IBM DATA SCIENCE CAPSTONE PROJECT

IMPACT OF VENUES ON COVID-19 CASES RATE IN NEW YORK CITY

1. Introduction

1.1 Background

The Corona Virus pandemic has had a devastating impact on the livelihoods of people and claimed lives all over the world. This project focuses on the spread of Covid-19 in New York City. New York is the most populous city in the United States, with an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles (784 km2). New York City is also the most densely populated major city in the United States. New York City is widely considered the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports in the United States and all over the world. Given its overall demographics, New York is one of the city's most susceptible to the rapid spread of Covid-19 globally[1]. According to some studies, the most populous boroughs in New York City, Queens, and Brooklyn, likely served as the major hubs of coronavirus disease (COVID-19) spread in the spring of 2020[2].

1.2 Problem

This study intends to assess the spread of COVID-19 in New York City across its neighborhoods. It has been argued that some of the most populous boroughs in New York City contributed significantly to the spread of COVID-19 at the onset of the pandemic. In this context, this project intends to cluster neighborhoods in New York City and map out venues in the specific clusters to see if there are any common features in the neighborhoods or clusters with high COVID-19 infections. The World Health Organisation notes that current evidence suggests that the virus spreads mainly between people who are in close contact with each other, and the virus can also spread in poorly ventilated and/or crowded indoor settings, where people tend to spend longer periods of time. In view of this, the initial thinking is that neighborhoods with a lot of restaurants, coffee shops and other venues where people crowd together are likely to have higher cumulative cases. In the end, the project considers choropleth maps showing COVID-19 red zones for neighborhoods in New York, and maps venues around centroids of calibrated clusters to assess the spread of COVID-19 based on venues associated with each cluster, as well as venues associated with each Borough.

1.3 Interest

This study is most useful for Health Care Policy makers and Medical Practitioners looking to understand how venues might influence the spread of a disease like COVID-19 for future and current use. It might also be useful for individuals looking to protect themselves by understanding how frequenting certain venues might increase chances of infection. An understanding of such dynamics is vital for coming up with targeted policy measures for most vulnerable neighborhoods such as, providing necessary medical advice and assistance and enforcing movement restrictions.

2. Data

The project will utilise data outlined below:

i. Download New York City COVID-19 Data by ZIP Code from url:https://raw.githubusercontent.com/nychealth/coronavirus-data/master/totals/data-by-modzcta.csv . This data presents information about COVID-19 infections in each

[1] See https://en.wikipedia.org/wiki/New_York_City
[2] See https://www.sciencedaily.com/releases/2021/05/210520145332.htm

neighborhood corresponding to Zip codes, as well as geographical coordinates of the neighborhoods, among others. The center of the clusters obtained will present geographical coordinates and the corresponding COVID-19 cases rate.

ii. Download New York neighborhood data from url: https://cocl.us/new_york_dataset (JSON file). This data contains Federal Information Processing Standard (FIPS) code for New York Counties as well as geographical coordinates for each county.

iii. Use Geolocator package to obtain coordinates of New York City and of neighborhoods in New York City.

iv. Foursquare API: to explore venues around centroids or centers for clustered neighborhoods in New York City.

FourSquare API will return lists of venues using the co-ordinates for the centroids of the clusters. It will also return the latitude and longitude for each venue.

The study uses the explore API call to identify venues within a walking distance, which is assumed to be 1 kilometre, from the center of each of the clusters.

3. Methodology

The study employs K-Means clustering to cluster neighborhoods in New York. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. A cluster is generally defined as a collection of data points aggregated together because of certain similarities. This method involves defining a target number K, which specifies the number of centroids you need in the dataset. A centroid is an imaginary or real location representing the center of the cluster. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. This means that the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

The purpose of this study is to explore the spread of COVID-19 in New York City with specific focus on differences in levels of infection across clusters, and how dissimilarities in clusters influence the spread of the COVID-19 disease, regarding popular venues in each cluster, by assessing the differences in total COVID-19 cases reported against locations of centroids in each cluster. The initial stage of the analysis is the clustering of neighborhoods in New York using an arbitrary K (K = 6), which in this case is equal to the number of Boroughs in the New York COVID-19 Data. After this stage, centers of the clusters are obtained and their coordinates are used to explore neighborhoods withing a walking distance (which is assumed to be 1 kilometre) of each center. The information obtained from Foursquare explore API call is then used to identify top 10 most popular venues in each cluster. The data on the popular venues is assessed to see how dissimilarities in venues influence COVID-19 infections. In addition, centers' locations are also evaluated against the number of COVID-19 cases to see if centers located in fundamentally different parts of the city have significantly different levels of infection. The project intends to see if the chosen number of clusters would mimic the dynamics observed in the data for the New York City Boroughs. Therefore, in the end, the COVID-19 rates for each centroid are assessed against the respective Boroughs.

4. Results

4.1 Data Analysis

Queens has the most neighborhoods according to the New York City COVID-19 Data and Staten Island has the lowest number of neighborhoods (see Figure 1). The data also shows that Queens has the highest rate of COVID-19 infection, with a total COVID-19 cases rate of 564,932, which is almost double the infection rate in the second highest county, Brooklyn (see Figure 2).

Figure 1: Neighborhoods per Borough

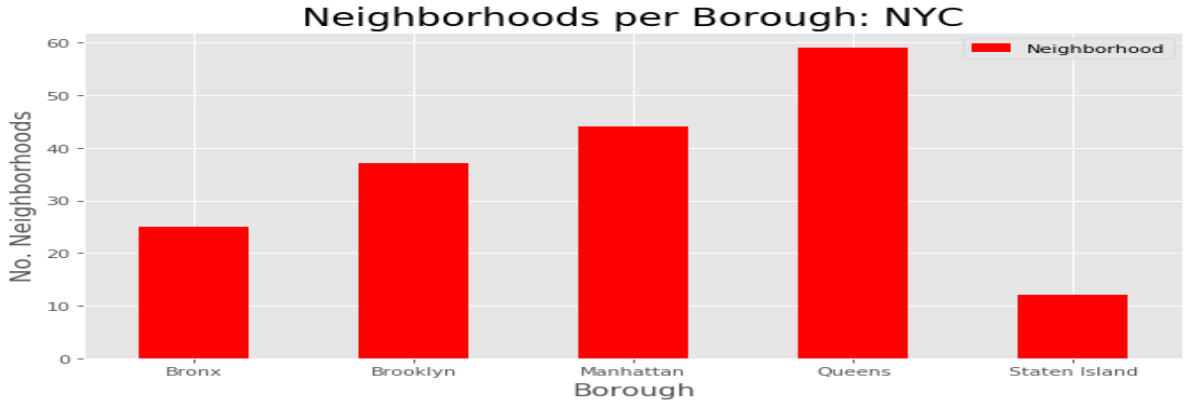
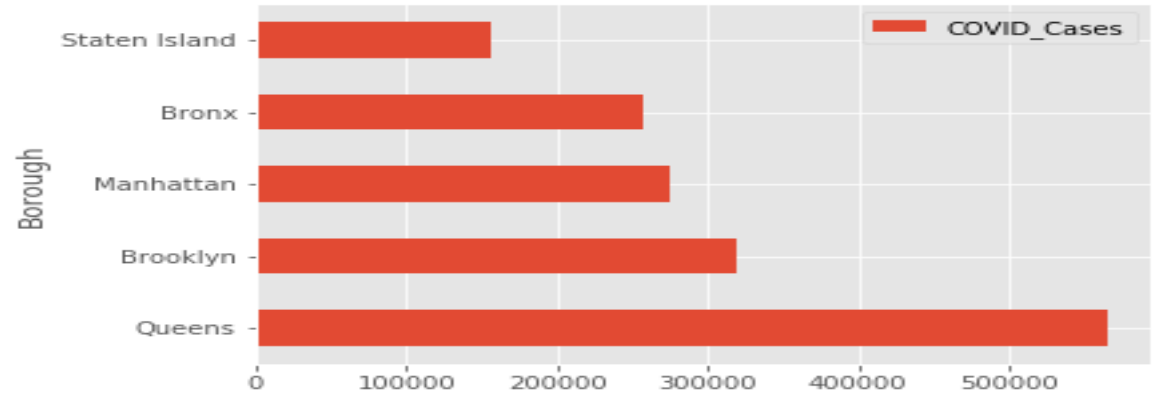


Figure 2: COVID-19 Cases in Each Borough

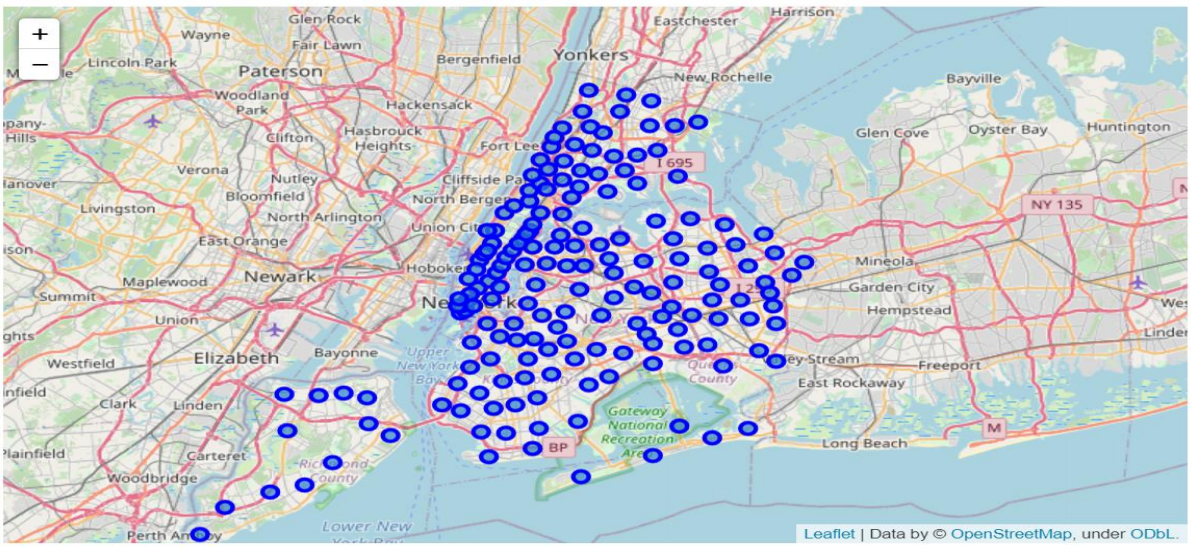


4.2 Clustering

The K Means machine learning technique is used to cluster Neighborhoods in New York City. After clustering, the centers or centroids of each cluster are computed. The computed centroids give geographical coordinates as well as the number of COVID-19 cases corresponding to those coordinates. The next step is the mapping of popular venues corresponding to those centers, to assess the location of the center of clusters against the number of recorded COVID-19 cases corresponding to those coordinates.

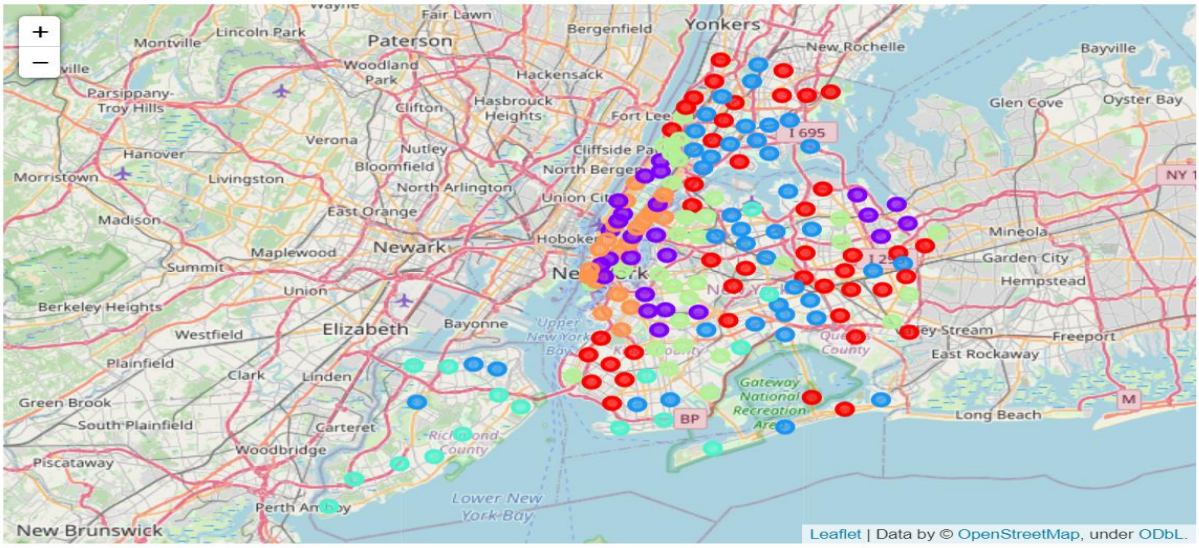
Map in Figure 3 shows the mapping of neighborhoods in New York before clustering.

Figure 3: Mapping of New York City Neighborhoods



The map below shows the neighborhoods in New York City after clustering using K Means algorithm, with K equal to 6.

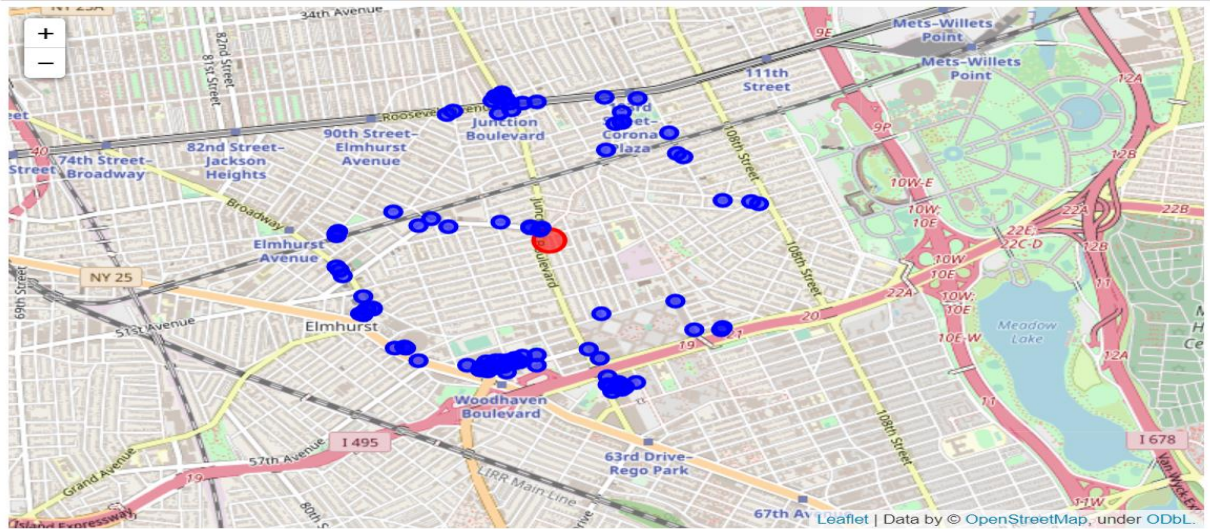
Figure 4: Clusters of Neighborhoods in New York City



Cluster 1

The center of cluster 1 corresponds to Junction Boulevard, with a significantly high number of venues with a walking distance from the center. This center recorded about 9,000 COVID-19 cases.

Figure 5: Mapping of Nearby Venues from Center of Cluster 1



Cluster 2

Cluster 2 (see Figure 6) around Queens Boulevard. Queens Boulevard is a major thoroughfare in the New York City borough of Queens connecting Midtown Manhattan, via the Queensboro Bridge, to Jamaica, and it is a significantly busy location with many venues around the center of the cluster[3]. The neighborhood at the center of this cluster has a COVID-19 cases rate of just above 6,000.

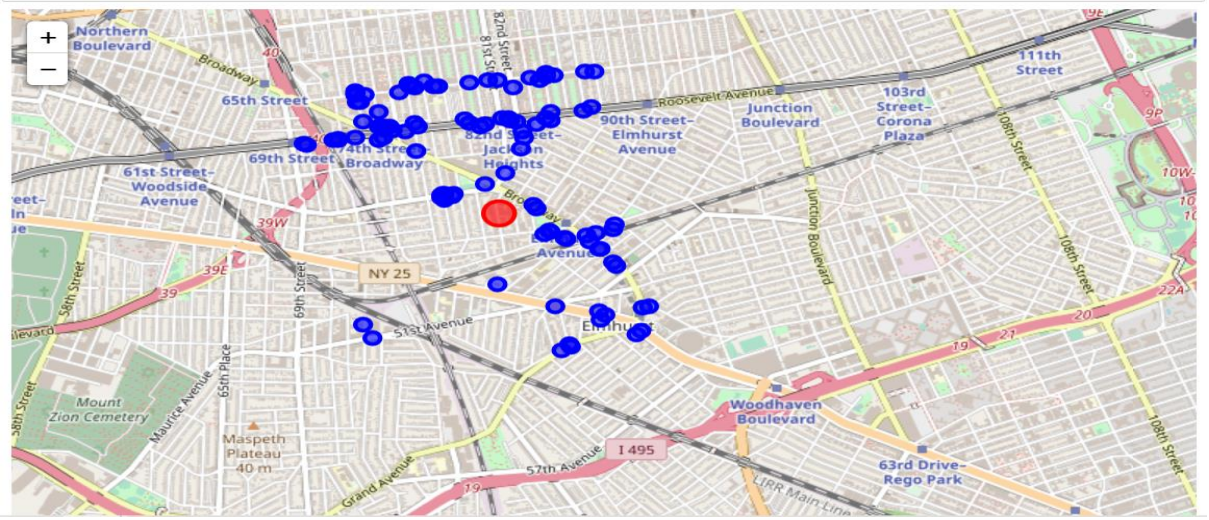
[3] https://en.wikipedia.org/wiki/Queens_Boulevard

Figure 6: Mapping of Nearby Venues from Center of Cluster 2

Cluster 3

The center of cluster 3 (see Figure 7) is around Broadway in Manhattan and as expected, it has lots of venues within a walking distance, given that Broadway is the busiest street in New York City. This centroid has a COVID-19 case rate of around 11,000.

Figure 7: Mapping of Nearby Venues from Center of Cluster 3



Cluster 4

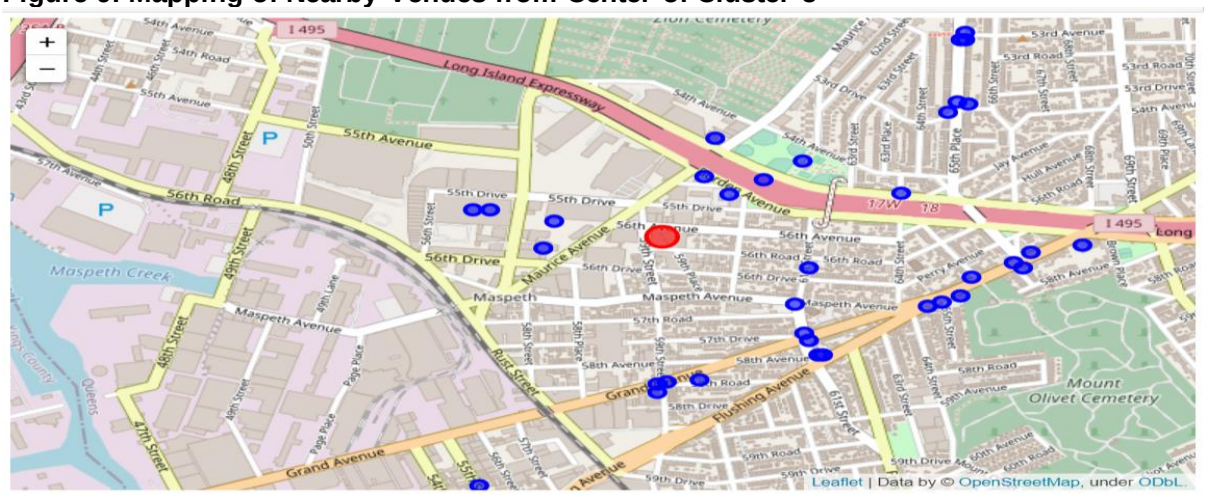
This centroid lies at a beach in Staten Island by the Verrazano Narrows Bridge (see Figure 8). This area is not busy in terms of nearby venues. However, given its location, it is a popular tourist site. Contrary to the initial thinking that a location surrounded by restaurants, coffee shops and other common venues frequented by people on a day-to-day basis should have higher COVID-19 cases, this neighborhood recorded the highest number of COVID-19 infections at just above 13,400.

Figure 8: Mapping of Nearby Venues from Center of Cluster 4

Cluster 5

The center of this cluster is around the Long Island Expressway and as expected, it also has many venues within a walking distance (see Figure 9). This neighborhood recorded just above 6,000 COVID-19 cases.

Figure 9: Mapping of Nearby Venues from Center of Cluster 5



Cluster 6

The center of this cluster is around the Union Square area. Even though stalls of the Union Square Greenmarket draw crowds for local produce and artisanal food, the neighborhood recorded the lowest number of COVID-19 cases at just below 5,000.

Figure 10: Mapping of Nearby Venues from Center of Cluster 6

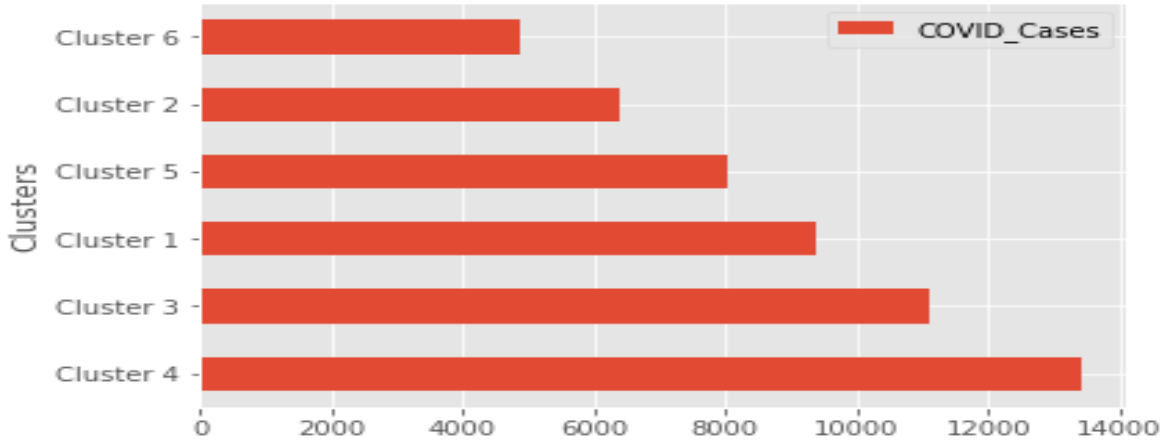
4.3 Summary of Results

The results show that the highest COVID-19 cases rate corresponds to a neighborhood in Staten Island whose distinctive features include a beach, hotels, tourism activity, and government buildings. A neighborhood corresponding to Broadway in Manhattan corresponds to the second highest infection zone. Cluster 1 is a neighborhood in Queens County, coming at third in terms of COVID-19 infections. The data shows that Queens has the highest cumulative number of infections. However, the center of the cluster in or around Queens neighborhoods comes at third with most popular venues including mostly restaurants and coffee shops, clothing stores, pharmacies, and phone shops-coming within a walking distance from the center. Interestingly, Queens being the largest Borough in New York City has another centroid corresponding to neighborhoods around Maurice Avenue (Cluster 5) with the fourth highest COVID-19 case rate. Common venues within a walking distance in this neighborhood include restaurant and other eateries, as well as a park, pubs, grocery stores, pharmacies, and home service centres. Cluster 2 also corresponds to a neighborhood in Queens, coming a fifth. This cluster has no distinct features compared to Cluster 5, in terms of popular nearby venues from the center. The cluster with the lowest COVID-19 cases rate represents neighborhoods around Union Square in Manhattan, whose most popular venues within a walking distance are also restaurants, coffee shops and grocery stores. In summary, the only purely distinctive center corresponds to a neighborhood in Staten Island, which also has the highest number of infections.

Table 1: Top 10 Most Common Venues Vs COVID-19 Cases

Clusters	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	COVID_Cases
Cluster 4	Park	Hotel	Lighthouse	Bus Stop	Government Building	Food Truck	Tourist Information Center	Historic Site	Donut Shop	Dog Run	13414.234375
Cluster 3	Thai Restaurant	Mexican Restaurant	Food Truck	Indian Restaurant	South American Restaurant	Vietnamese Restaurant	Bakery	Grocery Store	Filipino Restaurant	Indonesian Restaurant	11090.194500
Cluster 1	Mexican Restaurant	Coffee Shop	Cosmetics Shop	Pizza Place	Clothing Store	Mobile Phone Shop	Argentinian Restaurant	Bakery	Pharmacy	South American Restaurant	9359.625581
Cluster 5	Deli / Bodega	Pizza Place	Bakery	Grocery Store	Donut Shop	Pharmacy	Pub	Home Service	Peruvian Restaurant	Park	8031.434667
Cluster 2	Pizza Place	Bar	Donut Shop	Bakery	Mexican Restaurant	Coffee Shop	Grocery Store	Chinese Restaurant	Discount Store	Deli / Bodega	6367.787200
Cluster 6	Mediterranean Restaurant	Italian Restaurant	Juice Bar	Coffee Shop	Park	Pizza Place	Cosmetics Shop	American Restaurant	Grocery Store	Gourmet Shop	4891.963478

Figure 11: COVID-19 Cases for Each Centroid



5. Discussion

From the results, most cluster centers are remarkably similar in terms of nearby venues, dominated by restaurants, coffee shops, shopping areas and other types of eateries including pubs and food trucks. The neighborhood with the most distinctive location corresponds to the highest recorded COVID-19 cases rate. This neighborhood in Staten Island is a prime area for tourists and boasts one of the most popular beaches, the South Beach, as well as the Franklin D. Roosevelt Boardwalk. Beachgoers can sunbathe while taking in a lovely view of the Verrazano Bridge[4]. In addition, the neighborhood has government offices, hotels, a Tourist Information Centre, and a Historical Site, as well as other tourist attractions. Given the fact that COVID-19 was most carried across borders, it is plausible to argue that this area was most prone to COVID-19 infections, and probably played a role in spreading the Corona virus across counties in New York City as well as to the rest of the country.

Contrary to the initial thinking that restaurants and other eateries are likely to be the hubs for transmitting infections, from this analysis it appears that tourist areas are most vulnerable. It is also important to note that the K Means clustering method does not consider clusters within clusters, and as a result most centers picked by the algorithm in this exercise mostly correspond to City Centres. This presents a problem for the analysis since most testing and recording of infections is mostly likely done in residential neighborhoods due to the associated movement restrictions. The data shows that Queens has the most cumulative number of infections. This can be explained by the fact that it is the most populous Borough, and at the same time failure to capture the dynamics in the data pertaining to Queens County might be due to the clustering method used. Despite this shortcomings, higher infections in tourist areas presents a plausible thesis for formulating policy to deal with the COVID-19 pandemic by targeting these areas, specifically crowded beaches, parks, and historical sites. As a result, the study suggests that Healthcare Medical Practitioners, Policy makers and individuals should consider popular tourist attractions as hubs for rapid transmission of the pandemic and take appropriate actions to curb the scourge.

6. Conclusion

The goal of this study was to assess how dissimilarities in nearby venues in neighborhoods influence the transmission of COVID-19. The project finds that neighborhoods with higher numbers of venues like restaurants and coffee shops are not the most susceptible, and that popular tourist attractions present a higher risk of transmission. However, a more refined clustering of neighborhoods might likely capture the dynamics in Boroughs with the higher number of infections presented in the data.

[4] https://www.nycgovparks.org/parks/fdr-boardwalk-and-beach/facilities/beaches