Capstone Project - The Battle of Neighborhoods

Part 1: Introduction/Business Problem

New York City's demographics show that it is a large and ethnically diverse metropolis. It is the largest city in the United States with a long history of international immigration. New York City has been a major point of entry for immigrants and as many as 800 languages are spoken in New York. According to the U.S. Census Bureau 2018 American Community Survey 1-Year Estimates, as of 2018, 684,345 U.S. residents identify themselves as being of Peruvian origin. A high concentration of Peruvians resides in New York City. Peruvian migrants, in particular, have generally been successful in opening restaurants and creating jobs in New York. Peruvian food is one of the defining characteristics of Peruvian culture and have a special niche in serving distinct flavors and ingredients of each region of the country and the culinary traditions of each.

This capstone project examines the best locations for Peruvian restaurants in New York City. Intricate and intimate migrant networks have developed over the years between New York City and specific communities in Peru. As family members and friends from these communities move back and forth between the two countries, they facilitate the flow of information. Nevertheless, like any other business, opening a new restaurant requires serious considerations and is more complicated than it seems at first. Location influences the success or failure of a restaurant in a host of ways, from attracting enough initial customer interest to being convenient to visit.

The objective of this capstone project is to analyze and select the best locations in New York City to open a new Peruvian restaurant by analyzing and visualizing data. The project is particularly useful to new Peruvian migrants looking to open and invest on new restaurant in New York City. New York City was home to over 8.3 million people in 2019 and keeps growing as diverse city in terms of culture and food offering.

Part 2: Data

To accomplish this capstone project, it is required the following data from New York City:

1.New York City data related to neighborhoods and boroughs.

2.Neighborhoods latitude and longitude coordinates required for mapping plot and obtain venue data.

3.Venue data related to existing restaurants in New York City for further analysis of neighborhoods.

New York City data related to neighborhoods and boroughs can be obtained from the open data source (https://cocl.us/new_york_dataset) and geographical coordinates of neighborhoods, such as latitude and longitude, from Python Geocoder package. Foursquare API will be necessary to obtain venue data of New York City neighborhoods. Currently Foursquare's database contains more than 105 million places including venue data such as Peruvian restaurants in New York City.

The capstone project will show how to convert addresses into their equivalent latitude and longitude values. Also, how to use the Foursquare API to explore neighborhoods in New York City, use the explore function to get the most common venue categories in each neighborhood. Finally, the importance of using Folium library to visualize the neighborhoods in New York City.

Part 3: Methodology

For this capstone project the methodology will be the following:

1.Data will be obtained from the open data source (https://cocl.us/new_york_dataset) and cleaned/processed into a data frame.

2.Foursquare will be used to for venue data and filtered by Peruvian restaurants (ratings, tips, and likes by users will be counted and added to the data frame).

3.Data will be sorted based on rankings.

4.Data be will be visualized using Python Folium library.

Preparation - Downloading and importing all required libraries

```
import numpy as np # library to handle data in a vectorized manner
import pandas as pd # library for data analsysis
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
import json # library to handle JSON files
#!conda install -c conda-forge geopy --yes # uncomment this line if you haven't completed the Foursquare API lab
from geopy.geocoders import Nominatim # convert an address into latitude and longitude values
import requests # library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe
# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors
# import k-means from clustering stage
from sklearn.cluster import KMeans
#!conda install -c conda-forge folium=0.5.0 --yes # uncomment this line if you haven't completed the Foursquare API lab
!pip install folium
import folium # map rendering library
import urllib
import seaborn as sns
from matplotlib import pyplot as plt
print('Libraries imported.')
```

Credentials

```
CLIENT_ID = '[redacted]' # your Foursquare ID
CLIENT_SECRET = '[redacted]' # your Foursquare Secret
VERSION = '20180605' # Foursquare API version

print('Credentials:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

Functions to repeat the same process to all the neighborhoods in New York City

```
def geo_location(address):
    # get geo location of address
    geolocator = Nominatim(user_agent="foursquare_agent")
    location = geolocator.geocode(address)
    latitude = location.latitude
    longitude = location.longitude
    return latitude,longitude


def get_venues(lat,lng):
    #set variables
    radius=400
    LIMIT=100
    #url to fetch data from foursquare api
    url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)
    # get all the data
    results = requests.get(url).json()
    venue_data=results['response']['groups'][0]['items']
    venue_details=[]
    for row in venue_data:
        try:
            venue_id=row['venue']['id']
            venue_name=row['venue']['name']
            venue_category=row['venue']['categories'][0]['name']
            venue_details.append([venue_id,venue_name,venue_category])
        except KeyError:
            pass
    column_names=['ID','Name','Category']
    df = pd.DataFrame(venue_details,columns=column_names)
    return df
```

```
def geo_location(address):
    # get geo location of address
    geolocator = Nominatim(user_agent="foursquare_agent")
    location = geolocator.geocode(address)
    latitude = location.latitude
    longitude = location.longitude
    return latitude,longitude


def get_venues(lat,lng):
    #set variables
    radius=400
    LIMIT=100
    #url to fetch data from foursquare api
    url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)
    # get all the data
    results = requests.get(url).json()
    venue_data=results['response']['groups'][0]['items']
    venue_details=[]
    for row in venue_data:
        try:
            venue_id=row['venue']['id']
            venue_name=row['venue']['name']
            venue_category=row['venue']['categories'][0]['name']
            venue_details.append([venue_id,venue_name,venue_category])
        except KeyError:
            pass
    column_names=['ID','Name','Category']
    df = pd.DataFrame(venue_details,columns=column_names)
    return df
```

New York City neighborhoods and boroughs data

```
ny_data = get_new_york_data()
ny_data.head()
```

[9]:

	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

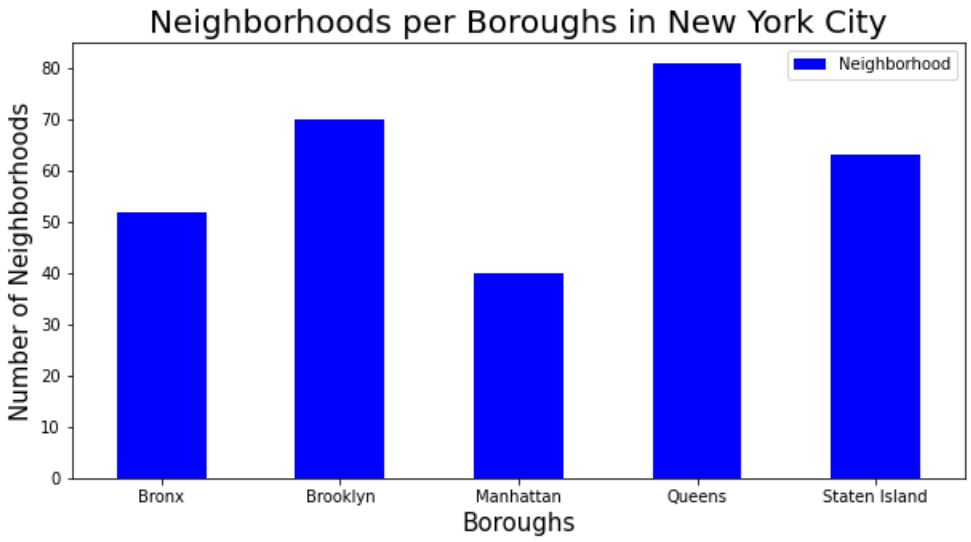
```
ny_data.shape
```

10]: (306, 4)

There are total of 306 different neighborhoods in New York City.

Data Analysis of New York City neighborhoods and boroughs

```
clr = "blue"
ny_data.groupby('Borough')['Neighborhood'].count().plot.bar(figsize=(10,5), color=clr)
plt.title('Neighborhoods per Boroughs in New York City', fontsize = 20)
plt.xlabel('Boroughs', fontsize = 15)
plt.ylabel('Number of Neighborhoods',fontsize = 15)
plt.xticks(rotation = 'horizontal')
plt.legend()
plt.show()
```



According to graph Queens is the borough in New York City with more number of neighborhoods

Peruvian restaurants that are in each neighborhood and borough in New York City

```
# preparing neighborhood list that contains peruvian restaurants
column_names=['Borough', 'Neighborhood', 'ID','Name']
peruvian_rest_ny=pd.DataFrame(columns=column_names)
count=1
for row in ny_data.values.tolist():
    Borough, Neighborhood, Latitude, Longitude=row
    venues = get_venues(Latitude,Longitude)
    peruvian_restaurants=venues[venues['Category']=='Peruvian Restaurant']
    print('(',count,'/',len(ny_data),')','Peruvian Restaurants in '+Neighborhood+', '+Borough+':'+str(len(peruvian_restaurants)))
    print(row)
    for resturant_detail in peruvian_restaurants.values.tolist():
        id, name , category=resturant_detail
        peruvian_rest_ny = peruvian_rest_ny.append({'Borough': Borough,
                                                'Neighborhood': Neighborhood,
                                                'ID': id,
                                                'Name' : name
                                               }, ignore_index=True)
    count+=1
```

```
# Saving the information so far to a .csv file due to limited calls on FourSquare
peruvian_rest_ny.to_csv('peruvian_rest_ny_tocsv1.csv')
```

```
peruvian_ny = pd.read_csv('peruvian_rest_ny_tocsv1.csv')
peruvian_rest_ny
```

2]:

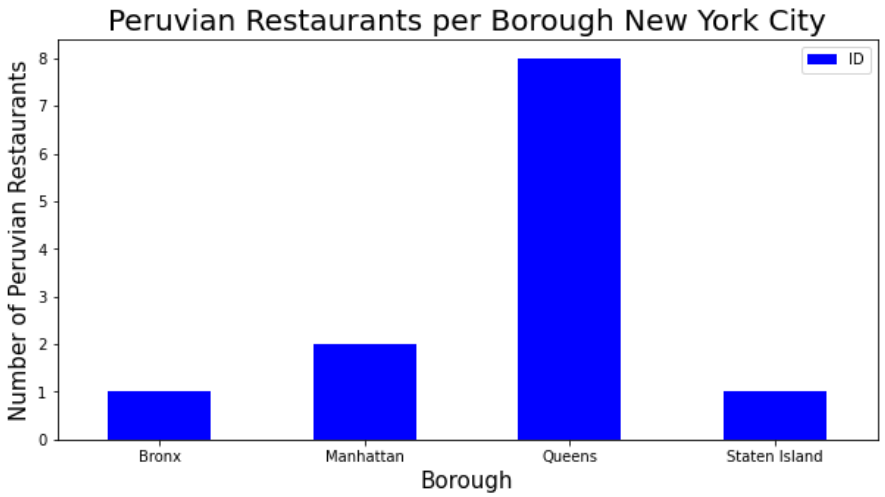
	Borough	Neighborhood	ID	Name
0	Bronx	Mott Haven	4bb28791715eef3b9f5e85bb	Pio Pio
1	Manhattan	Upper West Side	4a7a543df964a5202ee91fe3	Flor de Mayo
2	Manhattan	Clinton	4b1b1f52f964a52074f823e3	Pio Pio
3	Queens	Jackson Heights	4b9ece55f964a520590337e3	Urubamba
4	Queens	Jackson Heights	514cae4ae4b08e5e6fb50538	Don Alex Restaurant
5	Queens	Sunnyside	51ec2df8498ee2a4dc8ee843	Don Pollo II
6	Queens	Rego Park	4ede24237ee5f354d5122374	Don Alex
7	Queens	Rego Park	4b37cf54f964a520924625e3	Cuzco Peru
8	Queens	Little Neck	52d1e8f4498e56474f235066	Lima 33 Restaurant
9	Staten Island	Concord	536eb74e498e9a6b05cc9939	Inca's Grill Peruvian Cuisine
10	Queens	Hunters Point	52b46aec11d2522f8646e332	Jora
11	Queens	Sunnyside Gardens	4b942c0cf964a5209f6c34e3	Riko

```
peruvian_rest_ny.shape
```

1]: (12, 4)

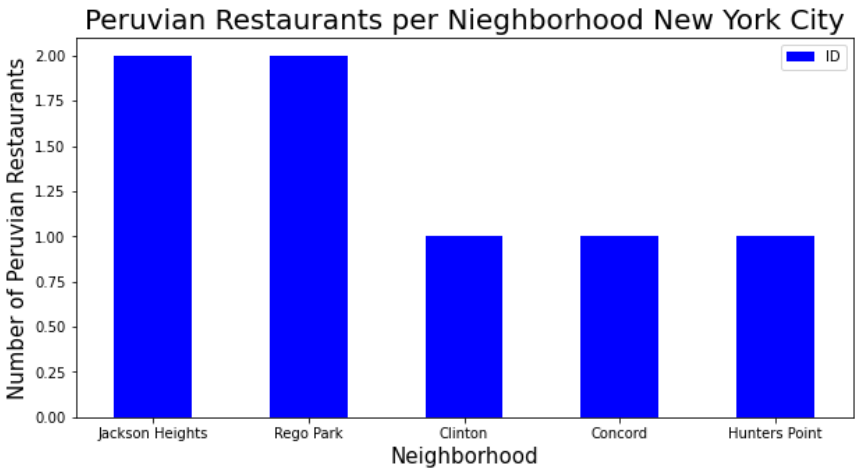
We obtained 12 Peruvian Restaurants across the New York City.

```
peruvian_rest_ny.groupby('Borough')['ID'].count().plot.bar(figsize=(10,5), color=clr)
plt.title('Peruvian Restaurants per Borough New York City', fontsize = 20)
plt.xlabel('Borough', fontsize = 15)
plt.ylabel('Number of Peruvian Restaurants', fontsize=15)
plt.xticks(rotation = 'horizontal')
plt.legend()
plt.show()
```



Queens borough has the greatest number of Peruvian restaurants by far in New York City.

```
NOofNeigh = 5
peruvian_rest_ny.groupby('Neighborhood')['ID'].count().nlargest(NOofNeigh).plot.bar(figsize=(10,5), color=clr)
plt.title('Peruvian Restaurants per Nieghborhood New York City', fontsize = 20)
plt.xlabel('Neighborhood', fontsize = 15)
plt.ylabel('Number of Peruvian Restaurants', fontsize=15)
plt.xticks(rotation = 'horizontal')
plt.legend()
plt.show()
```



Jason Heights and Rego Park neighborhood have the double of Peruvian restaurants than other neighborhoods in all New York City and both are located in Queens Borough.

Ranking of each restaurant for further analysis

```
column_names=['Borough', 'Neighborhood', 'ID','Name','Likes','Rating','Tips']
peruvian_rest_stats_ny=pd.DataFrame(columns=column_names)
count=1
for row in peruvian_rest_ny.values.tolist():
    Borough,Neighborhood,ID,Name=row
    try:
        venue_details=get_venue_details(ID)
        print(venue_details)
        id,name,likes,rating,tips=venue_details.values.tolist()[0]
    except IndexError:
        print('No data available for id=',ID)
        # we will assign 0 value for these resturants as they may have been
        #recently opened or details does not exist in FourSquare Database
        id,name,likes,rating,tips=[0]*5
    print('(',count,'/',len(peruvian_rest_ny),')','processed')
    peruvian_rest_stats_ny = peruvian_rest_stats_ny.append({'Borough': Borough,
                                            'Neighborhood': Neighborhood,
                                            'ID': id,
                                            'Name' : name,
                                            'Likes' : likes,
                                            'Rating' : rating,
                                            'Tips' : tips
                                           }, ignore_index=True)
    count+=1
peruvian_rest_stats_ny
```

	Borough	Neighborhood	ID	Name	Likes	Rating	Tips
0	Bronx	Mott Haven	4bb28791715eef3b9f5e85bb	Pio Pio	67	8.7	23
1	Manhattan	Upper West Side	4a7a543df964a5202ee91fe3	Flor de Mayo	201	8.3	92
2	Manhattan	Clinton	4b1b1f52f964a52074f823e3	Pio Pio	1125	8.9	330
3	Queens	Jackson Heights	4b9ece55f964a520590337e3	Urubamba	98	8.7	51
4	Queens	Jackson Heights	514cae4ae4b08e5e6fb50538	Don Alex Restaurant	6	7.5	3
5	Queens	Sunnyside	51ec2df8498ee2a4dc8ee843	Don Pollo II	7	7.8	2
6	Queens	Rego Park	4ede24237ee5f354d5122374	Don Alex	25	7.5	13
7	Queens	Rego Park	4b37cf54f964a520924625e3	Cuzco Peru	14	6.8	7
8	Queens	Little Neck	52d1e8f4498e56474f235066	Lima 33 Restaurant	27	8.3	4
9	Staten Island	Concord	536eb74e498e9a6b05cc9939	Inca's Grill Peruvian Cuisine	18	7.5	8
10	Queens	Hunters Point	52b46aec11d2522f8646e332	Jora	99	8.7	23

```
# Continuing to save data to a .csv file
peruvian_rest_stats_ny.to_csv('peruvian_rest_stats_ny_csv.csv')
```

```
peruvian_rest_stats_ny.shape
```

```
6]: (12, 7)
```

We obtained information such as number of likes, ratings, and tips for all 12 Peruvian Restaurants located in New York City.

Pio Pio restaurant in Clinton Neighborhood in Manhattan has the greatest number of likes, best rating, and tips among all Peruvian restaurants in the whole city.

```
peruvian_rest_stats_ny.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12 entries, 0 to 11
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Borough       12 non-null     object
 1   Neighborhood  12 non-null     object
 2   ID            12 non-null     object
 3   Name          12 non-null     object
 4   Likes         12 non-null     object
 5   Rating        12 non-null     float64
 6   Tips          12 non-null     object
dtypes: float64(1), object(6)
memory usage: 800.0+ bytes
```

```
#Converting string values to float
peruvian_rest_stats_ny['Likes'] = peruvian_rest_stats_ny['Likes'].astype('float64')
peruvian_rest_stats_ny['Tips'] = peruvian_rest_stats_ny['Tips'].astype('float64')
peruvian_rest_stats_ny.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12 entries, 0 to 11
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Borough       12 non-null     object
 1   Neighborhood  12 non-null     object
 2   ID            12 non-null     object
 3   Name          12 non-null     object
 4   Likes         12 non-null     float64
 5   Rating        12 non-null     float64
 6   Tips          12 non-null     float64
dtypes: float64(3), object(4)
memory usage: 800.0+ bytes
```

```
peruvian_rest_stats_ny.describe()
```

59]:

	Likes	Rating	Tips
count	12.000000	12.000000	12.000000
mean	144.750000	8.041667	49.333333
std	313.767176	0.654298	92.158296
min	6.000000	6.800000	2.000000
25%	17.000000	7.500000	6.250000
50%	38.500000	8.050000	18.000000
75%	98.250000	8.700000	39.750000
max	1125.000000	8.900000	330.000000

```
# Neighborhood with the maximum average rating of restaurants

ny_neighborhood_stats=peruvian_rest_stats_ny.groupby('Neighborhood',as_index=False).mean()[['Neighborhood','Rating']]
ny_neighborhood_stats.columns=['Neighborhood','Average Rating']
ny_neighborhood_stats.sort_values(['Average Rating'],ascending=False)
```

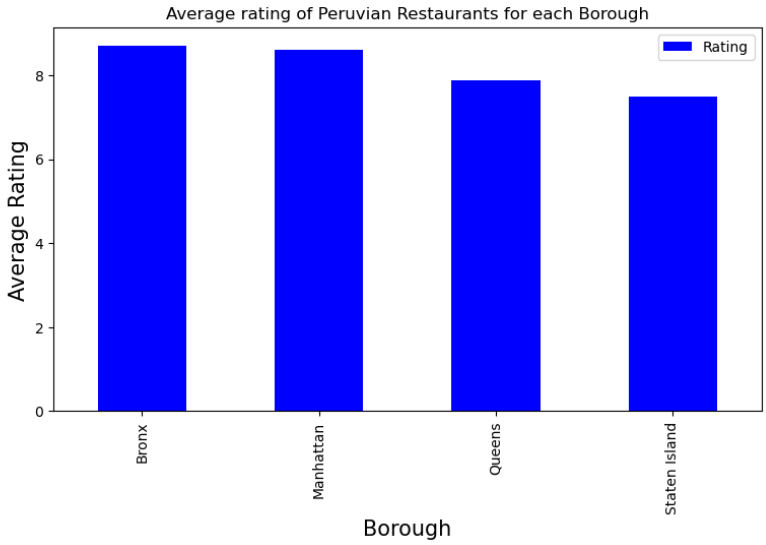
37]:

	Neighborhood	Average Rating
0	Clinton	8.90
2	Hunters Point	8.70
5	Mott Haven	8.70
4	Little Neck	8.30
9	Upper West Side	8.30
3	Jackson Heights	8.10
7	Sunnyside	7.80
8	Sunnyside Gardens	7.80
1	Concord	7.50
6	Rego Park	7.15

Clinton neighborhood has the highest average ratings of Peruvian restaurants in all New York City

```
#Visualization

plt.figure(figsize=(9,5), dpi = 100)
plt.title('Average rating of Peruvian Restaurants for each Borough')
plt.xlabel('Borough', fontsize = 15)
plt.ylabel('Average Rating', fontsize=15)
peruvian_rest_stats_ny.groupby('Borough').mean()['Rating'].plot(kind='bar', color=clr)
plt.legend()
plt.show()
```



```
# Now let's consider the top neighborhoods with average rating higher that 8.0 for map visualization
ny_neighborhood_stats=ny_neighborhood_stats[ny_neighborhood_stats['Average Rating']>=8.0]
ny_neighborhood_stats
```

39]:

	Neighborhood	Average Rating
0	Clinton	8.9
2	Hunters Point	8.7
3	Jackson Heights	8.1
4	Little Neck	8.3
5	Mott Haven	8.7
9	Upper West Side	8.3

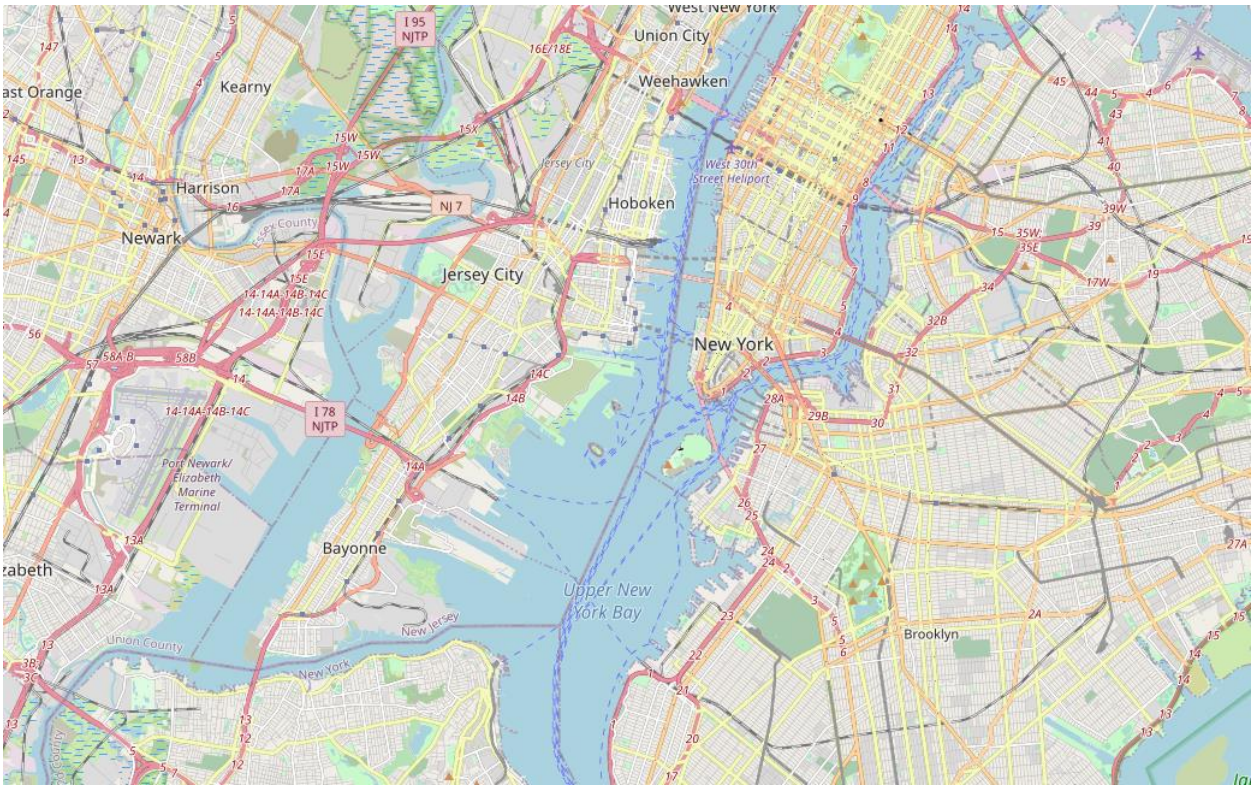
```
# Joining dataset to original New York data for longitude and latitude values
ny_neighborhood_stats=pd.merge(ny_neighborhood_stats,ny_data, on='Neighborhood')
ny_neighborhood_stats=ny_neighborhood_stats[['Borough','Neighborhood','Latitude','Longitude','Average Rating']]
ny_neighborhood_stats
```

30]:

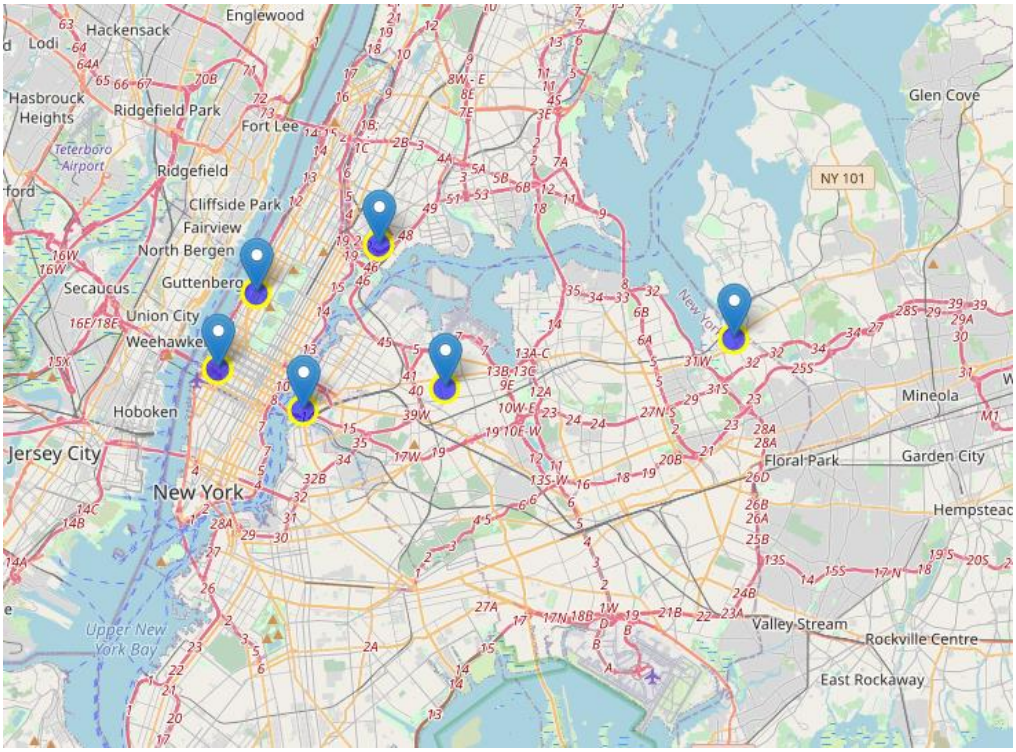
	Borough	Neighborhood	Latitude	Longitude	Average Rating
0	Manhattan	Clinton	40.759101	-73.996119	8.9
1	Queens	Hunters Point	40.743414	-73.953868	8.7
2	Queens	Jackson Heights	40.751981	-73.882821	8.1
3	Queens	Little Neck	40.770826	-73.738898	8.3
4	Bronx	Mott Haven	40.806239	-73.916100	8.7
5	Manhattan	Upper West Side	40.787658	-73.977059	8.3

```
# Map creation
ny_map = folium.Map(location=geo_location('New York City'), zoom_start=12)
# instantiate a feature group for the ratings in the dataframe
rating = folium.map.FeatureGroup()

# loop through the ratings and add each to the neighborhood feature group
for lat, lng, in ny_neighborhood_stats[['Latitude','Longitude']].values:
    rating.add_child(
        folium.CircleMarker(
            [lat, lng],
            radius=10, # define how big you want the circle markers to be
            color='yellow',
            fill=True,
            fill_color='blue',
            fill_opacity=0.6
        )
    )
ny_map
```




```
#Adding a new field to dataframe for labeling purpose
ny_neighborhood_stats['Label']=ny_neighborhood_stats['Neighborhood']+', '+ny_neighborhood_stats['Borough']+'('+ny_neighborhood_stats['Average Rating'].map(str)+')'
# add pop-up text to each marker on the map
for lat, lng, label in ny_neighborhood_stats[['Latitude','Longitude','Label']].values:
    folium.Marker([lat, lng], popup=label).add_to(ny_map)
# add ratings to map
ny_map.add_child(rating)
```



Part 4: Results and Discussion

According to the analysis there are only 12 Peruvian restaurants in all New York City where Bronx borough is the top-rated followed by Manhattan borough, leaving the other boroughs far behind in the average rating. Queens borough has a large number of Peruvian restaurants in New York City. Clinton neighborhood located in Manhattan borough has the highest average rating in all New York City followed by Hunter Point and Mott Haven, which both share same average rating, and are located in Queens and Bronx boroughs respectively. Pio Pio restaurant located in Clinton Neighborhood, in Manhattan borough, has the greatest number of likes, best rating, and tips among all Peruvian restaurants in New York City even though it has another restaurant in Bronx borough with high rating.

Based on the above analysis Manhattan borough, specifically Clinton neighborhood, would be the best location in New York City to open a new Peruvian restaurant by analyzing and visualizing data. However, since Pio Pio restaurant is located in this neighborhood and being the most popular restaurant in terms of likes, best rating, and tips; I would suggest also looking to open and invest on new restaurant in other boroughs such as Brooklyn which there are no records of Peruvian restaurants.

Finally, it is important to take into consideration that more extensive analysis and more data is required for better decision-making since for purposes of this capstone project only Foursquare API was used to explore Peruvian restaurants in New York City, excluding other external databases where other popular venues are registered such as Yelp or Google Maps. Also, small and more traditional Peruvian Restaurants might prefer not to announce their venues in local search-and-discovery mobile apps to keep themselves more local among their community.

Part 5: Conclusion

This capstone project gave me a taste of what data scientists go through in real life situations while working with data and data providers, such as Foursquare. Also, how to make API calls to the Foursquare API to retrieve data about venues, in this case Peruvian restaurants, in different neighborhoods around New York City. Furthermore, the capstone project allowed me to learn how to be creative in situations where data is not available by scraping web data and parsing HTML code. Last but not least, how to utilize Python and its pandas and folium libraries to manipulate data for exploring, analyzing and visualizing data.