IBM – Applied Data Science Capstone Project

Optimal allocation of Funding/Resources among the NYC Hospitals during COVID 19 Pandemic

*** Manna Samyal, August, 2020**

1. INTRODUCTION

1.1 Background

COVID-19 is an infectious disease caused by the most recently discovered corona virus. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019. COVID-19 is now a pandemic affecting many countries globally. WHO announced COVID-19 outbreak as a pandemic on 11 March 2020.

213 Countries and Territories around the world have reported a total of 23,586,511 confirmed cases of the corona virus COVID-19 and a death toll of 812,527 deaths. Five countries most affected by this pandemic as of Aug 23, 2020 are shown in Table 1 –

COUNTRY	CASES	DEATHS
United states	5,874,146	180,604
Brazil	3,605,783	114,772
India	3,106,348	57,692
Russia	956,749	16,383
South Africa	609,773	13,059

Table 1: Five worst affected countries due to COVID 19 pandemic (till Aug 23, 2020)

1.2 Covid-19 situation in NYC

New York City, which was once the epicentre of the pandemic in the US, has reported a total of 236,822 cases and 23,658 confirmed deaths (Source: The New York Times-Aug 23, 2020). It hit its peak in terms of confirmed daily deaths from the virus on April 7, with 597 deaths.

1.3 Business Problem

In this project, I am going to analyse COVID-19 situation in NYC. Number of cases in each neighbourhood along with hospitals located in the neighbourhood will be used to analyse which hospitals possibly have higher need for resources. During a pandemic, when health resources are likely to be limited, setting priorities and rationing resources should be justified. Hospitals in areas with higher number of cases are more likely to face shortages of resources like supplies, protective gear for clinical staff, oxygenation, respiratory equipment, direct patient care including additional services and staffing.

1.4 Target Audience

Government and non-government organisations, non-profit and charity organisations, any volunteer willing to help via donations in terms of resources or time.

2. DATA COLLECTION

Data was collected from the following sources:

2.1 NYC COVID-19 data set

CSV file was downloaded from *https://www1.nyc.gov/site/doh/covid/covid-19-data.*page. The dataset comprises 5 Boroughs and 122 Neighbourhoods, number of Covid-19 cases in the past 4 weeks (August, 2020), ZIP codes of each neighbourhood and Covid-19 case rate for different weeks.

2.2 NYC data set

Using neighbourhood names from Covid-19 dataset, latitudes and longitudes for each neighbourhood were extracted using geocoder. The coordinates were converted into a data frame and added to Covid-19 dataset.

2.3 NYC Hospitals

Hospitals in each neighbourhood are fetched from foursquare API. Latitude, longitude and name of each neighbourhood are passed to foursquare API along with radius size of 1km and limit 100. The results were converted into a data frame.

3. METHODOLOGY & DISCUSSIONS

STEP – 1: Dataset Cleaning and Merging

The Covid-19 dataset was clean without any null values. ZIP code, case rate columns were removed since they were not required. Many rows comprised data of same neighbourhood so they were grouped together. The final dataset had 122 unique neighbourhoods and their respective Covid-19 cases, belonging to 5 boroughs.

After extracting coordinates from geocoder using neighbourhood names and adding columns, the dataset now comprises of names of neighbourhood, borough, and total cases of past 4 weeks along with coordinates of each neighbourhood.

To use Foursquare API, developer account was created. Foursquare ID and secret key were obtained which were used to make API calls passing in the geographical coordinates of each neighbourhood from Covid-19 dataset in a python loop. In order to filter results to get only hospital data, hospital ID was passed along with limit of 100 within radius 1000 meters. Foursquare returned the call with hospital data in JSON format, from which hospital name, hospital category, neighbourhood name and its geographical coordinates were extracted into a data frame.

Total 1171 hospital names were provided by foursquare of different categories like hospital, doctor’s office, eye doctor, hospital ward, pet service, veterinarian, emergency room, co working space, mental health office, medical centre, medical school, college science building, pharmacy, dentist’s office, building, it services, office, police station, and medical supply store.

Out of these, 986 belonged to category hospital, were kept and rest were discarded.

Names of hospital per neighbourhood were converted into dummy variables and their total sum gave the number of hospitals in each neighbourhood.

Covid-19 cases and hospital datasets were merged to get final data frame.

Top five rows of Table 2 are shown as below -

	Neighborhood	Borough	Total cases (4 weeks)	Latitude	Longitude	No. of Hospitals
0	Airport/South Jamaica/Springfield Gardens/St. ...	Queens	28	40.667350	-73.776709	1
1	Allerton/Baychester/Pelham Gardens/Williamsbridge	Bronx	55	40.863930	-73.843390	7
2	Allerton/Norwood/Pelham Parkway/Williamsbridge	Bronx	99	40.857969	-73.851246	8
3	Alphabet City/East Village/Stuyvesant Town-Coo...	Manhattan	31	40.732440	-73.977710	48
4	Arrochar/Midland Beach/Shore Acres/South Beach...	Staten Island	38	40.609550	-74.066170	1

Table 2: Areas wise Covid-19 cases and hospital dataset (top five rows)

Different columns of the table –

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 122 entries, 0 to 121
Data columns (total 6 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Neighborhood           122 non-null    object
 1   Borough                122 non-null    object
 2   Total cases (4 weeks)  122 non-null    int64
 3   Latitude               122 non-null    float64
 4   Longitude              122 non-null    float64
 5   No. of Hospitals       122 non-null    int64
dtypes: float64(2), int64(2), object(2)
memory usage: 6.7+ KB
```

Table 3: Colum view of Areas wise Covid-19 cases and hospital dataset

STEP - 2: Exploratory Data Analysis

The final compiled dataset was used to plot Number of Hospitals vs. Total cases for each Borough to analyse which neighbourhoods are possible outliers in the data. For this Matplot and Seaborn libraries were used. To visualise outliers scatter plot was used.

The following graph represents Borough – Queens. Here most of the neighbourhoods have three or less number of hospitals.

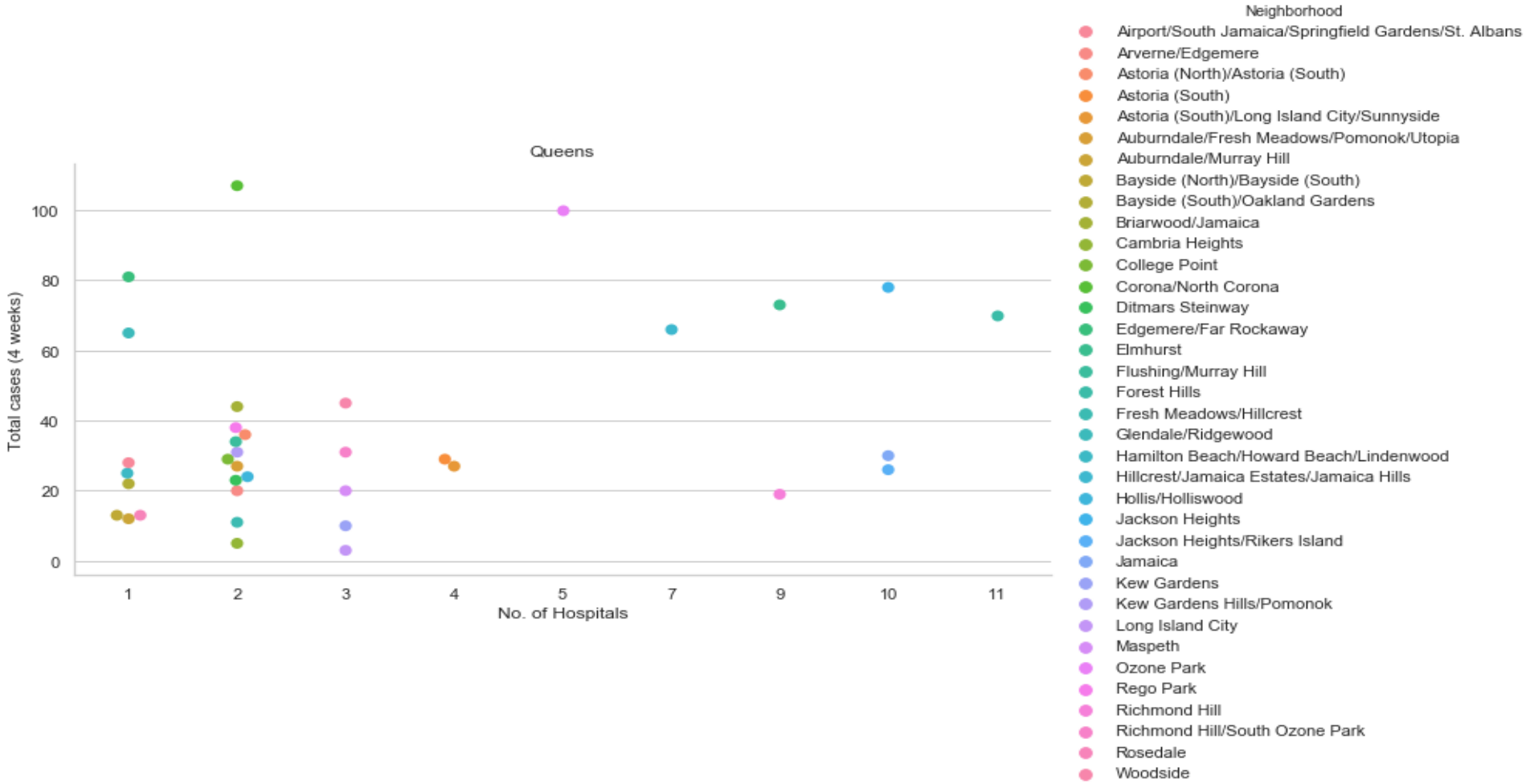

Figure 1: Covid-19 cases and number of hospitals in Borough-"Queens"

The following graph represents Borough – Bronx.

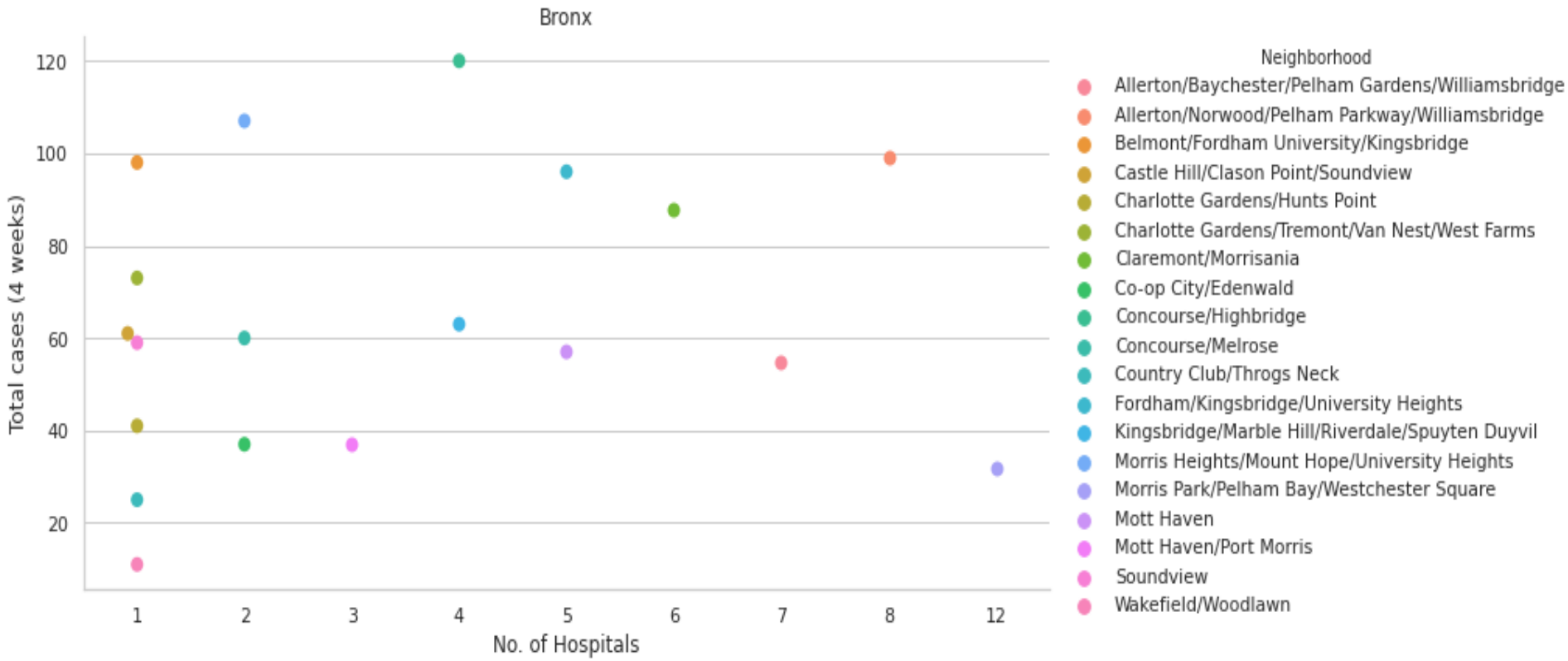

Figure 2: Covid-19 cases and number of hospitals in Borough-"Bronx"

The following graph represents Borough – Staten Island. The number of cases as well as number of hospitals is less since the island is least populated compared to other boroughs.

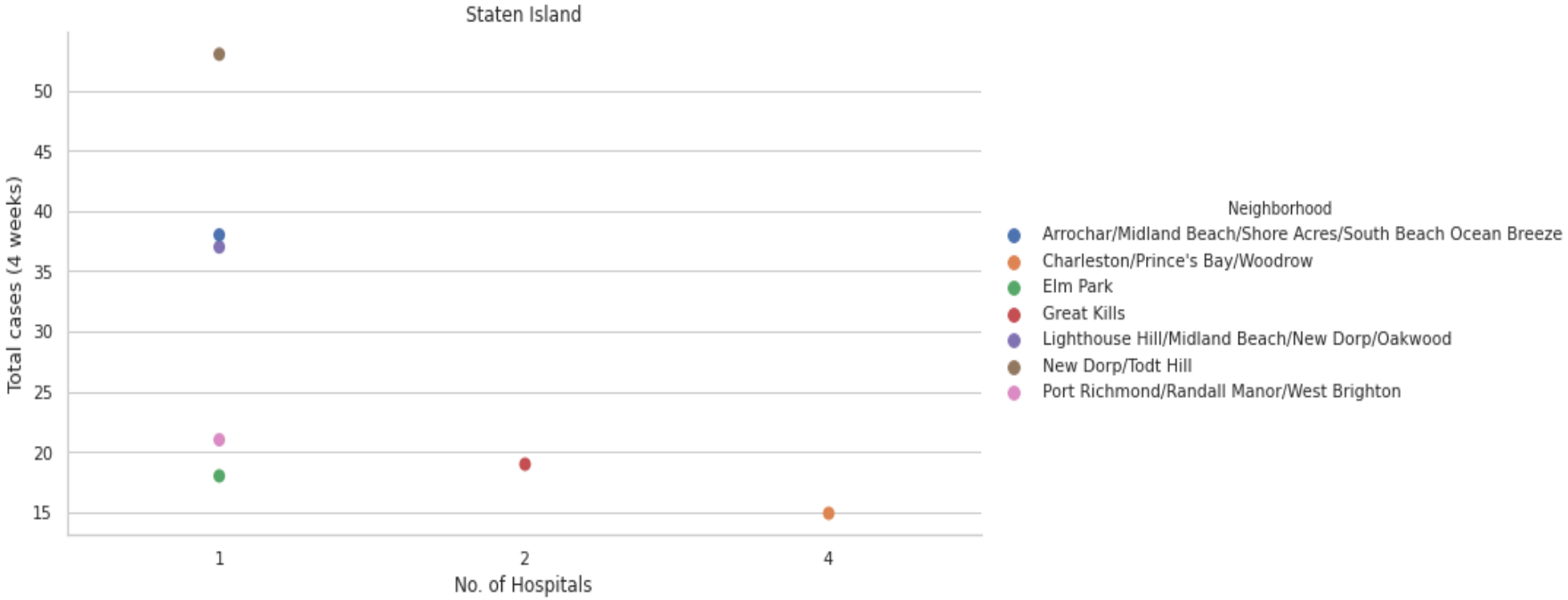

Figure 3: Covid-19 cases and number of hospitals in Borough-"Staten Island"

The following graph represents Borough – Manhattan. It's observed that neighbourhood East Harlem has 200 cases for 3 hospitals and is an outlier. There are also several neighbourhoods with very high number of hospitals, which can also be due to higher population in these areas compared to other neighbourhoods or other boroughs.

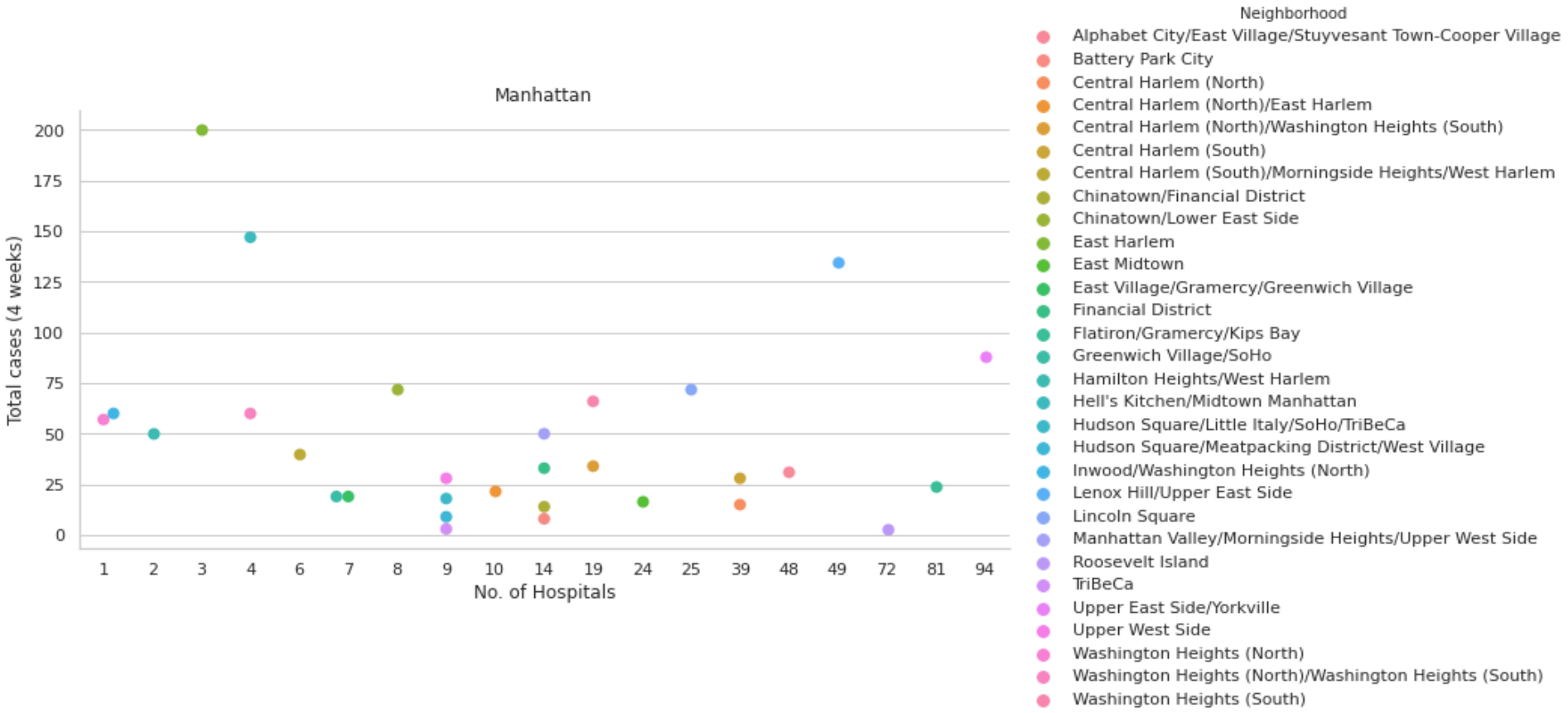

Figure 4: Covid-19 cases and number of hospitals in Borough-"Manhattan"

The following graph represents Borough – Brooklyn. In this borough the number of cases is high and number of hospitals is low. Sunset Park clearly is an outlier with over 400 cases and 10 hospitals.

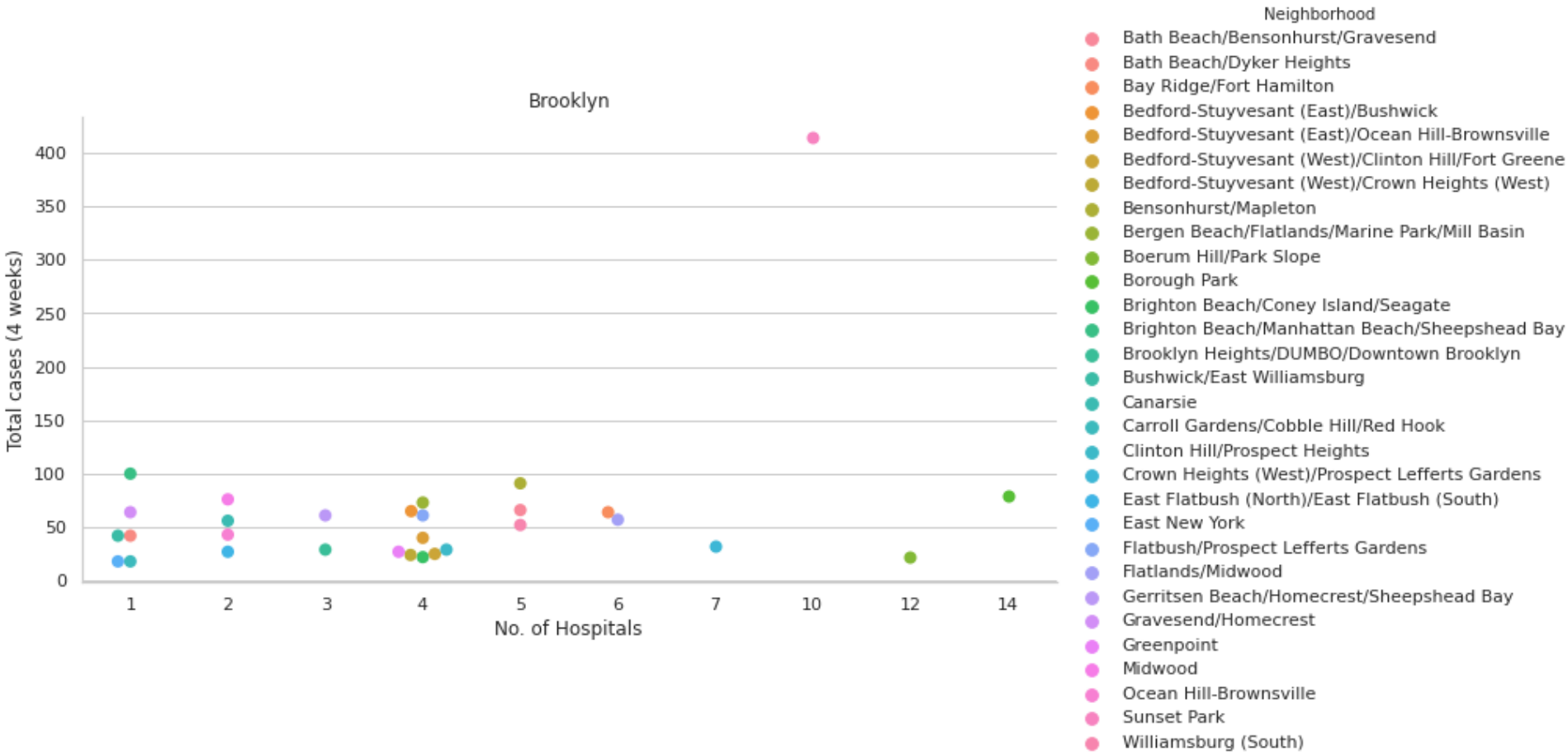

Figure 5: Covid-19 cases and number of hospitals in Borough-"Brooklyn"

For comparative analysis between each borough, bar graph was plotted between number of hospitals and total cases, where each bar's area is represented by percentage of hospitals.

In the graph, Brooklyn clearly requires attention with over 1750 Covid cases and 12.8% of total hospitals in NYC. Manhattan has higher number hospitals than other boroughs combined.

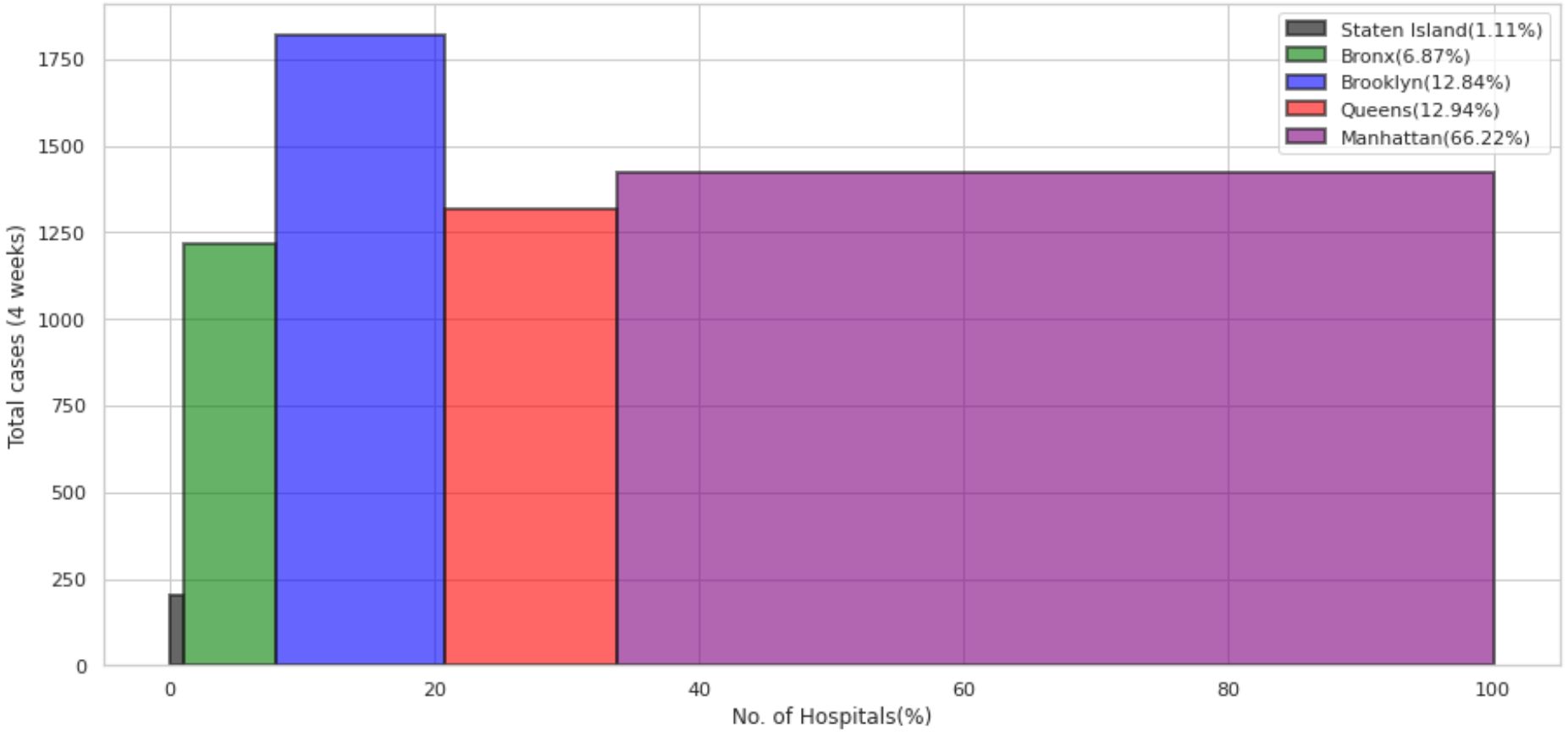

Figure 6: Graph depicting status of hospital availability vis a vis covid 19 cases &need assessment of resources (Borough Wise)

STEP - 3: Clustering Using K-Means

To analyse the data, K-Means clustering from Scikit-learn was performed. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms.

In order to do clustering, new dataset was made comprising of total number of cases and number of hospitals. This data was scaled using sklearn standard scaler to get best results from k-means.

Elbow method helps to select the optimal number of clusters by fitting the model with a range of values for K. The 'elbow' (the point of inflection on the curve) is a good indicator that the underlying model fits best at that point. From the graph below, value of K taken is 3.

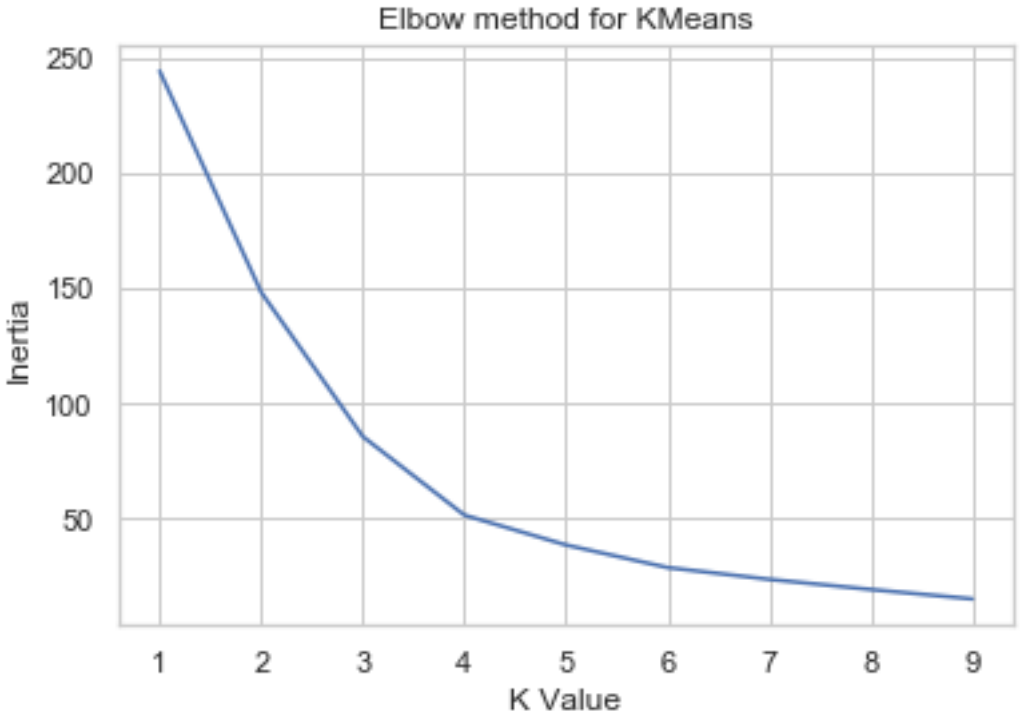

Figure 7: Elbow method for KMeans depicting the desired K value

K-Means was used to cluster data into 3 clusters as shown in Figure 8:

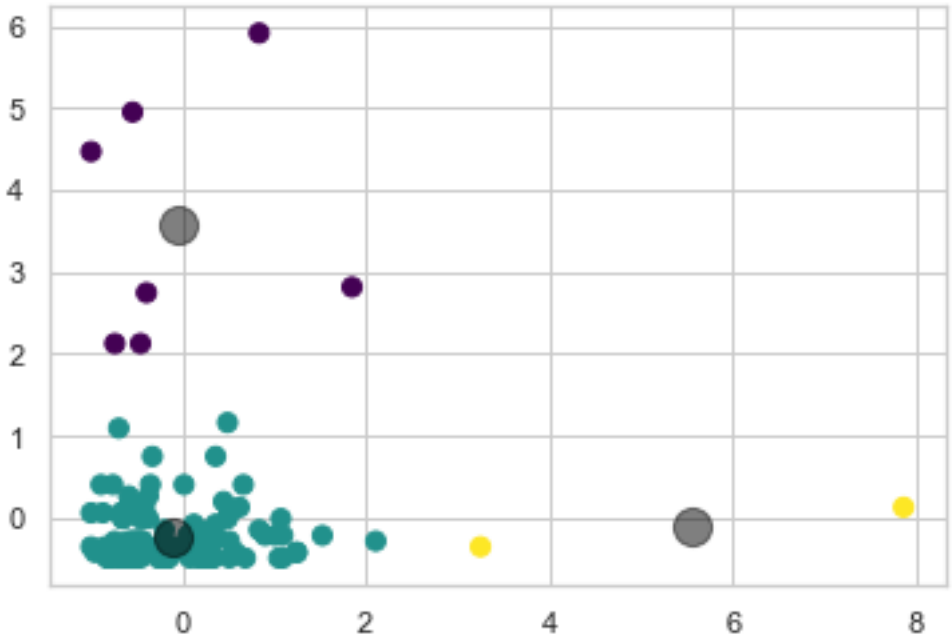

Figure 8: Three clusters depicting boroughs with high, low and medium need for resource allocation

The points represent various neighbourhoods of NYC with respective number of hospitals and covid cases, grey circles are the centroids of the clusters. Since the values were scaled, the axes do not represent the actual values. The model has divided the data into 3 clusters-

Yellow cluster represents neighbourhoods with high number of cases and low number of hospitals.

Cyan cluster represents the neighbourhoods where numbers of hospitals are acceptable with the number of cases.

Purple cluster represents the neighbourhoods with low number of cases and high number of hospitals.

4. RESULTS

The following point plot represents the outliers in the dataset, marked by green points. These are the neighbourhoods which have lower number of hospitals and higher number of cases.

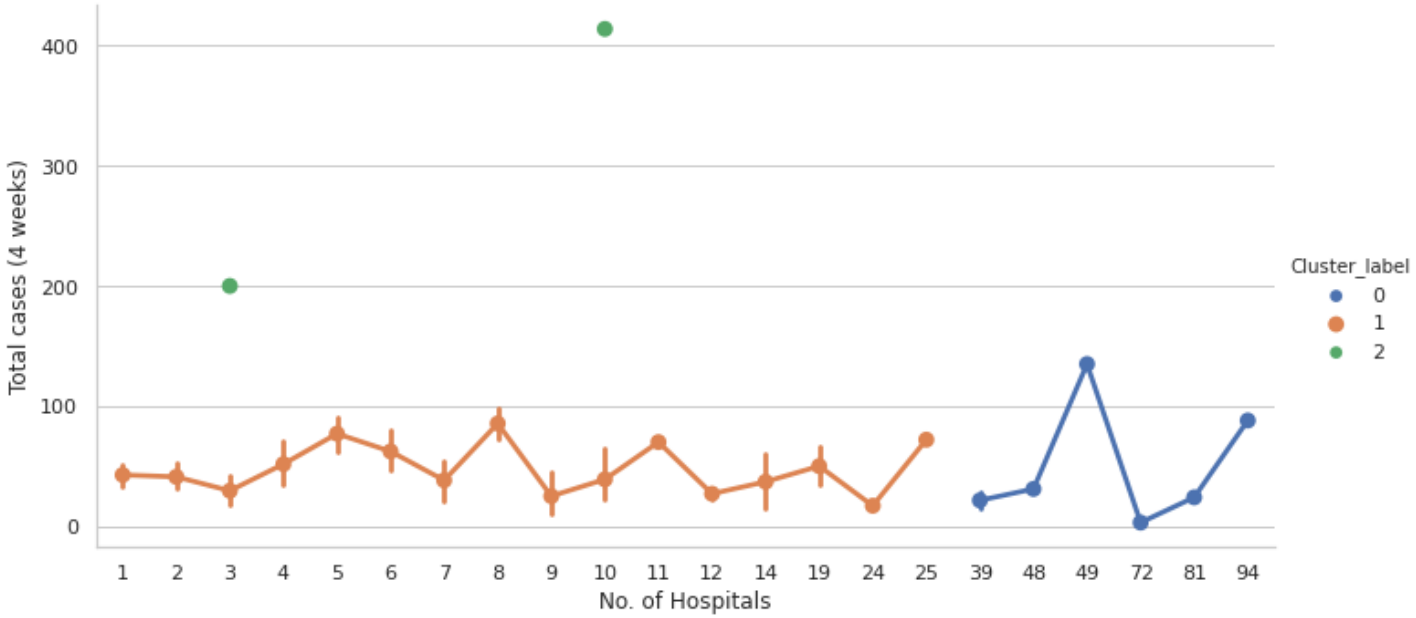

Figure 9: Dataset Outliers

The results of the clustering are visualized in the map below as shown in Figure 10 with cluster 2 in red colour, cluster 1 in purple colour, and cluster 0 in mint green colour.

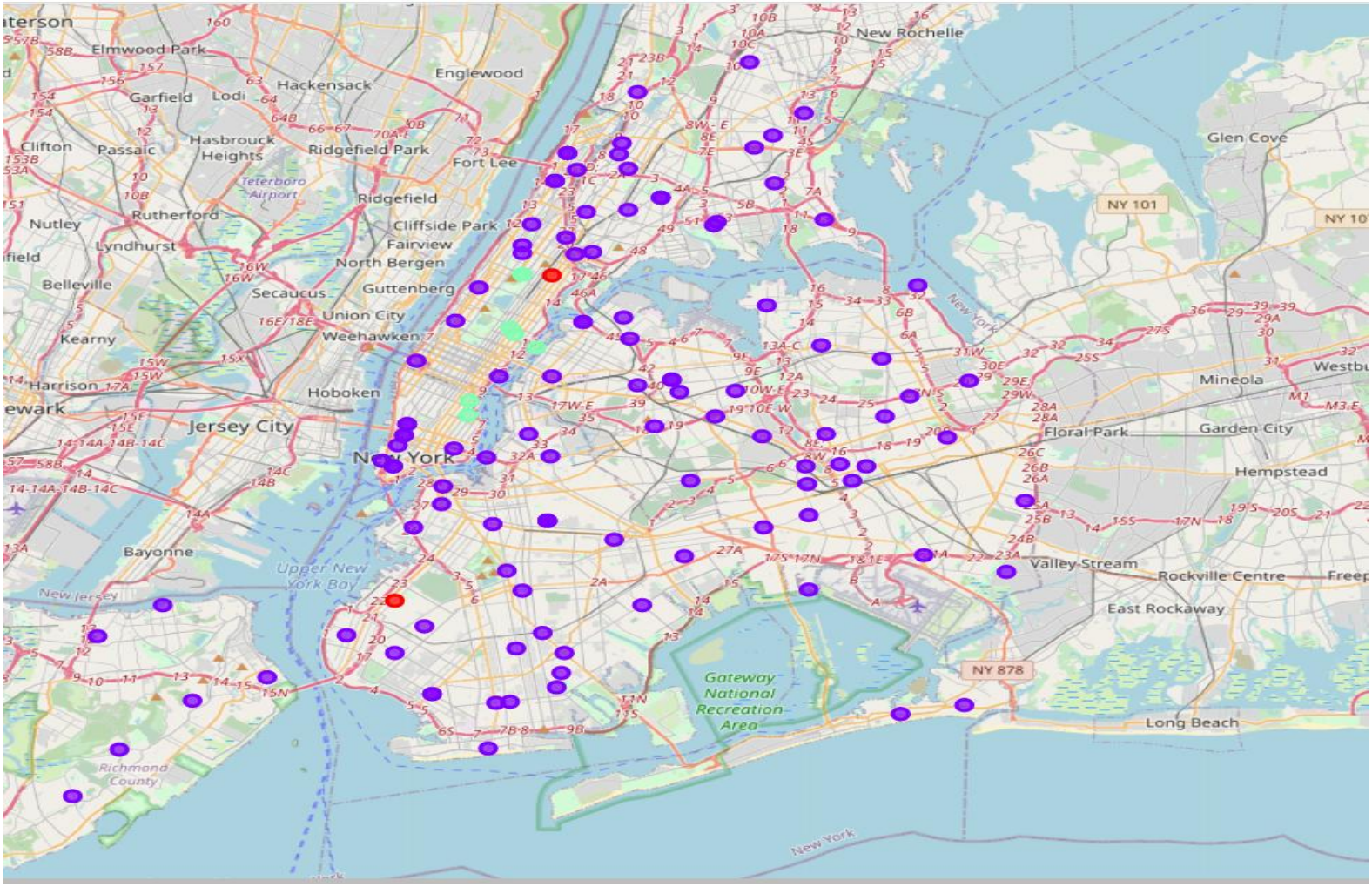

Figure 10: Map visualization of Clusters

5. CONCLUSION

Cluster 2 dataset was merged with hospital dataset to obtain names of hospitals which require donation of resources or funding. The names of the hospitals are -

1. *East Harlem, Manhattan* - North General Hospital, VA Harlem Veterans Centre, New York City Health and Hospitals Corporation.

2. *Sunset Park, Brooklyn* - CityMD Sunset Park Urgent Care, ModernMD Urgent Care, New York Centre for Special Surgery, NYU Lutheran Medical Centre, NYU Lutheran SICU, LMC 5th FLR, Calvary Hospital, NYU Lutheran Annex, NYU Langone Hospital, Maimonides Medical Centre.

	Cluster_label	Neighborhood	Borough	Total cases (4 weeks)	Latitude	Longitude	No. of Hospitals	Unnamed: 0	Neighborhood Latitude	Neighborhood Longitude	Name	Category
0	2	East Harlem	Manhattan	200	40.79828	-73.94081	3	414	40.79828	-73.94081	North General Hospital	Hospital
1	2	East Harlem	Manhattan	200	40.79828	-73.94081	3	415	40.79828	-73.94081	VA Harlem Veterans Center	Hospital
2	2	East Harlem	Manhattan	200	40.79828	-73.94081	3	416	40.79828	-73.94081	New York City Health and Hospitals Corporation...	Hospital
3	2	Sunset Park	Brooklyn	414	40.64558	-74.00982	10	994	40.64558	-74.00982	CityMD Sunset Park Urgent Care - Brooklyn	Hospital
4	2	Sunset Park	Brooklyn	414	40.64558	-74.00982	10	995	40.64558	-74.00982	ModernMD Urgent Care	Hospital
5	2	Sunset Park	Brooklyn	414	40.64558	-74.00982	10	996	40.64558	-74.00982	New York Center for Special Surgery	Hospital
6	2	Sunset Park	Brooklyn	414	40.64558	-74.00982	10	997	40.64558	-74.00982	NYU Lutheran Medical Center	Hospital
7	2	Sunset Park	Brooklyn	414	40.64558	-74.00982	10	998	40.64558	-74.00982	NYU Lutheran SICU	Hospital
8	2	Sunset Park	Brooklyn	414	40.64558	-74.00982	10	999	40.64558	-74.00982	LMC 5th FLR	Hospital
9	2	Sunset Park	Brooklyn	414	40.64558	-74.00982	10	1000	40.64558	-74.00982	Calvary Hospital	Hospital
10	2	Sunset Park	Brooklyn	414	40.64558	-74.00982	10	1001	40.64558	-74.00982	NYU Lutheran Annex	Hospital
11	2	Sunset Park	Brooklyn	414	40.64558	-74.00982	10	1002	40.64558	-74.00982	NYU Langone Hospital Brooklyn	Hospital
12	2	Sunset Park	Brooklyn	414	40.64558	-74.00982	10	1003	40.64558	-74.00982	Maimonides Medical Center / Pre Admission Testing	Hospital

Table 4: Hospitals in urgent need of resources and funding

6. SCOPE FOR FUTURE STUDY

- Hospital information for this project has been obtained from foursquare API, which may differ from the actual situation as there are many clinics, temporary wards or other facilities for covid patients. Thus, area wise hospitals specifically designated as covid centres can be considered for future studies.
- The project does not consider hospital beds or ICU beds in each hospital, which can help to further narrow down the search to determine which hospitals are in need for funding or resources.
- Not all covid positive patients require hospital facilities. Population density in a neighbourhood, age of patients, etc. are factors that influence requirement of resources for a particular hospital. Such factors can be taken into consideration for future study in the domain area.
- Number of clinical and non-clinical staff available in each hospital can also be taken into account.