Final Capstone's Report: Where to open my bar?

1. Introduction

1.1. Background

New York City

The City of New York, usually called either New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2017 population of 8,622,698 distributed over a land area of about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States. Located at the southern tip of the state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass and one of the world's most populous megacities, with an estimated 20,320,876 people in its 2017 Metropolitan Statistical Area and 23,876,155 residents in its Combined Statistical Area. A global power city, New York City has been described as the cultural, financial, and media capital of the world, and exerts a significant impact upon commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

1.2. Problem

Let us suppose that someone wants to open a new bar in a city (take New York for our example of city). The first reflex that comes is to find where there is a high and low density of Bars to look for a place where he can benefit from his new Bar. For example, the majority will choose to open a new bar where there is no high density of Bars. In this report, we are going to use you will use the Foursquare API to explore neighborhoods in Toronto.

We will group the neighborhoods into clusters and use them to choose where can someone open a new Bar (for example) to get the maximum benefits. Of course, it will depend of how many other bars exists in every neighborhoods, which means that if a bar is isolated from the others, it will give more benefits from people who are visiting this neighborhood.

2. Data acquisition and cleaning

2.1. Data sources

Using data in our previous lab, we are import it from https://cocl.us/new_york_dataset. The data we are going to explore is composed

5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

Our first data set is composed of four columns:

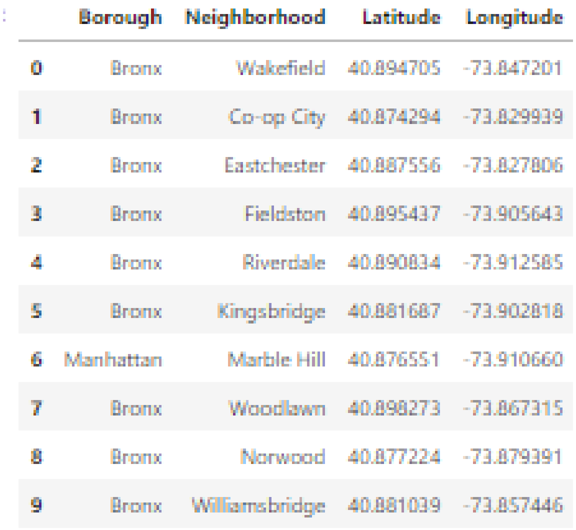
- Borough
- Neighborhood
- Latitude
- Longitude

Then we'll use Foursquare api to get venues in a radius of 500 m for every neighborhood and do some statistical methods to classify neighborhoods and make a choice of where can this person open a new Bar to fit the maximum benefits from it

New York City map will be used for this project. The map is on the link: https://geo.nyu.edu/catalog/nyu_2451_34572 Publisher: New York (City). Department of City Planning Place(s)

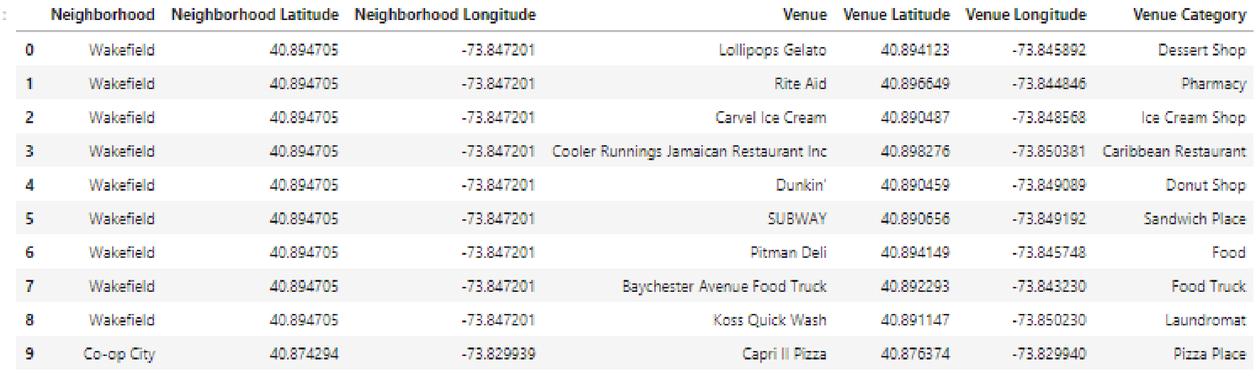
2.1. Data structuring and cleaning

After getting our data, we can will structure it as a dataframe as showed below:



	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
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

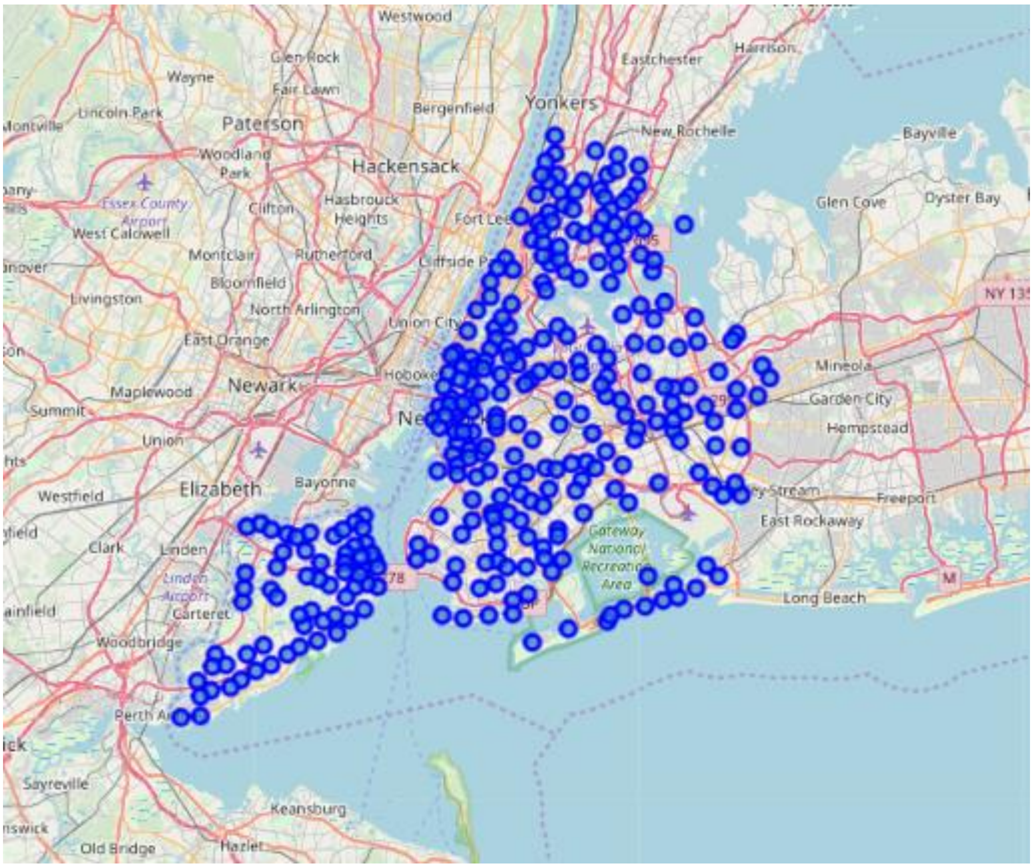
After getting the data of every Borough and Neighborhood on New York with Their Latitude and Longitude, we will use Foursquare API to get venues in every Neighborhod. The result of calling the Api is shown here :



	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	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898276	-73.850381	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop
5	Wakefield	40.894705	-73.847201	SUBWAY	40.890656	-73.849192	Sandwich Place
6	Wakefield	40.894705	-73.847201	Pitman Deli	40.894149	-73.845748	Food
7	Wakefield	40.894705	-73.847201	Baychester Avenue Food Truck	40.892293	-73.843230	Food Truck
8	Wakefield	40.894705	-73.847201	Koss Quick Wash	40.891147	-73.850230	Laundromat
9	Co-op City	40.874294	-73.829939	Capri II Pizza	40.876374	-73.829940	Pizza Place

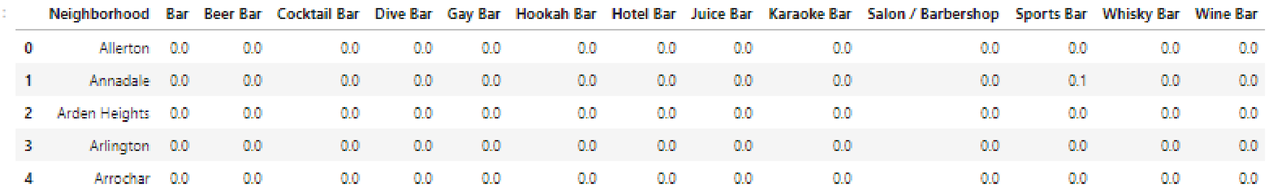
3. Exploratory Data Analysis

Let us first create a map where we can render the neighborhoods in New York:



The map shows all the neighborhoods of New York with all venues. We will filter our data in the next steps.

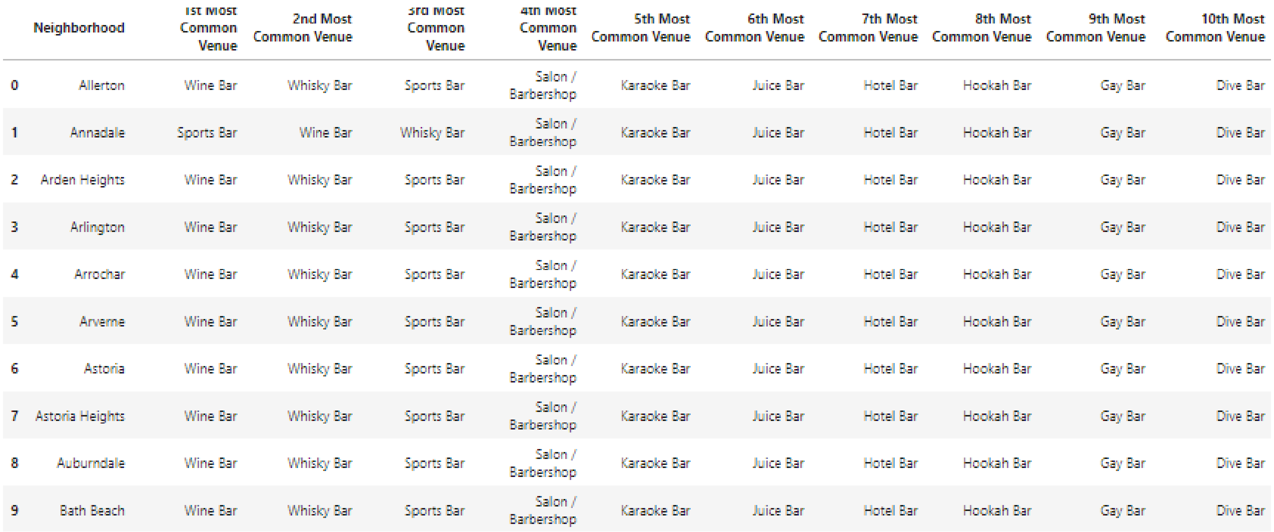
We will group rows by neighborhood and by taking the mean of the frequency of occurrence of each category where we will find the name 'Bar' on it. To select columns with 'Bar', we will use Regex for regular expressions. The result is shown here:



	Neighborhood	Bar	Beer Bar	Cocktail Bar	Dive Bar	Gay Bar	Hookah Bar	Hotel Bar	Juice Bar	Karaoke Bar	Salon / Barbershop	Sports Bar	Whisky Bar	Wine Bar
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

As you can see, the dataframe show every neighborhood in New York with the frequency of every type of Bars

Now let us create the new dataframe and display the top 10 venues that contain Bars. The result is shown here:



	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allerton	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
1	Annadale	Sports Bar	Wine Bar	Whisky Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
2	Arden Heights	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
3	Arlington	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
4	Arrochar	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
5	Arverne	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
6	Astoria	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
7	Astoria Heights	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
8	Auburndale	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
9	Bath Beach	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar

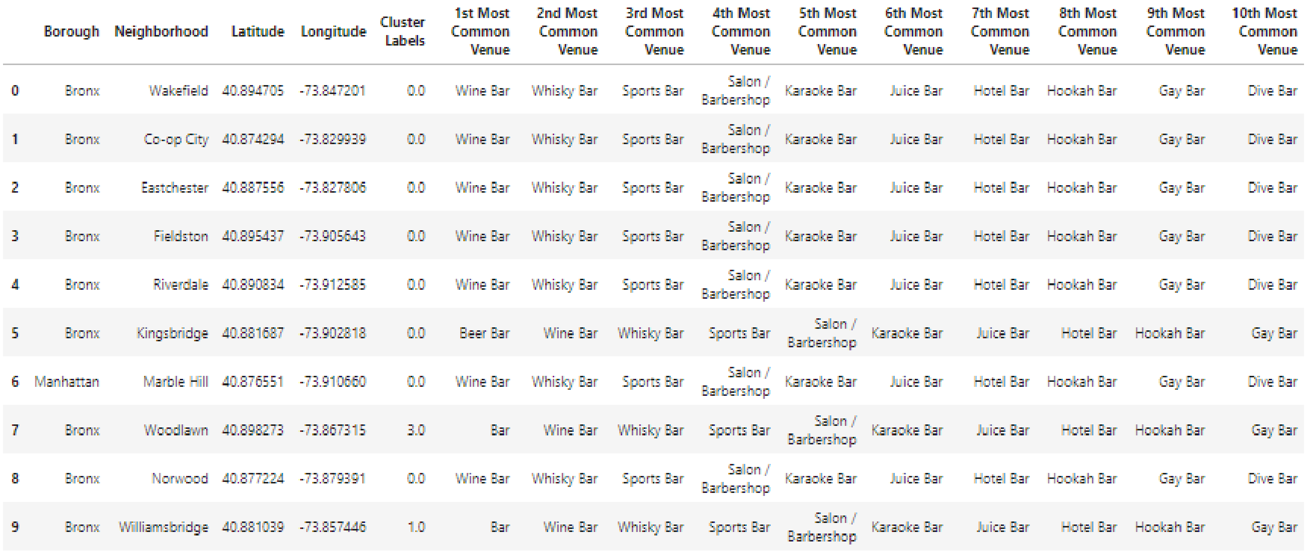
Our Data is now ready to fit k-means algorithm that we will use in the next steps

4. Modeling

Let us use k-means to cluster the neighborhood into 5 clusters. It means that the value of k for this exemple is 5.

k-means is one of the simplest and popular unsupervised machine learning algorithms. In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

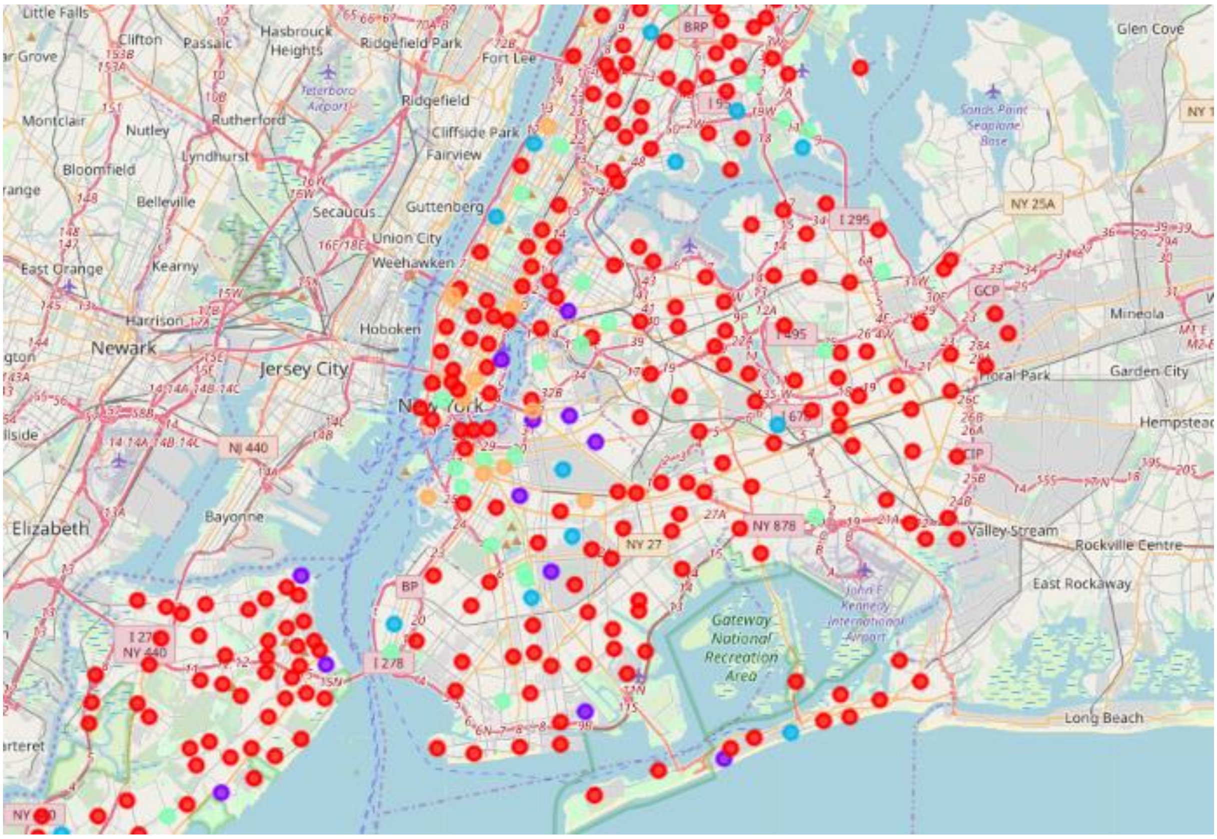
Here is the result of clustering with 5 labels after fitting our k-means algorithm:



	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bronx	Wakefield	40.894705	-73.847201	0.0	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
1	Bronx	Co-op City	40.874294	-73.829939	0.0	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
2	Bronx	Eastchester	40.887556	-73.827806	0.0	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
3	Bronx	Fieldston	40.895437	-73.905643	0.0	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
4	Bronx	Riverdale	40.890834	-73.912585	0.0	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
5	Bronx	Kingsbridge	40.881687	-73.902818	0.0	Beer Bar	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar
6	Manhattan	Marble Hill	40.876551	-73.910660	0.0	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
7	Bronx	Woodlawn	40.898273	-73.867315	3.0	Bar	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar
8	Bronx	Norwood	40.877224	-73.879391	0.0	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar	Dive Bar
9	Bronx	Williamsbridge	40.881039	-73.857446	1.0	Bar	Wine Bar	Whisky Bar	Sports Bar	Salon / Barbershop	Karaoke Bar	Juice Bar	Hotel Bar	Hookah Bar	Gay Bar

5. Result

Let us show our clusters on a map using Folium to get a visual version of the clusters:



As you can see on the map, we have five clusters of every type of bars. These clusters can be labeled depending on the density of bars in the cluster and help the person who want to open a bar to find the best neighborhood for that