Comparing Venue Popularity in New York City Neighborhoods Affected by COVID-19

Taylor Flowers

August 10, 2020

1. **Introduction: Business Problem**

New York city has been one of the most impacted cities in the United States by COVID-19. In the last six months, there have been over 220,000 confirmed cases in the city itself. Currently, the city is reopening. With the reopening of the city, the population may have different needs or wants in terms of available businesses. Prospective business owners are curious as to which venues are most popular in each neighborhood and if there is a relation between the number of COVID cases and the type of popular venues.

1. **Data**

The data necessary to solve this problem includes New York neighborhood data, New York COVID-19 data, and Foursquare location data.

First, we will need data to determine the prevalence of the virus in each neighborhood. We need a dataset containing the number of Covid cases by the neighborhood. This data is available here: <https://github.com/nychealth/coronavirus-data/blob/master/data-by-modzcta.csv>.

Additionally, we will need to determine the latitude and longitude values in each ZCTA to create maps and use Foursquare. This information is available from the census bureau here: <https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2016&layergroup=ZIP+Code+Tabulation+Areas>

Finally, we will use the Foursquare location data to determine which nearby venues are most popular in the different neighborhoods.

1. **Methodology**

The first part of solving the problem was creating a Python dataframe that would allow me to properly solve the problem. To do this, I first loaded the New York City Coronavirus Health information into a dataframe. From this dataframe, I created a new dataframe from the following columns: Borough Group, Neighborhood Name, Modified ZCTA, Covid Case Count, Covid Case Rate, and Covid Death Count. This was the only information that was relevant to the problem, so the other information from the NYC health data was ignored.

Next, to run a Foursquare API request, I knew that I would need latitude and longitude values for each location. I created a dataframe from census information that included ZCTA, latitude, and longitude values. Next, I merged this dataframe with the dataframe that I had created from the NYC health data. As a result, I had a dataframe with all of the data necessary to cluster the neighborhoods and run Foursquare API requests.

After having an informative dataframe ready, I clustered the neighborhoods according to Covid severity. I created five clusters according to their similarity of Covid Case Count, Covid Case Rate, and Covid Death Count. These clusters were renamed based on their Covid severity. To do this, I printed the values from each cluster to view them and I decided which clusters were more severe by looking at the Covid numbers. I added each neighborhood’s cluster label back into the original NY dataframe.

Now that I had a dataframe that clusters neighborhoods and includes location and Covid data, I began to make Foursquare API requests. I established my Foursquare credentials which included client ID and client secret. Next, I created a function that would make Foursquare API requests that would return nearby venues for each location that I had needed. I used this function to create a dataframe that had nearby venues for each neighborhood in my NY dataframe. Then, I used one hot encoding to analyze each neighborhood and then establish a dataframe with the mean frequency of each venue in each neighborhood.

After having the mean frequency of each venue for every neighborhood, I needed to create a dataframe that would give the ten most common venues in each neighborhood. I created a function that would return the ten most common venues. I used this function to create a dataframe that included each neighborhood name along with its ten most common venues.

With the neighborhoods and most common venues, I clustered the neighborhoods based on similarity of common venues. I created ten clusters with arbitrary cluster numbers as labels and added the labels back into the NY dataframe.

Using the NY dataframe, I created a folium map to visualize the clusters of neighborhoods according to most common venues.

1. **Results**

The results of the map visualizations of the clusters for venue and Covid clusters were as follows:

A close up of a map

Description automatically generated

Figure 1: NYC Neighborhoods Clustered by Covid Prevalence

A close up of a map

Description automatically generated

Figure 2: NYC Neighborhoods Clustered by Venue Popularity

A picture containing drawing

Description automatically generatedA picture containing crossword, drawing

Description automatically generated

Figure 4: 2nd Most Severe Covid Neighborhood Venues Histogram

Figure 3: Most Severe Covid Neighborhood Venues Histogram

A picture containing drawing

Description automatically generatedA picture containing drawing

Description automatically generatedA picture containing drawing

Description automatically generated

Figure 7: Least Severe Covid Neighborhood Venues Histogram

Figure 6: 4th Most Severe Covid Neighborhood Venues Histogram

Figure 5: 3rd Most Severe Covid Neighborhood Venues Histogram

1. **Discussion**

The results of the map visualizations reveal the geographical factor in the Covid and venue clusters. For the most part, the neighborhoods in both the Covid map and the venues map are close in proximity to other neighborhoods in the same cluster. It makes sense that neighborhoods in which Covid is prevalent will have surrounding neighborhoods that have a similar severity. It also makes sense that adjacent neighborhoods will have a similar selection of popular venues. This information indicates that while there may be a relation between venue popularity and Covid prevalence, this relation may be due to geographical factors. Furthermore, relational data between Covid prevalence and venue popularity in New York City may not be extrapolated to neighborhoods of other cities in the nation. In other words, any suggestion to start a new company in an NYC neighborhood using Covid and venue data may not be applicable in other US neighborhoods with similar Covid prevalence.

The histograms for each Covid cluster indicate that there are certain kinds of venues that are more popular in NYC neighborhoods with similar Covid prevalence. In the 3rd and 4th most prevalent clusters of Covid, the histograms are highly similar with almost all neighborhoods being in venue cluster eight. Likewise, the most severe and the 2nd most severe Covid neighborhoods have similar histograms with the top three venue clusters being one, five, and eight. However, the histogram for the least severe cluster of neighborhoods appears more similar to the most severe histogram. We would expect that the least severe cluster of neighborhoods would have more similar venue popularity to the 4th most severe cluster than the most severe cluster. Having the opposite result indicates that the Covid prevalence does not have as strong of an effect on venue popularity than expected. It also indicates that the venue popularity of a neighborhood has more factors than Covid severity and proximity.

1. **Conclusion**

To conclude, it is possible to suggest new venues for business owners to open while Covid is prevalent in NYC. For the neighborhoods with a severe Covid presence, venues consistent with clusters one, five, and eight will get more business. As their cases of Covid lessen, we would suggest starting business consistent with cluster eight. When the threat of Covid lessens enough, we would suggest that prospective business owners would start a variety of business.