# Clusters of New York City

Coursera Capstone Project Report By Xuemei Xu

## Introduction

### Background

As one of the cosmopolitan cities in the world, New York City has the largest population, is the most diverse city, and the highest GDP in the USA. With these characteristics, New York City becomes a dream city to start a new business. In New York City, if people take enough efforts to work hard for their goals, there will always be a path to success.

### Problem

However, before starting any business, vendors need to understand the environment of the neighborhood where they are going to open their businesses, it does not matter if vendors want to open a restaurant, a gym or any other small businesses. For example, which neighborhoods are good to open a gym?

This project will analyze neighbourhoods of 5 boroughs in New York City and help people who want to open a small business find out go-to neighbourhoods whenever they are looking for a specific type of business.

### Stakeholders

Since the project will explore and analyze neighbourhoods in New York City, the target stakeholders can be current residents of New York City, tourists who are planning to come to or currently in New York City, any individual wants to start a new business and organizations that are aiming to conduct any market research about different clusters of neighborhoods.

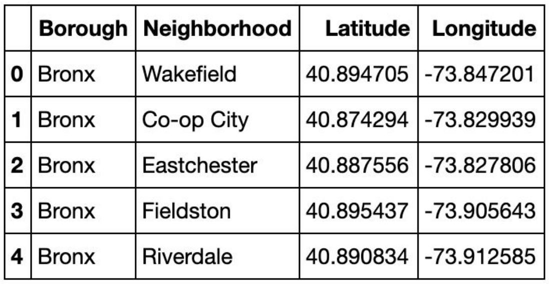
From this project, stakeholders can have an idea of which neighbourhoods are known for what kind of characteristics in order to save their time of researching, and have an overall view of the distribution of different clusters in New York City.

# Data

## Data  Description

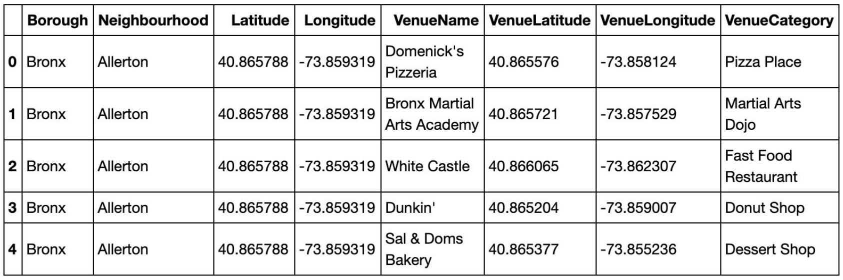
Two datasets will be used in this project.

One is the dataset of New York City boroughs and neighborhoods. This dataset will help to provide information of all neighborhoods and each neighborhood belongs to which borough. From this information, we can know the places we are going to analyze. It’s the basic information the project needs. For example, below is a part of the New York City neighborhoods chart. (Information will be called from newyork_data json file). In the json file, we can also get location data (latitude and longitude) of each neighborhood. Latitude and longitude information can help to explore neighborhoods based on the other dataset mentioned next.



| | 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 |

The other one is the dataset of New York City neighborhoods. This dataset is the main part of this project. Based on the above location data, we can explore restaurants in neighborhoods by using FourSquare API, a location provider. For example, we can explore the neighborhood in Allenton based on its location data. The results will have the venue names, latitudes and longitudes, and categories. After getting the results, we can analyze what kinds of categories have the top frequency in this neighborhood.



| | Borough | Neighbourhood | Latitude | Longitude | VenueName | VenueLatitude | VenueLongitude | VenueCategory |
|---|---|---|---|---|---|---|---|---|
| **0** | Bronx | Allerton | 40.865788 | -73.859319 | Domenick's Pizzeria | 40.865576 | -73.858124 | Pizza Place |
| **1** | Bronx | Allerton | 40.865788 | -73.859319 | Bronx Martial Arts Academy | 40.865721 | -73.857529 | Martial Arts Dojo |
| **2** | Bronx | Allerton | 40.865788 | -73.859319 | White Castle | 40.866065 | -73.862307 | Fast Food Restaurant |
| **3** | Bronx | Allerton | 40.865788 | -73.859319 | Dunkin' | 40.865204 | -73.859007 | Donut Shop |
| **4** | Bronx | Allerton | 40.865788 | -73.859319 | Sal & Doms Bakery | 40.865377 | -73.855236 | Dessert Shop |

## Data Sources

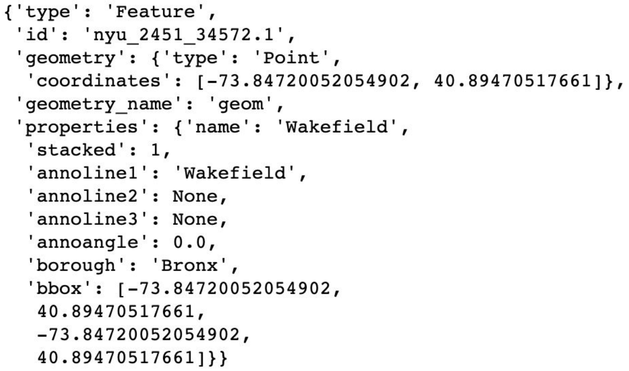
New York City neighborhoods dataset will be retrieved from a json file: 'newyork_data.json' https://cocl.us/new_york_dataset.

Location and restaurants data will be called from FourSquare API: https://foursquare.com/developers/apps

# Methodology

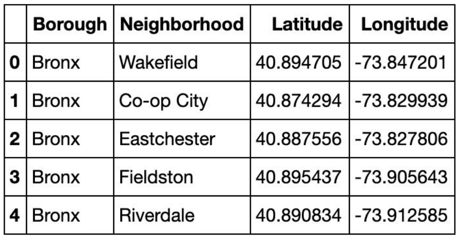
## Exploratory Data Analysis

To achieve the project goal, firstly, we need to download the New York City neighborhoods information. From the previous courses, there is a New York City json dataset, named “new york_data.json” retrieved from https://cocl.us/new_york_dataset. Once getting the dataset, we can analyze how the dataset looks, which is below:



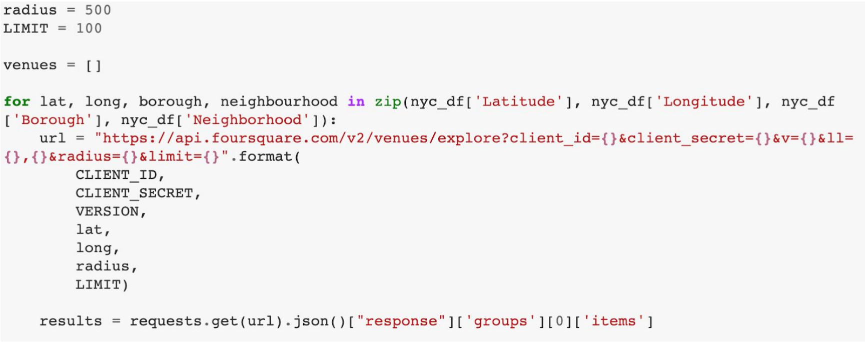
```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```

Since we cannot analyze the data in above format, we need to transform the dataset into pandas dataframe. But not all the information is needed. The main necessary features are borough, neighborhood, latitude and longitude. The transformed and cleaned final data frame looks like below:



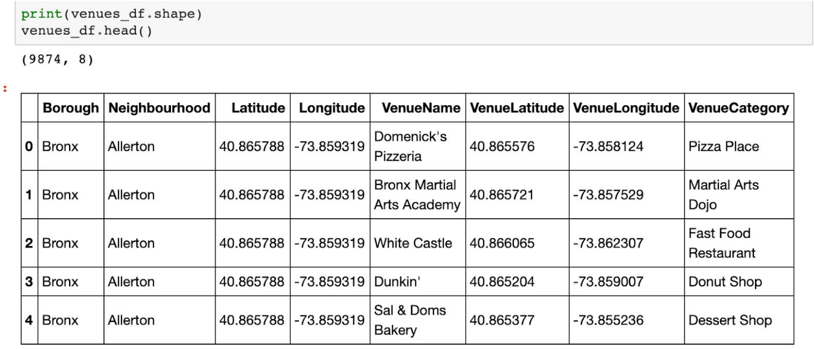
| | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|
| **0** | Bronx | Wakefield | 40.894705 | -73.847201 |
| **1** | Bronx | Co-op City | 40.874294 | -73.829939 |
| **2** | Bronx | Eastchester | 40.887556 | -73.827806 |
| **3** | Bronx | Fieldston | 40.895437 | -73.905643 |
| **4** | Bronx | Riverdale | 40.890834 | -73.912585 |

The next step is to explore each neighborhood and analyze how kinds of places and stores are in the neighborhoods. The results (venues) are called by accessing FourSquare API. To access FourSquare API, client ID and secret are needed to generate URLs. After creating a FourSquare API account, each account has its unique ID and secret. Below is how the venues are called.



```
radius = 500
LIMIT = 100

venues = []

for lat, long, borough, neighbourhood in zip(nyc_df['Latitude'], nyc_df['Longitude'], nyc_df
['Borough'], nyc_df['Neighborhood']):
    url = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll=
{},{}&radius={}&limit={}".format(
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        lat,
        long,
        radius,
        LIMIT)

    results = requests.get(url).json()["response"]['groups'][0]['items']
```

Total called venues are 9,874 in 306 neighborhoods, 5 boroughs. The results are also in json format and need to be transformed into pandas data frame to analyze. The major features we need here are venues' names, latitude, longitude and categories. After the transforming and cleaning processes, examples look like below:

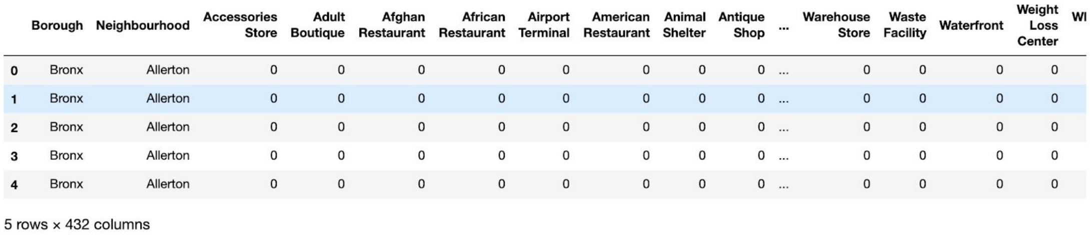


```
print(venues_df.shape)
venues_df.head()
```

```
(9874, 8)
```

| | Borough | Neighbourhood | Latitude | Longitude | VenueName | VenueLatitude | VenueLongitude | VenueCategory |
|---|---|---|---|---|---|---|---|---|
| 0 | Bronx | Allerton | 40.865788 | -73.859319 | Domenick's Pizzeria | 40.865576 | -73.858124 | Pizza Place |
| 1 | Bronx | Allerton | 40.865788 | -73.859319 | Bronx Martial Arts Academy | 40.865721 | -73.857529 | Martial Arts Dojo |
| 2 | Bronx | Allerton | 40.865788 | -73.859319 | White Castle | 40.866065 | -73.862307 | Fast Food Restaurant |
| 3 | Bronx | Allerton | 40.865788 | -73.859319 | Dunkin' | 40.865204 | -73.859007 | Donut Shop |
| 4 | Bronx | Allerton | 40.865788 | -73.859319 | Sal & Doms Bakery | 40.865377 | -73.855236 | Dessert Shop |

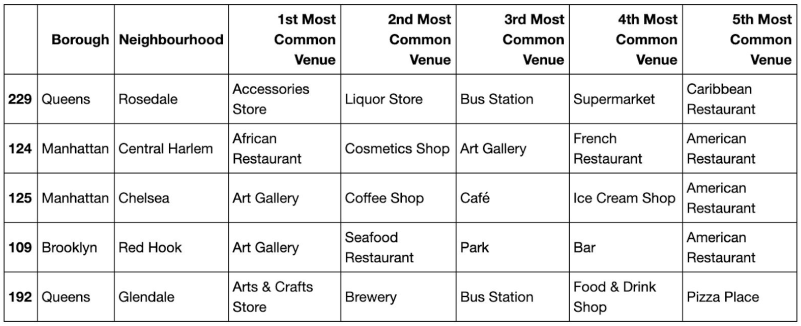
Finally, to analyze venues' categories in the next step, one hot function is used to transform strings to numeric variables. By using numeric variables, it's easier for us to count the frequency of each category in each neighborhood. As below, each category is converted to integer.



| | Borough | Neighbourhood | Accessories Store | Adult Boutique | Afghan Restaurant | African Restaurant | Airport Terminal | American Restaurant | Animal Shelter | Antique Shop | ... | Warehouse Store | Waste Facility | Waterfront | Weight Loss Center | Wl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Bronx | Allerton | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | |
| **1** | Bronx | Allerton | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | |
| **2** | Bronx | Allerton | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | |
| **3** | Bronx | Allerton | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | |
| **4** | Bronx | Allerton | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | |

5 rows × 432 columns

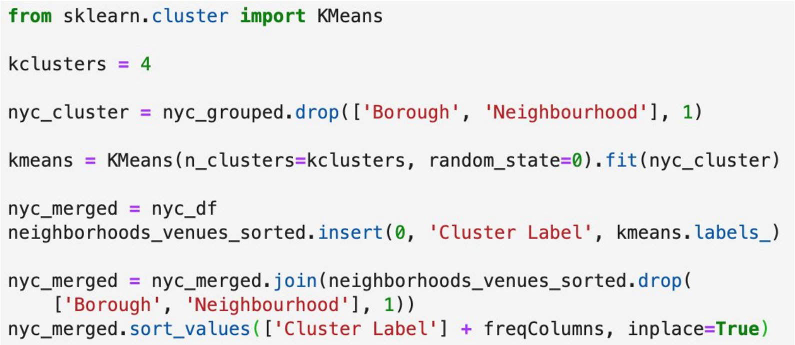
Then, we can count the categories that have the top frequency. Due to high volume of result, we decided to retrieve the top five categories in each neighborhood as below:



| | Borough | Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|---|
| **229** | Queens | Rosedale | Accessories Store | Liquor Store | Bus Station | Supermarket | Caribbean Restaurant |
| **124** | Manhattan | Central Harlem | African Restaurant | Cosmetics Shop | Art Gallery | French Restaurant | American Restaurant |
| **125** | Manhattan | Chelsea | Art Gallery | Coffee Shop | Café | Ice Cream Shop | American Restaurant |
| **109** | Brooklyn | Red Hook | Art Gallery | Seafood Restaurant | Park | Bar | American Restaurant |
| **192** | Queens | Glendale | Arts & Crafts Store | Brewery | Bus Station | Food & Drink Shop | Pizza Place |

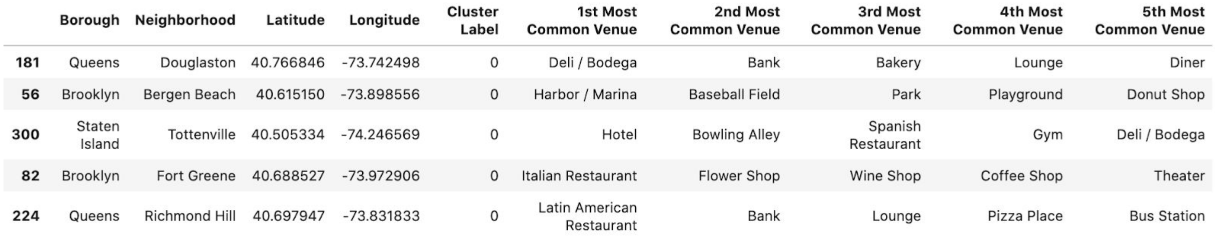
## Machine Learning Model

With the above chart, we are ready to prepare all the data we need to analyze neighborhoods in New York City. In order to segment neighborhoods into different clusters, we used K-means clustering, which is an unsupervised machine learning algorithm. Based on labs we did in previous courses, the number of clusters sets is 4. Below is how we generated clusters.



```
from sklearn.cluster import KMeans

kclusters = 4

nyc_cluster = nyc_grouped.drop(['Borough', 'Neighbourhood'], 1)

kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(nyc_cluster)

nyc_merged = nyc_df
neighborhoods_venues_sorted.insert(0, 'Cluster Label', kmeans.labels_)

nyc_merged = nyc_merged.join(neighborhoods_venues_sorted.drop(
    ['Borough', 'Neighbourhood'], 1))
nyc_merged.sort_values(['Cluster Label'] + freqColumns, inplace=True)
```

After generating the clusters, we need to insert the cluster results to the chart of our venues_df data frame so that we can clearly visualize each neighborhood belonging to which cluster, as below:

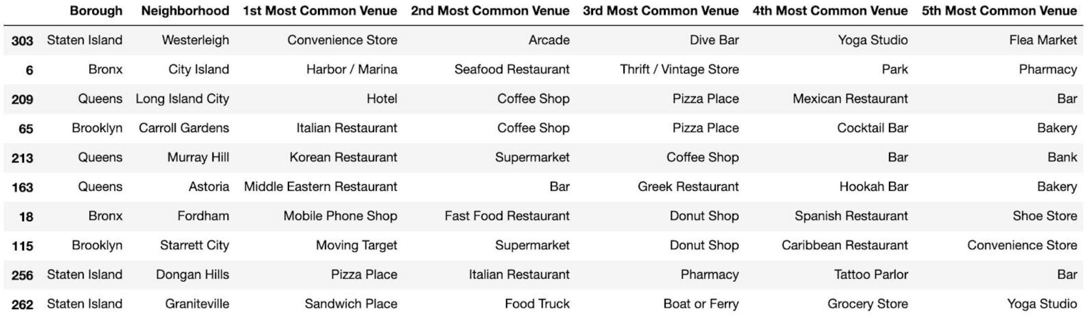


| | Borough | Neighborhood | Latitude | Longitude | Cluster Label | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|
| **181** | Queens | Douglaston | 40.766846 | -73.742498 | 0 | Deli / Bodega | Bank | Bakery | Lounge | Diner |
| **56** | Brooklyn | Bergen Beach | 40.615150 | -73.898556 | 0 | Harbor / Marina | Baseball Field | Park | Playground | Donut Shop |
| **300** | Staten Island | Tottenville | 40.505334 | -74.246569 | 0 | Hotel | Bowling Alley | Spanish Restaurant | Gym | Deli / Bodega |
| **82** | Brooklyn | Fort Greene | 40.688527 | -73.972906 | 0 | Italian Restaurant | Flower Shop | Wine Shop | Coffee Shop | Theater |
| **224** | Queens | Richmond Hill | 40.697947 | -73.831833 | 0 | Latin American Restaurant | Bank | Lounge | Pizza Place | Bus Station |

# Results

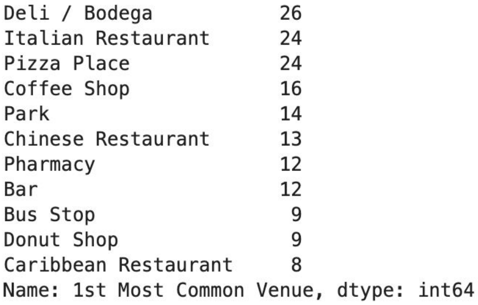
Based on above information, New York City Neighborhoods are segmented into 4 different clusters below.

Cluster 0 (Red)
From “1st common venue” to “5th common venue”, except some restaurants, there are various kinds of places for entertainment,  such as beach, bowling alley, lounge, theater, bar, night club, etc. These neighborhoods can be summarized as entertainment areas, which most likely weekend-to-go neighborhoods.



| | Borough | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|---|
| **303** | Staten Island | Westerleigh | Convenience Store | Arcade | Dive Bar | Yoga Studio | Flea Market |
| **6** | Bronx | City Island | Harbor / Marina | Seafood Restaurant | Thrift / Vintage Store | Park | Pharmacy |
| **209** | Queens | Long Island City | Hotel | Coffee Shop | Pizza Place | Mexican Restaurant | Bar |
| **65** | Brooklyn | Carroll Gardens | Italian Restaurant | Coffee Shop | Pizza Place | Cocktail Bar | Bakery |
| **213** | Queens | Murray Hill | Korean Restaurant | Supermarket | Coffee Shop | Bar | Bank |
| **163** | Queens | Astoria | Middle Eastern Restaurant | Bar | Greek Restaurant | Hookah Bar | Bakery |
| **18** | Bronx | Fordham | Mobile Phone Shop | Fast Food Restaurant | Donut Shop | Spanish Restaurant | Shoe Store |
| **115** | Brooklyn | Starrett City | Moving Target | Supermarket | Donut Shop | Caribbean Restaurant | Convenience Store |
| **256** | Staten Island | Dongan Hills | Pizza Place | Italian Restaurant | Pharmacy | Tattoo Parlor | Bar |
| **262** | Staten Island | Graniteville | Sandwich Place | Food Truck | Boat or Ferry | Grocery Store | Yoga Studio |

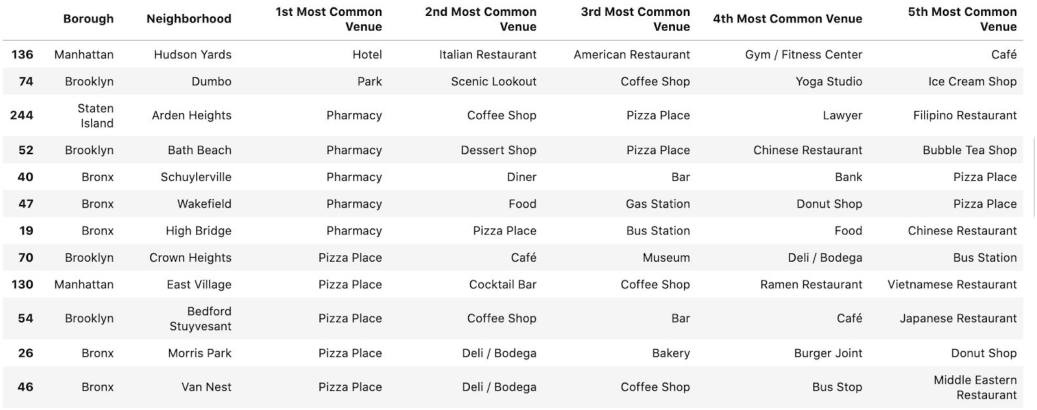
Cluster 1 (Purple)
Cluster 1 has the largest volume of the results. In order to analyze this cluster better, we calculated the top 10 frequently occurred categories as below:



```
Deli / Bodega           26
Italian Restaurant      24
Pizza Place             24
Coffee Shop             16
Park                    14
Chinese Restaurant      13
Pharmacy                12
Bar                     12
Bus Stop                 9
Donut Shop               9
Caribbean Restaurant     8
Name: 1st Most Common Venue, dtype: int64
```

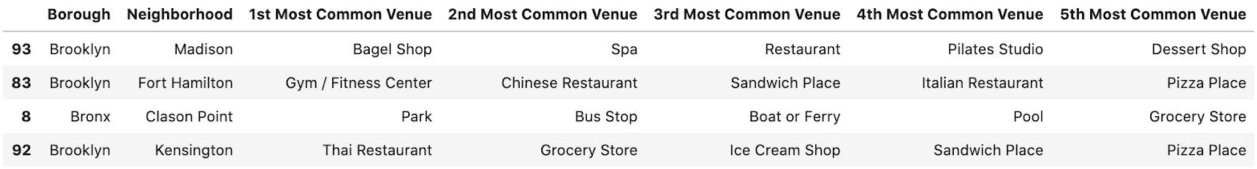
Except the most commonly occurred restaurants, there are mainly coffee shops, parks, deli, donut shops and pharmacies. All kinds of categories occur evenly in this cluster, which can be regarded as neighborhoods for living and business. These neighborhoods are active neighborhoods in each borough.

Cluster 2 (Light Blue)
In cluster 2, there are various kinds of food. For example, there are restaurants in different cuisines, such as ramen, steakhouse, spanish food, mexican food, breakfast stores, etc. For breakfast, there are bakeries, cafes, coffee shops, and donut shops. This cluster’s categories are mainly food.



| | Borough | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|---|
| 136 | Manhattan | Hudson Yards | Hotel | Italian Restaurant | American Restaurant | Gym / Fitness Center | Café |
| 74 | Brooklyn | Dumbo | Park | Scenic Lookout | Coffee Shop | Yoga Studio | Ice Cream Shop |
| 244 | Staten Island | Arden Heights | Pharmacy | Coffee Shop | Pizza Place | Lawyer | Filipino Restaurant |
| 52 | Brooklyn | Bath Beach | Pharmacy | Dessert Shop | Pizza Place | Chinese Restaurant | Bubble Tea Shop |
| 40 | Bronx | Schuylerville | Pharmacy | Diner | Bar | Bank | Pizza Place |
| 47 | Bronx | Wakefield | Pharmacy | Food | Gas Station | Donut Shop | Pizza Place |
| 19 | Bronx | High Bridge | Pharmacy | Pizza Place | Bus Station | Food | Chinese Restaurant |
| 70 | Brooklyn | Crown Heights | Pizza Place | Café | Museum | Deli / Bodega | Bus Station |
| 130 | Manhattan | East Village | Pizza Place | Cocktail Bar | Coffee Shop | Ramen Restaurant | Vietnamese Restaurant |
| 54 | Brooklyn | Bedford Stuyvesant | Pizza Place | Coffee Shop | Bar | Café | Japanese Restaurant |
| 26 | Bronx | Morris Park | Pizza Place | Deli / Bodega | Bakery | Burger Joint | Donut Shop |
| 46 | Bronx | Van Nest | Pizza Place | Deli / Bodega | Coffee Shop | Bus Stop | Middle Eastern Restaurant |

Cluster 3 (Yellow)
Cluster 3 has the fewest results in four clusters. The most categories belong to exercises. For example, except some food related categories, there are gym/fitness center, pool and pilates studio. These neighborhoods can be summarized as gym centers.

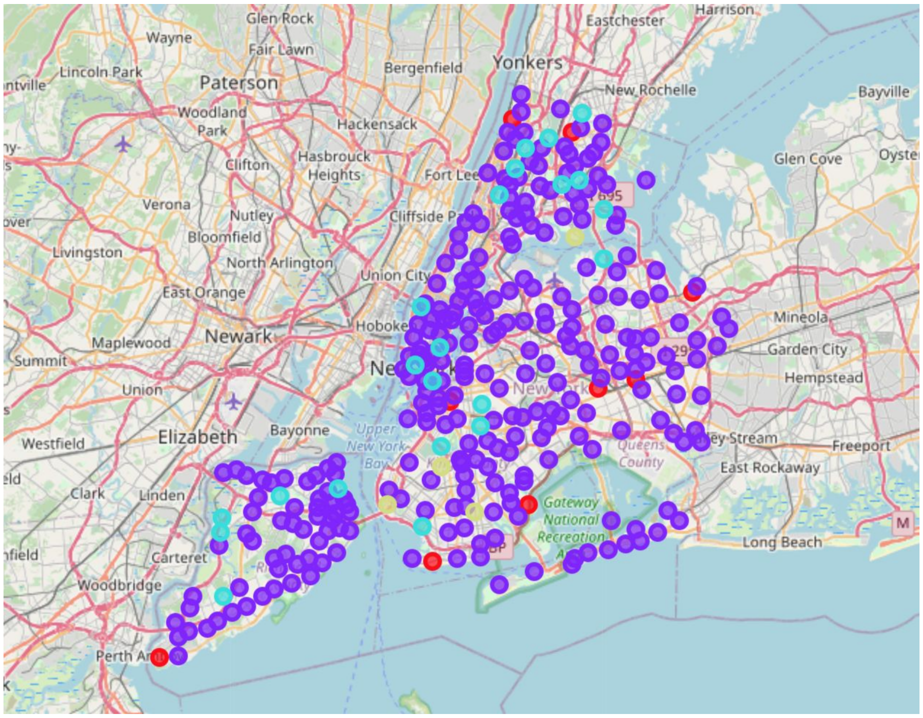


| | Borough | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|---|
| 93 | Brooklyn | Madison | Bagel Shop | Spa | Restaurant | Pilates Studio | Dessert Shop |
| 83 | Brooklyn | Fort Hamilton | Gym / Fitness Center | Chinese Restaurant | Sandwich Place | Italian Restaurant | Pizza Place |
| 8 | Bronx | Clason Point | Park | Bus Stop | Boat or Ferry | Pool | Grocery Store |
| 92 | Brooklyn | Kensington | Thai Restaurant | Grocery Store | Ice Cream Shop | Sandwich Place | Pizza Place |

# Discussion

## Observation

To visualize how each cluster is distributed, we used Folium library to map four clusters at top of New York City map as below:



Neighborhoods in cluster 1 (purple) are evenly spreaded in each borough. In these neighborhoods, categories are integrated. It can be considered as residential areas, and also can be seen as business centers. There are food places, parks, public transportations, grocery stores, shopping centers, etc.

Cluster 0 (red) is mainly located in the border of each borough, which is mostly next to the river or ocean and accords with American lifestyle. Most neighborhoods in cluster 3 (yellow) are in Brooklyn borough, which means residents in Brooklyn are more likely to work out.

Neighborhoods in cluster 2 (light blue) are located in four different boroughs. Queens borough has no cluster 2 neighborhoods. Queens borough has a large population and can be a potential opportunity for new food businesses.

## Recommendation

Due to a large population, each borough in New York City has lots of active living and businesses areas. Most small businesses have strong advantages to succeed. Further research needs to be conducted to serve people who are active in each neighborhood.

In Manhattan, there is no specific neighborhood known for fitness and entertainment. In Queens, there are no neighborhoods specifically famous for food. Staten Island has no fitness neighborhoods. These can be an opportunity for new related businesses. Brooklyn is the most integrated borough. For some experienced vendors, Brooklym can be a great choice to compete with other existing businesses. For residents and tourists, Brooklyn is also a good borough to stay.

# Conclusion

Neighborhoods of five boroughs in New York City are segmented into four clusters. Due to diversity and large population, most neighborhoods are segmented in the same cluster, cluster 1. The neighborhoods in this cluster are combined with food places, entertainment places, fitness centers, grocery stores, etc. Each category in each neighborhood is serving its specific customer segments. The main point to successfully start a new business in this cluster is to understand active people in each neighborhood. This needs further market research and consumer preferences.

However, in the other three clusters, there are three corresponding businesses that can be practiced and get results of a high possibility of success. Neighborhoods in cluster 0 are a good choice to open businesses for entertainment. Residents and tourists also can visit these neighborhoods for entertainment; cluster 2 is a good place for food, and cluster 3 is known for exercise and fitness businesses.