A. Introduction

As someone who has studied, worked, and lived in Manhattan for the past twelve years, moving away due to COVID-19 has been a big turning point in my life. Alongside this change in location is the decision to shift my career towards data science, so it was only natural for me to focus my capstone project for IBM Data Science Professional Certificate on the spread of COVID-19 in Manhattan.

As one of the five boroughs of New York City, Manhattan is home to roughly 1.63 million people all living in just 23 square miles. This translates to a population density of 70,826 people per square mile, or 27,346 per square kilometer, which makes Manhattan or New York County the most densely populated county in the United States. Sadly, this statistic is also one of the key reasons why Manhattan was hit hard by COVID-19 in the first half of 2020.

Beyond just residents of Manhattan, during a typical workday, commuters would push these figures to 3.9 million people or a population density of over 170,000 people per square mile. Combine these office hotspots with tourist hotspots like Time Squares and Central Park and cultural hotspots like Lincoln Center and Broadway, it is pretty clear Manhattan is particularly vulnerable to infectious diseases such as COVID-19.

What is unclear or left unexplored is how different venues within smaller neighborhoods inside Manhattan contributed or impacted the spread of COVID-19. So in this project, we will utilize the Foursquare API to first breakdown the neighborhoods of Manhattan by their venue types and then cross reference this data with COVID-19 data from NYC Department of Health and Mental Hygiene. With this initial analysis, we can then dive deeper by using supplementary data such as MTA transit data for buses and subways and daily weather data to discover underlying trends that contributed to the devastating spread of COVID-19 in metropolises such as Manhattan.

Ultimately, this project will utilize the data science techniques I learned in the course to shed light on the transmission of COVID-19 in cities like Manhattan and hopefully pass on insight to help other cities to take the right measures to curb the effects of COVID-19 on our livelihood. Lastly, I hope to continue to develop my data science skills and expand on this project beyond the scope of this project in order to help alleviate the impact of future infectious diseases both in terms of disease prevention and city planning.

B. Data Acquisition and Cleaning

B.1 Data Sources

To tackle our problem, I used the following data sources:

- Foursquare API was utilized to pull venue data for all Manhattan neighborhoods.

- NYC Health Coronavirus Data was utilized to sort COVID-19 data by ZCTA (Zip Code Tabulation Areas) Neighborhoods.
- ZIP Code Longitude and Latitude data from the US Census

Using these three data sources, we can perform many tests to solve our problem. First, we can cross reference the three data sources by using zip code to check for correlations of high COVID-19 spread to certain venues. Once this is established, we can further this analysis by using K-mean clusters to first group similar zip codes together in order to check if the COVID-19 spread is uniform across members of the same cluster or if any factors discovered in the clustering process can attribute to an increase in the impact of COVID-19.

B.2 Data Cleaning

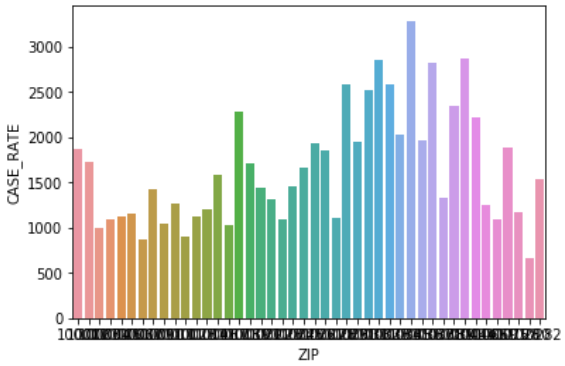
The first step of data cleaning for us is to merge the longitude and latitude data from the US Census with the NYC Health Data so we can eventually use the Foursquare API to pull venue data using the longitude and latitude that is unable on the NYC Health data.

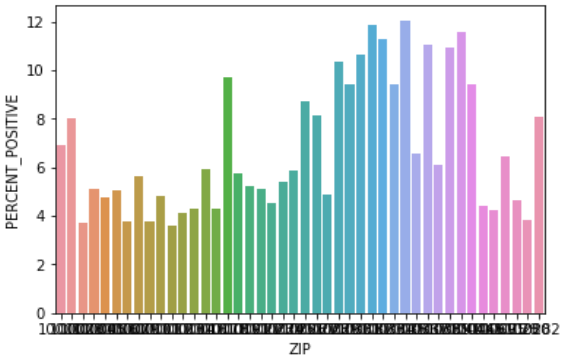
Then after creating this merged data frame, descriptive statistics and correlations between all the COVID data were reviewed to see if any irrelevant data categories can be removed.

While some correlations were expected such as those between cases and population and total tests, one thing became clear. We did not need to worry about death rate or death count as both of those figures are derivative to total cases and the medical attention and underlying health issues that were all irrelevant to this analysis. So after cleaning up some of the headers for consistency, these two data categories were removed.

C. Exploratory Data Analysis

After cleaning up the data, graphs were constructed to help better visualization and explore the data.





The two most important metrics for our analysis on the spread of COVID-19 are case

rate and percent positive which represents the both the breadth and depth of the degree of COVID-19 spread and while there are a lot of similarities between these two graphs, there are some heighten focus on a few peaks that are highlighted better by the percentage positive graph which really reduce a lot of noise that are introduced by differences in population between some of the zip codes.

D. Methodology - K Means Clustering

The problem at hand requires the zip codes within Manhattan to be grouped into similar neighborhoods before analyzing these similar neighborhoods for the level of COVID-19 impact and thus determining if the makeup of the most common venues within them have anything to do with the severity or level of COVID-19 spread.

To this end, we first utilized the Foursquare API to gather up all the top 100 venues within a 500 meter radius from all the zip codes. Then the list was compiled together by the frequency of each category type to arrive at the top five most common venue types for each zip code. From this data frame, we were able to conduct our K Means Clustering to divide up the 44 different zip codes within Manhattan to K clusters.

Now the biggest issue with K Clustering is determining the appropriate number for K. Too few clusters could hide some important trends while too many clusters could lead to overfitting that can prevent any trends from being discovered at all.

Test runs were done with K = 5, which reveals that certain zip codes were so unique that they alone occupied a cluster. This created a problem with 5 clusters as two of them ended up being one member clusters with the rest all crowded in three clusters. So after some further testing, I settled with K = 10. It was large enough to create meaningful clusters while small enough to not become too narrowly defined for the 44 zip codes.

In the end, these 10 clusters can be classified by the following characteristics.

1. Governor’s Island: A single member cluster due to the unique makeup and geographical location of Governor’s Island.
2. Poor residential neighborhoods that are geographically distant (Central Harlem, Chinatown, and Alphabet City), but they all represent large areas of project apartments with residents representing the lower class.
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 with luxury apartments.
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 is made up of three zip codes dominated by pizza places, restaurants, and bars.
9. Neighborhoods around Harlem with a heavy hispanic population with Mexican restaurants as the most common venue.
10. A singular and predominantly Chinese neighborhood in Washington Heights.

E. Results & Analysis

E.1. Analysis

Since the goal of this analysis was to spot trends in historical data and not be 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 results is that common venues are often very telling of the socioeconomic standings of the local population and that is sometimes the most important indicator in the spread of COVID.

E.2. 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.

E.3. Final 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 reflect 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.

F. Further Direction

Since the scope of this project required the usage of the Foursquare API, the analysis was limited to using clustering with venues to spot trends and the impact of COVID-19 among the different zip codes in Manhattan which was in particular heavily hit by COVID-19 in the early months of 2020.

While this analysis has yielded meaningful results, it has also revealed other underlying causes such as the socioeconomic level of the zip codes which are reflected through the most common venues but can be better analyzed with the help of other data sources like income level, rent level, and housing price level of each zip code. Population density per room and apartment can also play a key role with the assumption that the project apartments in certain areas of the city such as Chinatown, Alphabet City, and Harlem will have a higher density due to the lower income and bigger families of these families.

Another area of exploration is the lack of predictive nature in our analysis because COVID-19 is still ongoing and data is lacking, but with the Manhattan analysis done, we could treat this as a training set of data and apply the model established here to test for similar cluster patterns in other megacities like New York City. Since this is my first independent attempt at a data science problem, I am quite satisfied with the process and the results, but I am also eager to reapproach this problem down the line with new skills, experiences, data, and insight in order to improve upon this.