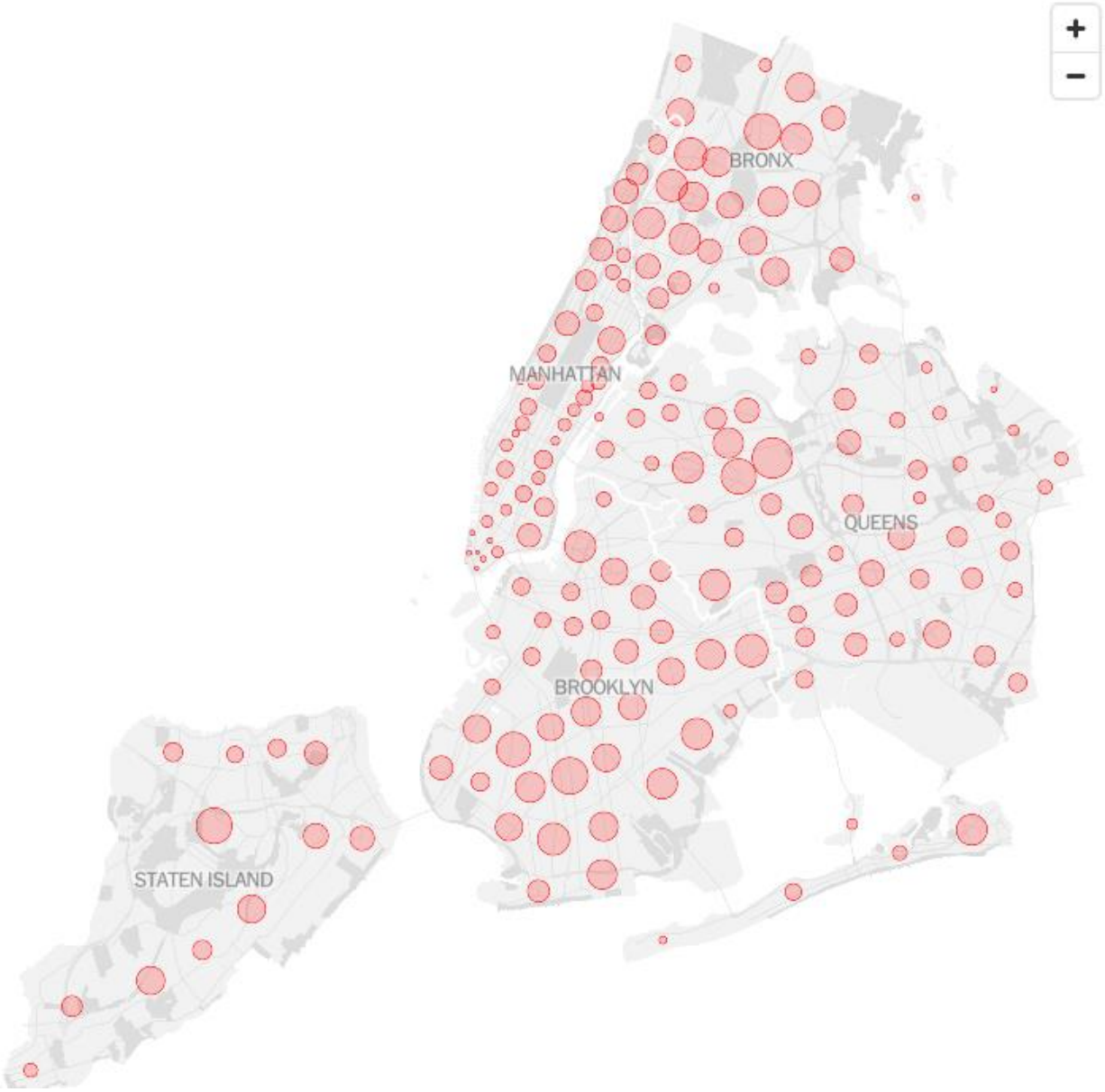
Capstone Project: The Battle of Neighborhoods

What are the common features of areas with high COVID case rate in New York?



Introduction

On March 11, 2020, the World Health Organization declared that COVID-19 was a global pandemic, indicating a significant global spread of an infectious disease [1]. At that point, there were 118,000 confirmed cases of the coronavirus in 110 countries [2]. China was the first country to report the spread of this new disease with an outbreak in January 2020. South Korea, Iran, and Italy followed in February with their own outbreaks. Weeks later, the virus was in all continents and over 177 countries. Unfortunately, the United States has the highest number of confirmed cases and, sadly, the most deaths. The virus was extremely contagious and led to death in the most vulnerable, particularly those older than 60 and those with underlying conditions. A lot of cities suffered from overwhelming local health care systems.

As the deaths rose from the virus that had no effective treatment or vaccine, countries shut their borders, banned travel to other countries, and began to issue orders for their citizens to stay at home. Schools and universities closed their physical locations and moved education online. Sporting events were canceled, airlines cut flights, tourism evaporated, restaurants, movie theaters and bars closed, theater productions canceled, manufacturing facilities, services, and retail stores closed. In some businesses and industries, employees have been able to work remotely, but in others, workers have been laid off, furloughed, or had their hours cut. However, the virus didn't disappear like a miracle during the summer. And the situation bounced back in this fall and winter. This long-lasting situation has largely changed people's daily lives.

In this capstone project, I did a tiny study to explore the distribution of COVID cases in New York and tried to understand what were the common features of those neighborhoods with relatively high COVID confirmed cases. With the venues data collected from Foursquare, neighborhoods in New York were clustered into 5 groups. The results showed that 70% of the top 20 areas with the highest confirmed case rate were from the same group. And most venues in these areas with the highest amounts were in food-related categories, such as Pizza Place, Deli/Bodega, Restaurant, Bagel Shop, Ice Cream Shop, Grocery Store, Bakery, and Coffee Shop. The result suggests that we need to take extra caution when we have to visit these kinds of places during the pandemic.

This pandemic has demonstrated the interconnected nature of our world and that no one is safe until everyone is safe. Only by acting in solidarity can communities save lives and overcome the devastating impacts of the virus. I hope this little finding in this study can serve as a caution for everyone, helping them protect themselves, protect their families, and protect people around them.

Data

COVID data

The COVID case data of New York come from NYC Health [3,4]. The data contains cumulative totals since the start of the COVID-19 outbreak in New York City, which the Health Department defines as the diagnosis of the first confirmed COVID-19 case on February 29, 2020 [4]. The

data used in this study were grouped by Modified Zip Code Tabulation Areas (MODIFIED_ZCTA).

The geography information is reported using MODIFIED_ZCTA because it can be challenging to map data that are reported by ZIP Code. A ZIP Code doesn't actually refer to an area, but rather a collection of points that make up a mail delivery route. Furthermore, there are some buildings that have their own ZIP Code, and some non-residential areas with ZIP Codes. To deal with the challenges of ZIP Codes, the Health Department uses ZCTAs which solidify ZIP codes into units of area. Often, data reported by ZIP code are actually mapped by ZCTA. The ZCTA geography was developed by the U.S. Census Bureau. The modified ZCTA geography combines census blocks with smaller populations to allow more stable estimates of population size for rate calculation.

In this dataset, one MODIFIED_ZCTA may contains one or more neighborhoods. And in some cases, one neighborhood is also separated into two or more MODIFIED_ZCTAs. Since this will not influence my search for features in areas with high COVID rate, I will perform the analysis using MODIFIED_ZCTA instead of neighborhood.

The data contain both number of confirmed cases by MODIFIED_ZCTA (COVID_CASE_COUNT) and rate of confirmed cases per 100,000 people by MODIFIED_ZCTA (COVID_CASE_RATE). To minimize the influence of different populations in different area, I only use COVID_CASE_RATE for my study.

The COVID data are updated daily. The data used in this study are based on the version of 12/16/2020.

Geographical coordinate

The latitude and longitude for each MODIFIED_ZCTA are collected from GeoPy using geocoders.nominatim. These geographical coordinates will be used to search for venues in each area.

Venue data

Venue data are collected from Foursquare. Due to the limitation of requests that can be made, I only collect 100 venues for each MODIFIED_ZCTA with a radius of 1000m. The constraint here may have influence the final result. But it is still possible to obtain some preliminary understandings from the venue data. The venue data will be used to clustered different areas of New York City into groups. And areas with cluster labels will be compared with COVID rates in order to find out whether areas with high COVID rates have common features.

Methodology

Retrieving and processing COVID data

The COVID case data of New York City are retrieved from NYC Health. The dataframe contains 10 columns, including MODIFIED_ZCTA, neighborhood name, total case count, and case rate. And it contains case counts for 177 areas. Here are the first 10 rows of the dataframe, showing parts of the columns.

	MODIFIED_ZCTA	NEIGHBORHOOD_NAME	BOROUGH_GROUP	COVID_CASE_COUNT	COVID_CASE_RATE	POP_DENOMINATOR	COVID_DEATH_COUNT	COVID_DEATH_RATE
0	10001	Chelsea/NoMad/West Chelsea	Manhattan	614	2223.58	27613.09	23	83.29
1	10002	Chinatown/Lower East Side	Manhattan	2093	2778.71	75322.71	163	216.40
2	10003	East Village/Gramercy /Greenwich Village	Manhattan	1060	1963.77	53977.81	33	61.14
3	10004	Financial District	Manhattan	93	3129.08	2972.12	1	33.65
4	10005	Financial District	Manhattan	168	1918.42	8757.23	0	0.00
5	10006	Financial District	Manhattan	73	2158.61	3381.80	1	29.57
6	10007	TriBeCa	Manhattan	143	2045.36	6991.45	4	57.21
7	10009	Alphabet City/East Village/Stuyvesant Town-Coo...	Manhattan	1412	2470.64	57151.12	66	115.48
8	10010	Flatiron/Gramercy/Kips Bay	Manhattan	688	2062.78	33353.00	23	68.96
9	10011	Chelsea	Manhattan	1107	2225.31	49745.99	47	94.48

Some areas include two or more neighborhoods. For example, 10001 contains 3 neighborhoods, Chelsea, NoMad, and West Chelsea. In some other cases, one neighborhood can be divided into several MODIFIED_ZCTA. For example, Financial District are separated into 10003, 10004, and 10005; while both 10001 and 10011 show Chelsea. Since there is no way to attribute the COVID case count into each neighborhood and I only need to focus on the common features, I will treat these neighborhoods together as one single area to simplify the process. And we will continue the following analysis based on MODIFIED_ZCTA.

Since different MODIFIED_ZCTAs have different populations, using actual COVID case counts will bias the result. To exclude the influence of population in each area, the following analysis will use COVID_CASE_RATEs instead, which represent rate of confirmed cases per 100,000 people by MODIFIED_ZCTA.

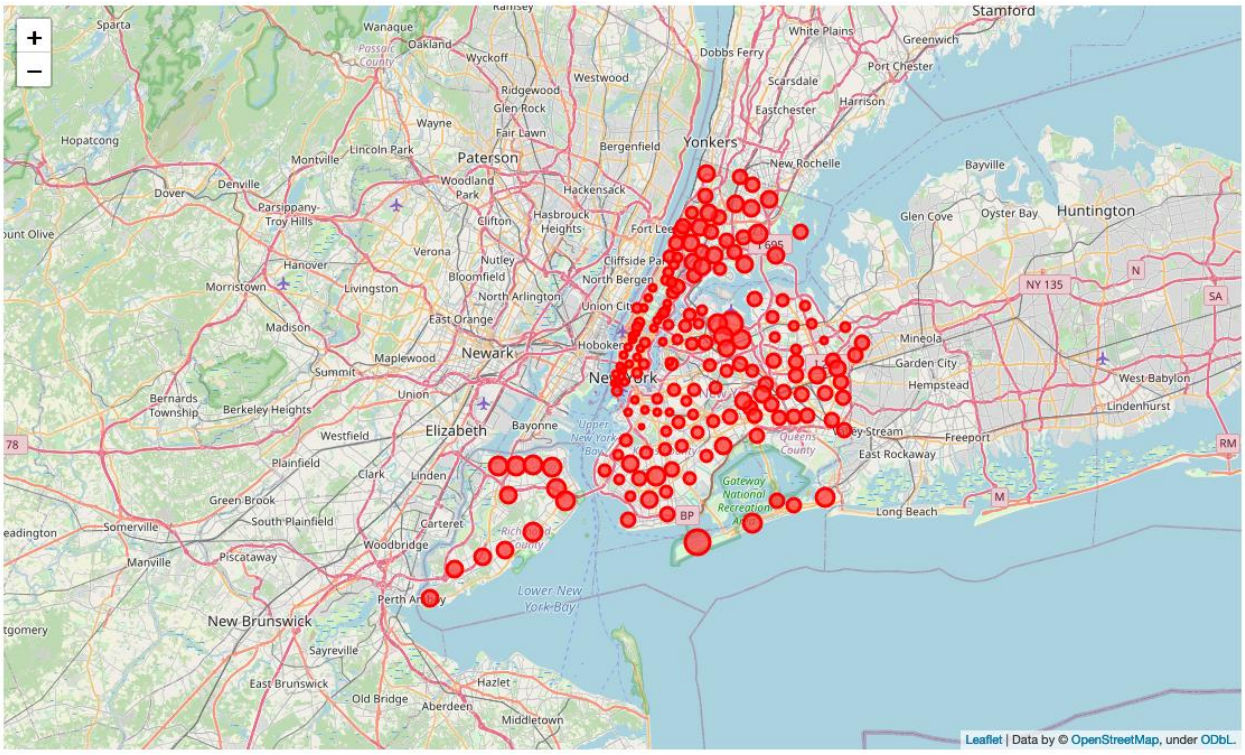
Combining geographical coordinates

In order to collect features for each area, geographical coordinates are first retrieved for each MODIFIED_ZCTA using GeoPy. Now, the data of interest look like this (showing first 5 rows only).

	MODIFIED_ZCTA	NEIGHBORHOOD_NAME	BOROUGH_GROUP	LATITUDE	LONGITUDE	COVID_CASE_RATE
0	10001	Chelsea/NoMad/West Chelsea	Manhattan	40.748418	-73.994114	2223.58
1	10002	Chinatown/Lower East Side	Manhattan	40.717058	-73.989325	2778.71
2	10003	East Village/Gramercy/Greenwich Village	Manhattan	40.731588	-73.988525	1963.77
3	10004	Financial District	Manhattan	40.700769	-74.013464	3129.08
4	10005	Financial District	Manhattan	40.720503	-74.006704	1918.42

Visualization of COVID data

With the latitudes and longitudes of all areas, we can get a basic understanding of the distribution of COVID cases by creating a map of New York with overlapping bubbles. The sizes of the circles represent the relative case rates of the areas.



Collecting venue information

To cluster these 177 neighborhoods, venue information is collected for each MODIFIED_ZCTA from Foursquare. However, we only collect 100 venues for each area due to the limitation, and we set the radius to 1000m.

A total number of 13589 venues for these 177 areas are collected, with their names and categories listed. Here is an example of the first few venues.

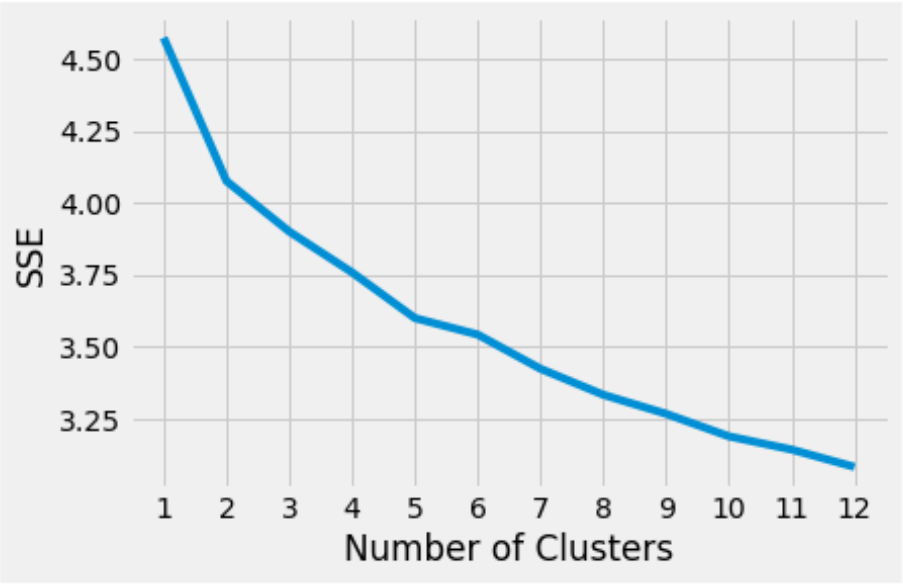
	MODIFIED_ZCTA	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	10001	40.748418	-73.994114	New York Pizza Suprema	40.750124	-73.994992	Pizza Place
1	10001	40.748418	-73.994114	iLoveKickboxing	40.746340	-73.992900	Boxing Gym
2	10001	40.748418	-73.994114	ALT: A Little Taste	40.746854	-73.992449	Café
3	10001	40.748418	-73.994114	Delta Sky360° Club	40.750564	-73.992824	Lounge
4	10001	40.748418	-73.994114	Marcelo Garcia Brazilian Jiu-Jitsu Academy	40.746565	-73.996275	Martial Arts School

And there are 444 unique categories curated from all the returned venues. One hot encoding is used to represent the results. And data are further grouped by MODIFIED_ZCTA and by taking the mean of the frequency of occurrence of each category. In this way, we can easily sort our data and figure out the most common or popular venues in each area. Here is an example showing the top 10 most common venues for each area. These venues will serve as features of areas and will be used to cluster these areas.

	MODIFIED_ZCTA	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	10001	Hotel	Korean Restaurant	Gym / Fitness Center	Coffee Shop	Italian Restaurant	Theater	Japanese Restaurant	Boxing Gym	Cuban Restaurant	Park
1	10002	Ice Cream Shop	Coffee Shop	French Restaurant	Bakery	Chinese Restaurant	Pizza Place	Bar	Wine Bar	Cocktail Bar	American Restaurant
2	10003	Ice Cream Shop	Coffee Shop	Pizza Place	Wine Shop	Juice Bar	Bar	Bakery	Bagel Shop	Hotel	Chinese Restaurant
3	10004	Coffee Shop	Mexican Restaurant	Bar	Gym	American Restaurant	Hotel	Café	Heliport	Park	Monument / Landmark
4	10005	Café	Clothing Store	Men's Store	Hotel	American Restaurant	Coffee Shop	Sushi Restaurant	Italian Restaurant	Spa	Gym / Fitness Center

Clustering neighborhoods

Using the venue information collected above, we will cluster these 177 areas into several groups. In this study, the k-means method is applied. So, we need to find out the most suitable number of clusters, k. Here, we are using the elbow method.

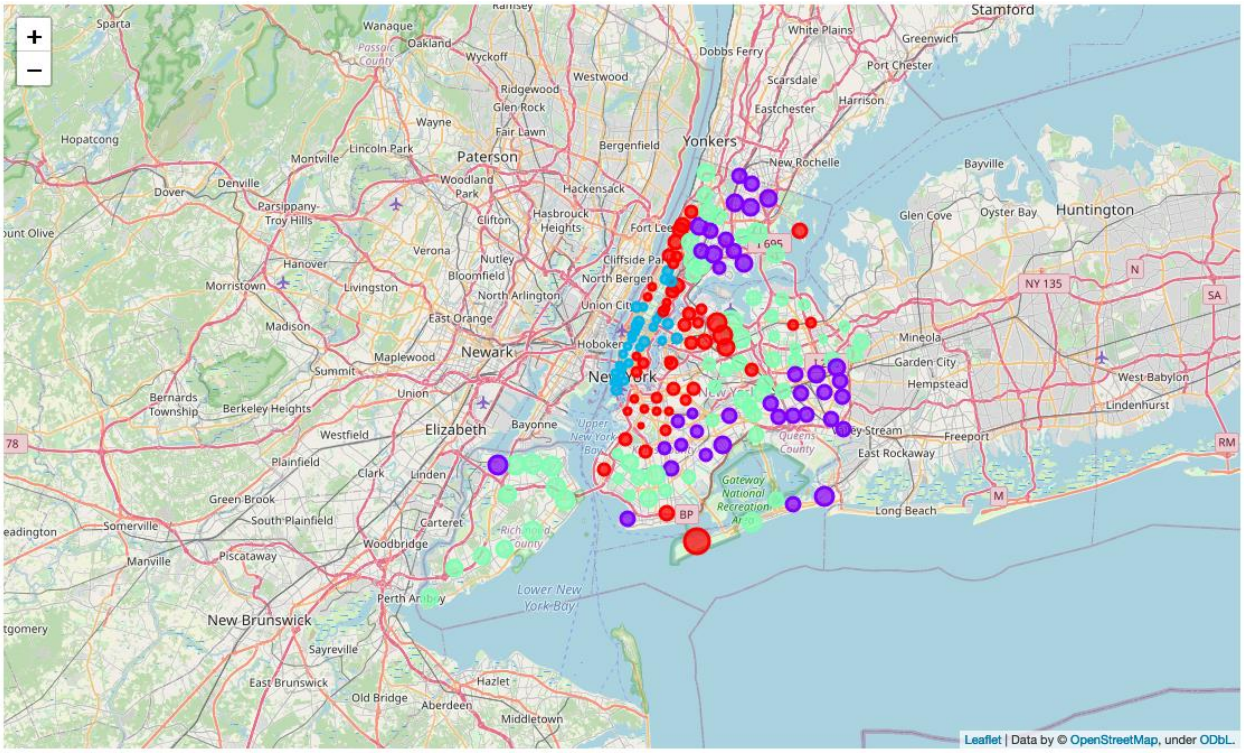


From the number of cluster vs. the sum of the squared distance between centroid and each member of the cluster (SSE), we can find that there are 2 obvious turns. Using only 2 clusters in this study is not enough, so I will use k = 5.

With a suitable number of clusters, we further apply the clustering to our data and add the cluster labels to our previous dataframe (only a small part of the dataframe is shown).

	MODIFIED_ZCTA	NEIGHBORHOOD_NAME	BOROUGH_GROUP	LATITUDE	LONGITUDE	COVID_CASE_RATE	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	10001	Chelsea/NoMad/West Chelsea	Manhattan	40.748418	-73.994114	2223.58	2	Hotel	Korean Restaurant	Gym / Fitness Center	Coffee Shop	Italian Restaurant
1	10002	Chinatown/Lower East Side	Manhattan	40.717058	-73.989325	2778.71	0	Ice Cream Shop	Coffee Shop	French Restaurant	Bakery	Chinese Restaurant
2	10003	East Village/Gramercy /Greenwich Village	Manhattan	40.731588	-73.988525	1963.77	0	Ice Cream Shop	Coffee Shop	Pizza Place	Wine Shop	Juice Bar
3	10004	Financial District	Manhattan	40.700769	-74.013464	3129.08	2	Coffee Shop	Mexican Restaurant	Bar	Gym	American Restaurant
4	10005	Financial District	Manhattan	40.720503	-74.006704	1918.42	2	Café	Clothing Store	Men's Store	Hotel	American Restaurant

With the clustering result, we can generate a map of New York City with these labels again.



Results and discussions

If we select the top 20 areas with the highest COVID case rate, we can find that 70% of them are labeled as cluster 3, which means those areas suffered from high COVID case rates do share some common features. Cluster 3 is shown as green circle in the map above.

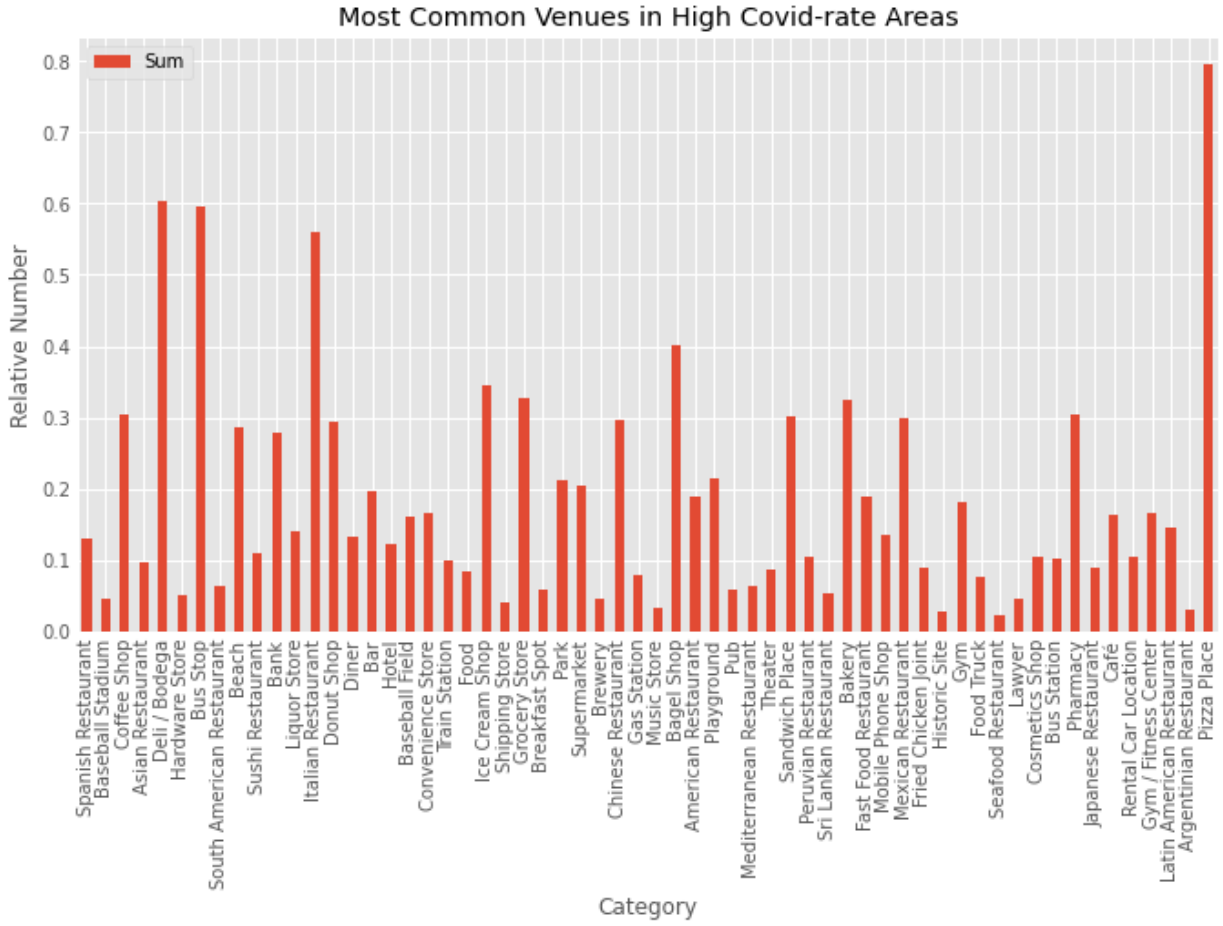
	MODIFIED_ZCTA	NEIGHBORHOOD_NAME	BOROUGH_GROUP	LATITUDE	LONGITUDE	COVID_CASE_RATE	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
176	11697	Breezy Point	Queens	40.559684	-73.915031	7928.98	0	Bar	Clothing Store	Café
140	11369	Airport/East Elmhurst	Queens	40.762499	-73.872825	6890.14	3	Hotel	Pizza Place	Rental Ca Location
49	10306	Lighthouse Hill/Midland Beach/New Dorp/Oakwood	Staten Island	40.569002	-74.116771	5876.14	3	Bank	Mexican Restaurant	Italian Restauran
45	10302	Elm Park	Staten Island	40.630756	-74.137176	5799.31	3	Chinese Restaurant	Deli / Bodega	Pizza Place
142	11372	Jackson Heights	Queens	40.751377	-73.883138	5777.59	0	Thai Restaurant	Bakery	Mexican Restauran
116	11230	Midwood	Brooklyn	40.621525	-73.965615	5716.10	3	Pizza Place	Ice Cream Shop	Bagel Shop
47	10304	New Dorp/Todt Hill	Staten Island	40.609770	-74.088140	5674.81	3	Deli / Bodega	Grocery Store	Playgroun
141	11370	Jackson Heights/Rikers Island	Queens	40.762385	-73.890552	5664.41	0	Bar	Bakery	Latin American Restauran
139	11368	Corona/North Corona	Queens	40.748072	-73.860616	5554.96	3	Mexican Restaurant	Latin American Restaurant	Pizza Place
66	10461	Morris Park/Pelham Bay/Westchester Square	Bronx	40.845863	-73.841032	5452.38	3	Pizza Place	Sandwich Place	Donut Shop
172	11691	Edgemere/Far Rockaway	Queens	40.601540	-73.757811	5390.67	1	Pizza Place	Beach	Chinese Restauran
53	10310	Port Richmond/Randall Manor/West Brighton	Staten Island	40.631926	-74.116506	5387.87	3	Bus Stop	Pizza Place	Fast Food Restauran
48	10305	Arrochar/Midland Beach/Shore Acres/South Beach...	Staten Island	40.598444	-74.076018	5378.45	3	Italian Restaurant	Pharmacy	Deli / Bodega
44	10301	Silver Lake/St. George	Staten Island	40.629516	-74.093853	5318.09	3	Bus Stop	Deli / Bodega	Sri Lankar Restauran
175	11694	Belle Harbor-Neponsit/Rockaway Park	Queens	40.577426	-73.846705	5303.33	3	Beach	Deli / Bodega	Pizza Place
46	10303	Graniteville/Mariner's Harbor/Port Ivory	Staten Island	40.630273	-74.158773	5295.60	1	Fast Food Restaurant	Discount Store	Spanish Restauran
50	10307	Tottenville	Staten Island	40.507954	-74.242519	5230.63	3	Italian Restaurant	Deli / Bodega	Grocery Store
55	10314	Bloomfield/Freshkills Park	Staten Island	40.604026	-74.147107	5225.68	3	Bus Stop	Italian Restaurant	Baseball Field
125	11239	East New York	Brooklyn	40.649444	-73.882771	5221.81	1	Department Store	Chinese Restaurant	Pizza Place
73	10468	Fordham/Kingsbridge /University Heights	Bronx	40.865528	-73.900228	5189.47	3	Pizza Place	Spanish Restaurant	Donut Shop

Now, let's further exam the venues of these 14 areas labeled as cluster 3 in the top20 list. And all these 140 categories will be weighted and merged together into one dataframe. In this way, we can find out the most common venues in the high COVID case rate areas.

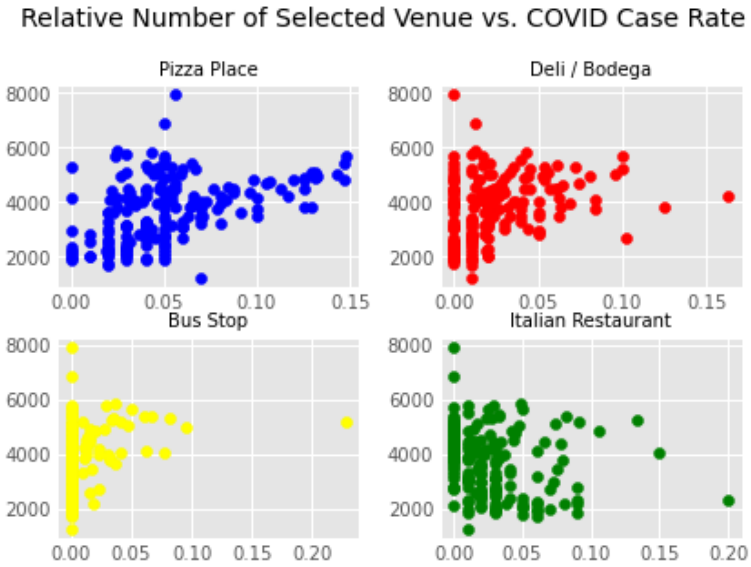
Among these 14 areas, there are 60 unique venue categories. The top 20 most common categories are shown below:

	Sum
Pizza Place	0.794616
Deli / Bodega	0.604513
Bus Stop	0.594735
Italian Restaurant	0.560006
Bagel Shop	0.401005
Ice Cream Shop	0.345016
Grocery Store	0.326751
Bakery	0.324681
Coffee Shop	0.303441
Pharmacy	0.303344
Sandwich Place	0.302170
Mexican Restaurant	0.299889
Chinese Restaurant	0.296871
Donut Shop	0.295130
Beach	0.285714
Bank	0.279552
Playground	0.214252
Park	0.212269
Supermarket	0.205637
Bar	0.197258

After summing up all the relative numbers of each categories, we can find out the top 20 most common categories in the high COVID case rate areas. The result shows that 70% of these popular venues are related to food. And here is bar plot to better show the relative amount of these venues.



It clearly shows that Pizza Place, Deli/Bodega, Bus Stop, and Italian Restaurant have much larger relative amount than other categories in these areas. The result implies that these 4 categories may have positive relationship with COVID case rates.



If we apply this guess to all 177 MODIFIED_ZCTAs, we can generate the scatter plots above. The x-axis represents the relative number of selected categories in each area, while the y-axis is the COVID case rate. It seems that there is a strong positive relationship between Pizza Place and COVID case rate. Deli/Bodega also shows a weaker relationship, while there is basically no relationship between Bus Stop or Italian Restaurant and case rate. This is probably because these two categories are selected from cluster label 3, and Bus Stop or Italian Restaurant are not features or common categories in other clusters.

Conclusions

From this preliminary study, we notice that most of the MODIFIED_ZCTAs with high COVID case rates are clustered into same groups (Cluster 3), implying that they share similar features. Closer examination of 14 areas labeled with cluster 3 in the top 20 highest COVID-case-rate areas shows that food related categories are quite common. 70% of the top 20 most common venues in these 14 areas are in food-related categories, such as Pizza Place, Deli/Bodega, Restaurant, Bagel Shop, Ice Cream Shop, Grocery Store, Bakery, and Coffee Shop. The result suggests that we need to take extra caution when we have to visit these kinds of places during the pandemic. Although this study is quite limited by the data, I still hope this little finding can serve as a caution for everyone, helping people to protect themselves, protect their families, and protect people around them from the virus.

References

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7205668/

[3] https://www1.nyc.gov/site/doh/covid/covid-19-data.page

[4] https://github.com/nychealth/coronavirus-data