IBM Data Science Capstone Project

Title: Relationship between Covid-19 and neighborhood venues

Objective:

The goal of this project is to explore the relationships, if any, between neighborhood venues and Covid-19 infection and death rates.

New York City is a diverse city with many distinct neighborhoods. Neighborhood venues are a reflection of the life-style within each neighborhood, and life-style may have a connection to how Covid-19 infected people through community spread. For example, several states that saw spiking Covid-19 infection rates have decided to shutdown bars and indoor dinning. Additionally, life-style choices have health implications on the community living in each neighborhood. Community health, in turn, has an impact on infection rate and death rate.

Target Audience:

This analysis is intended to inform public health officials and hopefully trigger beneficial policy changes. For example, one category of venues used in this analysis is the “Take_ out_food” category (e.g., McDonald’s, cake shops) . These venues have little or no seating space, therefore customers generally order something and take it away to consume. So far, public health officials have not identified this type of venues as high risk. However, these venues generally have high volume of traffic, so more people can potentially come into contact with an infected employee, for example. Moreover, take out foods are often associated with bad health outcomes which can cause the neighborhood population to be more vulnerable to infection and death.

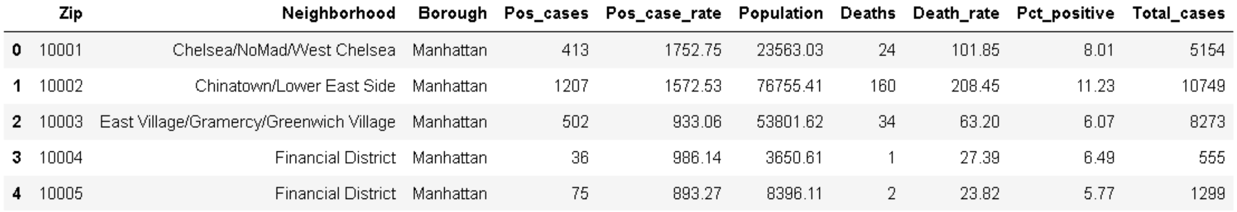
Data

The New York City (NYC) Covid-19 data is organized by Modified Zip Code Tabulation Areas (MODZCTA). This is also the method used by the US Census Bureau. Therefore, the data required are:

- NYC Covid-19 infection and death data by MODZCTA (ctl-click here).
- A mapping from MODZCTA to longitude and latitude (geospatial data) for extracting venue data from FourSquare. This data is available by state, ctl-click here for New York State data.
- Venues data from FourSquare extracted by MODZCTA.

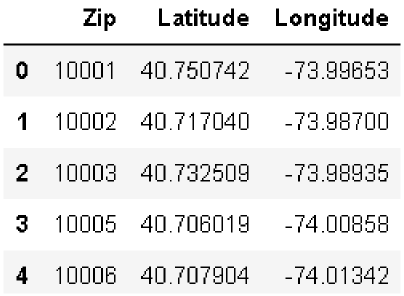
Going forward, let’s use Zipcode instead of MODZCTA.

Here is a sample of the NYC Covid-19 data:



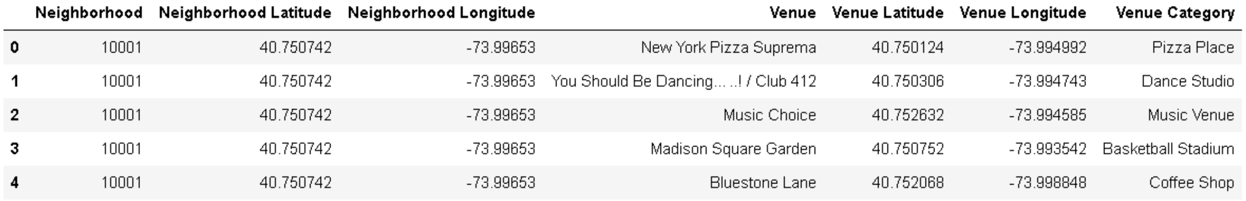
	Zip	Neighborhood	Borough	Pos_cases	Pos_case_rate	Population	Deaths	Death_rate	Pct_positive	Total_cases
0	10001	Chelsea/NoMad/West Chelsea	Manhattan	413	1752.75	23563.03	24	101.85	8.01	5154
1	10002	Chinatown/Lower East Side	Manhattan	1207	1572.53	76755.41	160	208.45	11.23	10749
2	10003	East Village/Gramercy/Greenwich Village	Manhattan	502	933.06	53801.62	34	63.20	6.07	8273
3	10004	Financial District	Manhattan	36	986.14	3650.61	1	27.39	6.49	555
4	10005	Financial District	Manhattan	75	893.27	8396.11	2	23.82	5.77	1299

Here is a sample of the longitude and latitude data:



	Zip	Latitude	Longitude
0	10001	40.750742	-73.99653
1	10002	40.717040	-73.98700
2	10003	40.732509	-73.98935
3	10005	40.706019	-74.00858
4	10006	40.707904	-74.01342

Here is a sample of venues data downloaded from FourSquare:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	10001	40.750742	-73.99653	New York Pizza Suprema	40.750124	-73.994992	Pizza Place
1	10001	40.750742	-73.99653	You Should Be Dancing... ..! / Club 412	40.750306	-73.994743	Dance Studio
2	10001	40.750742	-73.99653	Music Choice	40.752632	-73.994585	Music Venue
3	10001	40.750742	-73.99653	Madison Square Garden	40.750752	-73.993542	Basketball Stadium
4	10001	40.750742	-73.99653	Bluestone Lane	40.752068	-73.998848	Coffee Shop

Data Cleansing and Transformation

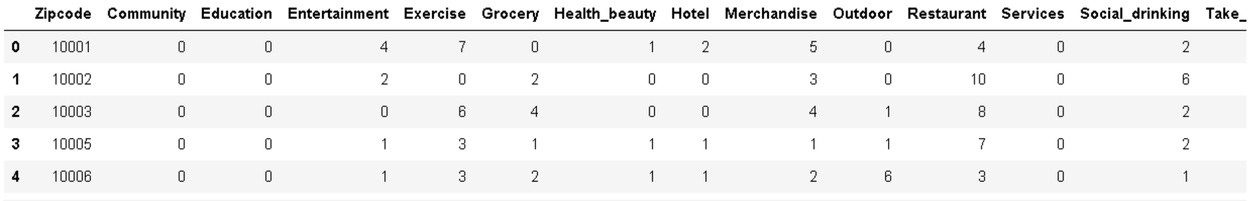
After merging the NYC Covid-19 data with the NY State geospatial data, it was discovered that four Zipcodes did not have longitude and latitude mappings. These were removed from the dataset. Additionally, one Zipcode's (11237) geospatial data was rejected by FourSquare for unknown reasons. This is also removed from the dataset.

One transformation was needed to make the data more useable. FourSquare provided a total of 332 categories of venues across NYC. This is too granular. These 332 categories were condensed into the following 14 categories:

Category Name	Examples
Community	Community center, church, synagogue
Education	School, university, professional training center

Category Name	Examples
Entertainment	Movie theater, music hall, museum
Exercise	Sports center, gym, tennis facility
Grocery	Supermarket, grocery store, liquor store
Health_beauty	Hospital, doctor’s office, beauty salon, barber shop
Hotel	Hotel, hostel
Merchandise	Arts and craft store, gift shop, department store
Outdoor	Parks, squares, attractions
Restaurant	Sit down restaurants, Italian, Japanese, African
Services	Bike rental, pet service, post office
Social_drinking	Bars, lounges, beer hall
Take_out_food	McDonald’s, café, cake shop, take out burger
Transportation	Bus station, train station, airport

After condensing the categories and using one-hot to create a separate column for each category, the data looks like:



	Zipcode	Community	Education	Entertainment	Exercise	Grocery	Health_beauty	Hotel	Merchandise	Outdoor	Restaurant	Services	Social_drinking	Take_
0	10001	0	0	4	7	0	1	2	5	0	4	0	2	
1	10002	0	0	2	0	2	0	0	3	0	10	0	6	
2	10003	0	0	0	6	4	0	0	4	1	8	0	2	
3	10005	0	0	1	3	1	1	1	1	1	7	0	2	
4	10006	0	0	1	3	2	1	1	2	6	3	0	1	

Lastly, the venues data will be transformed into per-capita data through division by the population for each Zipcode. Population data is available from the NYC Covide-19 dataset. The resulting data is of the same format as above.