Neighborhood Selection for a New Restaurant in New York City

Introduction

Finding an optimal location for a new restaurant in a major city like New York City can be a complex process with many factors involved. Indeed, the location of the restaurant could determine its success. Therefore, this analysis could provide useful data and insights for business owners and restaurant owners opening a new location.

In order to find the ideal location, we will assume that we want to:

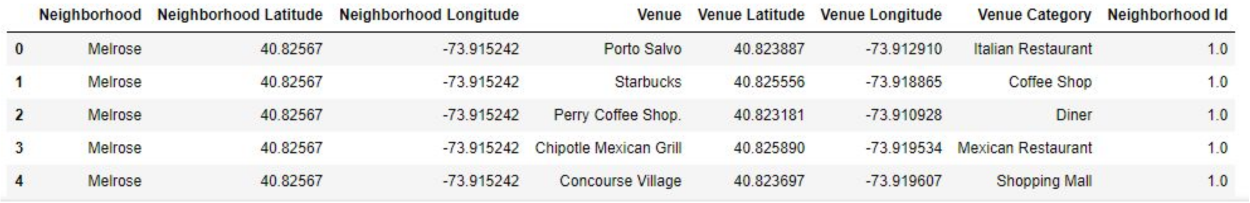
- Select the neighborhood that would maximize its proximity to high traffic venues
- Select the neighborhood with fewer competing restaurants of the same type

Data

We will gather all the neighborhoods in New York City from all 5 boroughs from the Wikipedia web page. We will then use the Python geopy module to geocode the neighborhoods into latitude and longitude coordinates. Then, we will use the FourSquare API to get venues from those latitude, longitude coordinates. The top eating venues and non-eating venues in each neighborhood in New York City will be found for each neighborhood and then used as the data to input to the clustering analysis.

Methodology

New York City's neighborhoods were gathered from the table in the Wikipedia web page. These neighborhoods were then geocoded into latitude and longitude coordinates. Then, we passed those coordinates to the FourSquare API to gather venue data from each neighborhood. This created a table of the form of:



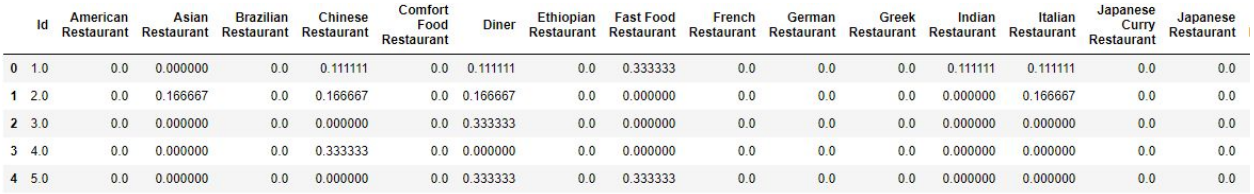
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Neighborhood Id
0	Melrose	40.82567	-73.915242	Porto Salvo	40.823887	-73.912910	Italian Restaurant	1.0
1	Melrose	40.82567	-73.915242	Starbucks	40.825556	-73.918865	Coffee Shop	1.0
2	Melrose	40.82567	-73.915242	Perry Coffee Shop.	40.823181	-73.910928	Diner	1.0
3	Melrose	40.82567	-73.915242	Chipotle Mexican Grill	40.825890	-73.919534	Mexican Restaurant	1.0
4	Melrose	40.82567	-73.915242	Concourse Village	40.823697	-73.919607	Shopping Mall	1.0

This venues table was then sorted so that two new tables were created: one for restaurant venues and one for non-eating venues. If a neighborhood had no restaurants or no non-eating venues, a new empty entry was created for that neighborhood.

Restaurant venues were defined as those that were of the categories: American, Diner, Chinese, Italian, Japanese, Korean, French, Vietnamese, Jewish, Comfort Food, Thai, Turkish, Greek, Mexican, Sushi, Southern, Spanish, Puerto Rican, Russian, Indian, Middle Eastern, Greek, German, Fast Food, Ethiopian, Carribean, Asian, Brazilian

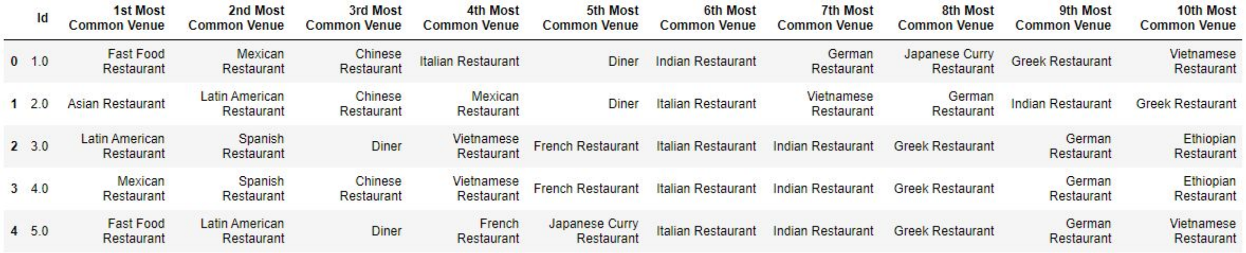
Non-eating venues were defined as those that were of the categories: Art, Zoo, Hotel, College, Gym, Theme Park, Theater, Stadium, Court, Field, Beach, Park, Plaza, River, Mall, Museum, Concert, Marina, Event Space, Trail, Auditorium, Amphitheater, Auditorium, Beach, Skating Rink, Garden, Bowling Alley, Landmark, Bookstore

Next, a new table was created by one-hot encoding the columns such that all venue categories were enumerated. Then, the rows were grouped together and the mean value of the one-hot encoded table was used as the data values.



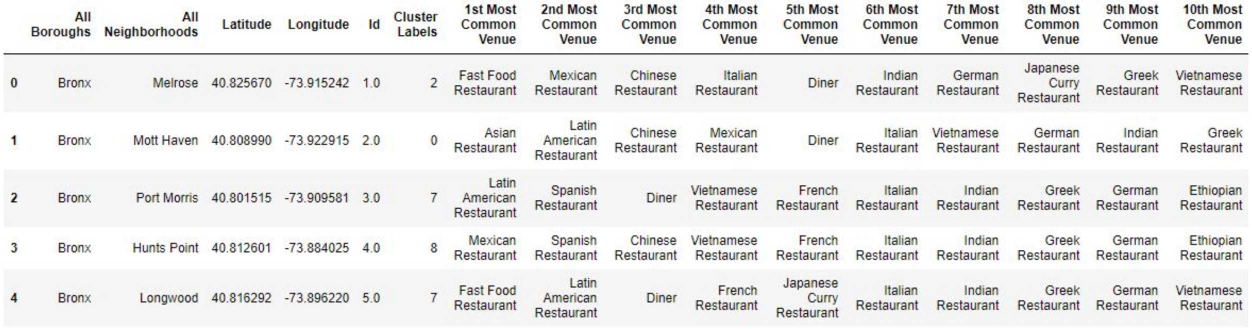
	Id	American Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Diner	Ethiopian Restaurant	Fast Food Restaurant	French Restaurant	German Restaurant	Greek Restaurant	Indian Restaurant	Italian Restaurant	Japanese Curry Restaurant	Japanese Restaurant
0	1.0	0.0	0.000000	0.0	0.111111	0.0	0.111111	0.0	0.333333	0.0	0.0	0.0	0.111111	0.111111	0.0	0.0
1	2.0	0.0	0.166667	0.0	0.166667	0.0	0.166667	0.0	0.000000	0.0	0.0	0.0	0.000000	0.166667	0.0	0.0
2	3.0	0.0	0.000000	0.0	0.000000	0.0	0.333333	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0
3	4.0	0.0	0.000000	0.0	0.333333	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0
4	5.0	0.0	0.000000	0.0	0.000000	0.0	0.333333	0.0	0.333333	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0

From these data values, we could find the most frequent venues in each neighborhood and list them in a table.



	Id	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	1.0	Fast Food Restaurant	Mexican Restaurant	Chinese Restaurant	Italian Restaurant	Diner	Indian Restaurant	German Restaurant	Japanese Curry Restaurant	Greek Restaurant	Vietnamese Restaurant
1	2.0	Asian Restaurant	Latin American Restaurant	Chinese Restaurant	Mexican Restaurant	Diner	Italian Restaurant	Vietnamese Restaurant	German Restaurant	Indian Restaurant	Greek Restaurant
2	3.0	Latin American Restaurant	Spanish Restaurant	Diner	Vietnamese Restaurant	French Restaurant	Italian Restaurant	Indian Restaurant	Greek Restaurant	German Restaurant	Ethiopian Restaurant
3	4.0	Mexican Restaurant	Spanish Restaurant	Chinese Restaurant	Vietnamese Restaurant	French Restaurant	Italian Restaurant	Indian Restaurant	Greek Restaurant	German Restaurant	Ethiopian Restaurant
4	5.0	Fast Food Restaurant	Latin American Restaurant	Diner	French Restaurant	Japanese Curry Restaurant	Italian Restaurant	Indian Restaurant	Greek Restaurant	German Restaurant	Vietnamese Restaurant

And, most importantly, because we had data values, we could perform a k-means cluster analysis on the one-hot encoded table with a row for each neighborhood. We selected a k value of 12 by performing the cluster analysis for various values of k and examining the error or distance of each value from the centroid. We found the elbow point where increasing the value of k did not cause much more of a decrease in error. The cluster labels could then be associated with each neighborhood in a final table.

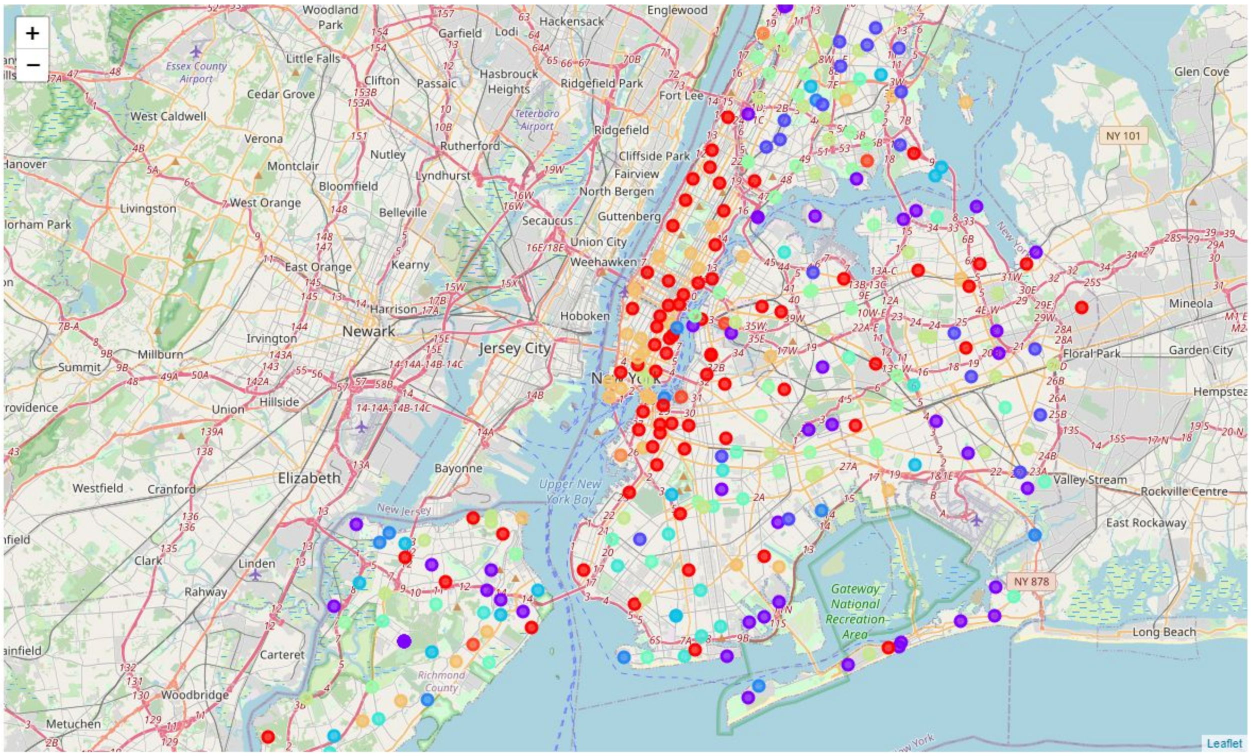


	All Boroughs	All Neighborhoods	Latitude	Longitude	Id	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bronx	Melrose	40.825670	-73.915242	1.0	2	Fast Food Restaurant	Mexican Restaurant	Chinese Restaurant	Italian Restaurant	Diner	Indian Restaurant	German Restaurant	Japanese Curry Restaurant	Greek Restaurant	Vietnamese Restaurant
1	Bronx	Mott Haven	40.808990	-73.922915	2.0	0	Asian Restaurant	Latin American Restaurant	Chinese Restaurant	Mexican Restaurant	Diner	Italian Restaurant	Vietnamese Restaurant	German Restaurant	Indian Restaurant	Greek Restaurant
2	Bronx	Port Morris	40.801515	-73.909581	3.0	7	Latin American Restaurant	Spanish Restaurant	Diner	Vietnamese Restaurant	French Restaurant	Italian Restaurant	Indian Restaurant	Greek Restaurant	German Restaurant	Ethiopian Restaurant
3	Bronx	Hunts Point	40.812601	-73.884025	4.0	8	Mexican Restaurant	Spanish Restaurant	Chinese Restaurant	Vietnamese Restaurant	French Restaurant	Italian Restaurant	Indian Restaurant	Greek Restaurant	German Restaurant	Ethiopian Restaurant
4	Bronx	Longwood	40.816292	-73.896220	5.0	7	Fast Food Restaurant	Latin American Restaurant	Diner	French Restaurant	Japanese Curry Restaurant	Italian Restaurant	Indian Restaurant	Greek Restaurant	German Restaurant	Vietnamese Restaurant

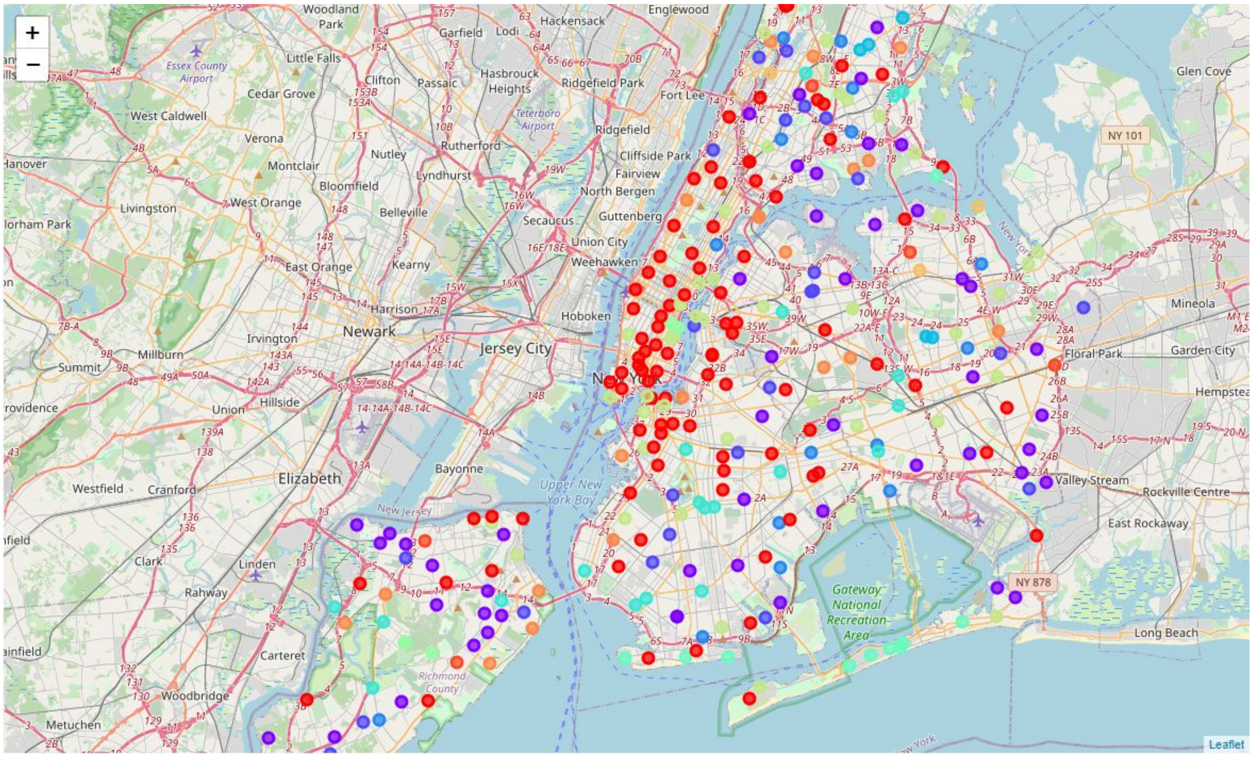
With these clusters, we could color code and plot them on a geographic map. We could also analyze cluster members and see what type of restaurants and non-eating venues are associated with the cluster.

Results

Restaurant Clustered Neighborhoods



Non-eating Venues Clustered Neighborhoods



Legend

#ff0000	Cluster 0
#8000ff	Cluster 1
#5247fc	Cluster 2
#2489f5	Cluster 3
#0ac0e8	Cluster 4
#3ae8d6	Cluster 5
#68fcc1	Cluster 6
#96fca7	Cluster 7
#c4e88a	Cluster 8
#f4c069	Cluster 9
#ff8947	Cluster 10
#ff4724	Cluster 11

The clustering results for restaurant data and non-eating data are marked on the geographic maps. The clusters were analyzed to determine its members. Two pieces of data can be extracted from the results. One is what type of restaurants will be most present in the surrounding area, the other is what type of non-eating venues will be most present in the area.

Discussion

Based on the data and the results of the analysis, we can deduce a few things.

From restaurant data:

Restaurant Cluster 1 members were analyzed and were shown to be those neighborhoods that do not have any restaurants in the neighborhood. FourSquare categorizes restaurants as those in which you sit down and are served. This excluded pizza joints, cafes, and other such establishments. However, we know that neighborhoods in cluster 1 have no restaurants in the area, which could be an opportunity to be the only sit down food establishment in the nearby area.

Restaurant Cluster 9 members were those in which Italian Restaurants were the most prevalent restaurants in the area. Therefore, it could be wise to avoid starting a new Italian Restaurant in these neighborhoods to avoid competition.

Restaurant Cluster 2 members were those neighborhoods in which Fast Food were the most common. These locations could be good locations to open a new restaurant to serve as an alternative to the widely prevalent fast food.

Restaurant Cluster 5 was shown to be the neighborhoods that have the highest relative frequency of Sushi Restaurants. So, these could be areas to avoid if one is trying to open a sushi place.

From non-eating venue data:

Non-eating Cluster 1 are areas in which there were no high-traffic venues. These are areas to avoid opening a restaurant, because we want locations to be desirable and have other businesses that can drive traffic to the restaurant.

Non-eating Cluster 5 had neighborhoods in which there were a high frequency of gyms compared to other non-eating establishments. There may be a positive relationship between gym-goers and the need for sit-down restaurants with healthy options.

Non-eating Cluster 8 had many non-eating venues with traffic among which parks were some of the most common. This cluster could be a good opportunity to open a business.

Conclusion

In this analysis, we have clustered neighborhoods in New York City based on restaurants present and non-eating venues present in the area for the purpose of determining optimal locations for a new restaurant. We have gathered a few insights as to where to locate restaurants to be in proximity to high traffic venues and away from competing restaurants.

This may be sufficient for many cases, however, there is room for improvement in the analysis. For example, extra weight could be placed on certain venues. A location next to Yankee Stadium, a historic landmark, or the Empire State Building could have higher value. Or better yet, another dataset that contains visitor numbers to the venues could be integrated to this analysis so that the weighting values can be accurately quantified.

In the end, we have found several insights that can help business owners and restaurant owners with the placement of their new restaurants.