Find the Best Neighborhood in NYC To Settle in During and After the Pandemic

Geetanshu Grover

July 20th,2020

Introduction

Background

New York is one of the worst hit state by COVID-19 in USA. New York city was at the center of the disaster. The hospitals are already stretched thin with patients overflowing. According to New York Times report, (at the moment of writing) death toll was 22872, case count topped 226,104. I was motivated by this to create something useful which would give some insight on this situation determine which neighborhood is best equipped for this pandemic and any other health issues, by finding out the best ratio of hospital beds per person for each neighborhood in this city.

Problem

Due to Covid-19 Pandemic most of us scared to move to a new location before any vaccine is developed, but there are circumstances where one has to move and knowing which area has the best hospital capacity to deal with mass hospitalizations is crucial. This project aims 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.

Interest

The report here should not be used as a measuring tool, because the situation has been changed a lot since COVID-19 has hit the city. And keeps changing day by day as some states are opening and some are closing back again. This report is intended to help people who are planning to move to New York during or after this pandemic to start their new lives (school or jobs).

Data acquisition and cleaning

Data was collected data from following sources:

1. New York City data that contains borough, neighborhoods along with their latitudes and longitudes. o Data source: NYC data set.
2. We are going to get population data from Scraping Wikipedia. o Data source: Wikipedia page of NYC neighborhood. o We are going to go through each of the links of neighborhood and find the population of each of them.
3. Hospital information is going to be fetched from foursquare API. o Data source: foursquare API

4. Hospital bed information is going to be fetched from NYS Health Profile website. o Data source: NYS Health Profile.

Data Preparation

Data downloaded or scraped from multiple sources were combined into one table. I decided to data with latest information and some of the data pages sued to data collection were dynamic which means daily were updated regularly. First, I removed any unnecessary data and missing values from the NYC neighborhood data set collected by web scrapping of their Wikipedia page which gave us the population of each neighborhood.

Second, longitudes and latitudes were extracted from the NYC data set removing all other unwanted and missing values and combined both the data sets were merged into one giving us each neighborhood's population and its longitudes and latitudes. Then using Foursqaure hospital data was collected and then from each hospital profile the data reading their beds was extracted. Next hospital bed data was merged with the earlier neighborhood dataset.

Methodology

The first step was to get JSON data from NYC dataset, for that I used request function to get the data and store it into a data frame. In this data frame I stored the longitudes and latitudes of New York City neighborhoods.

```
ny_df.head()
```

t[5]:

	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

In the next New York City data was extracted. Then we can use **BeautifulSoup** to scrape boroughs from Wikipedia. Then we 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 right hand side table.

In [8]:

```
nyc_population_df.head()
```

Out[8]:

	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

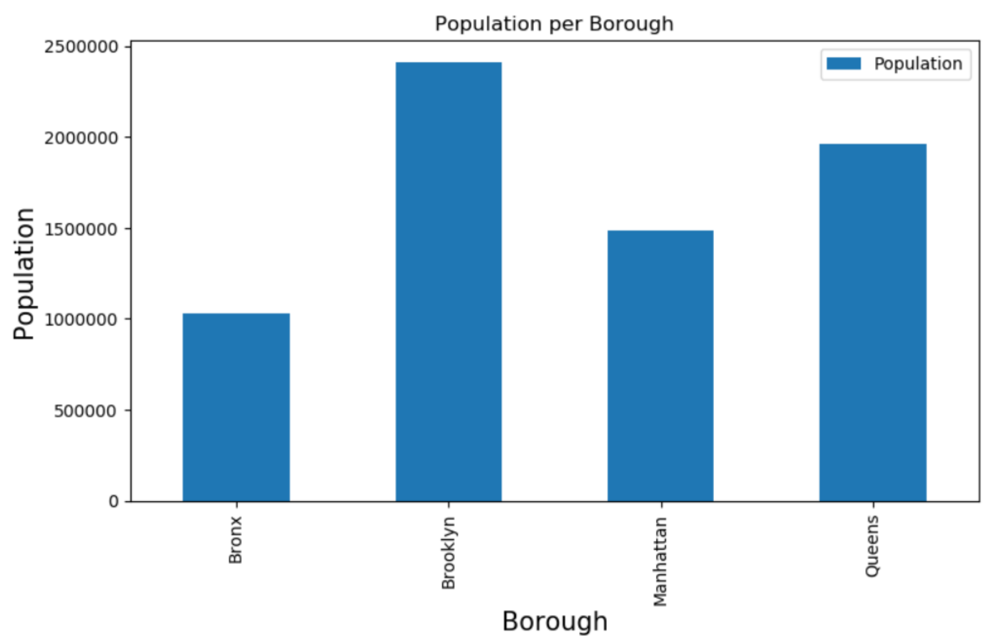
Next we can combine data frames from previous steps into one based on “neighborhood” and “borough”:

```
# Combine NYC Geo data with Population data
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()
```

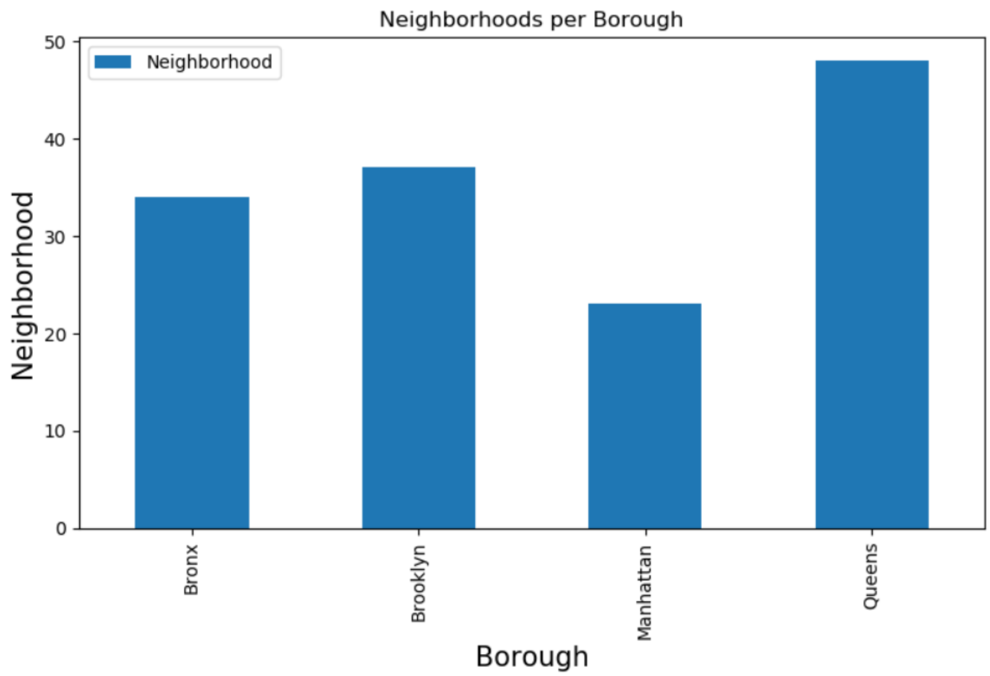
[9]:

	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

Box plot per borough and neighborhood



Second box plot for neighborhood per borough



Next hospital data from foursquare was collected. After collecting population data, now it is time to collect the hospital data. We can use the **Foursquare** API to fetch hospital data for latitude and longitude of each neighborhood from the previous dataset.

```
# Now let us use the above function
hospital_df = get_hospital_per_neighborhood_borough(nyc_df)
hospital_df.head()
```

15]:

	ID	Name	Latitude	Longitude	Borough	Neighborhood
0	59832a7bfe37406ea7eb3a79	Statcare Urgent & Walk-In Medical Care (Bronx ...	40.870056	-73.828316	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 Zicklin Jewish Hospice Resid...	40.888119	-73.910217	Bronx	Fieldston
4	5158ddffe4b086af71ca90c7	The Mollie & Jack Zicklin Jewish Hospice Resid...	40.888119	-73.910217	Bronx	Riverdale

We can also collect hospital bed related data from NYS Health Profile website. We can scrap data by using **Selenium** with **BeautifulSoap**. We 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:

Out[30]:

Neighborhood	Borough	Bed Number	ICU Bed Number
Bensonhurst	Brooklyn	204	8
Borough Park	Brooklyn	711	40
Briarwood	Queens	671	24
Brownsville	Brooklyn	600	28
Bushwick	Brooklyn	324	16

Now we are going to combine data from step four and step five. We are going to internally join the data frame based on "neighborhood" and "borough". In this step after combining data frames based on neighborhood and borough, we will get hospital names with their total beds and number of ICU beds, we will also get the name of the borough and neighborhood in which the hospital is located.

In [28]:

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

Out[28]:

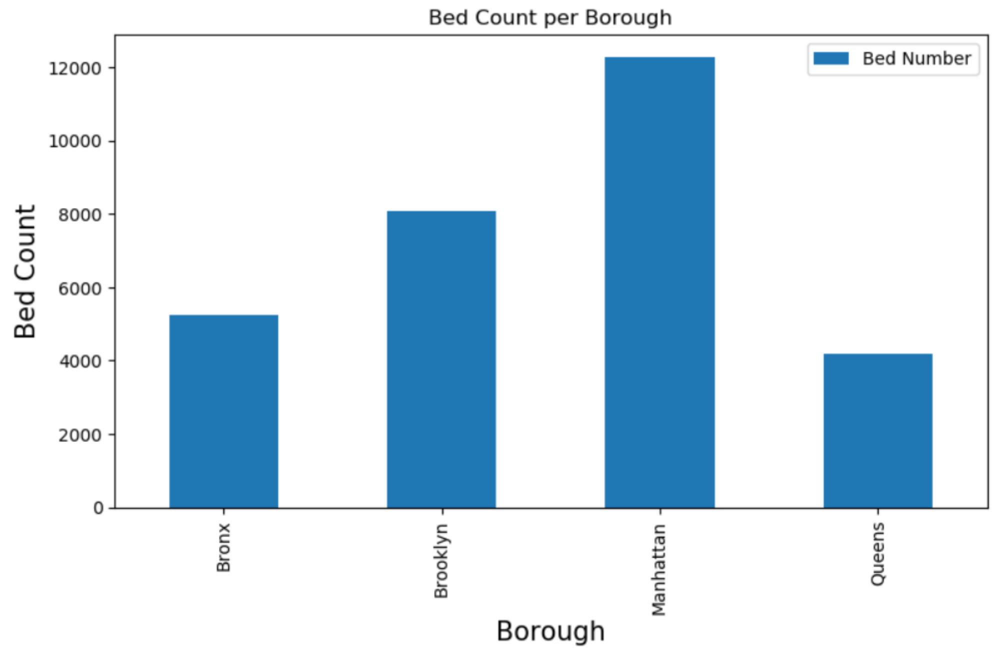
	Hospital Name	Bed Number	ICU Bed Number	Borough	Neighborhood
0	Jamaica Hospital Medical Center	402	8	Queens	Briarwood
1	New York Community Hospital of Brooklyn, Inc	134	7	Brooklyn	Fort Greene
2	Mount Sinai Hospital	1139	85	Manhattan	East Harlem
3	Nassau University Medical Center	530	22	Manhattan	Turtle Bay
4	Richmond University Medical Center	448	20	Manhattan	Turtle Bay

Now we need to clean the data frame so that we know number of regular beds and ICU beds each neighborhood and a brough has

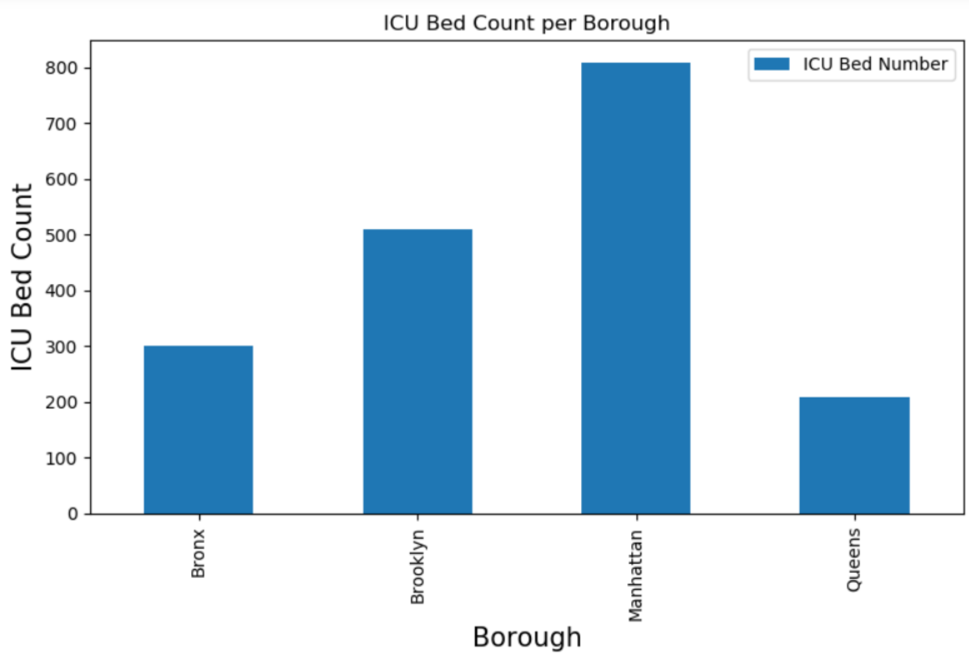
Out[30]:

Neighborhood	Borough	Bed Number	ICU Bed Number
Bensonhurst	Brooklyn	204	8
Borough Park	Brooklyn	711	40
Briarwood	Queens	671	24
Brownsville	Brooklyn	600	28
Bushwick	Brooklyn	324	16

Now, we plot the number of beds per brough so that it’s easier to understand



We also plot ICU beds per borough to compare it with the above plot



Now, we need to combine all the data that we have collected and cleaned till now so that we can see all the required information in one place. For that we create a data frame which has borough and neighborhood name, regular bed number and ICU bed number also the latitude and longitude with ach neighborhood's population. The merging is done on borough and neighborhood name. The final data frame looks like this

Out[33]:

	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population
0	Brooklyn	Bensonhurst	204	8	40.611009	-73.995180	151705
1	Brooklyn	Borough Park	711	40	40.633131	-73.990498	106357
2	Queens	Briarwood	671	24	40.710935	-73.811748	53877
3	Brooklyn	Brownsville	600	28	40.663950	-73.910235	58300
4	Brooklyn	Bushwick	324	16	40.698116	-73.925258	129239

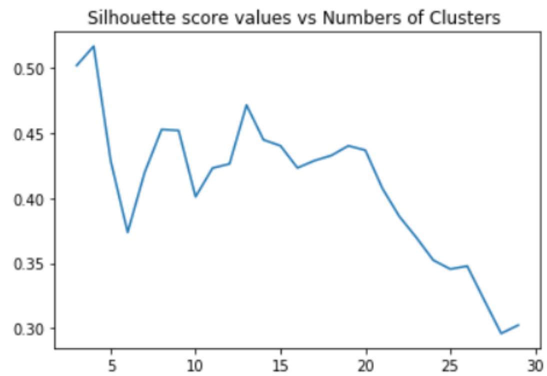
The above information is still too general, we need data that is little more focused or detailed. So, for that we are going to calculate bed per hundred people based on two rows: Population and Bed Number. Then add this to the data frame. Similarly, we are going to add ICU data to data frame:

Out[34]:

	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	Brooklyn	Borough Park	711	40	40.633131	-73.990498	106357	0.037609	0.668503
2	Queens	Briarwood	671	24	40.710935	-73.811748	53877	0.044546	1.245429
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

Since now we have all that data we need and it has also been cleaned, so now we build a model to do our calculations. We are going to use k-means clustering to partition the data into **k** groups. we will be using **elbow method** 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 3.

In [38]:
```
# Performing k-means clustering
plot_kmeans(df_clusters)
```




```
Optimal number of components is:
4
```

Now we merge cluster labels with groups of data frame

In [40]:
```
# Combining cluster data with dataframe
df.insert(0, 'Cluster Labels', kmeans.labels_)
df.head()
```

Out[40]:

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Hundred People	Bed Per Hundred People
0	1	Brooklyn	Bensonhurst	204	8	40.611009	-73.995180	151705	0.005273	0.134472
1	1	Brooklyn	Borough Park	711	40	40.633131	-73.990498	106357	0.037609	0.668503
2	0	Queens	Briarwood	671	24	40.710935	-73.811748	53877	0.044546	1.245429
3	0	Brooklyn	Brownsville	600	28	40.663950	-73.910235	58300	0.048027	1.029160
4	1	Brooklyn	Bushwick	324	16	40.698116	-73.925258	129239	0.012380	0.250698

Next, we find which borough belongs to which cluster

Borough belonging to cluster 0

Out[49]:

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Hundred People	Bed Per Hundred People
2	0	Queens	Briarwood	671	24	40.710935	-73.811748	53877	0.044546	1.245429
3	0	Brooklyn	Brownsville	600	28	40.663950	-73.910235	58300	0.048027	1.029160
6	0	Manhattan	Chinatown	392	25	40.715618	-73.994279	47844	0.052253	0.819329
7	0	Manhattan	Clinton	296	12	40.759101	-73.996119	45884	0.026153	0.645105
10	0	Bronx	East Tremont	282	14	40.842696	-73.887356	43423	0.032241	0.649425
13	0	Queens	Far Rockaway	257	8	40.603134	-73.754980	60035	0.013326	0.428084
14	0	Bronx	Fordham	1029	70	40.860997	-73.896427	43394	0.161313	2.371296
15	0	Queens	Forest Hills	312	28	40.725264	-73.844475	83728	0.033442	0.372635
16	0	Brooklyn	Fort Greene	598	31	40.688527	-73.972906	28335	0.109405	2.110464
18	0	Brooklyn	Gravesend	371	22	40.595260	-73.973471	29436	0.074738	1.260361
19	0	Brooklyn	Homecrest	306	17	40.598525	-73.959185	44316	0.038361	0.690496
20	0	Manhattan	Inwood	196	6	40.867684	-73.921210	58946	0.010179	0.332508
23	0	Manhattan	Morningside Heights	495	24	40.808000	-73.963896	31884	0.075273	1.552503
24	0	Bronx	Morris Heights	444	28	40.847898	-73.919672	36779	0.076130	1.207211
25	0	Bronx	Morrisania	170	0	40.823592	-73.901506	16863	0.000000	1.008124
27	0	Bronx	Norwood	1169	80	40.877224	-73.879391	40494	0.197560	2.886847
28	0	Bronx	Pelham Bay	200	0	40.850641	-73.832074	11931	0.000000	1.676305
29	0	Bronx	Pelham Parkway	421	22	40.857413	-73.854756	30073	0.073155	1.399927
30	0	Brooklyn	Prospect Heights	203	8	40.676822	-73.964859	67645	0.011826	0.300096

Borough Belonging to cluster 1

In [50]:
```
df[(df['Cluster Labels'] == 1)]
```

Out[50]:

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Hundred People	Bed Per Hundred People
0	1	Brooklyn	Bensonhurst	204	8	40.611009	-73.995180	151705	0.005273	0.134472
1	1	Brooklyn	Borough Park	711	40	40.633131	-73.990498	106357	0.037609	0.668503
4	1	Brooklyn	Bushwick	324	16	40.698116	-73.925258	129239	0.012380	0.250698
8	1	Brooklyn	Crown Heights	287	13	40.670829	-73.943291	143000	0.009091	0.200699
9	1	Manhattan	East Harlem	2902	193	40.792249	-73.944182	115921	0.166493	2.503429
12	1	Brooklyn	Erasmus	591	36	40.646926	-73.948177	135619	0.026545	0.435780
21	1	Queens	Jackson Heights	545	20	40.751981	-73.882821	108152	0.018492	0.503920
31	1	Brooklyn	Prospect Lefferts Gardens	2080	197	40.658420	-73.954899	99287	0.198415	2.094937
36	1	Brooklyn	Sunset Park	364	24	40.645103	-74.010316	126000	0.019048	0.288889
38	1	Manhattan	Upper East Side	632	15	40.775639	-73.960508	124231	0.012074	0.508730
39	1	Manhattan	Upper West Side	514	33	40.787658	-73.977059	214744	0.015367	0.239355

Borough belonging to cluster 2

Out[51]:

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Hundred People	Bed Per Hundred People
5	2	Brooklyn	Carroll Gardens	535	29	40.680540	-73.994654	12853	0.225628	4.162452
11	2	Manhattan	East Village	2140	210	40.727847	-73.982226	63347	0.331507	3.378218
17	2	Queens	Glen Oaks	1781	106	40.749441	-73.715481	29506	0.359249	6.036060
22	2	Bronx	Melrose	1118	59	40.819754	-73.909422	24913	0.236824	4.487617
26	2	Manhattan	Murray Hill	1426	60	40.748303	-73.978332	10864	0.552283	13.125920
37	2	Manhattan	Turtle Bay	1840	127	40.752042	-73.967708	24856	0.510943	7.402639
41	2	Brooklyn	Windsor Terrace	839	40	40.656946	-73.980073	20988	0.190585	3.997522
43	2	Manhattan	Yorkville	1438	103	40.775930	-73.947118	35221	0.292439	4.082792

So far, we have analyzed dataset for neighborhoods with hospitals. Now, we should look into neighborhoods without hospital data:

Out[52]:

	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
15	Bronx	West Farms

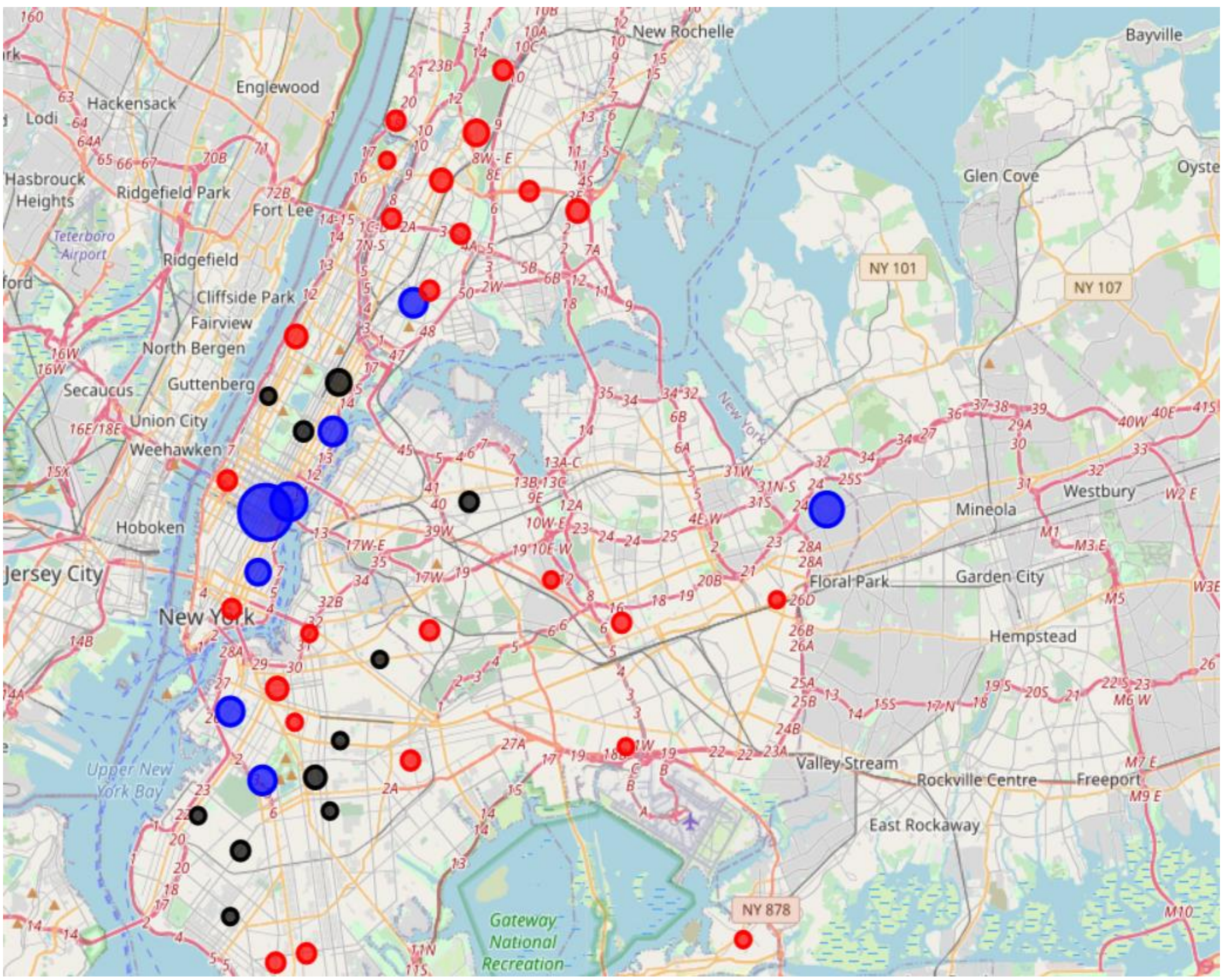
If we see the index count of neighborhoods with and without hospital, it should look like this:

```
Neighborhood without hospital count: 98
Neighborhood with hospital count: 44
```

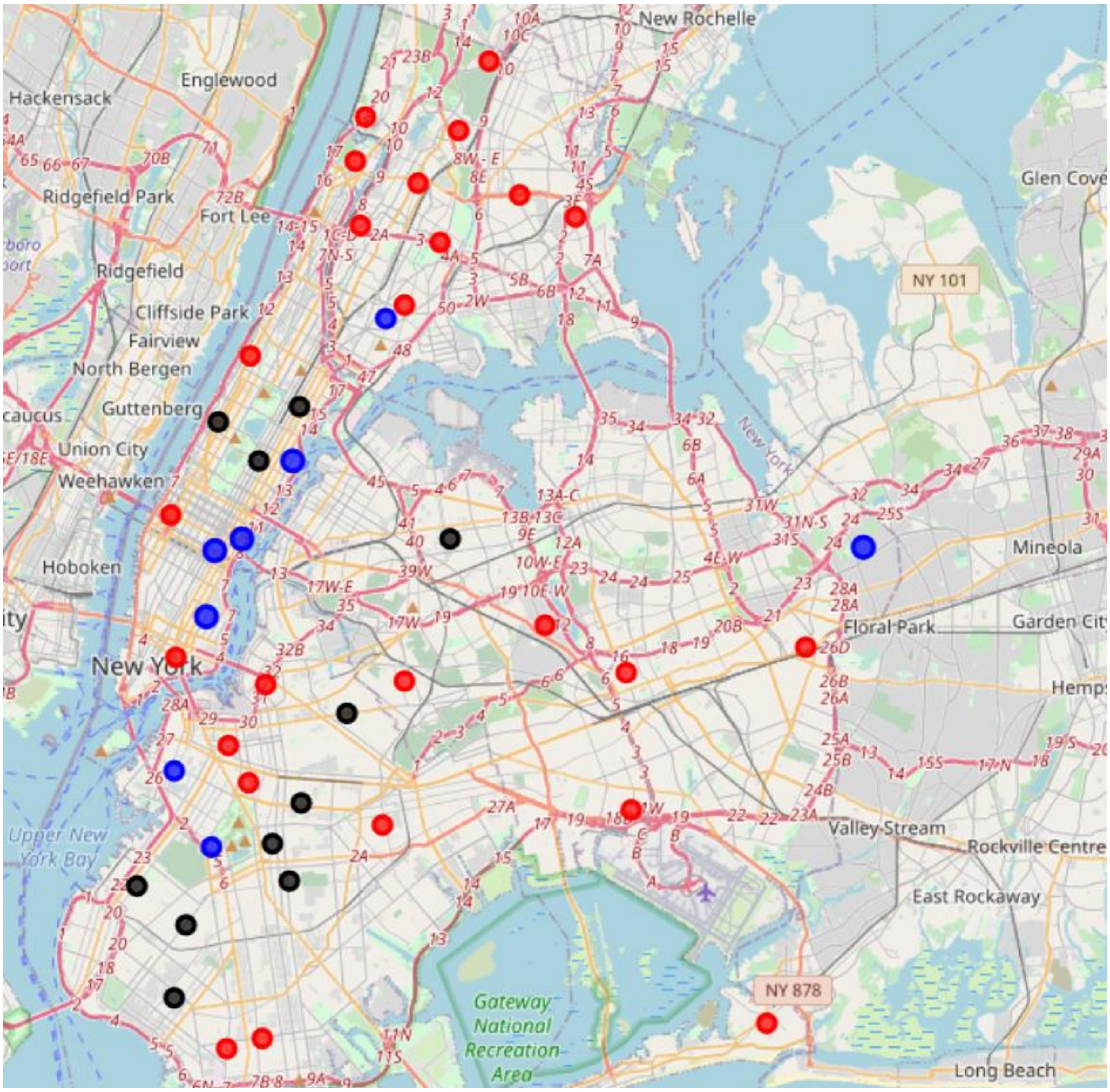
We can see that there are 98 neighborhoods that are without hospitals, so anyone who is moving to NYC should try to avoid these neighborhoods of they can.

Visualize with Folium

Now, we are going to use **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.



The second map illustrates the clusters where the radius of the Circle marker is proportional to ICU beds per hundred people.

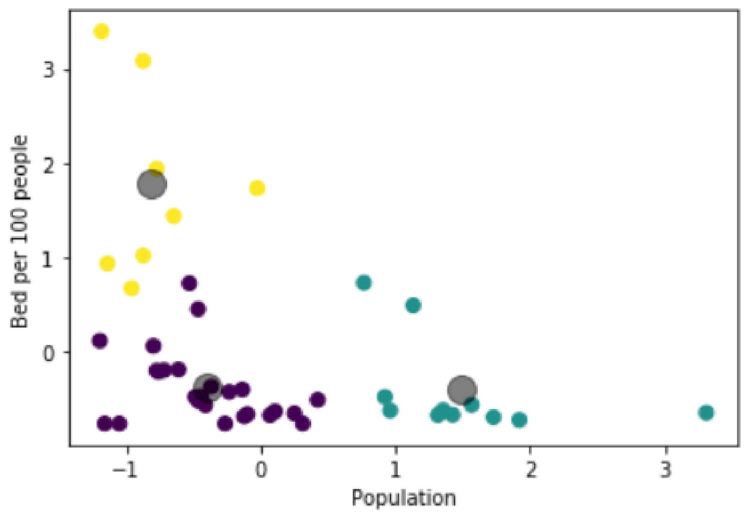


We can see that one of the clusters (blue circle) consists in one borough - **Manhattan**.

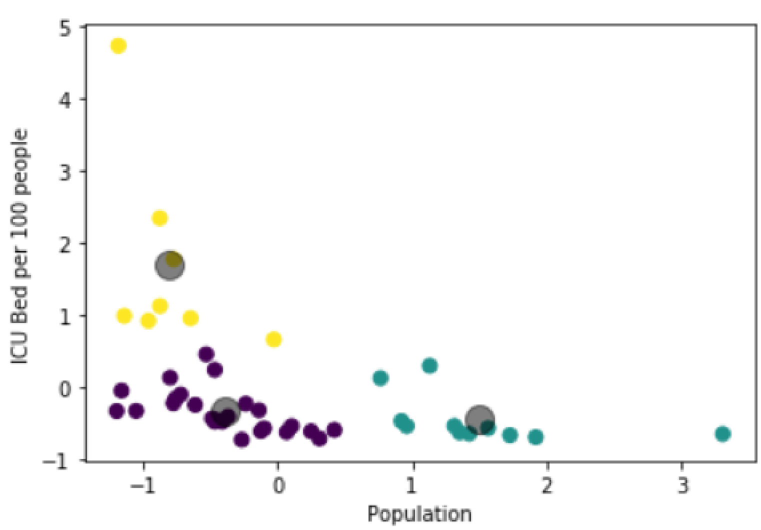
Scatter Plot

Let's look at the scatter plots of our data and define our clusters with colors. The grey circle marker is representing the centroid of each cluster. Don't forget that our data is normalized, so the axes do not deliver real values.

Beds per 100 people



ICU Beds per 100 people



We can observe the obvious outlier here. This neighborhood has a high number of beds per people ratio. From maps above we can easily say that it is **East Harlem**.

Results and Discussions

During the analysis, three clusters were defined. One cluster (cluster 2), 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 98 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. For that reason, other maps could be utilized such as Google map or Open Street map.

NYS Health Profile website lacks the latest information regarding hospital information. It could lack information regarding new hospitals. Also, hospital ids were extracted manually from NYS, which could have missing hospitals. We also dropped neighborhoods which did not have any hospital data matching in **NYS Health Profile website**. For this project, we are only using data from 74 hospitals in NYC.

We are using fuzzy-wuzzy 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.

We are also only considering hospital data. We did not consider other medical facilities like nursing home or health clinic.

We used population data from 2010(as per Wikipedia pages), which are not accurate currently. We should have used the latest population data.

Finally, to battle COVID-19, we should have had patient data for the neighborhood. Unfortunately, we could not find 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-East Harlem** was chosen as the most well equipped(as per hospital bed count and ICU bed count) area to battle pandemic.