Searching food waste recycling business opportunities by segmenting and clustering neighborhoods in Brooklyn, New York

By

Atasi Bhattacharjee

1. Introduction

1.1 **Background:** US restaurants generate an estimated 22 to 33 billion pounds of food waste each year(1). Only five percent of food is composted in the US and as a result, uneaten food is the single largest component of municipal solid waste (2). In landfills, food component breaks down and forms methane, a greenhouse gas that is up to 86 times more toxic than carbon dioxide. So, along with reduce and reuse, organic food waste recycling is of utmost importance to uphold the objective of the "environmental sustainability" goals.

There are many uses of recycling organic food waste and so, various opportunities lie in recycling business. Green waste disposal can be converted into compost, a natural source of fertilizer (plant nutrient) for agriculture. The biogas released from decomposed organic food can be collected and used as alternative source of energy to fossil fuels. Moreover, waste cooking oil recycling is also one of the easiest among recycling businesses. Restaurants produce a vast amount of waste cooking oil that if not recycled end up in our waterways where it creates an environmental hazard. Waste cooking oil can be reprocessed into a number of secondary products such as biofuels, animal feed products, detergents or soap, paints and industrial lubricants.

1.2 **Business Problem:** To exploit the potential of recycling business opportunities, entrepreneurs need to locate an economically feasible and effective area for collection of the waste materials i.e waste food and/or cooking oil. To discover the possible zones where food waste production is high such as restaurants, fast food shops, groceries; I am using the knowledge of machine learning. Clustering the probable waste generators and segmenting them into zones; I will have a better understanding of my ideal neighborhoods.

1.3 **Target Audience:** Food waste recycling startups that are looking for a suitable location and customers to whom they can provide their recycling services.

2. Data

Since Brooklyn is the most populous borough of New York city and the second-most densely populated county in the United States, I decided to explore the neighborhood of Brooklyn in my study. 2 sets of data have been used for this project.

- New York city has a total of 5 boroughs and 306 neighborhoods. New York neighborhoods data along with the latitude and longitude coordinates of each neighborhood can be found here.
- **Foursquare API** (3) where API is a RESTful set of addresses and in Foursquare platform one can send requests and use their data online without downloading anything into own server. With the help of Foursquare location data, I will be able to explore the whole neighborhood and the diversity of venues such as those that are essential target customer for me, restaurant, grocery stores etc.

3. Methodology

3.1 Exploring Data

First step for any analysis is to prepare the data. For that I downloaded the New York neighborhood data from the above-mentioned site. All the relevant data came in the *features* key, which is basically a list of the neighborhoods including latitude and longitude. After extracting the relevant information from the web page, I stored the result in a *pandas* dataframe. The dataset has all 5 boroughs and 306 neighborhoods. To get exact location of each neighborhood I used *geopy* library to get the latitude and longitude values of New York City. Then with the help of *Folium* library, I was able to visualize all the 306 neighborhoods of New York city.

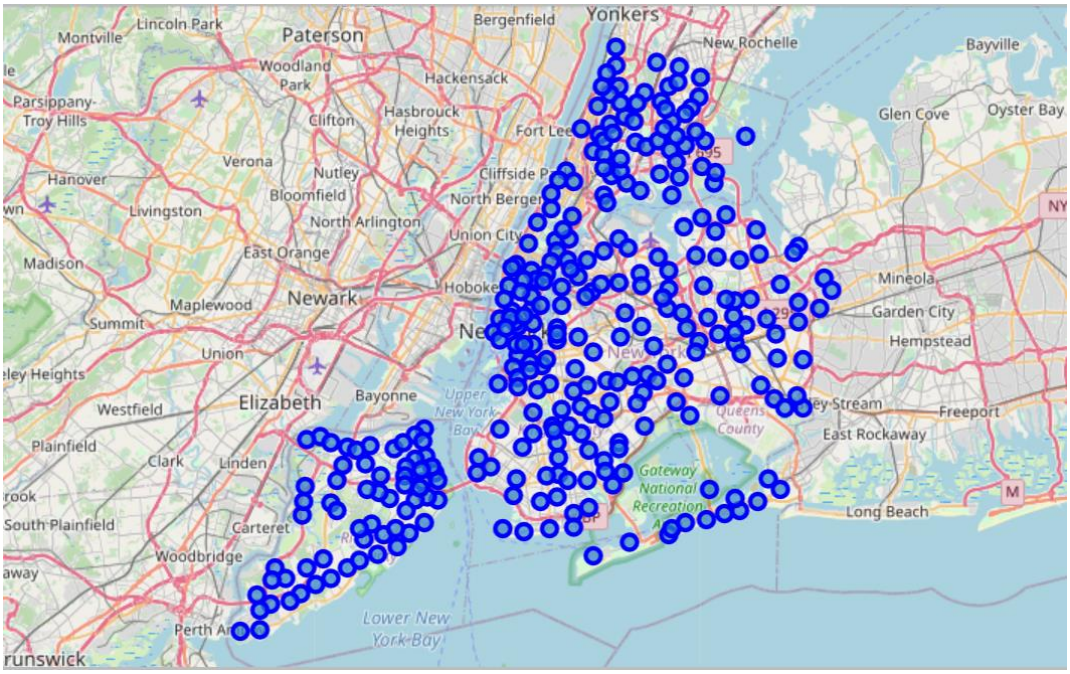

Figure 1 306 neighborhoods of New York city

However, the aim of this study is to explore food stores in Brooklyn borough. So, for illustration purposes, I simplified the above map and brought it down to show the neighborhoods in Brooklyn only. The geographical coordinate of Brooklyn is 40.6501038, -73.9495823 and Brooklyn has 66 neighborhoods.

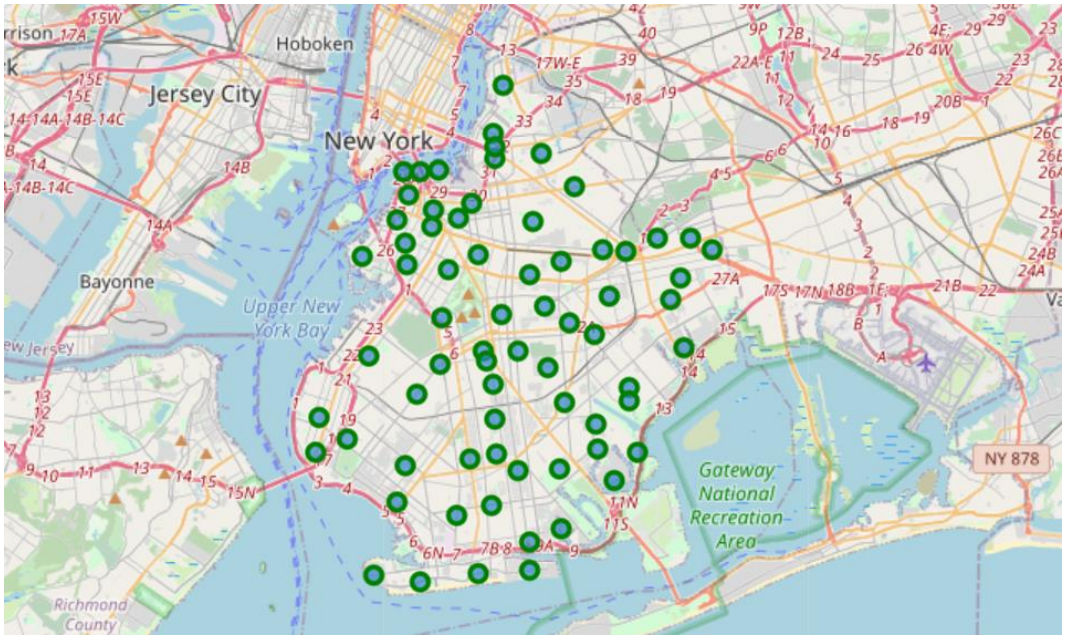

Figure 2 66 neighborhoods of Brooklyn

To get the venues by Foursquare API, I had to define Foursquare credentials with Client ID and Secret. Using 'explore' call with the coordinate of each neighborhood, and the filtered columns as venue: name, category, latitude and longitude 288 unique categories were returned. I targeted the limit of 100venus within 500 radius. When the API call is made, we receive the data as JSON (JavaScript Object Notation) object.

Table 2 shows the head of top 10 common venues of each neighborhood that is returned by Foursquare. It is noticeable that among the top 10, food stores such as restaurants, fast foods, tea/coffee shops, bakeries and groceries were more dominating.

Table 1 Head of top 10 venues of each neighborhood

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bath Beach	Pharmacy	Pizza Place	Chinese Restaurant	Italian Restaurant	Bubble Tea Shop	Fast Food Restaurant	Sushi Restaurant	Cantonese Restaurant	Gas Station	Women's Store
1	Bay Ridge	Spa	Italian Restaurant	Pizza Place	Pharmacy	Bagel Shop	American Restaurant	Greek Restaurant	Grocery Store	Bar	Sandwich Place
2	Bedford Stuyvesant	Coffee Shop	Pizza Place	Bar	Café	Japanese Restaurant	Gift Shop	New American Restaurant	Gourmet Shop	Juice Bar	Bagel Shop
3	Bensonhurst	Chinese Restaurant	Pizza Place	Donut Shop	Italian Restaurant	Sushi Restaurant	Dessert Shop	Ice Cream Shop	Cha Chaan Teng	Liquor Store	Bakery
4	Bergen Beach	Harbor / Marina	Baseball Field	Donut Shop	Playground	Hockey Field	Athletics & Sports	Event Space	Factory	Falafel Restaurant	Farm

3.2 Clustering Neighborhoods

I used Kmeans clustering to cluster the neighborhoods. K-means is a simple unsupervised machine learning algorithm which is used for unlabeled data in which each observation belongs to the cluster with the nearest mean, acting as a sample of the cluster. Kmeans groups the data into number (k) of clusters and to get the optimum K, “elbow” method was performed.

I iterated the values of k within 1 to 15 to get the optimum k at the “elbow” i.e the point after which the distortion/inertia started to decrease in flatter fashion. Figure 3 shows different Ks of “elbow” method.

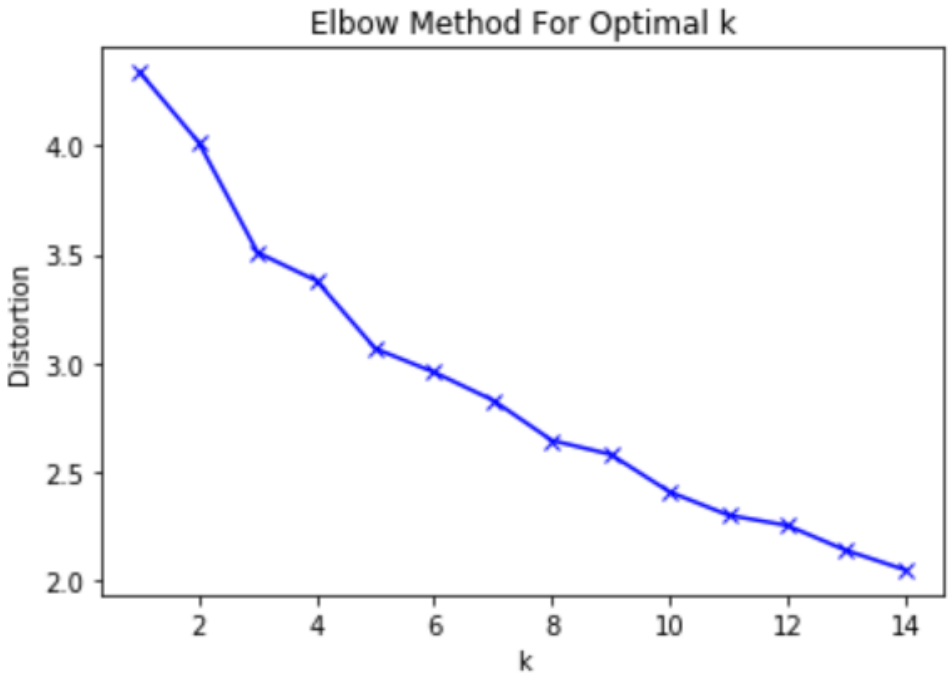

Figure 3 Elbow Method for Optimal K

For this given data, the optimal number of clusters for the data was checked with k= **5 and 7**. The map with 5 and 7 clusters are shown in Figure 4 and 5.

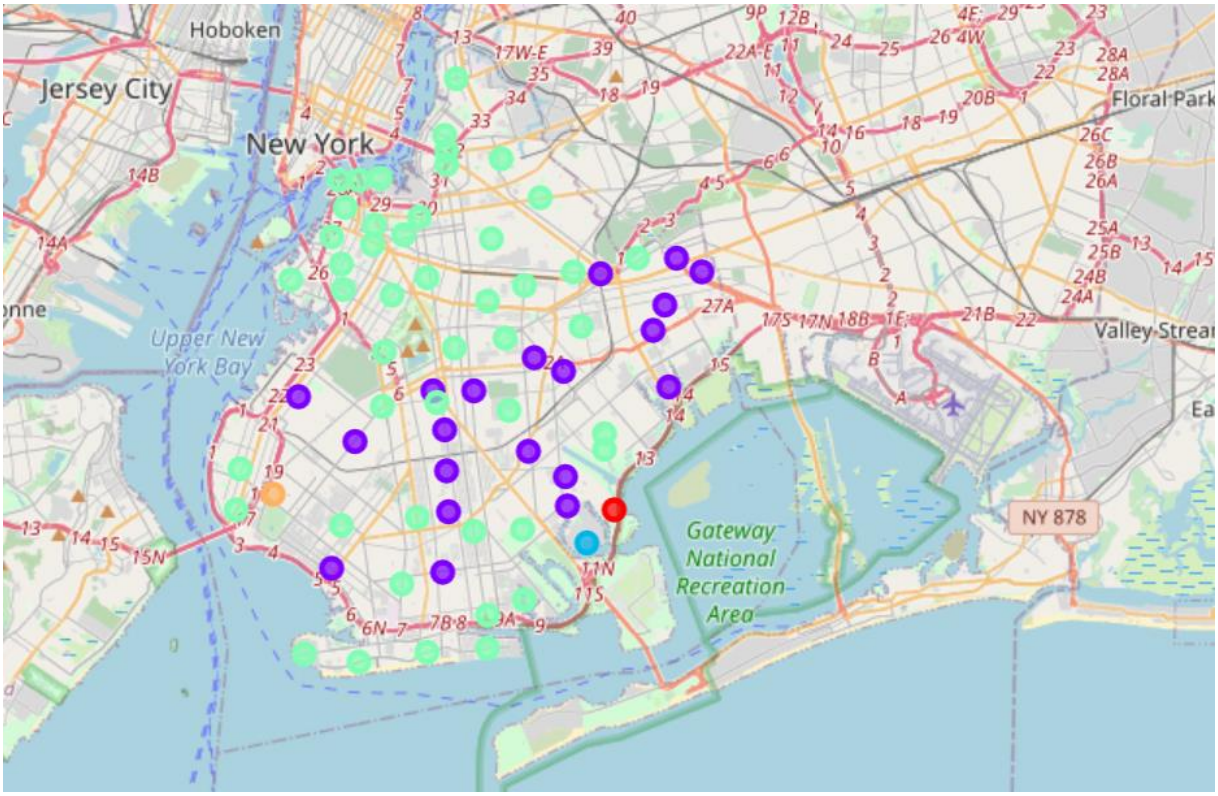

Figure 4 Clusters of venues in Brooklyn(k=5)

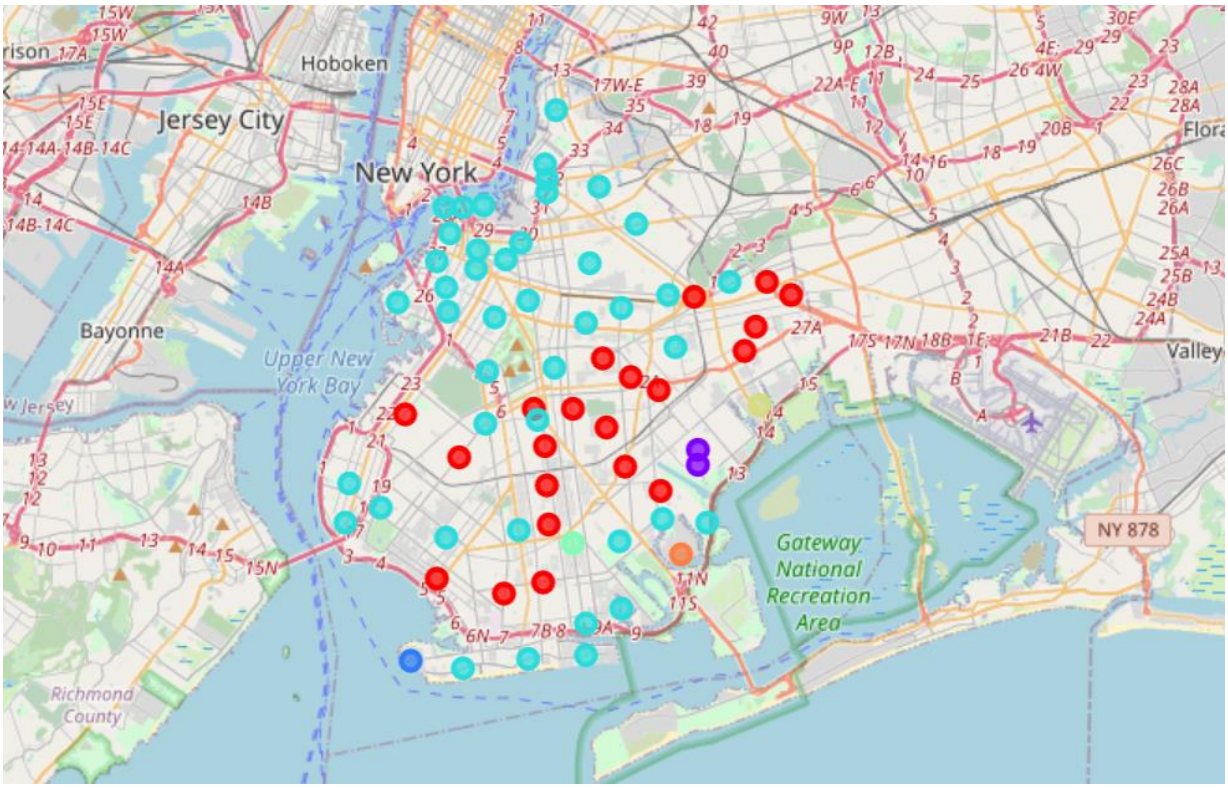

Figure 5 Clusters of venues in Brooklyn(k=7)

4 Results

Since there was no sharp elbow in this data, 2 values of K were examined. k=7 did not show any significant change in the clusters in terms of food category. So, K=5 has been taken as optimal number of clusters to examine the neighborhood with. Clustering on the basis of venue actually spreads food categories all over Brooklyn. However, important consideration could be the frequency of each category.

Here we can characterize the clusters by looking at venue categories in figure 5:

- Cluster 2 (purple) has consistently high scores for all venue categories. This is the heart of the city and so vibrant with restaurants, fast food and coffee shops.
- Cluster 4 (Paste) is also dominated by food store including fish market and farmer markets
- Cluster 1, 3 and 5 are single neighborhood clusters. Even though they have restaurants and grocery stores but in a very limited frequency.

5 Conclusion

Finally, the neighborhoods led with food stores were located in clusters 2 and 4. In this project, the main objective was to find areas where organic food waste produces in bulk. To meet this criterion; restaurants, fast food stores, grocery stores, bakery shops and even farmer's markets are all considered as potential customers. By clustering and segmenting one of the busiest counties in New York, I identified Brooklyn as a lucrative area to start food waste recycling business. Certainly, a more in-depth study needs to be done to get clusters according to the required venues which is limited with Foursquare API.

This project is useful for stakeholders who want to trace and visualize regions where potential customers are extended. This methodology of segmenting and clustering can be performed in any city in the USA. However, there are more market studies need to be performed for any new start-up.

1. Gunders, Dana. "Wasted: How America is Losing Up to 40 Percent of Its Food from Farm to Fork to Landfill." *Natural Resources Defense Council,* 2017. Retrieved March 7, 2019, from https://www.nrdc.org/sites/default/files/wasted-2017-report.pdf

2.US Environmental Protection Agency. "Sustainable Management of Food Basics." *EPA,* 2017. Retrieved March 7, 2019, from https://www.epa.gov/sustainable-management-food/sustainable-management-food-basics

3. Foursquare. https://foursquare.com/