2. Data Acquisition and Cleaning

2.1 Data Acquisition:

The Data Acquired for this project is a combination of a few sources, the most used of them being from the nychealth's github account- https://github.com/nychealth/coronavirus-data/blob/master/data-by-modzcta.csv

2.2 Data Description:

--From nychealth-github

Case reporting

NYC COVID-19 data include people who live in NYC or who live in another country but are being treated in NYC. The data do not include people who live outside of NYC but in the United States.

Rates vs. Cases

The Health Department is reporting rates of cases, hospitalizations, and deaths in addition to counts. We report rates to give clear comparisons between different groups — such as borough, sex, or age — with differently sized populations. For example, we may report that the rate of confirmed COVID-19 cases is 100 per 100,000 population in NYC. That means for every 100,000 people living in NYC, there are 100 people diagnosed with COVID-19.

Changes to Reported Data

We update data for earlier dates after we resolve testing and reporting delays. Reported data reflect what we know at the time of the report, not what occurred in real time. For example, we may find that a person who was originally reported to live in NYC no longer does. This person would be removed from our dataset after their address is updated, and our case count would decrease by one.

data-by-modzcta.csv

This file contains data by modified ZIP code tabulation areas (ZCTA). This unit of geography is similar to ZIP codes but combines census blocks with smaller populations to allow more stable estimates of population size for rate calculation. Please see description of modified ZCTAs in the technical notes section (Geography: Zip codes and ZCTAs).

This file contains the following cumulative indicators by modified ZCTA:

- Count of confirmed cases
- Rate of confirmed cases per 100,000 people by ZCTA
- Population denominators for ZCTAs derived from intercensal estimates by the Bureau of Epidemiology Services (see “Rates per 100,000 people” for more details)
- Count of confirmed deaths
- Rate of confirmed deaths per 100,000 people by ZCTA

- Percentage of people ever tested for COVID-19 who tested positive

This file includes the corresponding neighborhood and borough for each modified ZCTA.

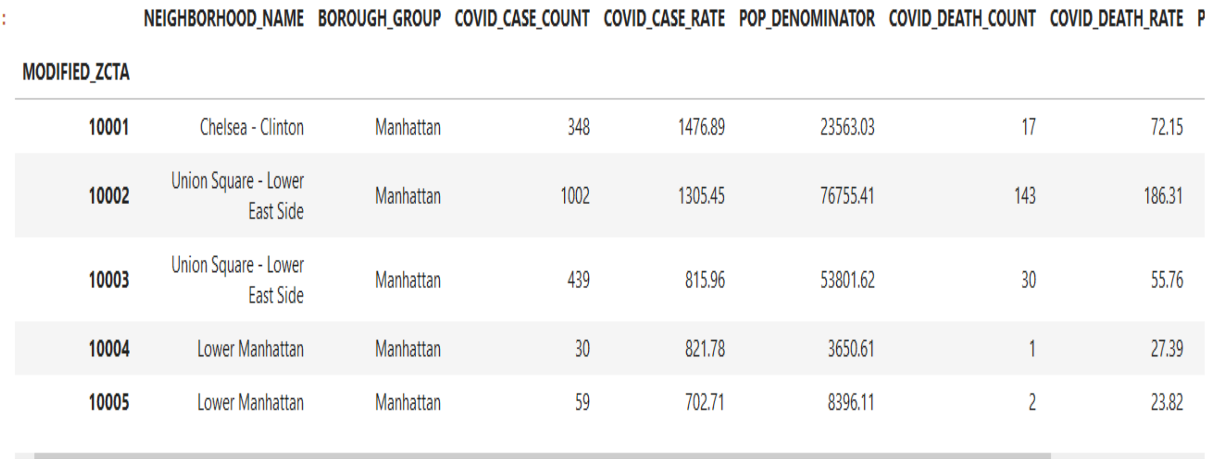
- Modified ZCTA
- Neighborhood name
- Borough name

Neighborhood names represent the 42 NYC United Hospital Fund (UHF) neighborhood. All cases are assigned to a UHF neighborhood based on ZCTA. Borough names are assigned according to the UHF neighborhood.

Note that sum of counts in this file may not match values in Citywide tables because of records with missing geographic information. This file does not currently contain information on probable deaths.

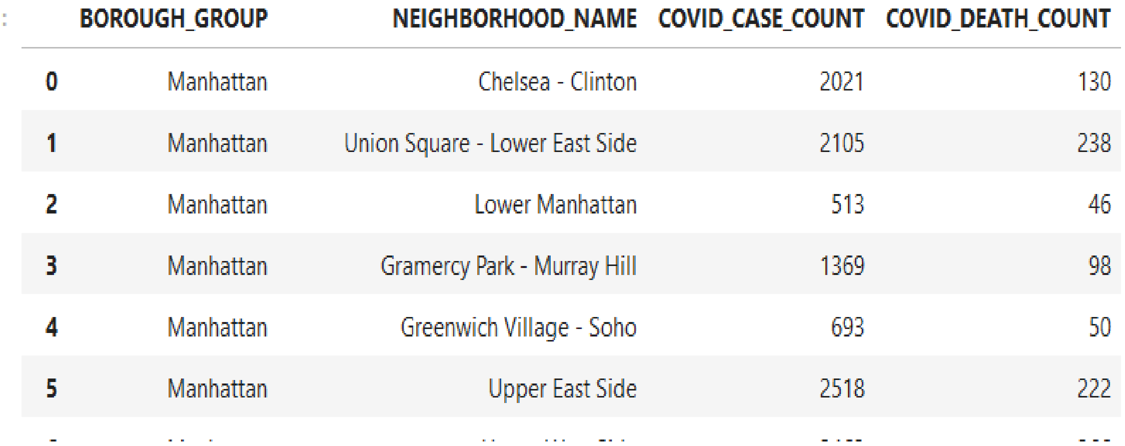
2.3 Data Cleaning:

Since the data acquired is in its raw form, it is required to clean it.



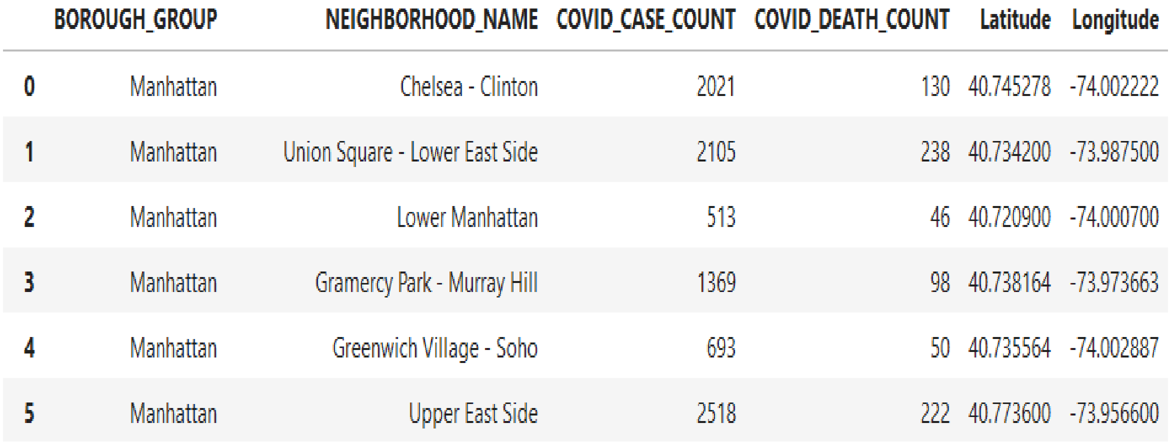
	NEIGHBORHOOD_NAME	BOROUGH_GROUP	COVID_CASE_COUNT	COVID_CASE_RATE	POP_DENOMINATOR	COVID_DEATH_COUNT	COVID_DEATH_RATE	P
MODIFIED_ZCTA								
10001	Chelsea - Clinton	Manhattan	348	1476.89	23563.03	17	72.15	
10002	Union Square - Lower East Side	Manhattan	1002	1305.45	76755.41	143	186.31	
10003	Union Square - Lower East Side	Manhattan	439	815.96	53801.62	30	55.76	
10004	Lower Manhattan	Manhattan	30	821.78	3650.61	1	27.39	
10005	Lower Manhattan	Manhattan	59	702.71	8396.11	2	23.82	

We require the data per neighborhood, and that is done with the help of pandas's group by option.



	BOROUGH_GROUP	NEIGHBORHOOD_NAME	COVID_CASE_COUNT	COVID_DEATH_COUNT
0	Manhattan	Chelsea - Clinton	2021	130
1	Manhattan	Union Square - Lower East Side	2105	238
2	Manhattan	Lower Manhattan	513	46
3	Manhattan	Gramercy Park - Murray Hill	1369	98
4	Manhattan	Greenwich Village - Soho	693	50
5	Manhattan	Upper East Side	2518	222

Finally whats left is to link each neighborhood to its respective location(coordinates)



	BOROUGH_GROUP	NEIGHBORHOOD_NAME	COVID_CASE_COUNT	COVID_DEATH_COUNT	Latitude	Longitude
0	Manhattan	Chelsea - Clinton	2021	130	40.745278	-74.002222
1	Manhattan	Union Square - Lower East Side	2105	238	40.734200	-73.987500
2	Manhattan	Lower Manhattan	513	46	40.720900	-74.000700
3	Manhattan	Gramercy Park - Murray Hill	1369	98	40.738164	-73.973663
4	Manhattan	Greenwich Village - Soho	693	50	40.735564	-74.002887
5	Manhattan	Upper East Side	2518	222	40.773600	-73.956600