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

Coursera Capstone Project

Introduction/ Business Problem

- Our health and that of our planet depends greatly on how cities are planned
- We want cities where people can live well and be healthy
- New York City (NYC) is a bustling place – Department of City Planning of NYC needs to *focus on making the city a healthy place to live in for its residents!*
- Need to increase ease of accessibility to health-improvement centers such as medical centers, emergency rooms, parks, gyms, trails, tracks …
- This immensely *benefits the general health of the public!*

HOW DO WE DO THIS??

Data

- One solution: *use location data*

- Collect information about boroughs and neighborhoods of NYC from https://geo.nyu.edu/catalog/nyu_2451_34572 (5 boroughs, 306 neighborhoods)

- Process the data from the source above – put it into a data frame using *pandas*

- Use the *Foursquare API location data* to explore all neighborhoods of NYC

- From the Foursquare data, extract information about venue types around neighborhoods and their frequency of occurrence

- Cleaned and processed data has *481 distinct venue categories* across all neighborhoods

Methodology

1. Processing the data

All relevant data is in the *features* key, which is a list of all the neighborhoods -- put all this data into a data frame

```
neighborhoods.head()
```

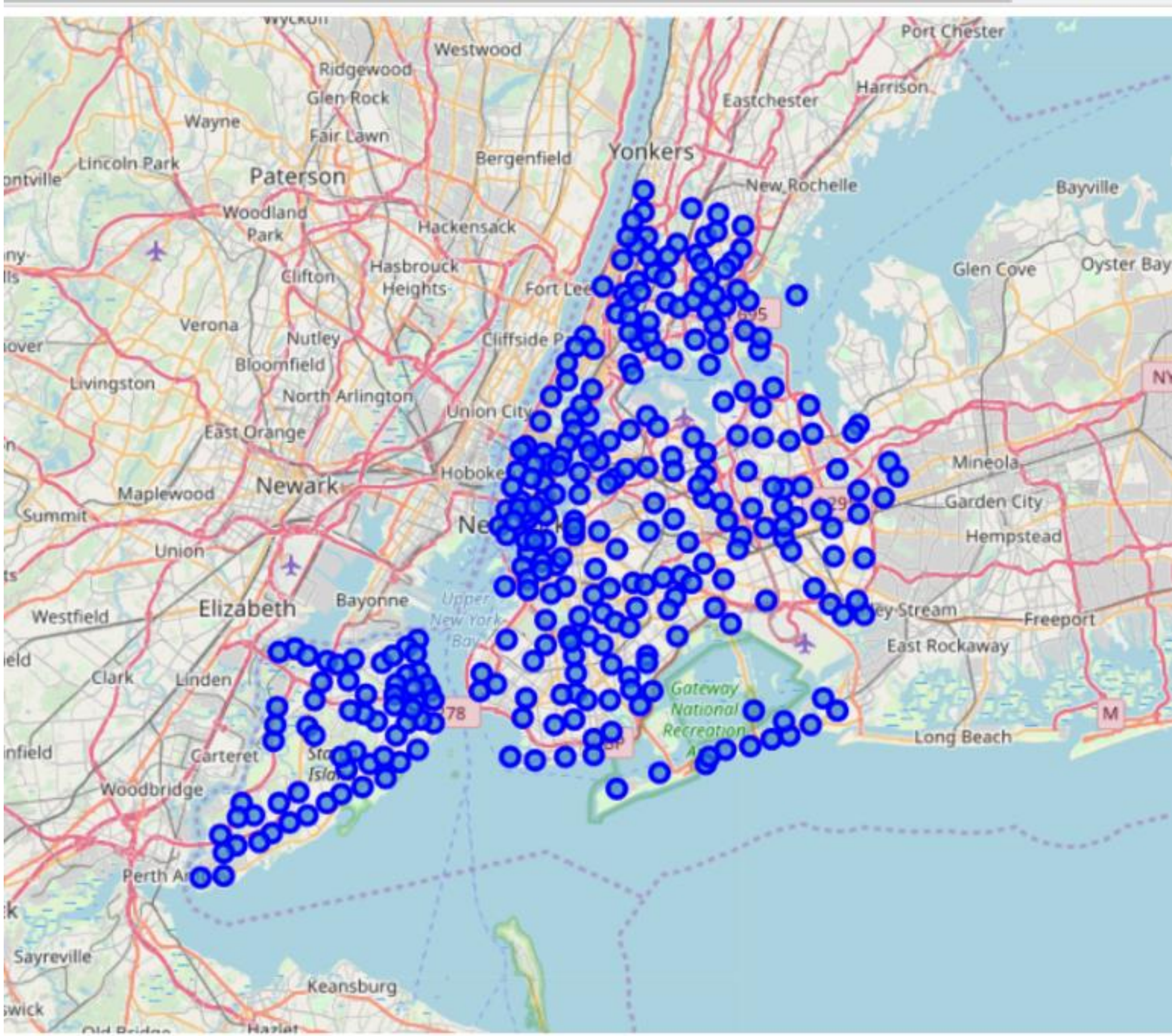
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Methodology

2. <u>Creating an initial map of NYC</u>

- Use the *geopy* library to get the latitude and longitude values of NYC

- Then use the *folium* library to create a map of NYC with its neighborhoods superimposed on top



Methodology

3. Using the Foursquare API

- Use Foursquare API to explore the neighborhoods of NYC – we want to extract information about the different types of venues around each neighborhood
- Limit the number of venues returned by the Foursquare API to **100 venues** within a radius of **1000 meters**
- **481 unique categories** could be curated from all the venues returned by the Foursquare API

```
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

Methodology

4. Analyzing each neighborhood

- Need to determine the frequency of occurrence of each venue type returned by the Foursquare API for every neighborhood
- Calculate this frequency of occurrence and put it into a data frame

	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

Methodology

5. Analysis of health facilities in NYC neighborhoods

- Use Foursquare API location data to examine the distribution of:
- Category 1: Basic Medical Facilities : Doctor's Office, Emergency Room, Eye Doctor, Medical Center, Pharmacy, Physical Therapy
- Category 2: Physical and Mental Health Improvement Facilities: Yoga Studio, Gym, Recreation Center, Park, Playground, Track, Trail, ….
- Put this data into a data frame which has the total frequency of occurrence of category 1 and 2 venues for each neighborhood

e it	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

Methodology

5. Analysis of health facilities in NYC neighborhoods

- Determine neighborhoods that need more Category 1 facilities by finding all neighborhoods that have a frequency of occurrence of Health Category 1 venues ≤ 0.001
- Determine neighborhoods that need more Category 2 facilities by finding all neighborhoods that have a frequency of occurrence of Health Category 2 venues ≤ 0.01
- Thresholds based on physical intuition -- since parks, fitness centers etc. are more abundant than ERs, medical centers, the threshold for category 2 venues is higher

Collate this data into data frames

	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

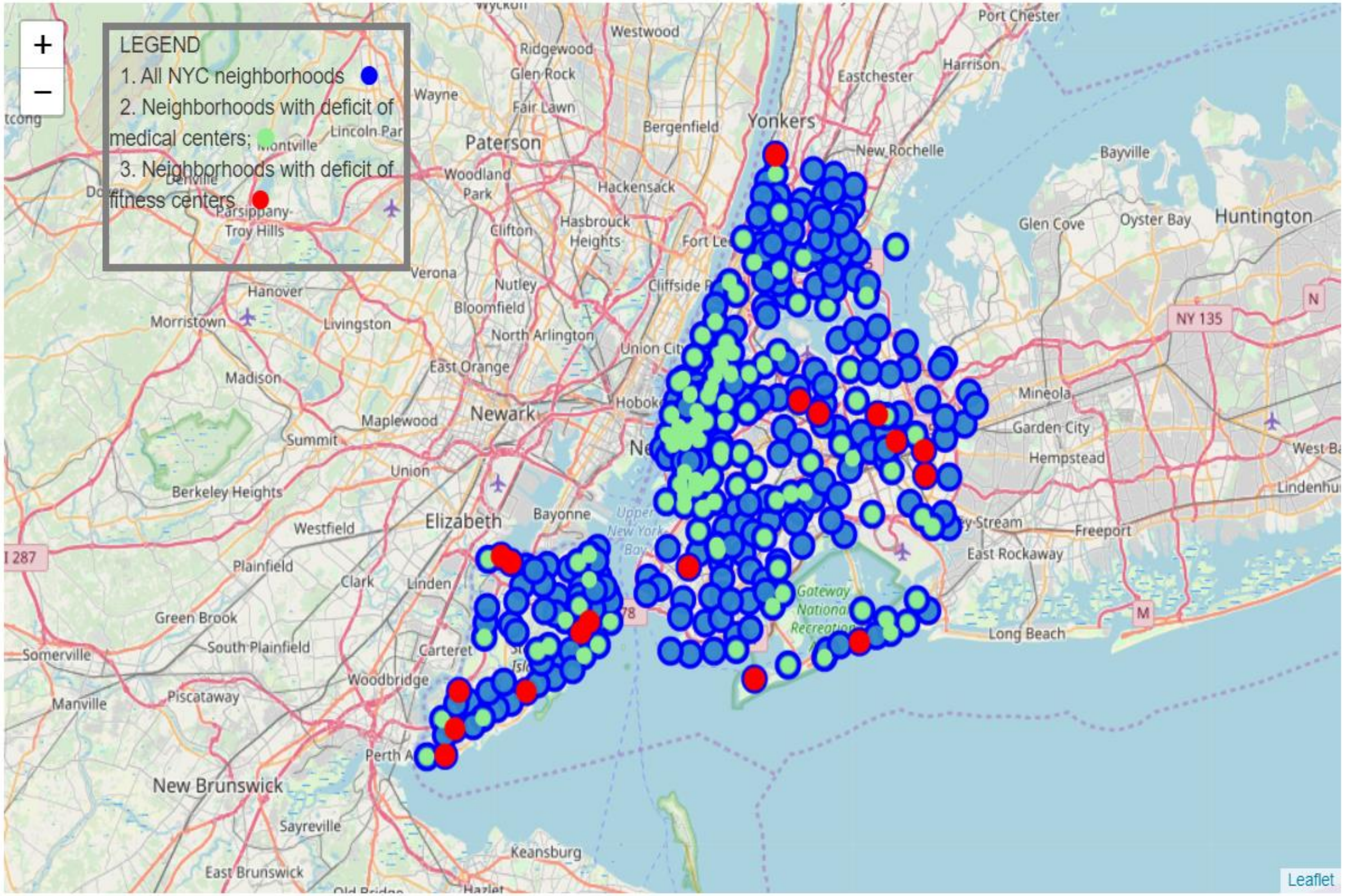
Methodology

6. <u>Visualizing neighborhoods with health facility deficits on the map</u>

- Use 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 is a good eye-balling aid: allows us to develop perspective on the areas of NYC that need more of these facilities to improve the overall well-being of its citizens

- Useful to formalize the notion of “distribution of neighborhoods” by employing some *machine learning clustering algorithms*



Methodology

7. <u>Machine learning clustering algorithm to cluster neighborhoods</u>

- *Clustering* (or cluster analysis) is a technique that allows us to find groups of similar objects

- Want to know how neighborhoods belonging to either category fall into "geographical pockets" based on their location

- We use the unsupervised learning K-means algorithm -- because it is good at identifying clusters with spherical shapes

- Can use this to make recommendations to the *Department of City Planning of NYC* about which areas (defined based on the clusters obtained) need more attention

Methodology

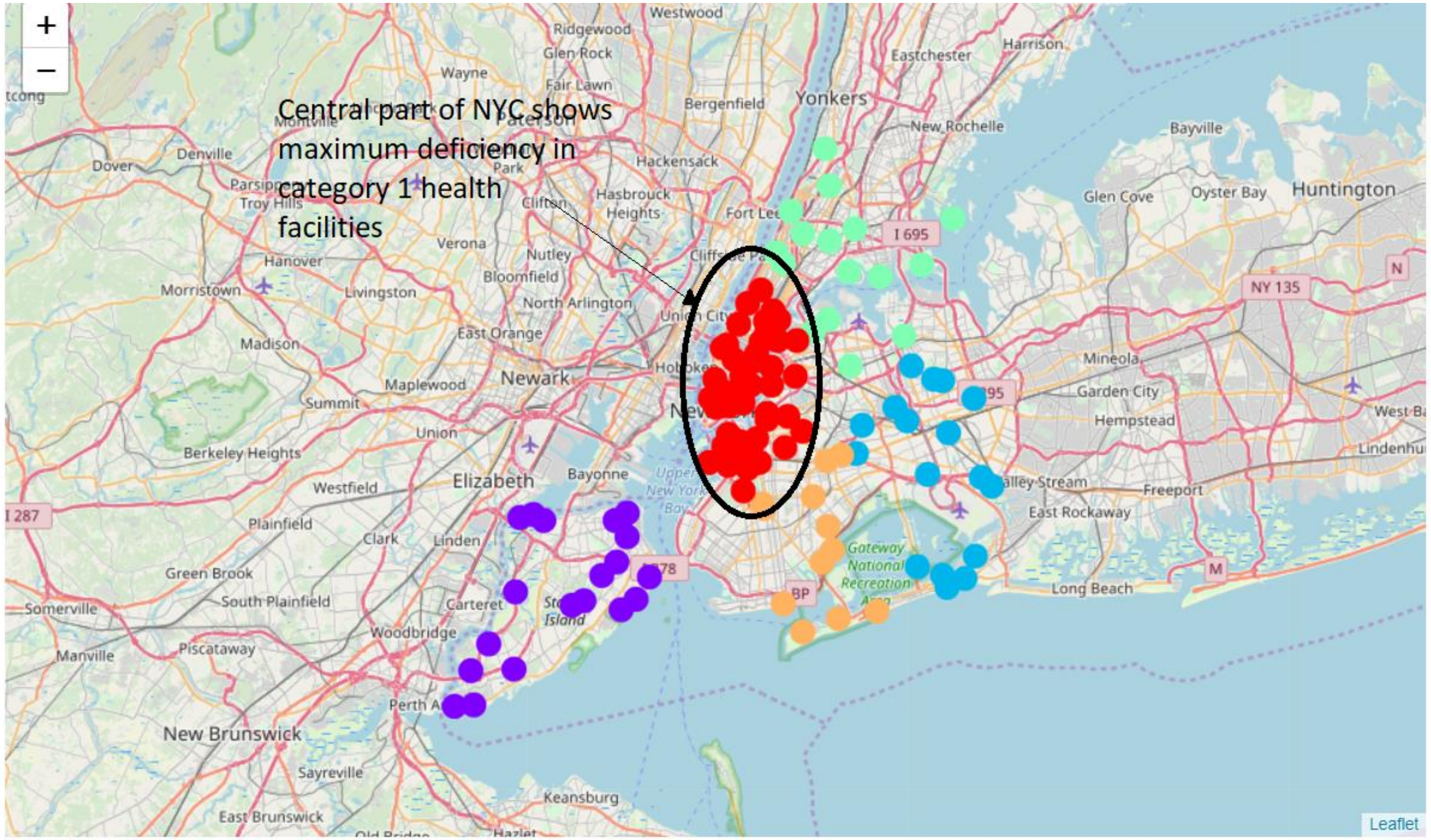
Data frames showing category-1- and category-2-deficient neighborhoods

	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

Category-1-deficient neighborhood clusters

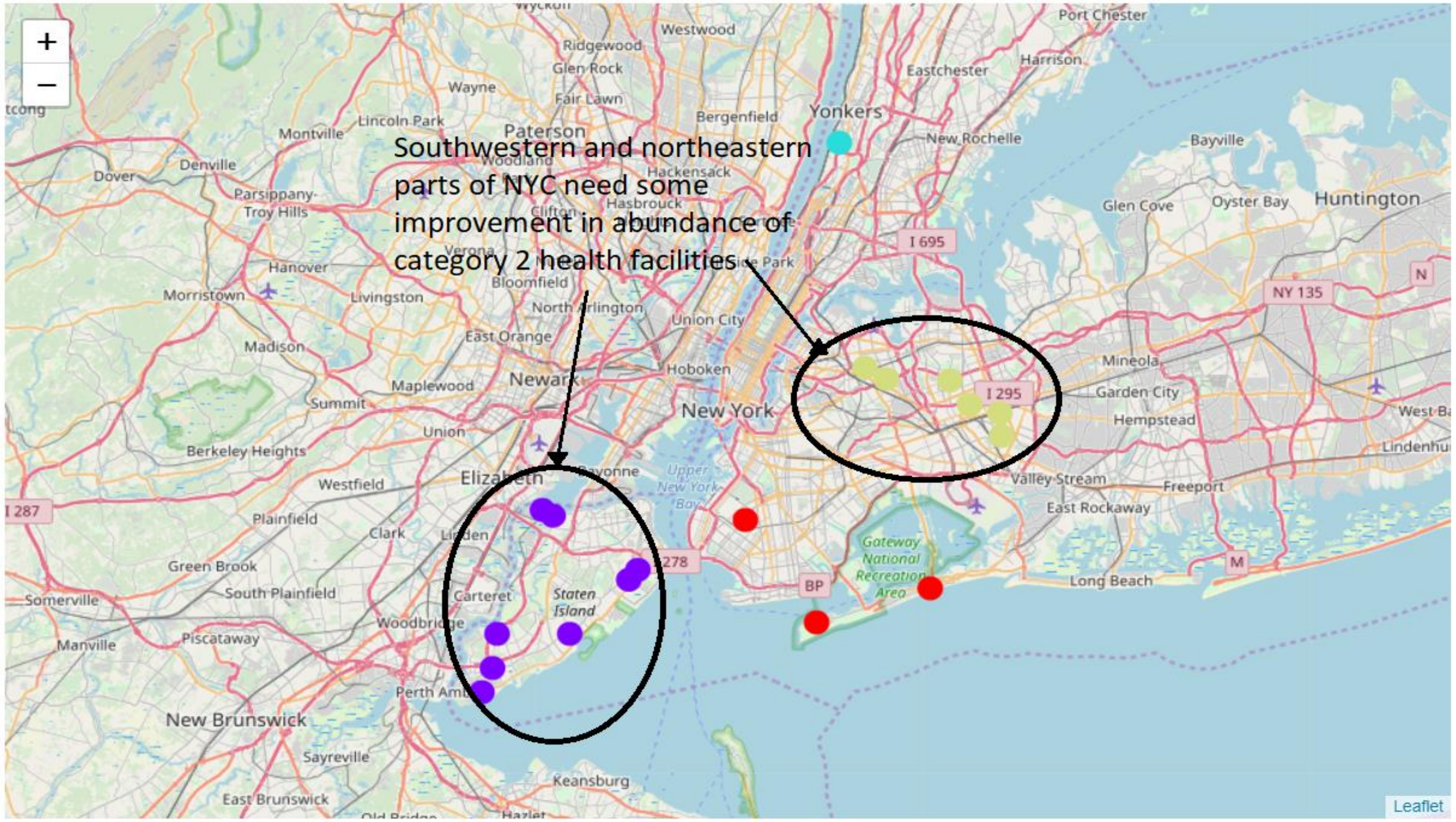
	Cluster Labels	Neighborhood	Borough	Latitude	Longitude
0	1	Arlington	Staten Island	40.635325	-74.165104
1	0	Borough Park	Brooklyn	40.633131	-73.990498
2	0	Breezy Point	Queens	40.557401	-73.925512
3	1	Butler Manor	Staten Island	40.506082	-74.229504
4	1	Dongan Hills	Staten Island	40.588673	-74.096399
5	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

Category-2-deficient neighborhood clusters



Results

- We visualize the clusters of category-1-deficient neighborhoods on a map of NYC
- We see that neighborhoods lacking in category 1 health facilities are segmented based on geographic proximity
- The central part of NYC seems to be most crowded with neighborhoods lacking category 1 health facilities



Results

- We visualize the clusters of category-2-deficient neighborhoods on a map of NYC
- We see that neighborhoods lacking category 2 health facilities are segmented based on geographic proximity
- The <u>southwestern and northeastern parts of NYC might be lacking a bit on category 2 health facilities</u>
- NYC seems to be doing much better on this front as compared with its performance on category 1 facilities

Discussion of results

- Roughly 1/3rd of the neighborhoods of NYC could use more medical centers, emergency rooms, and hospitals (category 1)
- The central part of NYC seems to be most crowded with neighborhoods lacking category 1 health facilities
- Roughly 6% of the neighborhoods of NYC could use more parks, public gyms, tracks, bike paths (category 2) etc.
- The southwestern and northeastern parts of NYC might be lacking a bit on category 2 health facilities
- NYC seems to be doing much better on this front as compared with its performance on category 1 facilities

Recommendations to the Department of City Planning of New York City

- 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 will need additional funding to employ more doctors and physicians -- Creating more medical facilities will also create more jobs
- 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
- Category 2 requires less skilled labor and has less maintenance costs as compared to category 1

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*

- 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 *machine learning 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

- If implemented, *our recommendations will have numerous health benefits* in adults, such as a reduction of stress, a longer life or better general and mental health