Capstone Project: Find Best Neighborhood to Fight Pandemic in New York City

## **Business problem**

Right now, New York is one of the worst hit state by COVID-19 in USA. New York city is at the center of the pandemic. The hospitals are already stretched thin with patients overflowing. According to New York Times report and the WHO (at the moment of writing) death toll was 28,134, case count topped 358,099.

I wanted to create something useful which would give some insight on this situation. In this project we are 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.

This would not only give insights of the current situation but also help the government to plan further actions. COVID-19 has become a problem not only for US but for whole world. So, these analyses particularly focus on New York City but can further be used to many other cities for analysis or data of beds in a hospital and the changes required as such.

## **Data**

We will be collecting data from following sources:

1. 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).

1. We are going to get population data from Scraping Wikipedia.

* Data source: Wikipedia page of NYC neighborhood.
* We are going to go through each of the links of neighborhood and find the population of each of them.

1. Hospital information is going to be fetched from foursquare API.

* Data source: foursquare API

1. Hospital bed information is going to be fetched from NYS Health Profile website.

* Data source: NYS Health Profile.

## **Approach**

This is our approach to resolve issue:

* Collect the New York city data from here.
* Collect population data for each neighborhood by scraping Wikipedia.
* Using Foursquare API, we will get hospitals for each neighborhood.
* Collect 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, get the json data from here, which will contain borough, neighborhood, latitude and longitude information.
* Neighborhood data in New York City will be collected from scraping the Wikipedia page. links given in the neighborhood section of the table will be visited via scraper, and find the population for each of them. Then data will be cleaned up and used to create a data frame containing borough, neighborhood and population.
* Hospitals per neighborhood information will be collected from foursquare API.
* We will collect bed and ICU capacity information from NYS Health Profile website. Will be using selenium-based scraping as this is a dynamic site.

## **Source code**

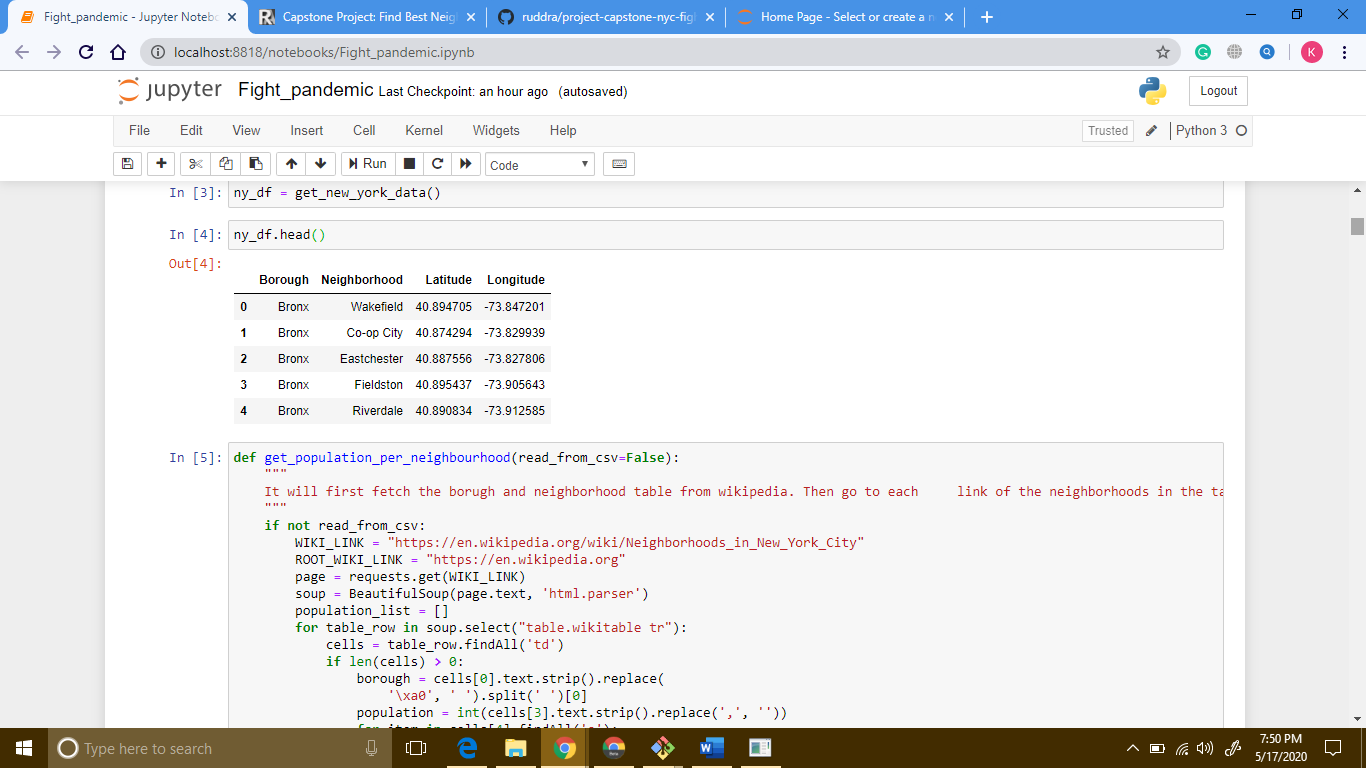
<http://localhost:8818/notebooks/Fight_pandemic.ipynb>

<https://github.com/K-THALA/Pandemic_Analysis/blob/master/Fight_pandemic.ipynb>

## **Methodology**

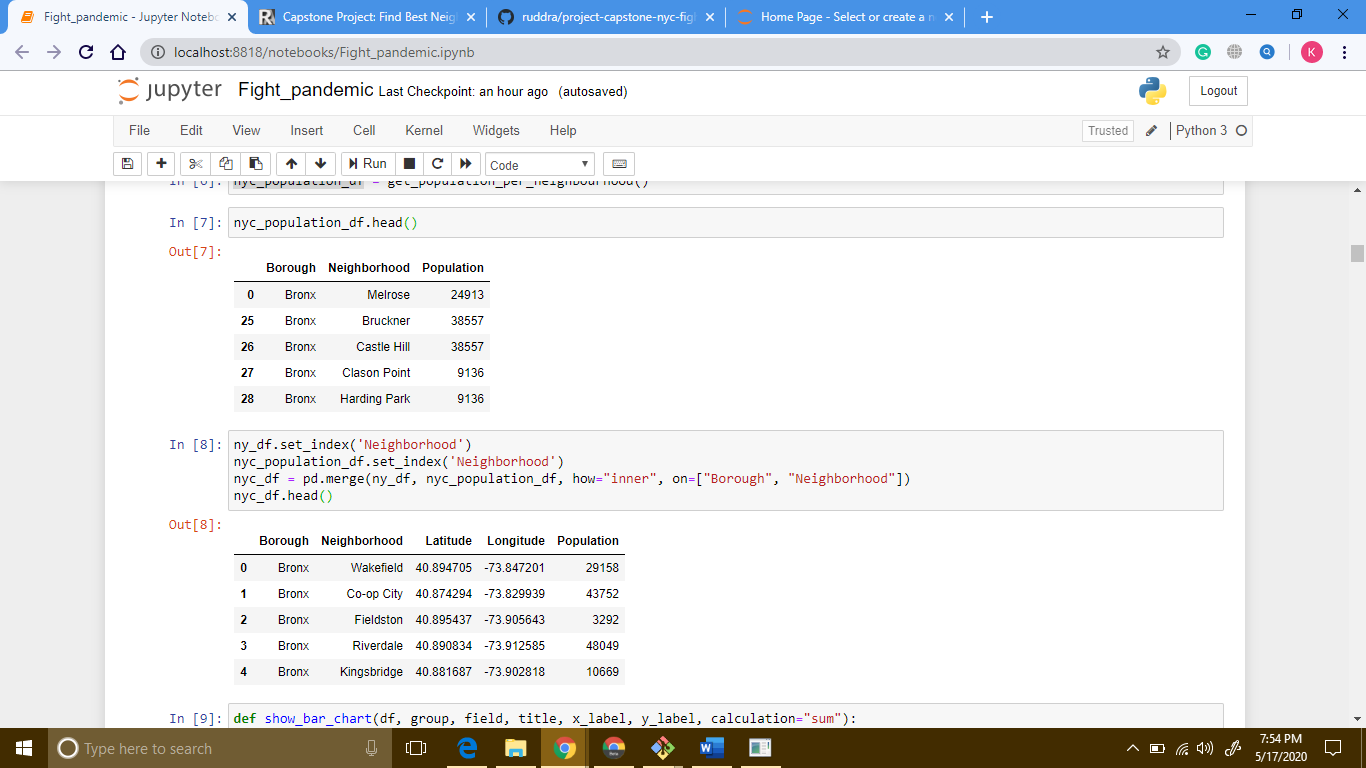
1. **Step one: New York city data with latitude and longitude**

We are using [**requests**](https://requests.readthedocs.io/en/master/) to get the json data from [NYC dataset](https://cocl.us/new_york_dataset) and stored it in a data frame.



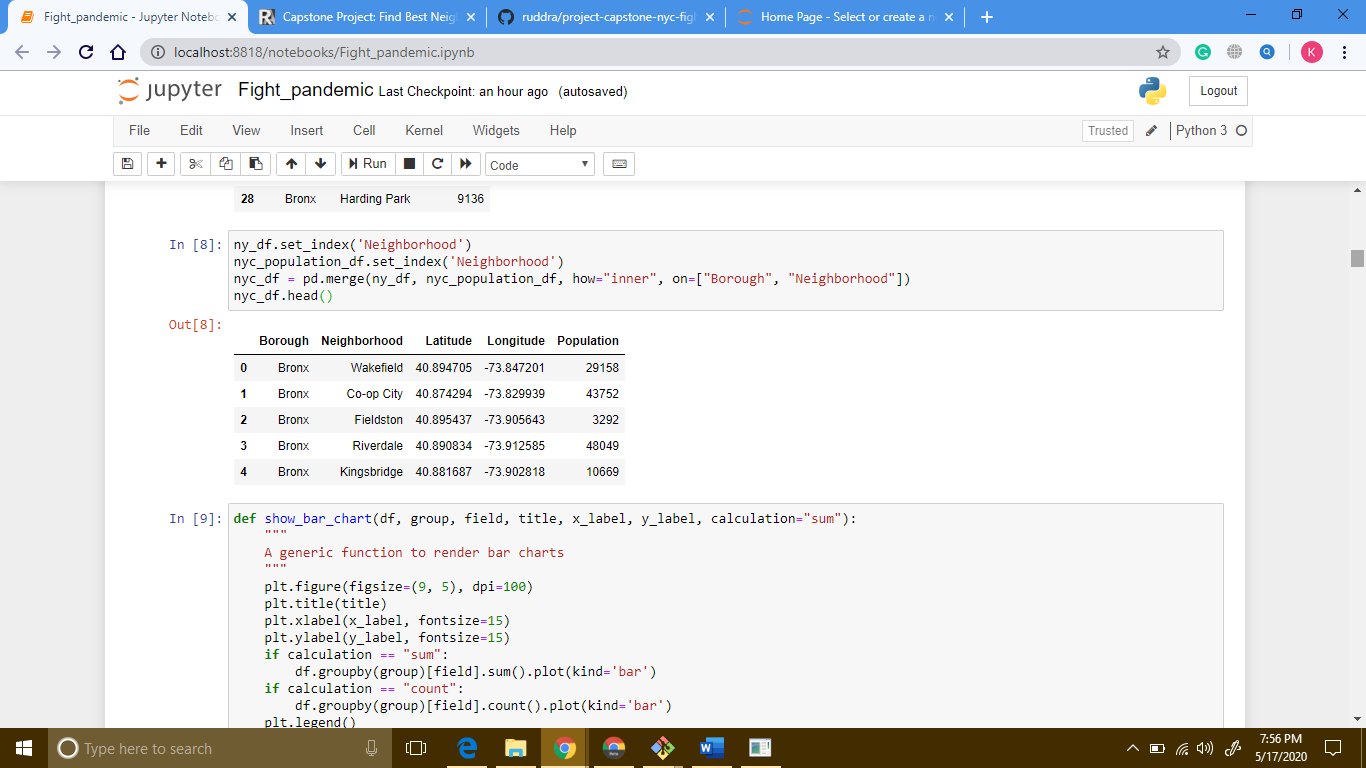
1. **Step two: New York city data with population**

Then we can use [**BeautifulSoup**](https://www.crummy.com/software/BeautifulSoup/bs4/doc/) 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.

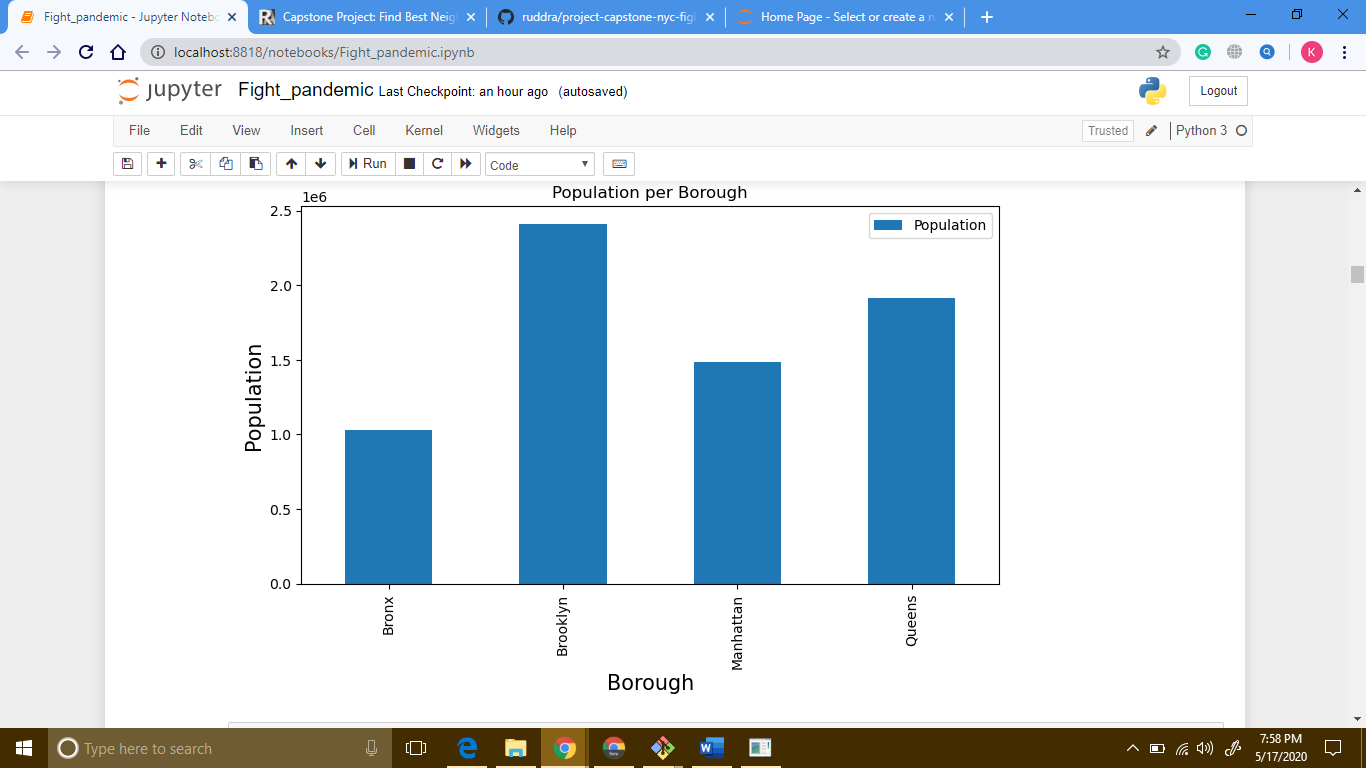


1. **Step three: combine step one and step two**

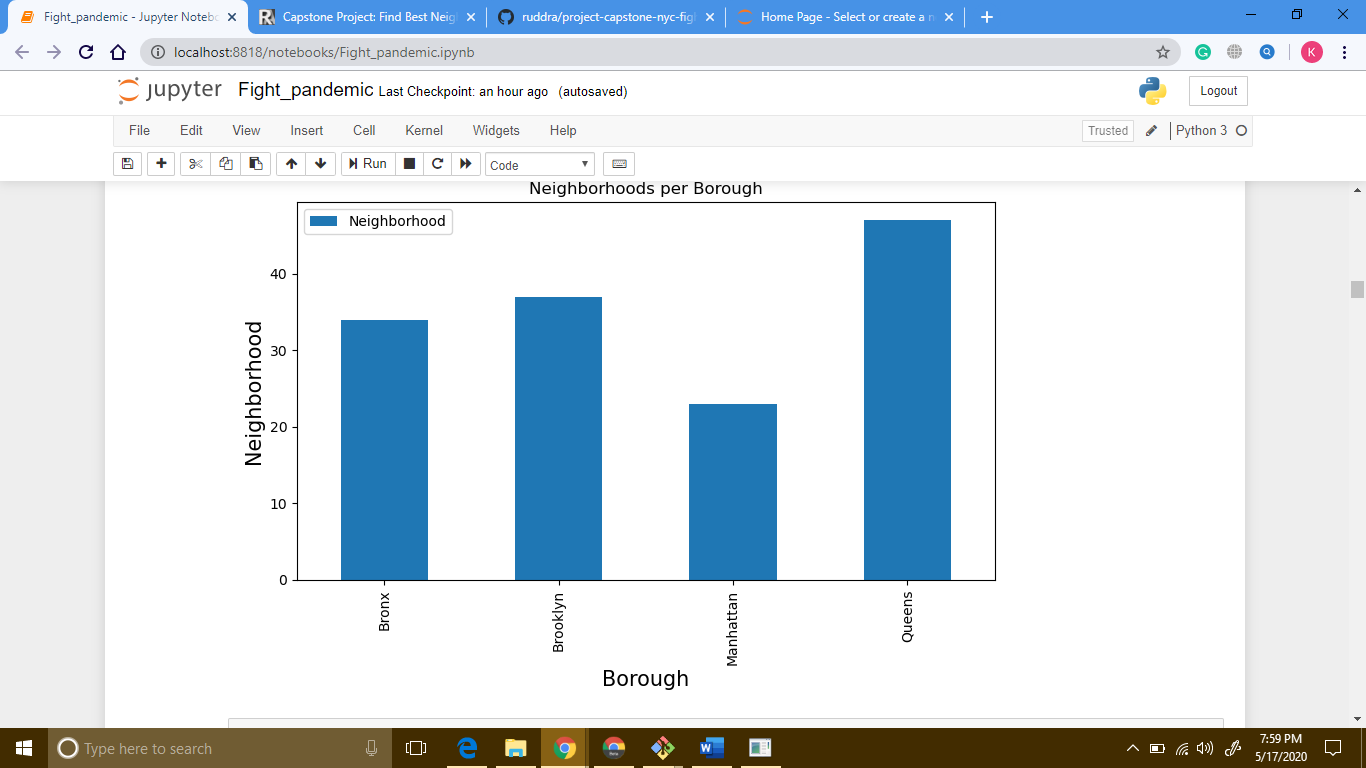
We can combine data frames from previous steps into one based on “neighborhood” and “borough”:



Here is a box chart of “Population” per “borough”:



Also, another box chart of “neighborhood” per “borough”:



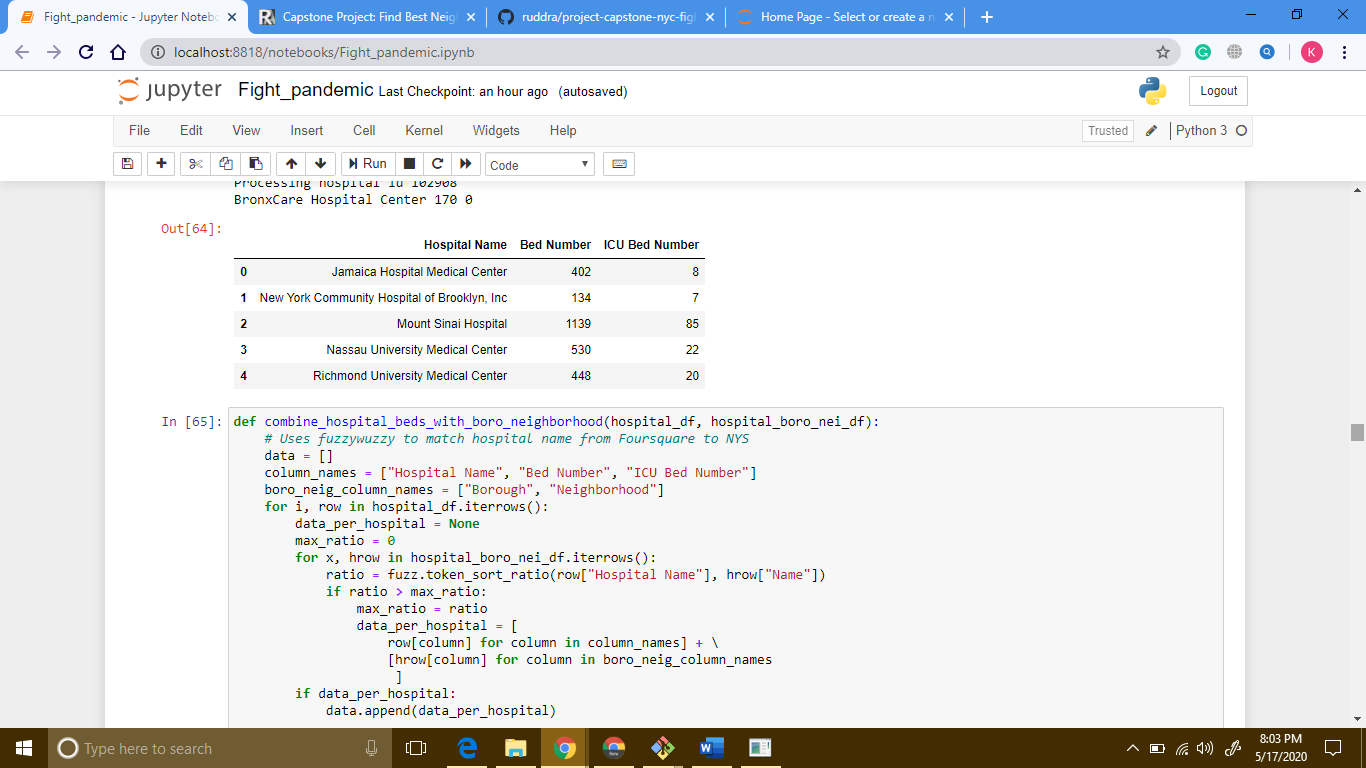
1. **Step four: collect hospital data from Foursquare**

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.



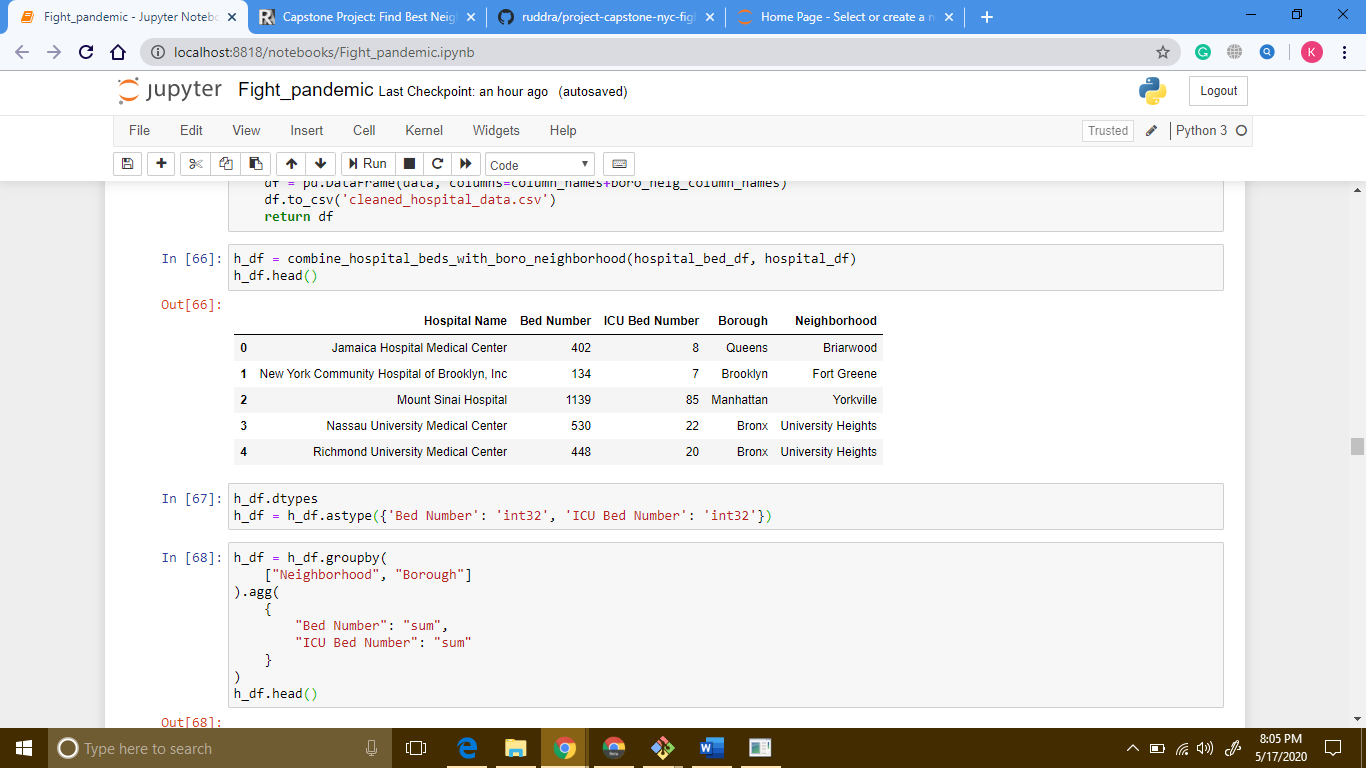
1. **Step five: collect hospital bed data from NYS Health Profile**

We can also collect hospital bed related data from [NYS Health Profile website](https://profiles.health.ny.gov/). We can scrap data by using [**Selenium**](https://www.selenium.dev/) with **Beautiful Soap**. 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:

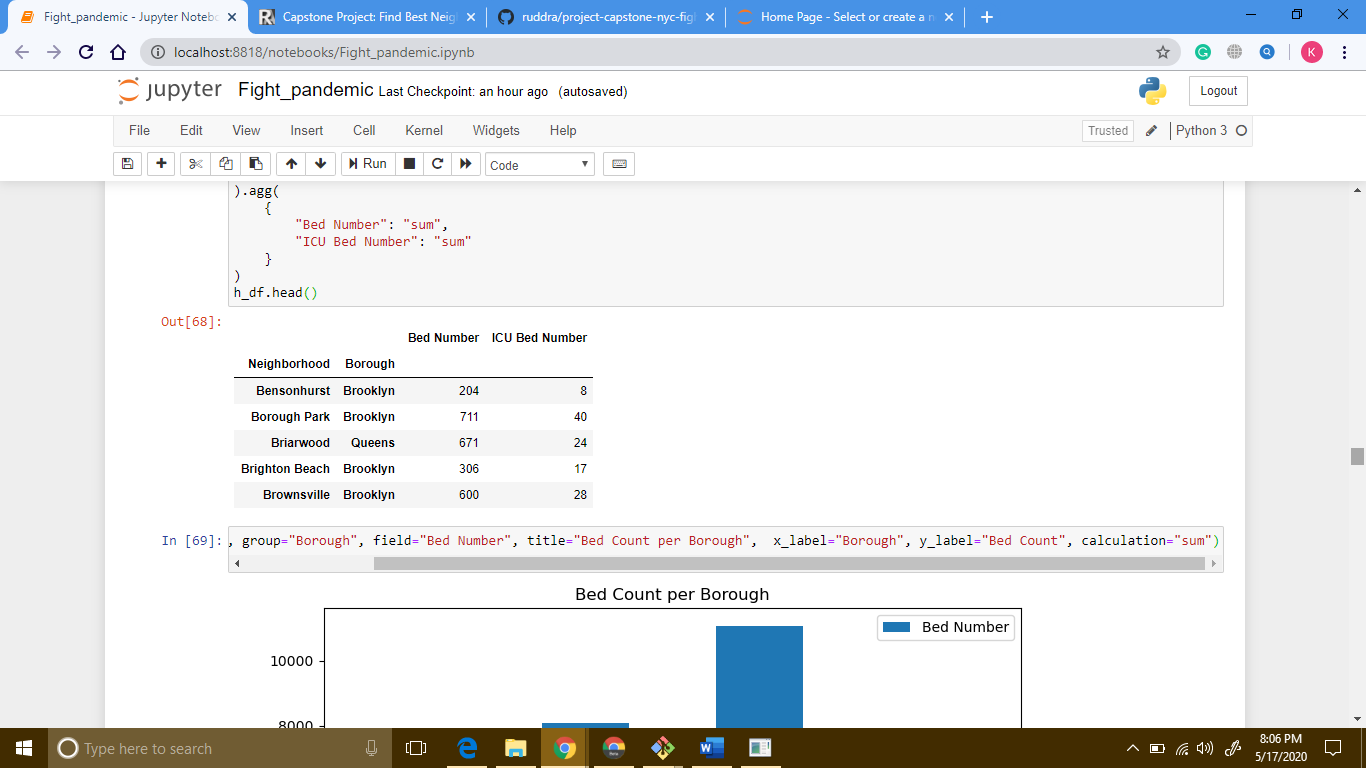


1. **Step six: combine step four and step five**

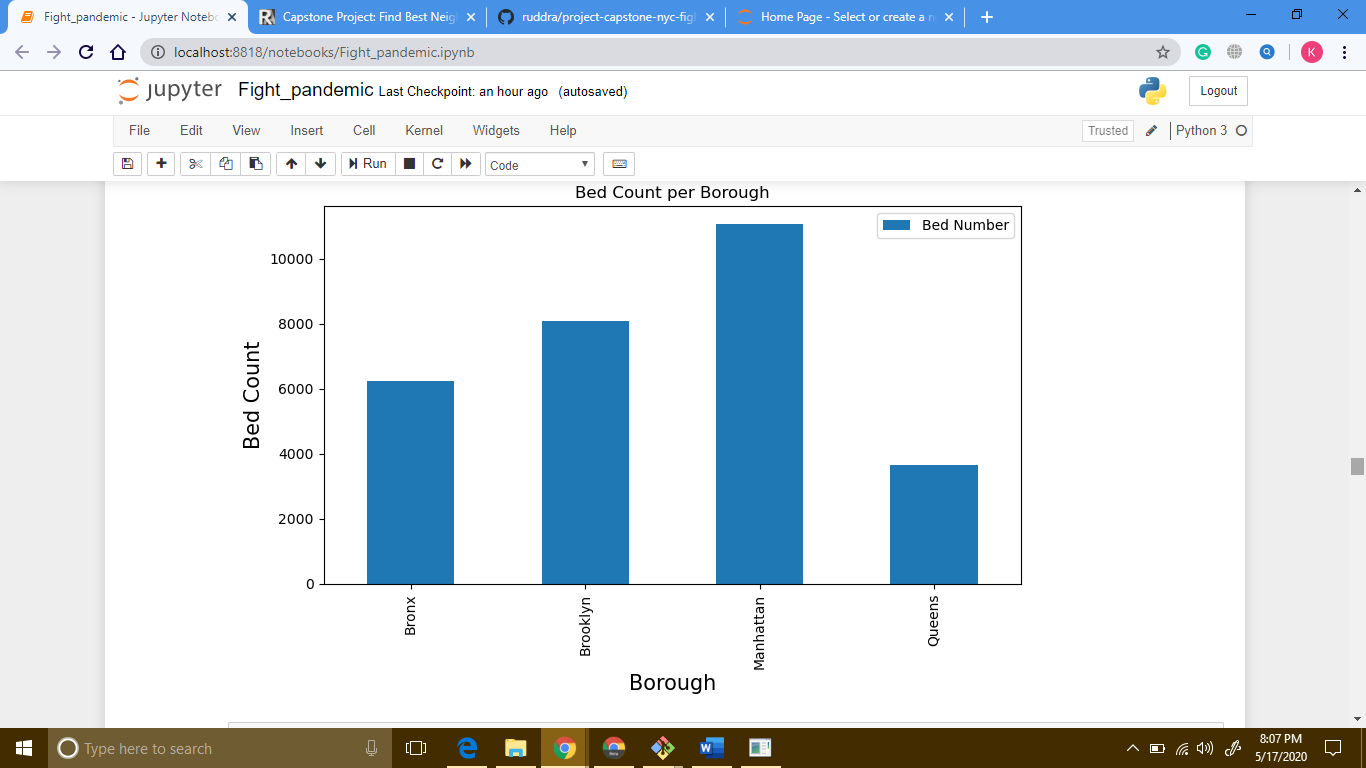
Now we are going to combine data from step four and step five. We are going to internally join the data frame based on “neighborhood” and “borough”.



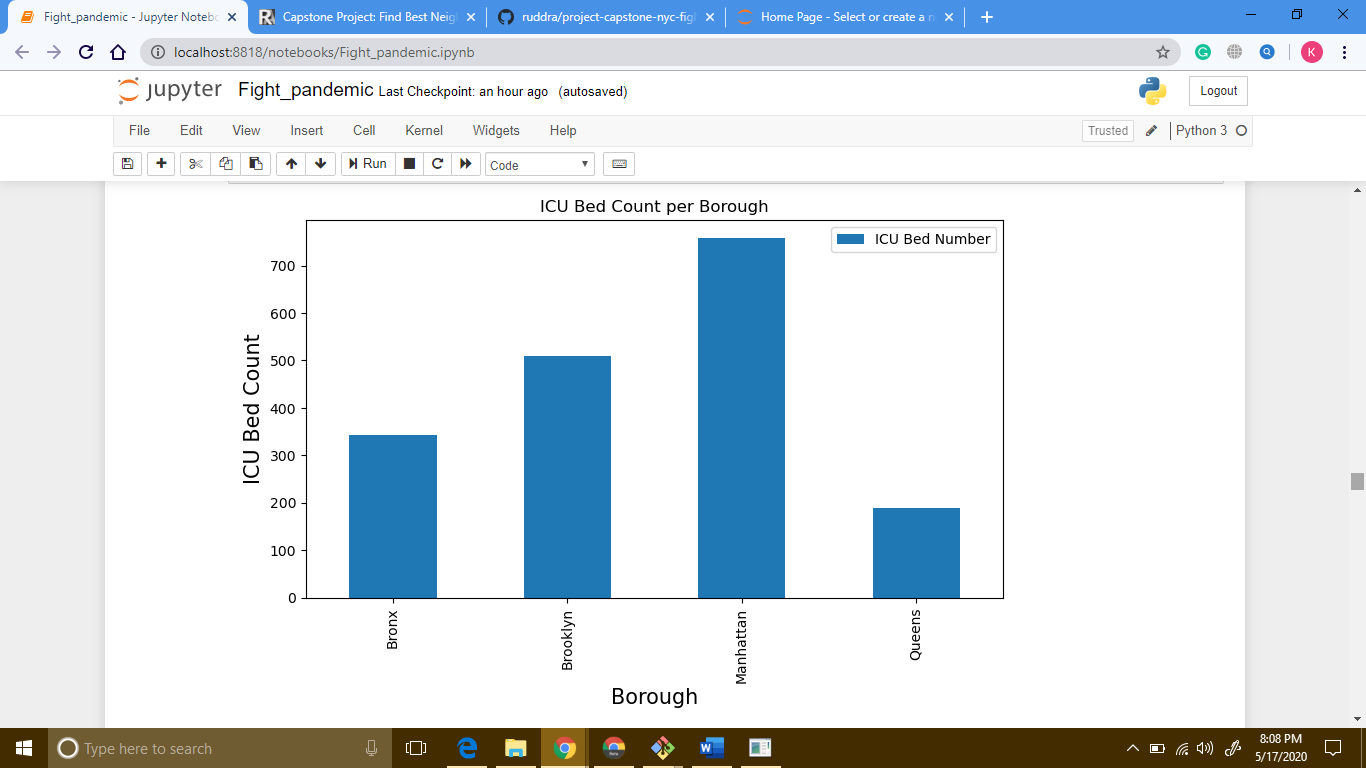
We are going to clean up the data a little bit and sum up bed count and icu bed count grouping by “neighborhood” and “borough”:



Here is a box charts of “bed count” per “borough”:

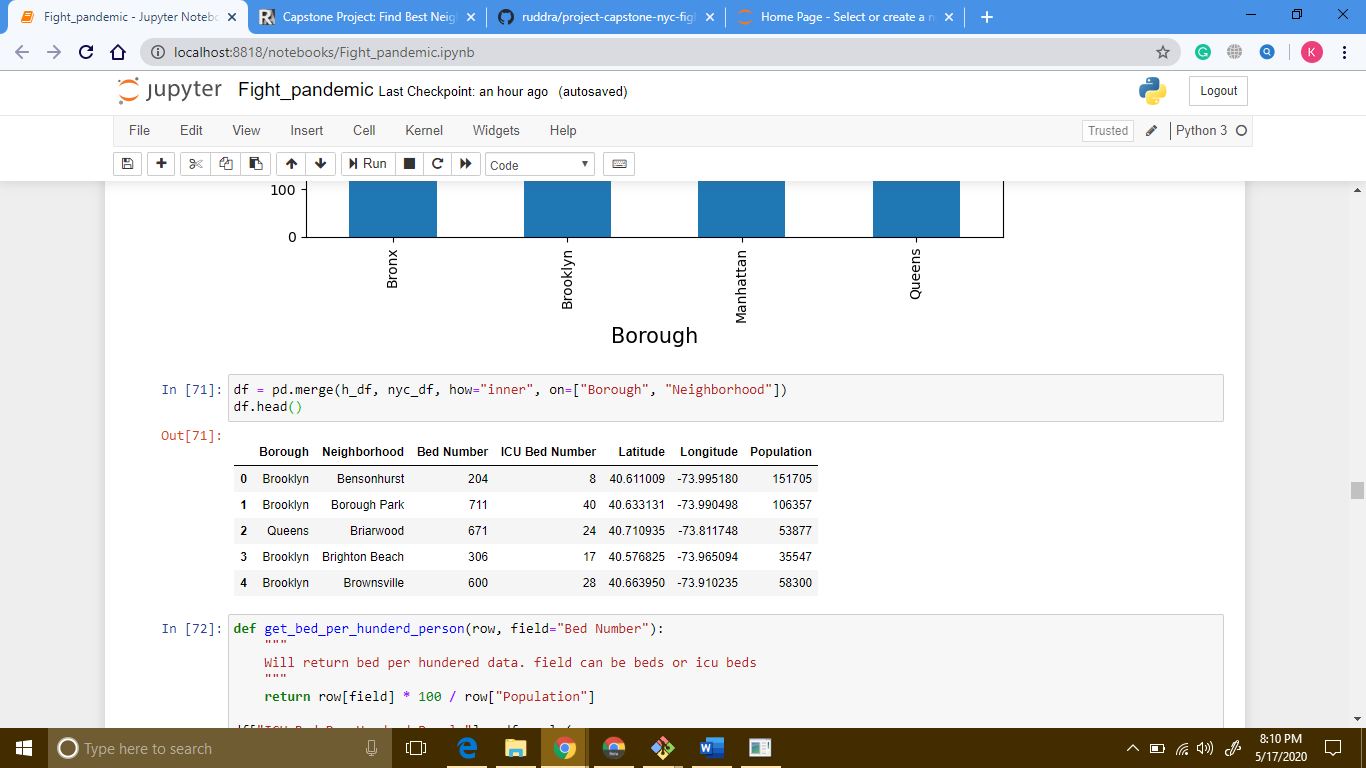


Also, another box charts of “ICU bed count” per “borough”:



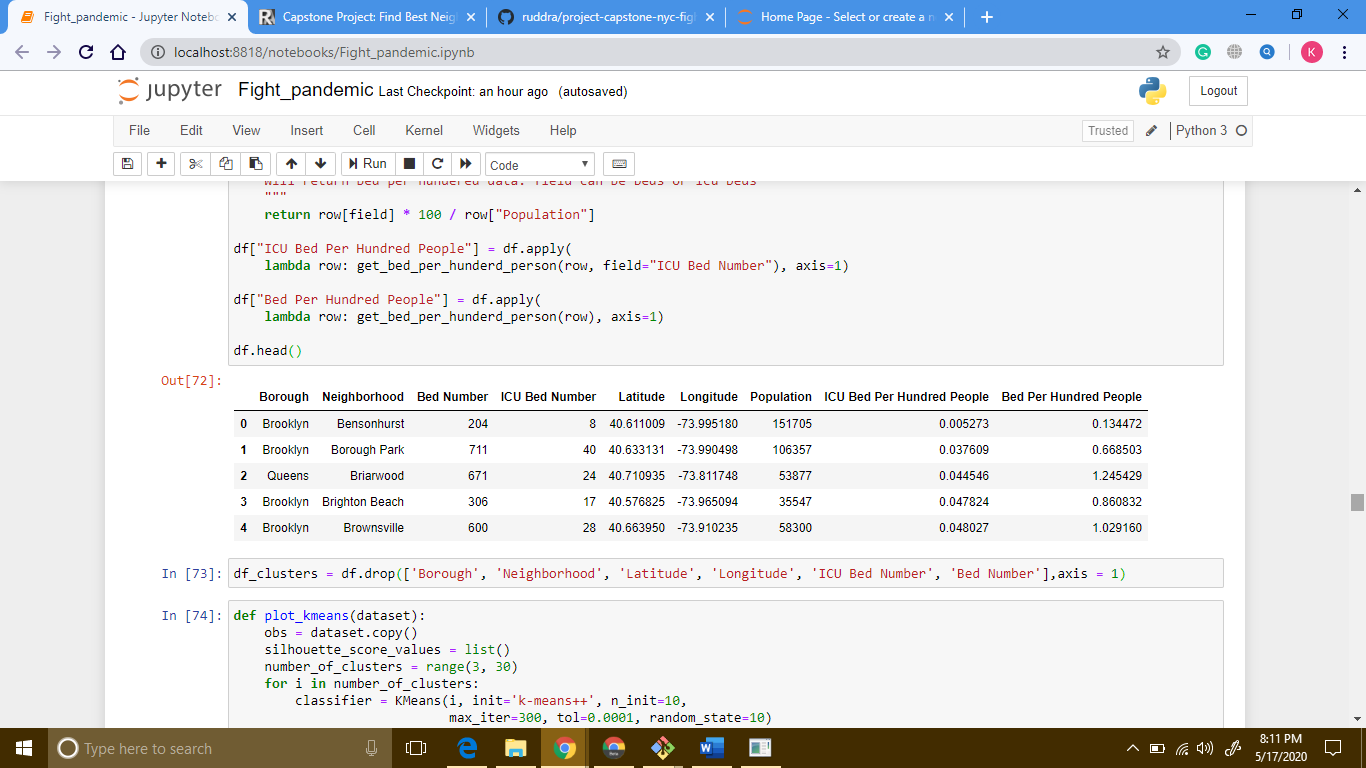
1. **Step seven: combine data from step three and step six**

Now we are going to combine data from step three and step six. Means, we are going to combine the population data with hospital bed count data. We are going to merge two data frames based on “neighborhood” and “borough”. New data frame looks like this:



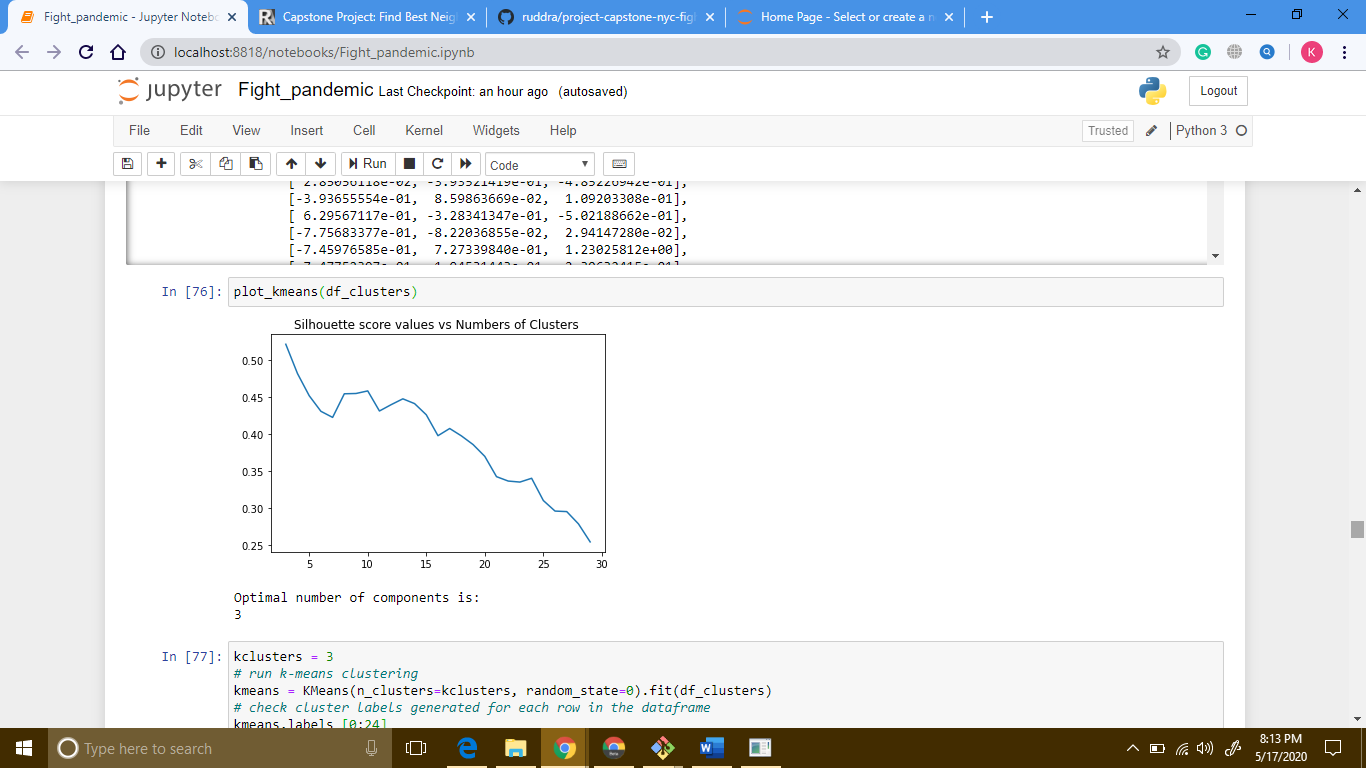
1. **Step eight: add bed and icu per hundred people to data frame**

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:



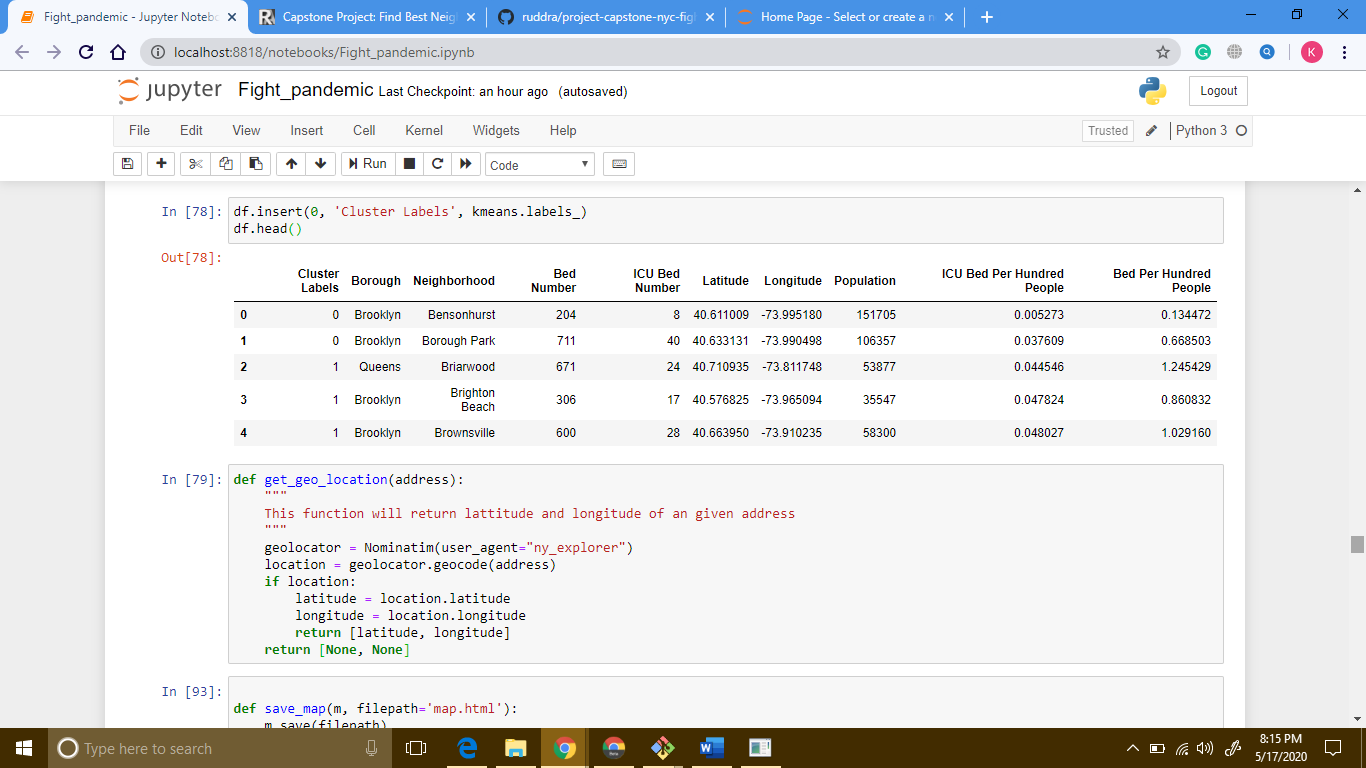
1. **Step nine: K-means clustering**

Now we are going to use k-means clustering to partition the data into **k** groups. we will be using [**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 3.



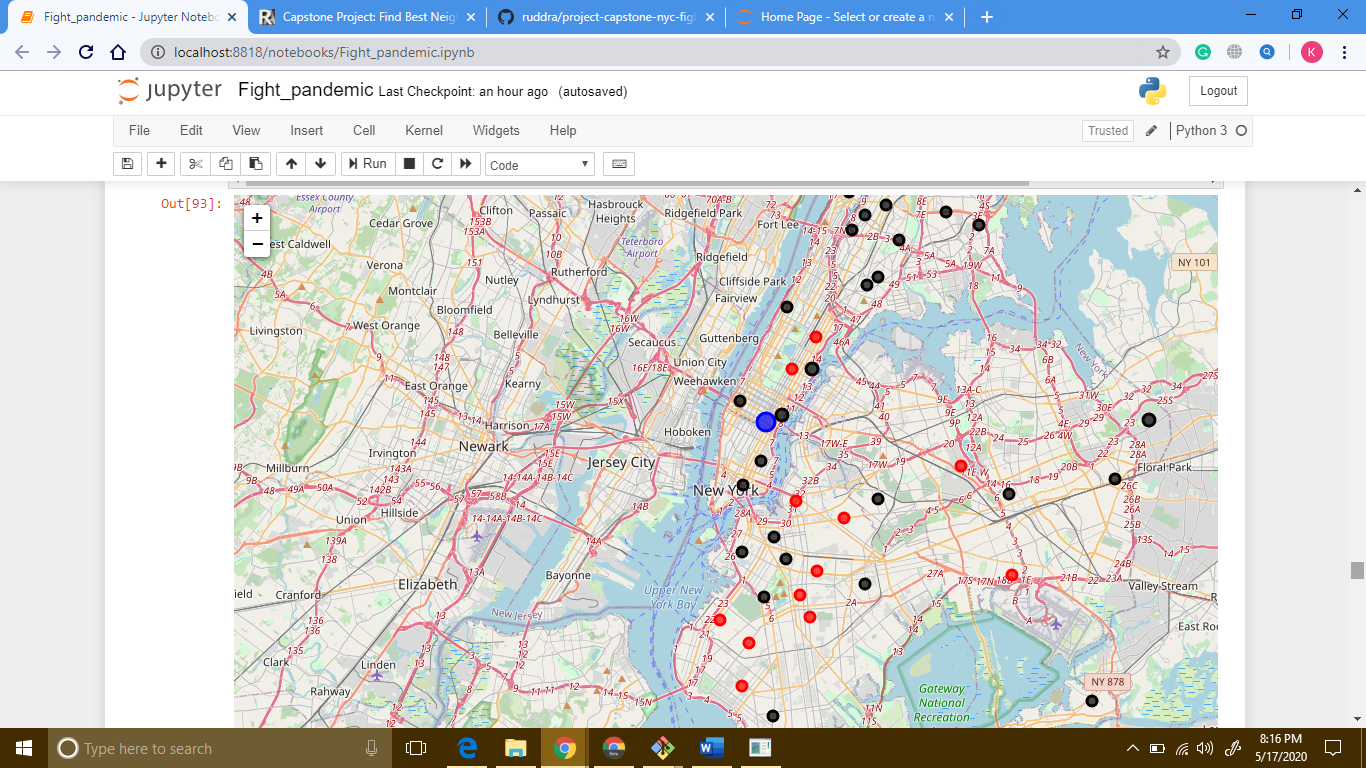
1. **Step ten: merge cluster labels with dataset**

After that, we are going to merge cluster labels of groups with data frames. The data frame looks like this:

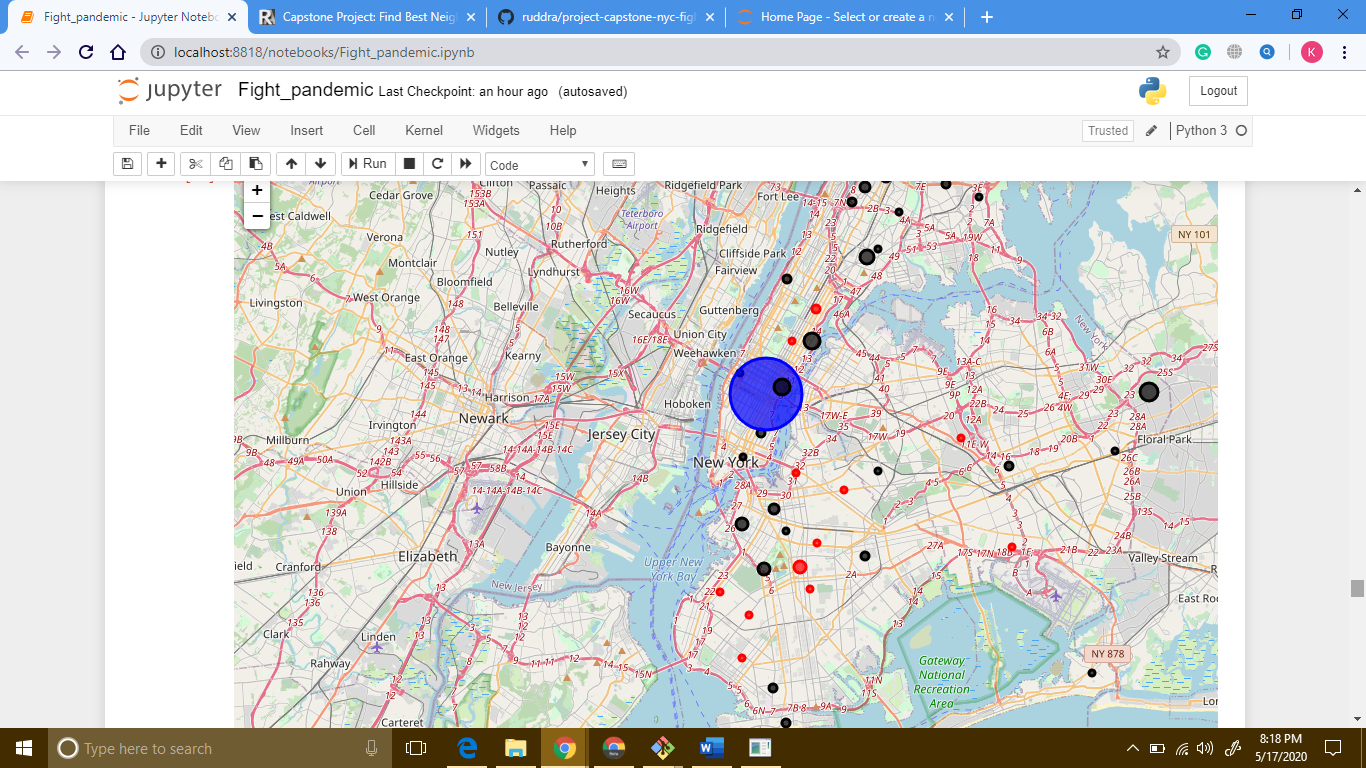


1. **Step eleven: visualize with folium**

Now, we are going to use [**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.

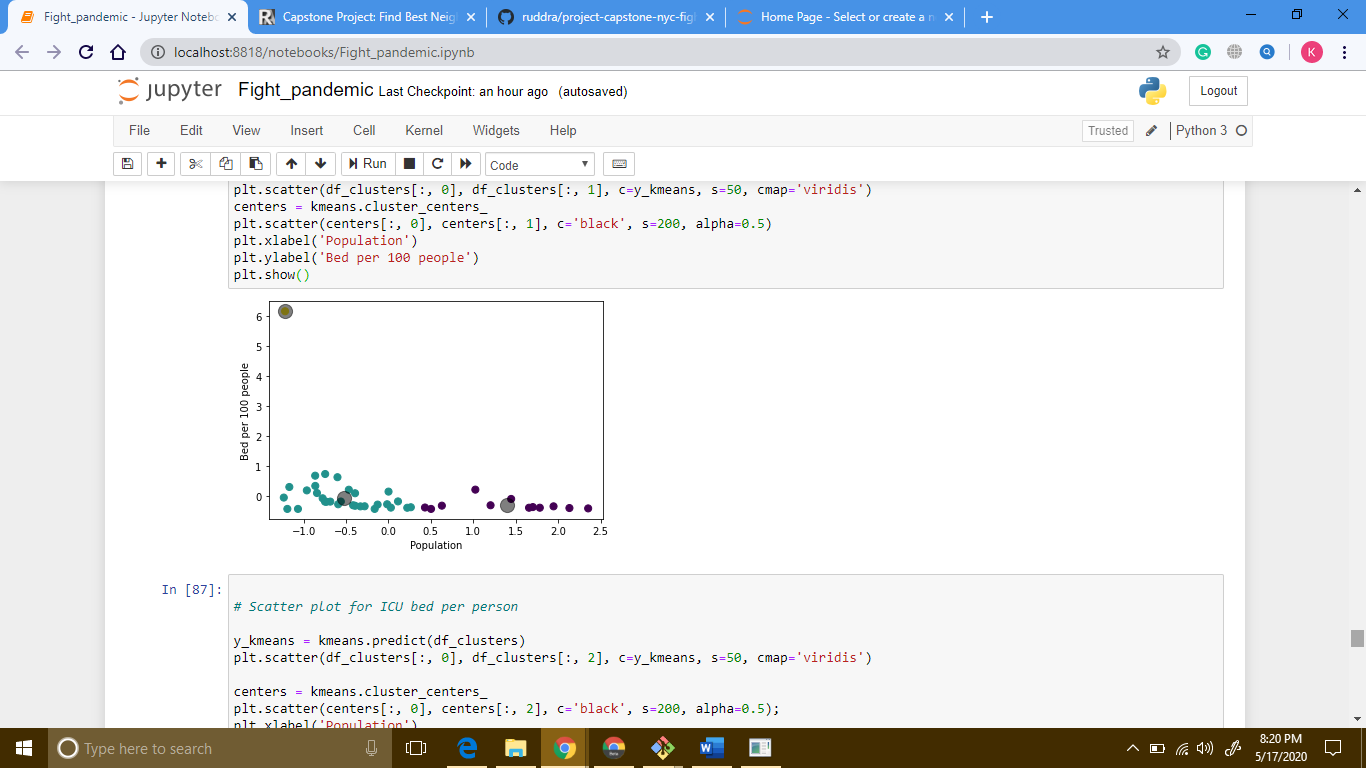


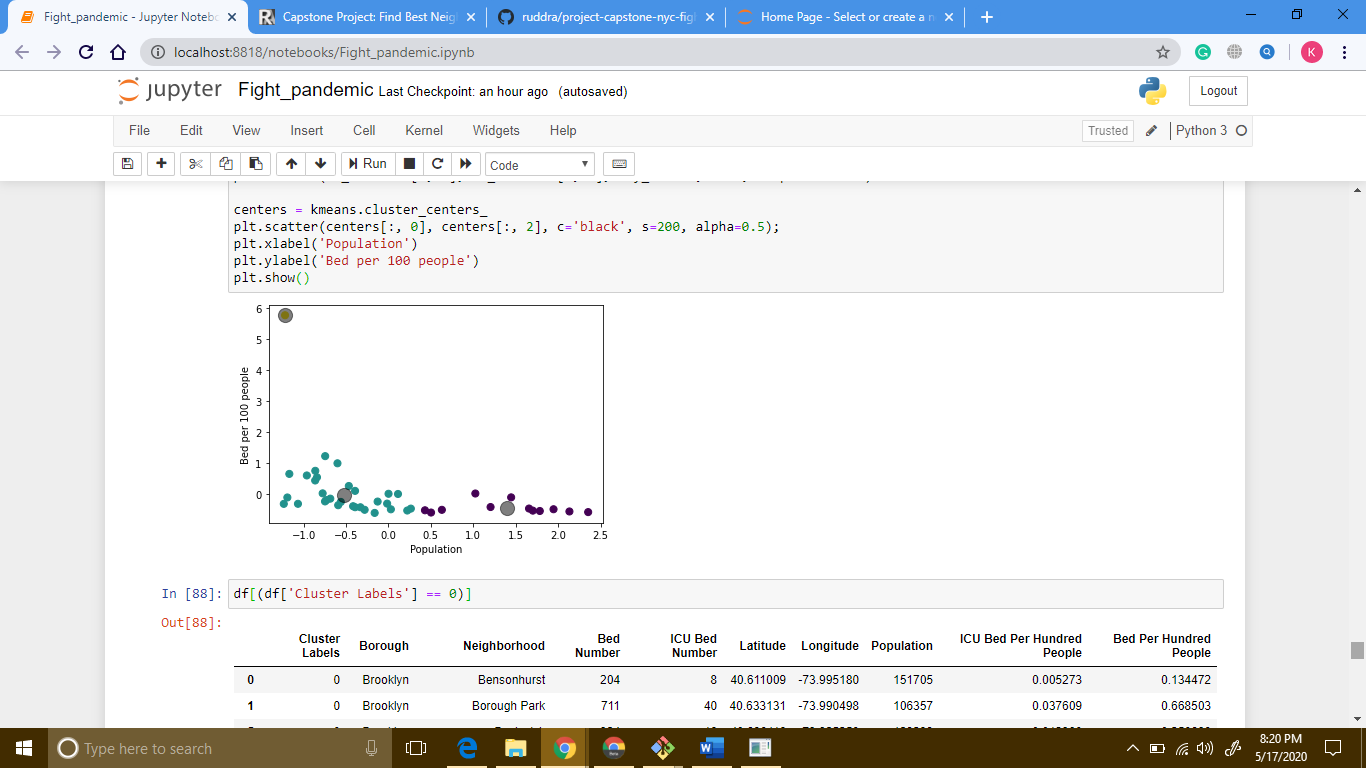
The second map illustrates the clusters where the radius of the Circle marker is proportional to icu beds per hundred people.



1. **Step twelve: use 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.



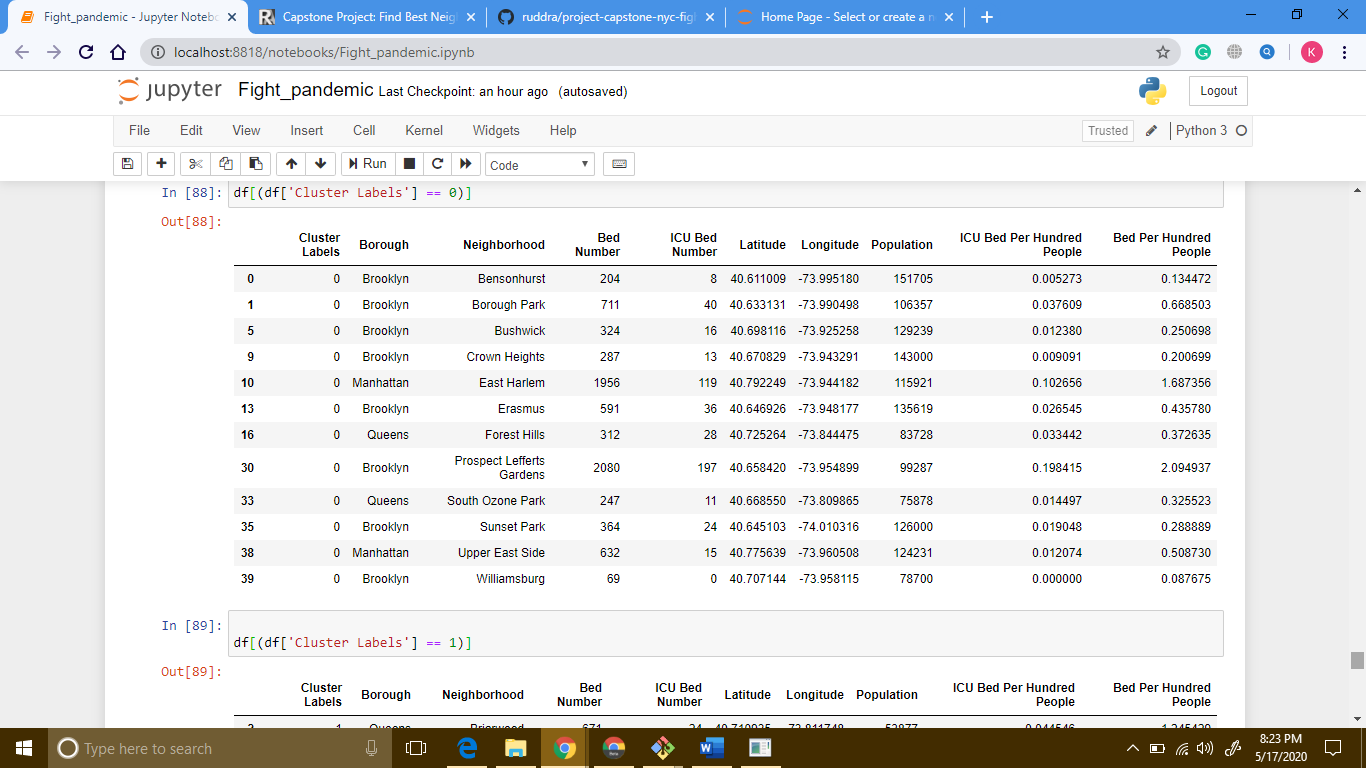


We can observe the obvious outlier here. This neighborhood has a high number of beds per people ratio. From maps above we can easily say that it is **Murray Hill**.

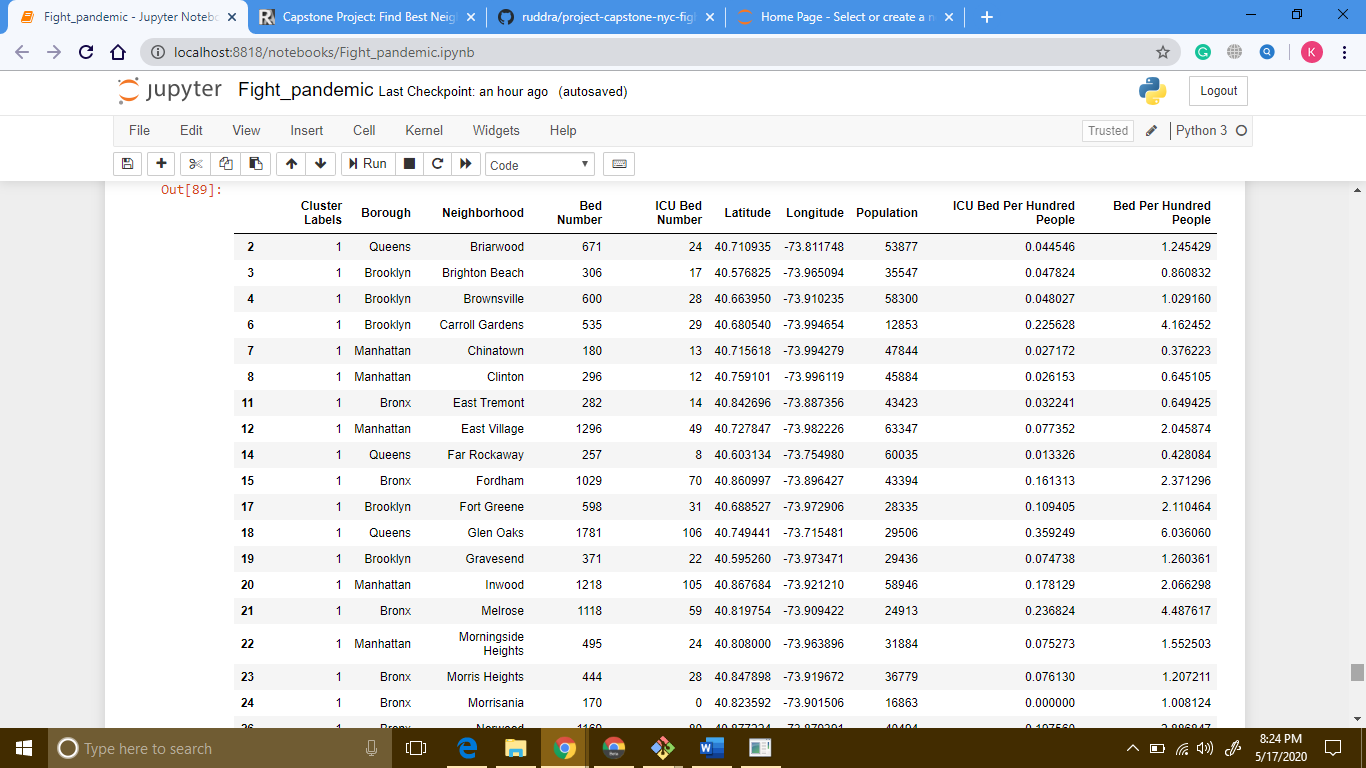
1. **Step thirteen: see which borough goes to which cluster**

Let us see which boroughs belong to which clusters.

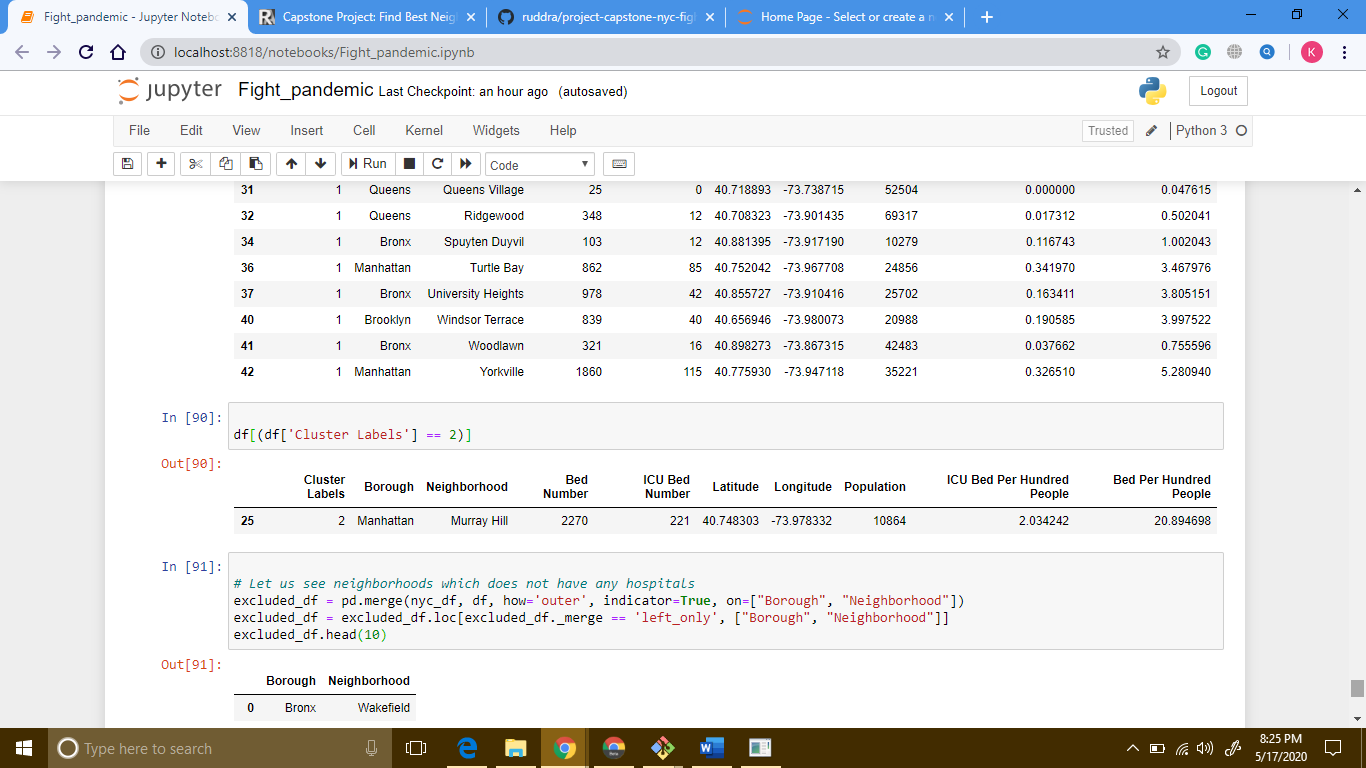
Here is the dataset for cluster 0:



Here is the dataset for cluster 1:

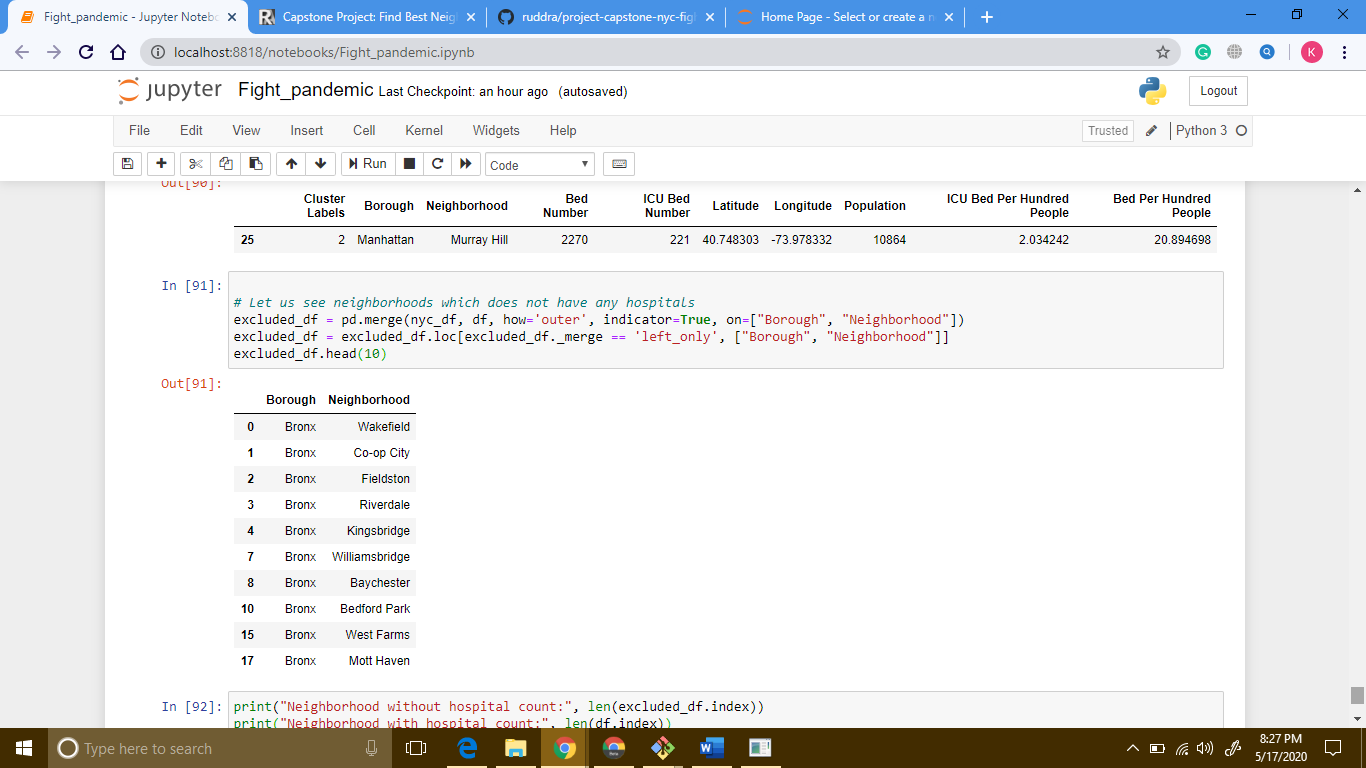


Here is the dataset for cluster 2:

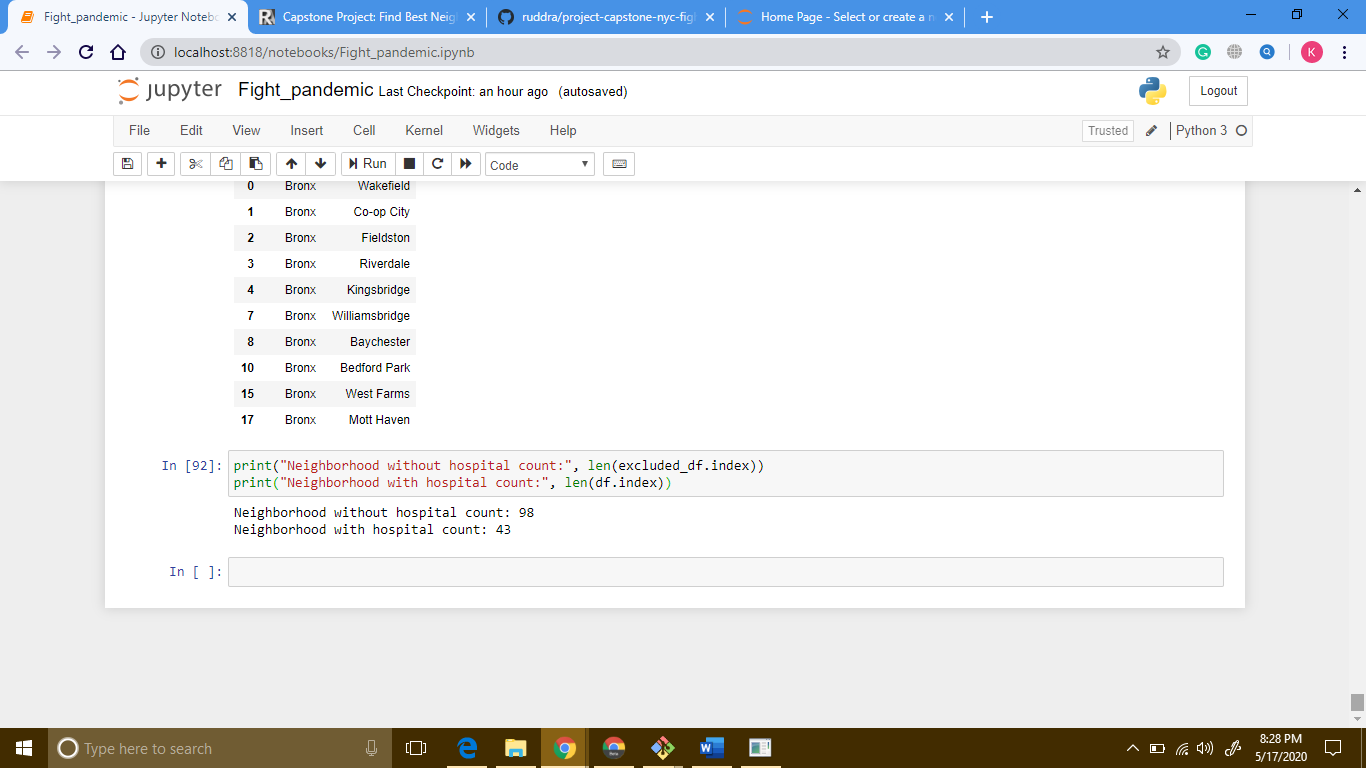


1. **Step fourteen: see neighborhoods without hospital**

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



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



## **Results and discussion**

During the analysis, three 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. For that reason, other maps could be utilized such as Google map or Open Street map.

**NYS Health Profile website** might 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 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.

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-Murray Hill** was chosen as the most well equipped (as per hospital bed count and icu bed count) area to battle pandemic.