Capstone Project: Battle of Neighborhoods

August 3, 2020

1 Introduction

With the rise of the present COVID-19 pandemic there has a been a lot of strain on how countries are handling the ongoing crisis,and how people are adapting to the change in lifestyle. Since travel restriction have been imposed on almost all countries and the virus spreading as rapidly as ever it has come to a point where people have started to accept a reality in which they live along with the virus and hence neighborhood planning comes into play. Hence there comes a question of whether which neighbourhoods or localities are ideal for one to start rehabilitating both for an individual safety perspective and that of the overall commonwealth in which city officials can structure or remap cities to be more COVID-19 resistant or to plan reallocation of patients and the non infected in such a away that the overall population is safe.

This could be studied on how certain neighborhoods have an underlying similarity which might contribute to how communicable disease are spread. Along with this, the current study with existing neighborhoods correlating the spread of the diseases to that of the neighbourhood similarities will be essential to study the spread of the virus in similar neighborhoods as a prevention mechanism for future cities and for city officials to plan out for the future. Identifying neighborhoods with potential infrastructure that are inclined to the urgent and essential physical medical care to score neighborhoods based on how COVID resistant/equipped they are.

New York has a long tradition of multicultural immigrants and is the largest city in the United States. In 2019 New York City was home to more than 8.3 million residents making it the most populous city in the US,it also is the financial capital of USA .It is also by far the hardest-COVID-19-hit U.S. city as a result New York seems to be the ideal to be modelled around for this study.

2 Data

For this study we'll be using venue data sourced from the Foursquare API,which will give the nearest venues and even specific venues like hospitals for each neighborhood in New York.Along with this data I'll be using data from the official NY city gov data for COVID-19 github:`https://github.com/nychealth/`

`coronavirus-data` for data regarding COVID-19 cases,deaths,testing etc and also the geojason file for the geographical boundaries for the choropleth map.The coordinates of each neighbourhood was sourced from a python package called uszipcode by utilizing the zip codes of each neighborhood.

```
search = SearchEngine(simple_zipcode=True)
l=pd.DataFrame()
for i,zipc in enumerate(covid_df['MODIFIED_ZCTA']):
    zipcode = search.by_zipcode(zipc)
    k=zipcode.values()
    l=l.append({'lat':k[7],'long':k[8]},ignore_index=True)
```

```
l.head()
```

	lat	long
0	40.750	-73.990
1	40.720	-73.990
2	40.730	-73.990
3	40.700	-74.020
4	40.705	-74.005

Figure 1: Coordinates sourced from zip codes using uszipcodes packages

3 Methodology

3.1 Exploratory Data Analysis

In the sourced data there were 177 neighbourhoods which belong to 5 different boroughs: Bronx,Brooklyn,Staten Island,Queens and Manhattan

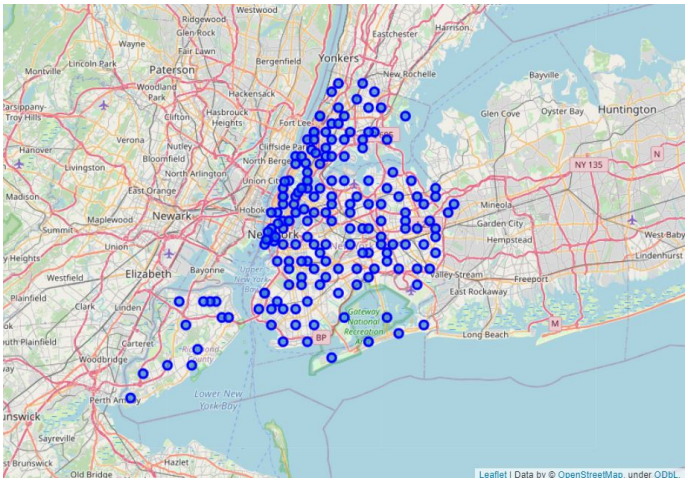

Figure 2: 177 Neighbourhoods of New York

A choropleth map was made of New York by dividing on the bases of the different neighborhoods,from this it was evident that ironically the most hit neighbourhood was Corona/North Corona having 4917 cases as shown below

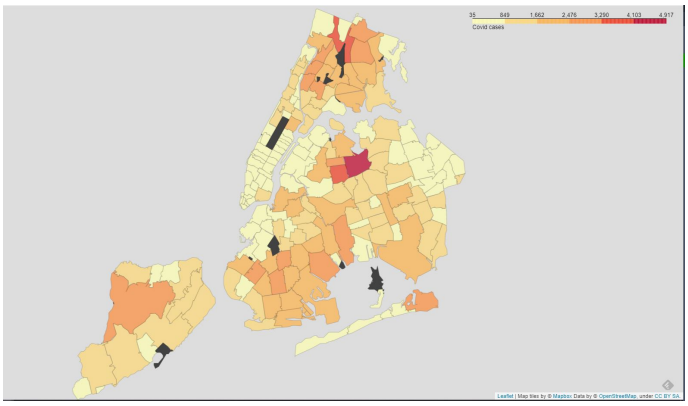

Figure 3: Distribution of COVID cases in terms of neighborhoods

On analysing the different boroughs it was found that Queens had reported the most number of COVID positive cases of 65219 (as of 29/07/2020) It has a rich demographic compared to the rest of the borough and is also the largest borough geographically.It is the second largest borough in terms of population in 2019 with an projected population of 2,253,858. Hence we will be using queens as the model for the study.

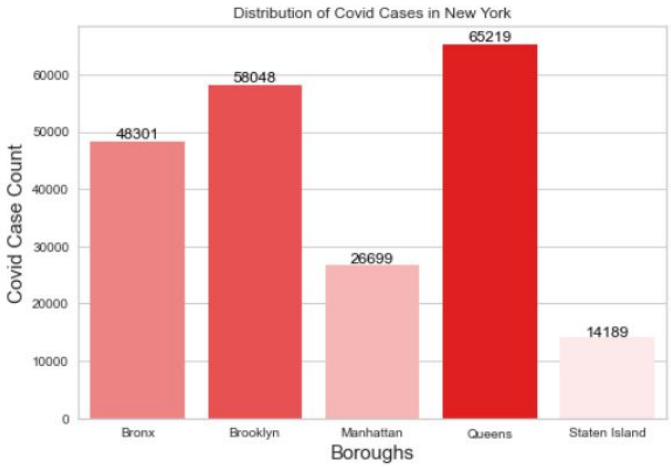

Figure 4: Distribution of COVID cases according to boroughs

Using the foursquare API the top 100 venues were scrapped within a radius of 500 m,out of this the top 10 most common venues were isolated for each

neighborhood and studied.

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Airport/East Elmhurst	Bus Station	Gas Station	Pizza Place	Donut Shop	Bakery	Supermarket	Mexican Restaurant	Chinese Restaurant	Peruvian Restaurant	Bank
Airport/South Jamaica/Springfield Gardens/St. ...	Southern / Soul Food Restaurant	Japanese Restaurant	Laundromat	Breakfast Spot	Chinese Restaurant	Gym	Pharmacy	Sandwich Place	Market	Bus Station
Arverne/Broad Channel	Beach	Surf Spot	Garden Center	Food	Metro Station	Construction & Landscaping	Playground	Pizza Place	Deli / Bodega	Chinese Restaurant
Arverne/Edgemere	Surf Spot	Bus Stop	Beach	Metro Station	Café	Board Shop	Taco Place	Supermarket	Brewery	Caribbean Restaurant
Astoria (North)	Pizza Place	Deli / Bodega	Italian Restaurant	Grocery Store	Peruvian Restaurant	Chinese Restaurant	Sandwich Place	Restaurant	Residential Building (Apartment / Condo)	Diner

Figure 5: Top 10 venues for each neighbourhood

3.2 Clustering

3.2.1 K-means Clustering

The COVID data from the New York health department had data related to the case count,the case rate,population of neighborhood,death count,percent positive and the total number of COVID tests performed. Out of the above said data points the population count,the total number of tests were removed from the data as these data would impact the clustering process due to its high values and which inherently do not have an effect on the spread of the disease. The above mentioned data and the frequency of the 100 venues nearest to each location were merged and normalized before being passed on to the k means algorithm. Now since the inherent number of clusters were not known elbow method was used to find the optimum K value using the distortion score,This was done using a package called yellow-brick.

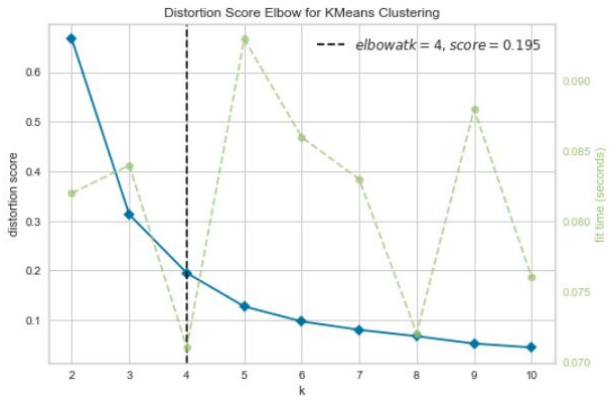

Figure 6: Elbow method for k means using Distortion score

4 Results

From the elbow method it was identified that the optimum K value was 4 ie;there will be 4 clusters.

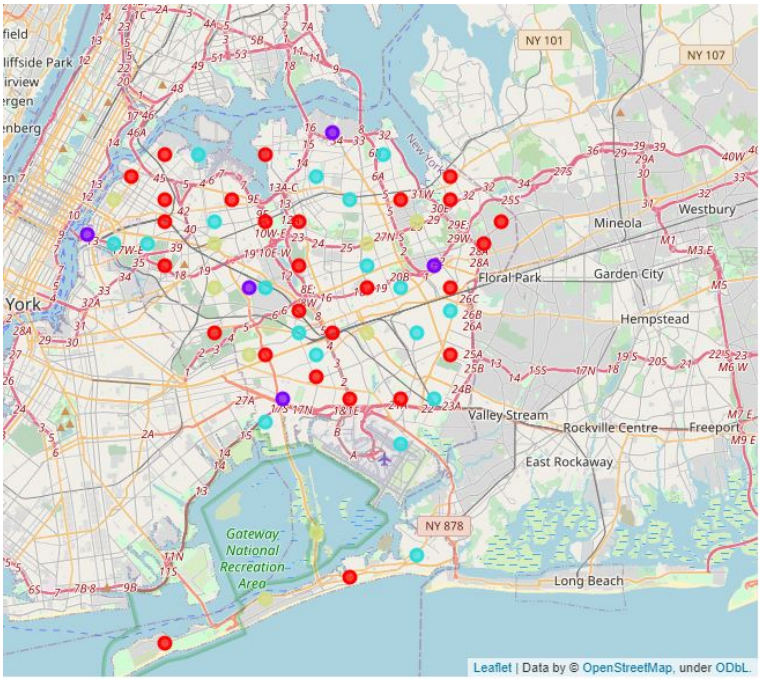

Figure 7: Clustered neighborhoods in Queens

The resulting clusters formed were then analysed and found that each cluster could be divided into 4 clusters of high cases-moderate rate,moderate cases-high rate,low case-moderate rate,moderate case-low rate

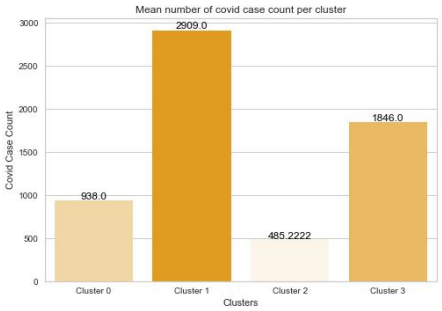

Figure 8: Mean no. of cases per cluster

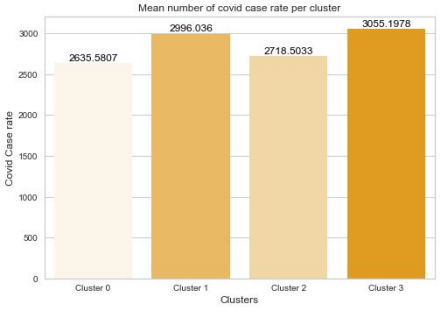

Figure 9: Mean case rate per cluster

As shown in the graphs it can be seen that cluster 1(purple) has a high COVID risk having the highest mean COVID cases count and also a significantly high COVID case rate as compared to other clusters, but the said cluster has only 5 neighborhoods out of the total 59 neighbourhoods in queens which is just about 8 percent of the overall number of neighborhoods in queens same goes for the other high risk cluster;cluster 3 which also has a smaller proportion of the total number of the neighbourhoods (9 neighbourhoods) but as for the low risk clusters they contain majority of all neighbourhoods in queens.

Now to understand the clustering based on venues and to what ways do venues contribute into such high risk clusters the venues of each cluster was analysed and from all the venues there was correlation between venues which are hot spots for people to gather and mingle and the amount of COVID cases in the said cluster. Hence such gather venues were isolated such as nightclubs,movie theaters,stadiums etc and their frequency of occurrence was tallied across each neighbourhood,restaurants were not considered for this as restaurants were the most popular venues rated using foursquare and hence would skew the results and not provide an accurate representation.

Post the 'scoring' the mean value of each cluster was identified and an association was clearly visible between the occurrence of gathering venues and the number of COVID cases in each neighbour hood as shown below.

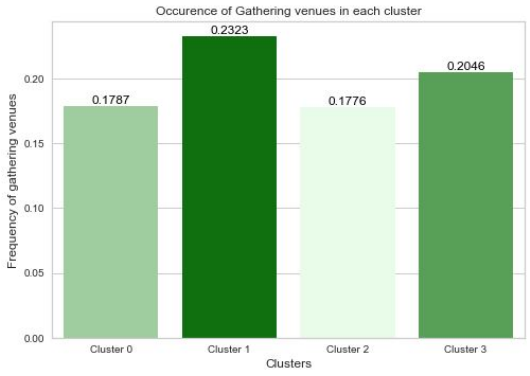

Figure 10: Frequency of gathering venues in each cluster

As seen above the high risk clusters(cluster 1 and cluster 3) have higher gathering spots hence can be a reason to the spread of the virus in such neighbourhoods similarly the opposite can be seen with the low risk clusters(cluster 0 and cluster 2) where they have fewer gathering spots and hence overall this shows a strong correlation between gathering spots and the spread of the virus.

As the disease grows one ones to wonder whether if the said person is to contract the disease how accessible are the healthcare facilities hence a similar way of scoring of gathering spots another metric was calculated using the frequency of emergency care and physical health care which can help for after contracting the infection. This was calculated after scraping the neighbourhoods with the

medical care category in the foursquare API,focusing on the essential medical care for the infection out of all the categories within the overall medical centre category like emergency care,doctors office etc.As seen below it was calculated and tallied according to each cluster.

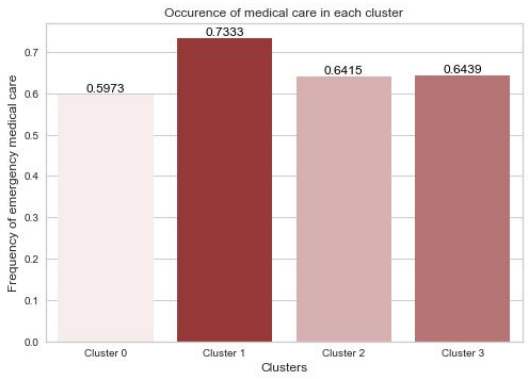

Figure 11: Frequency of medical care in each cluster

On analysing its evident that the high risk clusters have a higher medical care score but at the same time cluster 2 has a comparable and high value for medical care of 0.64 almost the second highest of all the clusters.

5 Discussion

- From the study we identifying there is correlation between gathering venues and the rate of COVID cases,even though correlation cannot be associated to causation there is a high chance neighbours having higher venues where people had close interactions have a high chance of being more vulnerable to be infected by the disease.

- Now if a family were to choose on which neighbourhood to live in queens, The ideal neighbourhood would be, any in cluster 2.There is a significantly less chance of acquiring the disease maybe due to being the cluster having the least number of venues of close gathering proximity. This is also an ideal neighbourhood selection,since even if one were to contract the disease there is significant healthcare infrastructure to handle it as compared to low risk cluster like cluster 0. Thus making neighbourhoods in cluster 2 to be the ideal neighbourhoods for family of high risk to move into.

- The Gathering venues and its correlation between the spread of the virus can be used to model neighbourhoods having a high risk population of contracting the disease(Elderly people) and hence be used in other models.

6 Conclusion

In conclusion, Neighbourhoods of New York city were studied using various python libraries and clustering algorithms by utilizing data sourced from Foursquare API and New York city health department.The scope of the analysis is somewhat limited maybe due to the dated data of the foursquare API and how many actual current hospitals were rated. The study was isolated to a particular borough of New York city and hence there can be an issue of scaling the same model for a completely different environment. With the data analysed and procured for queens I stand by the analysis and the recommendation for ideal neighbourhood for safety in queens made.