Capstone Project – Packaging free supermarket in NYC

Elisabeth A. Duijnstee

Table of contents:

1. Introduction

Business Problem & Background

There exist a huge plastic waste pile that grows every day, and primarily consists of single-use plastics. One of the main contributors to this problem is the food sector due to grocery packaging. Therefore, there is need for circular grocery shopping with recycable packaging in grocery stores.

Jarly, a circular, packaging-free grocery store is exploring the opportunity to open a venture in Manhatten, Brooklyn or Queens, New York. Those are the urban core of the New York metropolitan area, and the most densely populated of the five boroughs of New York City. It has been described as the cultural, financial, media and entertainment capital of the world. Outdoor dining used to be in the culture of the city, but COVID_19 has changed that. People tend to do more home-cooking nowadays and seek high-quality food which does not do harm to the environment. I will therefore provide an analytical view that will support decision making for the optimal supermarket location, considering the viable options discovered, and based on specific data and criteria discussed below.

2. Data Requirements

1. New York City

Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segement the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and logitude coordinates of each neighborhood.

This dataset exists for free on the web. Link to the dataset is: https://geo.nyu.edu/catalog/nyu_2451_34572

2. Food Stores

Foursquare API, which allows application developers to interact with the Foursquare Platform, will be used for finding supermarkets and farmers markets. Foursqueare is a social location service that allows users to explore the world around them. The API itself is a RESTful set of addresses to which you can send request and get responses. The API allows querying places and users, exploring popular places and checking out reviews and photographs of these places. New York City geographical coordinates data will be used as input for the Foursquare API, that will be leveraged to provision information on venues for each neighborhood in NYC.

To gather information of farmers markets and add this to the supermarket data frame, we will use:https://data.cityofnewyork.us/dataset/DOHMH-Farmers-Markets-and-Food-Boxes/8vwk-6iz2 and https://www.grownyc.org/greenmarketco/foodbox. A farmers' market is often defined as a public site used by two or more local or regional producers for the direct sale of farm products to consumers. In addition to fresh fruits and vegetables, markets may sell dairy products, fish, meat, baked goods, and other minimally processed foods.

After obtaining the neighborhoods data of New york City and the the supermarkets and farmers markets in each neighborhood using Foursquare API, data science techniques can be applied. One hot encoding will be used on the obtained dataset to find the 10 most common supermarkets in each neighborhood. The returned venues, are than clustered using k-means clustering. The optimal number of clusters can be obtained using silhouette score. The obtained clusters can be analyzed to find the major type of supermarket / food market in each cluster. Folium visualization library can be used to visualize the clusters superimposed on the map of New York City.

2. Population

For extra analysis on the population/geographics, the data from the following wikipedia can be used:

New York Population: https://en.wikipedia.org/wiki/New_York_City

New York City Demographics: https://en.wikipedia.org/wiki/Economy_of_New_York_City, https://en.wikipedia.org/wiki/Portal:New_York_City

There is a total of 5 boroughs and 305 neighborhoods. We narrow down and focus on the neighborhoods in Manhattan, Brooklyn and Queens. The following data are obtained from the Foursquare API for all three boroughs:

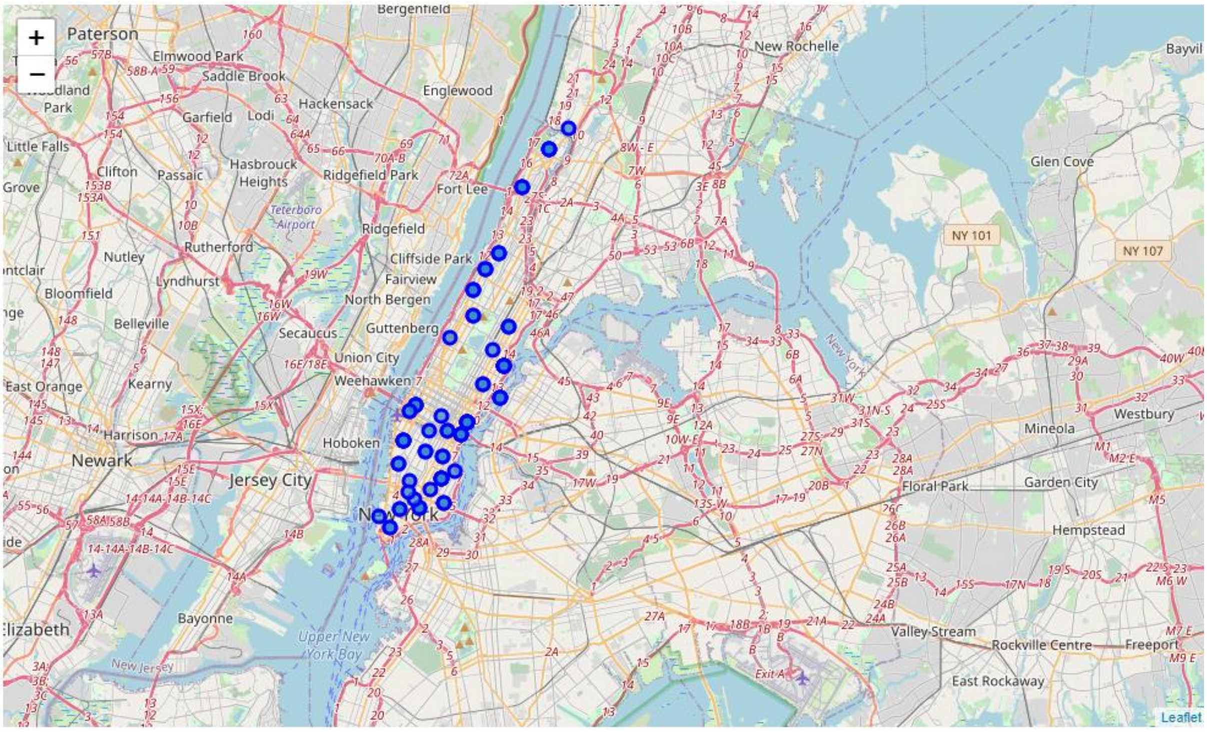
- Venue
- Venue Latitude
- Venue Longitude
- Venue Category data

3. Methodology

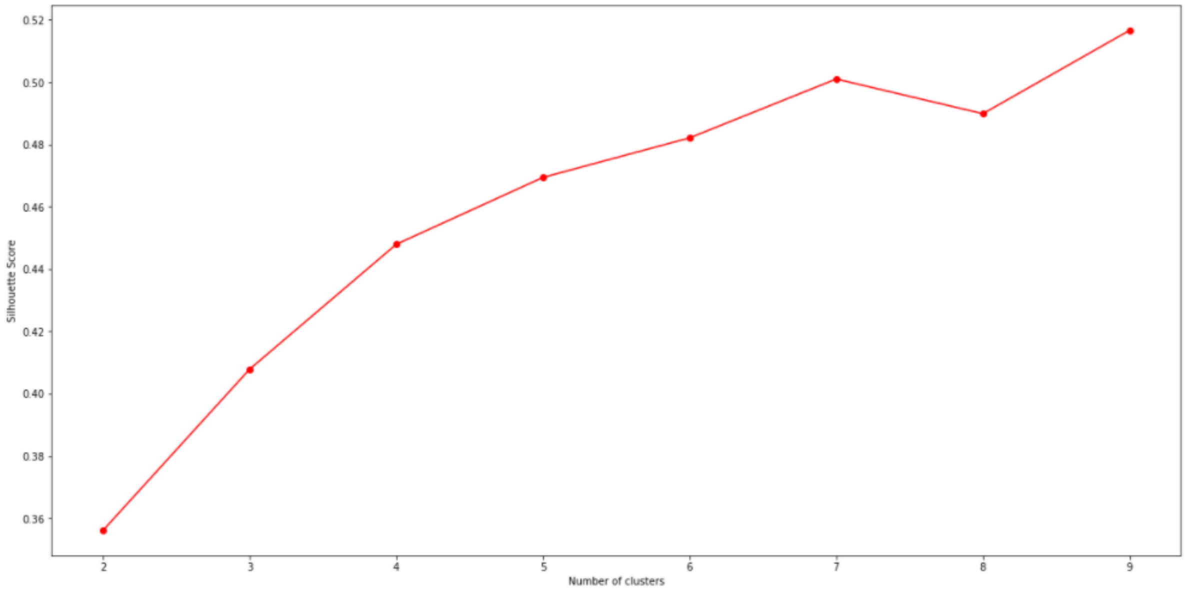
Now, we have the neighborhoods data of Manhattan, Brooklyn and Queens. Using Fousquare API we obtain all venues in each borough. We already find 3817 venues in Manhattan alone. We create a map of each borough and add markers for all venues. Because we want to open a supermarket, we look for supermarkets, organic markets, farmers markets, health food stores, convenience stores, deli/bodega's, fruit & vegetable stores, cheese shops, dairy stores, fish markets, butchers, and markets in all neighborhoods. We find 86 venues in Manhattan, 130 in Brooklyn and 151 in Queens. We will focus on these for our analysis.

We than perform one hot encoding on the obtained data sets and use it to find the 7 most common venue categories within our search in each neighborhood. We do this for the three boroughs and than merge the data frames.

As shown here, we plot all food related venues in manhatten (and do so for Brooklyn and queens too).



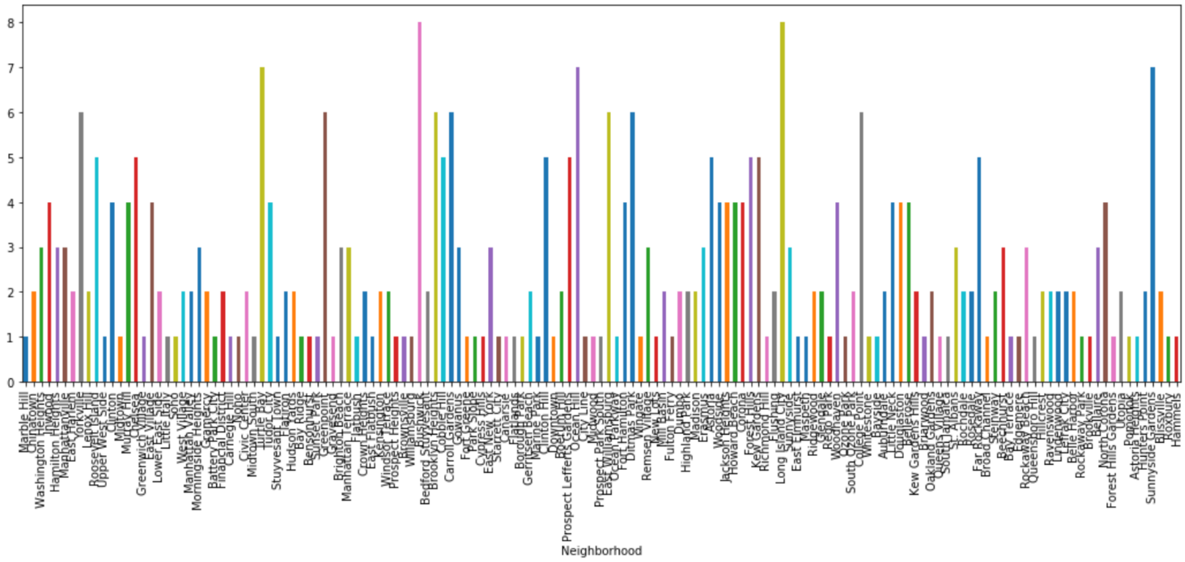
Than clustering can be performed on the dataset. Here k-nearest neighbor clustering technique is used. And the optimal number of clusters is found via silhouette score metric technique.



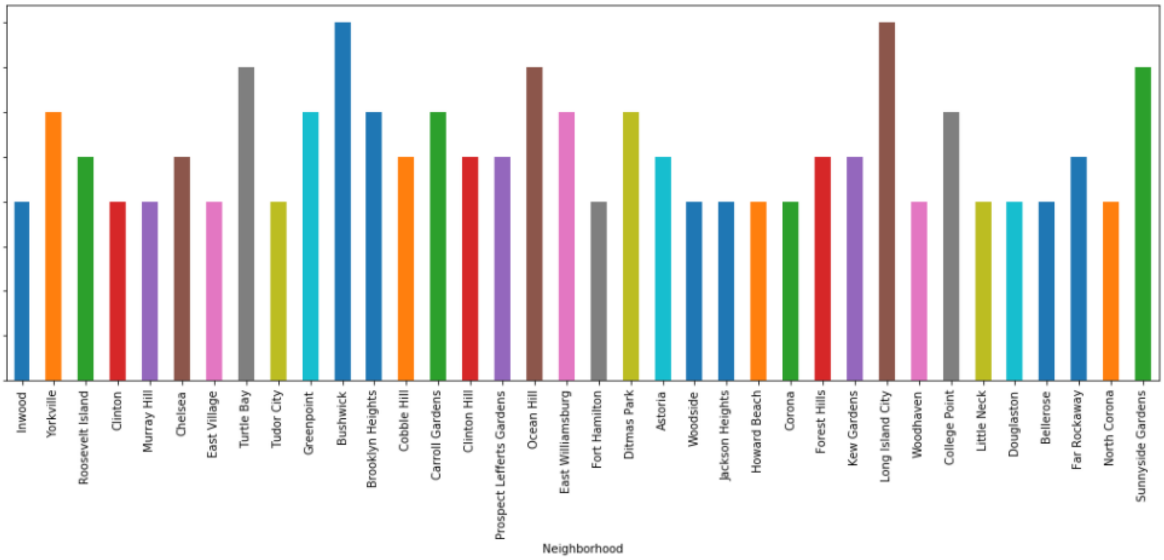
The clusters obtained can be analyzed to find the major type of venue category in each cluster. This data can be used to suggest suitable locations based on the category for our packaging free supermarket.

4. Analysis

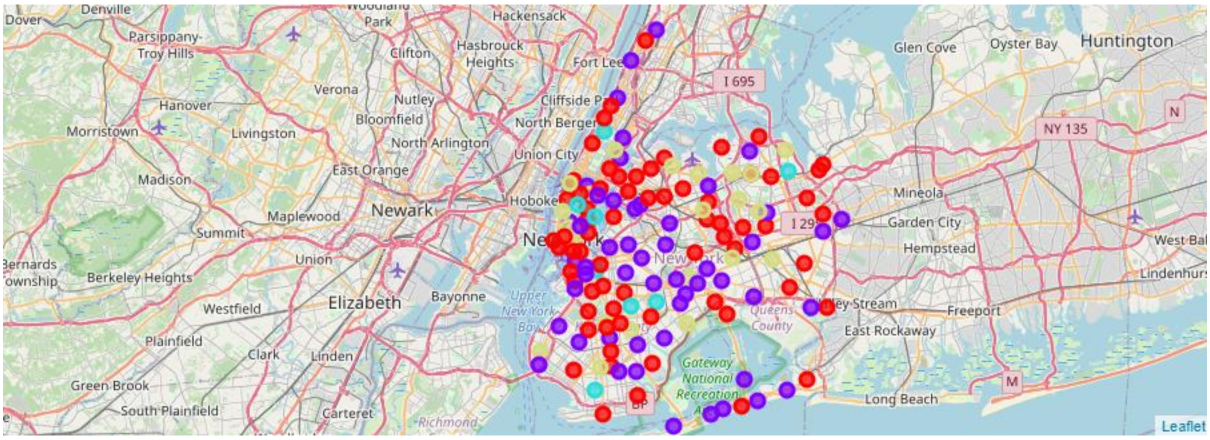
We plot the number of venues obtained in all the nieghborhoods in the 3 boroughs:



And narrow down to look at the neighborhoods with more than 3 vanues only:

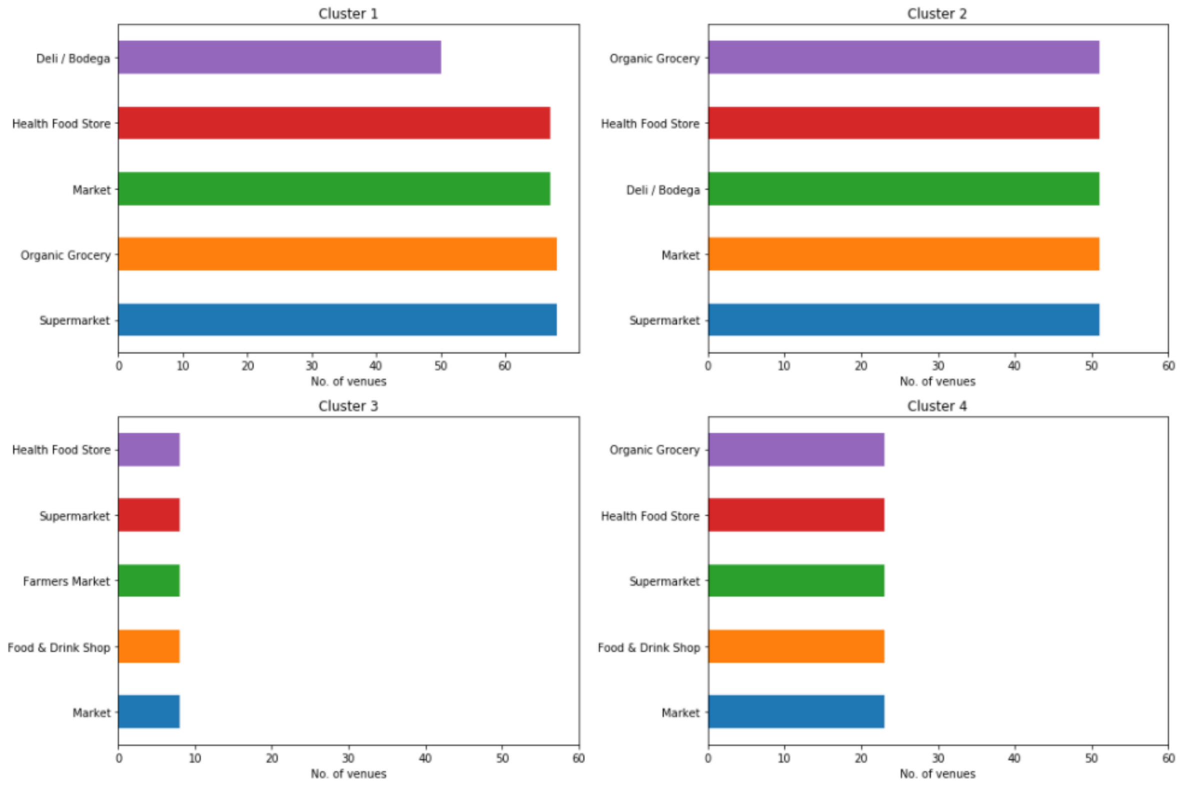


Next, we will perform one hot encoding on the filtered data to obtain the venue categories in each neighborhood. Then group the data by neighborhood and take the mean value of the frequency of occurrence of each category. This is used to obtain the top 7 most common venues in each neighborhood i.e. the 7 venues with the highest mean of frequency of occurrence. The resultant dataset can be used for the clustering algorithm. Here, the K-Nearest Neighbor (KNN) clustering algorithm is used. It is an unsupervised machine learning technique that clusters the given data into K number of clusters. For optimal result we need to select the best value for K. Here, the silhouette score is used to find the best value for K. A range of values from 2 to 9 was considered, KNN clustering was performed on the dataset and the silhouette score was calculated and plotted on a line plot. We use a value of K=4. This K value is used for the K-Means Clustering Technique. The K-Means labels obtained were included in the top neighborhoods dataset for examining the characteristics of each cluster.



5. Results & Discussion

We visualize the top 5 most common food venue categories in each of the clusters:



The neighborhoods in cluster 3 have the least number of food related venues. This could make it an attractive area to open our supermarket as it is less competitive. However, there is probably a reason for the little amount of venues in this cluster. Cluster 1 and 2 have the most food related venues, and we therefore think there is too much compeitition. Especially in cluster 2, as there is mostly organic markets there. Therefore we choose to locate the store in cluster 4 which has (1) the least competition in terms of total venues, but (2) interest in organic markets and healthy foods (given the current venues).

6. Conclusion

The pupose of this project is to analyze the neighborhoods of NYC and create a clustering model to suggest the best location for a packaging free circular grocery shop. The neighborhoods data was obtained from an online source and the Foursquare API was used to find the major venues in each neighborhood. We merged the three most popular boroughs, Manhattan, Brooklyn and Queens, and build a data science model. We selected our targets related to food markets and organic grocery shopping in all neighborhoods in the respective boroughs and used this as input to create a clustering model. The best number of clusters was obtained using the silhouette score. Each cluster was examined to find the most venue categories present, that defines the characteristics for that particular cluster.

A few examples for the applications that the clusters can be used for have also been discussed. A map showing the clusters have been provided. Both these can be used by stakeholders to decide the location for grocery store. Our suggestion would be cluster 4, as it has (1) the least competition in terms of total venues, but (2) interest in organic markets and healthy foods (given the current venues). A major drawback of this project was that the Foursquare API returned only few venues in each neighborhood. As a future improvement, better data sources can be used to obtain more venues in each neighborhood. This way the neighborhoods that were filtered out can be included in the clustering analysis to create a better decision model.