Utilizing Machine Learning Unsupervised Learning K-Mean Algorithm in Analyzing Sao Paulo's Neighborhoods Potential for a New Health Market Venue

Juliano Garcia

June 19th, 2019

Executive Summary

The present analysis aims to give entrepreneurs and interested parties a recommendation on the best neighborhoods in Sao Paulo, Brazil to establish a new Health Foods Market. Based on data scraped from the web and geospatial services like Foursquare and analysis supported by an unsupervised machine-learning algorithm for clustering the neighborhoods data by similarity, the K-Means model. Resulting in an overview of the attributes of each neighborhood, such as the correlation between the area development (HDI) and the number of health venues already established on each borough. The study could conclude that the higher HDI, the more health conscious the population is, therefore, increasing the possibility of success for a new venue. The analysis finally comes up with a top five list of neighborhoods in Sao Paulo and a recommend optimal spot on the top-1 analyzed neighborhood, Itaim Bibi.

Table of Contents

1. Introduction

1.1. Background

The city of Sao Paulo is considered an alpha global city by the Globalization and World Cities Research Network (GaWC) and the most populous city in Brazil and in the southern hemisphere. The city is the capital of the surrounding state of Sao Paulo, the most populous and wealthiest state in Brazil. It currently has an estimated total population of 12 million people and Human Development Index (HDI) of 0.829 and a GDP per capita of $39,624.

Natural and organic food in Brazil was not a remarkable market until the year of 2007, however that that period this sector is now becoming a trend in the country, being led by the city of Sao Paulo. The organic food market in Brazil experienced a period of remarkable growth between 2007 and 2013. During this period, the revenue of these products in Brazil went from BRL 118 million to BRL 700 million, representing over 0,5% of the total revenue of the Brazilian food industry - including food exports. The main contributing factor is the increase in the country's GDP in the 2000s, which boosted the middle-class economy enabling the consumption of more expensive products with a high cost of production, which is the case for natural and healthy products. However, that increased organic market development is still in its initial phase, by 2014 the whole city of Sao Paulo had only 20 organic farmer's market.

1.2. Problem Case

Social and geospatial data available across multiple sources might bring insights for an entrepreneur that is seeking to establish a new healthy foods market in the city of Sao Paulo. This study aims to leverage the available data online to determine what would be the best spots to establish a new health market venue on the metropolitan region of Sao Paulo by conducting a neighborhood focused analysis, leveraging on social data such as the Human Development Index (HDI) and current venue distribution for each analyzed neighborhood.

1.3. Benefits

The present study aims to be a guide for entrepreneurs how seek to understand the geographical distribution of health venues and the natural and organic market in the city of Sao Paulo, Brazil.

2. Data

2.1. Data Sources

The data used for the present analysis comes from four main different repositories and web servers. They are as follows.

2.1.1. Wikipedia

That source will provide the information required for the neighborhoods database as well as the main social indicator used for this study, the Human Development Index. All data used for this analysis is pulled from one specific page, *List of Sao Paulo's Boroughs by Human Development Index*, which is only available in Portuguese - Brazil's official language.

Index	Zone	Position	Neighborhood	HDI
0	Região Central	1	Consolação	0,950
1	Região Central	2	Bela Vista	0,940
2	Região Central	3	Liberdade	0,936
3	Região Central	4	Santa Cecília	0,930
4	Região Central	5	Cambuci	0,903
5	Região Central	6	República	0,858
6	Região Central	7	Sé	0,854
7	Região Central	8	Bom Retiro	0,847
8	Leste 1	1	Penha	0,865
9	Leste 1	2	Vila Matilde	0,86

Table 1: Neighborhoods scrapped form Wikipedia webpage

2.1.2. Geopy Library

The Geopy is a Python based library that contains pre-defined functions that converts geographical searches by name into coordinates – latitudes and longitudes. This library was used in order to come by the geographical coordinates for the city of Sao Paulo, as well as each neighborhood pulled from the Wikipedia base.

Index	Zone	Position	Neighborhood	HDI	Latitude	Longitude
55	West	12	Jaguara	0.863	-23.507446	-46.755315
22	East 2	2	Vila Jacuí	0.779	-23.500294	-46.458717
3	Central	4	Santa Cecília	0.930	-23.529660	-46.651892
42	Southeast	14	Sacomã	0.839	-23.631090	-46.595618
14	East 1	7	São Mateus	0.804	-23.599843	-46.477605

Table 2: Neighborhoods table filled with geographical coordinates from Geopy

2.1.3. NYU Geospatial Data Repository

The New York University hosts a spatial data repository open for developer, from there we can get *json* files that contain geographical coordinates for the major cities or any border tier. This study simply grabs the file containing the borders of the city of Sao Paulo.

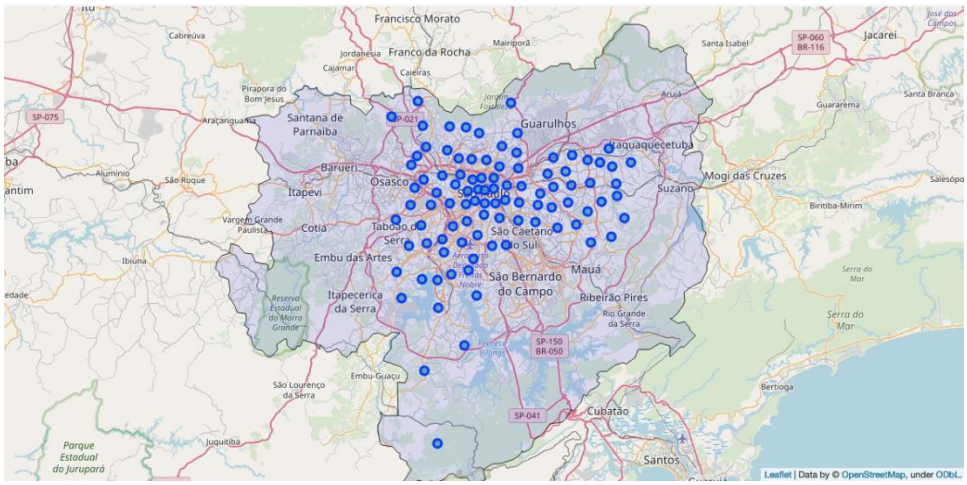

Figure 1: City of Sao Paulo with delimited borders and neighborhoods as blue markers

2.1.4. Foursquare API

By connecting the Python source code directly to the query URL of the API, this analysis pulls data regarding venues for each of the listed neighborhoods, in order to perform clustering analysis based on the similarities of each neighborhood. For the present query, it was use a radius of 500 meters and a limit of 100 venues per neighborhood. On the final set, we got 281 different venue categories.

Index	Neighborhood	Neigh. Latitude	Neigh. Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Consolação	-23.54808	-46.660029	Carlota	-23.546694	-46.660780	Restaurant
1	Consolação	-23.54808	-46.660029	Bráz Pizzaria	-23.547989	-46.657645	Pizza Place
2	Consolação	-23.54808	-46.660029	Ici Bistrô	-23.549389	-46.658190	French Restaurant
3	Consolação	-23.54808	-46.660029	Loja Mod	-23.546138	-46.658954	Furniture / Home Store
4	Consolação	-23.54808	-46.660029	Corrientes	-23.546014	-46.660163	Argentinian Restaurant

Table 3: Venues queried from Foursquare API by neighborhood

2.2. Data Cleaning & Transforming

The data used for this analysis were scraped from the multiple sources, as presented on the previous section of this report. After each data scrape step the resulted DataFrames were merged and converted into one that consolidates all the neighborhood attributes and each venue category as a separated attribute for each neighborhood (allocated on a separated columns), containing numeric values that accounts for the sum of venues in that specific category.

	Zone	Neighborhood	Latitude	Longitude	HDI	Acai House	...	Yoga Studio
0	West	Alto de Pinheiros	-23.549550	-46.712155	0.955	0	...	0
1	Northwest	Anhangüera	-23.432908	-46.788534	0.774	0	...	0
2	Southeast	Aricanduva	-23.572630	-46.518321	0.885	0	...	0
3	East 1	Artur Alvim	-23.539221	-46.485265	0.833	0	...	0
4	West	Barra Funda	-23.522709	-46.672928	0.917	0	...	0

Table 4: Preprocessed DataFrame

Finally based on the preprocessed DataFrame as shown in Table 4, we performed a feature selection in order to assess each neighborhood's health market score. For that, we selected only the columns that account for venues related to a "healthy lifestyle" as well as the HDI, coordinates and region of each neighborhood.

Kept Features

Attribute	**Venue Categories**	**Reason**
Health Markets	Farmers Market	Market that usually sells fresh organic food
	Health Food Store	Venue for health foods
	Organic Grocery	Sells organic products
	Fruit & Vegetable Store	Related to healthy lifestyle
Traditional Markets	Market	Access market competition on the neighborhood
	Supermarket	Access big market chains that might be a competition
	Grocery Store	Access market competition on the neighborhood
Restaurants	Gluten-free Restaurant	Related to "healthy" lifestyle
	Vegetarian / Vegan Restaurant	
Gyms	College Gym	
	Gym	
	Gym / Fitness Center	
	Gym Pool	
	Gymnastics Gym	

Table 5: Kept features with explained reasons

All the other venue categories were dropped for the analysis since we are seeking to aggregate the neighborhoods based on their relation to a healthy lifestyle, which would increase the acceptability of the new venue. In addition, the position attribute of each neighborhood on their zone based on HDI were dropped too since it would not add any extra useful information for the present analysis.

2.3. Feature Engineering

In order to classify the neighborhoods based on the desired and undesired attributes a new index was calculated, which were entitled the "Healthy-Human Development Index" (HH Index). That index is a calculation based on the desired attributes such as high HDI, Gyms and Healthy Restaurants and undesired attributes such as Healthy Markets and Traditional Markets, which might represent competition for the new venue.

*HH Index = (HDI * 10 + #Gyms + #Healthy Restaurants) / (#Healthy Markets * 5 + #Traditional Markets)*

3. Methodology

For this specific case problem, the sequence of the Data Science methodology was deliberately changed, the machine-learning algorithm was executed before the exploratory data analysis since it represented a new source of extra information for the analysis.

3.1. Machine Learning Algorithm

For the present analysis, the chosen algorithm was an unsupervised learning model for data clustering. The reason for the present choice is due to the fact that the analysis aims to identify similar neighborhoods that could be potential site for the new venue, based on HDI and the presence of health venues on those neighborhoods. Therefore, the K-means algorithm was the choice to identify such clusters.

The K-Means algorithms, as stated, is a model of unsupervised learning where it tries to minimize the Euclidean distance between each observation and their cluster's centroid, while maximizes the distances between them and observations from other clusters, all that based on the number of clusters that the user specified.

That presents a challenge for the current analysis on how many clusters should be used to gather the maximum information without overfitting the model. For that matter, we performed the *elbow* method. This method consists on running the model many times with a different number of clusters (as known as *k*) and storing the accumulated distortion – measured by the Euclidean distance between the points and their cluster's centroid – so we check when the information gain starts to be minimal, that would define the optimal k.

Finally, the model was created using the following attributes: 1) HDI, 2) Number of healthy markets, 3) Number of traditional markets, 4) Number of gyms and 5) Number of gyms. A standard scale of said attributes were performed before the model execution to ensure no bias taken towards a specific attribute.

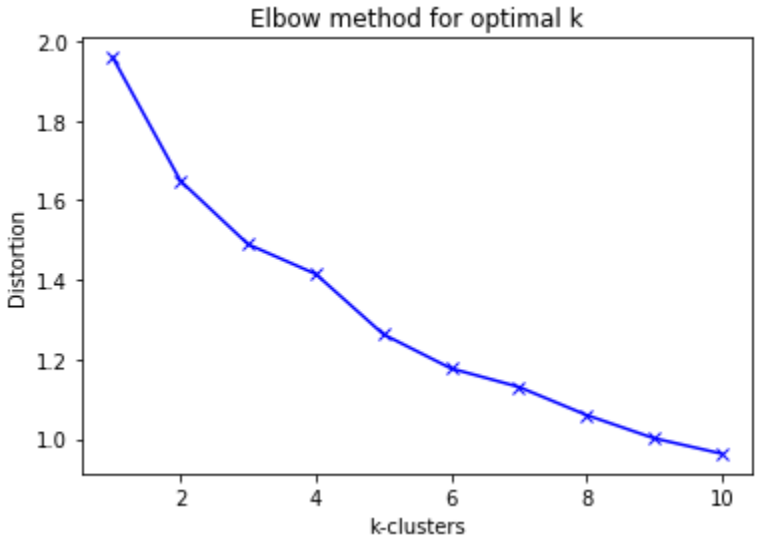

Figure 2: The elbow method, distortion loss by number of clusters in the model

When performing the *elbow* method to find the optimal *k* number of clusters, no clear point could be made from the curve – Figure 2. That happened because the data is too sparse (too many zeros in the dataset), however we can notice that the distortion loss after 8 clusters started to be a little flatter than before, therefore we continued the analysis with 8 clusters in total.

Cluster Label	Number of Neighborhoods in the Cluster
0	16
1	33
2	9
3	16
4	3
5	1
6	1
7	14

Table 6: Counts of neighborhoods in each cluster

3.2. Exploratory Data Analysis

Now based on the clusters output from the machine learning algorithm an analysis on the clusters' characteristics could be conducted, first an outlook on each cluster's means for the selected attributes were executed.

Cluster Labels	HDI	Healthy Markets	Traditional Markets	Gyms	Restaurants
5	0.960000	3.000000	2.000000	5.000000	2.0000
2	0.932222	0.444444	1.333333	3.555556	1.0000
4	0.928333	3.000000	3.333333	5.333333	0.0000
7	0.912143	0.142857	0.357143	1.857143	0.0000
3	0.886937	1.187500	0.812500	0.750000	0.0625
6	0.858000	1.000000	0.000000	0.000000	4.0000
0	0.832250	0.250000	2.437500	1.375000	0.0000
1	0.796758	0.090909	0.393939	0.484848	0.0000

Table 7: Clusters' attribute's mean

In terms of Healthy Markets, we have two clusters that contains a high average – 5 and 4 – and some clusters with a reasonably low average – 1, 0 and 7. Now visualizing the venues attributes on a bar chart we have as shown in Figure 3.

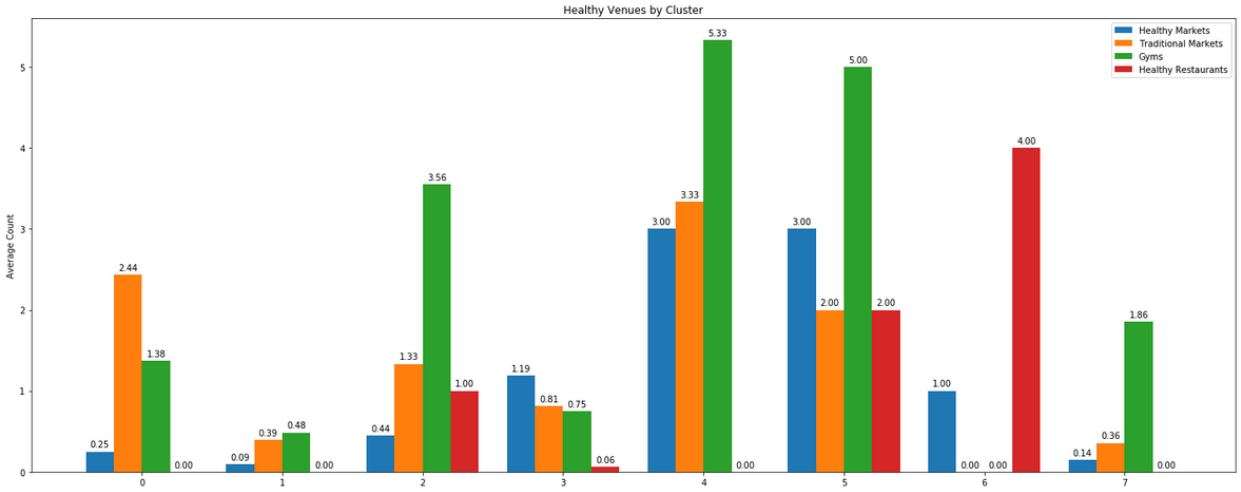

Figure 3: Clusters' Venue attributes

Some observations can arise from the chart in Figure 3:

- Clusters 4 and 5 are very healthy conscious, not only they have the highest average of Healthy Markets, but also for gyms, and a quite high traditional markets count;

- Clusters 0, 2 and 7 although got a high count of gyms they do not seem to have many health foods markets. Cluster 0 however have a very high count of traditional markets;
- Cluster 3 have a higher count of health markets than any other related venues, it seems to be already saturated;
- Cluster 6 seems to concentrate the healthy gourmet neighborhoods with a high count of healthy restaurants and markets and no count of gyms or traditional stores;
- Cluster 1 seems to lack any health-related venues.

In order to drawn better conclusions, we should analyze the observations from the Figure 3 with Figure 4 where we can see HDI distributions for each cluster.

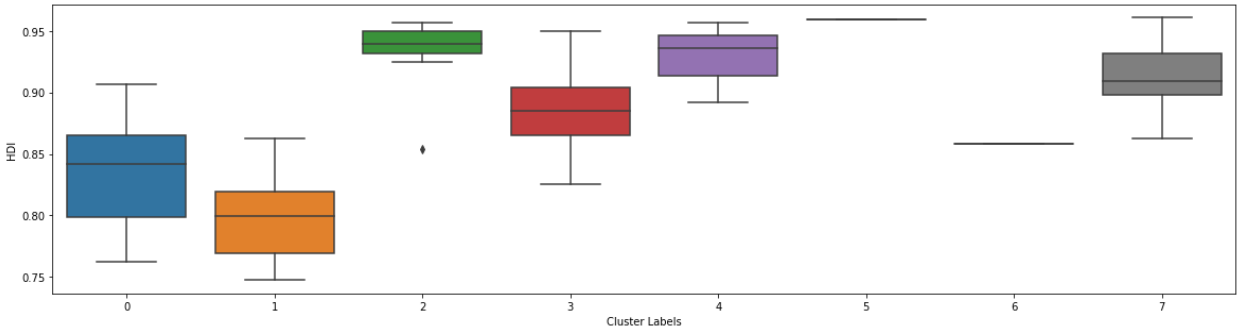

Figure 4: Boxplot distributions of HDI for each cluster

The data seems to show a correlation between the number of healthy venues and the HDI score of the neighborhoods. For instance, the clusters with the highest counts of healthy venues – 2, 4 and 5 – are also the clusters with the highest HDI scores, Table 8 shows correlation matrix and Figure 5 shows a scatterplot of HDI and Healthy Venues count, based on the observation of those two visual there is a conclusion that they are, in fact, correlated.

	HDI	Healthy Venues
HDI	1.00000	0.59436
Healthy Venues	0.59436	1.00000

Table 8: Correlation matrix of HDI and number of healthy venues

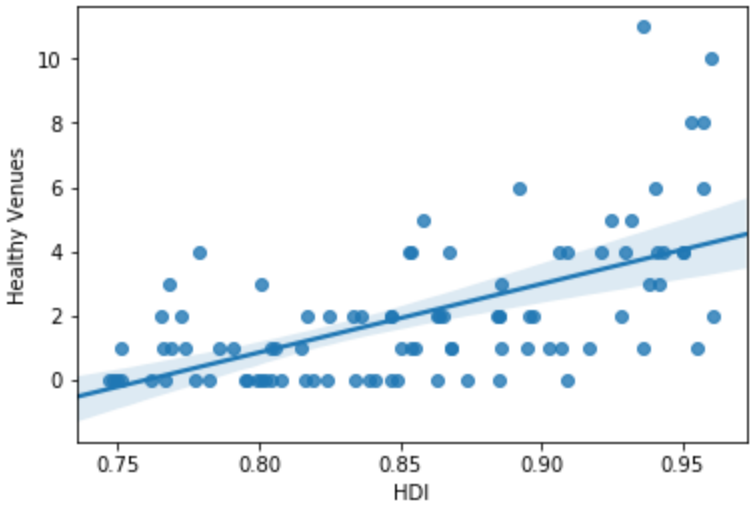

Figure 5: Scatter plot of neighborhoods' HDI and number of healthy venues

That observation proves a concept common on region, which is that healthy food and lifestyle is usually more expensive and, therefore, only people with a higher monetary power can afford. Another concept behind that 0.6 correlation is the fact that with a higher education and development comes a higher sense on the need to be healthy, therefore, it is safe to say that the higher the HDI of a given neighborhood the higher is the likelihood of the population embracing a new Health Foods Market. So, the potential neighborhoods should have a high HDI.

Based on that two clusters stand out from the rest with potential neighborhoods for the new venue, they the 2 and 7 clusters, the reasons are shown below and the neighborhood spread across the map on Figure 6.

- Cluster 2 concentrates a lot of high HDI neighborhoods (ranging around 0.95), has a lot of gyms which might means that the population on those neighborhoods are very healthy conscious, however the amount of Healthy Markets there seems very low scoring only 0.4 on average, which gives a potential growth possibility of the health sector;
- Cluster 7, although it's got a wider HDI distribution – we might be able to check some of the neighborhoods in the 3rd or 4rth quartile to stay in the high HDI range – this cluster contains neighborhoods with a high amount of gyms and no Healthy venues, Markets or even Restaurants, making them new territory for the health sector.

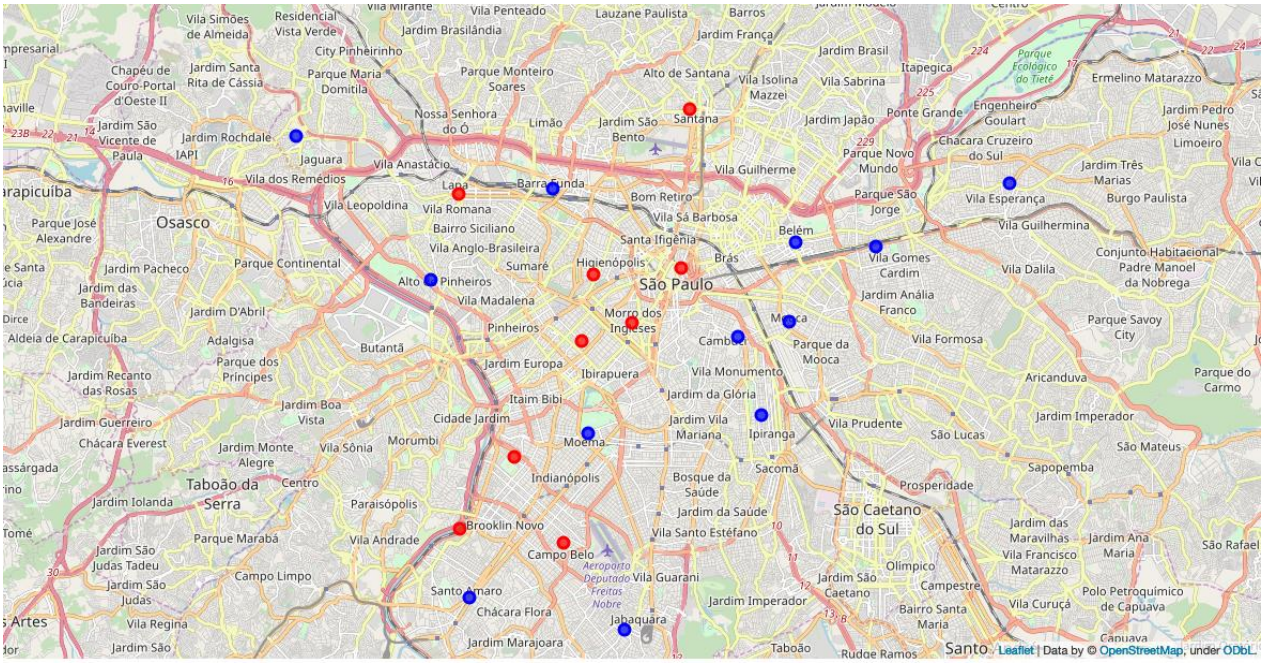

Figure 6: Clusters 2 (red) and 7 (blue)

4. Results

Finally, based on the described methodology and analysis process for the present case, the top neighborhoods for the new venue lays in the clusters 2 and 7, utilizing the engineered feature, the HH Index, to gather the top 5 neighborhoods we end up with the list shown in Table 9.

	Neighborhood	Zone	Cluster Labels	HDI	Healthy Venues	HH Index
9	Itaim Bibi	West	2	0.953	8	175.3
13	Lapa	West	2	0.941	4	134.1
15	Moema	Center-South	7	0.961	2	116.1
0	Alto de Pinheiros	West	7	0.955	1	105.5
22	Tatuapé	Southeast	7	0.936	1	103.6

Table 9: Top 5 HH Index neighborhoods in clusters 2 and 7

And as a recommendation spot, by taking the centroid between the gyms in the best HH Index scoring neighborhood – Itaim Bibi – the optimal spot would be as illustrated on Figure 7. That assumption takes as premise the fact that the population usually prefer to go to the gym that is closer to their homes, therefore a region with many gyms would mean a great number of health conscious residents.

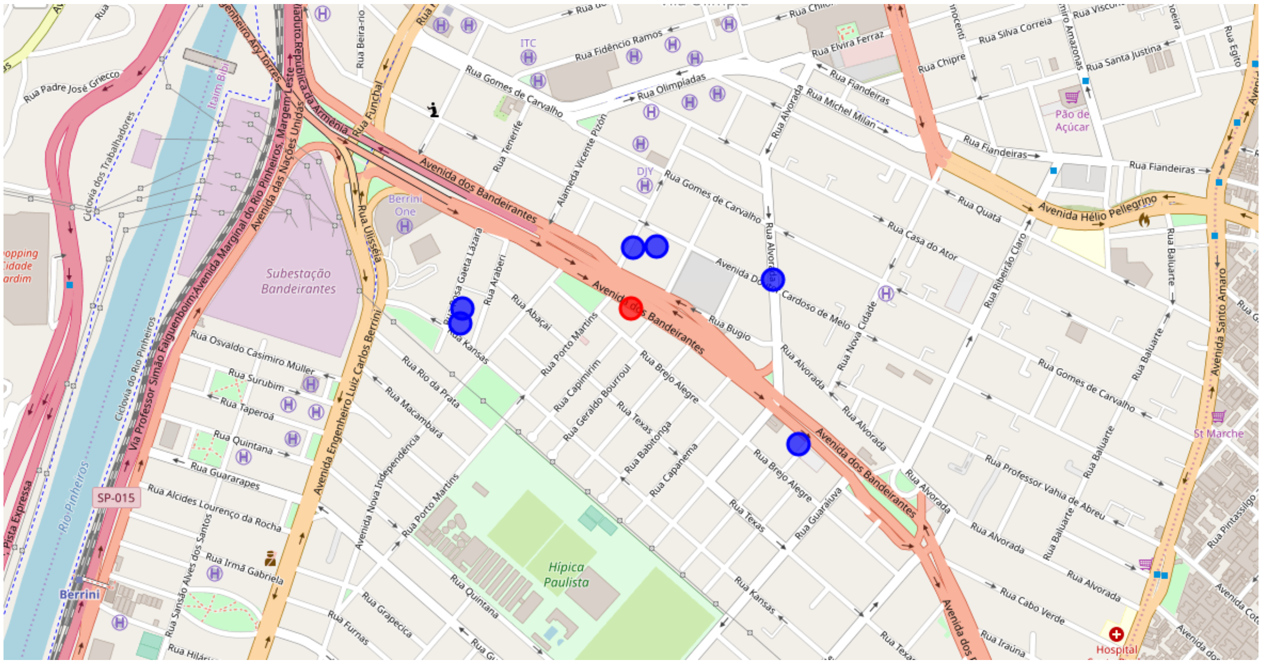

Figure 7: Optimal spot based on best HH neighborhood's gyms

5. Discussion

Based on the observed data and conducted analysis' insights we might infer that – due to the correlation between the Human Development index and the amount of healthy venues in the neighborhood – for an optimal spot to establish a new Health Foods Market, we might look for a high HDI. In addition, due to the beginning stage that the health sector is in Brazil, especially in Sao Paulo, we can look for the lack of those kind of venues in the selected neighborhoods to assess how saturated the market is for the given area.

Finally, the present analysis can recommend that the entrepreneurs interested in establishing a new Health Food venue can focus their attentions to neighborhoods with high HDI and low number of already established health markets, such as the top five recommended by the model 1) Itaim Bibi, 2) Lapa, 3) Moema, 4) Alto de Pinheiros and 5) Tatuape.

6. Conclusion

In summary, the presented analysis could achieve the desired information, which was to come up with the best spots for establishing a new Health Food Market. The recommendation was to focus on high HDI and low already established Health Market count, especially the top five listed in the previous sections.

7. References

- The Brazil Business <https://thebrazilbusiness.com/article/organic-food-market-in-brazil>;
- NPR.org <https://www.npr.org/sections/thesalt/2014/01/05/259455757/in-sao-paulo-organic-markets-are-beginning-to-take-off>;
- NYU Spatial Data Repository < https://geo.nyu.edu/catalog/>;
- Wikipedia: List of Sao Paulo Boroughs <https://pt.wikipedia.org/wiki/Lista_dos_distritos_de_S%C3%A3o_Paulo_por_%C3%8Dndice_de_Desenvolvimento_Humano>
- Foursquare