Analyzing COVID-19 Data For New York City Boroughs

Nilesh Koratpallikar

April 2020

1. Introduction

1.1 Background

COVID-19 has been the most devastating pandemic that has affected the entire world, since 1918. Several countries are struggling to effectively fight against the pandemic, which has already led to a tremendous loss of lives, across the globe. In US, several states have been already affected. In particular, New York City has been the most severely impacted. It is important to understand how this pandemic has affected people within various boroughs of New York City across a spectrum of gender, age, demographics and derive meaning insights from the data.

1.2 Problem

The goal of this project is to analyze the COVID-19 data for New York City and identify potential correlations between the number of COVID cases and other variables including gender, age, demographics, availability of hospital/medical cities, **across various Boroughs of New York City**. Also, given the availability of data for number of COVID cases for each calendar day since the beginning of this pandemic, would like to predict the curve for number of COVID cases.

1.3 Interest

Analyzing the COVID-19 data for New York City, would be very beneficial and the key insights could help several groups including the medical community to come up with remediation steps in order to minimize the impacts of this pandemic, potentially in other places.

2 Data Acquisition

2.1 Data needs & Data sources

- Most of the data sets are available at COVID-19 NY website which is updated on a daily basis.

- For choropleth Data Visualization, the latitude and longitudes of various New York City Boroughs would be needed. This geojson dataset is available at New York City Boroughs.

- Foursquare location API would be used to search/explore and to get the points of interest such as medical facilities, within each New York City Borough.

- NY City Borough latest Census data

  https://www.census.gov/quickfacts/fact/table/newyorkcountymanhattanboroughnewyork,bronxcountybronxboroughnewyork,queenscountyqueensboroughnewyork,kingscountybrooklynboroughnewyork,richmondcountystatenislandboroughnewyork,newyorkcitynewyork/HSG010218

- Additional data sets such as New York City demographic data would also be sourced.

The data sets would be cleansed, as required for exploratory data analysis and described in detail, within subsequent sections of this report.

2.2 Data Cleaning

The data from the NY COVID-19 was more in graphical format. I had to reverse engineer the data from the graphs and save then as individual data sets. The data sets chosen were as follows, for boroughs around New York City

- Daily COVID -19 case counts, starting from March 3rd, when the data collection initially began

- COVID-19 cases in various NY boroughs [Bronx, Brooklyn, Manhattan, Staten Island, Queens]

- COVID-19 cases based on gender

- COVID-19 cases based on age

- Additionally, for rendering subsequent choropleth graphs, NY city area geojson was also downloaded from available sources online.

The following data issues were resolved as part of data scrubbing:

- Date format of MM/DD, was incompatible with the some of the pandas data frame methods. The data in the csv had to be updated to MM/DD/YYYY

- When the csv files were read using pandas ***pd.read_csv*** method, default index [un labeled index] got added to the resulting data frame by default. This caused issues with rendering bar plots. To avoid the issue, explicit column name was passed as the index column [**index_col=<column_name>]** in the pandas ***pd.read_csv***

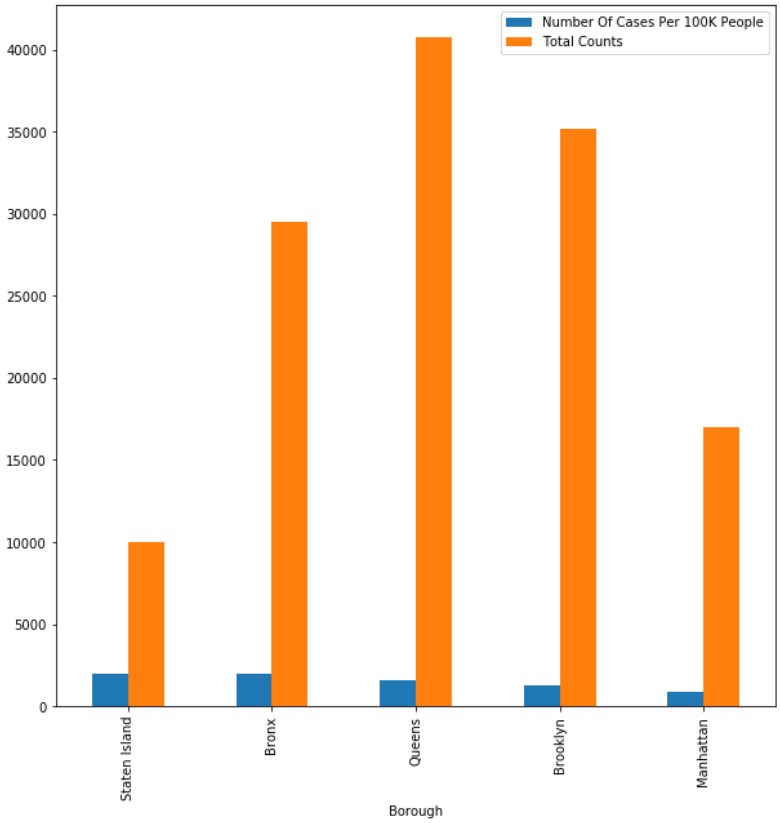
- The Logistic Regression model for modeling and curve fitting the total COVID cases data [sigmoid function], needed the data inputs as a ***float*** data types. This became a problem with the independent variable: the calendar date [3rd March etc.]. I had to convert the MM/DD/YYYY date into Date time object initially and then had to convert them into *milliseconds [float data type]*

3 Exploratory Data Analysis

3.1 Target Variable

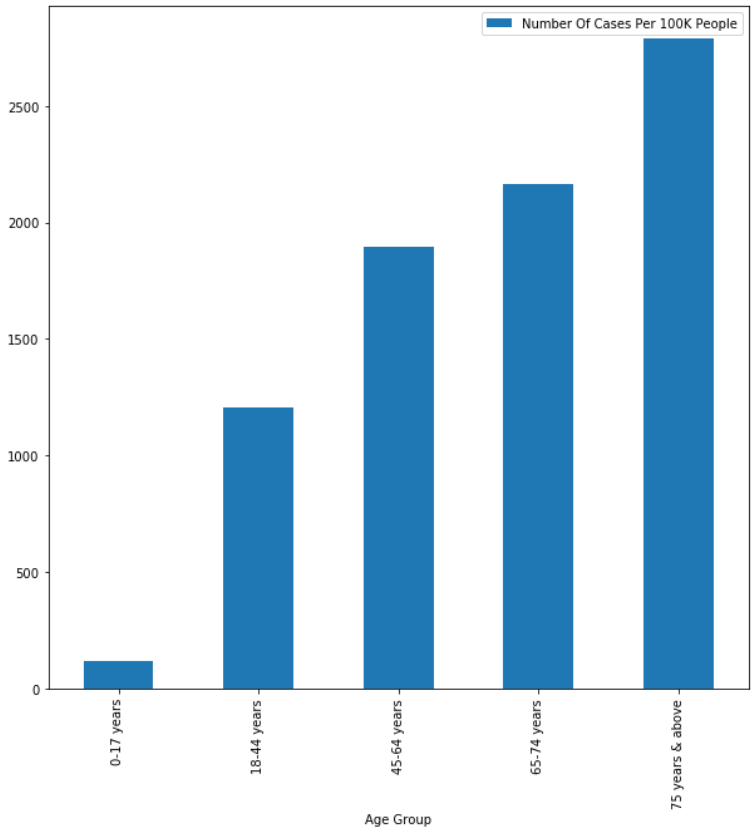
The target variable was obviously the number of COVID-19 cases. There was no calculation to derive the target variable. Data was already in the feature dataset that was collected from the relevant data source.

3.2 Distribution Of COVID cases across NY City Boroughs



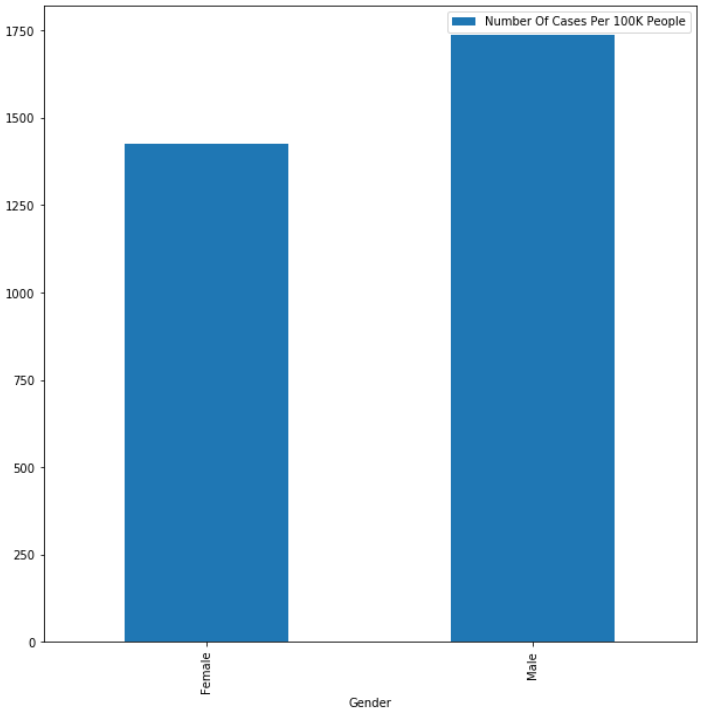
As seen from the Histogram above, Queens and Brooklyn Boroughs recorded the highest # of COVID cases. In subsequent, analysis, we will try to look into the potential reasons for this distribution of cases, including the demographic of the population, such as # of people over 65 years of age.

3.3 Relationship between COVID cases and patient age



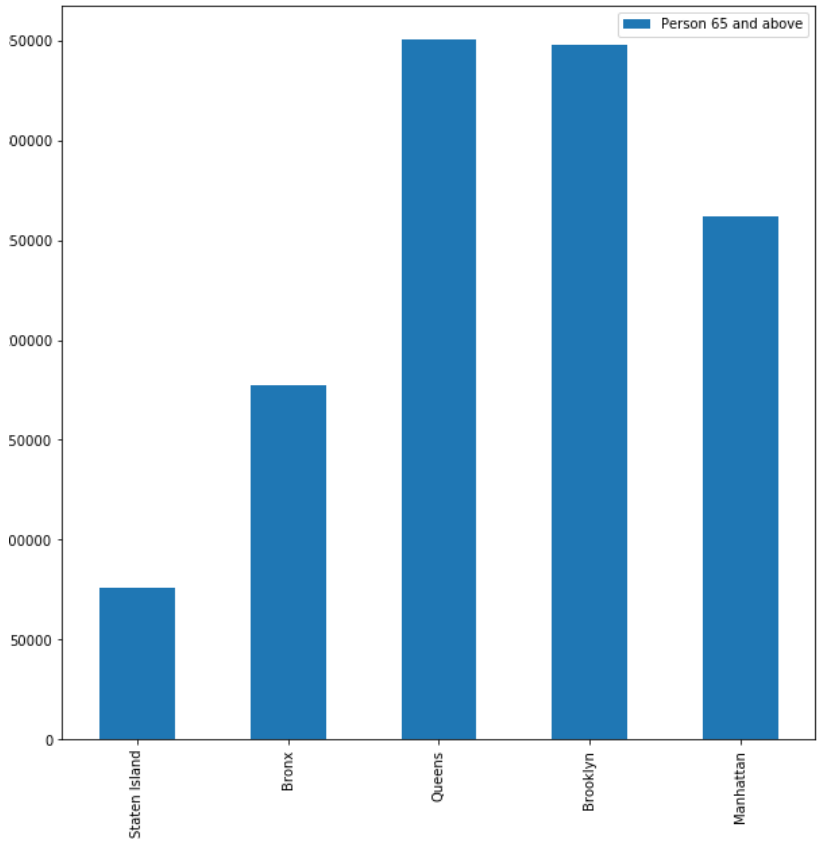
Based on the above histogram, # of cases is highly correlated with age group of patients who are 65 years and older. This was the possible scenario as mentioned by several medical experts, but the actual numbers validate and backup the argument that elder and people over 65 years of age were disproportionally affected by this epidemic. Interestingly, age group of 0-17 years, showed remarkably low rate of infections. The infections progressively went up, as the patient age increased.

3.4 Relationship between COVID cases and patient gender



Based on the above histogram, # of cases were marginally higher than females. But since the difference is not significantly higher, and there could be other reasons, which c.ould potentially be looked into

3.5 Relationship between COVID cases/Borough vs. # of people 65 & over



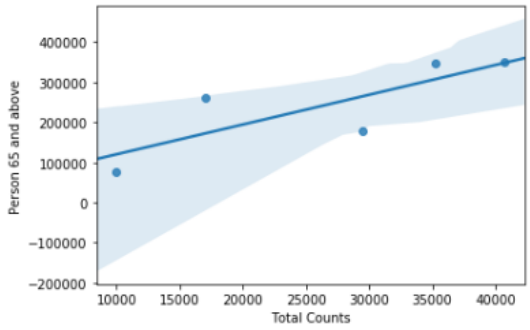
The histogram above shows the distribution of elderly people [65 years and above] in various Boroughs. As can be seen, Queens and Brooklyn have the highest number of elderly people. This demonstrates that there is a strong correlation between boroughs with number of elderly people and the number of COVID cases in that Borough. In other words, higher the number of elder people in a borough, higher the number of COVID cases, in that particular Borough.

Given that Brooklyn and Queens have the highest number of elderly people, they also have the highest number of COVID cases

Correlation

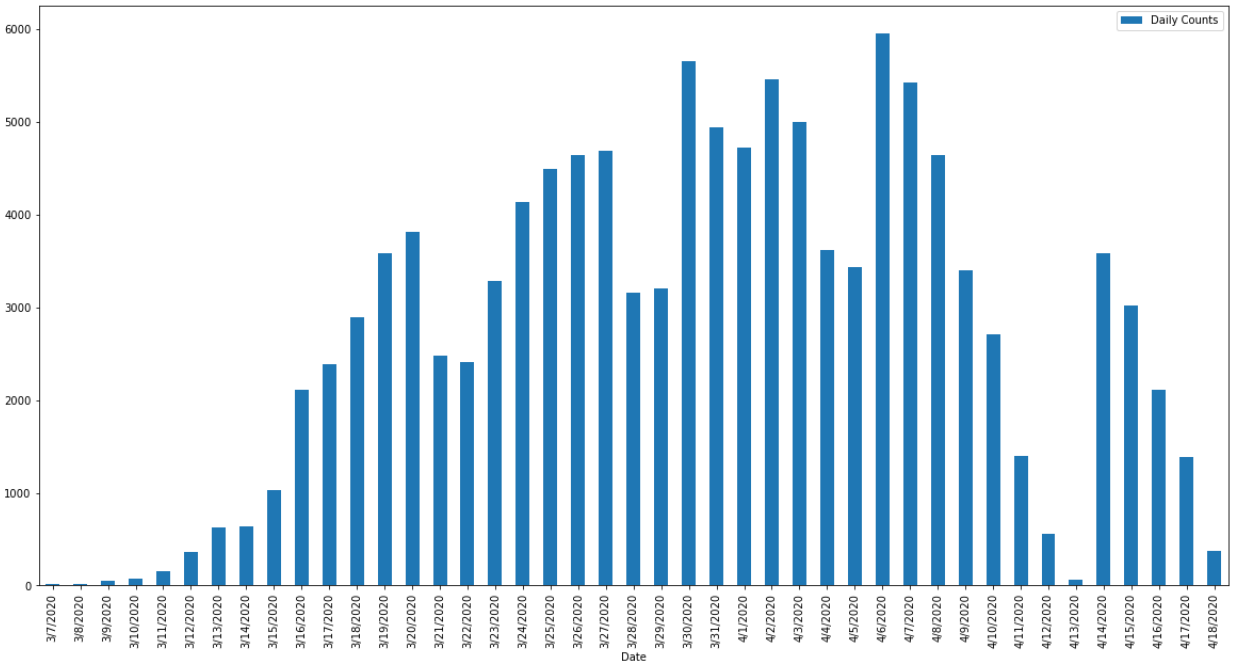
18]:

	Number Of Cases Per 100K People	Total Counts	Person 65 and above
Number Of Cases Per 100K People	1.000000	-0.035890	-0.605194
Total Counts	-0.035890	1.000000	0.805686
Person 65 and above	-0.605194	0.805686	1.000000



The Scatter plot above shows a strong correlation [0.805686] between the person aged 65 & above and the COVID case count for each of the 5 NY city Boroughs. The associated points are very close to the Linear Regression Line.

3.6 Total COVID cases over last month



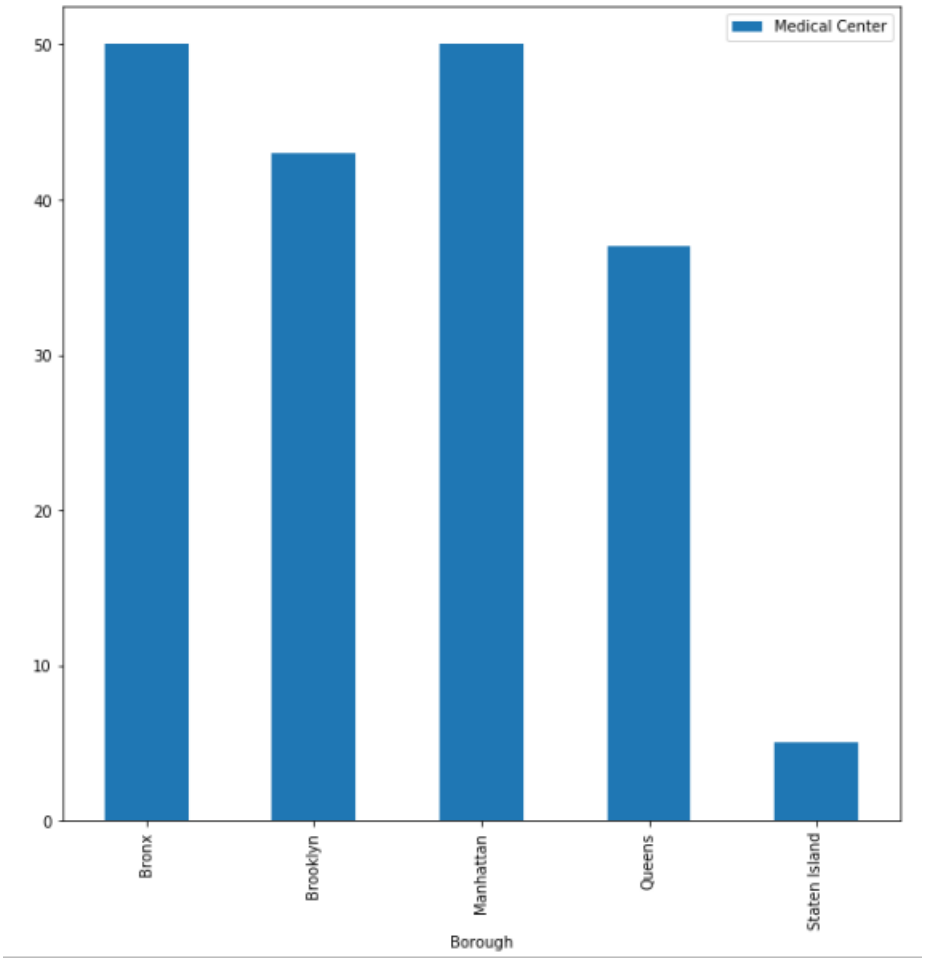
The histogram above shows the “curve” of the COVID cases progression over the last 2 months, starting 3/7/2020, the earliest date, from which the COVID data was collected. As can be seen

the number of cases initially gradually increased, until they reached the peak and curve "flattened" around first week of April, after which the case count gradually started to decrease. In the subsequent section, we will look at a Logistic Regression model [non-linear], that would fit this curve. We will use the appropriate Machine Learning model to train/fit the data and come up with a predictive model.

3.5 Distribution of Medical Center across various NY city Boroughs

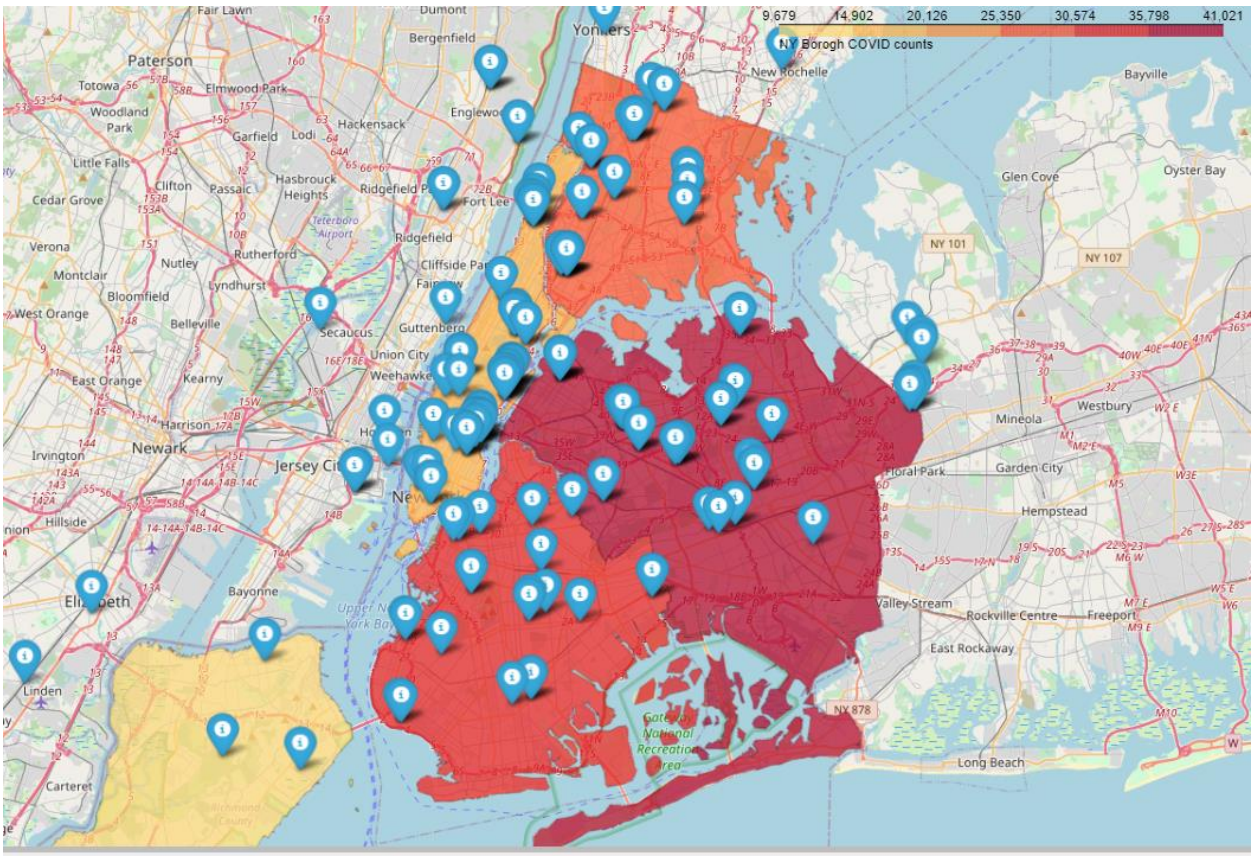
FourSquare API

```
Borough
Bronx            50
Brooklyn         43
Manhattan        50
Queens           37
Staten Island     5
Name: Medical Center, dtype: int64
```



- Foursquare API was used to get the distribution of "Emergency Medical Centers" in each of the Boroughs. "Emergency Medical Center" was used as a key to explore the various venues [hospitals] in each of these Boroughs.
- The number of Medical Centers, for each Borough was subsequently derived based on the response
- As can be seen from the histogram, Queens & Brooklyn do not have as many Medical centers as other Boroughs [e.g. Manhattan]
- Latitude/Longitude for each of the Medical centers was retrieved from the response and was plotted as Popup Markers in the folium Choropleth plot, described in the next section.

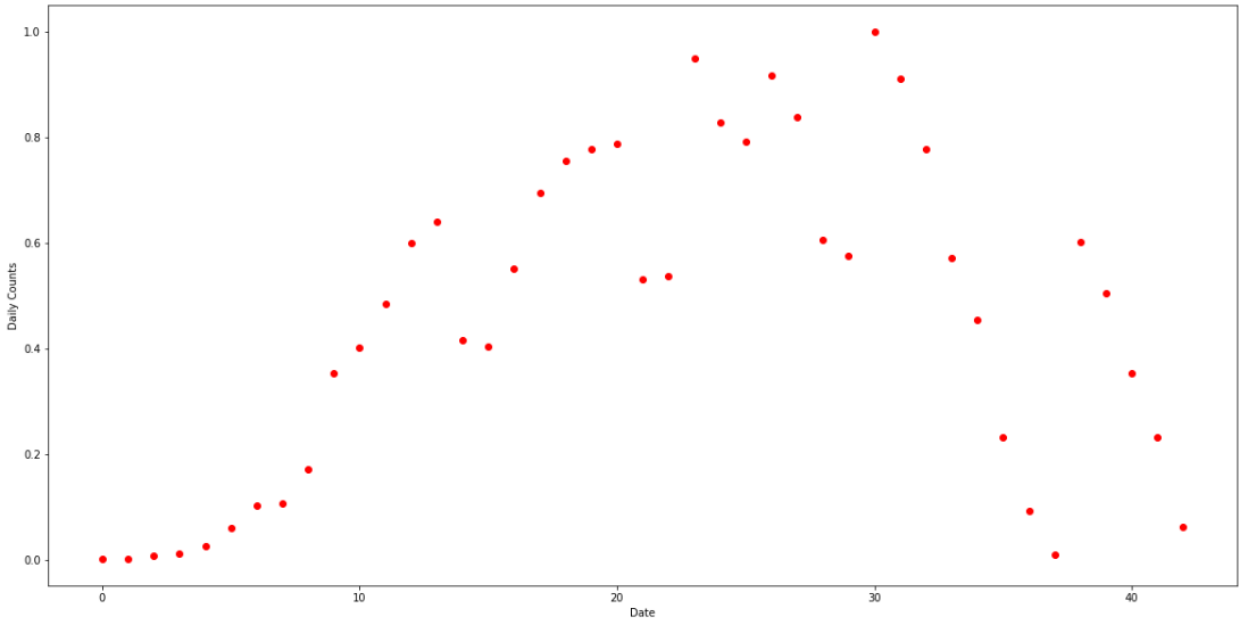
Choropleth Maps



Folium choropleth maps as shown above, was constructed with the following data points

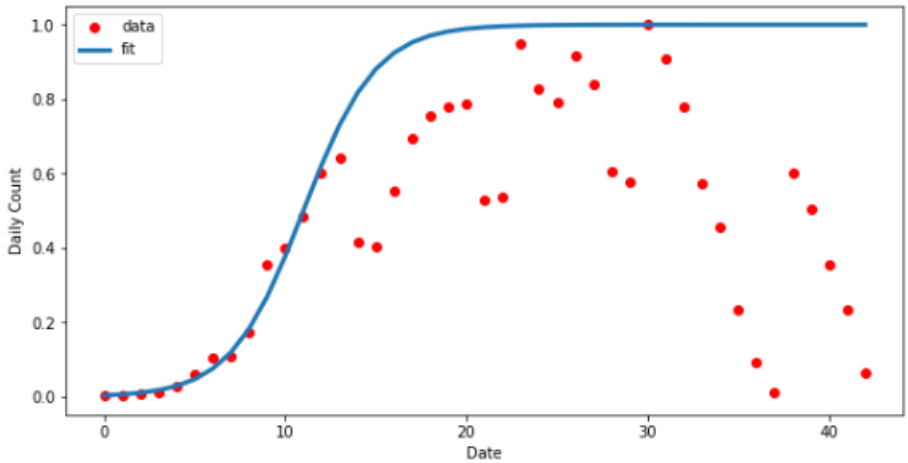
- Latitude/Longitude for each of the Medical centers, retrieved from the Foursquare API
- Total COVID case counts for each of the NY City Boroughs
- As seen from the plot, Queens and Brooklyn have the highest number of COVID cases

4 Predictive Modeling



Based on the distribution of the COVID cases, Logistic function could be a good approximation, since it has the property of starting with a slow growth, increasing growth in the middle, and then decreasing again at the end. In the subsequent section, we will model the distribution with a Logistic regression model for prediction.

4.1 Logistic Regression



A Logistic Regression model, was built based on the sigmoid function, as defined below.

```
def sigmoid(x, Beta_1, Beta_2):
     y = 1 / (1 + np.exp(-Beta_1*(x-Beta_2)))
     #y = 1 / (1 + np.exp(-(x-Beta_2)/Beta_1))

     return y
```

The accuracy of the model was determined based on the following parameters

```
Mean absolute error: 0.19
Residual sum of squares (MSE): 0.06
R2-score: -5.87
```

5 Conclusion

Based on the data, it can be concluded that the

- Queens & Brooklyn Boroughs had the highest number of COVID cases, since they also had the highest number of elderly people [65 + years]
- Based on the data, there is a strong correlation between the number of COVID cases and patient age. Higher the patient age in a given Borough, higher the case count in that Borough
- Distribution of COVID cases over a period of time, follows a Logistic Regression model [i.e. slow growth initially, gradual growth, flattening, followed by a decrease]
- Based on the Foursquare API, the number of venues [Emergency Hospitals] were relatively lesser in Queens & Brooklyn.