Applied Data Science Capstone Project Report

1. Introduction: Business Problem

1.1 Contextualization

Gym owners or new investors interested in building a new gym in the city of Rio de Janeiro .m face the challenge of having to choose between 163 different neighborhoods, each with its own characteristics and peculiarities. Popular neighborhoods already contain lots of gyms while others have very few. Average population age in traditional neighborhoods is usually higher, while neighborhoods containing universities typically have many young residents. The presence of different types of commercial establishments in each neighborhood also have an impact on the investment risks since it tells a lot about the neighborhoods realities. Any new gym investor or current gym owner might be interested in knowing what neighborhoods in the city will offer the best conditions for the success of his/her new gym.

1.2 Business Problem

What are the best neighborhoods in Rio de Janeiro to start a new gym? That's the business problem this project aims to solve. The final deliverable of the project should be a list of recommended neighborhoods.

2. Data

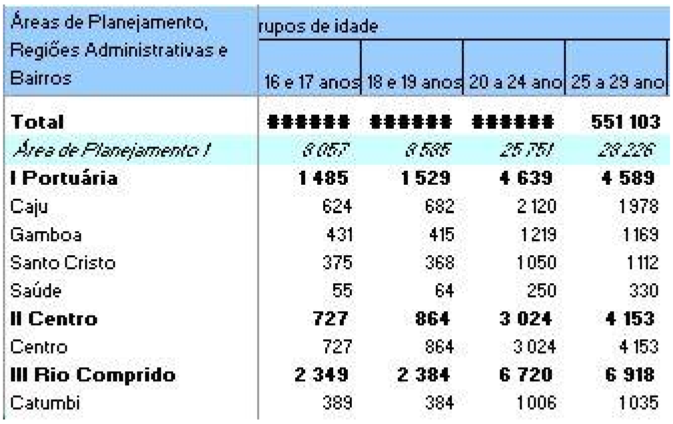
The analysed data sustaining the project had two main sources: Public data hosted at "DataRio", the official Rio de Janeiro Open Data Portal, and commercial data from the Foursquare Places API, which provides real-time access to Foursquare's global database of venue data and user content.

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 I	*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

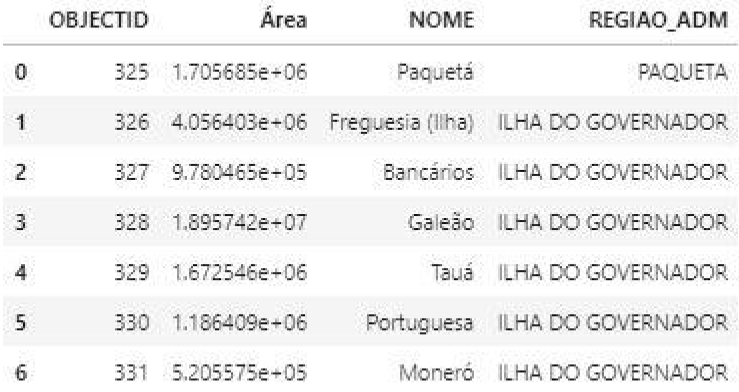
Neighborhood population table.

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.



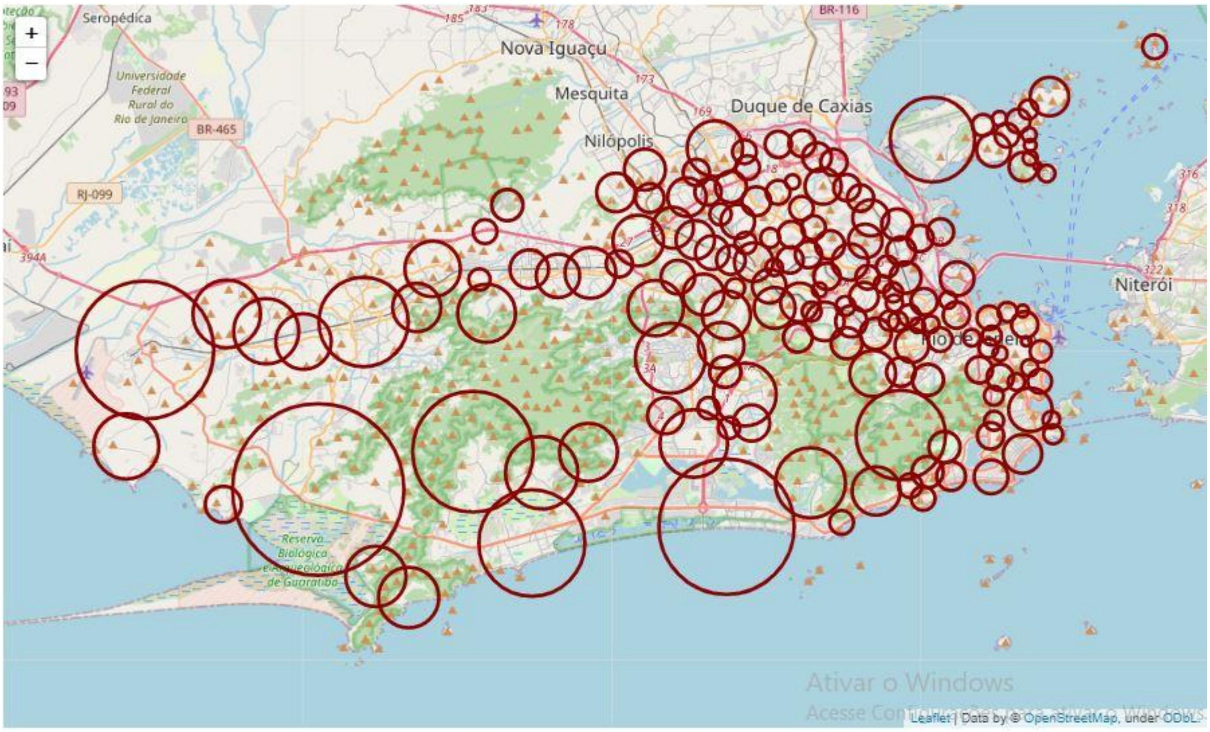
	OBJECTID	Área	NOME	REGIAO_ADM
0	325	1.705685e+06	Paquetá	PAQUETA
1	326	4.056403e+06	Freguesia (Ilha)	ILHA DO GOVERNADOR
2	327	9.780465e+05	Bancários	ILHA DO GOVERNADOR
3	328	1.895742e+07	Galeão	ILHA DO GOVERNADOR
4	329	1.672546e+06	Tauá	ILHA DO GOVERNADOR
5	330	1.186409e+06	Portuguesa	ILHA DO GOVERNADOR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


Seropédica
Universidade Federal Rural do Rio de Janeiro
BR-465
RJ-099
Nova Iguaçu
Mesquita
Nilópolis
Duque de Caxias
BR-116
Niterói
Rio de Janeiro
Reserva Biológica e Arqueológica de Guaratiba
Ativar o Windows
Leaflet | Data by © OpenStreetMap, under ODbL.

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.

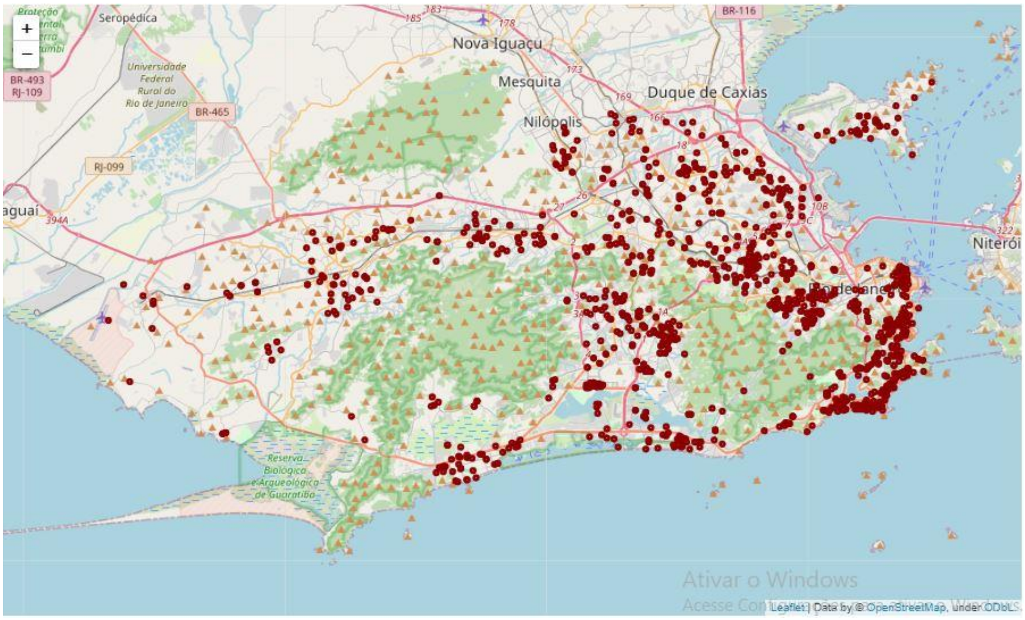
Each search resulted in a pandas dataframe where each row corresponded to a specific gym found inside the search area. For each venue found by the search the following information was kept, each in a column:

- 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.


+
−
Seropédica
Nova Iguaçu
Mesquita
Nilópolis
Duque de Caxias
BR-116
BR-493
RJ-109
BR-465
RJ-099
Universidade Federal Rural do Rio de Janeiro
Niterói
Reserva Biológica e Arqueológica de Guaratiba
Ativar o Windows
Leaflet | Data by © OpenStreetMap, under ODbL.

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 at: Income Dataset & Commercial Establishments Dataset

3. Methods

3.1 Data Cleaning

A total of ten datasets containing data specifically on the neighborhoods were downloaded, processed and combined into a single dataset.

The datasets are:

- Foursquare API gyms
- Demographics
- Neighborhoods geolocation and area
- Income (Ten years old or more)
- Income (Homeowners)
- Industry sectors salary mass
- Industry sectors establishments
- Industry sectors employees
- Energy consumption

These datasets combined composed the available features for the analysis. However, only relevant features were included and are going to be commented on.

The data processing stage involved many data manipulation techniques. Some of the original tables were far from the right format for analysis requiring major editions or even the entire reconstruction of the table while others required just a few adjustments.

Most frequently applied techniques were table formattings, such as index or column settings, table slicing and manipulation, dropping or replacing missing values and type conversions.

The major effort needed to combine the datasets was the uniformization of the neighborhood's names. There were lots of variants of the neighborhoods names among the datasets such as differences in the use of lower or upper case letters, accentuation and spelling. Therefore, a uniformization of the names was required so that the data could be correctly associated when merging the datasets.

Once the datasets were properly merged the resulting dataset was saved as a "csv" file for later importation to the analysis notebook.

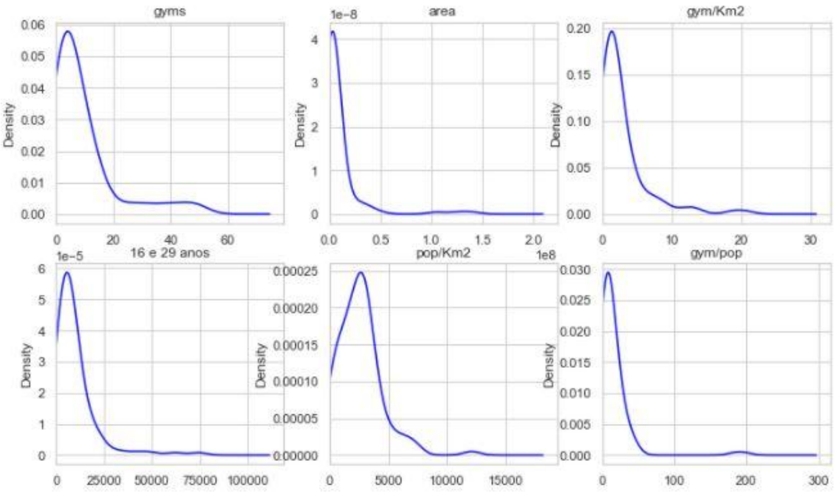
The Jupyter Notebook containing the additional cleaning and the merging steps can be downloaded at: Project_data_preparation.ipynb

3.2 Feature Engineering

The feature engineering step involved the estimation of three more features. These are the neighborhood's demographic densities of the target and adult populations, gym counts found by the foursquare API search per Km^2 and per 10000 inhabitants. Demographic density was calculated by dividing the mentioned population by the neighborhood area in square kilometers (Km^2). Similarly the second feature was calculated by dividing the number of gyms found by the search in a neighborhood by the area of the neighborhood in square kilometers (Km^2). The third new feature was estimated by dividing the number of gyms found by the search in a given neighborhood by the target population, and multiplying the result by 10000. The feature engineering steps can be checked at the end of the last Jupyter Notebook link. The remaining steps can be found either in the simplified or in the complete version of the main project notebook.

3.3 Exploratory Data Analysis

All six features studied are positively skewed as revealed by the plotting of their kernel density estimate functions. This means that they generally fall around low values, less frequently reaching high values.



Kernel Density Estimate Functions - Initial Variables

3.4 Data Transformation

All features fed to the segmentation models built in the project were scaled using the Standard Scaling method, which consists of dividing each sample point (x) value by the sample average (μ) and dividing the result by the sample standard deviation (σ).

$$x_s = \frac{(x - \mu)}{\sigma}$$

The process is also called normalization and reduces the skewness level of the features by converting them to normal distributions. It also ensures better functioning of the K-Means clustering algorithm since the feature's scales get more comparable.

3.5 Clustering Algorithm

Since the goal is to select the "best" neighborhoods for the new enterprise, a target neighborhood profile must be set, based on the available features, so neighborhoods that match this profile can be identified and recommended. The definition of this profile was achieved by interpreting each feature individually and defining whether the feature contributes to client acquisition by being maximal or minimal. For example, the "demographic concentration" feature generally contributes to client acquisition when increased. But the concentration of gyms in the neighborhood generally contributes to client acquisition by being minimal. The final target neighborhood profile was defined as follows

Features for maximization:

- Income
- Demographics
- Industry Sectors

Features for minimization:

- Gyms related features

One problem with this approach is that there may not be such neighborhoods that match the profile entirely, that is, minimize and maximize all the right features.

The k-means algorithm was chosen as the appropriate clustering tool to compartmentalize the sample because of the simple feature space geometries found within the sample as it'll be later exposed.

The process of actually selecting the neighborhoods that match the profile had four steps. Each step dealt with a group of features in order to simplify the interpretation and visualization of results. The groups of features are: Population, Income, Commerce and Gyms.

3.6 The optimization

The optimization procedure was developed specifically for this project although it can be applied for other clustering attempts as well. It's based on linear regression and the basic idea behind it is to measure how similar groups of clusters are by fitting each of them to associated lines or surfaces, depending on the relationship between the variables, and then calculating their respective mean squared error (R2) values for each number "n" of clusters. The intra-cluster distance between sample points starts to decrease more slowly once the number of clusters is sufficient to correctly distingh the natural clusters found in the sample. Therefore, the R2 value of the associated line or surface regressions and the average R2 stops to increase significantly and the "elbow method" can be applied in order to select the optimal value for the number of clusters.

For an example, consider the case where the feature space consists of three features, "x", "y" and "z" and

$$z = x/y$$

Once the number of segments of the feature space starts increasing, we can assume that the values of z will fall into very short intervals for each cluster, since the k-means algorithm naturally subdivides each feature axis into several intervals. Therefore we can assume that the values of z for each cluster belong to short intervals for large “n” values. If we plot x versus y on a plane and differentiate the points of each cluster with different colors we can observe that the clusters fall into straight lines. This is because z is the ratio between x and y, so, for the same cluster, the x and y proportion tend to be close to constant or belonging to a short interval, and the lines can be fitted to the following relationship:

$$y = a.x$$

where the “a” coefficient can be determined by the least square method for each cluster for each value of “n”. Therefore, the R2 measure tells how tight the interval of the values of z is for each cluster and the lines can be fitted to simple linear regression models and the average R2 metric calculated for each number “n” of clusters. The same principles can be applied to the fitting of z versus y, but for the curves with the following equation:

$$z = a.(1/y)$$

where the “a” coefficient can be determined by the least square method for each cluster for each value of “n”. The same logic applies here, where x is considered to be constant for each cluster when “n” is large enough, and the above relationship to be valid. This means that regressions between “z” and “y” can also measure the quality of the segmentation and the same goes for “z” versus “x” as in

$$z = a.x$$

The “a” coefficient can be interpreted as an approximation of the cluster average value for that supposedly constant feature that the “a” coefficient is replacing. For example, the “a” in the equation above replaces 1/y. Therefore the “1/a” value of each cluster represents their average “y” value, at each number “n” of clusters.

These three measures can be calculated and plotted independently, but they can also be average out into a final metric. Variations of these techniques were adopted for the optimization of the number of clusters when applying the K-Means algorithm.

In each step, the neighborhood sample is segmented into groups by inputting the step's features into the k-means algorithm. Optimization processes are applied in the first two steps in order to define the optimal number of clusters in each case.

In each step, once the optimal number of clusters is defined and the neighborhood sample is divided into similar groups, the group that most approximates the profile is then extracted and its neighborhood names stored.

This process is repeated for each of the first three feature groups using all neighborhoods each time. Finally, the intersection of the selected neighborhoods from each of the first three steps is extracted. This list of neighborhoods are considered generally recommended neighborhoods for enterprises, since "gym" features have not been analyzed yet.

The segmentation process is then repeated for the last group of features, the "gym" group, only for the selected neighborhoods of the previous steps. Again, the cluster most closely related to the profile is extracted resulting in the final list of recommended neighborhoods.

4. Results

4.1 Income Clusters

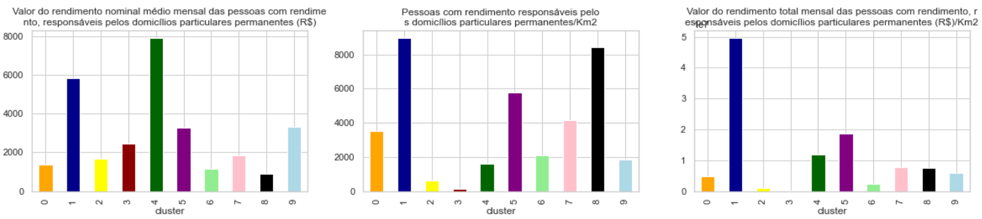
The first segmentation process divided the city neighborhoods into 10 groups. This number was chosen based on the optimization results which can be checked out on the Jupyter Notebook containing the analysis.

All three variables of the income group of features relate to home owners income only. In other words, this step of the analysis aims at the neighborhoods' residents income.

The first variable is the average homeowner monthly income of each neighborhood. The second is the area concentration of homeowners in each neighborhood and the last is the area concentration of the total income of homeowners per neighborhood.

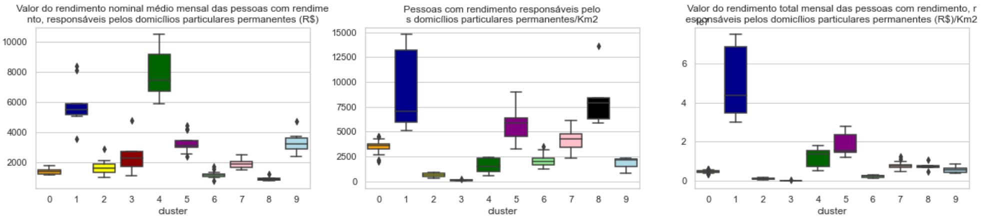
It's logical that companies should benefit from higher values of any of these variables, since they represent more concentration of money and therefore of the likelihood of potential customers. Therefore, this segmentation aimed at isolating the neighborhoods which maximize all three variables at once from neighborhoods which maxime only one, two or none of the variables.

In order to select the right group of neighborhoods, which maximized all three variables, the clusters profiles were compared either through bar plots or box plots. Besides, since this segmentation included only three variables, the clusters can actually be represented in three dimensions through a scatter plot. A color scheme was applied so clusters could be recognized from one plot to the next.



Income Clusters Averages Bar Plots

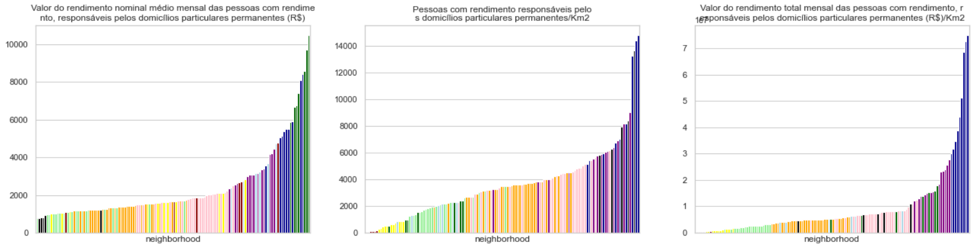
The bar charts above display the averages of the variables for each of the 10 clusters found. What's interesting about it is that the averages represent a profile of each cluster. For example, the neighborhoods of the cluster represented by the blue color has a high average for all three variables, while the light green cluster has a low average for all three variables.



Income Clusters Boxplots

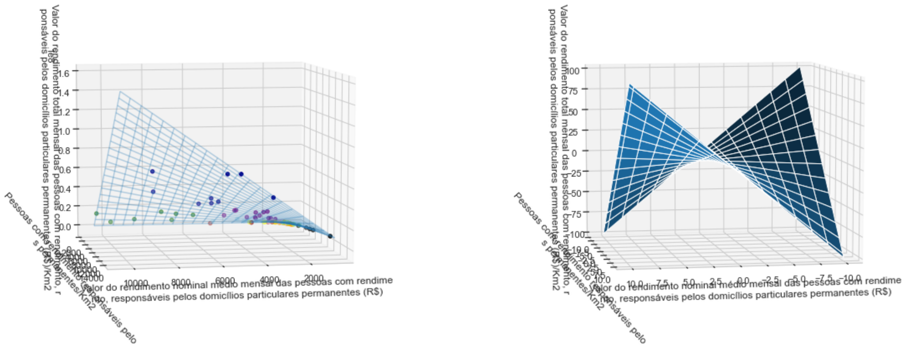
Furthermore, the boxplots above expose the variables' distributions of each group of neighborhoods found by the algorithm. In the first plot, it's evident that the dark green cluster includes the neighborhoods with the highest of homeowners average incomes. But at the same time, those neighborhoods do not have high area concentrations of homeowners. So even though homeowners in those neighborhoods have higher average income, they are too spread out in the neighborhood area even due to a big area or low number of homeowners.

And the same logic applies to the third variable, the area concentration of income. By selecting the dark blue cluster, we make sure that neighborhoods for recommendation have optimized conditions for all variables.



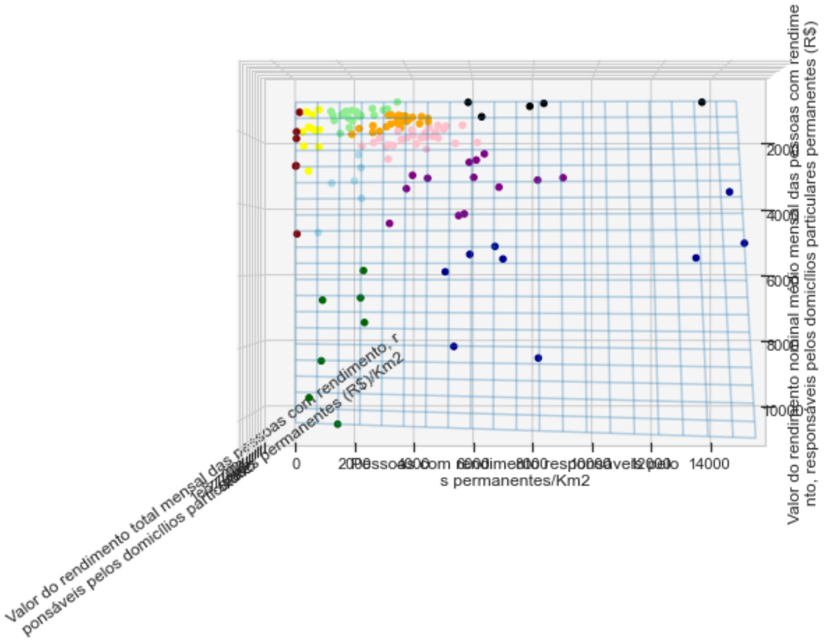
Income clusters bar plots

Another way of comparing the clusters is to sort the neighborhoods by each variable and represent each neighborhood with a bar. Note how the dark blue group is the only which is consistently found in the right side of each plot and how the dark green group is at the right side in the first plot only.



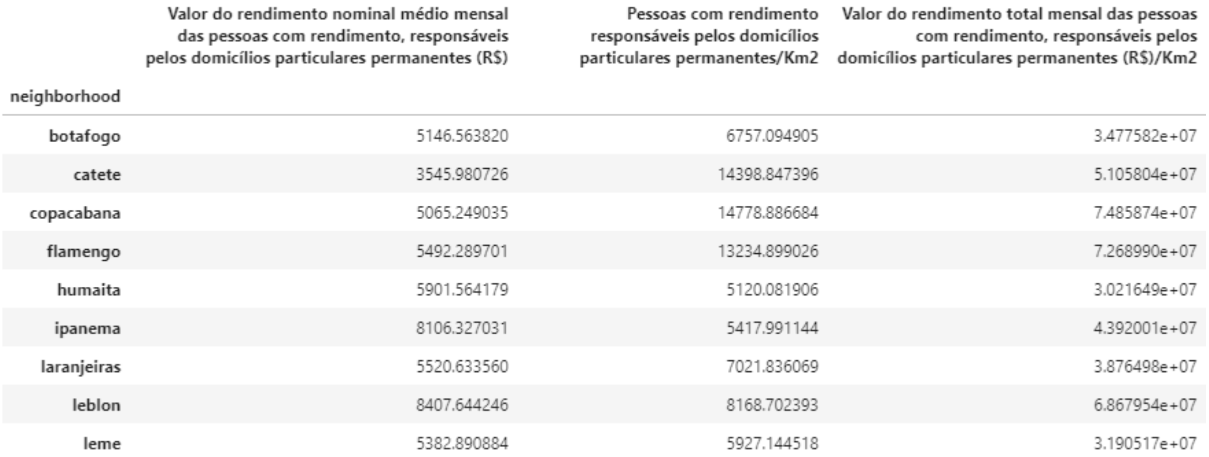
Income clusters 3D scatter and surface plots.

As mentioned, since the cluster model includes only three variables, the clusters can actually be represented in three dimensions. The points follow a surface geometry because of the relationship between the variables as best explained in the Jupyter Notebook containing the analysis. Note how the dark blue cluster is composed of dots (neighborhoods) which are the most distant to the origin.



Income clusters scatter plot - Top view

Finally, the neighborhoods belonging to the dark blue cluster can be selected and it's values compared.



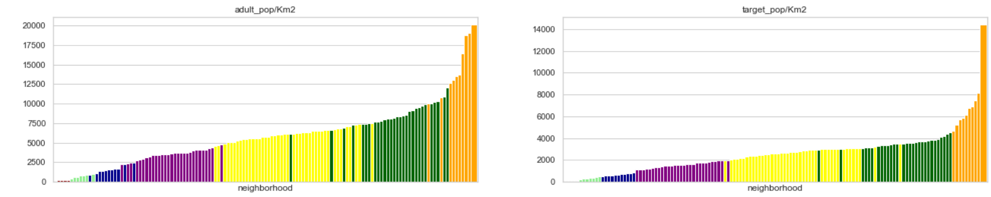
	Valor do rendimento nominal médio mensal das pessoas com rendimento, responsáveis pelos domicílios particulares permanentes (R$)	Pessoas com rendimento responsáveis pelos domicílios particulares permanentes/Km2	Valor do rendimento total mensal das pessoas com rendimento, responsáveis pelos domicílios particulares permanentes (R$)/Km2
neighborhood			
botafogo	5146.563820	6757.094905	3.477582e+07
catete	3545.980726	14398.847396	5.105804e+07
copacabana	5065.249035	14778.886684	7.485874e+07
flamengo	5492.289701	13234.899026	7.268990e+07
humaita	5901.564179	5120.081906	3.021649e+07
ipanema	8106.327031	5417.991144	4.392001e+07
laranjeiras	5520.633560	7021.836069	3.876498e+07
leblon	8407.644246	8168.702393	6.867954e+07
leme	5382.890884	5927.144518	3.190517e+07

Income based Neighborhood Recommendations.

The exact same steps were applied to the next two groups of features, the population and the commerce groups. But, for the gym group there were two main differences. The first is that the segmentation aims at isolating the neighborhoods which minimize all variables of the gym feature group. The second is that only neighborhoods which were selected in the first three segmentation processes were compared. Because of the low number of neighborhoods for analysis in this step, a visual análises was enough to select the final neighborhoods for recommendation and no clustering algorithm was applied. The next steps' results can be checked below.

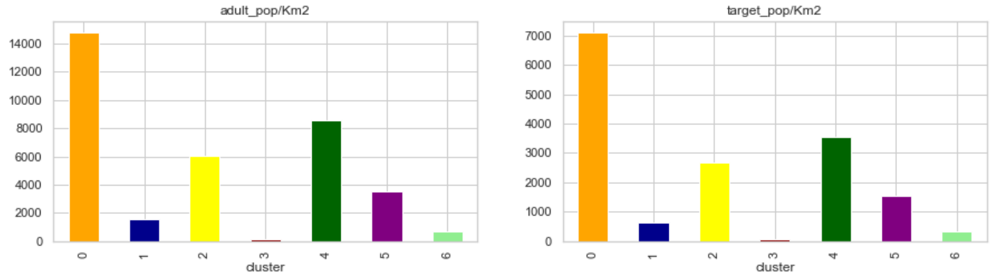
4.2 Demography Clusters

The two demographic features that composed the demographic cluster model were very similar. They are the demographic concentration of the adult population (18 to 65 years old) and the demographic concentration of the target population (16 to 29 years old). The reason for the similarity is a linear relationship between the variables. This means that it's impossible to isolate a group that has a high value for the first value but a low value of the second.



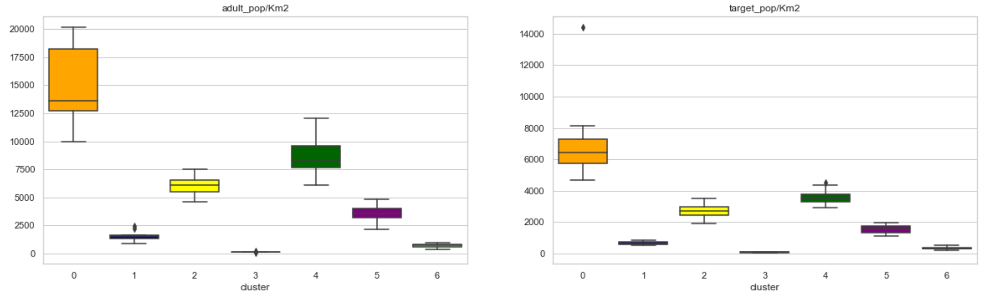
Population clusters bar plots.

The fact that the two variables are related is actually helpful, since the goal here is to detect and extract neighborhood groups which maxime both variables at the same time. The orange and green groups were then selected for recommendation based on adult and target population concentration. The choice to include the green group is that the orange group contains the most crowded neighborhoods in the city, which can actually be problematic. By choosing to include the green groups we get a more flexible result since we'll have a larger interval of recommendation. In this way, stakeholders can choose a demographic density level that suits their needs.



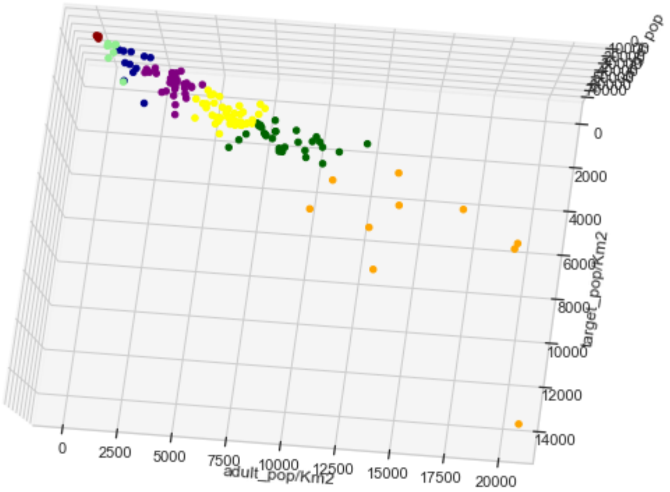
Population Clusters Averages Bar Plots.

The average bar plots again show the middle tendency of each neighborhood group. It's clear how the clusters represent gradual decreased demographic density.



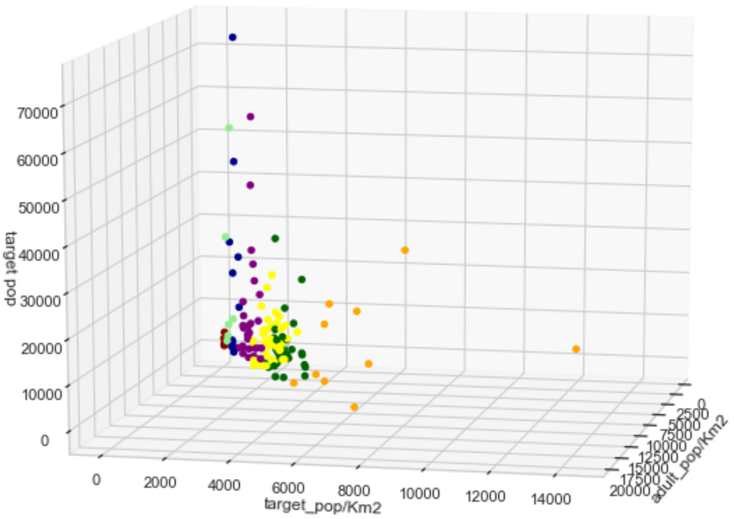
Population Clusters Boxplots.

The clusters boxplots again allow the comparison of the cluster's distributions. We see very little overlap between clusters in both variables because of the linear relationship between the variables.

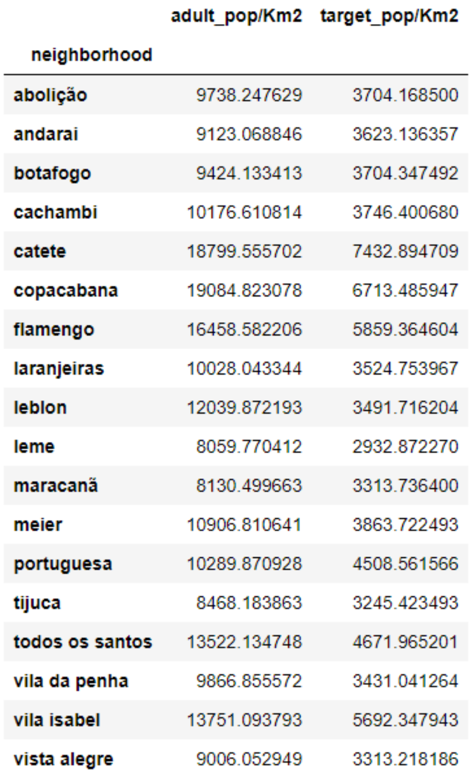


Population clusters 3D scatter plot - Top view.

The clusters can also be visualized in two dimensions as in the plot above. Despite the fact that the segmentation based on population features had only two variables, the clusters could be visualized in three dimensions with the inclusion of a third variable, the total target population.



Population clusters 3D scatter plot - 24°



	adult_pop/Km2	target_pop/Km2
neighborhood		
abolição	9738.247629	3704.168500
andarai	9123.068846	3623.136357
botafogo	9424.133413	3704.347492
cachambi	10176.610814	3746.400680
catete	18799.555702	7432.894709
copacabana	19084.823078	6713.485947
flamengo	16458.582206	5859.364604
laranjeiras	10028.043344	3524.753967
leblon	12039.872193	3491.716204
leme	8059.770412	2932.872270
maracanã	8130.499663	3313.736400
meier	10906.810641	3863.722493
portuguesa	10289.870928	4508.561566
tijuca	8468.183863	3245.423493
todos os santos	13522.134748	4671.965201
vila da penha	9866.855572	3431.041264
vila isabel	13751.093793	5692.347943
vista alegre	9006.052949	3313.218186

Population based neighborhood recommendations.

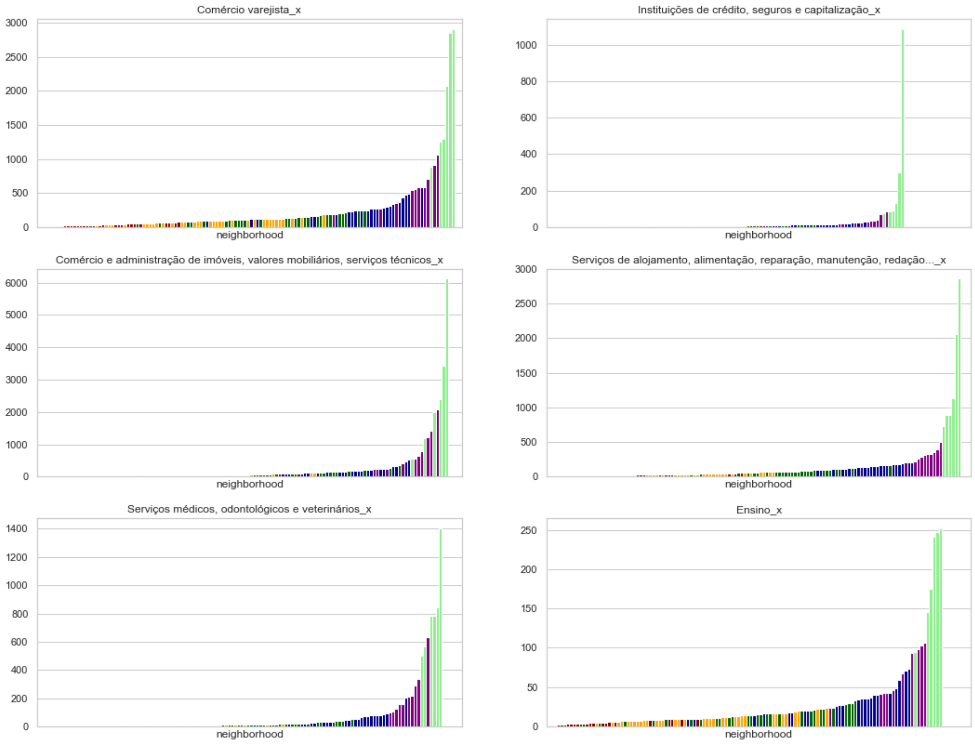
Despite the linearity between the variables, we see the discrepancy of their values. What's noticeable is that the two selected clusters here include many of the neighborhoods selected based on the income features, which means that some of the rich neighborhoods are also some of the most dense as well.

4.3 Comercial Clusters

The included establishment counts per sector, which are being called the commercial features, belong to the following commercial sectors :

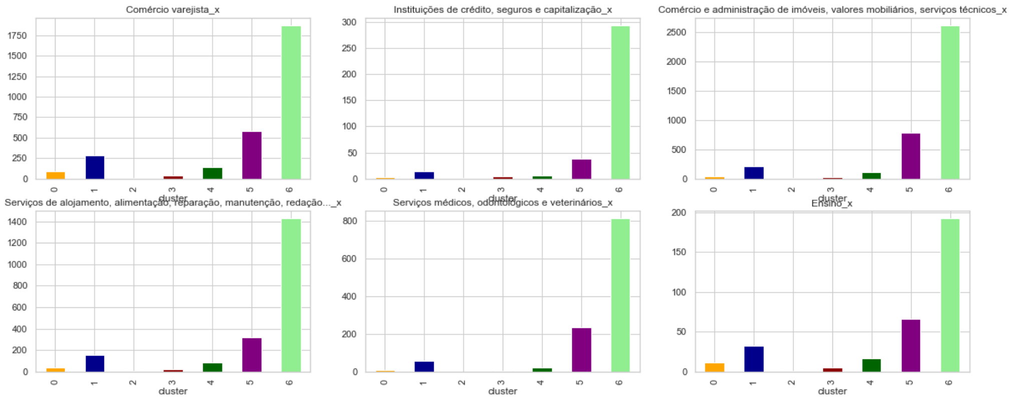
- Retailing
- Credit institutions, insurance and capitalization
- Trade and administration of real estate, securities, technical services
- Accommodation, food, repair, maintenance, writing services
- Medical, dental and veterinary services
- Education

The goal again is to select neighborhoods which present high establishment counts for every commercial sector included.



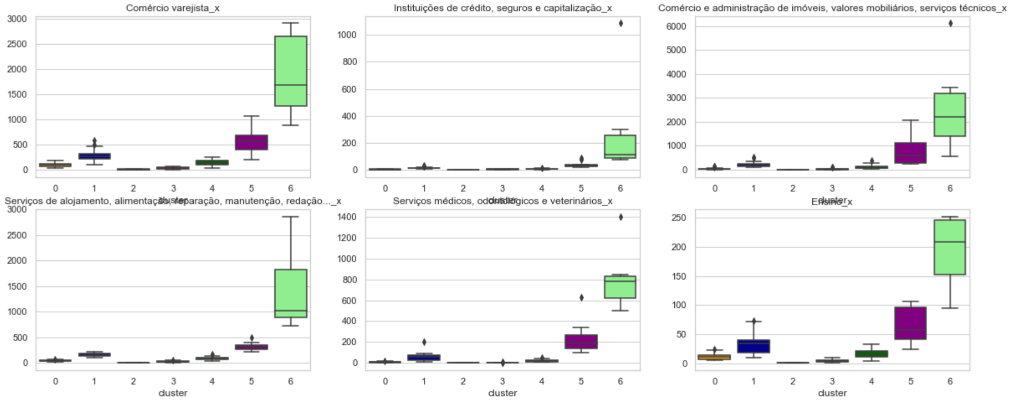
Commercial sectors clusters bar plots.

The number of clusters chosen for the K-Means algorithm was again selected based on trial and error until the natural clusters were correctly distinguished. It's evident how the light green and the purple clusters fill the right side of the plot, being the groups of neighborhoods that contain the highest numbers of establishments of the included Commercial sectors.



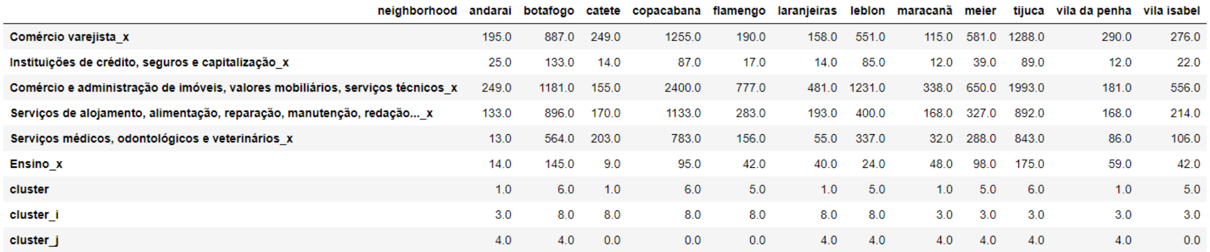
Commercial sectors clusters averages bar plots.

The difference in scale between the selected groups (light green and purple) and the rest of the groups is also very evident.



Commercial sectors clusters box plots.

It's clear how the algorithm has distinguished the city's neighborhoods which present the highest levels of commercial activity from those which present medium and low levels.



neighborhood	andarai	botafogo	catete	copacabana	flamengo	laranjeiras	leblon	maracanã	meier	tijuca	vila da penha	vila isabel
Comércio varejista_x	195.0	887.0	249.0	1255.0	190.0	158.0	551.0	115.0	581.0	1288.0	290.0	276.0
Instituições de crédito, seguros e capitalização_x	25.0	133.0	14.0	87.0	17.0	14.0	85.0	12.0	39.0	89.0	12.0	22.0
Comércio e administração de imóveis, valores mobiliários, serviços técnicos_x	249.0	1181.0	155.0	2400.0	777.0	481.0	1231.0	338.0	650.0	1993.0	181.0	556.0
Serviços de alojamento, alimentação, reparação, manutenção, redação..._x	133.0	896.0	170.0	1133.0	283.0	193.0	400.0	168.0	327.0	892.0	168.0	214.0
Serviços médicos, odontológicos e veterinários_x	13.0	564.0	203.0	783.0	156.0	55.0	337.0	32.0	288.0	843.0	86.0	106.0
Ensino_x	14.0	145.0	9.0	95.0	42.0	40.0	24.0	48.0	98.0	175.0	59.0	42.0
cluster	1.0	6.0	1.0	6.0	5.0	1.0	5.0	1.0	5.0	6.0	1.0	5.0
cluster_i	3.0	8.0	8.0	8.0	8.0	8.0	8.0	3.0	3.0	3.0	3.0	3.0
cluster_j	4.0	4.0	0.0	0.0	0.0	4.0	4.0	4.0	4.0	4.0	4.0	0.0

Commercial sectors based neighborhood recommendation table.

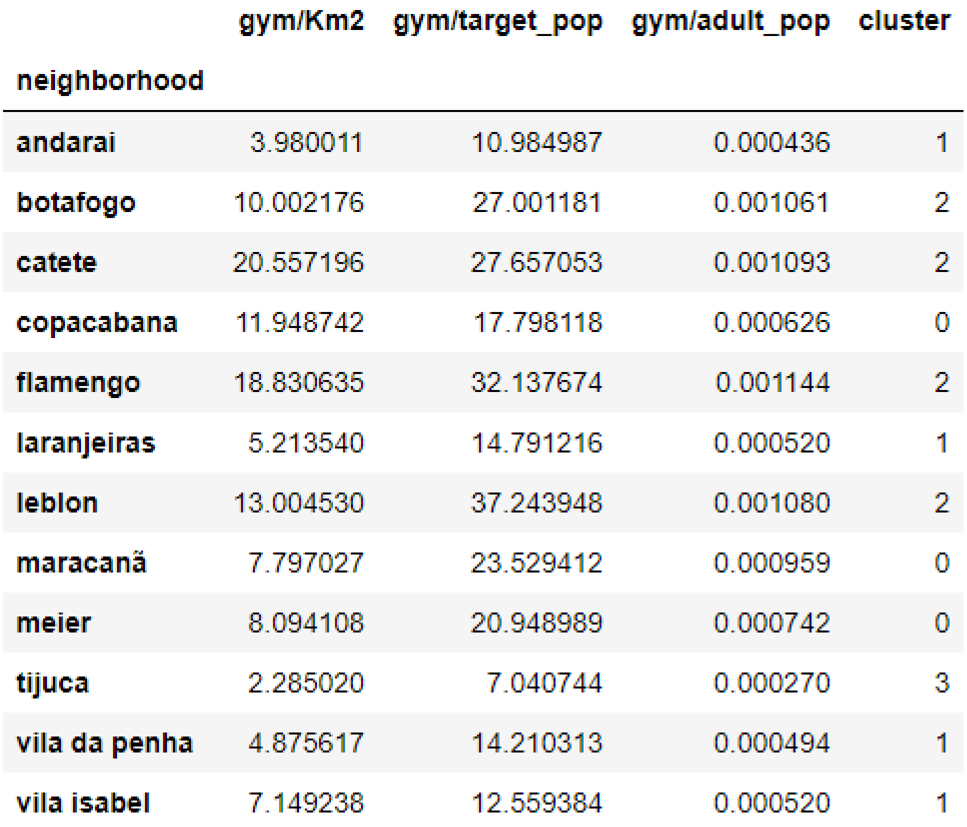
By selecting not only the light green but the purple cluster as well we make sure that the commerce based recommendation list matches a more diverse set of possible managers or stakeholders interests. Therefore the list can be consulted in order to select the best neighborhood profile for each scenario.

It's also evident that neighborhoods which were selected in the previous steps showed up again, meaning that there are neighborhoods which present optimal conditions for all features analysed until this point. This is a very promising result which generated a general purposes neighborhood recommendation list. That is, a list of neighborhoods recommended for many possible investments and enterprises.

This list led to the possibility of analysing the presence of gyms in these neighborhoods only.

4.4 Gyms

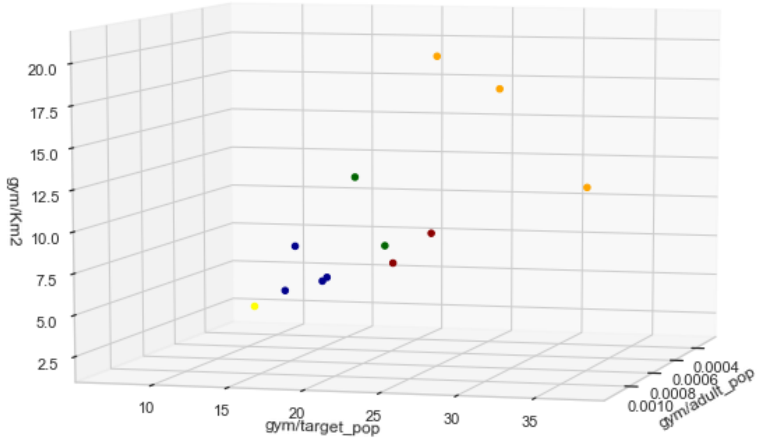
The general purpose recommendation list which composed the neighborhood sample for the study of the presence of gyms is exposed below, along with their values for the gym group of features.



	gym/Km2	gym/target_pop	gym/adult_pop	cluster
neighborhood				
andarai	3.980011	10.984987	0.000436	1
botafogo	10.002176	27.001181	0.001061	2
catete	20.557196	27.657053	0.001093	2
copacabana	11.948742	17.798118	0.000626	0
flamengo	18.830635	32.137674	0.001144	2
laranjeiras	5.213540	14.791216	0.000520	1
leblon	13.004530	37.243948	0.001080	2
maracanã	7.797027	23.529412	0.000959	0
meier	8.094108	20.948989	0.000742	0
tijuca	2.285020	7.040744	0.000270	3
vila da penha	4.875617	14.210313	0.000494	1
vila isabel	7.149238	12.559384	0.000520	1

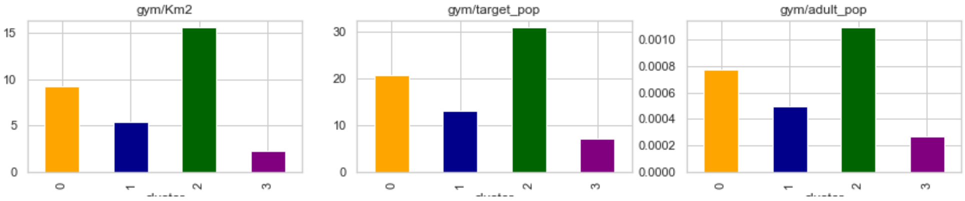
Neighborhoods included for the gym analysis - Gym group of features.

The three features obtained for the analysis of the presence of gyms all represent gym concentration. The first feature in the table above is the number of gyms per area in Km², the second per 10000 inhabitants of the target population (16 to 29 years old) and the third per 10000 inhabitants of the adult population (18 to 65 years old). The goal evidently is to select the neighborhoods presenting low concentration of gyms both per area and per population.



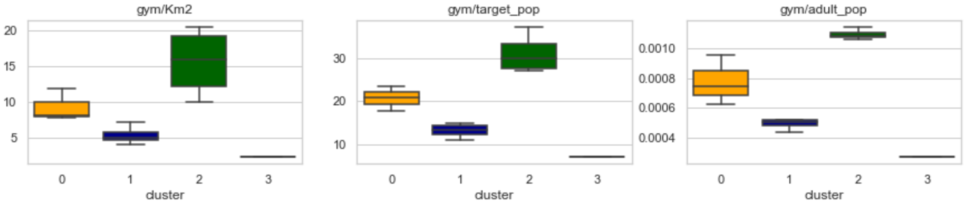
Gyms clusters 3D scatterplot.

We have a much smaller sample this time, but, luckily, the features variables have considerable amplitudes, meaning that there is indeed space to select neighborhoods with lower concentrations of gyms.



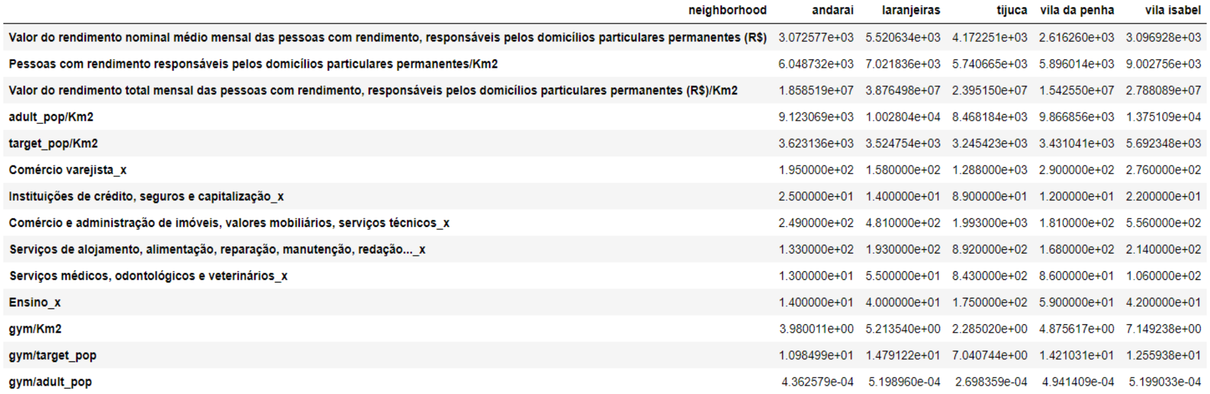
Gyms clusters average bar plots.

The bar plots highlight the discrepancy of the presence of gyms of the groups found by the algorithm.



Gyms clusters boxplots.

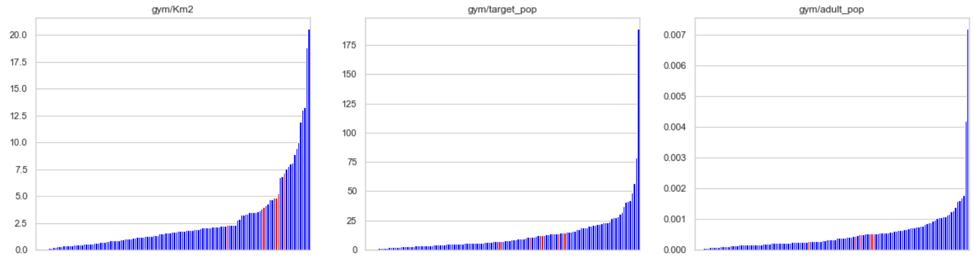
The boxplots above show that some neighborhoods are three to five times more concentrated with gyms than others. Clusters 1 and 3 which present the lowest concentrations of gyms both per area and population were then selected for recommendation.



neighborhood	andarai	laranjeiras	tijuca	vila da penha	vila isabel
Valor do rendimento nominal médio mensal das pessoas com rendimento, responsáveis pelos domicílios particulares permanentes (R$)	3.072577e+03	5.520634e+03	4.172251e+03	2.616260e+03	3.096928e+03
Pessoas com rendimento responsáveis pelos domicílios particulares permanentes/Km2	6.048732e+03	7.021836e+03	5.740665e+03	5.896014e+03	9.002756e+03
Valor do rendimento total mensal das pessoas com rendimento, responsáveis pelos domicílios particulares permanentes (R$)/Km2	1.858519e+07	3.876498e+07	2.395150e+07	1.542550e+07	2.788089e+07
adult_pop/Km2	9.123069e+03	1.002804e+04	8.468184e+03	9.866856e+03	1.375109e+04
target_pop/Km2	3.623136e+03	3.524754e+03	3.245423e+03	3.431041e+03	5.692348e+03
Comércio varejista_x	1.950000e+02	1.580000e+02	1.288000e+03	2.900000e+02	2.760000e+02
Instituições de crédito, seguros e capitalização_x	2.500000e+01	1.400000e+01	8.900000e+01	1.200000e+01	2.200000e+01
Comércio e administração de imóveis, valores mobiliários, serviços técnicos_x	2.490000e+02	4.810000e+02	1.993000e+03	1.810000e+02	5.560000e+02
Serviços de alojamento, alimentação, reparação, manutenção, redação..._x	1.330000e+02	1.930000e+02	8.920000e+02	1.680000e+02	2.140000e+02
Serviços médicos, odontológicos e veterinários_x	1.300000e+01	5.500000e+01	8.430000e+02	8.600000e+01	1.060000e+02
Ensino_x	1.400000e+01	4.000000e+01	1.750000e+02	5.900000e+01	4.200000e+01
gym/Km2	3.980011e+00	5.213540e+00	2.285020e+00	4.875617e+00	7.149238e+00
gym/target_pop	1.098499e+01	1.479122e+01	7.040744e+00	1.421031e+01	1.255938e+01
gym/adult_pop	4.362579e-04	5.198960e-04	2.698359e-04	4.941409e-04	5.199033e-04

Final neighborhood recommendation list for gym enterprises.

The final neighborhood recommendation list for new gym enterprises contains the values of every analysed feature so managers or stakeholders can choose the neighborhood with the profile which best suits their needs.



Recommended neighborhoods for gyms bar plot comparison against sample - Gym group of features

We can see how the selected neighborhoods for recommendation compare in terms of the presence of gyms against the whole sample. It's visible that all five neighborhoods have medium to high gym concentration per area. The neighborhoods which were not selected after the clustering based on the gym features were all at the extreme right of the area concentration plot. By excluding those neighborhoods, extreme area competition was avoided. Besides, we see that the concentration of gyms per 10000 inhabitants of the recommended neighborhoods are very reasonable, and range from medium to low values. Therefore, despite the fact that some recommended neighborhoods have a somewhat high area concentration of gyms, they all offer reasonable levels of market competition, meaning that new companies in these areas won't face issues such as lack of potential customers.

5. Discussion

The most sensitive part of this research is the count of gyms per neighborhood. This count was accomplished using the foursquare places API which offers places and geolocation data. The objective of this data collection step was not to get an accurate count of the number of gyms per neighborhood. The goal instead was to come up with a comparative measure of the quantity of gyms per neighborhood. Although these measures did not offer statistical significance (They could not represent the real number of gyms per neighborhood), they proved to be reliable for comparison across neighborhoods. They also proved to be valuable for insight and decision making. Mapping of the gyms across the neighborhoods was also performed which highlighted the quality of the final count result.

Another issue was that the gym search result came out empty for a few neighborhoods. The reason for these empty results was dubious. A possible explanation is that these neighborhoods do not have any gyms or have a small number of gyms. As this can not be made clear so easily, it was decided that these neighborhoods would get excluded from the analysis, in order to gain trust in the final result, that is, the list of recommended neighborhoods. The list of excluded neighborhoods can be found in the Jupyter Notebook containing the Foursquare API search located in the project repository.

The number of clusters in each segmentation was chosen based on the optimization processes applied and on trial and error, until the number of clusters was enough to distinguish the targeted groups.

The segmentation of the neighborhoods based on the income features was pretty straightforward since there was a clearly discernible group which maximized all three variables at once.

The same goes for the commerce segmentation. The neighborhoods selected after the commerce clusterization clearly represent a group of neighborhoods with more advanced infrastructure once it contains a much greater average number of schools, hospitals, credit agencies and food services than the other groups for example.

The segmentation of the neighborhoods by demographic factors represented a bigger challenge once the increase of one variable tends to cause a decrease in other variables. The way to solve this was to allow neighborhoods with lower values of population concentration to be selected once they maximized almost all other variables.

In summary, the selected neighborhoods present high demographic concentration in general and of the target population, which is the young and adult ages. Also, they present a high concentration of residents and the highest average income of homeowners and total income per Km². Besides, they also have advanced infrastructure such as many schools and hospitals, transport and diverse services. Finally, the remaining neighborhoods presented low to medium values of the number of gyms per Km² and per 10000 inhabitants, compared with the other neighborhoods.

6. Conclusion

The overall conclusion is that the analysis was successful at isolating the group of neighborhoods which offers the best or optimized conditions, based on the available data, to subside the investment of a new gym and other types of enterprises.

The clustering approach to segment and filter neighborhoods also proved to be a valuable resource. The challenge of comparing neighborhoods by many aspects (variables) at once was easily overcome since the clustering algorithms automatically detect and expose the aspects and patterns that make the neighborhoods similar or dissimilar, generating many insights into the city's aspects and conditions.

Remarkably, those few neighborhoods left for recommendation met every proposed criteria and presented optimal values from every single variable analysed. It's an understandable result once you consider that the recommended neighborhoods are among the most internationally popular spots of the Rio de Janeiro city and are usually recognized by.their great features.

The final result may not be the best for every case. The choice of the variables were arbitrary and intended for general purposes. Therefore, the neighborhoods for recommendation reflect the predetermined profile. Even so, for practical purposes the recommendation tables can be consulted to subside decision making in many levels and represent a solid match with the proposed profiles.

There is also a lot of room for further research. The city's areas presented great diversity and potential for many different projects. Many different neighborhoods profiles could be targeted and explored either through these or other methods. Besides, there's a lot of public data available specifically on neighborhoods from socioeconomics and demographics to energy consumption, crime and inequality.