Title: Determining the most effective neighbourhood to fight COVID-19 in New York City, USA

Introduction and Motivation:

New York is the worst hit city in the world, in terms of number of active cases and deaths due to the prevailing coronavirus pandemic in the world. The population density and challenges of law enforcement make the general population even more vulnerable to the deadly virus. I have undertaken this project to try coming up with a neighbourhood within New York that could be effective in combating the COVID-19 virus. The idea is to deploy the skillset I have acquired during the IBM Professional Certificate in order to tackle the epicentre of a major global crisis.

Business Problem:

As of now, it is clear that the state of New York is the worst hit by coronavirus in the world. To put things in perspective, currently, the state of New York has a whopping 295,000 confirmed cases and 17638 deaths, a significant proportion of which (12509), have happened in New York City. Therefore, as a matter of fact, it is critical to curtail the spread of the virus in New York City. This could bring the death toll down in New York State as well as the USA, which is the new COVID-19 hotspot of the world.

The problem gets compounded by the fact that the New York Medical Infrastructure is under immense strain due to the lack of hospital beds and the ever increasing number of patients requiring hospitalization. As a start, through this project, I intend to determine or identify the neighbourhood that is the best prepared to fight the pandemic.

This is proposed to be done by calculating the highest ratio of hospital beds per person in each neighbourhood of New York City. This would in turn give deeper insights into managing the limited and dwindling resources and also Personal Protective Equipment (PPEs) which are critical for the safety of the brave doctors and nurses. The resources of the neighbourhood with relatively better capacity could be deployed in the worst affected neighborhoods, thereby focusing efforts and attention towards the key COVID-19 hotspots.

Data Collection Plan:

The data shall be acquired/fetched from following sources:

- *New York City data that contains borough, neighborhoods along with their latitudes and longitudes (Data source: NYC data set) https://cocl.us/new_york_dataset*
- *The population data will be obtained from Scraping Wikipedia (Data source: Wikipedia page of NYC neighbourhood)*
- *Each of the links of neighbourhood will be accessed to obtain the population figures for each of them*
- *The Hospital data (location coordinates etc.) shall be fetched from foursquare API. Data source: Foursquare API*
- *Hospital bed information shall be fetched from NYS Health Profile website. Data source: https://profiles.health.ny.gov/*

Brief Methodology:

- Collect the New York city data from here
- Collect population data for each neighbourhood by scraping Wikipedia, using the 'Beautiful Soup Package'
- Using Foursquare API, generate the hospitals data (ID, Venue and location coordinates)for each neighbourhood.
- Collect hospital bed data by scraping data from **NYS Health Profile**.
- Perform Data Visualization on the Boroughs and Neighbourhoods and some statistical analysis to identify details such as the no. of neighbourhoods per Borough and the population per Borough
- Analyse the neighbourhoods using Clustering (Specially K-Means)
- Find the best value of K through the elbow-method (Inflection point of the graph)
- Visualize the neighbourhood which has the max density of hospital beds per 100 people, using the folium library
- Similarly, visualize the neighbourhood with max density of hospital ICU beds per 100 people.
- Examine the clusters and draw relevant inferences from these results and related conclusions

Data preparation

Data used in the analysis are listed below:

- First, get the json data from https://cocl.us/new_york_dataset, which will contain borough, neighbourhood, latitude and longitude information.
- Neighbourhood data in New York City will be collected from scraping the Wikipedia page. Links given in the neighbourhood section of the table will be visited via scraper, and the population statistics would be fetched for each of them. Then data will be cleaned up and used to create a data frame containing borough, neighbourhood and population.
- Hospitals per neighbourhood information will be collected from foursquare API URL call.
- Lastly, the 'bed and ICU capacity information from NYS Health Profile website will be obtained. Shall use selenium based scraping to adjust for the dynamic behaviour of the site

The source code can be obtained from github repository link shared on coursera

Detailed Methodology:

- ***Step one: New York city data with latitude and longitude***

Use requests to get the json data from nyc dataset and store it in a data frame.

```
ny_df.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

- ***Step two: New York city data with population***

Then use **'BeautifulSoup'** library to scrape boroughs from Wikipedia. Then collect every link given in neighbourhood column of the table. From each link, iterate via requests to visit those Wikipedia pages, and scrap population data from right hand side table.

```
nyc_population_df.head()
```

	Borough	Neighborhood	Population
0	Bronx	Melrose	24913
25	Bronx	Bruckner	38557
26	Bronx	Castle Hill	38557
27	Bronx	Clason Point	9136
28	Bronx	Harding Park	9136

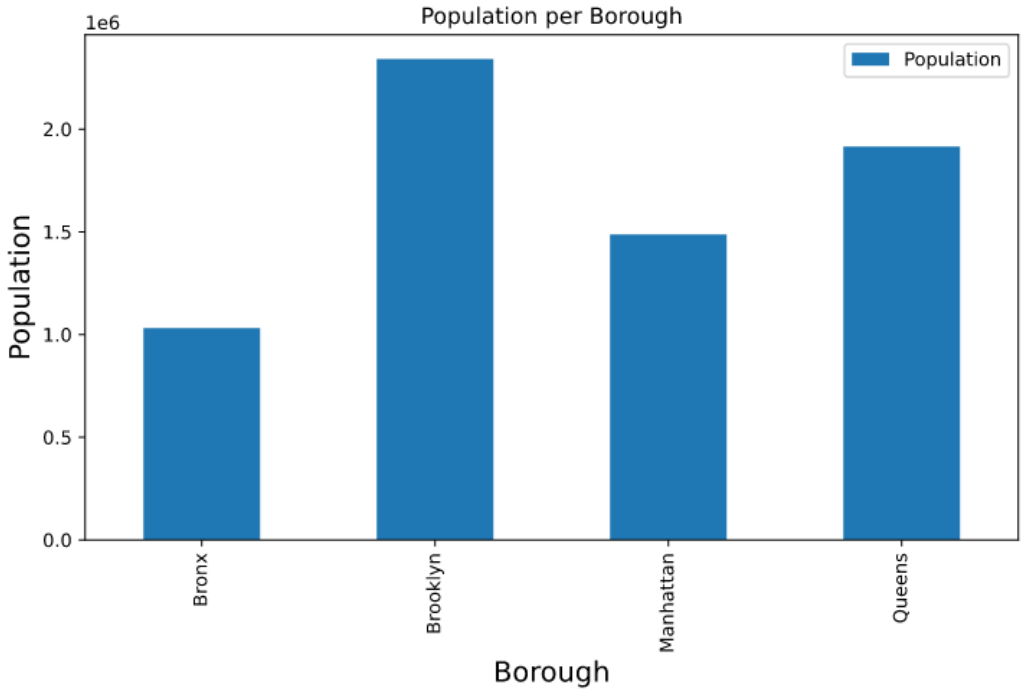
- ***Step three: Combine steps one and two***

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

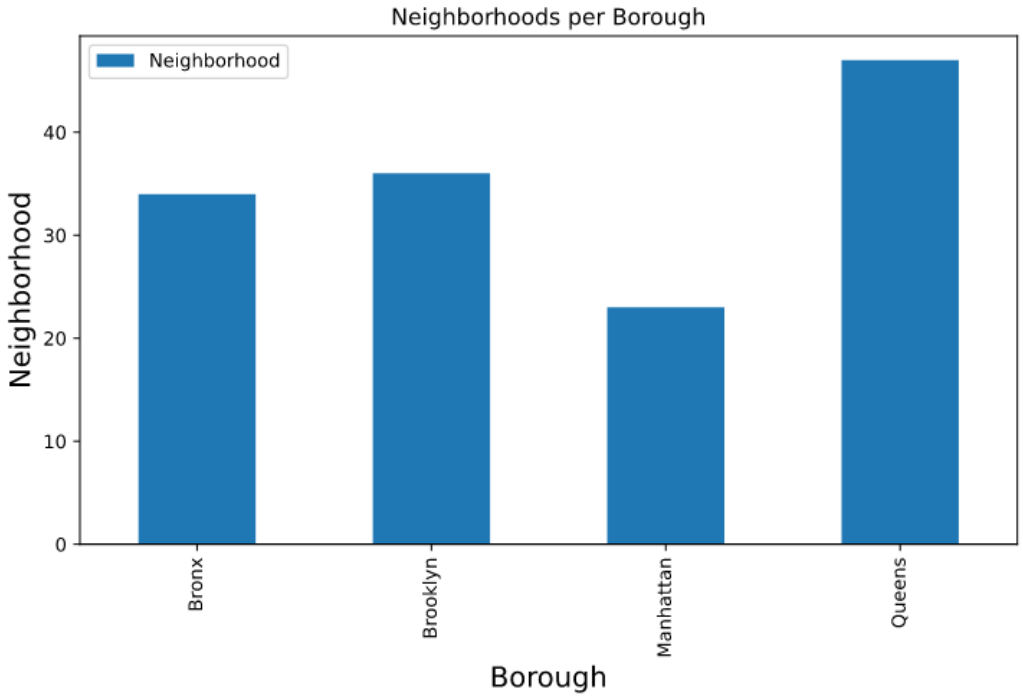
```
ny_df.set_index('Neighborhood')
nyc_population_df.set_index('Neighborhood')
nyc_df = pd.merge(ny_df, nyc_population_df, how="inner", on=["Borough", "Neighborhood"])
nyc_df.head()
```

	Borough	Neighborhood	Latitude	Longitude	Population
0	Bronx	Wakefield	40.894705	-73.847201	29158
1	Bronx	Co-op City	40.874294	-73.829939	43752
2	Bronx	Fieldston	40.895437	-73.905643	3292
3	Bronx	Riverdale	40.890834	-73.912585	48049
4	Bronx	Kingsbridge	40.881687	-73.902818	10669

Plotted the Normalized population for each Borough



Plotted the Number of Neighbourhoods for each Borough



- ***Step four: collect hospital data from Foursquare***

  After collecting population data, collect the hospital data. Here, I used the Foursquare API to fetch hospital data for latitude and longitude of each neighborhood from the previous dataset.

```
hospital_df = get_hospital_per_neighborhood_borough(nyc_df)
hospital_df.head()
```

```
Output was trimmed for performance reasons. To see the full output set the setting
"python.dataScience.textOutputLimit" to 0. ...
```

	ID	Name	Latitude	Longitude	Borough	Neighborhood
0	59832a7bfe37406ea7eb3a79	Statcare Urgent & Walk-In Medical Care (Bronx ...	40.870168	-73.828404	Bronx	Co-op City
1	568e86f5498ec6df53771448	CityMD Baychester Urgent Care - Bronx	40.866795	-73.827051	Bronx	Co-op City
2	50173409e4b0cfe38c43abf4	wellcare	40.874247	-73.837745	Bronx	Co-op City
3	5158ddffe4b086af71ca90c7	The Mollie & Jack Ziklin Jewish Hospice Resid...	40.888478	-73.910047	Bronx	Fieldston
4	5158ddffe4b086af71ca90c7	The Mollie & Jack Ziklin Jewish Hospice Resid...	40.888478	-73.910047	Bronx	Riverdale

- ***Step five: collect hospital bed data from NYS Health Profile***

  Now, collect hospital bed related data from NYS Health Profile website. Scrap the data using Selenium with BeautifulSoup. Manually collect the IDs of hospitals in NYC, and based on those IDs, scrape data from NYS Health Profile website. The data frame looks as follows:

	Hospital Name	Bed Number	ICU Bed Number
0	Jamaica Hospital Medical Center	402	8
1	New York Community Hospital of Brooklyn, Inc	134	7
2	Mount Sinai Hospital	1134	85
3	Nassau University Medical Center	530	22
4	Richmond University Medical Center	448	20

- ***Step six: combine step four and step five***

  Merge dataframes from steps four and five. Internally, join the data frame based on "neighborhood" and "borough".

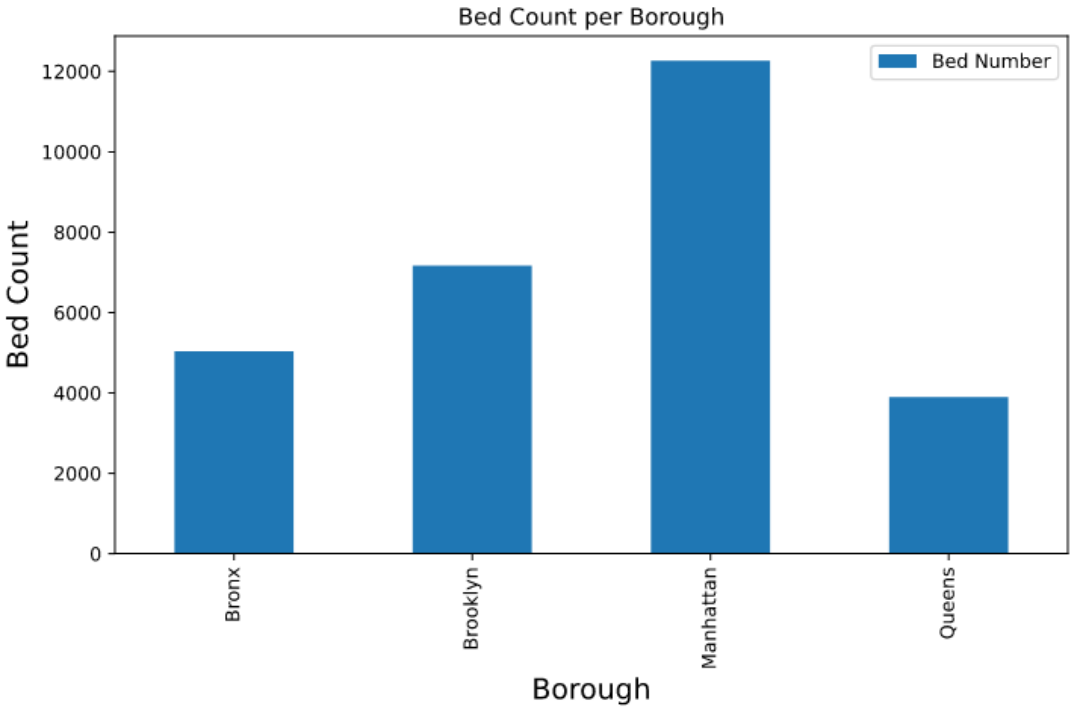
```
h_df = combine_hospital_beds_with_boro_neighborhood(hospital_bed_df, hospital_df)
h_df.head()
```

	Hospital Name	Bed Number	ICU Bed Number	Borough	Neighborhood
0	Jamaica Hospital Medical Center	402	8	Queens	Briarwood
1	New York Community Hospital of Brooklyn, Inc	134	7	Brooklyn	Fort Greene
2	Mount Sinai Hospital	1134	85	Manhattan	East Harlem
3	Nassau University Medical Center	530	22	Manhattan	Turtle Bay
4	Richmond University Medical Center	448	20	Manhattan	Turtle Bay

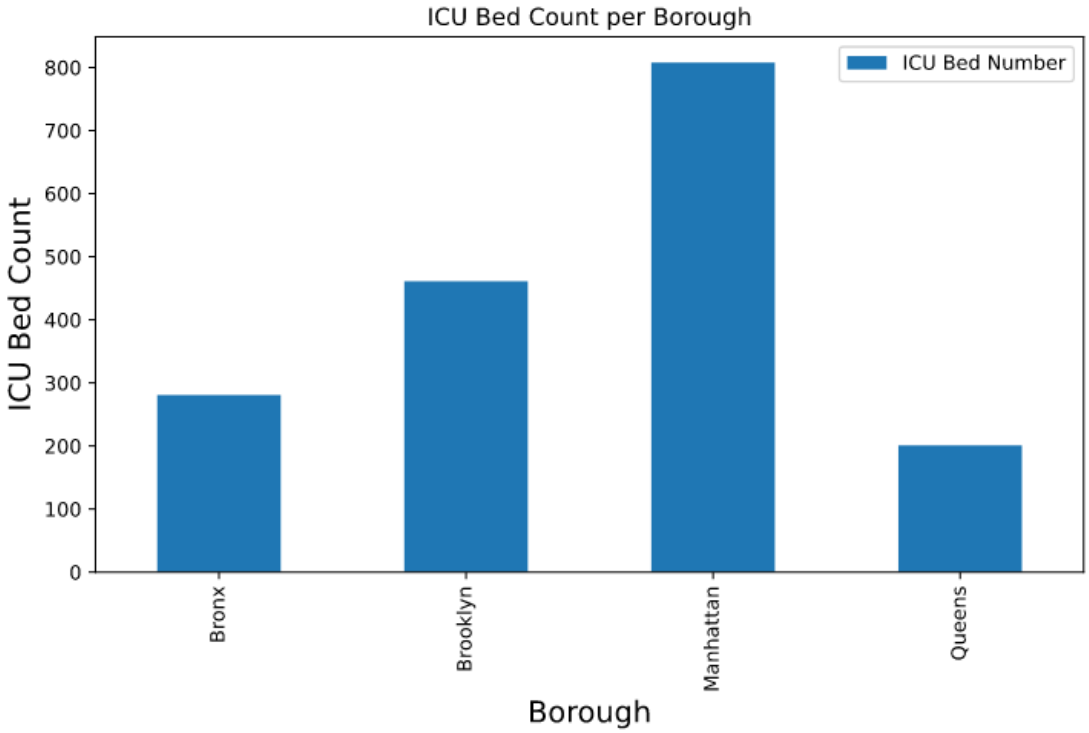
Clean up the data a little bit and sum up bed count and ICU- bed count, grouping by “neighborhood” and “borough”:

Neighborhood	Borough	Bed Number	ICU Bed Number
Bensonhurst	Brooklyn	204	8
Briarwood	Queens	671	24
Brighton Beach	Brooklyn	306	17
Brownsville	Brooklyn	600	28
Bushwick	Brooklyn	324	16

Plotted the Bed Count for each Borough



Plotted the ICU Bed Count for each Borough



- ***Step seven: combine data from step three and step six***
  Now, combine data from steps three and six. Basically, combine the population data with hospital bed count data. Merge two data frames based on "neighborhood" and "borough". New data frame looks as follows:

```
df = pd.merge(h_df, nyc_df, how="inner", on=["Borough", "Neighborhood"]
df.head()
```

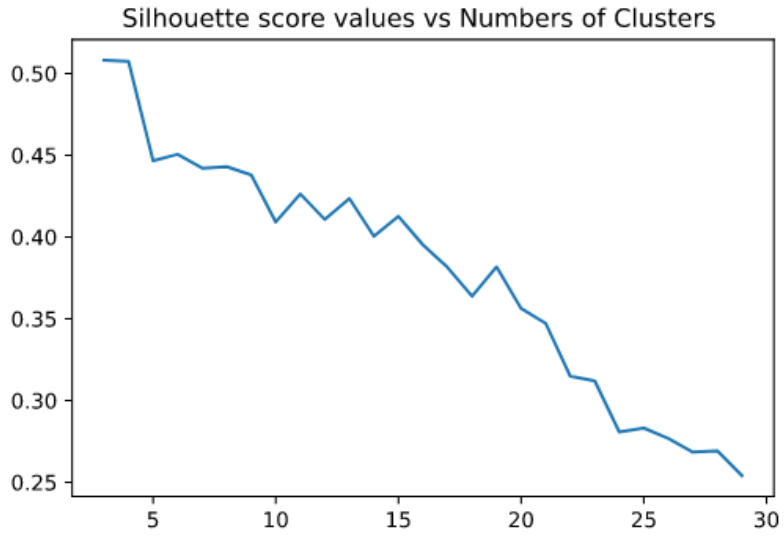
	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Populatior
0	Brooklyn	Bensonhurst	204	8	40.611009	-73.995180	151705
1	Queens	Briarwood	671	24	40.710935	-73.811748	53877
2	Brooklyn	Brighton Beach	306	17	40.576825	-73.965094	35547
3	Brooklyn	Brownsville	600	28	40.663950	-73.910235	58300
4	Brooklyn	Bushwick	324	16	40.698116	-73.925258	129239

- ***Step eight: add bed and ICU per hundred people to data frame***
  Now we are going to calculate bed per hundred people based on two rows: Population and Bed Number. Then add this to the data frame. Similarly, we are going to add ICU data to data frame:

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

- ***Step nine: K-means clustering***

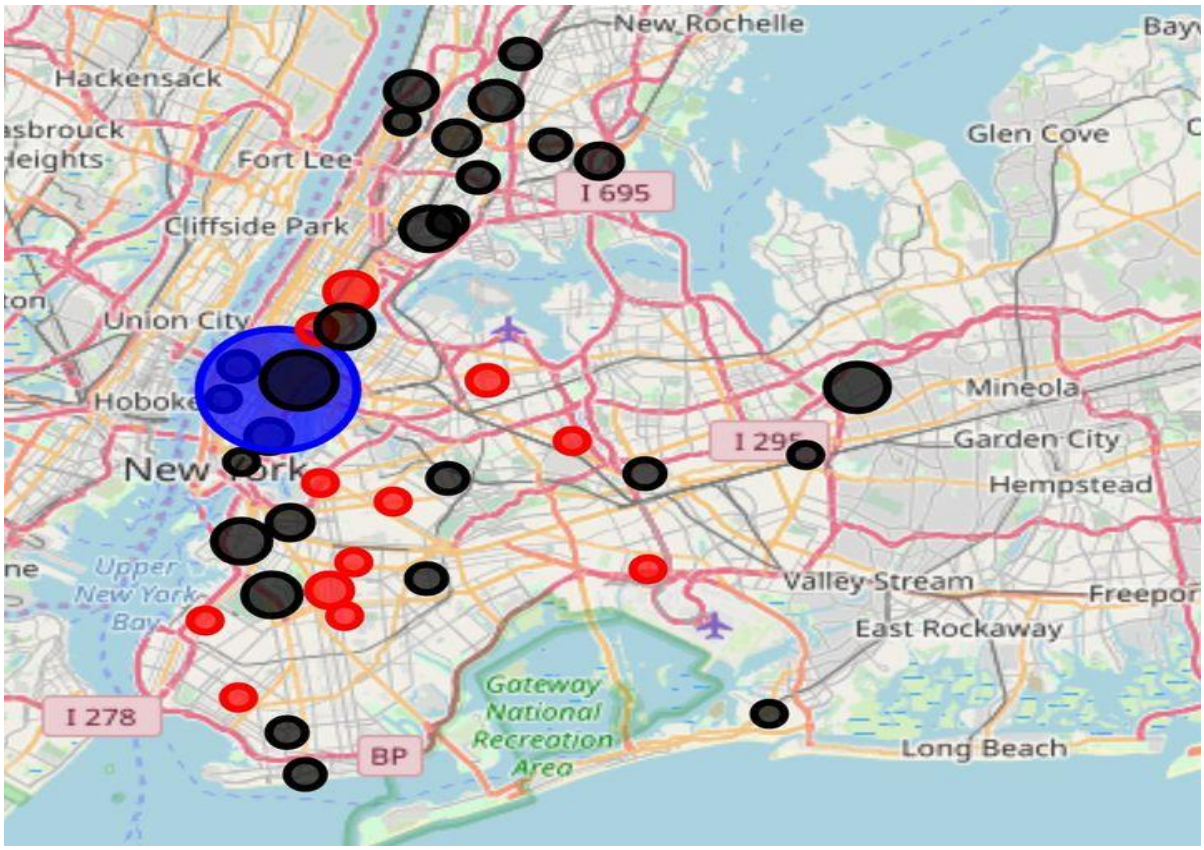
Now, use the k-means clustering algorithm to partition the data into 'k' groups. We use the elbow method to find the optimal value of k. The "elbow" (the point of inflection on the curve) is a reasonable indication that the underlying model fits best at that point. As seen in the visualizer "elbow", value of k is 3.



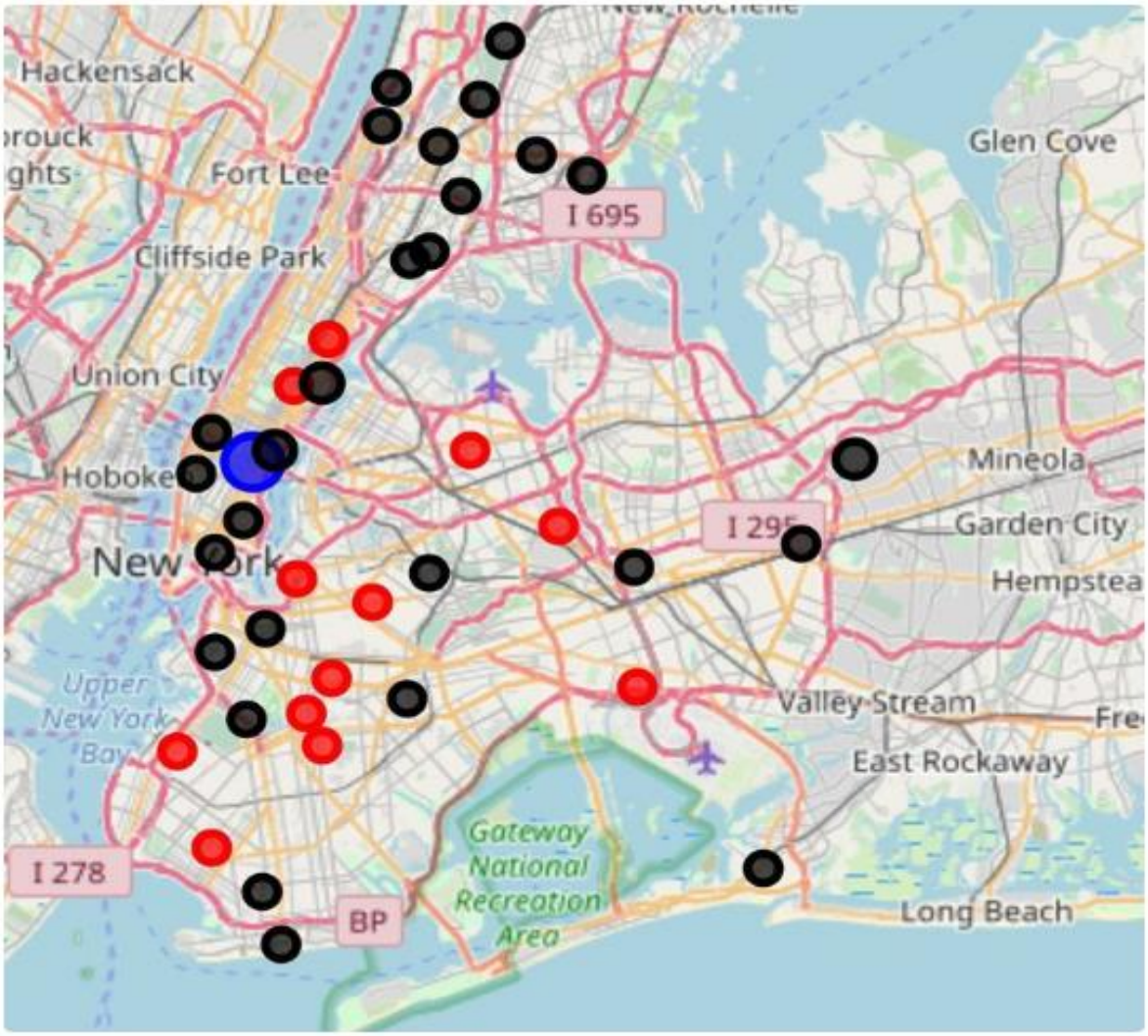
- ***Step ten: merge cluster labels with dataset***
  After that, we merge cluster labels of groups with data frames. The data frame looks as follows:

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

- ***Step eleven: visualize with folium***
  Now, we use folium to visualize the distribution. The first map illustrates the clusters where the radius of the Circle marker is directly proportional to hospital beds per hundred people



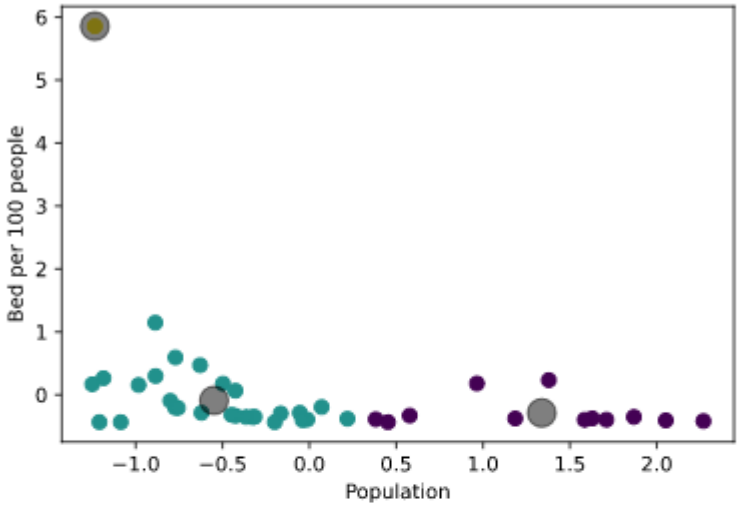
The second map illustrates the clusters where the radius of the Circle marker varies directly with the number of ICU beds per hundred people

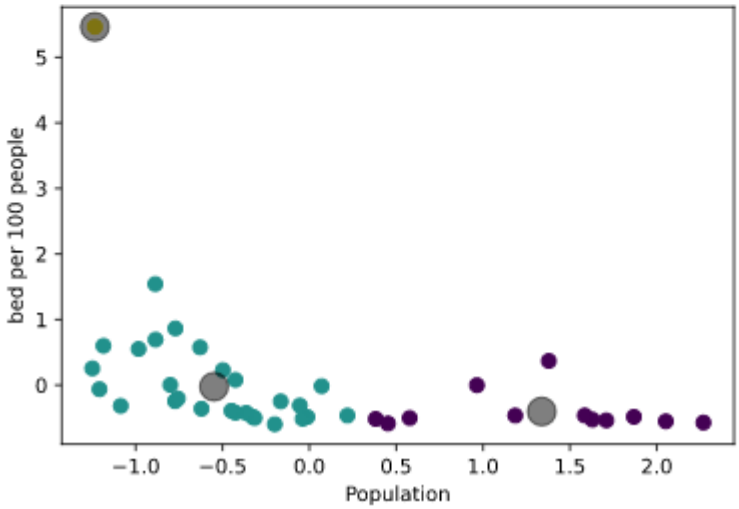


It is evident that one of the clusters (blue circle) consists of one borough - Manhattan.

- ***Step twelve: use scatter plot***

  Now, we look at the scatter plots of our data and define our clusters with colors. The grey circle marker represents the centroid of each cluster. Please note that our data is normalized, so the axes do not deliver real values.





We observe the obvious outlier here. This neighborhood has a high number of beds per people. From the maps above, we can easily say that it is Murray Hill.

- ***Step thirteen: Borough-Clustering***

  Identify which boroughs belong to which clusters.

  Below is the dataset for cluster 0:

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

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Hundred People	Bed Per Hundred People
0	0	Brooklyn	Bensonhurst	204	8	40.611009	-73.995180	151705	0.005273	0.134472
4	0	Brooklyn	Bushwick	324	16	40.698116	-73.925258	129239	0.012380	0.250698
9	0	Brooklyn	Crown Heights	287	13	40.670829	-73.943291	143000	0.009091	0.200699
10	0	Manhattan	East Harlem	3906	250	40.792249	-73.944182	115921	0.215664	3.369536
13	0	Brooklyn	Erasmus	591	36	40.646926	-73.948177	135619	0.026545	0.435780
16	0	Queens	Forest Hills	312	28	40.725264	-73.844475	83728	0.033442	0.372635
21	0	Queens	Jackson Heights	545	20	40.751981	-73.882821	108152	0.018492	0.503920
28	0	Brooklyn	Prospect Lefferts Gardens	2080	197	40.658420	-73.954899	99287	0.198415	2.094937
31	0	Queens	South Ozone Park	247	11	40.668550	-73.809865	75878	0.014497	0.325523
33	0	Brooklyn	Sunset Park	364	24	40.645103	-74.010316	126000	0.019048	0.288889
35	0	Manhattan	Upper East Side	632	15	40.775639	-73.960508	124231	0.012074	0.508730
36	0	Brooklyn	Williamsburg	69	0	40.707144	-73.958115	78700	0.000000	0.087675

Below is the dataset for cluster 1:

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

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Hundred People	Bed Per Hundred People
1	1	Queens	Briarwood	671	24	40.710935	-73.811748	53877	0.044546	1.245429
2	1	Brooklyn	Brighton Beach	306	17	40.576825	-73.965094	35547	0.047824	0.860832
3	1	Brooklyn	Brownsville	600	28	40.663950	-73.910235	58300	0.048027	1.029160
5	1	Brooklyn	Carroll Gardens	535	29	40.680540	-73.994654	12853	0.225628	4.162452
6	1	Manhattan	Chelsea	212	12	40.744035	-74.003116	47325	0.025357	0.447966
7	1	Manhattan	Chinatown	180	13	40.715618	-73.994279	47844	0.027172	0.376223
8	1	Manhattan	Clinton	296	12	40.759101	-73.996119	45884	0.026153	0.645105
11	1	Bronx	East Tremont	282	14	40.842696	-73.887356	43423	0.032241	0.649425
12	1	Manhattan	East Village	1296	49	40.727847	-73.982226	63347	0.077352	2.045874
14	1	Queens	Far Rockaway	257	8	40.603134	-73.754980	60035	0.013326	0.428084
15	1	Bronx	Fordham	1029	70	40.860997	-73.896427	43394	0.161313	2.371296
17	1	Brooklyn	Fort Greene	598	31	40.688527	-73.972906	28335	0.109405	2.110464
18	1	Queens	Glen Oaks	1497	98	40.749441	-73.715481	29506	0.332136	5.073544
19	1	Brooklyn	Gravesend	371	22	40.595260	-73.973471	29436	0.074738	1.260361
20	1	Manhattan	Inwood	196	6	40.867684	-73.921210	58946	0.010179	0.332508
22	1	Bronx	Melrose	1118	59	40.819754	-73.909422	24913	0.236824	4.487617
23	1	Bronx	Morrisania	170	0	40.823592	-73.901506	16863	0.000000	1.008124
25	1	Bronx	Norwood	1169	80	40.877224	-73.879391	40494	0.197560	2.886847
26	1	Bronx	Pelham Bay	225	0	40.850641	-73.832074	11931	0.000000	1.885844
27	1	Bronx	Pelham Parkway	421	22	40.857413	-73.854756	30073	0.073155	1.399927
29	1	Queens	Queens Village	25	0	40.718893	-73.738715	52504	0.000000	0.047615
30	1	Queens	Ridgewood	348	12	40.708323	-73.901435	69317	0.017312	0.502041
32	1	Bronx	Spuyten Duyvil	306	20	40.881395	-73.917190	10279	0.194571	2.976943
34	1	Manhattan	Turtle Bay	1840	127	40.752042	-73.967708	24856	0.510943	7.402639
37	1	Brooklyn	Windsor Terrace	839	40	40.656946	-73.980073	20988	0.190585	3.997522
38	1	Bronx	Woodlawn	321	16	40.898273	-73.867315	42483	0.037662	0.755596
39	1	Manhattan	Yorkville	1438	103	40.775930	-73.947118	35221	0.292439	4.082792

Below is the dataset for cluster 2:

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

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Hundred People	Bed Per Hundred People
24	2	Manhattan	Murray Hill	2270	221	40.748303	-73.978332	10864	2.034242	20.894698

- ***Step fourteen: see neighbourhoods without hospital***
  So far, we have analysed dataset for neighbourhoods with hospitals. Now, we can look into neighbourhoods without hospital data:

	Borough	Neighborhood
0	Bronx	Wakefield
1	Bronx	Co-op City
2	Bronx	Fieldston
3	Bronx	Riverdale
4	Bronx	Kingsbridge
7	Bronx	Williamsbridge
8	Bronx	Baychester
10	Bronx	Bedford Park
11	Bronx	University Heights
12	Bronx	Morris Heights

If we see the indexes of neighbourhoods with and without hospital, it should look like this:

```
Neighborhood without hospital count: 100
Neighborhood with hospital count: 40
```

We can see that there are 100 neighbourhoods without any hospital.

Results and discussion:

- During the analysis, three clusters have been identified. In particular, Cluster 2, consisting of only one area, has been defined as the outsider, due to the high number of hospital beds, implying that it is better equipped to handle this pandemic.

- The data was also clustered according to beds per hundred people and ICU beds per hundred people. The cluster with the lowest beds per person is the place where we should concentrate on providing beds and other medical equipment (Cluster 0).

- On the flip side, we also should look into conditions in Queens Village and Williamsburg as they have very low beds per hundred people. Furthermore, in hundred other neighbourhoods,

there is no hospital data. Hence, people living there are very vulnerable to infection and also fatality.

Potential Trade-Offs:

- Foursquare doesn't represent the full picture, since many hospitals are not on the list. For that reason, other maps could be utilized such as Google map or OpenStreet map. So, pairing the Foursquare API calls with other searching algorithms might help with the accuracy of the result.

- NYS Health Profile website might lack the latest hospital information. Also, since, hospital ids were extracted manually from NYS, some hospitals could have been missed on account of human/manual error. To maintain the integrity of the dataframe, I also dropped neighbourhoods which did not have any hospital data matching in NYS Health Profile website. For this project, I are only using data from 74 hospitals in NYC.

- I used fuzzy-wuzzy to match hospital data from Foursquare and NYS Health Profile. It might not give accurate results since the names may have changed and not captured in the NYS Health Profile.

- I am just considering hospital data and did not consider other medical facilities such as nursing home or health clinic etc.

- I used population data from 2010(as per Wikipedia sources), which is certainly inaccurate. Since, this is the latest data that is available online, I carried out the analysis using this.

- Finally, to battle COVID-19, I wished I had patient data for each neighborhood, for instance, number of patients per latitude/longitude. I couldn't find it online. Visualizations done this way, could have given much deeper insight into the spread of the pandemic- location-wise. Further, it could be updated LIVE and would have given a more realistic picture of the spread of the virus.

Conclusion:

To conclude, data analysis was performed to identify the most well equipped hospital in the NYC neighbourhoods. During the analysis, several important statistical features of the boroughs/neighbourhoods were explored and visualized. Furthermore, clustering helped to highlight the vulnerable areas and also pinpoint areas which are in a good position to tackle COVID virus. Finally, **Manhattan-Murray Hill** was chosen as the most well equipped (as per hospital bed count and ICU bed count) area to effectively battle the pandemic.