COVID-19 in New York neighbourhoods – Nearby venues and positivity rate

Alok Gupta

November 29, 2020

1 Introduction

1.1 Background

COVID-19 is caused by a coronavirus called SARS-CoV-2. The illness started to spread in early 2020 and is not yet in control as of date. There has been a re-emergence of Covid-19 in New York area this fall after an initial wave of high infection rate during the summer. With so many unknowns about the illness, there is curiosity about common features of persons or the environments they live in, which aid or prohibit spread of the virus. For example, data analysis indicates that older adults and people who have severe underlying medical conditions are at higher risk develop serious complications from COVID-19 illness. Similarly, it can help to understand if neighbourhoods where the virus has been spreading at a fast rate, have venues or locations which are affecting the spread.

1.2 Problem

This project attempts to relate venues (or their categories) in various New York neighbourhoods with COVID-19 percent positivity rate during November 15-21, 2020. It can help to know if neighbourhoods with certain categories of venues show high or low COVID-19 percent positivity rate or if COVID-19 percent positivity rate is independent of types of venues in a neighbourhood.

1.3 Interest

The information can be used by local authorities to implement stricter measures for controlling the spread of infection in venue categories with high new positive rate. It can be used by neighbourhood residents to avoid visits or take more precaution while visiting common venues with COVID-19 percent positivity rate.

1.4 Scope

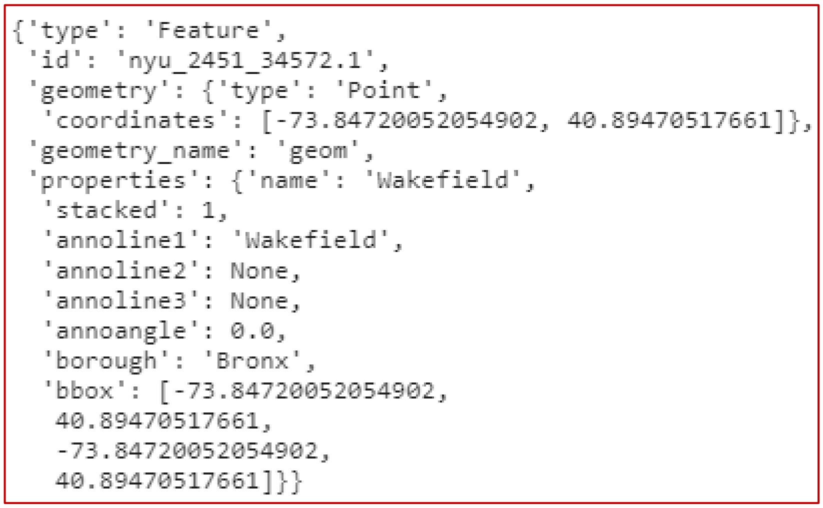
The scope of this project was restricted to data obtained during November 15-21, 2020 time period.

2 Data sources and data cleaning

Data from various sources was used to obtain location information about boroughs and neighbourhoods, COVID-19 testing data, and venues around neighbourhoods. Information about venue categories was analysed manually to obtain broader grouping such as retail or restaurant, indoor or outdoor, etc. The following subsections provide additional details about data sources.

2.1 New York boroughs and neighbourhoods

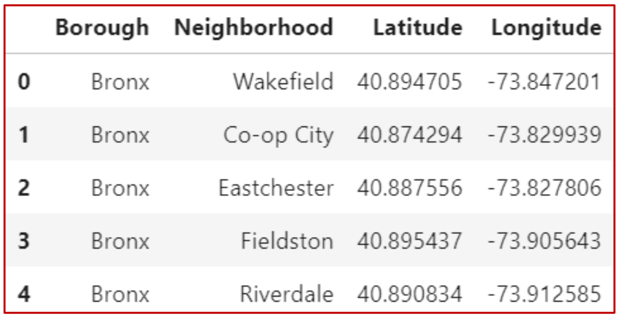
New York has a total of 5 boroughs and 306 neighbourhoods. Latitude and longitude coordinates of the neighbourhoods in the 5 boroughs is needed. This data is available through this link: https://geo.nyu.edu/catalog/nyu_2451_34572. A file downloaded to the server was used. A sample of this data available from the website is shown below (for one neighbourhood).



```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```

Figure 1 Raw data for New York neighbourhood

Relevant information from the raw data was used to construct a data table shown in Figure 2.

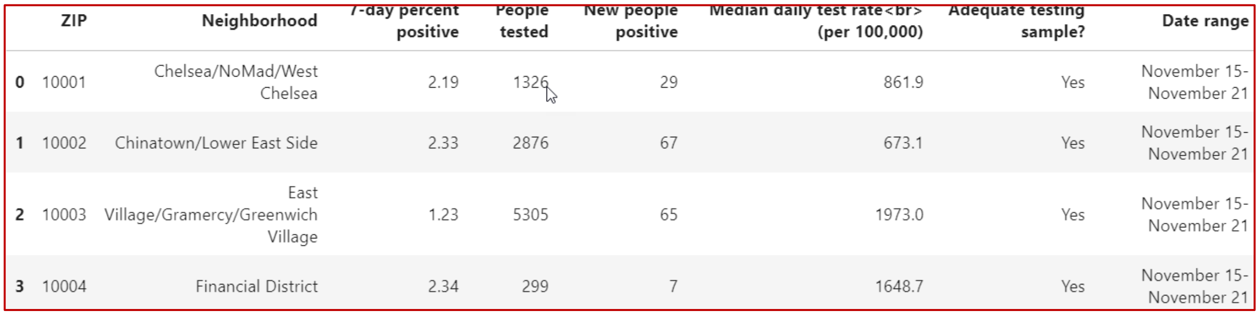


	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Figure 2 New York neighbourhood locations

2.2 COVID-19 test data

The percent of people who tested COVID-19 positive in each of the ZIP codes in New York is available from the website: https://www1.nyc.gov/site/doh/covid/covid-19-data.page#epicurve. The data is for November 15-21, 2020. It has been downloaded and placed in GitHub (https://github.com/alokanant/Coursera_Capstone/blob/master/data/Covid-data-8UcZR.csv) for use. The data also shows the rate of people tested during the most recent seven days. A neighbourhood is considered to have adequate testing when at least 260 residents per 100,000 have been tested in the past week. A sample of raw COVID-19 data is shown below.



	ZIP	Neighborhood	7-day percent positive	People tested	New people positive	Median daily test rate (per 100,000)	Adequate testing sample?	Date range
0	10001	Chelsea/NoMad/West Chelsea	2.19	1326	29	861.9	Yes	November 15-November 21
1	10002	Chinatown/Lower East Side	2.33	2876	67	673.1	Yes	November 15-November 21
2	10003	East Village/Gramercy/Greenwich Village	1.23	5305	65	1973.0	Yes	November 15-November 21
3	10004	Financial District	2.34	299	7	1648.7	Yes	November 15-November 21

Figure 3 COVID-19 raw test data

Raw COVID-19 data was cleaned and transformed primarily to:

a) Rename column headers.

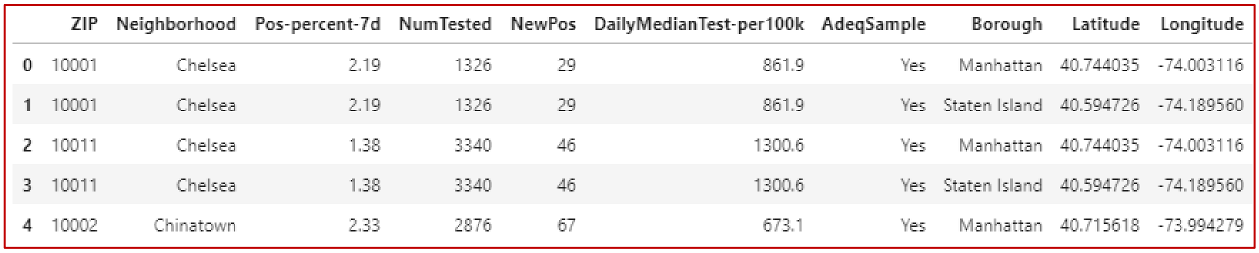
b) Split rows to convert multiple neighbourhoods in a row to one neighbourhood per row.

c) Add borough, latitude, and longitude information.

d) Remove *Date range* column as it has the same value for all records.

e) Remove rows that had inadequate testing samples.

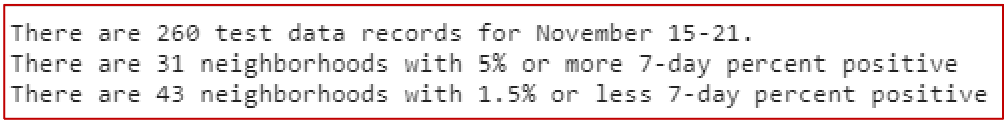
The transformed data is shown below.



	ZIP	Neighborhood	Pos-percent-7d	NumTested	NewPos	DailyMedianTest-per100k	AdeqSample	Borough	Latitude	Longitude
0	10001	Chelsea	2.19	1326	29	861.9	Yes	Manhattan	40.744035	-74.003116
1	10001	Chelsea	2.19	1326	29	861.9	Yes	Staten Island	40.594726	-74.189560
2	10011	Chelsea	1.38	3340	46	1300.6	Yes	Manhattan	40.744035	-74.003116
3	10011	Chelsea	1.38	3340	46	1300.6	Yes	Staten Island	40.594726	-74.189560
4	10002	Chinatown	2.33	2876	67	673.1	Yes	Manhattan	40.715618	-73.994279

Figure 4 Transformed COVID-19 data

Highest and lowest 7-day positive percentage (*Pos-percent-7d*) data was investigated further. Information from John Hopkins website (https://www.jhsph.edu/covid-19/articles/covid-19-testing-understanding-the-percent-positive.html) provides the following definition of percent positive – *the percentage of all coronavirus tests performed that are actually positive, or: (positive tests)/(total tests) x 100%*. The percent positive (sometimes called the "percent positive rate" or "positivity rate") helps public health officials understand current level of transmission, is enough testing being done, etc.
It was found that there were significant number of neighbourhoods with 5% or more and 1.5% or less positivity rate (see Figure 5). These numbers were selected after some trials and are referred as *high positivity rate* and *low positivity rate* in subsequent sections.



```
There are 260 test data records for November 15-21.
There are 31 neighborhoods with 5% or more 7-day percent positive
There are 43 neighborhoods with 1.5% or less 7-day percent positive
```

Figure 5 Analysis of records in COVID-19 data

2.3 Foursquare API

Foursquare API (`https://api.foursquare.com`) was used to search for specific type of venues, to explore a particular venue, to explore a geographical location, and to get trending venues (if required) around locations. More information about the data pulled using Foursquare API is available in the following sections.

3 Exploratory data analysis

3.1 Neighbourhoods with high and low positivity rates

In order to visualize Neighbourhoods with high positivity rate and low positivity rate the locations were displayed on a map. This indicates that the northern and southern neighbourhoods have a greater spread of COVID-19 compared to central neighbourhoods, in New York.

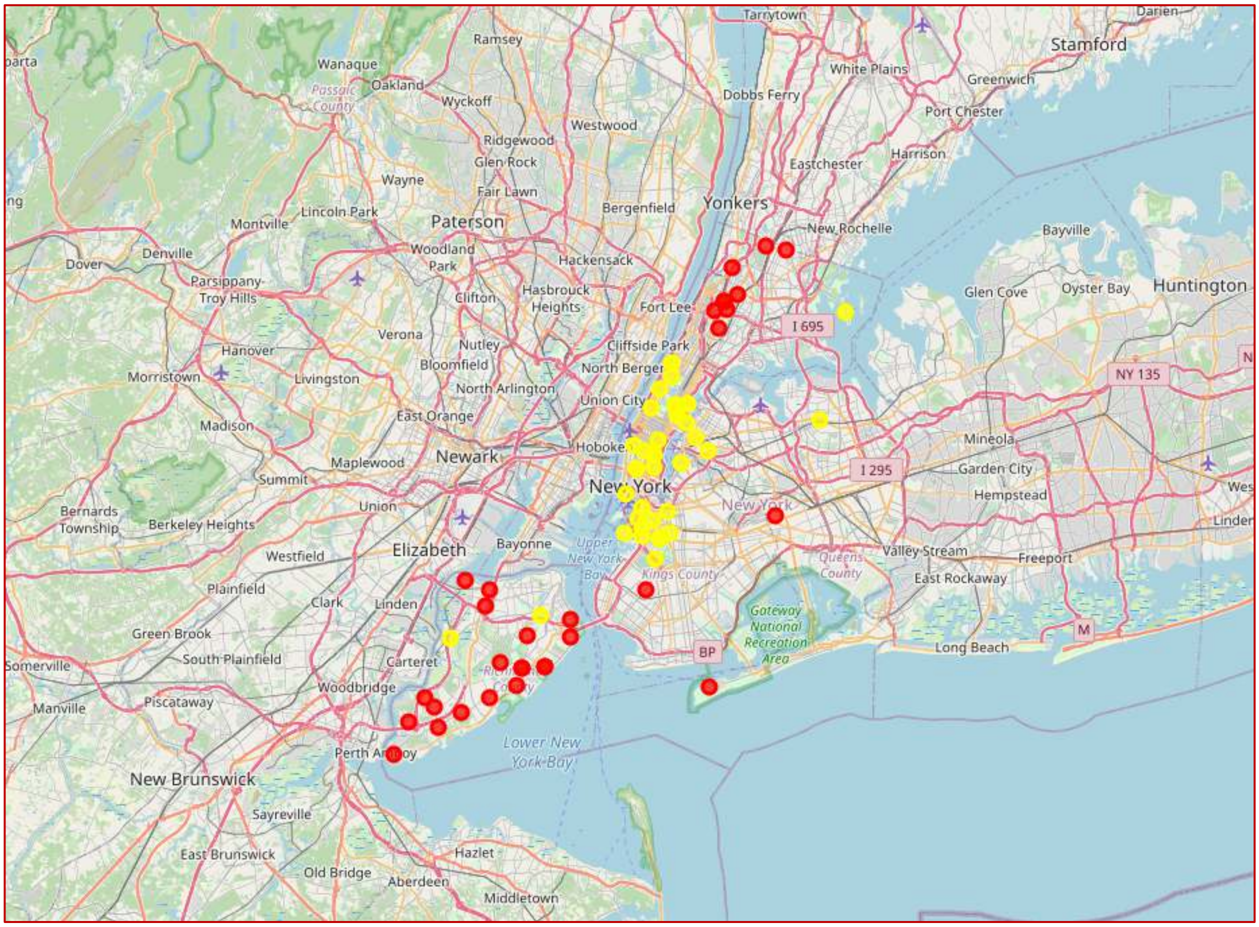
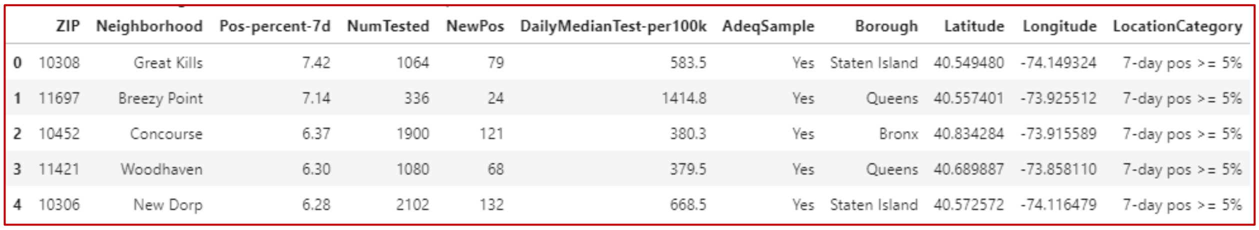


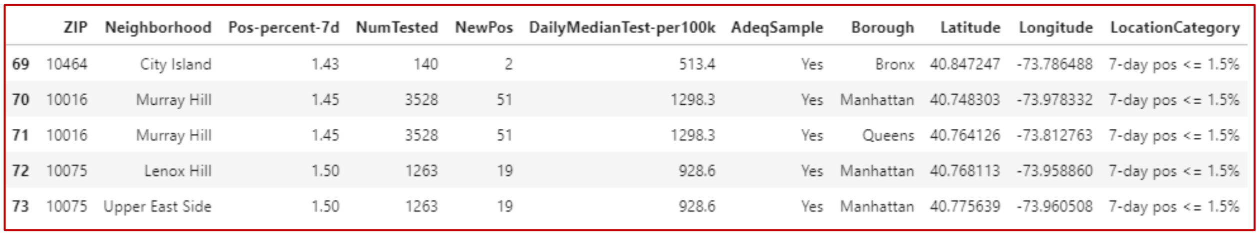
Figure 6 Locations with high (red) positivity rate & low (yellow) positivity rate

Figure 7 and Figure 8 show a sample of data for neighbourhoods with high positivity rate and low positivity rate, respectively.



	ZIP	Neighborhood	Pos-percent-7d	NumTested	NewPos	DailyMedianTest-per100k	AdeqSample	Borough	Latitude	Longitude	LocationCategory
0	10308	Great Kills	7.42	1064	79	583.5	Yes	Staten Island	40.549480	-74.149324	7-day pos >= 5%
1	11697	Breezy Point	7.14	336	24	1414.8	Yes	Queens	40.557401	-73.925512	7-day pos >= 5%
2	10452	Concourse	6.37	1900	121	380.3	Yes	Bronx	40.834284	-73.915589	7-day pos >= 5%
3	11421	Woodhaven	6.30	1080	68	379.5	Yes	Queens	40.689887	-73.858110	7-day pos >= 5%
4	10306	New Dorp	6.28	2102	132	668.5	Yes	Staten Island	40.572572	-74.116479	7-day pos >= 5%

Figure 7 High positivity rate neighbourhoods

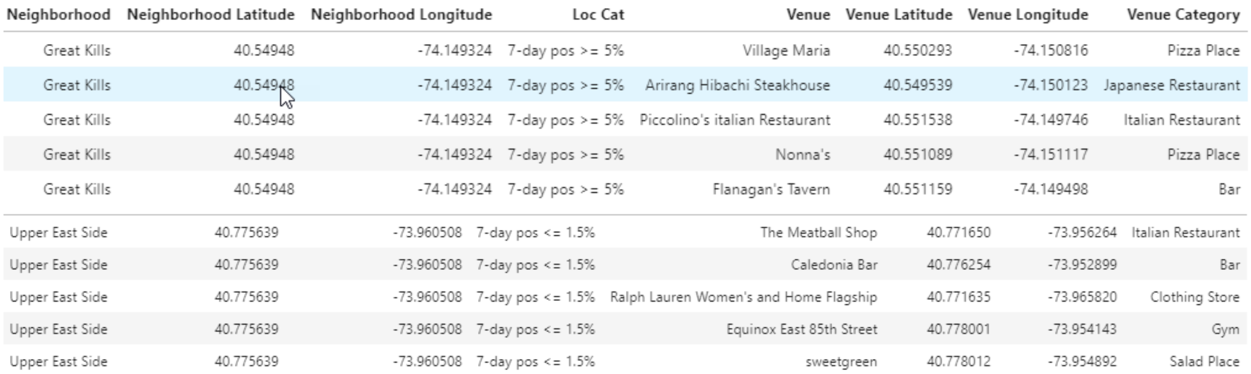


	ZIP	Neighborhood	Pos-percent-7d	NumTested	NewPos	DailyMedianTest-per100k	AdeqSample	Borough	Latitude	Longitude	LocationCategory
69	10464	City Island	1.43	140	2	513.4	Yes	Bronx	40.847247	-73.786488	7-day pos <= 1.5%
70	10016	Murray Hill	1.45	3528	51	1298.3	Yes	Manhattan	40.748303	-73.978332	7-day pos <= 1.5%
71	10016	Murray Hill	1.45	3528	51	1298.3	Yes	Queens	40.764126	-73.812763	7-day pos <= 1.5%
72	10075	Lenox Hill	1.50	1263	19	928.6	Yes	Manhattan	40.768113	-73.958860	7-day pos <= 1.5%
73	10075	Upper East Side	1.50	1263	19	928.6	Yes	Manhattan	40.775639	-73.960508	7-day pos <= 1.5%

Figure 8 Low positivity rate neighbourhoods

3.2 Venues near high/low positivity rate neighbourhoods

Foursquare API was used to explore the neighbourhoods determined in Section 3. Neighbourhoods were explored to find venues within 1000 m (1 km) near the neighbourhoods. There were around 5500 venues near the 74 neighbourhoods that had high/low positivity rates. Of these, around 4000 were near low positivity rate neighbourhoods and around 1500 were near high positivity rate neighbourhoods. A sample of the data is shown in Figure 9.

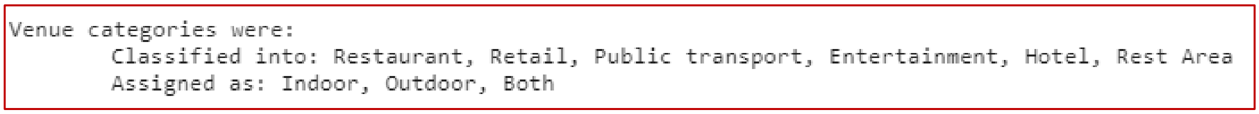


Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Loc Cat	Venue	Venue Latitude	Venue Longitude	Venue Category
Great Kills	40.54948	-74.149324	7-day pos >= 5%	Village Maria	40.550293	-74.150816	Pizza Place
Great Kills	40.54948	-74.149324	7-day pos >= 5%	Arirang Hibachi Steakhouse	40.549539	-74.150123	Japanese Restaurant
Great Kills	40.54948	-74.149324	7-day pos >= 5%	Piccolino's italian Restaurant	40.551538	-74.149746	Italian Restaurant
Great Kills	40.54948	-74.149324	7-day pos >= 5%	Nonna's	40.551089	-74.151117	Pizza Place
Great Kills	40.54948	-74.149324	7-day pos >= 5%	Flanagan's Tavern	40.551159	-74.149498	Bar
Upper East Side	40.775639	-73.960508	7-day pos <= 1.5%	The Meatball Shop	40.771650	-73.956264	Italian Restaurant
Upper East Side	40.775639	-73.960508	7-day pos <= 1.5%	Caledonia Bar	40.776254	-73.952899	Bar
Upper East Side	40.775639	-73.960508	7-day pos <= 1.5%	Ralph Lauren Women's and Home Flagship	40.771635	-73.965820	Clothing Store
Upper East Side	40.775639	-73.960508	7-day pos <= 1.5%	Equinox East 85th Street	40.778001	-73.954143	Gym
Upper East Side	40.775639	-73.960508	7-day pos <= 1.5%	sweetgreen	40.778012	-73.954892	Salad Place

Figure 9 Information about venues near neighbourhood

3.2.1 Venue Category

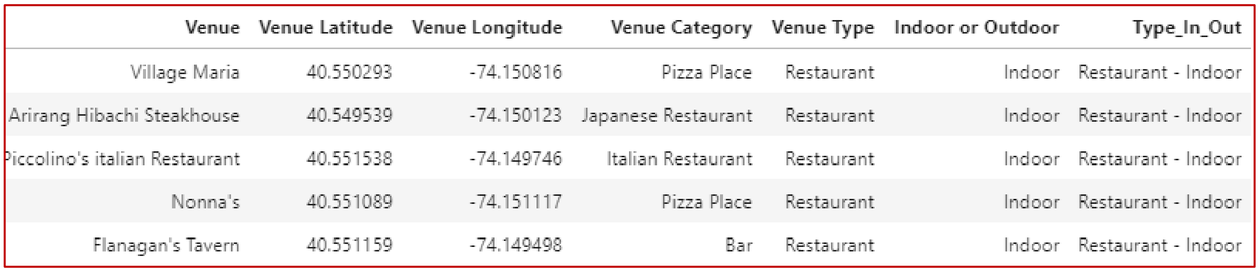
It was found that there are 355 unique venue categories provide by the Foursquare API. These unique categories were extracted and manually analysed in an Excel file to determine a broader grouping of the venue categories. The analysis used broader grouping parameters as shown in Figure 10.



```
Venue categories were:
        Classified into: Restaurant, Retail, Public transport, Entertainment, Hotel, Rest Area
        Assigned as: Indoor, Outdoor, Both
```

Figure 10 Broader grouping of venue categories

Information from Foursquare API was merged with the broader grouping data to obtain supplemented venue information, a sample of which is depicted in **Error! Reference source not found.**.



Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Type	Indoor or Outdoor	Type_In_Out
Village Maria	40.550293	-74.150816	Pizza Place	Restaurant	Indoor	Restaurant - Indoor
Arirang Hibachi Steakhouse	40.549539	-74.150123	Japanese Restaurant	Restaurant	Indoor	Restaurant - Indoor
Piccolino's italian Restaurant	40.551538	-74.149746	Italian Restaurant	Restaurant	Indoor	Restaurant - Indoor
Nonna's	40.551089	-74.151117	Pizza Place	Restaurant	Indoor	Restaurant - Indoor
Flanagan's Tavern	40.551159	-74.149498	Bar	Restaurant	Indoor	Restaurant - Indoor

Figure 11 Supplemented venue information

3.3 Venues near neighbourhoods with high and low positivity rate

Venues near neighbourhoods with high and low positivity rate were analysed using one hot encoding method on venue type with indoor and outdoor column together (column *Type_In_Out*). After encoding data was grouped by neighbourhoods to find the mean for each venue type in a neighbourhood. Common venues in each neighbourhood were determined by sorting the means of venue type for each column.

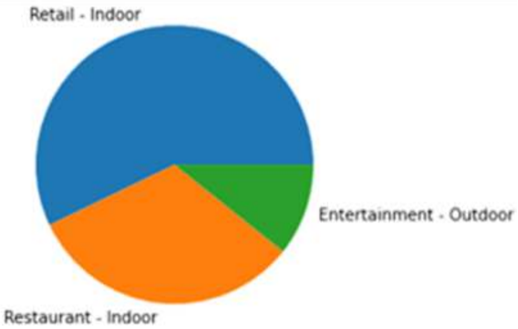
3.4 Comparison of venues near neighbourhoods

The following subsections provide a comparison of venues in high positivity rate and low positivity rate neighbourhoods, up to three most common venues. Data for up to eight most common venues is available in the python notebook.

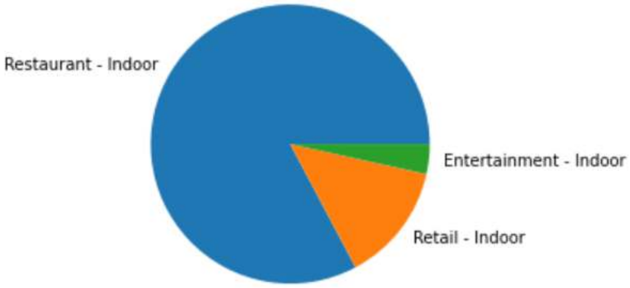
3.4.1 First most common venue

Majority of first most common venues within 1000 m of both high and low positivity rate neighbourhoods are retail stores or restaurants. Indoor restaurant venues are in high numbers in low positivity rate neighbourhoods. It is likely they had low occupancy. High positivity rate in neighbourhoods with large number of indoor retail venues can be due to people shopping more.

High positivity rate neighbourhood

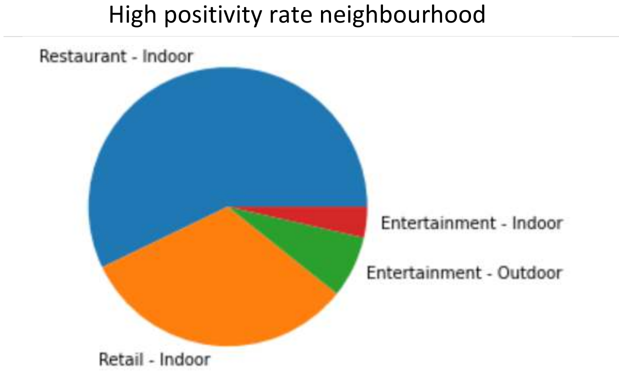


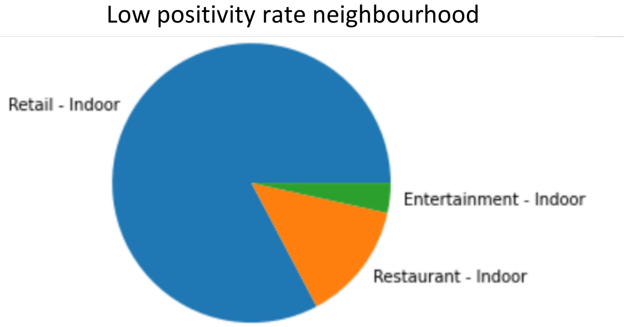
Low positivity rate neighbourhood



3.4.2 Second most common venue

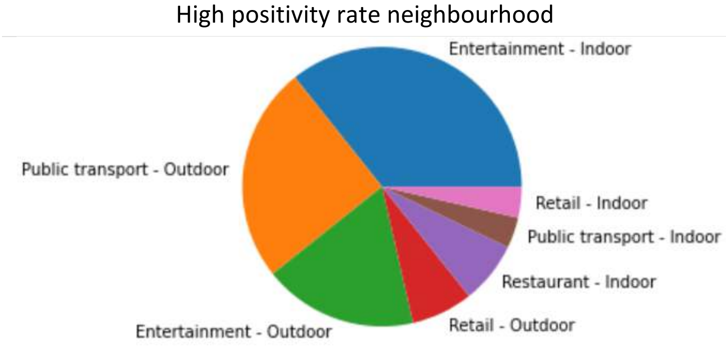
Majority of second most common venues within 1000 m of both high and low positivity rate neighbourhoods are retail stores or restaurants. Indoor restaurant venues are in high numbers in high positivity rate neighbourhoods. Low positivity rate in neighbourhoods have large number of indoor retail venues.

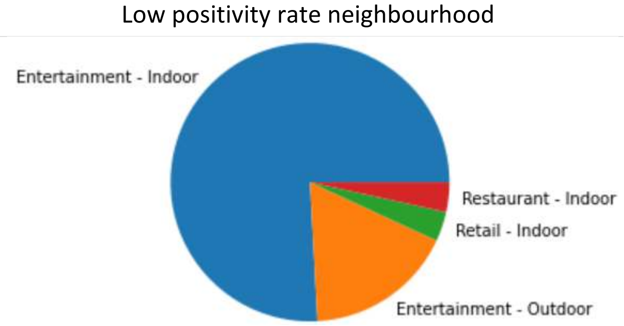

High positivity rate neighbourhood


Low positivity rate neighbourhood

3.4.3 Third most common venue

Majority of third most common venues within 1000 m of both high and low positivity rate neighbourhoods are entertainment stores or public transport. One outstanding feature of high positivity rate neighbourhoods is abundance of venues related to public transport. Public transport facilitates more mobility and could be a reason for high positivity rate.


High positivity rate neighbourhood

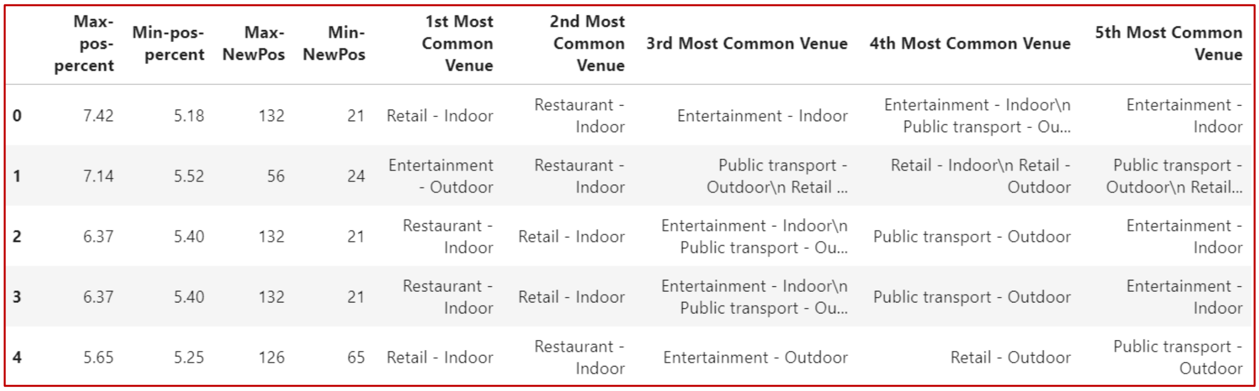

Low positivity rate neighbourhood

4 K-means clustering of venues near neighbourhoods

Five clusters of neighbourhoods were created using the KMeans clustering algorithm for observing distinct features in venues near high and low positivity rate venues.

4.1 High positivity rate venues

High positivity rate venue cluster parameters are shown in Figure 12 and the clusters are geographically depicted in Figure 13. One outstanding feature is the role played by public transport in all 5 high positivity rate venues. Except for the second cluster (index 0) all other clusters has number of new positive indications above 120.



	Max-pos-percent	Min-pos-percent	Max-NewPos	Min-NewPos	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	7.42	5.18	132	21	Retail - Indoor	Restaurant - Indoor	Entertainment - Indoor	Entertainment - Indoor\n Public transport - Ou...	Entertainment - Indoor
1	7.14	5.52	56	24	Entertainment - Outdoor	Restaurant - Indoor	Public transport - Outdoor\n Retail ...	Retail - Indoor\n Retail - Outdoor	Public transport - Outdoor\n Retail...
2	6.37	5.40	132	21	Restaurant - Indoor	Retail - Indoor	Entertainment - Indoor\n Public transport - Ou...	Public transport - Outdoor	Entertainment - Indoor
3	6.37	5.40	132	21	Restaurant - Indoor	Retail - Indoor	Entertainment - Indoor\n Public transport - Ou...	Public transport - Outdoor	Entertainment - Indoor
4	5.65	5.25	126	65	Retail - Indoor	Restaurant - Indoor	Entertainment - Outdoor	Retail - Outdoor	Public transport - Outdoor

Figure 12 Parameters of high positivity rate venue clusters

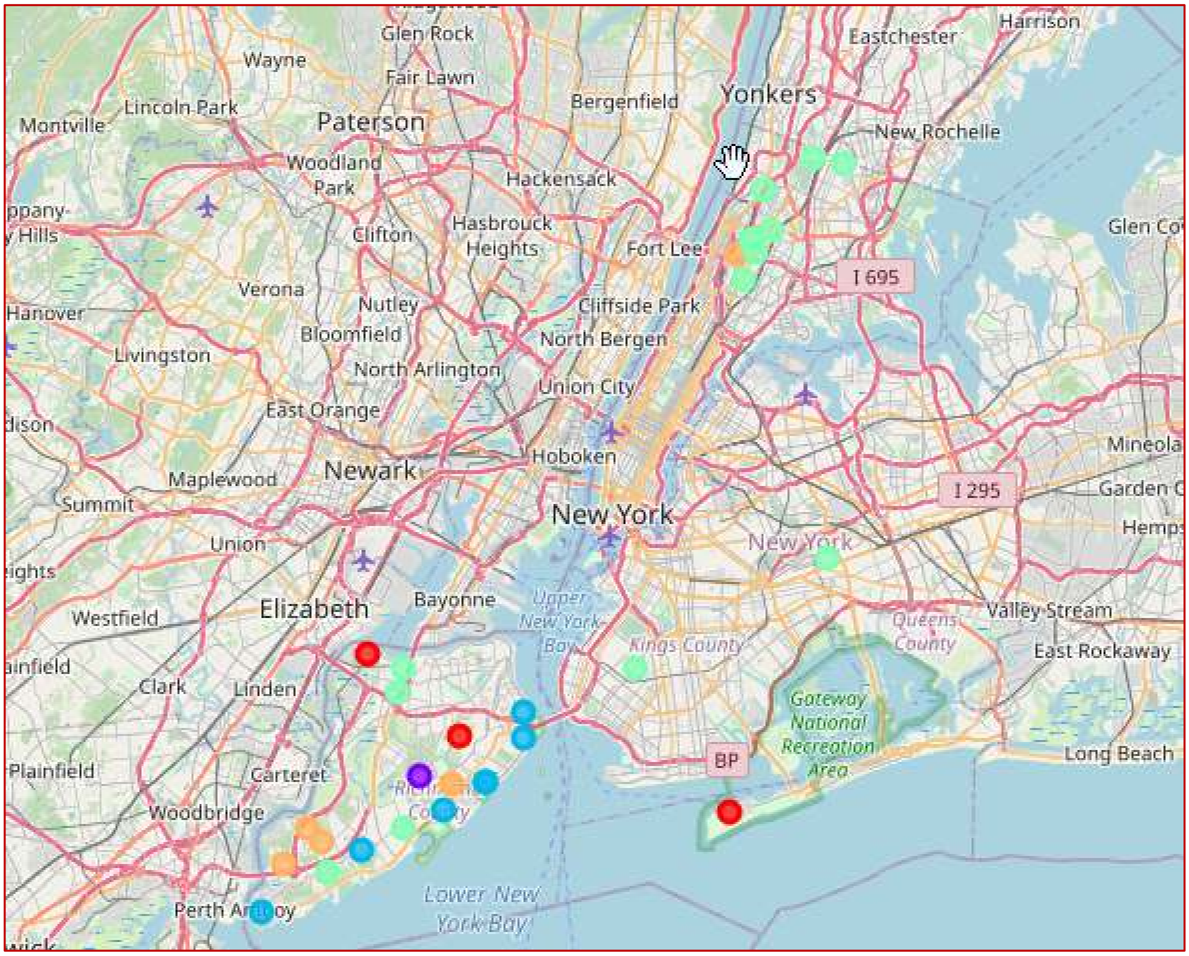
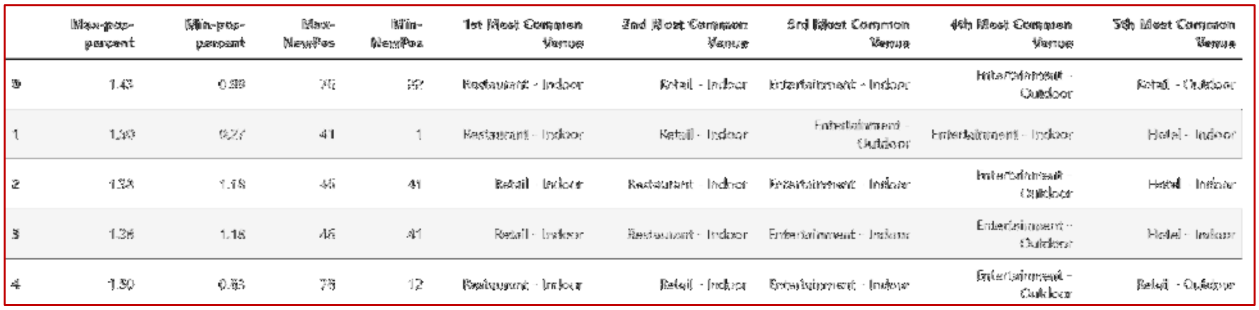


Figure 13 Cluster of neighbourhoods with high positivity rate

4.2 Low positivity rate venues

Low positivity rate venue cluster parameters are shown in Figure 14 and the clusters are geographically depicted in Figure 15. Although first and second most common venues in both high- and low-positivity rate clusters are retail and restaurant venues, the third most common venues are different. The third most common venues in high positivity rate neighbourhoods includes public transport locations for some clusters. In low positivity rate clusters, the third and fourth most common venues are Entertainment venues.



	Max-pos-percent	Min-pos-percent	Max-NewPos	Min-NewPos	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	1.43	0.89	76	39	Restaurant - Indoor	Retail - Indoor	Entertainment - Indoor	Entertainment - Outdoor	Retail - Outdoor
1	1.30	0.27	41	1	Restaurant - Indoor	Retail - Indoor	Entertainment - Outdoor	Entertainment - Indoor	Hotel - Indoor
2	1.33	1.18	45	41	Retail - Indoor	Restaurant - Indoor	Entertainment - Indoor	Entertainment - Outdoor	Hotel - Indoor
3	1.36	1.18	46	41	Retail - Indoor	Restaurant - Indoor	Entertainment - Indoor	Entertainment - Outdoor	Hotel - Indoor
4	1.30	0.63	75	12	Restaurant - Indoor	Retail - Indoor	Entertainment - Indoor	Entertainment - Outdoor	Retail - Outdoor

Figure 14 Parameters of low positivity rate venue clusters

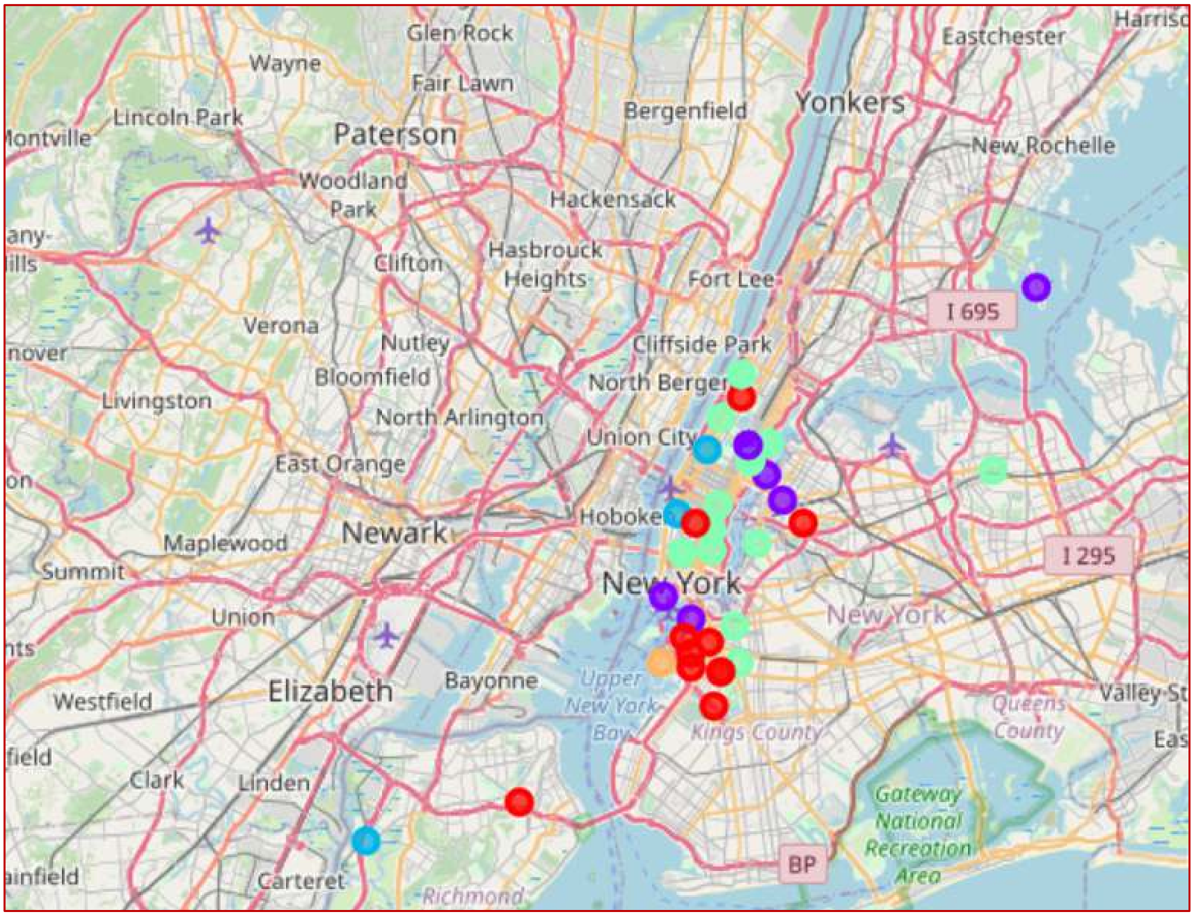


Figure 15 Cluster of neighbourhoods with low positivity rate

5 Conclusion

Relationship between venues (or their categories) in various New York neighbourhoods with COVID-19 percent positivity rate during November 15-21, 2020 indicates the following:

a) Northern and southern neighbourhoods have a greater spread of COVID-19 compared to central neighbourhoods, in New York.

b) There were around 5500 venues near the 74 neighbourhoods that had high/low positivity rates. Of these, around 4000 were near low positivity rate neighbourhoods and around 1500 were near high positivity rate neighbourhoods.

c) Majority of first and second most common venues within 1000 m of both high and low positivity rate neighbourhoods are retail stores or restaurants.

d) One outstanding feature of high positivity rate neighbourhoods is abundance of venues related to public transport. Public transport facilitates more mobility and could be a reason for high positivity rate.

e) Public transport venues were part of all 5 high positivity rate venues.

f) Although first and second most common venues in both high- and low-positivity rate clusters are retail and restaurant venues, the third most common venues are different. The third most common venues in high positivity rate neighbourhoods includes public transport locations for some clusters. In low positivity rate clusters, the third and fourth most common venues are Entertainment venues.

6 Future directions

The analysis can be extended to Covid-19 positive numbers from additional weeks (including those in summer) and/or neighbourhoods in additional US/global cities. Including information about number of persons visiting the venues, opening closing hours, trending locations at various times can be added to provide additional valuable information about spread of COVID-19.