Applied Data Science Capstone

A Study of COVID-19

Rita Lin

April 27, 2020

1. Introduction

1.1 Background

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by the severe acute respiratory syndrome coronavirus (SARS-CoV-2) that outbroke in December 2019. As of April 27th, 2020, more than 3 million people around the world were infected with coronavirus, among which more than 1 million people are from the United States. So far, the total number of death toll around the world has exceeded 200,000. Therefore, it is of importance for the public to understand what factors are related to the number of COVID-19 cases, so that effective measures can be taken to reduce the risk of exposure to coronavirus.

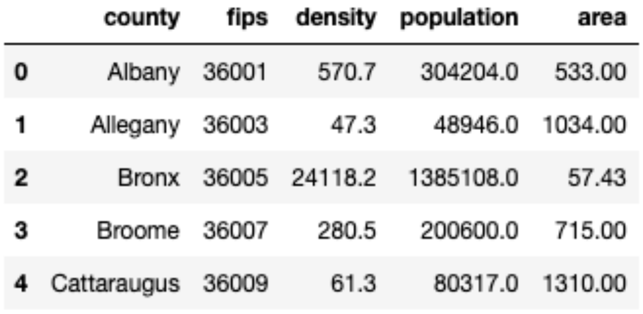
1.2. Problem and Interest

This work aims to answer if different *venue types* in the neighborhoods is related to the number of COVID-19 cases. The targeted audience of this work is the general public who are interested in learning the relationship between daily activities and COVID-19 infections. New York City, the epicenter of the coronavirus outbreak in the US, will be studied as an example. Besides *venue types* in in New York City neighborhoods, *population* and *population density* will also be explored as factors for COVID-19 infections on the county level across New York State.

2. Data

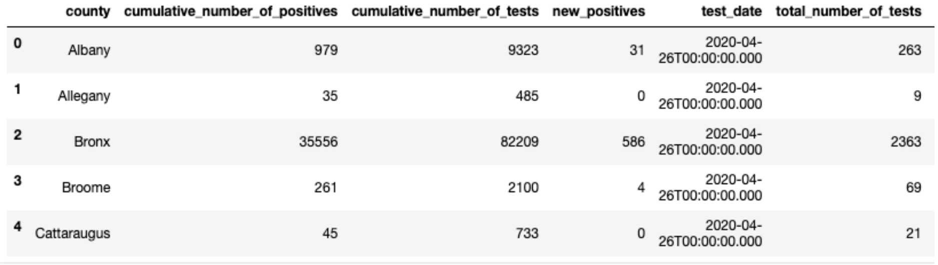
The data used for this work includes the following:

- County information in New York State, scraped from Wikipedia. *Population*, *population density* and *FIPS code* data is extracted from this source.
    - Data Sample:



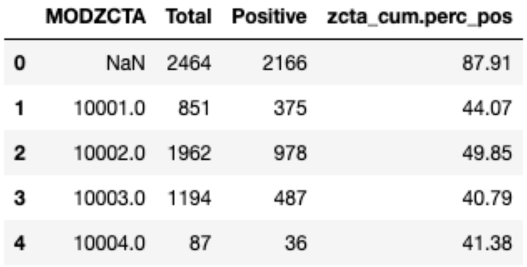
	county	fips	density	population	area
0	Albany	36001	570.7	304204.0	533.00
1	Allegany	36003	47.3	48946.0	1034.00
2	Bronx	36005	24118.2	1385108.0	57.43
3	Broome	36007	280.5	200600.0	715.00
4	Cattaraugus	36009	61.3	80317.0	1310.00

- COVID-19 statistics for New York State, obtained from the New York State government website.
  - Data Sample:



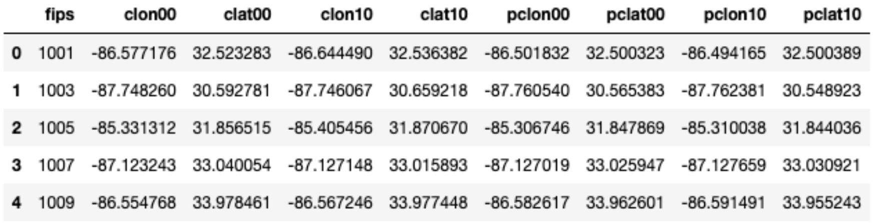
	county	cumulative_number_of_positives	cumulative_number_of_tests	new_positives	test_date	total_number_of_tests
0	Albany	979	9323	31	2020-04-26T00:00:00.000	263
1	Allegany	35	485	0	2020-04-26T00:00:00.000	9
2	Bronx	35556	82209	586	2020-04-26T00:00:00.000	2363
3	Broome	261	2100	4	2020-04-26T00:00:00.000	69
4	Cattaraugus	45	733	0	2020-04-26T00:00:00.000	21

- COVID-19 statistics for New York City, obtained from NYC Department of Health and Mental Hygiene GitHub.
  - Data Sample:



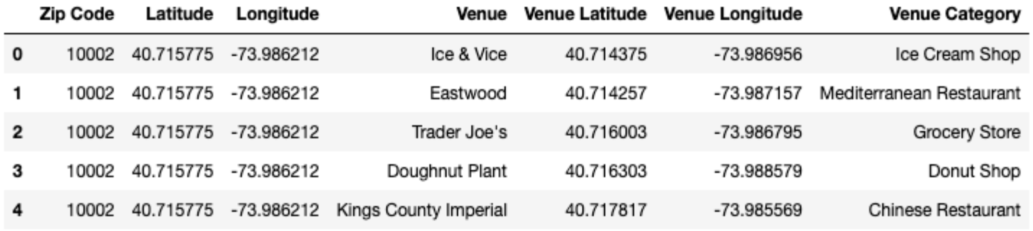
	MODZCTA	Total	Positive	zcta_cum.perc_pos
0	NaN	2464	2166	87.91
1	10001.0	851	375	44.07
2	10002.0	1962	978	49.85
3	10003.0	1194	487	40.79
4	10004.0	87	36	41.38

- County and zip code coordinate data, obtained from GitHub open sources.
  - Since geolocator from GeoPy has returned wrong coordinate information for multiple counties, public location files will be used to return coordinate data for better accuracy.
  - Data Sample:



	fips	clon00	clat00	clon10	clat10	pclon00	pclat00	pclon10	pclat10
0	1001	-86.577176	32.523283	-86.644490	32.536382	-86.501832	32.500323	-86.494165	32.500389
1	1003	-87.748260	30.592781	-87.746067	30.659218	-87.760540	30.565383	-87.762381	30.548923
2	1005	-85.331312	31.856515	-85.405456	31.870670	-85.306746	31.847869	-85.310038	31.844036
3	1007	-87.123243	33.040054	-87.127148	33.015893	-87.127019	33.025947	-87.127659	33.030921
4	1009	-86.554768	33.978461	-86.567246	33.977448	-86.582617	33.962601	-86.591491	33.955243

- Venue information in New York City, obtained using Foursquare API.
  - Data Sample:



	Zip Code	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	10002	40.715775	-73.986212	Ice & Vice	40.714375	-73.986956	Ice Cream Shop
1	10002	40.715775	-73.986212	Eastwood	40.714257	-73.987157	Mediterranean Restaurant
2	10002	40.715775	-73.986212	Trader Joe's	40.716003	-73.986795	Grocery Store
3	10002	40.715775	-73.986212	Doughnut Plant	40.716303	-73.988579	Donut Shop
4	10002	40.715775	-73.986212	Kings County Imperial	40.717817	-73.985569	Chinese Restaurant

3. Methodology

3.1 Exploratory Data Analysis

The total number of confirmed COVID-19 cases was first plotted versus the population density of each county in New York State on a scatter plot.

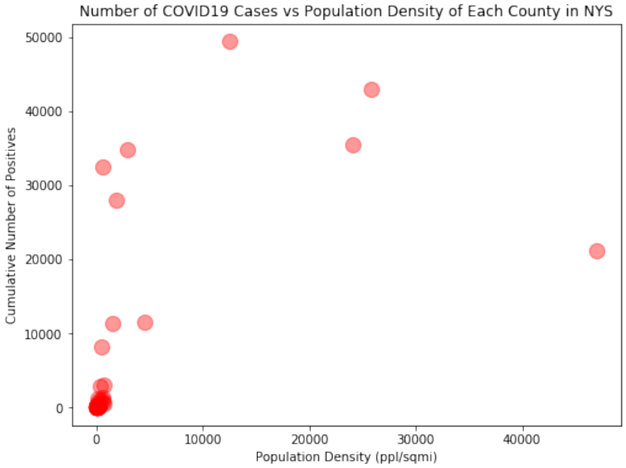


Figure 1. Number of COVID-19 Cases vs Population Density of Each County in NYS

According to the plot, a positive relationship can be observed between the total number of positive COVID-19 cases and the population density of each county. However, the data as a whole displays a weak linearity.

The total number of confirmed COVID-19 cases was then plotted versus the population of each county in New York State on a scatter plot.

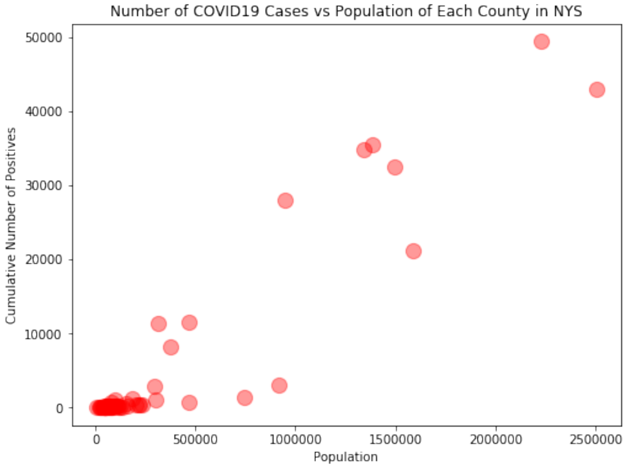


Figure 2. Number of COVID-19 Cases vs Population of Each County in NYS

Similar to the previous plot, a positive relationship can be observed between the total number of positive COVID-19 cases and the total population of each county. But unlike the previous plot, a relatively high linearity is shown by the data.

To investigate the relationship between number of tests performed and population, the two variables were plotted on a scatter plot.

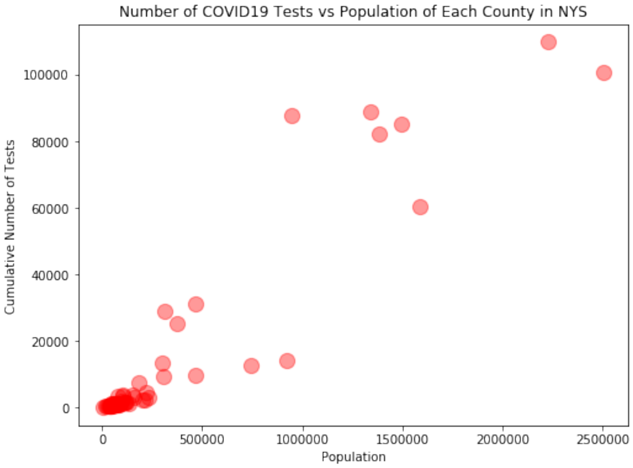


Figure 3. Number of COVID-19 Tests vs Population of Each County in NYS

This trend in the above plot is similar to that in Figure 2, which shows the number of COVID-19 cases vs population in New York State. The data in the bottom left corner in both plots even shows high similarity in both plots. This indicates that there might be a strong relationship between the number of COVID-19 cases and the number of COVID-19 tests.

Last but not least, the number of COVID-19 cases and the number of COVID-19 test are plotted on a scatter plot.

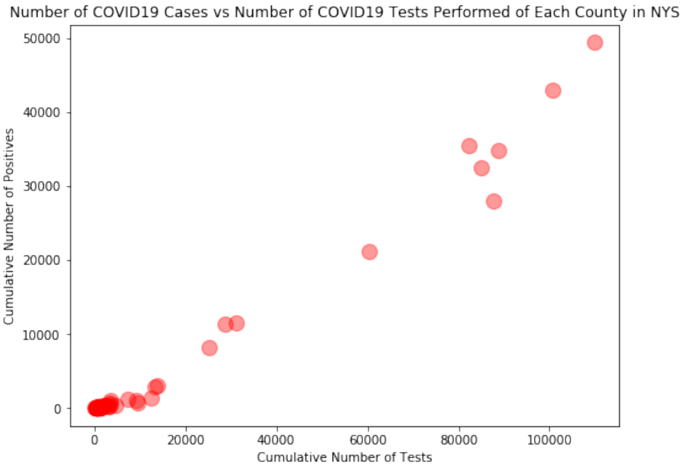


Figure 4. Number of COVID-19 Cases vs Number of COVID-19 Tests Performed of Each County in NYS

As expected, a strong linear relationship between the number of COVID-19 cases and the number of COVID-19 test can be seen from the plot, indicating that the more tests are performed, the more case confirmations there are.

3.2 COVID-19 Case Mapping

Population density and the cumulative number of COVID-19 cases in each county in New York State were mapped with Folium. Population density is represented with a choropleth map, and the cumulative number of COVID-19 cases is represented by red circular markers. Larger the red circular marker is, the higher the cumulative number of cases is.

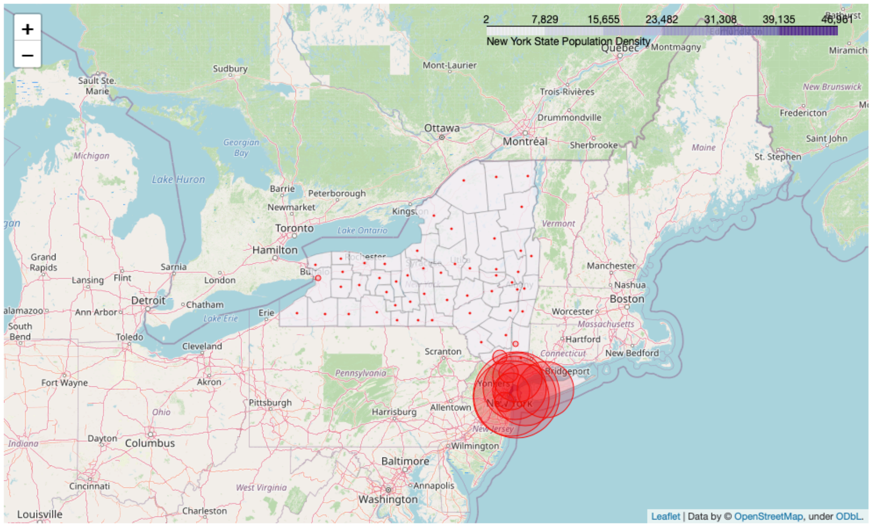

2
7,829
15,655
23,482
31,308
39,135
46,961
New York State Population Density
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 5. COVID-19 Case Mapping (New York State)

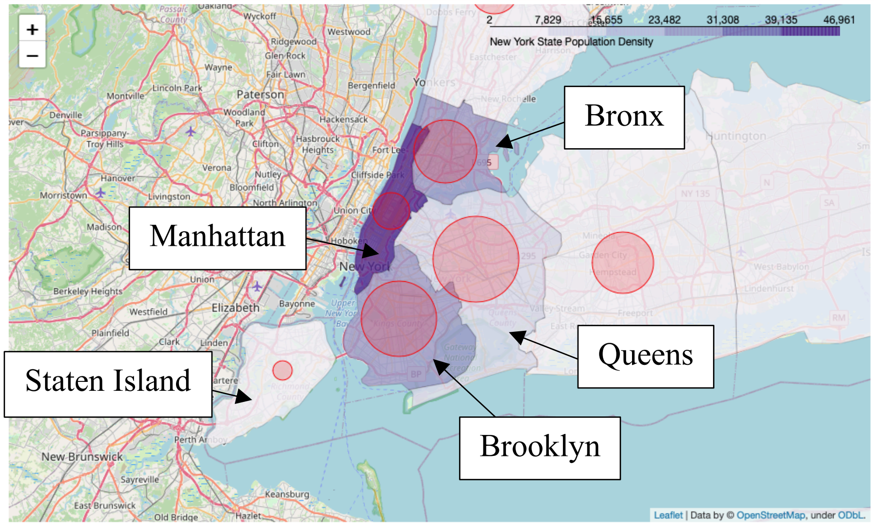

2
7,829
15,655
23,482
31,308
39,135
46,961
New York State Population Density
Bronx
Manhattan
Queens
Staten Island
Brooklyn
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 6. COVID-19 Case Mapping (New York City)

The above maps are alternative representations of the scatter plot in Figure.1. It can be seen from Figure 5 that the majority of the COVID-19 cases in New York State is in or close to New York City. Zooming in the map to take a closer look at New York City, we see that Queens, Kings (Brooklyn) and Bronx have highest numbers of COVID-19 cases in New York. What's interesting is that New York county (Manhattan), which has the highest population density in New York, surprisingly has a significantly lower number of COVID-19 cases comparing to the proximal counties.

3.3 New York City Neighborhood Clustering

To better understand what's going on in New York City, the relationship between venue type clusters and the number of COVID-19 cases was studied. The location of each zip code in New York City was first visualized on the map with Folium.

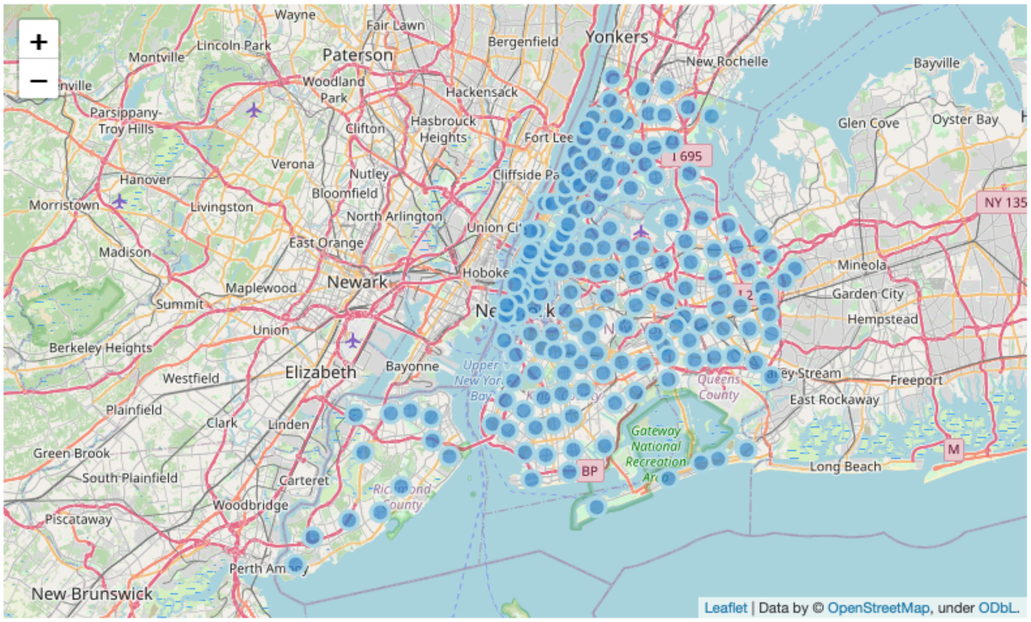
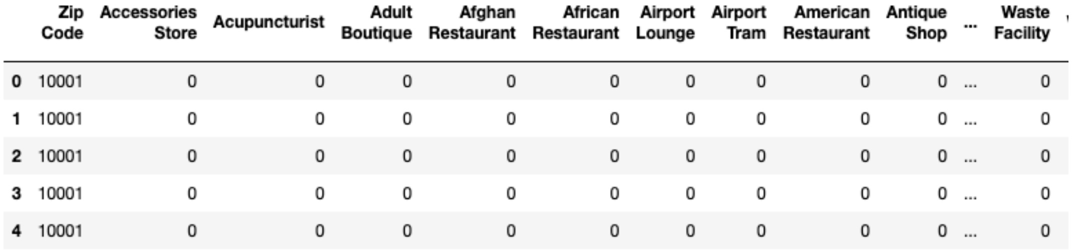


Figure 7. Locations of Zip Codes in New York City

Then Foursquare API was used to fetch venue data for each zip code address. After venue information was collected for all zip codes, one hot encoding technique was used to get the data ready for clustering.

Sample Data:



	Zip Code	Accessories Store	Acupuncturist	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Tram	American Restaurant	Antique Shop	...	Waste Facility
0	10001	0	0	0	0	0	0	0	0	0	...	0
1	10001	0	0	0	0	0	0	0	0	0	...	0
2	10001	0	0	0	0	0	0	0	0	0	...	0
3	10001	0	0	0	0	0	0	0	0	0	...	0
4	10001	0	0	0	0	0	0	0	0	0	...	0

Figure 8. Venue Data Processed Using One Hot Encoding Technique

K-Means clustering method was used to explore venue characteristics in different areas. The number of clusters was chosen to be 5. Venue clusters was then visualized on top of a COVID-19 choropleth map.

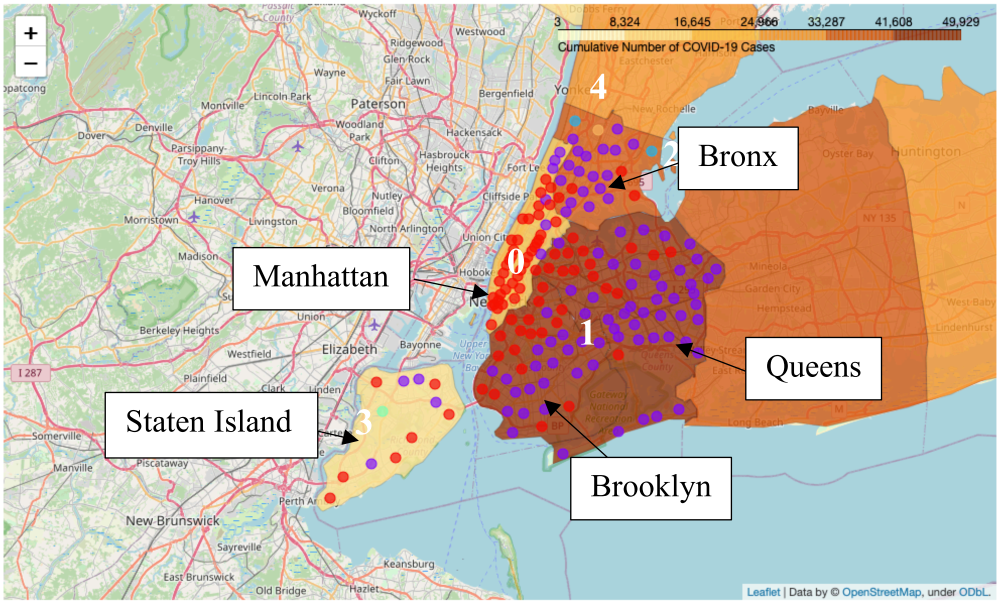
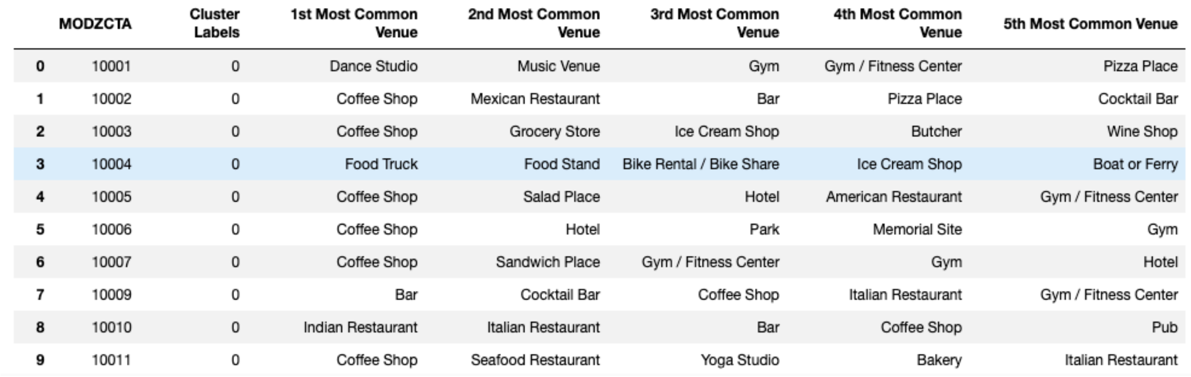


Figure 8. Venue Clusters and COVID-19 Case Distribution

4. Results

According to the choropleth map, venue cluster 0 and venue cluster 1 are the two main clusters in New York City. It can be seen from map that boroughs, or counties, with cluster 1 dominating (Bronx, Queens, Brooklyn) have significantly higher number of COVID-19 cases than boroughs with cluster 0 dominating (Manhattan, Staten Island). Cluster 0 and cluster 1 were then examined more closely for characteristics.

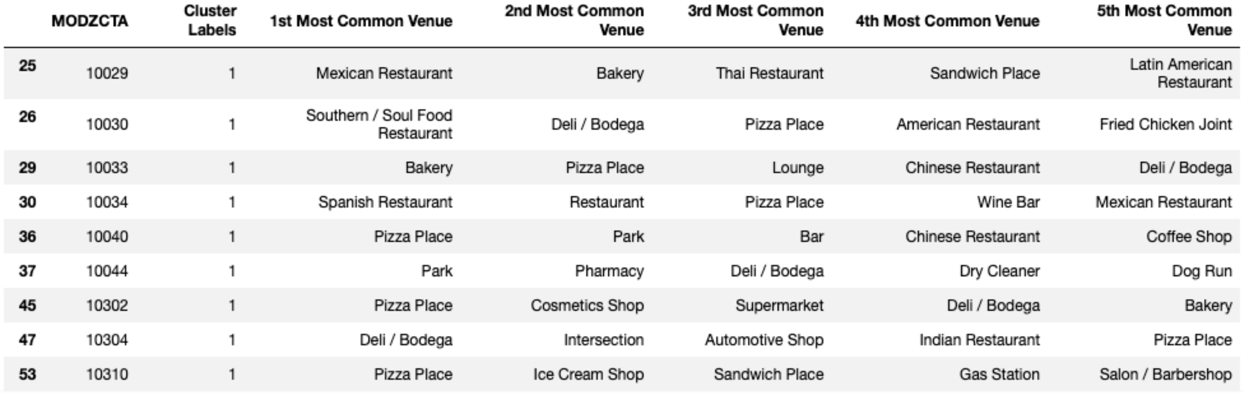
Cluster 0 Sample Data:



	MODZCTA	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	10001	0	Dance Studio	Music Venue	Gym	Gym / Fitness Center	Pizza Place
1	10002	0	Coffee Shop	Mexican Restaurant	Bar	Pizza Place	Cocktail Bar
2	10003	0	Coffee Shop	Grocery Store	Ice Cream Shop	Butcher	Wine Shop
3	10004	0	Food Truck	Food Stand	Bike Rental / Bike Share	Ice Cream Shop	Boat or Ferry
4	10005	0	Coffee Shop	Salad Place	Hotel	American Restaurant	Gym / Fitness Center
5	10006	0	Coffee Shop	Hotel	Park	Memorial Site	Gym
6	10007	0	Coffee Shop	Sandwich Place	Gym / Fitness Center	Gym	Hotel
7	10009	0	Bar	Cocktail Bar	Coffee Shop	Italian Restaurant	Gym / Fitness Center
8	10010	0	Indian Restaurant	Italian Restaurant	Bar	Coffee Shop	Pub
9	10011	0	Coffee Shop	Seafood Restaurant	Yoga Studio	Bakery	Italian Restaurant

After close examination, it was found that some of the most common venues in cluster 0 are coffee shop, bakery and bar.

Cluster 1 Sample Data:



	MODZCTA	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
25	10029	1	Mexican Restaurant	Bakery	Thai Restaurant	Sandwich Place	Latin American Restaurant
26	10030	1	Southern / Soul Food Restaurant	Deli / Bodega	Pizza Place	American Restaurant	Fried Chicken Joint
29	10033	1	Bakery	Pizza Place	Lounge	Chinese Restaurant	Deli / Bodega
30	10034	1	Spanish Restaurant	Restaurant	Pizza Place	Wine Bar	Mexican Restaurant
36	10040	1	Pizza Place	Park	Bar	Chinese Restaurant	Coffee Shop
37	10044	1	Park	Pharmacy	Deli / Bodega	Dry Cleaner	Dog Run
45	10302	1	Pizza Place	Cosmetics Shop	Supermarket	Deli / Bodega	Bakery
47	10304	1	Deli / Bodega	Intersection	Automotive Shop	Indian Restaurant	Pizza Place
53	10310	1	Pizza Place	Ice Cream Shop	Sandwich Place	Gas Station	Salon / Barbershop

After close examination, it was found that some of the most common venues in cluster 1 are restaurant, supermarket and bus station.

5. Discussion

Based on the results of exploratory data analysis, the number of COVID-19 cases in New York State is directly proportional to the population and the number of tests performed. This suggests the importance of getting enough number of COVID-19 tests. That way more cases can be identified. On the other hand, the relationship between the number of COVID-19 cases and population density is also positive, but displays a weak linearity. In counties with high population density, the number of COVID-19 cases is even inversely proportional to population density. This could be a result of strict stay-at-home order in county with high population density, suggesting the importance of physical distancing in preventing the spread of the viruses.

For venue clustering results in New York City, it can be seen that areas with restaurants, supermarkets and bus stations dominating have more COVID-19 cases (Bronx, Brooklyn and Queens). These venues share some common characteristics – these places serve people’s daily needs, so they tend to have large flows of people. Moreover, in these places, people tend to stay in a confined space for a relatively long time. Therefore, it is highly recommended use protective gears when visiting these areas or even avoid going in person. On the other hand, areas with coffee shops, bakeries and bars dominating have significantly smaller number of COVID-19 cases (Manhattan and Staten Island). One explanation could be that people tend to stay for a short amount of time in coffee shops and bakeries, and many people don't go to bars often.

6. Conclusion

To avoid spreading the coronavirus, it is strongly recommended to stick strictly with the stay-at-home order if the area has a significant number of COVID-19 cases. It is also important to make enough tests so that more cases can be identified. Based on the venue clustering results, dining at restaurants and shopping at supermarkets are not recommended in areas that has large number of COVID-19 cases. If one needs to go to restaurants and supermarkets in person, it is recommended to use protective gears to prevent the spread of viruses.