IBM Data Science Capstone Project

Title: Relationships between Covid-19 and neighborhood venues

Author: Michael Yeh

Date: August, 2020

Objective:

The goal of this project is to explore the relationships, if any, between neighborhood venues and Covid-19 infection / death rates.

New York City is a diverse city with many distinct neighborhoods. Neighborhood venues are a reflection of the life-style within each neighborhood, and life-style may have a connection to how Covid-19 infected people through community spread. For example, several states that saw spiking Covid-19 infection rates have shutdown bars and indoor dinning which implies a believe that these venues contribute to community spread risk. Additionally, life-style choices have health implications on the community living in each neighborhood. Community health, in turn, has an impact on Covid infections and deaths.

Target Audience:

This analysis is intended to inform public health officials and hopefully contribute to beneficial policy changes. For example, one category of venues used in this analysis is the "Take_ out_food" category (e.g., fast food, cake shops) . These venues have little or no seating space, therefore customers generally order something and take it away to consume. So far, public health officials have not identified this type of venues as high risk. However, these venues generally have high volume of traffic, so more people can potentially come into contact with an infected person. Moreover, take out foods are often associated with bad health outcomes which can cause the neighborhood population to be more vulnerable to infection and death.

Data

New York City (NYC) officials collect city-wide Covid-19 data and provide that data to the public via a Github depository. The data is organized by Modified Zip Code Tabulation Areas (MODZCTA), going forward let's refer to this as the "zipcode". This is also the method used by the US Census Bureau. Therefore, the data required for this analysis are:

- NYC Covid-19 infection and death data by MODZCTA (ctl-click here).
- A mapping from zipcode to longitude and latitude (geospatial data) for extracting venue data from FourSquare. This data is available by state, ctl-click here for New York State data.
- Size of each NYC borough for computing the average radius of a zipcode. This data is available from Wikipedia.
- Venues data from FourSquare extracted by zipcode.

Here is a sample of the NYC Covid-19 data:

	Zip	Neighborhood	Borough	Pos_cases	Pos_case_rate	Population	Deaths	Death_rate	Pct_positive	Total_cases
0	10001	Chelsea/NoMad/West Chelsea	Manhattan	413	1752.75	23563.03	24	101.85	8.01	5154
1	10002	Chinatown/Lower East Side	Manhattan	1207	1572.53	76755.41	160	208.45	11.23	10749
2	10003	East Village/Gramercy/Greenwich Village	Manhattan	502	933.06	53801.62	34	63.20	6.07	8273
3	10004	Financial District	Manhattan	36	986.14	3650.61	1	27.39	6.49	555
4	10005	Financial District	Manhattan	75	893.27	8396.11	2	23.82	5.77	1299

Here is a sample of the longitude and latitude data:

	Zip	Latitude	Longitude
0	10001	40.750742	-73.99653
1	10002	40.717040	-73.98700
2	10003	40.732509	-73.98935
3	10005	40.706019	-74.00858
4	10006	40.707904	-74.01342

Here is a sample of venues data downloaded from FourSquare:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	10001	40.750742	-73.99653	New York Pizza Suprema	40.750124	-73.994992	Pizza Place
1	10001	40.750742	-73.99653	You Should Be Dancing... ..! / Club 412	40.750306	-73.994743	Dance Studio
2	10001	40.750742	-73.99653	Music Choice	40.752632	-73.994585	Music Venue
3	10001	40.750742	-73.99653	Madison Square Garden	40.750752	-73.993542	Basketball Stadium
4	10001	40.750742	-73.99653	Bluestone Lane	40.752068	-73.998848	Coffee Shop

Data Cleansing and Transformation

After merging the NYC Covid-19 data with the NY State geospatial data, it was discovered that four zipcodes did not have longitude and latitude mappings. These were removed from the dataset.

FourSquare imposed a limit of 100 venues per retrieval, hence the venues dataset may be somewhat incomplete. The retrieval operations fetched 14,766 venue items across all zipcodes in NYC.

One transformation was needed to make the data more useable. FourSquare provided a total of 456 categories of venues across NYC. This is too granular. These categories were condensed into the following 13 categories:

Category Name	Examples
Community	Community center, church, synagogue
Education	School, university, professional training center
Entertainment	Movie theater, music hall, museum
Exercise	Indoor sports center, gym, indoor tennis facility
Markets	Supermarket, grocery store, liquor store
Housing	Hotel, hostel
Merchandise	Arts and craft store, gift shop, department store
Outdoor	Parks, squares, attractions
Restaurant	Sit down restaurants, Italian, Japanese, African
Services	Bike rental, pet service, post office
Social_drinking	Bars, lounges, beer hall
Take_out_food	McDonald's, café, cake shop, take out burger
Transportation	Bus station, train station, airport

Additionally, due to a dearth of data points, three categories were eliminated: Community, Education, Transportation. The remaining 10 categories were used in the analyses.

After condensing the categories and using one-hot to create a separate column for each category, the data looks like:

	Zipcode	Community	Education	Entertainment	Exercise	Grocery	Health_beauty	Hotel	Merchandise	Outdoor	Restaurant	Services	Social_drinking	Take_
0	10001	0	0	4	7	0	1	2	5	0	4	0	2	
1	10002	0	0	2	0	2	0	0	3	0	10	0	6	
2	10003	0	0	0	6	4	0	0	4	1	8	0	2	
3	10005	0	0	1	3	1	1	1	1	1	7	0	2	
4	10006	0	0	1	3	2	1	1	2	6	3	0	1	

Lastly, the venues data will be transformed into per-capita data and percentage data through division by the population for each zipcode or by total number of venues in each zipcode. Population data is available from the NYC Covide-19 dataset. The resulting data is of the same format as above.

Methodology

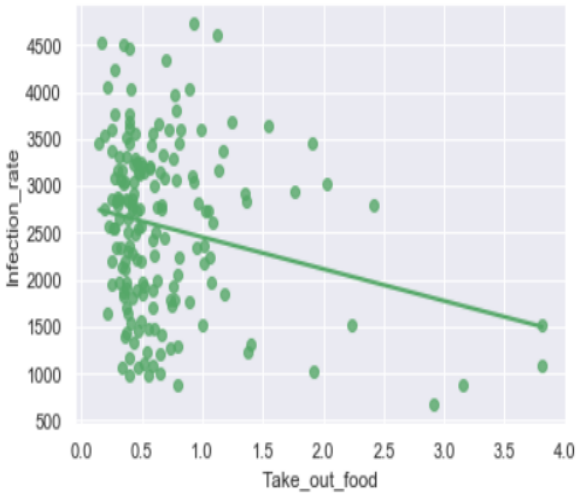
Three methods of analysis will be utilized:

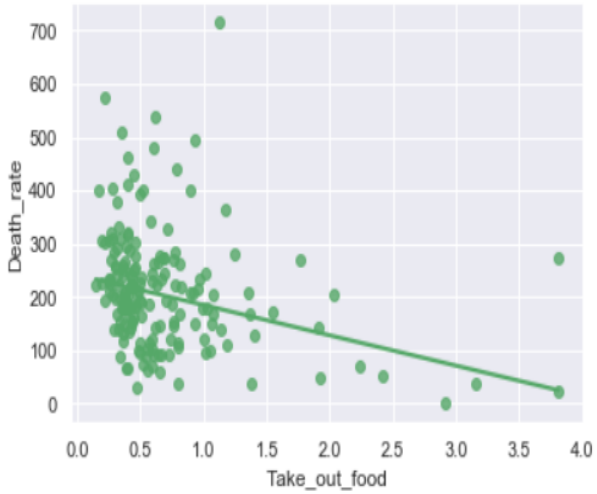
- Scatter plots for visualization
- Simple linear regressions of each category of venue data versus infection rates and death rates.
- Multivariate regression using a select subset of the ten venue categories versus infection rate and death rate.

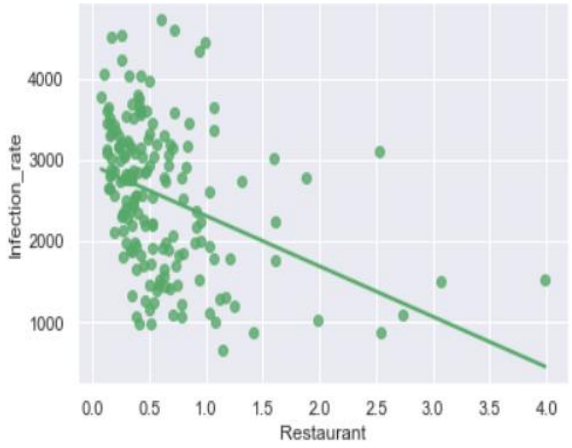
Note that the data is organized by zipcode.

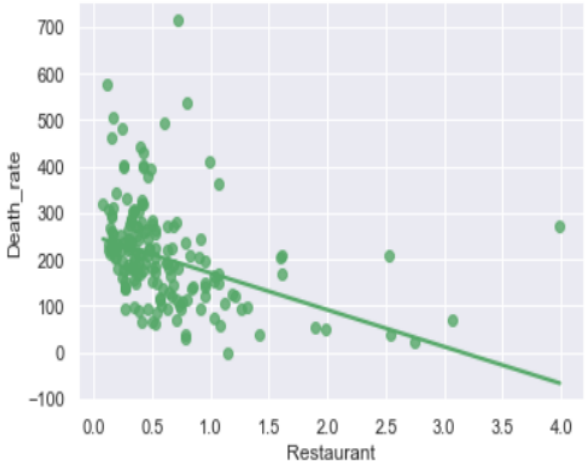
Scatter plots of per capita venue data

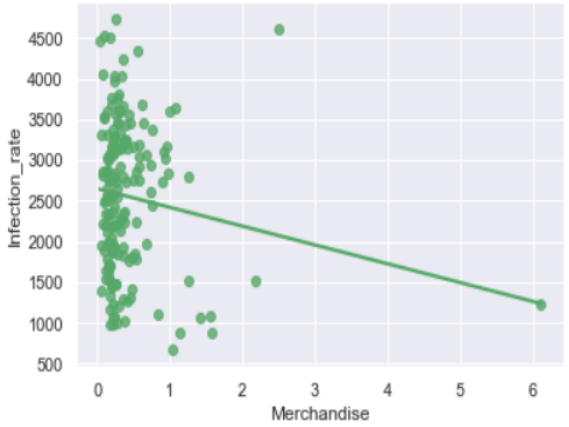
This sub-section presents scatter plots of per capita venue data versus infection rates and death rates. The x-axis labels show the venue categories. The order of presentation starts from the category with the most data points down to the category with the least data points. Note that the regression lines in the below plots are generated with the "robust" parameter set to True in order to minimize the effects of outliers.

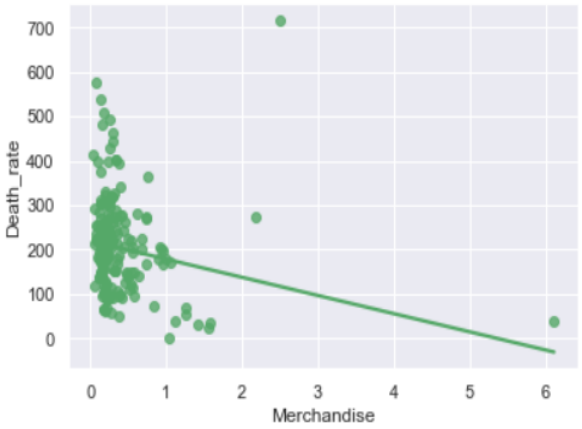


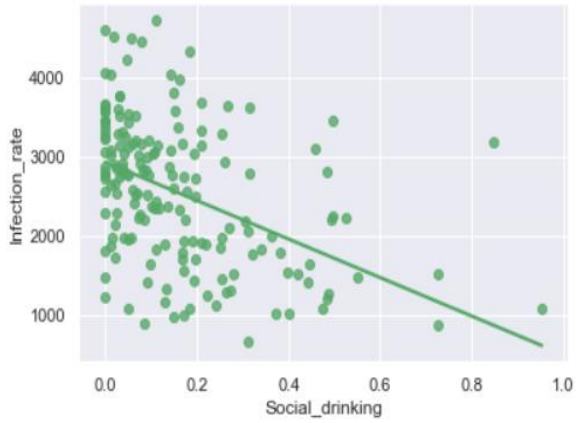


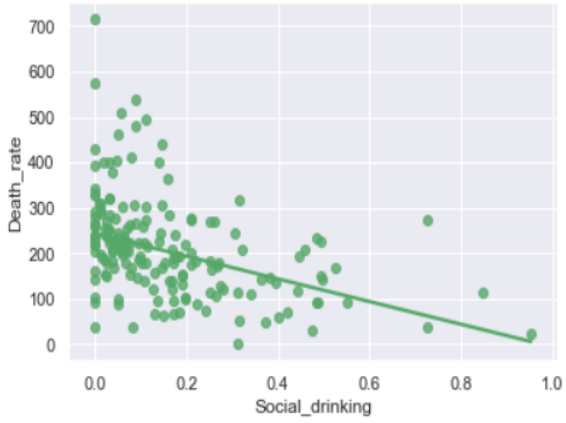

Infection_rate
4000
3000
2000
1000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Restaurant

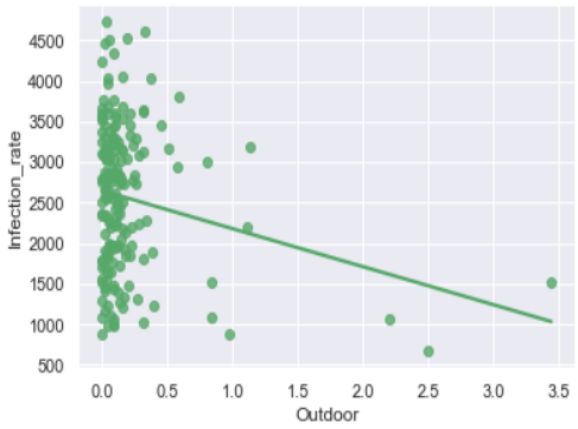

Death_rate
700
600
500
400
300
200
100
0
−100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Restaurant

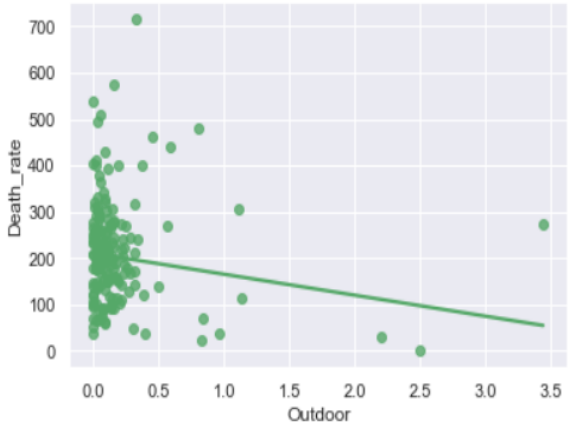

Infection_rate
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
Merchandise

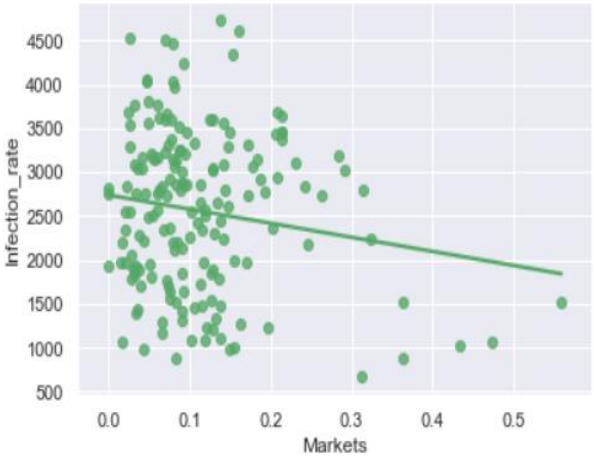

Death_rate
700
600
500
400
300
200
100
0
0
1
2
3
4
5
6
Merchandise

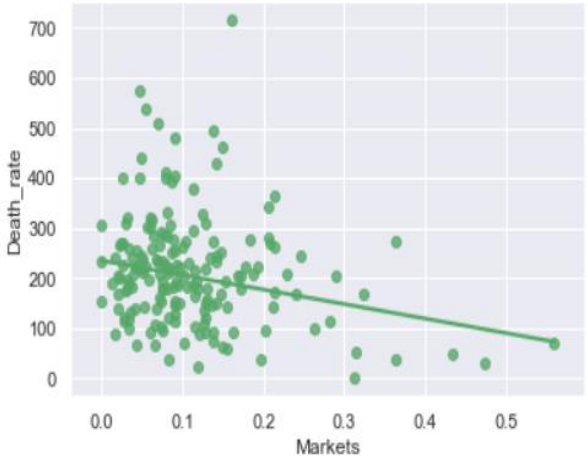

Infection_rate
4000
3000
2000
1000
0.0
0.2
0.4
0.6
0.8
1.0
Social_drinking

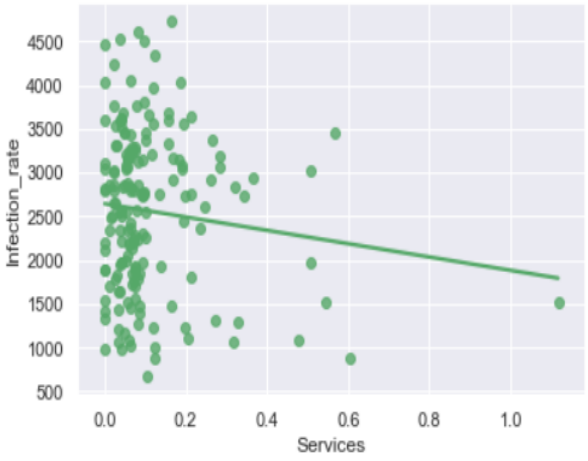

Death_rate
700
600
500
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0
Social_drinking

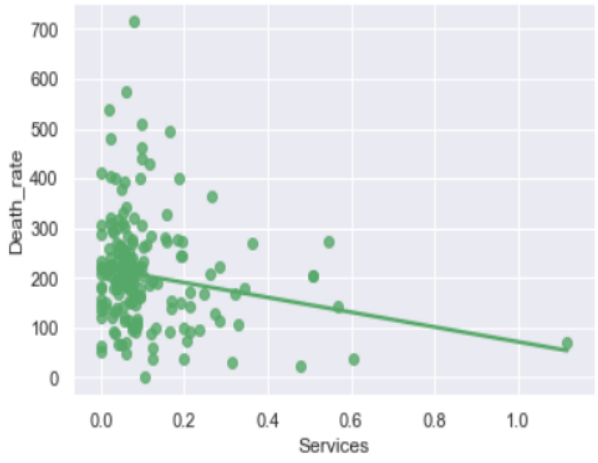

Infection_rate
Outdoor

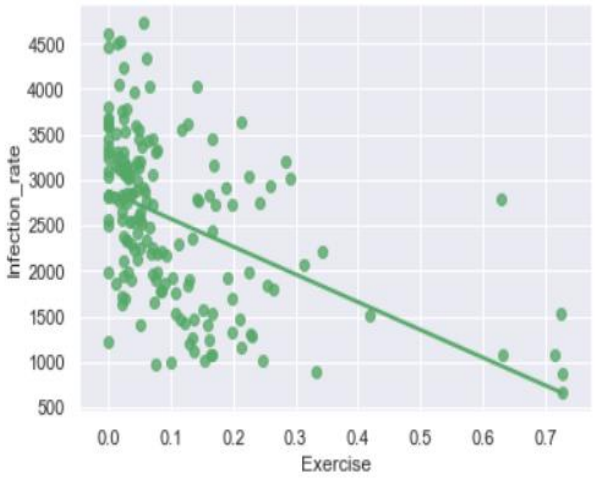

Death_rate
Outdoor

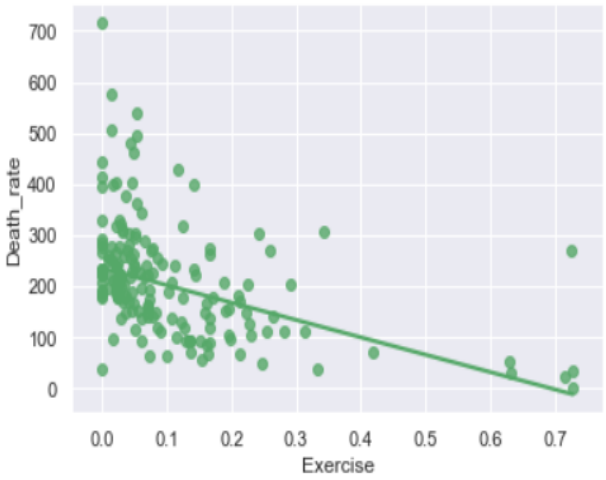

Infection_rate
Markets

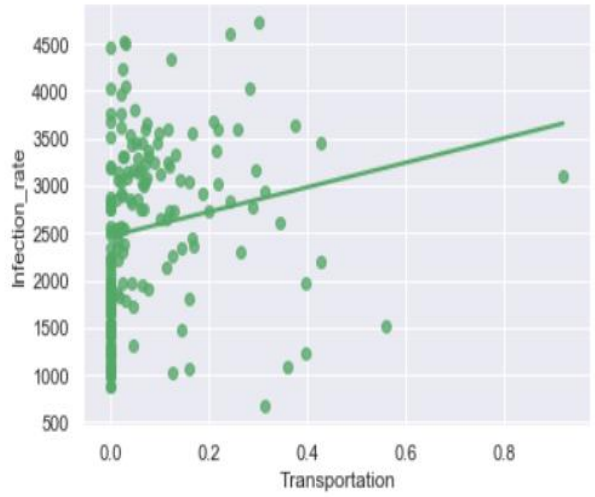

Death_rate
Markets

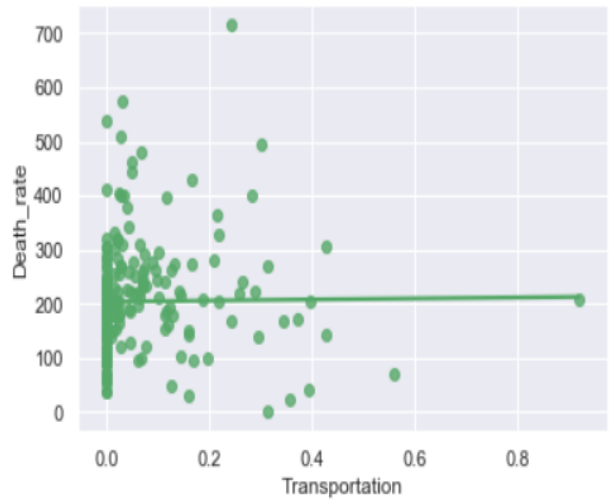

Infection_rate
Services

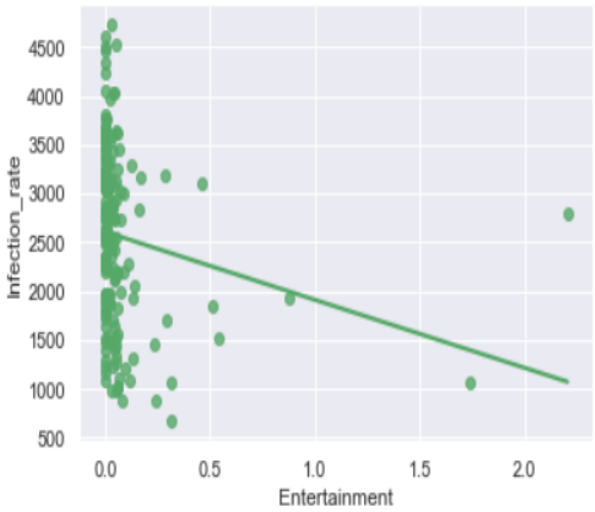

Death_rate
Services

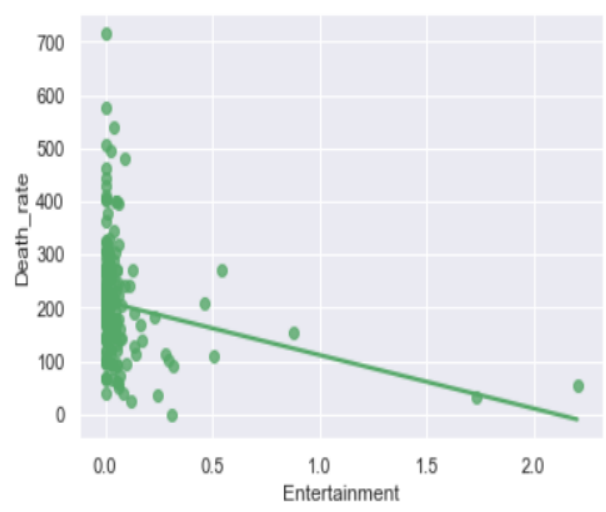

Infection_rate
Exercise


Death_rate
Exercise

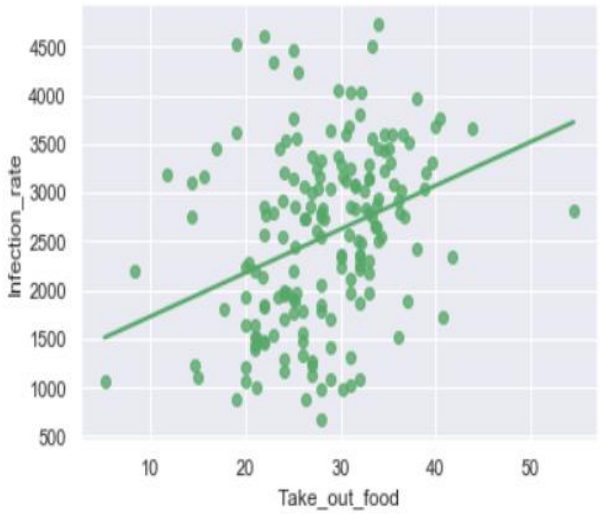

Infection_rate
Transportation

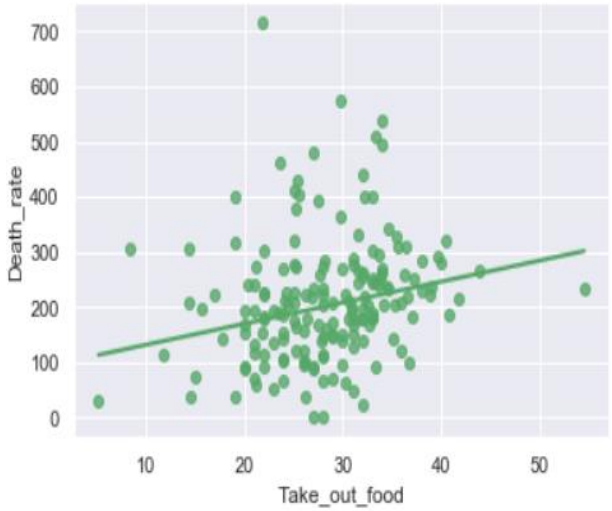

Death_rate
Transportation

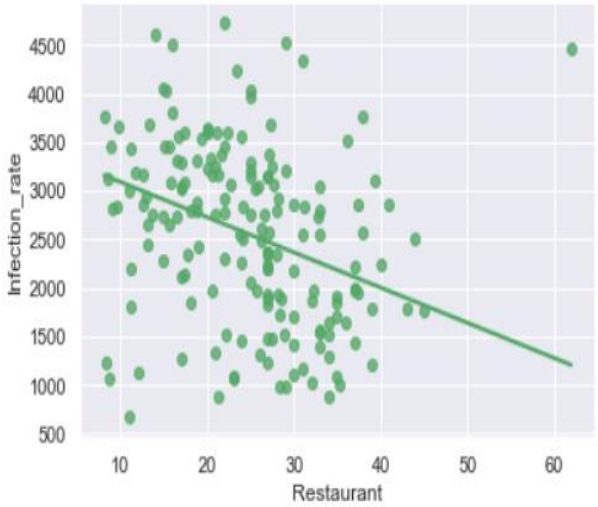

Infection_rate
Entertainment

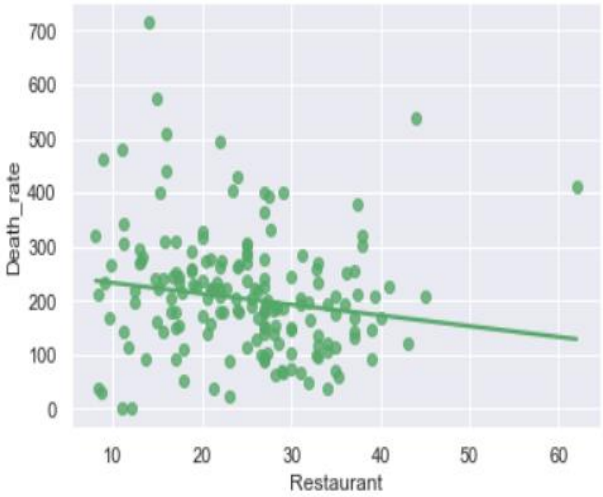

Death_rate
Entertainment

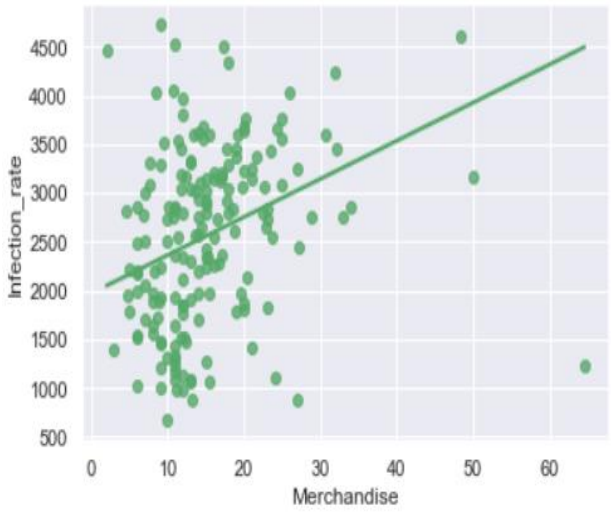
Scatter plots of percent venue data

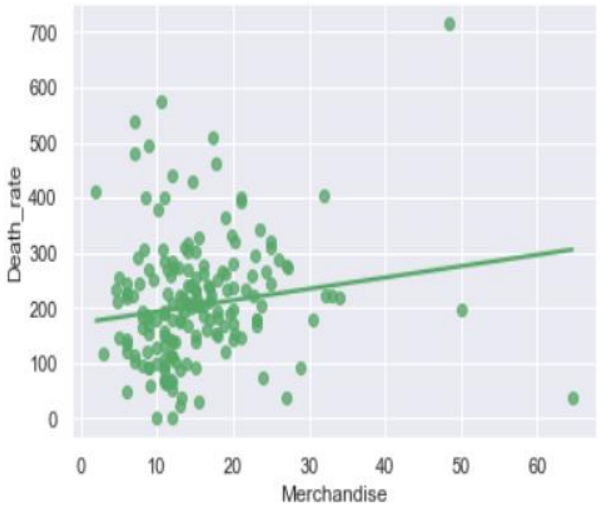


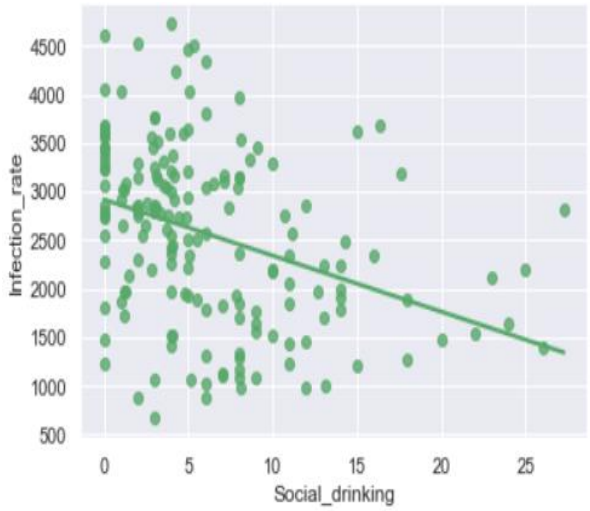


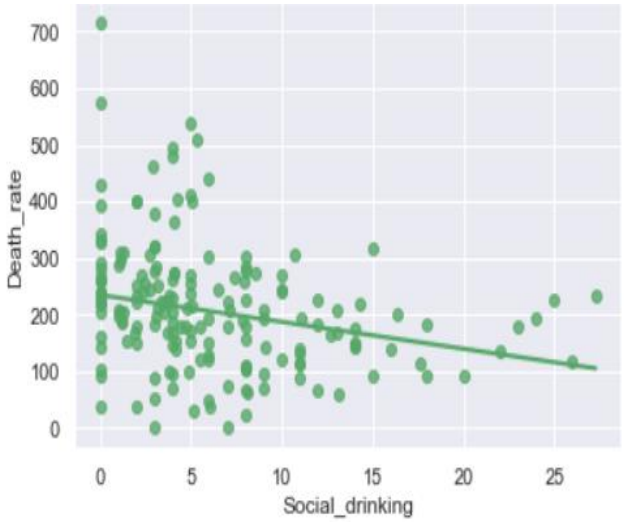




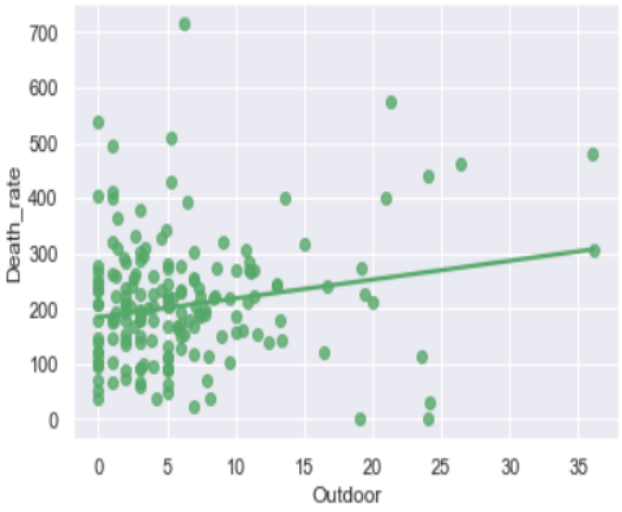


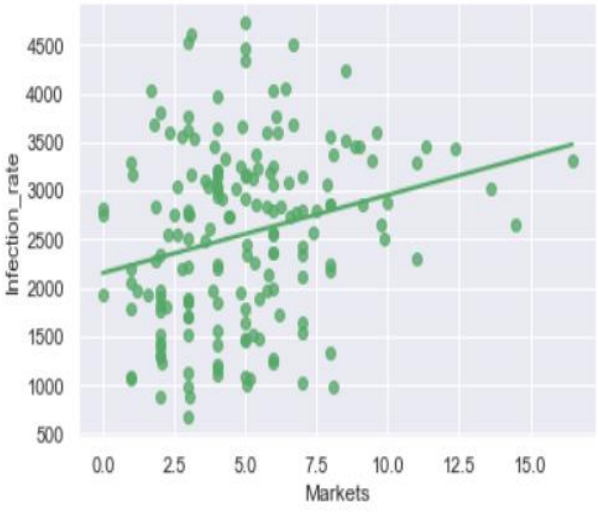


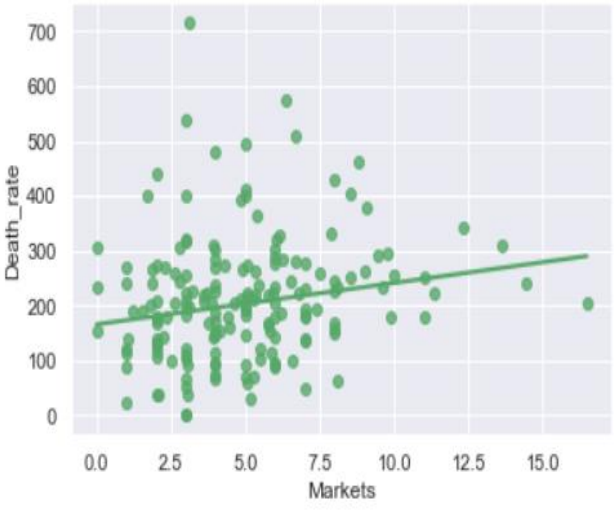

Infection_rate
Social_drinking

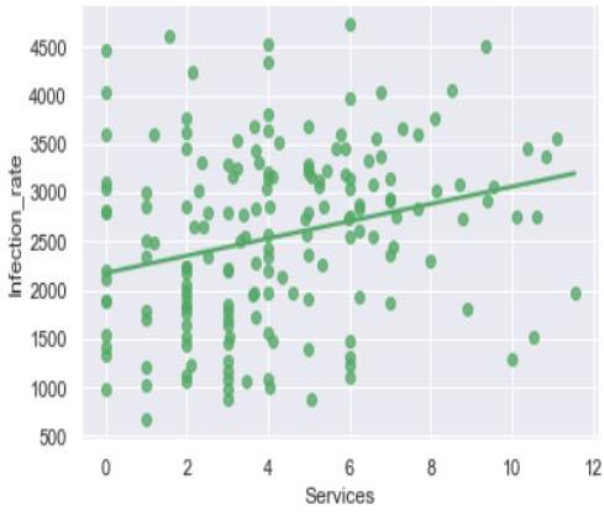

Death_rate
Social_drinking

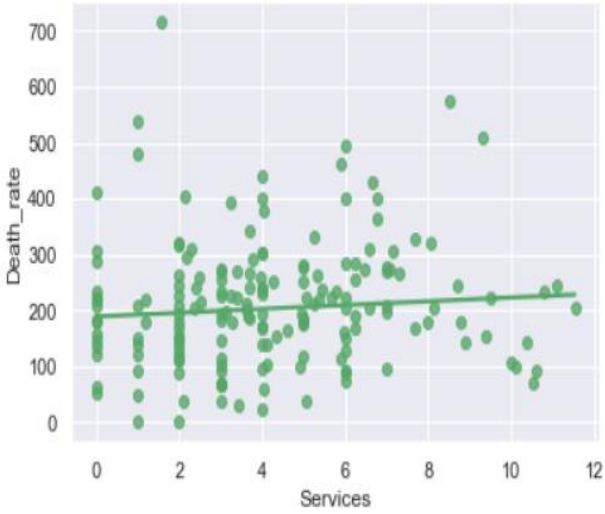

Infection_rate
Outdoor

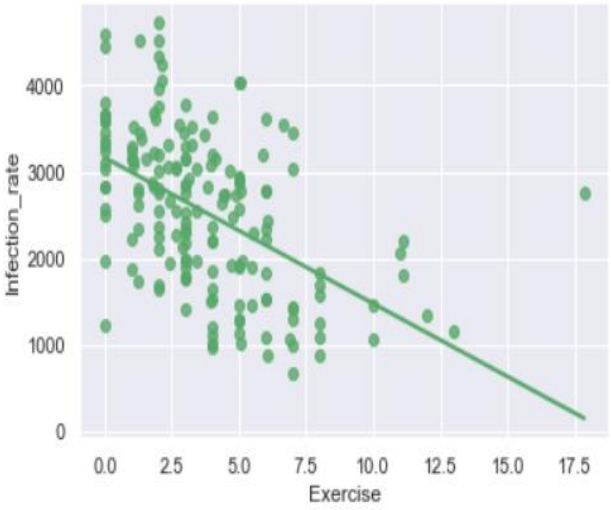

Death_rate
Outdoor

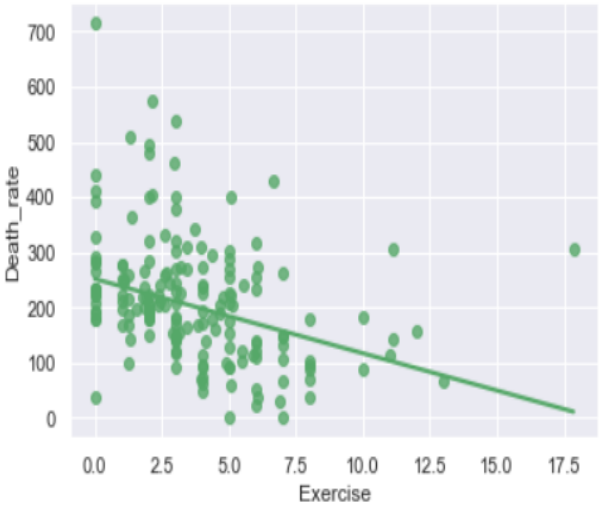

Infection_rate
Markets

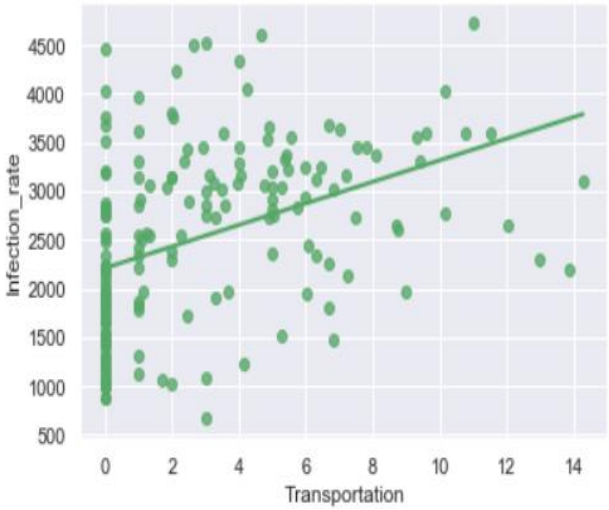

Death_rate
Markets

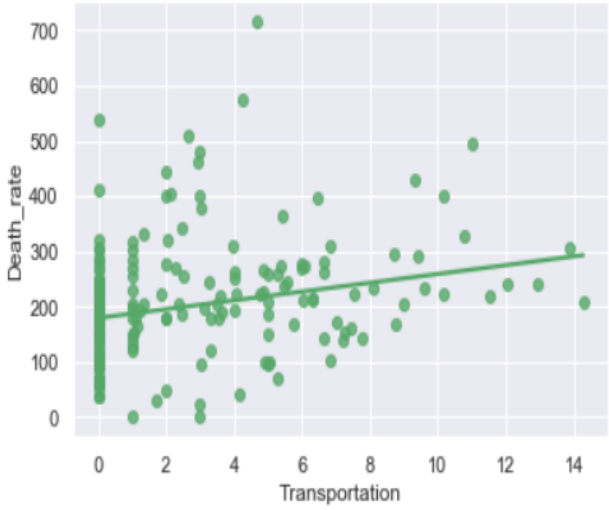

Infection_rate
Services

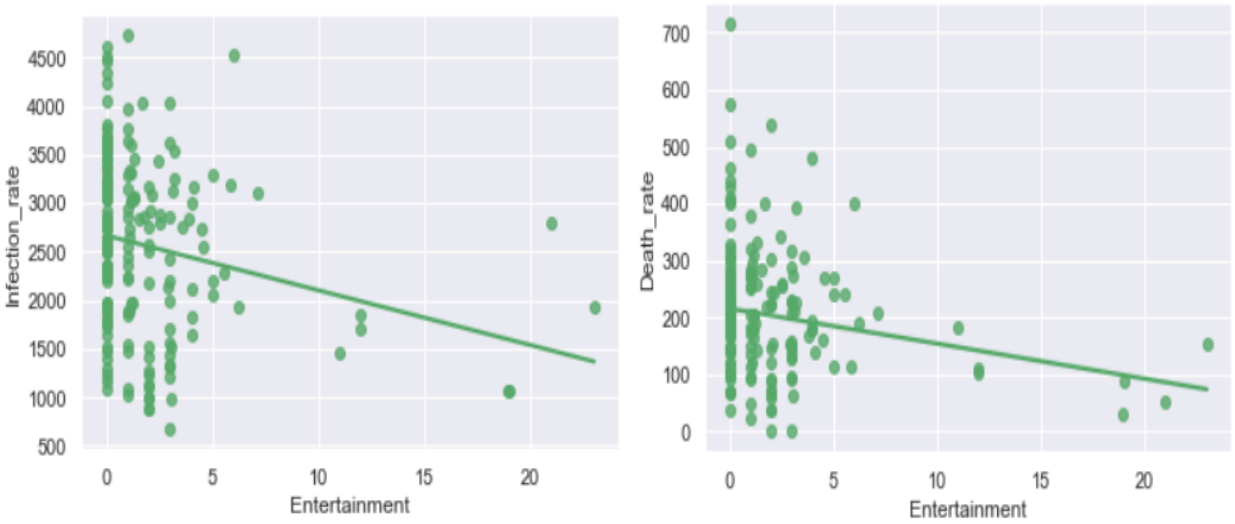

Death_rate
Services


Infection_rate
Exercise


Death_rate
Exercise


Infection_rate
Transportation


Death_rate
Transportation



Simple Linear Regression Analyses

The next analysis is to compute the regression statistics for the datasets shown in the above scatter plots. The table below shows the regression coefficients and R Squared values for each venue category vs infection rates (Inf) and death rates, for both per capita (PCap) data and percent (Pct) data.

	Venues	PCap Inf Coef	PCap Inf RSq	PCap Death Coef	PCap Death RSq	Pct Inf Coef	Pct Inf RSq	Pct Death Coef	Pct Death RSq
0	Entertainment	-651.82	0.03	-114.25	0.06	-56.71	0.05	-6.95	0.05
1	Exercise	-3007.04	0.21	-340.72	0.18	-153.52	0.23	-13.35	0.11
2	Markets	-1590.42	0.02	-295.62	0.05	75.71	0.05	7.23	0.03
3	Merchandise	-200.00	0.02	-28.90	0.02	26.69	0.06	1.79	0.02
4	Outdoor	-464.83	0.04	-28.21	0.01	7.38	0.00	2.87	0.03
5	Restaurant	-596.06	0.13	-67.37	0.11	-25.73	0.06	-1.58	0.02
6	Services	-758.44	0.01	-162.71	0.04	76.88	0.05	3.59	0.01
7	Social_drinking	-2204.13	0.17	-271.50	0.17	-54.41	0.12	-5.61	0.08
8	Take_out_food	-332.20	0.04	-50.18	0.07	40.56	0.09	3.30	0.04
9	Transportation	1074.26	0.02	15.03	0.00	103.94	0.15	8.55	0.07

Multivariate Linear Regression

The final analysis utilizes multiple venue categories in a multivariate regression to see if a statistically significant model can be found. The first step is to eliminate collinearity among the venue categories. This is accomplished by using the Variance Inflation Factor (VIF) method. The table below shows the VIF results:

	Venues	VIF Factor
0	Entertainment	1.494241
1	Exercise	3.070980
2	Markets	4.480379
3	Merchandise	4.131556
4	Outdoor	1.994871
5	Restaurant	7.258323
6	Services	3.827550
7	Social_drinking	2.979051
8	Take_out_food	12.487774
9	Transportation	2.496067

	Venues	VIF Factor
0	Entertainment	1.617017
1	Exercise	2.968173
2	Housing	1.424926
3	Markets	3.351609
4	Merchandise	3.510458
5	Outdoor	1.967708
6	Services	3.475242
7	Social_drinking	2.095293
8	Transportation	2.477045

The table on the left is the initial VIF run. It shows that the "Restaurant" and "Take_out_food" categories have significant collinearity characteristics, they should be removed. The table on the right shows the VIF results after the removal of the aforementioned venue categories. The VIF scores are now all under 5, a generally accepted cutoff value for VIF.

After performing the multivariate regression, the R Squared statistic were retrieved and shown below:

	Data Set	R Squared
0	Venues vs PCap Infection Rate	0.428951
1	Venues vs PCap Death Rate	0.334048
2	Venues vs Pct Infection Rate	0.402567
3	Venues vs Pct Death Rate	0.272350

Results, Observations, Discussions, Recommendations

- Systematically negative correlations between per capita venue data and infection / death rates:

  Nine out of 10 per capita venue categories exhibit negative linear relationships versus Covid infection and death rates. The only exception is the Transportation category. This is somewhat counter-intuitive. For example, if there are more Social_drinking venues per capita in a zipcode neighborhood, shouldn't infection and death rates higher. Furthermore, this behavior cuts across almost all venue categories.

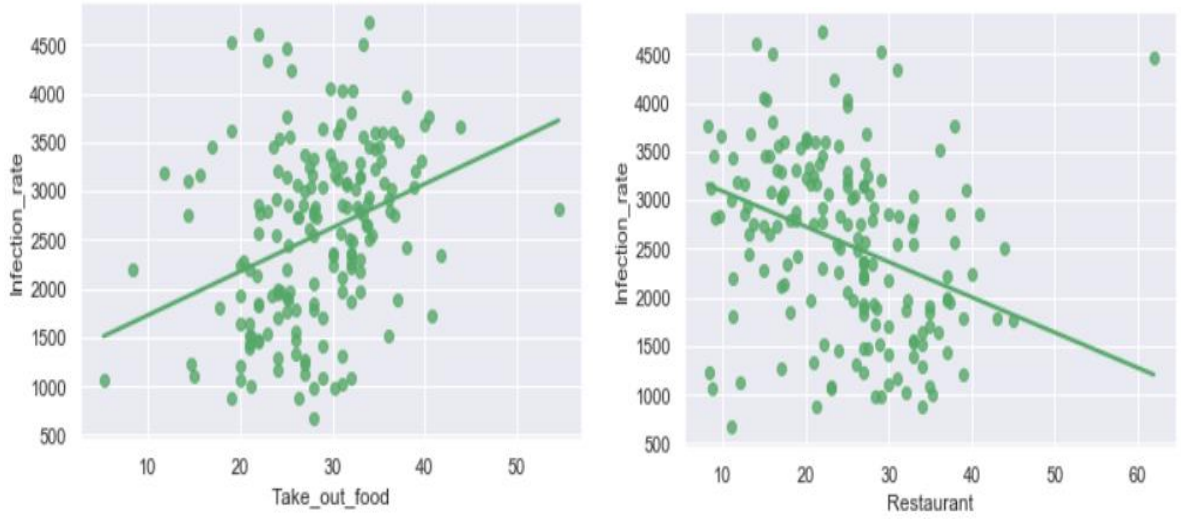
One possible explanation is that when there are more venues per resident (i.e., per capita) then there would be fewer people in each venue at any given time. People have more choices, so there is less crowding of venues. In other words, when there are more venues per resident, there is less density in each venue. Less density is associated with lower rates of infections which lowers the death rates as well.

For policy makers, this information can be useful when making re-opening or shutdown decisions on a local level. Governor Cuomo has said repeatedly that decisions must be made at the lowest levels of government where specifics of each locale's situations can be taken into consideration.

- Percent venue data yielded more differentiating results:

Percent venue data is compiled by dividing each venue's count within a zipcode by the total of all venues in the same zipcode. Surprisingly, when venues data is cast in this form, an added dimension of insights availed themselves.

For example, take a look at the two scatter plots below. The left plot is Take_out_food vs infection rates. The right plot is Restaurants vs infection rates.



Take out food venues is positively correlated with Covid infection rates, but Restaurants show a negative correlation to infection rates. Both categories are food venues, but they impact Covid in opposite ways.

Take out food venues are high volume, low cost, food options. People spend a minimal amount of time (low duration) in these venues, yet the more (percentage wise) of these there are the higher the infection rate! Perhaps the high traffic of take out venues lends itself to less cleanliness, and the smaller spaces may contribute to higher transmission risk (both higher risk for each individual and potential for larger number of people to get infected due

to higher volume of traffic). Lastly, take out food is associated with unhealthy life-style choices. Therefore, higher percentage of this type of venue in a neighborhood implies generally lower health levels of residents. Lower health leads to higher risk of infections and deaths.

Restaurants are sit-down eating venues where there is lower volume of traffic and higher cost which should lead to lower infection rates. However, patrons of sit-down restaurants also spend more time on premise (mostly indoors), often exceeding one hour (before Covid) which implies a higher infection rate should be the case. But, the data does not support that conclusion. Perhaps restaurants are cleaner, or perhaps restaurant patrons have higher income therefore they have better health care and are in better health.

Let's look at all venue categories and separate them by positive versus negative correlation to infection and death rates:

- Positively Correlated: Take_out_food, Markets, Merchandise, Services, Transportation. Higher percentage of these venues is associated with higher Covid-19 rates.

- Negatively Correlated: Entertainment, Exercise, Restaurants, Social_drinking. Higher percentage of these venues is associated with lower Covid-19 rates.

Let's extract common traits amongst venues of each group. Members of the Positively Correlated group share these traits: high volume, short duration, low cost. Members of the Negatively Correlated group share these traits: low volume, long duration, high cost.

Explanations for these traits and their relationships to Covid following those outlined for Take_out_food and Restaurants. Nevertheless, it is surprising and counter-intuitive that short duration is positively associated with Covid while long duration is negatively associated with Covid. Perhaps it is the types of people who visit these venues that makes the difference. More research on this would be recommended and it is potentially very impactful.

For policy makers, these findings can guide the shut-down and re-opening of neighborhoods. When closing or opening venues, it would be beneficial to move the mix towards the Negatively Correlated group.

It is worth noting that the Social_drinking category is in the Negatively Correlated group. This category includes bars which have been identified by public health officials as high risk and prioritized for shutdown. This results in this study indicate that when drinking venues are reduced through forced shutdown, Covid infection rates may increase. However, there are numerous documented cases where rapid community spread occurred in bars. So this particular type of venue will require more careful scrutiny.

- Simple regressions have low R Squared values:

  The table in the "Simple Linear Regression Analyses" section above shows that regressing individual venue categories against infection and death rates yielded very low R Squared values. Values are generally below 0.10 with a few exceptions. This should be noted when using the results of these analyses.

- Multivariate regression yields significant improvement in R Squared:

  It is expected that more independent variables lead to higher explanatory capability. The highest multivariate R Squared is 0.42 while the highest R Squared for an individual venue is 0.23 for the Exercise category. More research should be performed to identify the group of venue categories that offer the highest explanatory capability. Policy makers can use this information to prioritize venues for re-opening or lock-down.

Conclusions

Several surprising and counter-intuitive results came out of this analysis. These are the two most surprising:

1. Neighborhoods with a high per capita number of venues, regardless of the type of venue, have lower Covid infection and death rates.
2. The mix, or composition, of venues is important. Neighborhoods with higher percentage of positively correlated venues have higher Covid infection and death rates. These are venues with high traffic, low duration, low cost (e.g., take out food). On the other hand, negatively correlated venues can help to reduce Covid rates. These are low traffic, high duration, high cost venues.

When it comes to stopping Covid-19, there is nothing better than a complete lockdown. Any contact between people is a chance for Covid-19 to spread. Therefore, the findings in this study is not to suggest that we should open up more venues. Rather, these findings can be most effectively used in a gradual re-opening or a gradual lockdown should that be necessary again.

Finding #1 indicates that increasing the number of venues (of all types) available to each person can reduce the infection and death rates. This is because of reduced venue density. However, Finding #2 tells us that certain types of venues (negatively correlated) help to reduce infection rates, therefore these venues should be prioritized.

Said another way, it is important to pay attention to the mix of venues available in a neighborhood. When re-opening or gradually locking down, we should attempt to increase the percentage of negatively correlated venues.

More analyses should be performed on the findings in this study before the concepts are put into practice. Hopefully these findings will lead to a more safe and efficient approach to re-opening the economy.