CAPSTONE PROJECT

RESTAURANTS IN NYC

Shan Tang

INTRODUCTION: BUSINESS PROBLEM

In this project we will try to find an optimal location for a restaurant. Specifically, this report will be targeted to stakeholders interested in opening an **Chinese restaurant** in **Manhattan**, New York.

Since there are lots of restaurants in NYC we will try to detect **locations that are not already crowded with restaurants**. We are also particularly interested in **areas with no Chinese restaurants in vicinity**. We would also prefer locations **as close to city center as possible**, assuming that first two conditions are met.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

ANALYTICAL APPROACH

Our problem is clearly a clustering problem. We will therefore rely on a clustering model to solve it. Clustering models are numerous, with the two most popular being K-means clustering and hierarchical clustering. Fortunately, most clustering algorithms are already implemented in open source libraries for the language we will use (Python), therefore we won't have to do much coding. The most critical and the most tedious part of this project, as with most data science projects, will be to collect and clean the data.

DATA

Based on definition of our problem, factors that will influence our decision are:

- number of existing restaurants in the neighborhood (any type of restaurant)
- number of and distance to Italian restaurants in the neighborhood, if any
- distance of neighborhood from city center

We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

- centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using **Google Maps API reverse geocoding**
- number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**
- coordinate of NYC center will be obtained using **Google Maps API geocoding** of well known NYC location

NEIGHBORHOOD CANDIDATES

- New York City data that contains Borough, Neighborhoods along with their latitudes and longitudes
- Data Source: https://cocl.us/new_york_dataset
- Description: This data set contains the required information. And we will use this data set to explore various neighborhoods of New York city.
- Chinese restaurants in Manhattan neighborhood of New York city.
- Data Source: Foursquare API

- Description: By using this API we will get all the venues in Manhattan neighborhood. We can filter these venues to get only Chinese restaurants.

METHOLOGY

1. *Import library and collect data for neighborhoods in NYC*

```
import pandas as pd
import json
import requests # library to handle requests

def get_new_york_data():
    url='https://cocl.us/new_york_dataset'
    resp=requests.get(url).json()
    # all data is present in features label
    features=resp['features']
    # define the dataframe columns
    column_names = ['Borough', 'Neighborhood', 'Latitude', 'Longitude']
    # instantiate the dataframe
    new_york_data = pd.DataFrame(columns=column_names)
    for data in features:
        borough = data['properties']['borough']
        neighborhood_name = data['properties']['name']
        neighborhood_latlon = data['geometry']['coordinates']
        neighborhood_lat = neighborhood_latlon[1]
        neighborhood_lon = neighborhood_latlon[0]
        new_york_data = new_york_data.append({'Borough': borough,
                                              'Neighborhood': neighborhood_name,
                                              'Latitude': neighborhood_lat,
                                              'Longitude': neighborhood_lon}, ignore_index=True)
    return new_york_data


ny_data = get_new_york_data()
ny_data
```

	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
...	...	...	...	...
301	Manhattan	Hudson Yards	40.756658	-74.000111
302	Queens	Hammels	40.587338	-73.805530
303	Queens	Bayswater	40.611322	-73.765968
304	Queens	Queensbridge	40.756091	-73.945631
305	Staten Island	Fox Hills	40.617311	-74.081740

306 rows × 4 columns

2. Obtain Manhattan data from NYC dataset

Obtain Manhtattan data then use folium to map it

```
manhattan_data = ny_data[ny_data['Borough'] == 'Manhattan'].reset_index(drop=True)
manhattan_data
```

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688
5	Manhattan	Manhattanville	40.816934	-73.957385
6	Manhattan	Central Harlem	40.815976	-73.943211
7	Manhattan	East Harlem	40.792249	-73.944182
8	Manhattan	Upper East Side	40.775639	-73.960508
9	Manhattan	Yorkville	40.775930	-73.947118
10	Manhattan	Lenox Hill	40.768113	-73.958860
11	Manhattan	Roosevelt Island	40.762160	-73.949168
12	Manhattan	Upper West Side	40.787658	-73.977059
13	Manhattan	Lincoln Square	40.773529	-73.985338
14	Manhattan	Clinton	40.759101	-73.996119
15	Manhattan	Midtown	40.754691	-73.981669
16	Manhattan	Murray Hill	40.748303	-73.978332
17	Manhattan	Chelsea	40.744035	-74.003116
18	Manhattan	Greenwich Village	40.726933	-73.999914
19	Manhattan	East Village	40.727847	-73.982226
20	Manhattan	Lower East Side	40.717807	-73.980890
21	Manhattan	Tribeca	40.721522	-74.010683
22	Manhattan	Little Italy	40.719324	-73.997305
23	Manhattan	Soho	40.722184	-74.000657
24	Manhattan	West Village	40.734434	-74.006180
25	Manhattan	Manhattan Valley	40.797307	-73.964286
26	Manhattan	Morningside Heights	40.808000	-73.963896
27	Manhattan	Gramercy	40.737210	-73.981376
28	Manhattan	Battery Park City	40.711932	-74.016869
29	Manhattan	Financial District	40.707107	-74.010665
30	Manhattan	Carnegie Hill	40.782683	-73.953256
31	Manhattan	Noho	40.723259	-73.988434

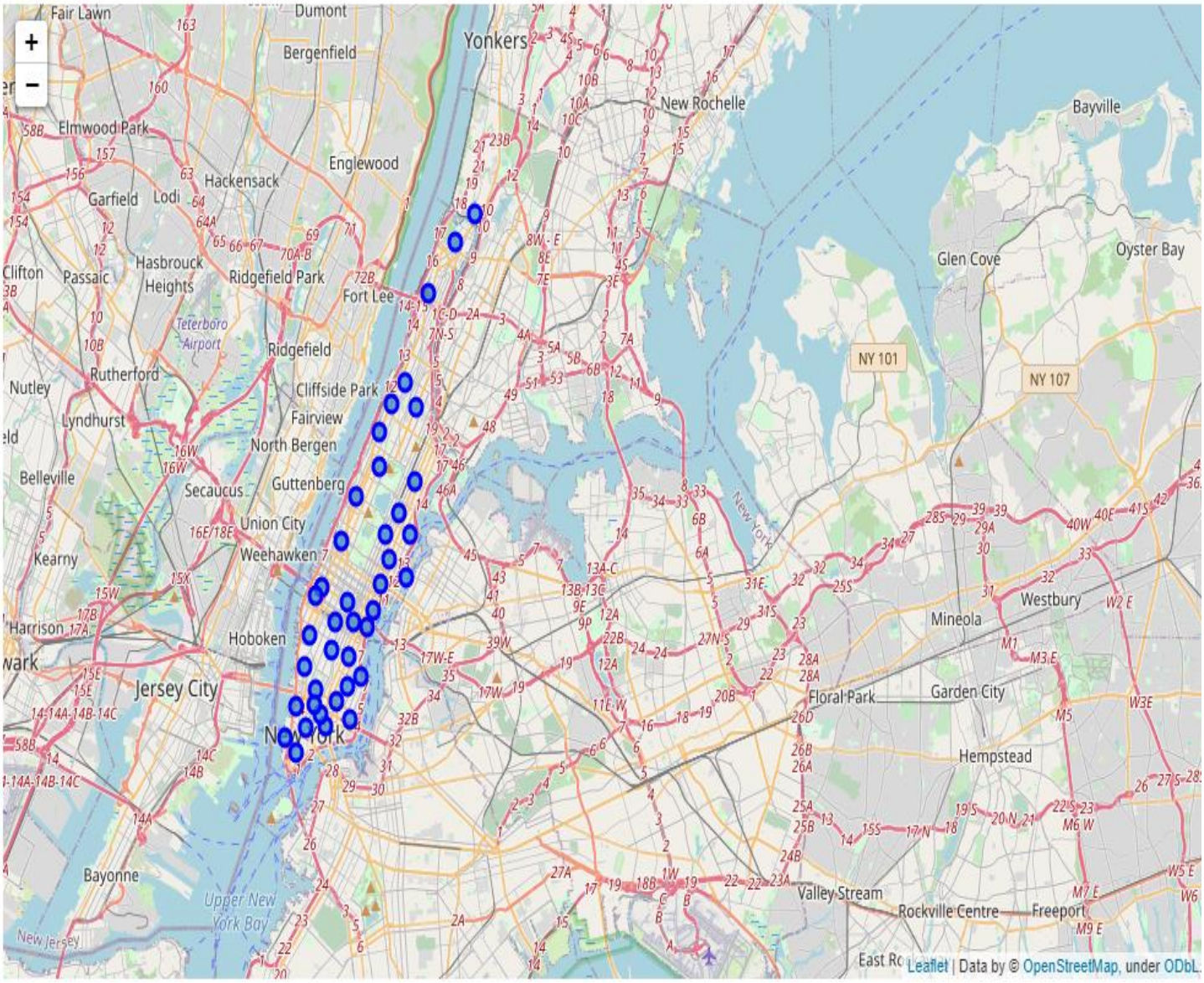
3. Visualize neighborhoods on map

```
import folium

# create map of Manhattan using latitude and longitude values
map_manhattan = folium.Map(location=[manhattan_latitude, manhattan_longitude], zoom_start=11)

# add markers to map
for lat, lng, label in zip(manhattan_data['Latitude'], manhattan_data['Longitude'], manhattan_data['Neighborhood']):
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_manhattan)

map_manhattan
```



4. Explore neighborhoods use Foursquare API

```
manhattan_venues = getVenues(names=manhattan_data['Neighborhood'],
                                latitudes=manhattan_data['Latitude'],
                                longitudes=manhattan_data['Longitude']
                               )
manhattan_venues.head()
```

```
Marble Hill
Chinatown
Washington Heights
Inwood
Hamilton Heights
Manhattanville
Central Harlem
East Harlem
Upper East Side
Yorkville
Lenox Hill
Roosevelt Island
Upper West Side
Lincoln Square
Clinton
Midtown
Murray Hill
Chelsea
Greenwich Village
East Village
Lower East Side
Tribeca
Little Italy
Soho
West Village
Manhattan Valley
Morningside Heights
Gramercy
Battery Park City
Financial District
Carnegie Hill
Noho
Civic Center
Midtown South
Sutton Place
Turtle Bay
Tudor City
Stuyvesant Town
Flatiron
Hudson Yards
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

5. Find neighborhoods with most food venues after clean up data

```
result=manhattan_grouped_sorted.loc[:,['Neighborhood','Total Visited Frequecy']]
result.head(10)
```

	Neighborhood	Total Visited Frequecy
0	East Village	0.590000
1	Upper West Side	0.555556
2	Manhattanville	0.533333
3	Turtle Bay	0.520000
4	Greenwich Village	0.520000
5	Central Harlem	0.511111
6	Chinatown	0.480000
7	Hamilton Heights	0.457627
8	Inwood	0.456140
9	West Village	0.450000

6. Find top 10 neighborhoods without Chinese Restaurant

```
no_ch_neighborhood = manhattan_grouped_sorted[manhattan_grouped_sorted['Chinese Restaurant']==0].reset_index(drop=True)


food_neighborhoods = no_ch_neighborhood.drop(columns='Total Visited Frequecy')
food_neighborhoods.head(10)
```

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Video Store	Vietnamese Restaurant	Volleyball Court
0	Turtle Bay	0.0	0.0	0.0	0.0	0.020000	0.00	0.0	0.00	0.000000	...	0.0	0.000000	0.0
1	East Harlem	0.0	0.0	0.0	0.0	0.000000	0.00	0.0	0.00	0.000000	...	0.0	0.000000	0.0
2	Manhattan Valley	0.0	0.0	0.0	0.0	0.000000	0.00	0.0	0.00	0.000000	...	0.0	0.021739	0.0
3	Noho	0.0	0.0	0.0	0.0	0.020000	0.00	0.0	0.01	0.040000	...	0.0	0.000000	0.0
4	Gramercy	0.0	0.0	0.0	0.0	0.044444	0.00	0.0	0.00	0.011111	...	0.0	0.000000	0.0
5	Civic Center	0.0	0.0	0.0	0.0	0.020000	0.01	0.0	0.00	0.010000	...	0.0	0.000000	0.0
6	Flatiron	0.0	0.0	0.0	0.0	0.030000	0.00	0.0	0.00	0.010000	...	0.0	0.000000	0.0
7	Financial District	0.0	0.0	0.0	0.0	0.040000	0.00	0.0	0.00	0.000000	...	0.0	0.000000	0.0
8	Morningside Heights	0.0	0.0	0.0	0.0	0.073171	0.00	0.0	0.00	0.000000	...	0.0	0.000000	0.0
9	Hudson Yards	0.0	0.0	0.0	0.0	0.053571	0.00	0.0	0.00	0.000000	...	0.0	0.000000	0.0

7. Use K-means cluster neighborhoods:

```
# import k-means from clustering stage
from sklearn.cluster import KMeans

# set number of clusters
kclusters = 5

manhattan_grouped_clustering = food_neighborhoods.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(manhattan_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]

# add clustering labels
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

manhattan_merged = manhattan_data

manhattan_merged = manhattan_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

manhattan_merged = manhattan_merged.dropna().reset_index(drop=True)

manhattan_merged.head() # check the last columns!
```

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	4.0	Coffee Shop	Gym	Yoga Studio	Big Box Store	Supplement Shop	Steakhouse	Shopping Mall
1	Manhattan	East Harlem	40.792249	-73.944182	3.0	Mexican Restaurant	Bakery	Thai Restaurant	Deli / Bodega	Spa	Latin American Restaurant	Sandwich Place
2	Manhattan	Roosevelt Island	40.762160	-73.949168	2.0	Park	Restaurant	Residential Building (Apartment / Condo)	Sandwich Place	Dry Cleaner	Liquor Store	Outdoors & Recreation
3	Manhattan	Manhattan Valley	40.797307	-73.964286	1.0	Mexican Restaurant	Bar	Thai Restaurant	Pizza Place	Park	Coffee Shop	Yoga Studio
4	Manhattan	Morningside Heights	40.808000	-73.963896	1.0	Park	Bookstore	American Restaurant	Coffee Shop	Burger Joint	Deli / Bodega	Sandwich Place

8. Visualize plot on map

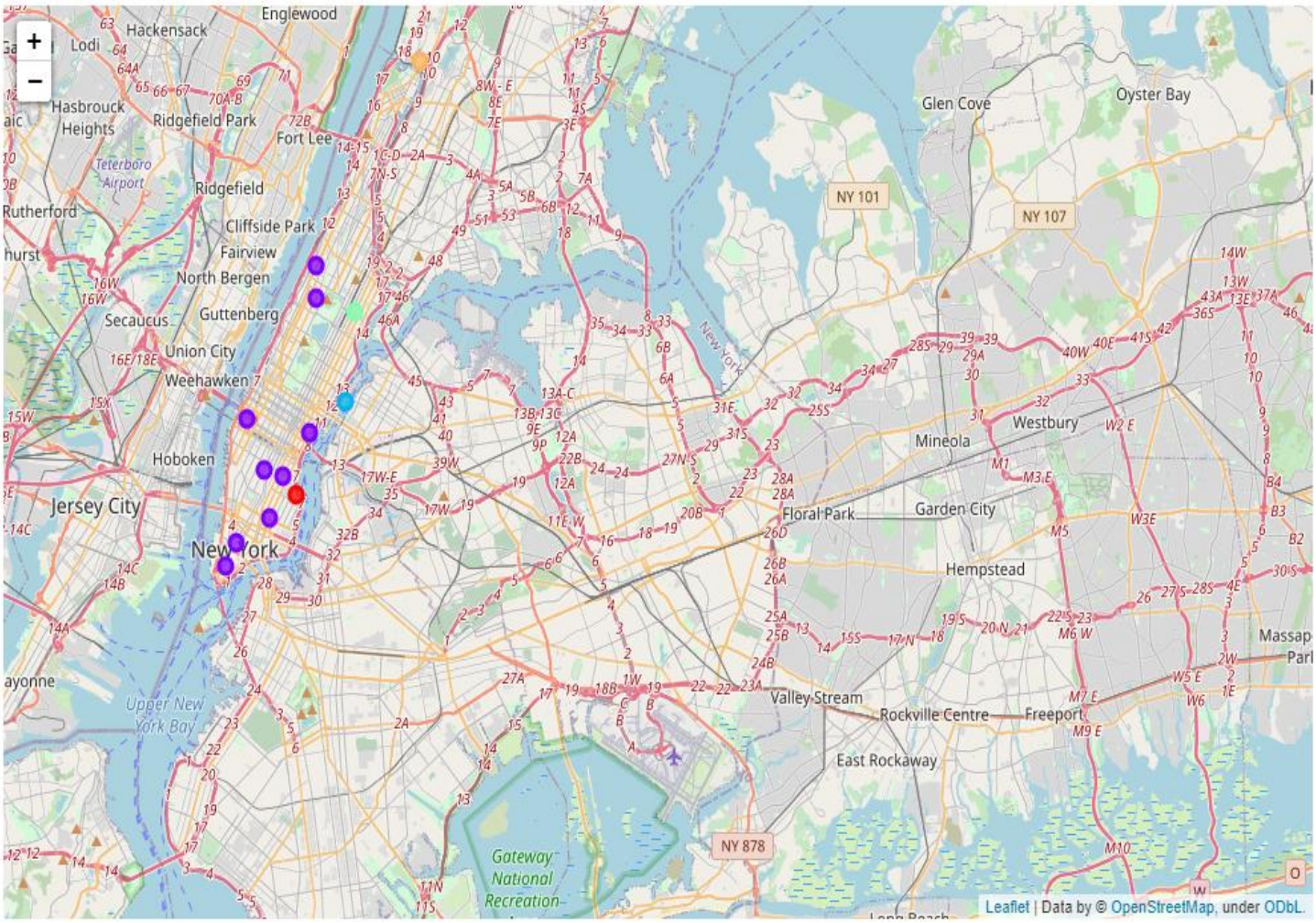
```
# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors

# create map
map_clusters = folium.Map(location=[manhattan_latitude, manhattan_longitude], zoom_start=11)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

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

map_clusters
```



9. Exam clusters:

```
tan_merged.loc[manhattan_merged['Cluster Labels'] == 2, manhattan_merged.columns[[1] + list(range(5, manhattan_merged.shape[1]))
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Roosevelt Island	Park	Restaurant	Residential Building (Apartment / Condo)	Sandwich Place	Dry Cleaner	Liquor Store	Outdoors & Recreation	Coffee Shop	Supermarket	Baseball Field

```
tan_merged.loc[manhattan_merged['Cluster Labels'] == 3, manhattan_merged.columns[[1] + list(range(5, manhattan_merged.shape[1]))
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	East Harlem	Mexican Restaurant	Bakery	Thai Restaurant	Deli / Bodega	Spa	Latin American Restaurant	Sandwich Place	Taco Place	Donut Shop	Cocktail Bar

```
tan_merged.loc[manhattan_merged['Cluster Labels'] == 4, manhattan_merged.columns[[1] + list(range(5, manhattan_merged.shape[1]))
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Marble Hill	Coffee Shop	Gym	Yoga Studio	Big Box Store	Supplement Shop	Steakhouse	Shopping Mall	Seafood Restaurant	Sandwich Place	Donut Shop

```
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 0, manhattan_merged.columns[[1] + list(range(5, manhattan_merged.shape
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
10	Stuyvesant Town	Boat or Ferry	Park	Baseball Field	Heliport	Gas Station	Skating Rink	Farmers Market	Bistro	Gym / Fitness Center	Cocktail Bar

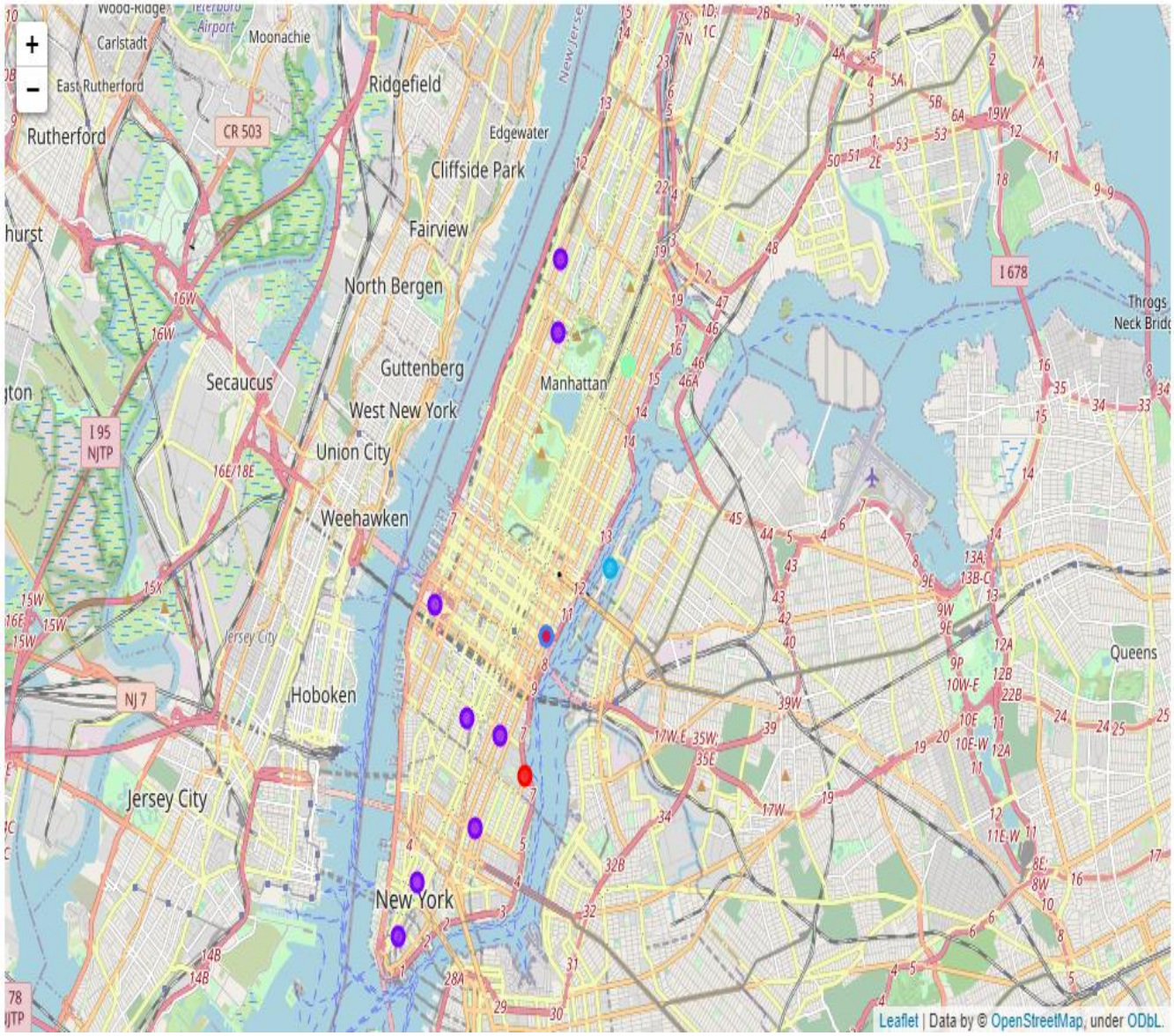
```
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 1, manhattan_merged.columns[[1] + list(range(5, manhattan_merged.shape
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	Manhattan Valley	Mexican Restaurant	Bar	Thai Restaurant	Pizza Place	Park	Coffee Shop	Yoga Studio	Fried Chicken Joint	Clothing Store	Ice Cream Shop
4	Morningside Heights	Park	Bookstore	American Restaurant	Coffee Shop	Burger Joint	Deli / Bodega	Sandwich Place	Pub	Supermarket	Mediterranean Restaurant
5	Gramercy	Bar	Pizza Place	Italian Restaurant	American Restaurant	Coffee Shop	Playground	Cocktail Bar	Mexican Restaurant	Bagel Shop	Grocery Store
6	Financial District	Coffee Shop	Pizza Place	American Restaurant	Café	Cocktail Bar	Gym	Gym / Fitness Center	Italian Restaurant	Bar	Park
7	Noho	Italian Restaurant	Coffee Shop	Pizza Place	Art Gallery	French Restaurant	Grocery Store	Sandwich Place	Rock Club	Sushi Restaurant	Mexican Restaurant
8	Civic Center	Coffee Shop	Hotel	Cocktail Bar	Gym / Fitness Center	French Restaurant	Italian Restaurant	Yoga Studio	Spa	Park	Sushi Restaurant
9	Turtle Bay	Sushi Restaurant	Italian Restaurant	Coffee Shop	Wine Bar	French Restaurant	Park	Deli / Bodega	Japanese Restaurant	Seafood Restaurant	Café
11	Flatiron	Gym / Fitness Center	New American Restaurant	Spa	Italian Restaurant	Mediterranean Restaurant	Furniture / Home Store	Vegetarian / Vegan Restaurant	Japanese Restaurant	Gym	American Restaurant
12	Hudson Yards	Hotel	Gym / Fitness Center	Italian Restaurant	American Restaurant	Coffee Shop	Dog Run	Nightclub	Gym	Park	Spanish Restaurant

10. Visualize and plot most visit venues for food (dot has red color covered with light blue)

```
for lat, lon, poi, cluster in zip(top_10_neigh_df['Latitude'], top_10_neigh_df['Longitude'], top_10_neigh_df['Neighborhood'], top
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        fill=True,
        fill_color=rainbow[4],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```



RESULT

- Turtle Bay is the most promising neighborhood shareholder should consider for a new Chinese restaurant.
- Cluster 1 is the most visited venues that covered with light blue.

CONCLUSION:

- Manhattan has totally 40 neighborhoods
- Top 10 neighborhoods with most food venues are:
    - East Village *freq=0.59*,
    - Upper West Side *freq=0.56*,
    - Manhattanville *freq=0.53*,
    - Turtle Bay *freq=0.52*,
    - Greenwich Village *freq=0.52*,
    - Central Harlem *freq=0.51*,
    - Chinatown *freq=0.48*,
    - Hamilton Heights *freq=0.46*,
    - Inwood *freq=0.46*,
    - West Village *freq=0.45*
- Turtle Bay is the best location with most food venues and no Chinese restaurant