New York City hospital segregation

**1.Introduction/Business Problem:**

**1.1 Background**

Since we know healthcare is a very important aspect of our day to day life as we never know what kind of problems one can face. A hospital is a very necessary component of any city and if people have to choose to in live an area, they should make sure that it has the necessary health care options. Statistically it was shown that about a few thousand people die every year on the way to a hospital or due to the time delay in receiving treatment. Graphs show that the count of people dying like this has increased, so its very important to select a borough which have high and large number of hospitals in a close by range.

**1.2 Problem**

In this project I am going to find all the hospitals in New York city, the ones that are considered the best and group them according to the Borough, which will help people pick a borough to stay in, since having a decent hospital close by is always a great factor of safety for people in family ranging from infant babies to senior citizens who all have their own needs and needs to be taken care of.

**2. Data:**

**2.1 Data sources**

Here I’m going to use the Foursquare to get the map of New York city, Ill be using foursquare to locate all the hospitals using the latitude and longitude of those hospitals from a dataset which ill explain in a bit. Foursquare will help me visualise the map of New York city with the points of hospitals marked, it’ll let me group those points with K-means and other machine learning ways.

The second one is the dataset of all the well-known hospitals in New York. I found this dataset in the link below

<https://data.cityofnewyork.us/Health/NYC-Health-Hospitals-patient-care-locations-2011/f7b6-v6v3>

**2.2 Data cleaning**

The downloaded or scraped data had a lot of hospitals with no address or a few boroughs with no hospitals using data cleaning techniques I have deleted all such entries.

Then the next problem was that the column names had extra spaces and white spaces at the end and beginning, using the replace function I removed all the extra spaces and cleaned it so it can be perfectly used.

**2.3 Feature selection**

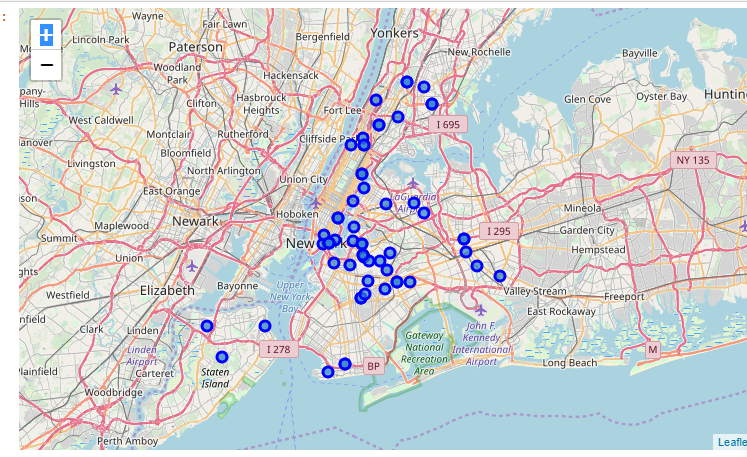
There were many useless columns present when I was first scraped the data, those columns had no use for my purpose, for example the postal code and number of patients and etc were redundant data in my case hence I eliminated them and only used data which was necessary for me like borough name, latitude, longitude and name of the hospital.

After all the above steps the final cleaned and useful data looked something like this.



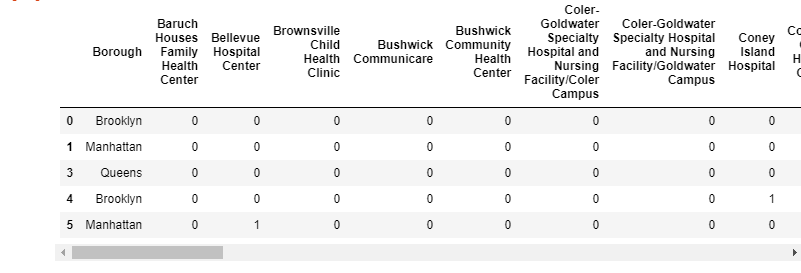
**3.Methodology:**

After scaping and cleaning the data, I took the clean data set and with required columns and used foursquare to plot a map of New York city with the all the hospitals in the data set plotted on them.

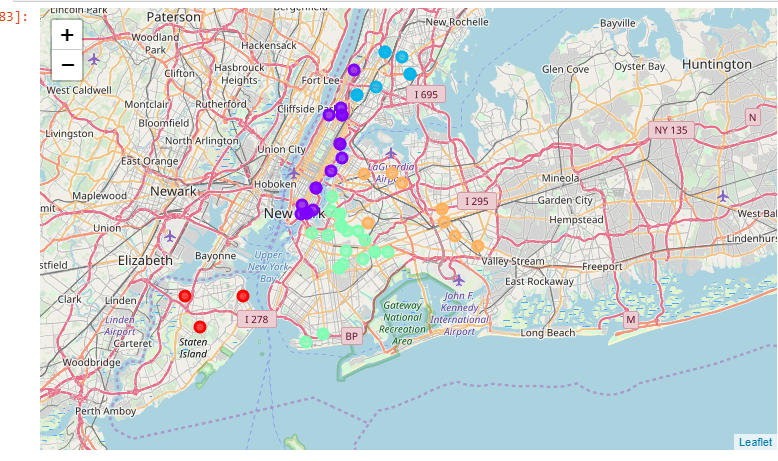


This is done using Four square’s developer app which allows you to use their map data for a data visualising.

Now my job was to cluster the points into a group according to its respective borough which will let people choose which borough to choose for better health care availability. In order to that we first have to one hot code the given data which will convert the non-categorical data into numbers which will then be easier to cluster.



After that I used the K-means technique to cluster data and which resulted in the following map.



**Results:**

According to my visualization we see that Manhattan is a borough which has more hospitals close to each other. I would suggest people who need medical attention to consider staying their as if we see the closeness and number of hospitals Manhattan is clearly the better choice.

**Discussion:**

Here an important thing to note is that I haven’t considered the cost of living factor and based my whole research just on health care availability. In future I can add cost of living, cost of operations as well to this study to make it more accurate.

**Conclusion:**

From this study of mine we can see the no of hospitals and their distances between them belonging to a particular borough. People can use this data and figure out where they want to stay if they have a requirement of medical attention. Although I haven’t considered cost as a factor which can be essential for certain group of people. But I have assumed my audience to be rich middle class or higher middle class.