Covid-19 and NYC Venues

Shiyun Zhang

November 2020

1. Introduction

1.1 Business Problem

The coronavirus has infected people's lives since March 2020 and the Covid-19 cases are still increasing daily. In this project, I will try to find out if there are some kind of relationships between Covid-19 and Venues in New York City. For example, If there are more cases when more restaurants? If there are more cases when more coffee shops?

1.2 Interest

This report will be targeted to stakeholders like the government to help people reduce their chances to get Covid-19 if people know the relationship between covid-19 and venues and avoid going to certain places. We will use our data science powers to analyze and visualize the data and then the result will be clearly displayed and expressed to the stakeholders.

2. Data acquisition and cleaning

2.1 Data needed

New York City is a large place, first of all, we need to divide New York City into neighborhoods and then get the venue data as well as Covid-19 data for each neighborhood so that we can compare the similarities and differences between those neighborhoods.
Therefore, data we need:

- number of Covid-19 cases in the neighborhood
- number of venue in the neighborhood

2.2 Data Sources

- Covid-19 data including total cases, borough, neighborhood, zip code can be found here.
- New York City population data by zip code which can be found here.
- New York City Zip Code list which can be found here.
- Venue data including type and location in every neighborhood/zipCode Area will be obtained using Foursquare API.
- Coordinates of neighbourhood/zipCode centers will be obtained using Google Maps geocoding.

2.3 Data Cleaning and Preparation

Data downloaded or scraped from multiple sources were combined into one table. There were five me missing values on the population, I simply found the data on the internet and added it to the table. The result as the following:

Table 1 zip code, neighborhood, borough, covid cases, population and percentage in one table

	ZIPCODE	Neighborhood	Borough	Covid-Cases	Population	Covid-Cases Percentage
0	10001	Chelsea/NoMad/West Chelsea	Manhattan	485	22924	0.021157
1	10002	Chinatown/Lower East Side	Manhattan	1486	74993	0.019815
2	10003	East Village/Gramercy/Greenwich Village	Manhattan	740	54682	0.013533
3	10004	Financial District	Manhattan	60	3028	0.019815
4	10005	Financial District	Manhattan	121	8831	0.013702

So we have the zipcode, neighborhood, borough, covid cases, population and covid cases percentage in one table. Covid cases percentage is calculated using covid-cases and population so that we have more accurate data to show how serious each area is. Sometimes we will think an area with more than 5000 cases is very serious but what if the population is very large. Now we have the percentage. However we have lots of different values. A better way to deal with the percentage data is to use data binning to group numbers of more or less continuous values into a smaller number of "bins". The result as the following:

Table 2 Table with CovidCases-binned After Data Binning

	ZIPCODE	Neighborhood	Borough	Covid-Cases	Population	Covid-Cases Percentage	CovidCases-binned
0	10001	Chelsea/NoMad/West Chelsea	Manhattan	485	22924	0.021157	Low
1	10002	Chinatown/Lower East Side	Manhattan	1486	74993	0.019815	Low
2	10003	East Village/Gramercy/Greenwich Village	Manhattan	740	54682	0.013533	Low
3	10004	Financial District	Manhattan	60	3028	0.019815	Low
4	10005	Financial District	Manhattan	121	8831	0.013702	Low

3. Exploratory Data Analysis

3.1 Explore Venues of the High, Medium, Low Covid Area

In order to find out the relationship between Covid-19 and Venues, we need to get the venues data in each area. Since we already get the coordinates for each zip code, we simply use the Foursquare API by sending the latitudes and longitudes for the zip codes to get nearby venues for each zip code in the high, medium, and low covid area. After some operations and calculation, we find out the 10 most common venues for each area. To make the results more visualiable, we use pie charts:

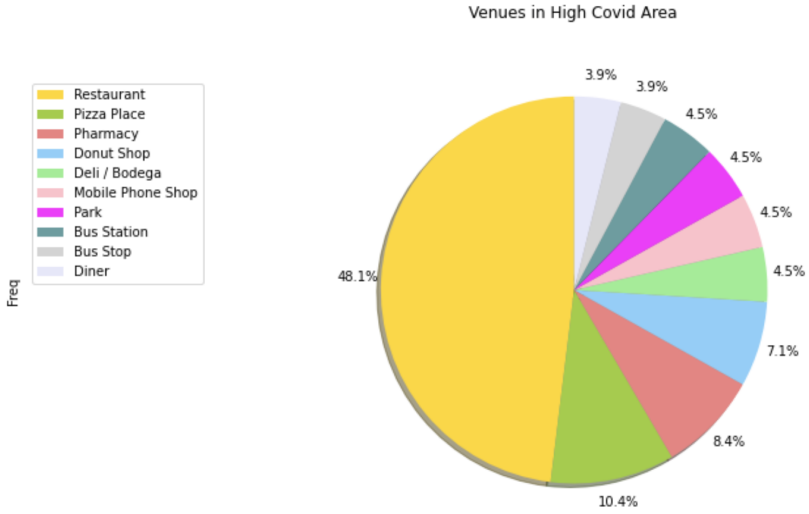

Figure 1 Venues in High Covid Area

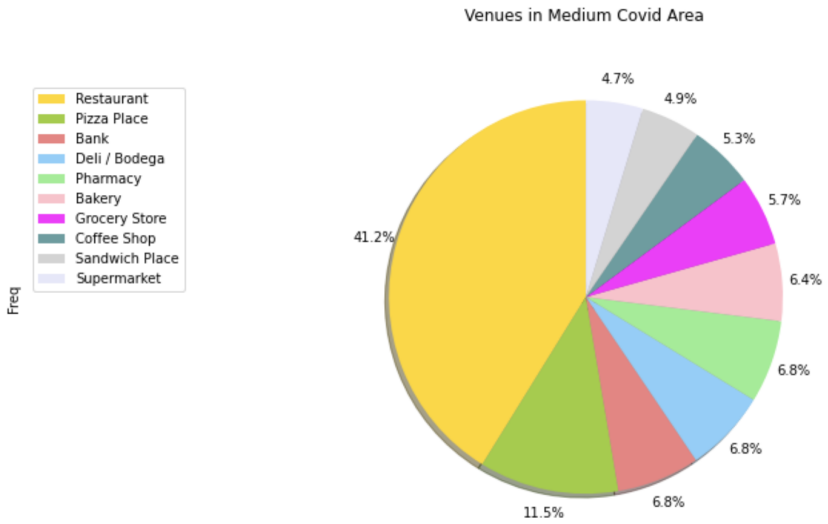

Figure 2 Venues in Medium Covid Area

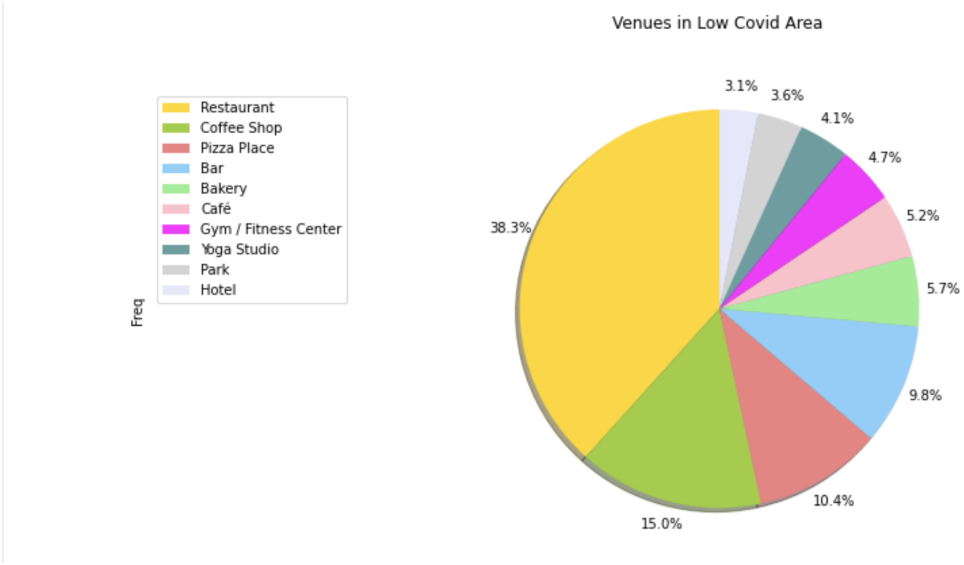

Figure 3 Venues in Low Covid Area

3.2 Compare the high, medium and low covid area

To make it more clear, we combine the three area to one table as following:

Table 3 Venues in High, Medium, Low Covid Areas

	High Covid Venue	Medium Covid Venue	Low Covid Venue
0	Restaurant	Restaurant	Restaurant
1	Pizza Place	Pizza Place	Coffee Shop
2	Pharmacy	Bank	Pizza Place
3	Donut Shop	Deli / Bodega	Bar
4	Deli / Bodega	Pharmacy	Bakery
5	Mobile Phone Shop	Bakery	Café
6	Park	Grocery Store	Gym / Fitness Center
7	Bus Station	Coffee Shop	Yoga Studio
8	Bus Stop	Sandwich Place	Park
9	Diner	Supermarket	Hotel

4. Methodology

In this project we will direct our efforts on comparing the venues in the high-covid area, medium-covid area and the low-covid area of New York City so that we can explore the relationship between venues and Covid-19. In the first step we have collected the required data: zip code, the number of covid cases in each zip code area, the population in each zip code area, percentage of covid cases based on population.

Second step in our analysis will be exploration of ' venues categories' across the high-covid area, the medium-covid area and the low-covid area. We will use the Foursquare API to get the venues data for each zip code in each area. And then we will gather all the venues information and do calculations on the high-covid area, the medium-covid area and the low-covid area so that we will have venues data based on the high, medium and low covid categories.

In the third and final step we will focus on comparing the venues in the high-covid area, medium-covid area and the low-covid area of New York City. We will present the data in one

table and compare those three areas venues using pie charts so that our stakeholder will have a better understanding of the data and also a much more clear view on the data.

5. Results and Discussion

As you can see on section 3 table 3, all the three areas have similar types of Venues which are 'Restaurant', 'Pizza Place', 'Pharmacy'. However those places are necessary in people's daily life, we should not consider them. The difference we can see between the High Covid Area and the Medium and Low Covid Area is the 'Bus Station' and 'Bus Stop'.

To make improvement on this project, we should remove the most common venues in people’s daily life so that we can get more accurate data.

6. Conclusion

The analysis shows that the difference between High Covid Area Venues and the Medium and Low Covid Area Venues are Bus Station and Bus Stop.

As we know, Covid-19 is a contagious respiratory and vascular disease. Contagious means spread from one person or organism to another by direct or indirect contact. Therefore people will have more chances to be affected if they travel to different places and talk to different people. There is a lot of traffic at the bus station. People come and go. More Bus Stations will increase the propagation speed and cause more Covid-19 cases.

To protect others and ourselves, we should stay home as much as we can, at least we should avoid taking public transit like bus and subway.