Using location data to improve the overall health of the residents of New York City, NY

1. Introduction/Business Problem

In this section, we describe the business problem in detail. In particular, we clearly discuss the motivation for addressing this problem, the group of stakeholders, and our intended audience and why they should care about our problem.

1.1 Background

Our health and that of our planet depends greatly on how cities are planned. Ideally, we would want a city where people can live well and be healthy. We know that New York City (NYC) has a high population density, is the financial capital of the US, and attracts more and more people to it every year. Therefore, it is important that there should be ample focus on making the city a healthy place to live in for its residents by creating more hospitals/health care facilities/medical centers, and physical spaces like urban parks, gyms, and fitness centers among others to simply increase ease of accessibility to these health-improvement centers. Needless to say, these have numerous health benefits in adults, such as a reduction of stress, a longer life or better general and mental health.

1.2 Problem

The broad goal of this project is to use location data to improve the overall health of the residents of New York City. Therefore, it is important to analyze the neighborhoods of NYC to see which areas need more (1) basic health facilities like medical centers, hospitals, emergency rooms, and so on, and (2) other facilities to improve physical and mental health. In order to do so, we will have to explore various types of venues in all the neighborhoods of NYC so as to identify all those areas that are in need of category (1) or category (2) facilities and make a recommendation to the Department of City Planning of NYC.

We will use the Foursquare API location data to explore all neighborhoods of NYC and examine the distribution of (i) hospitals/medical centers, and (ii) parks/gyms/fitness centers across the city. By segmenting, clustering, and plotting neighborhoods that lack either of these facilities on the map of NYC, and by comparing against a map with all the neighborhoods of the city, we can determine which areas of NYC are in dire need of more facilities to improve the health of its citizens. This will allow us to make recommendations to the Department of City Planning of NYC as to which areas of NYC need more health-improvement facilities.

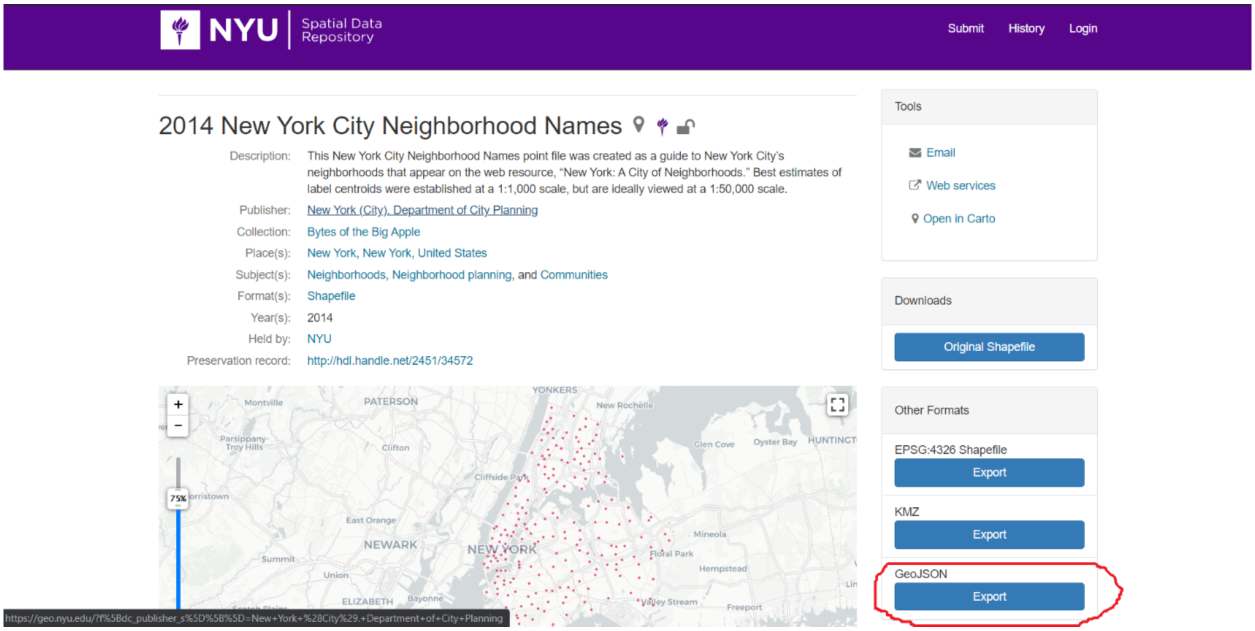
1.3 Interest

It is clear that the Department of City Planning of NYC is a stakeholder here because it is very useful for them to know exactly how they can improve each neighborhood to make NYC a healthier city to live in. Moreover, such an analysis will directly benefit the residents of NYC (target audience), because as stated in Section 1.1, installing more category (1) and/or category (2) facilities has numerous health benefits in adults, such as a reduction of stress, a longer life or better physical and mental health.

2. Data

In this section, we describe the data that we will be using to solve our problem and execute our idea. We will follow the steps below to obtain the data that we need to carry out our analysis.

a) First, we will gather data about the neighborhoods of New York City (NYC). NYC has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and all the neighborhoods that exist in each borough as well as the coordinates of each neighborhood. Luckily, this dataset exists for free on the web. We will obtain it from the following link: https://geo.nyu.edu/catalog/nyu_2451_34572 . We then download the JSON file (indicated by the red circle in the image below) to the local computer and read that data into Python.



2014 New York City Neighborhood Names

Description: This New York City Neighborhood Names point file was created as a guide to New York City's neighborhoods that appear on the web resource, "New York: A City of Neighborhoods." Best estimates of label centroids were established at a 1:1,000 scale, but are ideally viewed at a 1:50,000 scale.

Publisher: New York (City). Department of City Planning

Collection: Bytes of the Big Apple

Place(s): New York, New York, United States

Subject(s): Neighborhoods, Neighborhood planning, and Communities

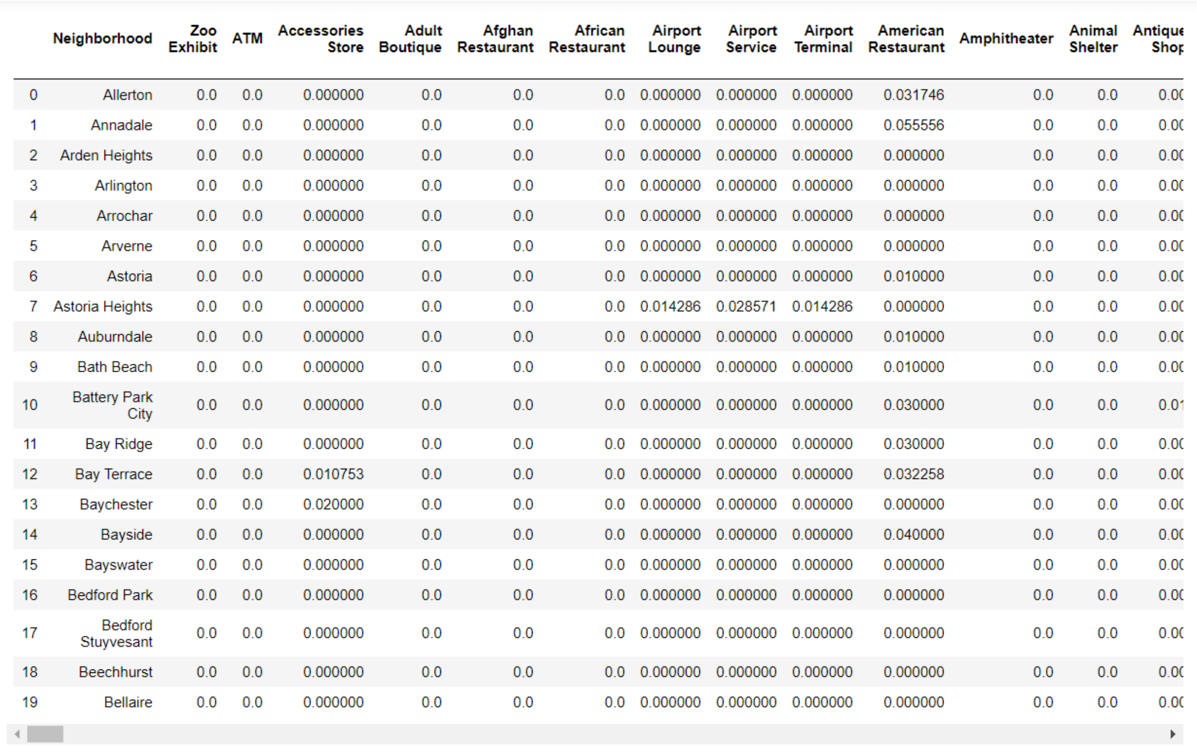
Format(s): Shapefile

Year(s): 2014

Held by: NYU

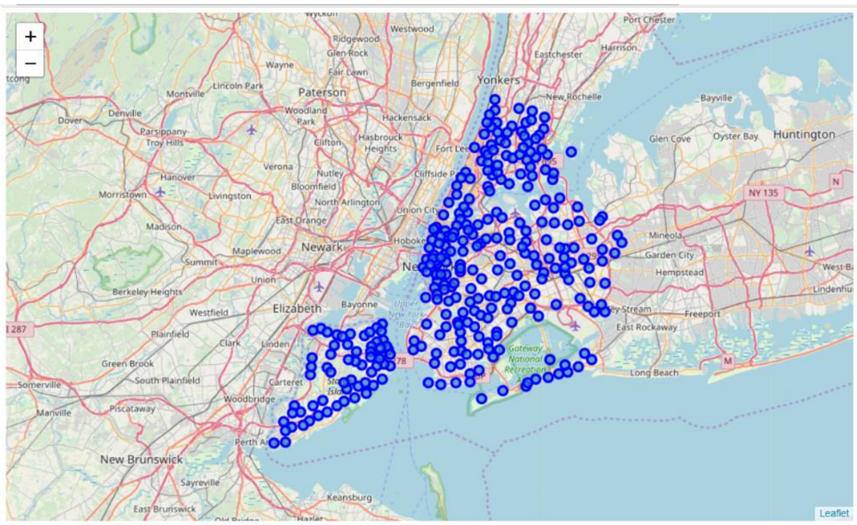
Preservation record: http://hdl.handle.net/2451/34572

b) We will use the **Foursquare API location data** to explore all neighborhoods of NYC. This will give us all the different types of venues (be it coffee shops, restaurants, bars, parks, yoga studios, hospitals, or schools …) and their frequencies of occurrence in each neighborhood. We will put this data into a data frame and then create a map of NYC with all the neighborhoods. The data frame with the different venue types and their frequencies of occurrence will look like:

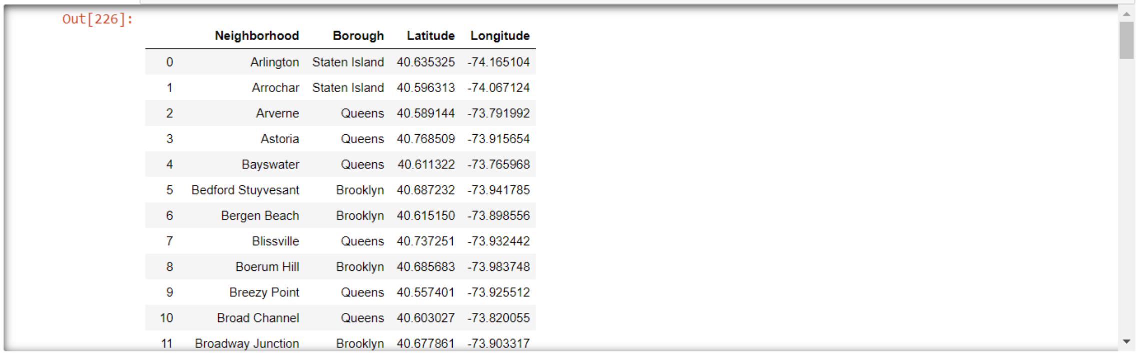


	Neighborhood	Zoo Exhibit	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphitheater	Animal Shelter	Antique Shop
0	Allerton	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.031746	0.0	0.0	0.00
1	Annadale	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.055556	0.0	0.0	0.00
2	Arden Heights	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.00
3	Arlington	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.00
4	Arrochar	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.00
5	Arverne	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.00
6	Astoria	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.010000	0.0	0.0	0.00
7	Astoria Heights	0.0	0.0	0.000000	0.0	0.0	0.0	0.014286	0.028571	0.014286	0.000000	0.0	0.0	0.00
8	Auburndale	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.010000	0.0	0.0	0.00
9	Bath Beach	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.010000	0.0	0.0	0.00
10	Battery Park City	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.030000	0.0	0.0	0.01
11	Bay Ridge	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.030000	0.0	0.0	0.00
12	Bay Terrace	0.0	0.0	0.010753	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.032258	0.0	0.0	0.00
13	Baychester	0.0	0.0	0.020000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.00
14	Bayside	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.040000	0.0	0.0	0.00
15	Bayswater	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.00
16	Bedford Park	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.00
17	Bedford Stuyvesant	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.00
18	Beechhurst	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.00
19	Bellaire	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.00

The map of NYC with all the neighborhoods superimposed on top looks like:



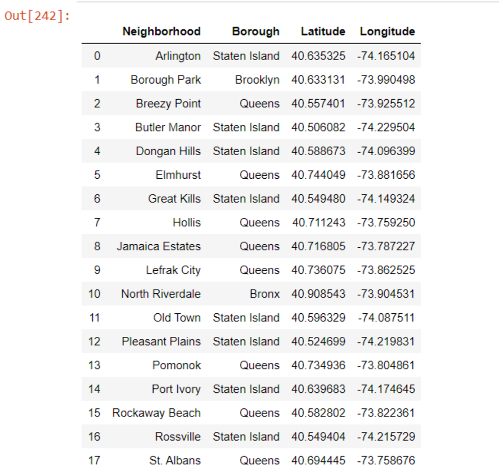
c) Next, we will extract all those neighborhoods that lack hospitals/medical centers (Category 1 health facilities), put them into a data frame, and examine their distribution on a map of NYC (after some segmenting and/or clustering if needed). By comparing against the previous map, we can determine which areas need more hospitals. The “processed” data with the list of neighborhoods that need more category 1 health facilities looks like:



Out[226]:

	Neighborhood	Borough	Latitude	Longitude
0	Arlington	Staten Island	40.635325	-74.165104
1	Arrochar	Staten Island	40.596313	-74.067124
2	Arverne	Queens	40.589144	-73.791992
3	Astoria	Queens	40.768509	-73.915654
4	Bayswater	Queens	40.611322	-73.765968
5	Bedford Stuyvesant	Brooklyn	40.687232	-73.941785
6	Bergen Beach	Brooklyn	40.615150	-73.898556
7	Blissville	Queens	40.737251	-73.932442
8	Boerum Hill	Brooklyn	40.685683	-73.983748
9	Breezy Point	Queens	40.557401	-73.925512
10	Broad Channel	Queens	40.603027	-73.820055
11	Broadway Junction	Brooklyn	40.677861	-73.903317

d) We will repeat the above procedure for parks/gyms/fitness centers/yoga studios etc. (Category 2 health facilities) across the city. The “processed” data with the list of neighborhoods that need more category 2 health facilities looks like:



Out[242]:

	Neighborhood	Borough	Latitude	Longitude
0	Arlington	Staten Island	40.635325	-74.165104
1	Borough Park	Brooklyn	40.633131	-73.990498
2	Breezy Point	Queens	40.557401	-73.925512
3	Butler Manor	Staten Island	40.506082	-74.229504
4	Dongan Hills	Staten Island	40.588673	-74.096399
5	Elmhurst	Queens	40.744049	-73.881656
6	Great Kills	Staten Island	40.549480	-74.149324
7	Hollis	Queens	40.711243	-73.759250
8	Jamaica Estates	Queens	40.716805	-73.787227
9	Lefrak City	Queens	40.736075	-73.862525
10	North Riverdale	Bronx	40.908543	-73.904531
11	Old Town	Staten Island	40.596329	-74.087511
12	Pleasant Plains	Staten Island	40.524699	-74.219831
13	Pomonok	Queens	40.734936	-73.804861
14	Port Ivory	Staten Island	40.639683	-74.174645
15	Rockaway Beach	Queens	40.582802	-73.822361
16	Rossville	Staten Island	40.549404	-74.215729
17	St. Albans	Queens	40.694445	-73.758676

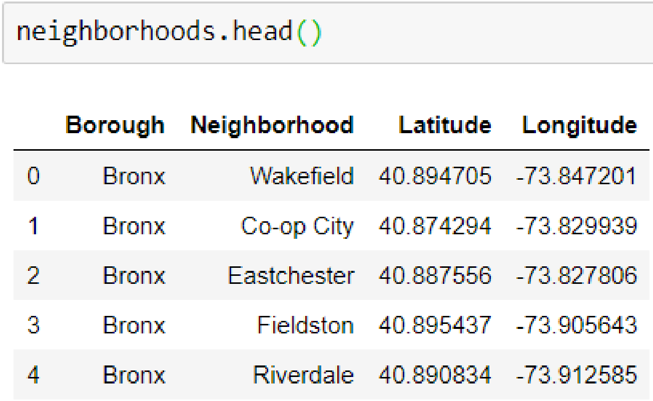
This will then allow us to make recommendations to the Department of City Planning of NYC as to which areas of NYC need more health-improvement facilities.

3. Methodology

This is the section where we carry out an in-depth analysis of our data, implement machine learning algorithms to further segment and cluster the neighborhood data. We first need to do some processing of the data (we alluded to this a bit in Section 2), but we will discuss the details here.

3.1 Processing the data

If we take a quick look at the data from https://geo.nyu.edu/catalog/nyu_2451_34572 , we can see that all the relevant data is in the “features” key, which is basically a list of all the neighborhoods. Putting all this data containing information such as *Neighborhood*, *Borough*, *Latitude*, and *Longitude*, into a data frame that we call “neighborhoods” (see below).

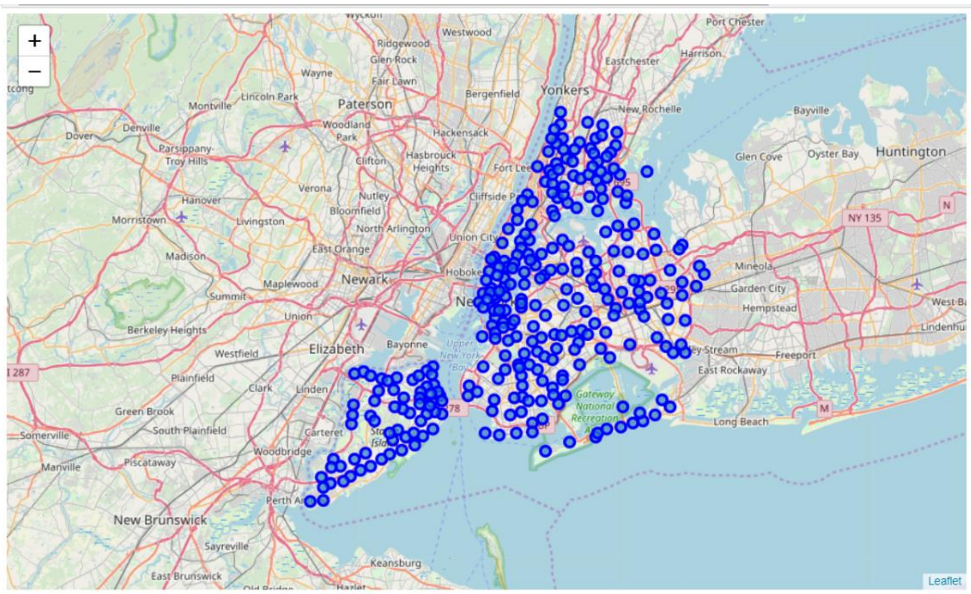


```
neighborhoods.head()
```

	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

3.2 Creating an initial map of NYC

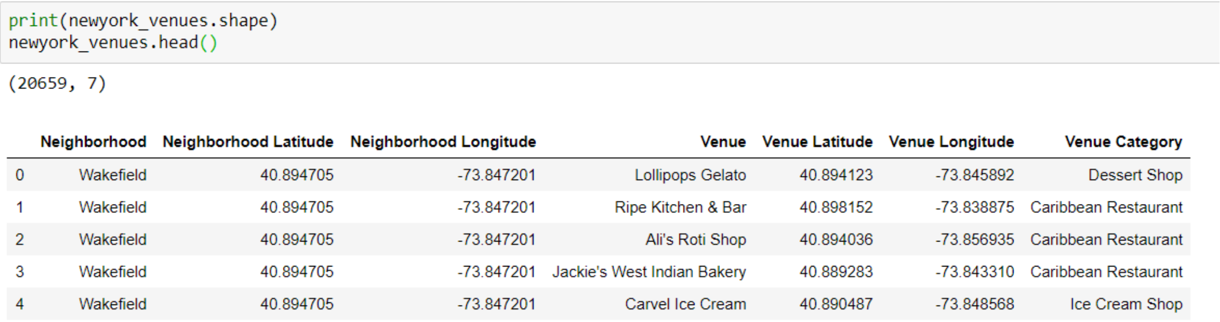
We first use the **geopy** library to get the latitude and longitude values of NYC. We then use the **folium** library to visualize the geographic details of NYC and create a map of NYC with its neighborhoods superimposed on top. This gives us the following map:



This map gives us an idea of the total number of neighborhoods in NYC, their locations, and their spatial distribution. This visual will be useful later when we want to see what portion of NYC is deficient in category 1 or category 2 health facilities.

3.3 Using the Foursquare API

Next we use the Foursquare API to explore the neighborhoods of NYC – in particular, we want to extract information about the different types of venues around each neighborhood (we limit the number of venues returned by the Foursquare API to **100 venues**) within a radius of **1000 meters**. The resulting data frame containing up to 100 venues within 1000 meters of a neighborhood for all neighborhoods had dimensions (20659, 7). The resulting data frame looks something like:



```
print(newyork_venues.shape)
newyork_venues.head()
```

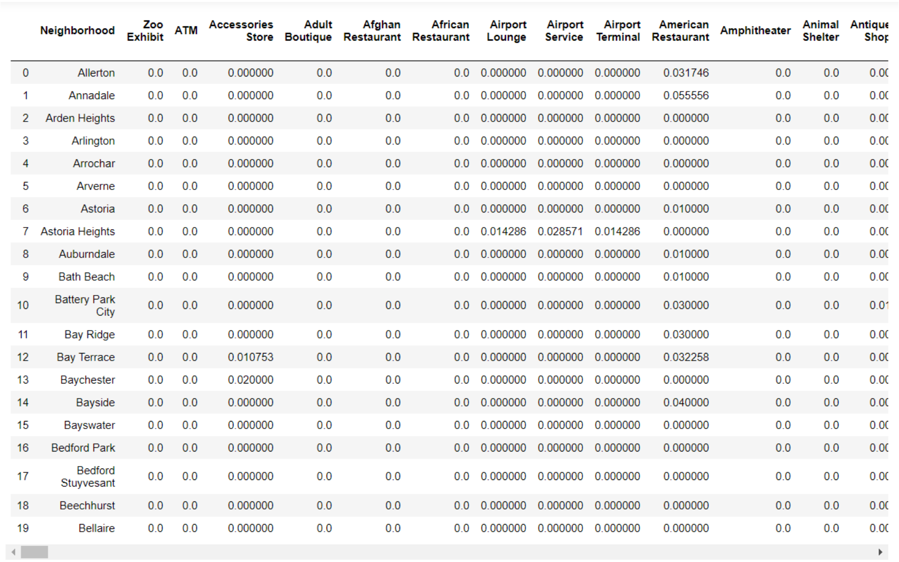
```
(20659, 7)
```

	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	Jackie's West Indian Bakery	40.889283	-73.843310	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop

We determined that **481 unique categories** could be curated from all the venues returned by the Foursquare API.

3.4 Analyzing each neighborhood

In order to determine which neighborhoods need more health facilities, we need to determine the frequency of occurrence of each venue type returned by the Foursquare API for every neighborhood – we calculate this frequency of occurrence and put it into a data frame as shown below:



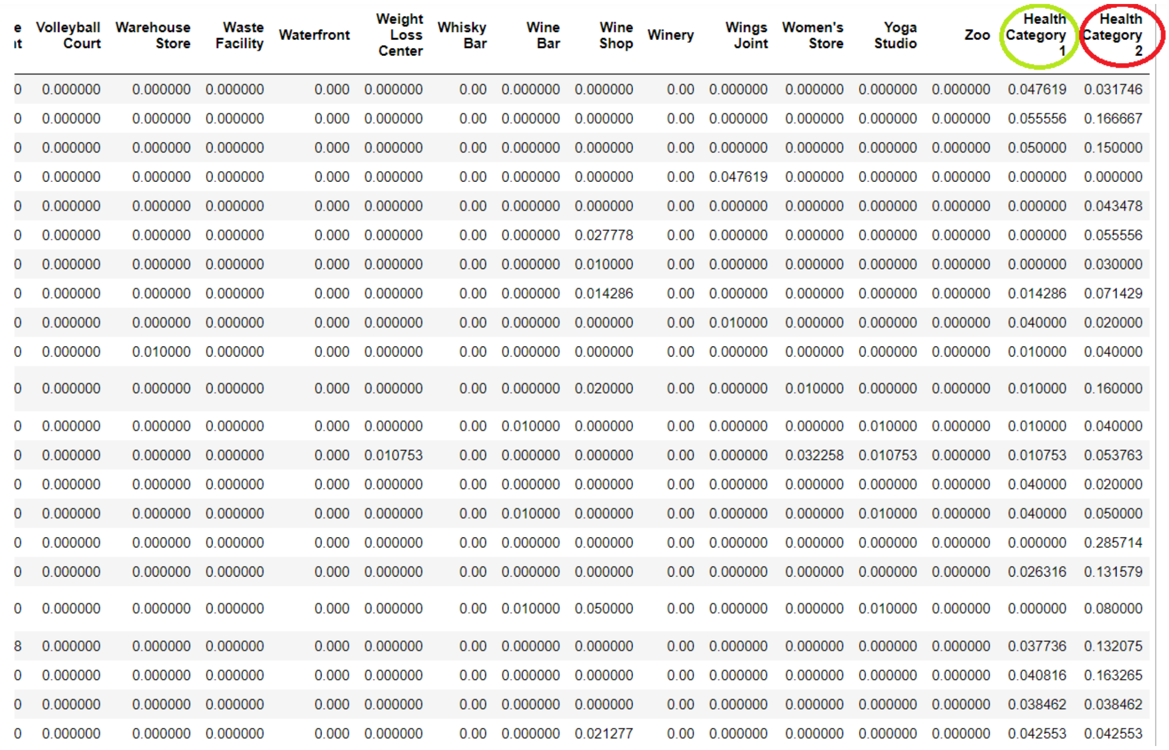
	Neighborhood	Zoo Exhibit	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphitheater	Animal Shelter	Antique Shop
0	Allerton	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.031746	0.0	0.0	0.0
1	Annadale	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.055556	0.0	0.0	0.0
2	Arden Heights	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
3	Arlington	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
4	Arrochar	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
5	Arverne	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
6	Astoria	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.010000	0.0	0.0	0.0
7	Astoria Heights	0.0	0.0	0.000000	0.0	0.0	0.0	0.014286	0.028571	0.014286	0.000000	0.0	0.0	0.0
8	Auburndale	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.010000	0.0	0.0	0.0
9	Bath Beach	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.010000	0.0	0.0	0.0
10	Battery Park City	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.030000	0.0	0.0	0.01
11	Bay Ridge	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.030000	0.0	0.0	0.0
12	Bay Terrace	0.0	0.0	0.010753	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.032258	0.0	0.0	0.0
13	Baychester	0.0	0.0	0.020000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
14	Bayside	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.040000	0.0	0.0	0.0
15	Bayswater	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
16	Bedford Park	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
17	Bedford Stuyvesant	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
18	Beechhurst	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
19	Bellaire	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0

3.5 Analysis of the health facilities in the neighborhoods of NYC

Now we get to the part where we will use the Foursquare API location data to explore all neighborhoods of NYC and examine the distribution of **(i) hospitals/medical centers (call this "Health Category 1")**, and **(ii) parks/gyms/fitness centers (call this "Health Category 2")** across the city. As suggested earlier, by segmenting, clustering, and plotting neighborhoods that lack either of these facilities on the map of NYC, and by comparing against a map with all the neighborhoods of the city, we determine which areas of NYC are in dire need of more facilities to improve the health of its citizens.

Let us begin by defining a list with the venue types (from among all the different venue types retuned by the Foursquare API) that fall under **"Category 1: Basic Medical Facilities"**. These venues are **(i) Doctor's Office, (ii) Emergency Room, (iii) Eye Doctor, (iv) Medical Center, (v) Pharmacy, (vi) Physical Therapy**. Similarly, we define a list with the venue types that fall under **"Category 2: Physical and Mental Health Improvement Facilities"**. Examples of such venues are **yoga studio, gym, recreation center, park, playground, track, trail, and so on**.

We then create a new data frame called "newyork_data_freq" which has the total frequency of occurrence of category 1 and category 2 venues for each neighborhood. The last two columns of the data frame contain information about the frequencies of occurrence of health venues of categories 1 and 2. The resulting data frame look like:

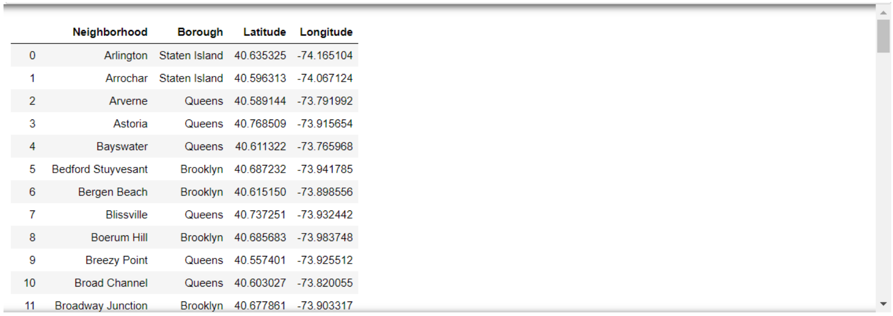


e ıt	Volleyball Court	Warehouse Store	Waste Facility	Waterfront	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Winery	Wings Joint	Women's Store	Yoga Studio	Zoo	Health Category 1	Health Category 2
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.047619	0.031746
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.055556	0.166667
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.050000	0.150000
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.047619	0.000000	0.000000	0.000000	0.000000	0.000000
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.043478
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.027778	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.055556
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.010000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.030000
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.014286	0.00	0.000000	0.000000	0.000000	0.000000	0.014286	0.071429
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.010000	0.000000	0.000000	0.000000	0.040000	0.020000
0	0.000000	0.010000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.010000	0.040000
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.020000	0.00	0.000000	0.010000	0.000000	0.000000	0.010000	0.160000
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.010000	0.000000	0.00	0.000000	0.000000	0.010000	0.000000	0.010000	0.040000
0	0.000000	0.000000	0.000000	0.000	0.010753	0.00	0.000000	0.000000	0.00	0.000000	0.032258	0.010753	0.000000	0.010753	0.053763
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.040000	0.020000
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.010000	0.000000	0.00	0.000000	0.000000	0.010000	0.000000	0.040000	0.050000
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.285714
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.026316	0.131579
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.010000	0.050000	0.00	0.000000	0.000000	0.010000	0.000000	0.000000	0.080000
8	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.037736	0.132075
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.040816	0.163265
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.038462	0.038462
0	0.000000	0.000000	0.000000	0.000	0.000000	0.00	0.000000	0.021277	0.00	0.000000	0.000000	0.000000	0.000000	0.042553	0.042553

Now that we have information about how well-equipped each neighborhood is with these two types of facilities, we determine which neighborhoods need more of these two types of health facilities.

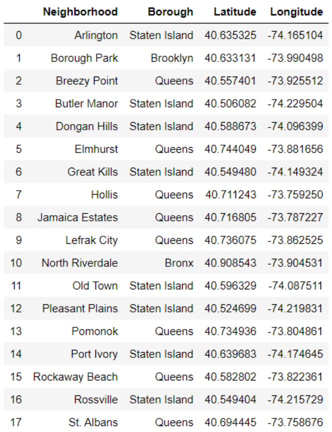
We determine neighborhoods that need more Category 1 facilities by finding all neighborhoods that have a frequency of occurrence of Health Category 1 venues of less than or equal to 0.001. And we determine neighborhoods that need more Category 2 facilities by finding all neighborhoods that have a frequency of occurrence of Health Category 2 venues of less than or equal to 0.01. These are some thresholds based on physical intuition. Of course, since parks, gyms, fitness centers etc. are more abundant than doctor's offices, medical centers, the threshold for category 2 venues is higher.

We call the data frames containing the names of neighborhoods with deficit in category 1 and category 2 venues "nbhds_needing_cat1" and "nbhds_needing_cat2" respectively. We also determine the latitude and longitude coordinates of these neighborhoods and put this information into data frames as follows:



	Neighborhood	Borough	Latitude	Longitude
0	Arlington	Staten Island	40.635325	-74.165104
1	Arrochar	Staten Island	40.596313	-74.067124
2	Arverne	Queens	40.589144	-73.791992
3	Astoria	Queens	40.768509	-73.915654
4	Bayswater	Queens	40.611322	-73.765968
5	Bedford Stuyvesant	Brooklyn	40.687232	-73.941785
6	Bergen Beach	Brooklyn	40.615150	-73.898556
7	Blissville	Queens	40.737251	-73.932442
8	Boerum Hill	Brooklyn	40.685683	-73.983748
9	Breezy Point	Queens	40.557401	-73.925512
10	Broad Channel	Queens	40.603027	-73.820055
11	Broadway Junction	Brooklyn	40.677861	-73.903317

Category 1 deficit neighborhoods



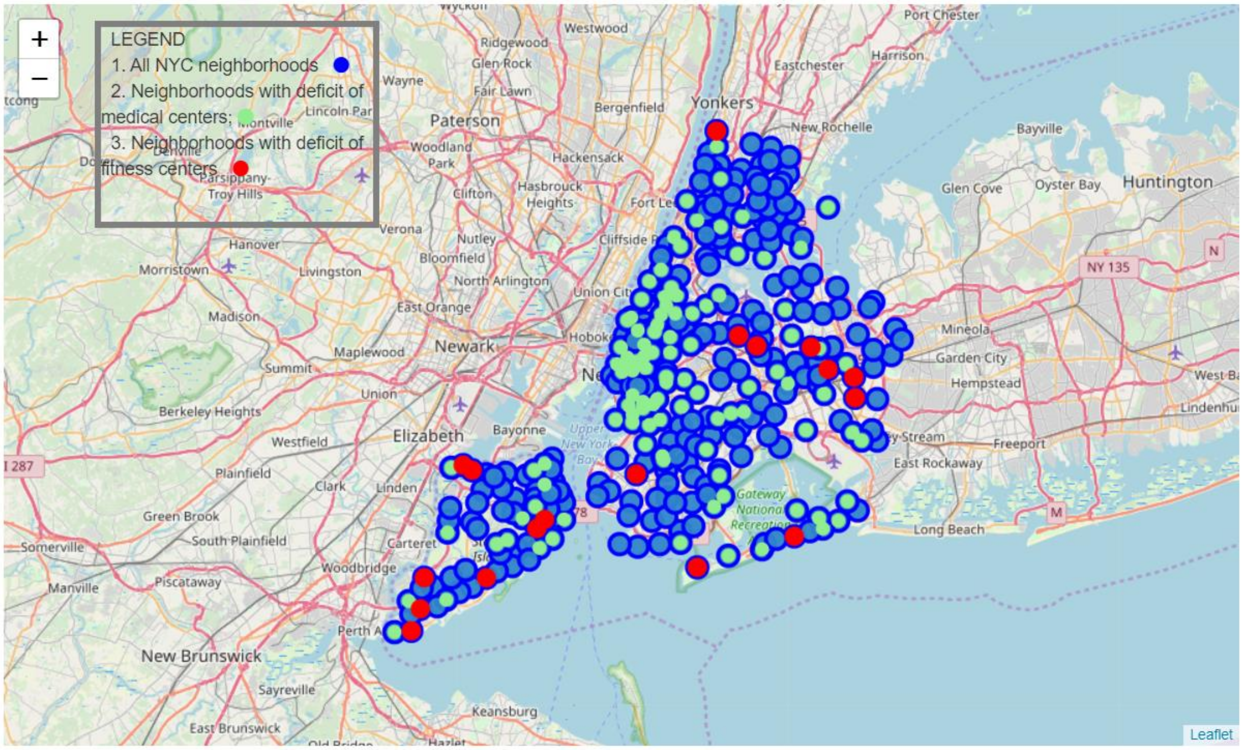
	Neighborhood	Borough	Latitude	Longitude
0	Arlington	Staten Island	40.635325	-74.165104
1	Borough Park	Brooklyn	40.633131	-73.990498
2	Breezy Point	Queens	40.557401	-73.925512
3	Butler Manor	Staten Island	40.506082	-74.229504
4	Dongan Hills	Staten Island	40.588673	-74.096399
5	Elmhurst	Queens	40.744049	-73.881656
6	Great Kills	Staten Island	40.549480	-74.149324
7	Hollis	Queens	40.711243	-73.759250
8	Jamaica Estates	Queens	40.716805	-73.787227
9	Lefrak City	Queens	40.736075	-73.862525
10	North Riverdale	Bronx	40.908543	-73.904531
11	Old Town	Staten Island	40.596329	-74.087511
12	Pleasant Plains	Staten Island	40.524699	-74.219831
13	Pomonok	Queens	40.734936	-73.804861
14	Port Ivory	Staten Island	40.639683	-74.174645
15	Rockaway Beach	Queens	40.582802	-73.822361
16	Rossville	Staten Island	40.549404	-74.215729
17	St. Albans	Queens	40.694445	-73.758676

Category 2 deficit neighborhoods

3.6 Visualizing neighborhoods with health facility deficits on the map

Next, we use the folium library to plot neighborhoods needing category 1 facilities in green, those needing category 2 facilities in red against all NYC neighborhoods (blue circles). This allows us to develop some perspective on the areas of NYC that need more of these facilities to improve the overall well-being of its citizens.

The resulting map is shown below. From the map we can clearly see that there is a **pocket near the center that is in general deficient in category 1 health facilities**. **Most neighborhoods seem to be well-equipped with some form of category 2 facility**, but as one might expect, the category 1 health facilities seem to be deficient in more areas because they need more funding and a lot of skilled labor.



So far, we have only identified which neighborhoods lack category 1 and category 2 health facilities and plotted them on a map. This has helped us develop some intuition for how these neighborhoods are distributed across NYC and provided us with a visual aid for what to expect. It will be immensely useful to formalize the notion of "distribution of neighborhoods" by employing some machine learning clustering algorithms. This is what we get to in the next section.

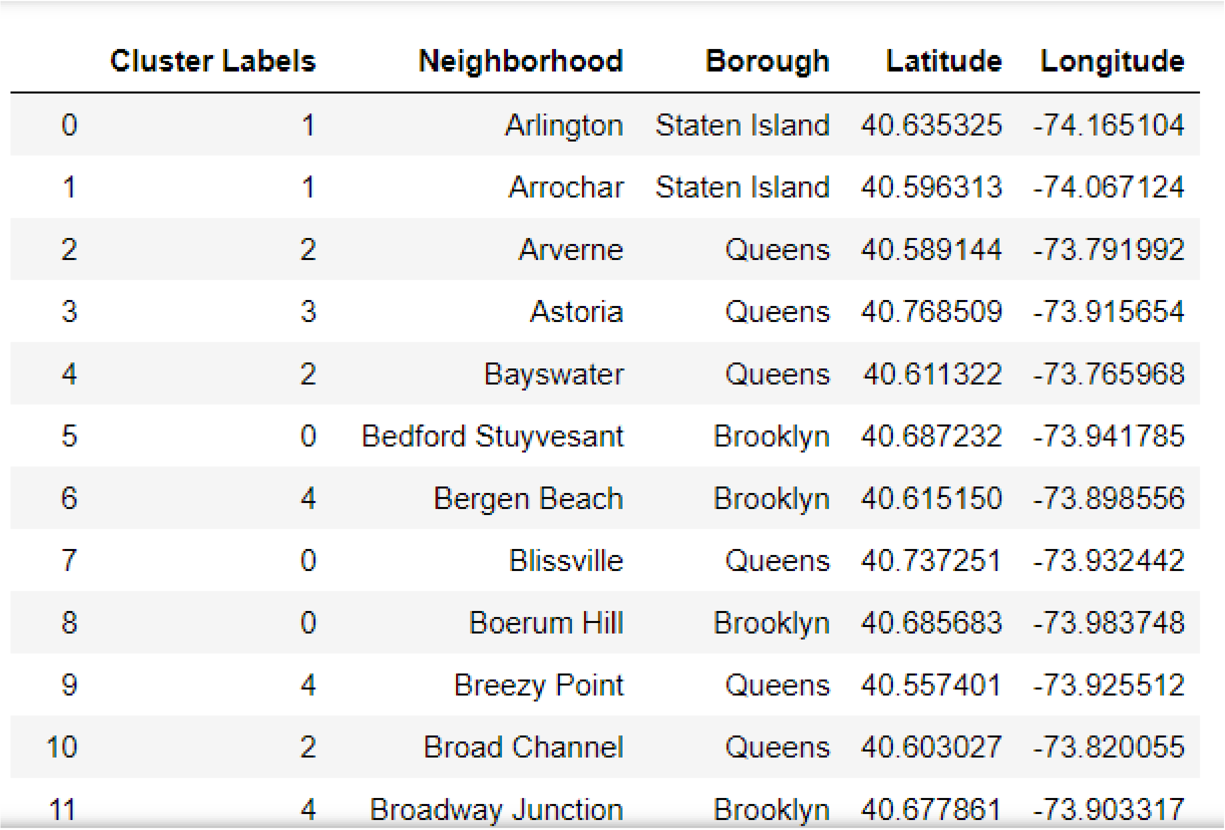
3.7 Machine learning clustering algorithm to cluster neighborhoods

Clustering (or cluster analysis) is a technique that allows us to find groups of similar objects, objects that are more related to each other than to objects in other groups. Here we would like for the clustering algorithms to tell us how neighborhoods belonging to either category fall into "geographical pockets" based on their location.

For this purpose, we use the unsupervised learning **K-means algorithm** to cluster the neighborhoods of category 1 and then those of category 2. K-means algorithm is one of the most commonly used cluster methods of unsupervised learning, and it will be good for us to work with because the K-means algorithm is particularly good at identifying clusters with spherical shapes – this will make it practical to make recommendations to the Department of City Planning of NYC about which areas (loosely defined based on the clusters obtained) need more attention.

3.7.1 Clustering category-1-deficient neighborhoods

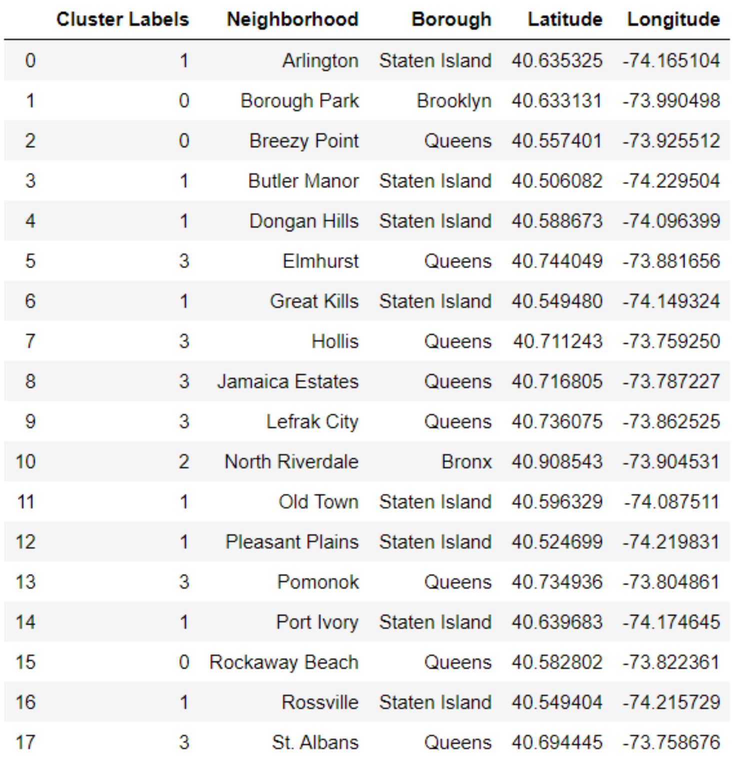
Now let us try to cluster the neighborhoods with a deficit of medical centers/emergency rooms/hospitals/doctor's offices... (what we have called "Category 1" so far) and see if there are any patterns. Upon running the K-means algorithm, the data frame of category-1-deficient neighborhoods with their cluster labels is shown below:



	Cluster Labels	Neighborhood	Borough	Latitude	Longitude
0	1	Arlington	Staten Island	40.635325	-74.165104
1	1	Arrochar	Staten Island	40.596313	-74.067124
2	2	Arverne	Queens	40.589144	-73.791992
3	3	Astoria	Queens	40.768509	-73.915654
4	2	Bayswater	Queens	40.611322	-73.765968
5	0	Bedford Stuyvesant	Brooklyn	40.687232	-73.941785
6	4	Bergen Beach	Brooklyn	40.615150	-73.898556
7	0	Blissville	Queens	40.737251	-73.932442
8	0	Boerum Hill	Brooklyn	40.685683	-73.983748
9	4	Breezy Point	Queens	40.557401	-73.925512
10	2	Broad Channel	Queens	40.603027	-73.820055
11	4	Broadway Junction	Brooklyn	40.677861	-73.903317

3.7.2 Clustering category-2-deficient neighborhoods

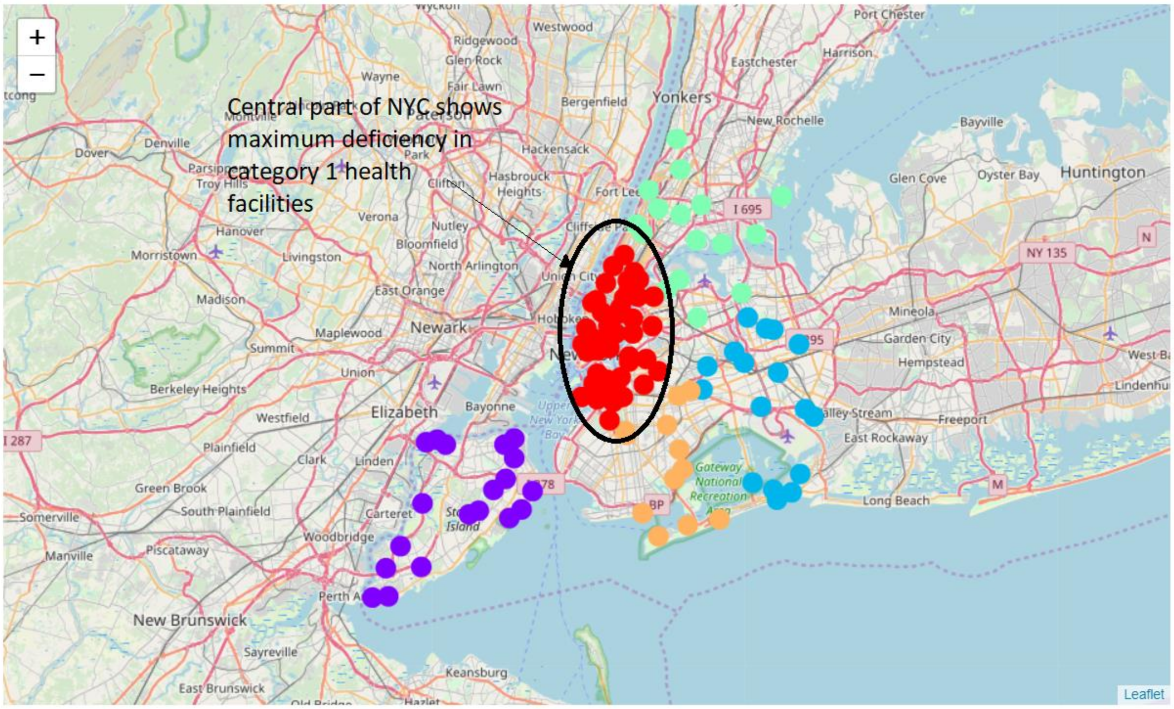
Now let us try to cluster the neighborhoods with a deficit of physical fitness centers /parks/trails/tracks... (what we have called "Category 2" so far) and see if there are any patterns. Upon running the K-means algorithm, the data frame of category-2-deficient neighborhoods with their cluster labels is shown below:



	Cluster Labels	Neighborhood	Borough	Latitude	Longitude
0	1	Arlington	Staten Island	40.635325	-74.165104
1	0	Borough Park	Brooklyn	40.633131	-73.990498
2	0	Breezy Point	Queens	40.557401	-73.925512
3	1	Butler Manor	Staten Island	40.506082	-74.229504
4	1	Dongan Hills	Staten Island	40.588673	-74.096399
5	3	Elmhurst	Queens	40.744049	-73.881656
6	1	Great Kills	Staten Island	40.549480	-74.149324
7	3	Hollis	Queens	40.711243	-73.759250
8	3	Jamaica Estates	Queens	40.716805	-73.787227
9	3	Lefrak City	Queens	40.736075	-73.862525
10	2	North Riverdale	Bronx	40.908543	-73.904531
11	1	Old Town	Staten Island	40.596329	-74.087511
12	1	Pleasant Plains	Staten Island	40.524699	-74.219831
13	3	Pomonok	Queens	40.734936	-73.804861
14	1	Port Ivory	Staten Island	40.639683	-74.174645
15	0	Rockaway Beach	Queens	40.582802	-73.822361
16	1	Rossville	Staten Island	40.549404	-74.215729
17	3	St. Albans	Queens	40.694445	-73.758676

4. Results

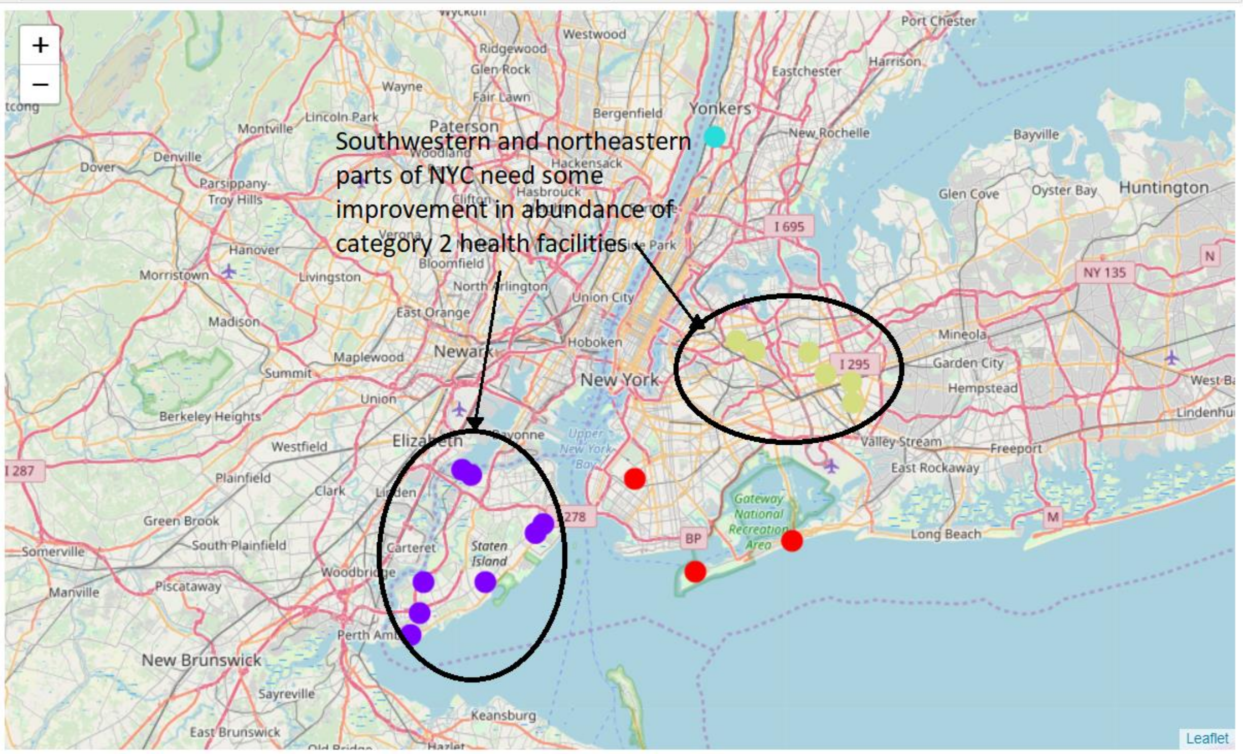
We will visualize the clusters of category-1-deficient neighborhoods on a map of NYC that we plot using the folium library. We get the following plot:



Central part of NYC shows maximum deficiency in category 1 health facilities

We see from the above map that the neighborhoods lacking in category 1 health facilities are segmented based on geographic proximity, and this gives us a nice picture of the pockets of NYC that need attention from the Department of City Planning as far as the construction of basic health facilities is concerned. The central part of NYC seems to be most crowded with neighborhoods lacking category 1 health facilities.

Next, we visualize the clusters of category-2-deficient neighborhoods on a map of NYC that we plot using the folium library. We get the following plot:


+
−
Southwestern and northeastern parts of NYC need some improvement in abundance of category 2 health facilities
New York
Yonkers
New Rochelle
Newark
Elizabeth
Staten Island
Hoboken
Paterson
Huntington
Mineola
Hempstead
Long Beach
New Brunswick
Gateway National Recreation Area
Leaflet

Once again, we see from the map (above) that the neighborhoods lacking category 2 health facilities are segmented based on geographic proximity, and this gives us a nice picture of the pockets of NYC that need attention from the Department of City Planning as far as having "tier 2" health facilities is concerned -- the map above suggests that the southwestern and northeastern parts of NYC might be lacking a bit on category 2 health facilities. But NYC seems to be doing much better on this front as compared with its performance on category 1 facilities.

5. Discussion

In this project, we have used data pertaining to the boroughs and neighborhoods of NYC and the Foursquare API location data to identify neighborhoods that need more health-improvement facilities of two categories: Category 1 corresponds to Basic Medical Facilities such as medical centers, hospitals, doctor’s offices, emergency rooms, etc., and Category 2 which corresponds to Physical and Mental Health Improvement Facilities like parks, gyms, trails, tracks, yoga studios etc.

We collated the neighborhood data with the data on venues around each neighborhood that we obtained from the Foursquare API to arrive at a list of neighborhoods deficient in each category. We then used the K-means clustering algorithm to see how they fell into geographic pockets and made the following observations:

(a) The data frame with the list of neighborhoods requiring more category 1 facilities produced 110 entries. This means that roughly 1/3rd of the neighborhoods of NYC could use more medical centers, emergency rooms, and hospitals. In particular, the central part of NYC seems to be most crowded with neighborhoods lacking category 1 health facilities.

(b) The data frame with the list of neighborhoods requiring more category 2 facilities produced 17 entries. This means that roughly 6% of the neighborhoods of NYC could use more parks, public gyms, tracks, bike paths etc. The southwestern and northeastern parts of NYC might be lacking a bit on category 2 health facilities. But NYC seems to be doing much better on this front as compared with its performance on category 1 facilities.

Based on these observations, we can make the following recommendations to the Department of City Planning of New York City:

i. To reserve funds and allocate resources for the construction of more medical centers with advanced facilities in the pockets of NYC (given by the clusters in our map). This means that additional funding will be needed to employ more doctors, physicians, etc. This seems to be more of an urgent requirement than category 2, especially because the population of NYC is growing with each year given that it offers so many opportunities. Creating more medical facilities will also create more jobs.

ii. To allocate funds towards the construction of parks and community gyms so that the residents of NYC can avail these facilities to improve their general health. This category requires less skilled labor and has less maintenance costs as compared to category 1.

6. Conclusions

In this project, we used data analysis and machine learning to determine how location data can be used to improve the overall health of NYC residents. We thoroughly went over the venues surrounding each neighborhood and determined the frequency of occurrence of two categories of health facilities: category 1 pertaining to basic medical facilities, and category 2 pertaining to other amenities to improve public health.

Our in-depth analysis aided by map visuals and clustering algorithms determined pockets of NYC that were suffering from lack of ease of accessibility to health facilities. Based on our analysis, we were able to make some important and useful observations that would immediately improve the health of NYC residents, and we collated and organized our data and results in a manner that would be suited for making recommendations to the Department of

City Planning of NYC. We also laid out some ideas for possible recommendations that immediately follow from our inferences. If implemented, these changes will have numerous health benefits in adults, such as a reduction of stress, a longer life or better general and mental health.