IBM Applied Data Science Capstone

Capstone Project

Comparison of New York Boroughs and the impact on outbreak of COVID 19

By: Usama Tariq

May, 2020

Contents

Introduction

The outbreak of COVID 19 pandemic has created a havoc around the global. Apart from infecting over 4 million people and causing almost 300 thousand deaths worldwide, the world economy is at a standstill. While researchers are still struggling to find a vaccine the world leaders have resorted to enforce social distancing through lockdowns. In the wake of these events, the need to break the transmission chain is of critical importance. Consequently, researchers need to identify the hotspots for the transmission to minimize the impact of COVID 19 on businesses and society at large.

Business Problem and Target Audience

The objective of my project is to compare the New York boroughs with highest and lowest number of COVID 19 cases and identify the neighborhood clusters which are existent in most affected borough only. This will help to highlight the hotspots in terms of the most common venues which might have contributed towards surge of COVID 19 cases in the respective boroughs. Once the hotspots are identified the analysis can be used effectively to control the outbreak and minimize the social and economic impact of COVID 19. The analysis can be used by governments to design public health policy by allocating medical resources in such neighborhoods which are at risk of the COVID 19 outbreak as well to lockdown the hotspots or to enforce strict prevention measures in these areas. Imagine if a location is at the risk of COVID 19 outbreak and the policy makers are already aware of the target areas in which the resources are to be deployed. Not only this will help the governments financially but will probably save a lot of lives too.

Data Requirements and Sources

In order to identify the boroughs with highest and lowest number of COVID 19 cases in New York, we will be using the data provided by NYC Department of Health and Mental Hygiene. website. The data reference is attached as Annexure 1. Once the two boroughs are identified we will then use the New York Json file available in week 3 of Coursera Capstone to acquire the geolocation of the neighborhoods in the New York boroughs. Since this data has already

been tried and tested in Week 3 lab the data file is quite comprehensive and will successfully fetch the venue details from the foursquare API.

Further, Foursquare API calls will be used to identify the common venues in the neighborhoods, and they will be clustered using K-means. For visualization purpose, GEOJSON data available on GITHUB will be used to segregate the New York boroughs in the folium map.

Annexure 1

Rates by Borough

This chart shows the number of positive cases per 100,000 people in each borough. It indicates the spread of COVID-19 relative to each borough's population.

	Rate per 100,000 people	Count
The Bronx	2,915	41,746
Staten Island	2,674	12,733
Queens	2,497	56,899
Brooklyn	1,939	50,079
Manhattan	1,398	22,771
Citywide		184,319

Methodology

Since we have limited API calls from Foursquare so we will be identifying and comparing the two New York boroughs with highest and lowest COVID 19 cases. To serve this purpose we used the exploratory data analysis of plotting a bar plot which clearly visualized the two boroughs were Bronx (highest) and Manhattan (lowest).

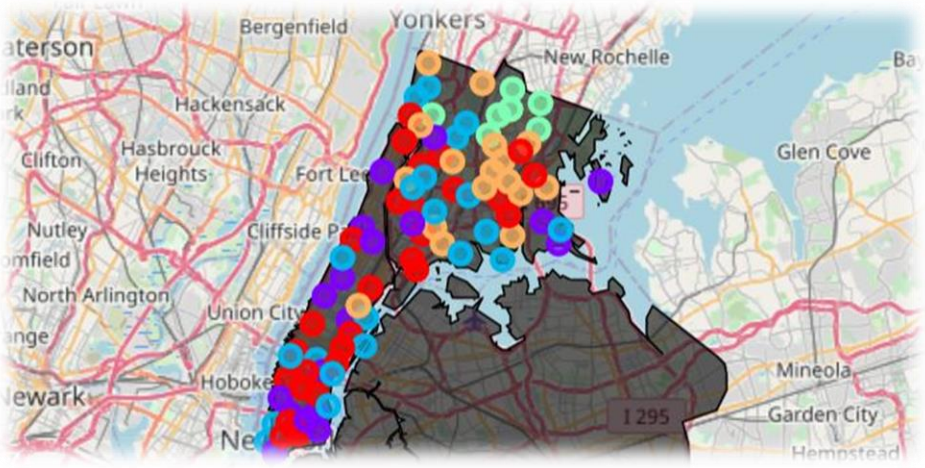
Once the boroughs were identified, we then used some basic data cleaning techniques to filter the geolocation data frame of New York for Manhattan and Bronx. The Foursquare API was then called to find the 5 venues within 1000 mete radius around the surrounding neighborhoods. After the venues were identified, they had to go through some data

preprocessing before we could finally apply the K means machine learning technique to segregate the similar neighborhoods in xx clusters. The venue categories were transformed to dummy values against neighborhoods by using one hot encoding technique. In order to normalize this dataframe we converted the categories count into mean values and grouped them by neighborhoods. Based on the mean, top 3 common venue categories were sorted to finally apply the K means clustering.

K means clustering has been chosen because it simply discovers the underlying patterns and group similar data points. This aided in identifying the similar and dissimilar neighborhoods in Manhattan and Borough.

The clustered data was then visualized using the folium map library to conclude our analysis.

Results



1. Cluster 1 = Red
2. Cluster 2 = Purple
3. Cluster 3 = Blue
4. Cluster 4 = Sea Green
5. Cluster 5 = Orange

The clustered data visualization makes it clear that cluster 4 is only existing in the Bronx. The remaining four clusters are existing in Manhattan and Bronx. Therefore, cluster 4 is an outlier

and it is implied that it may be the cause of epidemic spread of COVID 19 in the Bronx as compared to Manhattan.

Discussion

The results clearly depict that cluster four is an outlier between Manhattan and Bronx so it may be the reason for the spread of COVID 19. The top venue categories in this cluster are as follows.

Park	12.0
Pizza Place	3.0
Plaza	4.0

The analysis reveals that the public places like parks and plazas are the hotspots for the spread of COVID 19. This analysis can be used by governments to design the public health policy. For instance, places where COVID 19 has not spread exponentially, the government can close the parks and plazas to keep the situation from worsening. Also, they can implement strict prevention control measures near parks and plazas to avoid the spread of COVID 19.

In places where COVID 19 is already in full swing, governments can isolate the areas closest to the parks and plazas so that the spread is contained in these areas. The same can be implemented for businesses. Markets and offices near parks and plazas may be closed till the spread is controlled.

Limitations and Suggestions

The model could be tested on limited data only as Foursquare API calls are limited. If access to this API was unrestricted the same model could be tested for entire of United States or any other country. Further, there might be other factors like behavior of people which result in the spread of COVID 19. These factors have not been leveraged in this analysis.

Conclusion

This project is an attempt to analyze the hotspots leading to the spread of COVID 19 in the boroughs of New York. The analysis was inline with the consensus that the virus is spread through public places. The study revealed that the outlier in Bronx which has the highest number of COVID 19 cases are parks and plazas. This study can be used to minimize the spread of COVID 19 cases by implementing proactive health policies by governments and social members of the society.

References

https://news.google.com/covid19/map?hl=en-PK&gl=PK&ceid=PK%3Aen

https://www1.nyc.gov/site/doh/covid/covid-19-data.page