Joana Soares Blanco Enes

Capstone Project - The Battle of Neighborhoods

Where do I open a Bar in Tijuca, Rio de Janeiro?

Rio de janeiro - RJ

2021

Introduction

According to Wikipedia, Tijuca is a neighborhood in the North Zone of the city of Rio de Janeiro. It is among the oldest, most traditional and populous neighborhoods in the capital of Rio. Its quality of life index in 2000 was 0.887, the 18th best in the city, among 126 neighborhoods evaluated, considered high. According to data from 2010, it had 163,805 inhabitants, the largest in the North Zone. In the ranking of the most valued neighborhoods in the municipality, Tijuca occupies the 20th position.

Due to its varied transport options, it has become a lodging option for national and foreign tourists.

In gastronomy, the Tijuca is not far behind either, it has 3 traditionally adapted gastronomic poles : Praça Varnhagen, Rua Uruguai and Rua Mariz e Barros, and there is also a Saens Peña Square with many bars, patisseries and restaurants.

Such a populous neighborhood with a rich infrastructure is not difficult to imagine that it will be considered for opening new venues.

Business Problem

As everyone knows, choosing a location for a new business is an extremely difficult and challenging decision, the difference between achieving or not achieving positive results in a new venture can come from the selection of the commercial point. It is opportune to always think about having a favorable and more strategic place as possible. But how do you know where this place is?

This work proposes to bring, through the data extracted from Foursquare and the unsupervised machine learning algorithm 'k-means clustering', a better understanding of how the bars in the Tijuca neighborhood are distributed and which places would be more suitable for opening a new one.

The objective to be achieved here is to be able to provide enough information so that investors can clearly have the most interesting points of Tijuca for the opening of a new bar.

Data

To start data collection, we extracted information from wikipedia about the name of the neighborhoods that make up the “grande tijuca”.

With the help of the 'geopy' library, we found the latitudes and longitudes of these neighborhoods, and this information was used to search the foursquare API for locations in this neighborhood within a pre-established radius. Then we include a column to categorize the data that is of interest to us for study and save it to a specific dataframe.

	Borough	Latitude	Longitude	Venue	Venue_ID	Venue_Category	Venue_Latitude	Venue_Longitude	Prime_Category
3	Andaraí, RJ	-22.929084	-43.253486	Boteco Vou Demais	50ae3886e4b004e6d186cef9	Dive Bar	-22.926412	-43.253423	Bar
7	Andaraí, RJ	-22.929084	-43.253486	Bar Santo Remédio	4ef782f94fc6f26c26151ee3	Bar	-22.924775	-43.256872	Bar
10	Andaraí, RJ	-22.929084	-43.253486	Buteco do Barão	5467f32d498ea0e865527fa0	Bar	-22.926348	-43.247458	Bar
12	Andaraí, RJ	-22.929084	-43.253486	Boteco do Raoni	5a78c084491be712e2b4666b	Beer Bar	-22.924216	-43.254592	Bar
13	Andaraí, RJ	-22.929084	-43.253486	Bardot Vinhos e Artes	4d9659d20caaa143ec9975b3	Wine Bar	-22.923575	-43.256350	Bar
...	...	...	...	...	...	...	...	...	...
397	Vila Isabel, RJ	-22.915222	-43.247263	Yeasteria Ponto Cervejeiro	534186f8498e8a823c5665a3	Bar	-22.918708	-43.239297	Bar
401	Vila Isabel, RJ	-22.915222	-43.247263	Bar Gente Nossa	4cd716eafb718eecf7964c88	Bar	-22.913641	-43.244185	Bar
437	Vila Isabel, RJ	-22.915222	-43.247263	Edson Freitas Haute Coiffeur	4bc4e94e5e0ab7f3a7aa45eb	Salon / Barbershop	-22.919365	-43.250501	Bar
459	Vila Isabel, RJ	-22.915222	-43.247263	Varandão MT73	4cf0315111d18a143ebb74aec	Bar	-22.919006	-43.253883	Bar
470	Vila Isabel, RJ	-22.915222	-43.247263	Nosso Bar	4c818dd0d34ca14325ce2380	Bar	-22.912829	-43.238636	Bar

78 rows × 9 columns

The dataframe has 78 venues.

Clustering

With the necessary information in hand, we will work on the venues based on their location using K-means.

K-means is an unsupervised machine learning algorithm for clustering data. It aims to separate data into “k” clusters, based on the distance from each point to be centroid.

According to Hugo Honda, “The task of the algorithm is to find the nearest centroid (using some distance metric) and assign the point found to that cluster. After this step, the centroids are specialized, always taking the average value of all points in that cluster. For this method, numerical values are considered for the distance calculation, the nominal values can then be mapped into binary values for the same calculation. In case of success, the data are separated organically and can be labeled and centroid referenced to classify new data. "

To evaluate the ideal “K”, the Elbow and Silhouette method will be used.

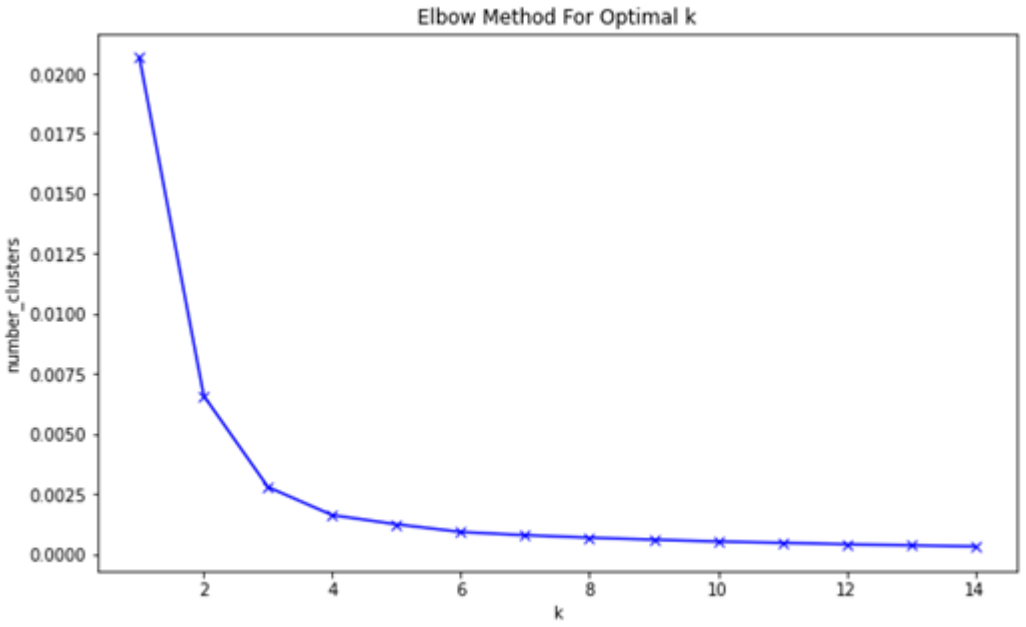
The elbow method is a very simple and widely used method to assess the best number of clusters needed. Tests the variation of the data in relation to the number of clusters. It is considered an ideal k value when the increase in the number of clusters does not represent a significant gain value.

About Silhouette, based on wikipeadia, it refers to a method of interpretation and validation of consistency within data clusters. The technique provides a succinct graphical representation of how well each object has been classified.

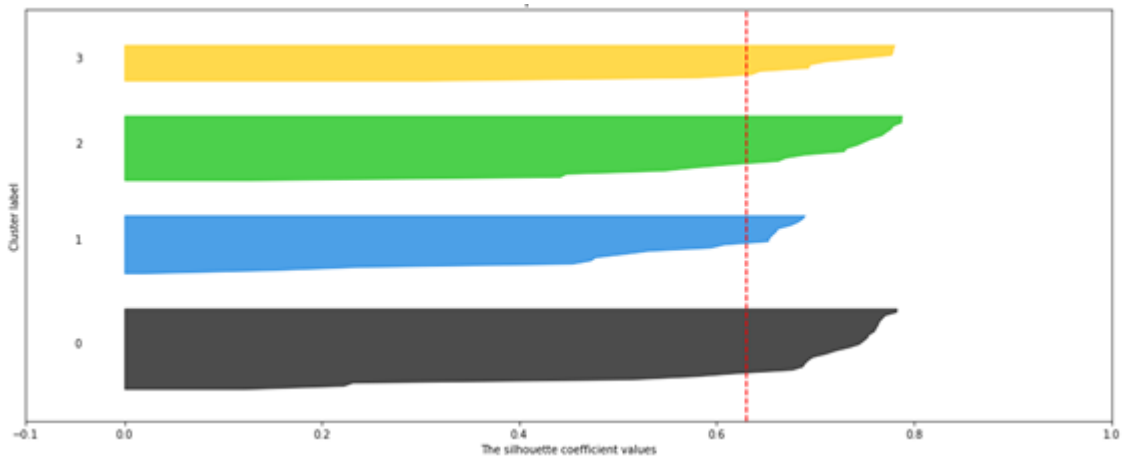
The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from - to +1, where a high value indicates that the object blends well with its own cluster and hardly matches neighboring clusters. If most objects have a high value, the cluster configuration is appropriate. If many points have a low or negative value, the cluster configuration can have many or few clusters.

The silhouette can be calculated with any distance metric, such as Euclidean distance or Manhattan distance.

Follow the results of the evaluations:

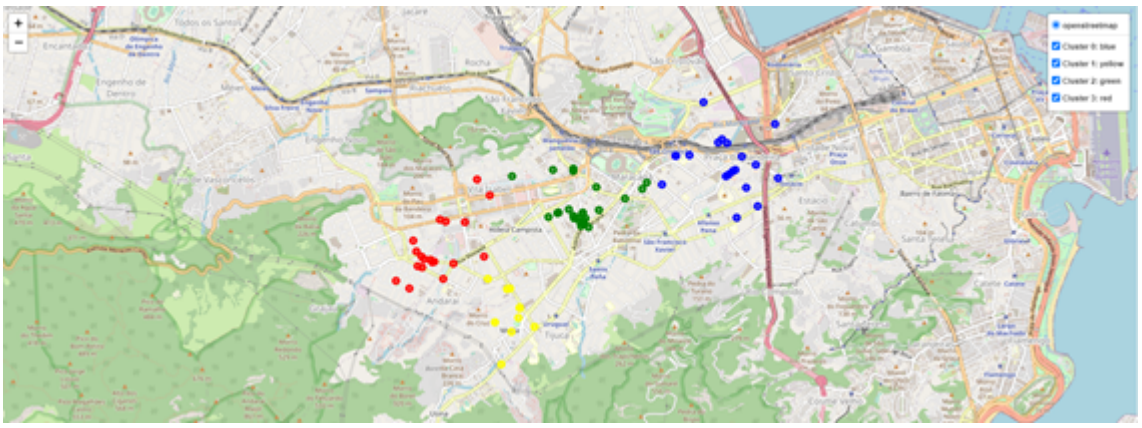



```
For n_clusters = 3 The average silhouette_score is : 0.5964544700229589
For n_clusters = 4 The average silhouette_score is : 0.6301132504093082
For n_clusters = 5 The average silhouette_score is : 0.5868906454055728
```



Based on these two results, we conclude that k = 4 is the best option.

We plotted the information about the clusters on a map for a better understanding of the distribution as you can see below:



Analysis of the results

Below you will individually evaluate the clusters found:

1. ***Cluster 0*** - *21 bars*

   It is the second most populous cluster and stands out from the others for having the highest rating and the best lowest score.

2. **Cluster 1** - *12 bars*

   It is the cluster with the least amount of bars. It has no prominence as to their ratings.

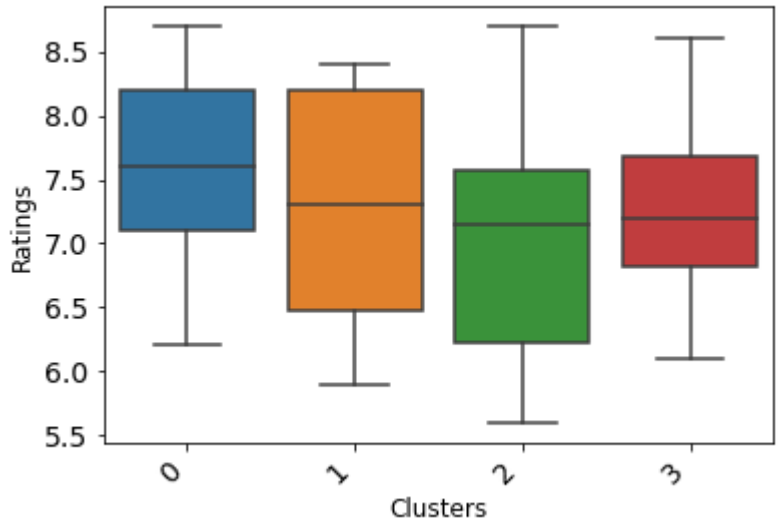
3. **Cluster 2** - *26 bars*

   It is the most populous cluster and has the greatest oscillation of counts. This cluster has the best and worst score of all locations.

4. **Cluster 3** - *19 bars*

   It is a cluster very similar to cluster 0 in terms of the number of bars and note distribution. It had the second highest average of evaluations.

For a better visualization of this information, a boxplot for each cluster follows and its average rating:



Ratings
7.604762
7.266667
7.023077
7.283333

Conclusion

It can be concluded that the best place to open a bar in Tijuca based on the data is the location inserted in cluster 0. This result is interesting because this region is formed by the neighborhoods Praça da Bandeira and Maracanã, fleeing a little from the most populated areas close to gastronomic centers.

We believe that this result is the result of a work to revitalize the neighborhoods and improve the infrastructure that solved a problem of years in this region with constant floods.

The place mixes traditional and new bars and is undoubtedly an aspiration to become another gastronomic hub of Tijuca that further reinforces our discoveries.