NEW HEALTH FOOD MARKET @ SAO PAULO/BR

UTILIZING MACHINE LEARNING UNSUPERVISED K-MEANS ALGORITHM IN ANALYZING SÃO PAULO’S NEIGHBORHHODS’ POTENTIAL FOR A NEW HEALTH VENUE

AUTHOR: JULIANO GARCIA <JULIANO.GARCIA@PROTONMAIL.COM>

AGENDA

Problem Case

Data Gathering

Selected Venues

Machine Learning Algorithm

Data Mining

Results

Conclusion & Recommendations

PROBLEM CASE

- The city of Sao Paulo is the most populous city in Brazil and in the Southern hemisphere, its currently got an estimated 12 million of total population and a Human Development Index (HDI) of 0.829;
- The health food Market in Brazil is at a great expansion but is still in its early stages, by 2014 São Paulo only had 20 farmer's markets;
- The present analysis aims to give insights on potential neighborhoods for the establishment of a new Health Foods Market in the city of Sao Paulo;
- The data analyzed consists of information scrapped from multiple sources and is based on social and geospatial indicators.

City of Sao Paulo / Brazil

DATA GATHERING

Wikipedia

Provided the information required for the neighborhoods database as well as the main social indicator used for this study, the Human Development Index.

GeoPy

Provided geographical coordinates for the city of Sao Paulo, as well as each neighborhood pulled from the Wikipedia base.

NYU Spatial Repository

Provided geojson file containing details on the administrative borders of the city of Sao Paulo.

Foursquare API

Provided data regarding venues for each of the listed neighborhoods. The query used a radius of 500 meters and a limit of 100 venues per neighborhood.

SELECTED VENUES

Health Markets

- Farmer’s Market
- Health Food Store
- Organic Grocery
- Fruit & Vegetable Store

Traditional Markets

- Market
- Supermarket
- Grocery Store

Health Restaurants

- Gluten-free Restaurant
- Vegetarian / Vegan Restaurant

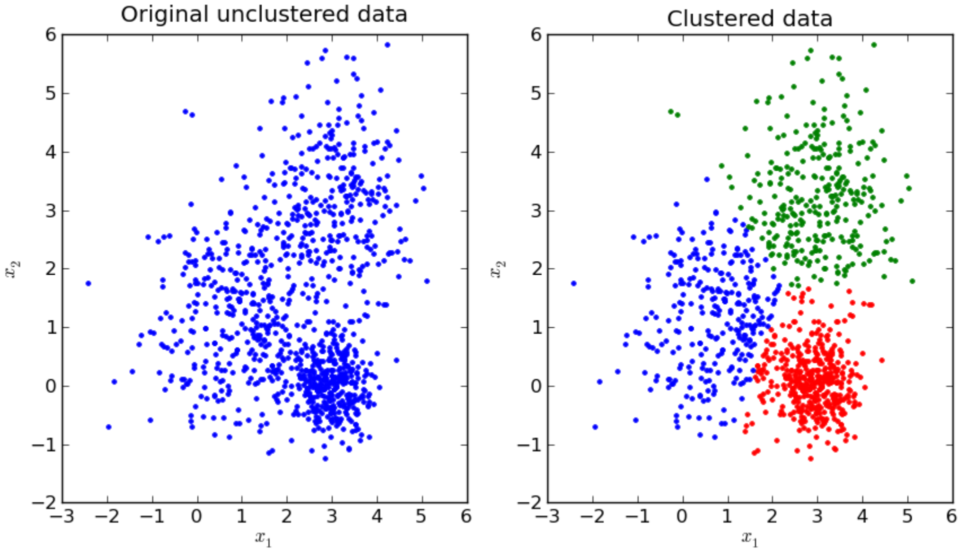
Gyms

- Gym
- Fitness Center
- Pool Gym
- Gymnastics Gym

- We performed a feature selection in order to assess each neighborhood's health market score. For that, we selected only the venues related to a ***healthy lifestyle*** as well as the ***HDI***, coordinates and region of each neighborhood;
- All the other venue categories were dropped for the analysis since we are seeking to aggregate the neighborhoods based on their relation to the health sector, which would increase the acceptability of the new venue.

MACHINE LEARNING ALGORITHM

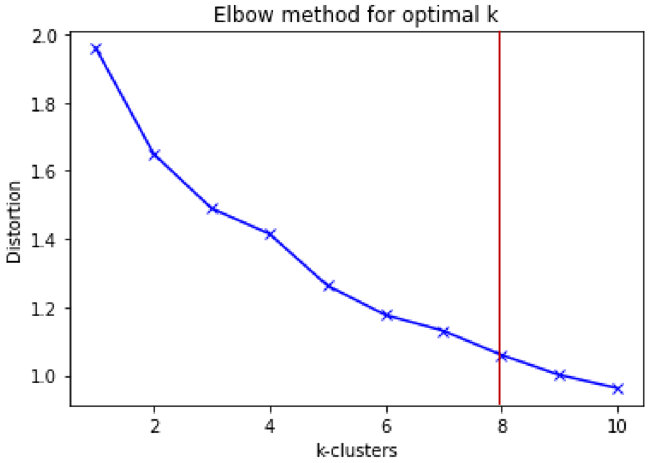
Selected algorithm: K-Means



K-Means clustering example

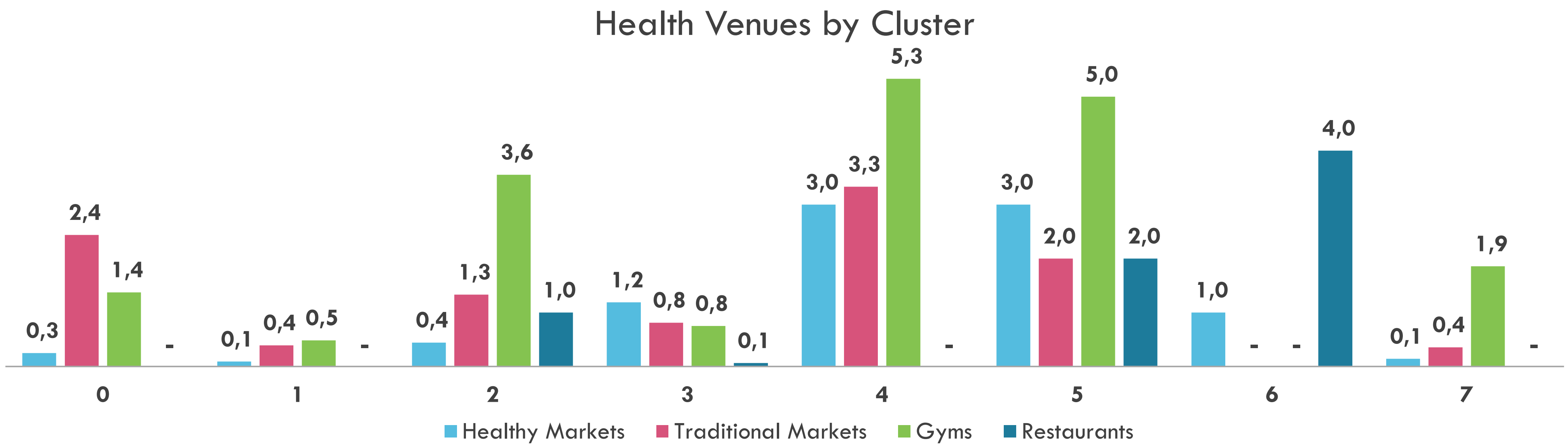
Cluster Labels	# of Neighborhoods in Cluster
0	16
1	33
2	9
3	16
4	3
5	1
6	1
7	14

- The K-Means algorithms is a model of unsupervised learning where it tries to minimize the Euclidean distance between each observation and their cluster's centroid based on the number of clusters that the user specified.
- For finding the optimal k number of clusters we conducted an ***elbow method*** analysis.



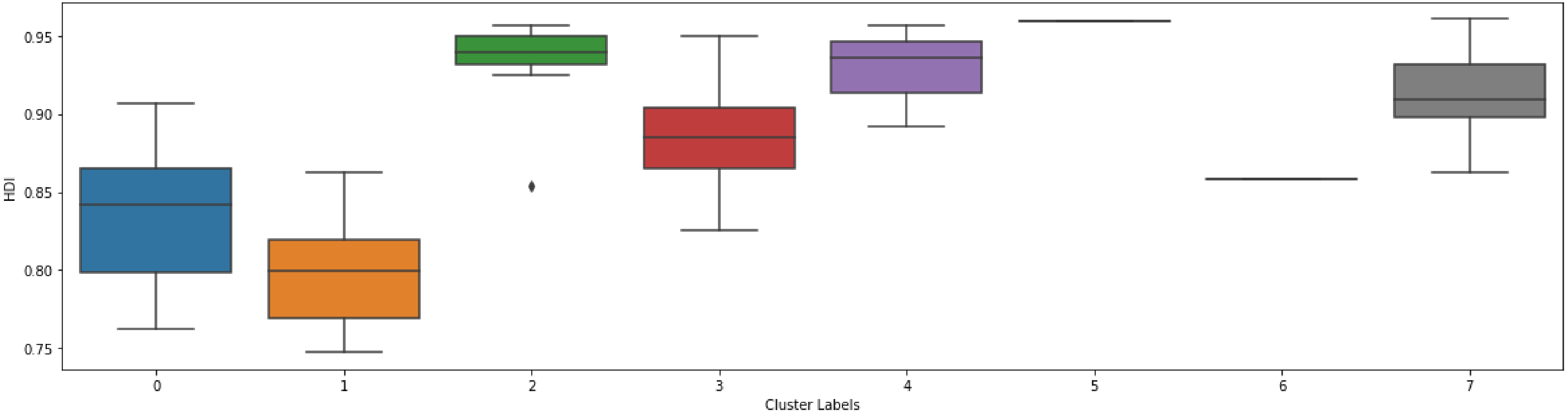
- No clear elbow could be made from the curve because the data is too sparse, however we can notice that the distortion loss after 8 clusters started to be a little flatter than before, therefore we continued the analysis with ***8 clusters*** in total.

DATA MINING – CLUSTER’S VENUES



- **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.

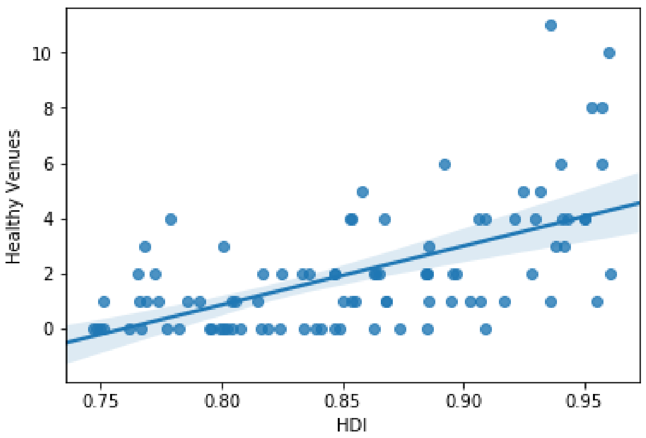
DATA MINING – CLUSTER’S HDI



- 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.

Correlation: 0.59

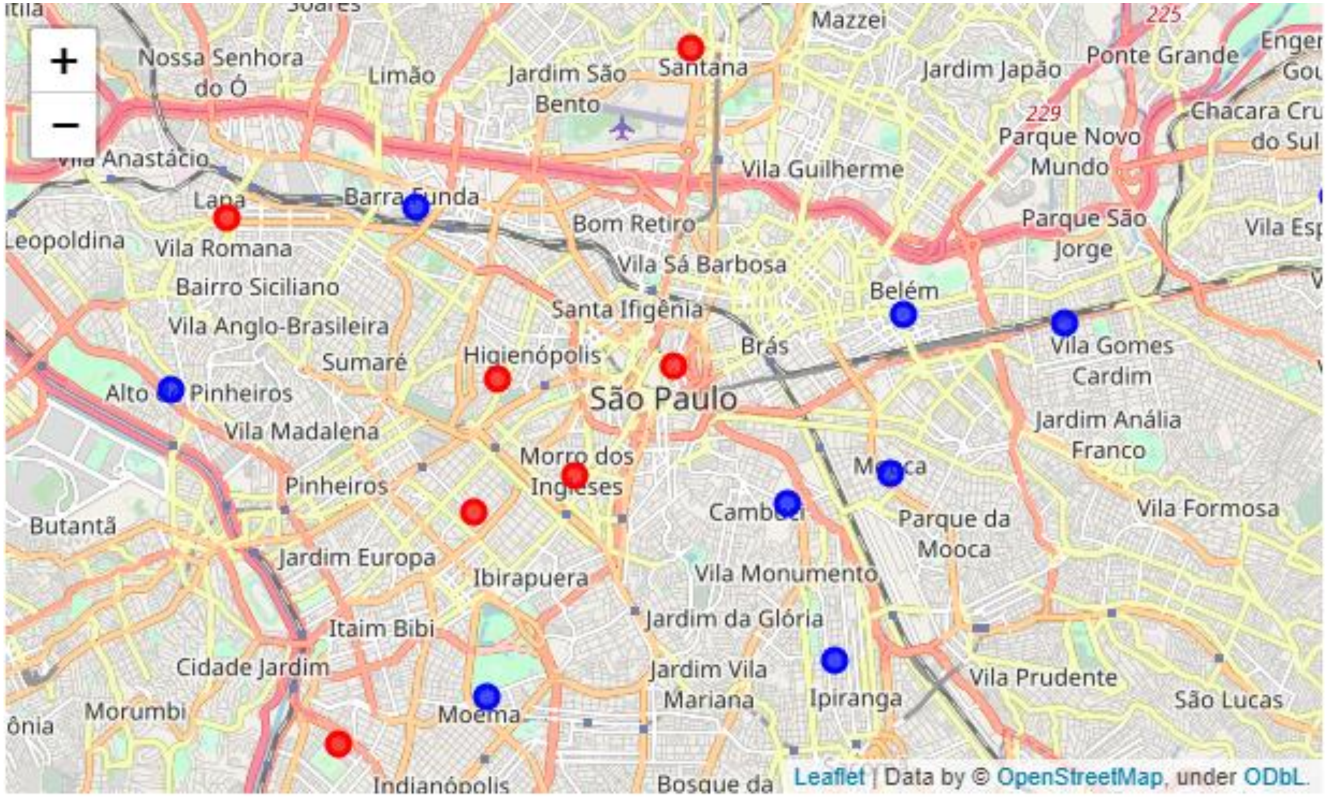
HDI vs Health Venues



- 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.

DATA MINING – CHOSEN CLUSTERS

- **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.



Sao Paulo map with clusters 2 (red) and 7 (blue)

RESULTS

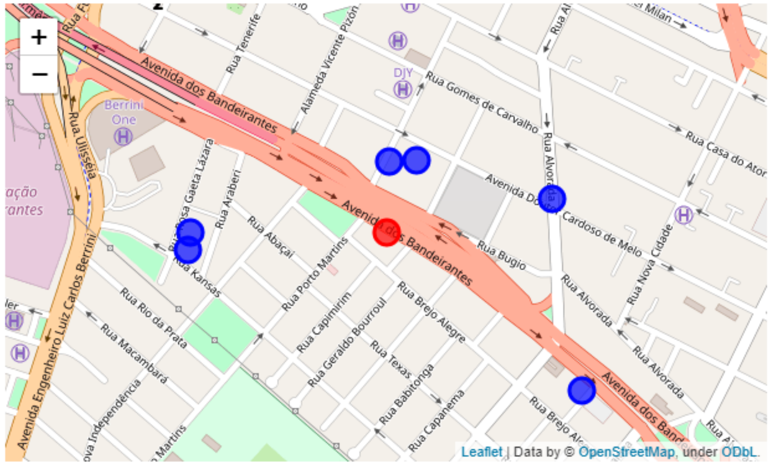
To come up with the top 5 neighborhoods in our analysis we will utilize an engineered feature, the *Healthy-Human Development Index* (HH Index). Its is based on the **desired attributes** such as high HDI, Gyms and Healthy Restaurants and **undesired attributes** such as Healthy Markets and Traditional Markets, which might represent competition for the new venue.

$$HH\ Index = \frac{HDI*10+\#Gyms+\#Healthy\ Restaurants}{\#Healthy\ Markets*5+\#Traditional\ Markets}$$

Neighborhood	Zone	Cluster Labels	HDI	Healthy Markets	Traditional Markets	Gyms	Healthy Restaurants	Total H. Venues	HH Index
Itaim Bibi	West	2	0.953	0	0	6	2	8	175.3
Lapa	West	2	0.941	0	0	3	1	4	134.1
Moema	Center-South	7	0.961	0	0	2	0	2	116.1
Alto de Pinheiros	West	7	0.955	0	0	1	0	1	105.5
Tatuapé	Southeast	7	0.936	0	0	1	0	1	103.6

Top 5 Neighborhoods for new venue based on selectes clusters

- 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 as the red spot on the following map;
- 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.



Recommended spot in top-1 neighborhood

CONCLUSION & RECOMMENDATIONS

- Due to the correlation between the Human Development index and the amount of healthy venues in the neighborhood, we might look for a **high HDI**;
- Due to the beginning stage that the health sector is in Brazil, we can look for the **lack of those kind of venues** in the selected neighborhoods to assess how saturated the market is in the area;
- Finally, the present analysis can recommend that the entrepreneurs interested in establishing a new Health Food venue can **<u>focus their attentions to neighborhoods with high HDI and low number of already established health markets</u>**, such as the top five recommended by the model
  - Itaim Bibi
  - Lapa
  - Moema
  - Alto de Pinheiros
  - Tatuape

Itaim Bibi's lounge park

Lapa's public market

Moema street view

Alto de Pinheiros from above

Tatuape neighborhood