Open an Italian Restaurant in Manhattan

Final Report

Introduction/Business Problem

Manhattan is the most densely populated of the five boroughs of New York City. Manhattan serves as the city's economic and administrative center. The borough consists mostly of Manhattan Island, bounded by the Hudson, East, and Harlem rivers; as well as several small adjacent islands. Manhattan has been described as the cultural, financial, media, and entertainment capital of the world and thousands of people work in the offices of Manhattan every day.

The Italian cousin is one of the most appreciate cousin in the world. This is due to its tasty and simplicity but also to the quality of the ingredients and to the great variety of dishes.

In this project we will analyze which is the best neighborhood of Manhattan where to open an Italian restaurant. The project is addressed to business people that want to exploit the great amount of people in Manhattan that will for sure be attracted from the Italian cousin.

To select the best neighborhood, we will extract the number of Italian Restaurant in each neighborhood of Manhattan in order to select the neighborhood with few Italian Restaurant already present. Moreover, we will choose the most “central” neighborhood that in this way will be easy to reach from every point of Manhattan.

Data

I will use the New York City dataset that contains Borough, Neighborhoods, Latitudes and Longitudes information. Data were downloaded trough a .csv file from https://data.cityofnewyork.us From this dataset the neighborhood of Manhattan will be extracted.
With the Foursquare API I will get all the venues in Manhattan neighborhood. I will then filter these venues to get only Italian restaurants. With the Italian restaurant dataset I will analyze the best value of k to perform a clustering analysis to extract the cluster of neighborhood with less Italian restaurant. At the end, from this neighborhood I will select the one closest to the center of Manhattan.

Data preparation and methodology

As first step I imported the .csv file with the Borough and Neighborhood of New York to ghether with their Latitude and Longitude coordinates and I transform it into a pandas dataframe (Figure 1)

```
In [1]: # import pandas libriary
        import pandas as pd

In [2]: # import .csv file of New York data and convert it to Pandas Dataframe
        column_names = ['Borough', 'Neighborhood', 'Latitude', 'Longitude']
        df= pd.read_csv("C:/Coursera/NYD.csv", header = None, names = column_names)
        df
```

	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
...	...	...	...	...
294	Staten Island	Lighthouse Hill	40.576506	-74.137927
295	Staten Island	Richmond Valley	40.519541	-74.229571
296	Queens	Malba	40.790602	-73.826678
297	Brooklyn	Highland Park	40.682486	-73.890281
298	Brooklyn	Madison	40.609378	-73.948415

299 rows × 4 columns

Figure 1

After that I defined my foursquare credential and the I extracted from the New York dataset only the Neighborhoods of Manhattan (Figure 2)

```
In [13]: # from the New York data dataframe extract only the neighborhood of Manhattan
         manhattan_data = df[df['Borough'] == 'Manhattan'].reset_index(drop=True)
         manhattan_data.head()
```

Out[13]:

	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

Figure 2

I then define a function to extract all the venues present in the Manhattan neighborhoods using Forsquare API and then I created a pandas dataframe called manhattan venues that contain the venues (Figure 3) and then I grouped the dataframe by neighborhood (Figure 4).

In [18]:
```
print(manhattan_venues.shape)
manhattan_venues.head()
```

```
(3075, 7)
```

Out[18]:

	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

Figure 3

In [19]:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Battery Park City	65	65	65	65	65	65
Carnegie Hill	87	87	87	87	87	87
Central Harlem	45	45	45	45	45	45
Chelsea	100	100	100	100	100	100
Chinatown	100	100	100	100	100	100
Civic Center	99	99	99	99	99	99
Clinton	100	100	100	100	100	100
East Harlem	40	40	40	40	40	40
East Village	100	100	100	100	100	100
Financial District	100	100	100	100	100	100
Flatiron	100	100	100	100	100	100
Gramercy	82	82	82	82	82	82
Greenwich Village	100	100	100	100	100	100
Hamilton Heights	61	61	61	61	61	61
Inwood	58	58	58	58	58	58
Lenox Hill	100	100	100	100	100	100
Lincoln Square	98	98	98	98	98	98
Little Italy	100	100	100	100	100	100
Lower East Side	47	47	47	47	47	47
Manhattan Valley	40	40	40	40	40	40
Manhattanville	47	47	47	47	47	47
Marble Hill	25	25	25	25	25	25
Midtown	100	100	100	100	100	100
Midtown South	100	100	100	100	100	100
Morningside Heights	42	42	42	42	42	42
Murray Hill	85	85	85	85	85	85
Noho	100	100	100	100	100	100
Roosevelt Island	31	31	31	31	31	31
Soho	96	96	96	96	96	96
Stuyvesant Town	18	18	18	18	18	18
Sutton Place	100	100	100	100	100	100
Tribeca	75	75	75	75	75	75
Tudor City	75	75	75	75	75	75
Turtle Bay	100	100	100	100	100	100
Upper East Side	89	89	89	89	89	89
Upper West Side	82	82	82	82	82	82
Washington Heights	88	88	88	88	88	88
West Village	100	100	100	100	100	100
Yorkville	100	100	100	100	100	100

Figure 4

After that I created a dataframe that displays the number of venue for each neighborhood divides for the types of venue and then I calculated the mean (Figure 5).

```
In [22]: # calculate the mean venues for eahc neighborhood
         manhattan_grouped = manhattan_onehot.groupby('Neighborhood').mean().reset_index()
         manhattan_grouped
```

Out[22]:

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Video Store	Vietnamese Restaurant	Volleyball Court
0	Battery Park City	0.000000	0.00	0.00	0.000000	0.015385	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
1	Carnegie Hill	0.000000	0.00	0.00	0.000000	0.011494	0.000000	0.000000	0.000000	0.011494	...	0.00	0.022989	0.000000
2	Central Harlem	0.000000	0.00	0.00	0.066667	0.044444	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
3	Chelsea	0.000000	0.00	0.00	0.000000	0.030000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
4	Chinatown	0.000000	0.00	0.00	0.000000	0.030000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.020000	0.000000
5	Civic Center	0.000000	0.00	0.00	0.000000	0.040404	0.010101	0.000000	0.000000	0.000000	...	0.00	0.010101	0.000000
6	Clinton	0.000000	0.00	0.00	0.000000	0.030000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
7	East Harlem	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
8	East Village	0.000000	0.00	0.00	0.000000	0.010000	0.000000	0.000000	0.010000	0.010000	...	0.00	0.020000	0.000000
9	Financial District	0.000000	0.00	0.00	0.000000	0.040000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
10	Flatiron	0.000000	0.00	0.00	0.000000	0.010000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
11	Gramercy	0.000000	0.00	0.00	0.000000	0.036585	0.000000	0.012195	0.000000	0.000000	...	0.00	0.000000	0.000000
12	Greenwich Village	0.000000	0.00	0.00	0.000000	0.010000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.020000	0.000000
13	Hamilton Heights	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
14	Inwood	0.000000	0.00	0.00	0.000000	0.034483	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
15	Lenox Hill	0.000000	0.00	0.01	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
16	Lincoln Square	0.000000	0.00	0.00	0.000000	0.030612	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.000000
17	Little Italy	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.010000	0.000000
18	Lower East Side	0.000000	0.00	0.00	0.000000	0.021277	0.000000	0.000000	0.000000	0.021277	...	0.00	0.021277	0.000000

Figure 5

From the above dataset I then extracted a new dataframe with only the Manhattan neighborhoods and the Italian restaurants (Figure 6).

```
In [23]: # create a dataframe that contains only the mean of Italian Restaurant for each neighborhood
         Italian_manhattan = manhattan_grouped[['Neighborhood','Italian Restaurant']]
         Italian_manhattan
```

Out[23]:

	Neighborhood	Italian Restaurant
0	Battery Park City	0.015385
1	Carnegie Hill	0.034483
2	Central Harlem	0.000000
3	Chelsea	0.020000
4	Chinatown	0.010000
5	Civic Center	0.010101
6	Clinton	0.050000
7	East Harlem	0.000000
8	East Village	0.020000
9	Financial District	0.040000
10	Flatiron	0.040000
11	Gramercy	0.036585
12	Greenwich Village	0.100000
13	Hamilton Heights	0.016393
14	Inwood	0.000000
15	Lenox Hill	0.050000
16	Lincoln Square	0.051020
17	Little Italy	0.040000
18	Lower East Side	0.021277
19	Manhattan Valley	0.025000
20	Manhattanville	0.042553

Figure 6

After that I calculated the best k value for clustering using the elbow method (Figure 7).

```
In [26]: # calculate the best value of k
         Sum_of_squared_distances = []
         K = range(1,15)
         for k in K:
             km = KMeans(n_clusters=k)
             km = km.fit(b)
             Sum_of_squared_distances.append(km.inertia_)
```

```
In [27]: # plot the k graph to use the elbow method
         import pandas as pd
         from sklearn.preprocessing import MinMaxScaler
         from sklearn.cluster import KMeans
         import matplotlib.pyplot as plt
         import matplotlib
         plt.plot(K, Sum_of_squared_distances, 'bx-')
         plt.xlabel('k')
         plt.ylabel('Sum_of_squared_distances')
         plt.title('Elbow Method For Optimal k')
         plt.show()
```

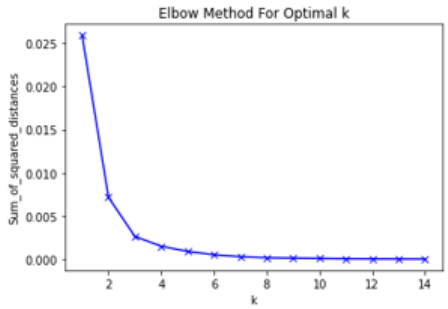

Figure 7

The best value of k is 3 and we apply it to cluster the dataset and to create a new dataframe with also the Cluster Label included

```
In [29]: # 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]
```

```
Out[29]: array([1, 0, 1, 1, 1, 1, 0, 1, 1, 0])
```

```
In [31]: # add clustering labels
         Italian_manhattan.insert(0, 'Cluster Labels', kmeans.labels_)

         manhattan_merged = Italian_manhattan

         # merge manhattan_merged with New york data to add latitude/longitude for each neighborhood
         manhattan_merged = manhattan_merged.join(df.set_index('Neighborhood'), on='Neighborhood')

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

Out[31]:

	Cluster Labels	Neighborhood	Italian Restaurant	Borough	Latitude	Longitude
0	1	Battery Park City	0.015385	Manhattan	40.711932	-74.016869
1	0	Carnegie Hill	0.034483	Manhattan	40.782683	-73.953256
2	1	Central Harlem	0.000000	Manhattan	40.815976	-73.943211
3	1	Chelsea	0.020000	Manhattan	40.744035	-74.003116
3	1	Chelsea	0.020000	Staten Island	40.594726	-74.189560

Figure 8

I then explore each cluster and I found that that cluster 0 contains neighborhood with intermidiate number of Italian restaurant (Figure 9), cluster 1 neighborhood with low number of restaurants (Figure 10) and cluster 2 neighborhood with high number of Italian restaurants (Figure 11)

Explore clusters

```
In [41]: cluster_0 = manhattan_merged.loc[manhattan_merged['Cluster Labels']== 0]
         cluster_0
```

Out[41]:

	Cluster Labels	Neighborhood	Italian Restaurant	Borough	Latitude	Longitude
1	0	Carnegie Hill	0.034483	Manhattan	40.782683	-73.953256
6	0	Clinton	0.050000	Manhattan	40.758334	-73.996408
9	0	Financial District	0.040000	Manhattan	40.707107	-74.010665
10	0	Flatiron	0.040000	Manhattan	40.739673	-73.990947
11	0	Gramercy	0.036585	Manhattan	40.737210	-73.981376
15	0	Lenox Hill	0.050000	Manhattan	40.768113	-73.958860
16	0	Lincoln Square	0.051020	Manhattan	40.773529	-73.985338
17	0	Little Italy	0.040000	Manhattan	40.719324	-73.997305
20	0	Manhattanville	0.042553	Manhattan	40.816934	-73.957385
26	0	Noho	0.040000	Manhattan	40.723259	-73.988434
30	0	Sutton Place	0.040000	Manhattan	40.760280	-73.963556
31	0	Tribeca	0.053333	Manhattan	40.721522	-74.010683
33	0	Turtle Bay	0.050000	Manhattan	40.752042	-73.967708
35	0	Upper West Side	0.048780	Manhattan	40.787658	-73.977059
38	0	Yorkville	0.060000	Manhattan	40.775930	-73.947118

Figure 9

```
In [42]: cluster_1=manhattan_merged.loc[manhattan_merged['Cluster Labels']== 1]
         cluster_1
```

Out[42]:

	Cluster Labels	Neighborhood	Italian Restaurant	Borough	Latitude	Longitude
0	1	Battery Park City	0.015385	Manhattan	40.711932	-74.016869
2	1	Central Harlem	0.000000	Manhattan	40.815976	-73.943211
3	1	Chelsea	0.020000	Manhattan	40.744035	-74.003116
3	1	Chelsea	0.020000	Staten Island	40.594726	-74.189560
4	1	Chinatown	0.010000	Manhattan	40.715618	-73.994279
5	1	Civic Center	0.010101	Manhattan	40.715229	-74.005415
7	1	East Harlem	0.000000	Manhattan	40.792249	-73.944182
8	1	East Village	0.020000	Manhattan	40.727847	-73.982226
13	1	Hamilton Heights	0.016393	Manhattan	40.823604	-73.949688
14	1	Inwood	0.000000	Manhattan	40.867684	-73.921210
18	1	Lower East Side	0.021277	Manhattan	40.717807	-73.980890
19	1	Manhattan Valley	0.025000	Manhattan	40.797307	-73.964286
21	1	Marble Hill	0.000000	Manhattan	40.876551	-73.910660
22	1	Midtown	0.000000	Manhattan	40.754691	-73.981669
23	1	Midtown South	0.010000	Manhattan	40.748510	-73.988713
24	1	Morningside Heights	0.000000	Manhattan	40.808000	-73.963896
25	1	Murray Hill	0.011765	Manhattan	40.748303	-73.978332
25	1	Murray Hill	0.011765	Queens	40.764126	-73.812763
27	1	Roosevelt Island	0.000000	Manhattan	40.762404	-73.949471
29	1	Stuyvesant Town	0.000000	Manhattan	40.731000	-73.974052
32	1	Tudor City	0.013333	Manhattan	40.746917	-73.971219
36	1	Washington Heights	0.011364	Manhattan	40.851903	-73.936900

Figure 10

```
In [43]: cluster_2= manhattan_merged.loc[manhattan_merged['Cluster Labels']== 2]
         cluster_2
```

Out[43]:

	Cluster Labels	Neighborhood	Italian Restaurant	Borough	Latitude	Longitude
12	2	Greenwich Village	0.100000	Manhattan	40.726933	-73.999914
28	2	Soho	0.083333	Manhattan	40.722184	-74.000657
34	2	Upper East Side	0.078652	Manhattan	40.775639	-73.960508
37	2	West Village	0.070000	Manhattan	40.734434	-74.006180

Figure 11

For this reason we select cluster number 1 and we plot it in a bar graph to visualize nighborhoods of this cluster with less Italian restaurant (Figure 12)

```
In [48]: # create bar graph
         graph.plot(kind = 'bar')
         plt.xlabel('Neighborhood')
```

Out[48]: Text(0.5, 0, 'Neighborhood')

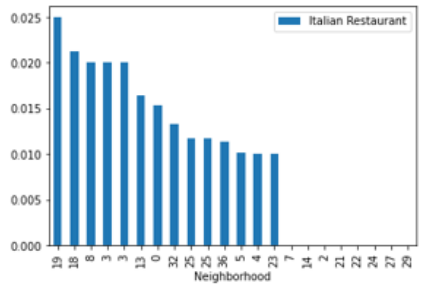

Figure 12

Results

The result of the analysis show that the neighborhoods in Manhattan with less Italian restaurant are: East Harlem, Inwood, Central Harlem, Marble Hill, Midtown, Morningsite Heights, Rooswel Islands, Stuyvesant Town and Rooswelt Island. If we look at the map (Figure 13) we see that from the selected neighborhood the Midtown Neighborhood is the more central one and more easy to reach from every poin of Manahattan. It woul be good to open an Italian Restaurant there

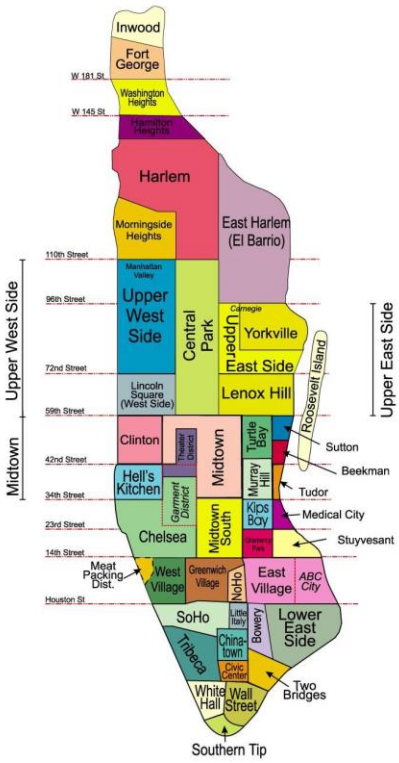

Figure 13

Discussion¶

According to the results, Midtown will provide the best place to open an Italian restaurant in Manhattan. Indeed in this neighborhood there are no Italian Restaurants ath the moment so no competition is expected and moreover it is in the middle of Manhattan, so easy to reach from every point. Howver, even if with this analysis we have a general idea of the best place to open an Italian Restaurant in Manhattan, I think that more analysis are required since also land price or distance from station could a major role.

Conclusion¶

In this project we analyzed the best nighborhood in Manhattan where to open an Italian Restaurant. The analysis was carry out combining several python libriaries,

clustering analsysis and Foursquare data. We found that Midtown coul be a good palce where to opena new Italian Restaurant.