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

Neighborhood exploration helps in decision-making process

Business problem

- What are the best candidate neighborhoods to open a store with healthy food?

Target audience

- Businessmen or contractors that would like to start a successful healthy food store in a new area

Assumptions

- People with an active lifestyle use facilities like gyms, pools, other sport facilities or parks
- Candidate neighborhoods shouldn't be rich in facilities like supermarkets and groceries
- Candidate neighborhoods shouldn’t have many restaurants of different kinds, fast food, pizza since these are indicators of social and cultural life, not sport activities
- Neighborhood clustering based on abundance of venues of different categories enables decisions whether the neighborhood is a good candidate or not

Data sources and pre-processing

Data sources

1. New York City neighborhood data (NYU Spatial Data Repository):
   - neighborhood name
   - borough name
   - neighborhood latitude
   - neighborhood longitude

2. Location data from Foursquare API:
   - venues and their categories

Data pre-processing

1. JSON files converted to pandas dataframes

2. Exploration of datasets to reveal:
   - potential issues in naming
   - invalid venue categories

Methods

Standard K-Means Clustering to cluster neighborhoods based on their similarities measured in terms of different venue categories and their abundance in a neighborhood:

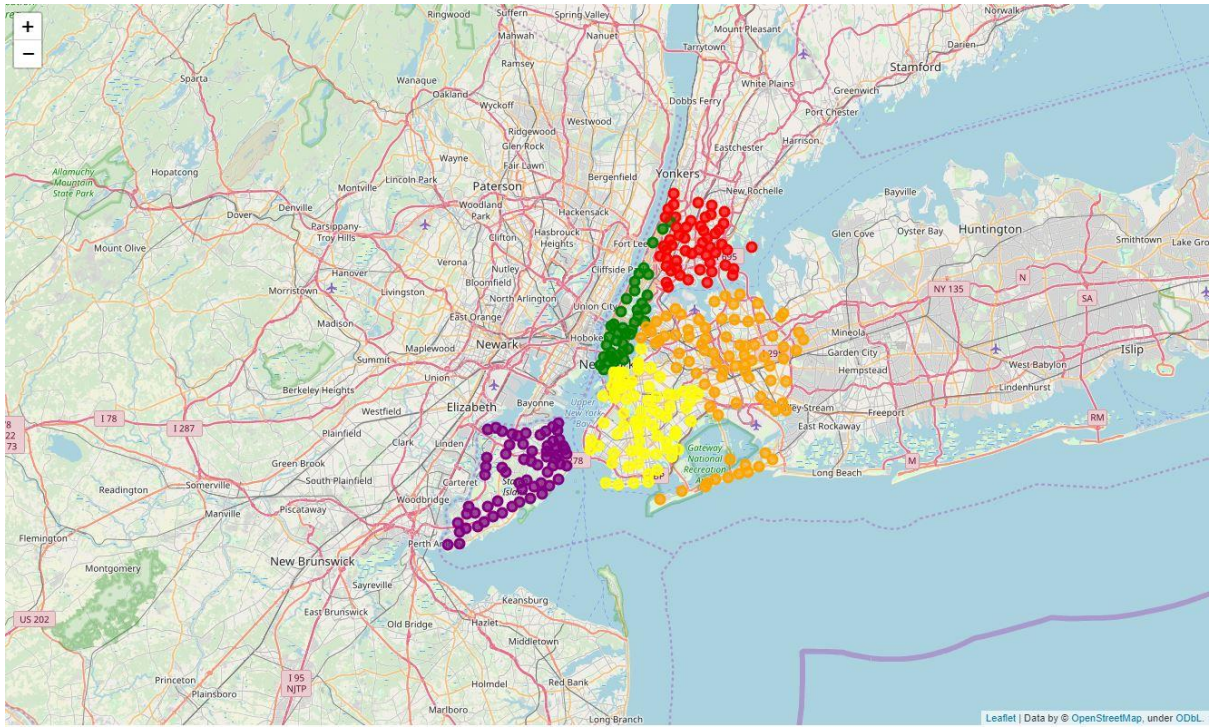
- cluster all neighborhoods in New York City with K=5 and K=10
- cluster neighborhoods within each borough with K=5 and K=8
- compare the results

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

Exploratory data analysis 1

New York city has 306 neighborhoods that belong to 5 boroughs



Exploratory data analysis 2

The overall number of venues: 10200

The number of different categories: 432

The most abundant categories (top 20):

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

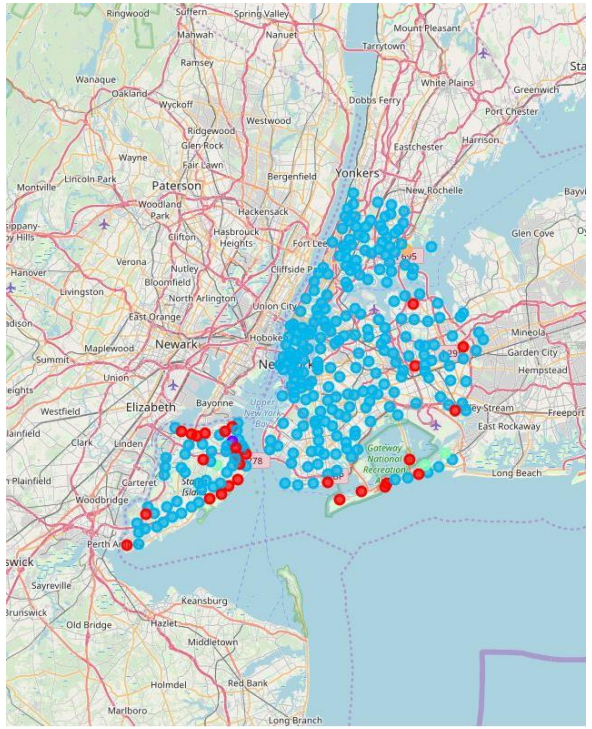
Clustering of NYC neighborhoods

Clustering with K=5:

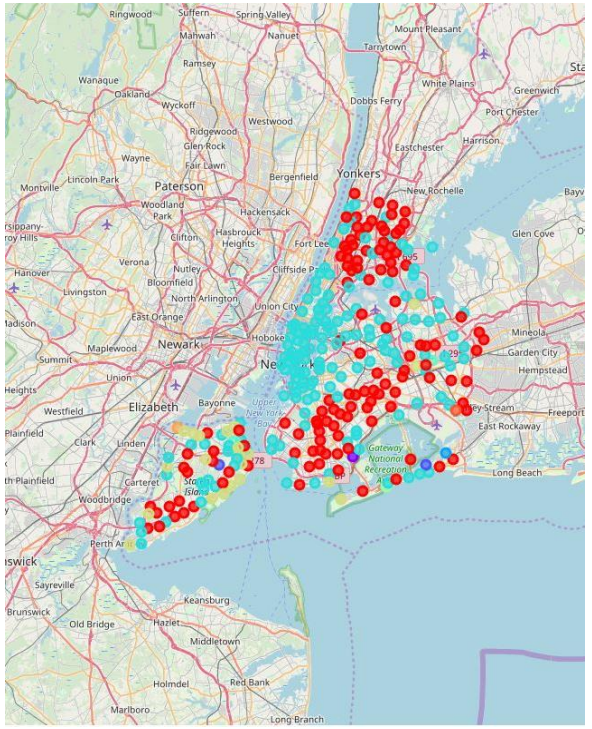
- not sufficient

Clustering with K=10:

- more distinguished clusters, still too general
- two large clusters with 153 and 121 members (typical venues: Pizza Place, Deli/Bodega and Italian Restaurant)
- smaller clusters (including one-member clusters) have sport facilities



K = 5



K = 10

Clustering within boroughs provides better segmentation of neighborhoods

Bigger clusters are similar across all boroughs:

- restaurants, pizza places, fast foods, coffee shops/cafes, corner shops
- sport and leisure-time facilities are not common
- pure sport cluster not discovered

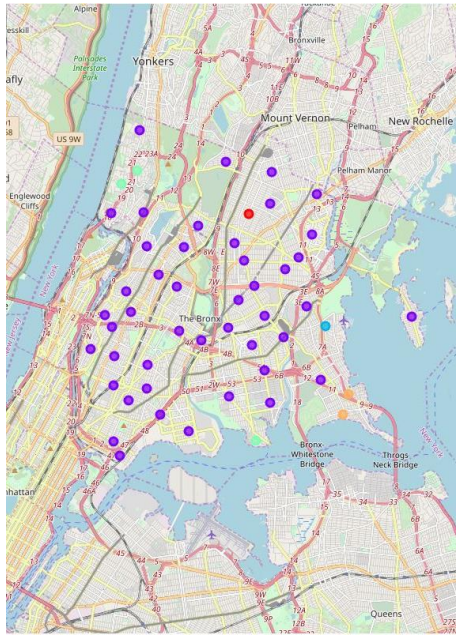
Small clusters or clusters containing only one or two neighborhoods:

- some have sport facilities (gym, yoga studio, pool, fitness)
- good candidates for a healthy food store come from all boroughs

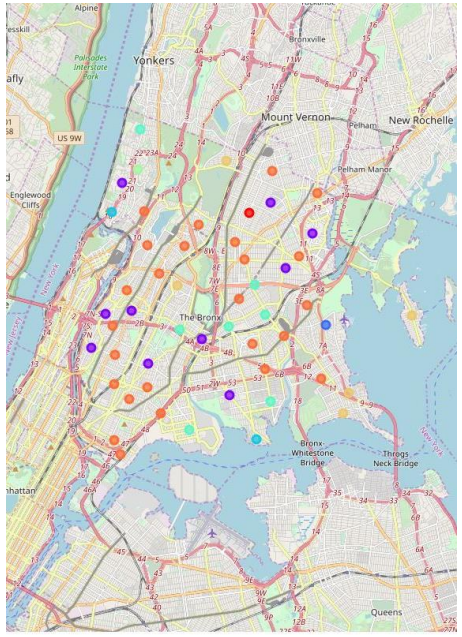
Clustering with K=5 vs. K=8
Bronx and Brooklyn

Clustering with K=8 reveals more details and helps in identification of proper candidate neighborhoods

Bronx (52 neighborhoods)

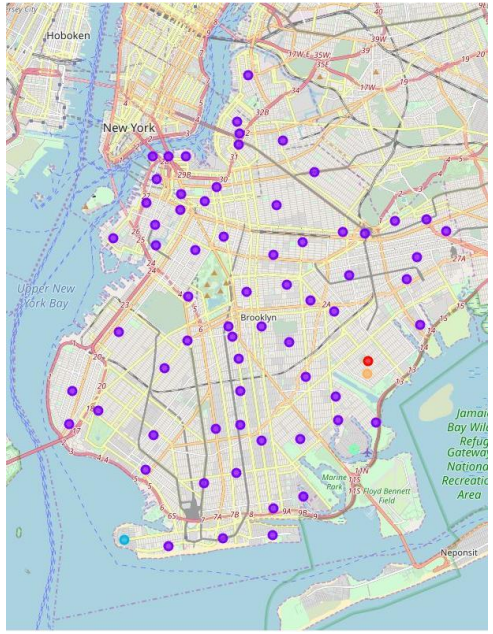


K=5

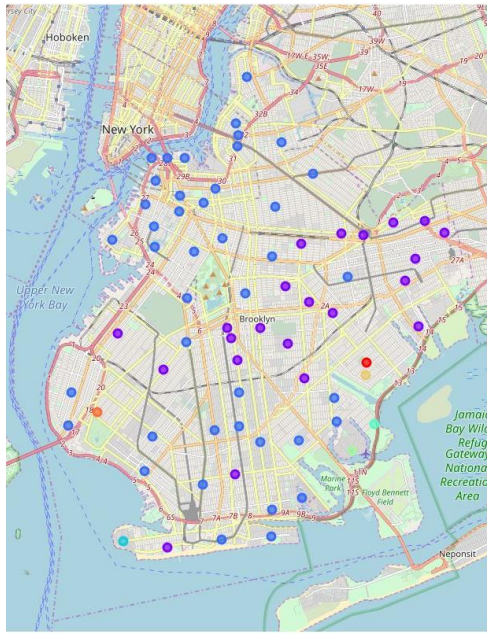


K=8

Brooklyn (70 neighborhoods)



K=5

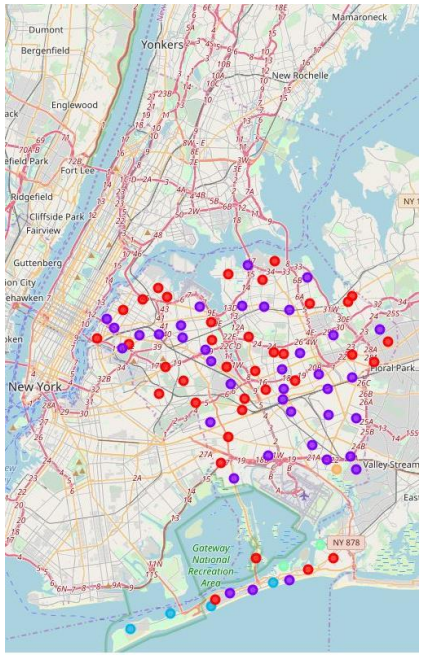


K=8

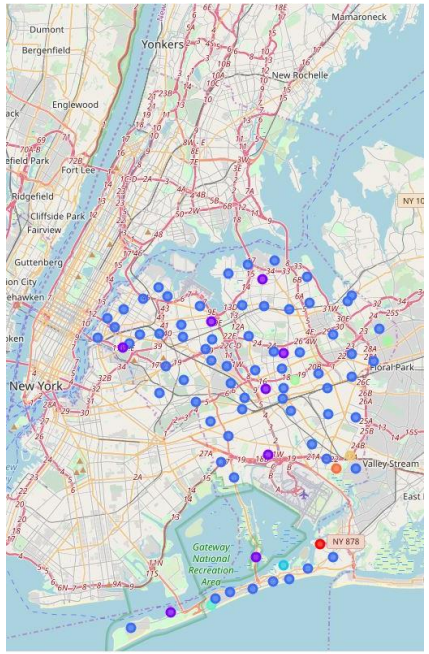
Clustering with K=5 vs. K=8
Queens and Staten Island

Clustering with K=8 doesn't bring much improvement

Queens (81 neighborhoods)



K=5



K=8

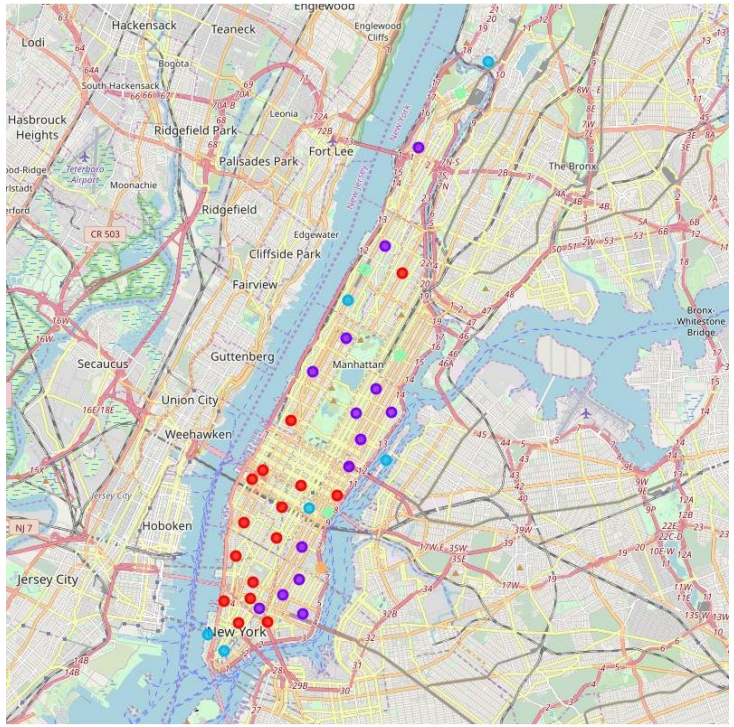
Staten Island (63 neighborhoods)

K=5

K=8

Clustering with K=5 Manhattan

Only clustering with K=5 performed due to small number of neighborhoods (40)



Conclusion

All neighborhoods for clustering (irrespective of the boroughs):

- high-level overview of the neighborhoods
- with both K=5 and K=10, the most of neighborhoods form one big cluster characterized by the most frequent venue categories in NYC
- higher K helps to reveal relevant candidates
- better to cluster within boroughs

Cluster analysis of neighborhoods within boroughs:

- the largest clusters within each borough are similar and include venues like restaurants, pizza places, sandwich places, coffee shops/cafes
- bigger neighborhood cluster of type sport/leisure time not identified
- identified smaller clusters or individual neighborhoods that do not fall under the common category "Restaurant/Pizza/Coffee" → appropriate candidates to open a healthy food store

Future recommendations

Other data sources:

- data on population density in the area
- data including information about the character of the area (industrial, business, living)

Try different algorithms

Algorithm tuning (parameters)