

**Capstone Applied Data Science: Battle of neighborhoods Rio de Janeiro**

Business Understanding and Data: December, 2018 – Marcelo Tueiv

## Introduction

Natural, sustainable and transparent consumption will no longer be an exclusive and luxury philosophy to become daily and widespread. Personal and planet care is the strongest trend in response to the stressful and frenetic modern life. And it is what will also determine the new demands of the customers. In the city of Rio de Janeiro this trend will become even stronger, where people have a greater concern with appearance and sports, because their bodies are more exposed on the beaches and because of the heat, where the themperatures reach 45 degrees in the summer.

## Objective

The objective of this project is to explore the possibilities of where to open a healthy food restaurant in the city of Rio de Janeiro and have a greater chance of success due to the characteristics of the place (transportation, proximity to gyms and competition from other restaurants). To do that we will analyze other successful healthy food restaurants characteristics and suggest places to open a new one following these characteristics.

### Stakeholders

Future entrepreneurs who wants to opent healthy food restaurant in Rio de Janeiro

### Selection criteria

The location of the establishment can guarantee the success of sales and, for this, following some behaviors identified in Rio de Janeiro we have, as points indicated for the opening of a restaurant of this type some factors that increase the chance of consumer frequency. We imagine that, some factors to be validated by the result of the approach we will implement can be (to be evaluated by the project):  
- Must be located within 500 meters of the subway  
- High income neighborhood  
- Proximity of more number of gyms  
- Proximity of the smallest possible number of restaurants of this type, with good evaluation

Initially, for this, we must obtain 3 clusters by grouping the neighborhoods as follows:  
- Cluster 0: neighborhoods that indicate a high chance of success of opening a restaurant of this type  
- Cluster1: neighborhoods that indicate little chance of success opening a restaurant of this type  
- Cluster 2: neighborhoods that are neither attractive nor negative for opening a restaurant of this type

## Data

This project will require the following data sources, namely:

**Rio de Janeiro Subway stations**

To identify proximity to subway stations in the city of Rio de Janeiro we will consult page table in Wikipedia with list of all the subway stations of the city and their references of longitude and latitude:

<https://pt.wikipedia.org/wiki/Lista_de_esta%C3%A7%C3%B5es_do_Metr%C3%B4_do_Rio_de_Janeiro>

Sample:



**Neighborhoods and socio-economics**

To identify the neighborhoods and socio-economic data as average income we will consult this other list, also existing in Wikipedia:

<https://pt.wikipedia.org/wiki/Lista_de_bairros_do_Rio_de_Janeiro_por_IDH>

Sample:



**Commercial Establishments Nearby**

Foursquare is a location data provider with information and evaluation about restaurants, hotels, gyms and other venues and events within an area of interest. Such information includes venue names, locations, evaluation, latitude and longitude. Foursquare information of these venues will be obtained via Foursquare API.

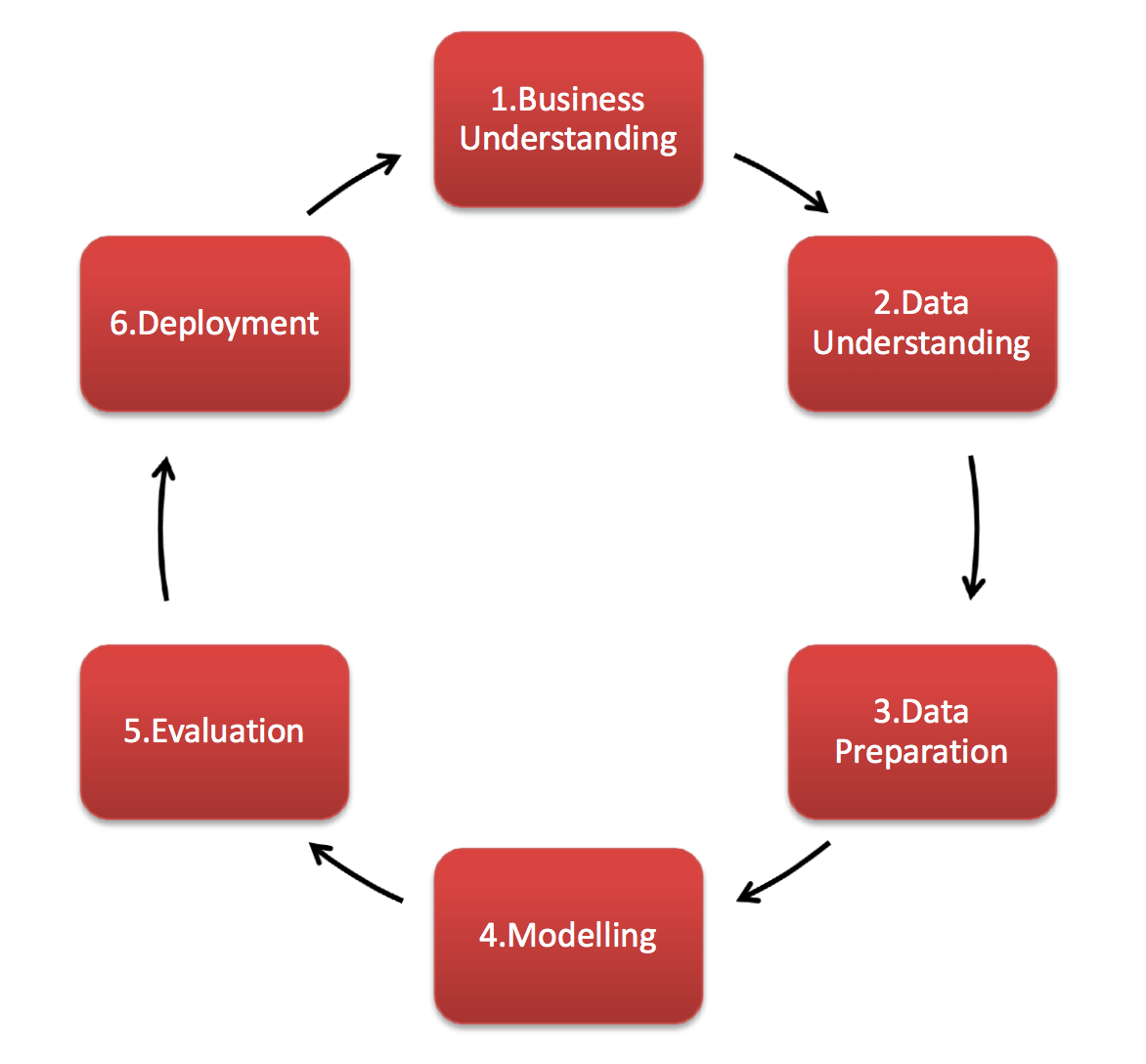
**Other Data Source, if necessary**

Another rich source of information is the open data site of the city of Rio de Janeiro, where we can obtain information such as rate of violence and commercial value of real estate:

<http://www.data.rio/>

## Methodology

For this project, it will be used CRIP-DM methodology as reference of activities to performed. In this methodology, the following steps must be performed:



**1.** **Business Understanding**

Focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition and a preliminary plan.

**2.** **Data Understanding**

Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

**3.** **Data Preparation**

The data preparation phase covers all activities to construct the final dataset from the initial raw data.

**4.** **Modeling**

Modeling techniques are selected and applied.  Since some techniques like neural nets have specific requirements regarding the form of the data, there can be a loop back here to data prep.

**5.** **Evaluation**

Once one or more models have been built that appear to have high quality based on whichever loss functions have been selected, these need to be tested to ensure they generalize against unseen data and that all key business issues have been sufficiently considered.  The end result is the selection of the champion model(s).

**6.** **Deployment**

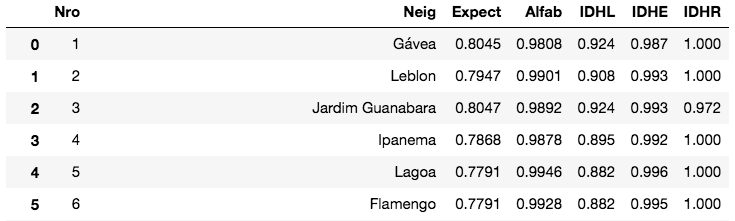
Generally this will mean deploying a code representation of the model into an operating system to score or categorize new unseen data as it arises and to create a mechanism for the use of that new information in the solution of the original business problem.  Importantly, the code representation must also include all the data prep steps leading up to modeling so that the model will treat new raw data in the same manner as during model development.

Reference: <https://www.datasciencecentral.com>

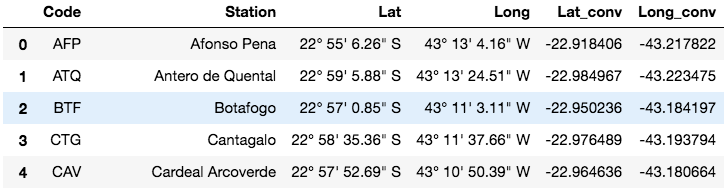
## Data Preparation

The following data cleaning, normalization and aggregation were needed to allow correct use of the data mentioned above:

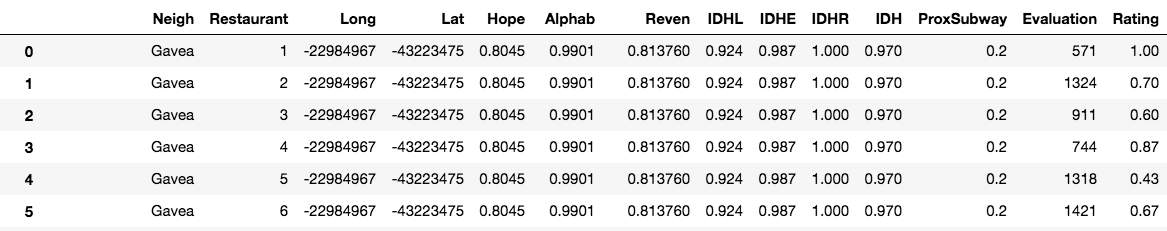
1. Socio Economic Data:
   1. It was necessary to split to change the numbers to ranges of 0 to 1, in this way it would be possible to cluster information and compare between then
   2. Unnecessary columns removed



1. Subway table
   1. It was necessary to convert longitude and latitude from degrees, minutes and seconds to decimal, and in this way, it could be plotted in the graphic
   2. It was necessary to create functions to obtain the neighborhoods of an specific geolocation of Venues



1. Get Venue Information
   1. Foursquare accessed to get information of Vegan and Salad restaurants in Rio de Janerio, based in the above lists
   2. Foursquare accessed to get detail information of ranking and like counts, to enrich restaurant information
   3. Join this information with Socio Economic Information



## Data Analysis

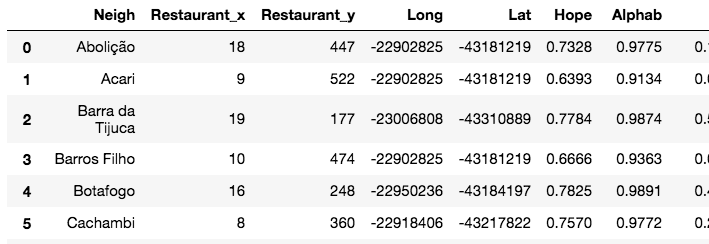
Some analysis were made in the data to understand the distribution of restaurant in the neighborhood. For example, how many Salad and Vegan restaurants were returned for each neighborhood



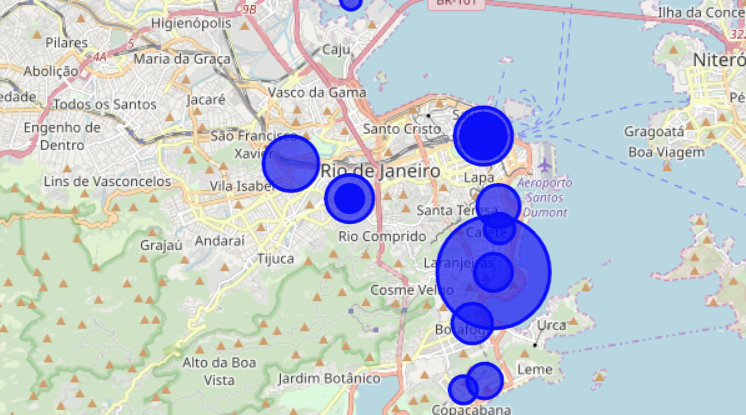
## Visualization

The concentration of restaurant by neighborhood were presented in the map.

To do that, first it was necessary to aggregate gps information to neighbohoods:

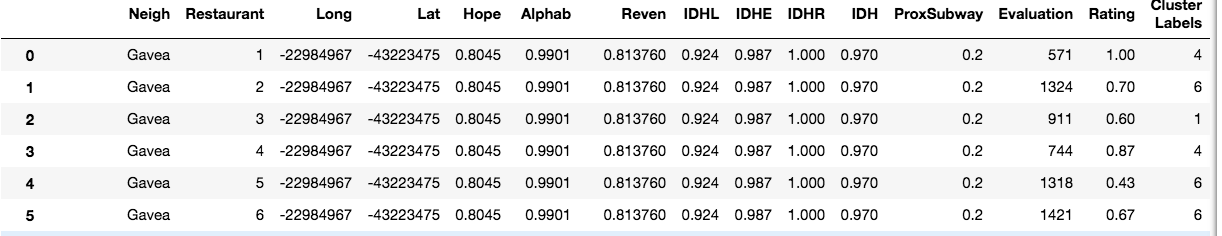


And then, present in circles the concentration of restaurants in each neighborhood:



## Clustering

Finally, we clustered the restaurant in groups with the characteristics of the best features to have a restaurant (ex.: proximity to subway, customer evaluation and socio economic information.



## Conclusion

It was possible to classify the restaurant in groups that allow to analyze what characteristics influentiate in the success (evaluation and revenue) of them, example, proximity of subway and socio economic factors. The model will allow to entrepreneurs evaluate what characteristics are important to have and where in Rio is the best place to open a restaurant of this category.