Analysis of Covid-19 positive rate variations among different postcode areas in NYC

Introduction

Positive Covid rates vary greatly among different areas in New York City. By using demographic data together with Foursquare venue location data, we may get some clues to explain the variations.

This analysis is not designed for any commercial purpose, but to show public/government the possible explanations behind the pandemic.

Data Acquisition

In this section, I will explain all types of data I would consider in this study, and the resources to get them. Since I want to compare the positive rate variations in different areas, the data I need will be exhibited by small areas, like community neighborhood or zip code area. I choose to use zip code area data for this study.

Covid data

NYC Covid-19 data is essential as the final target to understand. Covid data can be accessed through the website of NYC Department of Health and Mental Hygiene. The department has a github account to update the data daily: https://github.com/nychealth/coronavirus-data. The data show covid positive rates in each postal code area.

Demographic data

Population density could possibly cause different virus spreading rates. Also, since different age groups have their different social activity habits and show quite different reaction to the pandemic, it's interesting to know if the age percentage has played a role to cause different covid positive rates. The information can be found on data.census.gov and was collected from American Community Survey 2018.

Geographic data

uszipcode is a python zip code database. Information like the area of the post code areas, latitude and longitude can be easily loaded from the library. The library also offers economic information like median household income, which can also be used in our analysis.

Venue location data

For the venue data, first, I am curious to know if more restaurants in an area is related to a higher covid rate.

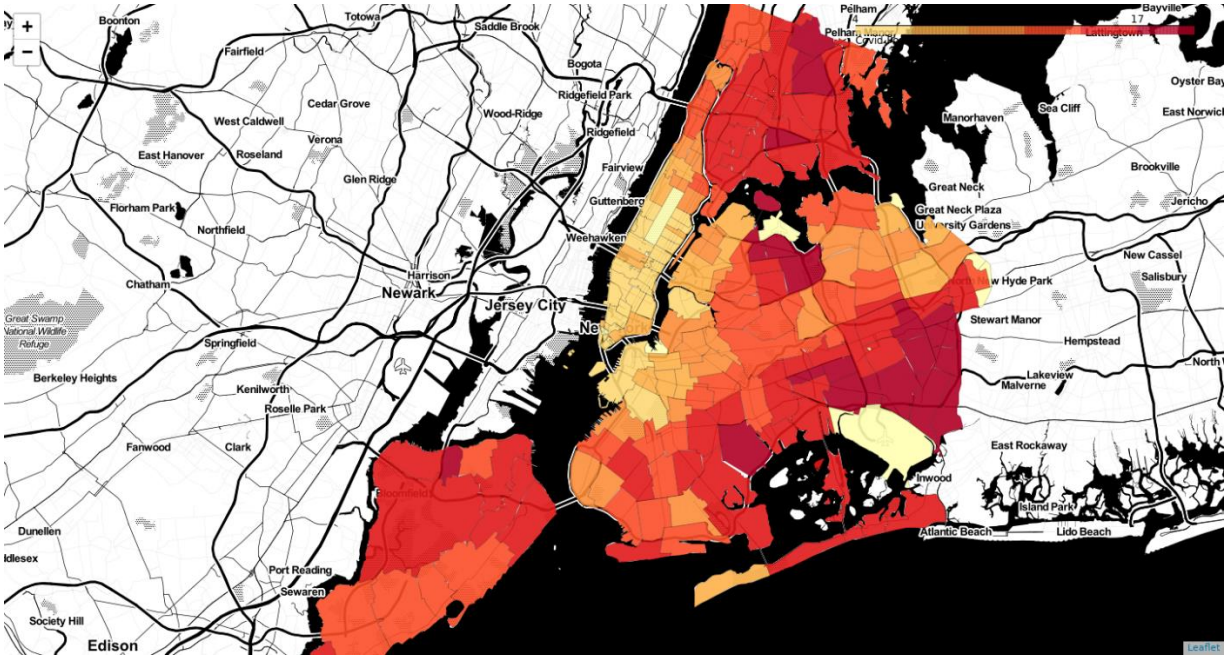
Second, different types and different quantities of venues can be potential reasons to cause higher or lower rates. With the venue information from Foursquare, we can cluster the venues based on zip code, and see if the cluster is related to different covid rates.

METHODOLOGY

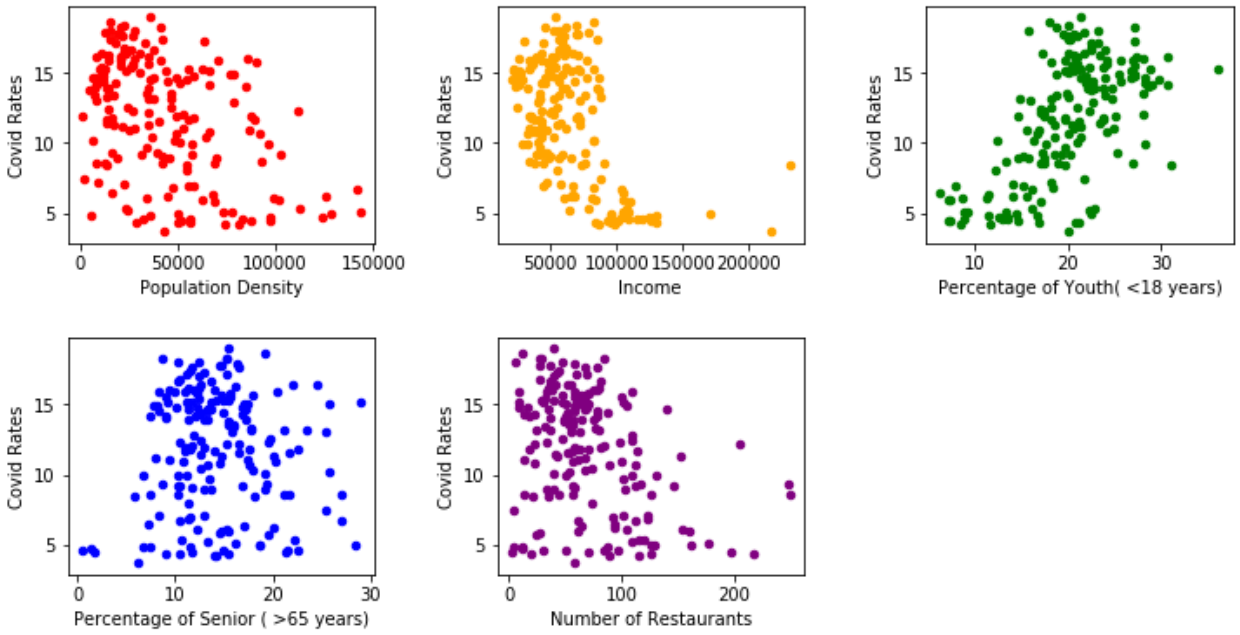
1. Plot covid rates on map with Folium and choropleth.
2. After importing data from different data sets, conduct some exploratory data analysis to see possible relationships between all types of factors and covid rates variation.
3. Use k-means method to cluster zip code areas based on venue types and check the covid rates distributions with histogram for each cluster.
4. Use word cloud to show the feature venues of different covid rate areas.

RESULTS

Covid positive rates vary a lot in different zip code area in NYC as shown below:



With some exploratory data analysis, we checked a list of potential factors that may affect positive covid rates.

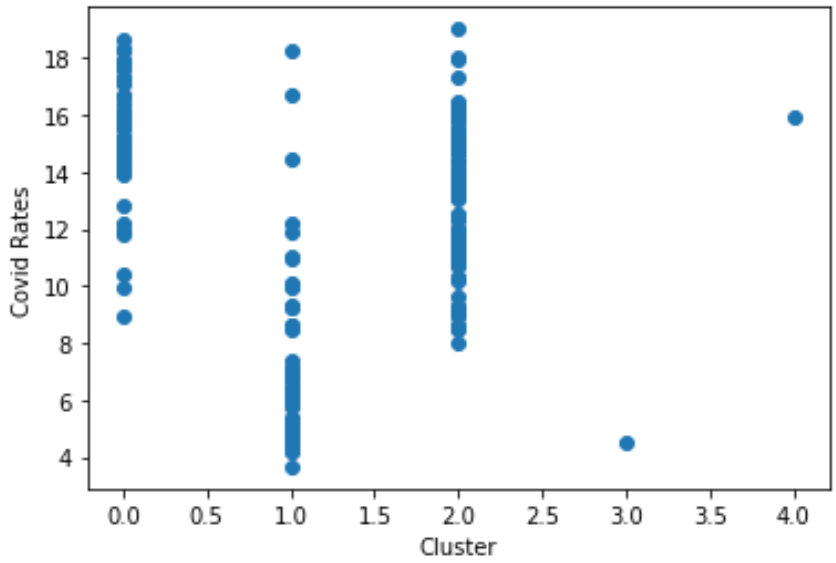


We all know that senior people are more vulnerable when facing covid-19 virus, however, the percentage of older people in the area has no direct correlation to covid rates. population density also show no direct relationship with covid rates.

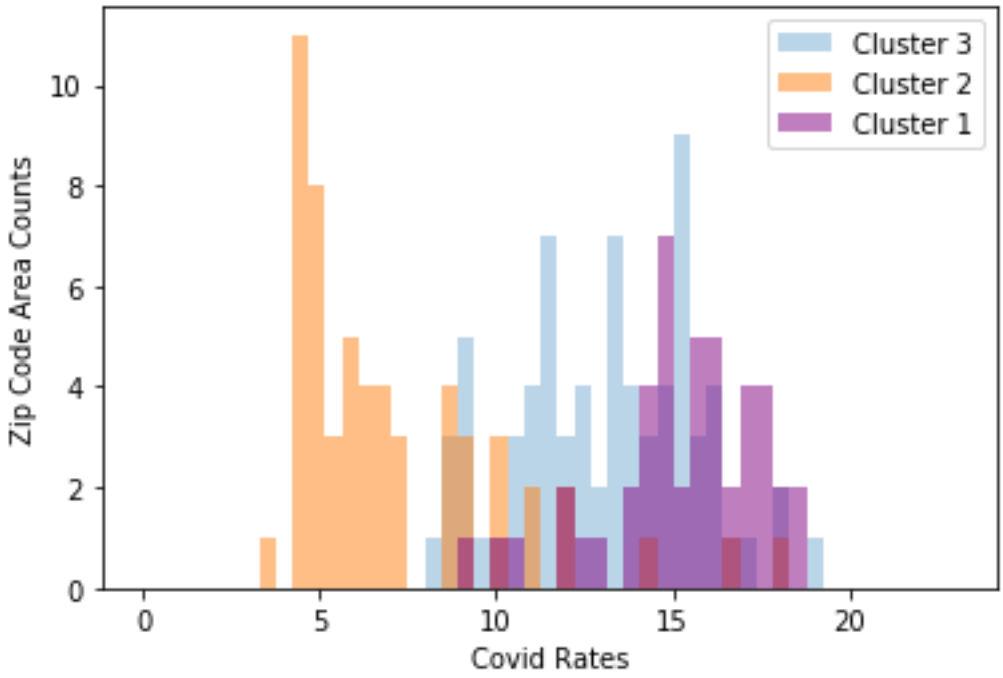
Higher median household income areas show much lower covid rates compare with low income areas. Covid Rates were found higher in areas with higher Youth percentage.

Surprisingly, restaurants density appears to be negatively correlated to covid rates. This implies a possible relationship between venues distribution and covid rates. Thus, we decide to dig deeper to explain the phenomenon.

For venue data, we adopted Foursquare to collect the information. K-means method was used to cluster the venues in 5 groups:



We can see different clusters show different covid rate trends. To better have a look at the distributions, we can plot the data in histograms:



Since cluster 4 and 5 has only 1 zipcode area, to simplify the figure, I ignored them in the histogram. We can see that in each cluster, covid rates are either mainly high (>10) or low (<10). Even for cluster 4 and 5, with only 1 zip code area in the cluster, they have clear high/low covid rate readings.

Then, we checked the feature venues for low/high covid rates areas:

	CovidCat	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	high	Pizza Place	Donut Shop	Deli / Bodega	Pharmacy	Chinese Restaurant	Sandwich Place	Bank	Bakery	Grocery Store	Italian Restaurant
1	low	Coffee Shop	Bar	Italian Restaurant	Pizza Place	Bakery	Café	American Restaurant	Mexican Restaurant	Park	Cocktail Bar

And plotted the word cloud for each category:

Popular Venues in Low Covid Rate Areas

Popular Venues in High Covid Rate Areas

Discussion

From the exploratory data analysis, the data implied that the **population density** has no direct relationship with covid rate variations. However, this value can only reflect people density based on residential area. Business area and transportation hubs can also have high people density every day, but without residential buildings, the registered population could be low.

Percentage of senior has no contribution to covid rates, while **youth ratio** is greatly correlated with covid rates. I think the main difference between the two groups is their social activity modes. Young people have a higher possibility to go to a bar, a restaurant or a movie theater with a group of friends, while senior people are most likely to have relative quiet activities with their families in a relative private spot. More vulnerable to virus does not make senior people more easily to get the disease, but just after getting the virus, their possibility to have severe symptoms is higher. Also, young people, as students, also need to go to school. This situation also put them under a higher risk to get infected.

Furthermore, higher median household **income** communities show lower positive covid rates. This could be due to more options that rich people have when facing this pandemic: they are willing to pay more to get all types of good-quality self-protection equipment; they can choose not to be exposed to public transportations; or they can stay in their suburban houses instead of small apartments in the city.

Surprisingly, higher **density of the restaurant** in an area gives a lower covid rate. Maybe people don't usually eat in their neighborhood, or there could be other explanations. To dig deeper into the venue information, I clustered the zip code areas based on their venue types and found that areas in the same

cluster show the similar covid rates. This is not surprising to me, because similar types venue combination implies similar type of activities in the area. In the last step, word cloud was used to present the most featured venues for low covid rate areas and high covid rate areas. This doesn't necessarily mean venues types in low covid rate plot are safer than others. It is just to see the features of the areas.

Conclusion

This study briefly analyzed some possible covid influence factors in the zip code areas in New York City. Possible explanations were given for each individual observation.

Higher or lower covid rate in an area does not mean how good or bad the area is. The rate is just to send a reminder for people to still stay alert and stay safe. We can win this battle together NYC!