Tuesday, 18 August 2020

# Analysing COVID-19 Deaths in New York City

James Curtiss

## 1. Introduction

### 1.1 Background

COVID-19 overwhelmed the state of New York, and on Thursday 12th March 2020, a state of emergency would be declared to better manage the pandemic. In recent months, the global lockdown has halted the spread of COVID-19, resulting in a reduced number of new cases and number of deaths related to the virus; the positive test rate in New York City has been reducing and now fluctuates between 1-2%[i]. As the U.S. Government enters the period of reopening New York City, research needs to be undertaken to identify the areas most at risk of further waves of the virus.

### 1.2 Problem

This study will identify at risk zip code areas in New York City which have a high proportion of venues where social distancing is difficult (i.e. pubs, nightclubs, beauty salons), and therefore could result in transmission of COVID-19.

### 1.3 Interest

The audience of this analysis will be the U.S. Government and the policy makers of New York City: by clustering the zip codes according to similar types of venues and population density, it will be possible to spot areas more at risk of further infections. As a result, the state decision makers will be more informed about the venues in each area that will have to close if infections begin to rise once more.

## 2. Data

### 2.1 Data

FourSquare location data will be used to attain the geographical coordinates of the zip codes in New York City and the most common venues in each. Furthermore, the number of deaths in each zip code will be displayed, helping to show what areas have been affected the most severely and the relation between the venues and the number of deaths. Finally, another factor important in the spread of COVID-19 is population density; for each zip code, population density will be included with the venues types in the clustering to avoid densely populated areas skewing the results.

### 2.2 Data Sources

- FourSquare API: venues of each zip code[ii]
- The New York Times: number of deaths in each zip code[iii]
- USA.com: population density of each zip code in New York[iv]

### 2.3 Features

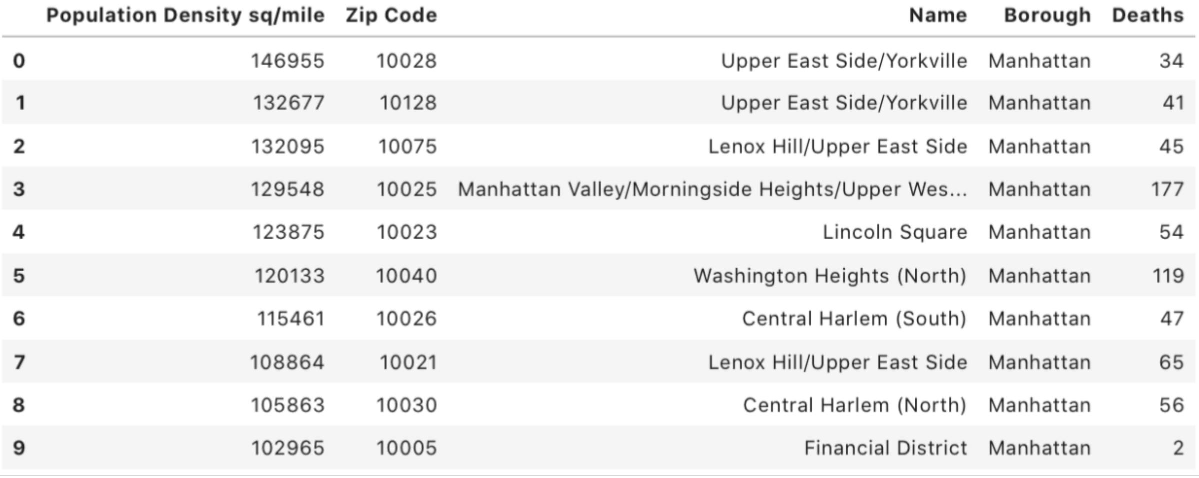
- Population density (/square mile) – this selection may need some cleaning.
- Number of deaths per zip code.
- Latitude and longitude of each zip code
- Venues of each zip code

# 3. Methodology

## 3.1 Data Cleaning

There were two main sources of data used in this study, and the resulting data frames required cleaning to ensure the values and format of the tables were suitable for further analysis. The final table of information scraped from the two websites shows the population density (per square mile) and the number of deaths related to COVID-19 in each zip code. The borough and the name of the predominant neighbourhood are also shown:

*Table 1: Final table of data collected from two website sources*

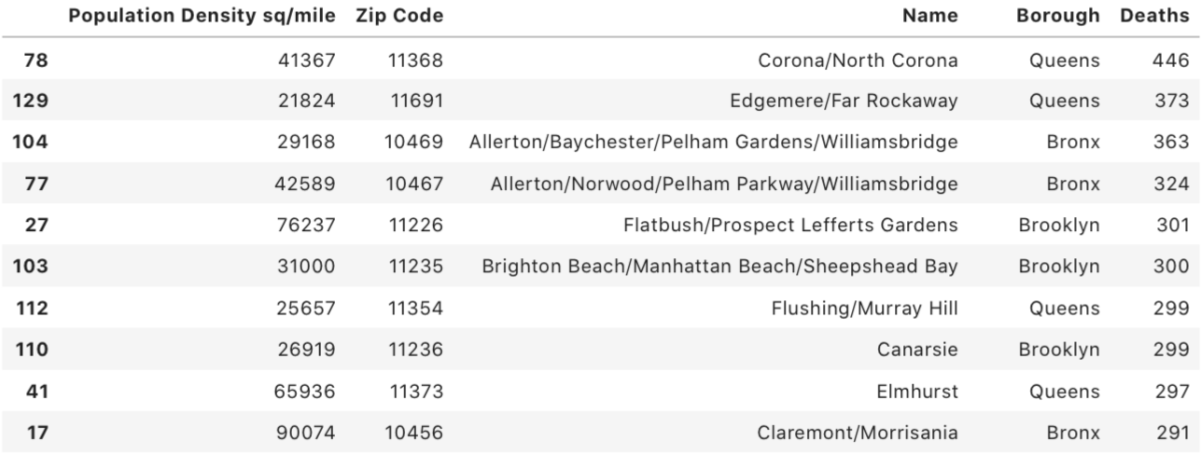


| | Population Density sq/mile | Zip Code | Name | Borough | Deaths |
|---|---|---|---|---|---|
| **0** | 146955 | 10028 | Upper East Side/Yorkville | Manhattan | 34 |
| **1** | 132677 | 10128 | Upper East Side/Yorkville | Manhattan | 41 |
| **2** | 132095 | 10075 | Lenox Hill/Upper East Side | Manhattan | 45 |
| **3** | 129548 | 10025 | Manhattan Valley/Morningside Heights/Upper Wes... | Manhattan | 177 |
| **4** | 123875 | 10023 | Lincoln Square | Manhattan | 54 |
| **5** | 120133 | 10040 | Washington Heights (North) | Manhattan | 119 |
| **6** | 115461 | 10026 | Central Harlem (South) | Manhattan | 47 |
| **7** | 108864 | 10021 | Lenox Hill/Upper East Side | Manhattan | 65 |
| **8** | 105863 | 10030 | Central Harlem (North) | Manhattan | 56 |
| **9** | 102965 | 10005 | Financial District | Manhattan | 2 |

## 3.2 Exploratory Analysis

Exploratory analysis is required to understand the data in further detail before the detailed analysis is undertaken. The data was ordered by highest number of deaths:

*Table 2: Zip codes ordered by highest number of deaths*

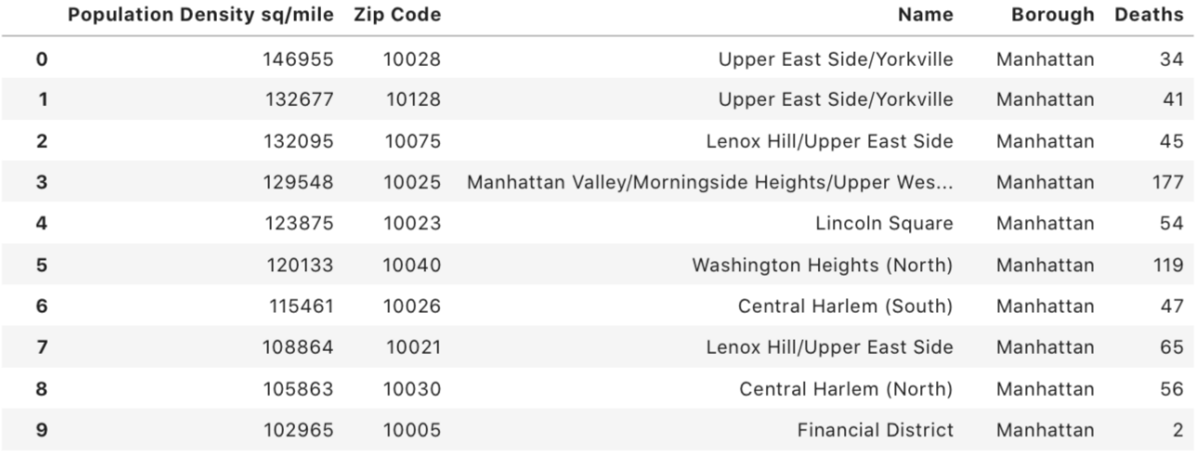


| | Population Density sq/mile | Zip Code | Name | Borough | Deaths |
|---|---|---|---|---|---|
| **78** | 41367 | 11368 | Corona/North Corona | Queens | 446 |
| **129** | 21824 | 11691 | Edgemere/Far Rockaway | Queens | 373 |
| **104** | 29168 | 10469 | Allerton/Baychester/Pelham Gardens/Williamsbridge | Bronx | 363 |
| **77** | 42589 | 10467 | Allerton/Norwood/Pelham Parkway/Williamsbridge | Bronx | 324 |
| **27** | 76237 | 11226 | Flatbush/Prospect Lefferts Gardens | Brooklyn | 301 |
| **103** | 31000 | 11235 | Brighton Beach/Manhattan Beach/Sheepshead Bay | Brooklyn | 300 |
| **112** | 25657 | 11354 | Flushing/Murray Hill | Queens | 299 |
| **110** | 26919 | 11236 | Canarsie | Brooklyn | 299 |
| **41** | 65936 | 11373 | Elmhurst | Queens | 297 |
| **17** | 90074 | 10456 | Claremont/Morrisania | Bronx | 291 |

The zip codes relating to Corona/North Corona show a much larger number of deaths at 446 – 73 higher than the nearest of Edgemere/Far Rockaway at 373. These top ten zip codes relating to number of deaths will be looked for once the detailed analysis has taken place to see if there is any relation between them.

Next, the table was ordered by population density – this is an important factor in the analysis as crowded areas and overpopulated neighbourhoods would be more susceptible to the transmission of COVID-19.

*Table 3: Zip codes ordered by population density (square mile)*



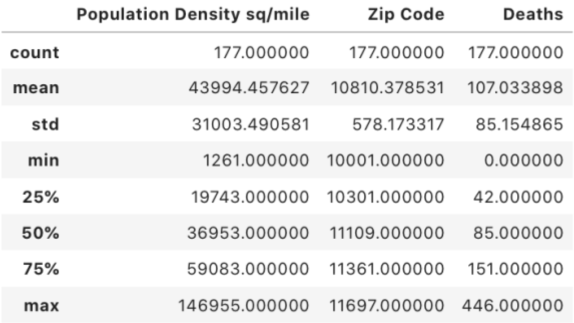
| | Population Density sq/mile | Zip Code | Name | Borough | Deaths |
|---|---|---|---|---|---|
| **0** | 146955 | 10028 | Upper East Side/Yorkville | Manhattan | 34 |
| **1** | 132677 | 10128 | Upper East Side/Yorkville | Manhattan | 41 |
| **2** | 132095 | 10075 | Lenox Hill/Upper East Side | Manhattan | 45 |
| **3** | 129548 | 10025 | Manhattan Valley/Morningside Heights/Upper Wes... | Manhattan | 177 |
| **4** | 123875 | 10023 | Lincoln Square | Manhattan | 54 |
| **5** | 120133 | 10040 | Washington Heights (North) | Manhattan | 119 |
| **6** | 115461 | 10026 | Central Harlem (South) | Manhattan | 47 |
| **7** | 108864 | 10021 | Lenox Hill/Upper East Side | Manhattan | 65 |
| **8** | 105863 | 10030 | Central Harlem (North) | Manhattan | 56 |
| **9** | 102965 | 10005 | Financial District | Manhattan | 2 |

As expected, the borough of Manhattan dominates the top ten zip codes due to it being one of the most populous areas of New York City. This will still be of interest later on as the relationship between population density, deaths and venues is analysed.

### 3.3 Inferential Statistical Testing

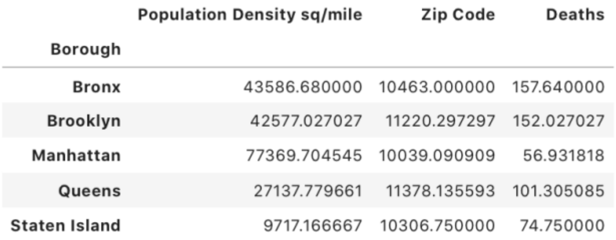
The audience of this analysis will be the U.S. Government and the policy makers of New York City: by clustering the zip codes according to similar types of venues and population density, it will be possible to spot areas more at risk of further infections. As a result, the state decision makers will be more informed about the venues in each area that will have to close if infections begin to rise once more.

*Table 4: Using the describe function on the columns*



| | Population Density sq/mile | Zip Code | Deaths |
|---|---|---|---|
| **count** | 177.000000 | 177.000000 | 177.000000 |
| **mean** | 43994.457627 | 10810.378531 | 107.033898 |
| **std** | 31003.490581 | 578.173317 | 85.154865 |
| **min** | 1261.000000 | 10001.000000 | 0.000000 |
| **25%** | 19743.000000 | 10301.000000 | 42.000000 |
| **50%** | 36953.000000 | 11109.000000 | 85.000000 |
| **75%** | 59083.000000 | 11361.000000 | 151.000000 |
| **max** | 146955.000000 | 11697.000000 | 446.000000 |

*Table 5: The column values when grouped by Borough*



| | Population Density sq/mile | Zip Code | Deaths |
|---|---|---|---|
| **Borough** | | | |
| **Bronx** | 43586.680000 | 10463.000000 | 157.640000 |
| **Brooklyn** | 42577.027027 | 11220.297297 | 152.027027 |
| **Manhattan** | 77369.704545 | 10039.090909 | 56.931818 |
| **Queens** | 27137.779661 | 11378.135593 | 101.305085 |
| **Staten Island** | 9717.166667 | 10306.750000 | 74.750000 |

Figure 5 utilises the *describe()* function to display the basic statistics of each column. From this, it is possible to understand the data further: 75% of the zip codes have a number of deaths at or below 151 and the mean of the number of deaths is 107, which shows that some zip codes have suffered far worse than others. Figure 5 shows the average column values for each Borough and it attained using the *groupby()* function. As it shows, Brooklyn and the Bronx have had a much higher number of deaths, 152 and 157 respectively, than the other three Boroughs. Interestingly, given the high population density of Manhattan, this Borough seems to have fared much better than its counterparts, registering an average deaths per zip code of 56 deaths.

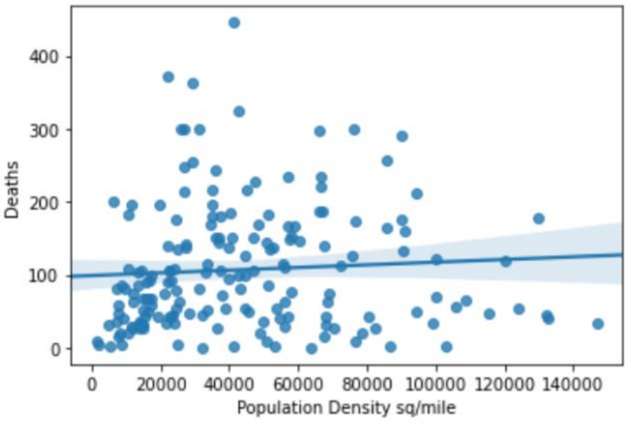
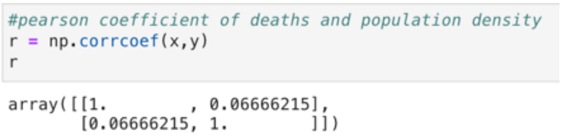


*Figure 1: Regression plot between number of deaths and population density for each zip code*

The correlation between deaths and population density was plotted on a regression plot to identify any relation between the two. As the plot shows, the relation between the two is very minimal if none. This also correlates to the Pearson Coefficient below of 0.066.



```
#pearson coefficient of deaths and population density
r = np.corrcoef(x,y)
r
```

```
array([[1.        , 0.06666215],
       [0.06666215, 1.        ]])
```

### 3.4 Geographical Exploratory Visualisation

To display the number of deaths in each zip codes, a choropleth map was chosen as it would be the easiest to visualise. The zip code polygon coordinates were obtained and added to the table.

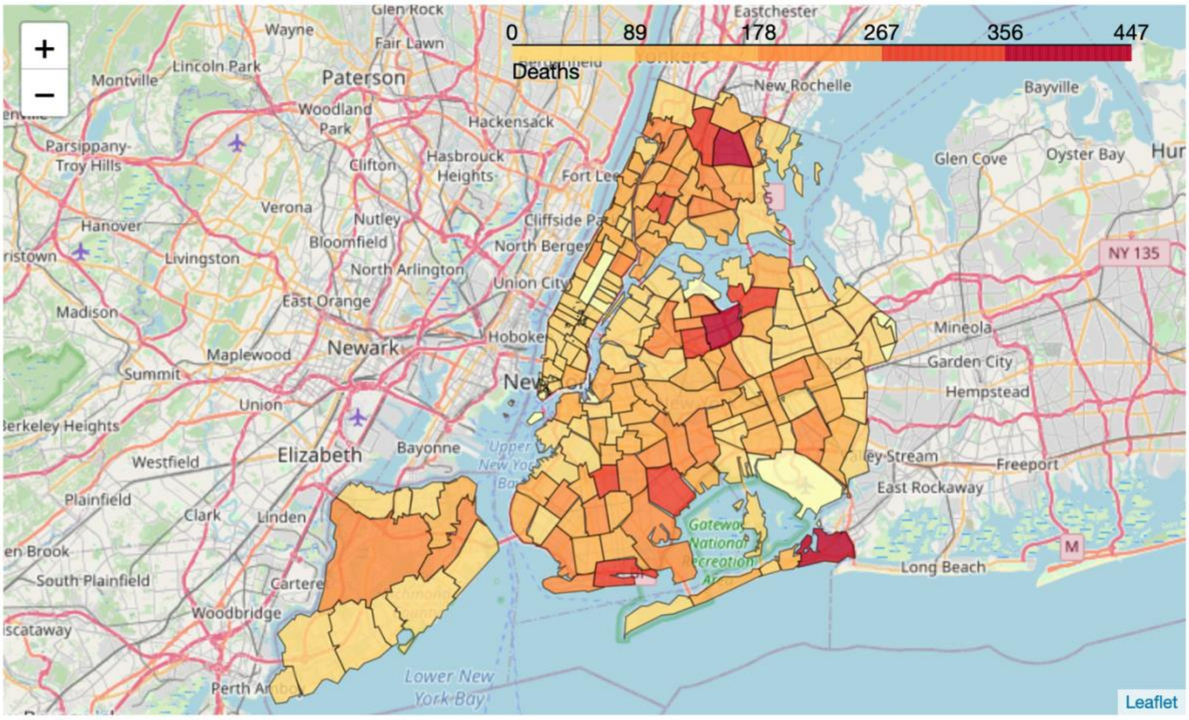


*Figure 2: Choropleth map showing the number of deaths per zip code*

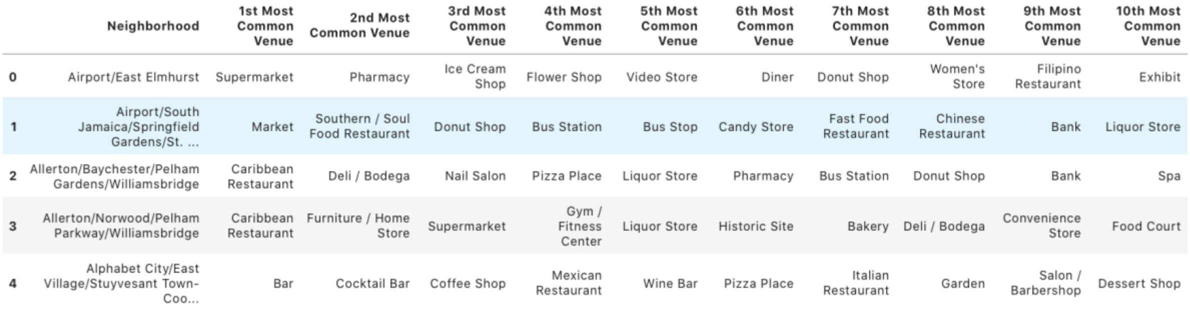
The above map displays the areas most affected by COVID-19, with there being several areas of dark red (high number of deaths) surrounded by dark orange areas. The map supports the idea of clusters of COVID-19 transmitted, with one badly affected area surrounded by areas also affected by COVID-19. Manhattan's low number of deaths is shown on the map in light yellow, whilst Statin Island is also less affected than the other Boroughs. The low number of deaths in Manhattan is surprising given the high population density. However, this could be due to the relative wealth, working age and class of the inhabitants.

Another map was created with markers added to the map showing the name of the Neighbourhood and the number of deaths – however this will be combined with the final map.

### 3.5 K-Means Clustering

After FourSquare was used to obtain the venue information for each respective zip code, it was possible to list the relative frequency of each type of venue using one-hot encoding. One hot encoding allows categorical variables to be listed as binary values for each venue category, and from here, the top ten most common venues can be found:

*Table 6: The top ten most common type of venues*



| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Airport/East Elmhurst | Supermarket | Pharmacy | Ice Cream Shop | Flower Shop | Video Store | Diner | Donut Shop | Women's Store | Filipino Restaurant | Exhibit |
| 1 | Airport/South Jamaica/Springfield Gardens/St. ... | Market | Southern / Soul Food Restaurant | Donut Shop | Bus Station | Bus Stop | Candy Store | Fast Food Restaurant | Chinese Restaurant | Bank | Liquor Store |
| 2 | Allerton/Baychester/Pelham Gardens/Williamsbridge | Caribbean Restaurant | Deli / Bodega | Nail Salon | Pizza Place | Liquor Store | Pharmacy | Bus Station | Donut Shop | Bank | Spa |
| 3 | Allerton/Norwood/Pelham Parkway/Williamsbridge | Caribbean Restaurant | Furniture / Home Store | Supermarket | Gym / Fitness Center | Liquor Store | Historic Site | Bakery | Deli / Bodega | Convenience Store | Food Court |
| 4 | Alphabet City/East Village/Stuyvesant Town-Coo... | Bar | Cocktail Bar | Coffee Shop | Mexican Restaurant | Wine Bar | Pizza Place | Italian Restaurant | Garden | Salon / Barbershop | Dessert Shop |

Next, we will use the k-means clustering algorithm to cluster the neighbourhoods into similar types. This will be based off the most common venues and the population density (this is a key factor in determining COVID-19 transmission likelihood. The venues with high numbers of bars, nightclubs, and gyms (places where the virus is likely to spread) will be clustered together whilst maintaining similar population density. These are of course not the only areas of high transmission.

K-means is used as a way to classify the neighbourhoods together which have not been explicitly labelled as similar already. It helps cluster similar neighbourhoods, with similar population density and a similar breakdown of common venues.

## 4. Results

### 4.1 Initial Observations

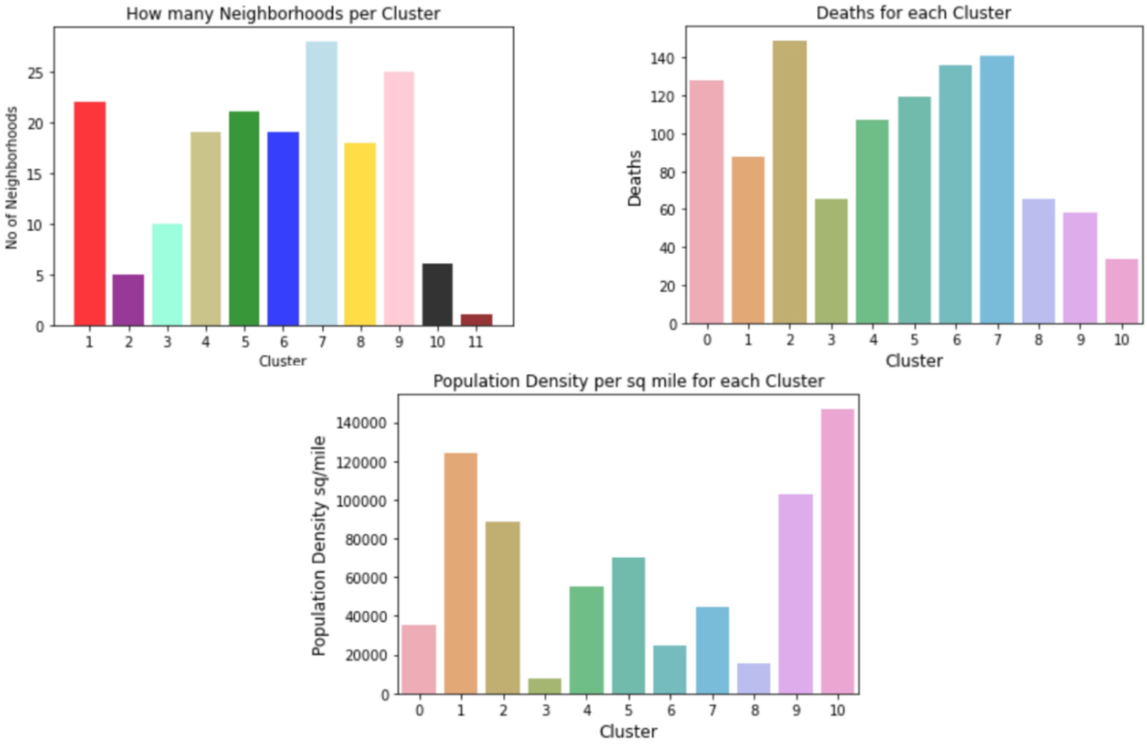


*Figure 3,4,5: Averages for each neighbourhood grouped by cluster label*

These plots above show the number of neighbourhoods per cluster, the average number of deaths per neighbourhoods in each cluster, and the average population density per neighbourhood in each cluster. It can be seen that clusters 0, 2, 5, 6 and 7 have a high average number of deaths whilst clusters 1, 2, 4, 5, 9, 10 have a high population density. Clusters 2 and 5 overlap in these areas.

### 4.2 Cluster Analysis

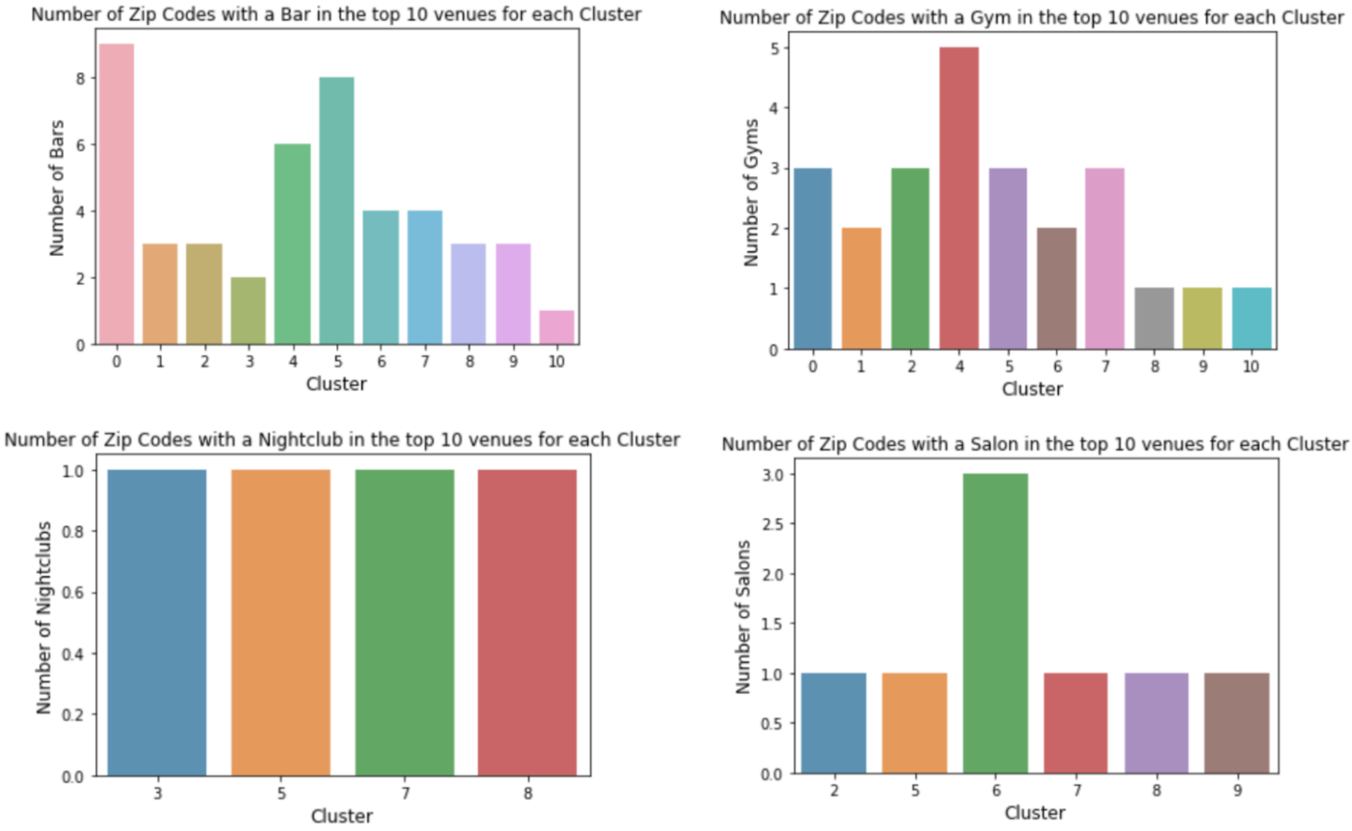


*Figure 6,7,8,9: Common venue categories grouped by cluster label*

From the above graphs, clusters 0, 2, 4, 5 and 7 show a high number of neighbourhoods with common venues being a bar or gym, while clusters 3, 5, 7 and 8 show a high number of neighbourhoods with nightclubs as a common venue. Cluster 6 also shows a much higher number of neighbourhoods with common venues being a salon.

## 5. Discussion

### 5.1 Observations

From the above graphs, clusters 4 and 5 show the neighbourhoods with the highest number of venues where COVID-19 transmission is likely. This can also be combined with the already high number of deaths for the clusters and an above average population density meaning neighbourhoods in these clusters will be susceptible to second waves of the virus. Another interesting cluster is number 7 as it appears near the top for all 4 of the types of venues and also already has a high number of deaths with an above average population density. Finally, cluster 0 shows a very high number of neighbourhoods with a bar as a top ten common venue with 9 – this indicates areas of nightlife and would be an area that should be analysed once the city comes out of lockdown.

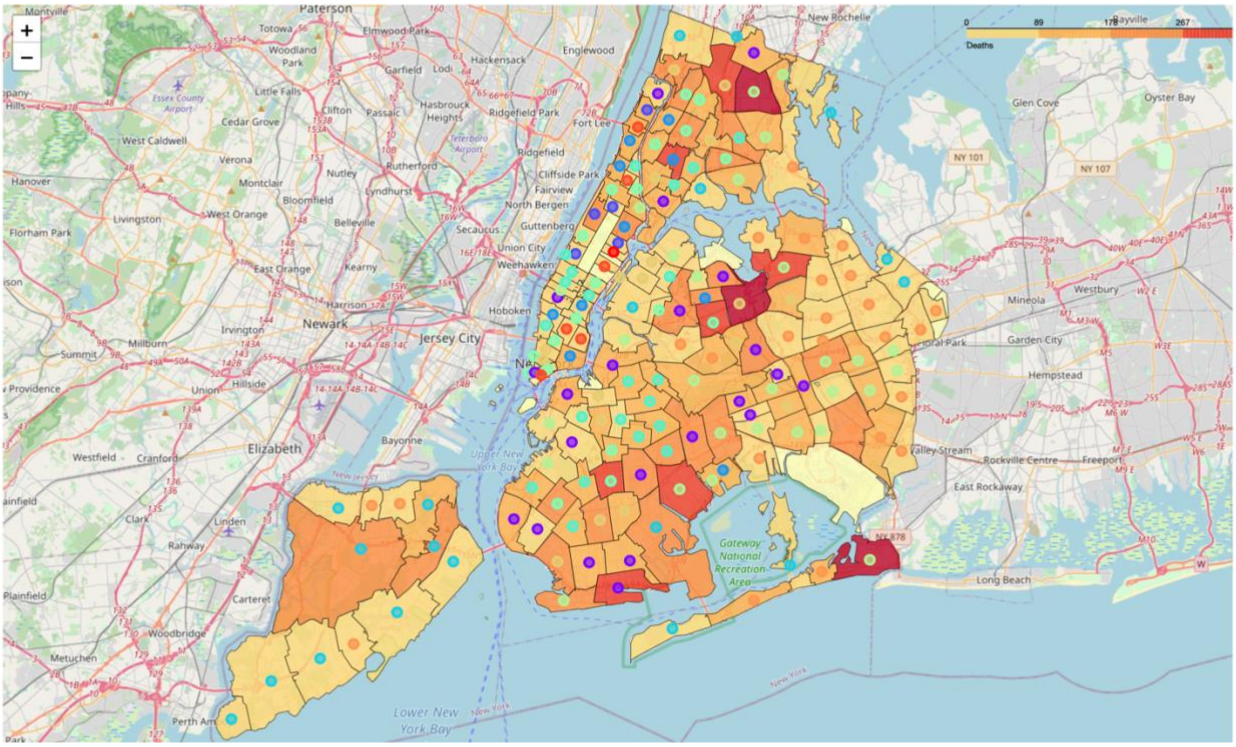


*Figure 10: Choropleth map showing the number of deaths, with the cluster markers overlapped to identify at risk areas*

In summary, clusters 0, 5 and 7 are clusters that should be cautiously re-opened as the makeup of their venues coupled with the death rate observed already make them high susceptible to further outbreaks. On the above map, the colours for clusters 0, 5 and 7 are: 🟣 🟢 🟡
The darker coloured areas indicating a higher number of deaths are often clustered in one of either 0, 5 or 7. The shows a correlation between the type of venue and the transmission of COVID-19. Obviously, there are many other factors that influence the COVID-19.

### 5.2 Recommendations

1. Clusters 0, 5 and 7 are at high risk of further waves of COVID-19 so will need to be observed for an increase in cases once New York City reopens.
2. As these clusters have a large number of venues that will need to be closed in the event in further lockdowns, it would be wise to investigate the economic impact these areas will suffer due to a large proportion of its shops closing.

## 6. Conclusion

This report has investigated the link between venue type of a neighbourhood and the transmission of COVID-19. It has shown partial evidence to support the theory that a higher number of bars, nightclubs, gyms and salons lead to a high number of COVID-19 related deaths. There are of course limitations to consider: other factors will influence such as demographics, wealth, other venue types and the primary types of work in each cluster. It would also be wise to consider further investigations into how public transport and other heavily used areas by the public affect the results. The economic impact of COVID-19 is not to be underestimated and a study into how each area will suffer would be useful when determining how and when shops will re-open. Another route to go is to map the re-opening of schools in each area in order to be able to track how COVID-19 cases are influenced by mass numbers of children in close contacts.

## References

[i] https://edition.cnn.com/2020/07/13/health/new-york-city-coronavirus-zero-deaths/index.html
[ii] https://foursquare.com
[iii] https://www.nytimes.com/interactive/2020/nyregion/new-york-city-coronavirus-cases.html
[iv] http://www.usa.com/rank/new-york-state--population-density--zip-code-rank.htm