IBM Data Science Specialization Capstone
Café shop location in NYC

By: Amged Elsheikh

1 Introduction

1.1 Motivation

New York City (NYC) is one of the biggest cities in the world with an area equal to 783.8 km² which contains 218 neighborhoods in 5 boroughs and a population of 8.419 million. Most office workers and tourists might not have time in the morning to make their own breakfast, instead, they have it in cafés. So, we want to know where to open a new café/coffee shop to target more customers using Airbnb location data and neighborhoods venues details.

1.2 Targeted audience

Investors who are interested in opening a new café/coffee shop in NYC.

1.3 Data

For this task, we will use NYC Airbnb open dataset which can be found in Kaggle (link). Neighborhoods information can be found using NYC Geojson data (link), using this file it's possible to draw each neighborhood boundary on folium map as a polygon, and getting the latitude and longitude by finding the centroid of each polygon. After getting each neighborhood geo-location I will use Foursquare API to analyze each neighborhood and find out where it is good to open a new café/coffee shop with the help of the NYC Airbnb data.

1.4 Why using two different datasets

1. Airbnb data to reduce the number of possible neighborhoods since we will assume that those who use Airbnb are usually do not make their own breakfast at home nor lunch.
2. Foursquare data to see which neighborhood members attend cafés/coffee shops since we are targeting the community of the neighborhood also.

2 Methodology

2.1 Airbnb Data Cleaning

The original NYC Airbnb dataset has 38843 rows and 16 columns. In our project, we are not interested in all columns, so we start first by choosing the columns that can help us process our data and convert the "last_review" column to show the years only.

```
There were 38843  Airbnb location in NYC in 2019
```

	Borough	Neighborhood	Latitude	Longitude	number_of_reviews	last_review	availability_365
0	Brooklyn	Kensington	40.64749	-73.97237	9	2018.0	365
1	Manhattan	Midtown	40.75362	-73.98377	45	2019.0	355
2	Brooklyn	Clinton Hill	40.68514	-73.95976	270	2019.0	194
3	Manhattan	East Harlem	40.79851	-73.94399	9	2018.0	0
4	Manhattan	Murray Hill	40.74767	-73.97500	74	2019.0	129

Now, we can clean our data by setting some criteria:

1. Remove all locations with less than or equal 40 reviews, the more reviews you get can show how good the place is (of course some reviews might be negative, but there is no way to tell from the current dataset).
2. Remove all locations that did not get any review in 2019 since we are looking for active places as much as possible.
3. Remove Neighborhoods that have less than 30 Airbnb locations.

The data will look like this:

	Borough	Neighborhood	Latitude	Longitude	number_of_reviews	last_review	availability_365
count	5536	5536	5536.000000	5536.000000	5536.000000	5536.0	5536.000000
unique	3	47	NaN	NaN	NaN	NaN	NaN
top	Brooklyn	Bedford-Stuyvesant	NaN	NaN	NaN	NaN	NaN
freq	2561	673	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	40.727621	-73.948223	105.179733	2019.0	196.262645
std	NaN	NaN	0.049338	0.039379	66.638485	0.0	102.990576
min	NaN	NaN	40.625800	-74.021960	40.000000	2019.0	30.000000
25%	NaN	NaN	40.685585	-73.978662	58.000000	2019.0	95.000000
50%	NaN	NaN	40.721125	-73.951140	85.000000	2019.0	207.500000
75%	NaN	NaN	40.762920	-73.932240	132.000000	2019.0	285.000000
max	NaN	NaN	40.858670	-73.764930	629.000000	2019.0	365.000000

We will only use about 14% of the initial dataset. We successfully reduced the number of brought from 5 to 3 and the number of neighborhoods from 218 to 47. Of course, it is still possible to reduce the size of the dataset, but we will process this data for now. Later we will drop the last three columns since they are not useful for us anymore.

2.2 Geographic data

After we clean the dataset, we need to visualize it in a way that is easy to understand. Since we are dealing with geographic data, we will use ***folium*** library to plot a choropleth map that shows Airbnb locations in NYC.

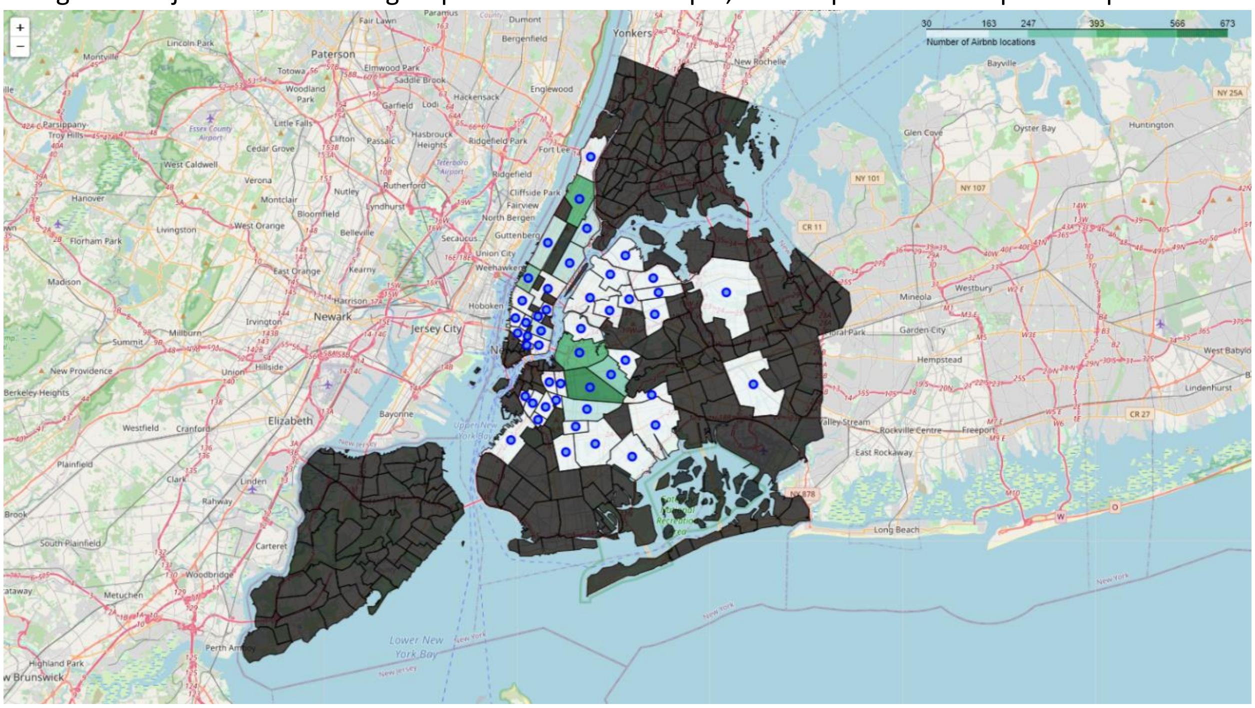
We will use NYC Geojson data to plot the boundary of each neighborhood in NYC. To do this we will take extra steps to make sure everything works fine:

1. Create a dataset from the Geojson data that contains the name of the neighborhood, it is polygon boundaries, centroid (latitude and longitude of the midpoint), and its borough name, we will call this dataset **poly_df**.
2. Using ***shapely*** library, we will loop in our Airbnb data frame, re-check the name of the neighborhood of each Airbnb location by converting the location into a shapely point object and confirm for which polygon it belongs.
3. We will change the name of the neighborhoods to the one we found in step 2. This might seem useless since we already get the name of the neighborhoods in our dataset, but this is a very good

practice to make sure the names of neighborhoods in Airbnb data matches the names in the Geojson data.

	Borough	Neighborhood	Latitude	Longitude	Polygons
0	Queens	Astoria	40.765187	-73.919746	[(-73.90160305064738, 40.76777029715587), (-73...
1	Brooklyn	Bedford-Stuyvesant	40.687068	-73.938201	[(-73.9411488595606, 40.70028115334914), (-73...
2	Brooklyn	Bushwick	40.695749	-73.918637	[(-73.90582150629088, 40.694113724380834), (-7...
3	Brooklyn	Canarsie	40.638840	-73.899707	[(-73.89034734693779, 40.64903360577805), (-73...
4	Brooklyn	Carroll Gardens	40.680540	-73.997064	[(-73.991332, 40.685448), (-73.98913989974679,...

4. Since we are interested to see the number of Airbnb locations in each neighborhood, we will group Airbnb data frame by the neighborhood names.
5. Using the Geojson data and the grouped dataset from step 4, we can plot the choropleth map.



Colors indicate the amount of Airbnb locations in each neighborhood as can be seen from the color bar at the top right of the map, black neighborhoods are the ones without any Airbnb location that met the criteria we set. Blue markers are interactive objects that will show the name of the borough and neighborhood. As can be seen from the map, Queens borough does not look like a good place to open, but we need to confirm first by grouping our data by borough name.

	Borough	Count
0	Brooklyn	2561
1	Manhattan	2268
2	Queens	707

As we expected, we can remove Queens borough Airbnb locations from our dataset. With this, we ended up with 4829 Airbnb locations in two boroughs only.

2.3 Neighborhoods' venues details:

After we narrowed down the number of neighborhoods to 36 neighborhoods, we shall explore those neighborhoods using Foursquare API. For this API, we need to find the exact location of each neighborhood. We have two options:

1. Using the centroid of the polygons as the latitude and longitude.
2. Using ***Geopy*** library to get the co-ordinate of the neighborhood using its address.

For the first option, we already got the center of each polygon in the **poly_df** data frame, this option might not sound like the optimal one, because the center of the neighborhood is not always the center of the polygon, but it can do the job. The second option requires high competition, in our case 36 neighborhood is not too much, but let us use the first option.

To use Foursquare API, you need to make a developer account in their developer webpage developer.foursquare.com and get your API credentials. I do not intend to make my credentials available for public use, I am sorry.

we set the searching radius to one kilometer from the center point of each neighborhood and set the limit number of venues to explore to 120. We can then save the results to a data frame named **ny_venues**, it has 304 unique shop categories, "venue id" is a unique key assigned by Foursquare for each venue, and in our case, it is very important as we will see later.

	Neighborhood	Venue	Venue Category	Venue id
0	Bedford-Stuyvesant	Saraghina	Pizza Place	4a593de0f964a52015b91fe3
1	Bedford-Stuyvesant	Bar Lunatico	Bar	5490f3d2498e4e2727ce17ac
2	Bedford-Stuyvesant	Do The Right Thing Crossing	Historic Site	4dbf2ef04b2221ec2d553767
3	Bedford-Stuyvesant	Saraghina Bakery	Bakery	53ff6b91498e916b5804dc9b
4	Bedford-Stuyvesant	Bar Camillo	Italian Restaurant	5e4567fa2eafa100085e9ec3

We are more interested in knowing what is popular in those neighborhoods, so we create a one-hot encoder using the "Venue Category" column, then sum the categories for each neighborhood. Lastly, we create a data frame that shows the top 10 categories in each neighborhood. To make data visualizing more proper, we will add a column called "Count" to show the number of Airbnb locations in each neighborhood. we will sort the data and show what are the most popular categories in the top 10 neighborhoods.

	Neighborhood	Count	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	Bedford-Stuyvesant	673	Bar	Coffee Shop	Café	Pizza Place	Chinese Restaurant	Deli / Bodega	Caribbean Restaurant	Playground	Discount Store	Juice Bar
1	Harlem	440	Southern / Soul Food Restaurant	Coffee Shop	Cocktail Bar	Jazz Club	Bar	Seafood Restaurant	Mexican Restaurant	French Restaurant	Café	Yoga Studio
2	Williamsburg	433	Bar	Coffee Shop	Pizza Place	Bakery	Italian Restaurant	Mexican Restaurant	Cocktail Bar	Café	Chinese Restaurant	Japanese Restaurant
3	Bushwick	301	Bar	Coffee Shop	Mexican Restaurant	Pizza Place	Bakery	Deli / Bodega	Gym	Italian Restaurant	Latin American Restaurant	Mediterranean Restaurant
4	Hell's Kitchen	284	Theater	Coffee Shop	Bar	Gym / Fitness Center	Gym	Wine Shop	Italian Restaurant	Thai Restaurant	Wine Bar	Gift Shop
5	Crown Heights	223	Caribbean Restaurant	Pizza Place	Southern / Soul Food Restaurant	Café	Bakery	Bar	Fried Chicken Joint	Bagel Shop	Discount Store	Juice Bar
6	East Village	210	Wine Bar	Japanese Restaurant	Bar	Italian Restaurant	Juice Bar	Dessert Shop	Coffee Shop	Ice Cream Shop	Korean Restaurant	Pizza Place
7	East Harlem	204	Mexican Restaurant	Bakery	Park	Pizza Place	Italian Restaurant	Café	Latin American Restaurant	Thai Restaurant	Cocktail Bar	Gym / Fitness Center
8	Upper East Side	178	Italian Restaurant	Coffee Shop	Sushi Restaurant	Ice Cream Shop	Gym / Fitness Center	Bar	Bakery	Dessert Shop	Thai Restaurant	Café
9	Upper West Side	169	Italian Restaurant	Bakery	Coffee Shop	Café	Mediterranean Restaurant	American Restaurant	Wine Bar	Gym	Bar	Park

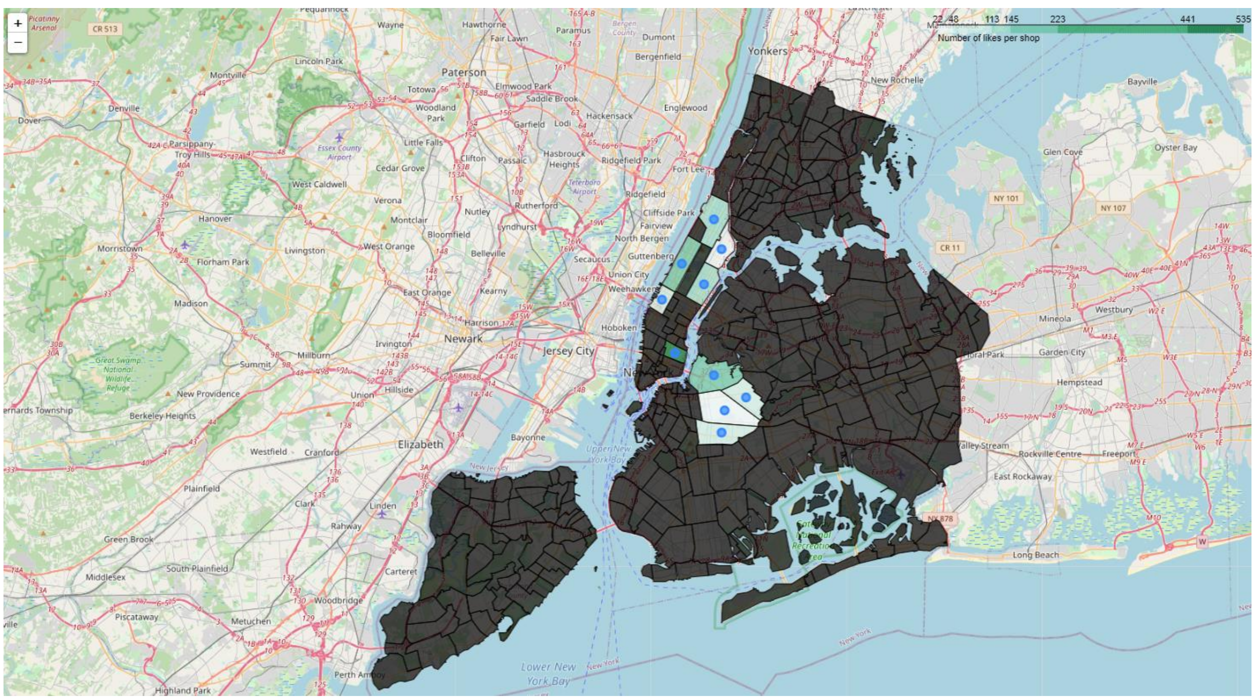
By looking at the table, we can easily tell that deciding café/coffee shop location using Airbnb data was a good idea since both types are common in the top 10 neighborhoods, in fact, we can also use these results for opening new bar also.

Now we know we can target one of those neighborhoods to open our new shop, but we need to know more about our competitors and the people since we want our shop to be popular. We can use Foursquare API again to get insights. I will explore café/coffee shop that appeared in the top 10 neighborhoods using its unique "venue id" to get the number of tips, likes, and price category for each shop.

	Neighborhood	Venue	Venue Category	likes	price_category	tips_count
0	Bedford-Stuyvesant	Brooklyn Kettle	Coffee Shop	24.0	Cheap	9.0
1	Bedford-Stuyvesant	Little Roy Coffee Co.	Coffee Shop	52.0	Cheap	6.0
2	Bedford-Stuyvesant	Zaca Cafe	Café	10.0	Cheap	6.0
3	Bedford-Stuyvesant	BoHaus Coffee and Flowers	Coffee Shop	27.0	Cheap	2.0
4	Bedford-Stuyvesant	Brown Butter	Café	30.0	Cheap	8.0
...	...	...	...	...	...	...
74	Williamsburg	Porto Rico Importing Co.	Coffee Shop	57.0	Cheap	25.0
75	Williamsburg	Pecoraro Latteria	Café	7.0	Cheap	2.0
76	Williamsburg	Think Coffee	Coffee Shop	46.0	Cheap	6.0
77	Williamsburg	Variety Coffee Roasters	Coffee Shop	433.0	Cheap	138.0
78	Williamsburg	The Flat's BK Speed Coffee	Coffee Shop	83.0	Cheap	16.0

We got 79 shops in 10 neighborhoods, but we are more interested in the neighborhoods, so we can group the number of tips and likes by neighborhood so that we can understand how well the shops in these areas do and this is also indicating if the people are active users (liking/tipping) in each neighborhood or no. To be popular you need to have a high number of reviews and ratings.

	Neighborhood	likes	tips_count	Number of coffee shops	likes per shop	Tips per shop
0	Bedford-Stuyvesant	334.0	83.0	14	23.857143	5.928571
1	Harlem	820.0	295.0	7	117.142857	42.142857
2	Williamsburg	1329.0	384.0	9	147.666667	42.666667
3	Bushwick	685.0	153.0	11	62.272727	13.909091
4	Hell's Kitchen	543.0	177.0	5	108.600000	35.400000
5	Crown Heights	380.0	93.0	7	54.285714	13.285714
6	East Village	2673.0	924.0	5	534.600000	184.800000
7	East Harlem	109.0	52.0	5	21.800000	10.400000
8	Upper East Side	1081.0	334.0	8	135.125000	41.750000
9	Upper West Side	1510.0	409.0	8	188.750000	51.125000



This map shows our top 10 targeted neighborhoods, markers are interactive, they show the name of the neighborhood, number of likes and tips.

If we consider Airbnb data only, Bedford will be our top priority, but the shops there do not get enough likes and tips. We can reduce the table rows by choosing only shops that get more than 100 like per shop and more than 20 tips per shop.

	Neighborhood	likes	tips_count	Number of coffee shops	likes per shop	Tips per shop
0	Harlem	820.0	295.0	7	117.142857	42.142857
1	Williamsburg	1329.0	384.0	9	147.666667	42.666667
2	Hell's Kitchen	543.0	177.0	5	108.600000	35.400000
3	East Village	2673.0	924.0	5	534.600000	184.800000
4	Upper East Side	1081.0	334.0	8	135.125000	41.750000
5	Upper West Side	1510.0	409.0	8	188.750000	51.125000

3 Results

1. We confirmed that cafes/coffee shops in the neighborhoods that have high Airbnb locations are common.
2. We were able to evaluate the performance of cafes/coffee shops in the top 10 neighborhoods.
3. We were able to reduce the number of neighborhoods where we can make the new shop to 6 neighborhoods only, instead of 218.

4 Discussion

1. While Bedford-Stuyvesant has the highest number of Airbnb locations, it's clear that cafes in that neighborhood don't get many likes and tips, this is not good because we want our cafe to be popular. On the other hand, maybe the shops there are popular among their community, so more investigations are needed.
2. Williamsburg is close to Bedford and Bushwick, so considering it is a good option. The drawback is that it already has 9 cafes.
3. Hell's Kitchen is good, but by reviewing the map Harlem is a better option.
4. East Village cafes have the highest ratings with only 5 cafes, the competition there is high so it's not highly recommended.
5. Upper West Side is close to Upper East Side, and it has more active users, but the drawback is the number of existing cafes.

5 Conclusions

We were able to narrow the possible locations for opening new cafe using Airbnb data and Foursquare API only. Making final decision will require field investigation to the top 6 nominated neighborhoods and Bedford neighborhood. It is possible to visualize each shop and their locations, but we will close our investigation here for now. The last decision for the location will depend on the budget, the type of cafe and the business plan.