRAS





Rio de Janeiro

# Contents

---

1. Business Problem
2. Data
3. Methodology
4. Results
5. Discussion
6. Conclusion

# 1. Business Problem

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, cultural 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 don’t 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.

So, the problem question is: **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.**

## 2. Data

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.

# Data

---

## 2.1. 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 the table below:

Field	Description
id	A unique string identifier for this venue.
name	The best known name for this venue.
location	An object containing none, some, or all of <code>address</code> (street address), <code>crossStreet</code> , <code>city</code> , <code>state</code> , <code>postalCode</code> , <code>country</code> , <code>lat</code> , <code>lng</code> , and <code>distance</code> . All fields are strings, except for <code>lat</code> , <code>lng</code> , and <code>distance</code> . Distance is measured in meters. Some venues have their locations intentionally hidden for privacy reasons (such as private residences). If this is the case, the parameter <code>isFuzzed</code> will be set to true, and the <code>lat/lng</code> parameters will have reduced precision.
categories	An array, possibly empty, of <code>categories</code> that have been applied to this venue. One of the categories will have a <code>primary</code> field indicating that it is the primary category for the venue. For the complete category tree, see <code>categories</code> .

**Figure 1.** Information contained in response to request towards “explore” endpoint

# Data

---

## 2.2. 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](https://pt.wikipedia.org/wiki/Lista_de_bairros_do_Rio_de_Janeiro_por_IDH)
- The table in Wikipedia doesn't have the location values (lat/long) so I need to geocode it through geopy library.

# 3. Methodology

---

To be done in Week 5



# 4. Results

---

To be done in Week 5

# 5. Discussion

---

To be done in Week 5

# 6. Conclusion

---

To be done in Week 5