2. Data

The analysed data sustaining the project had two main sources: Public data hosted at “DataRio”, the official Rio de Janeiro Open Data Portal, and commercial data from the Foursquare Places API, which provides real-time access to Foursquare’s global database of venue data and user content.

2.1 Population data

In addition to foursquare location data, population data of the city of Rio de Janeiro for the year of 2010 were collected. This dataset is composed of 199 rows, each representing a neighborhood or administrative region of the city, which are listed in the first column of the dataset, with the exception of the first row which contains the total population of the city. Administrative regions’ data was not included for analysis. The second column lists the total population of each neighborhood or region.

The remaining 28 columns list the population of each neighborhood or region by age group. The age groups are divided as follows: A separate column for each age between less than one year and 15 years (16 columns) and one column for each of the following age groups: 16-17, 18-19, 20-24, 25-29, 30-34, 35-39, 40-49, 50-59, 60-64, 65-69, 70-74, 75-79, 80 or more (12 columns).

Total population between the ages of 16 and 29 was obtained and considered as the target population. The other population feature considered was the adult population, between 18 and 65 years old. A snapshot with the first rows and the columns of interest of the original table is exposed below:

Áreas de Planejamento, Regiões Administrativas e Bairros	rupos de idade			
	16 e 17 anos	18 e 19 anos	20 a 24 ano	25 a 29 ano
Total	******	******	******	**551 103**
Área de Planejamento 1	*8 057*	*8 585*	*25 751*	*28 226*
I Portuária	**1 485**	**1 529**	**4 639**	**4 589**
Caju	624	682	2 120	1 978
Gamboa	431	415	1 219	1 169
Santo Cristo	375	368	1 050	1 112
Saúde	55	64	250	330
II Centro	**727**	**864**	**3 024**	**4 153**
Centro	727	864	3 024	4 153
III Rio Comprido	**2 349**	**2 384**	**6 720**	**6 918**
Catumbi	389	384	1 006	1 035

Neighborhood population table.

The Jupyter Notebook containing the processing of this table can be downloaded at: [population_cleaning.ipynb]

The dataset is available at: [Population Dataset].

2.2 Neighborhood Data

Another dataset from the same portal above mentioned was acquired containing data on the neighborhoods. This dataset contains information on each neighborhood such as area, length, administrative region, official name and more (The remaining columns are not going to be commented on as they do not relate to the project proposal). It has 163 rows (the number of the official count of neighborhoods) and 14 columns. From this dataset, only the name and area of each neighborhood were used. A snapshot of the first rows and selected columns of the original table can be seen below.

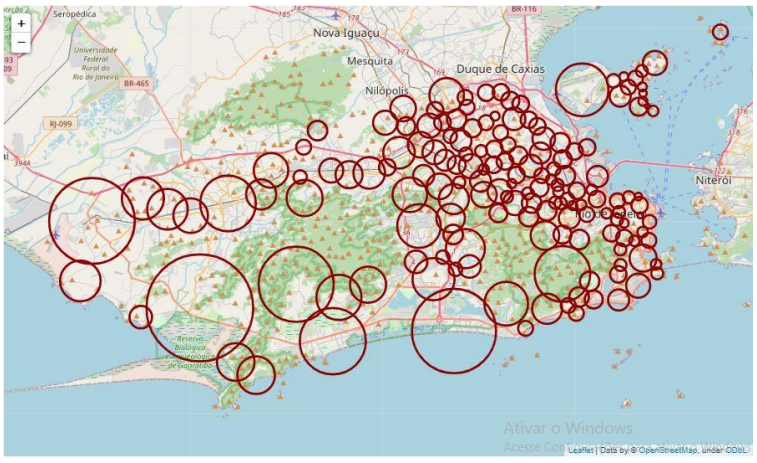
	OBJECTID	Área	NOME	REGIAO_ADM
0	325	1.705685e+06	Paquetá	PAQUETA
1	326	4.056403e+06	Freguesia (Ilha)	ILHA DO GOVERNADOR
2	327	9.780465e+05	Bancários	ILHA DO GOVERNADOR
3	328	1.895742e+07	Galeão	ILHA DO GOVERNADOR
4	329	1.672546e+06	Tauá	ILHA DO GOVERNADOR
5	330	1.186409e+06	Portuguesa	ILHA DO GOVERNADOR
6	331	5.205575e+05	Moneró	ILHA DO GOVERNADOR

Neighborhood table.

This dataset is available at: limite-de-bairros.

2.3 Neighborhoods Latitude, Longitude and Estimated Radius

The acquisition of the neighborhoods coordinates was done using the geopy library, by inputting the neighborhoods names to the library's geocode method. Using the acquired neighborhoods areas, radiuses values were estimated, which approximated the areas of the neighborhoods to circles so searches could be conducted in each neighborhood area. Manual corrections of the coordinates and radiuses were performed to improve results. Using the folium library, an interactive map was generated in order to visualize, evaluate and correct these values.



Rio de Janeiro neighborhood areas approximated to circles.

The Jupyter Notebook containing the collection and cleaning process of this data can be downloaded at: neighborhoods_data_collection.ipynb

2.4 Foursquare Gym Data

A search for gyms was performed inside the defined area of each neighborhood, using the foursquare places API 'Search' endpoint. For each search call, the parameters of the API request URL were the latitude and longitude values that approximate the neighborhood center, the estimated radius of the neighborhood which approximates the area of the neighborhood to a circle, and the id of the category 'gym'. This request configuration searches for venues categorized as gyms around the given coordinates, inside the given radius, and retrieves up to 50 venues per call. Only one search result reached the 50 venues limit, which did not compromise the overall quality of the search.

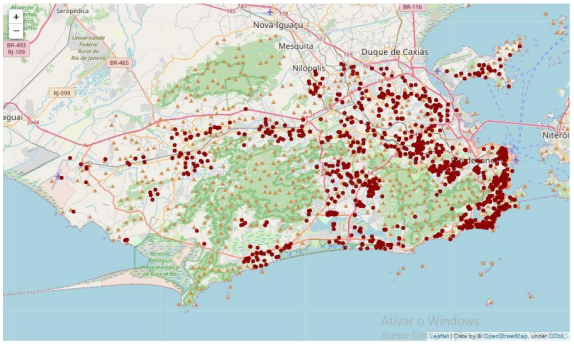
Each search resulted in a pandas dataframe where each row corresponded to a specific gym found inside the search area. For each venue found by the search the following information was kept, each in a column:

- venue name
- id
- category
- category id
- latitude
- longitude
- venue
- distance from search area center
- neighborhood

These data frames were generated and concatenated along the vertical axis iteratively, resulting in a final table composed of 1209 rows and 8 columns.

The search found a total of 972 gyms. The overlap between the search areas resulted in 237 gyms found more than once. These gyms were then considered to belong to more than one neighborhood as they are located near to the neighborhoods' borders.

Finally, the gyms were marked as dots on a map of the city so the result of the search could be visually evaluated.



Foursquare places API search for gyms in Rio de Janeiro.

The Jupyter Notebook containing the search can be downloaded at: gyms_data_collection.ipynb

2.5 Additional Data

Additional data on the neighborhoods was acquired in order to enrich the analysis and better select the neighborhoods for recommendation. These additional data consisted of two datasets, one on population income per neighborhood from 2010 and another containing counts of commercial establishments of industry sectors per neighborhood from 2016.

Only a few features from these two datasets were selected for analysis as follows,

Industry sectors features:

- Retailing
- Credit institutions, insurance and capitalization
- Trade and administration of real estate, securities, technical services
- Accommodation, food, repair, maintenance, writing services
- Medical, dental and veterinary services
- Education

Income features:

- Population with income responsible for the permanent private residences per Km2.
- Total monthly income of the population responsible for the permanent private residences per Km2.
- Average nominal income of the population responsible for the permanent private residences.

The Jupyter Notebooks containing the processing of these tables can be downloaded respectively at: renda_cleaning.ipynb & estabelecimentos_cleaning.ipynb

The datasets are available respectively at: Income Dataset & Commercial Establishments Dataset