Prediction of neighborhoods most likely to support multi-ethnic grocery stores

Introduction

Non-traditional multi-ethnic focused grocery stores are a growing segment in the grocery store industry. Over the next 20 years the United States will become a minority majority country with no one ethnic group composing the majority of the population. Migration is particularly strong among Asian and Hispanic ethnic groups. As a result, once niche Asian and Hispanic food markets are becoming mainstream opening up a bright new profit opportunity in the languishing brick and mortar grocery industry. Identification of prime locations for non-traditional multi-ethnic grocery stores will increase the likelihood of a new market becoming profitable. One possible marker of a neighborhood in which a non-traditional multi-ethnic grocery store may thrive is a high concentration of non-Western food focused restaurants. Such grocery markets do particularly well in urban centers with high immigrant populations. New York City has both a large immigrant population and a dense population density. As a consequence neighborhoods within New York City are likely to support a non-traditional multi-ethnic focused grocery store. To identify neighborhoods with the greatest demand for this type of grocery store and the ethnicity to focus upon, foursquare data will be used to identify a high concentration of Asian, Hispanic, and Middle Eastern Cluster analysis will be run to identify regional clusters of multi-ethnic eateries to identify neighborhoods most likely to support a non-traditional multi-ethnic grocery store.

Methodology

To conduct this analysis longitude and latitude data will be obtained for neighborhoods in New York City. This data was obtained previously in the course at https://cocl.us/new_york_dataset. Because of the data limitations of the free version of foursquare only the neighborhoods within the borough of Brooklyn will be analyzed here. However, the same methodology can be applied to the other New Yok city boroughs or to any region for which longitude and latitude information can be collated.

¶The Brooklyn neighborhoods and their longitude and latitude information was extracted from the New York City neighborhood data. Seventy longitude and latitude neighborhood locations were retrieved and fed into the foursquare api to identify 15 venues with a 1.5 km radius of each of the 70 neighborhood locations.

One hundred ninety unique venues were returned. The following categories were used for analysis: Arepa Restaurant, Asian Restaurant, Bubble Tea Shop, Burmese Restaurant, Cantonese Restaurant, Caribbean Restaurant, Chinese Restaurant, Cuban Restaurant, Dumpling Restaurant, Empanada Restaurant, Greek Restaurant, Halal Restaurant, Indian Restaurant, Israeli Restaurant, Japanese Restaurant, Latin American Restaurant, Lebanese Restaurant, Mediterranean Restaurant, Mexican Restaurant, Middle Eastern Restaurant, Noodle House, Peruvian Restaurant, Eastern European Restaurant, Russian Restaurant, South American Restaurant, Sushi Restaurant, Taco Place, Taiwanese Restaurant, Thai Restaurant, Tibetan Restaurant, Turkish Restaurant, Vietnamese Restaurant.   These restaurants were assigned to Asian, South/Central American, and Middle Eastern categories, and the number of restaurants in each category were calculated for each neighborhood.

This count data was fed into a k-means clustering algorithm. The data was not first standardized since the data was count based and there were no extreme differences in values. Visual inspection and domain knowledge was used to determine the k means cluster number best suited for this data.

Results

Number of Ethnic Restaurants per Neighborhood

An exploratory analysis of the restaurant count data revealed the following histogram
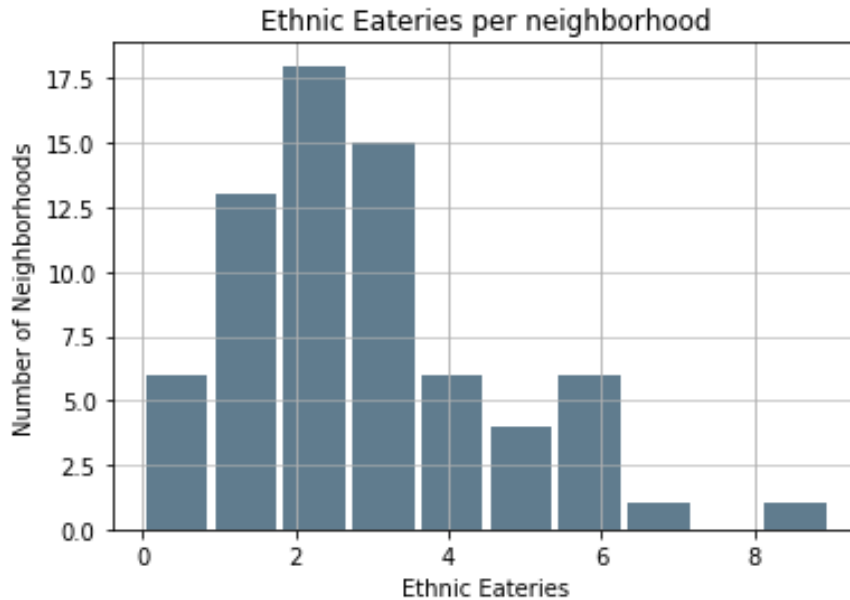

Figure 1 Histogram of Neighborhoods with increasing numbers Ethnic Restaurants

From the histogram it is apparent that the number of eateries per neighborhood ranges from 0-9 with 2 ethnic eateries being the most frequent number per neighborhood.

k-means clustering

The following count data was fed into the k-means clustering algorithm: number of Middle Eastern restaurants, number of Asian restaurants, number of Middle Eastern restaurants, and total number of non-Western European restaurants. Note that for this analysis Eastern European and Russian restaurants were included in the Middle Eastern group. This is due to the author's observation that the Middle Eastern grocery stores near her often have bakeries and other grocery sections dedicated to Russian and Eastern European customers.

Visual inspection of the cluster results using the author's domain knowledge was used to determine a cluster number that didn't compress the data too much but at the same time didn't over divide the data. A k-means cluster number of 5 was determined to return the most meaningful results. Below are the neighborhoods within each cluster and the number of restaurants in each.

Figure 2 contains cluster 1 and cluster 2. As seen in the figure 2 cluster one primarily contains neighborhoods with very few non-Western European/American restaurants. Similarly, cluster 2 contains primarily Asian and Middle Eastern restaurants, although there are relatively few total restaurants. Figure 3 contains clusters 3-5. Cluster 3 contains primarily Central and South American restaurants, while cluster 4 is similar to cluster 2, except that these neighborhoods contain a larger number of restaurants.

Cluster 5 identifies the 2 neighborhoods at the far end of the histogram in figure 1. These restaurants are primarily South and Central American.

These clusters are visualized overlaid on a map of Brooklyn, New York in Figure 4.

Neighborhood	ARestT	HRestT	MERestT
Greenpoint	0	2	0
Crown Heights	0	1	0
Bedford Stuyvesant	0	1	0
Cobble Hill	1	0	0
Red Hook	0	0	0
Gowanus	1	0	0
Park Slope	1	0	0
East New York	0	1	0
Starrett City	0	0	0
Canarsie	0	2	0
Mill Island	0	2	0
Coney Island	0	0	0
Marine Park	1	0	0
Clinton Hill	1	0	0
Sea Gate	0	0	0
Bergen Beach	0	1	0
Midwood	0	1	1
Georgetown	0	1	0
Wingate	0	2	0
New Lots	0	1	0
Paerdegat Basin	0	2	0
Fulton Ferry	0	1	0
Vinegar Hill	0	0	0
Weeksville	1	0	0
Dumbo	0	0	0

Cluster 1

Neighborhood	ARestT	HRestT	MERestT
Bay Ridge	0	1	2
Bensonhurst	2	0	0
Gravesend	0	0	2
Brighton Beach	0	0	2
Flatbush	1	1	1
Windsor Terrace	1	0	1
Williamsburg	0	1	2
Brooklyn Heights	2	0	0
Carroll Gardens	1	0	1
Bath Beach	2	0	0
Borough Park	1	0	1
Gerritsen Beach	1	0	2
Downtown	3	0	0
Boerum Hill	2	0	1
Ocean Hill	3	0	0
Prospect Park South	2	1	0
East Williamsburg	1	0	1
Fort Hamilton	2	0	0
Ditmas Park	1	1	0
Mill Basin	1	1	0
Broadway Junction	2	1	0
Madison	1	0	2

Cluster 2

Figure 2
Clusters 1 and 2. Note that cluster 1 appears to contain neighborhoods with few ethnic restaurants while cluster 2 contains neighborhoods with primarily Asian and Middle Eastern restaurants, although relatively few non-Western restaurants total.

A.

Neighborhood	ARestT	HRestT	MERestT
Sunset Park	2	3	0
Prospect Heights	1	3	0
Brownsville	0	3	0
Bushwick	0	4	0
Fort Greene	1	2	1
Cypress Hills	0	3	0
Flatlands	1	2	0
Prospect Lefferts Gardens	2	2	0
City Line	0	3	0
North Side	3	2	0
South Side	1	3	1
Rugby	2	4	0
Remsen Village	2	4	0
Highland Park	0	3	0

Neighborhood	ARestT	HRestT	MERestT
Sheepshead Bay	2	0	4
Manhattan Terrace	1	0	3
Kensington	4	0	1
Manhattan Beach	1	0	5
Dyker Heights	2	0	2
Ocean Parkway	3	0	3
Homecrest	5	1	1

B.

Neighborhood	ARestT	HRestT	MERestT
East Flatbush	2	7	0
Erasmus	0	6	0

C.

Figure 3.
A. Cluster 3 – Neighborhoods with primarily South/Central American restaurants
B. Cluster 4 – Neighborhoods with primarily Asian and Middle Eastern restaurants
C. Cluster 5 – Neighborhoods with the highest number of South/Central American restaurants.
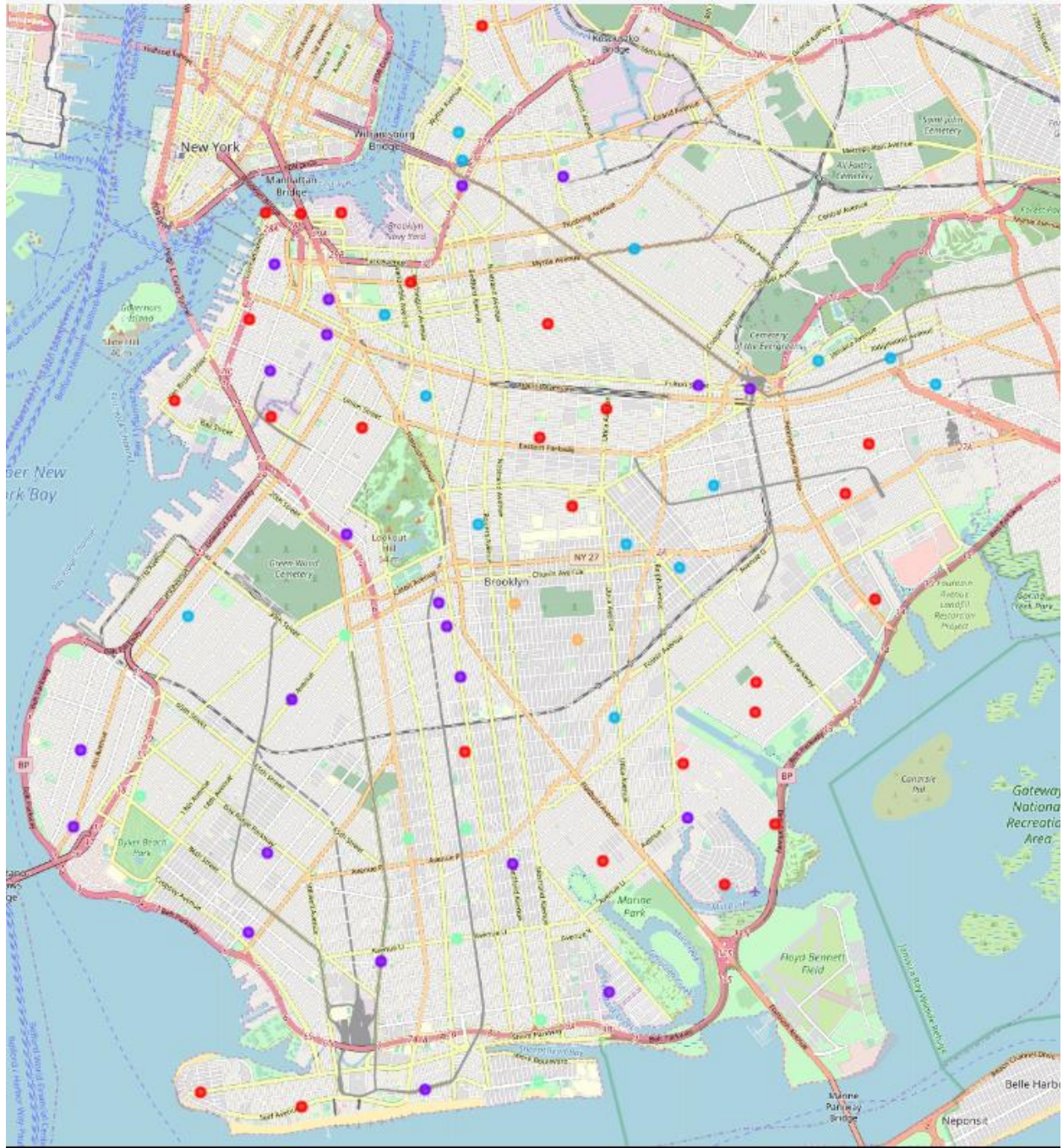
Figure 4 Map of Brooklyn overlaid with cluster results

Red dots are cluster 1

Purple dots are cluster 2

Turquoise dots are cluster 3

Green dots are cluster 4

Orange dots are cluster 5

Discussion

This was a preliminary analysis to check if neighborhood venue information can predict geographical locations for multi-ethnic grocery stores. Even this very elementary analysis can show regions where multi-ethnic grocery stores may thrive. For example, cluster 3 and cluster 5 identify neighborhoods where Hispanic grocery stores catering to patrons looking for Caribbean, Central, South American groceries and baked goods. In contrast, cluster 4 contains neighborhoods where a grocery store catering to populations looking for Asian and Middle Eastern groceries may thrive.

Additional analysis incorporating demographic information such as ethnicity, income, and property values may further refine these predictions.

This analysis can be repeated in other neighborhoods. The other New York City boroughs and other cities such as, London, Toronto, Los Angeles, Vancouver, and other capital and second cities would be logical neighborhood choices.