Exploring NYC Neighborhoods by Food Culture

JULY 4

Jakhongir Kobilov

Introduction

New York City is the main subject of this project, so it makes sense to take the time for a brief introduction to this truly magnetic and exhilarating place in our world. So, New York City is the most populous city in the US. The City has been recognized as the cultural, financial, and media capital of the world, significantly influencing world diplomacy, commerce, entertainment, science, technology, education, politics, tourism, art, fashion, sports, and so on.

New York City is truly a global symbol of freedom, and the greatest cultural hub in the US, has welcomed millions of immigrants long before the days of Ellis Island. Because of the different migrations from different places, bringing their own native dishes, NYC has become a place of great diversity. Without all the immigrant cultures brought here, NYC would not be a multicultural city, like how it is today.

Besides, the City is also a hotspot for both domestic and international tourism. In 2018 alone, NYC has welcomed over 65 million visitors. It is no secret that for any avid traveler, trying local cuisine is always at the top of their to-do list. If you don’t taste the local cuisine of the place you are traveling to, then you have never been there after all! No doubt, NYC’s food diversity is so rich that it will never disappoint.

“If you don’t taste the local cuisine of the place you are traveling to, then you have never been there after all!”

Problem statement

A good friend of mine is an entrepreneur and investor from another country who happens to be willing to invest in New York City's hospitality. Since I reside in NYC, he asked me for assistance with identifying what kind of restaurant to establish and which neighborhood to pick.

Undoubtedly, this is not an easy task, given New York City is comprised of five boroughs that geographically make up over 300 square miles. Within this area, 59 community districts define the economic profile of the City. As a result, there are so many unique neighborhoods that contribute to its demographic and cultural diversity.

This project will help to understand my entrepreneur friend in the understanding of the diversity of neighborhoods by leveraging venue data from Foursquare’s Places API and K-means clustering machine learning algorithm. Exploratory Data Analysis will help to discover further about the cultural diversities of NYC's neighborhoods. Also, this project can be used by food vendors that are willing to open a new restaurant.

Data

Following data sources will be used to examine the business problem:

1. New York City Dataset.
This New York City Neighborhood Names point file was created as a guide to New York City's neighborhoods that appear on the web resource, "New York: A City of Neighborhoods." Link: https://geo.nyu.edu/catalog/nyu_2451_34572

2. Foursquare API.
A location data provider Foursquare API will be used to make RESTful API calls to retrieve data about venues in different neighborhoods. Venues retrieved from all the neighborhoods are categorized broadly into "Arts & Entertainment", "College & University", "Event", "Food", "Nightlife Spot", "Outdoors & Recreation", etc.

Methodology

For this project, we need a dataset to segment the neighborhoods of New York City. The dataset should contain all the five boroughs and all the neighborhoods of each borough, with respective latitude and longitude coordinates. The dataset that matches our requirement was downloaded using the earlier mentioned URL.

After the .json file is downloaded, it is analyzed to understand the structure of the file. A python dictionary is returned by the URL and all the relevant data is found to be in the features key, which is basically a list of the neighborhoods. The dictionary is transformed into a pandas dataframe by looping through the data and filling the dataframe rows one at a time using the following depicted loop:

```
for data in neighborhoods_data:
    borough = data['properties']['borough']
    neighborhood_name = data['properties']['name']
    neighborhood_latlon = data['geometry']['coordinates']
    neighborhood_lat = neighborhood_latlon[1]
    neighborhood_lon = neighborhood_latlon[0]

    neighborhoods = neighborhoods.append({'Borough': borough,
                                          'Neighborhood': neighborhood_name,
                                          'Latitude': neighborhood_lat,
                                          'Longitude': neighborhood_lon}, ignore_index=True)
```

This creates a dataframe with Borough, Neighborhood, Latitude and Longitude details of the New York City's neighborhood:

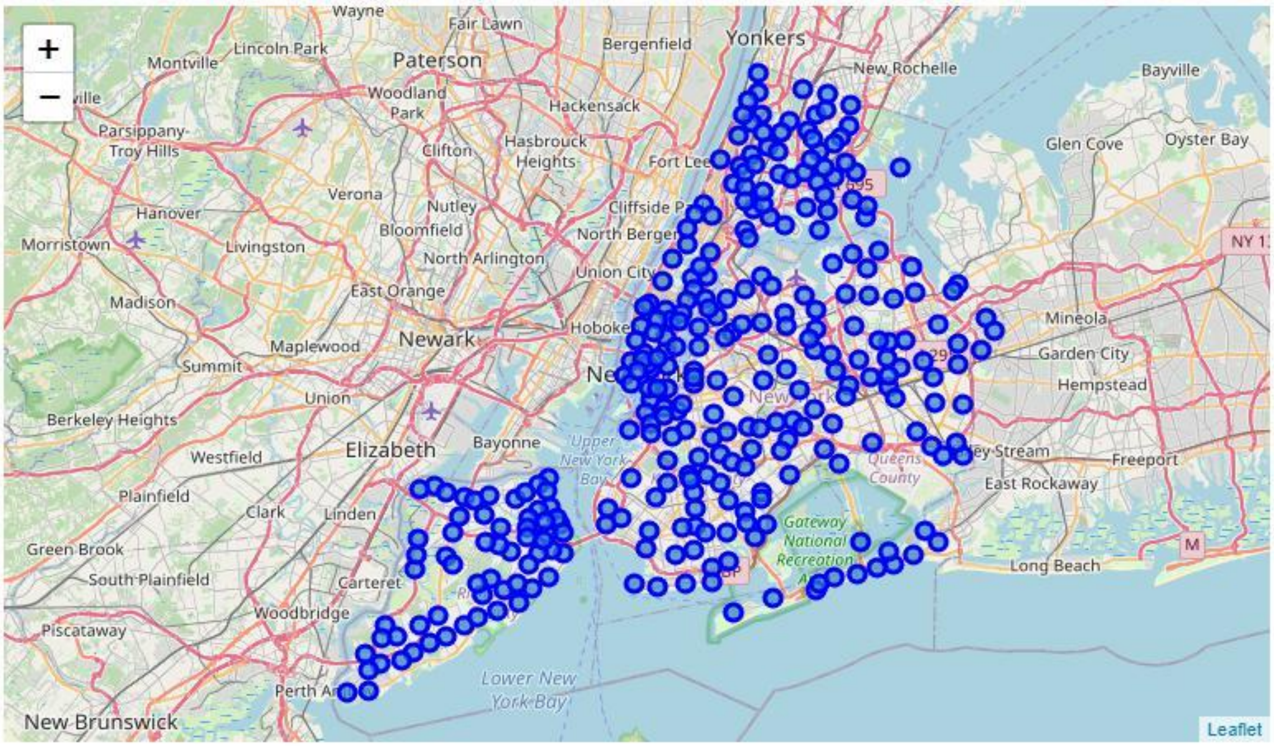
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Upon analysis, it is determined that the dataframe has 5 boroughs and 306 neighborhoods:

```
print('The dataframe has {} boroughs and {} neighborhoods.'.format(
        len(neighborhoods['Borough'].unique()),
        neighborhoods.shape[0]
    )
)
```

```
The dataframe has 5 boroughs and 306 neighborhoods.
```

Geopy library is used to get the latitude (40.7127281) and longitude (-74.0060152) coordinates of New York City. Then, the curated dataframe is then used to visualize by creating a map of New York City with neighborhoods superimposed on top. This map was generated using the ‘folium’ library:



RESTful API Calls to Foursquare

The Foursquare API is used to explore the neighborhoods and segment them. To access the API, ‘CLIENT_ID’, ‘CLIENT_SECRET’, and ‘VERSION’ are defined.
In this project, we are exploring NYC’s cuisines, so we are focusing on ‘Food’ category of GET requests. Foursquare Venue Category Hierarchy is retrieved using the following code:

Fetch Foursquare Venue Category Hierarchy

```
url = 'https://api.foursquare.com/v2/venues/categories?&client_id={}&client_secret={}&v={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION)
category_results = requests.get(url).json()
```

Upon analysis, there are 10 major or parent categories of venues, under which all the other sub-categories are included. Following depiction shows the 'Category ID' and 'Category Name' retrieved from API:

```
for data in category_list:
    print(data['id'], data['name'])
```

```
4d4b7104d754a06370d81259 Arts & Entertainment
4d4b7105d754a06372d81259 College & University
4d4b7105d754a06373d81259 Event
4d4b7105d754a06374d81259 Food
4d4b7105d754a06376d81259 Nightlife Spot
4d4b7105d754a06377d81259 Outdoors & Recreation
4d4b7105d754a06375d81259 Professional & Other Places
4e67e38e036454776db1fb3a Residence
4d4b7105d754a06378d81259 Shop & Service
4d4b7105d754a06379d81259 Travel & Transport
```

As mentioned earlier, the category of 'FOOD' is our matter of interest so, a function created to return a dictionary with 'Category ID' and 'Category Name' of 'Food' and its sub-categories.
To understand the result of GET Request better, the first neighborhood of the 'New York City' dataset is explored. The first neighborhood returned is 'Wakefield' with coordinates of 40.89 and -73.85.
Then, a GET request URL is created to search for Venue with proper category id for 'Food' and the radius was set to 500 meters

```
LIMIT = 1 # limit of number of venues returned by Foursquare API
radius = 500 # define radius
categoryId = '4d4b7105d754a06374d81259' # category ID for "Food"

# create URL

url = 'https://api.foursquare.com/v2/venues/search?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&c
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    neighborhood_latitude,
    neighborhood_longitude,
    radius,
    categoryId,
    LIMIT)
url # display URL
```

```
'https://api.foursquare.com/v2/venues/search?&client_id=LTVMCWFSTUNWWXKA2IPM2W0D1S5GVZ2IXSSMTJIPKC3HZO43&cl
ient_secret=QN11TFW5BC55D2UHIMO1ETQEBTMPFWTQ15EJ5HAYKYFFHUC0&v=20180605&ll=40.89470517661,-73.8472005205490
2&radius=500&categoryId=4d4b7105d754a06374d81259&limit=1'
```

The request returned the ‘Category Name’ of the venue as ‘Carvel Ice Cream’ which is in ‘Food’ category:

```
# Send the GET request and examine the resutls

results = requests.get(url).json()
results['response']['venues']
```


```
[{'id': '4c783cef3badb1f7e4244b54',
  'name': 'Carvel Ice Cream',
  'location': {'address': '1006 E 233rd St',
   'lat': 40.890486685759605,
   'lng': -73.84856772568665,
   'labeledLatLngs': [{'label': 'display',
     'lat': 40.890486685759605,
     'lng': -73.84856772568665},
    {'label': 'entrance', 'lat': 40.890438, 'lng': -73.848559}],
   'distance': 483,
   'postalCode': '10466',
   'cc': 'US',
   'city': 'Bronx',
   'state': 'NY',
   'country': 'United States',
   'formattedAddress': ['1006 E 233rd St',
    'Bronx, NY 10466',
    'United States']},
  'categories': [{'id': '4bf58dd8d48988d1c9941735',
    'name': 'Ice Cream Shop',
    'pluralName': 'Ice Cream Shops',
    'shortName': 'Ice Cream',
    'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/icecream_',
     'suffix': '.png'},
    'primary': True}],
  'referralId': 'v-1593208627',
  'hasPerk': False}]
```


The category name of the venue 'Carvel Ice Cream' is 'Food'.

The aim is to segment the neighborhoods of NYC with respect to the ‘Food’ in its neighborhood, it is further required to fetch this data from all the 306 neighborhoods’ venues. To make the task shorter, the ‘getNearbyFood’ function is created. What this function does is, it loops through all the neighborhoods of NYC and creates an API request URL with radius = 500, LIMIT = 100 (maximum number of nearby venues returned).

Pickle

Pickle is an easy to use at the same time very important library. It is used to serialize information retrieved from GET requests, to make a persistent .pkl file. This file can later be deserialized to retrieve an exact python object structure. This is a crucial step as it will counter any redundant requests to the Foursquare API, which is chargeable over the threshold limits.

Let's use pickle library to serialize the information retrieved from GET requests. This step will counter any redundant requests to the Foursquare API.

```
import pickle # to serialize and deserialize a Python object structure
try:
    with open('nyc_food_venues.pkl', 'rb') as f:
        nyc_venues = pickle.load(f)
    print("---Dataframe Existed and Deserialized---")
except:
    nyc_venues = getNearbyFood(names=neighborhoods['Neighborhood'],
                                    latitudes=neighborhoods['Latitude'],
                                    longitudes=neighborhoods['Longitude']
                                   )

    with open('nyc_food_venues.pkl', 'wb') as f:
        pickle.dump(nyc_venues, f)
    print("---Dataframe Created and Serialized---")
```

```
---Dataframe Existed and Deserialized---
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Central Deli	40.896728	-73.844387	Deli / Bodega
1	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898083	-73.850259	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Wakefield Deli	40.901998	-73.846910	Deli / Bodega
3	Wakefield	40.894705	-73.847201	Popeyes Louisiana Kitchen	40.889322	-73.843323	Fried Chicken Joint
4	Wakefield	40.894705	-73.847201	McDonald's	40.892779	-73.857473	Fast Food Restaurant

Up to this point, to python 'dataframe' are created:

1. 'neighborhood' which contains the Borough, Neighborhood, Latitude, and longitude values of NYC's neighborhoods
2. 'nyc_venues' which is a merger between 'neighborhoods' dataframe and its 'Food' category venues searched with 'Radius' = 500 meters and 'Limit' = 100. Also, each venue has its own Longitude, Latitude and Category.

Exploratory Data Analysis

The merged dataframe 'nyc_venues' contains all the required information. The size of this dataframe is determined, and it is found that there are 13724 venues in total.

```
print(nyc_venues.shape)
nyc_venues.head()
```

```
(13724, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Central Deli	40.896728	-73.844387	Deli / Bodega
1	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898083	-73.850259	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Wakefield Deli	40.901998	-73.846910	Deli / Bodega
3	Wakefield	40.894705	-73.847201	Popeyes Louisiana Kitchen	40.889322	-73.843323	Fried Chicken Joint
4	Wakefield	40.894705	-73.847201	McDonald's	40.892779	-73.857473	Fast Food Restaurant

Next, we need to find out how many unique categories can be curated from all the venues. There are 190 such categories

```
print('There are {} uniques categories.'.format(len(nyc_venues['Venue Category'].unique())))
nyc_venues.groupby('Venue Category')['Venue Category'].count().sort_values(ascending=False)
```

```
There are 190 uniques categories.
Venue Category
Deli / Bodega                   1266
Pizza Place                     1084
Coffee Shop                      938
Chinese Restaurant               684
Donut Shop                       644
Fast Food Restaurant             601
Bakery                           585
Italian Restaurant               447
Bagel Shop                       397
Café                             379
Mexican Restaurant               376
Ice Cream Shop                   332
Sandwich Place                   325
Caribbean Restaurant             322
Fried Chicken Joint              310
American Restaurant              305
Food                             242
```

Data Cleaning

Since the purpose of this project is to understand the cultural diversity of a neighborhood by clustering it categorically, using the venues' categories. Thus, it is important to remove all the venues from the 'dataframe' which have generalized categories (ex: Coffee shop, Café, etc.).

```
# manually create a list of generalized categories
general_categories = ['Dessert Shop','Food','Ice Cream Shop','Donut Shop','Bakery','Sandwich Place','Comfort Food Restaurant',
                      'Deli / Bodega','Food Truck','Bagel Shop','Burger Joint','Restaurant','Frozen Yogurt Shop','Coffee Shop',
                      'Diner','Wings Joint','Café','Juice Bar','Breakfast Spot','Grocery Store','Bar','Cupcake Shop',
                      'Pub','Fish & Chips Shop','Cafeteria','Other Nightlife','Arcade','Hot Dog Joint','Food Court',
                      'Health Food Store','Convenience Store','Food & Drink Shop','Cocktail Bar','Cheese Shop',
                      'Snack Place','Sports Bar','Lounge','Theme Restaurant','Buffet','Bubble Tea Shop','Building',
                      'Irish Pub','College Cafeteria','Tea Room','Supermarket','Hotpot Restaurant','Gastropub','Beer Garden',
                      'Fish Market','Beer Bar','Clothing Store','Music Venue','Bistro','Salad Place','Wine Bar','Gourmet Shop',
                      'Indie Movie Theater','Art Gallery','Gift Shop','Pie Shop','Fruit & Vegetable Store',
                      'Street Food Gathering','Dive Bar','Factory','Farmers Market','Mac & Cheese Joint','Creperie',
                      'Candy Store','Event Space','Skating Rink','Miscellaneous Shop','Gas Station','Organic Grocery',
                      'Pastry Shop','Club House','Flea Market','Hotel','Furniture / Home Store','Bookstore','Pet Café',
                      'Gym / Fitness Center','Flower Shop','Financial or Legal Service','Hotel Bar','Hookah Bar','Poke Place',
                      'Market','Gluten-free Restaurant','Smoothie Shop','Butcher','Food Stand','Beach Bar','Beach',
                      'Soup Place','Rock Club','Residential Building (Apartment / Condo)','Laundry Service',
                      'Government Building','Bowling Alley','Nightclub','Park','Moving Target']
```

Next, we subtract 'unique_categories' and "general_categories.' It leaves all the categories we need for further analysis:

```
# fetch all the required food categories
food_categories =  list(set(unique_categories) - set(general_categories))
print(', '.join(str(x) for x in food_categories))
```

```
Afghan Restaurant, South Indian Restaurant, Szechuan Restaurant, Thai Restaurant, Food Service, Lebanese Restaurant, Mexican Restaurant, Australian Restaurant, Russian Restaurant, Fried Chicken Joint, Venezuelan Restaurant, Cajun / Creole
Restaurant, Spanish Restaurant, Beer Store, Italian Restaurant, Jewish Restaurant, Vietnamese Restaurant, Taiwanese Restaurant, Ukrainian Restaurant, Brazilian Restaurant, Caucasian Restaurant, Tex-Mex Restaurant, Ramen Restaurant, Portug
uese Restaurant, Seafood Restaurant, Japanese Restaurant, Hobby Shop, Salvadoran Restaurant, Colombian Restaurant, Peruvian Restaurant, Record Shop, Austrian Restaurant, Pizza Place, Cuban Restaurant, Kebab Restaurant, Dim Sum Restaurant,
Modern Greek Restaurant, Hawaiian Restaurant, Moroccan Restaurant, Steakhouse, Turkish Restaurant, Yemeni Restaurant, Souvlaki Shop, Arepa Restaurant, Indian Restaurant, Paella Restaurant, Persian Restaurant, Falafel Restaurant, Sri Lanka
n Restaurant, Indoor Play Area, Egyptian Restaurant, Chocolate Shop, Modern European Restaurant, Mediterranean Restaurant, Cantonese Restaurant, Sushi Restaurant, Whisky Bar, Latin American Restaurant, French Restaurant, Taco Place, Ameri
can Restaurant, Southern / Soul Food Restaurant, German Restaurant, Filipino Restaurant, Noodle House, Poutine Place, Japanese Curry Restaurant, Korean Restaurant, Fast Food Restaurant, Argentinian Restaurant, Caribbean Restaurant, BBQ Jo
int, Eastern European Restaurant, Asian Restaurant, South American Restaurant, Middle Eastern Restaurant, Shanghai Restaurant, Israeli Restaurant, Empanada Restaurant, Greek Restaurant, Malay Restaurant, Pakistani Restaurant, Dosa Place,
Kosher Restaurant, African Restaurant, Varenyky restaurant, Vegetarian / Vegan Restaurant, New American Restaurant, North Indian Restaurant, Chinese Restaurant, Tapas Restaurant, Polish Restaurant, Dumpling Restaurant, English Restaurant,
Halal Restaurant, Indian Chinese Restaurant, Burrito Place, Bike Shop
```

The result is:

```
nyc_venues = nyc_venues[nyc_venues['Venue Category'].isin(food_categories)].reset_index()
nyc_venues.head(5)
```

	index	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	1	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898083	-73.850259	Caribbean Restaurant
1	3	Wakefield	40.894705	-73.847201	Popeyes Louisiana Kitchen	40.889322	-73.843323	Fried Chicken Joint
2	4	Wakefield	40.894705	-73.847201	McDonald's	40.892779	-73.857473	Fast Food Restaurant
3	5	Wakefield	40.894705	-73.847201	McDonald's	40.902645	-73.849485	Fast Food Restaurant
4	8	Wakefield	40.894705	-73.847201	McDonald's	40.889435	-73.843369	Fast Food Restaurant

Upon examining the unique categories, it is found that there are only 98 of them, as compared to 190 earlier. That means, almost 52% of the data was just a noise for the analysis. This essential step, data cleaning, helped to capture the data points of interest.

Feature Engineering

Now, each neighborhood is analyzed individually to understand the most common cuisine being served withing 500 meters of each neighborhood. This process is taken forth by using 'one hot encoding' function of python 'pandas' library. The function converts the categorical variables ('Venue Category') into a form that could be provided to Machine Learning algorithms to do a better job in prediction.

```
# one hot encoding
nyc_onehot = pd.get_dummies(nyc_venues[['Venue Category']], prefix="", prefix_sep="")
nyc_onehot.head()
```

	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	Beer Store	Bike Shop	Brazilian Restaurant	Burrito Place	Cajun / Creole Restaurant	Cantonese Restaurant	Caribbean Restaurant	Caucasian Restaurant	Chinese Restaurant	Chocolate Shop	Colombian Restaurant	Cuban Restaurant	Dim Sum Restaurant	Dosa Place	Dumpling Restaurant	Eastern European Restaurant
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Upon analyzing the size of data points all together, it was found that there are around 6493 data points in total.

Next, number of venues of each category in each neighborhood are counted.

```
venue_counts = nyc_onehot.groupby('Neighborhood').sum()
venue_counts.head(5)
```

Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	Beer Store	Bike Shop	Brazilian Restaurant	Burrito Place	Cajun / Creole Restaurant	Cantonese Restaurant	Caribbean Restaurant	Caucasian Restaurant	Chinese Restaurant	Chocolate Shop	Colombian Restaurant	Cuban Restaurant	Dim Sum Restaurant	Dosa Place	Dumpling Restaurant
Allerton	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5	0	0	0	0	0	0
Annadale	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Arden Heights	0	0	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Arlington	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
Arrochar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

The first five neighborhoods of the dataframe, ‘Annadale,’ ‘Arden Heights,’ and ‘Arlington’ has 3,3,2 ‘American Restaurant’ within its 500 meters proximity.

The top 10 ‘Venue Categories’ can also be found by counting their occurrences. This analysis is depicted below which shows that ‘Korean Restaurant,’ ‘Chinese Restaurant,’ ‘Caribbean Restaurant,’ ‘Italian Restaurant,’ ‘Fast Food Restaurant’ are among top 5:

```
venue_counts_described = venue_counts.describe().transpose()
venue_top10 = venue_counts_described.sort_values('max', ascending=False)[0:10]
venue_top10
```

	count	mean	std	min	25%	50%	75%	max
Korean Restaurant	302.0	0.238411	1.426883	0.0	0.0	0.0	0.0	22.0
Chinese Restaurant	302.0	2.264901	1.887895	0.0	1.0	2.0	3.0	13.0
Caribbean Restaurant	302.0	1.066225	2.393704	0.0	0.0	0.0	1.0	13.0
Italian Restaurant	302.0	1.480132	1.760947	0.0	0.0	1.0	2.0	11.0
Fast Food Restaurant	302.0	1.990066	1.941820	0.0	0.0	2.0	3.0	11.0
Indian Restaurant	302.0	0.311258	0.894319	0.0	0.0	0.0	0.0	10.0
Pizza Place	302.0	3.589404	1.995915	0.0	2.0	3.0	5.0	9.0
Seafood Restaurant	302.0	0.533113	0.906178	0.0	0.0	0.0	1.0	8.0
Fried Chicken Joint	302.0	1.026490	1.314105	0.0	0.0	1.0	2.0	7.0
Asian Restaurant	302.0	0.500000	0.857835	0.0	0.0	0.0	1.0	6.0

Data Visualization

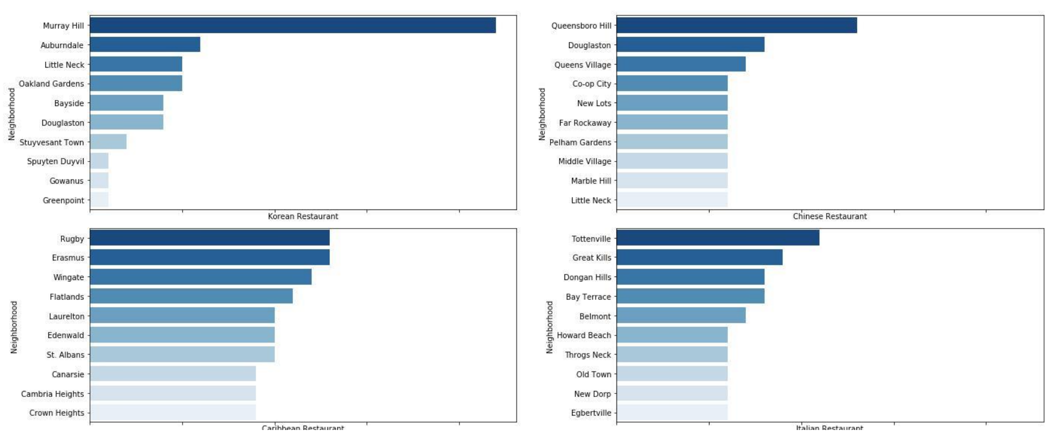
The top 10 categories are plotted individually on bar graph using python ‘seaborn’ library:

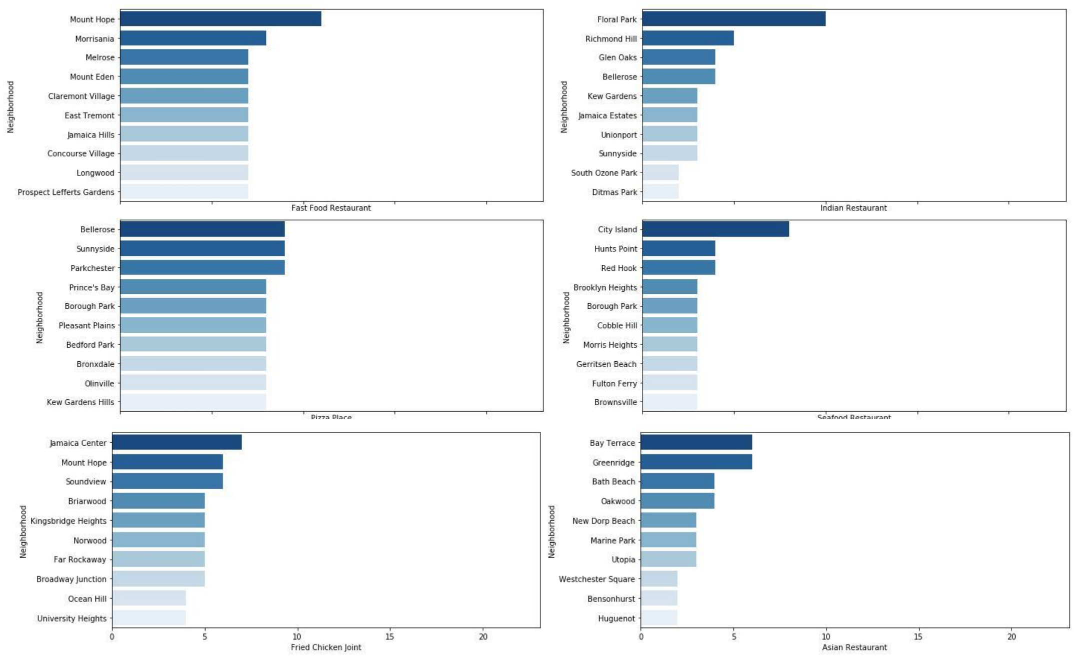
```
import seaborn as sns
import matplotlib.pyplot as plt

fig, axes =plt.subplots(5, 2, figsize=(20,20), sharex=True)
axes = axes.flatten()

for ax, category in zip(axes, venue_top10_list):
    data = venue_counts[[category]].sort_values([category], ascending=False)[0:10]
    pal = sns.color_palette("Blues", len(data))
    sns.barplot(x=category, y=data.index, data=data, ax=ax, palette=np.array(pal[::-1]))

plt.tight_layout()
plt.show();
```





Next, we group neighborhood rows to calculate the frequency of occurrence each category by taking the mean:

```
nyc_grouped = nyc_onehot.groupby('Neighborhood').mean().reset_index()
nyc_grouped.head()
```

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	Beer Store	Bike Shop	Brazilian Restaurant	Burrito Place	Cajun / Creole Restaurant	Cantonese Restaurant	Caribbean Restaurant	Caucasian Restaurant	Chinese Restaurant	Chocolate Shop	Colombian Restaurant	Cuban Restaurant	Dim Sum Restaurant	Dosa Place	Dum Resta
0	Allerton	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.045455	0.0	0.227273	0.0	0.0	0.0	0.0	0.0	
1	Annadale	0.0	0.0	0.166667	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.055556	0.0	0.0	0.0	0.0	0.0	
2	Arden Heights	0.0	0.0	0.157895	0.0	0.0	0.000000	0.0	0.0	0.052632	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.105263	0.0	0.0	0.0	0.0	0.0	
3	Arlington	0.0	0.0	0.111111	0.0	0.0	0.055556	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.055556	0.0	0.055556	0.0	0.0	0.0	0.0	0.0	
4	Arrochar	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.076923	0.0	0.0	0.0	0.0	0.0	

```
# Let's confirm the new size
nyc_grouped.shape
```

```
(302, 99)
```

Machine Learning

Unsupervised machine learning algorithm ‘k-means’ creates clusters of data points aggregated together because of certain similarities. The algorithm will be used to count neighborhoods for each cluster label for variable cluster size.

It is crucial to find out the optimal number of clusters to be able to successfully implement this algorithm. Thera are two most popular methods for the that, they are ‘The Elbow Method’ and ‘The Silhouette Method.’

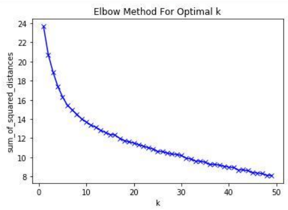
The Elbow Method

This method calculates the sum of squared distances of samples to their closest cluster center for different values of ‘k.’ The optimal number of clusters is the value after which there is no significant decrease in the sum of square distances:

```
sum_of_squared_distances = []
K = range(1,50)
for k in K:
    print(k, end=' ')
    kmeans = KMeans(n_clusters=k).fit(nyc_grouped_clustering)
    sum_of_squared_distances.append(kmeans.inertia_)
```

```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
```

```
plt.plot(K, sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('sum_of_squared_distances')
plt.title('Elbow Method For Optimal k');
```



In our case, the Elbow Method didn't give us required result. As, there is a gradual decrease in the sum of squared distances, optimal number of clusters can not be determined. To counter this, another method can be used.

The Silhouette Method

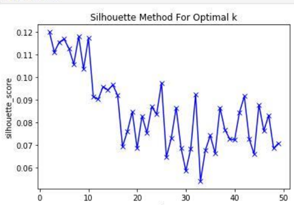
This method measures how similar a point is to its own cluster compared to other clusters. To do that, it requires minimum 2 clusters to define dissimilarity number of clusters will vary from 2 to 49:

```
from sklearn.metrics import silhouette_score

sil = []
K_sil = range(2,50)
# minimum 2 clusters required, to define dissimilarity
for k in K_sil:
    print(k, end=' ')
    kmeans = KMeans(n_clusters = k).fit(nyc_grouped_clustering)
    labels = kmeans.labels_
    sil.append(silhouette_score(nyc_grouped_clustering, labels, metric = 'euclidean'))
```

```
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
```

```
plt.plot(K_sil, sil, 'bx-')
plt.xlabel('k')
plt.ylabel('silhouette_score')
plt.title('Silhouette Method For Optimal k')
plt.show()
```



There is a peak at k = 2, k = 4 and k = 8. Two and four clusters will give a very broad classification of the venues.

k-Means

This code block runs the k-Means algorithm with number of clusters = 8 and prints the counts of neighborhoods assigned to different clusters:

Let's set number of clusters = 8

```
# set number of clusters
kclusters = 8

# run k-means clustering
kmeans = KMeans(init="k-means++", n_clusters=kclusters, n_init=50).fit(nyc_grouped_clustering)

print(Counter(kmeans.labels_))
```

```
Counter({6: 68, 2: 58, 1: 58, 4: 50, 5: 41, 3: 23, 0: 3, 7: 1})
```

Further the cluster labels curated are added to the dataframe to get the needed results of segmenting the neighborhood based upon the most common venues in its neighborhood:

```python
# add clustering labels
try:
    neighborhoods_venues_sorted.drop('Cluster Labels', axis=1)
except:
    neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

neighborhoods_venues_sorted.head(11)
```

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	4	Allerton	Pizza Place	Chinese Restaurant	Mexican Restaurant	Fried Chicken Joint	Fast Food Restaurant
1	2	Annadale	Pizza Place	American Restaurant	Italian Restaurant	Sushi Restaurant	Japanese Restaurant
2	2	Arden Heights	Pizza Place	American Restaurant	Italian Restaurant	Mexican Restaurant	Chinese Restaurant
3	2	Arlington	Pizza Place	American Restaurant	Fast Food Restaurant	Spanish Restaurant	Caribbean Restaurant
4	5	Arrochar	Italian Restaurant	Pizza Place	Japanese Restaurant	Polish Restaurant	Latin American Restaurant
5	6	Arverne	Pizza Place	Chinese Restaurant	American Restaurant	Asian Restaurant	Thai Restaurant
6	2	Astoria	Pizza Place	Greek Restaurant	Italian Restaurant	Fast Food Restaurant	Thai Restaurant
7	6	Astoria Heights	Greek Restaurant	Pizza Place	Chinese Restaurant	Italian Restaurant	Indian Restaurant
8	6	Auburndale	Korean Restaurant	Pizza Place	Greek Restaurant	Chinese Restaurant	Italian Restaurant
9	6	Bath Beach	Vietnamese Restaurant	Cantonese Restaurant	Chinese Restaurant	Asian Restaurant	Fast Food Restaurant
10	1	Battery Park City	Fast Food Restaurant	Seafood Restaurant	Mediterranean Restaurant	Italian Restaurant	American Restaurant

Now that 'neighborhoods_venues_sorted' is merged with 'nyc_data' to add the Borough, Latitude and Longitude for each neighborhood.

```python
# merge neighborhoods_venues_sorted with nyc_data to add latitude/longitude for each neighborhood
nyc_merged = neighborhoods_venues_sorted.join(neighborhoods.set_index('Neighborhood'), on='Neighborhood')
nyc_merged.head()
```

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
0	4	Allerton	Pizza Place	Chinese Restaurant	Mexican Restaurant	Fried Chicken Joint	Fast Food Restaurant	Bronx	40.865788	-73.859319
1	2	Annadale	Pizza Place	American Restaurant	Italian Restaurant	Sushi Restaurant	Japanese Restaurant	Staten Island	40.538114	-74.178549
2	2	Arden Heights	Pizza Place	American Restaurant	Italian Restaurant	Mexican Restaurant	Chinese Restaurant	Staten Island	40.549286	-74.185887
3	2	Arlington	Pizza Place	American Restaurant	Fast Food Restaurant	Spanish Restaurant	Caribbean Restaurant	Staten Island	40.635325	-74.165104
4	5	Arrochar	Italian Restaurant	Pizza Place	Japanese Restaurant	Polish Restaurant	Latin American Restaurant	Staten Island	40.596313	-74.067124

Next, NYC neighborhoods are visualized by using the code block which uses the python library 'folium.'

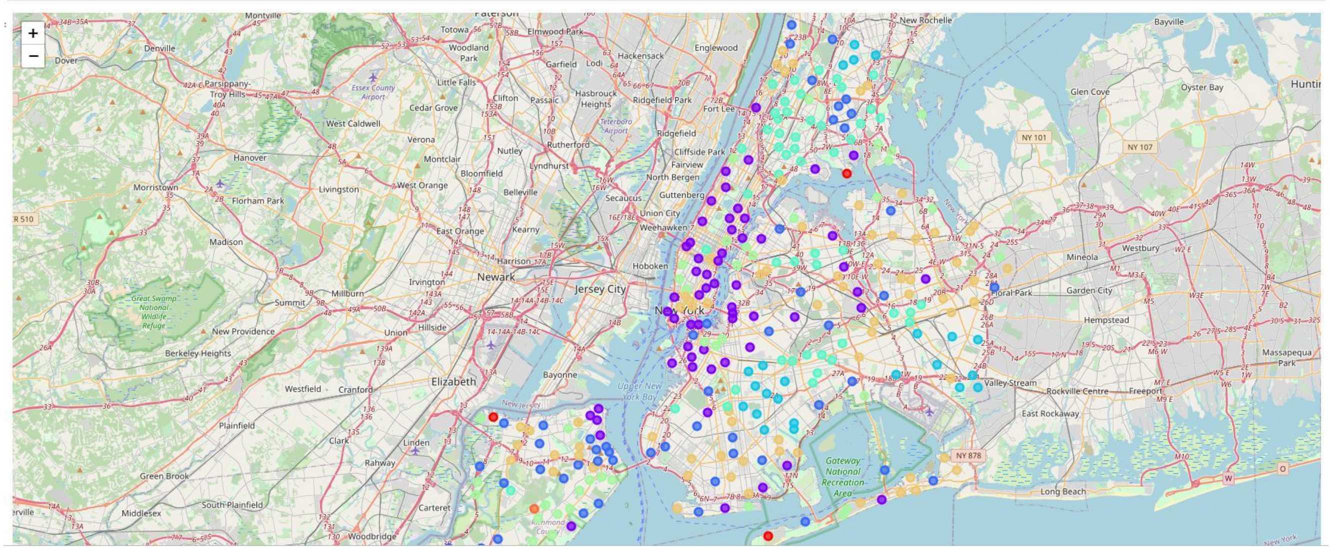
```python
# Finally, let's visualize the resulting clusters


# create map
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=10)

# set color scheme for the clusters
colors_array = cm.rainbow(np.linspace(0, 1, kclusters))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(nyc_merged['Latitude'], nyc_merged['Longitude'], nyc_merged['Neighborhood'], nyc_merged['Cluster Labels']):
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```



Results

Cluster – 0

```
cluster_0 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 0, nyc_merged.columns[1:12]]
cluster_0.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
5	Arverne	Pizza Place	Chinese Restaurant	American Restaurant	Asian Restaurant	Thai Restaurant	Queens	40.589144	-73.791992
7	Astoria Heights	Greek Restaurant	Pizza Place	Chinese Restaurant	Italian Restaurant	Indian Restaurant	Queens	40.770317	-73.894680
8	Auburndale	Korean Restaurant	Pizza Place	Greek Restaurant	Chinese Restaurant	Italian Restaurant	Queens	40.761730	-73.791762
9	Bath Beach	Vietnamese Restaurant	Cantonese Restaurant	Chinese Restaurant	Asian Restaurant	Fast Food Restaurant	Brooklyn	40.599519	-73.998752
13	Baychester	Chinese Restaurant	Fast Food Restaurant	Caribbean Restaurant	Pizza Place	Spanish Restaurant	Bronx	40.866858	-73.835798

```
for col in required_column:
    print(cluster_0[col].value_counts(ascending = False))
    print("_______________________________________________")
```

```
Chinese Restaurant       23
Pizza Place              12
Korean Restaurant         4
Sushi Restaurant          4
Caribbean Restaurant      1
Vietnamese Restaurant     1
Greek Restaurant          1
American Restaurant       1
Name: 1st Most Common Venue, dtype: int64
_______________________________________________
Chinese Restaurant       12
Pizza Place              11
Italian Restaurant        4
Fast Food Restaurant      3
Korean Restaurant         3
Cantonese Restaurant      2
BBQ Joint                 2
Russian Restaurant        2
Greek Restaurant          2
American Restaurant       2
Vietnamese Restaurant     1
Fried Chicken Joint       1
Japanese Restaurant       1
Taco Place                1
Name: 2nd Most Common Venue, dtype: int64
_______________________________________________
Queens           21
Brooklyn         11
Staten Island     7
Manhattan         5
Bronx             3
Name: Borough, dtype: int64
```

In cluster – 0, ‘Chinese Restaurant’ holds a big accountability with 23 occurrences in ‘1st Most Common Venue’ across different neighborhoods followed by ‘Pizza Place’ with 12 occurrences in ‘2nd Most Common Venue’. To add on, it is important to know that majority of these neighborhoods are in ‘Queens’ borough of New York City.

So, Cluster – 0 is a ‘Chinese Restaurant’ dominant cluster.

Cluster – 1

```
cluster_1 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 1, nyc_merged.columns[1:12]]
cluster_1.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
3	Arlington	Pizza Place	American Restaurant	Fast Food Restaurant	Spanish Restaurant	Caribbean Restaurant	Staten Island	40.635325	-74.165104
27	Boerum Hill	Pizza Place	New American Restaurant	Italian Restaurant	Seafood Restaurant	Cuban Restaurant	Brooklyn	40.685683	-73.983748
29	Breezy Point	American Restaurant	Pizza Place	Yemeni Restaurant	Greek Restaurant	Empanada Restaurant	Queens	40.557401	-73.925512
35	Brooklyn Heights	Pizza Place	Seafood Restaurant	Mediterranean Restaurant	French Restaurant	Italian Restaurant	Brooklyn	40.695864	-73.993782
44	Carroll Gardens	Pizza Place	Italian Restaurant	Seafood Restaurant	French Restaurant	Ramen Restaurant	Brooklyn	40.680540	-73.994654

```
for col in required_column:
    print(cluster_1[col].value_counts(ascending = False))
    print("_______________________________________________")
```

```
Pizza Place            14
American Restaurant     7
Seafood Restaurant      2
Spanish Restaurant      1
Name: 1st Most Common Venue, dtype: int64
_______________________________________________
American Restaurant        9
Pizza Place                6
Seafood Restaurant         3
Italian Restaurant         2
New American Restaurant    1
Sri Lankan Restaurant      1
Thai Restaurant            1
BBQ Joint                  1
Name: 2nd Most Common Venue, dtype: int64
_______________________________________________
Brooklyn         8
Staten Island    6
Queens           5
Bronx            3
Manhattan        2
Name: Borough, dtype: int64
_______________________________________________
```

‘Pizza Place’ holds a massive accountability for this cluster with 14 occurrences followed by ‘American Restaurant’ with 7 occurrences in ‘1st Most Common Venue’ across different neighborhoods. To add

on, it is inquisitive to know that majority of these neighborhoods are in ‘Brooklyn’ borough of New York City.
So, Cluster – 1 is a combination of ‘Pizza Place’ and ‘American Restaurant.’

Cluster – 2

```
cluster_2 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 2, nyc_merged.columns[1:12]]
cluster_2.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
36	Brookville	Caribbean Restaurant	Fried Chicken Joint	Pizza Place	Chinese Restaurant	Fast Food Restaurant	Queens	40.660003	-73.751753
37	Brownsville	Caribbean Restaurant	Pizza Place	Chinese Restaurant	Seafood Restaurant	American Restaurant	Brooklyn	40.663950	-73.910235
41	Cambria Heights	Caribbean Restaurant	Chinese Restaurant	African Restaurant	Pizza Place	Latin American Restaurant	Queens	40.692775	-73.735269
42	Canarsie	Caribbean Restaurant	Chinese Restaurant	Fast Food Restaurant	Pizza Place	Mexican Restaurant	Brooklyn	40.635564	-73.902093
68	Crown Heights	Caribbean Restaurant	Chinese Restaurant	Mexican Restaurant	Thai Restaurant	Fast Food Restaurant	Brooklyn	40.670829	-73.943291

```
for col in required_column:
    print(cluster_2[col].value_counts(ascending = False))
    print("_______________________________________________")
```

```
Caribbean Restaurant    22
Fast Food Restaurant     1
Name: 1st Most Common Venue, dtype: int64
_______________________________________________
Chinese Restaurant      8
Pizza Place             6
Fast Food Restaurant    4
Fried Chicken Joint     4
Caribbean Restaurant    1
Name: 2nd Most Common Venue, dtype: int64
_______________________________________________
Brooklyn    11
Queens       8
Bronx        4
Name: Borough, dtype: int64
_______________________________________________
```

‘Caribbean Restaurant’ dominates this cluster with 22 occurrences followed by ‘Fast Food Restaurant’ with 1 occurrence in the ‘1st Most Common Venue’ across different neighborhoods. Also, ‘Chinese Restaurant’ occurs 8 times followed by ‘Pizza Place’ occurrences of 6 times in ‘2nd Most Common Venue’. To add on, it is important to know that majority of these neighborhoods are in Brooklyn and Queens borough of New York City.
So, Cluster – 2 is a combination of Caribbean and Chinese restaurants

Cluster – 3

```
cluster_3 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 3, nyc_merged.columns[1:12]]
cluster_3.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
10	Battery Park City	Fast Food Restaurant	Seafood Restaurant	Mediterranean Restaurant	Italian Restaurant	American Restaurant	Manhattan	40.711932	-74.016869
11	Bay Ridge	Pizza Place	Fried Chicken Joint	Chinese Restaurant	Mexican Restaurant	Yemeni Restaurant	Brooklyn	40.625801	-74.030621
17	Bedford Stuyvesant	Caribbean Restaurant	Fried Chicken Joint	French Restaurant	Southern / Soul Food Restaurant	Pizza Place	Brooklyn	40.687232	-73.941785
25	Blissville	Pizza Place	Italian Restaurant	Chinese Restaurant	Mexican Restaurant	Fast Food Restaurant	Queens	40.737251	-73.932442
31	Brighton Beach	Eastern European Restaurant	Russian Restaurant	Pizza Place	Seafood Restaurant	Fast Food Restaurant	Brooklyn	40.576825	-73.965094

```
Pizza Place                    29
Mexican Restaurant             12
Fast Food Restaurant            4
Thai Restaurant                 2
Sushi Restaurant                1
Eastern European Restaurant     1
Italian Restaurant              1
Caribbean Restaurant            1
New American Restaurant         1
Vietnamese Restaurant           1
American Restaurant             1
Taco Place                      1
Indian Restaurant               1
Name: 1st Most Common Venue, dtype: int64
_______________________________________________
Mexican Restaurant           15
Fast Food Restaurant          7
Pizza Place                   7
Italian Restaurant            5
American Restaurant           4
Fried Chicken Joint           3
Chinese Restaurant            2
New American Restaurant       2
Seafood Restaurant            2
Latin American Restaurant     2
Asian Restaurant              1
Thai Restaurant               1
Middle Eastern Restaurant     1
Caribbean Restaurant          1
French Restaurant             1
Korean Restaurant             1
Russian Restaurant            1
Name: 2nd Most Common Venue, dtype: int64
_______________________________________________
Manhattan        23
Brooklyn         15
Queens           11
Staten Island     5
Bronx             2
Name: Borough, dtype: int64
_______________________________________________
```

‘Pizza Place’ dominates this cluster with 29 occurrences followed by ‘Mexican Restaurant’ with 12 occurrences in the ‘1st Most Common Venue’ across different neighborhoods. Also, ‘Mexican Restaurant’ occurs 15 times followed by ‘Fast Food Restaurant’ occurrences of 7 times in ‘2nd Most Common Venue’. To add on, it is inquisitive to know that neighborhoods in this cluster is spread mostly across ‘Manhattan’, and then ‘Brooklyn’ and ‘Queens’ with substantial number of neighborhoods in ‘Staten Island’.
So, Cluster – 3 is a combination of ‘Mexican Restaurant,’ ‘Pizza Place’ and ‘Fast Food Restaurant.’

Cluster – 4

```
cluster_4 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 4, nyc_merged.columns[1:12]]
cluster_4.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
1	Annadale	Pizza Place	American Restaurant	Italian Restaurant	Sushi Restaurant	Japanese Restaurant	Staten Island	40.538114	-74.178549
2	Arden Heights	Pizza Place	American Restaurant	Italian Restaurant	Mexican Restaurant	Chinese Restaurant	Staten Island	40.549286	-74.185887
6	Astoria	Pizza Place	Greek Restaurant	Italian Restaurant	Fast Food Restaurant	Thai Restaurant	Queens	40.768509	-73.915654
16	Bedford Park	Pizza Place	Chinese Restaurant	Fast Food Restaurant	Mexican Restaurant	Fried Chicken Joint	Bronx	40.870185	-73.885512
21	Bellerose	Pizza Place	Indian Restaurant	Chinese Restaurant	Halal Restaurant	American Restaurant	Queens	40.728573	-73.720128

```
for col in required_column:
    print(cluster_4[col].value_counts(ascending = False))
    print("___________________________________________")
```

```
Pizza Place           57
Asian Restaurant       1
Taco Place             1
Italian Restaurant     1
Name: 1st Most Common Venue, dtype: int64
___________________________________________
Chinese Restaurant            17
Italian Restaurant            11
American Restaurant            8
Pizza Place                    3
Fast Food Restaurant           3
Sushi Restaurant               3
Indian Restaurant              3
Mexican Restaurant             2
Greek Restaurant               2
Thai Restaurant                1
Eastern European Restaurant    1
BBQ Joint                      1
Caribbean Restaurant           1
Japanese Restaurant            1
Latin American Restaurant      1
Seafood Restaurant             1
Spanish Restaurant             1
Name: 2nd Most Common Venue, dtype: int64
___________________________________________
Staten Island    24
Queens           16
Brooklyn         11
Bronx             9
Name: Borough, dtype: int64
___________________________________________
```

In this cluster, ‘Pizza Place’ has taken over every other category with shooting 57 occurrences in ‘1st Most Common Venue’ across different neighborhoods followed by ‘Chinese Restaurant’ with 17 occurrences in ‘2nd Most Common Venue’. To add on, it is inquisitive to know that majority of these neighborhoods are spread across ‘Staten Island’, ‘Queens’ and ‘Brooklyn’ boroughs of New York City. So, Cluster – 4 can be termed as ‘Pizza Place’ dominant cluster.

Cluster – 5

```
cluster_5 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 5, nyc_merged.columns[1:12]]
cluster_5.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
0	Allerton	Pizza Place	Chinese Restaurant	Mexican Restaurant	Fried Chicken Joint	Fast Food Restaurant	Bronx	40.865788	-73.859319
15	Bayswater	Chinese Restaurant	Fast Food Restaurant	Pizza Place	Middle Eastern Restaurant	Caribbean Restaurant	Queens	40.611322	-73.765968
30	Briarwood	Fast Food Restaurant	Fried Chicken Joint	Pizza Place	Chinese Restaurant	Latin American Restaurant	Queens	40.710935	-73.811748
33	Broadway Junction	Fried Chicken Joint	Pizza Place	Fast Food Restaurant	Chinese Restaurant	Mexican Restaurant	Brooklyn	40.677861	-73.903317
47	Central Harlem	Pizza Place	Fast Food Restaurant	Southern / Soul Food Restaurant	American Restaurant	Fried Chicken Joint	Manhattan	40.815976	-73.943211

```
for col in required_column:
    print(cluster_5[col].value_counts(ascending = False))
    print("____________________________________________")
```

```
Fast Food Restaurant              27
Pizza Place                       12
Chinese Restaurant                 5
Fried Chicken Joint                4
Mexican Restaurant                 2
Southern / Soul Food Restaurant    1
Spanish Restaurant                 1
Name: 1st Most Common Venue, dtype: int64
____________________________________________
Fast Food Restaurant         12
Pizza Place                  10
Chinese Restaurant            9
Fried Chicken Joint           7
Mexican Restaurant            5
Latin American Restaurant     3
American Restaurant           2
Caribbean Restaurant          1
Middle Eastern Restaurant     1
Thai Restaurant               1
Steakhouse                    1
Name: 2nd Most Common Venue, dtype: int64
____________________________________________
Bronx            25
Queens           11
Brooklyn         10
Manhattan         5
Staten Island     1
Name: Borough, dtype: int64
____________________________________________
```

'Fast Food Restaurant' dominates this cluster with 27 occurrences followed by 'Pizza Place' with 12 occurrences in the '1st Most Common Venue' across different neighborhoods. C. To add on, it is inquisitive to know that neighborhoods in this cluster is spread mostly across 'Bronx', and then 'Queens' and 'Brooklyn.'

So, Cluster – 5 is a 'Fast Food Restaurant' dominant.

Cluster – 6

```
cluster_6 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 6, nyc_merged.columns[1:12]]
cluster_6.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
4	Arrochar	Italian Restaurant	Pizza Place	Japanese Restaurant	Polish Restaurant	Latin American Restaurant	Staten Island	40.596313	-74.067124
12	Bay Terrace	Italian Restaurant	Pizza Place	Asian Restaurant	American Restaurant	Chinese Restaurant	Queens	40.782843	-73.776802
12	Bay Terrace	Italian Restaurant	Pizza Place	Asian Restaurant	American Restaurant	Chinese Restaurant	Staten Island	40.553988	-74.139166
20	Belle Harbor	Seafood Restaurant	Italian Restaurant	Mexican Restaurant	BBQ Joint	Chinese Restaurant	Queens	40.576156	-73.854018
22	Belmont	Italian Restaurant	Fast Food Restaurant	Pizza Place	Mexican Restaurant	Chinese Restaurant	Bronx	40.857277	-73.888452

```
for col in required_column:
    print(cluster_6[col].value_counts(ascending = False))
    print("_________________________________________")
```

```
Italian Restaurant      32
Pizza Place              4
Seafood Restaurant       2
American Restaurant      2
Fast Food Restaurant     1
Indian Restaurant        1
Asian Restaurant         1
Name: 1st Most Common Venue, dtype: int64
_________________________________________
Pizza Place             16
Italian Restaurant      10
Mexican Restaurant       4
American Restaurant      3
Fast Food Restaurant     3
Chinese Restaurant       2
Asian Restaurant         2
Turkish Restaurant       1
Japanese Restaurant      1
Spanish Restaurant       1
Name: 2nd Most Common Venue, dtype: int64
_________________________________________
Staten Island    19
Queens            9
Bronx             6
Manhattan         5
Brooklyn          4
Name: Borough, dtype: int64
_________________________________________
```

'Italian Restaurant' holds a massive accountability for this cluster with 32 occurrences followed by 'Pizza Place' with 4 occurrences in '1st Most Common Venue' across different neighborhoods followed by 'Pizza Place' with 16 occurrences in '2nd Most Common Venue'. To add on, it is inquisitive to know that majority of these neighborhood are 'Staten Island' and 'Queens' boroughs of New York City.

It is known that, although pizza is an Italian cuisine, it is also a fast food. So, Cluster – 6 can be termed as 'Italian Restaurant' dominant cluster.

Cluster – 7

```
cluster_7 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 7, nyc_merged.columns[1:12]]
cluster_7.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
152	Lighthouse Hill	Italian Restaurant	Yemeni Restaurant	Greek Restaurant	Empanada Restaurant	English Restaurant	Staten Island	40.576506	-74.137927

```
for col in required_column:
    print(cluster_7[col].value_counts(ascending = False))
    print("_________________________________________")
```

```
Italian Restaurant    1
Name: 1st Most Common Venue, dtype: int64
_________________________________________
Yemeni Restaurant    1
Name: 2nd Most Common Venue, dtype: int64
_________________________________________
Staten Island    1
Name: Borough, dtype: int64
_________________________________________
```

It is clear, that only one neighborhood 'Lighthouse Hill' is curated under this cluster. This segmentation can be understood from the fact that 'Lighthouse Hill' is a tourist attraction for its heritage and is situated at the southernmost of the chain of hills that radiate from the northeast corner of Staten Island. This neighborhood has diverse cuisine in its top 5 most common venues list and hence a separate cluster. So, Cluster – 5 can be termed as exceptional as of now.

Discussion

To understand the clusters, three analysis were done, namely:

1. Count of ‘Borough’
2. Count of ‘1st Most Common Venue’
3. Count of ‘2nd Most Common Venue’

The above information speaks a lot about the ground reality of clustering based on the similarity metrics between the neighborhoods. Tabulating the results of the k-Mean unsupervised machine learning algorithm:

	Count of Occurrences within the Cluster		
Cluster	**1st Most Common Venue**	**2nd Most Common Venue**	**Borough**
0	Chinese Restaurant	Chinese Restaurant	Queens, Brooklyn
1	Pizza Place	American Restaurant	Brooklyn, Staten Island
2	Caribbean Restaurant	Chinese Restaurant	Brooklyn, Queens
3	Pizza Place	Mexican Restaurant	Manhattan, Brooklyn, Queens
4	Pizza Place	Chinese Restaurant	Staten Island, Queens, Brooklyn
5	Fast Food Restaurant	Fast Food restaurant	Bronx, Queens
6	Italian Restaurant	Pizza Place	Staten Island, Queens
7	Italian Restaurant	Yemeni Restaurant	Lighthouse Hill

Conclusion

On application of Clustering Algorithm, k-Means or others, to a multi-dimensional dataset, a very inquisitive results can be curated which helps to understand and visualize the data. The neighborhoods of New York City were very briefly segmented into eight clusters (0-7) and upon analysis it was possible to rename them basis upon the categories of venues in and around that neighborhood. Along with the Italian Restaurant, Caribbean and Chinese are dominant in New York City, and so is the diversity statistics. The scope of this project can be expanded further to understand the dynamics of each neighborhood and suggest my entrepreneur friend and a new food vendor a profitable location to open their restaurant.