Determining Candidate Neighborhoods for Health Food Store Expansion Using Cluster Analysis

Adam Horin
September 22, 2019

1. Introduction

1.1 Background
Sunac Natural Food is a health food business that offers organic, health-conscious foods to their customers [1]. With a rise in health and wellness awareness, it is not surprising health-related businesses are interested in expanding. This project will suggest candidate neighborhoods for Sunac Natural Food to expand to a new location in a Manhattan neighborhood. One factor that could contribute to health food markets' success in their current locations could be the similarities in surrounding venues. If they are looking to expand, they should likely look into neighborhoods that share these similarities. This project will use a k-means clustering method to cluster neighborhoods based on similarities in venues. Based on the outcomes of this analysis a suggestion can be made for candidate neighborhoods to look into for expansion locations.

1.2 Problem
Location expansion for a business has certain risks. There are many factors that may influence the success of a business in various locations, including the types of businesses that are also in the neighborhood. If Sunac Natural Food would like to expand to a new location, similarities in their current locations with other candidate neighborhoods should be considered. This can be accomplished using a k-means clustering analysis, a method of unsupervised machine learning that will cluster neighborhoods based on similarities of other businesses in their locations.

1.3 Interest
Business expansion is a logical next step in the growth of a company. It can be difficult to identify factors that influence the success of a business in its current locations. The analysis method used in this report allows unsupervised machine learning to be used to identify these factors and cluster neighborhoods based on these similarities. This method of analysis could be applied to many businesses that are looking to expand their business and would like to determine candidate neighborhoods for new locations.

2. Data

2.1 Data sources
Data for this analysis was obtained from the following locations. Data from the NYU repository of geospatial data for the neighborhoods and boroughs of New York City, NY, was obtained to identify neighborhoods and their locations in Manhattan [2]. Data from the Foursquare API on venues in the neighborhoods and boroughs of New York City, NY, was obtained in order to determine businesses and their locations in Manhattan [3].

2.2 Data cleaning

First, features were extracted from the data from the NYU repository. The features of interest from this dataset included borough, neighborhood, latitude, and longitude. These features allow us to determine the geospatial location of all neighborhoods in New York City. Next, the data was filtered to only include neighborhoods in the borough of Manhattan. This was to narrow the search of candidate neighborhoods to the borough of interest.

Data from the Foursquare API repository was then obtained. Data on the venues in each neighborhood in Manhattan were extracted. The features of interest from this dataset included venue and venue category.

After obtaining the venues and venue categories for each neighborhood in Manhattan, the data needed to be organized. In order to determine the frequency of different venue categories in each neighborhood, the data needed to be one hot encoded. This method allows us to compare non-ordered categorical variables for further analysis. After frequencies were determined, the top ten venue categories for each neighborhood were determined. This data was then used for the k-means clustering analysis.

After the k-means cluster analysis was performed (described below), cluster labels were then added to the data. Then based on the neighborhoods that Sunac Natural Food's current locations are located, the cluster of candidate neighborhoods could be determined. The clusters of neighborhoods could then be visualized on a map.

3. Methods

3.1 K-means cluster analysis

To cluster the neighborhoods based on similarities in venue categories, a k-means cluster analysis was conducted. This is a method of unsupervised machine learning that clusters the data into undefined groups based on their similarities in features. To determine the appropriate number of clusters, a range of k-means cluster analyses were run with 1-8 clusters, and the distortion for each analysis was calculated. The number of clusters versus their distortion was plotted and the optimal k number of clusters was determined using the elbow method.

First data from the NYU repository was loaded into a dataframe using the **pandas** library. The data was filtered to only include neighborhoods in the borough Manhattan (Table 1.)

	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

Table 1. Dataframe of New York City geospatial data, by borough and neighborhood.

A map representing these neighborhoods was then generated using the **folium** library (Figure 1). The Foursquare API was used to obtain data about the venues in each of these neighborhoods in Manhattan. A function was written that looped through each neighborhood in the NYU repository data to extract venue information. The data was then one hot encoded in order to analyze the frequency of the venue categories. The venue categories data are non-ordered categorical data and therefore need to be coded using one hot encoding to calculate the frequencies. Based on the average frequency of each venue category for each neighborhood, the top ten most common venues were determined for each neighborhood (Table 2). This data was then used for the k-means cluster analysis.

Figure 1. A map identifying the neighborhoods of Manhattan.

	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	Battery Park City	Park	Coffee Shop	Hotel	Memorial Site	Gym	Italian Restaurant	Wine Shop	Clothing Store	Plaza	Playground
1	Carnegie Hill	Coffee Shop	Pizza Place	Cosmetics Shop	Café	Japanese Restaurant	Grocery Store	French Restaurant	Yoga Studio	Bookstore	Bar
2	Central Harlem	African Restaurant	Chinese Restaurant	French Restaurant	Gym / Fitness Center	American Restaurant	Bar	Public Art	Seafood Restaurant	Fried Chicken Joint	Southern / Soul Food Restaurant
3	Chelsea	Coffee Shop	Italian Restaurant	Bakery	Ice Cream Shop	Nightclub	Hotel	Theater	American Restaurant	Seafood Restaurant	Tapas Restaurant
4	Chinatown	Chinese Restaurant	Cocktail Bar	Vietnamese Restaurant	Bakery	American Restaurant	Spa	Salon / Barbershop	Bubble Tea Shop	Korean Restaurant	Asian Restaurant

Table 2 Dataframe of the ten most common venues for each neighborhood in Manhattan.

4. Results

For the k-means cluster analysis, separate models were built ranging from 1-8 clusters, and model distortion, the sum of squared distances between each observation vector and their cluster centroid, was calculated for each. The number of clusters versus their model distortion was plotted and the optimal number of clusters was determined using the elbow method. The bend was identified, determining 6 was the optimal number of clusters for the model (Figure 2).

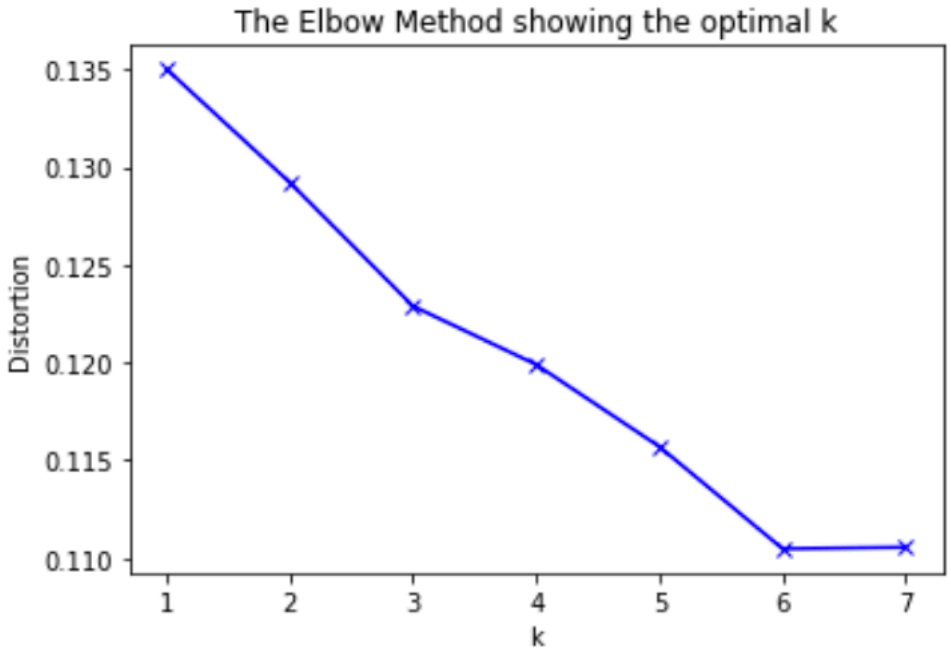

Figure 2 Plot of the model clusters their distortions to determine the optimal number of clusters.

The k-means cluster analysis was then run using 6 clusters. This model divided the Manhattan neighborhoods into 6 clusters based on similarities of their venues. The following table represents 'Cluster 0', including each neighborhood and their three most common venues (Table 3). This cluster contains the neighborhoods of the current Sunac Natural Food store locations (Table 4). It is also worth noting that many of the neighborhoods in this cluster contain a Gym/Fitness Center as one of their top three most common venues. These neighborhood clusters were then color coded and plotted on a map to represent the locations of these clusters.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
6	Central Harlem	African Restaurant	Gym / Fitness Center	French Restaurant
8	Upper East Side	Italian Restaurant	Exhibit	Art Gallery
13	Lincoln Square	Theater	Gym / Fitness Center	Café
14	Clinton	Theater	Gym / Fitness Center	American Restaurant
15	Midtown	Hotel	Coffee Shop	Cocktail Bar
16	Murray Hill	Coffee Shop	Hotel	Japanese Restaurant
17	Chelsea	Coffee Shop	Italian Restaurant	Bakery
18	Greenwich Village	Italian Restaurant	Clothing Store	Sushi Restaurant
21	Tribeca	Italian Restaurant	American Restaurant	Park
22	Little Italy	Bakery	Café	Ice Cream Shop
23	Soho	Clothing Store	Boutique	Art Gallery
24	West Village	Italian Restaurant	Cosmetics Shop	New American Restaurant
28	Battery Park City	Park	Coffee Shop	Hotel
29	Financial District	Coffee Shop	Pizza Place	Wine Shop
31	Noho	Italian Restaurant	Art Gallery	Cocktail Bar
32	Civic Center	Gym / Fitness Center	Italian Restaurant	Coffee Shop
34	Sutton Place	Gym / Fitness Center	Italian Restaurant	Indian Restaurant
38	Flatiron	Gym / Fitness Center	Yoga Studio	American Restaurant
39	Hudson Yards	American Restaurant	Italian Restaurant	Café

Table 3 Cluster 0 from the results of the k-means cluster analysis.

	Neighborhood	Venue	Venue Category
666	Lenox Hill	A Matter of Health	Health Food Store
1011	Clinton	Sunac Natural Food	Health Food Store
3297	Hudson Yards	Sunac Natural Food	Health Food Store
2997	Turtle Bay	The Health Nuts	Health Food Store
2089	Manhattan Valley	The Nutbox	Health Food Store

Table 1 Locations of Health Food Stores in Manhattan neighborhoods, including the current locations of Sunac Natural Food.

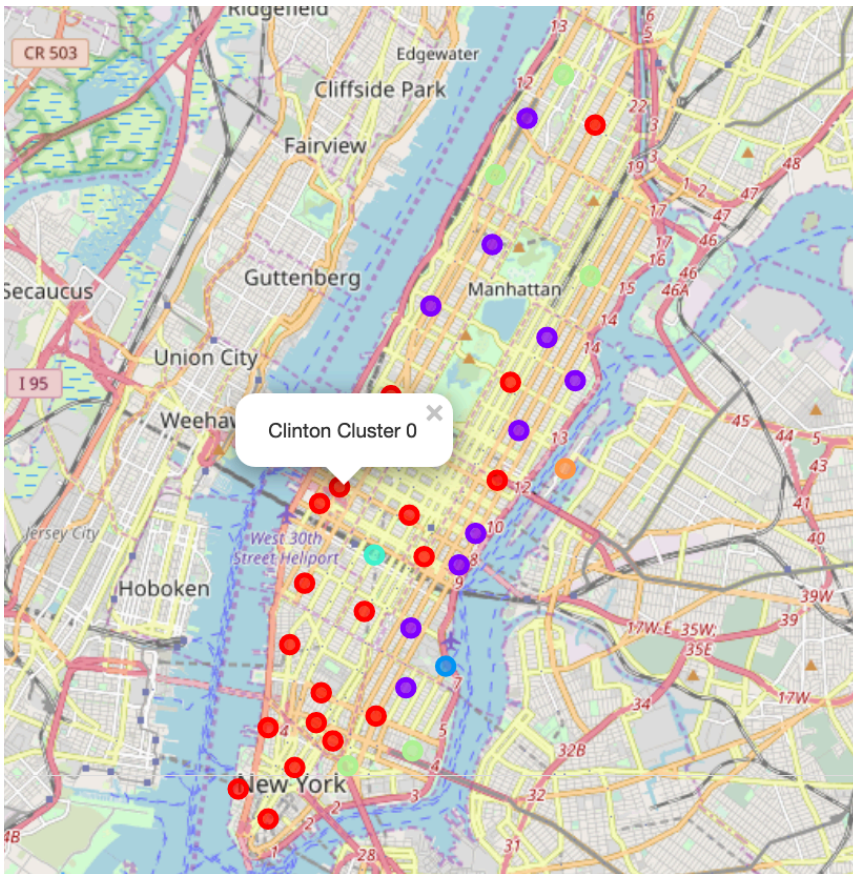

Figure 3 Map representing the clusters of the Manhattan neighborhoods based on the results of the k-means cluster analysis.

5. Discussion

Based on the clustering it could be recommended that Sunac Natural Food opens an expansion location in another neighborhood in Cluster 0. The current Sunac Natural Food

Locations are in the Clinton and Hudson Yards neighborhoods, which are both in cluster 0. Many of the neighborhoods in this cluster have a gym/fitness center in their top three most common venues, which would make sense as to why a health food store would likely be successful in these neighborhoods. The remaining neighborhoods in this cluster do not contain health food stores, so over saturating that market would not be as much of a concern. Therefore the following neighborhoods could be considered candidate neighborhoods for Sunac Health Food's next expansion location: Tribeca, West Village, Flatiron, Greenwich Village, Chelsea, Murray Hill, Midtown, Lincoln Square, Little Italy, Soho, Financial District, Upper East Side, Central Harlem, Noho, Civic Center, Sutton Place, and Battery Park City.

5.1 Limitations

There are limitations to this data analysis. First, the outcomes are only based on types of venues in each neighborhood of Manhattan. There are many other factors that should be considered when picking a business location, such as market retail prices and population demographics of the neighborhood. While similarities in venues can provide some evidence of potential success, other factors should be examined in order to narrow the search to fewer candidate locations. Since this cluster is large and provided many candidate locations, further analysis will be necessary in order to narrow the candidate locations.

5.2 Conclusions

Based on the results of the cluster analysis and the current locations of health food stores in Manhattan, the following neighborhoods could be considered candidate neighborhoods for Sunac Natural Food's expansion location.

5.3 Future directions

These candidate locations should be further investigated using real estate and demographic data to further narrow the selection for the most appropriate expansion location for Sunac Natural Food.

6. References

[1] https://www.sunacnaturalmarket.com/

[2] https://cocl.us/new_york_dataset

[3] https://developer.foursquare.com/