Capstone project

Sofia Nord

Introduction

Manhattan Island is in the center of New York City and is a center of attention for tourism, education, residential, shopping, sports activities and job employment. Manhattan is well known for the various activities you can do when visiting the area.

During the last years, the interest in a healthy lifestyle has increased and therefore a friend of mine has hired me as a contractor to investigate what neighborhood on Manhattan that could be a suitable area to establish a healthy food store in.

Objective

What neighborhood on Manhattan is most suitable for establishment of a new healthy food store?

In this project, the Manhattan area will be studied in details by using data from Foursquare and using machine learning techniques for segmentation and clustering. By using segmentation and clustering, this project aim to determine in which neighborhood on Manhattan it is most suitable to establish a new healthy food store when having in mind competition and location of gyms, metro stations, residential buildings/apartments and offices.

Data

The data of the different neighborhoods on Manhattan and their longitude and latitude data are acquired via a link that was used in one of the labs from week 3 in the capstone course:

https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json

From Foursquare following venue data for all neighborhood on Manhattan will be used for segmentation and clustering:

- Location of existing healthy food stores
- Location of gym/fitness centers
- Location of offices
- Location of metro stations
- Location of residential buildings/apartments

Please note, the amount and accuracy of data captured can not 100% determine correct classification in real world. The data will be leveraged in order to determine which neighborhood is the most appropriate for establishment of a new healthy food store.

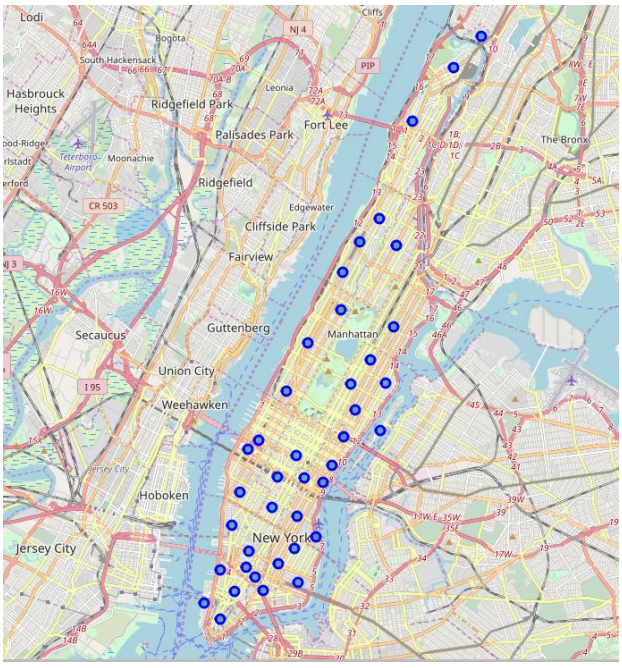
Methodology

In this project, the basic methodology taught in week 3 labs in the Capstone course will be used.

- Focus will be on Manhattan of all boroughs in New York City, therefore only the neighborhoods on Manhattan will be segmented and clustered.

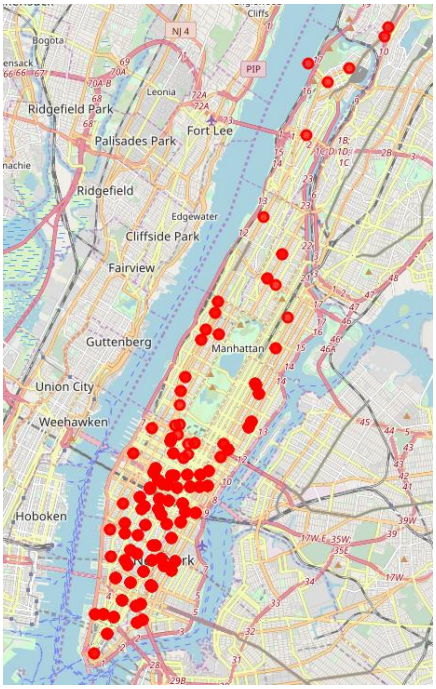
	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688
5	Manhattan	Manhattanville	40.816934	-73.957385
6	Manhattan	Central Harlem	40.815976	-73.943211
7	Manhattan	East Harlem	40.792249	-73.944182
8	Manhattan	Upper East Side	40.775639	-73.960508
9	Manhattan	Yorkville	40.775930	-73.947118
10	Manhattan	Lenox Hill	40.768113	-73.958860
11	Manhattan	Roosevelt Island	40.762160	-73.949168
12	Manhattan	Upper West Side	40.787658	-73.977059
13	Manhattan	Lincoln Square	40.773529	-73.985338
14	Manhattan	Clinton	40.759101	-73.996119
15	Manhattan	Midtown	40.754691	-73.981669
16	Manhattan	Murray Hill	40.748303	-73.978332
17	Manhattan	Chelsea	40.744035	-74.003116
18	Manhattan	Greenwich Village	40.726933	-73.999914
19	Manhattan	East Village	40.727847	-73.982226
20	Manhattan	Lower East Side	40.717807	-73.980890
21	Manhattan	Tribeca	40.721522	-74.010683

Latitude and longitude data for some of the neighborhoods on Manhattan.

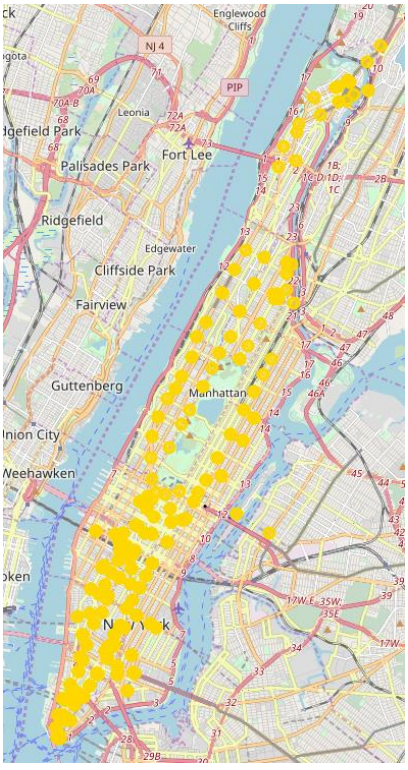


Neighborhoods on Manhattan

- Foursquare API will be used to explore existing healthy food stores, gym/fitness centers, offices, metro stations and residential buildings/apartments for each neighborhood on Manhattan. Pictures below shows example of exploring both existing healthy food stores and metro stations



Healthy food stores on Manhattan.



Metro stations on Manhattan.

- The sum of the venue categories for each neighborhood will be calculated.

	Borough	Neighborhood	Latitude	Longitude	Health food stores	Gym	Metro stations	Offices	Residential building
0	Manhattan	Marble Hill	40.876551	-73.910660	2.0	26.0	5.0	39.0	48.0
1	Manhattan	Chinatown	40.715618	-73.994279	9.0	50.0	15.0	50.0	50.0
2	Manhattan	Washington Heights	40.851903	-73.936900	1.0	22.0	5.0	37.0	50.0
3	Manhattan	Inwood	40.867684	-73.921210	3.0	13.0	11.0	38.0	50.0
4	Manhattan	Hamilton Heights	40.823604	-73.949688	2.0	39.0	6.0	45.0	50.0
5	Manhattan	Manhattanville	40.816934	-73.957385	1.0	46.0	5.0	46.0	50.0
6	Manhattan	Central Harlem	40.815976	-73.943211	3.0	48.0	14.0	49.0	50.0
7	Manhattan	East Harlem	40.792249	-73.944182	2.0	50.0	7.0	44.0	50.0
8	Manhattan	Upper East Side	40.775639	-73.960508	5.0	50.0	5.0	50.0	50.0
9	Manhattan	Yorkville	40.775930	-73.947118	4.0	50.0	3.0	49.0	50.0
10	Manhattan	Lenox Hill	40.768113	-73.958860	5.0	50.0	5.0	50.0	50.0
11	Manhattan	Roosevelt Island	40.762160	-73.949168	3.0	50.0	3.0	48.0	50.0
12	Manhattan	Upper West Side	40.787658	-73.977059	5.0	50.0	7.0	44.0	50.0
13	Manhattan	Lincoln Square	40.773529	-73.985338	10.0	50.0	10.0	50.0	50.0
14	Manhattan	Clinton	40.759101	-73.996119	17.0	50.0	15.0	50.0	50.0

Sum of the venue categories for each neighborhood on Manhattan.

- For each venue category, a weight (or penalty) has been defined based on what is considered most important when establishing a new healthy food store. These weights can be altered based on what is perceived as important when establishing a new healthy food store.
    - Existing healthy food stores have been weighted with -1 since we want to minimize the competition
    - Metro stations have been weighted with 1 as commuters prefer to have the store close
    - Offices have been weighted with 1 since many people want "fast food" but healthy
    - Gym/fitness centers have been weighted with 1.5, since many people exercising at the gym often want to combine training with health food
    - Residential buildings/ apartment have been weighted with 2, since people living in residential buildings want to be able to cook/buy and eat healthy.

```
# negative weight, because I want to open a healthy food store
weight_health = -1

# positive weight, because people prefer when communiting by m
weight_met = 1

# positive weight, because people exercisng at the gym want to
weight_gym = 1.5

# positive weight, because people working at offices want "fas
weight_off = 1.2

# positive weight because resedentials want to be able to eat
weight_res = 2
```

A weight for each venue category defined.

- A score for each neighborhood will be calculated to get the weighted sum of the amount of venues in each of the neighborhoods.

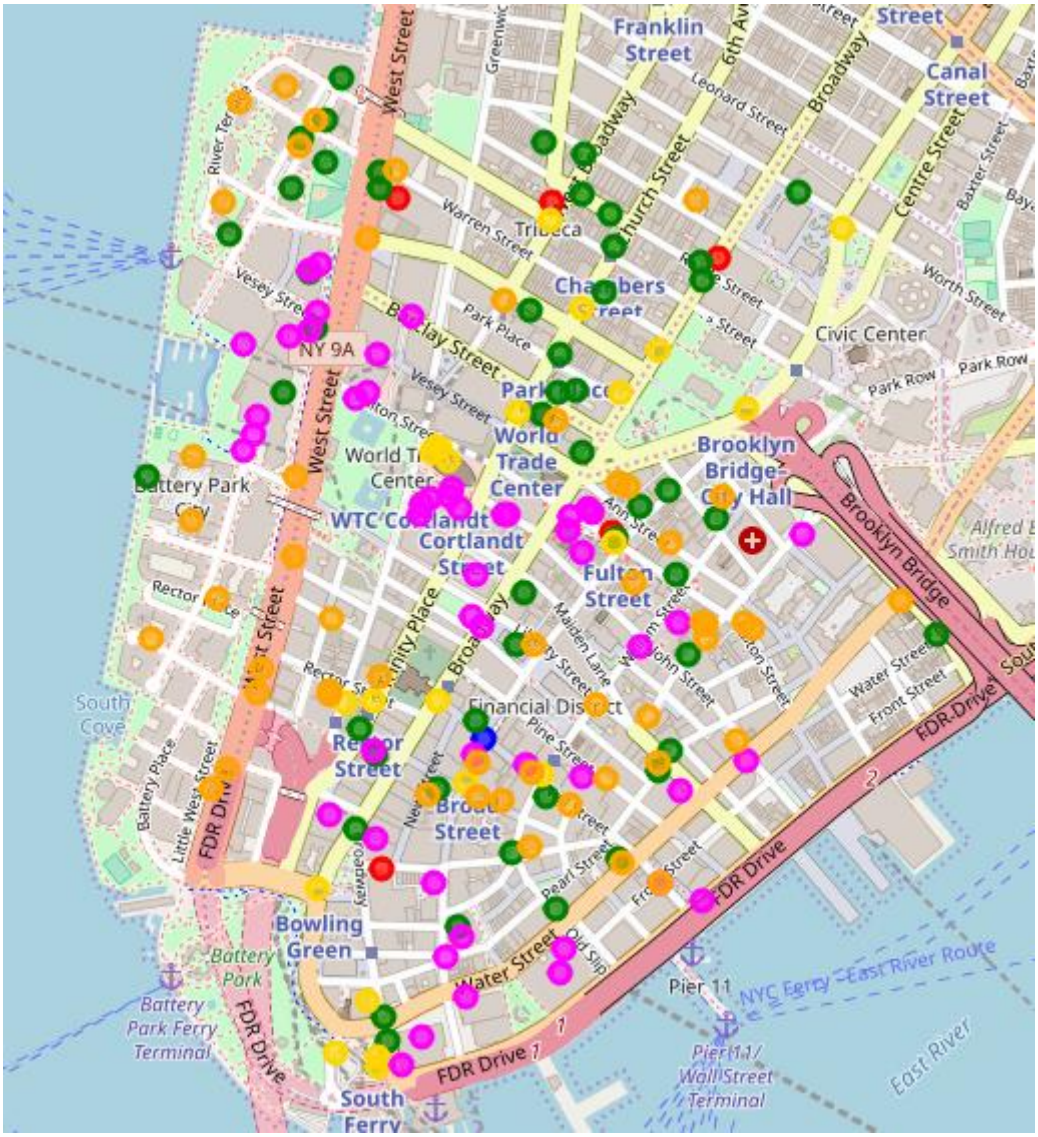
	Neighborhood	Score
29	Financial District	252.0
28	Battery Park City	251.0
32	Civic Center	250.0
22	Little Italy	248.0
21	Tribeca	246.0
31	Noho	244.0
18	Greenwich Village	243.0
6	Central Harlem	241.8
1	Chinatown	241.0
23	Soho	241.0
17	Chelsea	238.0
20	Lower East Side	238.0
24	West Village	237.0
34	Sutton Place	237.0
38	Flatiron	237.0
30	Carnegie Hill	235.8
8	Upper East Side	235.0

The score for each neighborhood.

Results

The result of the computing, visible in table above, shows that the neighborhood with the best score is "Financial district" and therefore it is the best option of a neighborhood to establish an healthy food store in. "Financial district" is closely followed by Battery Park City.

The result shows that the best option to maximize the amount of costumers from offices, metro stations, gym/fitness centers and residential buildings/apartments, and at the same time minimize the competition, is to establish the healthy food store in the "Financial district".



The Financial District.

Discussion

Using Foursquare API, we could capture data of different venues in New York City and specifically on Manhattan.

Based on the result, we were able to suggest a neighborhood on Manhattan that will be feasible to establish a new healthy food store in. At the same time, more perspectives are needed to have into account when deciding about establishment, for instance rent costs, crime rate, other establishments and details plans (infrastructure, buildings) for the neighborhood. Also, using smaller geographical area e.g. a neighborhood instead of whole Manhattan could improve the accuracy in the scores.

The classification method used in this project is a good start as support when looking at areas to establish a new venue. Further studies could be made within this area to increase the quality even more.

Conclusion

Using Foursquare API, we have been able to determine a suitable neighborhood where a healthy food store can be established store when having in mind competition and location of gyms, metro stations, residential buildings/apartments and offices.

The classification can become more qualitative if more perspective are included, for example rent costs, crime rate, other establishments and details plans (infrastructure, buildings) for the neighborhood.