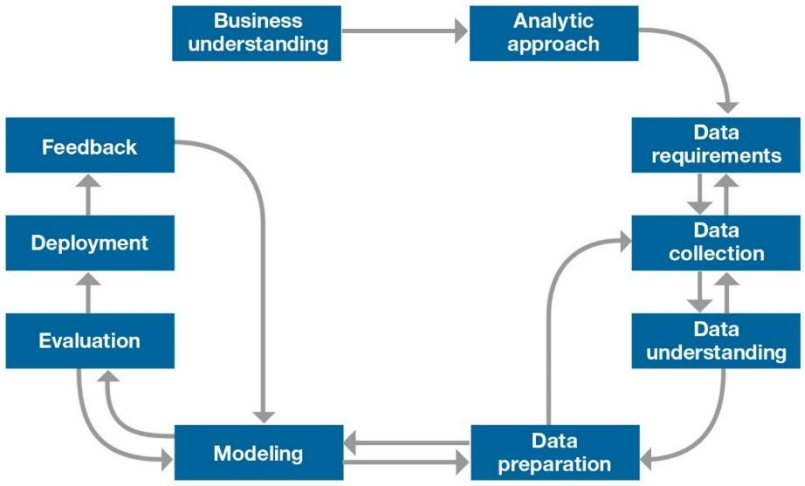
IBM/COURSERA CAPSTONE PROJECT

Exploring the neighborhoods & Boroughs in New-York City that are the best equipped to fight the Covid-19 Pandemic

To answer this question, we will follow the recommended IBM Methodology for Data Science:



1. Business Understanding.

1.1. Background.

The Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease was first identified in December 2019 in Wuhan, the capital of China's Hubei province, and has since spread globally, resulting in the ongoing 2019–20 coronavirus pandemic. As of 29 April 2020, more than 3.13 million cases have been reported across 185 countries and territories, resulting in more than 217,000 deaths.

Within this context, all countries have been hit, with some Regions of the World more or less severely impacted. New-York City in the US is one of the cities that have been strongly hit, with a death toll of approximately 12,509 as of today (out of a total of 58,864 deaths in the US) according to the WHO statistics.

1.2. Business Problem.

Hospitals across the state "have been urgently looking to expand capacity in advance of the continuing surge in the number of coronavirus patients and officials said they are planning to possibly shift confirmed virus patients from hospitals with dwindling numbers of available beds to hospitals elsewhere in the state" (ctmirror.org). At the time of writing this report (April 29th 2020) the situation has improved, but experts are dreading a second wave of infections. Within this context, it is important to have a closer overview of the neighborhoods in New-York City that are the best prepared to welcome infected patients and to fight against this pandemic, by looking at the hospital bed capacity at each neighborhood. To complete this analysis, I will also look at the cases of infected people by neighborhood in order to be able to compare with the hospitals bed capacity.

1.3. Target Audience.

The target audience would be medical and non-medical experts and analysts from the US Department of Health and Human Services (HHS) working for the City of New-York.

This study will give them a clear understanding of hospitals bed capacity at the different neighborhoods of NY City. The comparison with the number of infected cases at each neighborhood will give a better comprehension in the handling of the Covid-19 pandemic at the neighborhood level. This would also be helpful for making future predictions and for adopting a better crisis management approach in case of a second Covid-19 wave.

2. Analytic Approach.

We will adopt the following approach in an attempt to answer the problem:

- Collect the data about New York City.
- Collect the data about New-York City population for each neighborhood.
- Collect the data about the number of infected people for each neighborhood.
- Use the Foursquare API to get the list of hospitals at each neighborhood.
- Collect the hospital bed data.
- Perform Data Visualization statistical analysis.
- Analyze data by Clustering (using K-Means technique).
- 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.

- Look at the number of infected people at each neighborhood and compare with the hospitals bed capacity at each neighborhood, along with some visualizations.
- Provide feedback, draw conclusions and open on future implications or questions for research.

3. Data Requirements.

As mentioned, we need to collect a variety of data for this study.

- From public (online) data sources: we will collect data about New-York City, its neighborhoods, its population. We will also collect data about the hospitals-bed capacity and about the number of infected cases in New-York City. We might have to scrap data from these sources.
- From Fourthsquare API: we will get the list of hospitals at each neighborhood by calling the Foursquare API, using the "venues" parameter.

4. Data Collection.

- New York city data: from Json file: https://cocl.us/new_york_dataset
- Population data for each neighborhood: Wikipedia: https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City
- List of hospitals at each neighborhood: through the Foursquare API.
- Hospital bed data: from the NYS Health Profile: https://profiles.health.ny.gov/hospital/index#5.79/42.868/-76.809
- List of confirmed Covid-19 cases by Borough: Wikipedia: https://github.com/nychealth/coronavirus-data/blob/master/boro.csv

5. Methodology.

In this part, we will combine the remaining stages of the IBM guidelines.

- Step 1: we import all the required libraries to perform our analysis.

During this initial stage, we import all the packages and libraries that allow us to wrok with out data sets later. The packages we install incude geocoder, senium, selenium, fuzzywuzzy. We also import the matplotlib library to do visualizations.

- Step 2: get the New-York City dataset with the coordinates of each Neighborhood, and store it as a dataframe.

The New-York City dataset is publicly available at https://cocl.us/new_york_dataset. We use the "request" function to get the New-York data and store it into a dataframe (shown below).

	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

Head of dataframe

- Step 3: Let use the Beautifulsoup to scrap data from Wikipedia page to get the New-York City Boroughs.

We use the Wiki link https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City to get the data about the Boroughs. We scrap the Wikipedia page using Beautifulsoup. We store this data in another dataframe (shown below).

	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

Head of dataframe

- Step 4: we combine the 2 dataframes into a single one including all the data.

	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

Head of dataframe

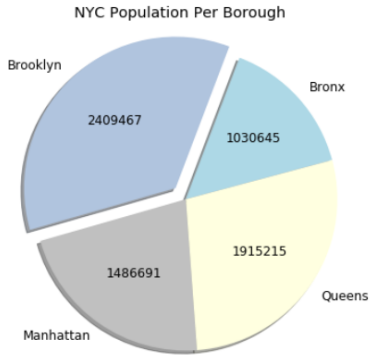
From the dataframe, we can print the total number of Boroughs and Neighborhoods, using the "print" function:

```
]: print('The dataframe has {} boroughs and {} neighborhoods.'.format(
        len(nyc['Borough'].unique()),
        nyc.shape[0]
    )
)
```

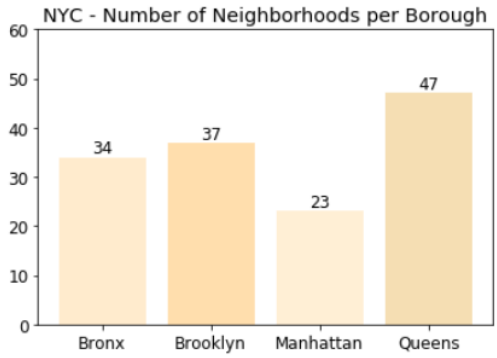
```
The dataframe has 4 boroughs and 141 neighborhoods.
```

Therefore, we see that there are 4 Boroughs in our data set and 141 Neighborhoods. We can also visualize the population per Neighborhood on a Folium Map, using some red colored markers to showing the name of the Neighborhood and the number of Population in the Neighborhood selected:

Next, we create a pie chart to display the Population per Borough. To do this, we use the "groupby" and "sum" functions to get the total population per Borough. Then, we plot a pie chart with Matplotlib:



After, we use once again the “groupby” function but with the “count” function, in order to get the number of Neighborhoods per Borough. And we use the Matplotlib library to plot a bar chart representing this data:



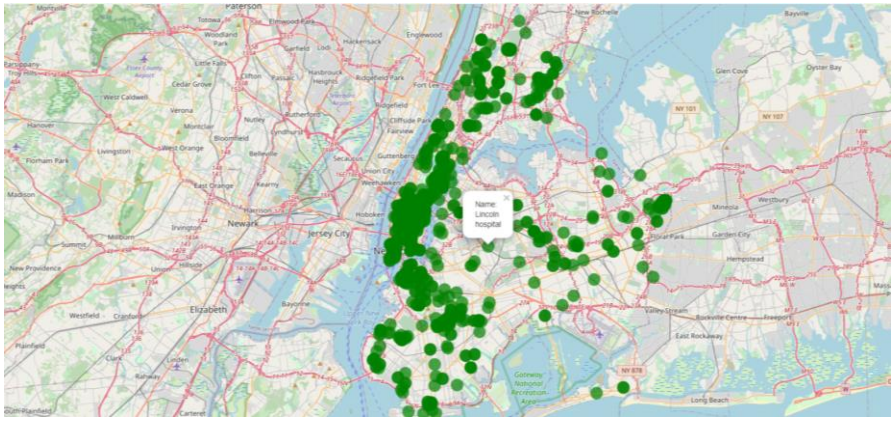
- Step 5: collect the hospitals data, using the Foursquare API.

We use the Foursquare API to get the hospital data with their latitude and longitude and we create a new dataframe to include those data into the previous dataframe that contains the Neighborhoods.

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

Head of dataframe

We create a Folium Map with green markers displaying each Hospital Name to view the Hospitals on a map:



- Step 6: we collect the Hospital-Beds data.

I collect these data from the NYS Health Profile website. I have downloaded and re-worked these data that I put in a CSV file hosted on my Github repository: https://raw.githubusercontent.com/ljulienne/Coursera_Capstone/master/beds_hospital.csv

Then, I make a dataframe containing the hospital-beds data along with Hospital Names:

	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

Head of dataframe

- Step 7: combine the Hospital-Beds data with the Neighborhoods and Boroughs data.

We combine the two dataframes by using the "join" method, based on the Neighborhoods and Boroughs:

	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	Yorkville
3	Nassau University Medical Center	530	22	Bronx	University Heights
4	Richmond University Medical Center	448	20	Bronx	University Heights

Head of dataframe

- Step 8: Make a new dataframe with the number of beds and ICU beds per Neighborhood per Borough.

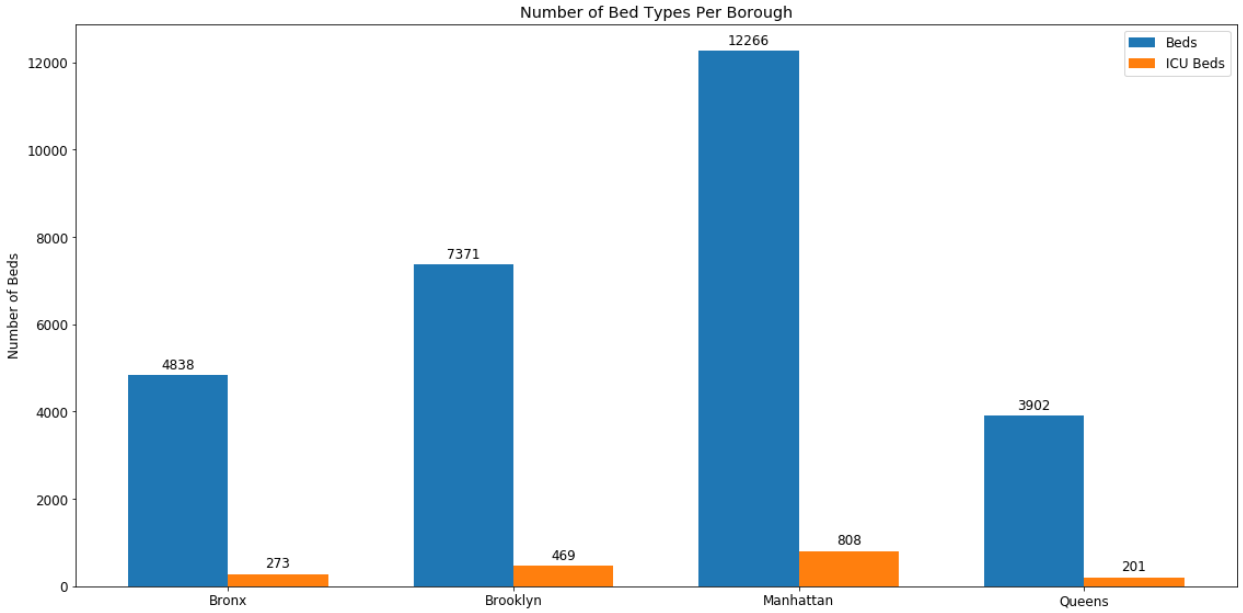
We make sure that the bed numbers and the ICU bed numbers are of type "int32" and we make the sum of beds number an ICU beds number, grouping them based on Neighborhood and Borough:

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

We can print the total of beds and the total of ICU beds per Borough:

```
In [32]: total_beds_borough=hosp_beds_br_nbh.groupby(['Borough'])['Bed Number'].sum()
         total_beds_borough.head()

Out[32]: Borough
         Bronx         5816
         Brooklyn      7371
         Manhattan    11288
         Queens        3902
         Name: Bed Number, dtype: int32

In [33]: total_icubeds_borough=hosp_beds_br_nbh.groupby(['Borough'])['ICU Bed Number'].sum()
         total_icubeds_borough.head()

Out[33]: Borough
         Bronx        315
         Brooklyn     469
         Manhattan    766
         Queens       201
         Name: ICU Bed Number, dtype: int32
```

Now, we can visualize the number of bed types available in each Borough, by plotting two bar charts (for beds and for ICU beds) for each Borough:



From this chart, we see that Manhattan is the Borough with the highest number of beds. Followed by Brooklyn, The Bronx and The Queens.

- Step 9: combine the New-York City data with the number of beds and ICU beds.

To achieve this, we will merge two dataframes: the one with the hospital-beds data and the one with the population data, based on Boroughs and Neighborhoods:

	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population
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

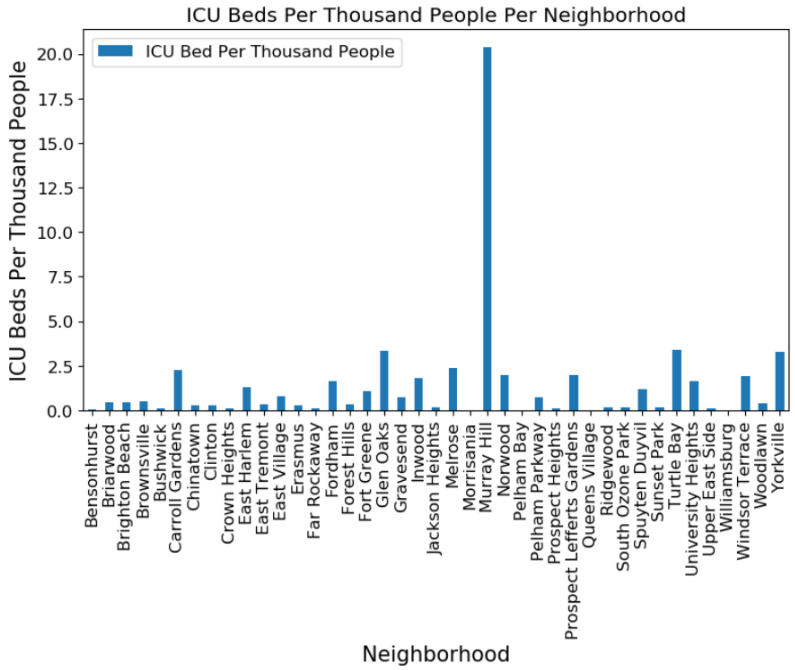
Head of dataframe

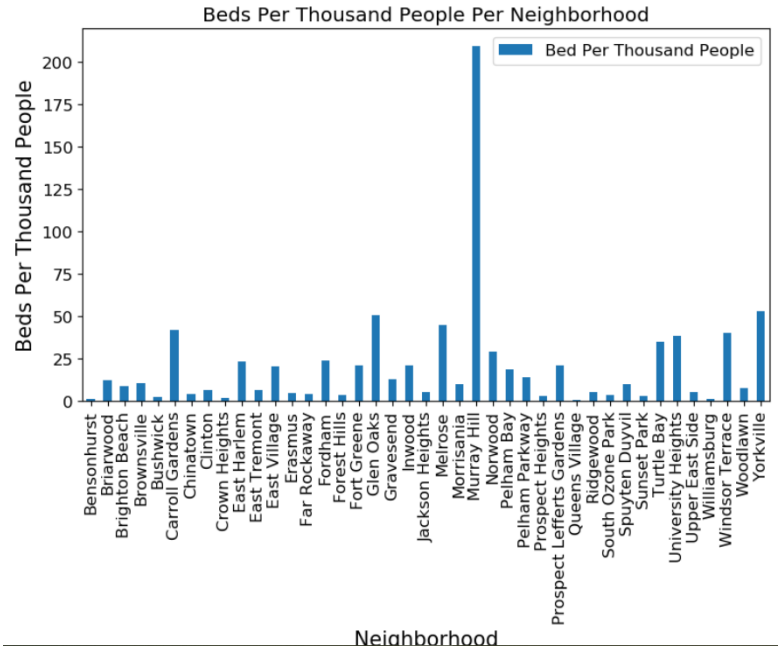
- Step 10: we add two columns to the dataframe, to include the number of beds per thousand people and the number of ICU beds per thousand people.

	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Thousand People	Bed Per Thousand People
0	Brooklyn	Bensonhurst	204	8	40.611009	-73.995180	151705	0.052734	1.344715
1	Queens	Briarwood	671	24	40.710935	-73.811748	53877	0.445459	12.454294
2	Brooklyn	Brighton Beach	306	17	40.576825	-73.965094	35547	0.478240	8.608321
3	Brooklyn	Brownsville	600	28	40.663950	-73.910235	58300	0.480274	10.291595
4	Brooklyn	Bushwick	324	16	40.698116	-73.925258	129239	0.123802	2.506983

Head of dataframe

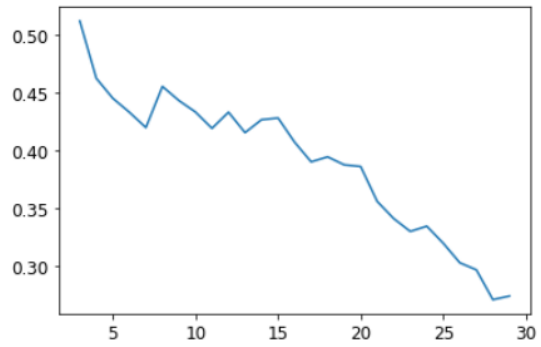
We can represent the number of ICU beds per thousand people and the number of beds per thousand people in each Neighborhood. I am still using bar charts to do this:





- Step 11: Prepare the data for K-Means Clustering.

We use K-Means clustering to partition data in k partitions. We use elbow method to find the optimum number of clusters. We normalize the data and we plot the score values vS the number of Clusters, so we can easily see on the chart the optimum k:



We see from the previous lines of code that the best value is for k = 3.

Therefore, we know that we can organize our data into 3 clusters. We run the K-Means Clustering and check the Clusters labels generated for each row in the dataframe:

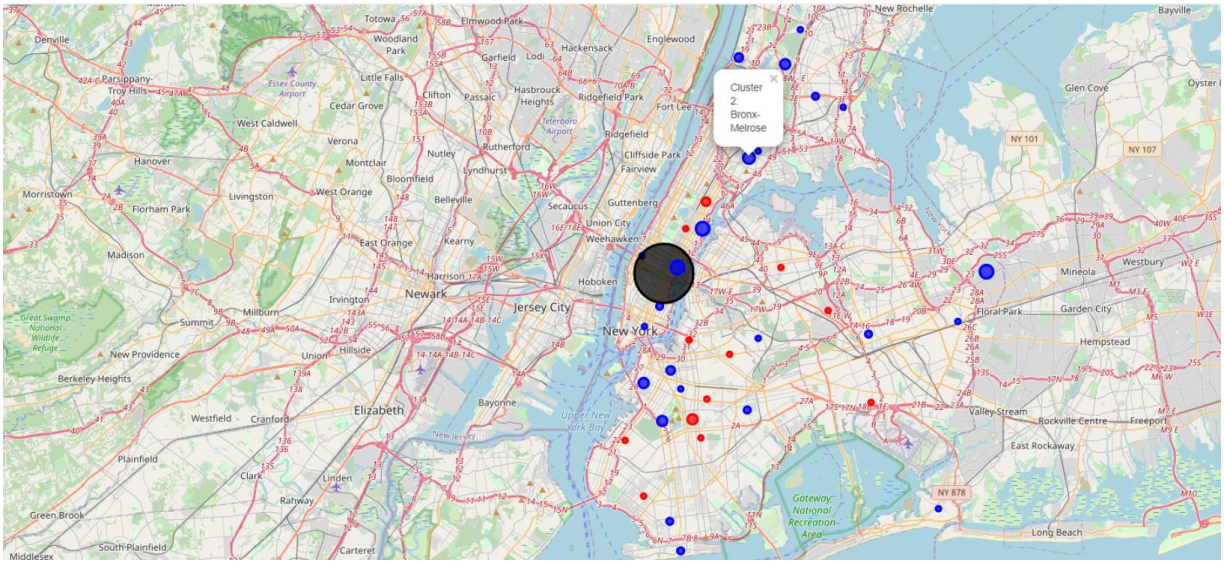
```
[44]: # From previously, we get the best value for k = 3
      kclusters = 3
      # run k-means clustering
      kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(df_clusters)
      # check cluster labels generated for each row in the dataframe
      kmeans.labels_[0:24]

Out[44]: array([0, 2, 2, 2, 0, 2, 2, 2, 0, 0, 2, 2, 0, 2, 2, 0, 2, 2, 2, 2, 0, 2,
                2, 1], dtype=int32)
```

After that, we can combine the Clusters data into a dataframe:

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Thousand People	Bed Per Thousand People
0	0	Brooklyn	Bensonhurst	204	8	40.611009	-73.995180	151705	0.052734	1.344715
1	2	Queens	Briarwood	671	24	40.710935	-73.811748	53877	0.445459	12.454294
2	2	Brooklyn	Brighton Beach	306	17	40.576825	-73.965094	35547	0.478240	8.608321
3	2	Brooklyn	Brownsville	600	28	40.663950	-73.910235	58300	0.480274	10.291595
4	0	Brooklyn	Bushwick	324	16	40.698116	-73.925258	129239	0.123802	2.506983

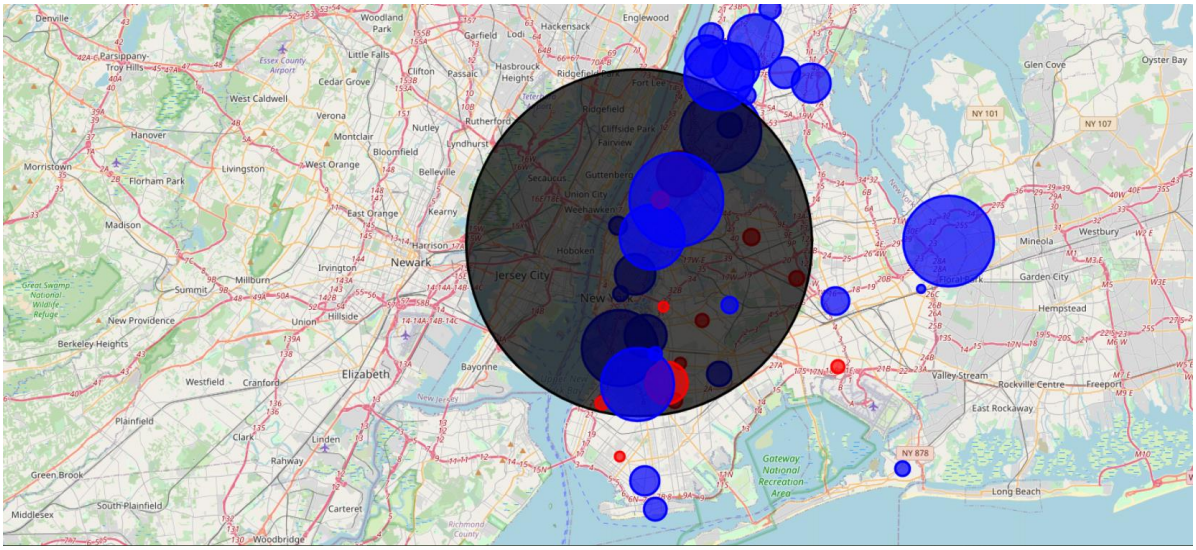
Head of dataframe

Let's define a map with the geocoder to later represent our clusters and then let's render the map:

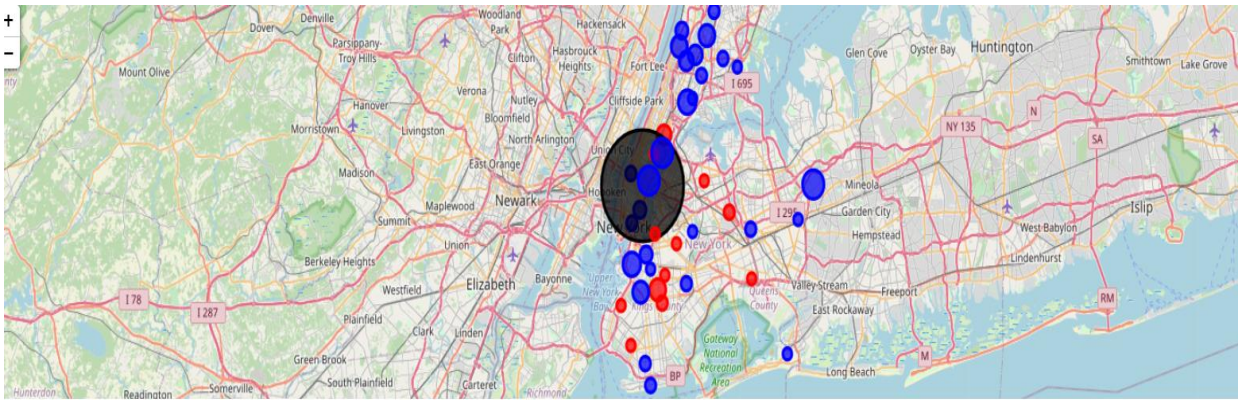


We can see that one of the clusters consists of Manhattan only (in black color).

Now, we render the map with the clusters to show the hospital-beds per thousand people. We use Folium to visualize the distribution. On the following map, we can the Clusters where the radius of the circle marker is proportional to the number of hospital-beds per thousand people:



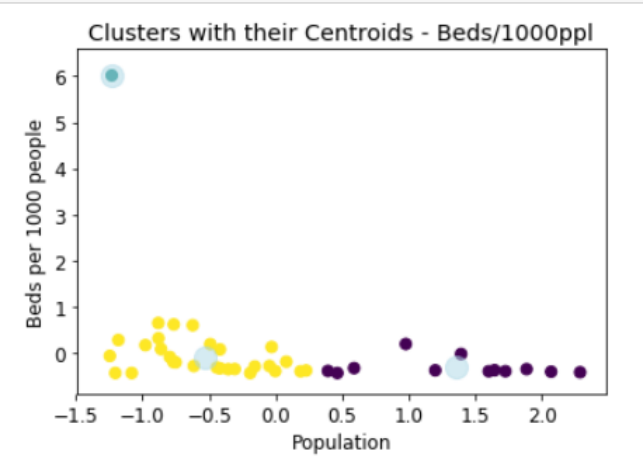
On the second map that we render, the radius of the circle marker is proportional to the number of ICU beds per thousand people:



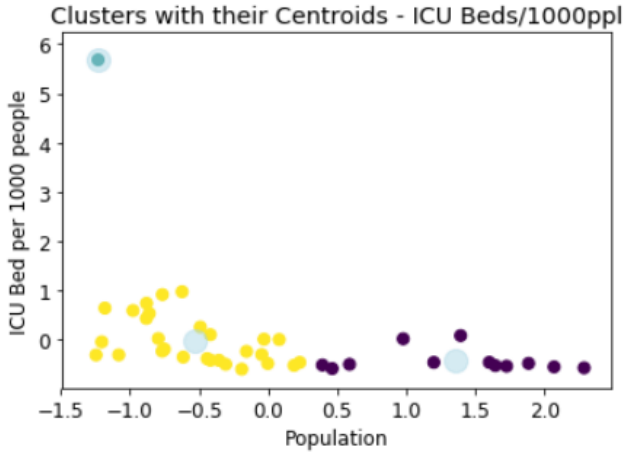
Once again, we see that one of the Clusters consists solely of one Borough, that is Manhattan (in black color).

Now, we can visualize the Clusters and their Centroids on a scatter plot. Each Cluster is represented with a different color, and the grey circles are the centroid of each cluster. The data was previously normalized, so the values on the axis do not represent actual values.

Let first see the Clusters and their Centroids for the Beds per thousand People:



Then, let see the Clusters and their Centroids for the Beds per thousand People:



So, on these two plots, we see our 3 clusters: one outlier (in light blue), another cluster in yellow and the other one in purple. Coming back to our previous analysis with dataframes and visualizations, we can easily deduct that the outlier (light blue) corresponds to Manhattan, and more specially to the Neighborhood of Murray Hill.

Let produce a dataframe to identify which Borough belongs to which Cluster. To do this, I see the dataframe based on each Cluster Label, so since we have three clusters, I will have three dataframes, one for “Cluster 0”, one for “Cluster 1” and one for “Cluster 2”:

Let see the dataframe for “Cluster 0”:

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Thousand People	Bed Per Thousand People
0	0	Brooklyn	Bensonhurst	204	8	40.611009	-73.995180	151705	0.052734	1.344715
4	0	Brooklyn	Bushwick	324	16	40.698116	-73.925258	129239	0.123802	2.506983
8	0	Brooklyn	Crown Heights	287	13	40.670829	-73.943291	143000	0.090909	2.006993
9	0	Manhattan	East Harlem	2679	151	40.792249	-73.944182	115921	1.302611	23.110567
12	0	Brooklyn	Erasmus	591	36	40.646926	-73.948177	135619	0.265450	4.357796
15	0	Queens	Forest Hills	312	28	40.725264	-73.844475	83728	0.334416	3.726352
20	0	Queens	Jackson Heights	545	20	40.751981	-73.882821	108152	0.184925	5.039204
28	0	Brooklyn	Prospect Lefferts Gardens	2080	197	40.658420	-73.954899	99287	1.984147	20.949369
31	0	Queens	South Ozone Park	247	11	40.668550	-73.809865	75878	0.144970	3.255225
33	0	Brooklyn	Sunset Park	364	24	40.645103	-74.010316	126000	0.190476	2.888889
36	0	Manhattan	Upper East Side	632	15	40.775639	-73.960508	124231	0.120743	5.087297
37	0	Brooklyn	Williamsburg	69	0	40.707144	-73.958115	78700	0.000000	0.876747

Now, let see the dataframe for “Cluster 1”:

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Thousand People	Bed Per Thousand People
23	1	Manhattan	Murray Hill	2270	221	40.748303	-73.978332	10864	20.342415	208.946981

Note: once again, we see the outlier here (Manhattan with the Neighborhood of Murray Hill) that has the highest number of beds and ICU beds.

To finish, let see the dataframe for “Cluster 2”:

	Cluster Labels	Borough	Neighborhood	Bed Number	ICU Bed Number	Latitude	Longitude	Population	ICU Bed Per Thousand People	Bed Per Thousand People
1	2	Queens	Briarwood	671	24	40.710935	-73.811748	53877	0.445459	12.454294
2	2	Brooklyn	Brighton Beach	306	17	40.576825	-73.965094	35547	0.478240	8.608321
3	2	Brooklyn	Brownsville	600	28	40.663950	-73.910235	58300	0.480274	10.291595
5	2	Brooklyn	Carroll Gardens	535	29	40.680540	-73.994654	12853	2.256283	41.624523
6	2	Manhattan	Chinatown	180	13	40.715618	-73.994279	47844	0.271716	3.762227
7	2	Manhattan	Clinton	296	12	40.759101	-73.996119	45884	0.261529	6.451050
10	2	Bronx	East Tremont	282	14	40.842696	-73.887356	43423	0.322410	6.494254
11	2	Manhattan	East Village	1296	49	40.727847	-73.982226	63347	0.773517	20.458743
13	2	Queens	Far Rockaway	257	8	40.603134	-73.754980	60035	0.133256	4.280836
14	2	Bronx	Fordham	1029	70	40.860997	-73.896427	43394	1.613126	23.712956
16	2	Brooklyn	Fort Greene	598	31	40.688527	-73.972906	28335	1.094053	21.104641
17	2	Queens	Glen Oaks	1497	98	40.749441	-73.715481	29506	3.321358	50.735444
18	2	Brooklyn	Gravesend	371	22	40.595260	-73.973471	29436	0.747384	12.603615
19	2	Manhattan	Inwood	1218	105	40.867684	-73.921210	58946	1.781291	20.662980
21	2	Bronx	Melrose	1118	59	40.819754	-73.909422	24913	2.368241	44.876169
22	2	Bronx	Morrisania	170	0	40.823592	-73.901506	16863	0.000000	10.081243
24	2	Bronx	Norwood	1169	80	40.877224	-73.879391	40494	1.975601	28.868474
25	2	Bronx	Pelham Bay	225	0	40.850641	-73.832074	11931	0.000000	18.858436
26	2	Bronx	Pelham Parkway	421	22	40.857413	-73.854756	30073	0.731553	13.999268
27	2	Brooklyn	Prospect Heights	203	8	40.676822	-73.964859	67645	0.118264	3.000961
29	2	Queens	Queens Village	25	0	40.718893	-73.738715	52504	0.000000	0.476154
30	2	Queens	Ridgewood	348	12	40.708323	-73.901435	69317	0.173118	5.020413
32	2	Bronx	Spuyten Duyvil	103	12	40.881395	-73.917190	10279	1.167429	10.020430
34	2	Manhattan	Turtle Bay	862	85	40.752042	-73.967708	24856	3.419697	34.679755
35	2	Bronx	University Heights	978	42	40.855727	-73.910416	25702	1.634114	38.051514
38	2	Brooklyn	Windsor Terrace	839	40	40.656946	-73.980073	20988	1.905851	39.975224
39	2	Bronx	Woodlawn	321	16	40.898273	-73.867315	42483	0.376621	7.555964
40	2	Manhattan	Yorkville	1855	115	40.775930	-73.947118	35221	3.265098	52.667443

We can also see which Neighborhoods do not have any hospital: I merge the dataframe with Boroughs and Neighborhoods with the one containing the hospital names, and then I use the “loc” method:

In [57]:
```
# Let us see neighborhoods which does not have any hospitals
no_hosp = pd.merge(nyc, df_beds_pop, how='outer', indicator=True, on=["Borough", "Neighborhood"])
no_hosp = no_hosp.loc[no_hosp._merge == 'left_only', ["Borough", "Neighborhood"]]
no_hosp
```

Out[57]:

	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
12	Bronx	Morris Heights
15	Bronx	West Farms
17	Bronx	Mott Haven
18	Bronx	Port Morris
19	Bronx	Longwood
20	Bronx	Hunts Point
22	Bronx	Soundview
23	Bronx	Clason Point
24	Bronx	Throgs Neck
25	Bronx	Parkchester
26	Bronx	Belmont
29	Bronx	Castle Hill
30	Bronx	Olinville
31	Bronx	Pelham Gardens

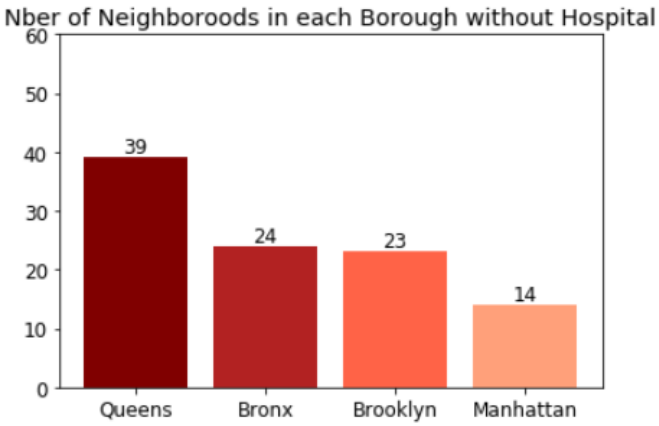
From the dataframe, we identify 100 Neighborhoods without hospital and therefore 41 Neighborhoods that have hospitals.

Let see the Boroughs that have the highest number of Neighborhoods without hospital:

]:
```
df = no_hosp['Borough'].value_counts().reset_index()
df.columns = ['Borough', 'count']
print (df)
```

```
      Borough  count
0      Queens     39
1       Bronx     24
2    Brooklyn     23
3   Manhattan     14
```

Let plot a bars chart to visualize this data:

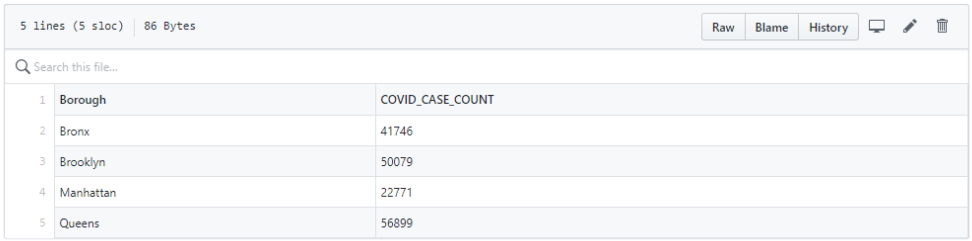


From the code and the chart, we see that The Queens is the Borough that has the highest number of Neighborhoods without hospital.

- Step 12: Get the number of confirmed Covid-19 Cases per Borough.

Note: from Step 12, to simplify my analysis, I focus on the Borough level.

I have got the data from the NYC Health website (https://www1.nyc.gov/site/doh/covid/covid-19-data.page), I created a simple Excel file with the Borough names and the number of Covid-19 positive cases per Borough. Then, I put this file on my Github repository.



Visualization of Excel file with Covid-19 positive case count on Github

I get the file and I make a dataframe:

	Borough	COVID_CASE_COUNT
0	Bronx	41746
1	Brooklyn	50079
2	Manhattan	22771
3	Queens	56899

Now, I produce a dataframe where I add the total number of beds, the total number of ICU beds, and the total population per Borough:

	Borough	COVID_CASE_COUNT	ICU_Beds	Beds	Latitude	Longitude	Population
0	Bronx	41746	273	4838	40.837048	-73.865433	1030645
1	Brooklyn	50079	469	7371	40.650002	-73.949997	2409467
2	Manhattan	22771	808	12266	40.783100	-73.971200	1486691
3	Queens	56899	201	3902	40.728200	-73.794900	1915215

Let have a close look to analyze the percentage of Covid-19 confirmed cases in relation to the population of each Borough. I first get the percentages of infected people compared to the population of each Borough and I add a column in my previous dataframe with those numbers:

```
: covid_borough3 = covid_borough2['COVID_CASE_COUNT']/covid_borough2['Population']*100
  covid_borough3
```

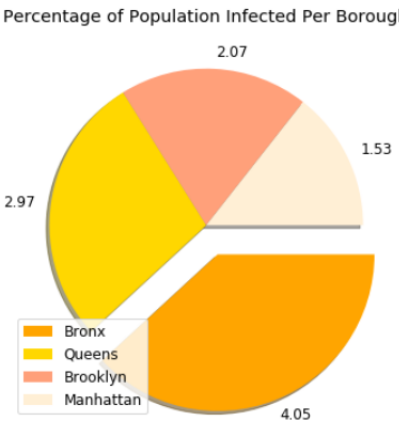
```
56]: 0    4.050473
     1    2.078426
     2    1.531657
     3    2.970894
     dtype: float64
```

```
: covid_borough4 = covid_borough2.assign(Percent_Pop_Infected = [4.05,2.07,1.53,2.97])
  covid_borough4
```

```
57]:
```

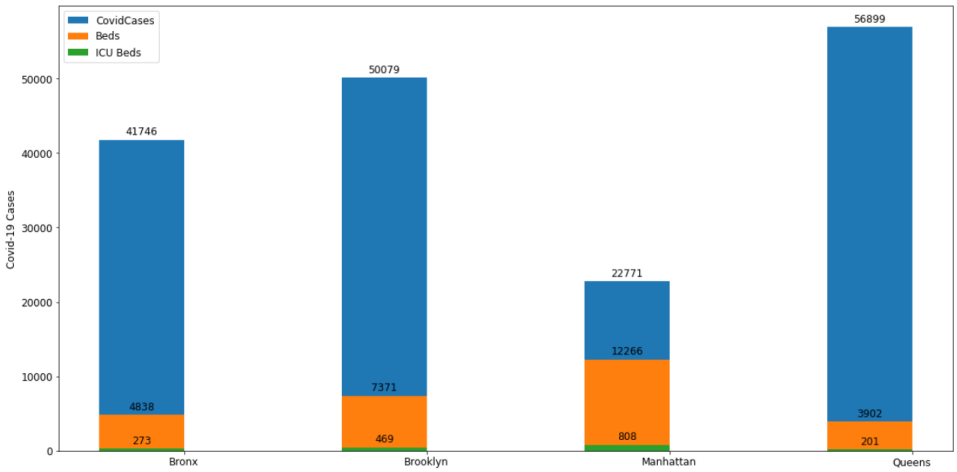
	Borough	COVID_CASE_COUNT	ICU_Beds	Beds	Latitude	Longitude	Population	Percent_Pop_Infected
0	Bronx	41746	273	4838	40.837048	-73.865433	1030645	4.05
1	Brooklyn	50079	469	7371	40.650002	-73.949997	2409467	2.07
2	Manhattan	22771	808	12266	40.783100	-73.971200	1486691	1.53
3	Queens	56899	201	3902	40.728200	-73.794900	1915215	2.97

To have a better idea, we visualize on a pie chart:



We see that the Borough with the highest percentage of its population being infected is The Bronx, followed by The Queens, then Brooklyn and Manhattan.

We can compare the number of bed types with the number of confirmed Covid-19 cases in each Borough:

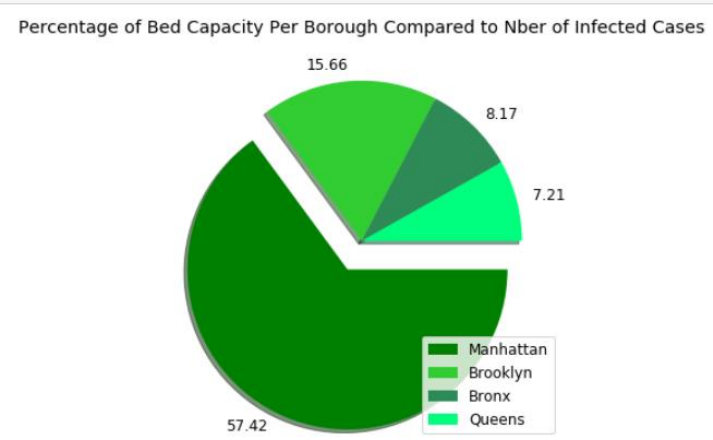


From this chart, it seems that, proportionally to the number of beds and of infected people, the best equipped Borough to fight Covid-19 is Manhattan (this comfort us with the previous resulsts that we found in our study). The Queens seem to be facing the most difficult situation with the highest number of population infected compared to its total beds capacity.

Let confirm our observations by calculating the percentages of total beds available in each Borough compared ot the number of infected people. Then, we make a dataframe with these data:

	Borough	Total Beds	Percent_Beds_Available
0	Bronx	5111	8.17
1	Brooklyn	7840	15.66
2	Manhattan	13074	57.42
3	Queens	4103	7.21

We can visualize on a pie chart to get a better idea:



The calculation of percentages and the visualization confirm our hypothesis: Manhattan is indeed the best equipped Borough in terms of beds capacity compared to the number of infected people. It is followed by Brooklyn, The Bronx and The Queens.

6. Results.

During our analysis, we have explored the New-York City data and we have made a comparison with the current number of COvid-19 infected population in order to get better insights. We have seen that New-York City has 4 Boroughs: The Bronx, Manhattan, The Queens and The Queens. We also learned that there are 140 Neighborhoods in these four Boroughs.

We looked at the population per Borough and at the number of Neighborhoods in each Borough. We got the hospital names located in each Borough through the Foursquare API and represented those on a map, in order to be able to better locate them. We realized that there are 41 Neighborhoods that have hospitals but there are 100 Neighborhoods that do not have any hospital: people living in those areas face a highest risk of not being treated. The Neighborhoods without hospital would obviously face a more challenging situation in case of a raise of Covid-19 infected people, and specially the Borough of The Queens, since it is the Borough with the highest number of Neighborhoods without hospital.

Then, we have created clusters using the K-Means method, and we identified 3 clusters with one outsider that is Murray Hill in Manhattan: this is the Neighborhood with the highest hospital-beds capacity and therefore the best equipped Neighborhood to fight the Covid-19 pandemic. The "Cluster 0" is the cluster with the lowest hospital-beds capacity: efforts should be made to provide more beds and equipment in this cluster, otherwise it will face a challenging situation to threat Covid-19 infected people who need to be hospitalized. Additionally, a particular attention should be paid to the Neighborhoods of Queens Village, Williamsburg, Bensonhurst and Crown Heights, because they have the lowest beds per thousand people capacity and the lowest ICU beds per thousand bed capacity.

Then, we focused at the Borough level and we compared data with the number of Covid-19 positive cases in each Borough. We analyzed the percentage of infected people in each Borough, compared to the population in each Borough and we noticed that The Queens is the Borough with the highest number of positive Covid-19 cases but the Bronx is the Borough with the highest number of positive cases compared to its population. Adversely, Manhattan is the Borough with the lowest percentage of positive Covid-19 cases. So, Manhattan is in a pretty good situation since it has the lowest number of positive cases and the highest hospital-beds capacity, but The Bronx and The Queens are not in a real good shape since they have a high number of positive cases and a relatively small hospital-beds capacity.

7. Limitations of this Analysis.

However, there are limitations to this analysis. The Covid-19 situation is not static and is always evolving with time, therefore this analysis only represent the situation at a particular instant.

Furthermore, another limitation comes from the data sets that we used: the data set used to collect the data with the New-York City Boroughs include four Boroughs, whereas now there is a new Borough named Staten Island that was not part of the data set. This means that this Borough and its Neighborhoods was not considered for this analysis. And the data with the New-York City population came from a Wikipedia page that might not contain the latest population data.

Another limitation comes from the Foursquare API and the list of hospitals that we obtain from it: we are unsure if all the hospitals are listed in Forusquare, and we might not have the latest hospital list in this analysis.

Also, at the end, we simplified our analysis by focusing on the number of positive Covid-19 cases at the Borough level, because of lack of data at the Neighborhood level. Focusing on the Neighborhood level could have give us a better idea of the local situations and challenges.

8. Recommendations for Future Research.

The first recommendation would be to use up-to-date data sets with the latest information about the New-York City population, total number of Boroughs and hospitals list, as well with the data about the number of Covid-19 infected people in each Neighborhood.

The second recommendation is in regards with the second part of our analysis: it would be more accurate to write a code that is not “static” where we can change the parameters in function of the evolution of the Covid-19 situation. This would give us the exact situation of the pandemic throughout the time, instead of having a picture of the situation at a specific moment, especially by creating sort of “interactive” maps.

9. Conclusion.

As a conclusion, we have performed an analysis to find the Neighborhoods that are best equipped to fight the Covid-19 pandemic, especially by clustering the Neighborhoods using the K-Means Clustering, which is an unsupervised machine learning algorithm. We have identified the best equipped Neighborhood in terms of hospital-beds capacity as being Murray Hill, located in the Borough of Manhattan, which constitutes itself a cluster. We have done several data analysis and statistics, accompanied with visualizations to help us represent the information.

We pushed further our analysis by making comparisons between the hospital-beds capacity in each Borough with the number of positive Covid-19 cases in each Borough in order to get a more realistic representation of the situation.