Business problem

Right now, New York is one of the worst hit state by [COVID-19](https://en.wikipedia.org/wiki/Coronavirus_disease_2019) in USA. New York city is at the center of the disaster. The hospitals are already stretched with patients overflowing. As on 9th June 2020, death toll was 17127, case count topped 207K.

I have motivated by this to create something useful which would give some insight on this situation. In this project I am going to determine which neighborhood is best prepared for this pandemic, by finding out the best ratio of hospital beds per person for each neighborhood in this city.

Data

For this project, I collected data from following sources:

Data source: [NYC data set](https://cocl.us/new_york_dataset) - New York City data that contains **borough**, **neighborhoods** along with their **latitudes** and **longitudes**.

To know the neighborhood population I will be using Wikipedia page of [NYC neighborhood](https://en.wikipedia.org/wiki/neighborhoods_in_New_York_City). Will collect the data from Scraping Wikipedia. data will be cleaned up and used to create a data frame containing borough, neighborhood and population.

Next, Hospitals per neighborhood information will be collected from foursquare API (using categoryid 4bf58dd8d48988d196941735) . I will collect bed and ICU capacity information from [NYS Health Profile website](https://profiles.health.ny.gov/). This data will give an insight which neighborhoods are is best prepared for this pandemic, by finding out the best ratio of hospital beds per person for each neighborhood in this city.

Approach

This was my approach to resolve issue:

* Collected the New York city data from [here](https://cocl.us/new_york_dataset).
* Collected population data for each neighborhood by scraping Wikipedia.
* Using Foursquare API got hospitals for each neighborhood.
* Collected hospital bed data by scraping data from **NYS Health Profile**.
* Data Visualization and some statistical analysis.
* Analyzing using Clustering (Specially K-Means).
* Find the best value of K
* Visualize the neighborhood max density of hospital beds per 100 people.
* Visualize the neighborhood max density of hospital ICU beds per 100 people.
* Inference From these results and related conclusions.

Data preparation

Data used in the analysis are listed below:

* First, got the json data from [here](https://cocl.us/new_york_dataset), which contain borough, neighborhood, latitude and longitude information.
* Neighborhood data in New York City was collected from scraping the [Wikipedia](https://en.wikipedia.org/wiki/neighborhoods_in_New_York_City) page.
* Then data was cleaned up and used to create a data frame containing borough, neighborhood and population.
* Used foursquare API to get neighborhood information.
* Collected Bed and ICU information [NYS Health Profile website](https://profiles.health.ny.gov/). Used selenium based scraping as this is a dynamic site and generated local copy and pushed to Github.

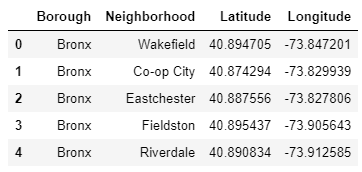
Source code

You can find the complete source in **[github](https://github.com/anandshaw214-source/appliedDataScience_Assignments/blob/Neighborhoods/The%20Battle%20of%20Neighborhoods%20-%20NYC%20Bed%20vs%20ICU.ipynb)** or in Watson Studio.

Methodology

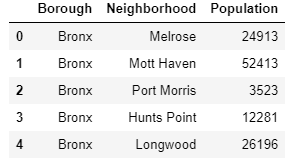
Step 1: Get New York city data with latitude and longitude

I used **requests** to get the json data from [nyc dataset](https://cocl.us/new_york_dataset) and stored it in a data frame.



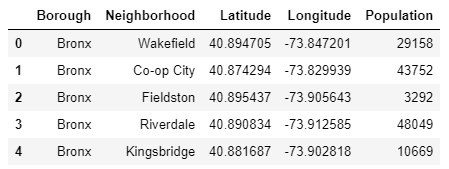
Step 2: Get New York city data with population

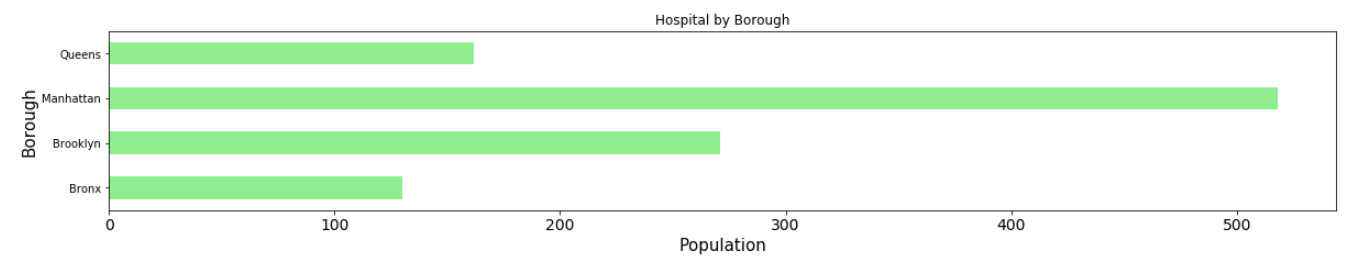
Used **BeautifulSoup** to scrape boroughs from Wikipedia. Then I have collected every link given in neighborhood column of the table. From each link, we can run iteration via requests to visit those Wikipedia pages, and scrap population data from table.



Step 3: Combine step one and step two

Combined data frames from previous steps into one based on “neighborhood” and “borough”:





Step 4: Get hospital data from Foursquare

After collecting population data, now it is time to collect the hospital data. I have used the **Foursquare** API to fetch hospital data for latitude and longitude of each neighborhood from the previous dataset.

Step 5: Get hospital Bed data from NYS Health Profile

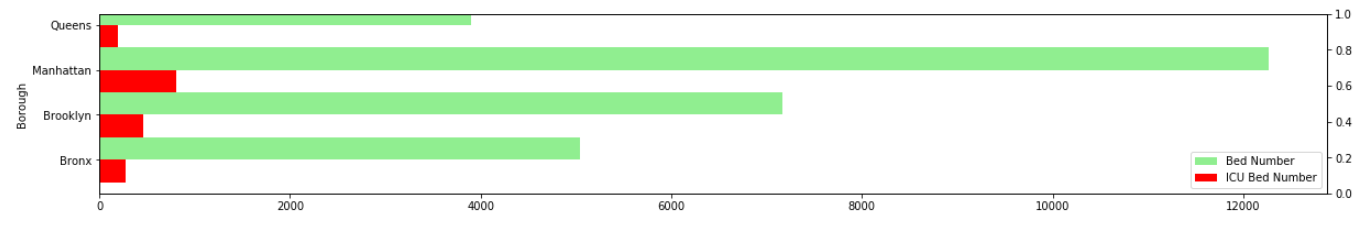
Collected hospital bed related data from [NYS Health Profile website](https://profiles.health.ny.gov/). Scraped data by using [**Selenium**](https://www.selenium.dev/) with **BeautifulSoap**. I have collected the IDs of hospitals in NYC manually, and based on those IDs, we have scraped data from **NYS Health Profile website**. The data frame looks like this:



Step 6: combine step four and step five

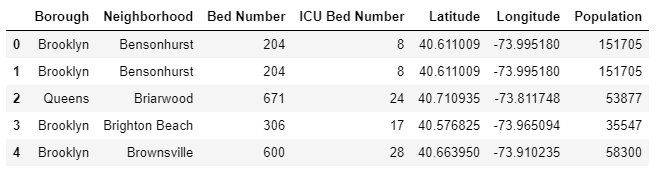
Combined the data from step four and step five. Cleaned up the data a little bit and sum up Bed count and ICU bed count grouping by “neighborhood” and “borough”:

Here is a bar chart of “bed count and ICU count” per “borough”:



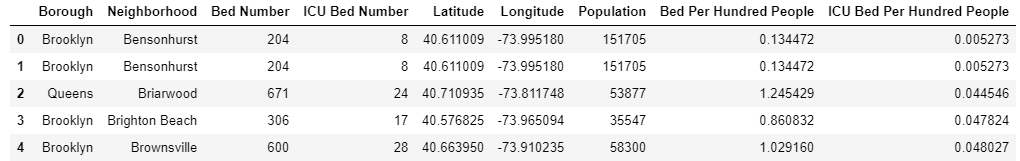
Step 7: Combine data from step three and step six

Combined data from step three and step six, combined the population data with hospital bed count data. After that merged two data frames based on “neighborhood” and “borough”. New data frame looks like this:



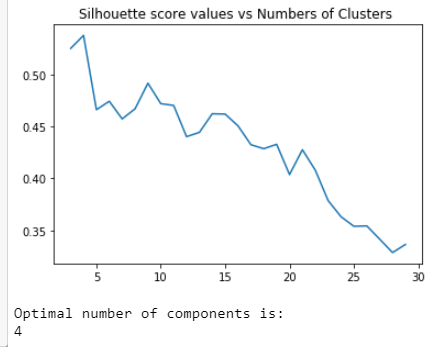
Step 8: Add Bed and ICU per hundred people to data frame

Calculated bed per hundred people based on two rows: Population and Bed Number. Then added this to the data frame. Similarly, added ICU data to data frame:



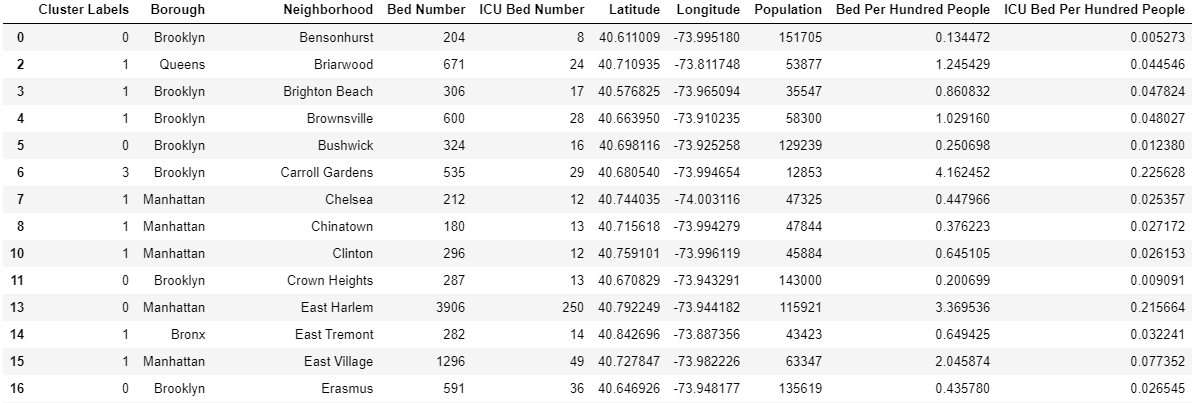
Step 9: K-means clustering

Used k-means clustering to partition the data into **k** groups. Have used [**elbow method**](https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/) to find the optimal number of **k**. The “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. In the visualizer “elbow”, value of **k** is 4.



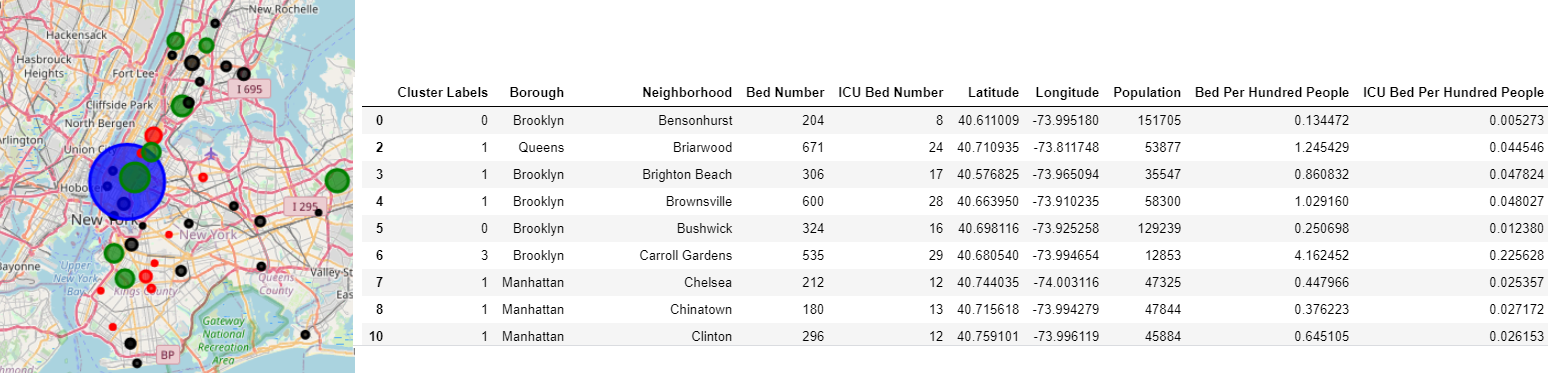
Step 10: Merge cluster labels with dataset

After that, merged the cluster labels of groups with data frames. The data frame looks like this:



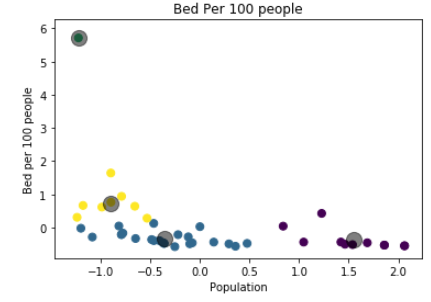
Step 11: Visualize with Folium

For Visualization, I used [**folium**](https://python-visualization.github.io/folium/) to visualize the distribution. The first map illustrates the clusters where the radius of the Circle marker is proportional to hospital beds per hundred people.



Step 12: Scatter plot

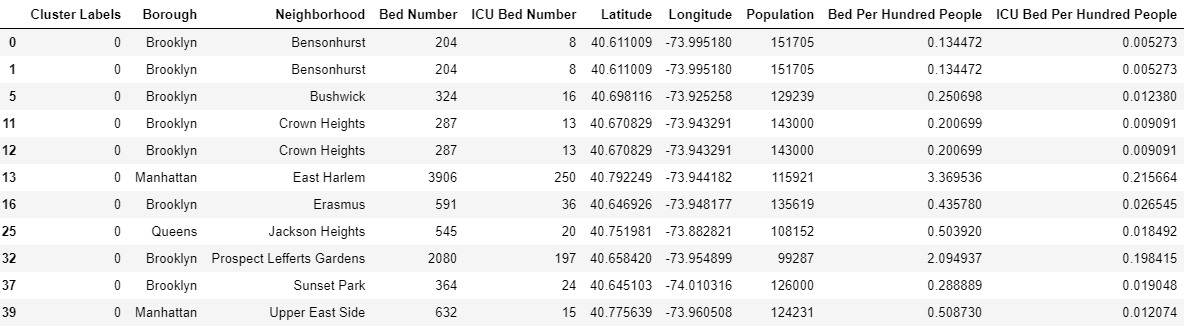
Plotted the scatter plots of my data and define my clusters with colors. The grey circle marker is representing the centroid of each cluster. Don’t forget that the data is normalized, so the axes do not deliver real values.



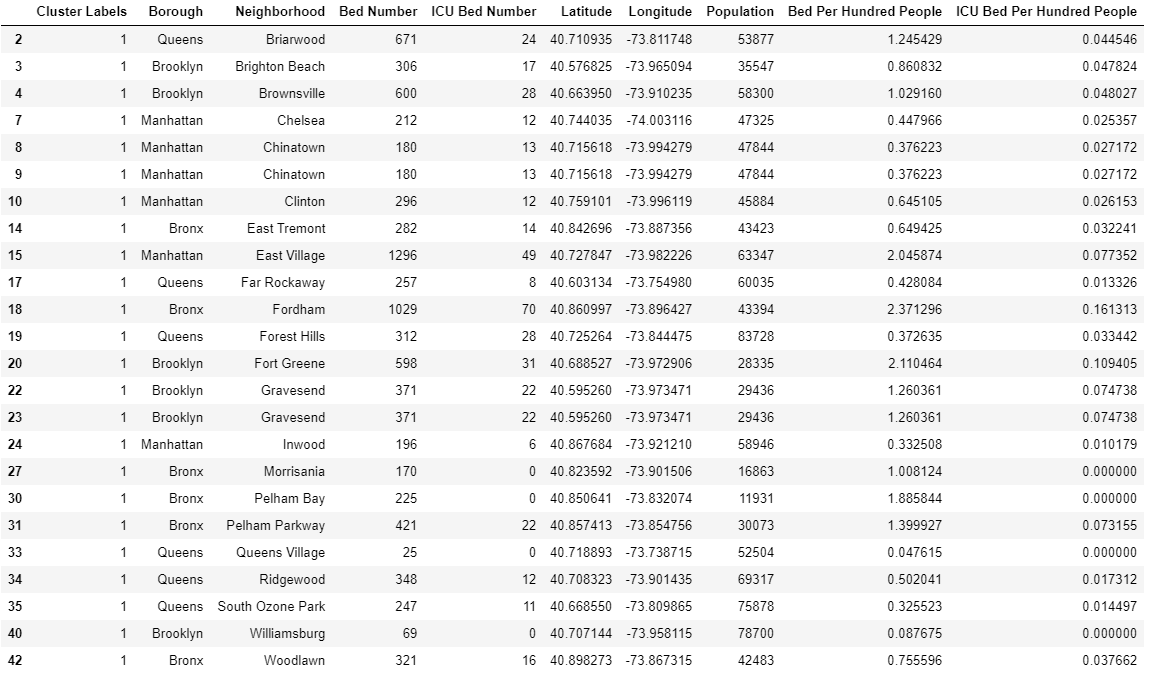
Step 13: see which borough goes to which cluster

Lets see which boroughs belong to which clusters.

Here is the dataset for *cluster 0*:



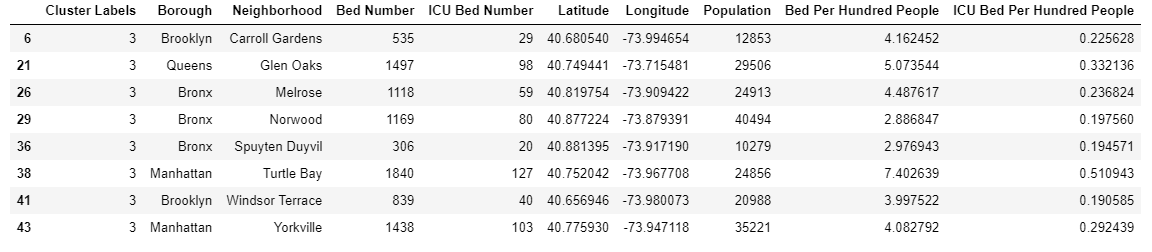
Here is the dataset for *cluster 1*:



Here is the dataset for *cluster 2*:



Here is the dataset for *cluster 3*:



Step 14: Check neighborhoods without hospital

Checked the neighborhoods with vs without hospital count:



Results and discussion

During the analysis, four clusters were defined. One cluster(cluster 2), that consists of only one area, has been defined as the outsider, due to the high number of hospital beds, which means it is better equipped to handle this pandemic. Two other groups were clustered according to bed per hundred people and ICU bed per hundred people. It is obvious that the cluster with the lowest beds per person is the place where we should concentrate on providing beds and other equipment(Cluster 0). We also should look into conditions in Queens Village and Williamsburg as they have very low beds per hundred people. Furthermore, in hundred other neighborhoods, there is no hospital data. Hence, people living there are at high risk of not being treated during pandemic.

**What could be done better**

Foursquare doesn’t represent the full picture, since many hospitals are not on the list. After data merge and cleansing, compromised lot of data. For that reason, other maps could be utilized such as Google map or OpenStreet map.

**NYS Health Profile website** might lacks the latest information regarding hospital information. It could lack information regarding new hospitals. Also, hospital IDs I have extracted manually from NYS, which could have missing hospitals. I also dropped neighborhoods which did not have any hospital data matching in **NYS Health Profile website**.

I have used fuzzywuzzy to match hospital data from Foursquare and NYS Health Profile. It is not a correct measure because we are matching the names nearest possible, it could be wrong in real life scenario.

I have only considered hospital data. Did not consider other medical facilities like nursing home or health clinic.

Lastly, to battle COVID-19, we should have had patient data for the neighborhood. Unfortunately, I could not enough research to get it like this(for example, get patient per latitude longitude) from any source, hence could not incorporate it.

Conclusion

To conclude, the basic data analysis was performed to identify the most well equipped hospital in the NYC neighborhoods. During the analysis, several important statistical features of the boroughs/neighborhoods were explored and visualized. Furthermore, clustering helped to highlight the group of optimal areas. Finally, **Manhattan-Murray Hill** is indicating as the most well equipped(as per hospital Bed count and ICU bed count) area to battle pandemic. This may not be true as data set was very limited. With better data set the conclusion will be better.