IBM Applied Data Science Capstone

Recommendations on Location and Food Type of a Restaurant Targeting Tourists in New York

Boyan Gu

1. Introduction

1.1 Background

New York is the most populous city in US, and is an international center of business, finance, arts, and culture. It is recognized as one of the most multicultural and cosmopolitan cities in the world. Its economy is highly diversified with strengths in technology, design, education, arts, tourism, and etc. Therefore, it will be of great chance to open a restaurant successfully and attractive customers, especially tourists.

1.2 Business Problem

The target audiences of this project are investors who want to open a new restaurant targeting tourists. The problems for the investors now are where to open a restaurant is least competitive and can make a great profit, and which food types are popular and attractive for tourists in New York. We could simply assume that restaurants visited most are those locate near tourism-related venues (such as Arts & Entertainment, Nightlife, Outdoors & Recreation), which are convenient and attractive for tourists to get there. Therefore, we could locate tourism venues in New York, and do clustering to determine tourism cluster centers, where there are more tourists. Then we analyze the competition levels of restaurants around those centers, and determine the trending type of food that are favorable by visitors.

2 Data

This report would use location data of tourism-related venues, and foot-traffic data in restaurants, and their types. New York json data would be scrapped from website, and Foursquare API would be used to get data of trending venues for analysis.

1. We scrap the json data including New York neighborhoods from the website: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json
Each feature in the json includes following information.

Down load json data using url

```
url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsN
r = requests.get(url)
data = r.json()
```

```
neighborhoods_data = data['features']
neighborhoods_data[0]
```

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

There are 5 boroughs and 306 neighborhoods in New York.

2. We use ***Categories*** call to check the main categories and sub-categories in Foursquare API, and then use ***Explore*** calls to query trending venues in each neighborhood of New York region. The following is example of result of Foursquare API call.

```
# create the API request URL
url = 'https://api.foursquare.com/v2/venues/categories?&client_id={}&client_secret={}&v={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION)

# make the GET request
results = requests.get(url).json()
```

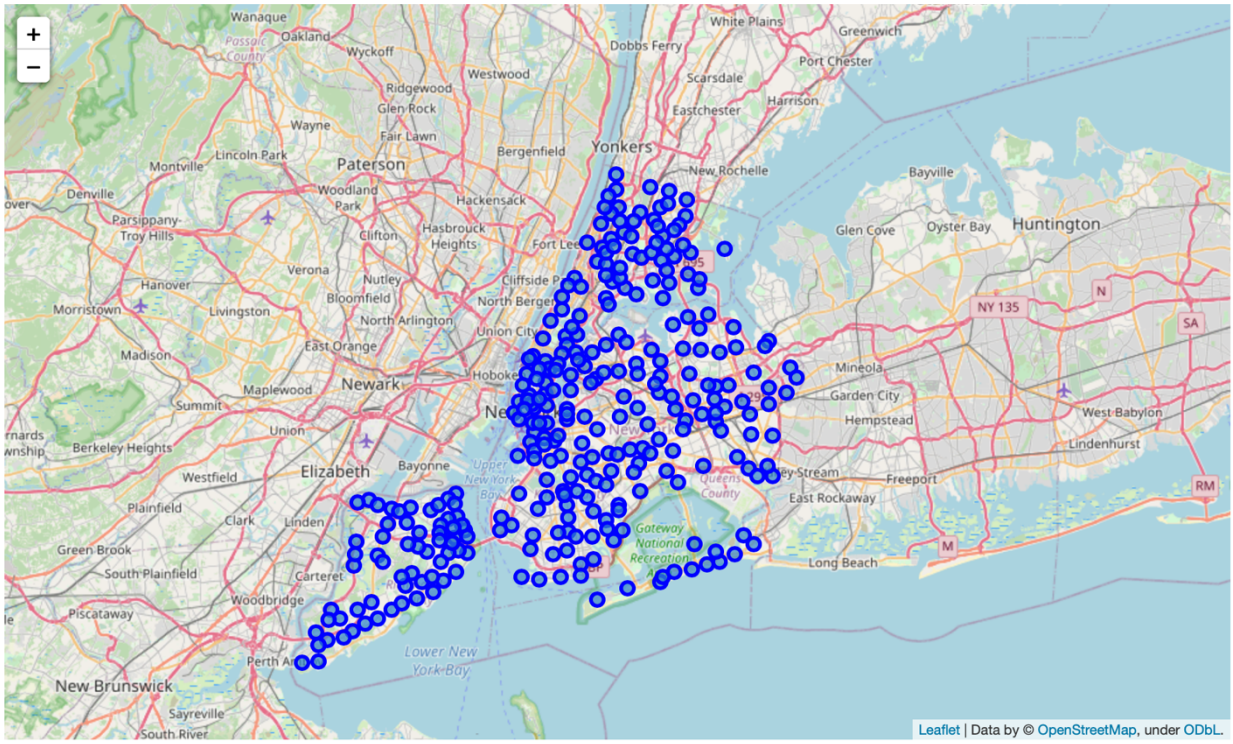
```
results
```

```
{'meta': {'code': 200, 'requestId': '6063f7682138947f202b9cc7'},
 'response': {'categories': [{'id': '4d4b7104d754a06370d81259',
    'name': 'Arts & Entertainment',
    'pluralName': 'Arts & Entertainment',
    'shortName': 'Arts & Entertainment',
    'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/arts_entertainment/default_',
     'suffix': '.png'},
    'categories': [{'id': '56aa371be4b08b9a8d5734db',
      'name': 'Amphitheater',
      'pluralName': 'Amphitheaters',
      'shortName': 'Amphitheater',
      'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/arts_entertainment/default_',
       'suffix': '.png'},
      'categories': []},
     {'id': '4fceea171983d5d06c3e9823',
      'name': 'Aquarium',
      'pluralName': 'Aquariums',
      'shortName': 'Aquarium',
      'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/arts_entertainment/aquarium_',
```

3 Methodology

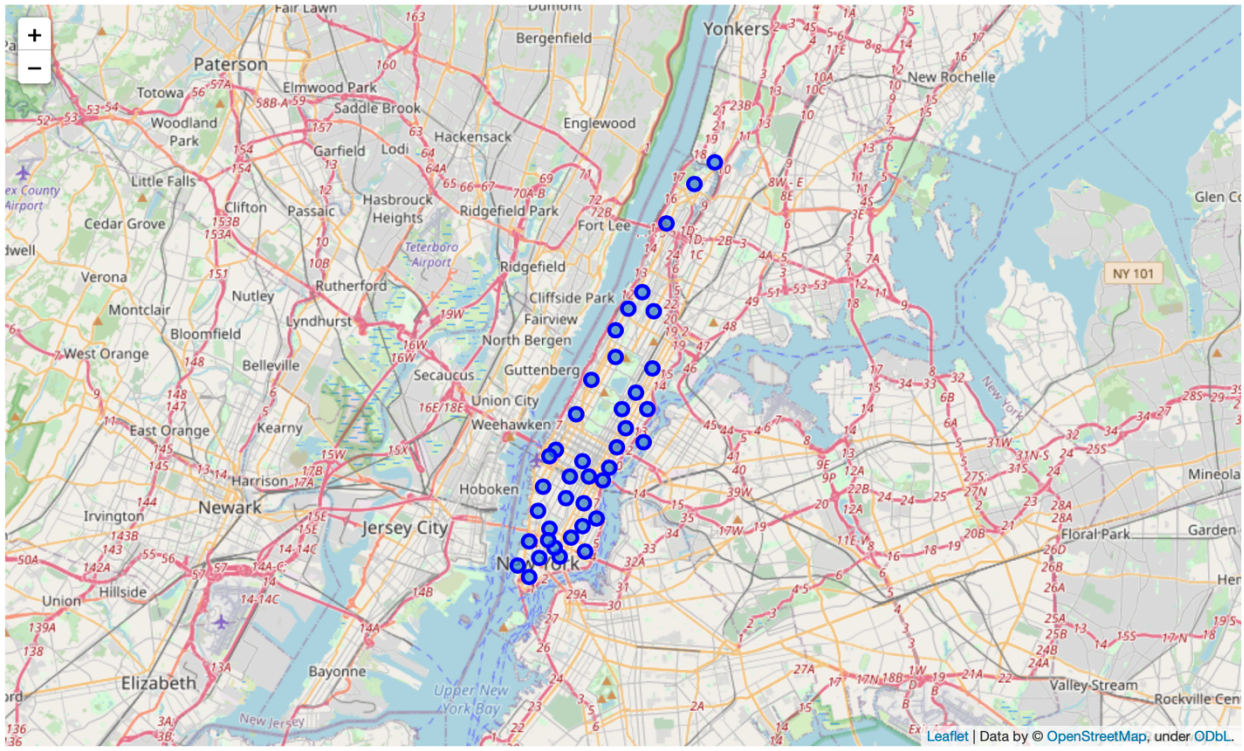
We first collect the data needed for this project (shown in previous section). We retrieve json data from website. There are 5 boroughs and 306 neighborhoods in New York. We then create a map of New York with neighborhoods superimposed on top.

```
map_newyork
```



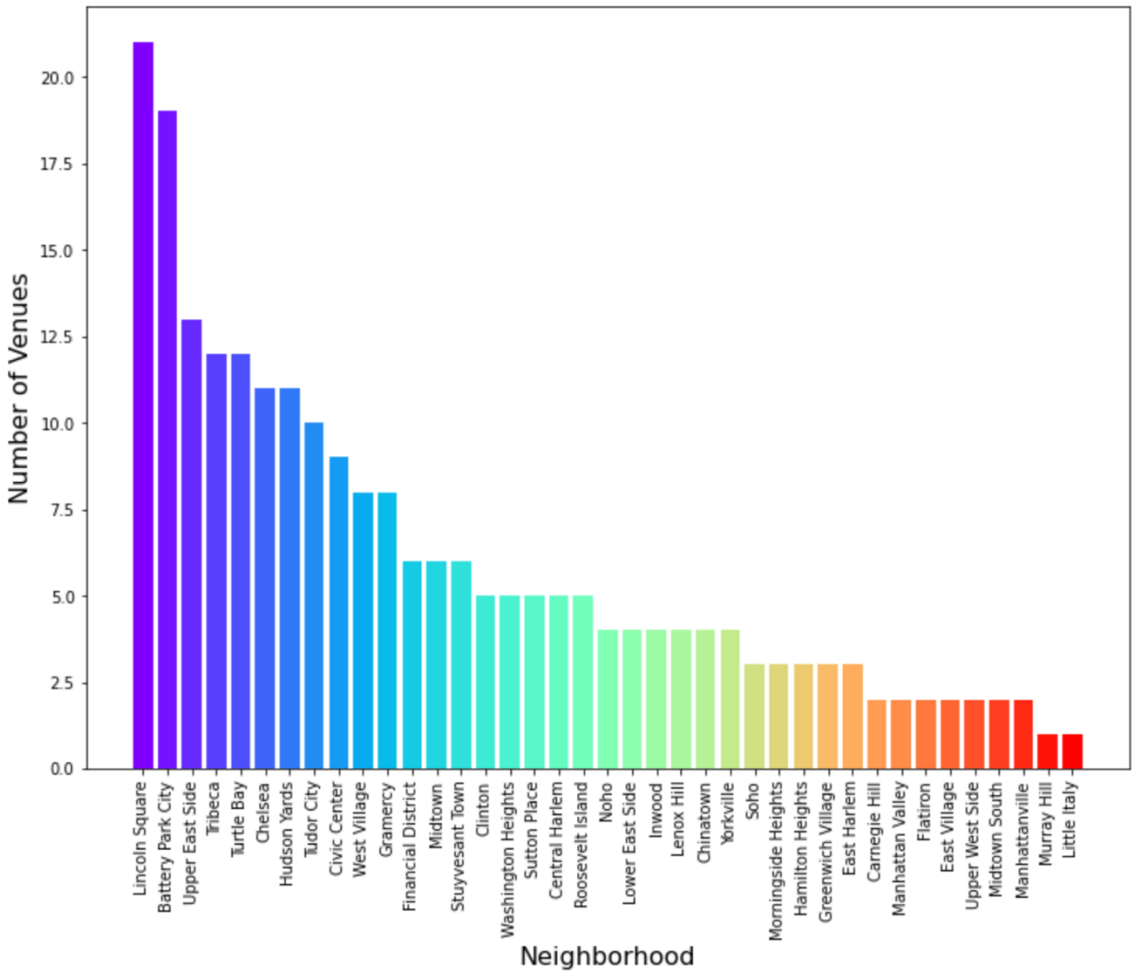
From this map, we can see that Manhattan Borough has dense neighborhoods. Based on our knowledge, there are also many tourist attractions in Manhattan, and therefore we choose Manhattan Borough for further analysis (due API call limit).

```
map_manhattan
```



Then we use Foursquare API calls to explore venues in Manhattan. We first make get request to check the number of main categories and sub-categories present in Foursquare API. The main cat egories include ‘Arts & Entertainment’, ‘College & University’, ‘Event’, ‘Food’, ‘Nightlife Spo t’, ‘Outdoors & Recreation’, ‘Professional & Other Places’, ‘Residence’, ‘Shop & Service’, and ‘Travel & Transport’. We divide these categories into groups: we believe Arts & Entertainment, Nightlife Spot, and Outdoors & recreations are related to tourism, and Food, Shop & Services, an d Travel & Transport are relevant tourist services, and others are not that related to tourism or to urists.

Next, we use ***Explore*** calls to retrieve the ambient tourism related venues within 500 m of neighborhoods in Manhattan. There are 232 venues and 34 unique categories. The following figure shows the number of nearby tourism venues in each neighborhood. The Lincoin Square has the most tourism venues (21 venues).



Then we further use one-hot encoding to analyze each neighborhood. We group rows by venues and by taking the mean of the frequency of occurrence of each category, and print each neighborhood along with the top 5 most common venues. The following figure shows top 5 most common venues in first two neighborhoods.

```
----Battery Park City----
           venue  freq
0           Park  0.42
1  Memorial Site  0.16
2           Plaza  0.11
3     Playground  0.11
4         Garden  0.05


----Carnegie Hill----
         venue  freq
0       Museum   0.5
1   Playground   0.5
2  Art Gallery   0.0
3         Pool   0.0
4         Park   0.0
```

We group venues into venue categories in each neighborhood.

```
venue_venue_category = manhattan_venues.groupby(['Neighborhood', 'Venue Category'], as_index=False).agg(lambda x: ",
venue_venue_category.head()
```

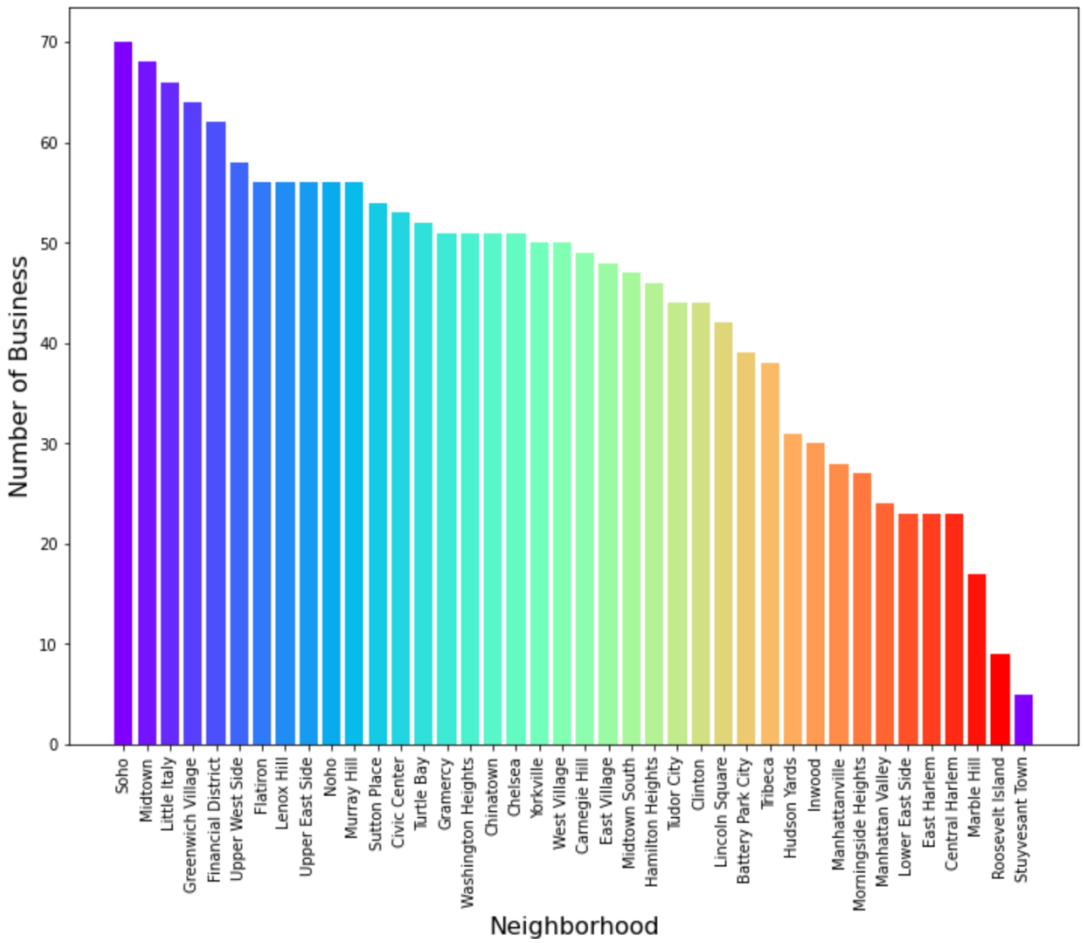
	Neighborhood	Venue Category	Venue
0	Battery Park City	Athletics & Sports	Asphalt Green Battery Park City
1	Battery Park City	Garden	Liberty Community Garden
2	Battery Park City	Memorial Site	9/11 Memorial North Pool, National September 1...
3	Battery Park City	Park	Battery Park City Esplanade, Hudson River Trai...
4	Battery Park City	Performing Arts Venue	Winter Garden Atrium

We also sort the most common venues according to venue category in each neighborhood.

```
neighborhoods_venues_sorted.head()
```

	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	Battery Park City	Battery Park City Esplanade, Hudson River Trai...	9/11 Memorial North Pool, National September 1...	Waterfront Plaza, Brookfield Place, Oculus Plaza	Kowsky Plaza Playground, West Thames Playground	One World Observatory	Winter Garden Atrium	Asphalt Green Battery Park City	Liberty Community Garden		
1	Carnegie Hill	Samuel Seabury Playground	The Jewish Museum								
2	Central Harlem	La masion d'Art, Tatiana Pagès Gallery	Big L Memorial Mural	St. Nicholas Park	Shrine World Music Venue						
3	Chelsea	David Zwirner Gallery, Milk Gallery, Gagosian ...	High Line, Clement Clarke Moore Park	PH-D at Dream Downtown, 1 OAK	London Terrace Gardens Courtyard	High Line 10th Ave Amphitheatre					
4	Chinatown	Sofar HQ	Chinatown Soup	The Crown	Museum at Eldridge Street						

It gives us some intuition that Battery Park City, Financial District, Hudson Yards, Lincoin Square, Midtown, Turtle Bay, Upper East Side might be good location to start a restaurant, since these places with many venues in different venue categories may be attractive to tourists. We then want to find the most common tourist-related business near tourism spots. We make another round Foursquare API calls to retrieve popular business within 500 m of neighborhood in Manhattan. The following figure shows the number of nearby tourist-related businesses in each neighborhood. Soho has the most businesses (70 venues).



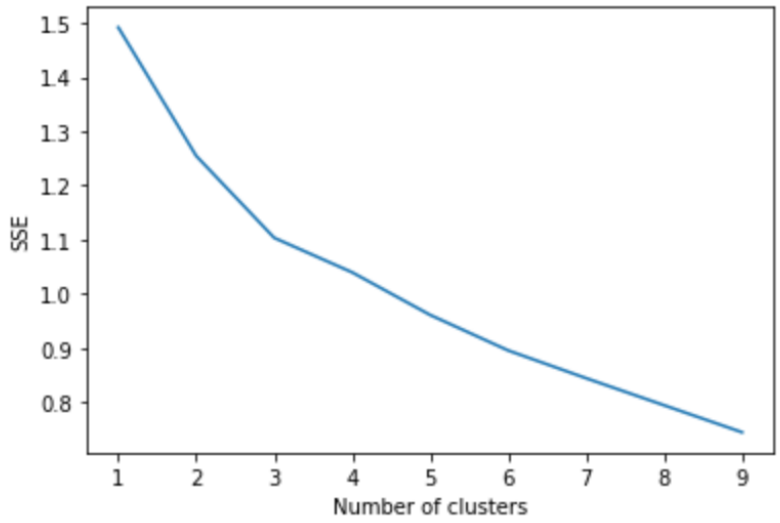
Similar one-hot encoding analysis also apply to nearby business in each neighborhood.

	Neighborhood	1st Most Common Business	2nd Most Common Business	3rd Most Common Business	4th Most Common Business	5th Most Common Business	6th Most Common Business	7th Most Common Business	8th Most Common Business	9th Most Common Business	10th Most Common Business
0	Battery Park City	Coffee Shop	Clothing Store	Hotel	Food Court	Sandwich Place	Pizza Place	Italian Restaurant	Burger Joint	BBQ Joint	Shopping Mall
1	Carnegie Hill	Coffee Shop	Café	French Restaurant	Bookstore	Cosmetics Shop	Italian Restaurant	Shipping Store	Bakery	Pizza Place	Bank
2	Central Harlem	African Restaurant	French Restaurant	Cosmetics Shop	American Restaurant	Seafood Restaurant	Caribbean Restaurant	Southern / Soul Food Restaurant	Spa	Pizza Place	Café
3	Chelsea	Coffee Shop	Bakery	Italian Restaurant	American Restaurant	French Restaurant	Hotel	Seafood Restaurant	Pet Store	Café	Market
4	Chinatown	Bakery	Salon / Barbershop	Spa	American Restaurant	Dessert Shop	Mexican Restaurant	Optical Shop	Bubble Tea Shop	Asian Restaurant	Dumpling Restaurant
5	Civic Center	Coffee Shop	Spa	French Restaurant	American Restaurant	Hotel	Italian Restaurant	Bakery	Falafel Restaurant	Wings Joint	Bubble Tea Shop

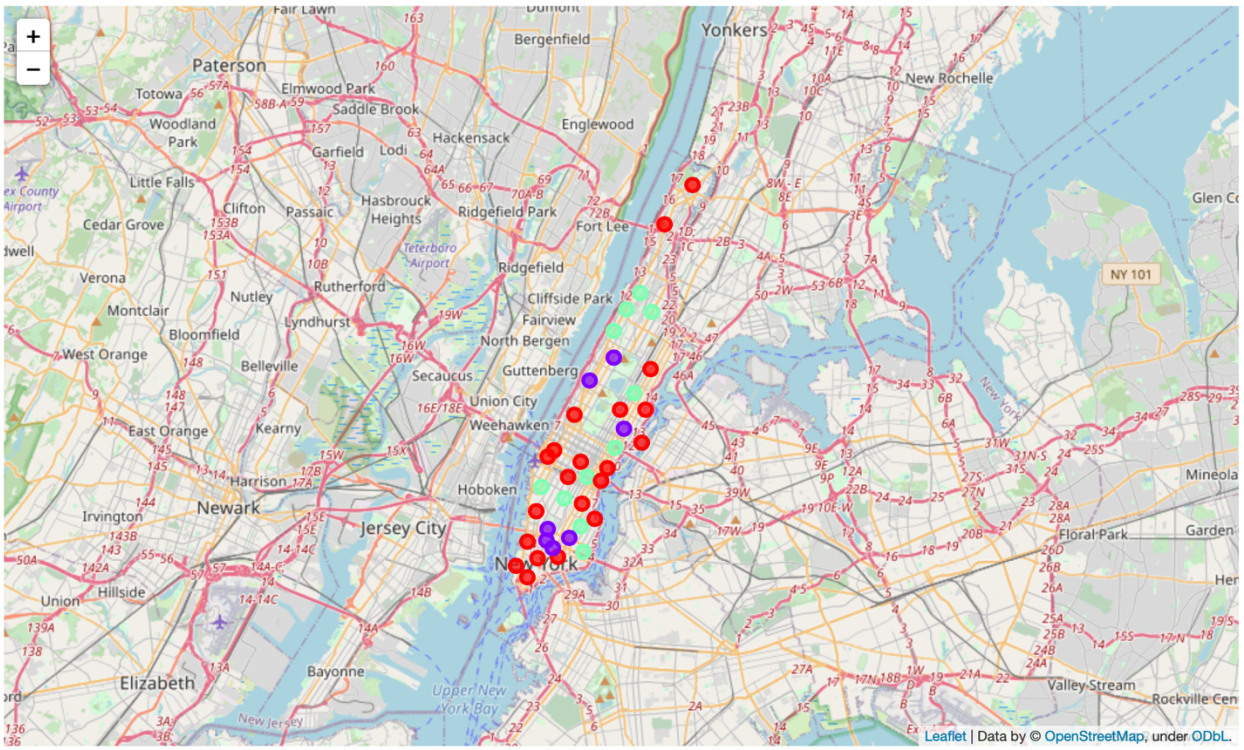
Furthermore, K-means Clustering is applied to cluster neighborhoods in Manhattan. We cluster the tourism attractions and tourist-related businesses separately.

4 Results

We use elbow method for optimal of k in K-means Clustering.

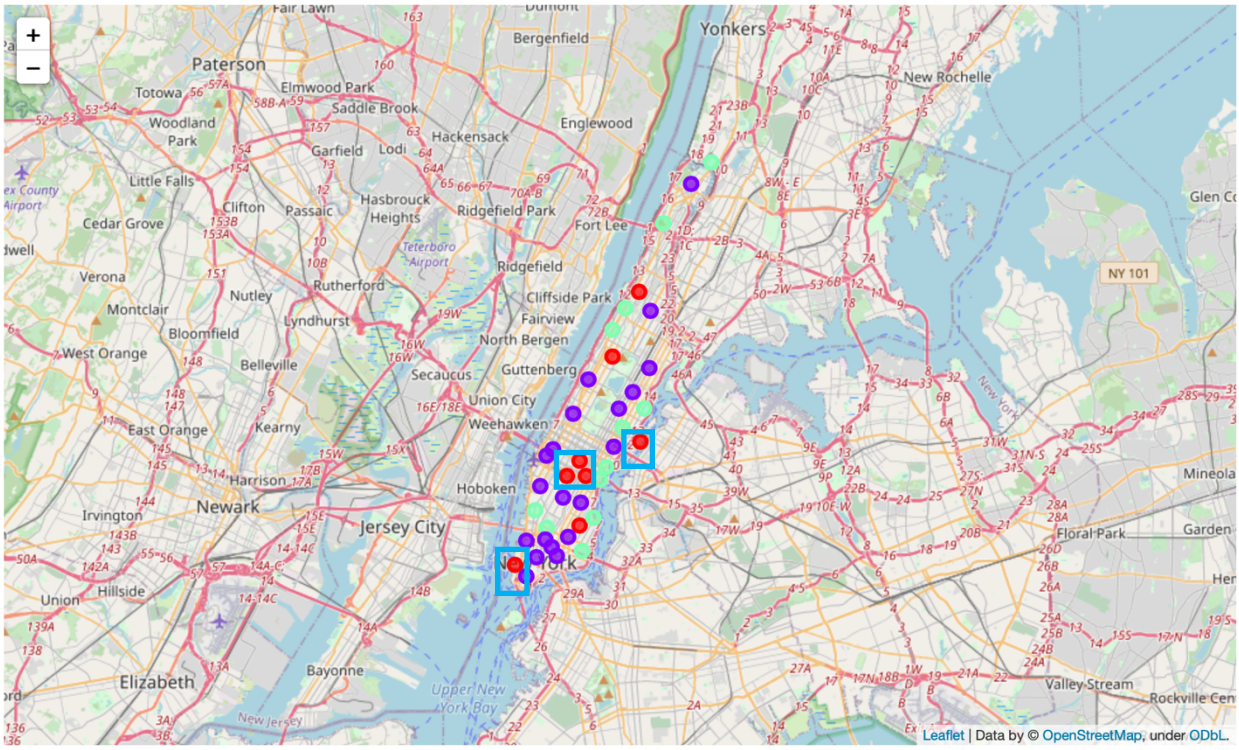


Then we choose k = 3 for clustering. The following figure is the result of clustering for tourism venues in Manhattan.



Cluster 1 (red) includes neighborhoods containing more categories and more tourism venues, especially museums, art centers, and parks, which may attract many tourists compared to the other two clusters. Cluster 2 (purple) and cluster 3 (green) have less tourism venues, and cluster 2 has neighborhoods with least tourism attractions. Intuitionally, neighborhoods in cluster 1 are popular among tourists.

We then look at clustering for tourist-related businesses. We also choose k = 3 for clustering due to elbow method.



The most popular food types of cluster 1 (red) are Mexican food, American food, and Italian food. Italian food, American food, and French food are welcomed in neighborhoods of cluster 2 (purple). Cluster 3 (green) has neighborhoods that prefer Italian food, fast food, American food, and seafood most. In general, Italian food and American food are popular in Manhattan region. Neighborhoods in cluster 2 have relatively fierce competition, and cluster 1 has relatively small pressure of competition.

Compared two clustering figures, we find that most neighborhoods in (southern part of) cluster 1 of tourist-related business locate at cluster 1 of tourism venues, where there are supposed to be more tourists and less competitors of Italian or American food. We recommend investors to open an Italian or American restaurant at those neighborhoods.

5 Conclusion

In this project, we use data retrieved from website and Foursquare API to determine applicable location and food type of a restaurant targeting tourists in New York. After data analysis and K-Means Clustering, we find good locations (blue squared) with low competence level near tourism venues. In the future, we may further use Foursquare API to get foot-traffic data of restaurants, and determine the popular locations and food types with more statistical data.

Libraries Which are Used to Develope the Project:
Pandas: For creating and manipulating dataframes.
Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.
Scikit Learn: For importing k-means clustering.
JSON: Library to handle JSON files.

Geocoder: To retrieve Location Data.
Matplotlib: Python Plotting Module.