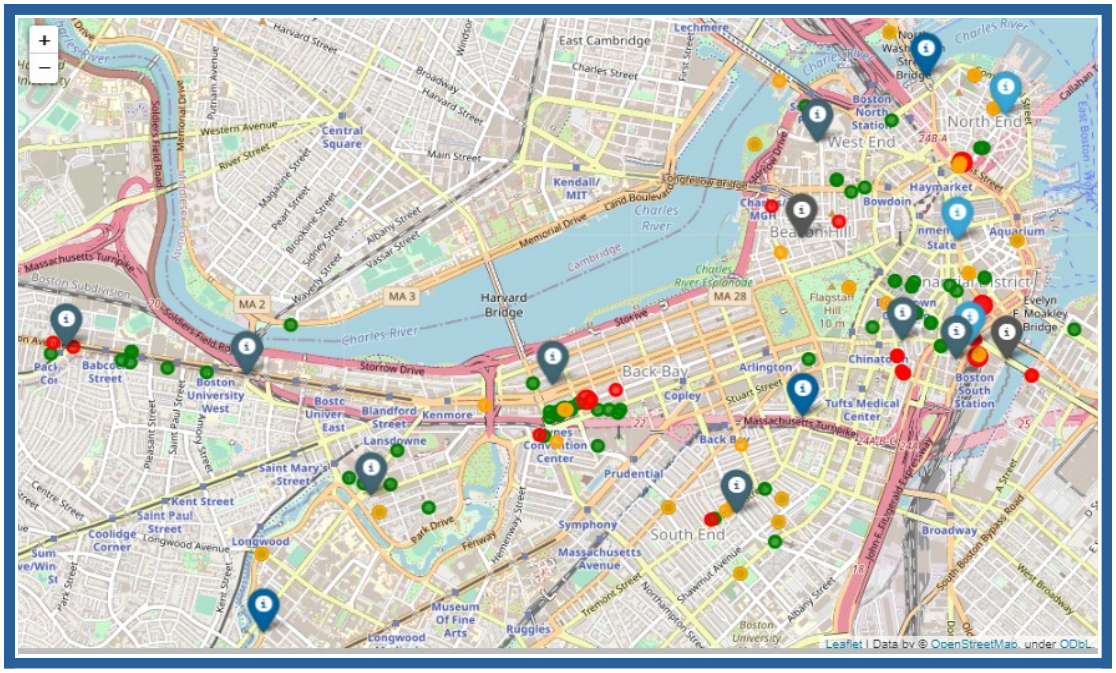
IBM Coursera Applied Data Science Specialization

Capstone Project

Location Analysis for a new Smoothie and Juice Shop in Boston Area



Maheesha Tennakoon

July 2020

1. Introduction

Now a days people consider more on healthy options when they select food and drinks. Lots of health and nutrition experts recommend eating fresh vegetables and fruits which helps our human body in many different ways. Fresh smoothies and juices are widely considered a healthy food option. People like to get smoothies and juices at any time of the day. However, there is a high tendency of people to seeking for a smoothies and juices while they are commuting to work, after an exercise/work out routine, in their shopping trips and visiting a park or participating other leisure activities.

When opening a smoothie and juice shop one of the most important factor is the location. It is important to find a neighborhood with a higher demand for the product as well as no sufficient amount shops serving smoothies and juices. A location in a neighborhood with a reach to shopping centers, gymnasiums, parks, universities, office places, transportation hubs and other higher people traffic areas will get more opportunity to establish a successful business.

This project will assist smoothie and juice business looking for a new opportunity in Boston neighborhoods. The location analysis and recommendations provided in this project will help their decision making process.

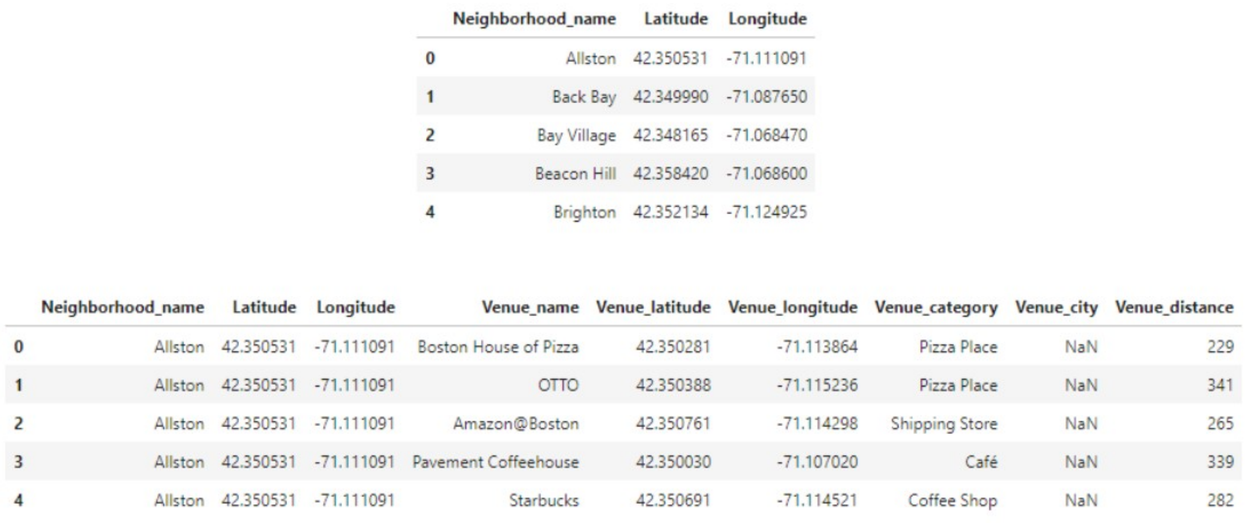
2. Business Problem

When establishing a new business, the business owners can benefit higher sales and a good profit if they open their shop in a neighborhood having a higher demand for the product and less supply. This project is intended to solve this business problem for a smoothie and juice shop to provide recommendations on the locations with higher business potential.

3. Data

The dateset used for the analysis was created using four sources listed below.

- List of neighborhoods in Boston, MA, USA.
  - Wikipedia page: https://en.wikipedia.org/wiki/Neighborhoods_in_Boston
- Latitude and longitude of each neighborhoods.
  - Python geocoder package
  - Project Link: https://github.com/DenisCarriere/geocoder
- Demographic Data for Boston's Neighborhoods
  - Boston Neighborhood Demographics, 2013-2017 American Community
  - Data Link: https://data.boston.gov/dataset/neighborhood-demographics
- Venue data about other business in each neighborhood.
  - Foursquare API
  - API Link: https://api.foursquare.com/v2/venues/explore



	Neighborhood_name	Latitude	Longitude
0	Allston	42.350531	-71.111091
1	Back Bay	42.349990	-71.087650
2	Bay Village	42.348165	-71.068470
3	Beacon Hill	42.358420	-71.068600
4	Brighton	42.352134	-71.124925

	Neighborhood_name	Latitude	Longitude	Venue_name	Venue_latitude	Venue_longitude	Venue_category	Venue_city	Venue_distance
0	Allston	42.350531	-71.111091	Boston House of Pizza	42.350281	-71.113864	Pizza Place	NaN	229
1	Allston	42.350531	-71.111091	OTTO	42.350388	-71.115236	Pizza Place	NaN	341
2	Allston	42.350531	-71.111091	Amazon@Boston	42.350761	-71.114298	Shipping Store	NaN	265
3	Allston	42.350531	-71.111091	Pavement Coffeehouse	42.350030	-71.107020	Café	NaN	339
4	Allston	42.350531	-71.111091	Starbucks	42.350691	-71.114521	Coffee Shop	NaN	282

Figure 1: Neighborhoods and venues datasets

4. Methodology

Extract neighborhood and venues data from Wikipedia and Foursquare API to analyze the areas in Boston to find suitable locations with higher business opportunity to open a smoothie and juice shop.

Getting Data

- Python web scraping library BeautifulSoup was used for web scraping the Wikipedia page https://en.wikipedia.org/wiki/Neighborhoods_in_Boston to obtain a list of neighborhoods in Boston, MA, USA.
- The **latitude and longitude of each neighborhood** were obtained using Python geocoder. Since this library failed in the local Jupyternotebook enviroenment, it is done in IBM Watson Studio. The results (neighborhood, latitude, longitude) stored as Pandas DataFrame as saved into csv file (Boston_Neighborhoods.csv) and downloaded from Watson Studio.
- Foursquare API endpoint explore was used to get **venues data near each neighborhood** location. This data consist of Venue_name, Venue_latitude, Venue_longitude, Venue_category and Venue_distance to the location. The venue dataset is saved to a csv file (Boston_Venues.csv).
- **Neighborhood deographics** were obtained from the City of Boston's open data hub named *Analyze Boston*. This data is downloaded as excel file presenting deographic data in neigborhood level in multiple sheets. The columns Median Income,Total Households, Total Population, and Median Age was manually picked and created a new dataset saved to a csv file (Boston_DemographicData.csv).

Creating Features for Clustering

- Out of 207 categories in the venue datasets 26 categories relevant for the business were selected. They were further identified as below.

Table 1: Identifying relevant venue categories

Highly Competitive venues	Juice Bar

	Smoothie Shop
Moderately Competitive	Ice Cream Shop Bubble Tea Shop Food & Drink Shop
Venues Increasing Business Opportunity	Gym / Fitness Center Gym Yoga Studio College Rec Center College Gym Boxing Gym
Shopping Venues	Big Box Store Grocery Store Department Store Clothing Store Sporting Goods Shop Mobile Phone Shop Automotive Shop Gift Shop Gift Shop Bookstore

Transportation	Bus Stop Bus Station Train Station Rental Car Location
Leisure	Park

- The above selected venue categories were grouped into the following ones.
    - Smoothie_Juice = {Smoothie Shop, Juice Bar}
    - Gym_Fitness = {Gym / Fitness Center, Gym, Yoga Studio, College Rec Center, College Gym, Boxing Gym}
    - Dessert_Drink = {Ice Cream Shop, Bubble Tea Shop, Food & Drink Shop}
    - Transportation = {Bus Stop, Bus Station, Train Station, Rental Car Location}
    - Shopping = {Big Box Store, Grocery Store, Department Store, Clothing Store, Sporting Goods Shop, Mobile Phone Shop, Automotive Shop}
    - Books_Gifts = {Gift Shop, Comic Shop, Bookstore}
- New features were created for clustering with a consideration of business opportunity in the area.
    - Smoothie_Juice_per_Shopping = Smoothie_Juice venues / Shopping venues
    - Smoothie_Juice_per_Gym_Fitness = Smoothie_Juice venues / Gym_Fitness venue
    - For the neighborhoods having no venues in the denominator were given -1.
- Venue count columns were scaled by the total venues for that neighborhood
- All the demographic features were scaled by the respective values for Boston City

Clustering Neighborhoods

- k- means clustering machine learning algorithm is used to create five neighborhood clusters using the features created on the extracted data.

Analysis of Neighborhood clusters

- Output of the clustering was analyzed along side with the neighborhood characteristics to provide recommendations on neighborhoods to open a new smoothie and juice shop.

5. Results

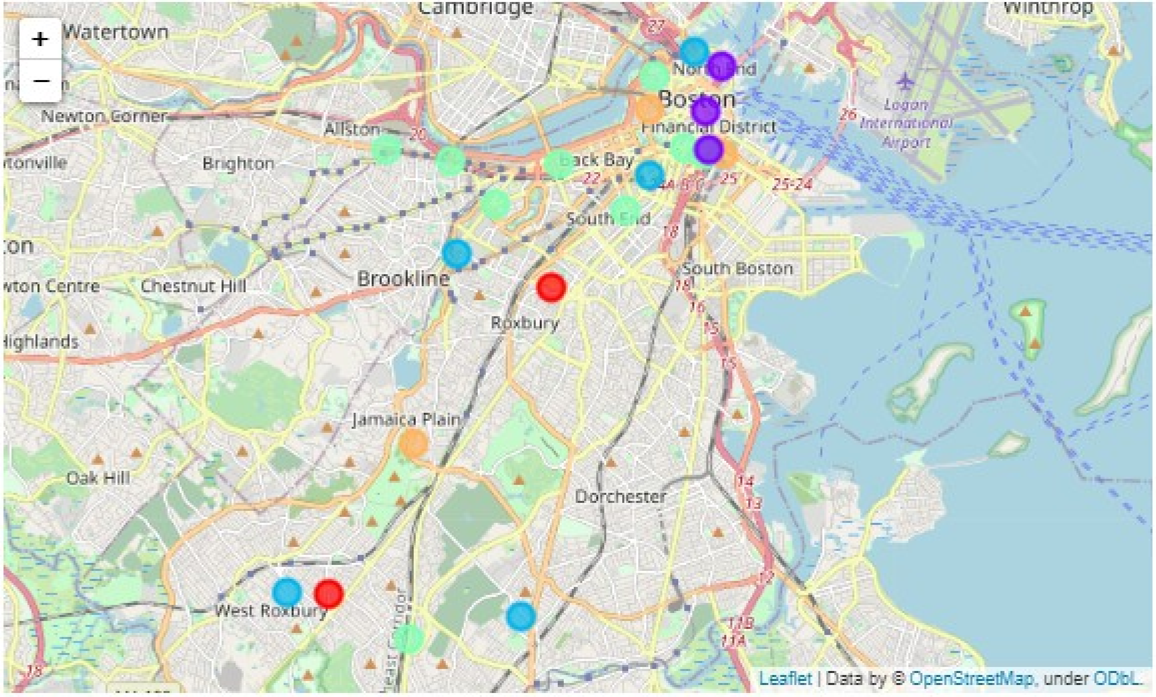
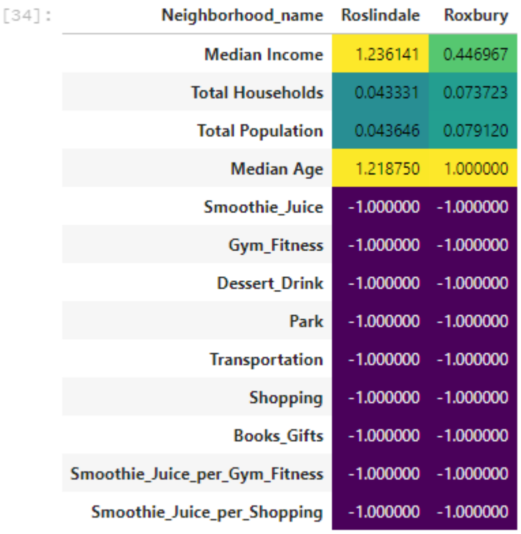


Figure 2 : Cluster Map: cluster 1 (red), cluster 2 (purple), cluster 3 (blue), cluster 4 (green), cluster 5 (orange)

Cluster Analysis

In cluster 1, no data about the focused venues are available. Therefore, it is hard to make a recommendation on the neighborhoods Roslindale and Roxbury in this cluster.

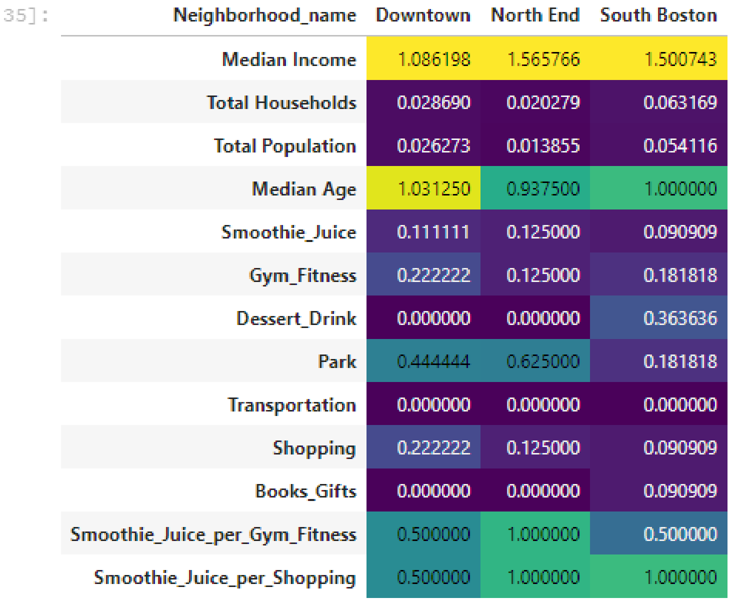


[34]:

Neighborhood_name	Roslindale	Roxbury
Median Income	1.236141	0.446967
Total Households	0.043331	0.073723
Total Population	0.043646	0.079120
Median Age	1.218750	1.000000
Smoothie_Juice	-1.000000	-1.000000
Gym_Fitness	-1.000000	-1.000000
Dessert_Drink	-1.000000	-1.000000
Park	-1.000000	-1.000000
Transportation	-1.000000	-1.000000
Shopping	-1.000000	-1.000000
Books_Gifts	-1.000000	-1.000000
Smoothie_Juice_per_Gym_Fitness	-1.000000	-1.000000
Smoothie_Juice_per_Shopping	-1.000000	-1.000000

Figure 3: Cluster 1 Neighborhoods

Cluster 2 neighborhoods have both types of the venues which can increase business opportunity as well as the competitive business such as Smoothie and Juice shops.

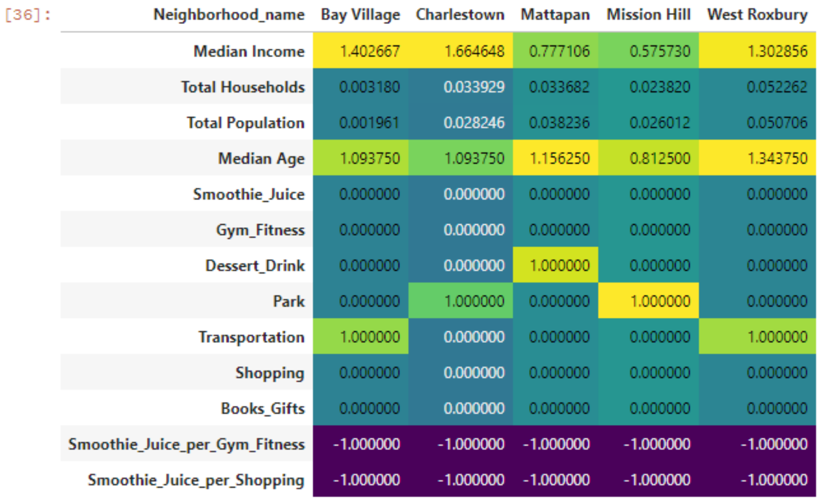


35]:

Neighborhood_name	Downtown	North End	South Boston
Median Income	1.086198	1.565766	1.500743
Total Households	0.028690	0.020279	0.063169
Total Population	0.026273	0.013855	0.054116
Median Age	1.031250	0.937500	1.000000
Smoothie_Juice	0.111111	0.125000	0.090909
Gym_Fitness	0.222222	0.125000	0.181818
Dessert_Drink	0.000000	0.000000	0.363636
Park	0.444444	0.625000	0.181818
Transportation	0.000000	0.000000	0.000000
Shopping	0.222222	0.125000	0.090909
Books_Gifts	0.000000	0.000000	0.090909
Smoothie_Juice_per_Gym_Fitness	0.500000	1.000000	0.500000
Smoothie_Juice_per_Shopping	0.500000	1.000000	1.000000

Figure 4: Cluster 2 Neighborhoods

Cluster 3 neighborhoods have no venues which can increase business opportunity such as Gym/Fitness Centers and Shopping Centers.

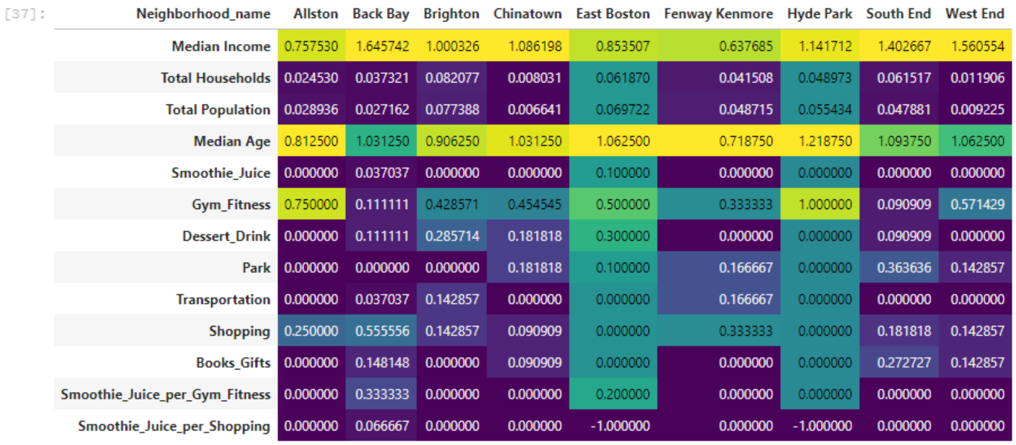


[36]:

Neighborhood_name	Bay Village	Charlestown	Mattapan	Mission Hill	West Roxbury
Median Income	1.402667	1.664648	0.777106	0.575730	1.302856
Total Households	0.003180	0.033929	0.033682	0.023820	0.052262
Total Population	0.001961	0.028246	0.038236	0.026012	0.050706
Median Age	1.093750	1.093750	1.156250	0.812500	1.343750
Smoothie_Juice	0.000000	0.000000	0.000000	0.000000	0.000000
Gym_Fitness	0.000000	0.000000	0.000000	0.000000	0.000000
Dessert_Drink	0.000000	0.000000	1.000000	0.000000	0.000000
Park	0.000000	1.000000	0.000000	1.000000	0.000000
Transportation	1.000000	0.000000	0.000000	0.000000	1.000000
Shopping	0.000000	0.000000	0.000000	0.000000	0.000000
Books_Gifts	0.000000	0.000000	0.000000	0.000000	0.000000
Smoothie_Juice_per_Gym_Fitness	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
Smoothie_Juice_per_Shopping	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000

Figure 5: Cluster 3 Neighborhoods

Cluster 4 neighborhoods have many venues including the ones can increase business opportunity such as Gym/Fitness Centers and Shopping Centers. The neighborhoods in this cluster can be highly considered to open a Smoothie and Juice shop. The neighborhoods Allston, Brighton, Fenway Kenmore, South End and West End are highly recommended to open a Smoothie and Juice shop based on the analysis with available data.

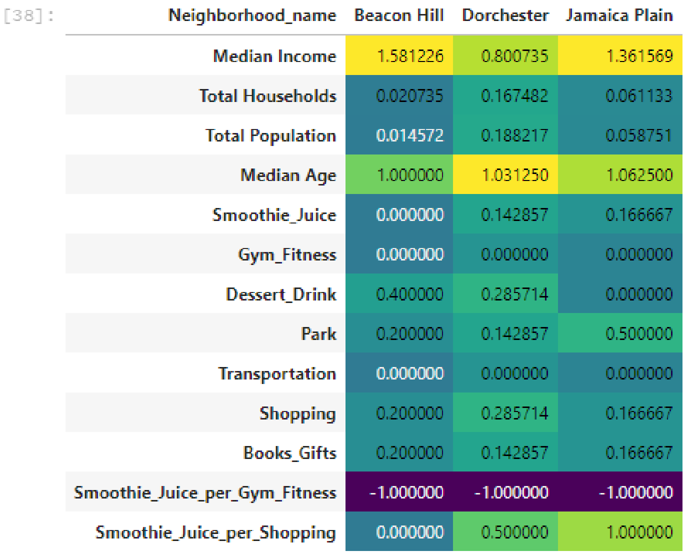


[37]:

Neighborhood_name	Allston	Back Bay	Brighton	Chinatown	East Boston	Fenway Kenmore	Hyde Park	South End	West End
Median Income	0.757530	1.645742	1.000326	1.086198	0.853507	0.637685	1.141712	1.402667	1.560554
Total Households	0.024530	0.037321	0.082077	0.008031	0.061870	0.041508	0.048973	0.061517	0.011906
Total Population	0.028936	0.027162	0.077388	0.006641	0.069722	0.048715	0.055434	0.047881	0.009225
Median Age	0.812500	1.031250	0.906250	1.031250	1.062500	0.718750	1.218750	1.093750	1.062500
Smoothie_Juice	0.000000	0.037037	0.000000	0.000000	0.100000	0.000000	0.000000	0.000000	0.000000
Gym_Fitness	0.750000	0.111111	0.428571	0.454545	0.500000	0.333333	1.000000	0.090909	0.571429
Dessert_Drink	0.000000	0.111111	0.285714	0.181818	0.300000	0.000000	0.000000	0.090909	0.000000
Park	0.000000	0.000000	0.000000	0.181818	0.100000	0.166667	0.000000	0.363636	0.142857
Transportation	0.000000	0.037037	0.142857	0.000000	0.000000	0.166667	0.000000	0.000000	0.000000
Shopping	0.250000	0.555556	0.142857	0.090909	0.000000	0.333333	0.000000	0.181818	0.142857
Books_Gifts	0.000000	0.148148	0.000000	0.090909	0.000000	0.000000	0.000000	0.272727	0.142857
Smoothie_Juice_per_Gym_Fitness	0.000000	0.333333	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.000000
Smoothie_Juice_per_Shopping	0.000000	0.066667	0.000000	0.000000	-1.000000	0.000000	-1.000000	0.000000	0.000000

Figure 6: Cluster 4 Neighborhoods

Cluster 5 neighborhoods have Shopping Centers but none of them have Gym/Fitness centers. These neighborhoods also have competition from other shops.



[38]:

Neighborhood_name	Beacon Hill	Dorchester	Jamaica Plain
Median Income	1.581226	0.800735	1.361569
Total Households	0.020735	0.167482	0.061133
Total Population	0.014572	0.188217	0.058751
Median Age	1.000000	1.031250	1.062500
Smoothie_Juice	0.000000	0.142857	0.166667
Gym_Fitness	0.000000	0.000000	0.000000
Dessert_Drink	0.400000	0.285714	0.000000
Park	0.200000	0.142857	0.500000
Transportation	0.000000	0.000000	0.000000
Shopping	0.200000	0.285714	0.166667
Books_Gifts	0.200000	0.142857	0.166667
Smoothie_Juice_per_Gym_Fitness	-1.000000	-1.000000	-1.000000
Smoothie_Juice_per_Shopping	0.000000	0.500000	1.000000

Figure 7: Cluster 5 Neighborhoods

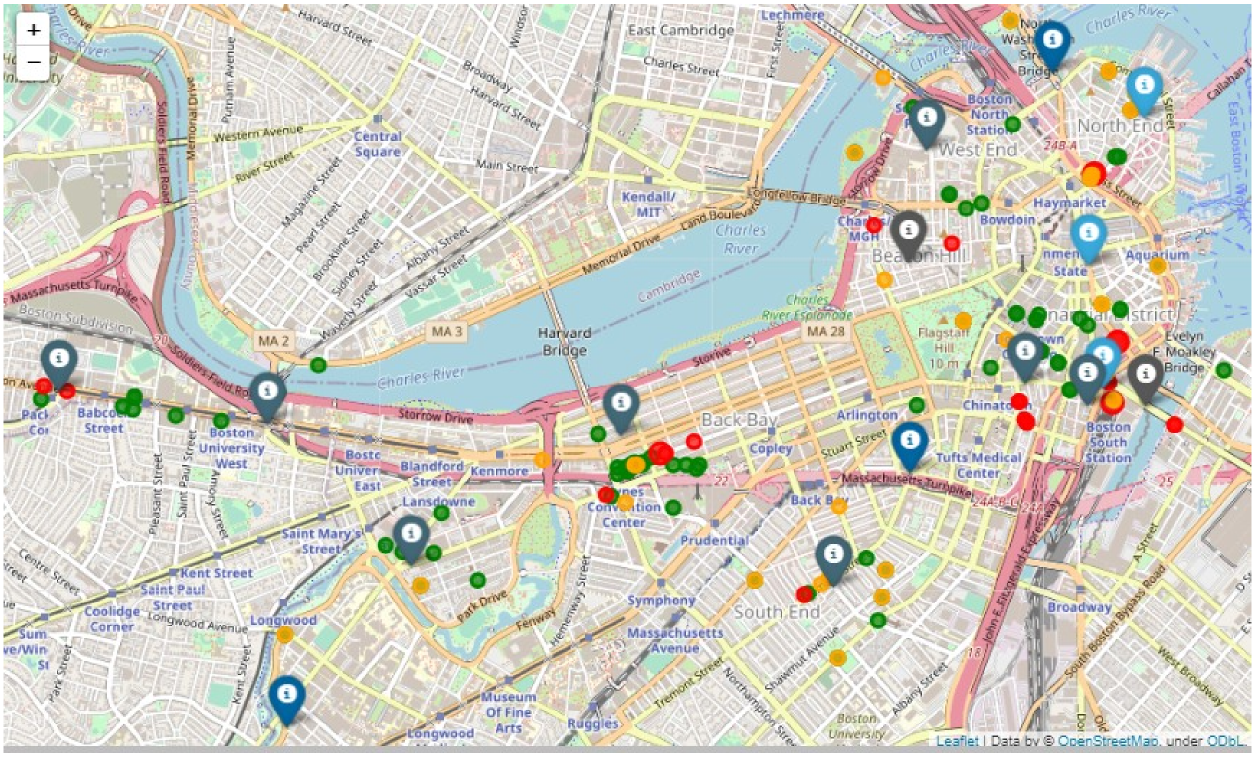


Figure 8: Adding Venues to Neighborhood Cluster map

6. Discussion

Finding a location with an elevated market opportunity is one of the most important things to consider when starting a business. In this project, neighborhoods and venues data were utilized with a machine learning approach to produce a data driven recommendation for a location to open a Smoothie and Juice shop in Boston, MA.

The impact from other business venues in the area will have negative impact based on the nature of their business and the market of their products. When considering a Smoothie and Juice shop, the general understanding is that the customers visiting business venues like Gym/Fitness centers, Shopping centers and Leisure activity locations are highly likely to become customers of Smoothie and Juice shop. However, need to be verified with data collected for a proper study. In other hand the competition coming from the business venues in similar business also needs to be considered when deciding a location. If someone opens a shop in a location with lower demand and higher competition, the risk of running a successful business is higher.

Looking at the median income in these neighborhoods can also help to business owner to determine the price limits of the products and decide on what items should be on the menu in different neighborhoods. Analyzing total population and total households also important to locate an area with a suitable population beneficial for the business. Selecting a populated area can serve many people. If the population is high, Smoothie and Juice business can expect a larger customer base and, it makes a good profit to the business. Looking at the median age can help to the business owner to create the menu according to the age groups of the neighborhood. Therefore, median income, total population, total households and median age are also good factors to determine a location for open a Smoothie and Juice shop in the Boston area.

7. Conclusion

In this project, 22 neighborhoods in Boston, MA were clustered using k- means clustering algorithm providing demographic and venues data. Neighborhoods and the clusters were analyzed to get insights into business opportunity to open and successfully run a Smoothies and Juice shop. The neighborhoods Allston, Brighton, Fenway Kenmore, South End and West End are clustered together and found those neighborhoods have venues favorable to a nearby Smoothies and Juice shops. The data also showed the absence of competitive businesses in the area. The demographic data also provided insights favorable to opening a Smoothies and Juice shop in any of these neighborhoods.

The current analysis has some limitations and it can be further improved by incorporating more demographic data and data, consumer trends, pedestrian traffic, etc. The free version of Foursquare API used has limitation of giving only100 venues per location. Therefore, many locations could have left out from the dataset. Therefore, using the full (paid) version to obtain an extended list of venues will be crucial for an analysis that can apply to a real business use. The current framework developed in this project can be reused for that extended study with more data.

8. References

1. "Segmenting and Clustering Neighborhoods in New York City" by Alex Aklson and Polong Lin, Cognitive Class.
2. Neighborhoods in Boston, https://en.wikipedia.org/wiki/Neighborhoods_in_Boston
3. Geocorder, https://github.com/DenisCarriere/geocoder
4. Boston Neighborhood Demographics, 2013-2017 American Community, https://data.boston.gov/dataset/neighborhood-demographics
5. Foursquare API Documentation, https://developer.foursquare.com/docs/api-reference/venues/search/
6. Scikit-Learn Kmean, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
7. Folium, https://python-visualization.github.io/folium/
8. Beautifulsoup, https://www.crummy.com/software/BeautifulSoup/bs4/doc/