Coursera Platform

IBM Data Science Professional Certificate

Applied Data Science Capstone

(Find similar business approach)

Eng. Mohamed Mostafa

engineer.m.mostafa@gmail.com

Jan 2019

1. Introduction

Many business seekers or startups always looking for new ideas and promising business opportunities. When new idea come alive all think how to get benefit of these idea or where I could apply the same idea so I could success same to the existing success story.

It's now clear after many cases/experiments that not always the new idea creator will only success but many of followers who decide to walk the same road also success if they study the market well. Sometimes followers achieve gain and profit more than the idea creator just because they avoid first attempt mistakes, choose the proper market area and provide better customer service/ experience.

So in this report I'll try to use what I learnt during these professional certificate track to help a businessman find appropriate neighborhoods to open the second branch of his bakery shop after the success of the first branch in Downtown, Brooklyn.

1.1 Business Problem

Allocating similar neighborhoods (market areas) for upcoming business. A businessman looking to expand his business by opening a second branch of his bakery shop after the success of the first branch located in Downtown, Brooklyn.

So the purpose is to use machine learning especially clustering algorism to cluster New York neighborhoods to define neighborhoods similar to Downtown, Brooklyn and within these cluster choose the promising neighborhood which has shortage of these service as a business opportunity for his second branch.

1.2 Target Audience of this project

This report is particularly useful to business developers and investors looking to open, expand or invest in new markets and looking for guidance in allocating the proper location for these business or service. Who know and understand the benefits of machine learning in helping target their investments in many disciplines where available data and statistics can push his progress.

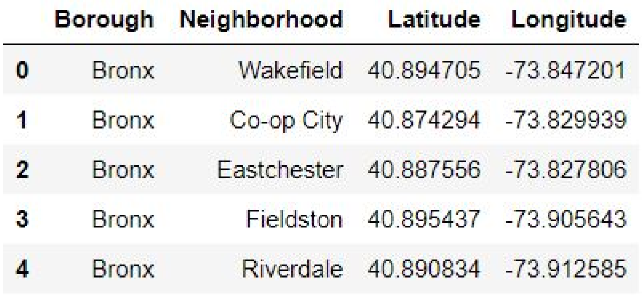
2. Data acquisition and cleaning

2.1 Data sources

Used neighborhoods data which provided at previous course module to build data frame of New York neighborhoods and its Coordinates, also I will use New York Geo JSON data from (Carto website & NYC Open data website) for visualization purposes. Then will use Foursquare API to retrieve venues data along all NY neighborhoods which will be used in neighborhoods clustering, especially bakery category which will be used to allocate poor neighborhoods with the same category (bakery).

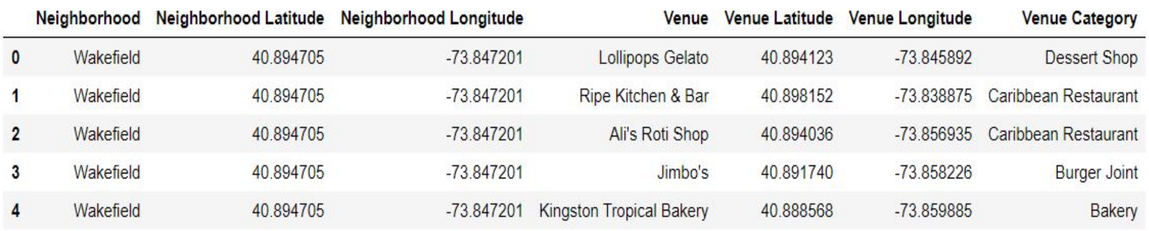
2.2 Data cleaning

Several data cleaning and formatting done using Python Pandas library to build the neighborhood data frame containing borough, neighborhood and Coordinates data as follows:



	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

Using Foursquare API to build venues data frame contains venue location and category type as follow:

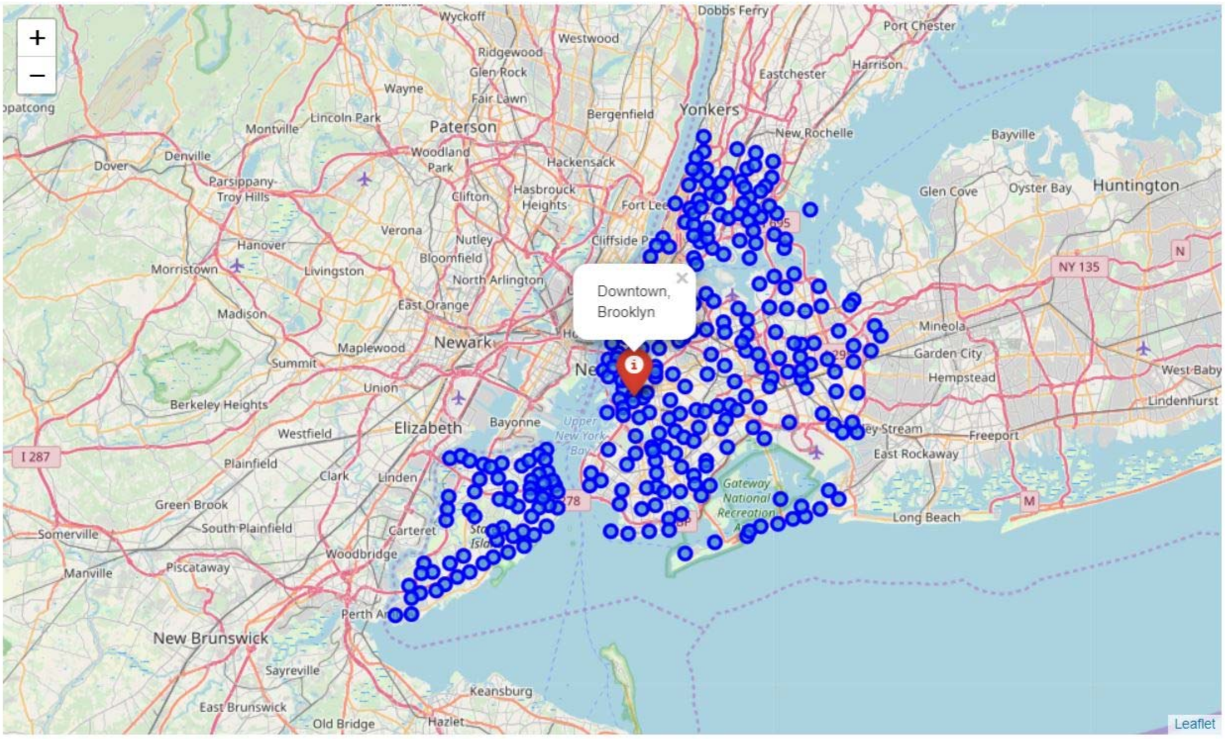


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Jimbo's	40.891740	-73.858226	Burger Joint
4	Wakefield	40.894705	-73.847201	Kingston Tropical Bakery	40.888568	-73.859885	Bakery

3. Exploratory Data Analysis

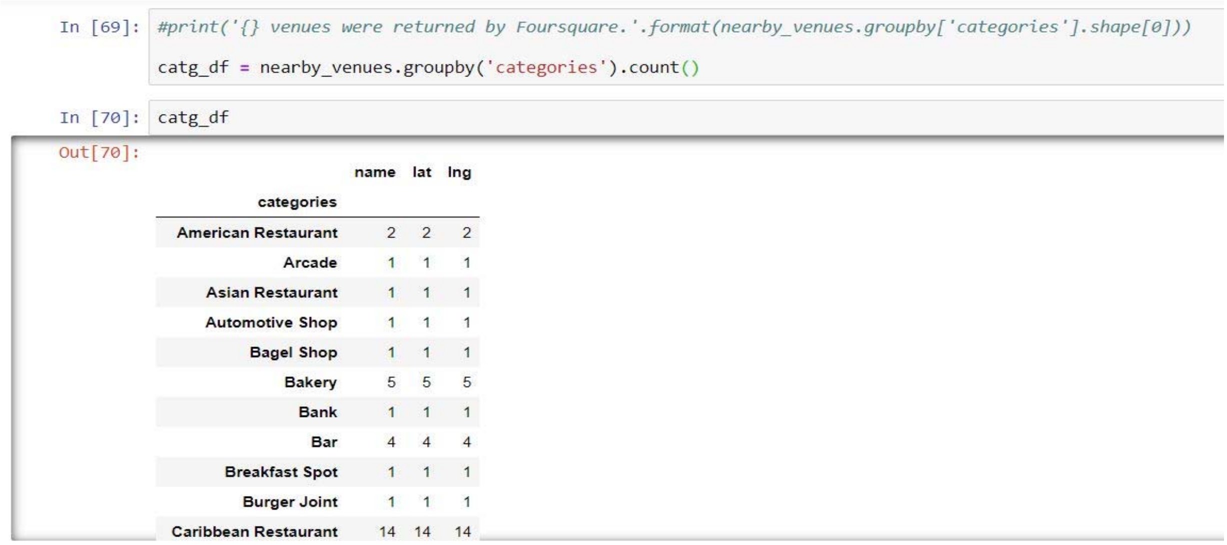
3.1 Visualize NYC Neighborhoods

Using formatted data, we can visualize NYC neighborhoods to better understand our data and get familiar with the first branch location (Downtown, Brooklyn) using Folium library.



3.2 Understand foursquare venues data

Iterate through JSON response, exclude important data in to pandas data frame and do basic analysis to get familiar with the data and its content:

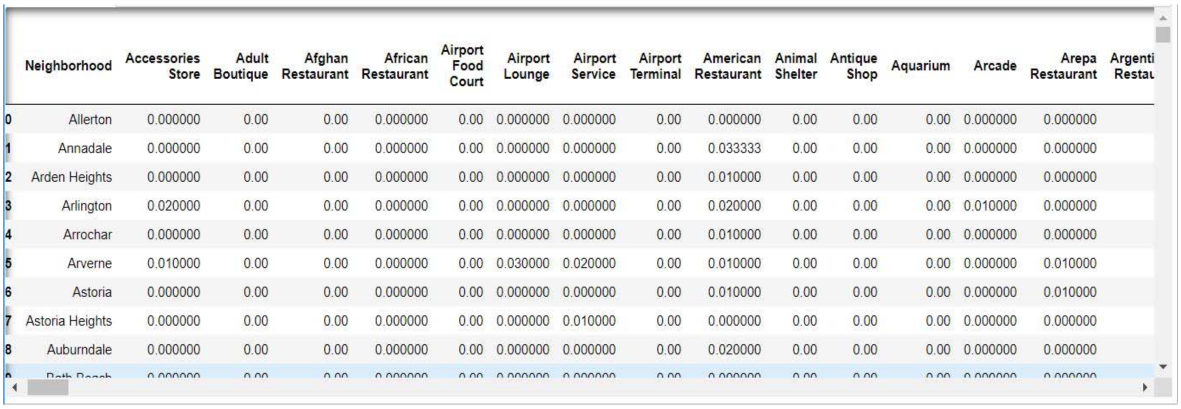


```
In [69]: #print('{} venues were returned by Foursquare.'.format(nearby_venues.groupby['categories'].shape[0]))

         catg_df = nearby_venues.groupby('categories').count()

In [70]: catg_df

Out[70]:
```

	name	lat	lng
categories			
American Restaurant	2	2	2
Arcade	1	1	1
Asian Restaurant	1	1	1
Automotive Shop	1	1	1
Bagel Shop	1	1	1
Bakery	5	5	5
Bank	1	1	1
Bar	4	4	4
Breakfast Spot	1	1	1
Burger Joint	1	1	1
Caribbean Restaurant	14	14	14

3.3 Analyze NYC Neighborhoods

Using gathered and formatted data, we can analysis NYC neighborhoods to define hot locations with many returned venues and popular venue categories in each neighborhood. So using hot encoding and sorting functions we could form coming tabular format

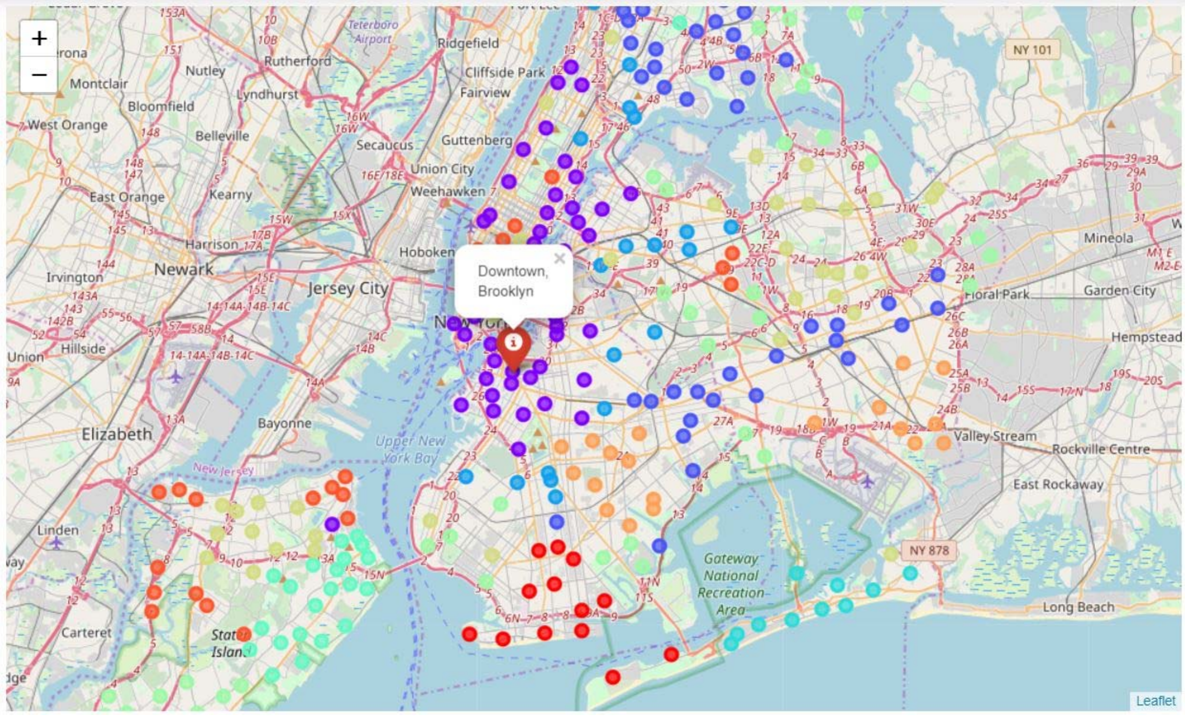


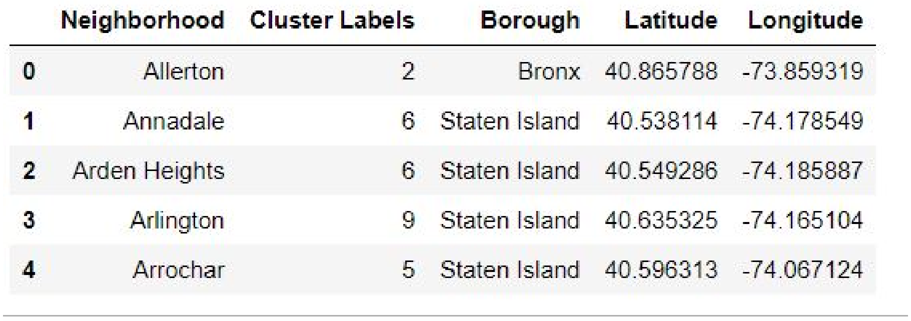
	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Arcade	Arepa Restaurant	Argenti Restau
0	Allerton	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.00	0.00	0.00	0.000000	0.000000	
1	Annadale	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.033333	0.00	0.00	0.00	0.000000	0.000000	
2	Arden Heights	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.010000	0.00	0.00	0.00	0.000000	0.000000	
3	Arlington	0.020000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.020000	0.00	0.00	0.00	0.010000	0.000000	
4	Arrochar	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.010000	0.00	0.00	0.00	0.000000	0.000000	
5	Arverne	0.010000	0.00	0.00	0.000000	0.00	0.030000	0.020000	0.00	0.010000	0.00	0.00	0.00	0.000000	0.010000	
6	Astoria	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.010000	0.00	0.00	0.00	0.000000	0.010000	
7	Astoria Heights	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.010000	0.00	0.000000	0.00	0.00	0.00	0.000000	0.000000	
8	Auburndale	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.020000	0.00	0.00	0.00	0.000000	0.000000	
9	Bath Beach	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.00	0.00	0.00	0.000000	0.000000	

4. Methodology

Use neighborhood and venues data to help investor locating the appropriate location for his bakery shop second branch after the success of the first branch in Downtown, Brooklyn.

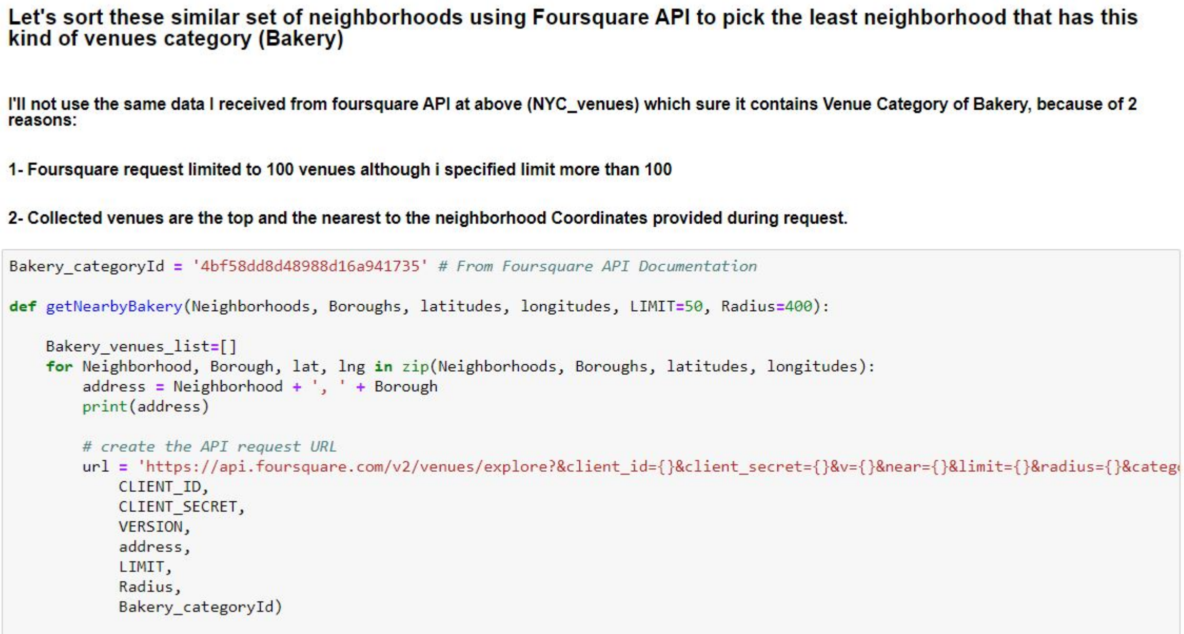
We will use machine learning to make clustering of New York neighborhoods according to foursquare venues data, then pick the cluster contain Downtown, Brooklyn which indicate the similarity of these neighborhoods.





	Neighborhood	Cluster Labels	Borough	Latitude	Longitude
0	Allerton	2	Bronx	40.865788	-73.859319
1	Annadale	6	Staten Island	40.538114	-74.178549
2	Arden Heights	6	Staten Island	40.549286	-74.185887
3	Arlington	9	Staten Island	40.635325	-74.165104
4	Arrochar	5	Staten Island	40.596313	-74.067124

Within the cluster neighborhoods contain Downtown, Brooklyn will again retrieve foursquare data for Bakery category only, sort data ascending just to output the neighborhoods lack of the required service as a promising business opportunity for the second branch.



Let's sort these similar set of neighborhoods using Foursquare API to pick the least neighborhood that has this kind of venues category (Bakery)

I'll not use the same data I received from foursquare API at above (NYC_venues) which sure it contains Venue Category of Bakery, because of 2 reasons:

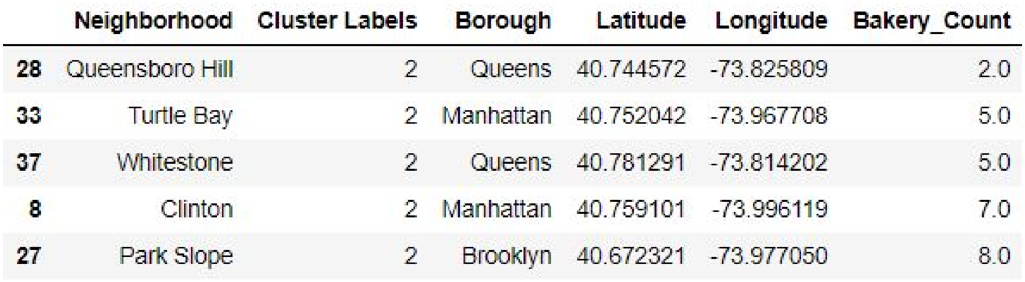
1- Foursquare request limited to 100 venues although i specified limit more than 100

2- Collected venues are the top and the nearest to the neighborhood Coordinates provided during request.

```
Bakery_categoryId = '4bf58dd8d48988d16a941735' # From Foursquare API Documentation

def getNearbyBakery(Neighborhoods, Boroughs, latitudes, longitudes, LIMIT=50, Radius=400):

    Bakery_venues_list=[]
    for Neighborhood, Borough, lat, lng in zip(Neighborhoods, Boroughs, latitudes, longitudes):
        address = Neighborhood + ', ' + Borough
        print(address)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&near={}&limit={}&radius={}&categ
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            address,
            LIMIT,
            Radius,
            Bakery_categoryId)
```

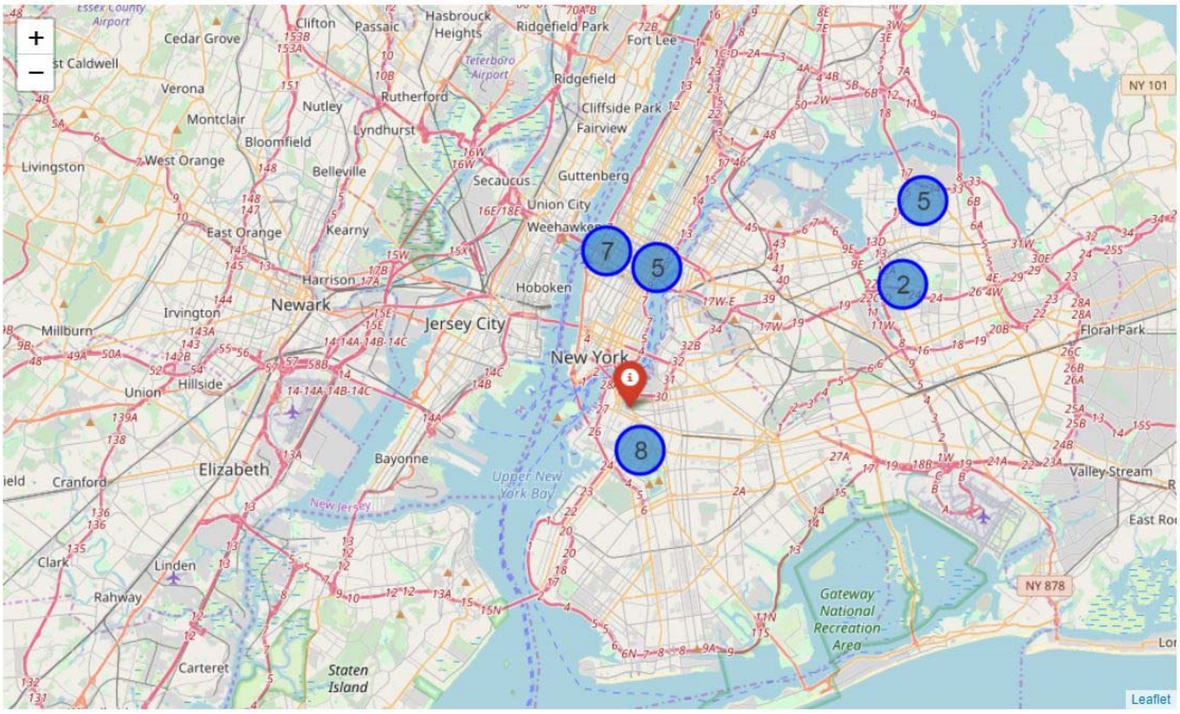
5. Results

Now we formed our output in a shape of data frame of New York neighborhoods similar to Downtown, Brooklyn and have the least number of bakery shops.



	Neighborhood	Cluster Labels	Borough	Latitude	Longitude	Bakery_Count
28	Queensboro Hill	2	Queens	40.744572	-73.825809	2.0
33	Turtle Bay	2	Manhattan	40.752042	-73.967708	5.0
37	Whitestone	2	Queens	40.781291	-73.814202	5.0
8	Clinton	2	Manhattan	40.759101	-73.996119	7.0
27	Park Slope	2	Brooklyn	40.672321	-73.977050	8.0

As per observations we noted based on the results we can visualize our output in a sake of simplicity with following map.



6. Conclusions

Now the investor has a scientific guide for his next step which increase of his potential success and decrease the effort of locating the next branch to specific areas of interest.

7. Area of Future improvements

This report will be more accurate if more data for clustering are available such neighborhood population, level of service which all make enhancements and increase the clustering criteria.