New York City Neighborhood Suitability for a Business Plan: Healthy Food Store

1. Introduction

New York City has 306 neighborhoods in 5 boroughs. While some neighborhoods can bear similar characteristics, other neighborhoods can be unique and appropriate for a specific business purpose. For example, a high density of coffee shops or cafes can be a distinguishing factor for the neighborhoods with high density of office buildings. Here it would be beneficial to open a business restaurant provided that the restaurant density is not extremely high or a special restaurant after careful analysis of the data. On the other hand, for example, neighborhoods with parks, sport and leisure time facilities, groceries and schools could be a good living choice for a family with children.

Characteristic features of these neighborhoods can be determined based on geolocation data and various statistics or machine learning techniques. In this project, Foursquare venues location data is used to explore New York City neighborhoods using K-Means clustering. Neighborhoods are divided into clusters based on their similarity, i.e. type and occurrence of venues.

Business problem

Problem/question to solve: What are the best candidate neighborhoods to open a store with healthy food?

Analysis of location data can be used to answer many questions. Here, we try to find the best neighborhood to open a shop with healthy food. We want to find the neighborhoods that would be the best candidates to open the shop that sells products for active lifestyle and healthy diet, such as bioproducts including fresh vegetables and fruits, wholegrain products, food wealthy on protein, special types of flours, cereals, grains, etc. It's a shop where people could find all they need for their nutrition needs. The project aims to determine the best neighborhood(s) to start a new healthy food store.

Target audience

The study's target audience are businessmen or contractors that would like to start a successful healthy food store in a new area. Thus, the study will help to assist in the decision making process where to start the new store in order to maximize profits and minimize risks related to opening of the new shop.

A properly selected location will help to gain a stable and potentially increasing number of target customers. It will also eliminate losses that could originate e.g. from insufficient abundance of target customers. The study aims to predict the adequate locations for the business purpose with respect to the characteristics of the place and people that will likely visit the place.

Assumptions and considerations

- Let's assume that people with an active lifestyle use facilities like gyms, pools, other sport facilities or parks. Neighborhoods with these features would be proper neighborhoods for such a healthy food shop.

- There is a high chance that products of healthy lifestyle are commonly sold in supermarkets and groceries. Our candidate neighborhoods shouldn't be rich in these facilities. We don't want to add another shop if there are many nearby shops, because it could reduce the profits.
- High abundance of restaurants of different kinds, fast food, pizza and other places might suggest that the neighborhood is not the best candidate for our business idea. Such neighborhoods might be rich in social and cultural life, and people wouldn't spend their time looking for healthy products here.
- Neighborhood clustering based on abundance of venues belonging to different categories enables decisions whether the neighborhood is a good candidate or not.

2. Data

Data sources

We use data from two sources:

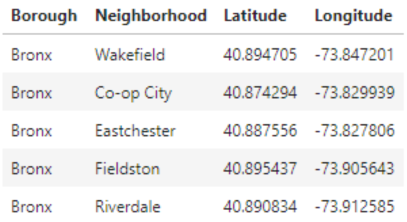
- New York City neighborhood data (available from here: NYU Spatial Data Repository) that contains following information about every neighborhood (the data fields are self-explanatory):
  - neighborhood name
  - borough name
  - neighborhood latitude
  - neighborhood longitude
- location data obtained from Foursquare API that include information about venues and their categories in the respective neighborhood

Both data is converted from JSON format to pandas dataframes to make it available for easy manipulation and analysis.

Location data will be used to cluster neighborhoods based on their similarities.

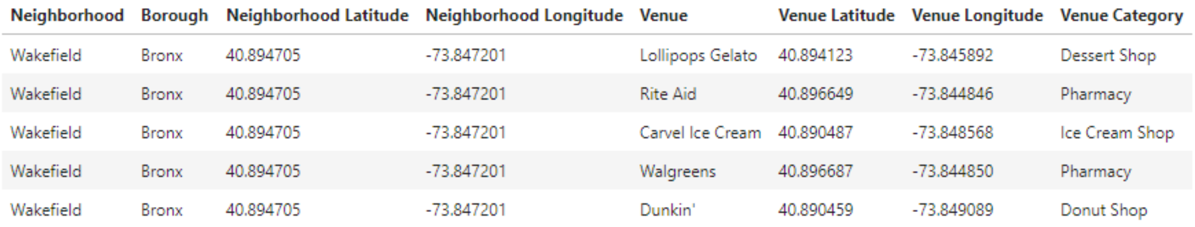
Examples of data:

1. New York City neighborhood data as a Pandas dataframe:



Borough	Neighborhood	Latitude	Longitude
Bronx	Wakefield	40.894705	-73.847201
Bronx	Co-op City	40.874294	-73.829939
Bronx	Eastchester	40.887556	-73.827806
Bronx	Fieldston	40.895437	-73.905643
Bronx	Riverdale	40.890834	-73.912585

2. Location data as a Pandas dataframe:



Neighborhood	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Wakefield	Bronx	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
Wakefield	Bronx	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
Wakefield	Bronx	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
Wakefield	Bronx	40.894705	-73.847201	Walgreens	40.896687	-73.844850	Pharmacy
Wakefield	Bronx	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Location data contain much more information but we will use only venues and their categories to cluster neighborhoods.

Data preparation and pre-processing

New York City neighborhoods data were obtained as a geo JSON file. As a first step, the JSON data were converted to a pandas dataframe. More precisely, JSON 'features' contain the relevant information in fields called 'properties.borough' (borough name), 'properties.name' (neighborhood name), and 'geometry.coordinates' (latitude and longitude). Those fields were transferred to 'Neighborhood', 'Borough', 'Latitude' and 'Longitude' columns in the pandas dataframe.

The neighborhood dataframe was examined to check the number of neighborhoods in total and in the boroughs. It was also found that some neighborhoods in different boroughs have the same name. Therefore, the neighborhood name alone does not uniquely identify the neighborhood, and a combination of neighborhood and borough was used for this purpose.

Venues data were obtained in a JSON format and were converted into a pandas dataframe. For each neighborhood, the venue limit was set to 100 and the radius of 500 meters was searched. That is, neighborhood latitude and longitude coordinates given in the New York City neighborhoods dataset were used as a central point. Foursquare API returns plenty of venue information – in this study, only venue name ('venue.name'), venue category ('venue.categories') and venue coordinates ('venue.location.lat', 'venue.location.lng') were used. All these fields are present in 'response.group.items' part of the response. The respective column names in pandas dataframe are following: 'Venue', 'Venue Latitude', 'Venue Longitude' and 'Venue Category'. Every row in the venues dataframe contains information about one venue in a certain neighborhood.

The venue dataframe was carefully examined. The venue categories were checked and any venues with category 'Neighborhood' were removed, because category 'Neighborhood' looks like an invalid category (there were 5 such venues in total).

In order to use K-Means clustering, the data was prepared as described in following steps:

- the venues dataset was used to create a new dataframe with one-hot encoded venue categories
- based on the one-hot encoded dataset, a grouped dataset with average frequencies of venue categories per neighborhood was created
- the grouped dataset with averaged venue categories was used as the input for the K-Means algorithm

For the purposes of evaluation of the results, another dataset containing the top ten most frequent venues for a neighborhood was created.

3. Methodology

We used standard K-Means Clustering to cluster neighborhoods based on their similarities measured in terms of different venue categories and their abundance in a neighborhood. To analyze neighborhoods and study the effects of clustering, following approach was used:

- cluster all neighborhoods in New York City, irrespective of the boroughs they belong to:
    - use K (number of clusters) 5 and 10, and compare the results
- cluster neighborhoods within each borough, i.e. take only neighborhoods belonging to one borough at a time:
    - use K 5 and 8, and compare the results

The goal of this approach is to:

- find a reasonable way to cluster neighborhoods
- determine the similarity of neighborhoods within boroughs and among boroughs
- recommend proper candidate neighborhoods to start a healthy food store

The results were evaluated by observation of the most common venue categories and their abundance in neighborhood clusters or individual neighborhoods.

4. Exploratory data analysis

New York city has 306 neighborhoods that belong to 5 boroughs (Figure 1):

- Bronx
- Brooklyn
- Manhattan
- Queens
- Staten Island

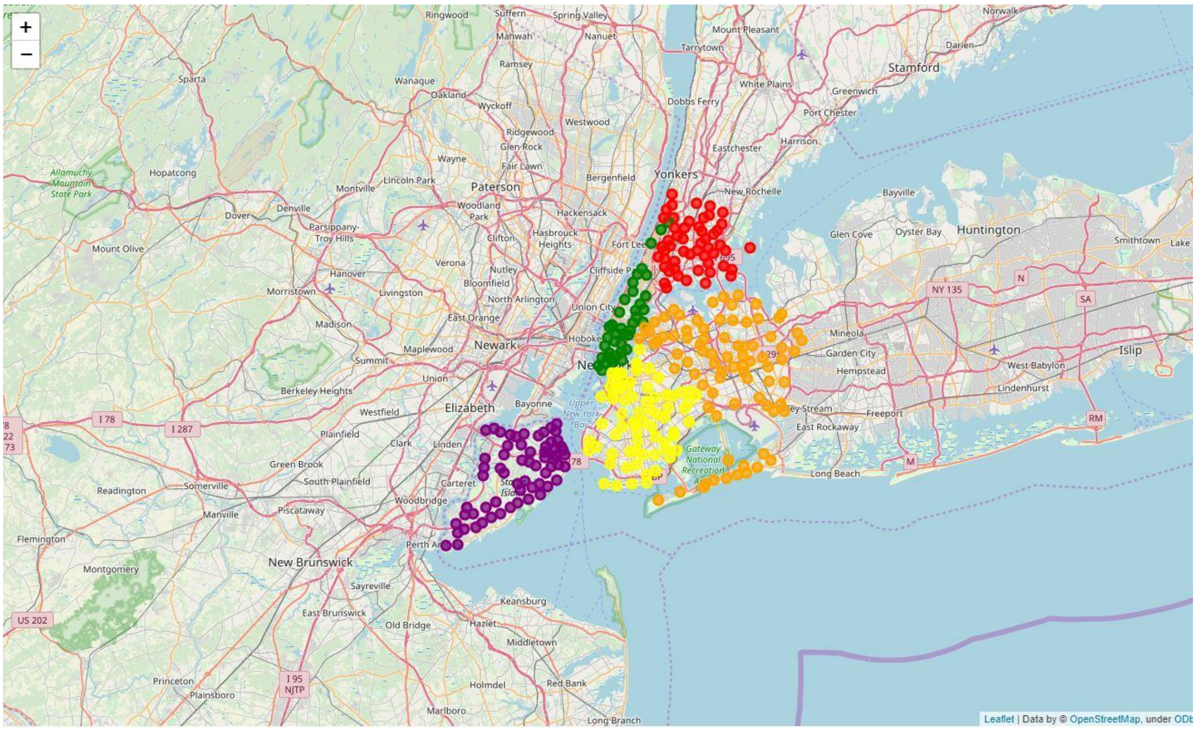


Figure 1 *Map of New York City with neighborhoods superimposed on top. Color legend: red – Bronx, yellow – Brooklyn, green – Manhattan, orange – Queens, purple – Staten Island.*

The overall number of venues obtained for all neighborhoods is about 10200, which would be about 33 venues per neighborhood. However, some of the neighborhood have more venues (Table 1), while the others can have as little as 1 venue (Table 2).

Neighborhood, Borough	Number of venues
Yorkville, Manhattan	100
Chinatown, Manhattan	100
Sunnyside Gardens, Queens	100
Lenox Hill, Manhattan	100
South Side, Brooklyn	100
Soho, Manhattan	100
Brooklyn Heights, Brooklyn	100
Lincoln Square, Manhattan	100
Little Italy, Manhattan	100
Carnegie Hill, Manhattan	100
Carroll Gardens, Brooklyn	100
Chelsea, Manhattan	100
Civic Center, Manhattan	100
Sutton Place, Manhattan	100
Clinton, Manhattan	100

Table 1 *Examples of neighborhoods with the highest number of venues*

Neighborhood, Borough	Number of venues
Brookville, Queens	1
Somerville, Queens	1
Port Ivory, Staten Island	1
Mill Island, Brooklyn	1
Todt Hill, Staten Island	1
Grymes Hill, Staten Island	1
Bayswater, Queens	2
Malba, Queens	3
Fieldston, Bronx	3
Country Club, Bronx	3

Table 2 *Examples of neighborhoods with the lowest number of venues*

Further exploration revealed, that 63 out of all neighborhoods have more than 50 venues, 58 neighborhoods have less than 10 venues, and 21 neighborhoods have less than 5 venues.

The venues belong to 432 different categories. The most abundant ones are Pizza Place, Italian Restaurant, Coffee Shop and Deli/Bodega with over 280 occurrences (Table 3).

Venue category	Abundance	Venue category	Abundance
Pizza Place	439	**Pharmacy**	175
Italian Restaurant	308	**American Restaurant**	173
Coffee Shop	294	**Café**	167
Deli / Bodega	286	**Donut Shop**	166
Bar	222	**Park**	163
Bakery	222	**Ice Cream Shop**	145
Chinese Restaurant	213	**Bank**	144
Sandwich Place	188	**Gym / Fitness Center**	128
Grocery Store	184	**Gym**	119
Mexican Restaurant	181	**Bagel Shop**	113

Table 3 *Examples of neighborhoods with the lowest number of venues*

5. Results and discussion

In this study, two New York City neighborhood clustering strategies were selected: 1) clustering of all neighborhoods at once, and 2) clustering of neighborhoods within boroughs. In both, two values of K (number of clusters) were chosen.

All neighborhoods

When taking all New York City neighborhoods into clustering with K=5, 271 out of 306 neighborhoods create one big cluster and other 28 neighborhoods create another cluster (Figure 2 left, Table 4). The remaining 3 clusters contain less than 5 neighborhoods. Clustering with K=5 is therefore not sufficient to segment the neighborhoods adequately - according to the results, most of the neighborhoods (more than 88%) would be similar.

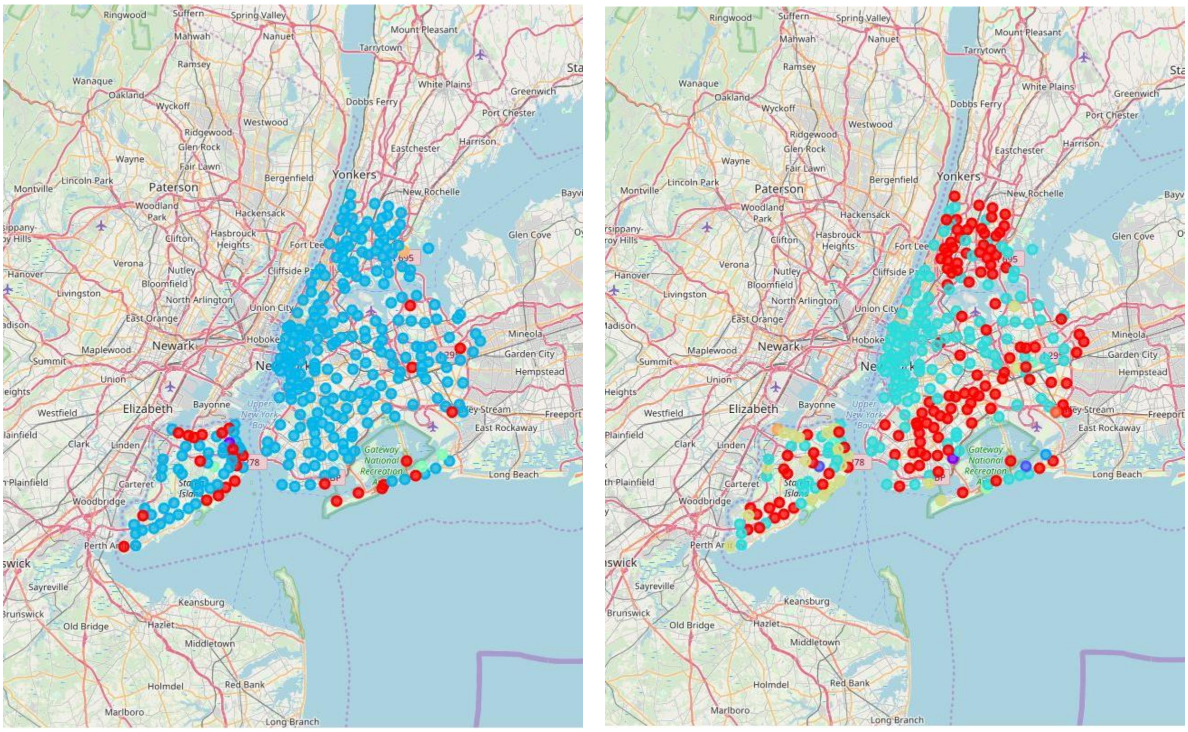


Figure 2 *Clusters of New York City neighborhoods with K=5 (left) and K=10 (right)*

K=5		**K=10**			
Cluster label	**Number of neighborhoods**	**Cluster label**	**Number of neighborhoods**	**Cluster label**	**Number of neighborhoods**
0	28	**0**	121	**5**	2
1	1	**1**	1	**6**	1
2	271	**2**	2	**7**	21
3	3	**3**	1	**8**	1
4	1	**4**	153	**9**	1

Table 4 *Numbers of New York City neighborhoods in different clusters with K=5 and K=10*

Clustering of all neighborhoods at once with K=10 results in more distinguished clusters (Figure 2 right, Table 4). Similarly to clustering with K=5, most of the neighborhoods (153 out of 306) form one big cluster that can be characterized by venues like Pizza Place, Deli/Bodega and Italian Restaurant. The second largest cluster consists of 121 neighborhoods and the most common venue is Pizza Place. The third largest cluster is formed only by 21 neighborhoods and the most abundant venues are Bus Stop and Deli/Bodega. The remaining clusters contain only 1 or 2 neighborhoods.

While the big clusters don't have any sport facility in the top common venue categories, the neighborhoods that don't belong to the large clusters can be described by venues like Yoga Studio, Pool or Park. This might suggest that these neighborhoods could be proper candidates for our business purpose. Because these neighborhoods don't form larger clusters, it's likely that there aren't any other neighborhoods where sport facilities "win". On the other hand, it doesn't necessarily mean that neighborhoods belonging to the large clusters don't have any sport facilities - they can just be overrun by restaurants, fast food places, coffee shops and other much more common venue categories.

It's obvious that most of the neighborhoods within every borough belong to the two largest clusters. One would expect that there would be a category that would cover at least half of the neighborhoods. Therefore, it can be concluded that the largest clusters are too general and are based on smaller contributions from many different venue categories. This leads to assumption that it would be more beneficial to perform clustering within each borough to reveal patterns in more detail.

In order to cluster neighborhoods within individual boroughs, number of clusters (K) was set to 5 and 8. Let's discuss the boroughs one by one.

Bronx (52 neighborhoods)

With K=5, 52 neighborhoods in Bronx were clustered into one big cluster with 45 members, and four one to three member clusters (Figure 3 left, Table 5). Venues like Pizza Place, Donut Shop and Deli/Bodega are typical for the big cluster. Some of the remaining neighborhoods have sport facilities such as Yoga Studio and Pool. In addition, restaurants are not a common venue category here.

The big cluster with 45 neighborhoods is broken down to smaller clusters when the target number of clusters is set to K=8 (Figure 3 right, Table 5). The three most abundant clusters contain 25, 10 and 7 members. However, the most common venues are venues like Pizza Place, Restaurant and Deli/Bodega. Closer look at neighborhoods that don't belong to these bigger clusters reveals that these ones have sport facilities. Examples of such neighborhoods are Country Club, Clason Point, Spuyten Duyvil. Increasing the number of clusters from 5 to 8 has helped to distinguish some of these neighborhoods.

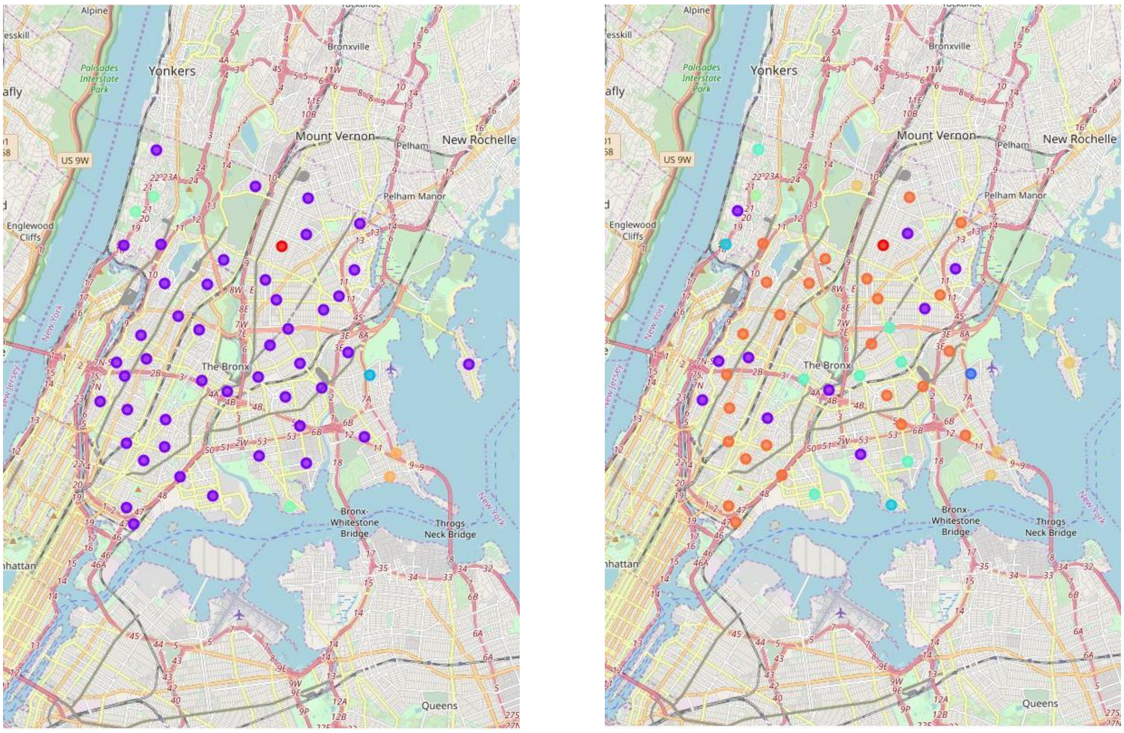


Figure 3 *Clusters of Bronx neighborhoods with K=5 (left) and K=8 (right)*

K=5		K=8			
Cluster label	**Number of neighborhoods**	**Cluster label**	**Number of neighborhoods**	**Cluster label**	**Number of neighborhoods**
0	1	**0**	1	**5**	1
1	45	**1**	10	**6**	5
2	1	**2**	1	**7**	25
3	3	**3**	2		
4	2	**4**	7		

Table 5 *Numbers of Bronx neighborhoods in different clusters with K=5 and K=8*

Brooklyn (70 neighborhoods)

Clustering of Brooklyn neighborhoods into 5 clusters leads to one big cluster with almost all neighborhoods (66) and four one-membered clusters (Figure 4 left, Table 6). In general, sport facilities are not very common in neighborhoods belonging to the largest cluster. The common venues mostly include Pizza Place, different types of Restaurants, and Coffee Shops. On the other hand, the neighborhoods not belonging to the large cluster would probably make good candidates for the shop with healthy food because of venues like Pool, Gym or Spa.

Using K=8 leads to two clusters with more than 20 members (Figure 4 right, Table 6). While sport facilities are not very common in neighborhoods in any of these clusters, the remaining neighborhoods have sport facilities in the top five most abundant venue categories. Similarly to case with K=5, there are neighborhoods not belonging to the above-mentioned abundant clusters that could make good candidates for the shop with healthy food because of sport and leisure time places. Examples of such neighborhoods would be: Sea Gate, Paerdegat Basin, Bergen Beach.

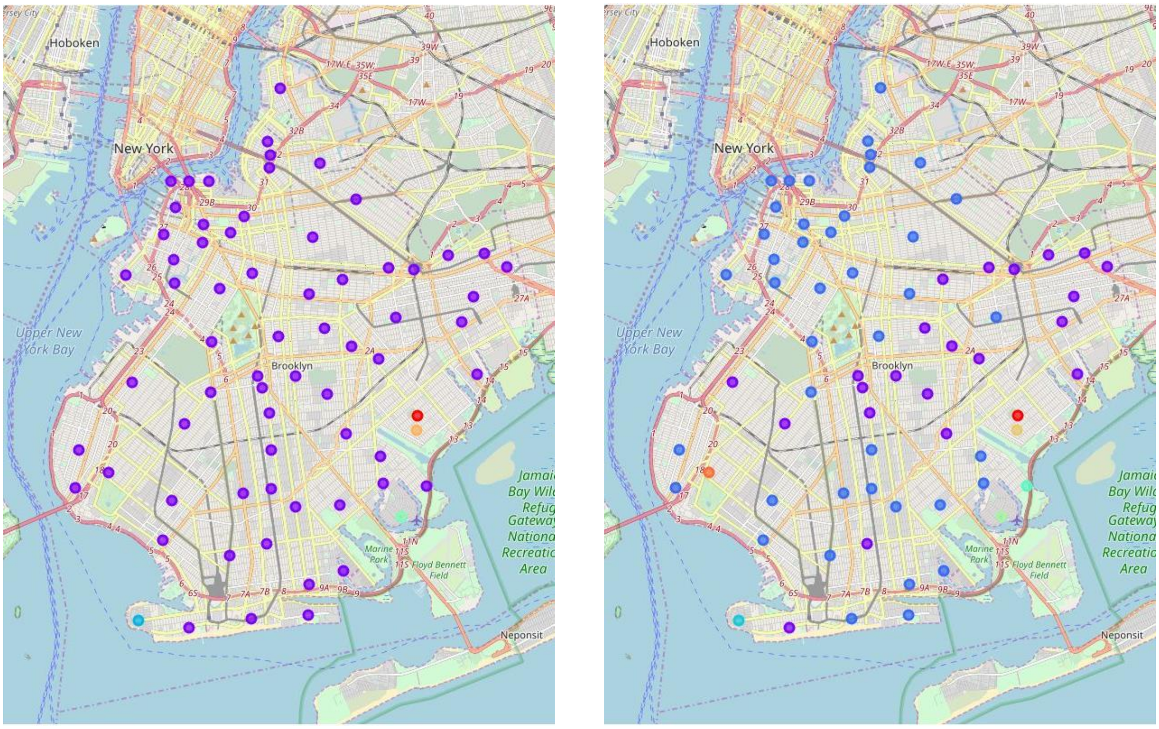


Figure 4 *Clusters of Brooklyn neighborhoods with K=5 (left) and K=8 (right)*

K=5		K=8			
Cluster label	**Number of neighborhoods**	**Cluster label**	**Number of neighborhoods**	**Cluster label**	**Number of neighborhoods**
0	1	**0**	1	**5**	1
1	66	**1**	22	**6**	1
2	1	**2**	42	**7**	1
3	1	**3**	1		
4	1	**4**	1		

Table 6 *Numbers of Brooklyn neighborhoods in different clusters with K=5 and K=8*

Manhattan (40 neighborhoods)

Because of lower number of neighborhoods in this borough, only clustering with K=5 was performed (Figure 5, Table 7). Comparing to other boroughs, Manhattan neighborhoods are much more equally distributed among the different clusters. The largest cluster contains 15 neighborhoods with venues such as Restaurant (Italian, American, Indian) and Café/Coffee Shop. Sport related venues are not very common. Similar observation applies for the second largest cluster with 14 neighborhoods. On the other hand, the third largest cluster (6 neighborhoods) could be described by venue categories like Coffee Shop, Sandwich Place, Park and Gym. Although Coffee Shop and Sandwich Place are above sport and leisure time places, the neighborhoods belonging to this cluster could be good candidates for our business purpose.

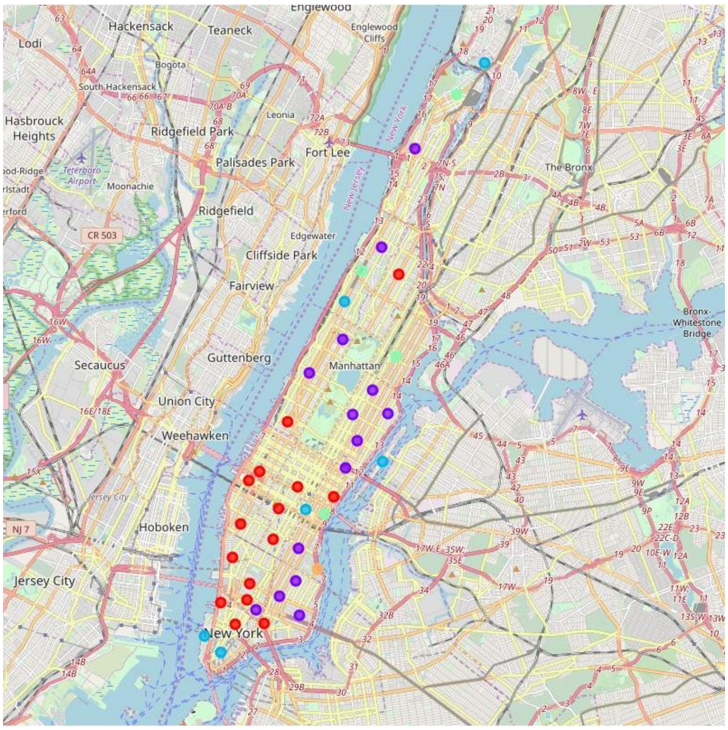


Figure 5 *Clusters of Manhattan neighborhoods with K=5*

K=5	
Cluster label	**Number of neighborhoods**
0	15
1	14
2	6
3	4
4	1

Table 7 *Numbers of Manhattan neighborhoods in different clusters with K=5*

Queens (81 neighborhoods)

Cluster analysis with K=8 doesn't look like improvement when compared to K=5 (Figure 6, Table 8). Therefore, we will look only at clustering with K=5. Almost all neighborhoods belong to two large clusters - neighborhoods have mostly non-sport facilities (Deli/Bodega, Pizza Place, Restaurants of different type, Bakery, Donut Shop). In contrast, neighborhoods outside these two clusters could be good candidates because of some sport and leisure time facilities.

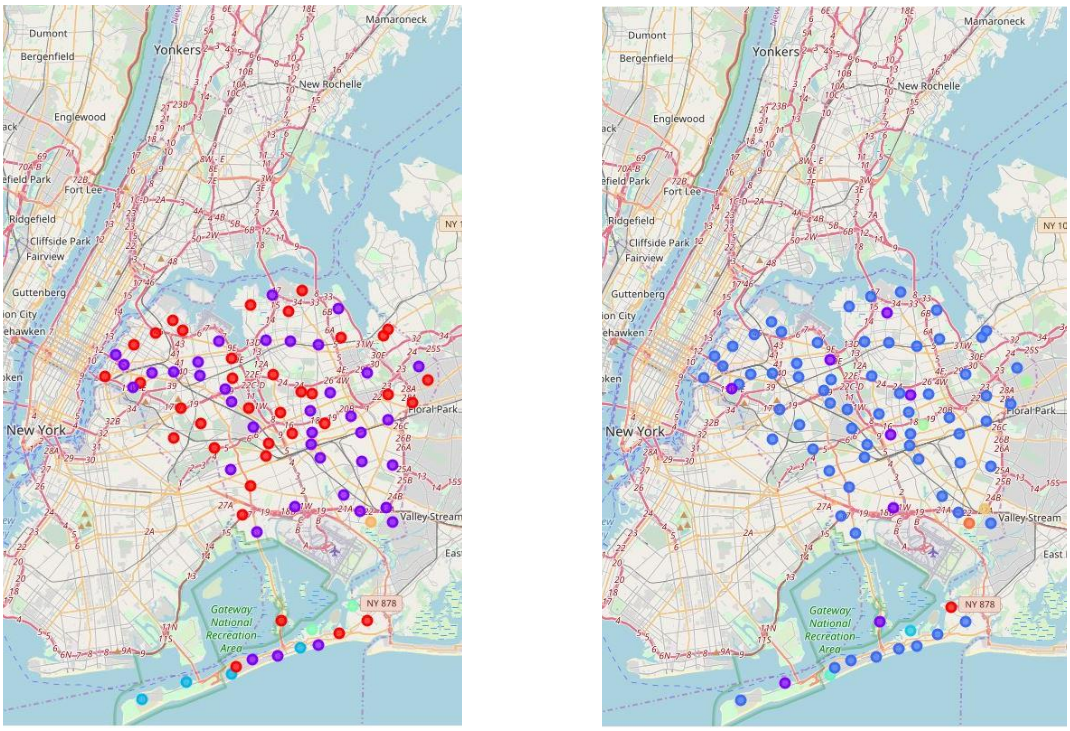


Figure 6 *Clusters of Queens neighborhoods with K=5 (left) and K=8 (right)*

K=5		K=8			
Cluster label	**Number of neighborhoods**	**Cluster label**	**Number of neighborhoods**	**Cluster label**	**Number of neighborhoods**
0	36	**0**	1	**5**	2
1	38	**1**	8	**6**	1
2	4	**2**	66	**7**	1
3	2	**3**	1		
4	1	**4**	1		

Table 8 *Numbers of Queens neighborhoods in different clusters with K=5 and K=8*

Staten Island (63 neighborhoods)

Similarly to Queens, clustering with K=8 is not a significant improvement when compared to clustering with K=5 (Figure 7, Table 9), and we will discuss only results for clustering with K=5. The two largest clusters contain 46 and 8 members. While neighborhoods from the biggest cluster with venues such as Italian Restaurant, Pizza Place and Deli/Bodega wouldn't make the best candidates for our business idea of a healthy food shop, neighborhoods belonging to the second largest cluster or other clusters with venues such as Gym and Yoga Studio would fit much better (examples of such neighborhoods: Grymes Hill, Park Hill).

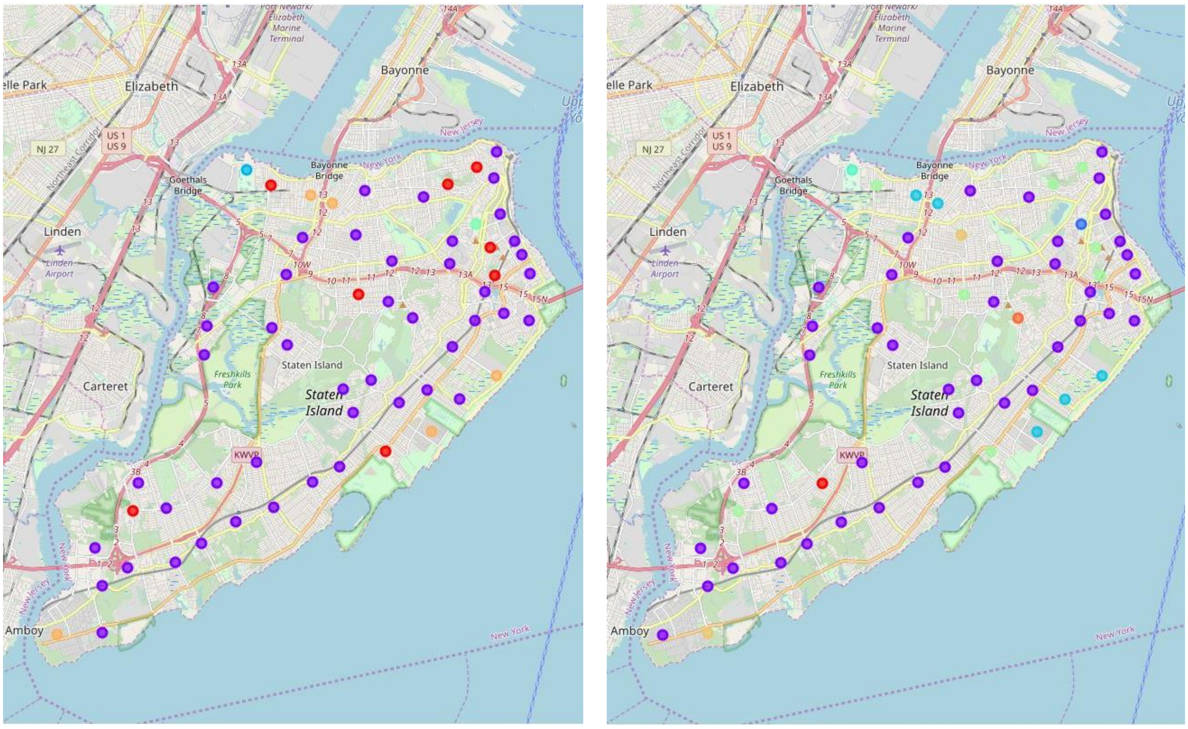


Figure 7 *Clusters of Staten Island neighborhoods with K=5 (left) and K=8 (right)*

K=5		K=8			
Cluster label	**Number of neighborhoods**	**Cluster label**	**Number of neighborhoods**	**Cluster label**	**Number of neighborhoods**
0	8	**0**	1	**5**	8
1	46	**1**	42	**6**	2
2	1	**2**	1	**7**	1
3	1	**3**	5		
4	5	**4**	1		

***Table 9** Numbers of Staten Island neighborhoods in different clusters with K=5 and K=8*

6. Conclusion

In this study, neighborhoods of New York City have been segmented and clustered in order to identify the best candidates for a business plan – opening a healthy food store in a new area. Neighborhoods were clustered based on the similarity and abundance of venues belonging to different categories using the standard K-Means clustering algorithm. Two approaches have been chosen: 1) clustering of all neighborhoods at the same time (with the target number of clusters K=5 and K=10), and 2) clustering of neighborhoods in every borough separately (with K=5 and K=8). The obtained clusters were examined in a detail to gain insights into the common features of the neighborhoods.

Using all neighborhoods for clustering (that is, irrespective of the boroughs) gives a high-level overview of the neighborhoods. With both K=5 and K=10, the most of neighborhoods form one big cluster that can be characterized by the most frequent venue categories. Although we can identify some neighborhoods with features in alignment with the business plan of starting a healthy food store, it might be possible that other suitable neighborhoods are simply hidden and couldn't be revealed due to prevailing venues that are not in our interest. On the other hand, further increase of K could help to distinguish the relevant neighborhoods.

Cluster analysis of neighborhoods within boroughs was performed with K=5 and K=8. In general, the largest clusters within each borough are similar – they mostly include venues like restaurants, pizza places, sandwich places, coffee shops/cafes. We didn't identify any bigger neighborhood cluster that could be defined as a sport/leisure time type of cluster. However, we have found some smaller clusters or individual neighborhoods that do not fall under the common category "Restaurant/Pizza/Coffee". These would be the appropriate candidates to open a healthy food store, based on the criteria defined above.

The study provides a basic understanding of neighborhood segmentation in New York City and aims to determine the best neighborhoods for the defined business plan. However, other sources of data and more detailed analyses would be required to gain better understanding of the problem. For example, it would be beneficial to use data on population density in the area or data including information about the character of the area (industrial, business, living). In addition to this, different algorithm or different selection of features could help in better understanding of the neighborhood similarities and segmentation.