Beijing Wu
Coursera: Applied Data Science Capstone
Capstone Project: The Battle of the Neighborhoods
November 18, 2020

Capstone Project: The Battle of the Neighborhoods

1. Background

New York City is one of the most densely populated cities in the United States. New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, fashion, and sports. New York City is composed of five boroughs – the Bronx, Manhattan, Queens, Brooklyn, and Staten Island. The city and its metropolitan area constitute the premier gateway for immigration to the United States.

Bubble Tea, also known as pearl milk team, or boba milk tea, is a tea-based drink originating in Taiwan. It includes chewy tapioca balls (pearl or boba) or a wide range of other toppings. Bubble Tea has become extremely popular among Asian immigrants in the US, and has become a signature flavor itself and inspired a variety of bubble tea flavored snacks. The highly increased demand in bubble tea drinks and their related industry provide opportunities for the market expansion.

As a Chinese-American immigrant, I am always in the hunt of the best bubble tea wherever I visit. My dream is to open my own bubble tea business one day in the heart of the New York City. Therefore, in this project, I will look into the five boroughs and their 306 neighborhoods in New York City, use different drink types as attributes to build up an unsupervised machine learning classifier model to classify the neighborhoods into clusters, study how neighborhood differs from each other, and find the right neighborhood to start my first bubble tea shop.

2. Data

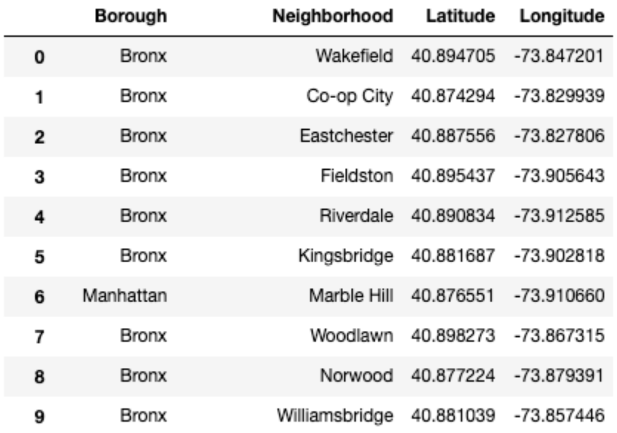
To segment New York City's neighborhoods and explore them, the dataset that contains 5 boroughs and the neighborhoods, along with the latitude and longitude coordinates of the neighborhoods, will be downloaded from the Internet (https://geo.nyu.edu/catalog/nyu_2451_34572).

The top 100 venues in each neighborhood within a radius of 500 meters will be requested and obtained using the Foursquare API. The data is then cleaned and preprocessed for explicability and understandability in order to meet business requirements purpose.

3. Methodology

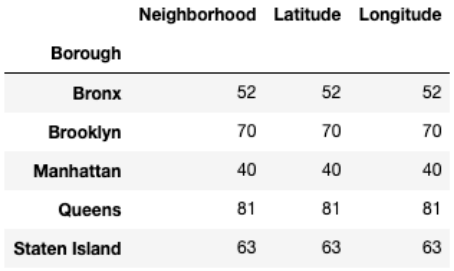
1) Explore the Data

After download the dataset as a json file, the relevant data, which is in the *features* key, is transformed into a *pandas* dataframe.



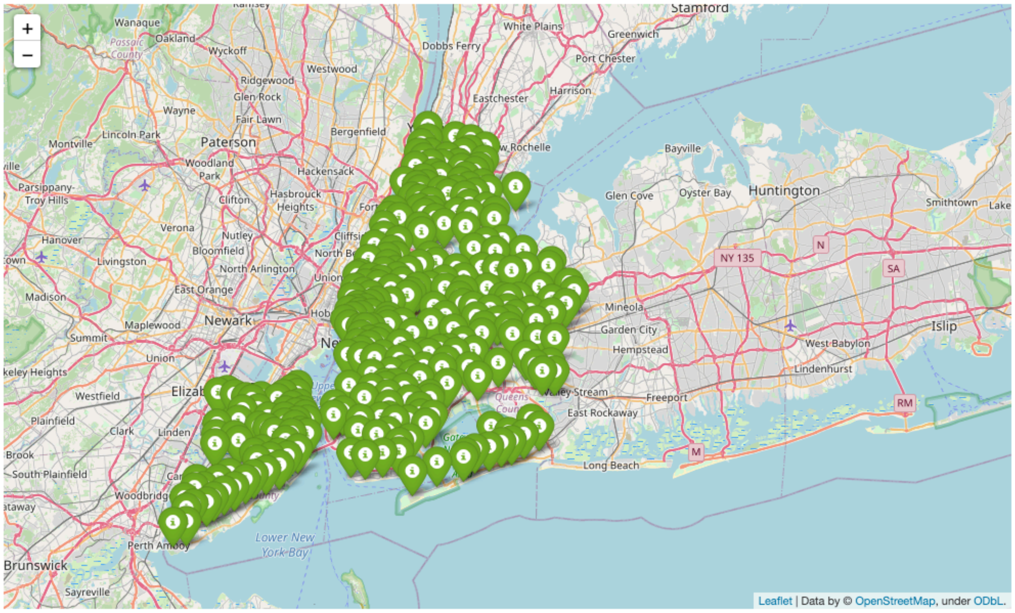
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

The dataframe is grouped by “Borough”, and the number of neighborhoods in each borough is counted. There are a total of 5 boroughs and 306 neighborhoods, as expected.

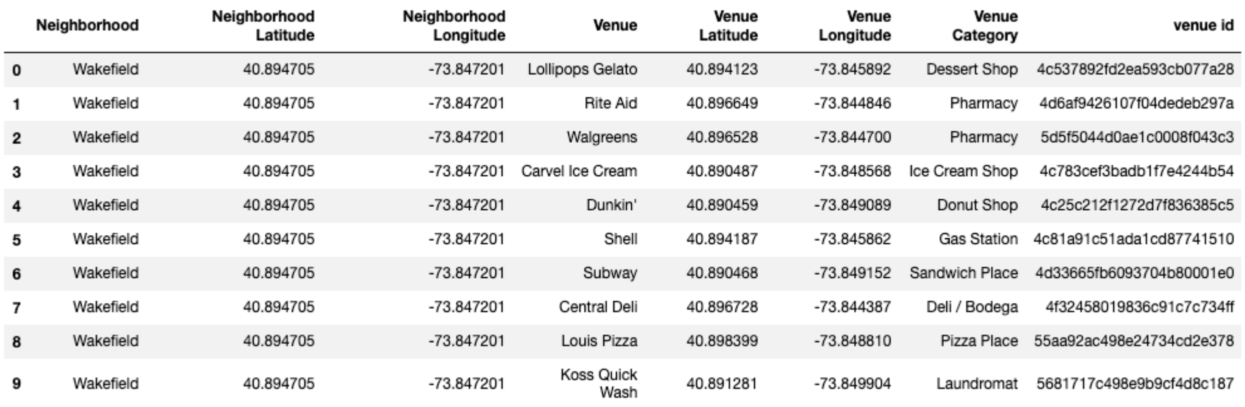


	Neighborhood	Latitude	Longitude
Borough			
Bronx	52	52	52
Brooklyn	70	70	70
Manhattan	40	40	40
Queens	81	81	81
Staten Island	63	63	63

The New York City’s coordinate is obtained using the geopy library, and a map of New York City with neighborhoods superimposed on top is created.

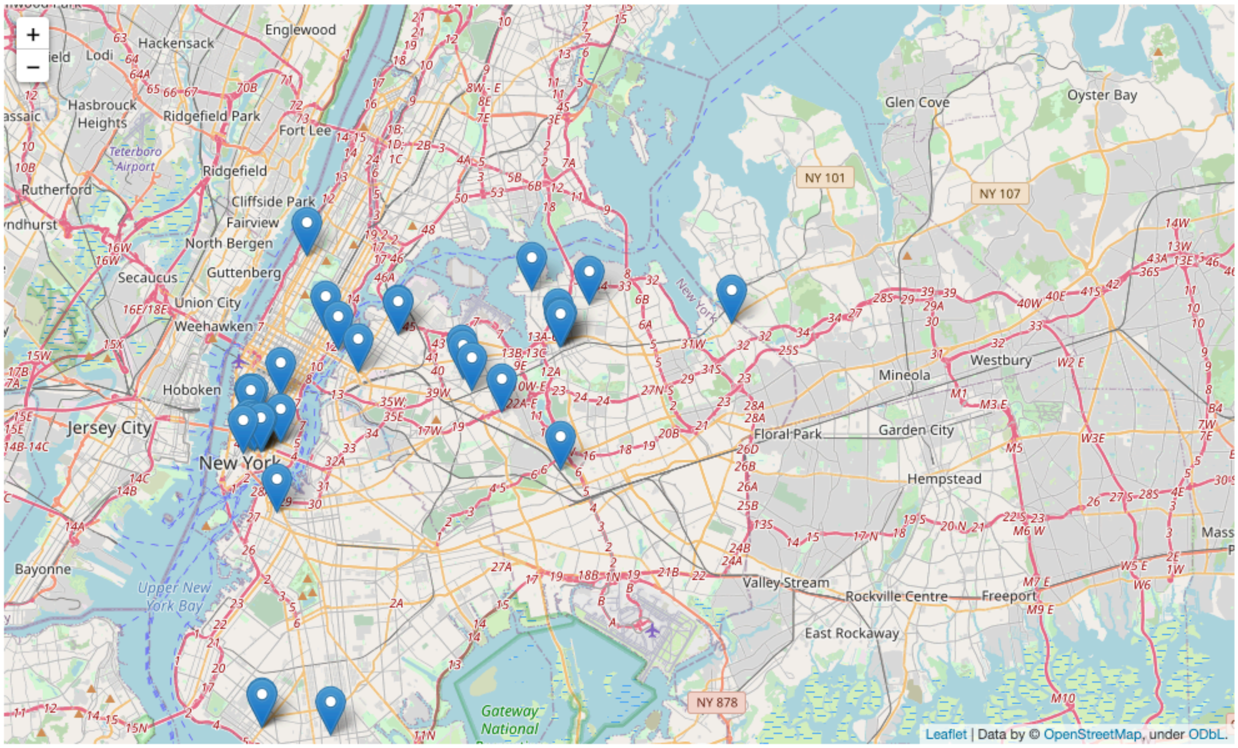


Then the top 100 venues that are in each of the neighborhood within a radius of 500 meters are requested via Foursquare API. The information is collected and transformed into dataframe ny_venues.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	venue id
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop	4c537892fd2ea593cb077a28
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy	4d6af9426107f04dedeb297a
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy	5d5f5044d0ae1c0008f043c3
3	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop	4c783cef3badb1f7e4244b54
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop	4c25c212f1272d7f836385c5
5	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station	4c81a91c51ada1cd87741510
6	Wakefield	40.894705	-73.847201	Subway	40.890468	-73.849152	Sandwich Place	4d33665fb6093704b80001e0
7	Wakefield	40.894705	-73.847201	Central Deli	40.896728	-73.844387	Deli / Bodega	4f32458019836c91c7c734ff
8	Wakefield	40.894705	-73.847201	Louis Pizza	40.898399	-73.848810	Pizza Place	55aa92ac498e24734cd2e378
9	Wakefield	40.894705	-73.847201	Koss Quick Wash	40.891281	-73.849904	Laundromat	5681717c498e9b9cf4d8c187

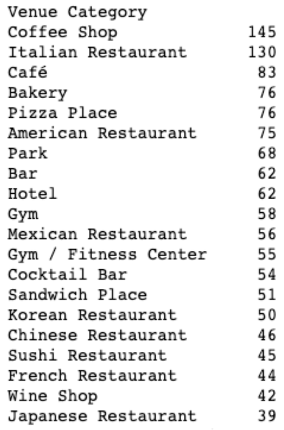
A new dataframe bubble_tea is created as a subset of ny_venues, where "Venue Category" is "bubble tea shop," and to help visualize the location of these bubble tea shops in New York City, an interactive visualization with a Folium map is created. The ratings of these bubble tea shops are requested using Foursquare API with the venue id. The average ratings of these bubble tea shops are 6.9, so there are plenty room to improve.



2) Analyze Manhattan Neighborhoods

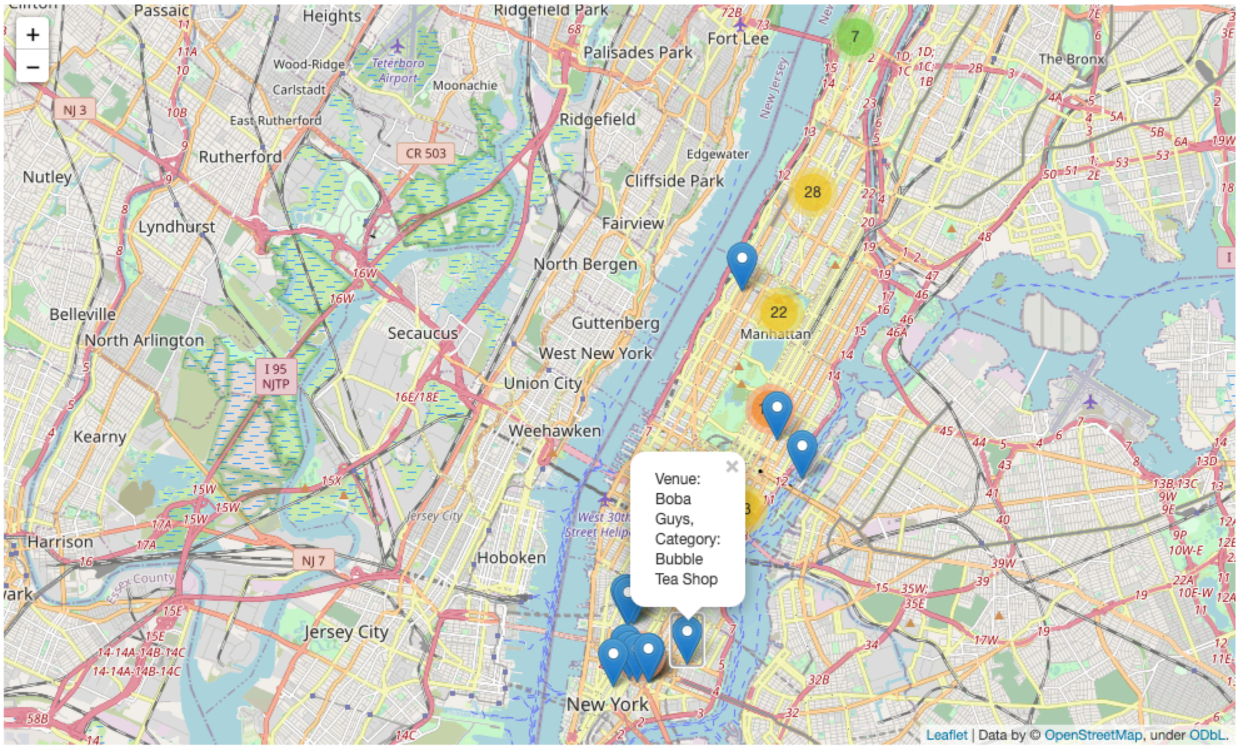
Manhattan is truly the heart of the New York City, with famous landmarks and household-name attractions found all over the island. Looking at the map above, two clusters of bubble tea shops can be easily spotted – one cluster in lower Manhattan, and one cluster in Queens. Manhattan would be my top choice because of its rich cultural, financial, and tourism background. Let's explore Manhattan to see what kind venues there are.

First, let’s glance what venues there are in Manhattan. The distribution of venues in Top 20 categories are:



```
Venue Category
Coffee Shop              145
Italian Restaurant       130
Café                      83
Bakery                    76
Pizza Place               76
American Restaurant       75
Park                      68
Bar                       62
Hotel                     62
Gym                       58
Mexican Restaurant        56
Gym / Fitness Center      55
Cocktail Bar              54
Sandwich Place            51
Korean Restaurant         50
Chinese Restaurant        46
Sushi Restaurant          45
French Restaurant         44
Wine Shop                 42
Japanese Restaurant       39
```

It looks like coffee shop is one of the most prevalent venue categories in Manhattan, along with a variety of restaurants, such as Italian, American, Mexican, and Asian. To make the category broader, restaurants, food trucks or stands, and diners are combined into one “food” group. Bars, wine shops, coffee shops, and bubble tea shops are combined into the “drink” group. Other venues are grouped into “dessert”, “store”, “park”, and “other”. A Folium map is used to visualize “drink” venues with bubble tea shops superimposed on top.

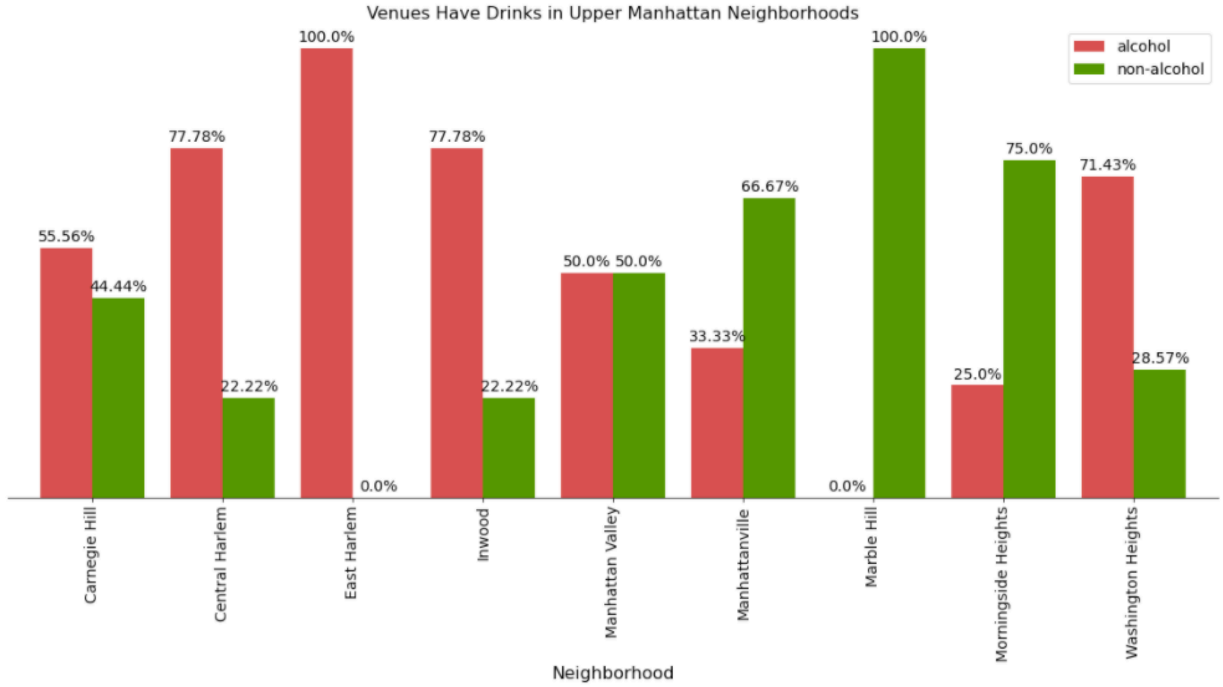


It is clear that most bubble tea shops are clustered in lower Manhattan area, whereas upper Manhattan has very few.

3) Explore Upper Manhattan Neighborhoods

Nine neighborhoods in upper Manhattan (i.e. Marble Hill, Inwood, Washington Heights, Manhattanville, Morningside Heights, Central Harlem, East Harlem, Carnegie Hill, and Manhattan Valley) are then selected. The “drink” venues in these neighborhoods are then grouped into 2 “drink groups”: venues with alcohol drinks (bars, wine shops, etc.) and venues with non-alcohol drinks, such as juice bars. There are 39 venues selling alcohol drinks and 28 venues with non-alcohol drinks.

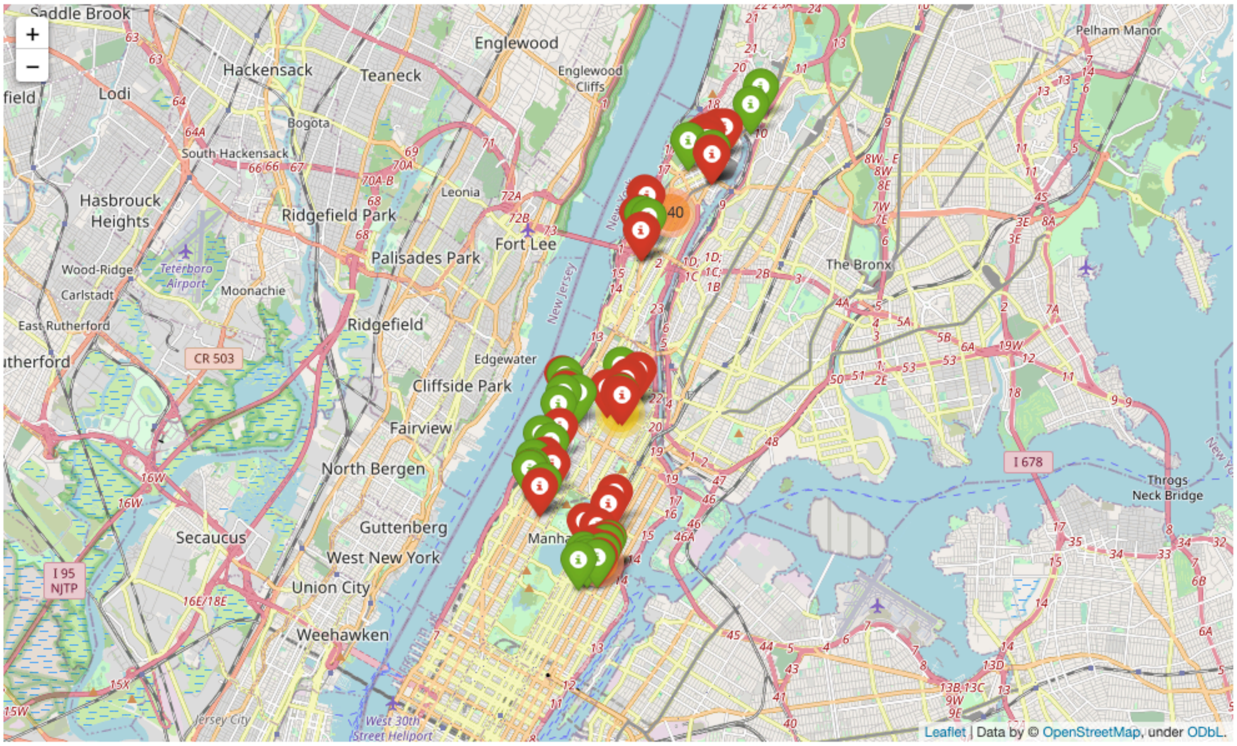
“Drink group” is then processed with one hot encoding to a dummy matrix with each unique value of category into a single attribute. The sum occurrence of each category within each neighborhood is calculated and plotted.



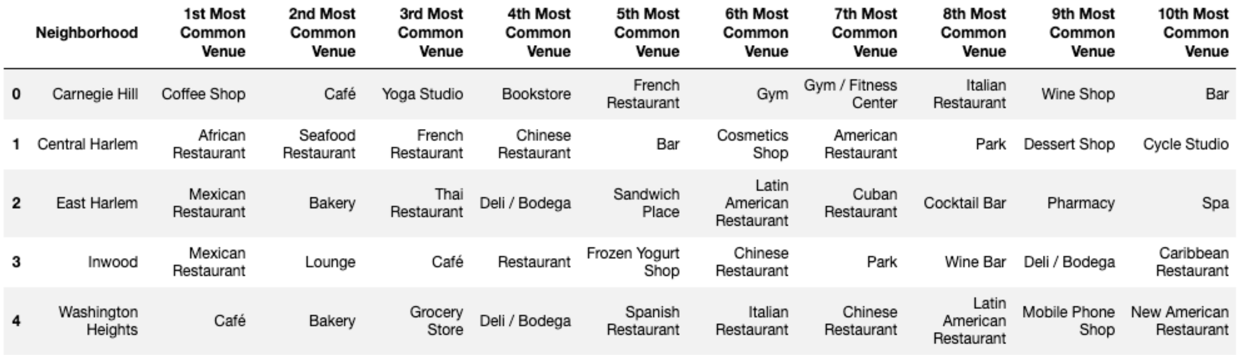
Neighborhoods with less non-alcohol venues are better options than those already saturated with non-alcohol drinks. Therefore, the candidate neighborhoods are narrowed down to Carnegie Hill, Central Harlem, East Harlem, Inwood, and Washington Heights.

4) Explore the Candidate Neighborhoods

A Folium map is generated to visualize the distribution of all venues in these candidate neighborhoods, along with the location of drink venues with alcohol drinks (red icon) and drink venues with non-alcohol drinks (green icon).



“Venue category” is then process with on hot encoding to a dummy matrix with each unique value of category into a single attribute. The sum occurrence of each category within each of these candidate neighborhoods is calculated. The top 10 most frequent venues are shown:



	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	Carnegie Hill	Coffee Shop	Café	Yoga Studio	Bookstore	French Restaurant	Gym	Gym / Fitness Center	Italian Restaurant	Wine Shop	Bar
1	Central Harlem	African Restaurant	Seafood Restaurant	French Restaurant	Chinese Restaurant	Bar	Cosmetics Shop	American Restaurant	Park	Dessert Shop	Cycle Studio
2	East Harlem	Mexican Restaurant	Bakery	Thai Restaurant	Deli / Bodega	Sandwich Place	Latin American Restaurant	Cuban Restaurant	Cocktail Bar	Pharmacy	Spa
3	Inwood	Mexican Restaurant	Lounge	Café	Restaurant	Frozen Yogurt Shop	Chinese Restaurant	Park	Wine Bar	Deli / Bodega	Caribbean Restaurant
4	Washington Heights	Café	Bakery	Grocery Store	Deli / Bodega	Spanish Restaurant	Italian Restaurant	Chinese Restaurant	Latin American Restaurant	Mobile Phone Shop	New American Restaurant

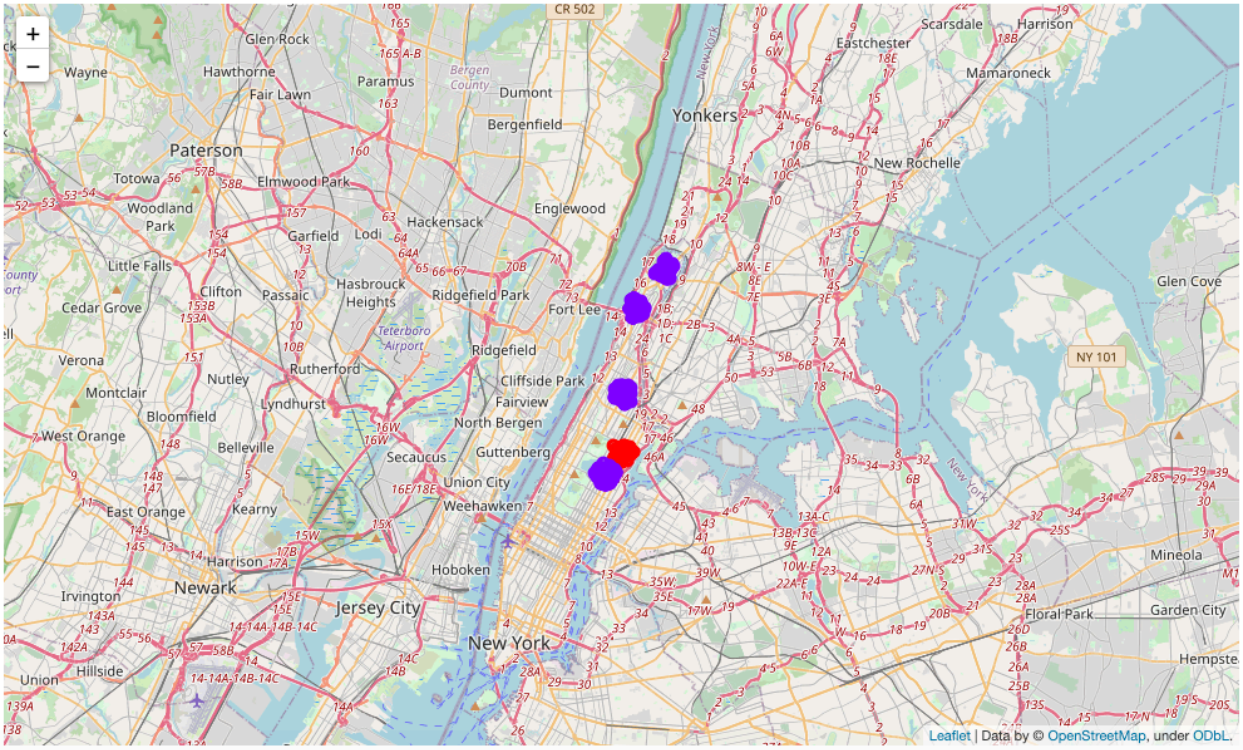
5) **KMeans**

K-means clustering is used as a feature learning step in the unsupervised learning. The goal of this algorithm is to find groups in the data, with the number of groups represented by variable K. The algorithm works iteratively to assign each data point to one of the K groups based on the features that are provided. Data points are clustered based on feature similarity. Therefore, the goal is to cluster the candidate neighborhoods together and find similarities of venue distribution in order to determine the best neighborhood for opening a brand new bubble team shop.

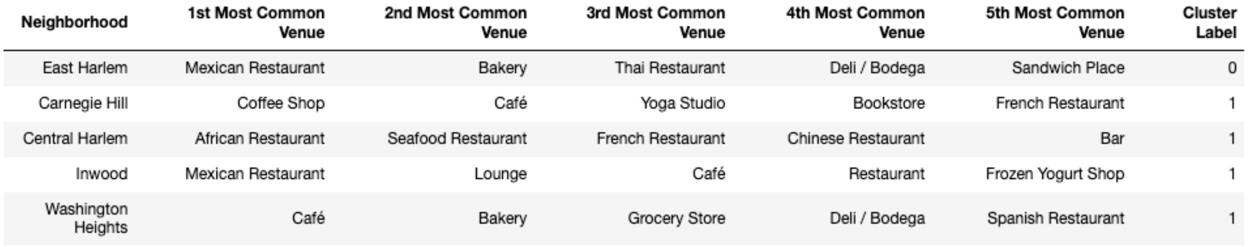
KMeans is performed with kcluster of 2.

4. Results

By clustering the candidate neighborhoods into 2, the distribution of the 2 clusters is shown below in the Folium map with red dots representing cluster 0 and purple dots representing cluster 1.



Four neighborhoods (Carnegie Hill, Central Harlem, Inwood, and Washington Heights) are clustered together based on similarity (purple) and East Harlem stands out by itself. Looking close into each cluster, Top 5 venue categories of each neighborhood is shown in the table below:



Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Cluster Label
East Harlem	Mexican Restaurant	Bakery	Thai Restaurant	Deli / Bodega	Sandwich Place	0
Carnegie Hill	Coffee Shop	Café	Yoga Studio	Bookstore	French Restaurant	1
Central Harlem	African Restaurant	Seafood Restaurant	French Restaurant	Chinese Restaurant	Bar	1
Inwood	Mexican Restaurant	Lounge	Café	Restaurant	Frozen Yogurt Shop	1
Washington Heights	Café	Bakery	Grocery Store	Deli / Bodega	Spanish Restaurant	1

In neighborhoods clustered to "cluster label" 1, at least one kind of drink venue is in the top 5 venue categories. For example, in Carnegie Hill, coffee shops/café are the two most common venues in the neighborhood. On the other hand, East Harlem stands out by itself, ending up in "cluster label" 0, because there is not any kind of drink venue that made to the top 5 common venue list, making it an ideal neighborhood for a bubble tea shop.

5. Discussion

With the help of k-means clustering, East Harlem is identified as the final candidate neighborhood in Manhattan to open a bubble tea shop. It is an ideal neighborhood because the number of venues selling non-alcohol drinks is significant lower than other neighborhoods in Manhattan. East Harlem is one of the largest predominantly Hispanic communities in New York City. Therefore, it is not too surprised that the 1st most common venue in the neighborhood is Mexican Restaurant. Although East Harlem has suffered from high crime rate historically, it has become one of the "New Hot Neighborhoods" in New York City.

To better serve the population in the neighborhood, besides the classic tea and flavors, my bubble tea shop could offer Asian-Hispanic fusion style drinks. For example, agua fresca is very popular among Hispanic population. Adding boba to agua fresca could potentially be a great selling point to Hispanic Americans.

In order to promote my bubble tea and to reach broader population, one option is to partner with food delivery company. In addition, my bubble tea shop can be mobile as well. A mobile bubble tea truck, which serves refreshing boba drinks and delicious snacks, is coming to East Harlem!

6. Conclusion

This project explores the neighborhoods in Manhattan, New York City, in order to start an Asian bubble tea business. The result suggests that East Harlem could be a great neighborhood for the business with a lot of potentials because of its lack of venues offering drinks. If you ever spot a mobile bubble tea truck that serves both classic boba milk tea and fusion horchata boba, don't forget to grab a drink!