# Analyzing COVID Case and Death Rates in New York City

Dustin Zhao

14 August 2020

## 1. Introduction

### a. Background

In early 2020, the COVID-19 pandemic swept the globe, causing numerous illnesses and deaths, as well as global economic devastation. As scientists all around the world race to develop vaccines and treatments for COVID-19, public health experts have been studying patterns of infections and outbreaks.

### b. Problem

Data of neighborhood composition in New York City can provide another perspective on infection patterns and outbreaks of COVID-19. This analysis will seek to understand the relationship between neighborhood composition and rates of infections and deaths from COVID-19.

### c. Interest

Public health researchers may be interested this analysis as it could shed new light on factors highly correlated with the spread of COVID-19, allowing for more effective policy to be crafted to contain the pandemic.

## 2. Data Acquisition and Cleaning

### a. Data Sources

Statistics for COVID-19 infection rates and death rates in New York City was provided by the New York City Department of Health and Mental Hygiene. This data is segmented by ZIP code.

Data on the venues in a neighborhood was provided by FourSquare.

### b. Data Cleaning

The data provided by the NYC Department of Health was very well-formatted and did not require any data cleaning. The data was downloaded on August 7, 2020.

The venue data from FourSquare required no cleaning, although some re-formatting was necessary to make it usable for this analysis. FourSquare provided a list of venues within a 500 meter radius of a given set of geolocational

coordinates. A table of the coordinates for each ZIP code area in New York City was used to collect these lists of venues, and then these lists were reformatted into a frequency table of every category of venue for each ZIP code.

## 3. Methodology

### a. Neighborhood Clusters: K-means clustering

The ZIP codes in New York City were clustered by venue categories using a K-means clustering algorithm. The optimal number of clusters was determined using the elbow method. A visual examination of a map of the neighborhood clusters revealed that Manhattan was its own cluster. This presented a problem because it suggested that our analysis would be unable to provide any meaningful data that was independent of geographic location or differences between boroughs in New York City.

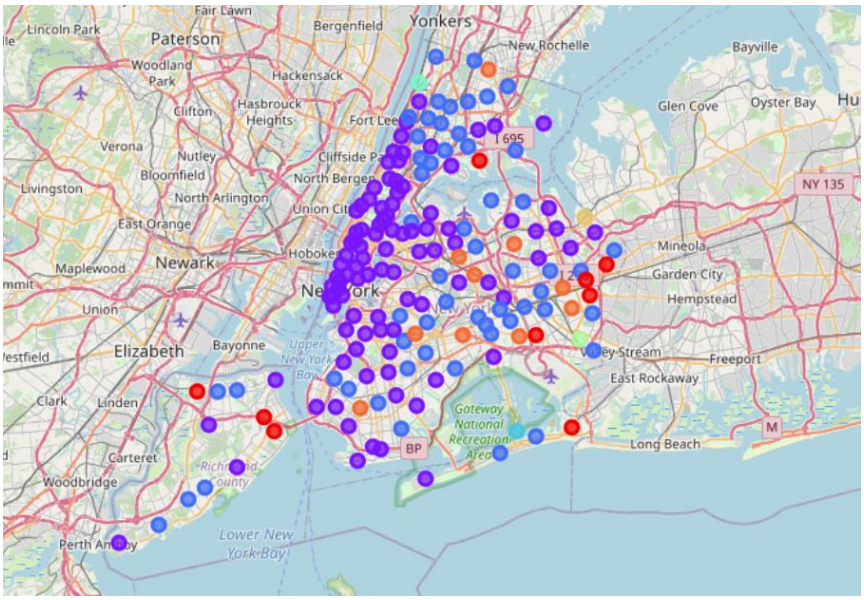


Figure 1: neighborhoods in NYC, color-coded by neighborhood cluster.

Therefore, I decided to conduct two different analyses: one on neighborhoods in Manhattan, and the other on NYC neighborhoods outside of Manhattan.

The data was segmented accordingly and again a K-means clustering algorithm was used to cluster the neighborhoods by venue types, using the elbow method to determine optimal number of clusters.

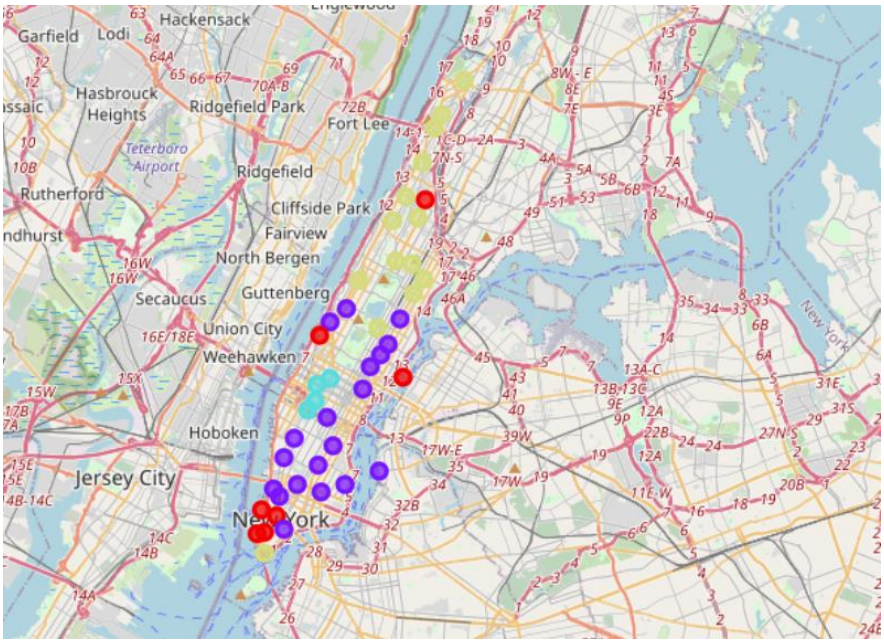

Figure 2: neighborhoods in Manhattan, color-coded by neighborhood cluster

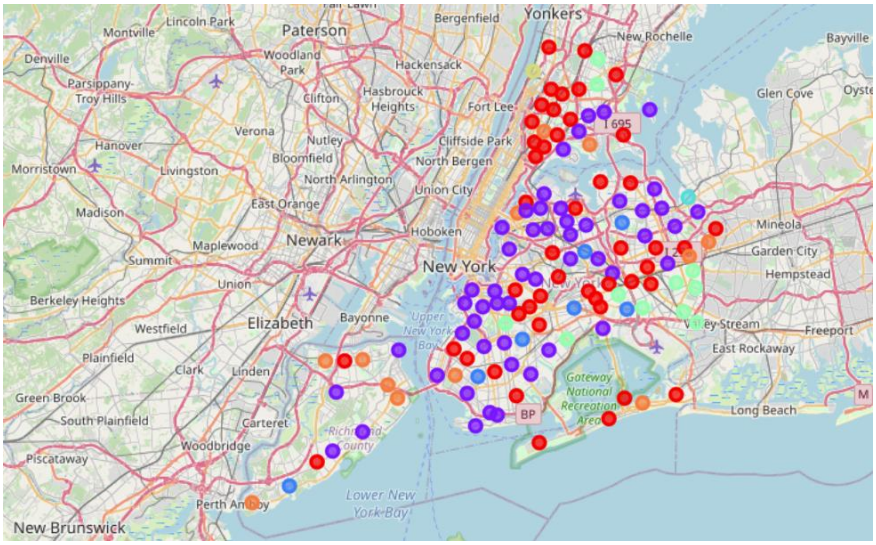

Figure 3: NYC neighborhoods outside of Manhattan, color-coded by neighborhood cluster

**b. Statistical Analysis: ANOVA and Tukey HSD**

Before running the ANOVA and Tukey HSD tests on the data, neighborhood clusters with very few neighborhoods in them were dropped from the data, as there were insufficient data points in these clusters to allow us to be confident in our analysis.

As an exploratory analysis, boxplots were constructed of the infection and death rates of COVID-19 in each cluster, as well as relevant data tables.

| | Case Rate | Death Rate |
|---|---|---|
| **Cluster Labels** | | |
| **0** | 1224.060000 | 107.926667 |
| **1** | 1278.781579 | 100.726842 |
| **3** | 2154.347857 | 205.475000 |

Table 1: average COVID-19 case and death rates by neighborhood cluster in Manhattan

| | Case Rate | Death Rate |
|---|---|---|
| **Cluster Labels** | | |
| **1** | 2428.391887 | 211.569434 |
| **2** | 2781.391250 | 228.296250 |
| **4** | 3355.629091 | 299.304545 |
| **6** | 2980.777500 | 230.997500 |
| **7** | 3077.333478 | 252.146087 |

Table 2: average COVID-19 case and death rates by neighborhood cluster outside Manhattan

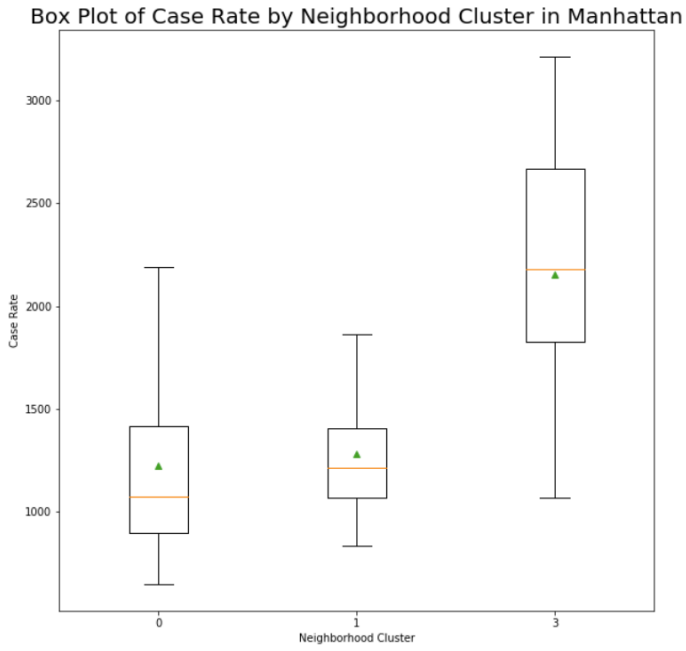


Figure 4: box plot of case rates by neighborhood cluster in Manhattan

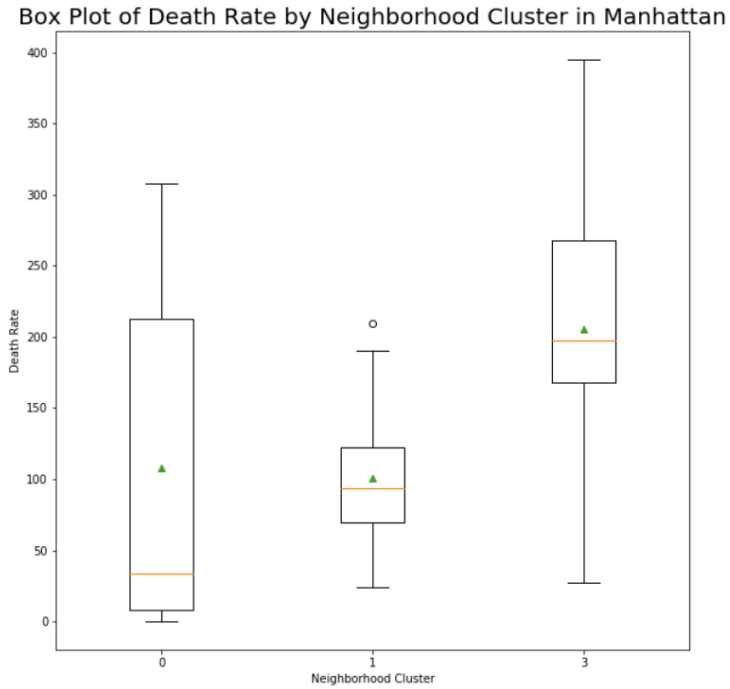


Figure 5: box plot of death rates by neighborhood cluster in Manhattan

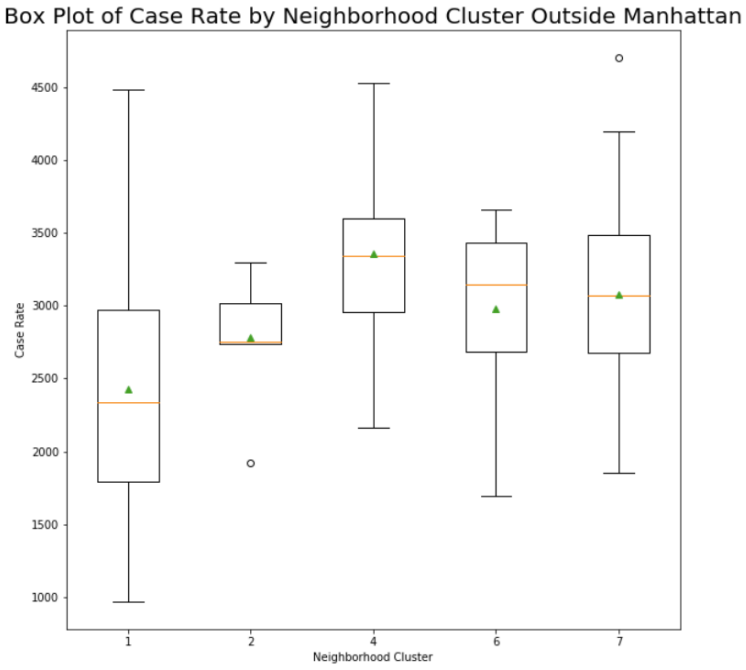


Figure 6: box plot of case rates by neighborhood cluster outside of Manhattan

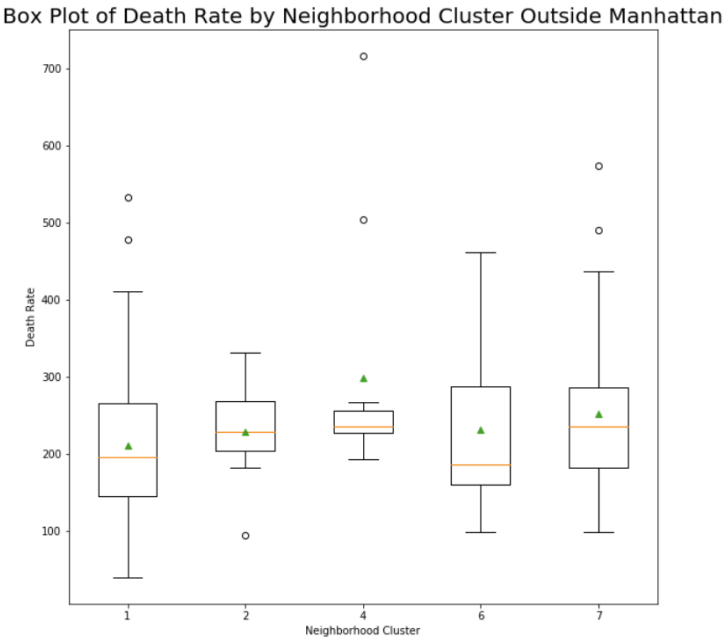


Figure 7: box plot of death rates by neighborhood cluster outside of Manhattan

An ANOVA test was conducted for the following categories: case rates in Manhattan, death rates in Manhattan, case rates outside of Manhattan, and death rates outside of Manhattan. The Shapiro Wilk test was used to check that residuals were normally distributed, and the Levene test was used to check that variances among neighborhood clusters were homogeneous. In the case of case rates in Manhattan, the Levene test found that variance among neighborhood clusters was not homogeneous, and therefore we cannot be confident in the results of the ANOVA test.

In addition to the ANOVA tests, a Tukey honestly significant difference test was run on each category to determine which neighborhood clusters were significantly different from each other.

## 4. Conclusions

### a. Case rates in Manhattan

The Levene test found that variance among neighborhood clusters was not homogeneous, therefore we cannot be confident that the conclusions of the ANOVA are not spurious.

The ANOVA found a significant difference in case rates among neighborhood clusters, and the Tukey HSD found specifically that neighborhood cluster 3 had a significantly higher case rate than other neighborhood clusters. We can conclude this with $\alpha = 0.05$.

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| C(manhattan_abridged['Cluster Labels']) | 7.101484e+06 | 2.0 | 15.035075 | 0.000018 |
| Residual | 8.501901e+06 | 36.0 | NaN | NaN |

Table 3: Results of ANOVA on case rates by neighborhood cluster in Manhattan

| group1 | group2 | meandiff | lower | upper | reject |
|---|---|---|---|---|---|
| 0 | 1 | 54.7216 | -501.4684 | 610.9116 | False |
| 0 | 3 | 930.2879 | 350.7511 | 1509.8246 | True |
| 1 | 3 | 875.5663 | 457.2336 | 1293.8989 | True |

Table 4: Results of Tukey HSD on case rates by neighborhood cluster in Manhattan

**b. Death rates in Manhattan**

The ANOVA found a significant difference in death rates among neighborhood clusters. The Tukey HSD revealed that only neighborhood clusters 1 and 3 differed significantly in death rates. We can conclude this with $\alpha = 0.05$.

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| C(manhattan_abridged['Cluster Labels']) | 95482.701004 | 2.0 | 6.593173 | 0.003633 |
| Residual | 260677.004094 | 36.0 | NaN | NaN |

Table 5: Results of ANOVA on death rates by neighborhood cluster in Manhattan

| group1 | group2 | meandiff | lower | upper | reject |
|---|---|---|---|---|---|
| 0 | 1 | -7.1998 | -104.5903 | 90.1907 | False |
| 0 | 3 | 97.5483 | -3.9302 | 199.0269 | False |
| 1 | 3 | 104.7482 | 31.4969 | 177.9994 | True |

Table 6: Results of Tukey HSD on death rates by neighborhood cluster in Manhattan

**c. Case rates outside of Manhattan**

The ANOVA found a significant difference in case rates among neighborhood clusters. The Tukey HSD revealed that the only meaningful differences were:

1. between clusters 1 and 4
2. between clusters 1 and 7

We can conclude this with $\alpha = 0.05$.

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **C(notManhattan_abridged['Cluster Labels'])** | 1.463687e+07 | 4.0 | 6.91829 | 0.000046 |
| **Residual** | 6.611490e+07 | 125.0 | NaN | NaN |

Table 7: Results of ANOVA on case rates by neighborhood cluster outside of Manhattan

| group1 | group2 | meandiff | lower | upper | reject |
|---|---|---|---|---|---|
| 1 | 2 | 352.9994 | -410.5813 | 1116.5801 | False |
| 1 | 4 | 927.2372 | 260.2329 | 1594.2416 | True |
| 1 | 6 | 552.3856 | -91.1922 | 1195.9634 | False |
| 1 | 7 | 648.9416 | 243.2708 | 1054.6123 | True |
| 2 | 4 | 574.2378 | -361.186 | 1509.6617 | False |
| 2 | 6 | 199.3863 | -719.4801 | 1118.2526 | False |
| 2 | 7 | 295.9422 | -475.2203 | 1067.1048 | False |
| 4 | 6 | -374.8516 | -1215.1817 | 465.4785 | False |
| 4 | 7 | -278.2956 | -953.9664 | 397.3751 | False |
| 6 | 7 | 96.556 | -555.9995 | 749.1114 | False |

Table 8: Results of Tukey HSD on case rates by neighborhood cluster outside of Manhattan

### d. Death rates outside of Manhattan

The ANOVA found no significant difference in death rates among neighborhood clusters, and the Tukey HSD confirmed this.

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **C(notManhattan_abridged['Cluster Labels'])** | 1.463687e+07 | 4.0 | 6.91829 | 0.000046 |
| **Residual** | 6.611490e+07 | 125.0 | NaN | NaN |

Table 9: Results of ANOVA on death rates by neighborhood cluster outside of Manhattan

| group1 | group2 | meandiff | lower | upper | reject |
|---|---|---|---|---|---|
| 1 | 2 | 16.7268 | -95.1026 | 128.5562 | False |
| 1 | 4 | 87.7351 | -9.9503 | 185.4205 | False |
| 1 | 6 | 19.4281 | -74.8264 | 113.6826 | False |
| 1 | 7 | 40.5767 | -18.8354 | 99.9887 | False |
| 2 | 4 | 71.0083 | -65.9882 | 208.0048 | False |
| 2 | 6 | 2.7013 | -131.8703 | 137.2728 | False |
| 2 | 7 | 23.8498 | -89.0899 | 136.7896 | False |
| 4 | 6 | -68.307 | -191.3767 | 54.7626 | False |
| 4 | 7 | -47.1585 | -146.1131 | 51.7962 | False |
| 6 | 7 | 21.1486 | -74.4207 | 116.7179 | False |

Table 10: Results of Tukey HSD on death rates by neighborhood cluster outside of Manhattan

**e. Caveats**

The usefulness of these conclusions is limited for several reasons. Firstly, the nature of the K-means clustering algorithm to form the neighborhood clusters is opaque by nature and, even having concluded some differences in case and death rates, the causes of these differences are difficult to elucidate. For example, a visual examination of the neighborhood cluster map of Manhattan shows a geographical correlation among neighborhood clusters, so perhaps any differences in case and death rates is a result of geography. Or, perhaps certain neighborhoods are home to certain venues due to the average income levels of residents of that neighborhood, and therefore any differences in case and death rates is the result of income correlations. There is no way to tell.

Secondly, due to the random nature of the K-means clustering algorithm, it is very likely that different clusters may result in different conclusions drawn from the ANOVA and Tukey tests.

Finally, the data set we are working with is somewhat limited. Some neighborhood clusters have very few data points, which means we cannot be confident in our conclusions.

## 5. Future Discussions

Available public health data already suggests that case and death rates correlate with socioeconomic factors and race. It is very likely that the types of venue common to a

neighborhood correlates with socioeconomic factors and race as well, and therefore our K-means clustering algorithm implicitly captured these correlations.

A more illuminating analysis of how types of venues affects infection and death rates may seek to separate out neighborhood characteristics from the frequency of venue types within a neighborhood. Perhaps an ordinary least squares regression may be more illuminating in this regard.

The fields of public health and epidemiology are complex and constantly changing, especially in the face of an unprecedented and unknown illness. Any data analysis at all can be illuminating in these fields and lead to better understanding and better policy when it comes to tackling this pandemic.