Rio de Janeiro's Neighborhoods Clustering and HDI Analysis

Enzo Casemiro Zuccaro

February 4, 2020

IBM Data Science Professional Certificate

Table of Contents

1 Introduction

1.1 Background

Rio de Janeiro is one of the biggest and probably the most famous brazilian city. With more than six million people, Rio de Janeiro is a common vacation destiny, attracting thousands of tourists every year. Most people don't realize how big the city is, many neighborhoods are unknown for the majority of visitors. Another aspect of the city that is not clear for many is Rio's social inequality, which can be exposed by making a Human Development Index (HDI) analysis of each neighborhood. The neighborhoods' HDI ranges from 0.970, higher than the HDI of countries like Norway and Switzerland, to 0.700, more or less equivalent to countries like Egypt, Gabon and Vietnam. Violence directly impacts on these values, it affects everyone's lives in the neighborhoods with the lowest HDI.

1.2 Problem

This work aims to verify the relationship between the HDI of Rio de Janeiro's neighborhoods and the number of venues of each one. Also, a cluster analysis will be made to observe similarities between some neighborhoods based on the kind of the venues present in each one.

1.3 Interest

Public authorities and academic institutions would be the most interested in this work.

2 Data Acquisition and Cleaning

2.1 Data Sources

All data related to Rio's neighborhoods was acquired in the following wikipedia article: `https://pt.wikipedia.org/wiki/Lista_de_bairros_do_Rio_de_Janeiro_por_IDH`. The venues information were acquired using the Foursquare API and the coordinates of each neighborhoods were acquired using GeoPy library.

2.2 Data Cleaning

The data cleaning consisted basically in removing the rows of the data frame found in the article that weren't neighborhoods, removing the neighborhood Mangueira, whose coordinates weren't correctly found by GeoPy, removing the duplicates in the venues data frame and removing the neighborhoods with no venues located by the Foursquare api for the cluster analysis. Also, the HDI values needed to be converted into float type and the columns name were translated to English.

2.3 Feature Selection

From the data frame of the article, only the neighborhoods' names and the HDI values were selected, it was also necessary to separate the rows containing multiple neighborhoods (the ones with the same HDI). After that, GeoPy was used to gather the coordinates of each neighborhood. Having the names of the neighborhoods and the coordinates, it was possible to find the venues in a 1km radius using the Foursquare API.

3 Methodology

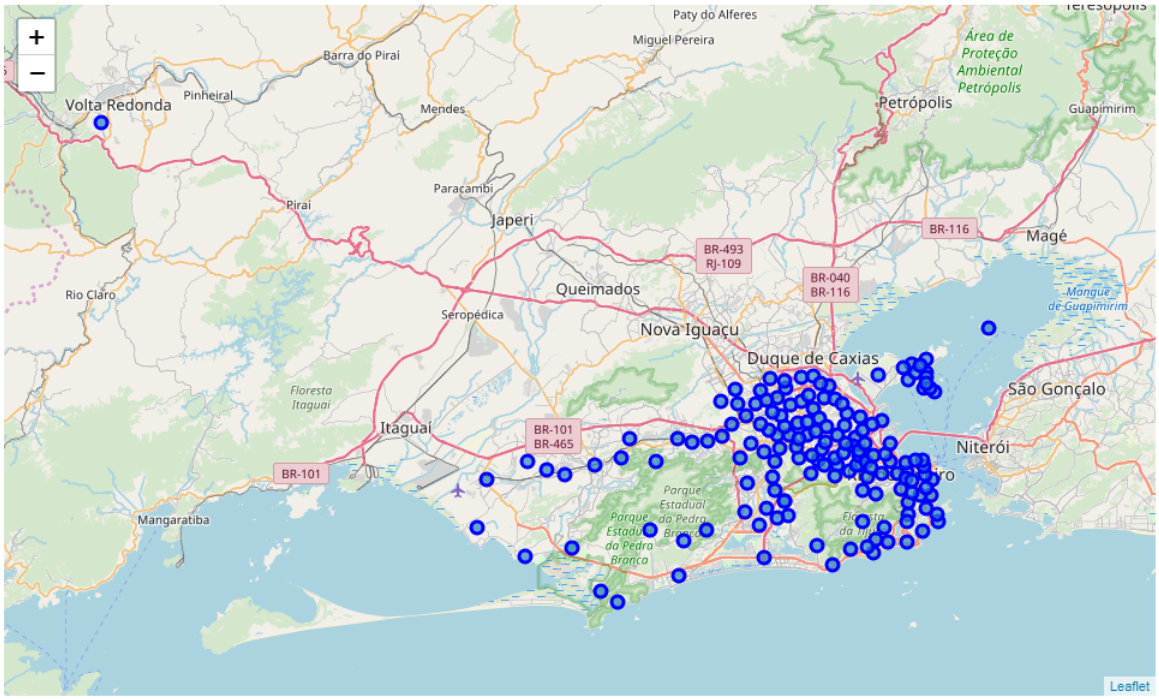
After the initial cleaning of the data set from the Wikipedia article, the following table containing the names and the HDI from each neighborhood was acquired.

	Neighborhood	HDI
0	Gávea	0.970
1	Leblon	0.967
2	Jardim Guanabara	0.963
3	Ipanema	0.962
4	Humaitá	0.959

The GeoPy library was then used to acquire the coordinates of the neighborhoods.

	Neighborhood	HDI	Latitude	Longitude
0	Gávea	0.970	-22.981424	-43.238324
1	Leblon	0.967	-22.983556	-43.224938
2	Jardim Guanabara	0.963	-22.812836	-43.200779
3	Ipanema	0.962	-22.983956	-43.202216
4	Humaitá	0.959	-22.954641	-43.200480

With the names and the coordinates is possible to visualize each neighborhood with Folium and verify if the coordinates are correct.



The neighborhood Mangueira is misplaced, GeoPy got the coordinates of a neighborhood with the same name of another city. Since it's not located in Rio de Janeiro, this neighborhood was removed from the data set.

It's also possible to acquire the venues using the coordinates of the 158 neighborhoods of Rio de Janeiro, for this the Foursquare API is used. For each neighborhood, a limit of 100 venues and 1km radius was used. The following data set contains the venues data, it shows venues names, coordinates and categories.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Gávea	-22.981424	-43.238324	Instituto Moreira Salles (IMS)	-22.981627	-43.239637	Cultural Center
1	Gávea	-22.981424	-43.238324	Empório Jardim	-22.981632	-43.239682	Breakfast Spot
2	Gávea	-22.981424	-43.238324	Sociedade Germania	-22.980419	-43.239628	Athletics & Sports
3	Gávea	-22.981424	-43.238324	Bosque da PUC	-22.980186	-43.234375	College Quad
4	Gávea	-22.981424	-43.238324	Parque da Cidade	-22.981210	-43.242434	Park

Since one of the objectives is to do a cluster analysis of the neighborhoods using the k-means algorithm, it was necessary to apply a one hot encoding method in the Category column, group by Neighborhood and take the mean of the values in order to obtain the frequency that each category appears in each neighborhood. The following table was obtained.

	Neighborhood	ATM	Acai House	Accessories Store	Adult Boutique	Airport Lounge	American Restaurant	Amphitheater	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Entei
0	Abolição	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	Acari	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	Alto da Boa Vista	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	Anchieta	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	Andaraí	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Using this data set it was possible to create a new table with the 10 most common venues of each neighborhood.

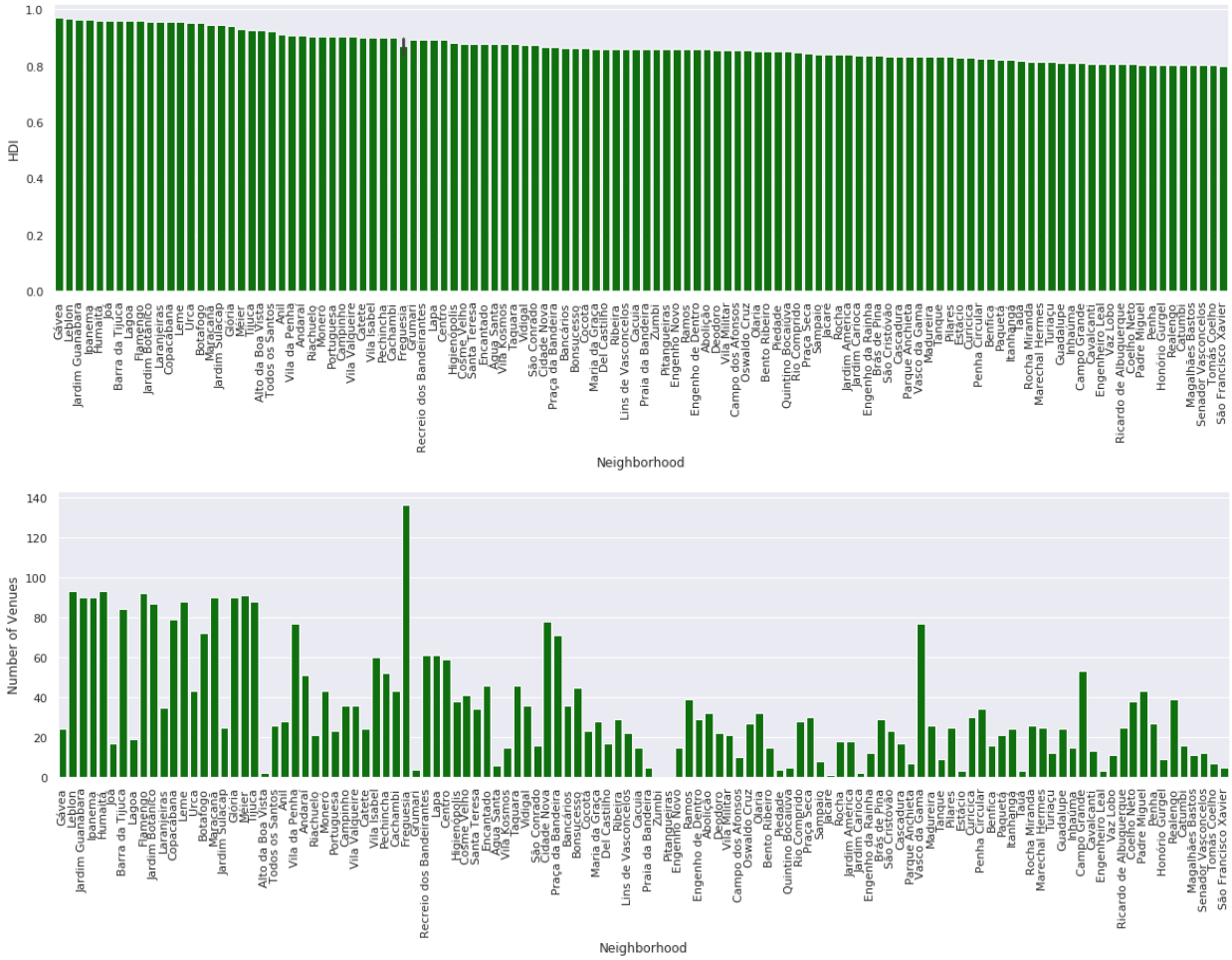
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abolição	BBQ Joint	Food Truck	Soccer Field	Bar	Deli / Bodega	Gym / Fitness Center	Gym	Music Venue	General Entertainment	Burger Joint
1	Acari	Market	Plaza	Burger Joint	Bus Station	Buffet	Supermarket	Brazilian Restaurant	Gym / Fitness Center	Soccer Field	Snack Place
2	Alto da Boa Vista	Mountain	Scenic Lookout	Zoo	French Restaurant	Food & Drink Shop	Food Court	Food Service	Food Stand	Food Truck	Football Stadium
3	Anchieta	Plaza	Gym / Fitness Center	Pizza Place	Fast Food Restaurant	Nightclub	Fruit & Vegetable Store	Gym	Pet Store	Church	General Entertainment
4	Andaraí	Bar	Bakery	Gym / Fitness Center	Food Truck	Japanese Restaurant	Music Venue	Gym	Pizza Place	Farmers Market	Plaza

4 Results

4.1 HDI Analysis

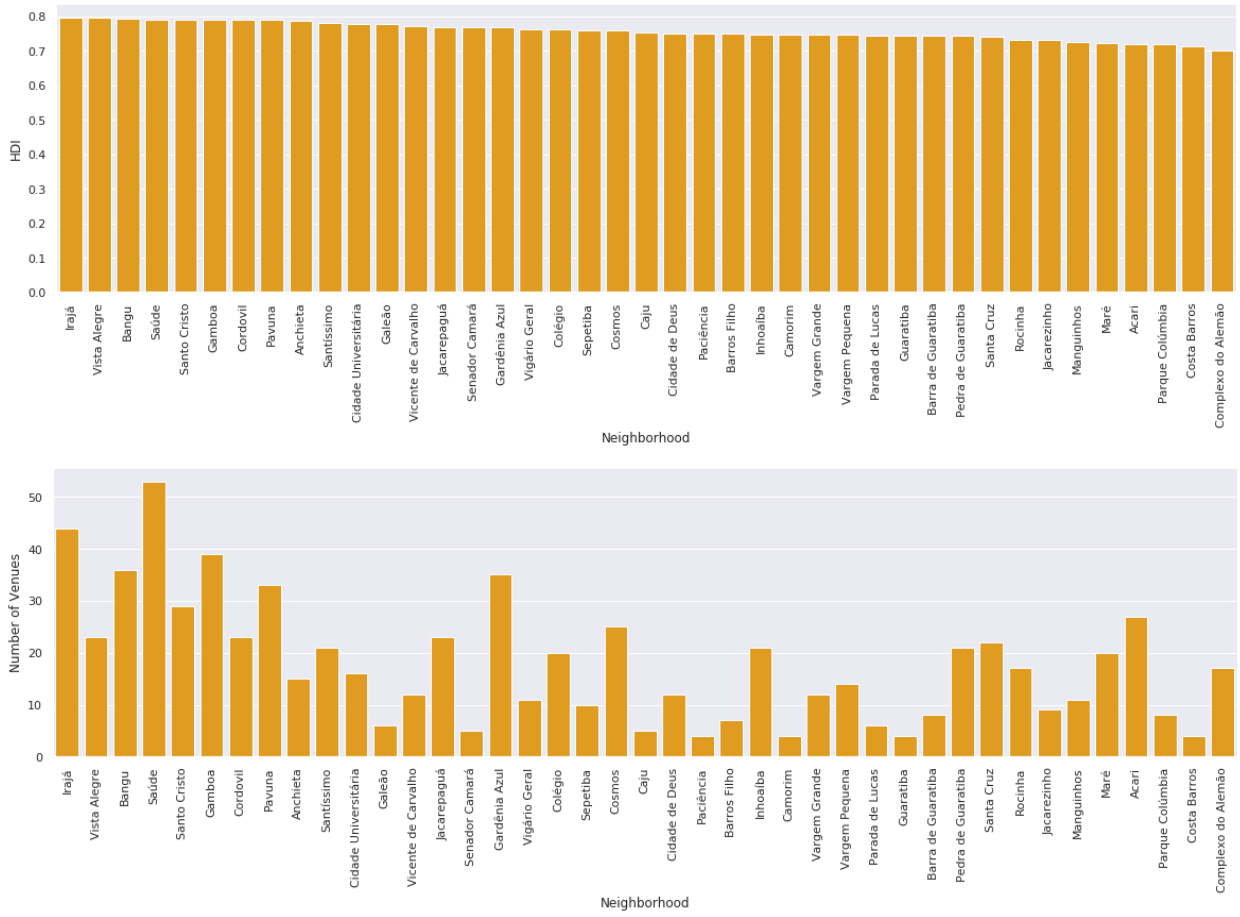
Before making the HDI analysis a new column was added to the first data set. The new column was the Number of Venues, making possible the comparison between this column and the HDI.

- **Neighborhoods with a HDI equal or higher than 0.8**



For these neighborhoods the average number of venues is 35.40

- **Neighborhoods with a HDI lower than 0.8**



For these neighborhoods the average number of venues is 17.85

4.2 Clustering

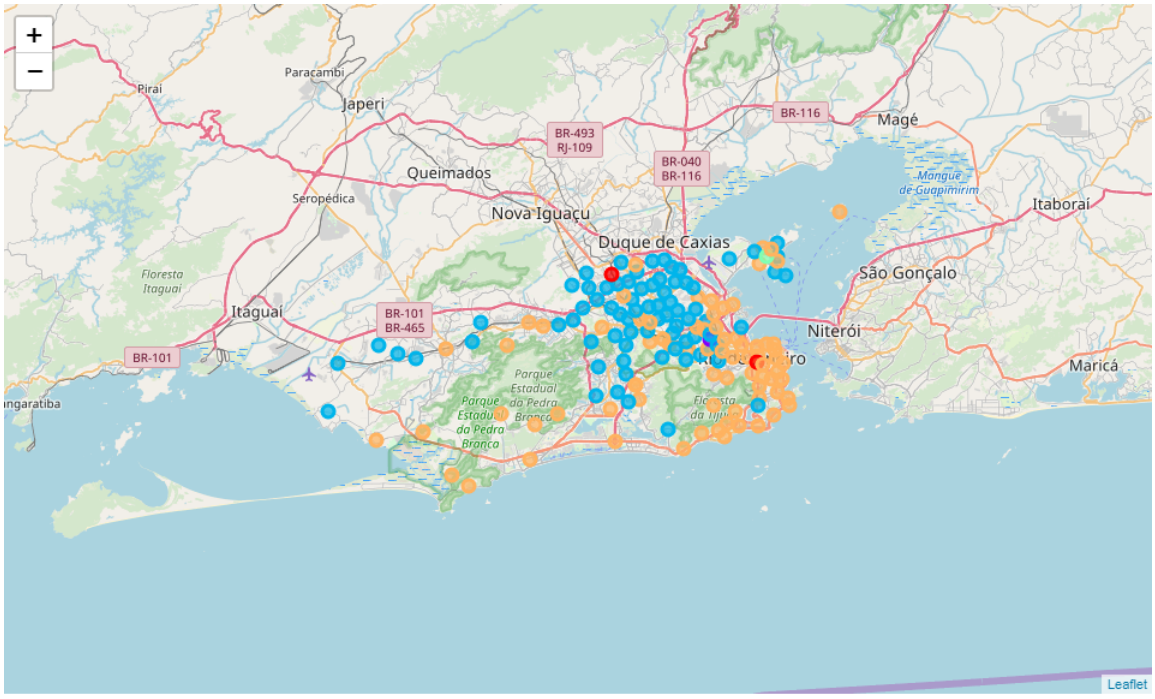
Five clusters were used to divide the neighborhoods using the k-means algorithm. Each neighborhood was assigned to a cluster and the result is shown in the data set containing

the 10 most common venues.

	Neighborhood	HDI	Latitude	Longitude	Number of Venues	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Mo Commo Venu
0	Gávea	0.970	-22.981424	-43.238324	24.0	4.0	Sushi Restaurant	Park	Cultural Center	Trail	Theater	Gym	Brazilian Restaurant	Fruit Vegetab Sto
1	Leblon	0.967	-22.983556	-43.224938	93.0	4.0	Brazilian Restaurant	Steakhouse	Bar	Restaurant	Pizza Place	Italian Restaurant	Juice Bar	Frenc Restaura
2	Jardim Guanabara	0.963	-22.812836	-43.200779	90.0	4.0	Brazilian Restaurant	Gym	Pizza Place	Burger Joint	Beach Bar	Japanese Restaurant	Bakery	B
3	Ipanema	0.962	-22.983956	-43.202216	90.0	4.0	Food Stand	Brazilian Restaurant	Pizza Place	Bar	Italian Restaurant	Japanese Restaurant	Hostel	Hot
4	Humaitá	0.959	-22.954641	-43.200480	93.0	4.0	Bar	Brazilian Restaurant	Café	Vegetarian / Vegan Restaurant	Nightclub	Plaza	Pie Shop	Restaura

The neighborhoods that have no venues got no cluster label assigned to them, so they were removed from the data set.

Folium was then used again to visualize how the neighborhoods were clustered.



- **Cluster 0**

	Neighborhood	HDI	Latitude	Longitude	Number of Venues	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9t Co
90	Estácio	0.829	-22.916626	-43.203630	3.0	0	Plaza	Tunnel	Zoo	Frame Store	Food & Drink Shop	Food Court	Food Service	Food Stand	
157	Costa Barros	0.713	-22.824700	-43.369839	4.0	0	Plaza	Border Crossing	Mountain	Liquor Store	Zoo	French Restaurant	Food Court	Food Service	

The average HDI for this cluster is 0.771

- **Cluster 1**

	Neighborhood	HDI	Latitude	Longitude	Number of Venues	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th M Comm Ve
76	Jacaré	0.839	-22.893117	-43.2572	1.0	1	Market	Zoo	Flower Shop	Food & Drink Shop	Food Court	Food Service	Food Stand	Food Truck	Foo Stac

The average HDI for this cluster is 0.839

- **Cluster 2**

	Neighborhood	HDI	Latitude	Longitude	Number of Venues	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Mo Commo Venu
7	Lagoa	0.959	-22.962466	-43.202488	19.0	2	Scenic Lookout	Gym / Fitness Center	Japanese Restaurant	Trail	Dive Bar	Pool	Plaza	Playgrour
16	Jardim Sulacap	0.944	-22.895680	-43.392997	25.0	2	Gym	Plaza	Restaurant	Food Truck	Toll Booth	Tunnel	Snack Place	Bi
21	Todos os Santos	0.922	-22.894563	-43.285060	26.0	2	Soccer Stadium	Bar	Furniture / Home Store	Pizza Place	Clothing Store	Tattoo Parlor	History Museum	Shoppir Ma
23	Vila da Penha	0.909	-22.843507	-43.310058	77.0	2	Bakery	Pizza Place	Brazilian Restaurant	Bar	Plaza	Fast Food Restaurant	Japanese Restaurant	Gym Fitnes Cent
25	Riachuelo	0.905	-22.902695	-43.255175	21.0	2	Gym	Restaurant	Brazilian Restaurant	Train Station	Farmers Market	Organic Grocery	Supermarket	Gym Fitnes Cent

The average HDI for this cluster is 0.824

- **Cluster 3**

	Neighborhood	HDI	Latitude	Longitude	Number of Venues	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9 C
80	Jardim Carioca	0.836	-22.805249	-43.193195	2.0	3	Samba School	Asian Restaurant	Zoo	French Restaurant	Food & Drink Shop	Food Court	Food Service	Food Stand	

The average HDI for this cluster is 0.836

- **Cluster 4**

	Neighborhood	HDI	Latitude	Longitude	Number of Venues	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Mo Commo Venu
0	Gávea	0.970	-22.981424	-43.238324	24.0	4	Sushi Restaurant	Park	Cultural Center	Trail	Theater	Gym	Brazilian Restaurant	Fruit Vegetab Sto
1	Leblon	0.967	-22.983556	-43.224938	93.0	4	Brazilian Restaurant	Steakhouse	Bar	Restaurant	Pizza Place	Italian Restaurant	Juice Bar	Frenc Restaura
2	Jardim Guanabara	0.963	-22.812836	-43.200779	90.0	4	Brazilian Restaurant	Gym	Pizza Place	Burger Joint	Beach Bar	Japanese Restaurant	Bakery	Ba
3	Ipanema	0.962	-22.983956	-43.202216	90.0	4	Food Stand	Brazilian Restaurant	Pizza Place	Bar	Italian Restaurant	Japanese Restaurant	Hostel	Hot
4	Humaitá	0.959	-22.954641	-43.200480	93.0	4	Bar	Brazilian Restaurant	Café	Vegetarian / Vegan Restaurant	Nightclub	Plaza	Pie Shop	Restaura

The average HDI for this cluster is 0.858

5 Discussion

The results shows that there is some relationship between the HDI and the number of venues present in a neighborhood in Rio de Janeiro. For neighborhoods with a HDI equal or higher than 0.8, which is considered a high HDI, the average number of venues is 35, twice the average number of venues with a HDI lower than 0.8 and higher than 0.5.

The k-means method divided the neighborhood into 2 large clusters (2 and 4), and 3 smaller ones (0, 1 and 3). The 2 largest clusters seems to be divided into groups of higher HDI neighborhoods (cluster 4) and lower HDI neighborhoods (cluster 2).

6 Conclusion and Future Directions

This work shows the relationship between the HDI of Rio de Janeiro's neighborhoods and the average number of venues. It was verified that there is tendency that a neighborhood

with a higher HDI will present more venues, due to higher safety, higher number of clients with a higher purchasing power, etc.

This kind of analysis can be done in other cities as well, and other variables can be considered in order to obtain clearer results, such as crime rates, average housing prices, etc.