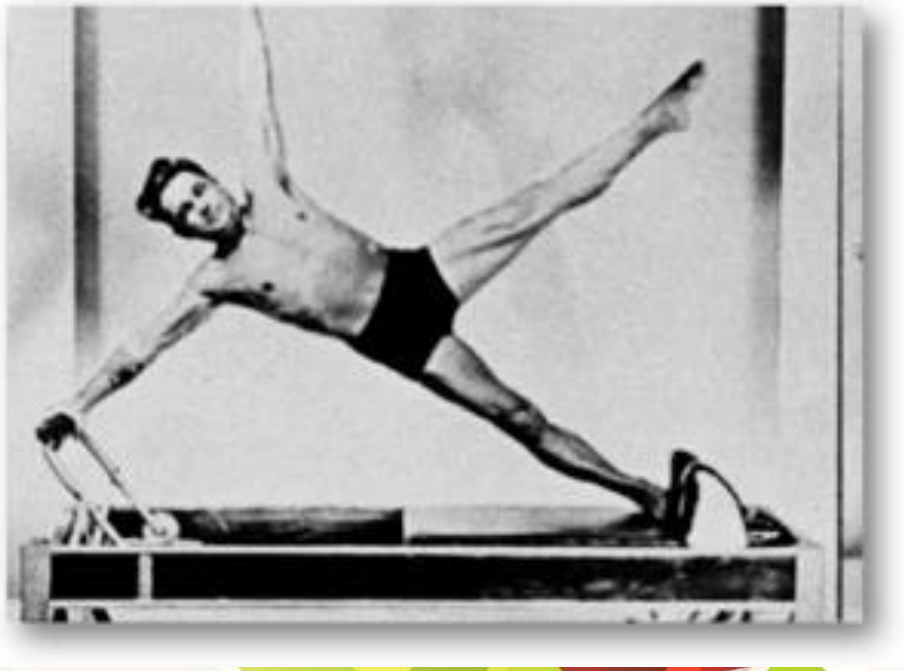


EXPLORING NEW SPOTS FOR PILATES STUDIO IN SÃO PAULO - BRAZIL

Author: Renato de Oliveira Souza

ROS
Science applied to management

Introduction

- The main objective of this work is finding a new possible spot to open a Pilates Studio within the city of São Paulo,
- Based on the customer records of the existing units, some parameters were previously defined to direct the search:
  - 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 analysis, we used 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 built 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.

Data source

Wikipedia	Provided information on socio-economic indicators, such as HDI and methodology.
GeoPy	Provided the geographic coordinates of the neighborhoods in the city of São Paulo.
Prefeitura da Cidade de São Paulo	Provided official geographic and demographic information for the city of São Paulo.
Agência IBGE Notícias	Provided socio-economic information about the city of São Paulo.
Rede Nossa São Paulo	Provided monthly family income information by neighborhood of the São Paulo city.
FSQ/developer	Provided geographic coordinates of the companies and type of activity, clustered by neighborhood and address.

Data

Raw data gathered example

Neighborhood	HDI	Area (Km2)	Population over 40	Density
Aricanduva	0.830	6.6	32288	0.42
Carrão	0.886	7.5	34431	0.40
Vila Formosa	0.884	7.4	36903	0.43
Butantã	0.928	12.5	21827	0.15
Morumbi	0.938	11.4	20644	0.16

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

Data transformation example

Formula transformation: $$HDDII = \frac{HDI + Density^{(min\ max)} + Family\ Income^{(min\ max)}}{3}$$

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

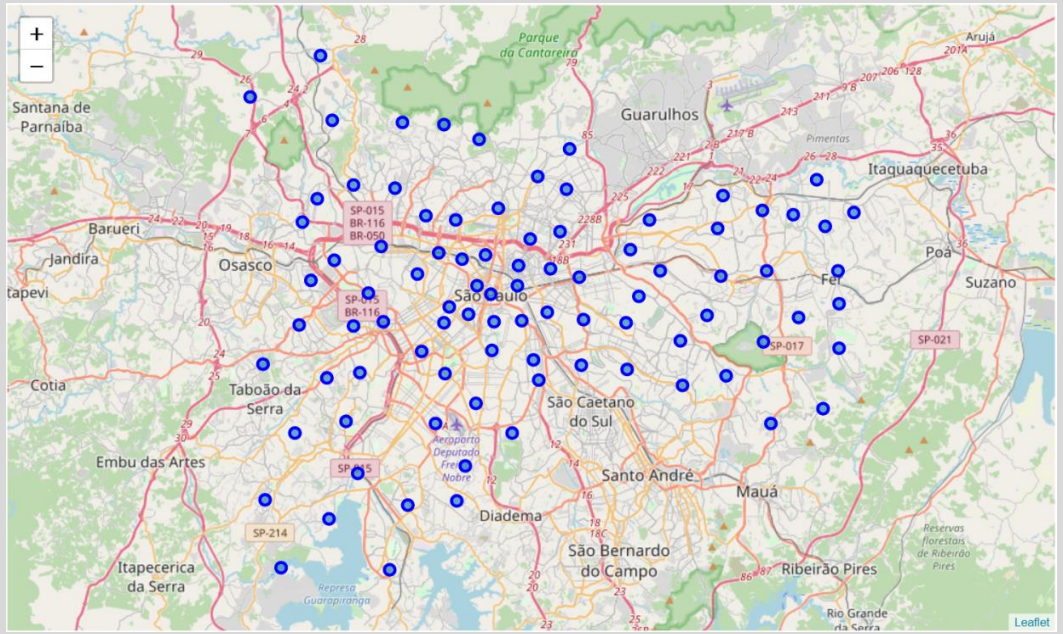
Geographic coordinates and features example

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

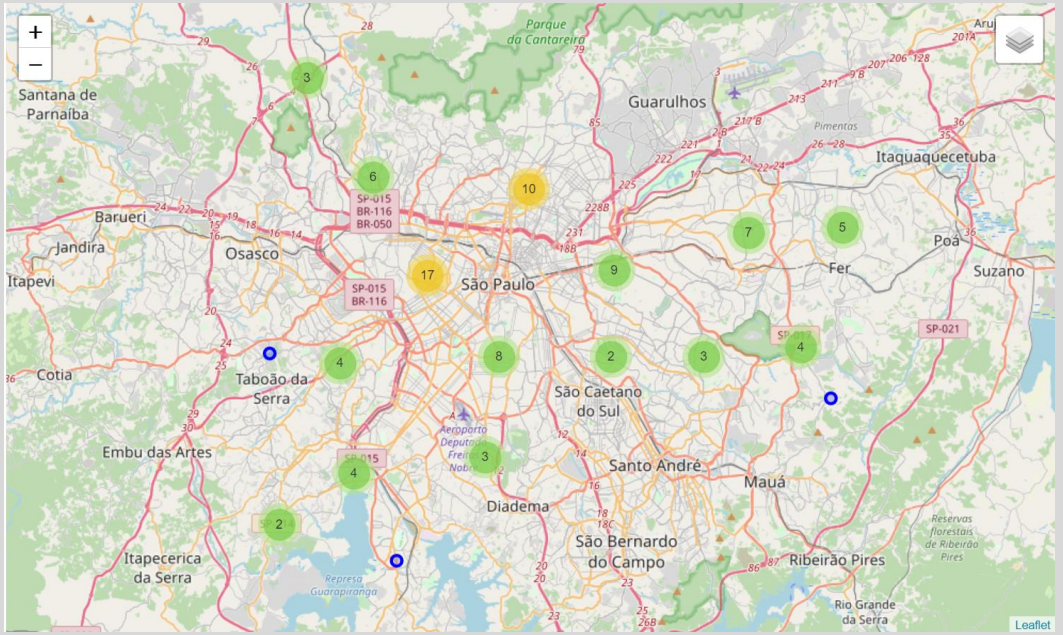
Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category
Moema	Arte de Viver	-23.596409	-46.666.665	Yoga Studio
Moema	Amadí SPA	-23.599936	-46.660.857	Spa
Moema	Smart Fit	-23.601100	-46.665.530	Gym / Fitness Center
Moema	Race Bootcamp	-23.595136	-46.671.162	Gym / Fitness Center
Moema	Needs Academia	-23.602955	-46.657.411	Gym

Data analysis - Maps

Neighborhood distribution



Neighborhood concentration

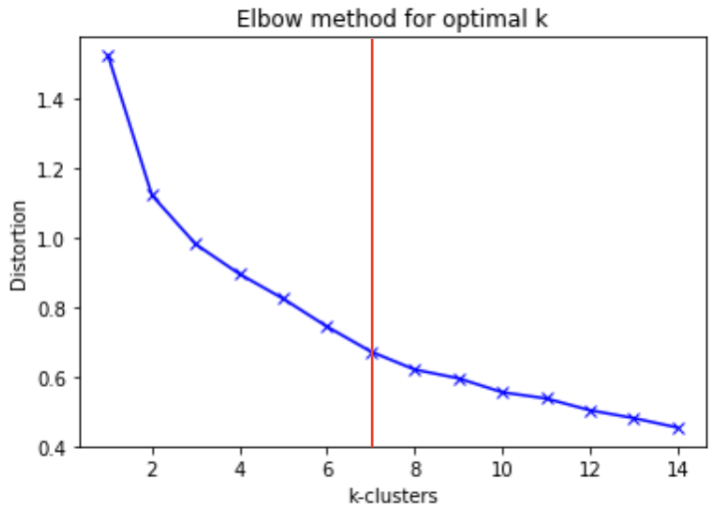


18/03/2021

Methodology – Machine learning algorithm

- 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.
- For this task, we've used 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.
- 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.
- We can notice that the distortion reduction starts from k-clusters 7, because the curve starts to flatten from that point, so we decide to use 7 clusters, which appear to represent the best option.

Elbow method graph



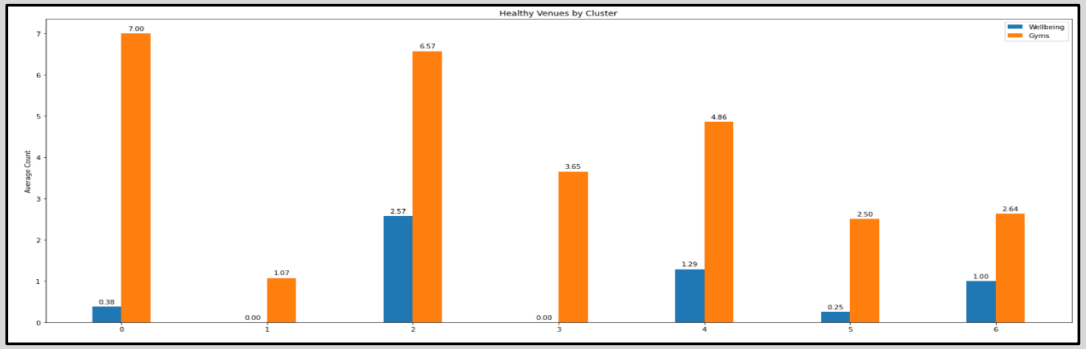
Methodology – Cluster analysis

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

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

Healthy venues quantity mean by cluster



- In terms of HDDII, we have only one cluster above 0.75. On the other hand, there is a great diversity in the number of wellbeing centers and gyms per cluster.
- It's clear on this distribution bar chart that not all clusters have wellbeing centers, and it can help us to select our focus cluster

Methodology – Model quality analysis

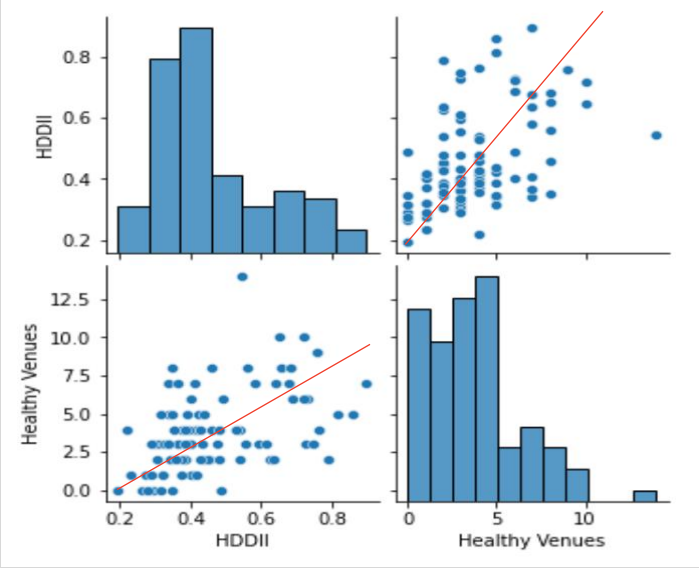
- To assess the quality of the model, it is interesting to analyze the correlation between the clusters and their characteristics, such as HDDII and the number of wellness spaces.

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

- 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.

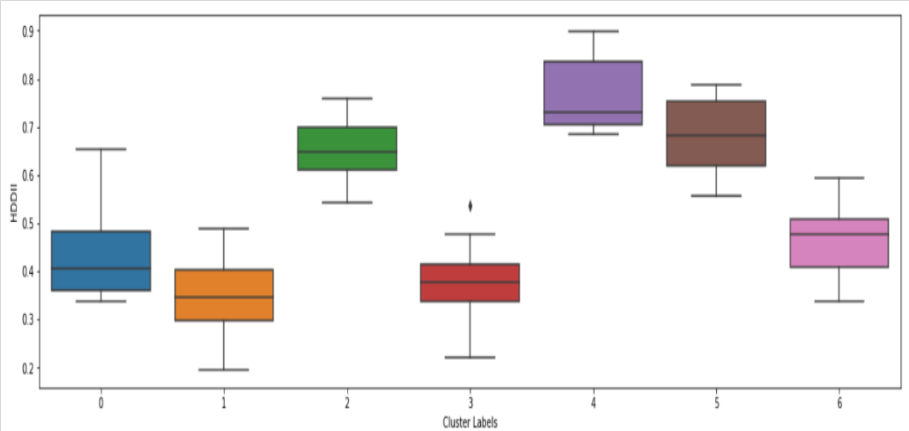
Correlation graph



Methodology – Cluster definition

- To choose our cluster, we can analyze the HDDII distribution for each cluster, using a box plot graph.

Box plot graph



- Based on the HDDII index, 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.

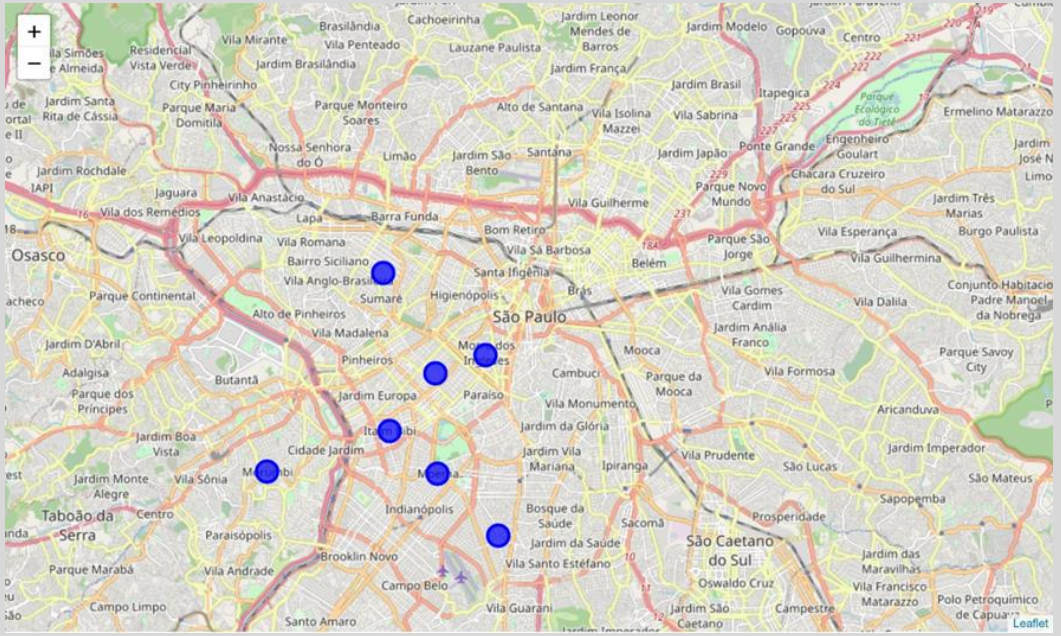
Methodology – Cluster 4 analysis

- 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.

Cluster 4 neighborhoods features

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

Cluster 4 neighborhoods map position



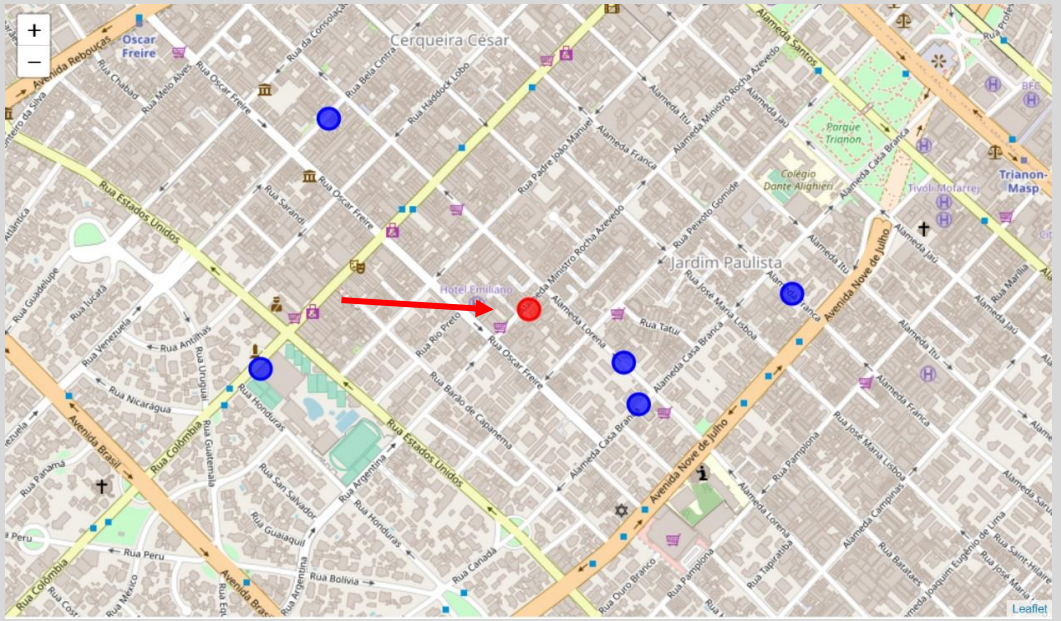
Result

- 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 as the first option.

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

Recommended point



18/03/2021

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.