Capstone Project - The Battle of Neighborhoods

1. Description of the Problem and Discussion of the Background

Introduction/Business Problem

Due to increasing popularity of drinking bubble tea (a new variety of tea) globally, an investor came to me with an idea of opening a bubble tea shop. Bubble Tea is a Taiwanese tea-based drink invented in Tainan and Taichung in the 1980s. Recipes contain tea of some kind, flavors or milk, as well as sugar. Toppings, such as chewy tapioca balls, popping boba, fruit jelly, grass jelly, agar jelly, and puddings are often added.

Investor's main concern is to get a recommendation of the most suitable location to open bubble tea shop in New York. Furthermore, rental price of the shop is not the main concern of the investors. Hence, we should consider a number of factors that could potentially affect how the business would go and also we have to the right neighborhood. With the help from the tools of data science, we can analyze the neighborhoods and decide the best spot to start the business.

Factors affecting the business

Below are some of the main factors that might affect how well the business goes.

Restaurants nearby

Most people would like to get a cup of bubble tea after they have their lunch or dinner. So it is best to set up the bubble tea shop near restaurants. Foursquare are able to provide us with the data of each neighborhood with the information on their top restaurants and even photos and comments.

Cuisine Type

People from different neighborhoods may have very different tastes in food which in turns affects the likelihood of them buying bubble tea after having a good meal. Hence, Cuisine type also an important factor to consider. Customers who enjoy a specific type of cuisine might also share some common interests and characteristics and these characteristics can be analyzed using the data science tools we have learned.

Demographic of the Neighborhoods and Facilities Nearby

Demographical data affects how well the restaurant runs. It is best to set up bubble tea shop in the area with the Most Frequent venues. For example, in an area near restaurants that have high popularity, near cinemas/theaters or even park and gym where people might want to get some drinks after undergoing training or exercises.

2. Data Preparation

Most of the data used in this project will be taken from Foursquare API. The data are crowd sourced, comprehensive geographical data source. With Foursquare API, we are able to get insight on the most popular venues in each neighborhood, ratings and customer comments on those venues.

For this project, I will only be analyzing neighborhood in Manhattan due to large amount of restaurants, cinemas and large diversity in population. I will also be using the information of these neighborhoods from Foursquare, especially the ratings and rankings for restaurants in different neighborhoods to get knowledge of the preferences of customers in each neighborhood. Their preferences will then help us decide the best location to set up the business.

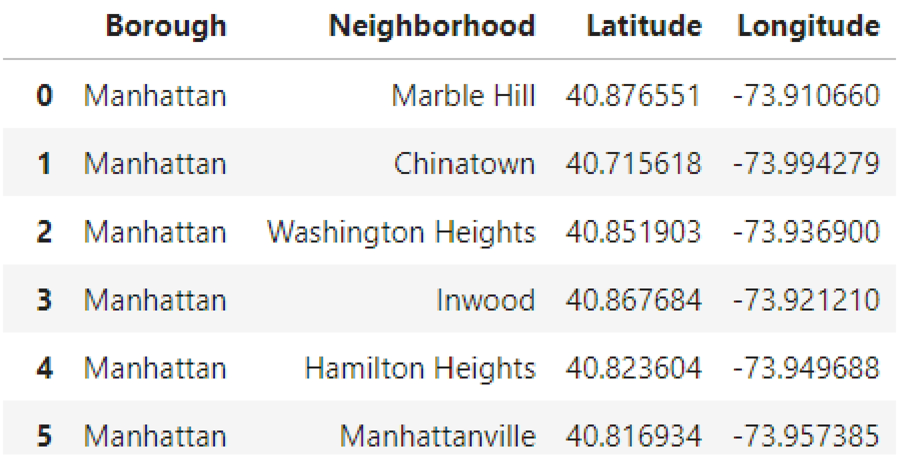
For New York City's Boroughs, Neighborhoods and their coordinate values, we use the below JSON file.

https://cocl.us/new_york_dataset

From the above JSON file, we extract only the Bronx Borough's Neighborhoods.

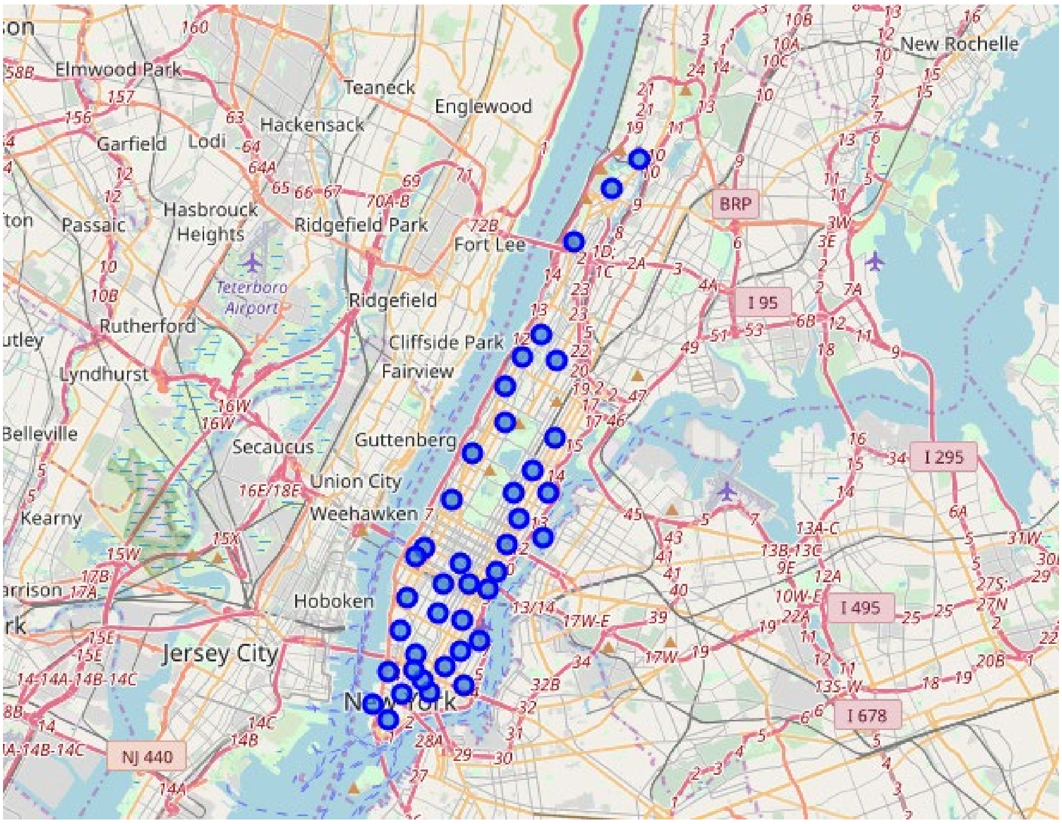
3. Methodology

First, I loaded the geographical data of neighborhoods in all NYC Next, I filter out the Manhattan from the dataframe whole dataset. A snapshot of the dataset looks like this:

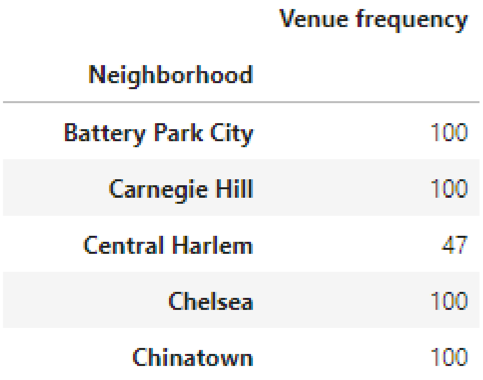


	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688
5	Manhattan	Manhattanville	40.816934	-73.957385

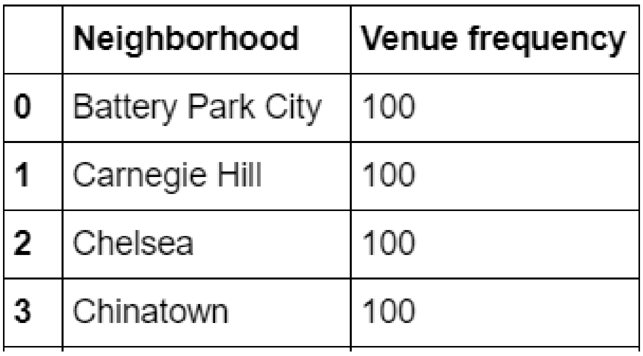
For visualization, we plotted these neighborhoods on the map of Manhattan



With the help of Foursquare API, we are able to explore the neighborhoods and segment them. Having retrieved restaurants' data for each of the neighborhood from FourSquare, we then check how many venues were returned for each neighborhood. To make our analysis easier, we will remove neighborhood with number of Venue Category less than the maximum count which is 100. This is because we are concern with the neighborhood which have quite a lot of activities. For instance, more activities near the neighbourhood (i.e gym training, watching movies in theatre, etc) might result in more people coming to get drinks.

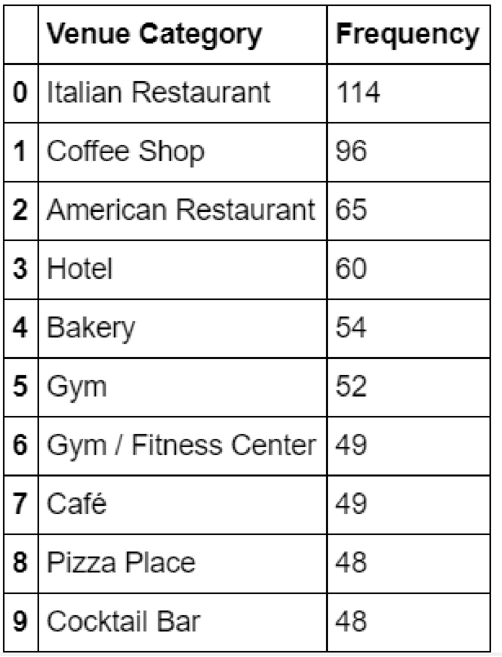


	Venue frequency
Neighborhood	
Battery Park City	100
Carnegie Hill	100
Central Harlem	47
Chelsea	100
Chinatown	100

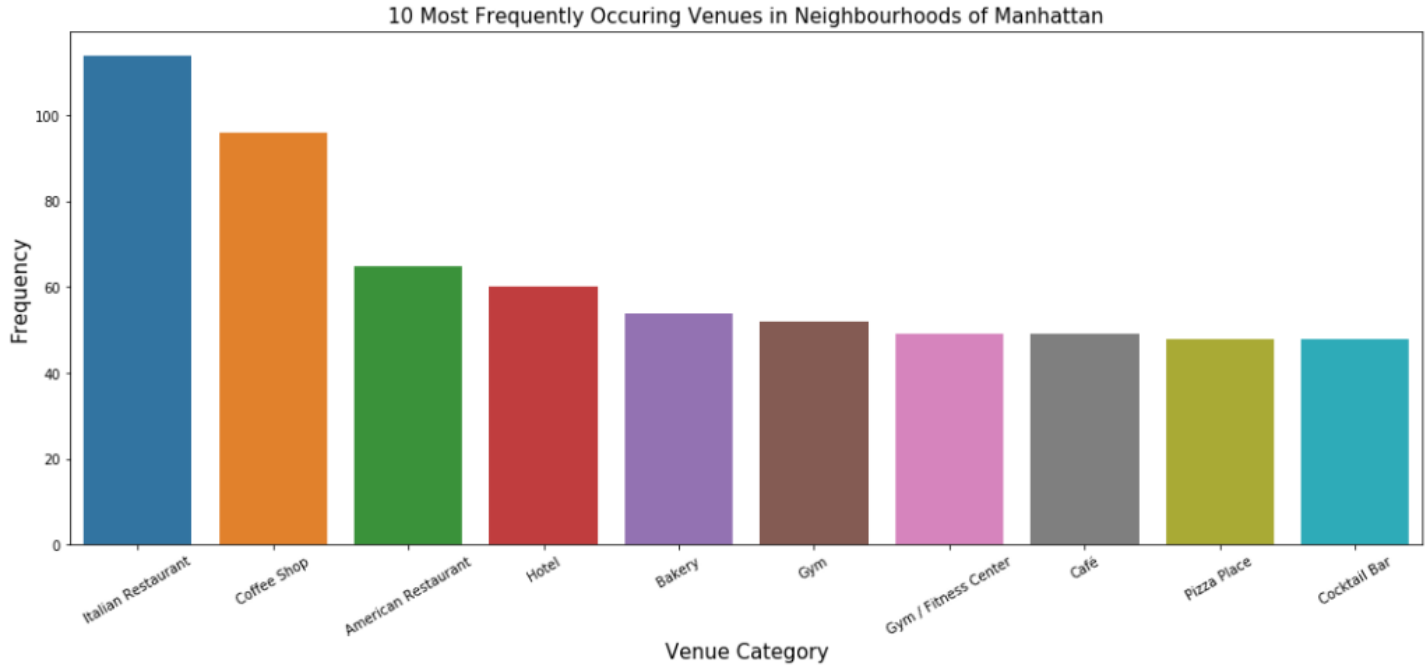


	Neighborhood	Venue frequency
0	Battery Park City	100
1	Carnegie Hill	100
2	Chelsea	100
3	Chinatown	100

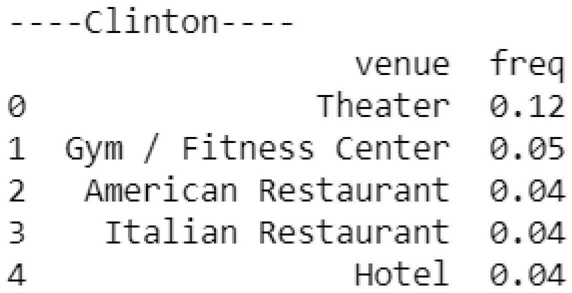
We then create a table that shows the 10 Most Frequently Occuring Venue_Category and also plotted a histogram for better visualization.



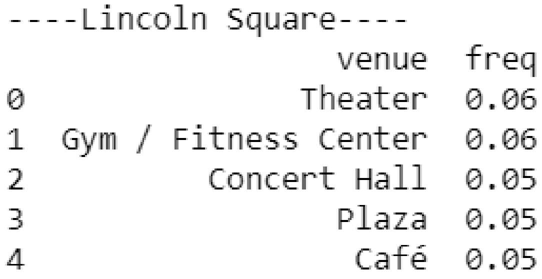
	Venue Category	Frequency
0	Italian Restaurant	114
1	Coffee Shop	96
2	American Restaurant	65
3	Hotel	60
4	Bakery	54
5	Gym	52
6	Gym / Fitness Center	49
7	Café	49
8	Pizza Place	48
9	Cocktail Bar	48



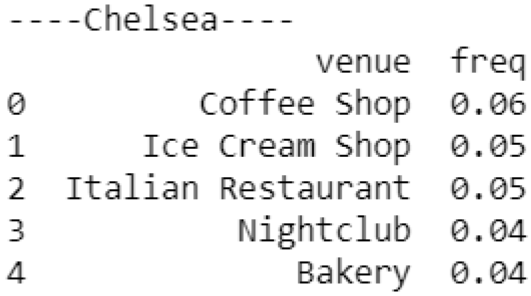
Here we have found out that Italian restaurants top the charts of most common venues in the 5 districts, followed by coffee shop, American Restaurant and Hotel. The Next step was to create the new dataframe and display the top 10 venues for each neighbourhood. A snippet of the dataframe is shown below:



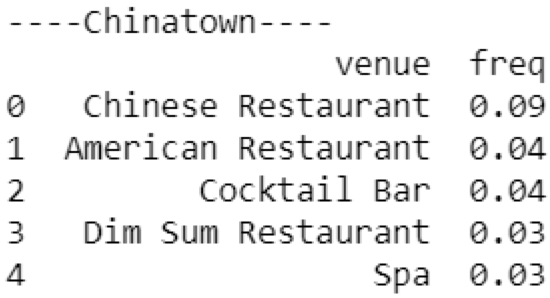
```
----Clinton----
                    venue  freq
0                 Theater  0.12
1  Gym / Fitness Center  0.05
2   American Restaurant  0.04
3    Italian Restaurant  0.04
4                   Hotel  0.04
```



```
----Lincoln Square----
                    venue  freq
0                 Theater  0.06
1  Gym / Fitness Center  0.06
2          Concert Hall  0.05
3                   Plaza  0.05
4                    Café  0.05
```



```
----Chelsea----
                  venue  freq
0           Coffee Shop  0.06
1        Ice Cream Shop  0.05
2  Italian Restaurant  0.05
3             Nightclub  0.04
4                Bakery  0.04
```



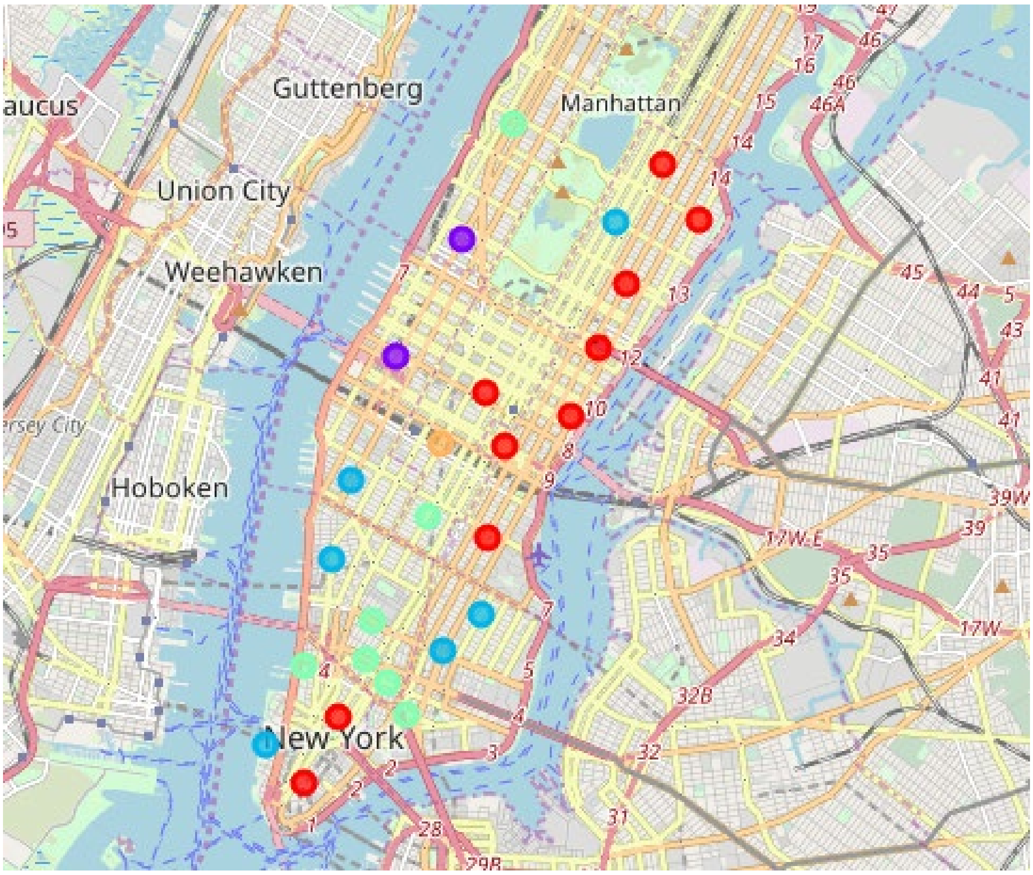
```
----Chinatown----
                   venue  freq
0   Chinese Restaurant  0.09
1  American Restaurant  0.04
2         Cocktail Bar  0.04
3   Dim Sum Restaurant  0.03
4                   Spa  0.03
```

From the above dataframe, we can conjecture that Clinton, Lincoln Square and Chinatown might be a good location to set up the business. This is because most of the activities such as theatre, gym, hotels and restaurants have some correlation with drinking bubble tea. For example, people might tend to buy some drinks after undergoing exhaustive training and exercise or tourist might want to try out the special drinks after they have their lunch or dinner near the restaurant. However, Chelsea might not be a good spot for the business since we will be facing many competitions from coffee shop and ice cream shop. Both coffee and ice cream do not go well with bubble tea.

Finally, we try to cluster the neighbourhoods into 5 clusters based on the frequency of venue categories and by K-Means clustering. So, our expectation would be based on the similarities of venue categories, these districts will be clustered. Using K-Means algorithm rom Scikit-learn library we obtain 3 clusters as shown below.

4. Results

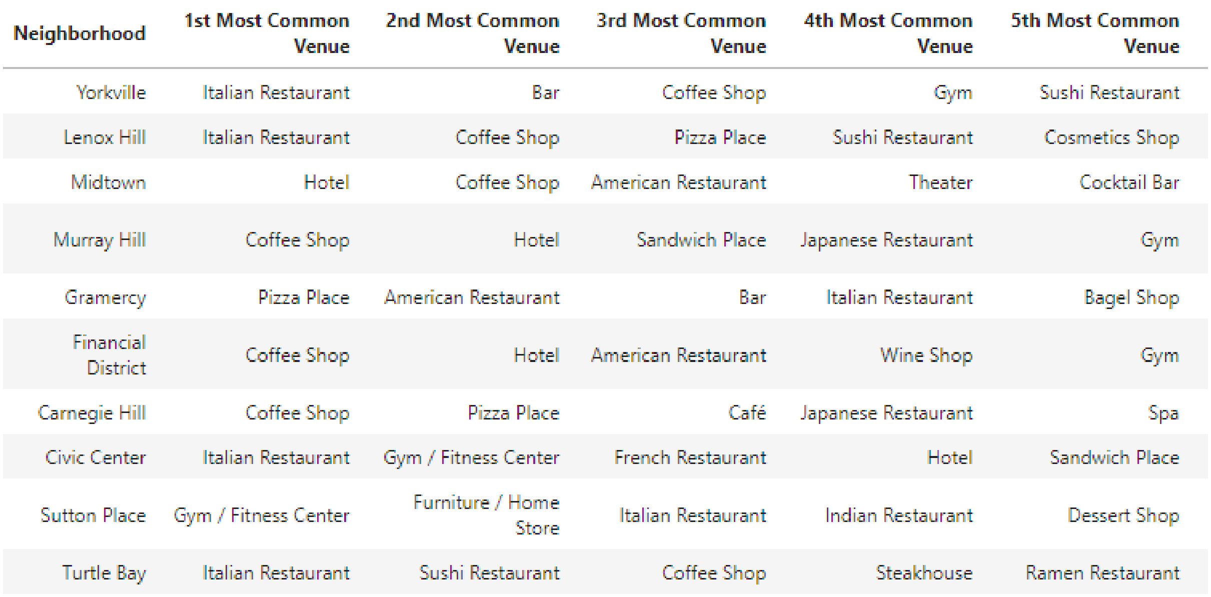
To visualize the clusters, we plotted the following map. The different colours of the big dot represent different clusters



Cluster	Colours
Cluster 1	Red
Cluster 2	Purple
Cluster 3	Blue
Cluster 4	Green
Cluster 5	Orange

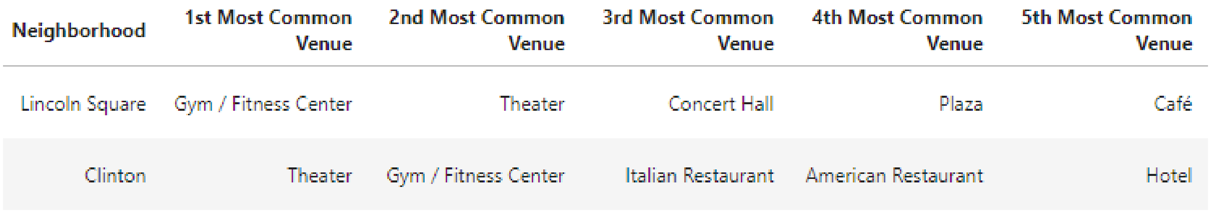
The table below shows the different neighbourhood in different clusters and also the most common venue.

Cluster 1:



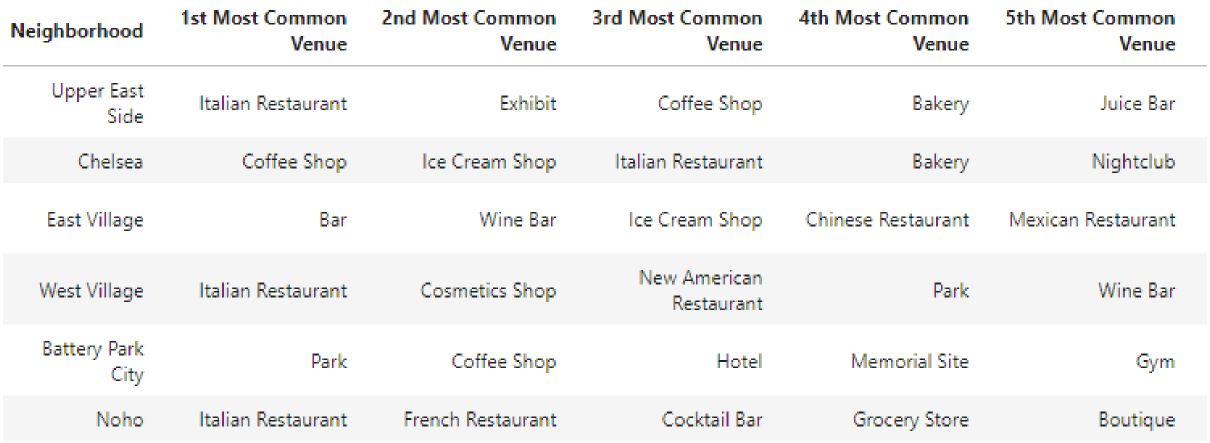
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Yorkville	Italian Restaurant	Bar	Coffee Shop	Gym	Sushi Restaurant
Lenox Hill	Italian Restaurant	Coffee Shop	Pizza Place	Sushi Restaurant	Cosmetics Shop
Midtown	Hotel	Coffee Shop	American Restaurant	Theater	Cocktail Bar
Murray Hill	Coffee Shop	Hotel	Sandwich Place	Japanese Restaurant	Gym
Gramercy	Pizza Place	American Restaurant	Bar	Italian Restaurant	Bagel Shop
Financial District	Coffee Shop	Hotel	American Restaurant	Wine Shop	Gym
Carnegie Hill	Coffee Shop	Pizza Place	Café	Japanese Restaurant	Spa
Civic Center	Italian Restaurant	Gym / Fitness Center	French Restaurant	Hotel	Sandwich Place
Sutton Place	Gym / Fitness Center	Furniture / Home Store	Italian Restaurant	Indian Restaurant	Dessert Shop
Turtle Bay	Italian Restaurant	Sushi Restaurant	Coffee Shop	Steakhouse	Ramen Restaurant

Cluster 2:



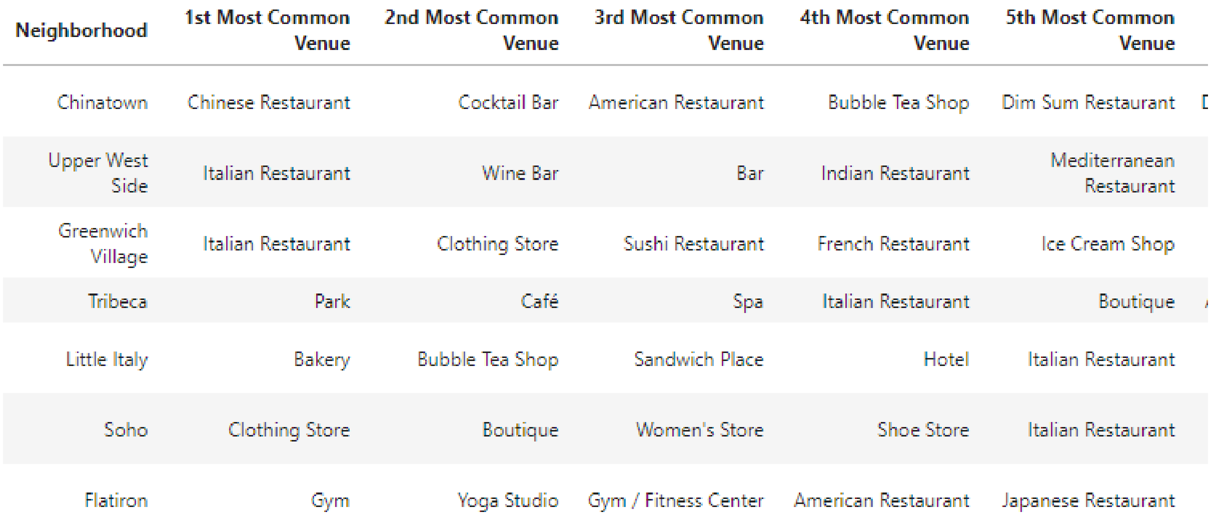
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Lincoln Square	Gym / Fitness Center	Theater	Concert Hall	Plaza	Café
Clinton	Theater	Gym / Fitness Center	Italian Restaurant	American Restaurant	Hotel

Cluster 3:



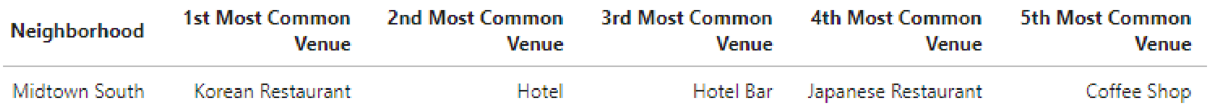
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Upper East Side	Italian Restaurant	Exhibit	Coffee Shop	Bakery	Juice Bar
Chelsea	Coffee Shop	Ice Cream Shop	Italian Restaurant	Bakery	Nightclub
East Village	Bar	Wine Bar	Ice Cream Shop	Chinese Restaurant	Mexican Restaurant
West Village	Italian Restaurant	Cosmetics Shop	New American Restaurant	Park	Wine Bar
Battery Park City	Park	Coffee Shop	Hotel	Memorial Site	Gym
Noho	Italian Restaurant	French Restaurant	Cocktail Bar	Grocery Store	Boutique

Cluster 4:



Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Chinatown	Chinese Restaurant	Cocktail Bar	American Restaurant	Bubble Tea Shop	Dim Sum Restaurant
Upper West Side	Italian Restaurant	Wine Bar	Bar	Indian Restaurant	Mediterranean Restaurant
Greenwich Village	Italian Restaurant	Clothing Store	Sushi Restaurant	French Restaurant	Ice Cream Shop
Tribeca	Park	Café	Spa	Italian Restaurant	Boutique
Little Italy	Bakery	Bubble Tea Shop	Sandwich Place	Hotel	Italian Restaurant
Soho	Clothing Store	Boutique	Women's Store	Shoe Store	Italian Restaurant
Flatiron	Gym	Yoga Studio	Gym / Fitness Center	American Restaurant	Japanese Restaurant

Cluster 5:



Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Midtown South	Korean Restaurant	Hotel	Hotel Bar	Japanese Restaurant	Coffee Shop

From the able table, we can clearly see that our conjecture where that Lincoln Square and Clinton might be the good place to start our bubble tea business is right. They are being clustered into the same group 2 since both the location have common venues that are similar and the venues have some correlation with drinking bubble tea.

On the other hand, cluster 3 might not be an ideal location for our bubble tea business as there are quit a lot of venue which might have very low correlation with drinking bubble tea i.e bar, coffee shop, ice cream shop, nightclub etc. This also suggest that our conjecture about Chelsea not being a good location for the business was right.

Cluster 1 and 4 can be considered to be an acceptable location but is still inferior compared to cluster 2 since we will be facing quite a lot of competition (drinks related) from coffee shop, bar and café .

Cluster 5 can also be considered another good spot for our business targeted to Asians or those with high interest in Asian food due to the presence of Korean and Japanese restaurants. Together with the presence of hotels, we can also assume that most of the tourists are Asians. So, we can make a delicious bubble with taste catered for Asians or people which favours Asians cuisines in this location.

5.Discussion

By clustering the neighborhoods in Manhattan, we were able to tell the taste differences between different neighborhoods and able to find out which neighborhood are similar to other ones. This gave us some insights in regard to choosing the optimal location for our new bubble tea shop. One drawback of this analysis is that the clustering is completely based on the most common venues obtained from Foursquare data

6.Conclusions

With the help of machine learning techniques i.e K-mean Clustering, we were able to merge different neighborhoods into clusters based on their similarities. This also provided insights for the problem we tried to solve in the beginning, that is to suggest a best place to the investor for setting up bubble tea business. By analyzing the clusters, we were able to identify neighbourhood 2 and 5 are the best location for the stakeholder’s business.