# Moving to Rio de Janeiro from São Paulo

Marcelo Rodrigues

November, 2020

## 1. Introduction

### 1.1 Background

The city of Rio de Janeiro/Brazil is very big and has more than 150 " neighborhoods", if one is moving from an specific neighborhood in São Paulo to Rio de Janeiro and wants to live in a neighborhood that at least looks like the one she/he lived before, where should they search places?

### 1.2 Problem

One is moving from the neighborhood "Moema", in São Paulo and never been in Rio de Janeiro before, this person works and study "Home Office" so he/she doesn't care for work/university distance. Which neighborhoods he/she should search in Rio de Janeiro? Which are the most looks like neighborhoods? So We will separate neighborhoods by Cluster and compare which cluster has more "Common Venues" to Moema, São Paulo. With this one will have a leaner list of neighborhoods to look.

### 1.3 Question

Which neighborhoods should be considered if someone moves from Moema to Rio de Janeiro?

## 2. Data Acquisition and Cleaning

### 2.1 Data Sources

Through the Foursquare API we are able to verify which "venues" are the most common in the Moema region. For that, we only need to obtain the Latitude and Longitude of the neighborhood through the "Geolocator" library. In a similar way, we can get the list of all neighborhoods in the city of Rio de Janeiro through Wikipedia, obtain their respective longitudes and latitudes through the "Geolocator" library and finally, discover the most common "venues" of each neighborhood.

### 2.2 Data Cleaning

It is necessary to clean the source data of the Neighborhood List obtained through Wikipedia. The information source has some neighborhoods on the same line, we need to "explode" those neighborhoods that share lines for new lines. In addition, several columns in the neighborhoods of Rio de Janeiro are not used, and we can use a "drop" on them. Finally, we found that some neighborhoods do not have any information about "venues" on Foursquare, so they do not have common "venues". These neighborhoods were removed from the database and disregarded in the analysis.

## 3. Metodology

### 3.1 Approach

As a solution to our problem, and to minimize the total of 150 potential neighborhoods that our client could live in, we chose to cluster the neighborhoods in Rio de Janeiro, based on the similarity of "venues". Once clustered, we could choose the cluster that would be closest to the customer's original neighborhood (Moema, São Paulo / SP). In this way we would be able to minimize the total number of neighborhoods to be studied by the client.

### 3.2 Strand

To cluster our neighborhoods we will use the "K-means" solution. “K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.”(GARBADE).

“To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

The centroids have stabilized — there is no change in their values because the clustering has been successful.

The defined number of iterations has been achieved.”(GARBADE).

## 4. Results and Discussions

We were able to divide the whole city of Rio de Janeiro in 5 differents clusters as we may seen above:

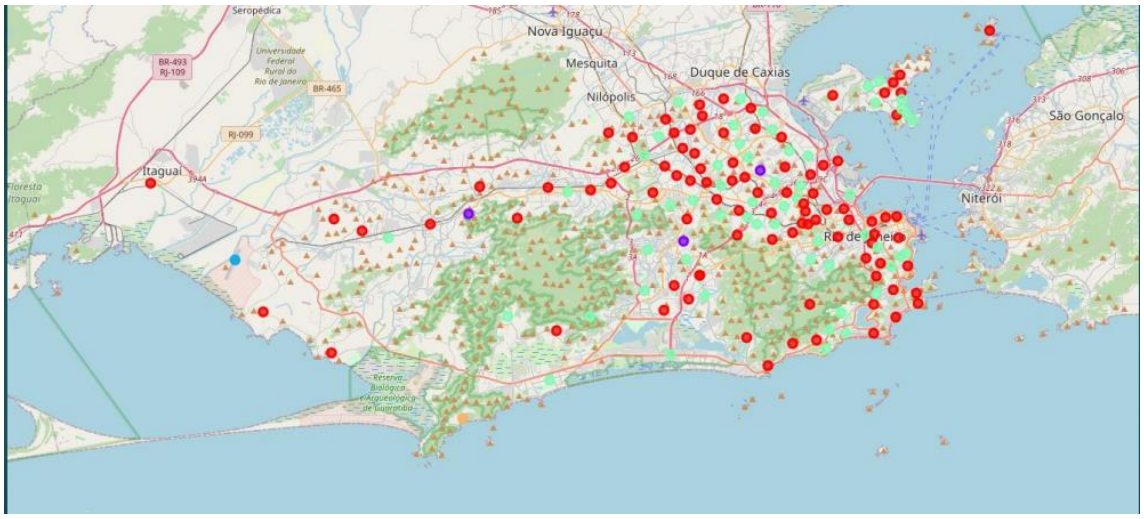


Image 1 – Rio de Janeiro Neighbourhoods Clustered

After analyse the results, we Checked that neighbourhoods from cluster 3 are the most similars to Moema, São Paulo/SP.

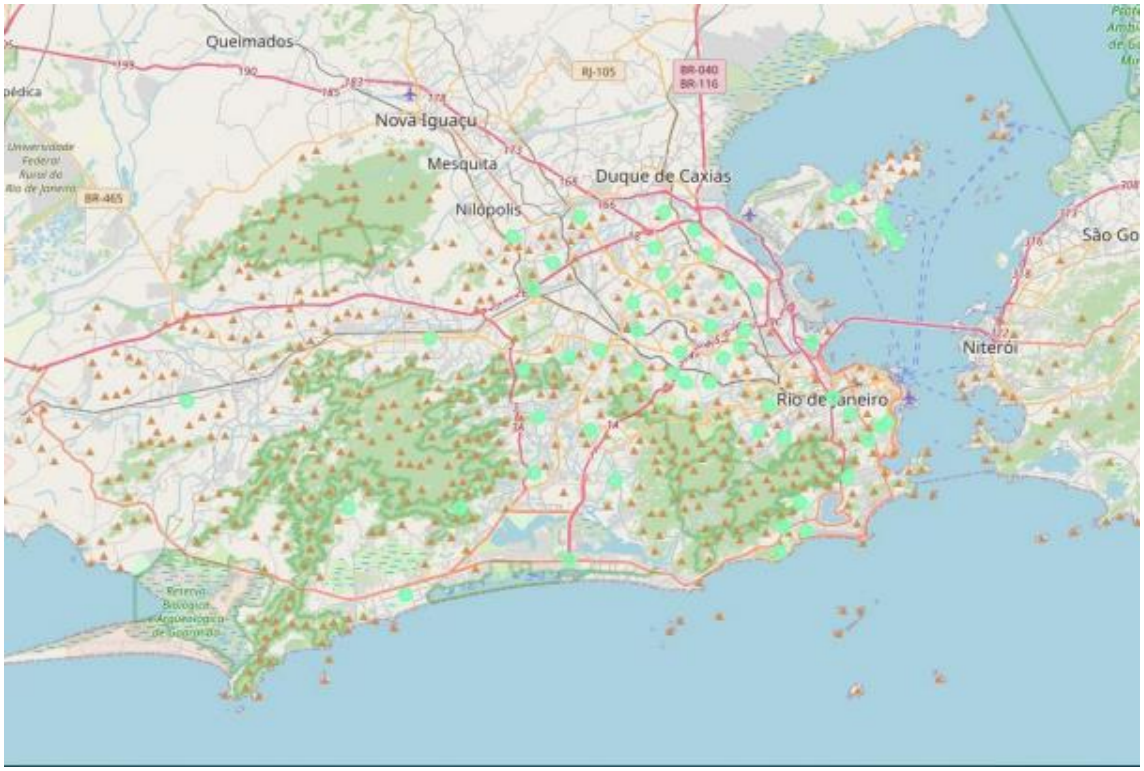


Image 2 – Most similars Neighbourhoods

Therefore, we were able to decrease the original Dataset from approximately 160 Neighbourhoods to just 55:

| | Neighbourhood | Latitude | Longitude | Cluster Labels |
|---|---|---|---|---|
| **0** | Barra da Tijuca | -22.999740 | -43.365993 | 3 |
| **2** | Gávea | -22.981424 | -43.238324 | 3 |
| **3** | Leblon | -22.983556 | -43.224938 | 3 |
| **4** | Jardim Guanabara | -22.812836 | -43.200779 | 3 |
| **8** | Humaitá | -22.954641 | -43.200480 | 3 |

Image 3 – Sample of most similars Neighbourhoods output

## 5. Bibliography

Understanding K-means Clustering in Machine Learning

(Garbade)

https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

## 6. Other Infos

Github: https://github.com/marceloasr94/Coursera_Capstone