Restaurant Battle of Manhattan Neighborhoods

Luke LaBonte

August 9, 2019

Table of Contents

INTRODUCTION

BACKGROUND

A pair of restaurant owners want to expand their operations to high density locations that are most ideal location to open up a profitable restaurant(s). The growth strategy is to be able to open a chain of sustainable franchises domestically in the United States northeast and then other major cities across the nation in the future. The owners want to target Manhattan in New York City first.

PROBLEM

Over the past couple decades, the food services industry has evolved significantly to accommodate a new wave of genres and eating habits. Boutique restaurants, fast-food franchises, and even grocery stores have gradually shifted to a wider range of appeal to everchanging modern consumer diets. The push for healthier foods and proof of provenance from farm to consumer has become increasingly vital - thus creating more market space for niche yet popular cuisines. Manhattan is one of the densest areas in the United States and is known for its fast pace and diverse culture with a variety of food options that the owners think will serve as a great starting point for our initiative. By exploring each neighborhood, their venues, and analyzing trends via Foursquare data, the owners will be able to effectively gauge optimal locations that will yield the best possible success, sustainability, and growth for their restaurant expansion.

DATA

DATA SOURCES

Dataset 1 - New York City borough and neighborhood data

First, I will used data from a webpage which provides information about list of different boroughs and their neighborhoods in New York City. I downloaded the file locally as a JSON file and uploaded it into the notebook. I then extracted the data into a table from JSON format. This table contains four columns: Borough, Neighborhood, Latitude, and Longitude. The link to the webpage with the data is: (https://cocl.us/new_york_dataset). A sample view of the dataset in a Pandas data frame can be seen below:

	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

Dataset 2 - Different venues in Manhattan, New York City

This dataset will be formed using the Foursquare API. I used the Foursquare location data to explore different venues in each Manhattan neighborhood. The types of these venues varied from Parks, Coffee Shops, Hotels, to Gyms, etc. Using the Foursquare location data, I was able to quickly get information about these venues and analyzed the neighborhoods they resided in. I used the geographical coordinates from the dataset to generate the location dataset. A sample view of the data set can be seen below:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop
5	Marble Hill	40.876551	-73.91066	Blink Fitness Riverdale	40.877147	-73.905837	Gym
6	Marble Hill	40.876551	-73.91066	TCR The Club of Riverdale	40.878628	-73.914568	Tennis Stadium
7	Marble Hill	40.876551	-73.91066	Land & Sea Restaurant	40.877885	-73.905873	Seafood Restaurant

PRE-PROCESSING

After obtaining the two datasets, I pre-processed the second dataset from Foursquare via one-hot encoding so that it could used in the K-Means clustering algorithm. Doing so, we one-hot encoded the *manhattan_venues* data frame and stored in a data frame named *manhattan_onehot* which is ready for clustering.

However, this dataset contains information about all the nearby venues like Park, Gym, Shops, etc. which is not necessary as we are only interested in venues in 'food' venues - venues like Park, Gym, Playground are discarded from the *manhattan_onehot* data frame. We are also only looking for venues that are proper restaurants. Venues such as coffee shops, pizza places, bakeries etc. are not direct competitors of the restaurant business, so we aren't interested in those as much and only include venues that have 'restaurant' in category name.

We make sure to include all the subcategories of different restaurants in the neighborhood, e.g. Afghan restaurant, Italian restaurant, etc. For this, we locate venues from *manhattan_onehot* data frame that are only restaurants and store them in a new data frame named *manhattan_restaurants*. This new data frame will now be used for clustering algorithm.
Also, a data frame named *neighborhood_venues_sorted* was also created to list all the Manhattan neighborhoods along with their respective five most common venues. This dataset would eventually help in visualizing the solution. First five rows of this data frame are depicted in figure below:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Battery Park City	Park	Coffee Shop	Hotel	Memorial Site	Gym
1	Carnegie Hill	Pizza Place	Coffee Shop	Cosmetics Shop	Café	French Restaurant
2	Central Harlem	African Restaurant	French Restaurant	Public Art	Cosmetics Shop	Seafood Restaurant
3	Chelsea	Coffee Shop	Italian Restaurant	Ice Cream Shop	Nightclub	Bakery
4	Chinatown	Chinese Restaurant	Cocktail Bar	Vietnamese Restaurant	American Restaurant	Salon / Barbershop

ANALYSIS

In the *manhattan_restaurants* data frame, I added a column calculating the total number of restaurants in that neighborhood. This helped us further analyze each cluster after running the K-Means algorithm. The K-Means clustering algorithm clustered the *manhattan_restaurants* data frame so we can further segment our analysis of the neighborhoods. For this, I set number of clusters to be five. This is a chosen parameter for the algorithm.

After the clusters were made, I merged the first dataset and the *manhattan_venues_sorted* data frame and inserted cluster labels. The result data frame was named *manhattan_merged* which looked like this:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	3	Sandwich Place	Coffee Shop	Discount Store	Yoga Studio	Supplement Shop
1	Manhattan	Chinatown	40.715618	-73.994279	0	Chinese Restaurant	Cocktail Bar	American Restaurant	Vietnamese Restaurant	Salon / Barbershop
2	Manhattan	Washington Heights	40.851903	-73.936900	4	Café	Bakery	Mobile Phone Shop	Grocery Store	Italian Restaurant
3	Manhattan	Inwood	40.867684	-73.921210	2	Café	Mexican Restaurant	Pizza Place	Lounge	Bakery
4	Manhattan	Hamilton Heights	40.823604	-73.949688	2	Mexican Restaurant	Pizza Place	Café	Coffee Shop	Yoga Studio

Next, we dug deeper into each cluster to compare the total number of neighborhoods and total number of restaurants. I then calculated the Restaurant/Neighborhood ratio to find which cluster had the lowest and continue with that cluster for further analysis. Cluster 3 yielded the lowest ratio and consisted of total four neighborhoods and 13 restaurants. Out of these four neighborhoods, one was discarded since it was the furthest neighborhood from Manhattan's center. The other three didn't have a high number of total restaurants, were relatively close to the center of Manhattan, and seemed ideal for a new restaurant(s).

The final dataset contains all the information about these three remaining neighborhoods:

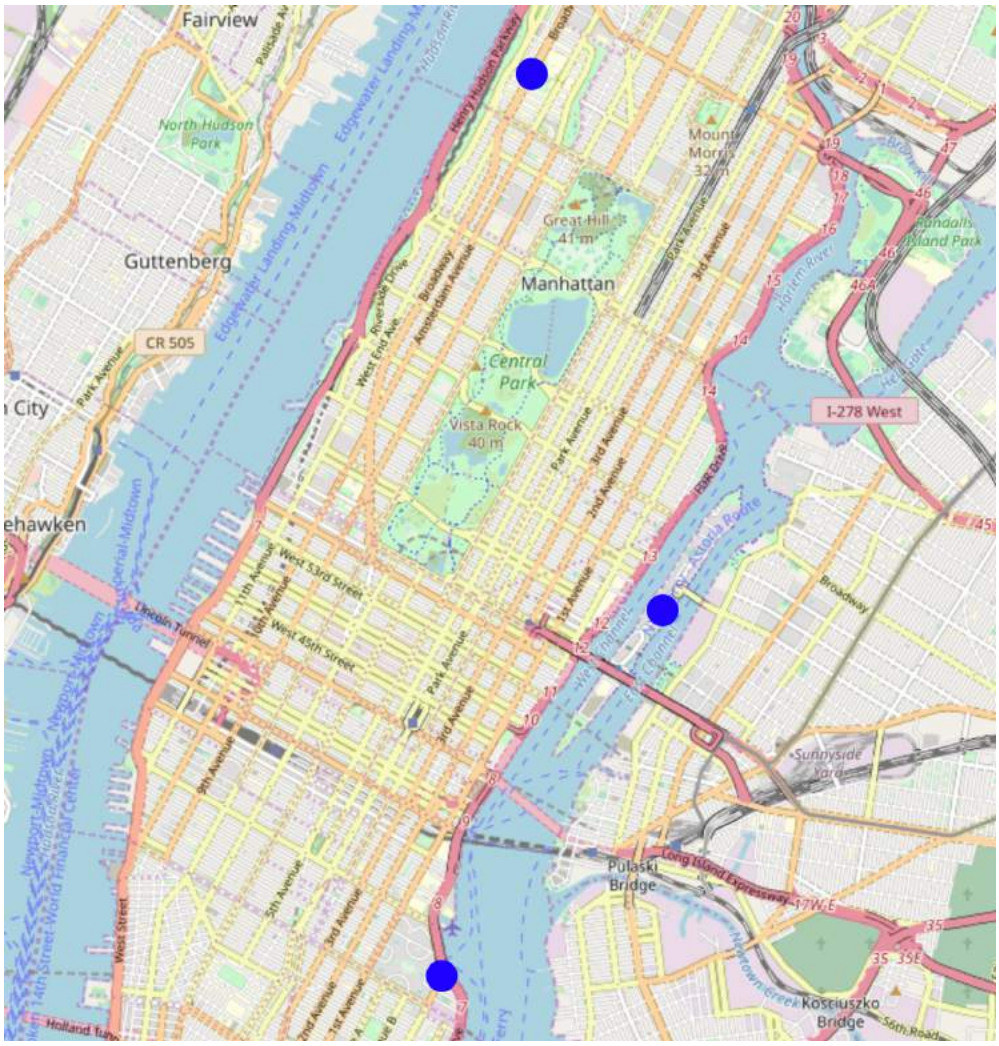
	Neighborhood	Borough	Latitude	Longitude
1	Roosevelt Island	Manhattan	40.76216	-73.949168
2	Morningside Heights	Manhattan	40.80800	-73.963896
3	Stuyvesant Town	Manhattan	40.73100	-73.974052

VISUALIZATION

A map of Manhattan was generated using *Folium*, a great visualization library. All 40 Manhattan neighborhoods are marked with blue circles on the map below from the first dataset.

After applying K-Means clustering, five unique clusters of neighborhoods were created based on *manhattan_restaurants* data frame. The new map of Manhattan neighborhood clusters can be seen below.

The final neighborhoods were also presented on a map:



The three Manhattan neighborhoods – Morningside Heights, Roosevelt Island, and Stuyvesant Town – are depicted by the blue dots in the map above.

RESULTS & DISCUSSION

Although there is a plethora of restaurants in Manhattan, our analysis showed pockets of low restaurant density fairly close to city center. To identify these pockets, we used a clustering algorithm and segmented our neighborhood dataset accordingly.

Using the K-means clustering algorithm, we constructed five clusters each containing a portion of Manhattan neighborhoods based on the number of restaurants in their vicinity. Next, we analyzed each cluster by calculating Restaurant/Neighborhood ratio of each cluster. We saw that Cluster 3 had lowest ratio, which means very there are fewer restaurants present within the

vicinity of each neighborhood compared to that of the other clusters. There were total 4 neighborhoods in Cluster 3. Upon further analysis, we found that 1 of the 4 neighborhoods, Marble Hill, was not a good location for opening up a new restaurant due it not being close to the center of Manhattan.

According to our analysis, we found three neighborhoods where new restaurant business might see increasing success and there are two reasons for that. First, we saw that these neighborhoods do not inhabit many restaurants which will lower the competition for the new restaurant(s) aspiring success and sustainable growth. Second, as seen in the above map, these three neighborhoods are located toward the center and more dense areas of Manhattan that are more populous which will provide more foot traffic and potential customers. And at the same time, they are spread out enough from each other that they can accompany different busy sections of Manhattan.

These final three resulting neighborhoods are ideal for opening a new restaurant and are stored in a data frame with their extended geographical contains information.

The owners can look to examine these three locations first and further determine the type of restaurant and cuisine that behooves their business strategy and growth.

CONCLUSION

The purpose of this project was to identify neighborhoods in Manhattan, New York with a low number of restaurants in order to aid stakeholders in narrowing down the search for optimal location(s) for a new restaurant. By analyzing restaurant density distribution from Foursquare data, we have first identified the five most common nearby venues of each neighborhood. Then with the help of clustering techniques and further analysis we were able to narrow down to three neighborhoods, Morningside Heights, Roosevelt Island, and Stuyvesant Town, that fit the density criteria yielding ideal candidate locations for opening up a new restaurant.