Identifying A Location for A New Restaurant In New York

Claudia Kudiabor

August 10, 2020

1. Introduction

1.1 Background

New York is home to different cultures which continue to be enriched with a diverse immigrant population which has influenced everything from art to food. It offers a unique opportunity for all people to showcase their culture especially through food. My aunt is an incredible chef and she would like to open a Ghanaian-inspired family style restaurant. With the large population of New York and the its high immigrant population she figured it would be a great place to open what would hopefully be the first of many restaurants with great exposure. However, with the restaurant industry being a high-risk business it is critically important to place the restaurant in a neighborhood which would provide high visibility and traffic but should not already be saturated with eateries to increase the probability for success.

1.2 Problem

The data required includes New York neighborhood data and Foursquare data on eateries in the neighborhoods. The Foursquare data will be overlaid on the neighborhood data to determine which boroughs/neighborhoods offer the most potential for her new restaurant opening. Neighborhoods which do not show an overly saturated market based on the number of eateries in the areas but shows a large number of highly trafficked areas, like parks and bookstores, will be the best candidates.

1.3 Interest

My aunt in definitely invested in the success of her new restaurant and would like to strategically choose a location that will increase the probability of success. Also, with the entire family contributing towards the upfront investment, we are all interested in making sure that the investments yield a strong payoff.

2. Data

2.1 Data sources

The NYU Spatial Data Repository's 2014 New York City Neighborhood Names data has information on the 5 borough and 306 neighborhood that make up the city. It also includes the longitudes and latitudes of each of the neighborhoods. Using the Foursquare APIs, the neighborhoods coordinates will be used to identify surrounding venues in these neighborhoods. The venues include details on the categories which will inform whether a neighborhood is highly trafficked but not saturated by restaurants.

2.2 Data cleaning

The data needed was downloaded and loaded as a json file from the NYU site which holds New York City neighborhoods data as of 2014. The data needed from the file is under the 'features' category. It includes information on the Borough, Neighborhood, Longitude, and Latitude. A pandas dataframe with these respective columns was created and the data looped into the empty dataframe row by row. With the newly created dataframe a few checks to make sure all the data was transferred correctly and get preliminary information was executed. It revealed that there were indeed five boroughs and 306 neighborhoods.

Upon sharing the preliminary information with the client, she requested that we narrow our search to the Queens borough which will be the most convenient because of its proximity to her home. Therefore, the dataframe was filtered to only include neighborhoods in Queens.

The data was clean without missing values or concerns about outlier and inconsistencies.

The Foursquare data on the surrounding neighborhoods was extracted using an API call on the coordinates from the pandas dataframe. To avoid overloading the call request the radius was set to 500 and the limit to 100. All the venues reported were put into a dataframe for easy manipulation and increased functionality. First, the number of venues were grouped by the respective neighborhoods, then a simple check was run to determine the number of unique venue categories. Finally using one hot encoding the categories were indexed by the neighborhood so that each venue category was displayed as a column and each neighborhood as row with a dummy variable (1 or 0) indicating whether or not a

venue category is present in the neighborhood. The dummy variable was then transformed to show the mean of the frequency of the occurrence of each venue category per neighborhood. The venues were then sorted at the top ten most common venues were extracted a placed into a new dataframe.

Now both the spatial data on Queens and the Foursquare data on the nearest and most common venues were in two dataframes ready for the clustering process.

3. Methodology

3.1 Clustering

The problem that needed a resolution with this analysis was to determine where in New York City, more specifically Queens, offered the best environment for the opening of a new restaurant. A cluster analysis offers the best methodology to group like neighborhoods based on the most common venues in the neighborhood. The hypothesis is that the neighborhood(s) with a saturated eatery market will appear together in a single cluster and those outside of that cluster will offer the right environment for the new restaurant.

3.2 Process

Following the data cleaning there were 40 neighborhoods and 330 venues ready for clustering.
The k-means from the clustering stage (*from sklearn.cluster import KMeans*) offer the opportunity to test and execute the cluster analysis quickly. For the initial run, the k-means was set to five to get five different clusters. Each neighborhood is assigned a cluster of zero to four to determine its cluster group. The cluster results with the venues is then joined to the neighborhood data which includes the information on the borough and the coordinates.

3.3 Evaluation

Each cluster was examined by filtering for its cluster label to determine which venue categories distinguished one from the either. The examination revealed inconclusive clusters for determining the right neighborhood or neighborhoods to open the new restaurant. They were not unique enough to make a conclusive decision. It appeared that there were too many clusters for the neighborhoods and therefore

the k-means should be reduced. The objective was to have a stand out cluster that was not like the rest. Therefore, the k-means was updated to three for the development of three clusters. The same process was then rerun.

4. Result

The three clusters included the following:

Cluster One included the Breezy Point, Neponsit, and Hammels neighborhoods. There are a few eateries in the area but the most common venue across all the neighborhoods was the beach.

```
queens_merged.loc[queens_merged['Cluster Labels'] == 0, queens_merged.columns[[1] + list(range(5, queens_merged.shape[1]))
```

	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
43	Breezy Point	Beach	Monument / Landmark	Trail	Empanada Restaurant	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Filipino Restaurant	Fish & Chips Shop
50	Neponsit	Beach	Lounge	Women's Store	Fish & Chips Shop	Farm	Farmers Market	Fast Food Restaurant	Filipino Restaurant	Fish Market	Event Space
78	Hammels	Beach	Food Truck	Deli / Bodega	Fried Chicken Joint	Gym / Fitness Center	Diner	Bus Station	Bus Stop	Dog Run	Shoe Store

Cluster Two only includes the Somerville neighborhood with park as the most common venue.

```
queens_merged.loc[queens_merged['Cluster Labels'] == 1, queens_merged.columns[[1] + list(range(5, queens_merged.shape[1]))
```

	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
63	Somerville	Park	Filipino Restaurant	Empanada Restaurant	Event Space	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Indian Restaurant

Cluster Three had the most neighborhoods represented here. There was a high concentration of restaurants which include Mexican, Italian, Thai and Latin American.

```
queens_merged.loc[queens_merged['Cluster Labels'] == 2, queens_merged.columns[[1] + list(range(5, queens_merged.shape[1]))]
```

	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
0	Astoria	Bar	Middle Eastern Restaurant	Greek Restaurant	Seafood Restaurant	Hookah Bar	Pizza Place	Indian Restaurant	Mediterranean Restaurant	Bakery
1	Woodside	Grocery Store	Thai Restaurant	Bakery	Bar	Filipino Restaurant	Latin American Restaurant	Donut Shop	Pub	American Restaurant
2	Jackson Heights	Latin American Restaurant	Peruvian Restaurant	South American Restaurant	Bakery	Mexican Restaurant	Mobile Phone Shop	Thai Restaurant	Grocery Store	Supplement Shop
3	Elmhurst	Thai Restaurant	Mexican Restaurant	Vietnamese Restaurant	Chinese Restaurant	Salon / Barbershop	Pizza Place	Bank	Bakery	Malay Restaurant
4	Howard Beach	Italian Restaurant	Bagel Shop	Pharmacy	Sandwich Place	Fast Food Restaurant	Chinese Restaurant	Deli / Bodega	Supermarket	Jewelry Store
5	Corona	Mexican Restaurant	Bakery	Supermarket	Convenience Store	Pizza Place	Donut Shop	Park	Restaurant	Sandwich Place
6	Forest Hills	Gym / Fitness Center	Gym	Yoga Studio	Pharmacy	Convenience Store	Park	Thai Restaurant	Pizza Place	Video Game Store
7	Kew Gardens	Chinese Restaurant	Deli / Bodega	Pizza Place	Bank	Donut Shop	Cosmetics Shop	Indian Restaurant	Bar	Pet Store
8	Richmond Hill	Latin American Restaurant	Pizza Place	Lounge	Bank	Gym / Fitness Center	Deli / Bodega	Moving Target	Supermarket	Caribbean Restaurant

5. Discussion

Each of the three clusters have eateries represented in the top five most common venues which is a positive representation of the Queens area as a welcoming market for eateries. However, the distinguishing factor across the clusters is the type of eateries and surrounding non-eatery venues. Cluster one is close to the beach and shows a few fast/casual eateries and a Filipino restaurant. It also has a Monument/Landmark, Trail and Lounge in the top three most common venues. There is a low saturation of eateries but attests to being a highly trafficked area. Cluster two offers a single neighborhood with the park as the most common venue but is followed by Filipino and Empanada restaurants. It has a high concentration of restaurants with less non-eateries to determine possible uninhibited traffic. Cluster three has the most eateries in the top three venues without much non-eatery options to drive traffic. There is a high concentration of eateries. Based on the proportion of the eateries to non-eateries in each cluster, cluster one offers a great opportunity to distinguish the new restaurant from the other eateries in the area. It has only a few number of restaurants across its neighborhoods in

the most common venues. Also, considering that the client will be opening a Ghanaian-inspired restaurant having a neighborhood with few non-American options will make it a standout in the neighborhood.

6. Conclusion

In this study, we analyzed New York city neighborhoods, specifically Queens, to determine the best neighborhoods to open a brand-new Ghanaian-inspired restaurant. We took a look at the neighborhood data to determine which neighborhoods made up the Queens borough and mapped out their respective coordinate to visualize their locations. We then transformed the Foursquare data into a dataframe which included the top ten venues across each of these neighborhoods. We then run a k-means cluster analysis to group the neighborhood into what turned out to be three informative clusters. The cluster chosen, cluster one, will be presented to the client to help her determine where in Queens to place the first of many restaurants. The clustering model can be replicated for any other restauranteurs interested in opening a restaurant in New York city.