# Advice Engine for building a Business in a Chosen Area

Mahdi Arfaoui

15, Nov 2020

## Introduction

One of the most pillars in a building a physical B2C Business, is to choose a convenient place, where the project will have all the assets it needs ( Client target, strategical location, provision, supplying, distributing … ).

So study location part is a must phase to do in the BPM of the Project.

## Problem

It is not an easy task to choose wisely a great location for a certain business, the factors are multiple and diversified, gathering data, processing it , understanding it, and extract knowledge from it, is a big issue and the key to accomplish this purpose.

## Data Processing

### 1- Data Acquisition :

to make things simpler, I worked on the Manhattan city data, there is a data set already prepared (!wget -q -O 'newyork_data.json' https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json)

### 2 – Data Cleaning

As you may notice from the link, it is a New York city data.

```
{'type': 'FeatureCollection',
 'totalFeatures': 306,
 'features': [{'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]}},
  {'type': 'Feature',
```

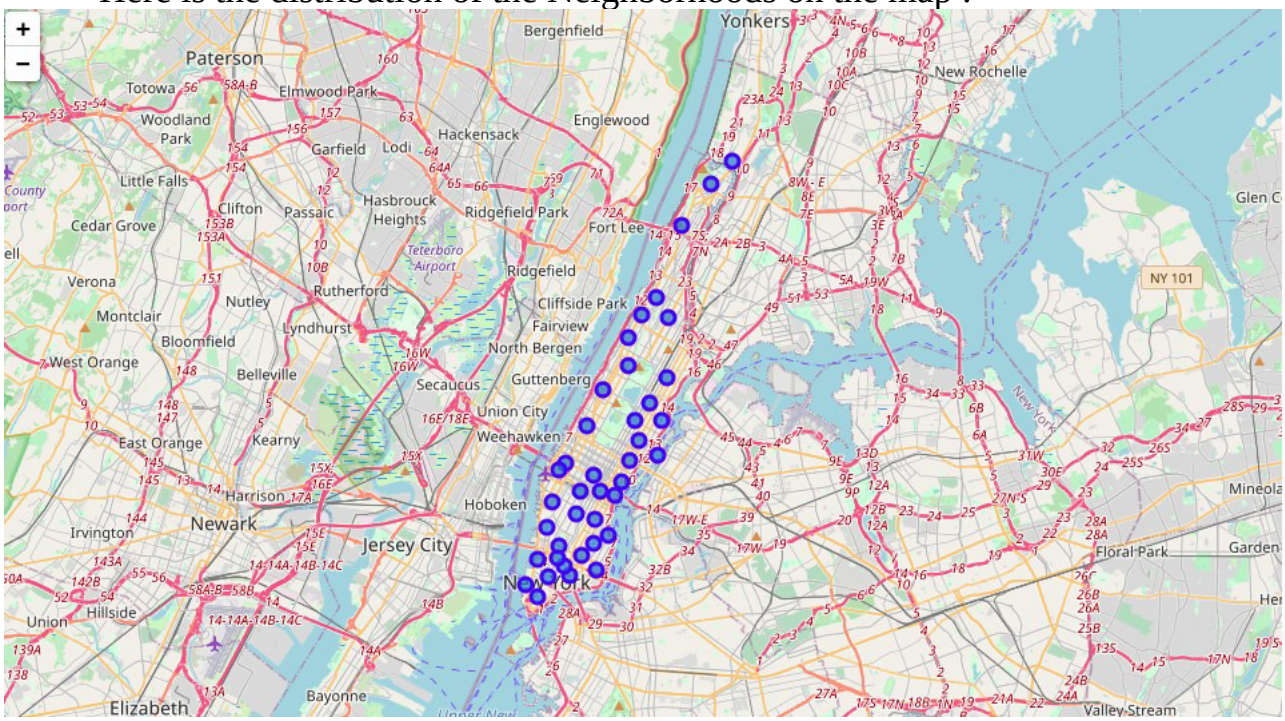

By simple look on the json data, we could see that this data is composed of numerous Borough with each one have multiple Neighborhoods. With simple instruction we could see the Data into a structured Data Frame.

| | 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 |

We extract then the Manhattan City data and placed into a Data Frame.

| | 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 |

Here is the distribution of the Neighborhoods on the map :




## 3 – Data Exploratory

Our main purpose is to determine the best places where a client want to build a project, there are multiple factors to choose such a place ( business nature, client target, ways of supplying and distributing, concurrent project …) .

Since I don't have a data set where all these factors are presented, I'm going to only make this study based on the other concurrent business.

For example, if a client want to build a Gym, we'll see if the region he want to build his project on, has multiple gyms or a little number, and make our decision upon the result.

For that purpose, we are going to load the visited venues for the specific area from **Foursquare** api with their location coordinates and their category.

| | 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 | Dunkin' | 40.877136 | -73.906666 | Donut Shop |
| 4 | Marble Hill | 40.876551 | -73.91066 | Starbucks | 40.877531 | -73.905582 | Coffee Shop |

we are going to assemble categories, for example if a category have the word "Restaurant" then it will become that name ("Fast Food Restaurant" will become "Restaurant")

To make things easier to process, I'll work only on 6 categories : "Gym" , "Restaurant" , "Shop" , "Hotel" , "Bar" and "Café".

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **61** | Chinatown | 40.715618 | -73.994279 | Sofar HQ | 40.713523 | -73.996289 | Music Venue |
| **62** | Chinatown | 40.715618 | -73.994279 | Yi Ji Shi Mo Noodle Corp | 40.718254 | -73.995930 | Restaurant |
| **63** | Chinatown | 40.715618 | -73.994279 | Bacaro | 40.714468 | -73.991589 | Restaurant |
| **64** | Chinatown | 40.715618 | -73.994279 | Hong Kong Supermarket 香港超級市場 | 40.717596 | -73.996173 | Supermarket |
| **65** | Chinatown | 40.715618 | -73.994279 | oo35mm.com | 40.716605 | -73.997890 | Shop |
| **66** | Chinatown | 40.715618 | -73.994279 | Simple | 40.718145 | -73.991988 | Restaurant |
| **67** | Chinatown | 40.715618 | -73.994279 | Jing Fong Restaurant 金豐大酒樓 | 40.715881 | -73.997209 | Restaurant |
| **68** | Chinatown | 40.715618 | -73.994279 | 99 Favor Taste 99號餐廳 | 40.717560 | -73.992580 | Restaurant |
| **69** | Chinatown | 40.715618 | -73.994279 | l'estudio | 40.715720 | -73.990332 | Restaurant |
| **70** | Chinatown | 40.715618 | -73.994279 | Nam Son Vietnamese Restaurant | 40.718215 | -73.994345 | Restaurant |
| **71** | Chinatown | 40.715618 | -73.994279 | Little Canal | 40.714317 | -73.990361 | Shop |
| **72** | Chinatown | 40.715618 | -73.994279 | Joe's Shanghai 鹿鳴春 | 40.715661 | -73.996693 | Restaurant |

We could see the change easily.

Now we will eliminate the venues that have not the categories we work on.

Out[34]:

| | index | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|---|
| **0** | 9 | Marble Hill | 40.876551 | -73.910660 | Land & Sea Restaurant | 40.877885 | -73.905873 | Restaurant |
| **1** | 24 | Chinatown | 40.715618 | -73.994279 | Spicy Village | 40.717010 | -73.993530 | Restaurant |
| **2** | 27 | Chinatown | 40.715618 | -73.994279 | Kiki's | 40.714476 | -73.992036 | Restaurant |
| **3** | 29 | Chinatown | 40.715618 | -73.994279 | Wah Fung Number 1 Fast Food 華豐快餐店 | 40.717278 | -73.994177 | Restaurant |
| **4** | 32 | Chinatown | 40.715618 | -73.994279 | Da Yu Hot Pot 大渝火锅 | 40.716735 | -73.995752 | Restaurant |
| **5** | 35 | Chinatown | 40.715618 | -73.994279 | Xi'an Famous Foods | 40.715232 | -73.997263 | Restaurant |
| **6** | 39 | Chinatown | 40.715618 | -73.994279 | Forgtmenot | 40.714459 | -73.991546 | Restaurant |
| **7** | 41 | Chinatown | 40.715618 | -73.994279 | Dimes | 40.714830 | -73.991719 | Restaurant |
| **8** | 44 | Chinatown | 40.715618 | -73.994279 | Ling Kee Malaysian Beef Jerky | 40.714713 | -73.991538 | Restaurant |
| **9** | 47 | Chinatown | 40.715618 | -73.994279 | Cervo's | 40.714763 | -73.991455 | Restaurant |

```
Entrée [78]: manhattan_reduced['Venue Category'].unique()
Out[78]: array(['Restaurant', 'Bar', 'Hotel', 'Café', 'Gym', 'Shop'], dtype=object)
```

We can see clearly that we only have the categories that we want in the DF.

Now we going to calculate how often a venue is visited, by giving each categories a dummy number then calculating the mean by each Neighborhood.

Our result is :

| | Neighborhood | Bar | Café | Gym | Hotel | Restaurant | Shop |
|---|---|---|---|---|---|---|---|
| 0 | Battery Park City | 0.041667 | 0.000000 | 0.166667 | 0.208333 | 0.166667 | 0.416667 |
| 1 | Carnegie Hill | 0.111111 | 0.074074 | 0.111111 | 0.018519 | 0.388889 | 0.296296 |
| 2 | Central Harlem | 0.206897 | 0.034483 | 0.068966 | 0.000000 | 0.517241 | 0.172414 |
| 3 | Chelsea | 0.096154 | 0.038462 | 0.038462 | 0.038462 | 0.403846 | 0.384615 |
| 4 | Chinatown | 0.134328 | 0.000000 | 0.000000 | 0.014925 | 0.597015 | 0.253731 |
| 5 | Civic Center | 0.105263 | 0.017544 | 0.140351 | 0.105263 | 0.403509 | 0.228070 |
| 6 | Clinton | 0.163636 | 0.018182 | 0.163636 | 0.072727 | 0.400000 | 0.181818 |
| 7 | East Harlem | 0.100000 | 0.050000 | 0.050000 | 0.000000 | 0.750000 | 0.050000 |
| 8 | East Village | 0.250000 | 0.013158 | 0.000000 | 0.000000 | 0.500000 | 0.236842 |
| 9 | Financial District | 0.175439 | 0.052632 | 0.087719 | 0.052632 | 0.350877 | 0.280702 |
| 10 | Flatiron | 0.049180 | 0.032787 | 0.098361 | 0.000000 | 0.524590 | 0.295082 |
| 11 | Gramercy | 0.214286 | 0.035714 | 0.000000 | 0.035714 | 0.392857 | 0.321429 |
| 12 | Greenwich Village | 0.041667 | 0.055556 | 0.041667 | 0.013889 | 0.652778 | 0.194444 |
| 13 | Hamilton Heights | 0.189189 | 0.108108 | 0.000000 | 0.000000 | 0.513514 | 0.189189 |
| 14 | Hudson Yards | 0.062500 | 0.062500 | 0.187500 | 0.156250 | 0.406250 | 0.125000 |
| 15 | Inwood | 0.129032 | 0.096774 | 0.000000 | 0.000000 | 0.580645 | 0.193548 |
| 16 | Lenox Hill | 0.131148 | 0.049180 | 0.098361 | 0.000000 | 0.459016 | 0.262295 |
| 17 | Lincoln Square | 0.052632 | 0.105263 | 0.157895 | 0.026316 | 0.368421 | 0.289474 |

A simple processing with this df, we could extract top 3 visited venues for each Neighborhood

```
----Battery Park City----
   venue  freq
0   Shop  0.42
1  Hotel  0.21
2    Gym  0.17


----Carnegie Hill----
        venue  freq
0  Restaurant  0.39
1        Shop  0.30
2         Bar  0.11


----Central Harlem----
        venue  freq
0  Restaurant  0.52
1         Bar  0.21
2        Shop  0.17
```

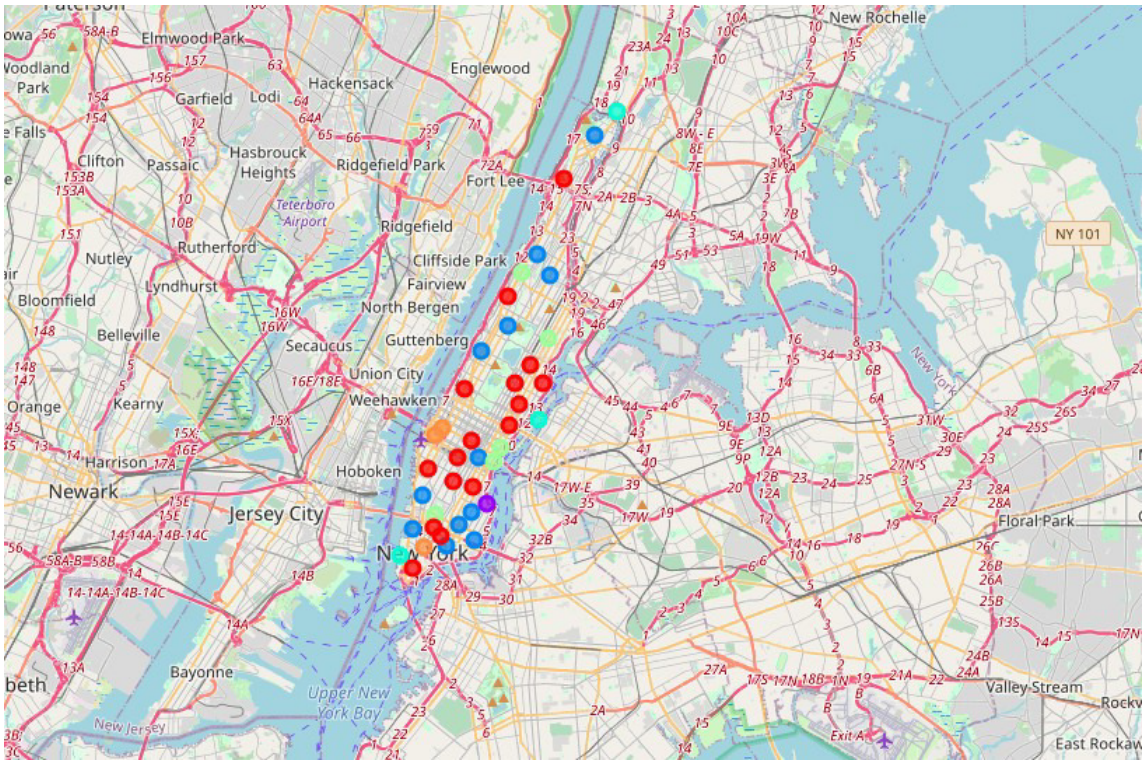
We'll make now a structured df to visualize this information on all Neighborhoods, and we'll use later to extract best location for a giving business.

| | 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 |
|---|---|---|---|---|---|---|---|
| 0 | Battery Park City | Shop | Hotel | Restaurant | Gym | Bar | Café |
| 1 | Carnegie Hill | Restaurant | Shop | Gym | Bar | Café | Hotel |
| 2 | Central Harlem | Restaurant | Bar | Shop | Gym | Café | Hotel |
| 3 | Chelsea | Restaurant | Shop | Bar | Hotel | Gym | Café |
| 4 | Chinatown | Restaurant | Shop | Bar | Hotel | Gym | Café |

Let's add the locations :

| | Borough | Neighborhood | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Manhattan | Marble Hill | 40.876551 | -73.910660 | 3 | Shop | Gym | Restaurant | Hotel | Café | Bar |
| 1 | Manhattan | Chinatown | 40.715618 | -73.994279 | 2 | Restaurant | Shop | Bar | Hotel | Gym | Café |
| 2 | Manhattan | Washington Heights | 40.851903 | -73.936900 | 0 | Restaurant | Shop | Café | Gym | Bar | Hotel |
| 3 | Manhattan | Inwood | 40.867684 | -73.921210 | 2 | Restaurant | Shop | Bar | Café | Hotel | Gym |
| 4 | Manhattan | Hamilton Heights | 40.823604 | -73.949688 | 2 | Restaurant | Shop | Bar | Café | Hotel | Gym |

You are must wondering why 'Cluster Label' there, well it just for more clear picture how the data is distributed, here's the plot in the next figure map :



## Extracting Target Information :

We'll extract the top 3 venues with theirs locations into our final df (3 is our threshold, we could choose any other number according to the client desire – to be included in the **future directions** section)

| | Neighborhood | Latitude | Longitude | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
|---|---|---|---|---|---|---|
| 0 | Marble Hill | 40.876551 | -73.910660 | Shop | Gym | Restaurant |
| 1 | Chinatown | 40.715618 | -73.994279 | Restaurant | Shop | Bar |
| 2 | Washington Heights | 40.851903 | -73.936900 | Restaurant | Shop | Café |
| 3 | Inwood | 40.867684 | -73.921210 | Restaurant | Shop | Bar |
| 4 | Hamilton Heights | 40.823604 | -73.949688 | Restaurant | Shop | Bar |
| 5 | Manhattanville | 40.816934 | -73.957385 | Restaurant | Shop | Bar |
| 6 | Central Harlem | 40.815976 | -73.943211 | Restaurant | Bar | Shop |
| 7 | East Harlem | 40.792249 | -73.944182 | Restaurant | Bar | Shop |
| 8 | Upper East Side | 40.775639 | -73.960508 | Restaurant | Shop | Hotel |
| 9 | Yorkville | 40.775930 | -73.947118 | Restaurant | Shop | Gym |
| 10 | Lenox Hill | 40.768113 | -73.958860 | Restaurant | Shop | Bar |
| 11 | Roosevelt Island | 40.762160 | -73.949168 | Shop | Restaurant | Gym |
| 12 | Upper West Side | 40.787658 | -73.977059 | Restaurant | Shop | Bar |

After creating a function to eliminate all the places where our client choice included in their top 3 visited venues we got this the example of "Bar"

```
#Let's polt the places where the licent want to build a Bar
input = "Bar"

target_df = adviceMe(input)
target_df.head()
```

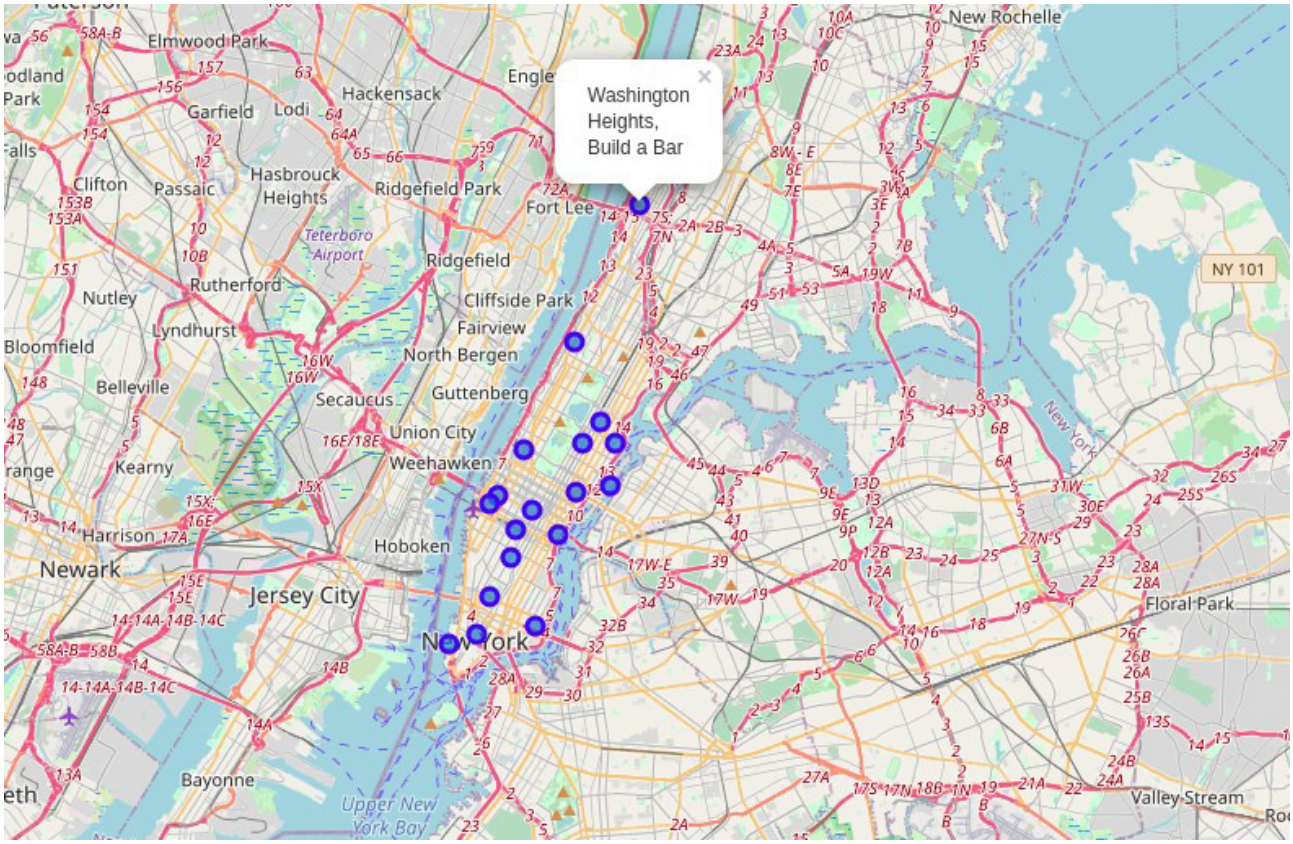
| | Neighborhood | Latitude | Longitude | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
|---|---|---|---|---|---|---|
| 0 | Marble Hill | 40.876551 | -73.910660 | Shop | Gym | Restaurant |
| 2 | Washington Heights | 40.851903 | -73.936900 | Restaurant | Shop | Café |
| 8 | Upper East Side | 40.775639 | -73.960508 | Restaurant | Shop | Hotel |
| 9 | Yorkville | 40.775930 | -73.947118 | Restaurant | Shop | Gym |
| 11 | Roosevelt Island | 40.762160 | -73.949168 | Shop | Restaurant | Gym |

```
target_df.shape
```

```
(19, 6)
```

We could easily see that we got 19 places (from total of 40 Neighborhoods) where a “Bar” is not often visited.

And here’s the map :




## Future Directions :

There is a great for improvements in this solution :

1- We could generalize the client input
2- We could generalize the area to be included in the input
3- We could make the search factors more diversified (that only need a larger and more diversified data-set(s)