Analysis of Manhattan COVID-19 Spread

A Study on the Impact on Neighborhood Venues on the Spread of Infectious Diseases

The Problem: Infectious Diseases in Mega-Cities

COVID-19 devastated New York City in the early months of 2020

As someone who lived in Manhattan during the height of the COVID-19 pandemic, this was a problem that had a profound impact on my life. My career was impacted. My close friend fell ill to the disease and had to fight for his life.

In the end, I eventually had to relocate away from Manhattan after studying, working, and living there for over a decade all due to COVID-19.

Data Acquisition

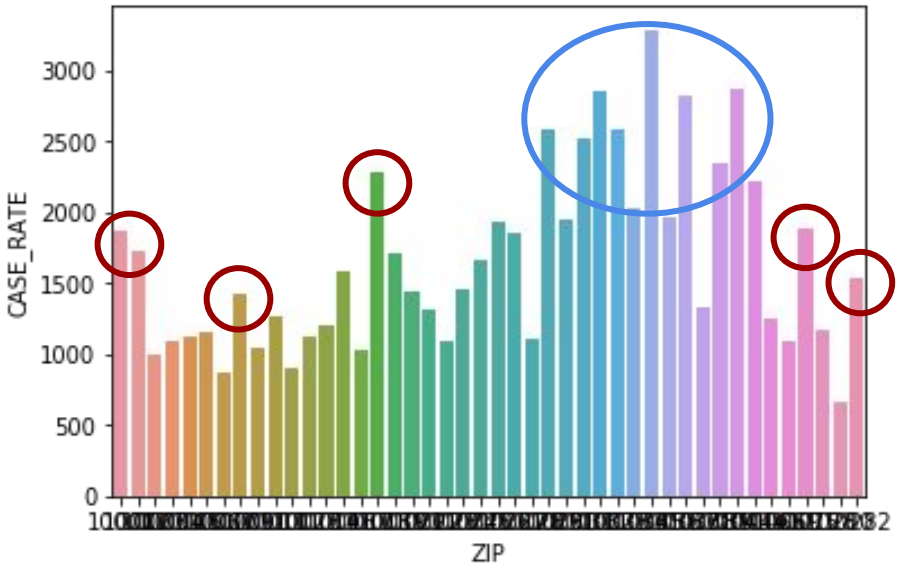
- Public Health Data for New York City (and more specifically Manhattan) was provided through NYC Health's website and their github: "https://raw.githubusercontent.com/nychealth/coronavirus-data/master/data-by-modzcta.csv"
- In addition to the data on COVID case count, case rate, population denominator, death rate, percent positive, and total tests by zip code, I needed longitude and latitude data corresponding to these zip codes so I can use the Foursquare API to look for the most common venues in these zip code. To this end, I used the US Census ZIP Code data from github: "https://gist.githubusercontent.com/erichurst/7882666/raw/5bdc46db47d9515269ab12ed6fb2850377fd869e/US%2520Zip%2520Codes%2520from%25202013%2520Government%2520Data"
- Lastly, the Foursquare API was used to find the most common venues within each of the zip code defined neighborhoods within Manhattan for our analysis.

Data Wrangling

- First, the data from NYC Health needed to be merged with the longitude and latitude data from US Census.
- Then the combined data frame was cleaned up for consistency as headers were renamed.
- Since the central problem is the analysis of the impact of venues on the spread of COVID-19, data like COVID death count and death rate were irrelevant to our analysis and thus removed from our data.
- Additionally, descriptive statistics of all the data were reviewed and correlations between the data also tested to see if any redundant data could be removed.

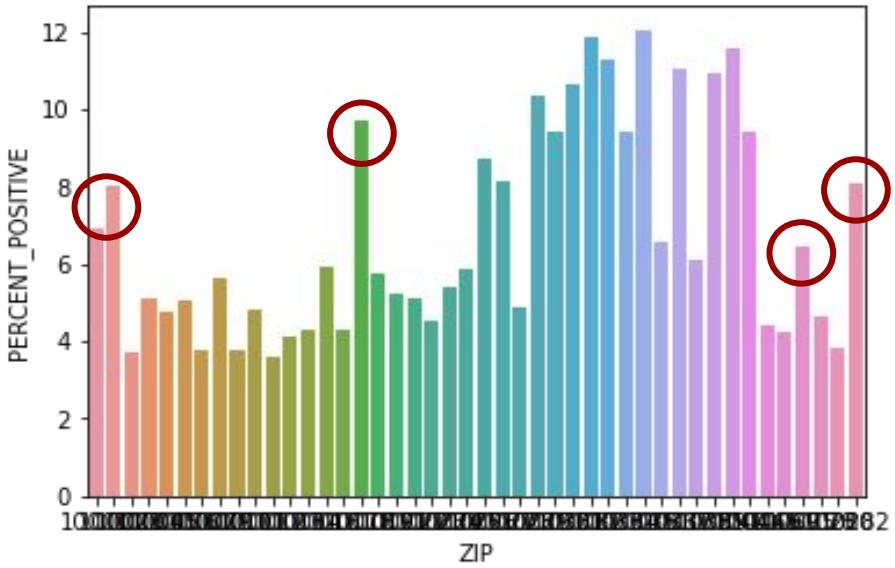
Data Visualization: COVID Case Rate by ZIP Code

Since the numerical closeness of ZIP Code also implies geographic closeness, it is unsurprising to see some clusters of high case rate form around each other, but there are some interesting individual spikes.



Data Visualization: Percent Positive by ZIP Code

Taking a different view of the data by using the percent positive of the testing done in each ZIP code, we can see a similar trend as the case rate graph but now some of those individual spikes are more clear.



Clustering the Neighborhoods

- The central element of the analysis is the clustering of the neighborhoods.
- To this end, the top 100 venues from a 500 meter radius of the zip code were collected and categorized by the venue type.
- Based on venue type frequency, top 5 most common venues within each zip code was selected.
- Then using this data, a K Means Clustering was used to determine k clusters of similar zip codes.

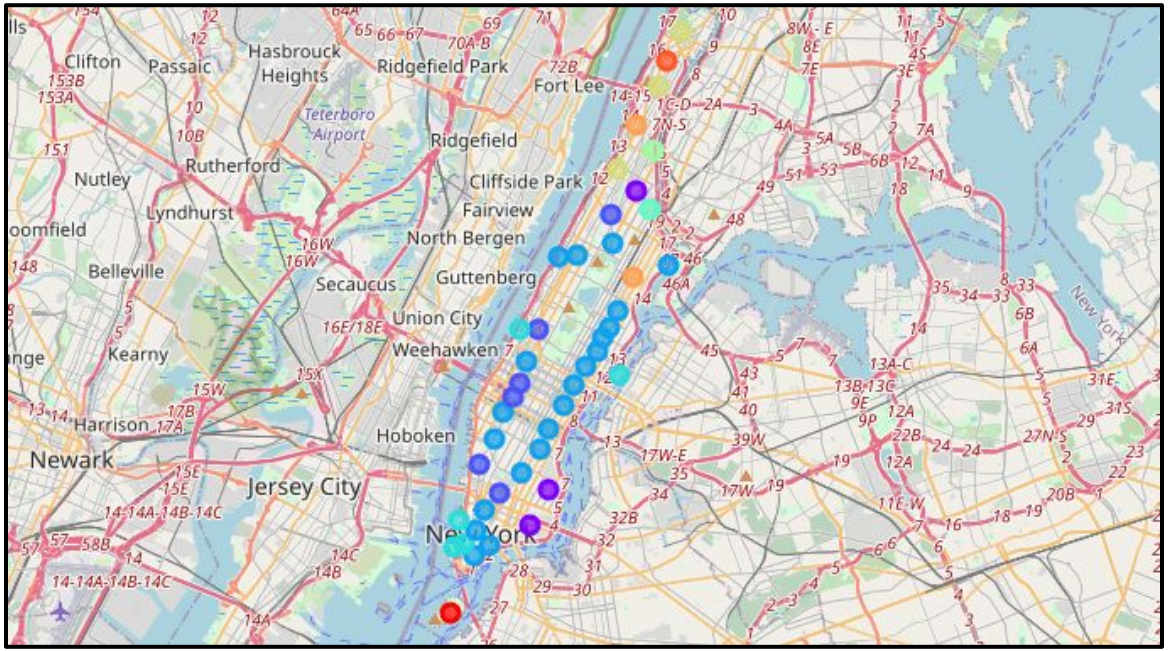
K Means Clustering Methodology

- With the top 5 most common venues picked out for each zip code, K Means Clustering applied starting with k = 5.
- The end result was a very skewed clustering with most of the zip codes grouping up into the first three clusters with only 1 zip code in each of the fourth and fifth cluster. This was no good.
- It was clear, k must be increased to create some more well defined clusters to aid our analysis, but k could not be too big as we only had 44 zip codes, so in the end after some testing, k = 10 was selected as it provides some well defined neighborhood through common venue types which would aid us immensely in our analysis.

Characteristics of the 10 Clusters

1. Governor's Island (minor tourist attraction)
2. Poor residential neighborhoods that are geographically distant (Central Harlem, Chinatown, and Alphabet City).
3. Trendy Neighborhoods with Italian restaurants, coffee shops, gyms, and stores.
4. The largest cluster that runs up and down the main avenues of both the east and west sides of Manhattan that mainly features bars, restaurants, and cafes.
5. Neighborhoods that are dominated by parks and public spaces and often residential.
6. A singular and unique neighborhood in Central/East Harlem.
7. Another singular and unique neighborhood in Central Harlem/Washington Heights
8. The Washington Heights neighborhood made up of three zip codes dominated by pizza places, restaurants, and bars.
9. Neighborhoods around Harlem with a heavy hispanic population with Mexican restaurant as the most common venue.
10. A singular and predominantly Chinese neighborhood in Washington Heights.

Visualization of the 10 Clusters in Manhattan



Results & Analysis

common venues are the reflections of people living there

Since the goal of this analysis was to spot trends in historical data and not predictive in nature. No regression analysis was used as the majority of the analysis comes from the K Means Clustering results and the high correlations between some of these clusters and high case rate and percent positive rate.

Overall, one of the most significant result is that common venues is often very telling of the socioeconomic standings of the local population and that is sometimes the most important indicator in the spread of COVID.

Interesting Observations

The highest percent positive and case rate was found in East Harlem at 12.06% and 3285.25. It ended up being buried in cluster 4, which is interesting since most of the cluster featured neighborhoods with shops and lots of restaurants along the main avenues in Manhattan, but East Harlem's most common venue is the supermarket followed by clothing store, trail, pizza place, and cafe which makes it feel out of place in this cluster.

Another slightly odd but interesting observation is the uniformity of the testing done throughout Manhattan in all the zip codes as the total test remained around 25% of the population denomination for each zip code with only some minor deviations. This made our analysis easier as it eliminated any potential issues from lack of testing over overtesting in any given zip code.

Key Takeaway

Cluster 6 through 10 all had the highest level of percent positive results and case rates in manhattan and after spending some time combing through the Foursquare data, the conclusion is simple. These neighborhoods (all situated within various regions of Harlem) are all densely populated, which are highlighted by their population data. The most common venues are often cheaper restaurants like pizza places or ethnic foods that reflects the population of the neighborhood (Hispanic, Chinese, African American). Other commonality is that bodega which functions as both a cheap supermarket and a cheap restaurant are also common throughout all these neighborhoods which were all very unique and diversified when compared to some of the bigger clusters in Manhattan that were all dominated by more upscale dining options, coffee shops, and retail stores. So in essence, in the case of an infectious disease like COVID-19, venues do not tell the whole picture but it does reveal the underlying issues such as the concentration of poorer neighborhoods and higher population density within them.