**This notebook is an implementation of the IBM Data Science Certificate Capstone project.**

**Author: Linh Nguyen**

# Battle of the Neighborhoods

# INTRODUCTION

## PROJECT MOTIVATION

As of 07/28/2020, the U.S. has 4.2 millions confirmed cases of COVID-19. New York City is the at the epicenter of the battle against the pandemics with more than 221,000 cases confirmed and approximately 56,000 cases of hospitalization.

This project aims to analyze the hospital bed density and the COVID-19 case rate for each neighborhood in the New York City. The hospital bed density is measured as the number of beds per 1,000 people. It can be considered as one of the measurement of the service availbility. When pitted against the COVID-19 case rate, it can provide a better picture of the pandemic.

This project is also motivated by lisu1222: The battle of the neighborhoods: hospital density (https://github.com/lisu1222/The-Battle-of-Neighborhoods-New-York-Hospital-Bed-Density)

and ruddra: capstone project (https://github.com/ruddra/project-capstone-nyc-fights-pandemic)

# Data

NYC Neighborhood Data is a json file dowloaded from New York City Neighborhoods Names (https://geo.nyu.edu/catalog/nyu_2451_34572). It includes latitude and longitude for each neighborhood of the 5 boroughs in NYC.

NYC Population Data includes population data per neighborhood that were scrapped from Wikipedia, and then integrated with New York Neighborhood Data.

Hospital Data is scrapped from New York State Department of Health. It contains the name, bed type and the corresponding bed numbers in each type.

NYC COVID 19 case rate data is taken from NYC health (https://www1.nyc.gov/site/doh/covid/covid-19-data.page). The data set contains case count, case rates, death rate and so on. These rate are calculated per 100,000 and thus will need to be processed.

These data sets will be processed, explored and integrated for further analysis.

# METHOD

## 1. NYC Neighborhood and Population Data

The data contained in New York City Neighborhoods Names (https://geo.nyu.edu/catalog/nyu_2451_34572) is in a json file. The latitude and longitude for each neighborhood of the 5 boroughs in NYC is extracted and saved into a dataframe. Beauty Soup was used to scrapped the wiki pages of each neighborhood to find the population data. These 2 datasets are then combined together as follow:

| | Unnamed: 0 | Borough | Neighborhood | Latitude | Longitude | Population |
|---|---|---|---|---|---|---|
| 0 | 0 | Bronx | Wakefield | 40.894705 | -73.847201 | 29158 |
| 1 | 1 | Bronx | Co-op City | 40.874294 | -73.829939 | 43752 |
| 2 | 2 | Bronx | Fieldston | 40.895437 | -73.905643 | 3292 |
| 3 | 3 | Bronx | Riverdale | 40.890834 | -73.912585 | 48049 |
| 4 | 4 | Bronx | Kingsbridge | 40.881687 | -73.902818 | 10669 |
| ... | ... | ... | ... | ... | ... | ... |
| 137 | 137 | Brooklyn | Dumbo | 40.703176 | -73.988753 | 1139 |
| 138 | 138 | Brooklyn | Homecrest | 40.598525 | -73.959185 | 44316 |
| 139 | 139 | Queens | Middle Village | 40.716415 | -73.881143 | 37929 |
| 140 | 140 | Brooklyn | Erasmus | 40.646926 | -73.948177 | 135619 |
| 141 | 141 | Manhattan | Hudson Yards | 40.756658 | -74.000111 | 70150 |

142 rows × 6 columns

## 2. Hospital Data

I used Foursquare API to explore the information of the hospital in each neighborhood, including name of the hospital, neighborhood, borough, longitude and latitude.
With ID of each hospital, I then used Beauty Soup to scrap the New York State Department of Health websites to obtain information on the name of the hospital, bed type and the corresponding bed numbers in each type.\ The package fuzzywuzzy was then used to match the information of the hospital in the dataset 1 and the dataset 2 on the name of the hospital.
After processed, this dataset is merged with the neighborhood and population data set. Then the ICU bed per Hundred people and bed per hundred people were added.

| | Borough | Neighborhood | Bed Number | ICU Bed Number | Unnamed: 0 | Latitude | Longitude | Population | ICU Bed Per Hundred People | Bed Per Hundred People |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Brooklyn | Bensonhurst | 204 | 8 | 34 | 40.611009 | -73.995180 | 151705 | 0.005273 | 0.134472 |
| 1 | Queens | Briarwood | 671 | 24 | 107 | 40.710935 | -73.811748 | 53877 | 0.044546 | 1.245429 |
| 2 | Brooklyn | Brighton Beach | 306 | 17 | 38 | 40.576825 | -73.965094 | 35547 | 0.047824 | 0.860832 |
| 3 | Brooklyn | Brownsville | 600 | 28 | 44 | 40.663950 | -73.910235 | 58300 | 0.048027 | 1.029160 |
| 4 | Brooklyn | Bushwick | 324 | 16 | 46 | 40.698116 | -73.925258 | 129239 | 0.012380 | 0.250698 |

# 3. Covid case rate Data

NYC COVID 19 case rate data is taken from [NYC health (https://www1.nyc.gov/site/doh/covid/covid-19-data.page)](https://www1.nyc.gov/site/doh/covid/covid-19-data.page). The data set contains case count, case rates, death rate and so on.

| | MODIFIED_ZCTA | NEIGHBORHOOD_NAME | BOROUGH_GROUP | COVID_CASE_COUNT | COVID_CASE_RATE | POP_DENOMINATOR | COVID_DEATH_COUNT | CC |
|---|---|---|---|---|---|---|---|---|
| 0 | 10001 | Chelsea/NoMad/West Chelsea | Manhattan | 420 | 1782.45 | 23563.03 | 25 | |
| 1 | 10002 | Chinatown/Lower East Side | Manhattan | 1229 | 1601.19 | 76755.41 | 160 | |
| 2 | 10003 | East Village/Gramercy/Greenwich Village | Manhattan | 515 | 957.22 | 53801.62 | 34 | |
| 3 | 10004 | Financial District | Manhattan | 39 | 1068.32 | 3650.61 | 1 | |
| 4 | 10005 | Financial District | Manhattan | 79 | 940.91 | 8396.11 | 2 | |

I decided to limit the relevant information to the case rates only.
This dataset is then integrated into the hospital data set to obtain the nyc final data set by matching on the neighborhood variable.
(At this stage, I applied a rather greedy matching process which result in duplicated data. This will be discussed as a limitation of the project)

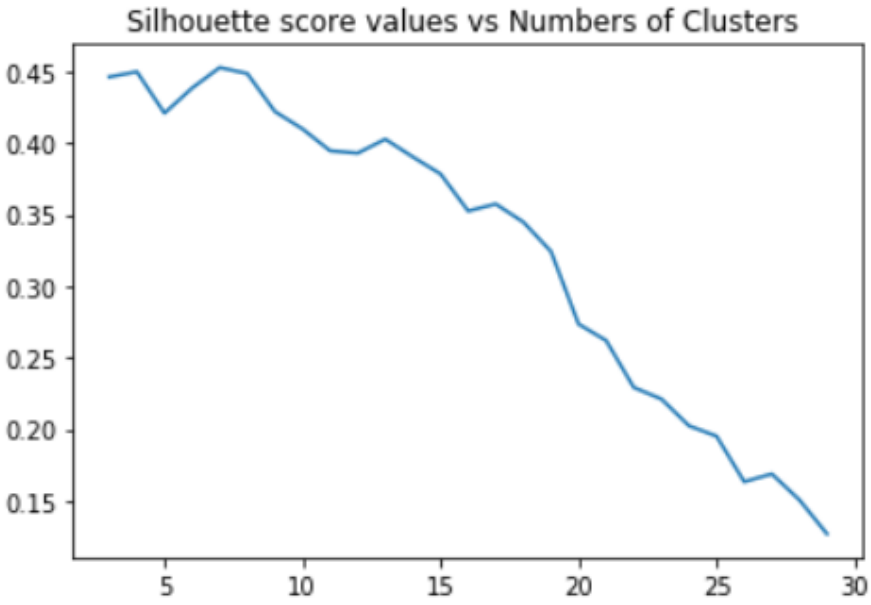
| | Borough | Neighborhood | Bed Number | ICU Bed Number | Unnamed: 0 | Latitude | Longitude | Population | ICU Bed Per Hundred People | Bed Per Hundred People | COVID_CASE_RATE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Brooklyn | Bensonhurst | 204 | 8 | 34 | 40.611009 | -73.995180 | 151705 | 0.005273 | 0.134472 | 2.70541 |
| 1 | Queens | Briarwood | 671 | 24 | 107 | 40.710935 | -73.811748 | 53877 | 0.044546 | 1.245429 | 2.98974 |
| 2 | Brooklyn | Brighton Beach | 306 | 17 | 38 | 40.576825 | -73.965094 | 35547 | 0.047824 | 0.860832 | 2.96039 |
| 3 | Brooklyn | Brownsville | 600 | 28 | 44 | 40.663950 | -73.910235 | 58300 | 0.048027 | 1.029160 | 2.59078 |
| 4 | Brooklyn | Bushwick | 324 | 16 | 46 | 40.698116 | -73.925258 | 129239 | 0.012380 | 0.250698 | 1.80722 |

# 4. Neighborhood Clustering Analysis

We will use K-means to cluster NYC neighborhoods.

First we select features for clustering: ICU Bed Per Hundred People and Bed Per Hundred People and COVID_CASE_RATE (per hundred). Data such as Population is already included in other variables and thus, I chose to leave it out. We then normalize the data since K-Means algorithm requires standardized dataset to calculate distances.

Then we use elbow method to find the optimum number of clusters and the output optimum number of k is 7.




```
Optimal number of components is:
7
```

### Examine clusters

```
nyc_final[(nyc_final['Cluster Labels'] == 0)]
```

| | Cluster Labels | Borough | Neighborhood | Bed Number | ICU Bed Number | Unnamed: 0 | Latitude | Longitude | Population | ICU Bed Per Hundred People | Bed Per Hundred People | COVID_CASE_RATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 0 | Brooklyn | Bushwick | 324 | 16 | 46 | 40.698116 | -73.925258 | 129239 | 0.012380 | 0.250698 | 1.80722 |
| 6 | 0 | Manhattan | Chelsea | 212 | 12 | 76 | 40.744035 | -74.003116 | 47325 | 0.025357 | 0.447966 | 1.78245 |
| 7 | 0 | Manhattan | Chinatown | 180 | 13 | 64 | 40.715618 | -73.994279 | 47844 | 0.027172 | 0.376223 | 1.60119 |
| 8 | 0 | Brooklyn | Crown Heights | 287 | 13 | 40 | 40.670829 | -73.943291 | 143000 | 0.009091 | 0.200699 | 2.36551 |
| 13 | 0 | Queens | Forest Hills | 312 | 28 | 89 | 40.725264 | -73.844475 | 83728 | 0.033442 | 0.372635 | 2.16008 |
| 14 | 0 | Brooklyn | Fort Greene | 598 | 31 | 49 | 40.688527 | -73.972906 | 28335 | 0.109405 | 2.110464 | 1.80855 |
| 15 | 0 | Brooklyn | Gravesend | 371 | 22 | 37 | 40.595260 | -73.973471 | 29436 | 0.074738 | 1.260361 | 1.80214 |
| 16 | 0 | Manhattan | Inwood | 196 | 6 | 66 | 40.867684 | -73.921210 | 58946 | 0.010179 | 0.332508 | 1.87818 |
| 26 | 0 | Queens | Ridgewood | 348 | 12 | 95 | 40.708323 | -73.901435 | 69317 | 0.017312 | 0.502041 | 2.29379 |
| 29 | 0 | Brooklyn | Sunset Park | 364 | 24 | 35 | 40.645103 | -74.010316 | 126000 | 0.019048 | 0.288889 | 1.81818 |
| 30 | 0 | Manhattan | Upper East Side | 632 | 15 | 70 | 40.775639 | -73.960508 | 124231 | 0.012074 | 0.508730 | 1.38991 |
| 31 | 0 | Brooklyn | Williamsburg | 69 | 0 | 45 | 40.707144 | -73.958115 | 78700 | 0.000000 | 0.087675 | 2.05622 |

```
nyc_final[(nyc_final['Cluster Labels'] == 1)]
```

| | Cluster Labels | Borough | Neighborhood | Bed Number | ICU Bed Number | Unnamed: 0 | Latitude | Longitude | Population | ICU Bed Per Hundred People | Bed Per Hundred People | COVID_CASE_RATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 1 | Manhattan | Murray Hill | 2270 | 221 | 75 | 40.748303 | -73.978332 | 10864 | 2.034242 | 20.894698 | 1.5193 |

```
nyc_final[(nyc_final['Cluster Labels'] == 2)]
```

| | Cluster Labels | Borough | Neighborhood | Bed Number | ICU Bed Number | Unnamed: 0 | Latitude | Longitude | Population | ICU Bed Per Hundred People | Bed Per Hundred People | COVID_CASE_RATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2 | Brooklyn | Bensonhurst | 204 | 8 | 34 | 40.611009 | -73.995180 | 151705 | 0.005273 | 0.134472 | 2.70541 |
| 1 | 2 | Queens | Briarwood | 671 | 24 | 107 | 40.710935 | -73.811748 | 53877 | 0.044546 | 1.245429 | 2.98974 |
| 2 | 2 | Brooklyn | Brighton Beach | 306 | 17 | 38 | 40.576825 | -73.965094 | 35547 | 0.047824 | 0.860832 | 2.96039 |
| 3 | 2 | Brooklyn | Brownsville | 600 | 28 | 44 | 40.663950 | -73.910235 | 58300 | 0.048027 | 1.029160 | 2.59078 |
| 19 | 2 | Bronx | Morrisania | 170 | 0 | 21 | 40.823592 | -73.901506 | 16863 | 0.000000 | 1.008124 | 2.93872 |
| 23 | 2 | Bronx | Pelham Parkway | 421 | 22 | 9 | 40.857413 | -73.854756 | 30073 | 0.073155 | 1.399927 | 3.11771 |
| 27 | 2 | Queens | South Ozone Park | 247 | 11 | 99 | 40.668550 | -73.809865 | 75878 | 0.014497 | 0.325523 | 2.52710 |
| 33 | 2 | Bronx | Woodlawn | 321 | 16 | 5 | 40.898273 | -73.867315 | 42483 | 0.037662 | 0.755596 | 3.40335 |

```
: nyc_final[(nyc_final['Cluster Labels'] == 3)]
```

:

| | Cluster Labels | Borough | Neighborhood | Bed Number | ICU Bed Number | Unnamed: 0 | Latitude | Longitude | Population | ICU Bed Per Hundred People | Bed Per Hundred People | COVID_CASE_RATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 3 | Brooklyn | Carroll Gardens | 535 | 29 | 48 | 40.680540 | -73.994654 | 12853 | 0.225628 | 4.162452 | 1.23531 |
| 10 | 3 | Manhattan | East Village | 1296 | 49 | 78 | 40.727847 | -73.982226 | 63347 | 0.077352 | 2.045874 | 0.95722 |
| 32 | 3 | Brooklyn | Windsor Terrace | 839 | 40 | 42 | 40.656946 | -73.980073 | 20988 | 0.190585 | 3.997522 | 0.95807 |
| 34 | 3 | Manhattan | Yorkville | 1438 | 103 | 71 | 40.775930 | -73.947118 | 35221 | 0.292439 | 4.082792 | 1.05381 |

```
nyc_final[(nyc_final['Cluster Labels'] == 4)]
```

| | Cluster Labels | Borough | Neighborhood | Bed Number | ICU Bed Number | Unnamed: 0 | Latitude | Longitude | Population | ICU Bed Per Hundred People | Bed Per Hundred People | COVID_CASE_RATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 18 | 4 | Bronx | Melrose | 1118 | 59 | 16 | 40.819754 | -73.909422 | 24913 | 0.236824 | 4.487617 | 3.51015 |
| 21 | 4 | Bronx | Norwood | 1169 | 80 | 6 | 40.877224 | -73.879391 | 40494 | 0.197560 | 2.886847 | 3.68015 |
| 28 | 4 | Bronx | Spuyten Duyvil | 306 | 20 | 27 | 40.881395 | -73.917190 | 10279 | 0.194571 | 2.976943 | 3.07488 |

```
nyc_final[(nyc_final['Cluster Labels'] == 5)]
```

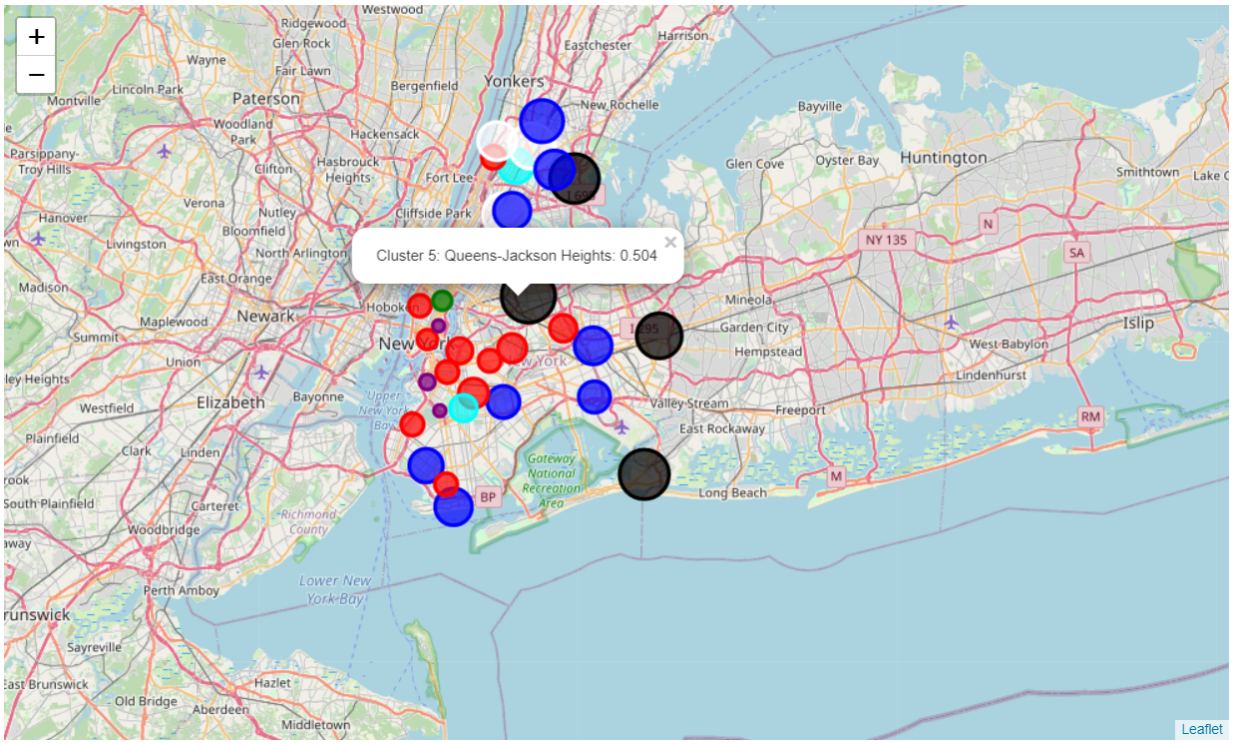
| | Cluster Labels | Borough | Neighborhood | Bed Number | ICU Bed Number | Unnamed: 0 | Latitude | Longitude | Population | ICU Bed Per Hundred People | Bed Per Hundred People | COVID_CASE_RATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | 5 | Queens | Far Rockaway | 257 | 8 | 116 | 40.603134 | -73.754980 | 60035 | 0.013326 | 0.428084 | 3.96702 |
| 17 | 5 | Queens | Jackson Heights | 545 | 20 | 85 | 40.751981 | -73.882821 | 108152 | 0.018492 | 0.503920 | 4.30115 |
| 22 | 5 | Bronx | Pelham Bay | 225 | 0 | 28 | 40.850641 | -73.832074 | 11931 | 0.000000 | 1.885844 | 3.93496 |
| 25 | 5 | Queens | Queens Village | 25 | 0 | 109 | 40.718893 | -73.738715 | 52504 | 0.000000 | 0.047615 | 3.62179 |

```
nyc_final[(nyc_final['Cluster Labels'] == 6)]
```

| | Cluster Labels | Borough | Neighborhood | Bed Number | ICU Bed Number | Unnamed: 0 | Latitude | Longitude | Population | ICU Bed Per Hundred People | Bed Per Hundred People | COVID_CASE_RATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 6 | Manhattan | East Harlem | 3906 | 250 | 69 | 40.792249 | -73.944182 | 115921 | 0.215664 | 3.369536 | 2.46317 |
| 12 | 6 | Bronx | Fordham | 1029 | 70 | 13 | 40.860997 | -73.896427 | 43394 | 0.161313 | 2.371296 | 2.74438 |
| 24 | 6 | Brooklyn | Prospect Lefferts Gardens | 2080 | 197 | 60 | 40.658420 | -73.954899 | 99287 | 0.198415 | 2.094937 | 2.14862 |

## Visualization of clusters

Finally, I visualized the clusters on the New York map with the radius corresponding to the Covid_case_rate of each neighborhood.




# Discussion

Looking into the results, we find that the algorithm classifies the neighborhoods into 7 clusters with extensive logics as follow:

| | ICU | Bed | COVID rate |
|---|---|---|---|
| 1 | low | low | low |
| 2 | high | high | low |
| 3 | low | medium | medium |
| 4 | low | high | low |
| 5 | low | high | high |
| 6 | low | low | high |
| 7 | low | high | medium |

However, even though the rate of group 1 is low, the rate of beds over the rate of cases (less than 1 in most cases) may indicate that group 1 can not meet the demand at some points. Thus, group 2, group 4 and group 7 may provide better service availability. Among them, group 2, Murray Hill of Manhattan has the highest availability.

*Limitations*

- The hospital beds data we collected from New York State Department of Health may not include the latest information.
- The NYC population data we collected from Wikipedia pages are from 2010, that is not very accurate.

- The COVID dataset initially contains an extensive neighborhood data. During the process, I essentially reduced this information to match with the neighborhood data from the hospital dataset using a strict approach. Some data was lost at this point. If there is a way to preserve this data, we can conduct a more extensive clustering experiment.
- In this project, I used K-mean clustering. The result was reasonable. However, there can be other methods to cluster the dataset that worth trying.

# Conclusion

This project aims at clustering the neighborhood in New York city based on the data of the ICU and hospital beds and COVID rate case of the neighborhoods. To do so, collecting, cleaning, transforming and processing the data was necessary. In the end, we identify 7 clusters of neighborhoods.