ROS
Science applied to management

Renato de Oliveira Souza

Exploring new spots for Pilates Studio in São Paulo - Brazil

IBM Data Science Capstone Project

São Paulo – SP – Brazil

2021

INTRODUCTION

São Paulo is one of the largest cities in the world, with more than 12 million inhabitants (5% of the Brazilian population) and occupies the first place in the ranking of municipal GDP in Brazil, with 10% of the Brazilian GDP (IBGE, 2020).

Thus, São Paulo is considered the main Brazilian market for several consumer items, including services focused on well-being and health.

Among the services focused on health and well-being are fitness centers, spas, yoga studios and, in growing relevance, pilates studios.

Pilates is an exercise method developed by Joseph Pilates in the 1920s, being a recognized technique for treating and preventing postural problems. The method works on concentration, body centralization, precision of movement, breathing, motor control and fluidity of movement (Wikipedia, 2021).

A client intends to expand its operations within the city of São Paulo, opening a new studio and defining some parameters, based on the registration of clients of the existing units.

The criteria are:

- People aged 40 or over;
- Higher education, such as minimum schooling;
- Live up to 900 meters away from the studio;
- High family income.

To support this decision, we will use indicators at the neighborhood level, such as the human development indicator (HDI), to capture schooling and health concerns. The population density of people aged 40 and over to capture public concentration and monthly family income to capture neighborhoods with a high family income.

These indicators will build an index that added to a survey of the geographic distribution of gyms, spas and pilates studios in the city of São Paulo, carried out through the API of the Foursquare website, will indicate the best location to open the new studio, using algorithms of clustering like K-means.

DATA

The data to be used comes from public sources. The Human Development Index (HDI) is a summary measure of long-term progress in three basic dimensions of human development: income, education and health. website: List of Sao Paulo's Boroughs by Human Development Index.

Table 1 – Neighborhood HDI example

Neighborhood	HDI
Moema	0.961
Pinheiros	0.960
Perdizes	0.957
Jardim Paulista	0.957
Alto de Pinheiros	0.955

To calculate the demographic density of people over 40 years of age, the area of each neighborhood in the city of São Paulo will be collected on the website: Demographic data of the city of São Paulo and the population of the neighborhoods by age group on the website: Population by age group, which provides the data in CSV format.

Equation 1- Population density

$$\text{Density} = \frac{\text{Population over 40}}{\text{Area (Km}^2\text{)}}$$

Table 2 – Population density over 40 examples

Neighborhood	Area (Km2)	Population over 40	Density
Aricanduva	6.6	32288	0.42
Carrão	7.5	34431	0.40
Vila Formosa	7.4	36903	0.43
Butantã	12.5	21827	0.15
Morumbi	11.4	20644	0.16

The data on the average monthly family income will be taken from the annual report “Map of inequality”, available on the website:www.nossasaopaulo.org.br.

Table 3 – Monthly family income example

Neighborhood	Monthly family income(R$)
Alto de Pinheiros	9591.93
Perdizes	9348.58
Jardim Paulista	9327.12
Moema	9248.43
Santo Amaro	9159.73

These variables will be normalized by min max, for a scale of 0 to 1 and the average value between them calculated, to create the index of place of interest (HDDII), as a reference.

Equation 2- Population density

$$\text{HDDII} = \frac{\text{IDH} + \text{Density}^{(\text{min max})} + \text{Family Income}^{(\text{min max})}}{3}$$

Table 4 – Neighborhood HDDI example

Neighborhood	HDDII
Perdizes	0.90
Jardim Paulista	0.86
Moema	0.81
Alto de Pinheiros	0.79
Consolação	0.76

Geographical location of São Paulo neighborhoods will be gotten from Geopy Python library that converts geographical searches by name into coordinates – latitudes and longitudes. The neighborhood venues will be retrieved within 900 meters radius from each neighborhood geographical center using the Foursquare API. This radius was selected because it can select the perfect distance to move from home to new site by foot. After retrieving the venues, the API also returns the latitude and longitude of each venue, to complete the analysis for the best location.

Table 5 – Neighborhood coordinates example

Neighborhood	Latitude	Longitude
Alto de Pinheiros	-23.549.461	-46.712.293
Anhanguera	-23.432.908	-46.788.534
Aricanduva	-23.578.024	-46.511.454
Artur Alvim	-23.539.221	-46.485.265
Barra Funda	-23.525.462	-46.667.513

Table 6 – Neighborhood venues example

Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category
Moema	Arte de Viver	-23.596.409	-46.666.665	Yoga Studio
Moema	Amadí SPA	-23.599.936	-46.660.857	Spa
Moema	Smart Fit	-23.601.100	-46.665.530	Gym / Fitness Center
Moema	Race Bootcamp	-23.595.136	-46.671.162	Gym / Fitness Center
Moema	Needs Academia	-23.602.955	-46.657.411	Gym

Based on the venue category returned from foursquare API, we selected the categories for our analysis and group by attributes:

- Wellbeing = Pilates Studio, Spa and Yoga Studio
- Gym = College Gym, Gym, Gym / Fitness Center, Gym Pool and Gymnastics Gym

With this step, we conclude our dataset compilation:

Table 7 – Neighborhood features example

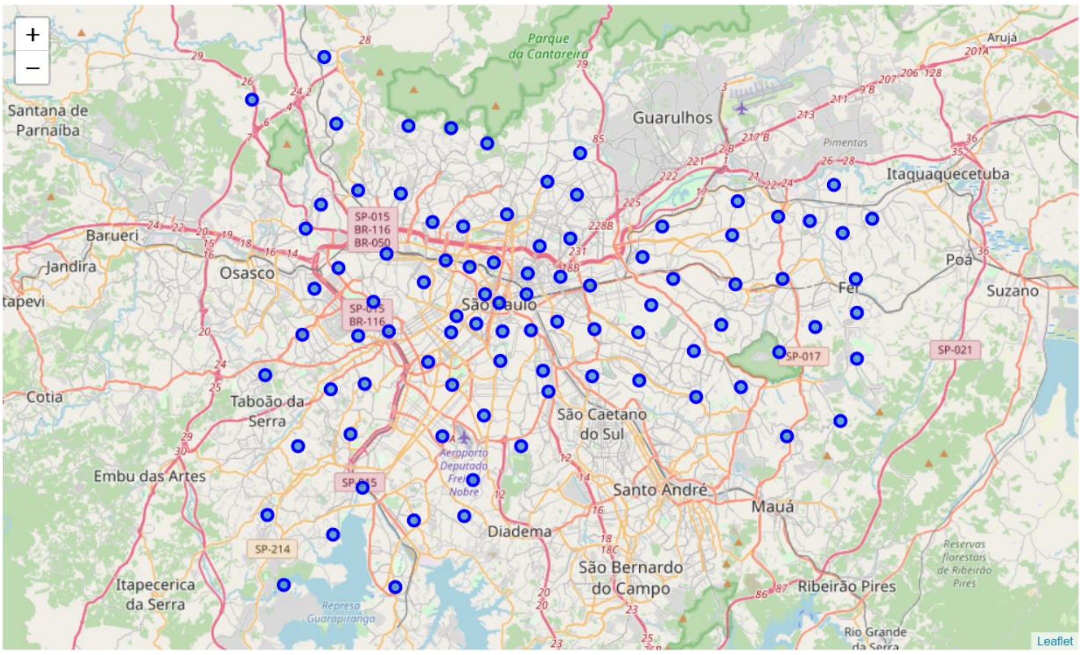
Neighborhood	Latitude	Longitude	HDDII	Wellbeing	Gyms
Alto de Pinheiros	-23.549461	-46.712293	0.788391	1	1
Anhanguera	-23.432908	-46.788534	0.232859	0	1
Aricanduva	-23.578024	-46.511454	0.371937	0	4
Artur Alvim	-23.539221	-46.485265	0.400723	0	3
Barra Funda	-23.525462	-46.667513	0.557293	0	3

METHODOLOGY

After defining the data, parameters, and indicators to be used in the analysis to choose the opening location of the pilates studio, we need to define the analysis methodology.

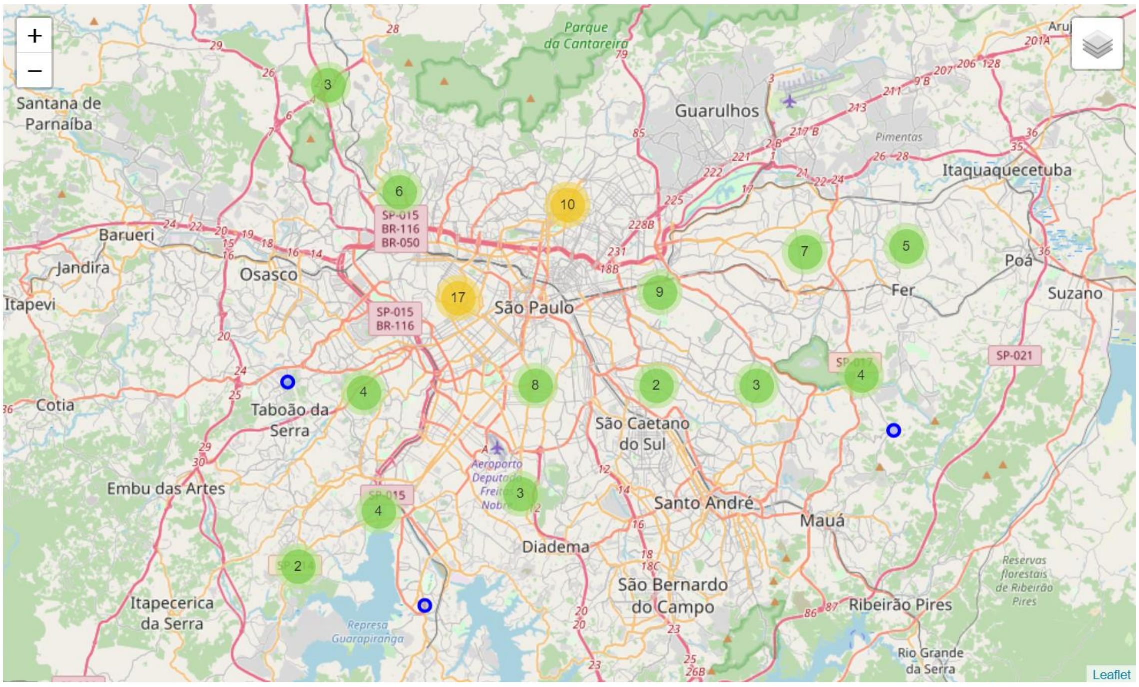
We started with the construction of maps to visualize the spatial distribution of the neighborhoods:

Map 1 – Neighborhood distribution map



We generate a clustered view of the neighborhoods to visualize the urban concentration of the city:

Map 2 – Neighborhood concentration map



As we have 95 neighborhoods, we need to use an algorithm that groups them into clusters with similar characteristics and that meets our business requirements. In this sense, we need to group the neighborhoods according to the HDDII index, which provides us with information on the demographic density of people over 40, human development level and family income level. Added to the number of health-related sites that were found in the Foursquare API.

For this task, we will use the K-Means algorithm of unsupervised learning, which evaluates and clusters the data according to their intragroup similarities, and as distinct as possible with the data of the other groups (Wikipedia, 2021).

As it works with the calculation of distances and averages from the points to the centroid (Center of each cluster), K-Means is one of the algorithms used for geospatial analysis.

The first challenge of this algorithm is to define the number of clusters that we will be working on. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use (Yellow brick, 2021). This means one should choose several clusters so that adding another cluster does not give much better modeling of the data. The graph was no so clear about the optimum K (number of clusters). This happened because the data is very dispersed (many zero values in the data set), however we can notice that the distortion reduction starts from 7 clusters, because the curve starts to flatten from that point, so we decide to use 7 clusters, which appear to represent the best option.

Figure 1 – Elbow method graph

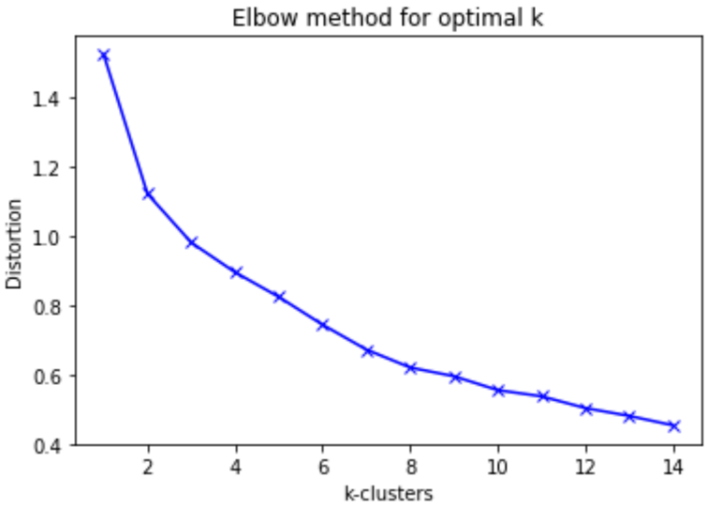


Table 8 – Healthy venues cluster

Cluster	Healthy venues
0	8
1	27
2	7
3	26
4	7
5	8
6	11
Total	**94**

After defining the cluster quantity, we will start analyzing their characteristics to evaluate how the algorithm worked. We have started summarizing by each cluster characteristics means.

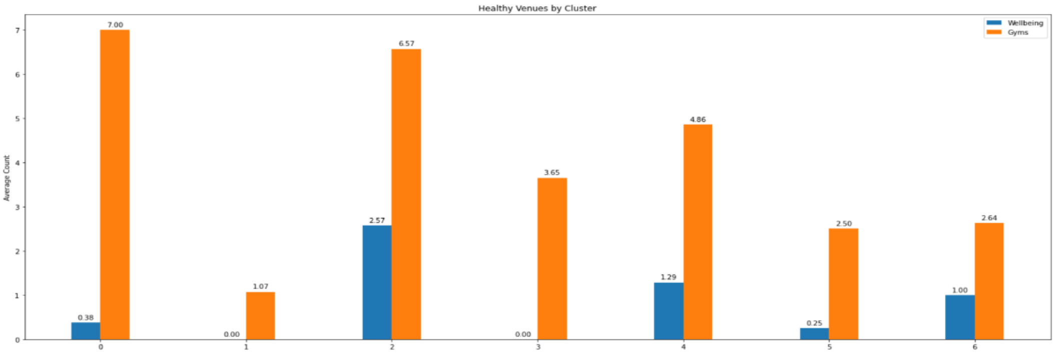
Table 9 – Neighborhood characteristics cluster

Cluster Labels	HDDII	Wellbeing	Gyms
0	0.44	0.38	7.00
1	0.35	0.00	1.07
2	0.65	2.57	6.57
3	0.38	0.00	3.65
4	0.77	1.29	4.86
5	0.68	0.25	2.50
6	0.46	1.00	2.64

In terms of HDDII, we have only one cluster above 0.75 and diversity in the number of wellness centers and gyms.

A good option to understand this behavior, is visualize this distribution on a bar chart.

Figure 2 – Healthy venues mean by cluster



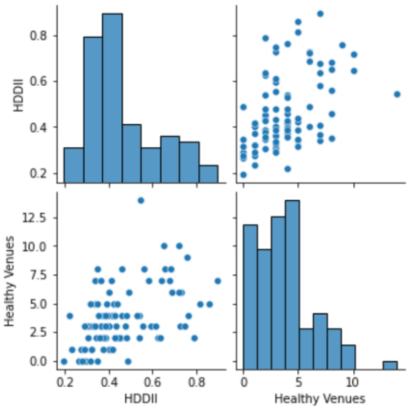
Now, it is clear that not all clusters have wellbeing sites, and it can help us to select our focus cluster.

It is also interesting to analyze whether there is a correlation between the clusters, the HDDII and the number of sites focused on wellbeing, to verify the quality of the variables chosen for the model. Then we built the matrix and the correlation graph.

Table 10 – Correlation matrix

Correlation	Cluster Labels	HDDII	Healthy Venues
Cluster Labels	1	0.399302	0.036562
HDDII	0.399302	1	0.476571
Healthy Venues	0.036562	0.476571	1

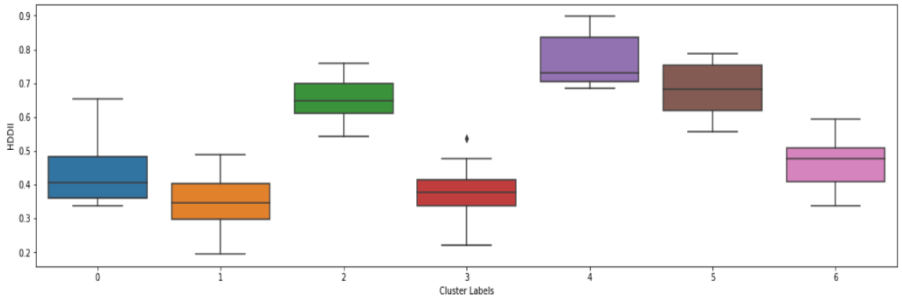
Figure 3 – Correlation graph



We can verify that the variables have a positive correlation, so we can conclude that the variables can help us in the process of choosing the best location for the opening of the new unit, because the higher the HDDII, the greater the number of spaces aimed at wellbeing, which indicates greater concern with health.

To draw better conclusions, we can analyze the HDDII distribution for each cluster, using a box plot graph.

Figure 4 – Box plot graph



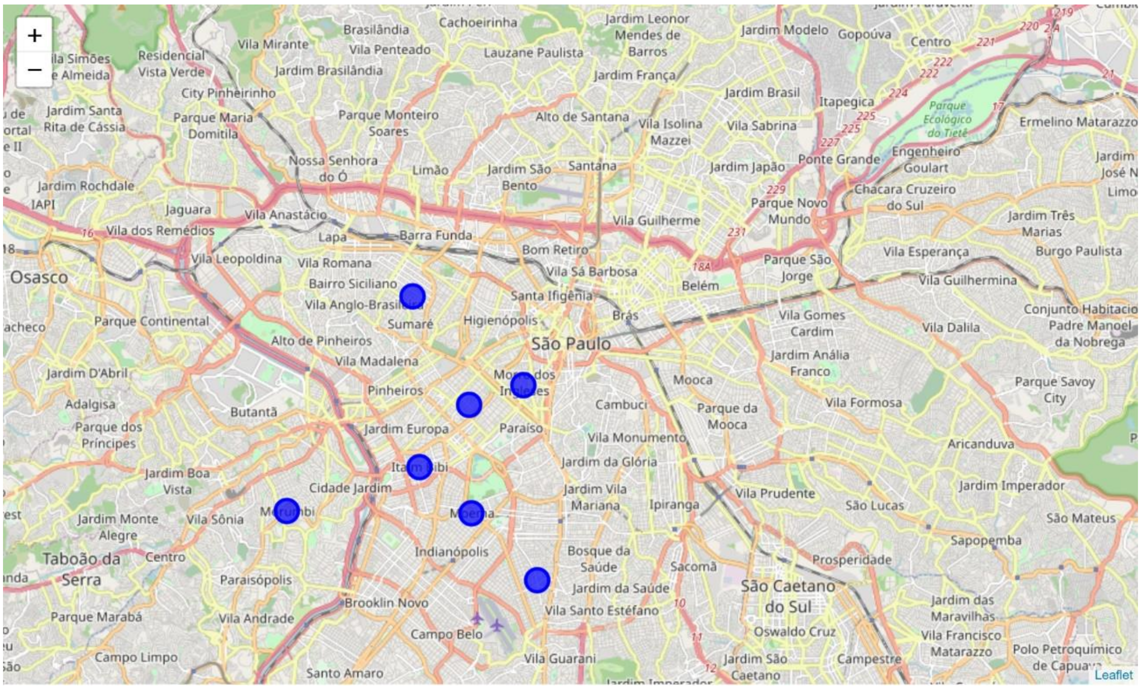
To meet the client's requirements, the HDDII index was built, so considering this criterion as the gold standard, we can define that cluster 4 is the chosen one for the continuity of the analysis, as it has the highest median and the highest value at the lower limit.

As cluster 4 was chosen, we will look for the best neighborhood for the pilates studio within that cluster. For that, we start with a list of the neighborhoods that make up the cluster and a map with the geographic locations.

Table 11 – Cluster 4 neighborhoods

Neighborhood	Latitude	Longitude	Cluster Labels	HDDII	Wellbeing	Gyms	Healthy Venues
Jardim Paulista	-23.567435	-46.663692	4	0.860440	2	3	5
Moema	-23.597085	-46.662888	4	0.813121	2	3	5
Bela Vista	-23.562210	-46.647766	4	0.730488	1	5	6
Itaim Bibi	-23.584381	-46.678444	4	0.722139	1	5	6
Morumbi	-23.596499	-46.717845	4	0.683930	1	7	8

Map 3 – Cluster 4 neighborhoods



RESULT

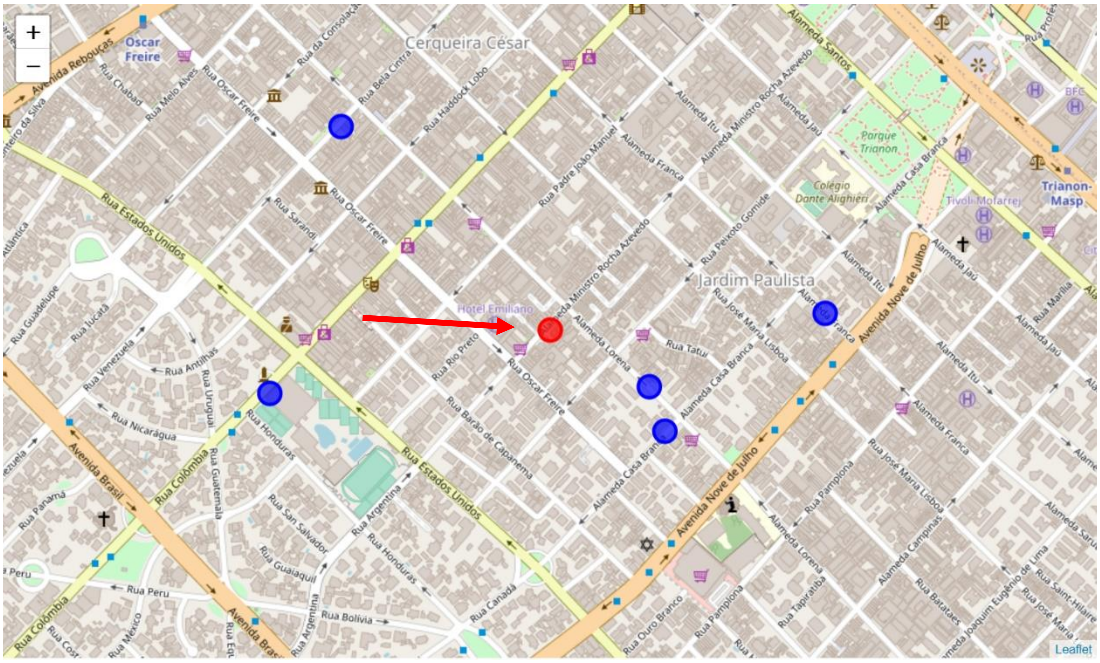
Finishing the analysis, we can conclude that the methodology used was correct and leads us to choose the Jardim Paulista neighborhood with the highest HDDII and the least amount of wellness spaces, such as the region to be indicated to the client as the first option.

For the recommendation, we can list the wellbeing spaces in the neighborhood and define the centroid between them as the optimum point. This definition is premised on the fact that, in general, people prefer to visit wellbeing spaces closer to their homes, so a region with many locations would mean a reference place for health-conscious residents.

Table 12 – Jardim Paulista neighborhood venues list

Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category
Jardim Paulista	Les Cinq Gym	-23.567124	-46.662261	Gym / Fitness Center
Jardim Paulista	Top Form Academia	-23.567911	-46.661955	Gym / Fitness Center
Jardim Paulista	My Yoga	-23.565829	-46.658869	Yoga Studio
Jardim Paulista	Academia CAP	-23.567254	-46.669574	Gym
Jardim Paulista	Spa L'Occitane	-23.562532	-46.668205	Spa

Map 4 – Recommended point



DISCUSSION/CONCLUSION

The construction of the HDDII index generated important information for future analysis and application in other studies, being an important reference for decision making in situations where we need to choose places according to human development, level of family income and demographic density. The index allows adjustments by age group, since for this study we work with people over 40 years old. In addition, the present study constructed a ranking of the neighborhoods, which can be explored by the client in the construction of performance strategies in different market niches, when evaluating the performance in other neighborhoods and clusters.

REFERENCES

- https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/29729-quase-metade-do-pib-do-pais-estava-concentrado-em-71-municipios-em-2018.
- https://pt.wikipedia.org/wiki/Pilates
- List of Sao Paulo's Boroughs by Human Development Index
- Demographic data of the city of São Paulo
- Population by age group
- www.nossasaopaulo.org.br
- https://developer.foursquare.com/
- https://en.wikipedia.org/wiki/K-means_clustering
- https://www.scikit-yb.org/en/latest/api/cluster/elbow.html