Spatial Summary of Outdoor Dining and COVID-19 Rates in NYC

Pratishta Yerakala

May 10 2021

Abstract

As restaurants began adjusting restrictions to allow for more outdoor dining options, the COVID-19 pandemic was still affecting NYC. This study is meant to determine whether there is a spatial relationship between the number of open restaurants and COVID-19 case and death rates per area unit of zip code. The data was collected through NYC OpenData for the restaurants and NYC Department of Health for the coronavirus data. First spatial diagnostics were taken and a positive and statistically significant Moran's I value determined that there is indeed spatial autocorrelation between these variables. Calculating the local Moran's I values for each zip code produced maps that showcased the clustered neighborhoods that had high-high and low-low observations. Finally a more fitting spatial lag model was used to obtain more accurate coefficients. This analytical approach found that for every restaurant approved for outdoor dining in a neighborhood, the case and death rate fall by 3 per 100,000 and 0.739 per 100,000 respectively. Ultimately, though this study shed light on a very particular set of data, it allows for a similar study to be done with more variables that could affect the spread and rate of COVID-19 as described in the literature.

Introduction

Since March of 2020 it seems that no discussion in or about New York City can be exempt from a COVID-19 caveat. Though the words 'unprecedented times' have become somewhat of a hollow and cringe-worthy phrase haphazardly thrown around in emails, the fact remains that this city became the epicenter of the COVID-19 outbreak in the United States shortly after the first national case. For all the unique problems that this pandemic has presented us with, it is not as though it is the first epidemiological crisis that the country has faced. The closest comparable outbreak would be the SARS epidemic in 2003. The symptoms of COVID-19 are milder than those of SARS, but the spread of the virus is a lot faster via human to human contact (Fani,Teimoori, Ghafari 2020). The result was federal governments, local governments and the institutions of authority on public health, such as the Center for Disease Control (CDC), all providing the general public with overwhelming and often conflicting guidelines.

Now, for the first time in recent history, everyone was responsible for doing their part in slowing the spread of this disease. Staying home, wearing a mask, social distancing, self monitoring - all precautions that were either enforced by the government, community, or the individual. Because individual responsibility can look like the combination of many factors, common practices such as social distancing and wearing a mask can help prevent the spread of the disease. But government mandated structural changes are in effect as well and, as opposed to the measures taken by the individual, these affect a wider number and range of people. There have already been incidents where social distancing practices were more enforced in certain neighborhoods than others.[1] Other examples of such mandates are those related to NYC restaurants and the option of adding sidewalk or roadside seating. With restaurants in particular taking a hit from the stay at home order and losing business, outdoor dining has become an increasingly appealing option, especially as the weather warms up. Naturally this brings up questions about whether the relaxed restrictions could invite a new wave of COVID cases and if so, whether there are some neighborhoods more affected than others. This project seeks to address a portion of those questions.

Just as the COVID-19 pandemic has produced a score of new and unique problems, there have been similarly novel technological responses to documenting and researching solutions.

[1] Southall, Ashley

Because the longitudinal studies necessary to determine the wide-spread and long-term effects of the pandemic and virus itself will require far more resources and time, quick and smaller scoped projects are far more prevalent. This study specifically does two things:

1) provide an exploratory summary via data visualization of COVID-19 related metrics
2) address the question: Is there a geospatial relationship between open restaurants approved for outdoor dining and COVID-19 cases?

The reason it is important to provide an exploratory summary about COVID-19 in NYC is because there is an abundance of data, all of which are being interpreted differently by various entities before being published for the public. The methodology for data visualization in this project is simply another tool to make sense of publicly available and accessible data. As mentioned earlier, the responsibility of the individual to both protect and be informed of the wider community was often emphasized during the pandemic. Adding to the library of references for an individual to replicate or expand with their own research interests would similarly empower the individual. This study is also important because along with other activities that have been halted or slowed down by COVID-19 (i.e. postal mail for the election, slower hospital visits for other non-COVID-19 related issues), restaurants have been severely impacted as well. In fact it's one of the biggest pulls for tourists and generates revenue for the city and the business that make it up. Having outdoor seating is one way that businesses in danger of closing can have agency to stay afloat during the pandemic. But as with opening schools and gyms, allowing restaurants to open in any capacity would still attract crowds and gatherings which could potentially lead to the spread of COVID-19. Doing a geospatial analysis would provide a brief picture of where restaurants with open outdoor seating and where COVID-19 cases are concentrated on a map. This project is not meant to read narratively. However, a geospatial analysis in combination with exploratory data visualization could possibly provide an individual reader with more personal context for a geographical location they are familiar with.

This paper is divided into the following parts: a literature review and theoretical background, followed by data summaries, explanation of variables, analytical approach, results, a discussion and finally a conclusion.

Literature review

The literature and resource review for this thesis will cover existing studies that performed geospatial analysis and used various models, then it will pivot to some background information about COVID-19 spread and epidemiological measurement. Many of these studies focus on one particular problem and thus, have a narrow scope. While this provides useful information, there are still many factors that could contribute to this pandemic that could not be considered within the scope of any one or even several papers. I will describe which aspect of these papers are helpful for this project and what I could improve upon and contribute to.

Demographic Overview and Possible Geospatial Factors

One of the most comprehensive papers written about this topic is "GIS-based spatial modeling of COVID-19 incidence rate in the continental United States" by Mollalo, Vahedi, and Rivera. The authors combined a list of 35 variables that range from demographic to environmental and performed local and global geospatial models to find if there was spatial dependence. This is a particularly good study to use as a reference because they break down the global (spatial lag and spatial error models) and local models (geographically weighted regression model). The four variables chosen to be in the final model were income inequality, median household income, the percentage of nurse practitioners and percentage of black female population at the county level. It seems that income inequality was most influential in COVID-19 incident rate in the tri-state-area and parts of the northeast. Though southern U.S. areas were not as well explained by income inequality, other variables such as nurse practitioner population and black female population had a bigger influence in that area. But because my thesis is focusing on NYC, the researchers' results in the northeast are most valuable.

Another study called "Spatial analysis of COVID-19 clusters and contextual factors in New York City" by Cordes and Castro hone in on testing specifically when looking for spatial dependence. They look at clustering specifically for testing rates, positivity rates, and proportions of tests that were positive. What's helpful from this study for my thesis is that the researchers looked at the data on a zip code level and organized it by creating a four quantile categorization. This ensures that there are an equal number of zip codes in each category. Additionally, they also calculate the Moran's I value which determines whether there is a spatial dependence present in the first place. Though they focus more on testing, there is still a significant amount of work put into

background demographic distributions as well. For example, Cordes and Castro produced joint spatial distributions of higher positive test cases with higher percentages of black populations and higher percentages of populations with higher education degrees. Already having access to what some of thes maps may look like is invaluable in setting expectations for the thesis.

Among the chaos that COIVD-19 has caused, researchers have honed in on several disparities that arose. One of those is the question of race and ethnicity and how that plays a role in how COVID-19 affects people of particular groups. Gross et. al in their paper describe how a lack of access to proper testing has produced adverse outcomes. Their paper focused on evaluating the completeness of race and ethnicity reporting. This is useful as a primer when beginning my research so that I could evaluate myself on the holistic nature of the data in OpenData. For example, there may be some discrepancies where testing rates may be higher or lower depending on whether certain neighborhoods have a higher number of immigrants and whether that population would be subject to the same kind of reporting that citizens would be. This is important to consider especially since geographic data is a proxy by nature and having many layers of inference could cloud clear results.

Almagro and Orane-Hutchinson performed regression analysis to determine whether factors such as occupation affects COVID-19 exposure. This paper is important because it could help highlight how certain specific determinants are worth exploring. For example, the researchers found that occupation is much more of a determinant than income or specific race factors.

The resulting demographics could also be interesting. The Center on Budget and Policy Priorities tracked the recession effects on food, housing and employment hardships. Because OpenData also has data on free meals locations directly due to COVID-19 and has location data, I could perform exploratory research to see if there is a connection.

Spatial Analysis

Harris describes in his working paper that the NYC subway played a critical role in the spread of COVID-19. He goes into depth about how the shutoff of the subway ridership has a strong correlation with the slowing down of cases since. The methodology of this paper is interesting since the subway turnstile data has point location data which is superimposed on zip code level

maps. This could be helpful for me especially if I want to map point data specific information such as Grab and Go free meals on top of zip code level data of most COVID-19 areas.

Another paper provides more insight into which neighborhoods are better equipped to take preventive measures such as social distancing. Researchers Jay et. al describe how physical distancing, though effective, might be subject to certain socioeconomic factors. They found that spending the day entirely at home was almost more than doubled by those in high income neighborhoods and those from low income neighborhoods were more likely to work outside their home. This paper could be used in conjunction with Almagro and Orane-Hutchinson's paper regarding occupation as well.

But both these papers use geographic data such as neighborhoods and zip code as a proxy or auxiliary factors whereas I want to perform a strictly geographic spatial regression if I find there to be a spatial dependence at all. Though it seems from most of the literature that there already is a spatial dependence and it might be due to various other factors that highly correlate with geography, honing in on certain neighborhood clusters and mapping demographic data on top of those clusters may provide the reader with a more explicit view of these results.

Outdoor transmission of disease

Senatore et. al write in their paper, "Indoor versus outdoor transmission of SARS-COV-2: environmental factors in virus spread and underestimated sources of risk" about the risks they researched that are associated with both indoor and outdoor transmission. This study can be immensely useful in providing context for the justification of allowing the reopening of restaurants as long as they only do outdoor dining. Though my dataset focuses on outdoor dining providing restaurants, this could also be applied in future studies or exploratory research for outdoor pivoted activities with social distancing in general. The most important parts of this study are detailed in the 'Underestimated outdoor risk sources' section where both older and newer research explains the spread of a virus via aerosol. This is fairly common knowledge especially among the general public. However, studies that report that smoke and dust can also contribute to the dispersion of the virus, especially since it can last for days on various surfaces. This brings up questions about NYC's air and pollution quality as well as the general disturbances and construction that tends to happen on sidewalk streets. If indeed there is a spatial correlation between high COVID-19 cases and high density of outdoor dining

restaurants, it might be worthwhile to look into this research further and possibly add other variables (i.e. air quality data, construction location and schedule data, etc).

Data

The two main data sources for this project are the Open Restaurant Applications[2] from NYC OpenData and NYC Coronavirus Disease 2019 (COVID-19) Data[3] from the NYC Health Department. For geospatial data, the shapefiles from the NYC Health Department GitHub repository was used. Finally, after all the datasets have been merged on the unifying key of Zip Code Tabulation Areas (ZCTAs), and the projection has been set, the data is ready for analysis and visualization.

The quality of this data is measured by how well the data is consistent for the period of time that is being queried and whether this data is publically available. The update frequency for the Open Restaurant Applications is every weekday. The update frequency for the interested data from NYC Coronavirus Disease Data is every day with a 3 day lag. The well documented and saturated nature of these two datasets make them of fairly good quality. They are also publically available which more easily allows for similar projects such as this using similar dataset sources.

Additionally, there is the matter of date that the data was queried. The governor announced increased capacity for indoor dining on March 19. For this project, I propose that as indoor dining becomes more available, as the weather warms up for outside dining to be just a viable option, this may be a good proxy for what an approximate patronage to restaurants looks like in the upcoming months. And because COVID-19 symptoms manifest anywhere from 2-14 days from initial exposure, an approximation of two weeks from March 19, 2021 is April 5, 2021.

Restaurant Data

The Open Restaurant Applications dataset contains applications from restaurants or "food service establishments" seeking authorization to re-open under Phase Two of NYC's Forward Plan and also provide outdoor seating for dining. The dataset is provided by the Department of Transportation (DOT) and is owned by NYC OpenData. A detailed description of the dataset says: The Open Storefronts program allows eligible businesses to conduct activity on sidewalks and roadways through the Open Streets: Restaurants program, or a combination of both. In addition, to businesses engaged in retail trade, repair stores, personal care services, and

[2] NYC OpenData and NYC DOT
[3] NYC Department of Health

dry-cleaning and laundry services are able to use outdoor space for seating, queuing, or display of dry goods. Each row in the dataset represents a business that applied to the Open Storefronts program beginning from June 6, 2020 and updated regularly until present day. The data used for this project was taken on April 5, 2021.

NYC OpenData allows for various methods of acquiring the data including via API, downloading JSON or CSV files. For this project, this dataset has been downloaded as a CSV file and then imported into Google Collaboratory notebook for data cleanup.

The schema includes the columns from Appendix Table 1. The most important columns from this raw CSV are the descriptive data (i.e. `Restaurant Name`, `Approved for Sidewalk Seating`, and `Approved for Roadway Seating`) and the geolocation data (i.e. `Latitude`, `Longitude`, `geometry`, etc).

COVID-19 Data

NYC Coronavirus Disease 2019 (COVID-19) Data is hosted and updated weekly on a GitHub repository. The data itself is sourced from NYC Department of Health and Mental Hygiene. The data spans from when the outbreak first started on February 29, 2020 to present day. The repository contains data in aggregates and sums for various metrics such as counting COVID-19 cases, hospitalizations and deaths. But for this particular project, because it is within the lens of geospatial data, only the geographic data will be queried from this repository. The geographic information is presented through Modified Zip Code Tabulation Areas (MODZTCAs) being the defining geographic feature. The data used for this project was taken on April 5, 2021.

Because the data is hosted on GitHub it is possible to use the raw file URLs but in order to preserve any possible breaks in hyperlinks (since there is no direct API access), the data has been downloaded in CSV format. It has then imported into Google Collaboratory notebook for data cleanup.

The schema includes the columns Appendix Table 2. The most important columns from this raw CSV are the `modzcta`, `NEIGHBORHOOD_NAME`, `lat`, `lon`, `COVID_CASE_COUNT`, `COVID_CASE_RATE`, `COVID_DEATH_COUNT`, and `COVID_DEATH_RATE`. Though there is only one independent variable that is within the scope of the research question for this project, the COVID-19 metrics would prove useful for the exploratory data research aspect.

NYC Geospatial Data

The geospatial shapefiles also come from the GitHub repository for NYC Coronavirus Disease 2019 (COVID-19) Data. However, an issue with data that comes along with geographical data analysis in general is the problem of how to measure geospatial data in the first place. This becomes a problem specifically when different variables are measured for different geographic features such as ZIP code vs. U.S. Census tracts. A member of the general public would be familiar with ZIP code as measurement of geographic area due to this data point being associated with residential and business addresses. That is because ZIP codes are primarily used by U.S. Postal Service to determine routes taken to deliver mail. Because ZIP codes don't necessarily contain geographical data – there can be a ZIP code for a whole building or a ZIP code for an area that doesn't have a residential population – the NYC Health Department uses ZCTAs. ZCTAs – which stands for Zip Code Tabulation Areas – consolidate ZIP codes into units of area. This geography has been created by the U.S. Census Bureau. However, NYC Health Department data actually further modifies these ZCTAs into Modified ZTCAs (MODZTCA). This combines census blocks with smaller populations to allow more stable estimates of population size for rate calculation.

Merging Data

In order to perform the analyses described below, the variables of interest need to be consolidated into one GeoPandas Dataframe. First all three files - NYC Shapefile, COVID-19 data, and open restaurants data - were loaded. The two CSV files were converted into GeoPandas Dataframes after identifying the appropriate geospatial variables (e.g. longitude, latitude and geometry). All three files were set to the same projection of EPSG 4326, also known as WGS84. This projection was chosen as a constant because it is the default provided by the Shapefile.

Next, the open restaurants data and NYC Shapefile data merged so that the restaurant observations were made into point data using GeoPandas sjoin function. Then restaurants that have only been approved for either sidewalk or roadway seating have been filtered. And since the geospatial unit of measurement is MODZTCA, the restaurants were grouped by MODZTCA and aggregated. Finally, the COVID-19 data can also be merged on MODZTCA. The final dataframe contains COVID-19 rates and open restaurant count by MODZTCA.

Descriptive Statistics and Exploratory Data Visualization

Because most of the data being analyzed is geospatial, descriptive statistics will primarily include choropleth maps. These maps can show where certain variables are concentrated or distributed over NYC. The main variables are:

1) Open restaurant count
2) COVID-19 case rate
3) COVID-19 death rate

Table 1. Descriptive statistics per variable aggregated for MODZTCA

	Open Restaurant Count	COVID-19 Case Rate	COVID-19 Death Rate
Count	174	174	174
Mean	57.160919540229884	8085.601379310347	293.13511494252884
Standard Deviation	71.56643278828233	2298.5683147532686	139.98479169558416
Min	1	3252.65	0
25% Quartile	10.25	6321.3725	198.0775
50% Quartile	27.5	8192.84	288.49
75% Quartile	81.5	9801.164999999999	360.92249999999996
Max	412	15415.83	906.14

The descriptive statistics shown above demonstrate that the number of restaurants with outdoor dining in a zip code area range between 1 nad 412. The case and death rate which are measured per 100,000 individuals as per Bureau of Epidemiology Services guidelines[4].

[4] Population Denominator Data Sources

Chart 1

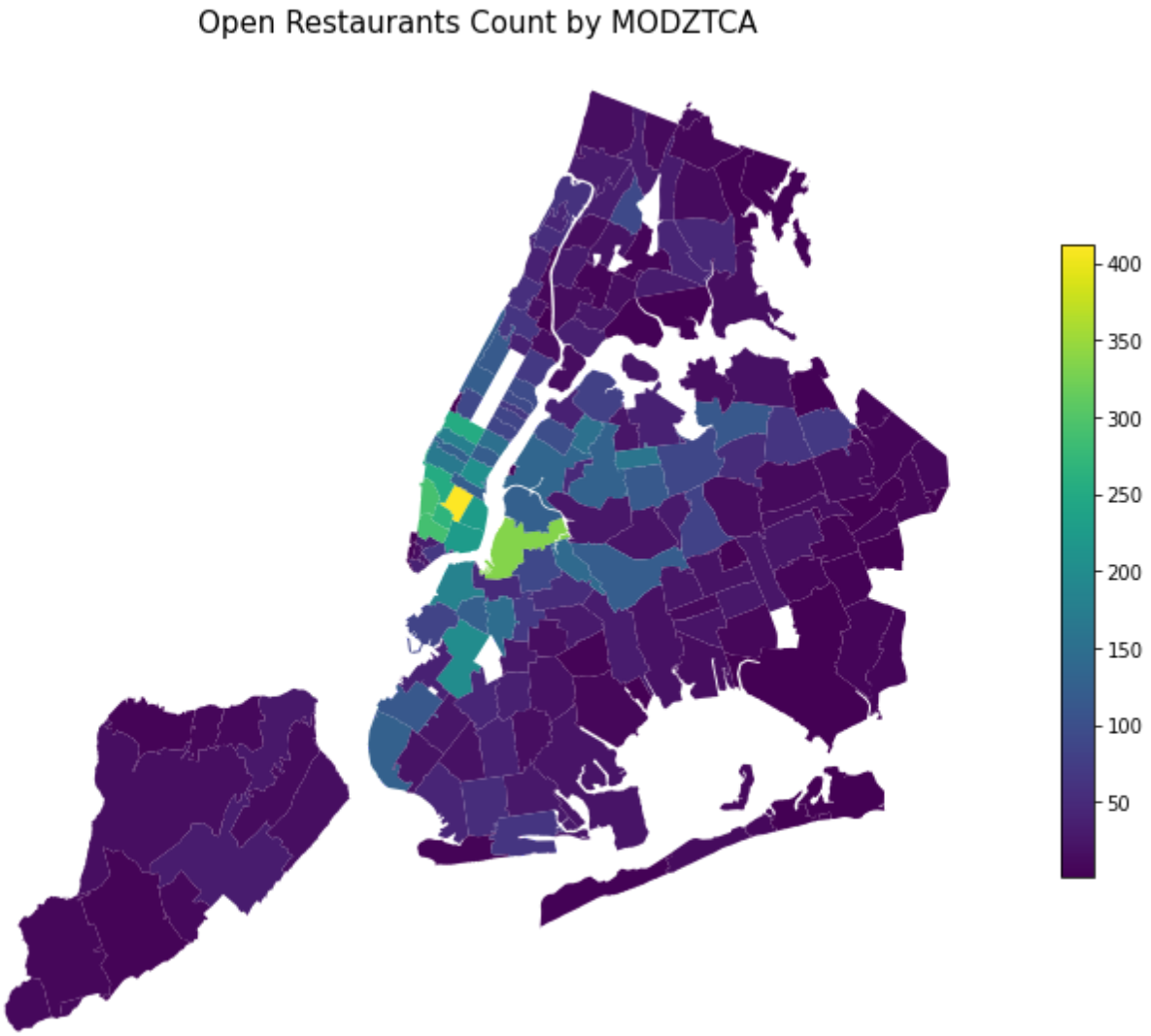
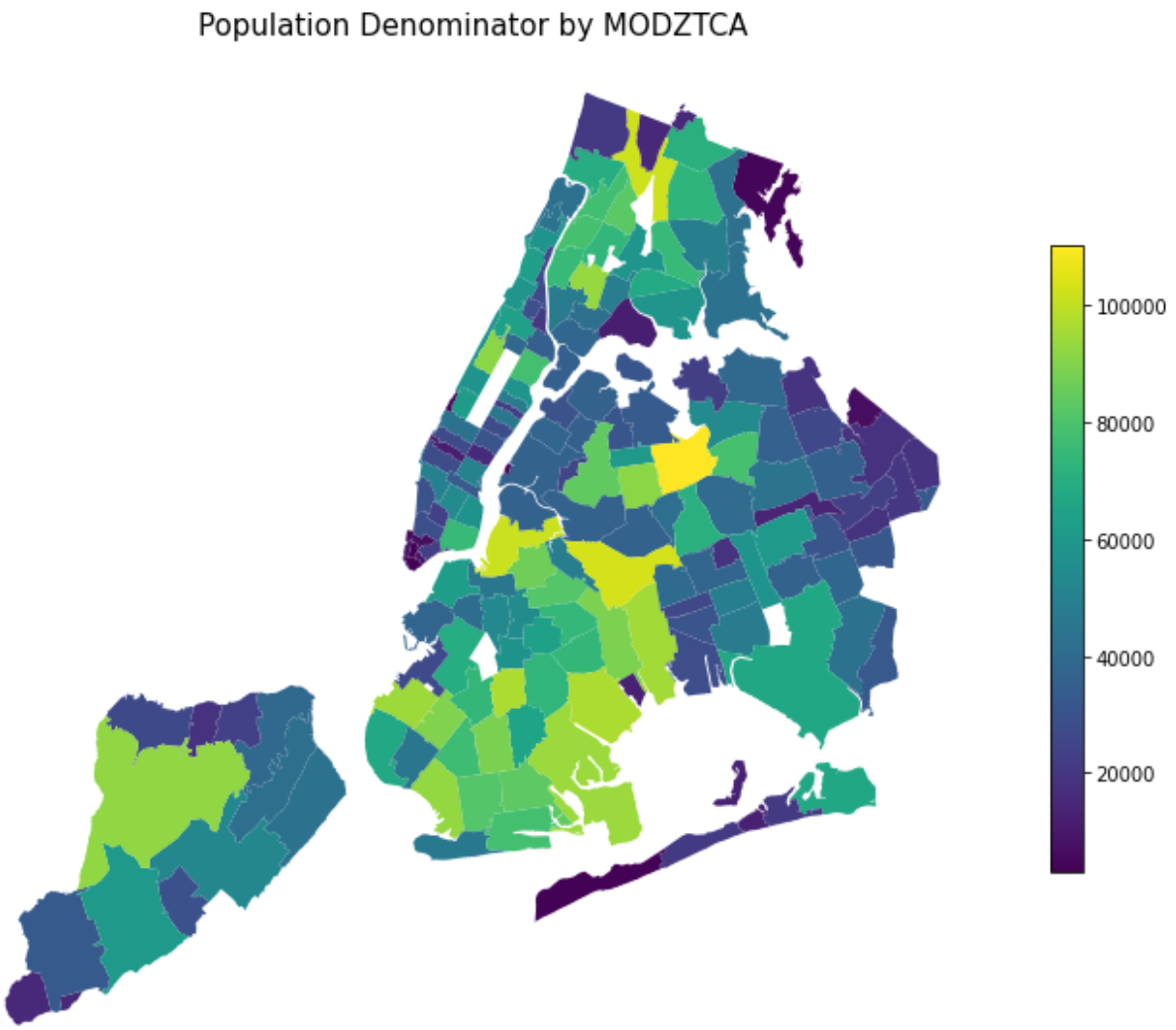


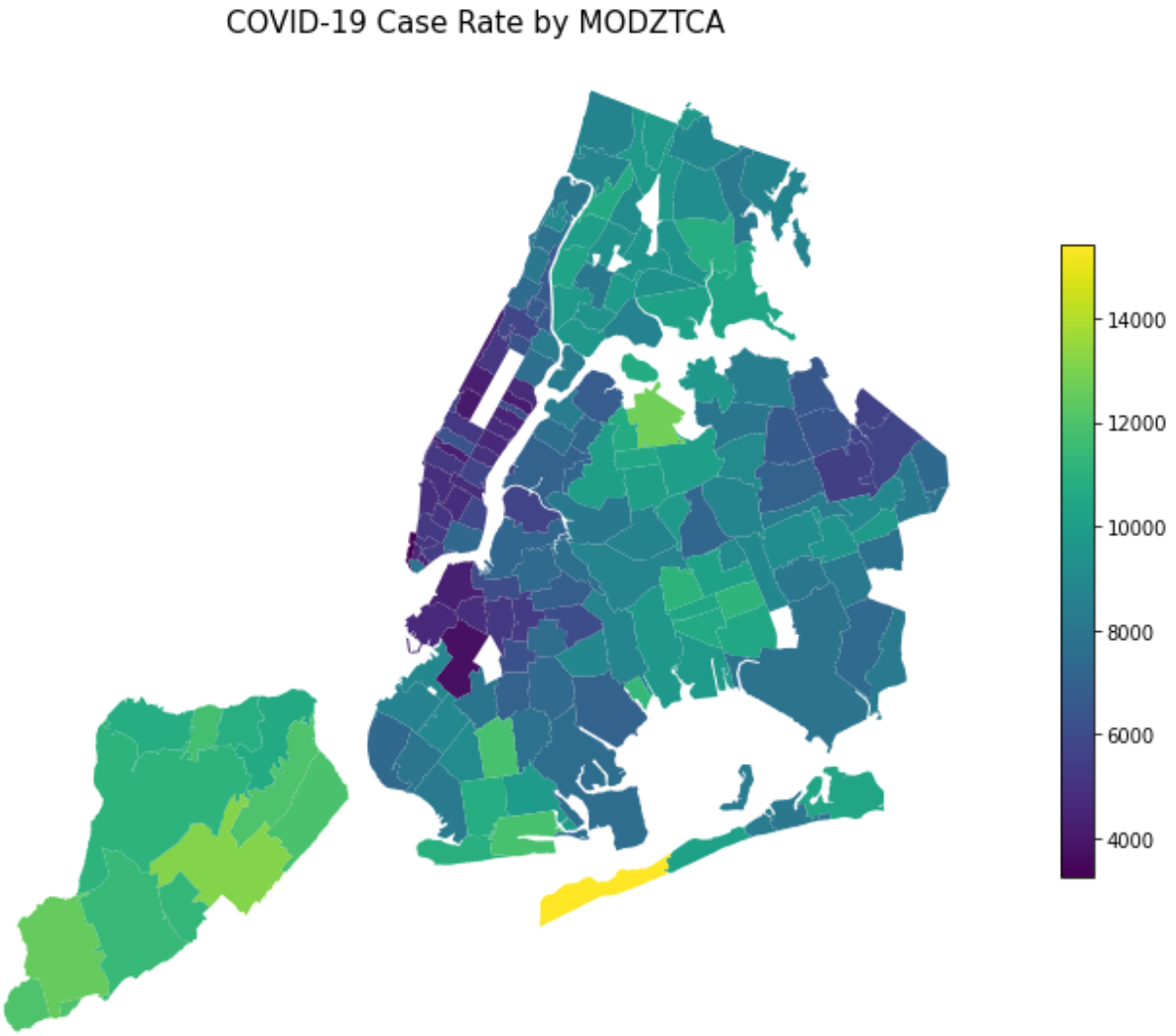
Chart 1 shows that most of the open restaurants with outdoor dining are concentrated in lower Manhattan and some parts of Brooklyn and Queens that are close to Manhattan.

Chart 2



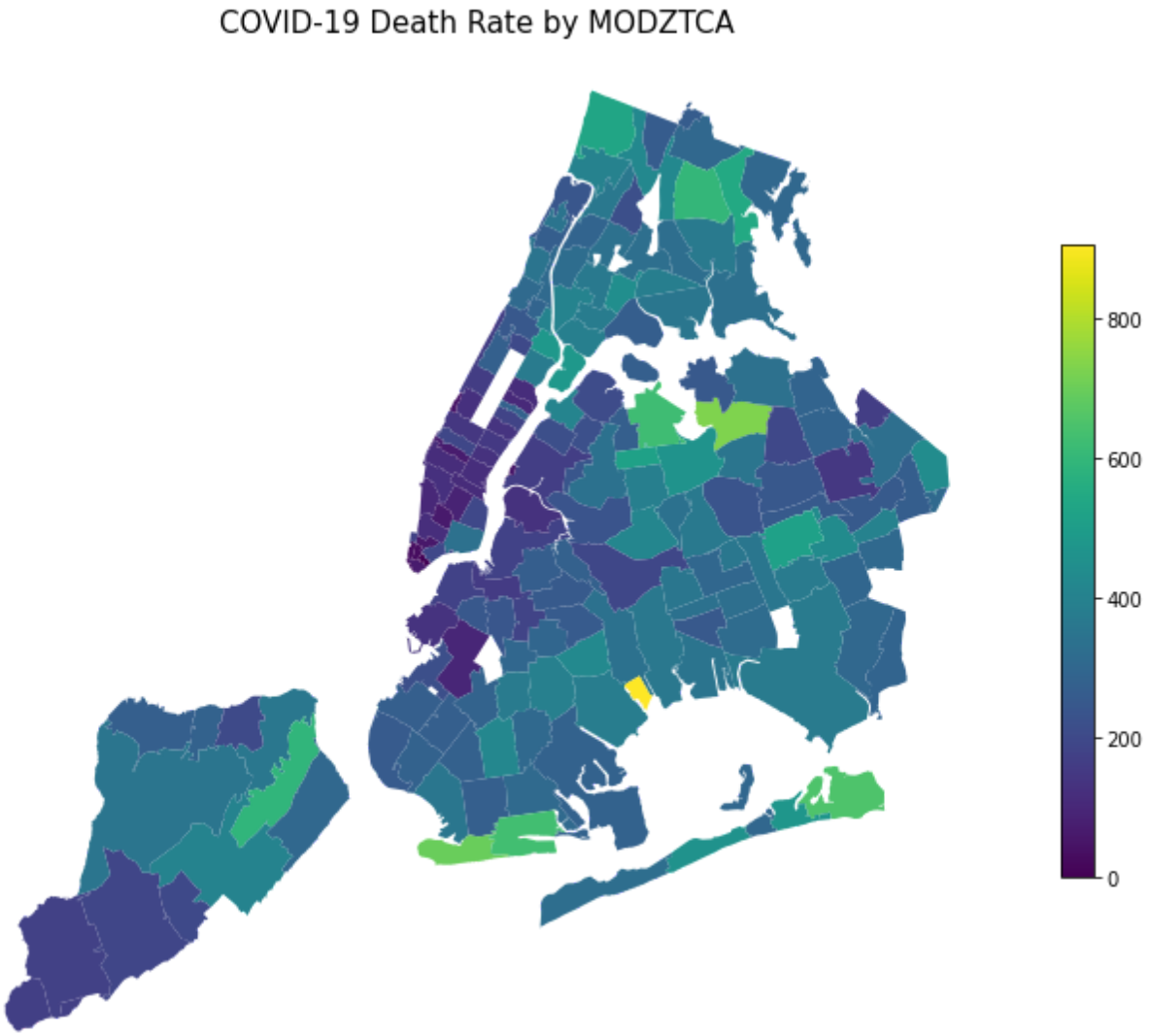
The population denominator in Chart 2 shows that the population that is at risk of contracting COVID-19 are fairly spread out and concentrated more toward the outer boroughs.

Chart 3



The case rate in Chart 3 seems to have lower rates in downtown Manhattan and parts of Brooklyn that are closer to lower Manhattan. The higher rates are seen in Staten Island and the outer parts of Queens.

Chart 4



The death rate in Chart 3 seems overall fairly consistent with higher rates in the outer boroughs and away from lower Manhattan.

Given these graphs, the distribution seems to be fairly straightforward and brings interest to lower Manhattan and outer borough areas.

Measures / Variables

Open restaurants

The term "open restaurants" here refers to a count of restaurants that have applied for outdoor dining and have been approved for either sidewalk or roadside seating.

MODZTCA

The final geographic feature of MODZTCAs are used to provide geographic data. This said, with a couple exceptions, namely the following zip codes: most other zip codes can be directly associated with their ZTCA and MODZTCA counterparts. The data wrangling process takes care of these conversions for analysis.

COVID-19 Case and Death Rate

When the term "COVID-19 rates" is used, it refers to both the dependent variables of COVID-19 case rate and COVID-19 death rate. As specified by the data source, the rates are calculated based on the population denominator variable included in the dataset and it is per 100,000 people.

Population denominator

The population denominator, a variable that directly affects the COVID-19 rates is defined by the CDC as "the sum of the time each person was observed, totaled for all persons. This denominator represents the total time the population was at risk of and being watched for disease." In effect, it is not representative of the actual census population of a MODZTCA.

Analytical approach

The analytical approach of this project takes place in three parts: 1) exploratory data visualization 2) analyzing for spatial autocorrelation, 3) checking for local indicators of spatial autocorrelation, i.e. identifying clusters, and 4) performing a spatial lag regression model.

Data visualization

The exploratory data visualization aspect of the analysis is used primarily to show the distribution of the variables being examined. Namely, the number of open restaurants, population denominators, COVID-19 case and death rates per 100,000 people. All of these variables are split up geographically by MODZTCA as a unit of geospatial analysis. The descriptive statistics section already shows the geographic distribution.

The reason exploratory data visualization is also being included in the analytical approach is because if there is visually an even distribution or any evidence of missing or corrupted data, the rest of the analysis would need to be adjusted. Having seen that there are not particularly hard stop outliers, the rest of the analysis is good to go.

Spatial analysis

Since spatial analysis is the primary focus of the research questions, there are several statistics and analyses that need to be taken into account.

Weights matrix

The weights matrix used here is the queen's contiguity matrix. Similar to how Saffir et. al calculated the Moran's I and LISA calculations on the Queen's contiguity as well, this weights matrix is also of order 1[5]. That is, to be considered "neighbors" for a particular MODZTCA, that MODZTCA must intersect with vertices. Order 1 specifies that it is only a neighbor that is in direct contact with the observed value; no "neighbors of neighbors."

Queen's contiguity is often compared to Rook's contiguity for which to be considered neighbors, the polygons need to share an edge of a certain length, not a single point vertex. And since the polygonal geometry of the MODZTCA areas are irregular in shape, the Queen's contiguity weights matrix is slightly more lenient in considering more areas to be neighbors. This more

[5] Saffary, T, et al

inclusive matrix allows for a more holistic and connective picture of NYC and its MODZTCAs, which is why it will be used.

Moran's Index

The Moran's I statistic determines whether there exists spatial autocorrelation in a dataset. It is normally used to rule out autocorrelation for linear regression models. In this case since the research questions ask whether spatial dependence exists, a positive statistic would confirm that there is. Calculating the weighted OLS regression will produce this and other diagnostic statistics. The p-values can be determined through conducting a series of simulations to determine whether the Moran's I statistic can geographically predict the same or similar map[6].

It is important to note however that this statistic is a global statistic that describes the dataset as a whole. It merely states whether there are areas that there is or is not clustering that is taking place. But it does not describe the nature of the clustering; for that, a more localized approach should be used.

Linear regression (OLS)

A preliminary linear regression analysis would determine whether there exists a relationship at all between the number of open restaurants and COVID-19 rates. The reason for this preliminary analysis is to determine with some significance that there is a relationship between the number of open restaurants and COVID-19 rates at all, and to produce some spatial and inferential statistics to more closely examine the spatial reltaionship. This way there will be a coefficient with a particular significance and fit that could be improved upon by subsequently examining LISA clusters, and ultimately performing a spatial lag regression model.

The basic linear regression model equation is as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n + \varepsilon$$

Where y represents the dependent variable in this case, COVID-19 rates. The β represents the coefficient, the statistic that shows how strongly the relationship between the independent and dependent variable is and the nature of the relationship (e.g. positive or negative). The X variable denotes the independent variables and the ε is the random error term which represents

[6] Local Spatial Autocorrelation — Geographic Data Science with Python.

the residuals. This analysis has one independent variable (open restaurant count) and two dependent variables (COVID-19 case rate and COVID-19 death rate). Because the research question only addresses both the dependent variables separately, it results in two linear regression model equations:

$$y_{cr} = \beta_{cr} X_{cr} + \varepsilon$$

and

$$y_{dr} = \beta_{dr} X_{dr} + \varepsilon$$

where β_{cr} and β_{dr} represent the coefficients for COVID-19 case rate and COVID-19 death rate.

Local indicators of spatial autocorrelation (LISA) clusters

LISA clusters are a localized version of Moran's I statistic. The way this statistic is calculated for each unit of area is by comparing the value of an observation to the average value of its neighbors. If the observed value is more similar or more dissimilar to the average value of its neighbors, it would determine if that area is a high-high or low-low cluster respectively. That is, if an observed value of a neighborhood was a high number of open restaurants, and the average number of open restaurants of that observed area's neighbors is also high, it would indicate that the observed neighborhood is in a high-high (HH) cluster.

Plotting the statistically significant LISA clusters on maps would showcase where there are clusters where there is a high open restaurant count surrounded by other neighborhoods with other high restaurant counts. Or conversely clustered with low-low (LL) restaurant counts. Producing LISA clusters maps for all the variables could showcase interesting distributions of clustering prior to doing a spatial lag regression analysis.

Spatial lag

After determining that there is a spatial autocorrelation between the number of open restaurants and COVID-19 rates, and seeing the corresponding clusters, a natural question may be: what exactly could be causing the spatial dependence? One explanation could be measurement error. For example, it could be the case that using MODZTCAs as mentioned above is not the ideal unit of geospatial measurement. Or it could also be that there is in fact a spatial reason that's rooted in a social or economic characteristic that could be producing the dependence. For example, the number of tourist monuments in a MODZTCA could contribute to the number of open restaurants, which in turn could affect COVID-19 rates.

Additionally, since the linear regression model is no longer a sound analysis now that spatial dependence violates assumptions that render the coefficients biased or inefficient, a spatial regression model is needed. After performing some initial spatial analysis and gaining some diagnostics, Anselin's diagnostic flowchart determines that spatial lag seems to be the best model to address the research questions[7]. The equations for spatial lag are as follows:

$$y_{cr} =_{cr} +\beta_{cr} X_{cr} + \varepsilon$$

and

$$y_{dr} =_{dr} +\beta_{dr} X_{dr} + \varepsilon$$

where ρ represents the spatial lag parameter, W_{cr} and W_{dr} represents the Queen's contiguity of order 1 weights matrix. The rest of the terms represent the coefficient and error term as explained before in the OLS regression model.

Research questions

1) Is there a spatial relationship between the number of open restaurants in a zip code area and the rate of COVID-19 cases in that area?
2) Is there a spatial relationship between the number of open restaurants in a zip code area and the rate of COVID-19 deaths in that area?
3) How are clusters, if any exist, distributed across NYC for open restaurant density and COVID-19 related cases?

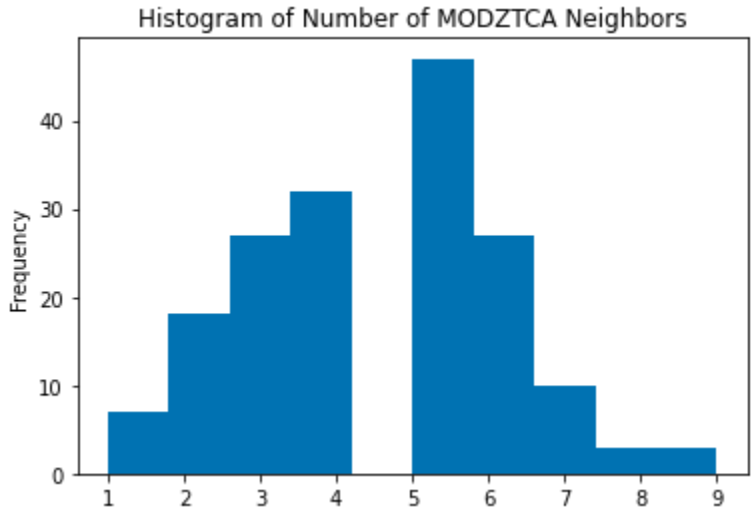
[7] Anselin, Luc, and Sergio Joseph Rey

Results

The results will be organized the same order that the analysis was performed.

Linear regression (OLS)

Chart 5



After applying the Queen's contiguity weights matrix of order 1, visualizing the connectivity between each MODZTCA becomes easier. Chart 5 shows that most MODZTCAs have about 5 neighbors and there is not a severe skew one way or the other.

Table 2

Dependent Variable	Variable	Coefficient	Std. Error	t-Statistic	Probability
COVID-19 Case Rate	**Open Restaurant Count**	-14.699	2.177	-6.750	0.0000000
COVID-19 Death Rate	**Open Restaurant Count**	-0.739	0.138	-5.353	0.0000003

Table 2 shows that there exists a relationship between COVID-19 rates and the number of open restaurants. Both the coefficients show a negative relationship and represent the relationship between the two variables across all MODZTCAs.

For COVID-19 case rates, on average the number of open restaurants with sidewalk or roadway seating by 1 would result in 14.699 fewer cases per 100,000 people in that MODZTCA.

For COVID-19 death rates, on average an increase in the number of open restaurants with sidewalk or roadway seating by 1 would result in 0.739 fewer deaths per 100,000 people in that MODZTCA.

Both these coefficients are statistically significant as the p-values are less than 0.005.

Table 3 Case rate spatial diagnostics

	DF	Value	Probability
Moran's I	0.6232	11.789	0.000
Lagrange Multiplier (lag)	1	146.473	0.000
Robust LM (lag)	1	16.855	0.000
Lagrange Multiplier (error)	1	130.086	0.000
Robust LM (error)	1	0.468	0.4939

A positive and statistically significant Moran's I demonstrates that there is spatial autocorrelation and clustering taking place. Next, Anselin's diagnostic tests[8] can be used to determine what kind of spatial dependence regression model could be ideal: spatial error or spatial lag? The Lagrange Multiplier for lag and the Lagrange Multiplier for error are both statistically significant. Now to look at the Robust Lagrange Multiplier statistics-- only the Robust Lagrange Multiplier for lag is statistically significant. This indicates that to have a more efficient and unbiased model, the spatial lag regression model is best for the dependent variable COVID-19 case rate.

Table 4 Death rate spatial diagnostics

	DF	Value	Probability
Moran's I	0.3177	6.097	0.000
Lagrange Multiplier (lag)	1	45.014	0.000
Robust LM (lag)	1	14.083	0.000
Lagrange Multiplier (error)	1	33.815	0.000
Robust LM (error)	1	2.884	0.895

A positive and statistically significant Moran's I demonstrates that there is spatial autocorrelation and clustering taking place. Next, Anselin's diagnostic tests[9] can be used to determine what kind

[8] Anselin, Luc, et al
[9] Anselin, Luc, et al

of spatial dependence regression model could be ideal: spatial error or spatial lag? The Lagrange Multiplier for lag and the Lagrange Multiplier for error are both statistically significant. Now to look at the Robust Lagrange Multiplier statistics-- only the Robust Lagrange Multiplier for lag is statistically significant. This indicates that to have a more efficient and unbiased model, the spatial lag regression model is best for the dependent variable COVID-19 death rate as well.

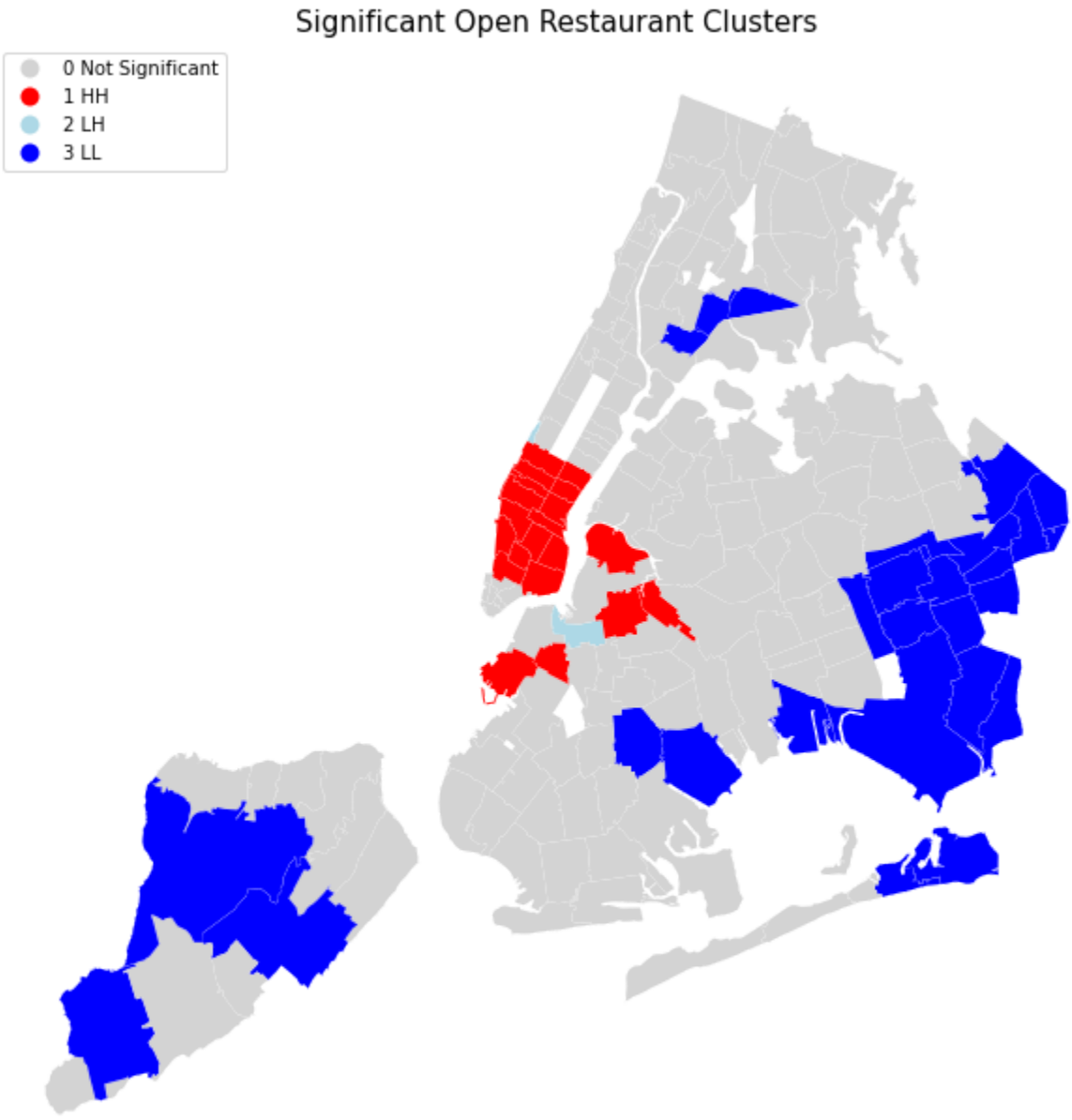
LISA clusters

Before running the spatial lag model, it would be helpful to put the Moran's I statistic to use by producing local Moran's I values to visualize the clustering that's taking place. There are 174 MODZTCAs that are considered, so instead of producing a Moran's I value for each, LISA clusters can be divided into 4 types: HH, LH, LL, HL[10]. The following LISA cluster visualizations help answer the third research question posed above: How are clusters, if any exist, distributed across NYC for open restaurant density and COVID-19 related cases?

[10] Local Spatial Autocorrelation — Geographic Data Science with Python.

Open Restaurants Clustering

Chart 6

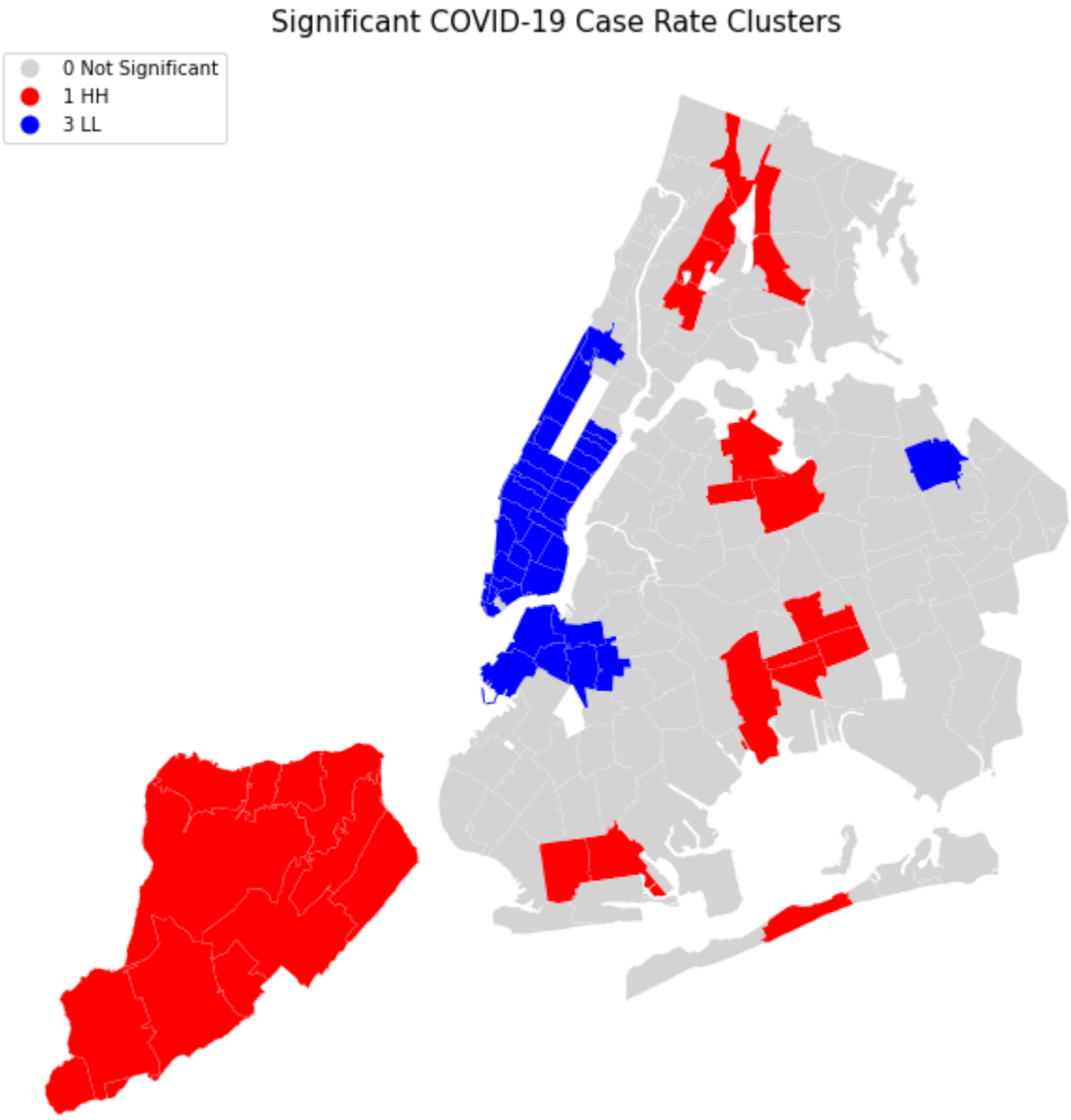


It looks like open restaurants are geospatially clustered to be high-high (HH) in downtown Manhattan and some parts of Brooklyn that are close to downtown Manhattan. This could show that the lower Manhattan area has a high count of open restaurants and is surrounded by other areas with high counts of open restaurants. The low-low (LL) clusters seem to be mostly focused on the outer boroughs. And for Queens and Brooklyn the LL clusters are even further away from downtown Manhattan with most other MODZTCAs LISA scores not having been

significant enough. In fact there is not a single "cold spot" or low-low (LL) cluster in Manhattan at all. This is fairly to be expected or intuited since there is a high density of restaurants and city life in general in Manhattan. But to see the LL clusters be focused to the outer edges of the boroughs is indicative of the spread of open restaurants and perhaps whether it is a priority for some restaurants in the LL clusters to even apply for outdoor dining.

COVID-19 Case Rate

Chart 7

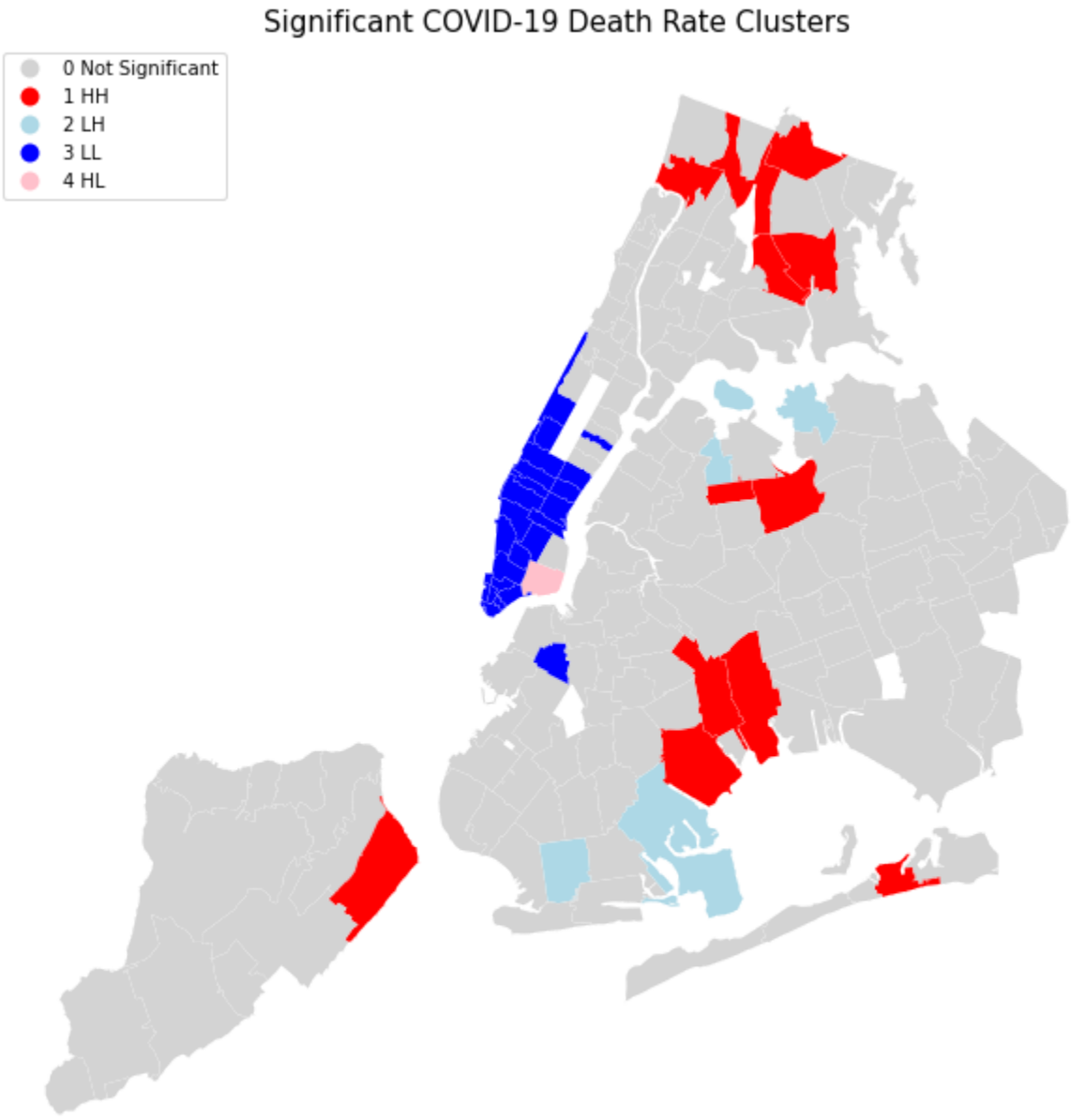


The statistically significant LISA clusters for case rate are only HH "hot spots" or LL "cold spots." They're polarized spatially as well with the LL clusters being in mid to lower Manhattan with parts of Brooklyn that are close to Manhattan. This represents MODZTCAs that have a low COVID-19 case rate per 100,000 are also surrounded by MODZTCAs with - on average - low case rates as well. Conversely, all of Staten Island, parts of the Bronx, Queens and Brooklyn are

HH clusters meaning not only are those areas high in case rate but their neighbors are also high in case rate.

COVID-19 Death Rate

Chart 8



The significant LISA clusters for COVID-19 death rates are disparately different in distribution than the previous two LISA cluster maps. There are more MODZTCAs that have a low open restaurant count but are surrounded by neighbors with high open restaurant counts, a low-high (LH) cluster. These areas are mostly subject to the outside parameters of Brooklyn and Queens. The LL clusters for death rate, similar to case rate, are focused in lower Manhattan. The HH clusters are not as widespread as the case rate cluster but are still more prevalent in the outer

boroughs. There is one HL case in lower Manhattan which may be an interesting MODZTCA case to examine more closely.

Spatial lag

From the spatial diagnostics in the OLS regression model, Anselin's diagnostic test points toward fitting a spatial lag recession model to both the COVID-19 case rate and COVID-19 death rate variables.

Table 5

Dependent Variable	Variable	Coefficient	Std. Error	t-Statistic	Probability
COVID-19 Case Rate	**Open Restaurant Count**	-3.9319879	1.2933898	-3.0400640	0.0023653
COVID-19 Death Rate	**Open Restaurant Count**	-0.3325239	0.1201694	-2.7671267	0.0056553

Including a spatial lag parameter for the analysis provides a statistically significant coefficient for COVID-19 case rate variable but not a statistically significant coefficient for the COVID-19 death rate variable.

For COVID-19 case rates, on average the number of open restaurants with sidewalk or roadway seating by 1 would result in 3.932 fewer cases per 100,000 people in that MODZTCA.

Discussion

The reason that this project is meant to be a summary is because the scope does not include finding a direct causal link between open restaurants and COVID-19 rates. The aim is to identify if there is a spatial correlation at all. If so, what is the nature of that relationship? And finally, how are these features clustered spatially?

To answer the first question, the Moran's I for both COVID-19 case rate and for COVID-19 death rate are both positive and statistically significant. This indicates that all three variables are indeed spatially dependent. As such the statistically significant LISA cluster visualizations show where exactly these clustering is happening and to what capacity. It was clear that Manhattan, especially lower Manhattan and the neighborhoods in both Brooklyn and Queens that were geographically close to lower Manhattan, had some of the lowest COVID-19 rates while having the highest open restaurant count. This is reflected in the negative coefficients for both dependent variables representing open restaurants.There were also inverse correlations in the outer boroughs that had low open restaurant count but high COVID-19 rates.

Though the spatial lag regression model was a better fit, the coefficient representing the number of open restaurants and its relationship to COVID-19 death rates was not significant. Future work could include mapping the residuals to ensure that there is indeed no residual clustering as well.

Beyond the scope of this project, other research to build on could include several things. For one, it would be enlightening and important to also consider population as a whole (not just as a denominator for estimating epidemiological infection spread). This is because it may relate to the concentration or density of restaurants with outdoor dining in a particular neighborhood. Additionally, there is other interesting data within the Open Restaurants dataset including that about the area and size of the proposed outdoor seating. It would be interesting to see if there's a geospatial relationship between the proposed seating and the area of a MODZTCA or the population. This would require some more research on population distribution over spread out neighborhoods, but it could lead to more complex research about COVID-19 rates in various types of environments. Another example would be to do multivariate analysis by including MODZTCA average income as well to examine if there's a relationship to open restaurant density.

Conclusion

It seems that though there are numerous variables that could affect COVID-19 related cases and deaths, the number of open restaurants with outdoor dining could be one of them. Looking at spatial diagnostics, all those features are spatially dependent and after running a weighted OLS regression and spatial lag regression model for both case and death rates, there are statistically significant coefficients that describe in which way they are related. Namely, the higher the number of open restaurants, the lower the COVID-19 rates. However, there are also localized metrics which are more visible through visualizations such as those of LISA clusters. Though these clusters and results seem to have some relationship with COVID-19 cases, there are many more features that are available to examine with a spatial lens. A similar approach as one from this project may shed light on future studies.

Bibliography

Anselin, Luc, and Sergio Joseph Rey. Modern Spatial Econometrics in Practice. Geoda Press LLC, 2014.

Anselin, Luc, et al. “Simple Diagnostic Tests for Spatial Dependence.” Regional Science and Urban Economics, no. 1, Elsevier BV, Feb. 1996, pp. 77–104. Crossref, doi:10.1016/0166-0462(95)02111-6.

Cordes, Jack, and Marcia C. Castro. “Spatial Analysis of COVID-19 Clusters and Contextual Factors in New York City.” Spatial and Spatio-Temporal Epidemiology, Elsevier BV, Aug. 2020, p. 100355. Crossref, doi:10.1016/j.sste.2020.100355.

Darmofal, D. "Spatial Analysis for the Social Sciences." University of South Carolina. Accessed 10 May 2021.

Fani, M., Teimoori, A., & Ghafari, S. (2020). Comparison of the COVID-2019 (SARS-CoV-2) pathogenesis with SARS-CoV and MERS-CoV infections. Future Virology, 10.2217/fvl-2020-0050. https://doi.org/10.2217/fvl-2020-0050

Harris, Jeffrey E. “The Subways Seeded the Massive Coronavirus Epidemic in New York City.” SSRN Electronic Journal, Elsevier BV, 2020. Crossref, doi:10.2139/ssrn.3574455.

“Local Spatial Autocorrelation — Geographic Data Science with Python.” Welcome! | Geographic Data Science with PySAL and the PyData Stack, https://geographicdata.science/book/notebooks/07_local_autocorrelation.html. Accessed 10 May 2021.

Mollalo, A., Vahedi, B., & Rivera, K. M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. The Science of the total environment, 728, 138884. https://doi.org/10.1016/j.scitotenv.2020.138884

NYC Department of Health. "NYC Coronavirus Disease 2019 (COVID-19) Data." NYC Department of Health, February 29, 2020. https://github.com/nychealth/coronavirus-data.

NYC OpenData and NYC Department of Transportation (DOT). "Open Restaurant Applications." NYC OpenData, June 29, 2020. https://data.cityofnewyork.us/Transportation/Open-Restaurant-Applications/pitm-atqc.

"Population Denominator Data Sources | U.S. Cancer Statistics Data Visualizations Tool Technical Notes | CDC.” Centers for Disease Control and Prevention, https://www.cdc.gov/cancer/uscs/technical_notes/data_sources/population.htm. Accessed 10 May 2021.

Saffary, T, et al. "Analysis of COVID-19 Cases' Spatial Dependence in US Counties Reveals Health Inequalities." Frontiers in Public Health, November 12, 2020.

Senatore, V., Zarra, T., Buonerba, A. et al. Indoor versus outdoor transmission of SARS-COV-2: environmental factors in virus spread and underestimated sources of risk. Euro-Mediterr J Environ Integr 6, 30 (2021). https://doi.org/10.1007/s41207-021-00243-w

Southall, Ashley. “Scrutiny of Social-Distance Policing as 35 of 40 Arrested Are Black.” The New York Times. The New York Times, May 7, 2020. https://www.nytimes.com/2020/05/07/nyregion/nypd-social-distancing-race-coronavirus.html.

Staff, Eater. “Coronavirus in NYC: Restaurants That Have Closed Permanently Due to the Pandemic - Eater NY.” Eater NY, Eater NY, 8 May 2020, https://ny.eater.com/2020/5/8/21248604/nyc-restaurant-closings-coronavirus.

Appendix

Appendix Table 1

	Open Restaurant Data Columns
0	objectid
1	globalid
2	Seating Interest (Sidewalk/Roadway/Both)
3	Restaurant Name
4	Legal Business Name
5	Doing Business As (DBA)
6	Building Number
7	Street
8	Borough
9	Postcode
10	Business Address
11	Food Service Establishment Permit #
12	Sidewalk Dimensions (Length)
13	Sidewalk Dimensions (Width)
14	Sidewalk Dimensions (Area)
15	Roadway Dimensions (Length)
16	Roadway Dimensions (Width)
17	Roadway Dimensions (Area)
18	Approved for Sidewalk Seating
19	Approved for Roadway Seating
20	Qualify Alcohol
21	SLA Serial Number
22	SLA License Type
23	Landmark District or Building
24	landmarkDistrict_terms
25	healthCompliance_terms
26	Time of Submission
27	Latitude
28	Longitude
29	Community Board
30	Council District
31	Census Tract

32	BIN
33	BBL
34	NTA
35	geometry

Appendix Table 2

	COVID-19 Data Columns
0	modzcta
1	NEIGHBORHOOD_NAME
2	BOROUGH_GROUP
3	label
4	lat
5	lon
6	COVID_CASE_COUNT
7	COVID_CASE_RATE
8	POP_DENOMINATOR
9	COVID_DEATH_COUNT
10	COVID_DEATH_RATE
11	PERCENT_POSITIVE
12	TOTAL_COVID_TESTS