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.

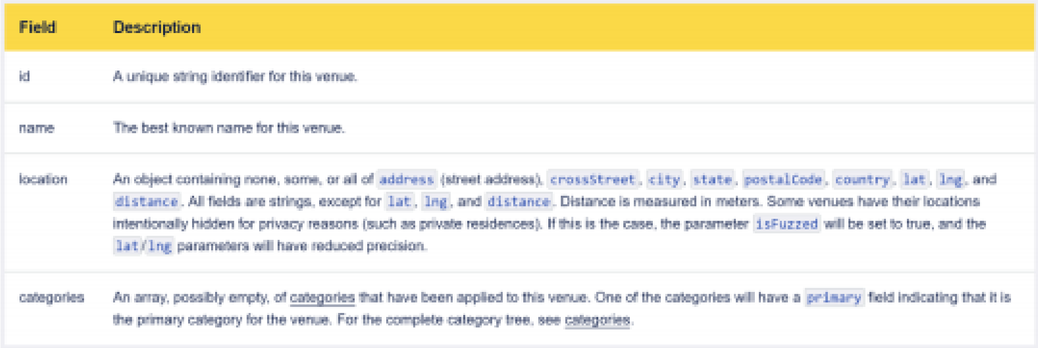
* Rio de Janeiro Neighborhood Census Data

- 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 `address` (street address), `crossStreet`, `city`, `state`, `postalCode`, `country`, `lat`, `lng`, and `distance`. All fields are strings, except for `lat`, `lng`, and `distance`. 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 `isFuzzed` will be set to true, and the `lat`/`lng` parameters will have reduced precision.
categories	An array, possibly empty, of categories that have been applied to this venue. One of the categories will have a `primary` field indicating that it is the primary category for the venue. For the complete category tree, see categories.

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
- Some feature that can be extracted includes: life expectancy, education rate, per capita income and some others.
- The table in Wikipedia doesn't have the location values (lat/long) so I need to geocode it through geopy library.

3. Methodology

The capstone project was separated in six parts to be more organized:

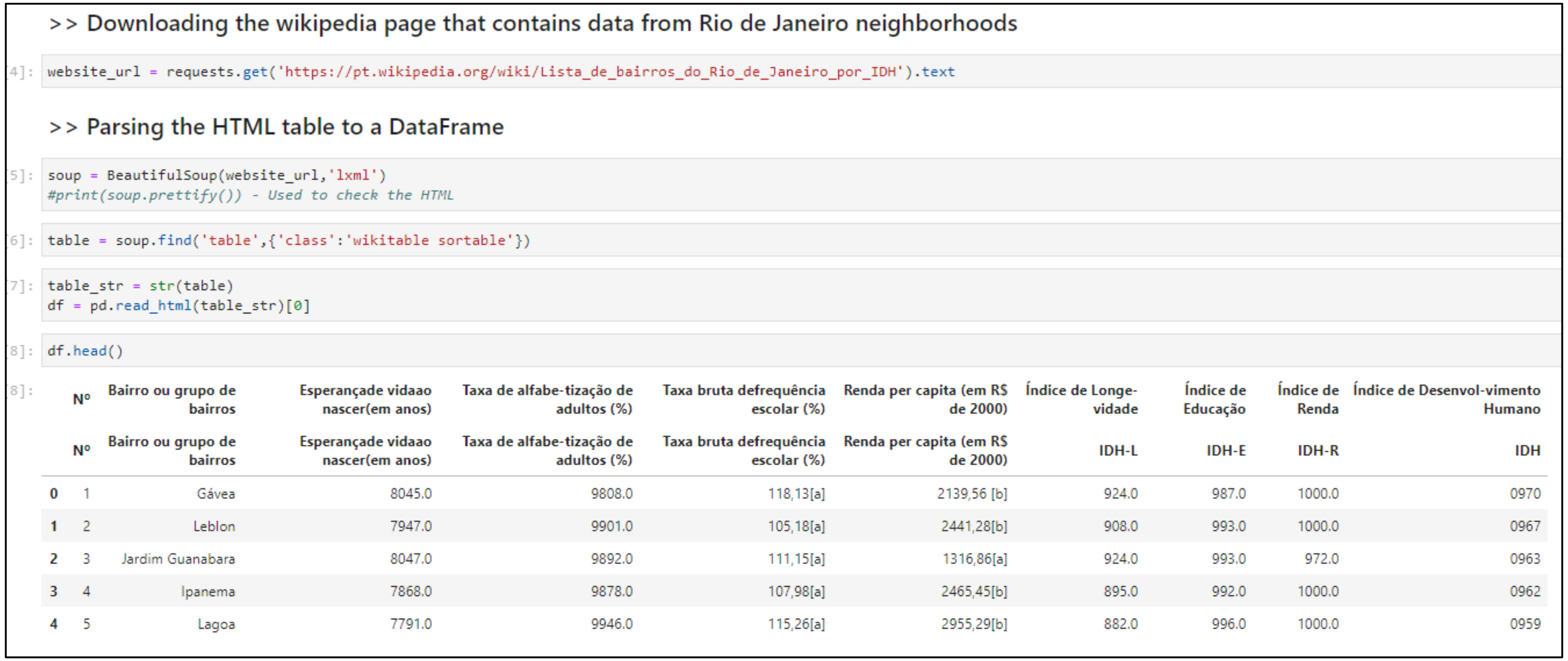
- Part 1: Data Loading and Data Wrangling
- Part 2: Geocoding the Data
- Part 3: Data Exploration and Data Visualization
- Part 4: Machine Learning – Clustering
- Part 5: Venues Evaluation using Foursquare API
- Part 6: Conclusion

This section will show all the steps done in the analysis: data exploration, charts and models. The results and discussions will follow in the next sections of the presentation

3. Methodology (Part 1)

→ Data Loading and Data Wrangling

Creating the data frame from Wikipedia about Rio de Janeiro census data that contains the HDI data and Per Capita Income that were used in the analysis



>> Downloading the wikipedia page that contains data from Rio de Janeiro neighborhoods

```
website_url = requests.get('https://pt.wikipedia.org/wiki/Lista_de_bairros_do_Rio_de_Janeiro_por_IDH').text
```

>> Parsing the HTML table to a DataFrame

```
soup = BeautifulSoup(website_url,'lxml')
#print(soup.prettify()) - Used to check the HTML
```

```
table = soup.find('table',{'class':'wikitable sortable'})
```

```
table_str = str(table)
df = pd.read_html(table_str)[0]
```

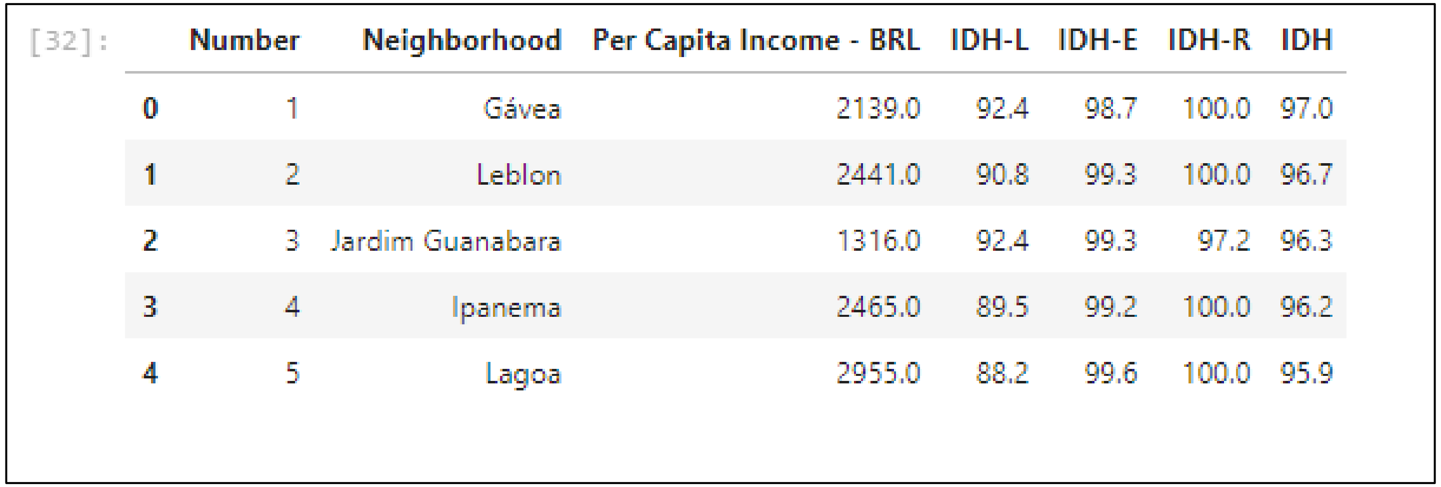
```
df.head()
```

	Nº	Bairro ou grupo de bairros	Esperançade vidaao nascer(em anos)	Taxa de alfabe-tização de adultos (%)	Taxa bruta defrequência escolar (%)	Renda per capita (em R$ de 2000)	Índice de Longe-vidade	Índice de Educação	Índice de Renda	Índice de Desenvol-vimento Humano
	Nº	Bairro ou grupo de bairros	Esperançade vidaao nascer(em anos)	Taxa de alfabe-tização de adultos (%)	Taxa bruta defrequência escolar (%)	Renda per capita (em R$ de 2000)	IDH-L	IDH-E	IDH-R	IDH
0	1	Gávea	8045.0	9808.0	118,13[a]	2139,56 [b]	924.0	987.0	1000.0	0970
1	2	Leblon	7947.0	9901.0	105,18[a]	2441,28[b]	908.0	993.0	1000.0	0967
2	3	Jardim Guanabara	8047.0	9892.0	111,15[a]	1316,86[a]	924.0	993.0	972.0	0963
3	4	Ipanema	7868.0	9878.0	107,98[a]	2465,45[b]	895.0	992.0	1000.0	0962
4	5	Lagoa	7791.0	9946.0	115,26[a]	2955,29[b]	882.0	996.0	1000.0	0959

3. Methodology (Part 1)

→ Data Loading and Data Wrangling

The final data frame after some fixes in values and categories. Also, some categories were drop.



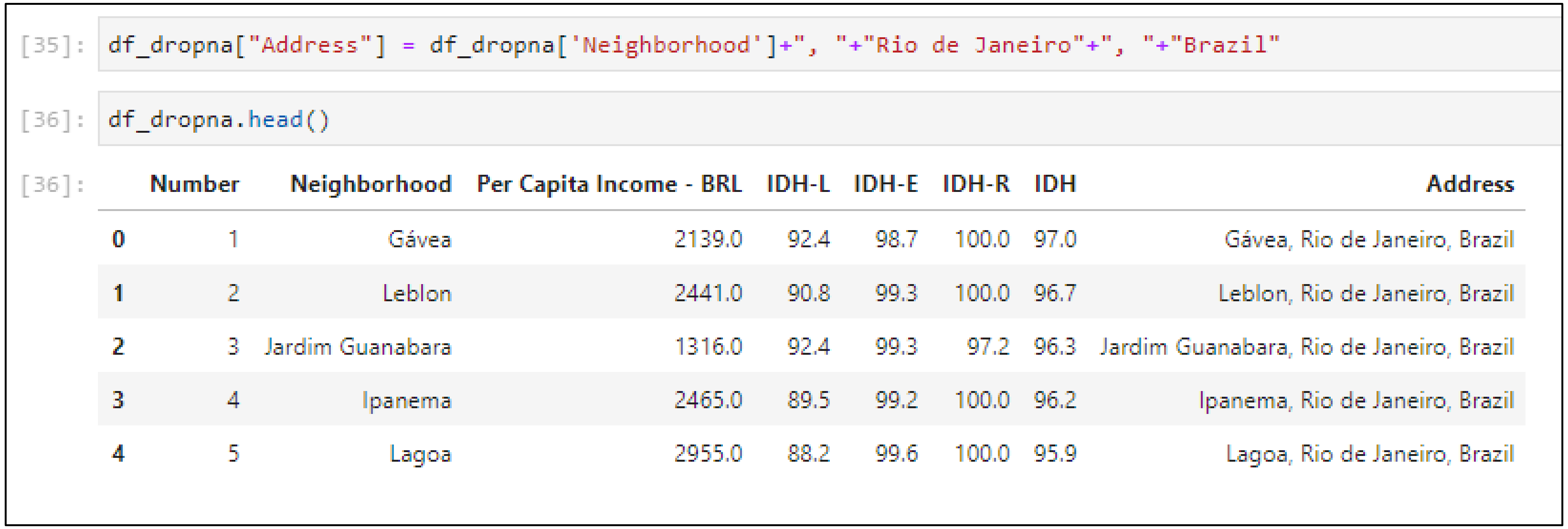
[32]:

	Number	Neighborhood	Per Capita Income - BRL	IDH-L	IDH-E	IDH-R	IDH
0	1	Gávea	2139.0	92.4	98.7	100.0	97.0
1	2	Leblon	2441.0	90.8	99.3	100.0	96.7
2	3	Jardim Guanabara	1316.0	92.4	99.3	97.2	96.3
3	4	Ipanema	2465.0	89.5	99.2	100.0	96.2
4	5	Lagoa	2955.0	88.2	99.6	100.0	95.9

3. Methodology (Part 2)

→ Geocoding

Geopy was used to geocode (OpenStreetMap API) the Address column that was created concatenating the neighborhood names and the city name



```
[35]: df_dropna["Address"] = df_dropna['Neighborhood']+", "+"Rio de Janeiro"+", "+"Brazil"
```

```
[36]: df_dropna.head()
```

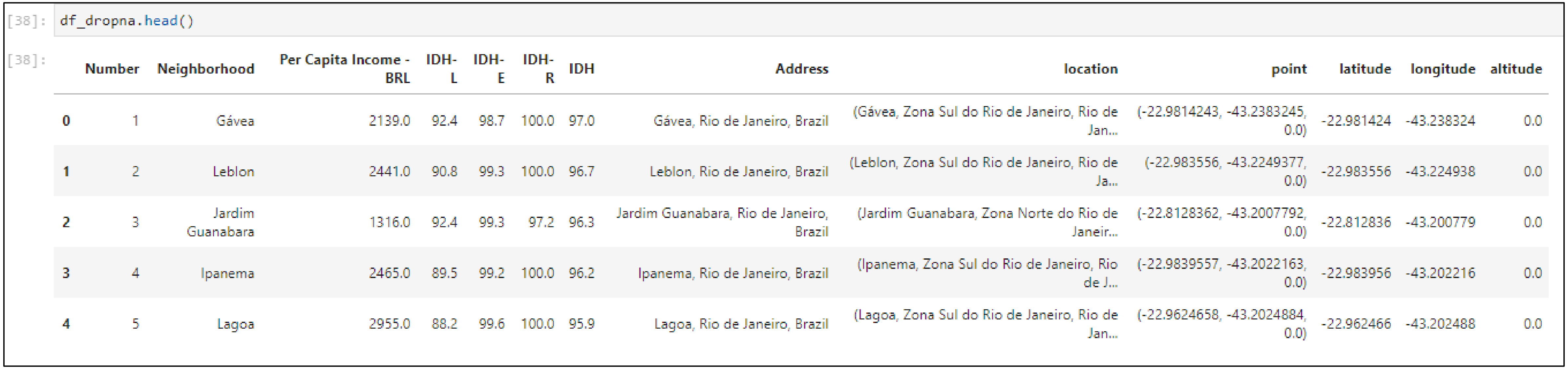
[36]:

	Number	Neighborhood	Per Capita Income - BRL	IDH-L	IDH-E	IDH-R	IDH	Address
0	1	Gávea	2139.0	92.4	98.7	100.0	97.0	Gávea, Rio de Janeiro, Brazil
1	2	Leblon	2441.0	90.8	99.3	100.0	96.7	Leblon, Rio de Janeiro, Brazil
2	3	Jardim Guanabara	1316.0	92.4	99.3	97.2	96.3	Jardim Guanabara, Rio de Janeiro, Brazil
3	4	Ipanema	2465.0	89.5	99.2	100.0	96.2	Ipanema, Rio de Janeiro, Brazil
4	5	Lagoa	2955.0	88.2	99.6	100.0	95.9	Lagoa, Rio de Janeiro, Brazil

3. Methodology (Part 2)

→ Geocoding

The geocoded data frame after processing through the Geopy function



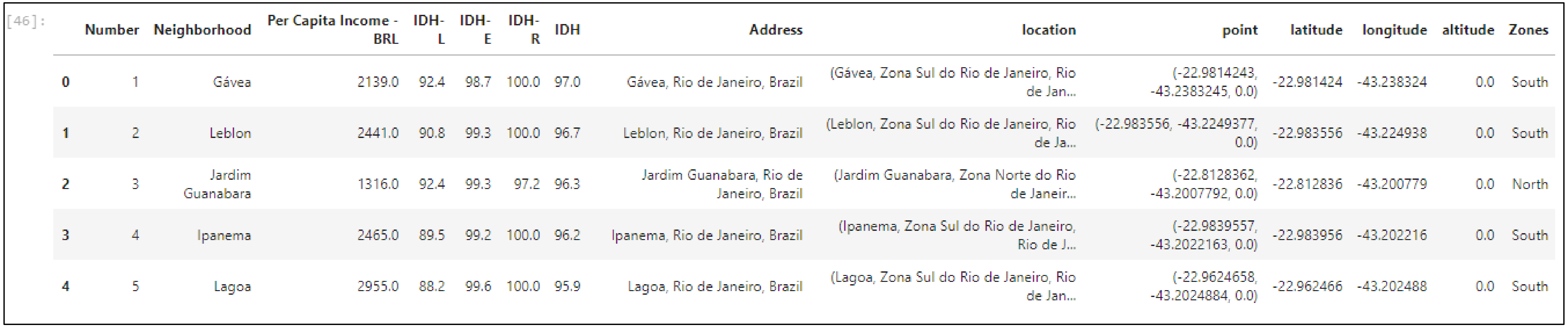
```
[38]: df_dropna.head()
```

[38]:

	Number	Neighborhood	Per Capita Income - BRL	IDH-L	IDH-E	IDH-R	IDH	Address	location	point	latitude	longitude	altitude
0	1	Gávea	2139.0	92.4	98.7	100.0	97.0	Gávea, Rio de Janeiro, Brazil	(Gávea, Zona Sul do Rio de Janeiro, Rio de Jan...	(-22.9814243, -43.2383245, 0.0)	-22.981424	-43.238324	0.0
1	2	Leblon	2441.0	90.8	99.3	100.0	96.7	Leblon, Rio de Janeiro, Brazil	(Leblon, Zona Sul do Rio de Janeiro, Rio de Ja...	(-22.983556, -43.2249377, 0.0)	-22.983556	-43.224938	0.0
2	3	Jardim Guanabara	1316.0	92.4	99.3	97.2	96.3	Jardim Guanabara, Rio de Janeiro, Brazil	(Jardim Guanabara, Zona Norte do Rio de Janeir...	(-22.8128362, -43.2007792, 0.0)	-22.812836	-43.200779	0.0
3	4	Ipanema	2465.0	89.5	99.2	100.0	96.2	Ipanema, Rio de Janeiro, Brazil	(Ipanema, Zona Sul do Rio de Janeiro, Rio de J...	(-22.9839557, -43.2022163, 0.0)	-22.983956	-43.202216	0.0
4	5	Lagoa	2955.0	88.2	99.6	100.0	95.9	Lagoa, Rio de Janeiro, Brazil	(Lagoa, Zona Sul do Rio de Janeiro, Rio de Jan...	(-22.9624658, -43.2024884, 0.0)	-22.962466	-43.202488	0.0

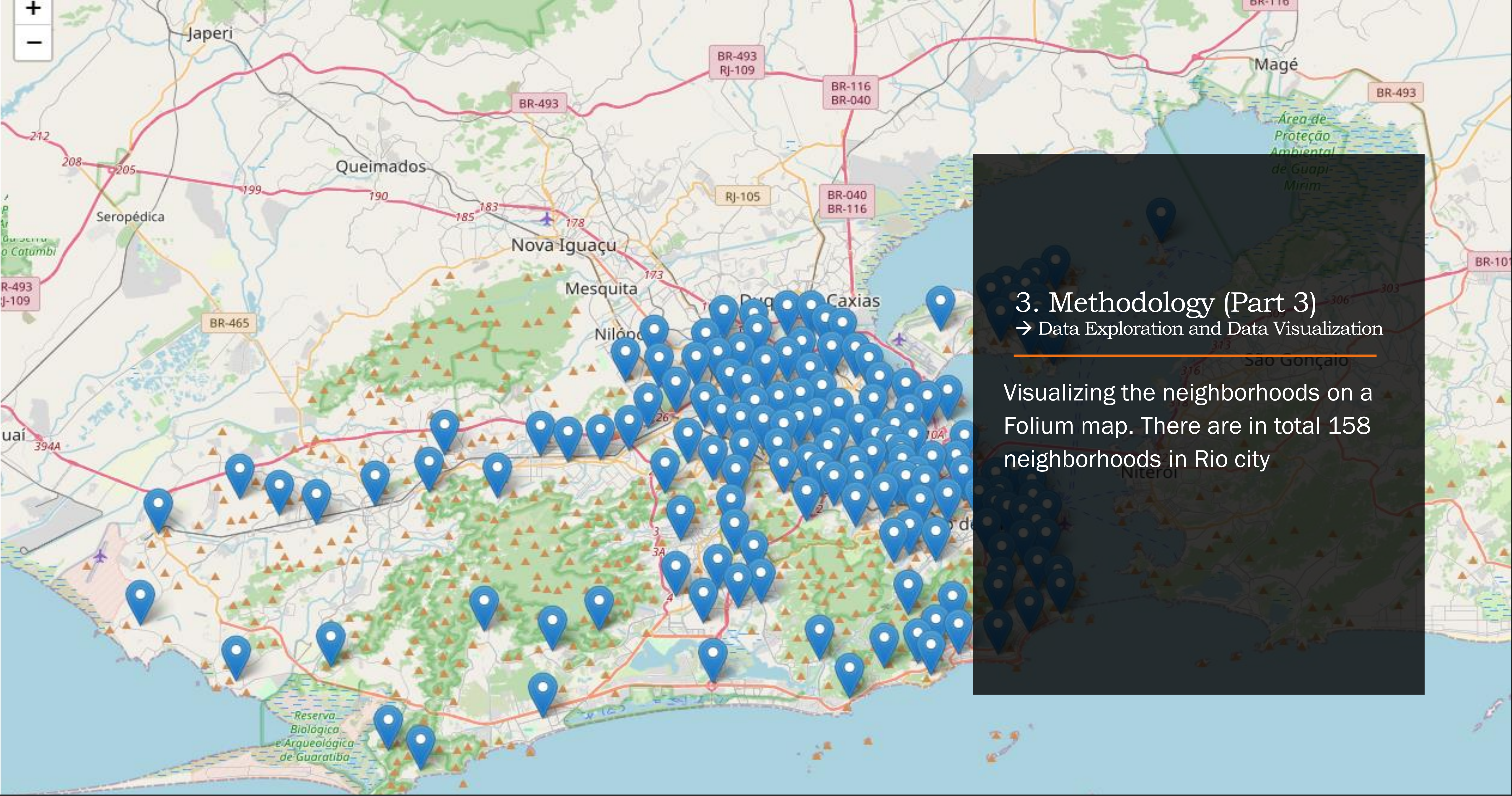
3. Methodology (Part 2)
→ Geocoding

Given the geocoded data, a new column (Zones) was added that says in which region is the neighborhood. This is the cleaned full data frame.



[46]:

	Number	Neighborhood	Per Capita Income - BRL	IDH-L	IDH-E	IDH-R	IDH	Address	location	point	latitude	longitude	altitude	Zones
0	1	Gávea	2139.0	92.4	98.7	100.0	97.0	Gávea, Rio de Janeiro, Brazil	(Gávea, Zona Sul do Rio de Janeiro, Rio de Jan...	(-22.9814243, -43.2383245, 0.0)	-22.981424	-43.238324	0.0	South
1	2	Leblon	2441.0	90.8	99.3	100.0	96.7	Leblon, Rio de Janeiro, Brazil	(Leblon, Zona Sul do Rio de Janeiro, Rio de Ja...	(-22.983556, -43.2249377, 0.0)	-22.983556	-43.224938	0.0	South
2	3	Jardim Guanabara	1316.0	92.4	99.3	97.2	96.3	Jardim Guanabara, Rio de Janeiro, Brazil	(Jardim Guanabara, Zona Norte do Rio de Janeir...	(-22.8128362, -43.2007792, 0.0)	-22.812836	-43.200779	0.0	North
3	4	Ipanema	2465.0	89.5	99.2	100.0	96.2	Ipanema, Rio de Janeiro, Brazil	(Ipanema, Zona Sul do Rio de Janeiro, Rio de J...	(-22.9839557, -43.2022163, 0.0)	-22.983956	-43.202216	0.0	South
4	5	Lagoa	2955.0	88.2	99.6	100.0	95.9	Lagoa, Rio de Janeiro, Brazil	(Lagoa, Zona Sul do Rio de Janeiro, Rio de Jan...	(-22.9624658, -43.2024884, 0.0)	-22.962466	-43.202488	0.0	South



3. Methodology (Part 3)

→ Data Exploration and Data Visualization

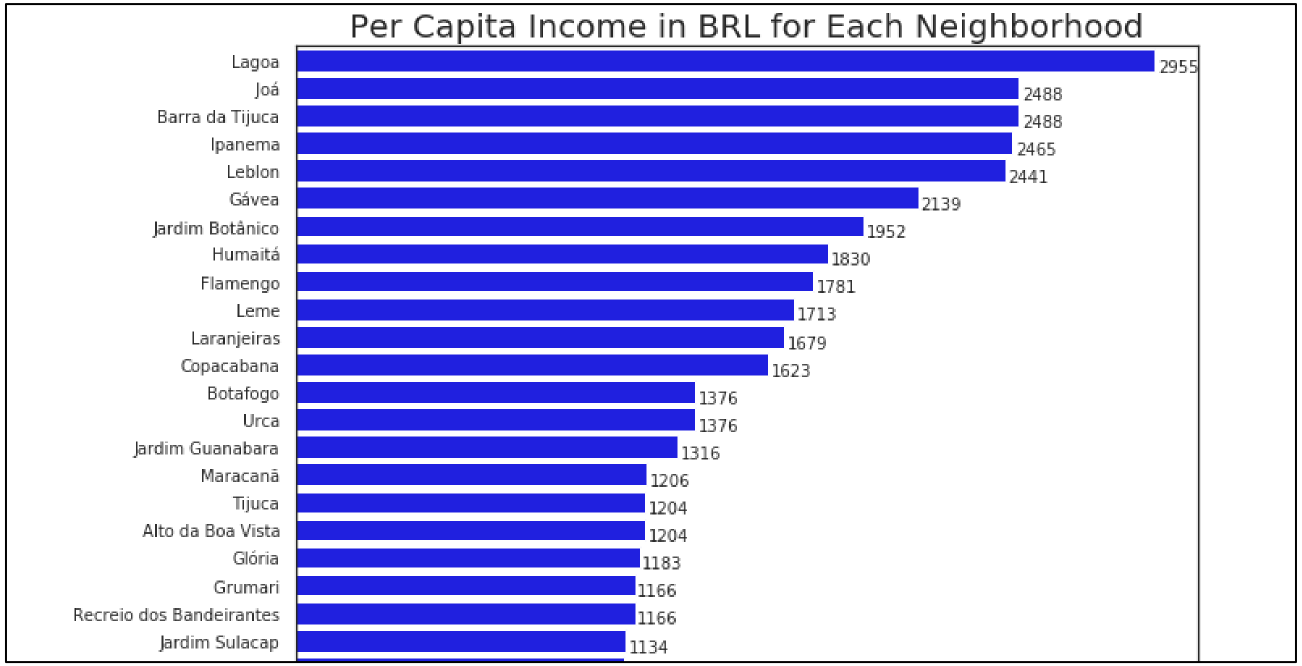
Visualizing the neighborhoods on a Folium map. There are in total 158 neighborhoods in Rio city

3. Methodology (Part 3)

→ Data Exploration and Data Visualization

A bar chart to evaluate the Per Capita Income for each neighborhood in Rio de Janeiro city

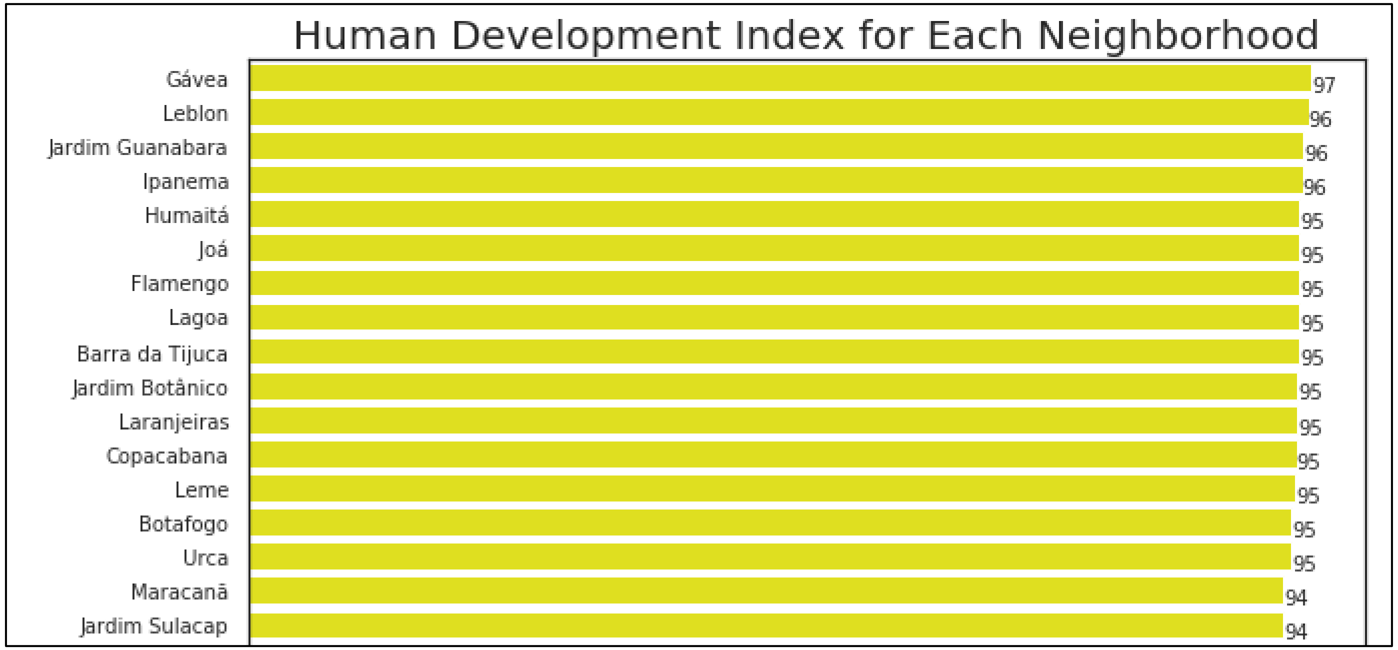
(I will take a section of the chart, for the complete one refer to the jupyter notebook)



3. Methodology (Part 3)
→ Data Exploration and Data Visualization

A bar chart to evaluate the Human Development Index for each neighborhood in Rio de Janeiro city

(I will take a section of the chart, for the complete one refer to the jupyter notebook)

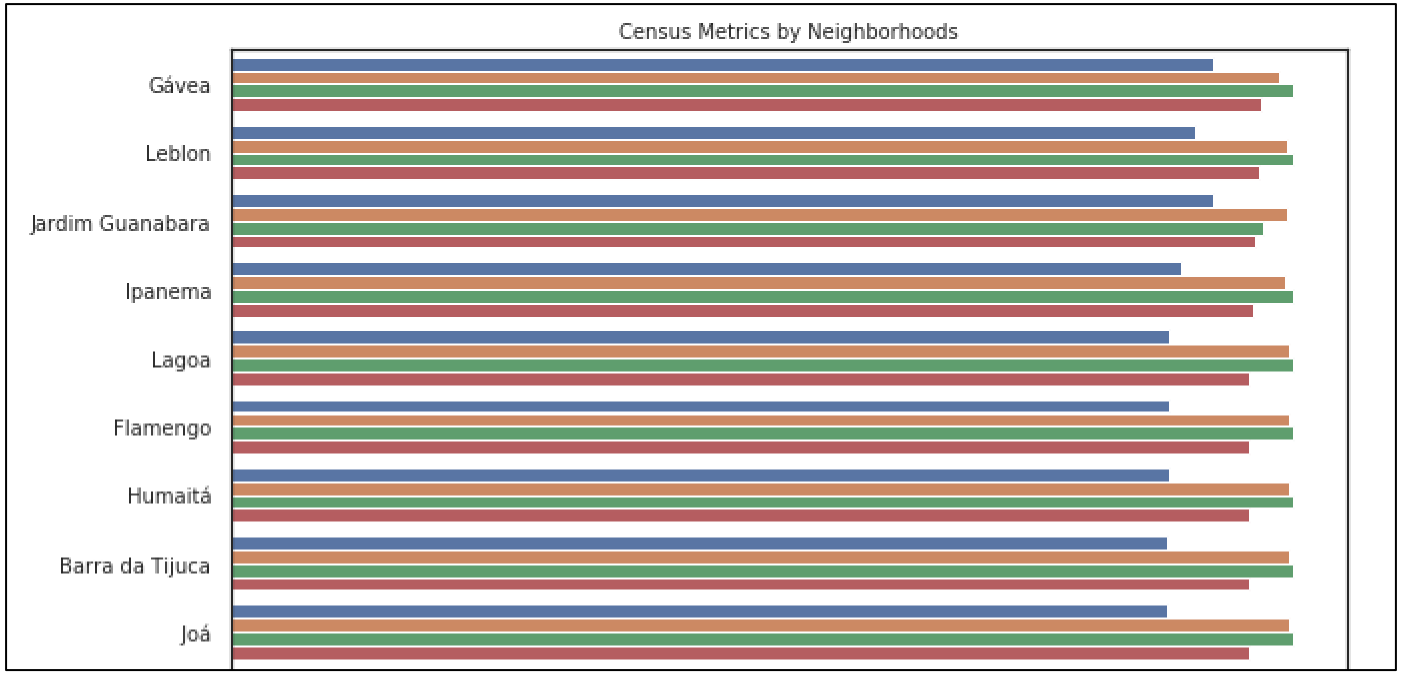


3. Methodology (Part 3)

→ Data Exploration and Data Visualization

A bar chart to compare the census metrics that compose the main HDI rate for each neighborhood

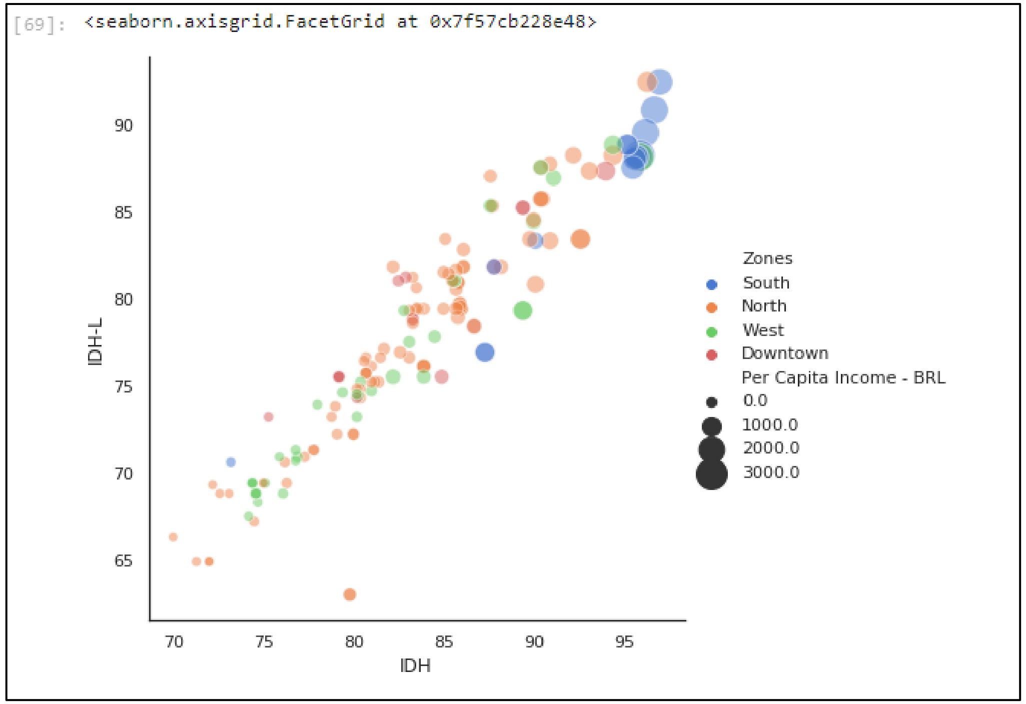
(I will take a section of the chart, for the complete one refer to the jupyter notebook)



3. Methodology (Part 3)

→ Data Exploration and Data Visualization

A scatter plot comparing Human Development Index (X) vs Life Longevity Rate (Y) for each neighborhood

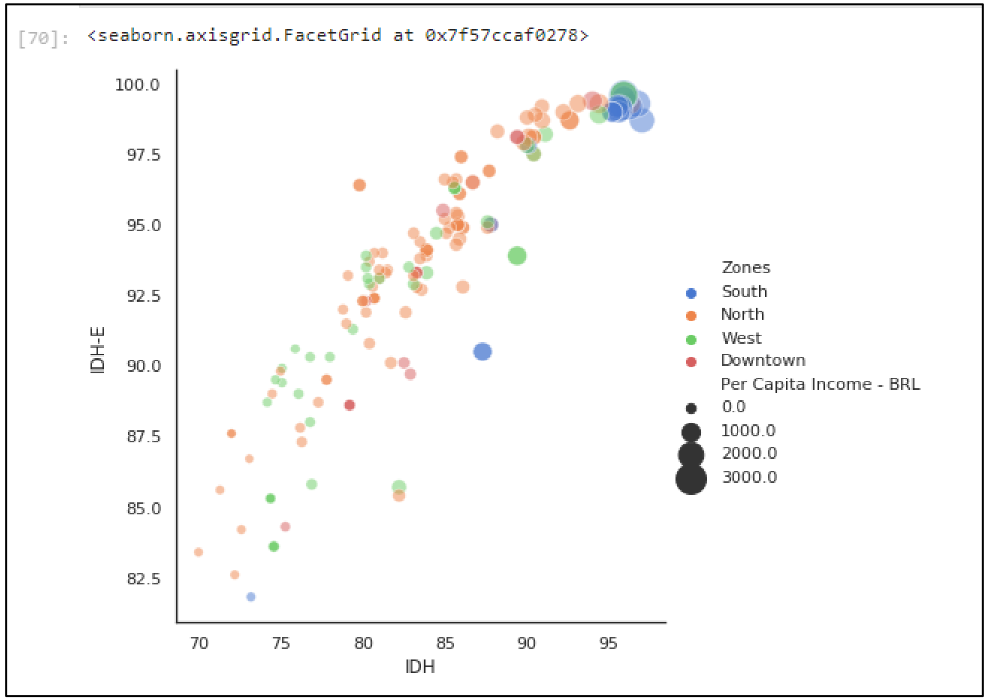


```
[69]: <seaborn.axisgrid.FacetGrid at 0x7f57cb228e48>
```

3. Methodology (Part 3)

→ Data Exploration and Data Visualization

A scatter plot comparing Human Development Index (X) vs Education Rate (Y) for each neighborhood

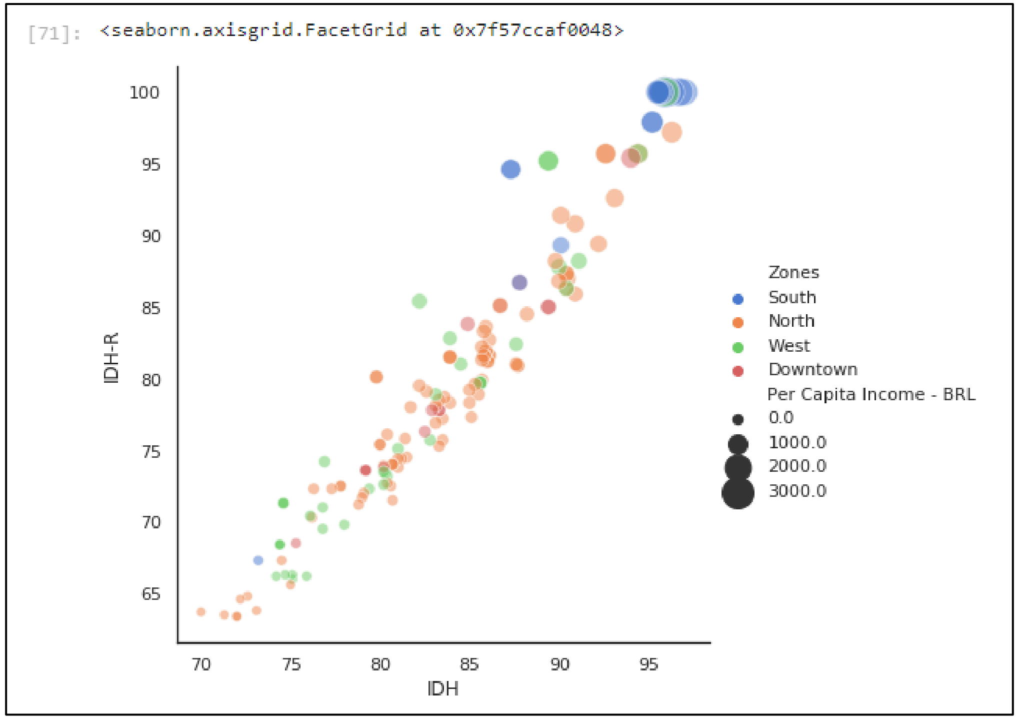


```
[70]:  <seaborn.axisgrid.FacetGrid at 0x7f57ccaf0278>
```

3. Methodology (Part 3)

→ Data Exploration and Data Visualization

A scatter plot comparing Human Development Index (X) vs Income Rate (Y) for each neighborhood

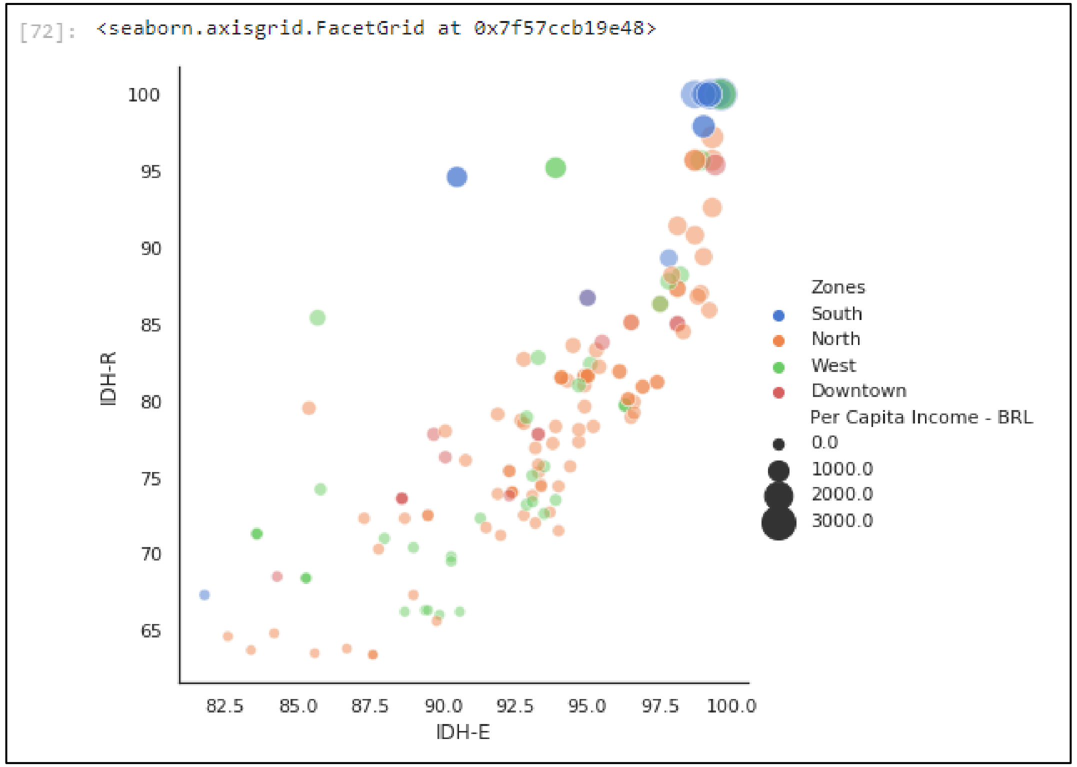


```
[71]:  <seaborn.axisgrid.FacetGrid at 0x7f57ccaf0048>
```

3. Methodology (Part 3)

→ Data Exploration and Data Visualization

A scatter plot comparing Education Rate (X) vs Income Rate (Y) for each neighborhood

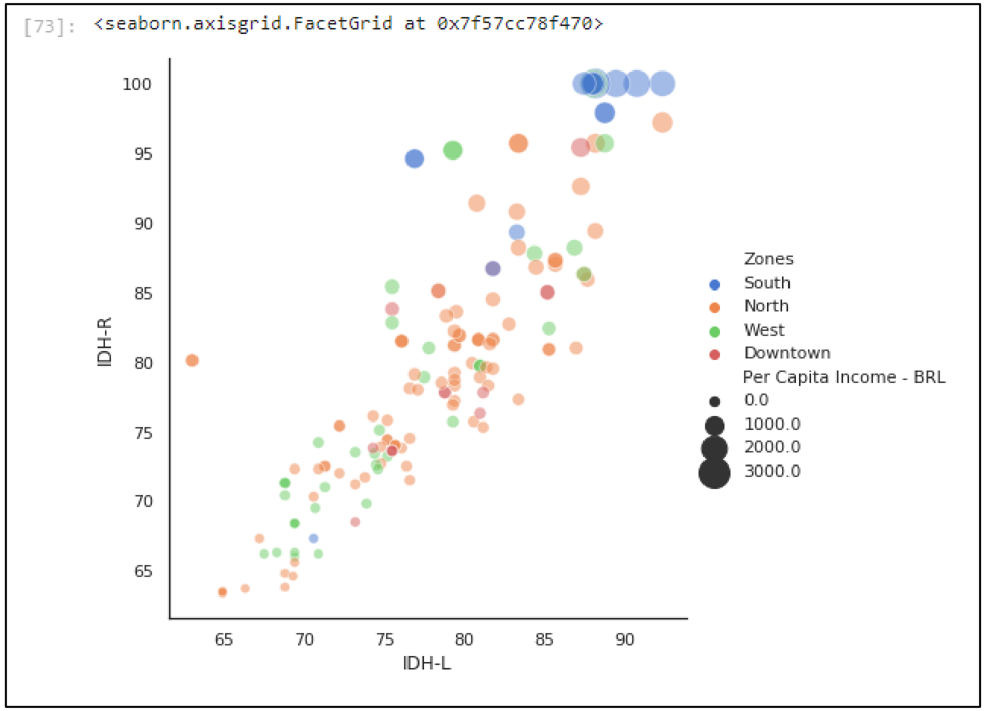


```
[72]: <seaborn.axisgrid.FacetGrid at 0x7f57ccb19e48>
```

3. Methodology (Part 3)

→ Data Exploration and Data Visualization

A scatter plot comparing Life Longevity Rate (X) vs Income Rate (Y) for each neighborhood

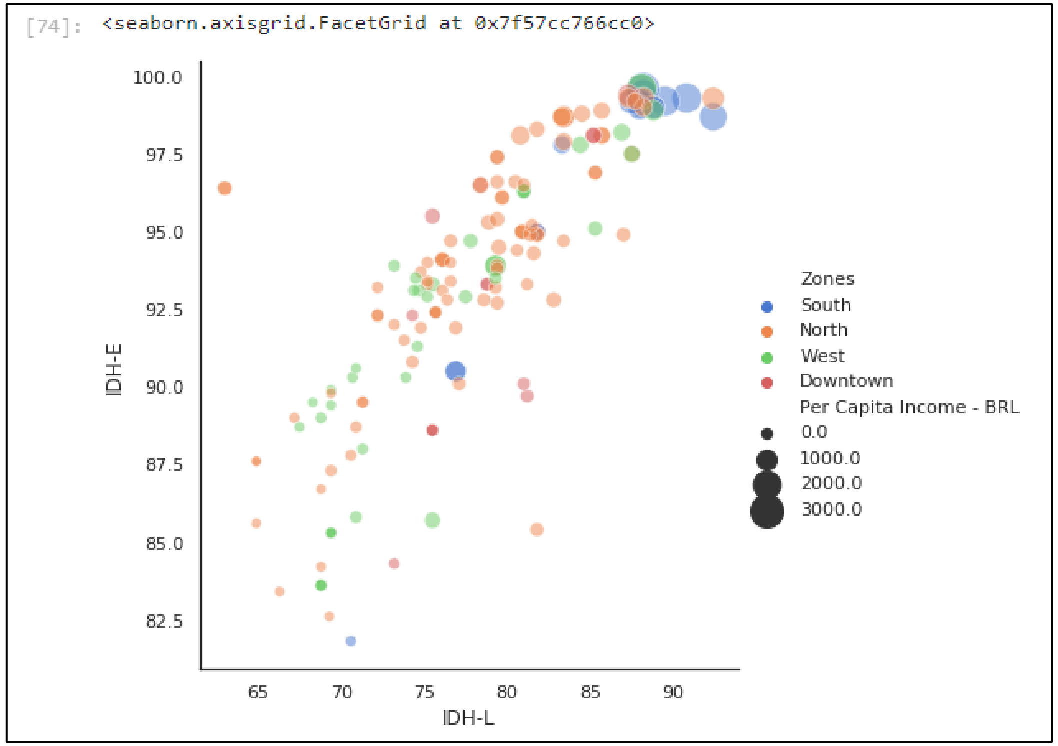


```
[73]: <seaborn.axisgrid.FacetGrid at 0x7f57cc78f470>
```

3. Methodology (Part 3)

→ Data Exploration and Data Visualization

A scatter plot comparing Life Longevity Rate (X) vs Education Rate (Y) for each neighborhood



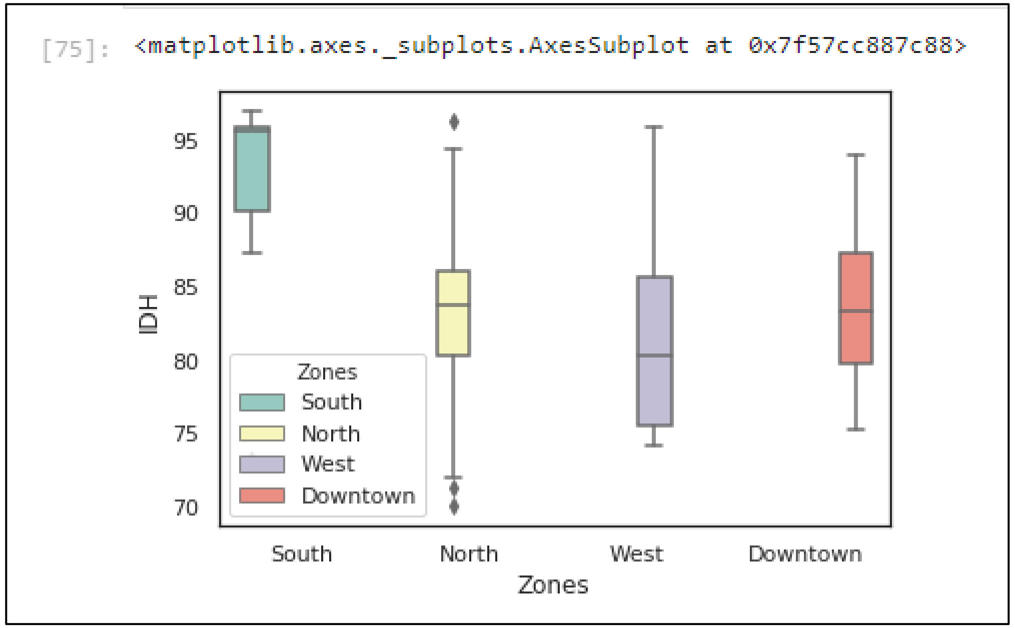
```
[74]: <seaborn.axisgrid.FacetGrid at 0x7f57cc766cc0>
```

3. Methodology (Part 3)

→ Data Exploration and Data Visualization

A box plot to evaluate how the data is distributed

Human Development Index by city Zones

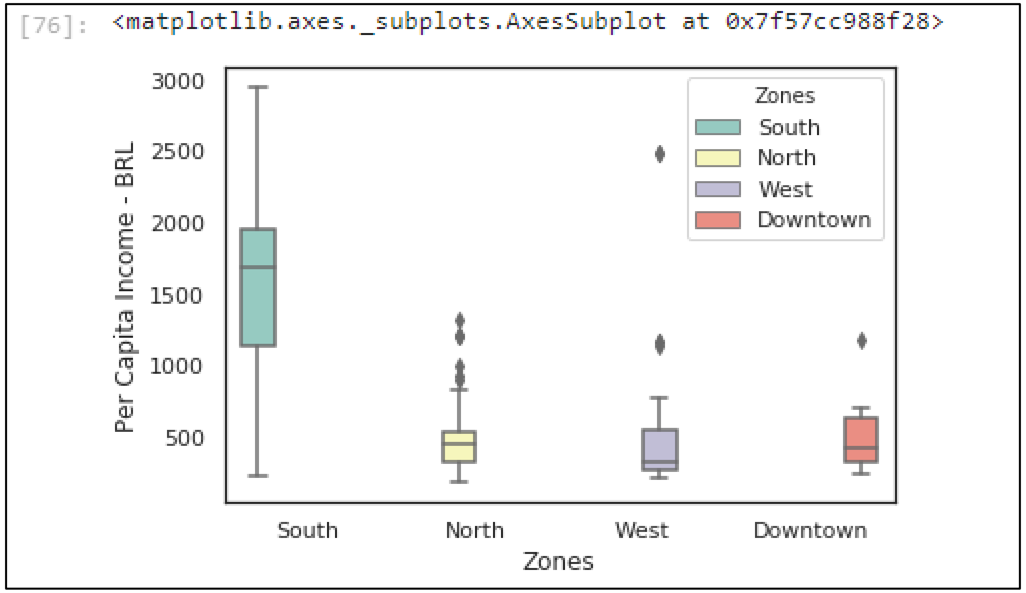


```
[75]: <matplotlib.axes._subplots.AxesSubplot at 0x7f57cc887c88>
```

3. Methodology (Part 3)
→ Data Exploration and Data Visualization

A box plot to evaluate how the data is distributed

Per Capita Income by city Zones

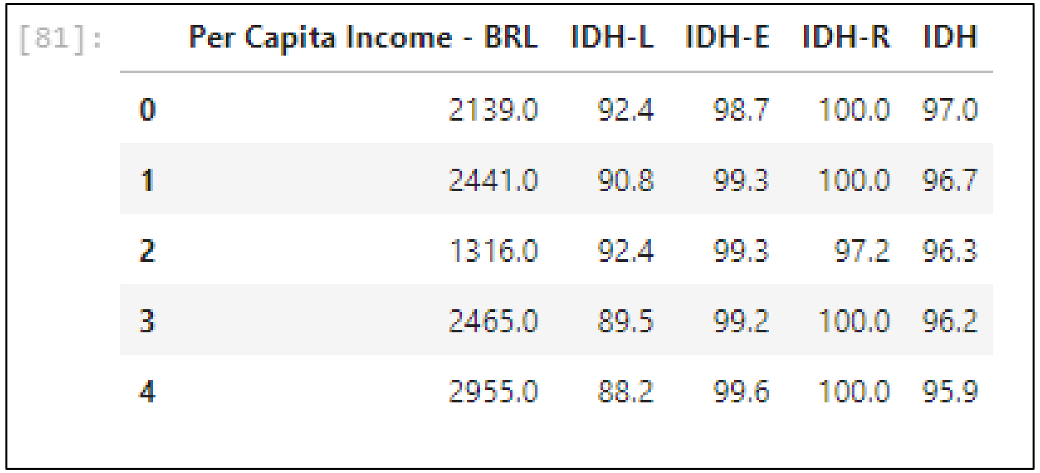


```
[76]: <matplotlib.axes._subplots.AxesSubplot at 0x7f57cc988f28>
```

3. Methodology (Part 4)

→ Machine Learning - Clustering

K-Means was used to cluster the neighborhoods and try to discover with one has more similarities between them. The data frame below sent to the algorithm containing only numerical variables.



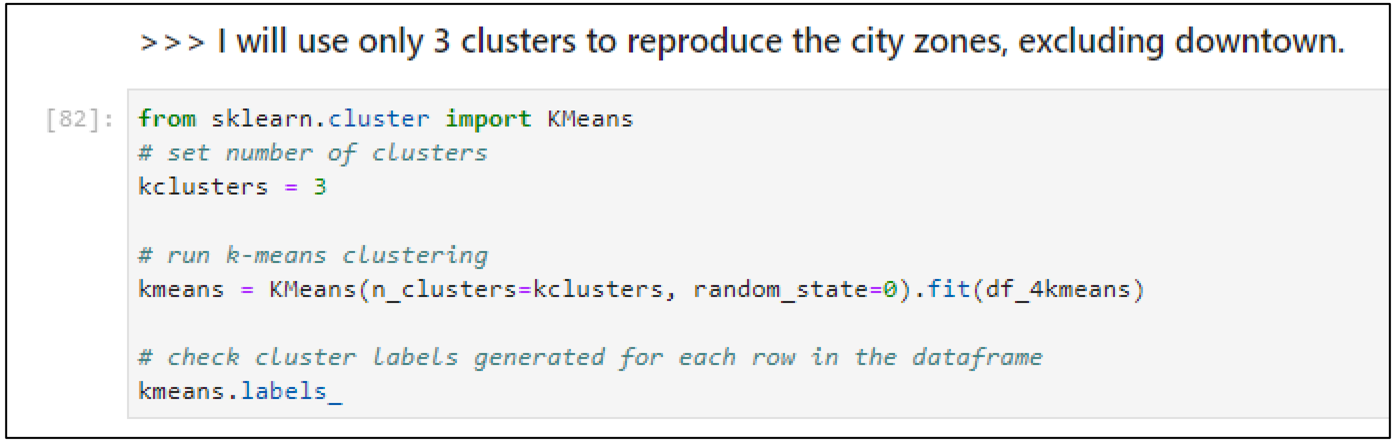
[81]:

	Per Capita Income - BRL	IDH-L	IDH-E	IDH-R	IDH
0	2139.0	92.4	98.7	100.0	97.0
1	2441.0	90.8	99.3	100.0	96.7
2	1316.0	92.4	99.3	97.2	96.3
3	2465.0	89.5	99.2	100.0	96.2
4	2955.0	88.2	99.6	100.0	95.9

3. Methodology (Part 4)

→ Machine Learning - Clustering

K-Means modeling



>>> I will use only 3 clusters to reproduce the city zones, excluding downtown.

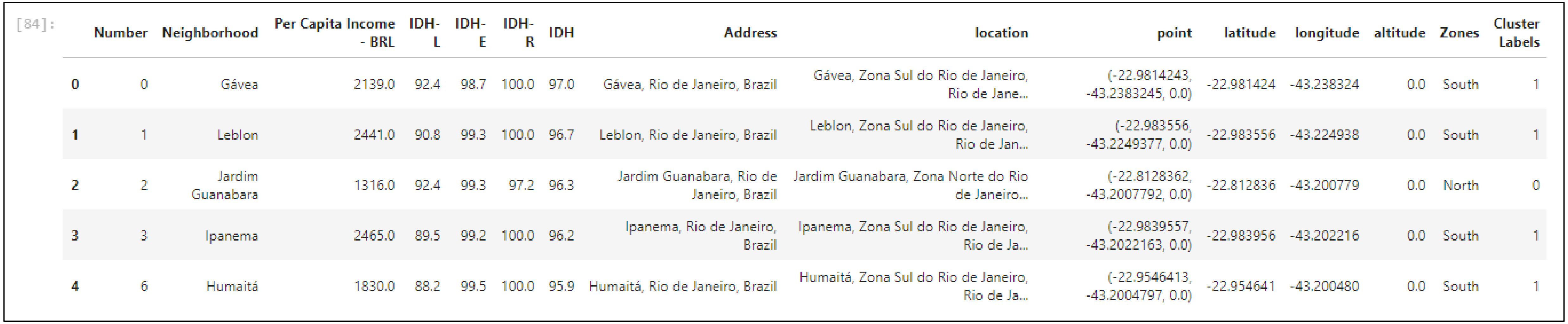
[82]:

```
from sklearn.cluster import KMeans
# set number of clusters
kclusters = 3

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(df_4kmeans)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
```

3. Methodology (Part 4)
→ Machine Learning - Clustering

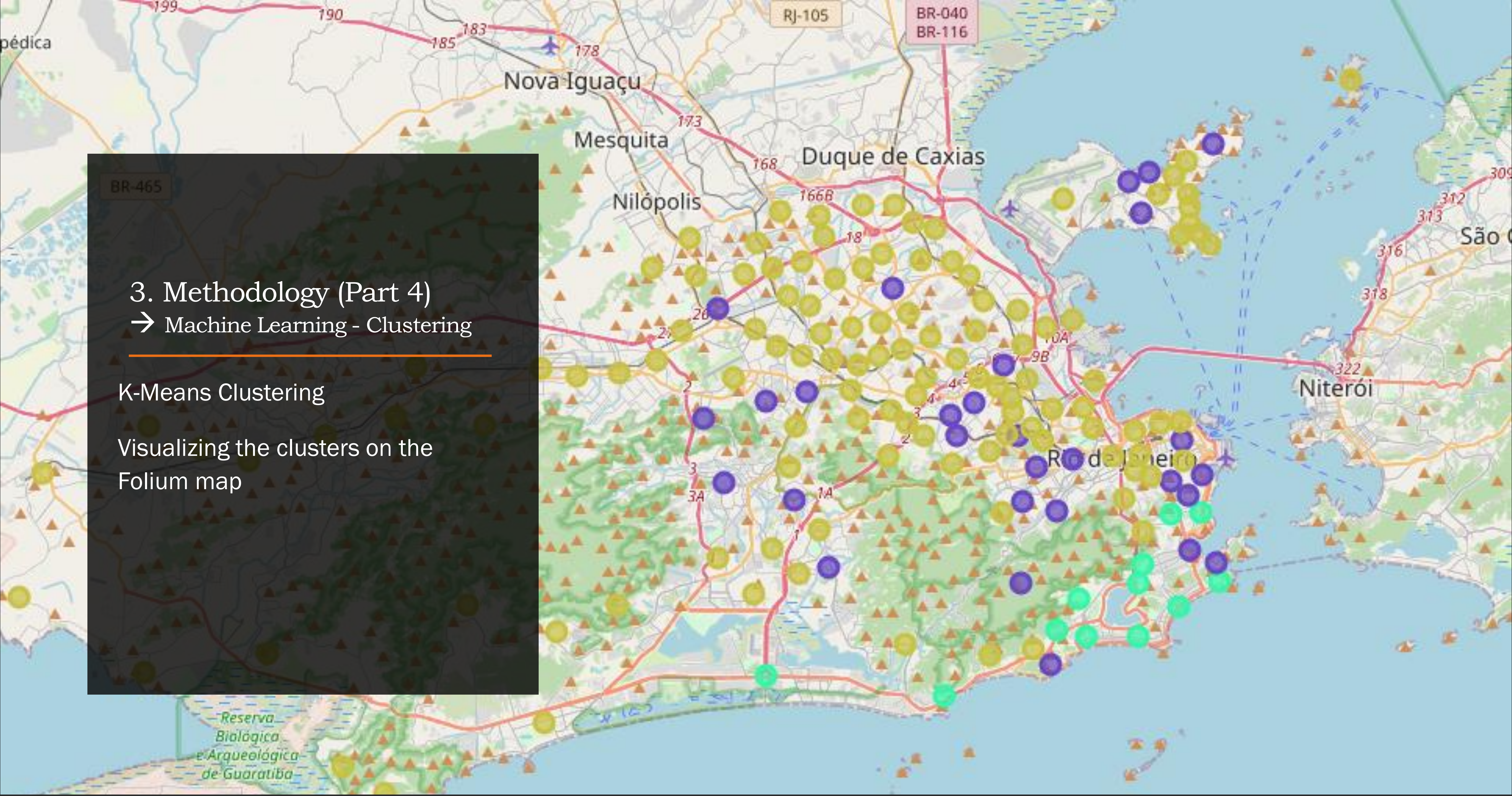
K-Means Clustering

Cluster Labels was added to the main data frame



[84]:

	Number	Neighborhood	Per Capita Income - BRL	IDH-L	IDH-E	IDH-R	IDH	Address	location	point	latitude	longitude	altitude	Zones	Cluster Labels
0	0	Gávea	2139.0	92.4	98.7	100.0	97.0	Gávea, Rio de Janeiro, Brazil	Gávea, Zona Sul do Rio de Janeiro, Rio de Jane...	(-22.9814243, -43.2383245, 0.0)	-22.981424	-43.238324	0.0	South	1
1	1	Leblon	2441.0	90.8	99.3	100.0	96.7	Leblon, Rio de Janeiro, Brazil	Leblon, Zona Sul do Rio de Janeiro, Rio de Jan...	(-22.983556, -43.2249377, 0.0)	-22.983556	-43.224938	0.0	South	1
2	2	Jardim Guanabara	1316.0	92.4	99.3	97.2	96.3	Jardim Guanabara, Rio de Janeiro, Brazil	Jardim Guanabara, Zona Norte do Rio de Janeiro...	(-22.8128362, -43.2007792, 0.0)	-22.812836	-43.200779	0.0	North	0
3	3	Ipanema	2465.0	89.5	99.2	100.0	96.2	Ipanema, Rio de Janeiro, Brazil	Ipanema, Zona Sul do Rio de Janeiro, Rio de Ja...	(-22.9839557, -43.2022163, 0.0)	-22.983956	-43.202216	0.0	South	1
4	6	Humaitá	1830.0	88.2	99.5	100.0	95.9	Humaitá, Rio de Janeiro, Brazil	Humaitá, Zona Sul do Rio de Janeiro, Rio de Ja...	(-22.9546413, -43.2004797, 0.0)	-22.954641	-43.200480	0.0	South	1



3. Methodology (Part 4)

→ Machine Learning - Clustering

K-Means Clustering

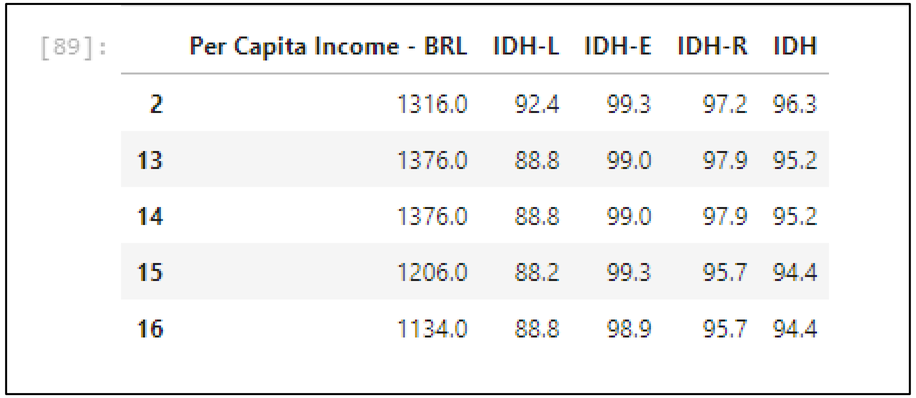
Visualizing the clusters on the Folium map

3. Methodology (Part 4)

→ Machine Learning - Clustering

Hierarchical Clustering (Dendrogram)

A dendrogram was applied to the cluster 0 of the K-Means output as was considered the best cluster to setup an english school. Below a data frame containg only the metrics from the neighborhoods in cluster 0. Already cleaned.

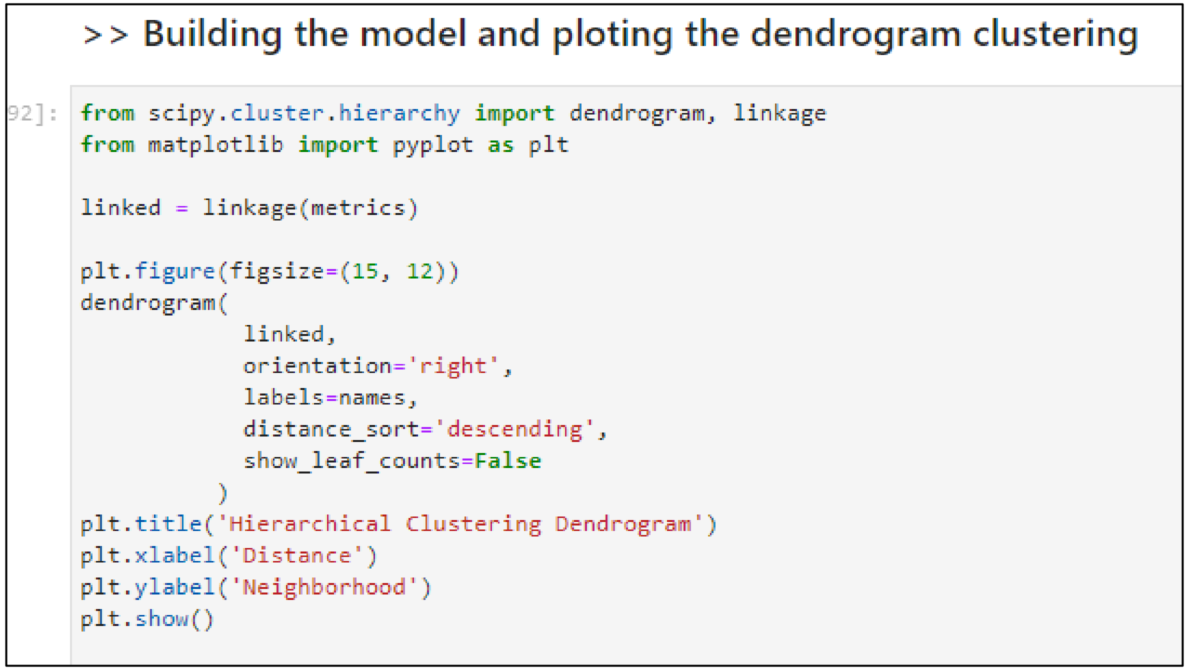


[89]:

	Per Capita Income - BRL	IDH-L	IDH-E	IDH-R	IDH
2	1316.0	92.4	99.3	97.2	96.3
13	1376.0	88.8	99.0	97.9	95.2
14	1376.0	88.8	99.0	97.9	95.2
15	1206.0	88.2	99.3	95.7	94.4
16	1134.0	88.8	98.9	95.7	94.4

3. Methodology (Part 4)
→ Machine Learning - Clustering

Hierarchical Clustering (Dendrogram)



>> Building the model and ploting the dendrogram clustering

92]:

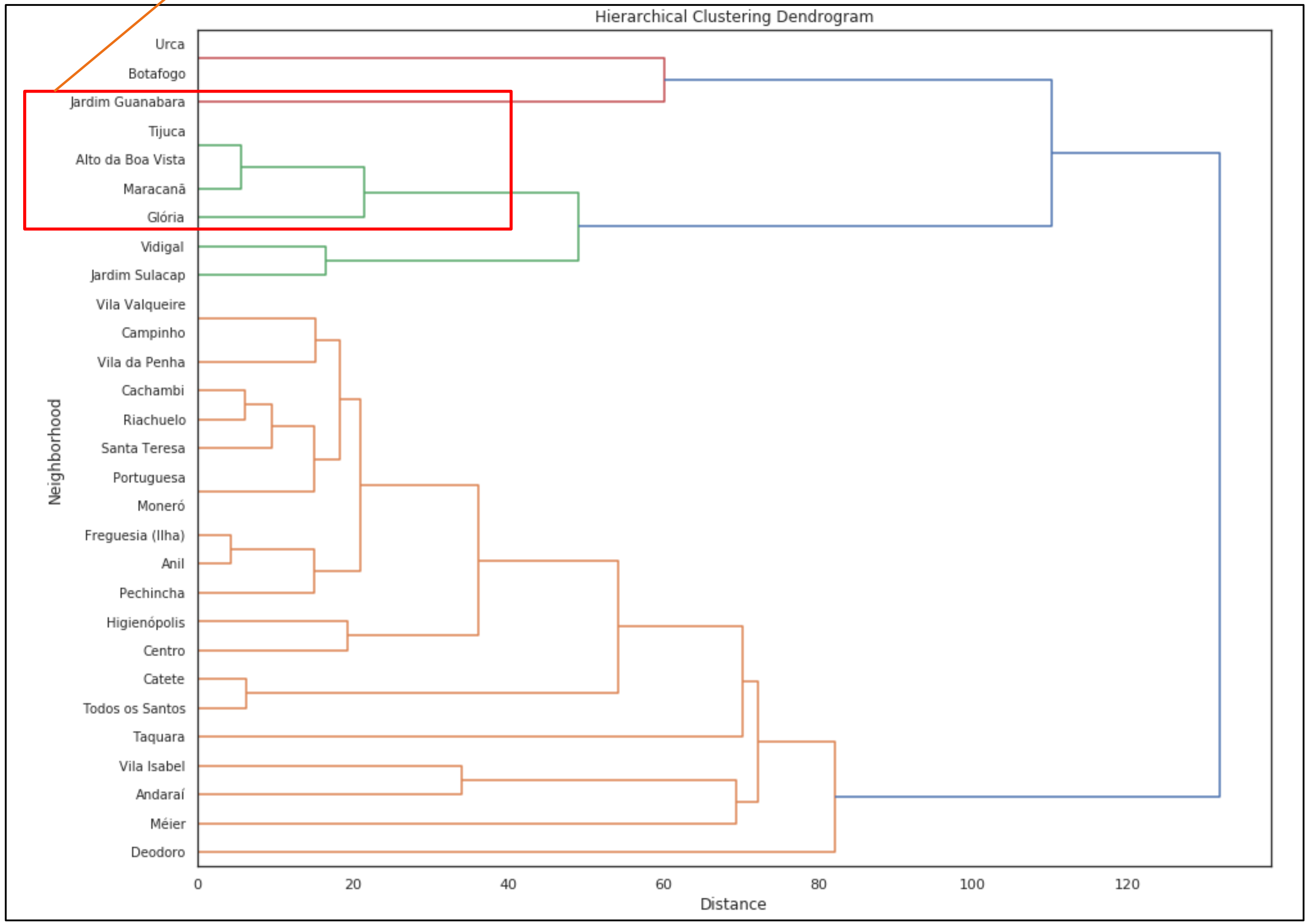
```
from scipy.cluster.hierarchy import dendrogram, linkage
from matplotlib import pyplot as plt

linked = linkage(metrics)

plt.figure(figsize=(15, 12))
dendrogram(
            linked,
            orientation='right',
            labels=names,
            distance_sort='descending',
            show_leaf_counts=False
          )
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('Distance')
plt.ylabel('Neighborhood')
plt.show()
```

3. Methodology (Part 4)

→ Machine Learning - Clustering

Hierarchical Clustering (Dendrogram)

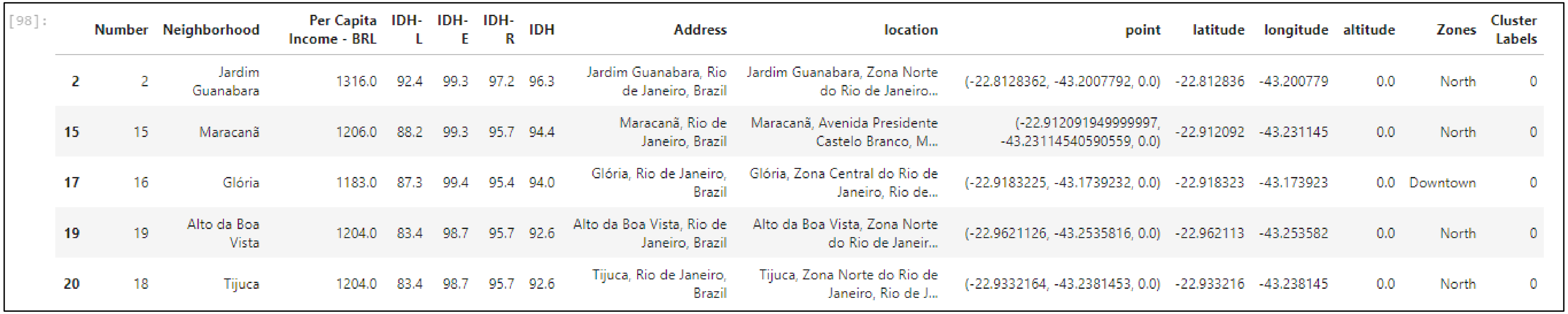
Chosen Neighborhoods



3. Methodology (Part 5)

→ Venues Evaluation using Foursquare API

To use the API, the data frame was filtered only to the chosen neighborhoods



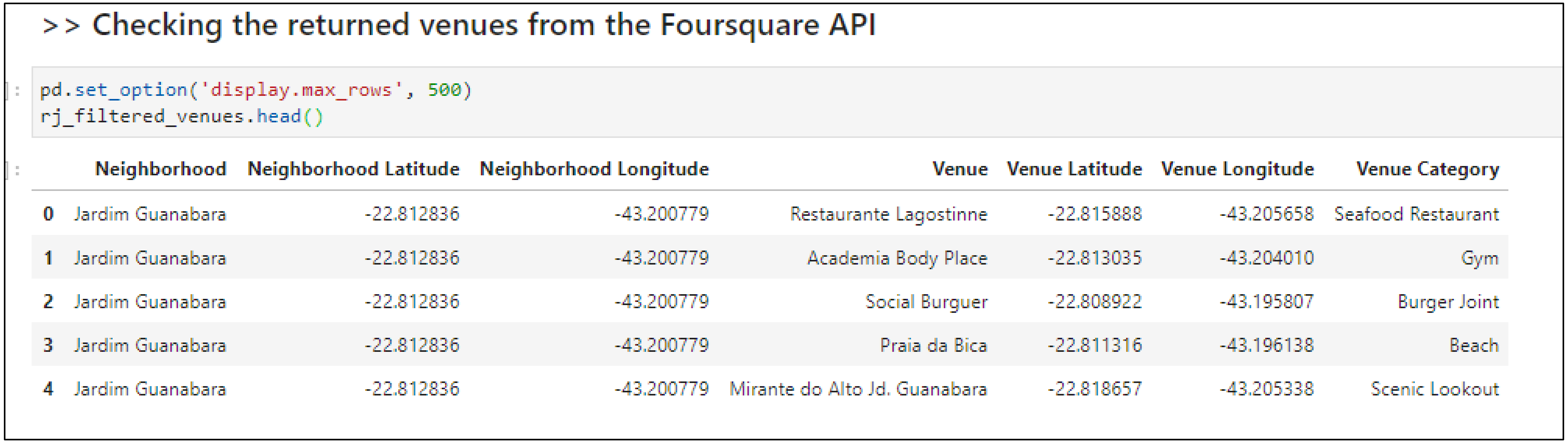
[98]:

	Number	Neighborhood	Per Capita Income - BRL	IDH-L	IDH-E	IDH-R	IDH	Address	location	point	latitude	longitude	altitude	Zones	Cluster Labels
2	2	Jardim Guanabara	1316.0	92.4	99.3	97.2	96.3	Jardim Guanabara, Rio de Janeiro, Brazil	Jardim Guanabara, Zona Norte do Rio de Janeiro...	(-22.8128362, -43.2007792, 0.0)	-22.812836	-43.200779	0.0	North	0
15	15	Maracanã	1206.0	88.2	99.3	95.7	94.4	Maracanã, Rio de Janeiro, Brazil	Maracanã, Avenida Presidente Castelo Branco, M...	(-22.912091949999997, -43.23114540590559, 0.0)	-22.912092	-43.231145	0.0	North	0
17	16	Glória	1183.0	87.3	99.4	95.4	94.0	Glória, Rio de Janeiro, Brazil	Glória, Zona Central do Rio de Janeiro, Rio de...	(-22.9183225, -43.1739232, 0.0)	-22.918323	-43.173923	0.0	Downtown	0
19	19	Alto da Boa Vista	1204.0	83.4	98.7	95.7	92.6	Alto da Boa Vista, Rio de Janeiro, Brazil	Alto da Boa Vista, Zona Norte do Rio de Janeir...	(-22.9621126, -43.2535816, 0.0)	-22.962113	-43.253582	0.0	North	0
20	18	Tijuca	1204.0	83.4	98.7	95.7	92.6	Tijuca, Rio de Janeiro, Brazil	Tijuca, Zona Norte do Rio de Janeiro, Rio de J...	(-22.9332164, -43.2381453, 0.0)	-22.933216	-43.238145	0.0	North	0

3. Methodology (Part 5)

→ Venues Evaluation using Foursquare API

Venues returned from the API for the neighborhoods indicated



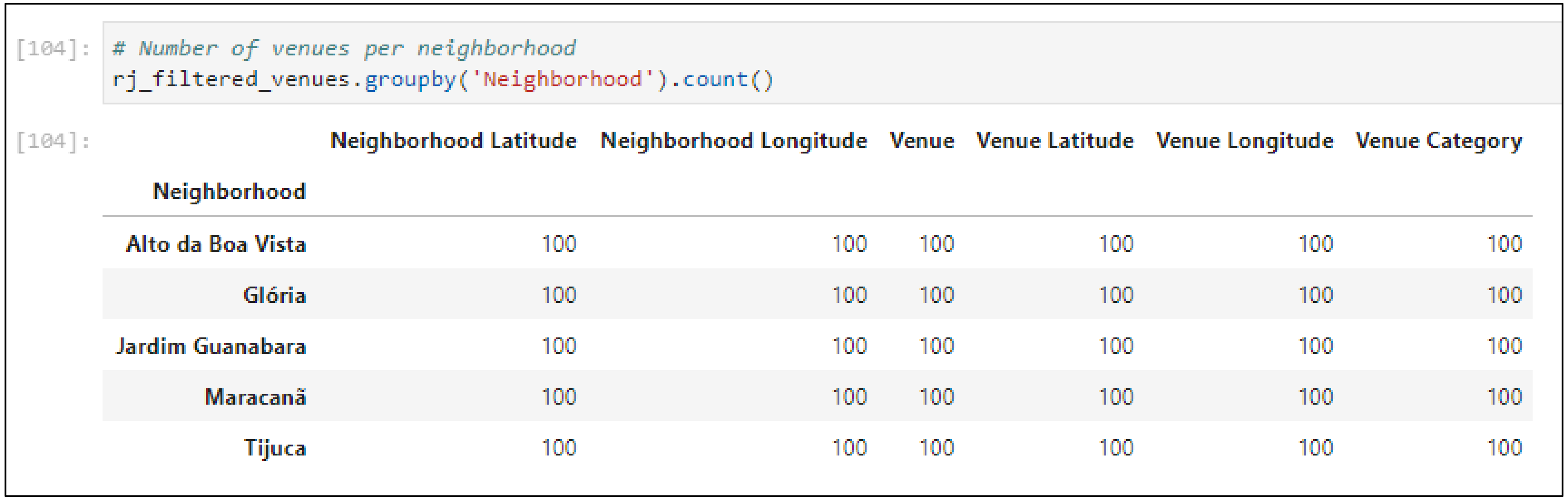
>> Checking the returned venues from the Foursquare API

```
pd.set_option('display.max_rows', 500)
rj_filtered_venues.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Jardim Guanabara	-22.812836	-43.200779	Restaurante Lagostinne	-22.815888	-43.205658	Seafood Restaurant
1	Jardim Guanabara	-22.812836	-43.200779	Academia Body Place	-22.813035	-43.204010	Gym
2	Jardim Guanabara	-22.812836	-43.200779	Social Burguer	-22.808922	-43.195807	Burger Joint
3	Jardim Guanabara	-22.812836	-43.200779	Praia da Bica	-22.811316	-43.196138	Beach
4	Jardim Guanabara	-22.812836	-43.200779	Mirante do Alto Jd. Guanabara	-22.818657	-43.205338	Scenic Lookout

3. Methodology (Part 5)
→ Venues Evaluation using Foursquare API

Number of venues returned from the API for the neighborhoods indicated



```
[104]: # Number of venues per neighborhood
       rj_filtered_venues.groupby('Neighborhood').count()
```

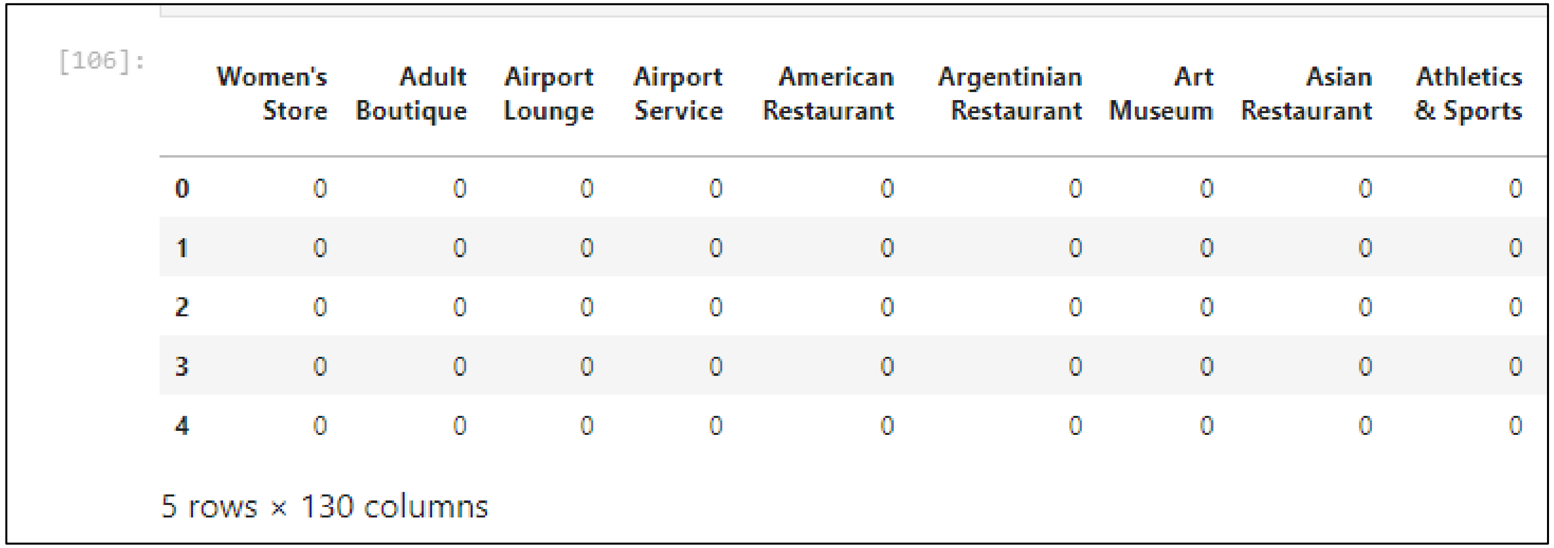
[104]:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Alto da Boa Vista	100	100	100	100	100	100
Glória	100	100	100	100	100	100
Jardim Guanabara	100	100	100	100	100	100
Maracanã	100	100	100	100	100	100
Tijuca	100	100	100	100	100	100

3. Methodology (Part 5)

→ Venues Evaluation using Foursquare API

One Hot Encoding (Venues data frame)



[106]:

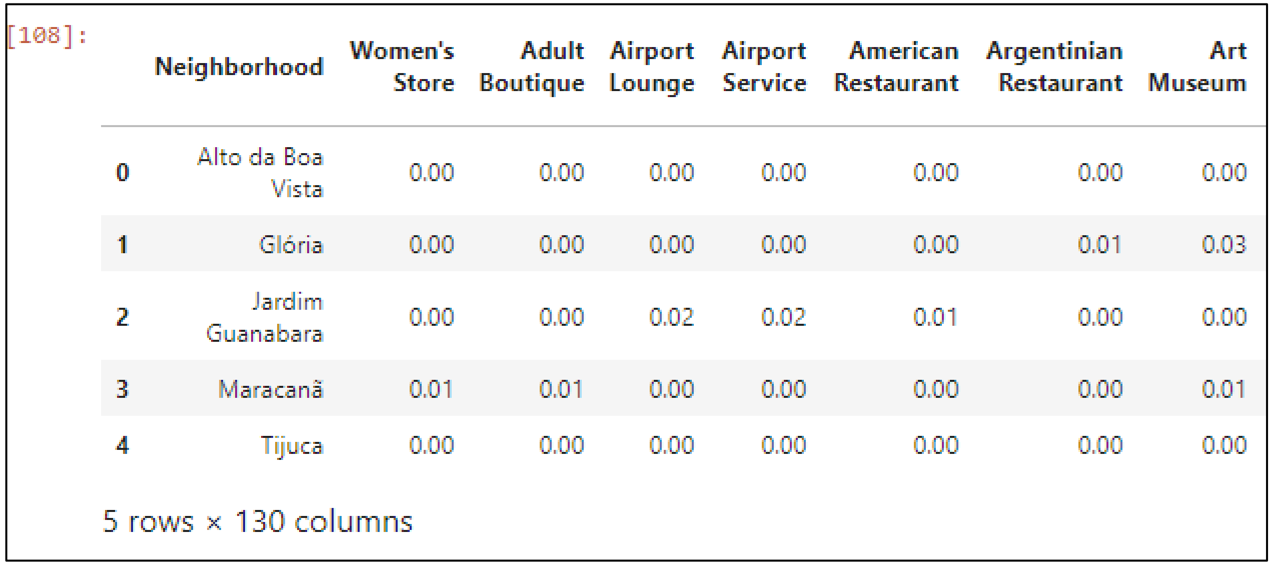
	Women's Store	Adult Boutique	Airport Lounge	Airport Service	American Restaurant	Argentinian Restaurant	Art Museum	Asian Restaurant	Athletics & Sports
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0

5 rows × 130 columns

3. Methodology (Part 5)

→ Venues Evaluation using Foursquare API

Grouping the venues by neighborhood



[108]:

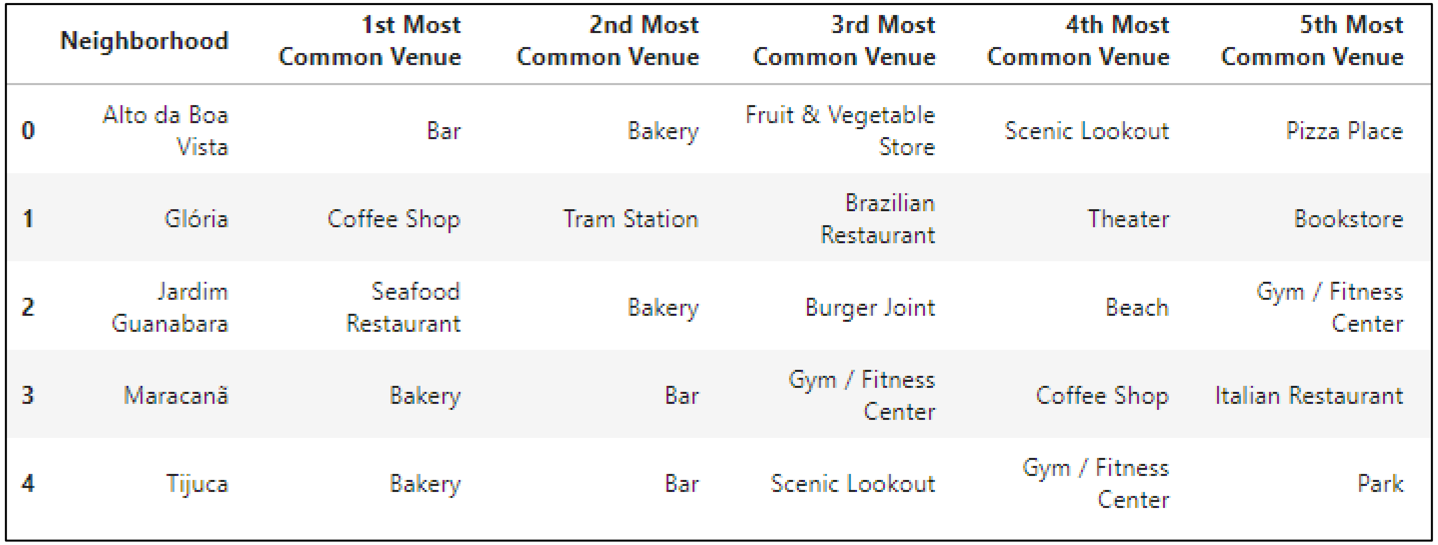
	Neighborhood	Women's Store	Adult Boutique	Airport Lounge	Airport Service	American Restaurant	Argentinian Restaurant	Art Museum
0	Alto da Boa Vista	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	Glória	0.00	0.00	0.00	0.00	0.00	0.01	0.03
2	Jardim Guanabara	0.00	0.00	0.02	0.02	0.01	0.00	0.00
3	Maracanã	0.01	0.01	0.00	0.00	0.00	0.00	0.01
4	Tijuca	0.00	0.00	0.00	0.00	0.00	0.00	0.00

5 rows × 130 columns

3. Methodology (Part 5)

→ Venues Evaluation using Foursquare API

Most common venues by neighborhood



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Alto da Boa Vista	Bar	Bakery	Fruit & Vegetable Store	Scenic Lookout	Pizza Place
1	Glória	Coffee Shop	Tram Station	Brazilian Restaurant	Theater	Bookstore
2	Jardim Guanabara	Seafood Restaurant	Bakery	Burger Joint	Beach	Gym / Fitness Center
3	Maracanã	Bakery	Bar	Gym / Fitness Center	Coffee Shop	Italian Restaurant
4	Tijuca	Bakery	Bar	Scenic Lookout	Gym / Fitness Center	Park

4. Results

On this section, will be presented all the data analysis about the neighborhoods in Rio de Janeiro city focusing in understanding and identify a good neighborhood to start an English School from a new franchise that coms to Brazil.

- 1st Discussion: In Rio de Janeiro, most of the people that has higher per capita income lives in the south area of the city and closer to the beach. That explains a lot for example why Copacabana, Ipanema and Leblon are neighborhoods highly looked for tourists to stay when they are in the city. Off course, crime rates in those areas are lower compared to other areas, like the suburb for example.

4. Results

- 2nd Discussion: Even with a big difference in per capita income in most of the neighborhoods, the city of Rio de Janeiro has a good Human Development Index and I think some reasons can explain that, two of them are:
  - 1) The city is not too big, that means most of activities are accessible by most of the people living in the city,
  - 2) As we can see in the other metrics the education rate and life longevity rate are very good even in areas where the per capita income are low. That values raises the average of the HDI (that is the calculated rate)

- 3rd Discussion: As said in the 2nd discussion, the HDI rate (IDH) is heavily influenced by the education rate (IDH-E) that raises the IDH rate average making the neighborhoods having not much difference in HDI. Because the city is not that big, the inequality rate is very high with slums present in almost every neighborhood of the city. In fact, I can conclude the problems that has in the city are not easily saw only looking to HDI metrics but the data is enough to see where people with more acquisition power lives.
  - To a company that is looking to know where to open an english school in Rio de Janeiro, the data is good enough.

4. Results

- 4th Discussion: If we look the box plots searching for inequality we can see the south area is more linear on that. In general, people that lives in the south area have better results in every metrics available and discussed here. The North area has some difference depending on the localization. How far north, worst is the conditions. The Downtown area usually only have commercial buildings, few people lives there compared to others area. The West zone is similar to the south area only around "Barra da Tijuca". The rest of the west area is more similar to the north area.

- 5th Discussion: As I suspected before looking the box plots, the distribution of south area is more linear when compared to the other areas. That means the inequality there is concentrated in some small slums around the neighborhoods. The others areas, such West and North has some outliers that comproves they are more unequal. There are some regions (outliers) that maybe have some opportunity there to establish an english school as some neighborhood has the same Per Capita Income and HDI from south area. Off course we will need to set a range for the price of the monthly subscription.

4. Results

- 6th Discussion: At least, the result of K-Means is very interesting! Look at the markers on the coast area (those are the south area) that are very asked from tourists to have a stay in RJ. Off course there are many english schools around and we can check on this later. But, the most interesting cluster is the "0" (zero) that point some area with similar metrics with good per capita income rate.

- 7th Discussion: The cluster 0 provided by K-Means was considered the "middle term" of the neighborhood in Rio de Janeiro. They are not totally in the south area that has many competition and the cost to start any business is high (ex: rentals) and they are not in the worse area of the city. To help to understand a bit more, I selected the cluster 0 result and applied a dendrogram to cluster the similar neighborhoods inside the cluster. The result was very good. The dendrogram creates 3 big clusters, the first one on the top is a kind of outlier. Urca and Botafogo are in the south area and it's any square is very disputed. Jardim Guanabara is a a good option as it is not in the south area and most of the people that lives there has good living conditions. The second cluster is a good option too, excluding Jardim Sulacap and Vidigal. The first one is too far from downtown and it's a kind of army village and the second one is a slum located in the south area. The other ones: Tijuca, Maracanã, Alto da Boa Vista and Glória can be considered a good option to start the business.

4. Results

- 8th Discussion: Looking at the data provided by the Foursquare API, it is possible to see, the 10 most frequents venues does not have any language school that can be a competitor against my customer. That's cool! Any of this neighborhood probably is a good location to start their business. To help to choose one location we will consider some human experience about the city and the neighborhoods. But, before that, I will check if there are any language school that it is not fitted in the tens and do some other checks looking for better infrastructure such metros, trams and restaurants.

- 9th Discussion: Some insights from the Foursquare data:
  - No english schools registered in the Foursquare API for those selected neighborhoods
    - But maybe there are some that is not registered in the Foursquare API (need a better check using other API or data source)
  - Maracanã and Glória has some Tram Stations
    - Quick note on this: Tijuca, Glória and Maracanã already have metro and train stations that wasn't returned by the Foursquare API
  - All of the selected neighborhoods has good restaurants to serve our customers during the classes period

5. Discussion

The data analysis using the census data from the Rio de Janeiro city is good enough to know and understand which neighborhood offer better conditions for some investment, such opening an English school.

Although, will be more interesting enrich the data sources using more data coming from another systems, such crime rates, metro usage, traffic hours and many others.

The Foursquare API is very nice but some places are not registered and the results based only on that can be imprecise.

6. Conclusion

Our Customer will be advised that three of the selected neighborhoods are good options to start their first english school in Rio de Janeiro according to the data source evaluated, they are:

- 1st: Glória
  - Glória is the 1st option because it is closer to the downtown and for people that works every day in offices, the proximity can be a good differential compared to the other neighborhoods. Glória is part of the south area of the city but because of the proximity with downtown the rental prices are not too high.

6. Conclusion

- 2nd: Tijuca
  - Tijuca is the 2nd option because it is the neighborhood in the north area that is more traditional. It is served by metro stations, the percapita income is similar with some neighborhoods of the south area (including Glória). The only cons is the distance from downtown compared to Glória.

6. Conclusion

- 3rd: Jardim Guanabara
  - Jardim Guanabara is the 3rd options because is too far from downtown and it is not good on transportations (only taxis and buses in there). The per capita income there is one of the best in the city also the HDI rate. It is a good option but only for people that lives there.

Thank You!

Rodrigo Eiras
rodrigoeiras.github.io