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

## **Business problem**

At the present time, New York is one of the most noticeably awful hit states by COVID-19 in USA. New York city is at the focal point of the pandemic. The emergency clinics are as of now extended dainty with patients flooding. As indicated by New York Times report and the WHO (right now of composing) loss of life was 28,134, case check beat 358,099.

I needed to make something helpful which would give some knowledge on this circumstance. In this venture we will figure out which neighborhood is best arranged for this pandemic, by discovering the best proportion of emergency clinic beds per individual for every area in this city.

This would give bits of knowledge of the present circumstance as well as help the administration to design further activities. COVID-19 has become an issue for US as well as for entire world. Along these lines, these examinations especially center around New York City yet can additionally be utilized to numerous different urban communities for investigation or information of beds in an emergency clinic and the progressions required all things considered

## **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**

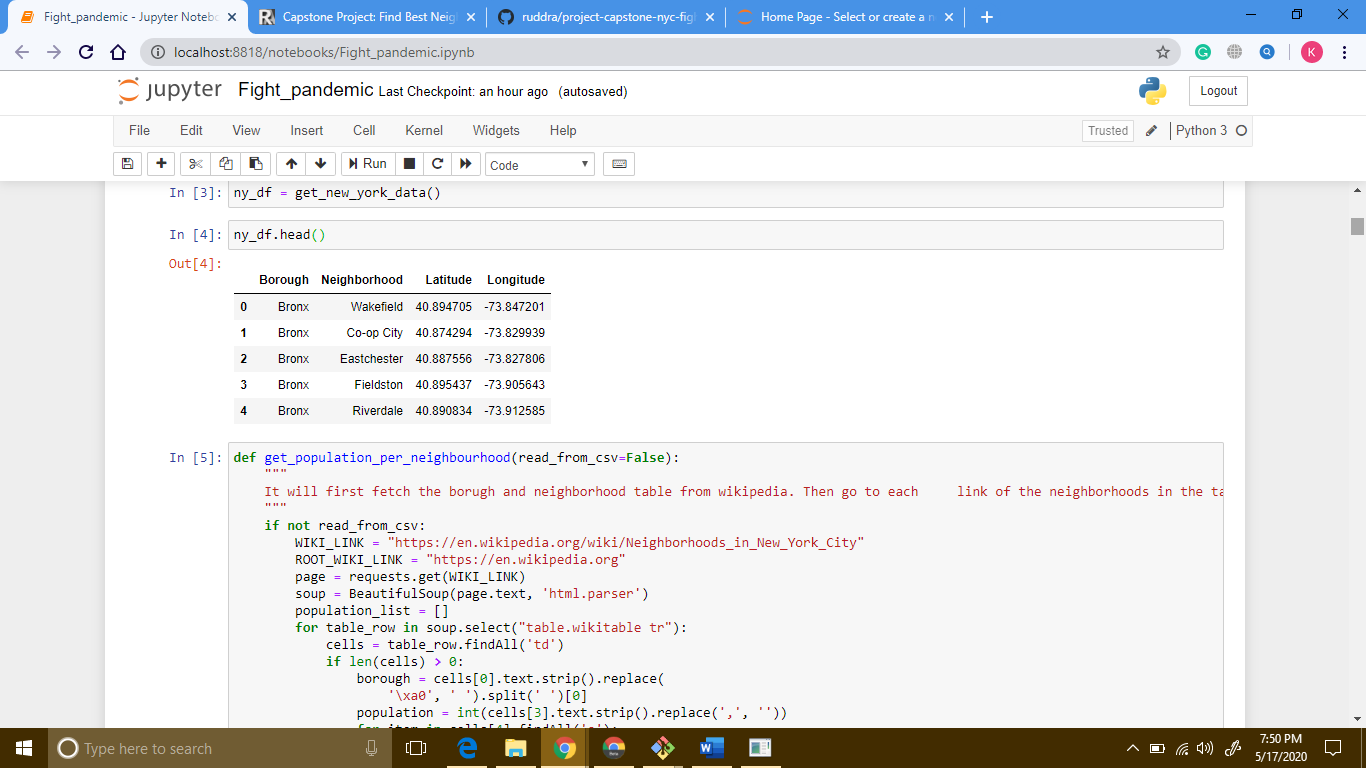
<http://localhost:8818/notebooks/Fight_pandemic.ipynb>

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

## **Methodology**

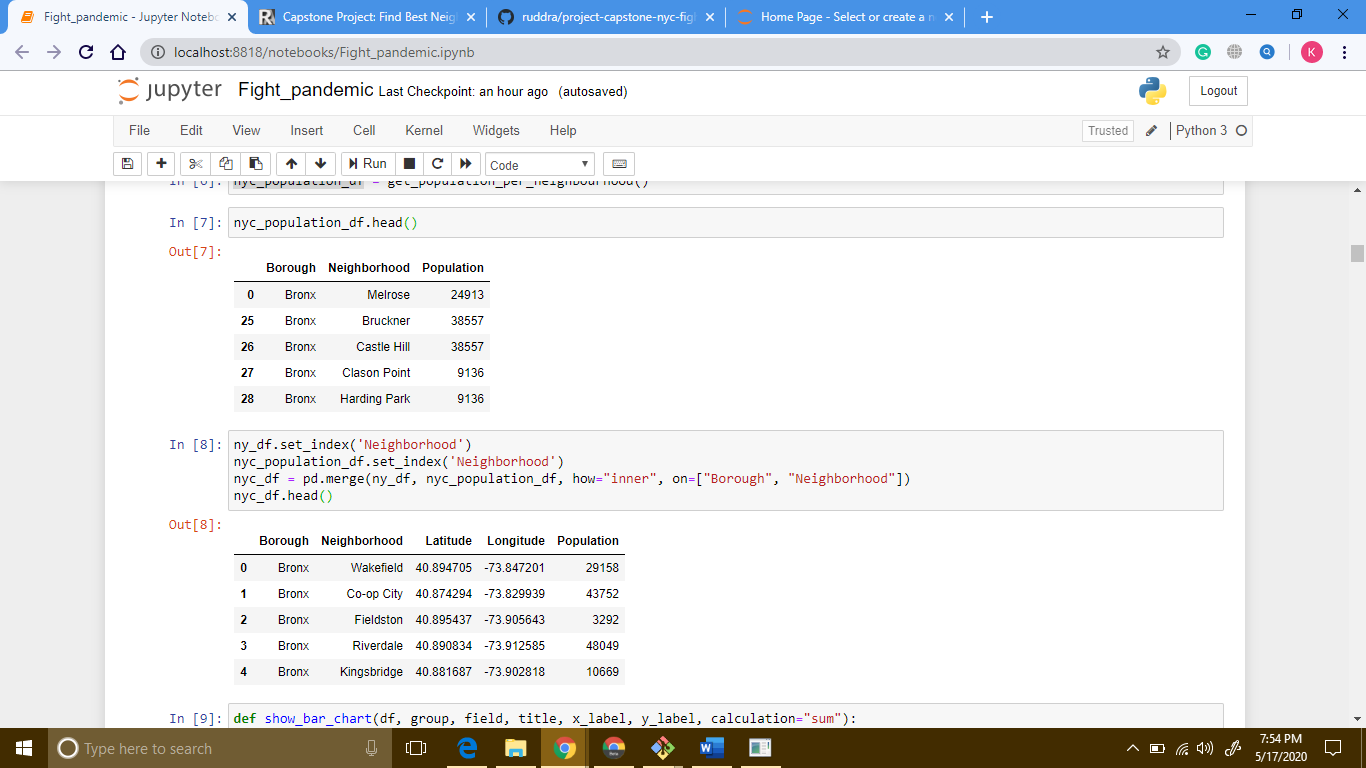
1. **Step one: New York city data with latitude and longitude**

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.



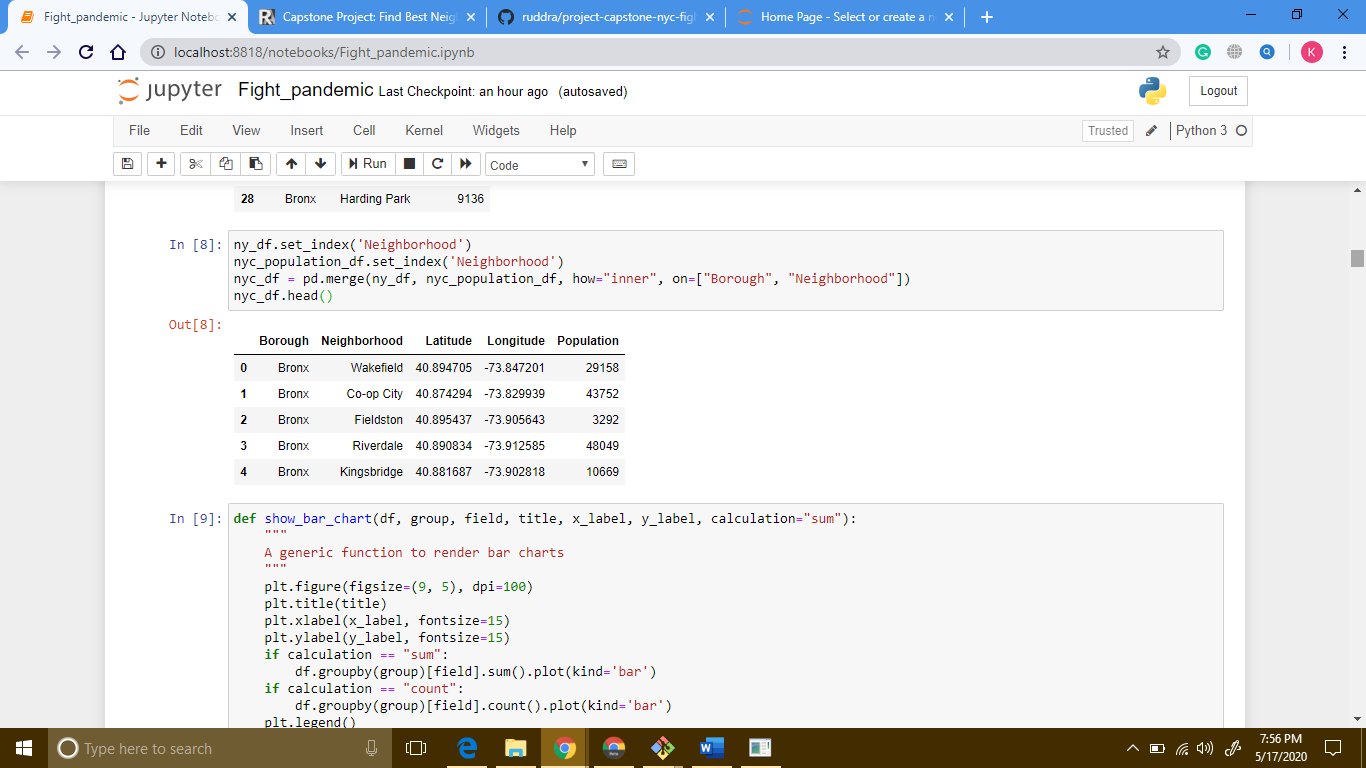
1. **Step two: New York city data with population**

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.

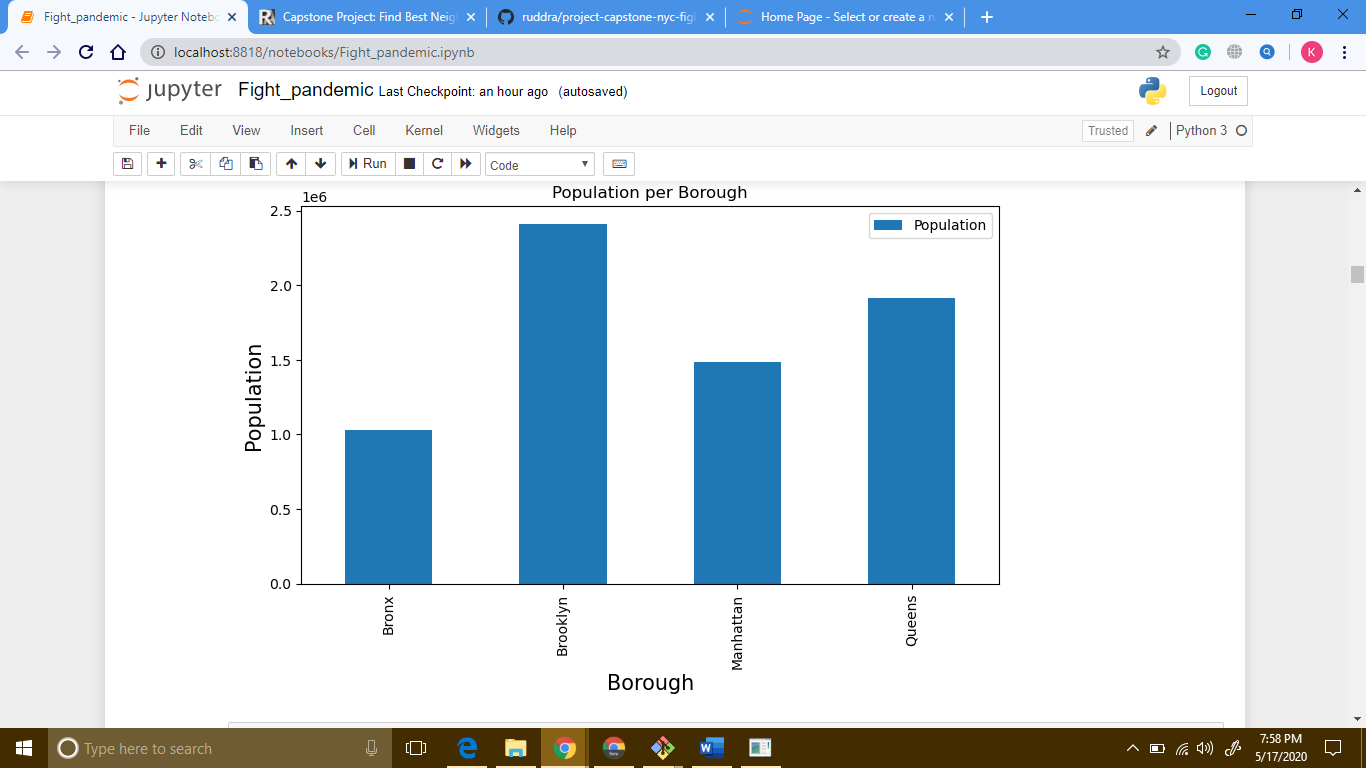


1. **Step three: combine step one and step two**

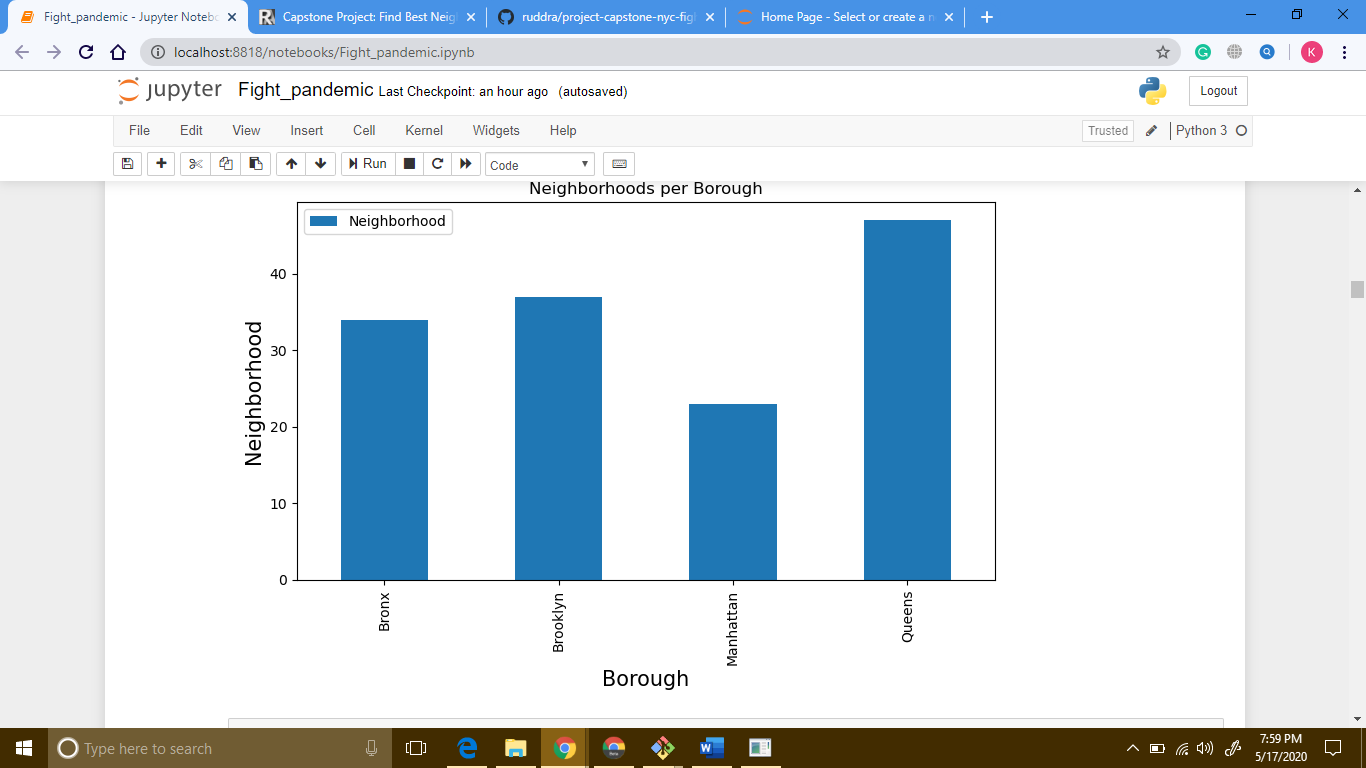
We can combine data frames from previous steps into one based on “neighborhood” and “borough”:



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



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



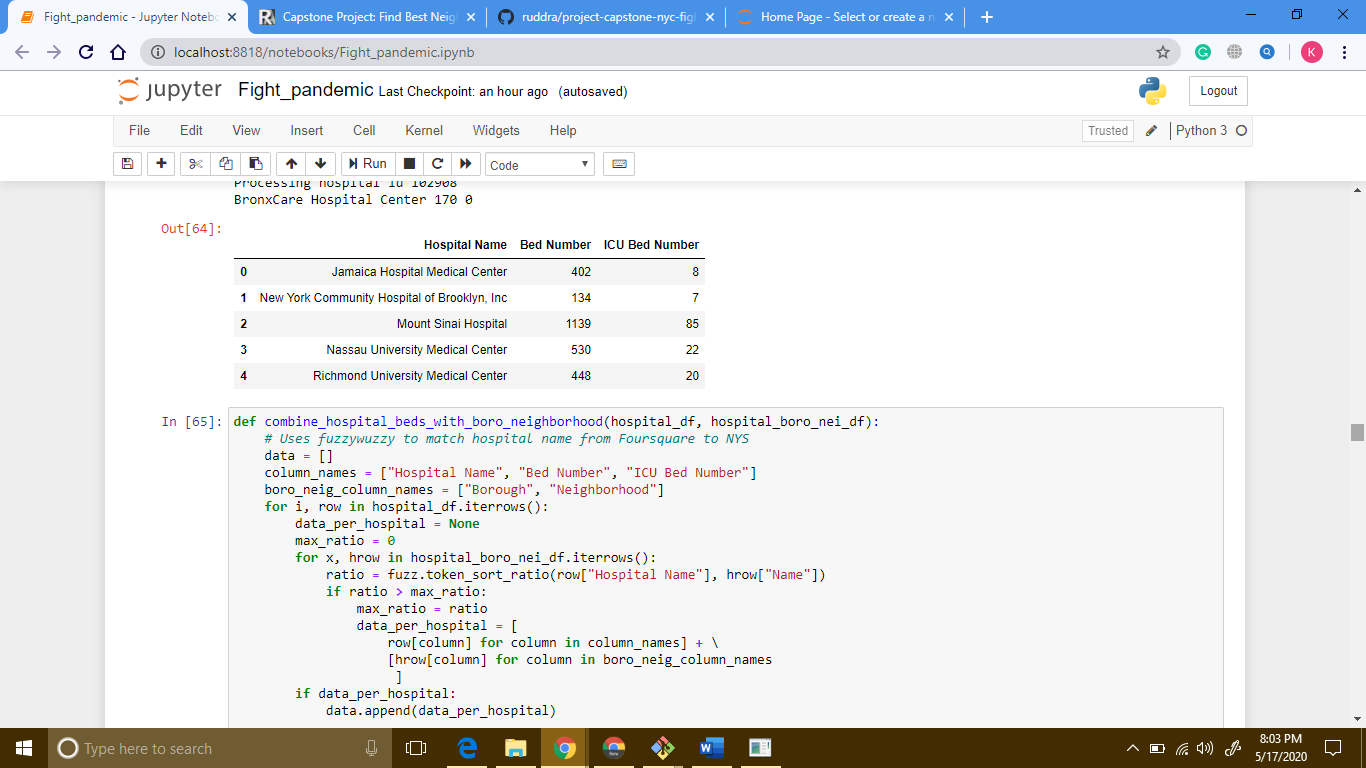
1. **Step four: collect hospital data from Foursquare**

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.



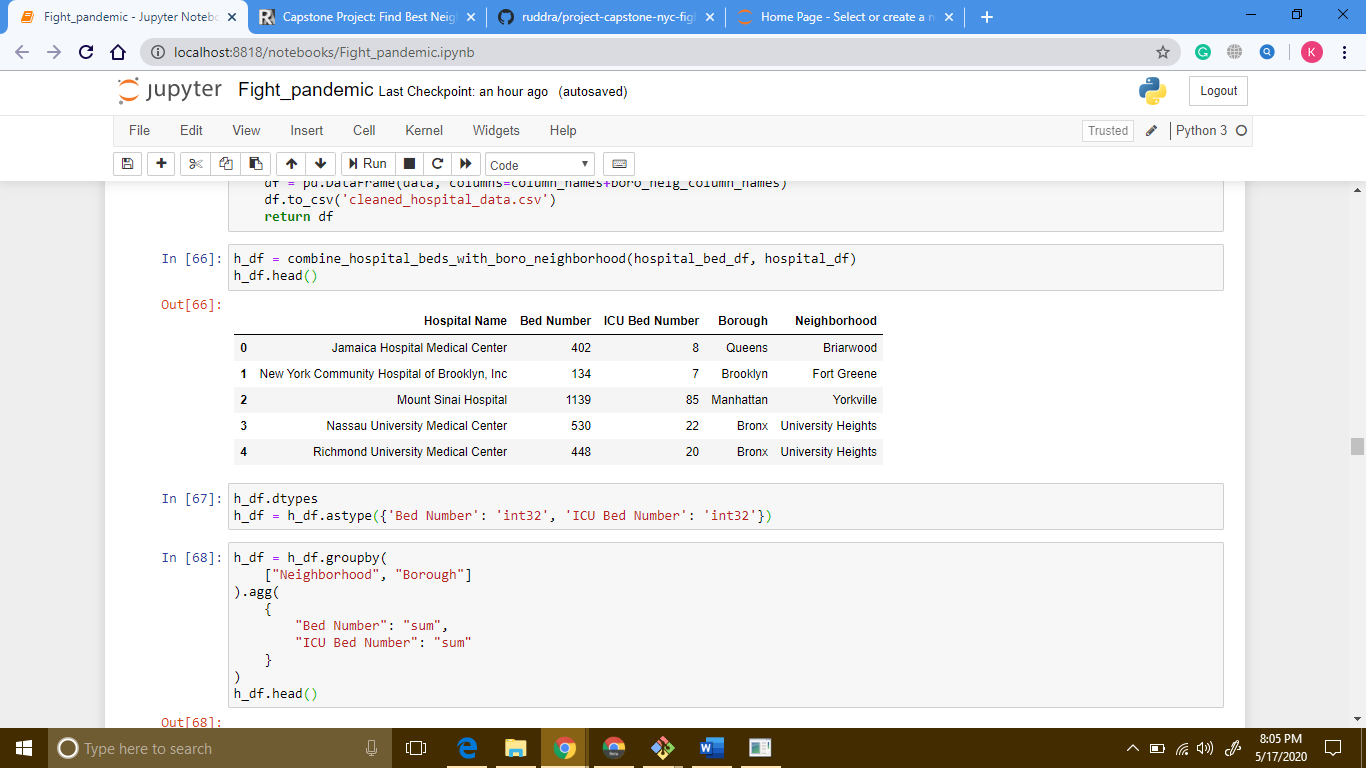
1. **Step five: collect hospital bed data from NYS Health Profile**

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:

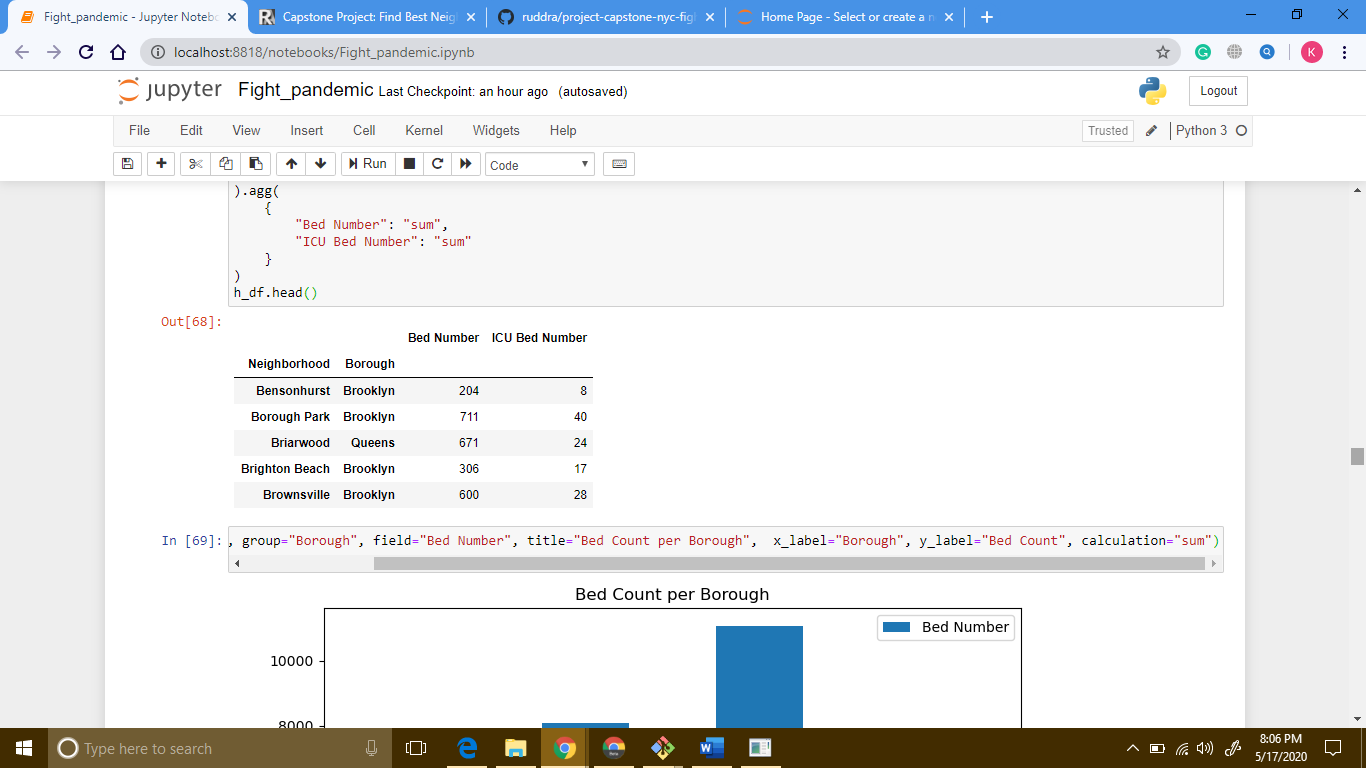


1. **Step six: combine step four and step five**

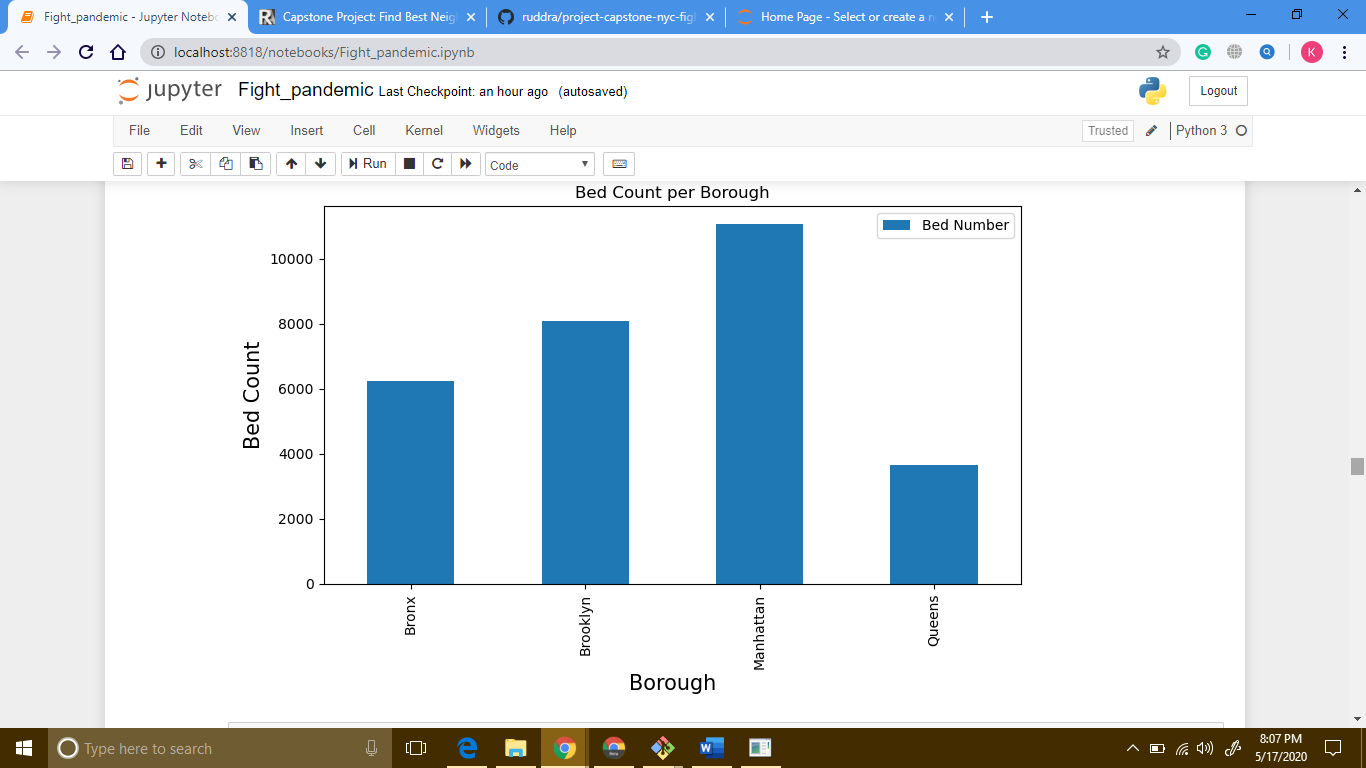
Now we are going to combine data from step four and step five. We are going to internally join the data frame based on “neighborhood” and “borough”.



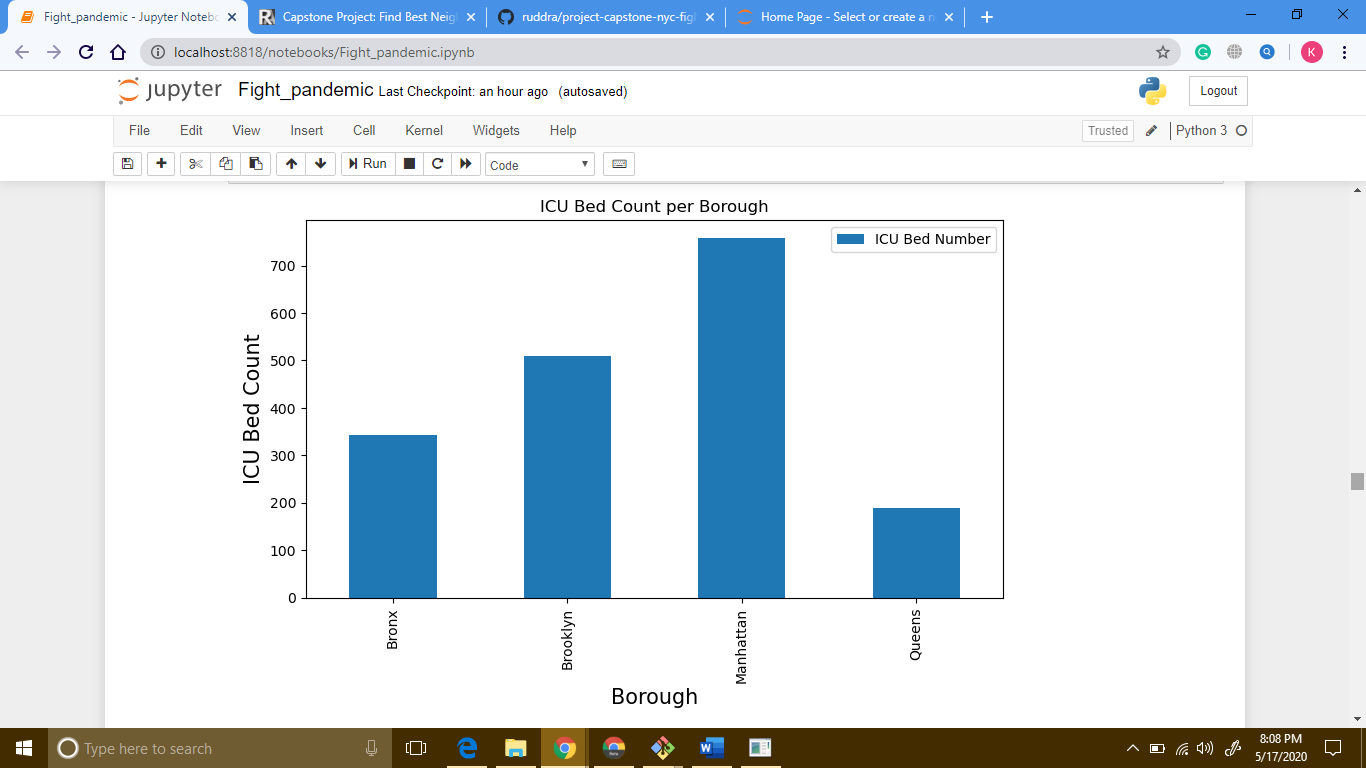
We are going to clean up the data a little bit and sum up bed count and icu bed count grouping by “neighborhood” and “borough”:



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

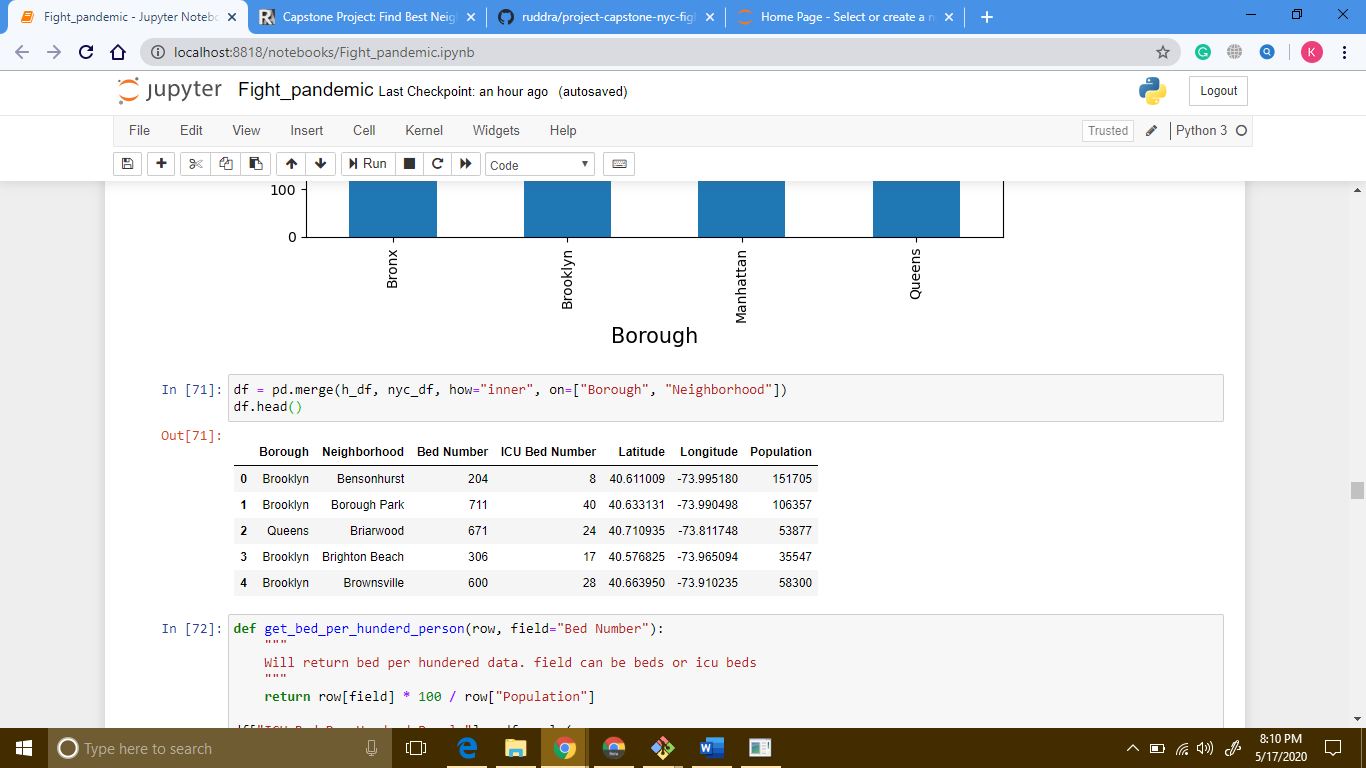


Also, another box charts of “ICU bed count” per “borough”:



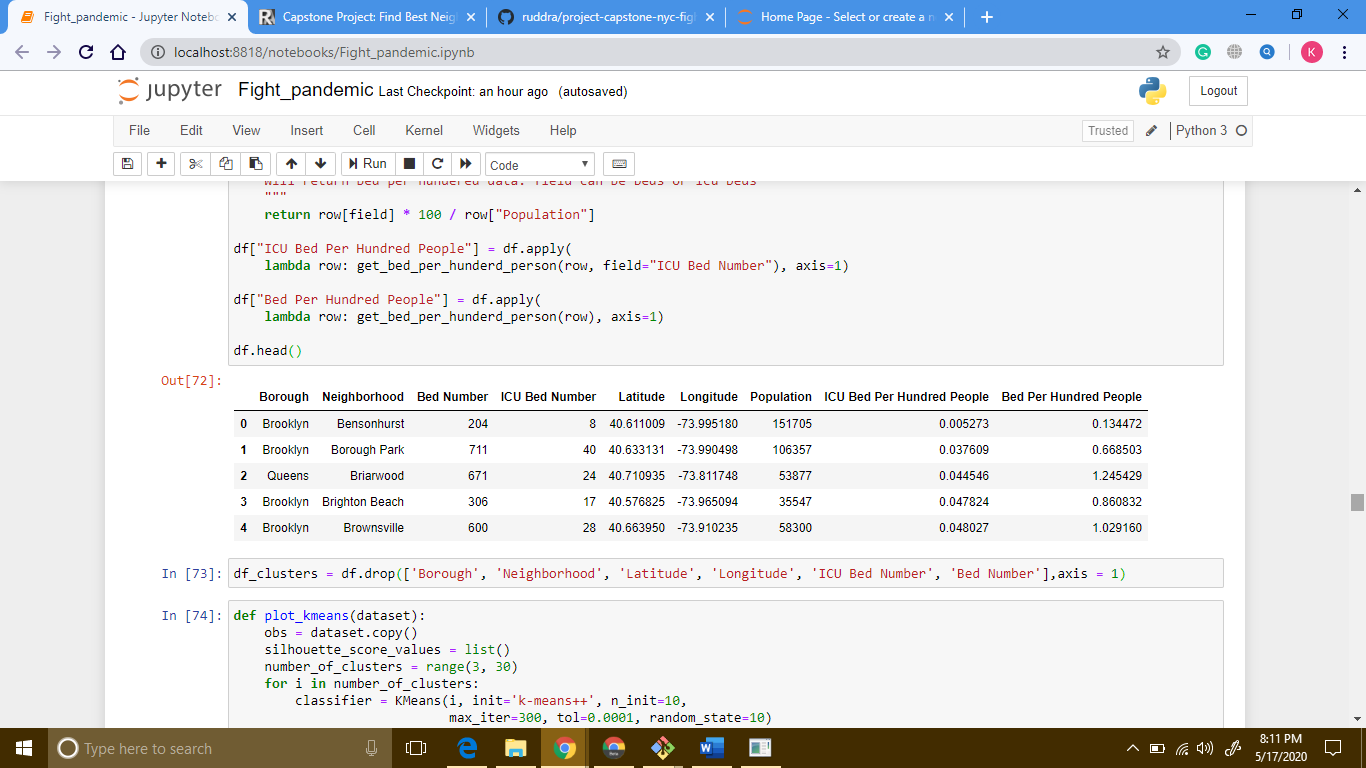
1. **Step seven: combine data from step three and step six**

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:



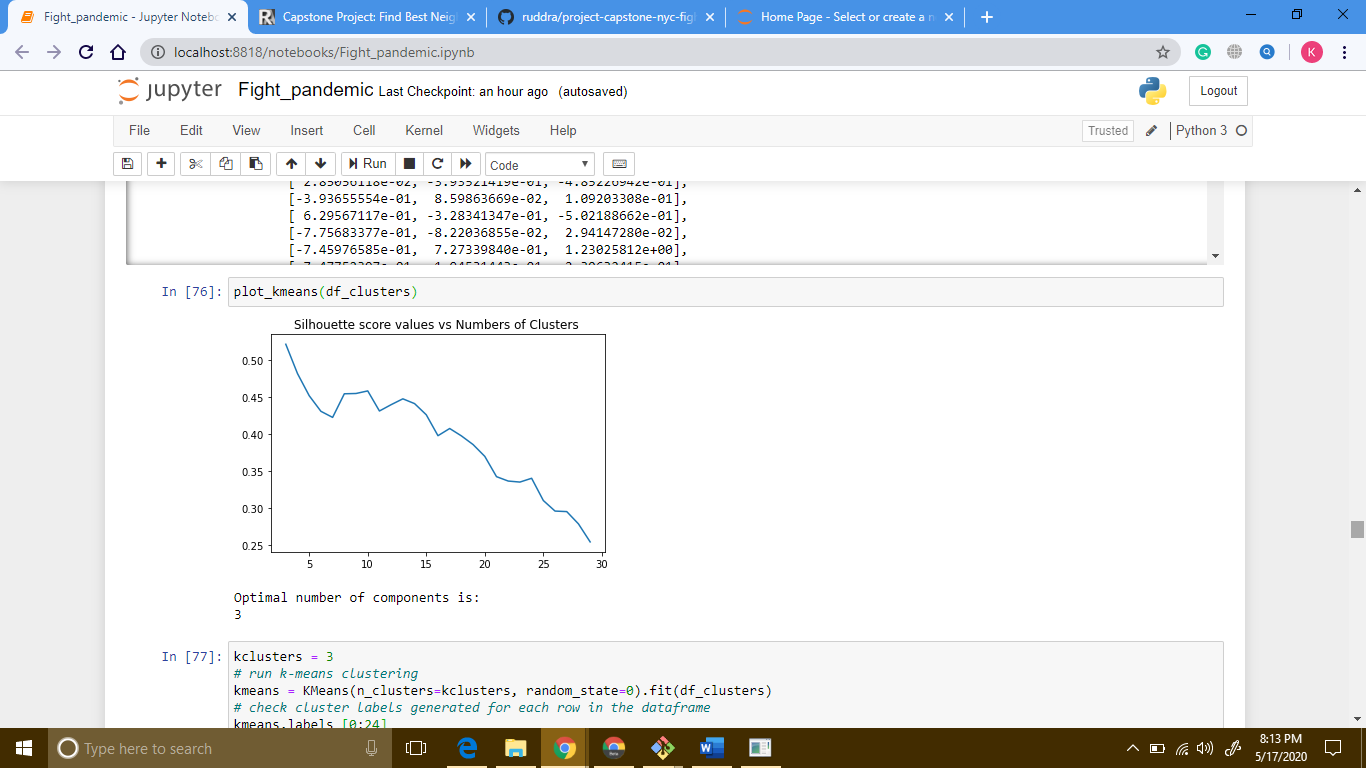
1. **Step eight: add bed and icu per hundred people to data frame**

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:



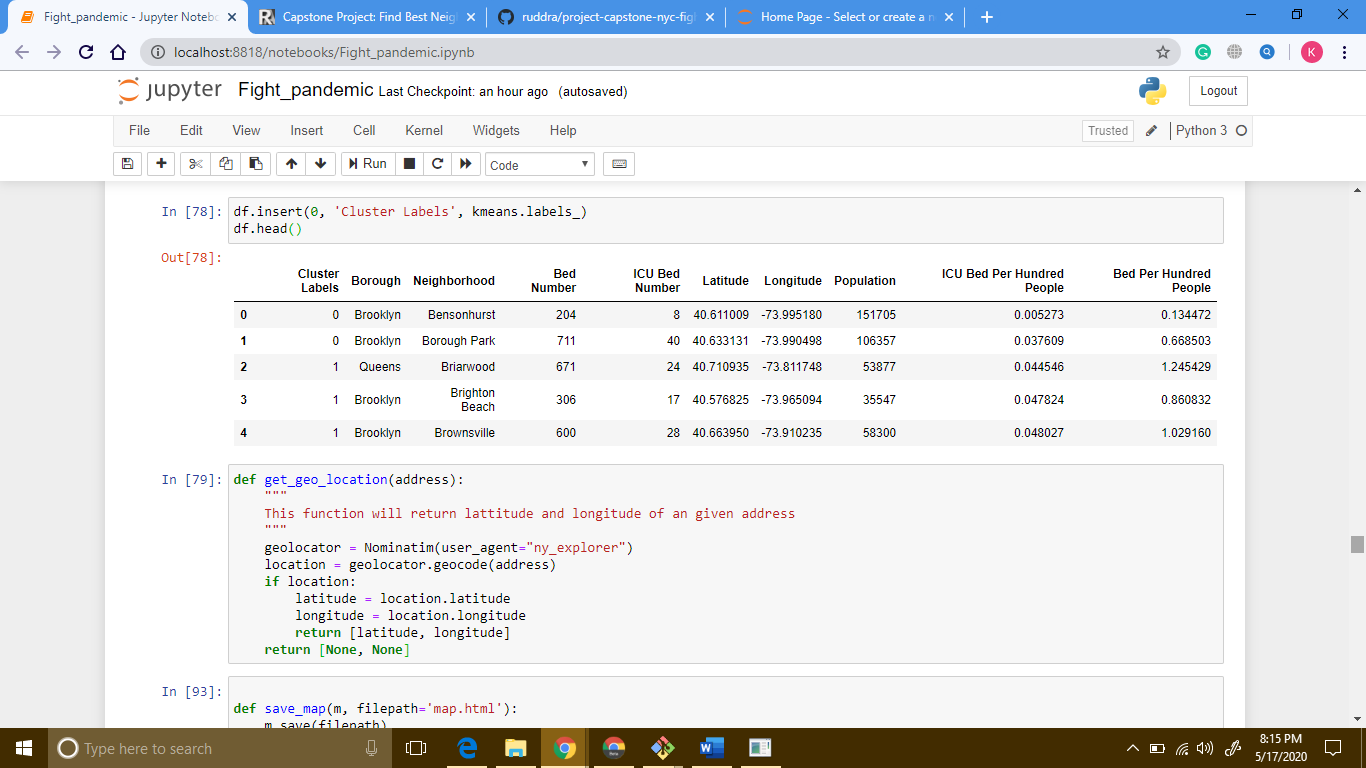
1. **Step nine: K-means clustering**

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.



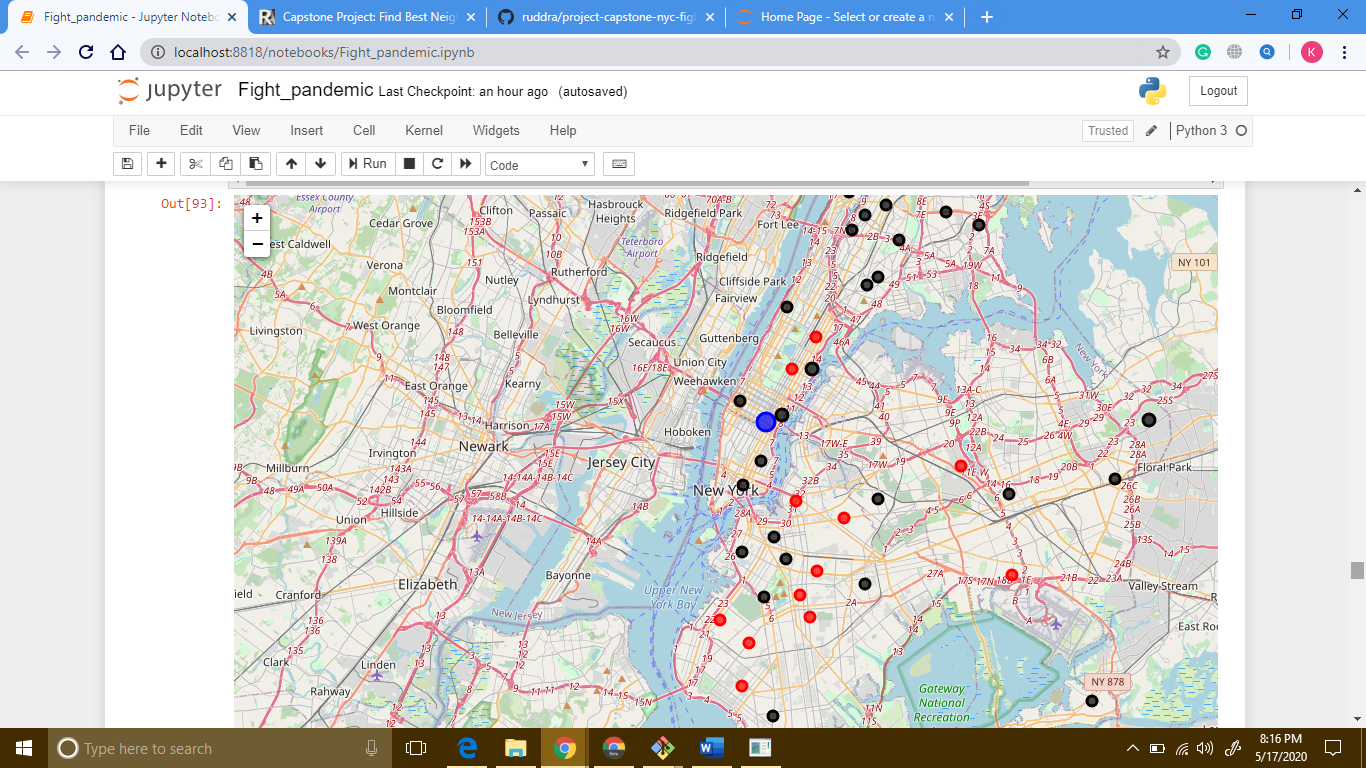
1. **Step ten: merge cluster labels with dataset**

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

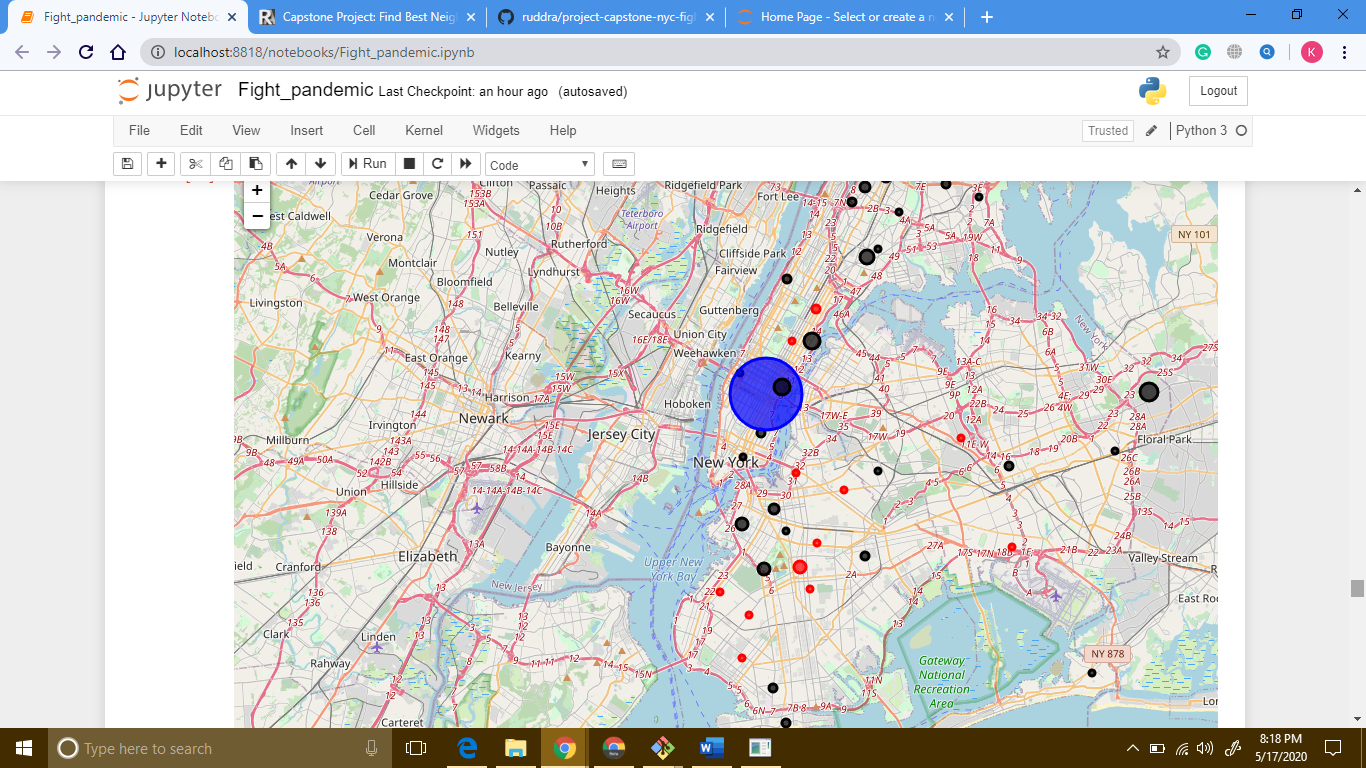


1. **Step eleven: visualize with folium**

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.

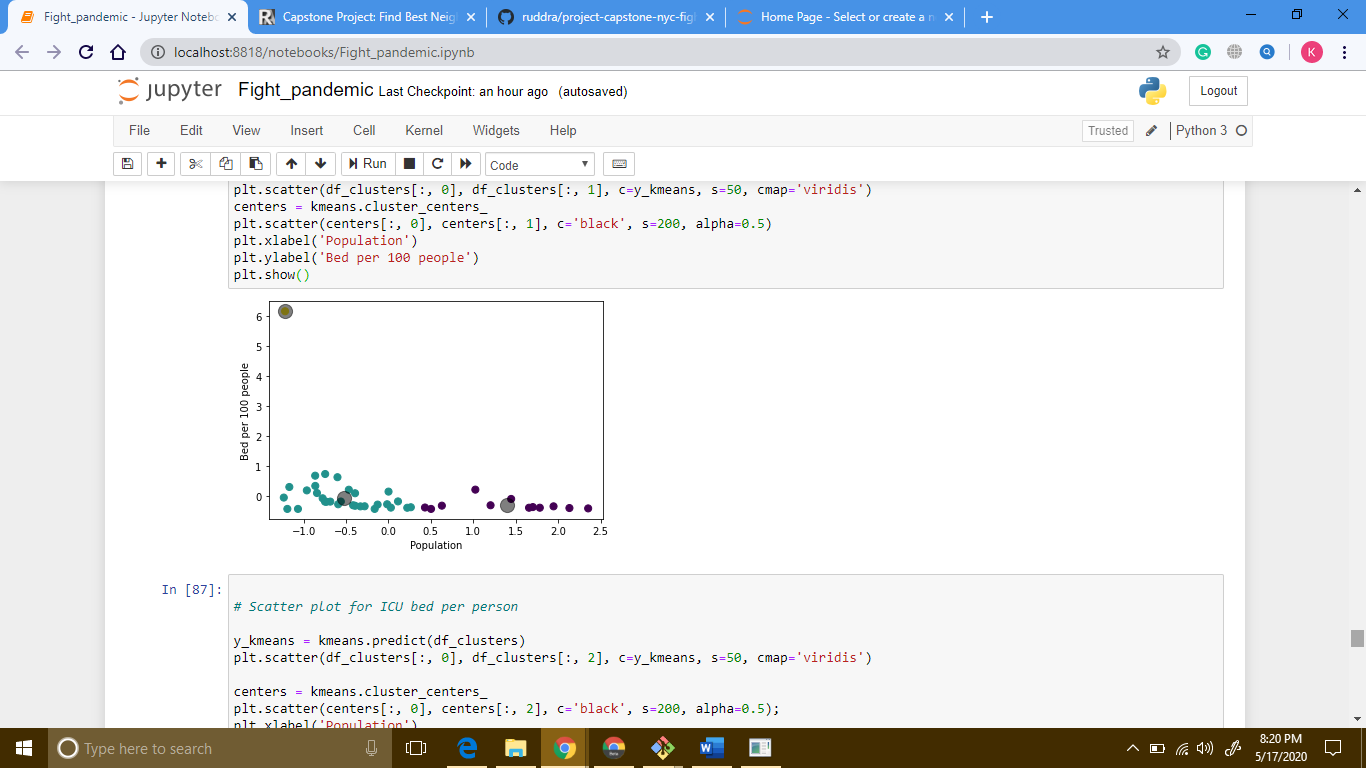


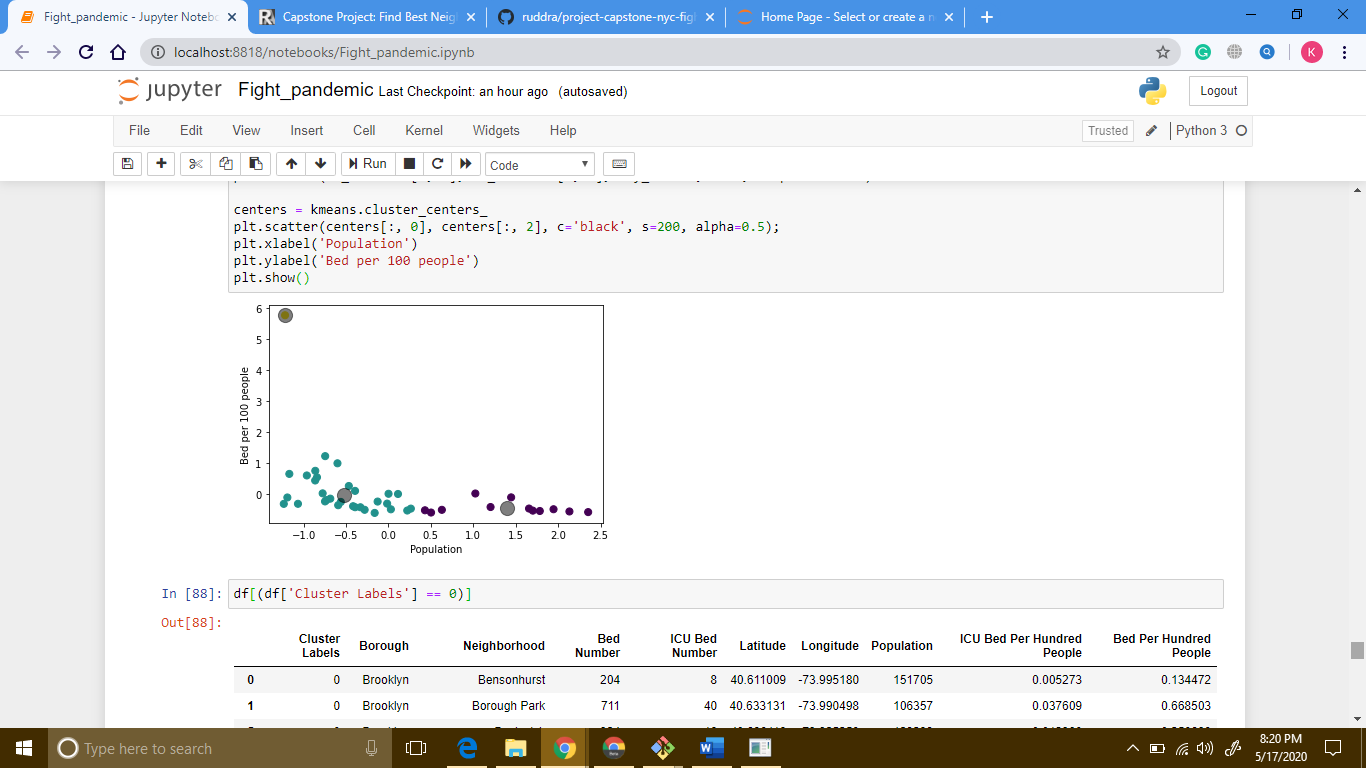
The second map illustrates the clusters where the radius of the Circle marker is proportional to icu beds per hundred people.



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.



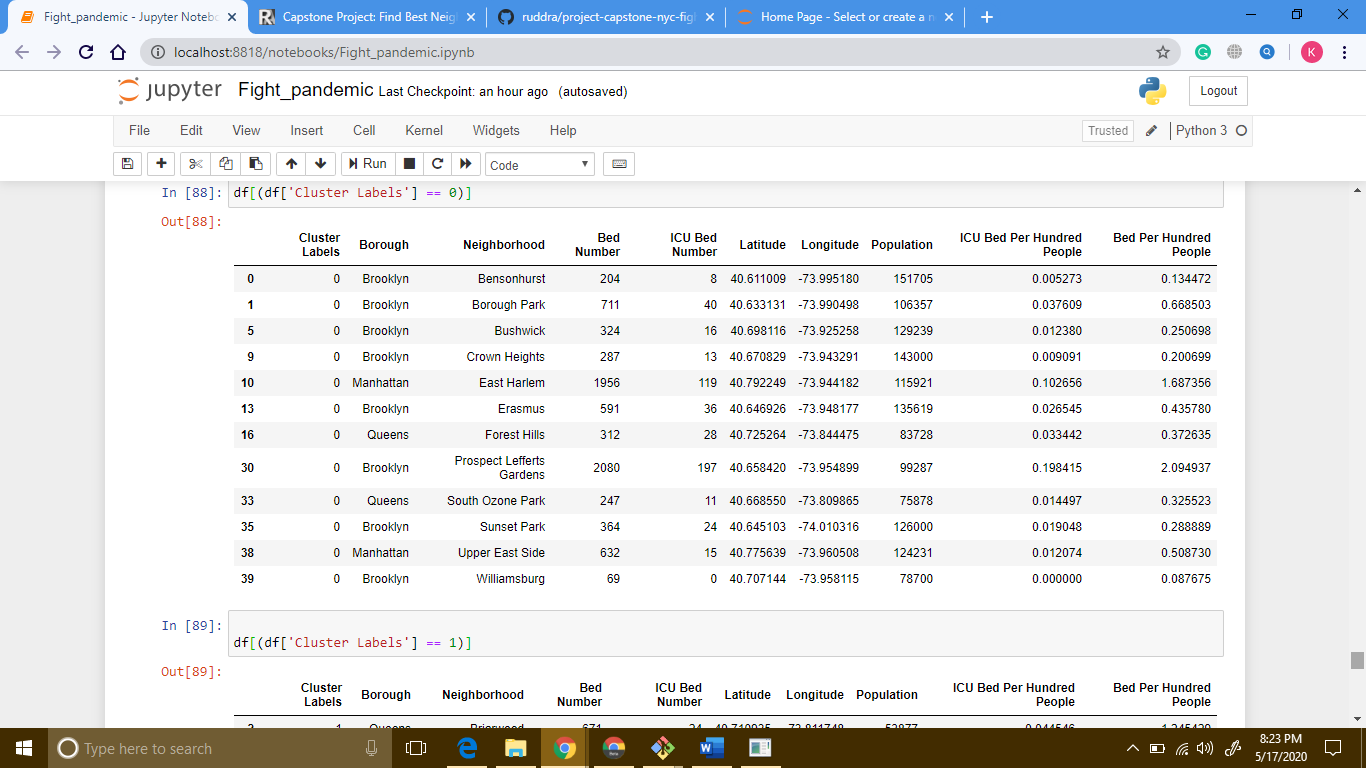


We can observe the obvious outlier here. This neighborhood has a high number of beds per people ratio. From maps above we can easily say that it is **Murray Hill**.

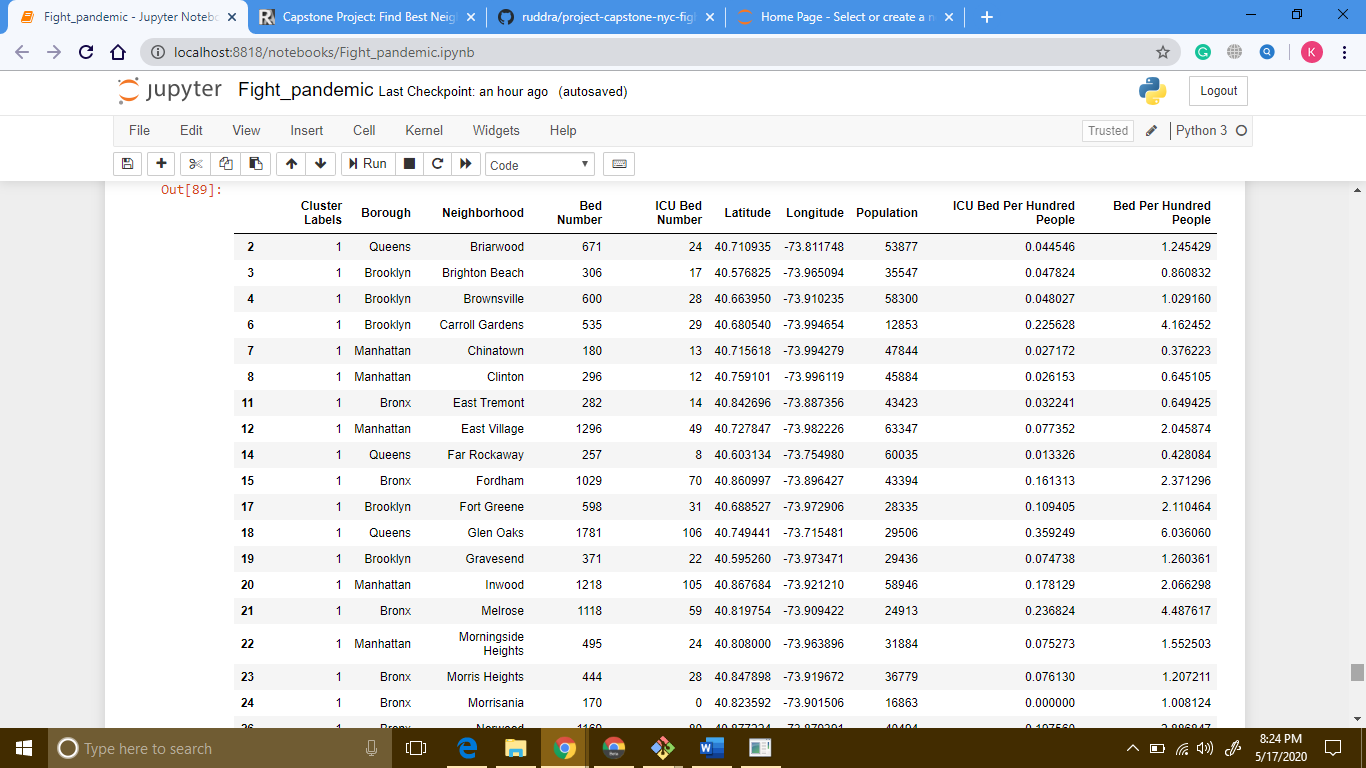
1. **Step thirteen: see which borough goes to which cluster**

Let us see which boroughs belong to which clusters.

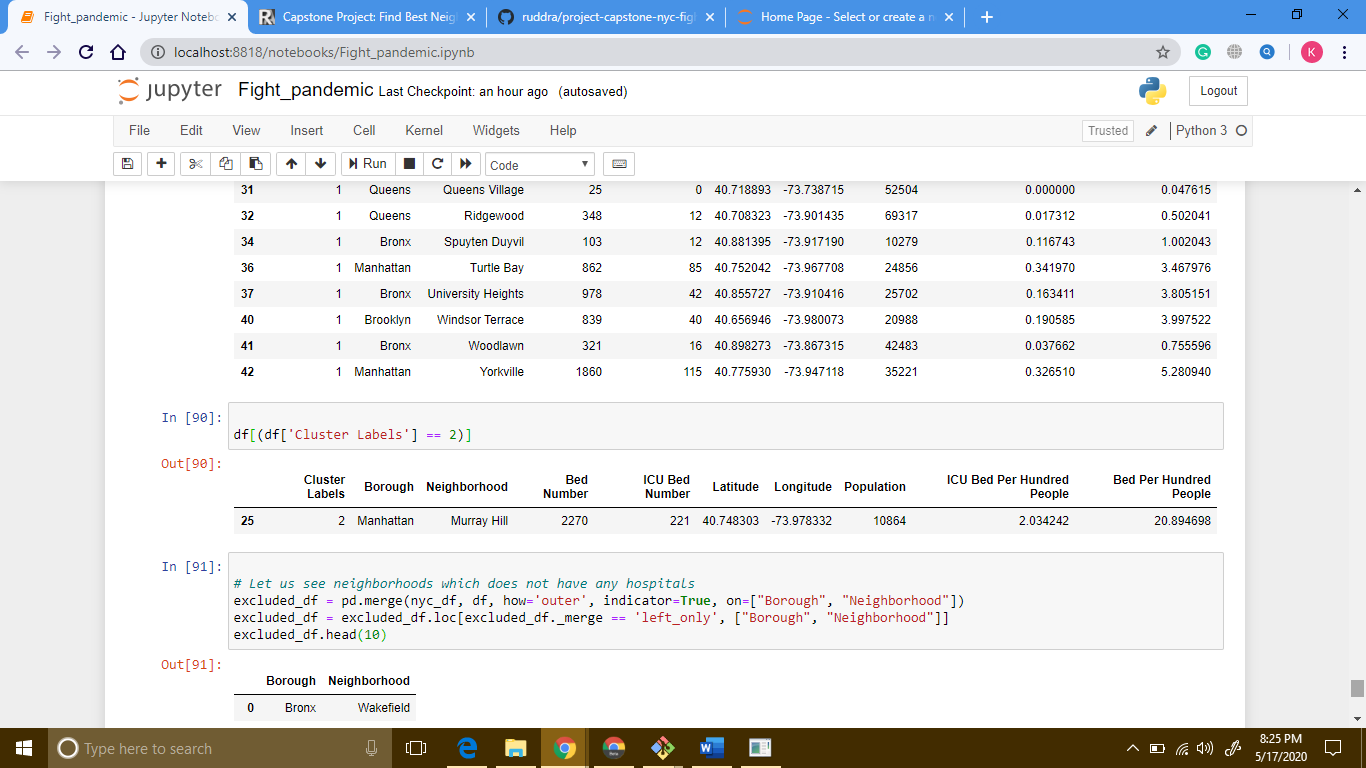
Here is the dataset for cluster 0:



Here is the dataset for cluster 1:

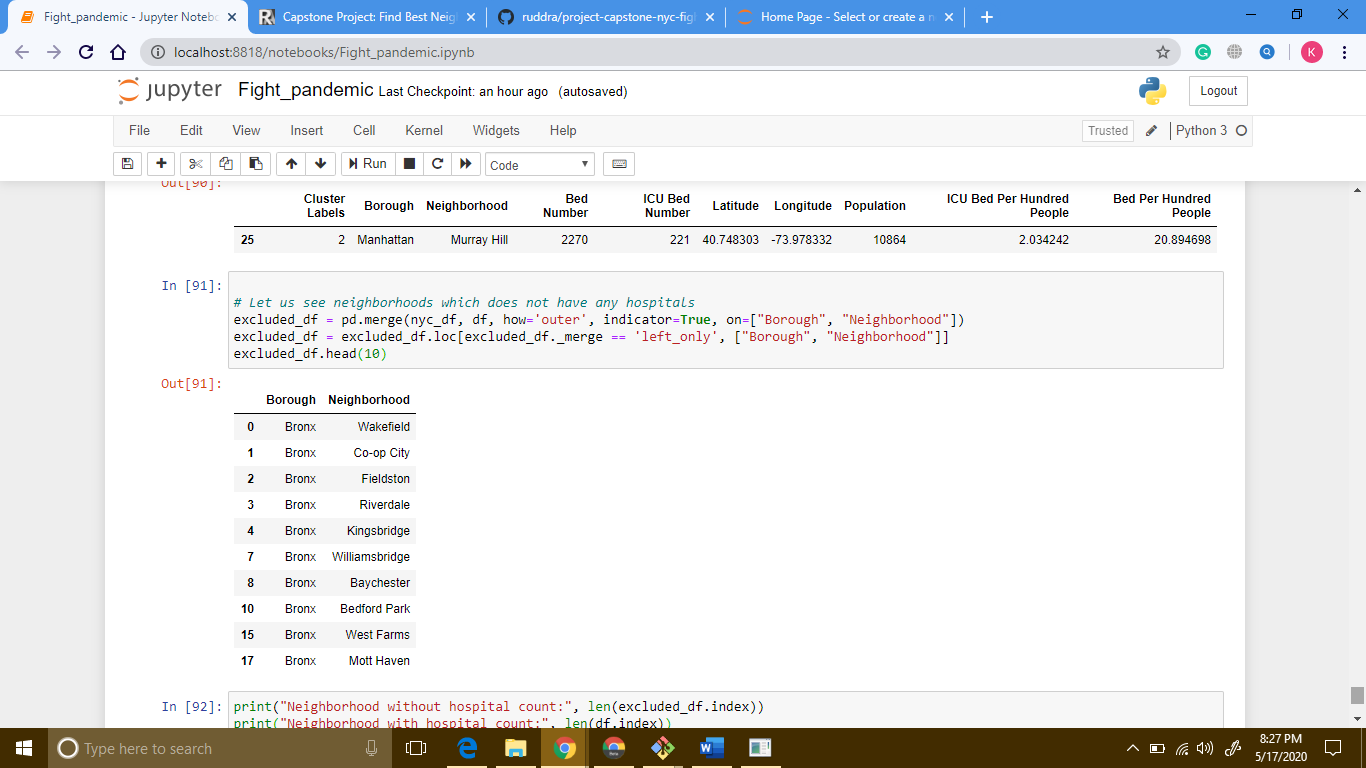


Here is the dataset for cluster 2:

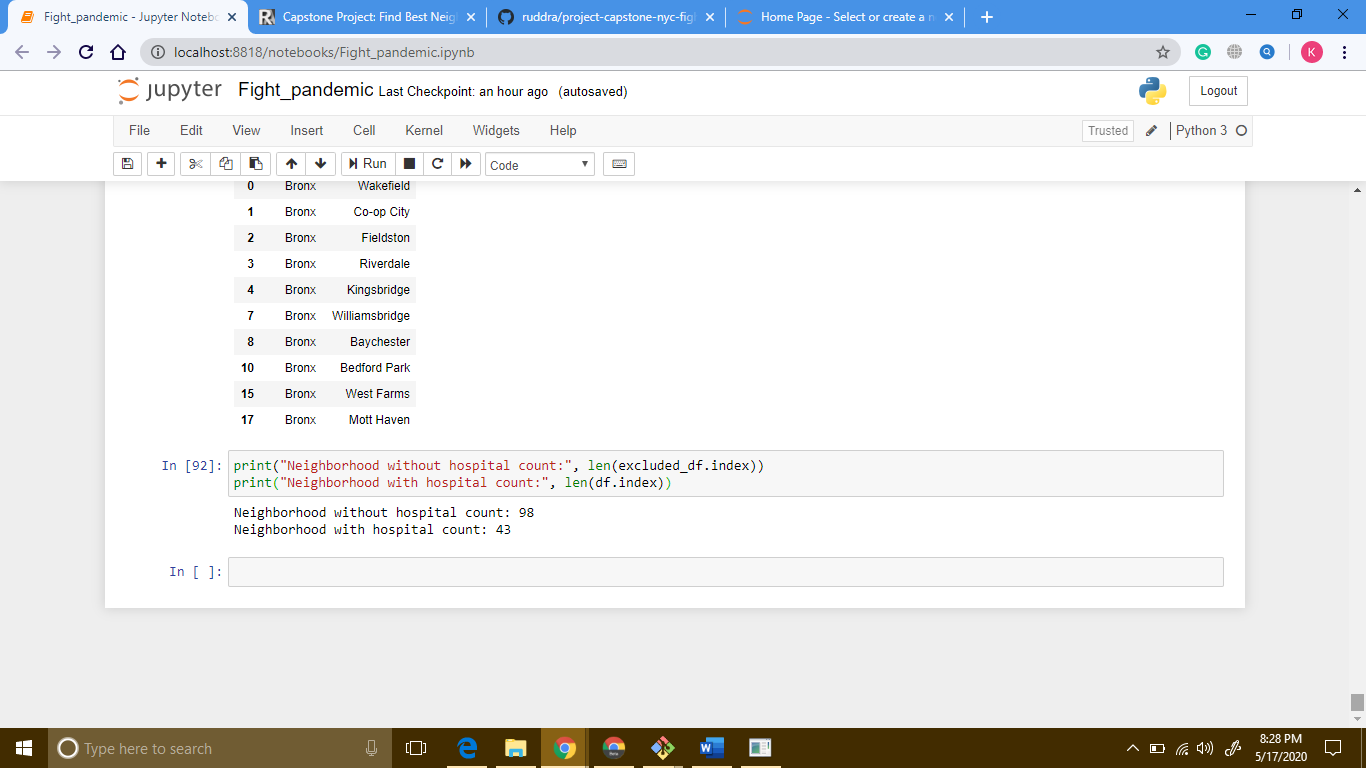


1. **Step fourteen: see neighborhoods without hospital**

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 examination, three bunches were characterized. One bunch (group 2), that comprises of just a single region, has been characterized as the outcast, because of the high number of medical clinic beds, which implies it is better prepared to deal with this pandemic. Two different gatherings were grouped by bed per hundred individuals and icu bed per hundred individuals. Clearly the bunch with the most reduced beds per individual is where we should focus on giving beds and other gear (Group 0). We likewise should investigate conditions in Sovereigns Town and Williamsburg as they have extremely low beds per hundred individuals. Moreover, in hundred different neighborhoods, there is no medical clinic information. Henceforth, individuals living there are at high danger of not being treated during pandemic.

**What could be done better**

Foursquare doesn't speak to the full picture, since numerous emergency clinics are not on the rundown. Therefore, different maps could be used, for example, Google guide or Open Road map.

NYS Wellbeing Profile site may come up short on the most recent data with respect to medical clinic data. It could need data with respect to new emergency clinics. Additionally, clinic ids were removed physically from NYS, which could have missing medical clinics. We additionally dropped neighborhoods which didn't have any emergency clinic information coordinating in NYS Wellbeing Profile site. For this task, we are just utilizing information from 74 emergency clinics in NYC.

We are utilizing fuzzywuzzy to coordinate medical clinic information from Foursquare and NYS Wellbeing Profile. It's anything but a right measure since we are coordinating the names closest conceivable, it could not be right, all things considered, situation.

We are additionally just thinking about medical clinic information. We didn't consider other clinical offices like nursing home or wellbeing center.

We utilized populace information from 2010(as per Wikipedia pages), which are not exact as of now. We ought to have utilized the most recent populace information.

At last, to fight COVID-19, we ought to have had tolerant information for the area. Tragically, we were unable to think that it’s like this (for instance, get tolerant per scope longitude) from any source, thus couldn't consolidate it.

## **Conclusion**

To close, the fundamental information examination was performed to distinguish the most well-prepared clinic in the NYC neighborhoods. During the investigation, a few significant factual highlights of the wards/neighborhoods were investigated and imagined. Moreover, bunching assisted with featuring the gathering of ideal regions. At last, Manhattan-Murray Slope was picked as the most well prepared (according to emergency clinic bed tally and icu bed check) region to fight pandemic.