Clustering of Neighborhoods in Manhattan Based on Average Rental and Nearby Features

The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

1. Introduction/Business Problem

The initial problem defined is to determine the **best location for a someone to stay** in **Manhattan**, one of the biggest metropolitan area in the world with high living standards and expenses.

It will be important for the audiences to understand which area they should be aiming for to strike for a balance between **rental price** and **proximity to the various facilities** in the town.

By clustering the neighborhoods in Manhattan, the audiences can decide the exact area they are looking for based on their own considerations.

2. Data

a) List of neighborhoods in Manhattan will be obtained from a data set that contains all of the neighborhoods in New York.

b) Coordinate of each neighborhood will be obtained using Geopy Client.

c) Average rent in Manhattan of each neighborhood will be extracted through web-scraping from https://www.rentcafe.com/average-rent-market-trends/us/ny/manhattan/.

d) Foursquare location data will be used together with the average rent to cluster the neighborhoods.

3. Methodology

a) A list of boroughs and neighborhoods in New York was extracted from a dataset.

```
In [2]: neighborhood = pd.read_csv(r'C:\Users\jiahao.lo\Desktop\projects\IBM-Data-Science\NeighborhoodDF.csv')

        neighborhood.head()
```

Out[2]:

	Borough	Neighborhood
0	Bronx	Wakefield
1	Bronx	Co-op City
2	Bronx	Eastchester
3	Bronx	Fieldston
4	Bronx	Riverdale

b) Neighborhoods from Manhattan is extracted from the data frame into a new data frame.

```
In [4]: manhattan_df = neighborhood[neighborhood['Borough'] == 'Manhattan'].reset_index(drop=True)
        manhattan_df
```

Out[4]:

	Borough	Neighborhood
0	Manhattan	Marble Hill
1	Manhattan	Chinatown
2	Manhattan	Washington Heights
3	Manhattan	Inwood
4	Manhattan	Hamilton Heights

c) Average rent in Manhattan of each neighborhood will be extracted through web-scraping using BeautifulSoup from https://www.rentcafe.com/average-rent-market-trends/us/ny/manhattan/.

```
In [7]: from bs4 import BeautifulSoup

        response_obj = requests.get('https://www.rentcafe.com/average-rent-market-trends/us/ny/manhattan/').text
        soup=BeautifulSoup(response_obj,'lxml')
        df_rent = soup.find('table')
```

```
In [8]: results = soup.find("div", attrs={'class':'table-neighborhood'})
```

d) Through some data cleaning and wrangling, the data frame with borough, neighborhood and data frame will be formed as below.

	Borough	Neighborhood	Average Rent
0	Manhattan	Marble Hill	$1,658
1	Manhattan	Chinatown	$4,875
2	Manhattan	Washington Heights	$2,217
3	Manhattan	Inwood	$2,305
4	Manhattan	East Harlem	$2,785

e) Coordinate of each neighborhood will be obtained using Geopy Client.

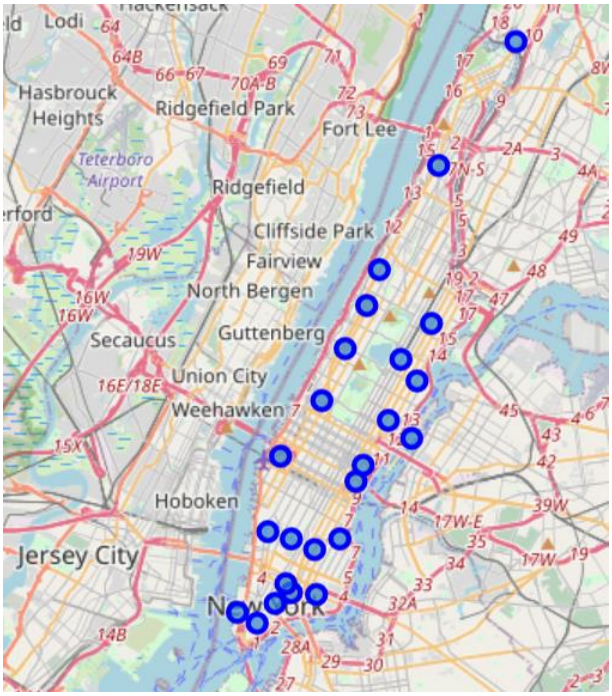
```
address = 'Manhattan, NY'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Manhattan are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Manhattan are 40.7896239, -73.9598939.
```

```
df_all['Latitude']= df_all['Neighborhood'].apply(geolocator.geocode).apply(lambda x: x.latitude)
df_all['Longitude']= df_all['Neighborhood'].apply(geolocator.geocode).apply(lambda x: x.longitude)
df_all
```

	Borough	Neighborhood	Average Rent	Latitude	Longitude
0	Manhattan	Marble Hill	$1,658	40.876298	-73.910429
1	Manhattan	Chinatown	$4,875	40.716491	-73.996250
2	Manhattan	Washington Heights	$2,217	40.840198	-73.940221
3	Manhattan	Inwood	$2,305	43.307926	-96.432184
4	Manhattan	East Harlem	$2,785	40.794722	-73.942500

f) Map of Manhattan with neighborhoods markers.



g) Dataset of nearby features around the neighborhoods is obtained through foursquare API.

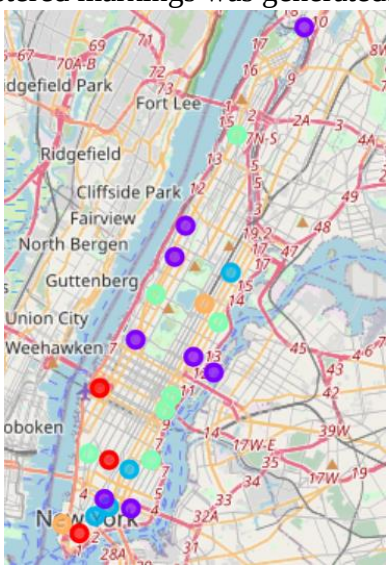
	Neighborhood	Accessories Store	Afghan Restaurant	American Restaurant	Animal Shelter	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auditorium	Austral Restaur
0	Battery Park City	0.00	0.00	0.010000	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.01	0
1	Carnegie Hill	0.00	0.00	0.034483	0.00	0.000000	0.00	0.000000	0.034483	0.00	0.000000	0.000000	0.00	0
2	Chelsea	0.00	0.00	0.021277	0.00	0.000000	0.00	0.021277	0.000000	0.00	0.000000	0.000000	0.00	0
3	Chinatown	0.00	0.00	0.000000	0.01	0.000000	0.00	0.000000	0.000000	0.00	0.020000	0.000000	0.00	0
4	Civic Center	0.00	0.00	0.000000	0.00	0.010753	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0
5	East Harlem	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0

h) K-means clustering was carried out to cluster the neighborhoods into 5 cluster based on average rental and Foursquare data.

```
kclusters = 5

manhattan_grouped_clustering = manhattan_grouped.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:-1]
```

```
array([3, 3, 1, 1, 1, 1, 1, 4, 4, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2,
       2, 2, 2, 2])
```

i) Map of Manhattan with clustered markings was generated.



4. Results and Discussions

a) Cluster 1

	Neighborhood	Average Rent	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	8th Most Common Venue
0	Marble Hill	$1,658	40.876298	-73.910429	0	Coffee Shop	Sandwich Place	Discount Store	Gym	Yoga Studio	Kids Store	Steakhouse	Shopping Mall
18	Morningside Heights	$4,142	40.810000	-73.962500	0	Café	Italian Restaurant	Chinese Restaurant	Coffee Shop	Mexican Restaurant	Pharmacy	Deli / Bodega	Park
3	Inwood	$2,305	43.307926	-96.432184	0	Bar	Food	Playground	Event Space	Empanada Restaurant	Drugstore	Dry Cleaner	Dumpling Restaurant
17	Manhattan Valley	$4,441	40.799776	-73.967772	0	Bar	Indian Restaurant	Yoga Studio	Thai Restaurant	Bagel Shop	Bank	Baseball Field	Bistro
15	Little Italy	$5,228	40.719273	-73.998215	0	Bakery	Café	Coffee Shop	Chinese Restaurant	Cocktail Bar	Vietnamese Restaurant	Bubble Tea Shop	Sandwich Place
6	Lenox Hill	$4,048	40.766437	-73.959017	0	Sushi Restaurant	Italian Restaurant	Coffee Shop	Café	Gym	Pizza Place	Burger Joint	Salon / Barbershop
7	Roosevelt Island	$3,313	40.761418	-73.950228	0	Gym	Park	Bus Line	Metro Station	Supermarket	Food & Drink Shop	Bubble Tea Shop	Soccer Field
14	Lower East Side	$4,559	40.715936	-73.986806	0	Mexican Restaurant	Coffee Shop	Bakery	Bar	Cocktail Bar	Café	American Restaurant	Chinese Restaurant
9	Lincoln Square	$4,211	40.772319	-73.984401	0	French Restaurant	Plaza	Coffee Shop	Theater	Café	Performing Arts Venue	Italian Restaurant	Concert Hall
10	Murray Hill	$3,754	38.290348	-85.588294	0	Golf Course	Food	Drugstore	Dry Cleaner	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Electronics Store

There are 10 neighborhoods in Cluster 1. In general, the neighborhoods in cluster 1 have good access to food and beverages and a relatively high rental cost, excepts for Marble Hill with a rental of $1658. If the audience is looking for a location with wide variety of food and beverages with low rental cost, Marble Hill might be a good choice, although the coordinates show that Marble Hill is quite some distance away from the downtown area.

b) Cluster 2

	Neighborhood	Average Rent	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	8th Most Common Venue
11	Chelsea	$3,840	51.487542	-0.168220	1	Bakery	Pub	English Restaurant	French Restaurant	Ice Cream Shop	Park	Japanese Restaurant	Supermarket
4	East Harlem	$2,785	40.794722	-73.942500	1	Bakery	Mexican Restaurant	Pizza Place	Taco Place	Donut Shop	Dance Studio	Steakhouse	Spanish Restaurant
1	Chinatown	$4,875	40.716491	-73.996250	1	Chinese Restaurant	Bakery	Vietnamese Restaurant	Bubble Tea Shop	Dessert Shop	Shanghai Restaurant	Spa	Dim Sum Restaurant
22	Civic Center	$3,493	40.713679	-74.002404	1	Chinese Restaurant	Bubble Tea Shop	Coffee Shop	Bakery	Dim Sum Restaurant	Dessert Shop	Vietnamese Restaurant	Park
13	East Village	$4,024	40.729269	-73.987361	1	Japanese Restaurant	Grocery Store	Dessert Shop	Coffee Shop	Pizza Place	Chinese Restaurant	Korean Restaurant	Wine Bar

There are five neighborhoods in Cluster 2. This cluster is very suitable for an Asian audience to live in as these area offers various Asian cuisines in Chinese, Vietnamese, Japanese and Korean style. Do note that the average rental fees are high while only East Harlem has an average rental less than $3,000.

c) Cluster 3

	Neighborhood	Average Rent	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	8th Most Common Venue
26	Stuyvesant Town	$3,511	40.731971	-73.978093	2	Bar	Pizza Place	Park	Bagel Shop	Coffee Shop	Grocery Store	Mexican Restaurant	Taco Place
8	Upper West Side	$4,249	40.787045	-73.975416	2	Italian Restaurant	Bar	Café	American Restaurant	Wine Bar	Bakery	Pizza Place	Indian Restaurant
5	Yorkville	$4,177	40.778007	-73.948202	2	Italian Restaurant	Bar	Coffee Shop	Gym	Wine Shop	Sushi Restaurant	Mexican Restaurant	Japanese Restaurant
16	West Village	$4,392	40.734186	-74.005580	2	Italian Restaurant	Cocktail Bar	New American Restaurant	American Restaurant	Park	Cosmetics Shop	Wine Bar	Jazz Club
2	Washington Heights	$2,217	40.840198	-73.940221	2	Pizza Place	Mexican Restaurant	Latin American Restaurant	Bakery	Thai Restaurant	Coffee Shop	Bookstore	Chinese Restaurant
24	Turtle Bay	$3,804	40.753467	-73.968866	2	Coffee Shop	Hotel	Sushi Restaurant	Italian Restaurant	French Restaurant	Greek Restaurant	Japanese Restaurant	Indian Restaurant
25	Tudor City	$3,752	40.748623	-73.971389	2	Café	Park	Coffee Shop	Sushi Restaurant	Japanese Restaurant	Deli / Bodega	Gym	American Restaurant

There are seven neighborhoods in Cluster 3. This cluster is very suitable for Caucasian audience to live in as these area offers various western cuisines in Italian, American, Latin American and Mexican style. Do note that the average rental fees are high while only Washington has an average rental less than $3,000.

d) Cluster 4

	Neighborhood	Average Rent	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	8th Most Common Venue
19	Battery Park City	$4,869	40.711017	-74.016937	3	Coffee Shop	Park	Hotel	Clothing Store	Memorial Site	Gym	Wine Shop	Plaza
21	Carnegie Hill	$4,177	40.784197	-73.954339	3	Café	Wine Shop	Gym / Fitness Center	Gym	Bar	Art Museum	Coffee Shop	French Restaurant

There are only two neighborhoods in the cluster, and the observation drawn might require further evidences. Both of the area has a high rental cost greater than $ 4,000, and they have good excess to coffee shops and cafes.

e) Cluster 5

	Neighborhood	Average Rent	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	8th Most Common Venue
20	Financial District	$3,354	40.707668	-74.009271	4	Coffee Shop	Pizza Place	Italian Restaurant	Café	Gym	Hotel	Falafel Restaurant	Juice Bar
12	Greenwich Village	$4,056	40.731980	-73.996566	4	American Restaurant	Coffee Shop	Italian Restaurant	Yoga Studio	Food Truck	Wine Bar	Bar	Salad Place
27	Hudson Yards	$3,809	40.755906	-74.000532	4	American Restaurant	Italian Restaurant	Gym / Fitness Center	Park	Burger Joint	Coffee Shop	Café	Hotel

There are only three neighborhoods in the cluster, and the observation drawn might require further evidences. The average rent is approximately around the range of $3,000 to $4,000, and the neighborhoods have good access to American restaurant and coffee shops.

5. Conclusion

In conclusion, the clustering of neighborhood in Manhattan was successfully carried out with K-means clustering based on the available data found online. We managed to extract some basics insights from this elementary exercise for clustering. The accuracy of this study can be improved by including a bigger dataset, and also experimenting other clustering algorithms.