Battle of the Neighborhoods

Report

Yvette Lans

A. Introduction

A.1. Description Problem & Discussion of the Background

My friend John came to me and asked me for help. This summer he will go to Columbia University, New York. He wants to know what is the best neighborhood to live in for him and his girlfriend.

John loves Foursquare. But it will take too much time for him to check each and every neighborhood in Manhattan. John and his girlfriend prefers to live in a neighborhood with a:

- Bakery
- Bar
- Breakfast Spot
- Café
- Cocktail Bar
- Coffee Shop
- Gym / Fitness Center
- Sandwich Place
- Shopping Mall
- Sushi Restaurant
- Pizza Place
- Yoga Studio

He asked me to answer the following question:
"Given the requirements, in what neighborhood of Manhattan should I start searching for a property?"

A.2. Data Description

To answer John's problem I've used the following datasets:

- New York Location Data (https://cocl.us/new_york_dataset)
  I've used this dataset to explore and store the venues for every neighborhood.
- Foursquare API
  I've used the Foursquare API to explore the venue details of the Manhattan neighborhoods.
- Geopy library
  I've used Geopy to retrieve the coordinates for every neighborhood.

B. Methodology

In the first step, I load all the data of New York. From this dataset I use:

- Borough, to extract Manhattan
- Neighborhood, since we're searching for the best neighborhood
- Latitude and longitude, so that we can create a map

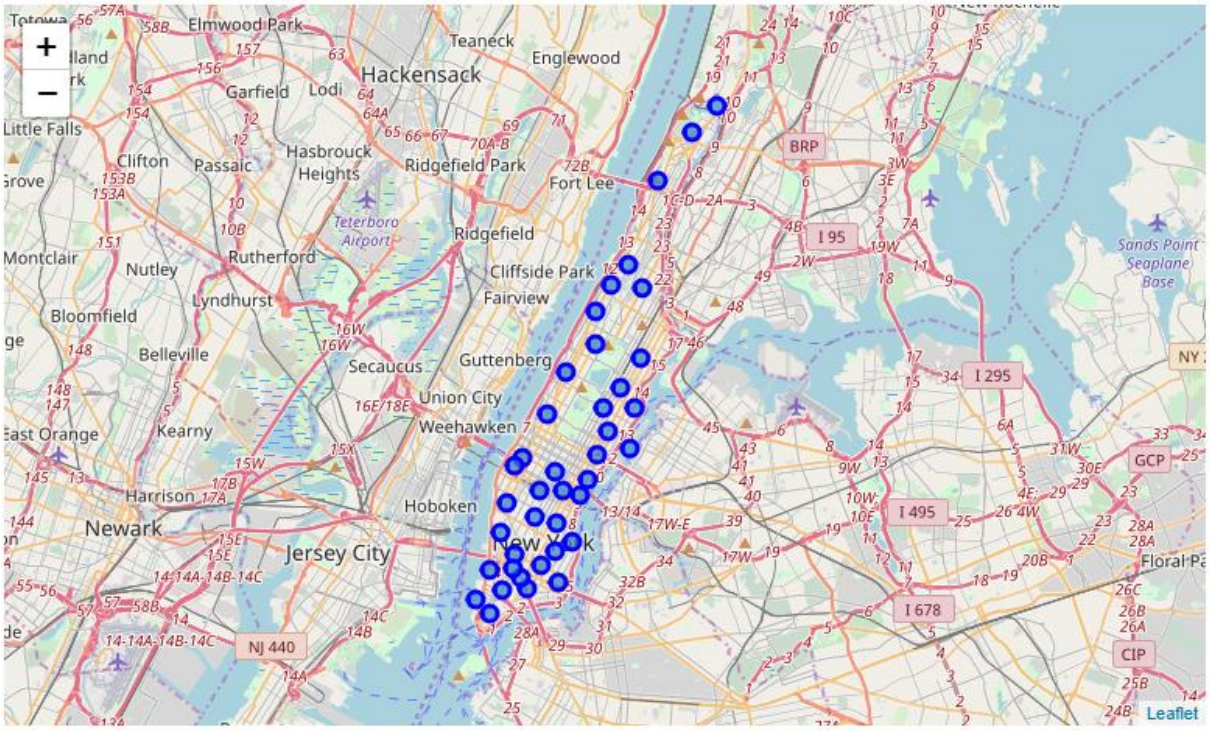
The result is the following dataframe (head only)

	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

After that, I've sliced the dataframe so that it only contains the data of the Manhattan neighborhoods.

	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

In the second step, I've used the python Folium library to create a map of Manhattan that shows the neighborhoods.



In the third step I used the Foursquare API to explore the venues of the neighborhoods of Manhattan. First I made a call for just one neighborhood and retrieved the top 100 venues in a radius of 500. I've stored the results in a dataframe.

	name	categories	lat	lng
0	Arturo's	Pizza Place	40.874412	-73.910271
1	Bikram Yoga	Yoga Studio	40.876844	-73.906204
2	Tibbett Diner	Diner	40.880404	-73.908937
3	Starbucks	Coffee Shop	40.877531	-73.905582
4	Land & Sea Restaurant	Seafood Restaurant	40.877885	-73.905873

Next I repeated the process for all neighborhoods and combined all information in a new dataframe.

	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	Land & Sea Restaurant	40.877885	-73.905873	Seafood Restaurant

Then I used one hot encoding to see which venues are present in the neighborhoods.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Arge Rest
0	Marble Hill	0	0	0	0	0	0	0	0	0	0
1	Marble Hill	0	0	0	0	0	0	0	0	0	0
2	Marble Hill	0	0	0	0	0	0	0	0	0	0
3	Marble Hill	0	0	0	0	0	0	0	0	0	0
4	Marble Hill	0	0	0	0	0	0	0	0	0	0

I created a list of requirements that matches John's preferences.

- Bakery
- Bar
- Breakfast Spot
- Café
- Cocktail Bar
- Coffee Shop
- Gym / Fitness Center
- Sandwich Place
- Shopping Mall
- Sushi Restaurant
- Pizza Place
- Yoga Studio

Then I've updated the dataframe so that it shows the required venues per neighborhood.

	Bakery	Bar	Breakfast Spot	Café	Cocktail Bar	Coffee Shop	Gym / Fitness Center	Sandwich Place	Shopping Mall	Sushi Restaurant	Pizza Place	Yoga Studio
Neighborhood												
Battery Park City	1	0	0	0	0	8	1	2	2	1	2	0
Carnegie Hill	2	3	1	4	1	5	2	0	0	1	6	3
Central Harlem	0	1	0	1	0	0	2	0	0	0	1	0
Chelsea	4	1	1	1	1	7	1	1	0	1	1	0
Chinatown	3	2	0	0	4	2	0	2	0	0	1	1

I used the K-means algorithm to cluster the neighborhoods into 5 clusters and I've calculated the centre of each cluster, sorted the data and stored it in a dataframe.

	Bakery	Bar	Breakfast Spot	Café	Cocktail Bar	Coffee Shop	Gym / Fitness Center	Sandwich Place	Shopping Mall	Sushi Restaurant	Pizza Place
G1	1.555556	3.000000	0.333333	1.666667	2.000000	4.222222	1.444444	1.000000	0.000000e+00	2.444444	2.777778
G4	3.250000	1.000000	0.000000	1.500000	1.000000	2.000000	5.500000	1.000000	0.000000e+00	1.250000	0.500000
G2	1.166667	1.166667	0.166667	1.333333	0.666667	6.666667	1.500000	1.666667	3.333333e-01	0.333333	1.333333
G3	2.625000	0.687500	0.187500	2.562500	1.562500	2.437500	0.625000	1.187500	1.387779e-17	1.000000	1.937500
G5	0.000000	1.200000	0.000000	0.400000	0.200000	1.000000	0.400000	0.600000	2.000000e-01	0.200000	0.600000

In the final step I analysed the results to answer Johns question. I mapped the cluster results with the neighborhoods to find the best neighborhoods.

	Neighborhood	Group
0	Battery Park City	2
1	Carnegie Hill	1
2	Central Harlem	5
3	Chelsea	2
4	Chinatown	3
5	Civic Center	4
6	Clinton	2
7	East Harlem	3
8	East Village	1
9	Financial District	2
10	Flatiron	4
11	Gramercy	1
12	Greenwich Village	3
13	Hamilton Heights	3
14	Hudson Yards	2
15	Inwood	3
16	Lenox Hill	1
17	Lincoln Square	4
18	Little Italy	3
19	Lower East Side	3
20	Manhattan Valley	3
21	Manhattanville	5
22	Marble Hill	5
23	Midtown	3
24	Midtown South	3
25	Morningside Heights	2
26	Murray Hill	1
27	Noho	1
28	Roosevelt Island	5
29	Soho	3
30	Stuyvesant Town	5
31	Sutton Place	4
32	Tribeca	3
33	Tudor City	3
34	Turtle Bay	1
35	Upper East Side	3
36	Upper West Side	1
37	Washington Heights	3
38	West Village	3
39	Yorkville	1

C. Results

The result is the following dataframe that shows the best neighborhoods for John and his girlfriend.

	Neighborhood	Group
2	Central Harlem	5
21	Manhattanville	5
22	Marble Hill	5
28	Roosevelt Island	5
30	Stuyvesant Town	5

Since the Columbia University is closeby Manhattanville, I advised John to start his search for a property in that neighborhood.

D. Discussion

To decide the best neighborhood, I created 5 clusters. More clusters can be used for more details when needed. The neighborhoods of Group 5 suited the preferences of John best. But off course John and his girlfriend will also have requirements for the actual house, which are outside of scope for this assignment. Let´s hope he can find a nice play to stay during this college time!

F. Conclusion

The result of this assignment is that I've learned how to gather data, clean data and visualize data. There are many ways to find a solution for his problem, but given the assignment I mainly used the foursquare data as a great practise for real world problems.

G. References

New York data set: https://cocl.us/new_york_dataset
Foursquare API: https://developer.foursquare.com/
Geopy library: https://pypi.org/project/geopy/