Data Science Capstone Project
Coursera | IBM

Using location data to understand the dynamics of COVID-19 spread

Giovanne B. Diniz, Ph.D.

This work is intended as the Capstone Project of the IBM-Coursera Data Science Professional Certificate Course

Abstract

As the COVID-19 pandemic develops, local and federal governments scramble to understand, limit, and mitigate its effects. Most countries adopted an economic slowdown approach, encouraging people to stay at home and limiting their movement. One of the most important questions that need to be addressed, therefore, is how to restart the economy without creating a new wave of public health crises. In the United States, a country of continental proportions, states are the main agents of change with respect to the economy, but that does not necessarily mean a one-fits-all approach for each state is the most efficient way to reopen. Therefore, in this study, the spread of COVID-19 was evaluated in a county by county basis, and the relationship between disease spread and the types of venues found in each county was taken into consideration. Using both regression and classification algorithms, a relationship between venues and the spread of the disease was found. The relative frequency of venues for each county proved itself a reliable indicator of the infection rate, suggesting this type of location data can be used to make better informed decisions about reopening and to prevent surges of infection. When applied to a different state, however, the models underperformed, suggesting a state- or regon-specific relationship between venues and COVID-19. The results obtained in this study may be of interest to governmental bodies and policymakers when gauging how to use data science to inform their decisions.

Summary

Introduction..... 3

Data..... 4

Methods..... 5

Results..... 8

Discussion..... 17

Conclusion..... 20

1. Introduction

The 2020 COVID pandemic is an event of historical proportions. In addition to its life toll, with more than 2.7 million people infected globally and over 190'000 fatalities (as of 04/24/2020)[1], the pandemic has imposed an extreme financial toll resulting from the necessary social distancing and consumer behavior changes. The United States is one of the hardest hit countries[2], with an economic impact that exceeds trillions of dollars[3], millions of job losses[4], and expected bankruptcies[5]. The continental proportion of the US, its large population, and its heterogeneity, led to multiple shelter-in-place orders and the complete stagnation of some states, while others were less affected[6].

Given that most of the economic activity had to be shut off, and the cost to keep it that away, the United States Federal government released a plan to reopen the country in three stages, going from lower risk activities to those that carry a larger risk to public health[7]. Despite these guidelines, the decision to and how to reopen was largely left to the states[8]. Not only it may occur in phases, but states have the power to also open specific counties and cities at different times. The question of how to restart the economy, therefore, is one of the most important decision each state will have to make.

A complicating factor in making reopening decisions is the lack of uniformity in the data. States and counties within states are not testing people at the same rate or following the same guidelines, and there is no guarantee that cases are being efficiently reported by the local health providers to the government, creating substantial heterogeneity in the number of cases and epidemiological data obtained[9]. This implicates that economical decisions made uniquely based in number of cases and other simple epidemiological data are likely to be biased, imperfect, and therefore, dangerous.

Given the above, finding auxiliary data that can help us understand the spread of COVID-19 would help feed models of reopening, allowing states to make more effective decisions in terms of what business and areas can be allowed to return to operation sooner. One possible data that could feed these models is venue data, here understood as parks, open areas, commerce, services, and buildings found in a determined location. Therefore, in this project, we will test the hypothesis that venue data can be used to predict the susceptibility of a county to COVID-19 spread regardless of the testing performed by that state.

[1] Source: https://en.wikipedia.org/wiki/Template:2019–20_coronavirus_pandemic_data
[2] Source: https://www.sciencemag.org/news/2020/04/united-states-leads-coronavirus-cases-not-pandemic-response
[3] Source: https://www.washingtonpost.com/business/2020/03/25/trump-senate-coronavirus-economic-stimulus-2-trillion/
[4] Source: https://www.washingtonpost.com/business/2020/04/16/unemployment-claims-coronavirus/
[5] Source: https://www.businessinsider.com/coronavirus-could-trigger-retail-bankruptcies-and-mass-store-closings-2020-4
[6] Source: https://www.bbc.com/news/world-us-canada-52103066
[7] Source: https://www.politico.com/news/2020/04/16/trump-reopening-plan-191959
[8] Source: https://www.politico.com/news/2020/04/16/trump-plan-for-reopening-economy-191073
[9] Source: https://www.washingtonpost.com/health/2020/04/21/kentucky-rhode-island-coronavirus-testing/

2. The data

To test our hypothesis, we will use geolocation data (based on zip codes), venue data (obtained through the Foursquare API), and epidemiological data for COVID-19. For model development, we will use the state of New York. The state of New York has several characteristics that make it a desirable target for our preliminary investigation. First, New York is the epicenter of COVID infections in the United States, alone corresponding to more than a quarter of all infections[1]. This led to a massive testing effort, with some of the most comprehensive data found in the country[10]. Second, in addition to New York City, New York state has a huge upstate area with numerous counties, allowing us some of the largest datasets possible. Third, New York state has a comprehensive and well-polished database system, with detailed populational[11], geographical[12], and epidemiological[13] data that can be obtained through the Socrata API (https://dev.socrata.com). Therefore, the data we will need can be easily imported and manipulated in the program.

Epidemiological data and populational data will be used in exploratory analysis to verify some of the premises of our hypothesis. In particular, we want to verify if: there is substantial variability between counties in the number of infected, even when corrected by population; there is substantial variability between counties in the number of tested, even when corrected by population; and if there is substantial variability between counties in the rate of infected (positive cases/tested cases). Substantial variability between counties is a prerequisite for this model. We will also verify correlation factor between these parameters, to understand how population density can affect the spread of COVID-19. If population correlates too strongly with COVID-19 infection, it will eclipse and render our model irrelevant.

If the assumptions for our model are verified, we will obtain geolocation data using zip codes as proxy. While NY city has detailed information for neighborhoods, including geolocation information, the rest of the state does not have such granular data. Therefore, we will use zip codes to obtain latitude-longitude pairs for each zipcode, verify data integrity, and devise a strategy to ensure maximum area coverage for each of these points while minimizing overlapping. Once these "query" areas are obtained, we will employ the Foursquare API (https://developer.foursquare.com) to obtain the types of venues found in each area, quantify them, and then obtain the relative presence of each venue type per zip code and per county. The relative presence of each venue will then be used as independent variables (feature set), while either the rate of infection (continuous variable) or a risk classification (categorical variable) will be used as dependent variables (target set). Models will be trained and tested using NY data and evaluated based on accuracy and recall metrics. Finally, if a model is found that has reliable scores, it will be trained on the whole set of NY data and then tested against data from another state (Illinois[14]) to evaluate how transposable the model is between states.

[10] Source: https://www.politico.com/interactives/2020/coronavirus-testing-by-state-chart-of-new-cases/
[11] Source: https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k
[12] Source: https://data.ny.gov/Government-Finance/New-York-State-ZIP-Codes-County-FIPS-Cross-Referen/juva-r6g2
[13] Source: https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Testing/xdss-u53e
[14] Source: https://www.dph.illinois.gov/covid19

3. Methodology

3.1. Data acquisition – model development

For model development, data was obtained from the NY Open Data Portal (https://data.ny.gov/) using the Socrata API. The whole "xdss-u53e" database was obtained from health.data.ny.gov using a non-tokened Socrata client, consisting of the following columns: "Test Date", "County", "New Positives", "Cumulative Numbers", "Total Number of Tests", and "Cumulative Number of Tests". The database is updated daily with the latest testing data. Since we are not interested in the day-to-day data, the database was filtered to only show the most recent cumulative data. To correct the data for population, we imported the whole "e9uj-s3sf" database from health.data.ny.gov, consisting of columns "Year", "Age Group Code", "Age Group Description", "Gender Code", "Gender Code Description", "Race Ethnicity Code", "Race/Ethnicity Description", "County Code", "County Name", and "Population". Since we are not interested in demographics, the dataset was filtered to only show data for the latest year for the total of the population, per county.

Since there is no single database with neighborhood data for the whole state of New York, zip codes were employed as proxies for neighborhoods. As described above, the whole "juva-r6g2" database was imported from data.ny.gov, consisting of "County Name", "State FIPS", "County Code", "County FIPS", "Zip Code", and "File Date". The database was filtered for "County Name" and "Zip Code" and saved as a data frame. All data frames used in the study were tested for null values and entire rows were dropped if a null value was found. When applicable, duplicate values were also dropped and/or merged. All the data manipulation performed in this study are publicly available at: https://github.com/GBaroni/Coursera_Capstone/blob/master/NY%20COVID%20Project.ipynb.

3.2. Data Exploration

Preliminary exploration of the data was performed using both epidemiological and populational data. In addition to the cumulative value of positive cases and tests, three metrics were obtained for each county: number of positive cases per 100.000 people (Positive/Population), number of tests per 100.000 people (Tests/Population), and ratio of positives over tests (Positive/Tests). Bar plots were used to assess variability between counties for each of those metrics. To further understand the data, we obtained Pearson standard correlation coefficients for all metrics to visualize interactions between the data. Given the close relationship between the metrics employed in this study, we expect a high correlation between multiple factors.

3.3. Geolocation data

For the location querying, latitude/longitude (LAT/LONG) coordinates are necessary. To obtain LAT/LONG data, the pgeocode 0.2.1 library was employed (https://pypi.org/project/pgeocode/). For NY state zip codes, the pgeocode library returned LAT/LONG coordinates for every zip code, but 528 LAT/LONG coordinates were found for more than one zip codes. To prevent overlapping in location queries, these duplicated zip codes were dropped, reducing the number of zip codes to 1684.

In addition to LAT/LONG coordinates, queries are made based on a radius, in meters. When there is substantial homogeneity between locations, a standard radius can be used. In this study, zip codes cover a wide range of areas, varying from very small neighborhoods, where different zip codes are found in proximity, to large areas of woods or farmland covered by a single zip. Therefore, a decision was made to refine the dataset and obtain custom radius for each zip code. Zip codes within 2km of another zip code (resulting in a radius < 1000m) were merged into a single zip code centered

in the Euclidian midpoint. To do that, a matrix of distances between each zip code was generated using the geopy 1.21.0 library (https://pypi.org/project/geopy/). Although the pgeocode library also offers distance calculation between LAT/LONG coordinates, the time to calculate distances was substantially higher for pgeocode than for geopy, resulting in substantial delays. Since large numbers of distances had to be calculated iteratively, the geopy library was chosen. To minimize computational time, instead of generating a new distance matrix every time two zip codes were merged, all pairs within 2km of each other were flagged, merged, and then a new matrix were computed and the process were repeated, until no zip codes were found within 2km of each other.

To obtain custom radii, a similar strategy was employed. A distance matrix was used to find the smallest distance between two zip codes, and half the distance was assigned to these zip codes as their radius. The next smallest distance was then identified, and new radius assigned, until all zip codes received a radius value, in meters. It's important to note that this method do not guarantee zero overlap between zip codes, but it substantially reduces the number of overlapping venues found.

3.4. Venue data

To obtain venue data for each location, the Foursquare Places API (https://developer.foursquare.com/docs/places-api/) version 20180605 was used. For each zip code-latitude-longitude-radius quartet an "explore" endpoint was called, returning all venues in the foursquare database within "radius" from the LAT/LONG provided. The return of each query was saved in a data frame and the whole response was saved a json file to prevent the need to call the API every time the code is run, since there is a daily limit to the number of calls a free plan can perform. The obtained data was evaluated in terms of number of returned venues for each zip code and duplicated. Instead of using individual venue data, the type of venue was computed and hot encoded. In the resulting dataframe, each zip code was assigned the relative contribution of each venue type to the total of venues found for that zip code. Finally, data for each zip code was grouped by county, and the ratio of positive cases/tests were added at the end of the data frame. The resulting dataframe (and variations) were used for model training and selection. In addition to modelling continuous data, the possibility of classifying counties based on risk group was also explored. To do that, counties were split into two, three, or four risk categories, with different delimiters, and the model weighed accuracy and recall were evaluated for each binning strategy.

Geolocation data was visualized using the Folium 0.5 library (https://python-visualization.github.io/folium/).

3.5. Model development

Several algorithms were used to model the data in this study, predominantly through the scikit learn library (https://scikit-learn.org/stable/). For prediction of the infection rate, linear regression was employed, with and without multicollinearity mitigation through ridge regression. Least absolute shrinkage and selection operator (Lasso) regression was also performed for feature selection using regularization. The performance of each model was tested both with standardized and non-standardized data. Given the small number of samples, the leave-one out cross-validation (LOOC) method was employed. For regression models, R^2 score was the main metric used for evaluation. The recall score of the model was also taken into consideration, given that classifying a higher-risk county as low-risk would have higher public health costs than classifying a lower-risk county as high risk. Model optimization was performed for ridge regression, with different values of regularization strength (alpha) used [10^{-3} - 10^{3}].

For classification, four algorithms were utilized: K nearest neighbors (KNN), decision tree (DT), support vector machine (SVM), and logistic regression (LogR). As described above, LOOC was employed for hyperparameter optimization and evaluation. For KNN, train-test sets were used to obtain the optimal number of neighbors. For SVM, four kernels were tested: linear, polynomial, radial basis function (rbf), and sigmoid. For LogR, both solver (newton-cg, lbfgs, liblinear, sag, and saga) and regularization strength (C: 10^{-2} - 10^{2}) were varied for optimization. In all cases, F1-score, Jaccard score, accuracy, and recall were evaluated. Logarithmic loss was also employed for LogR models. In specific cases, confusion matrices were also employed to better understand the model predictions.

3.6. Clustering

In addition to evaluating venue data for the whole county, the possibility that neighborhood profiles (rather than individual data) would be a better predictor was also entertained. Therefore, k-means clustering was used to group each zip code in different neighborhood clusters (the number of clusters was also varied to optimize results). Relative density of each cluster type was then used with the aforementioned models to predict risk classification.

3.7. Applying the models

To test if the model developed for one state can be applied to predict risk in other states, data from Illinois was used. As is the case with New York state, Illinois is a good case study for having test data available, similar territorial extension, and the presence of a large metropolis and rural counties. Epidemiological data was obtained from the Illinois Department of Public Health (https://www.dph.illinois.gov/covid19/covid19-statistics), and zip code data for Illinois was obtained from https://www.zip-codes.com/state/il.asp. Data was treated as described above, and after training the model on NY state data, the models were used to predict high-risk counties in Illinois based on venue frequency.

4. Results

4.1. Epidemiological data

New York state is comprised of 62 counties. The number of cases varied widely between counties, ranging from as low as 3 positive cases in Hamilton and as high as 53'640 cases in Queens, and an average of 5103 cases. The number of tests also varied substantially between counties, ranging between 44 and 132'650 tests, with an average of 15'901 tests performed. See Table. 1 for more descriptive statistics on this dataset. Given there was also substantial variability in population (Table. 1), it was necessary to investigate the relationship between population and COVID-19 spread.

Table 1

	Positives	Tests
count	62.000000	62.000000
mean	5103.467742	15901.790323
std	12384.411665	33173.865316
min	3.000000	44.000000
25%	62.000000	885.250000
50%	151.000000	1613.500000
75%	997.750000	8173.500000
max	53640.000000	132650.000000

Descriptive statistics of the epidemiological data available for New York State as of 05/03/2020.

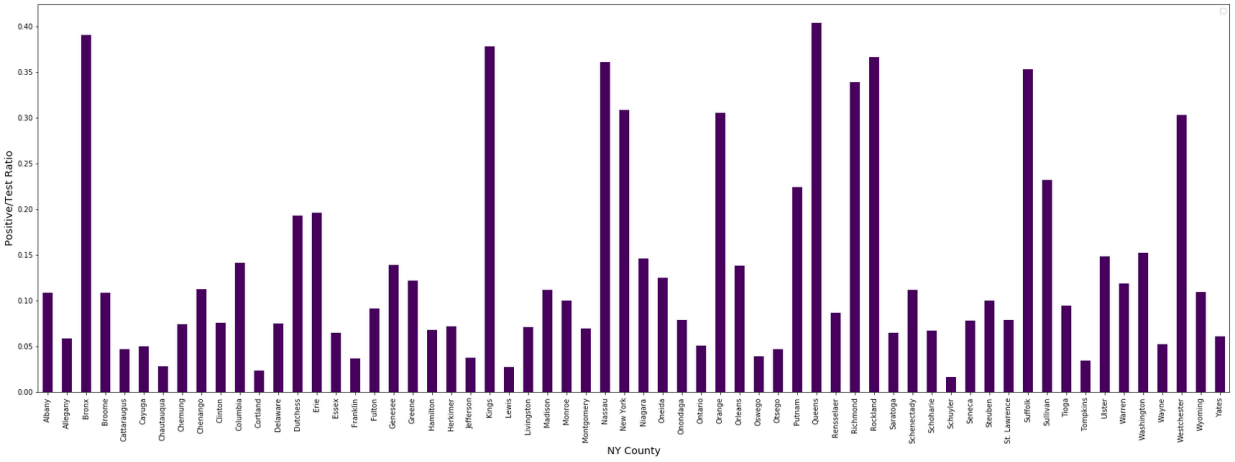
Table 2

	Positive/Population	Tests/Population	Positive/Tests
count	62.000000	62.000000	62.000000
mean	587.912257	2924.586209	0.133531
std	896.351984	2201.068169	0.107321
min	27.897029	798.184521	0.016393
25%	89.735854	1561.353143	0.065067
50%	193.434666	2008.328919	0.097389
75%	386.854634	3416.773656	0.147924
max	3656.482236	10059.740228	0.404372

Descriptive statistics of the epidemiological data available for New York State as of 05/03/2020.

All three metrics varied substantially between counties: positive/population – 587.91 (896.35) per 100'000 people, tests/population - 2924.58 (2201.06) per 100'000 people, and positive/tests ratio – 0.1335 (0.1073) (Table 2). The heterogeneity in epidemiological data between counties is best illustrated by Fig. 1 These observations confirm there is substantial variability between counties within a state, one of the requirements for a model similar to the one developed here to be deployed in a meaningful way. The second requirement is some degree of uncoupling between population and COVID-19 spread.

Figure 1



Ratio of positive cases over the number of tests per county.

The Pearson coefficient of correlation was obtained for the epidemiological metrics (Table 3). While a high correlation was observed between all metrics, as expected, the most relevant is the high correlation between population and the ratio of COVID-19 infection (0.782), because that stat indicates that population can be used as a predictor of infection rate. When splitting counties based on population, substantial variation was observed in the Pearson correlation factor (see notebook for more details).

Table 3

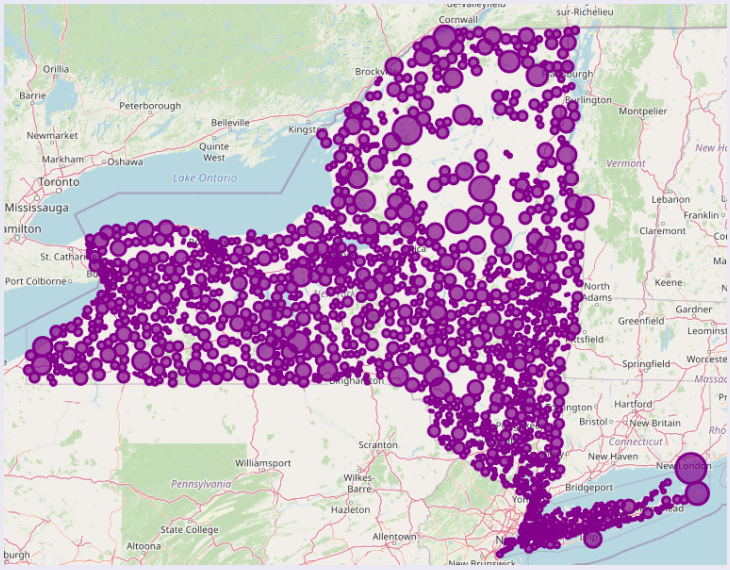
	Population	Positives	Tests	Positive/Population	Tests/Population	Positive/Tests
Population	1.000000	0.941282	0.948461	0.654579	0.541093	0.783299
Positives	0.941282	1.000000	0.992658	0.783019	0.666023	0.841190
Tests	0.948461	0.992658	1.000000	0.801171	0.701771	0.850768
Positive/Population	0.654579	0.783019	0.801171	1.000000	0.957257	0.927905
Tests/Population	0.541093	0.666023	0.701771	0.957257	1.000000	0.849318
Positive/Tests	0.783299	0.841190	0.850768	0.927905	0.849318	1.000000

Pearson correlation factor between different epidemiological metrics.

4.2. Geolocation data

A total of 1982 zip codes covering a large area of the state were included in this study (Fig. 2). A radius of 1km, however, results in substantial empty areas in the northern part of the state, while several zip codes overlap in urban areas. After merging zip codes and developing custom radii (see Methods for details), the area of coverage increased and the overlap between zip codes was minimized. After merging, a total of 1684 zip codes were used in the analysis. After querying for venues, a total of 36'151 venues were listed, with 131 duplicate values (0.36%). The parameters used provided a substantial return of venues. Only 3.2% of zip codes returned no venues, and over 72% returned 5 or more venues (Fig. 3).

Figure 2

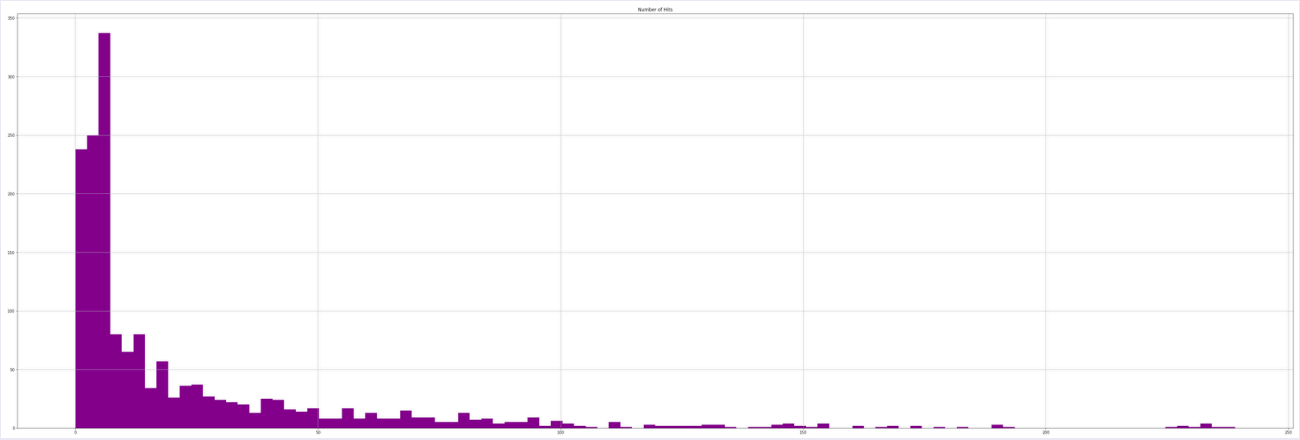


Total area coverage of the zip codes used in this study.

Venues obtained were split in 556 different types of venues, the main metric of analysis in this study. The most frequent venue types for the state of New York were: Pizza Place (898), American Restaurant (794), Convenience store (517), Bar (498), Discount Store (492), Sandwich Place (464), Italian Restaurant (464), Pharmacy (432), Diner (394), and Ice Cream Shop (384). The average number of returns per venue was 39.66, with a standard deviation of 95.42. The correlation between venue frequency and the rate of COVID-19 varied substantially. Ethnic restaurants, bagel shops, convention centers, and banks were the most highly

correlated venue types with COVID-19 spread, while outdoors areas, such as campgrounds, rivers and lookouts figured abundantly as negatively-correlated venues (Fig. 8).

Figure 3



Histogram depicting the frequency of hits per zip code for the state of New York, ranging from 0 hits to 250.

4.3. The models

4.3.1. Predicting COVID-19 spread

The first models developed aimed at identifying the ratio of COVID-19 spread for counties. Simple linear regression resulted in a very small mean squared error (MSE 0.006), but a value of R^2 of 0.45. Predicted values were plotted against real values, indicating that, in most cases, the infection rate was overestimated for counties with lower rates, and underestimated for counties with larger infection rates (Fig. 4). To account for collinearity, ridge regression was performed, resulting in a smaller MSE (0.006) and improved R^2 (0.65). This indicates some venues are highly correlated, but ridge regression cannot fully mitigate the effects of that collinearity. To reduce noise caused by the large number of features used in the model, Lasso feature selection was employed. Using 165 features out of the original 556, the linear regression model achieved an MSE of 0.003 and R^2 of 0.67. Plots of predicted versus real values can be found in Fig. 4. The list of selected features is available in Table 5. Using subsets of counties excluding the most populated counties did not improve the score of the regression models.

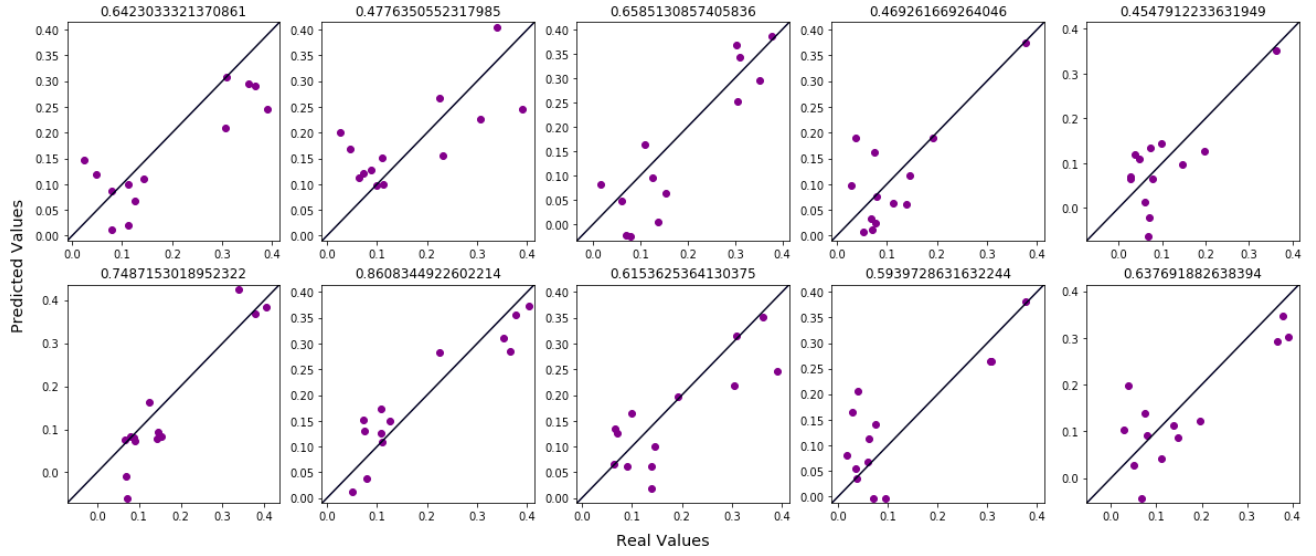
Table 4

Bagel Shop	0.791398	**River**	-0.288216
Bank	0.694208	**Diner**	-0.297713
Sushi Restaurant	0.658005	**Lake**	-0.316713
Spanish Restaurant	0.636337	**Convenience Store**	-0.328975
Latin American Restaurant	0.622984	**Business Service**	-0.338211
Caribbean Restaurant	0.614024	**Campground**	-0.350196
Spa	0.610117	**Brewery**	-0.382060
Mexican Restaurant	0.607224	**Construction & Landscaping**	-0.383008
Bakery	0.594169	**Discount Store**	-0.426543
Peruvian Restaurant	0.587165	**Hotel**	-0.467212

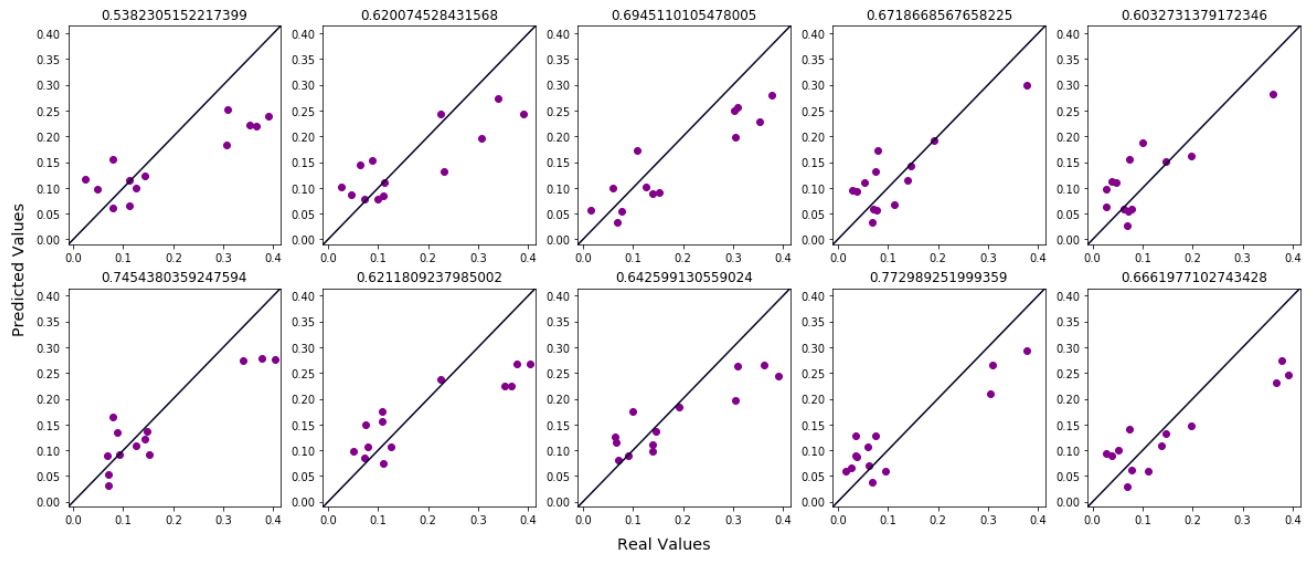
Venues with the highest positive (left) and negative (right) Pearson correlation factor with respect to infection rate.

Figure 4

A



B



Plots of real vs predicted values. The diagonal line represents a perfect match between predicted and real values. Numbers on top of each graph represent the R^2 of each simulation. In A, predicted values using Linear Regression on the whole set of venues. In B, using only features selected using the Lasso feature selection method.

4.3.2. Predicting COVID-19 risk

A more direct application for a COVID-19 prediction model is to classify counties in terms of their risk. To do that, counties must be classified in terms of their risk using binning strategies. Four classification algorithms and 13 binning strategies were employed, and strategies were evaluated based in five parameters: accuracy, precision, f1-score, recall, and ROC-AUC. Among these metrics, recall was used as main metric due to the public health cost of misidentifying a high-risk county. A summary of the results can be found in Table 6. The best results overall were obtained using KNN algorithm with high-risk counties defined as having an infection rate higher than 20%. In these conditions,

Table 5

0	Accessories Store	Airport	Airport Service	American Restaurant	Antique Shop
1	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Dealership	Auto Garage
2	Auto Workshop	BBQ Joint	Bagel Shop	Bakery	Bank
3	Bar	Baseball Field	Basketball Stadium	Bath House	Beach
4	Bed & Breakfast	Beer Store	Boat or Ferry	Bookstore	Border Crossing
5	Boutique	Bowling Alley	Boxing Gym	Breakfast Spot	Brewery
6	Bridge	Burger Joint	Burrito Place	Bus Station	Bus Stop
7	Business Service	Cable Car	Café	Candy Store	Caribbean Restaurant
8	Casino	Cheese Shop	Chinese Restaurant	Clothing Store	Coffee Shop
9	College Bookstore	College Cafeteria	College Gym	Construction & Landscaping	Convenience Store
10	Deli / Bodega	Dessert Shop	Diner	Dive Bar	Donut Shop
11	Dry Cleaner	Event Service	Event Space	Fabric Shop	Farmers Market
12	Fast Food Restaurant	Financial or Legal Service	Fireworks Store	Fish & Chips Shop	Fishing Store
13	Food	French Restaurant	Fried Chicken Joint	Frozen Yogurt Shop	Furniture / Home Store
14	Garden	Garden Center	Gas Station	Gastropub	Gay Bar
15	General Entertainment	German Restaurant	Gift Shop	Go Kart Track	Golf Course
16	Gourmet Shop	Greek Restaurant	Gun Range	Gun Shop	Gym / Fitness Center
17	Harbor / Marina	Hardware Store	Health & Beauty Service	Historic Site	History Museum
18	Hockey Arena	Home Service	Hot Dog Joint	Hotel	Hotel Bar
19	Ice Cream Shop	Indian Restaurant	Insurance Office	Intersection	Irish Pub
20	Jewelry Store	Juice Bar	Korean Restaurant	Kosher Restaurant	Lake
21	Latin American Restaurant	Light Rail Station	Lingerie Store	Liquor Store	Lounge
22	Massage Studio	Mattress Store	Mediterranean Restaurant	Men's Store	Mexican Restaurant
23	Middle Eastern Restaurant	Miscellaneous Shop	Motel	Motorcycle Shop	Mountain
24	Movie Theater	Museum	Music Venue	Nature Preserve	New American Restaurant
25	Nightclub	Office	Opera House	Outdoors & Recreation	Outlet Store
26	Paper / Office Supplies Store	Performing Arts Venue	Pet Store	Playground	Plaza
27	Pub	Rafting	Recreation Center	Resort	Rest Area
28	River	Salon / Barbershop	Scenic Lookout	Seafood Restaurant	Shipping Store
29	Shoe Store	Shopping Mall	Skating Rink	Ski Area	Smoke Shop
30	Snack Place	South American Restaurant	Southern / Soul Food Restaurant	Speakeasy	Sporting Goods Shop
31	Sports Bar	Sports Club	Steakhouse	Summer Camp	Supermarket
32	Supplement Shop	Surf Spot	Sushi Restaurant	Taco Place	Tea Room
33	Tex-Mex Restaurant	Thai Restaurant	Theme Park Ride / Attraction	Thrift / Vintage Store	Tourist Information Center
34	Toy / Game Store	Trail	Train Station	Video Game Store	Video Store
35	Vietnamese Restaurant	Vineyard	Waterfall	Wine Bar	Wings Joint
36	Yoga Studio	---	---	---	---

List of venues selected using the Lasso feature selection method. It is important to note that these venues are not necessarily positively correlated with infection rate. This table is rather a combination of positively and negatively correlated venues that, together, comprise the best set for Linear Regression.

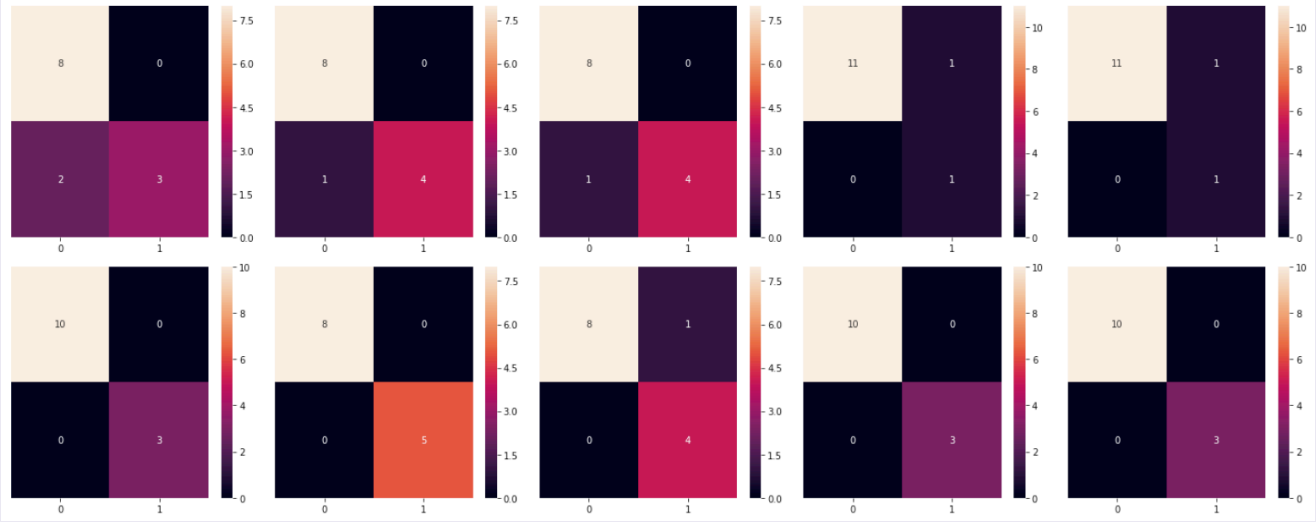
KNN had a recall of 0.91, an F1 score of 0.96, ROC-AUC of 0.94, accuracy of 0.96, and precision of 0.85. The decision tree algorithm underperformed, not exceeding a recall of 0.75. Support vector machine algorithms returned high recalls but unsustainably low F1 scores, and logistic regression failed to produce any effective models. A series of confusion matrices for the best performing model can be found in Fig. 5.

Table 6

Binning	KNN Acc	KNN Pre	KNN F1	KNN Recall	KNN ROC-AUC	DT Acc	DT Pre	DT F1	DT Recall	DT ROC-AUC
0.14	0.854839	0.632933	0.853580	0.722222	0.815657	0.725806	0.398327	0.717863	0.444444	0.642677
0.15	0.919355	0.734101	0.918389	0.800000	0.878723	0.758065	0.345161	0.748365	0.400000	0.636170
0.16	0.935484	0.766952	0.935484	0.857143	0.907738	0.838710	0.477727	0.834152	0.571429	0.744048
0.17	0.935484	0.766952	0.935484	0.857143	0.907738	0.838710	0.477727	0.834152	0.571429	0.744048
0.18	0.935484	0.766952	0.935484	0.857143	0.907738	0.790323	0.420981	0.797086	0.642857	0.738095
0.19	0.935484	0.766952	0.935484	0.857143	0.907738	0.758065	0.288274	0.737756	0.285714	0.590774
0.20	0.967742	0.856407	0.967742	0.916667	0.948333	0.903226	0.610887	0.903226	0.750000	0.845000
0.21	0.967742	0.856407	0.967742	0.916667	0.948333	0.790323	0.352867	0.798729	0.583333	0.711667
0.22	0.967742	0.856407	0.967742	0.916667	0.948333	0.838710	0.373320	0.825682	0.416667	0.678333
0.23	0.951613	0.773705	0.952425	0.909091	0.934938	0.887097	0.509971	0.884957	0.636364	0.788770
0.24	0.935484	0.714286	0.939570	1.000000	0.961538	0.854839	0.391789	0.857590	0.600000	0.751923
Binning	**SVM Acc**	**SVM Pre**	**SVM F1**	**SVM Recall**	**SVM ROC-AUC**	**LR Acc**	**LR Pre**	**LR F1**	**LR Recall**	**LR ROC-AUC**
0.14	0.290323	0.290323	0.130645	1.0	0.500000	0.419355	0.316703	0.390079	0.888889	0.558081
0.15	0.241935	0.241935	0.094261	1.0	0.500000	0.403226	0.257083	0.408040	0.800000	0.538298
0.16	0.225806	0.225806	0.083192	1.0	0.500000	0.774194	0.225806	0.675660	0.000000	0.500000
0.17	0.225806	0.225806	0.083192	1.0	0.500000	0.774194	0.225806	0.675660	0.000000	0.500000
0.18	0.225806	0.225806	0.083192	1.0	0.500000	0.774194	0.225806	0.675660	0.000000	0.500000
0.19	0.225806	0.225806	0.083192	1.0	0.500000	0.774194	0.225806	0.675660	0.000000	0.500000
0.20	0.193548	0.193548	0.062772	1.0	0.500000	0.935484	0.731183	0.930273	0.666667	0.833333
0.21	0.193548	0.193548	0.062772	1.0	0.500000	0.806452	0.193548	0.720046	0.000000	0.500000
0.22	0.193548	0.193548	0.062772	1.0	0.500000	0.806452	0.193548	0.720046	0.000000	0.500000
0.23	0.177419	0.177419	0.053469	1.0	0.500000	0.822581	0.177419	0.742506	0.000000	0.500000
0.24	0.532258	0.256410	0.580241	1.0	0.721154	0.838710	0.161290	0.765139	0.000000	0.500000

Metrics for different classifier algorithms. KNN – K-nearest neighbor; DT – decision tree; SVM – support vector machine; LR – logistic regression. Acc – accuracy; Pre – precision; F1 – f1-score; ROC-AUC – receiver operating characteristic area under the curve.

Figure 5



Confusion matrices for 10 different train-test splits using the KNN method with automatic selection of algorithm, 5 neighbors, a p of 2, and uniform weights.

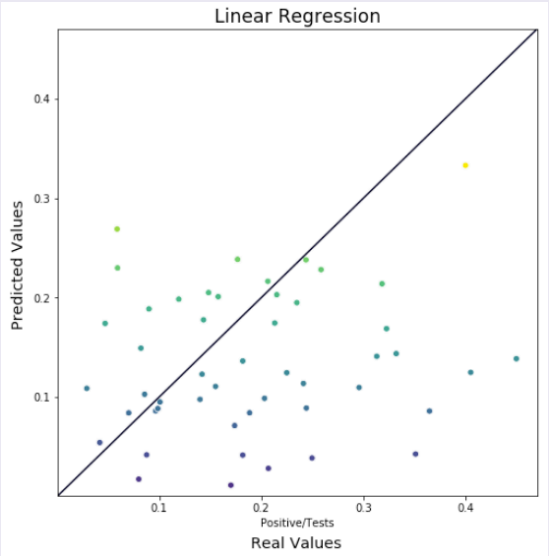
4.3.3. Cluster-based analysis

In addition to classifying risk based on the relative frequency of each venue, the possibility of classifying risk based on the macro composition of neighborhoods (here understood as zip codes) was also explored. Neighborhoods were clustered using the KMeans method in 1 to 20 different clusters, and a dataframe containing the relative distribution of each type of neighborhood per county was assembled. That data frame was used in the four classifier models described above and the results were evaluated following the same metrics. A summary of the results can be found in Table 7. Overall, the performance using clusters was inferior to that using the relative frequency of venues. As was the case with direct analysis of venue frequency, the KNN algorithm displayed the most consistent performance, achieving a maximum recall of 0.83, with an F1 score of 0.95, accuracy of 0.95, ROC-AUC of 0.90, and precision of 0.78.

4.4. Assessing the transposability of the data.

Once the best-performing models were selected, they were used to predict infection rate in the state of Illinois to evaluate how transposable the models are to different states. Unfortunately, the models did not perform well, resulting in negative R2 scores for regression and a recall of 0 using the KNN classifier. A list of correlation between venues and the infection rate in Illinois can be found in Table 8.

Figure 7



Real vs. predicted values for the state of Illinois. The R2 score for the Linear Regression model was negative.

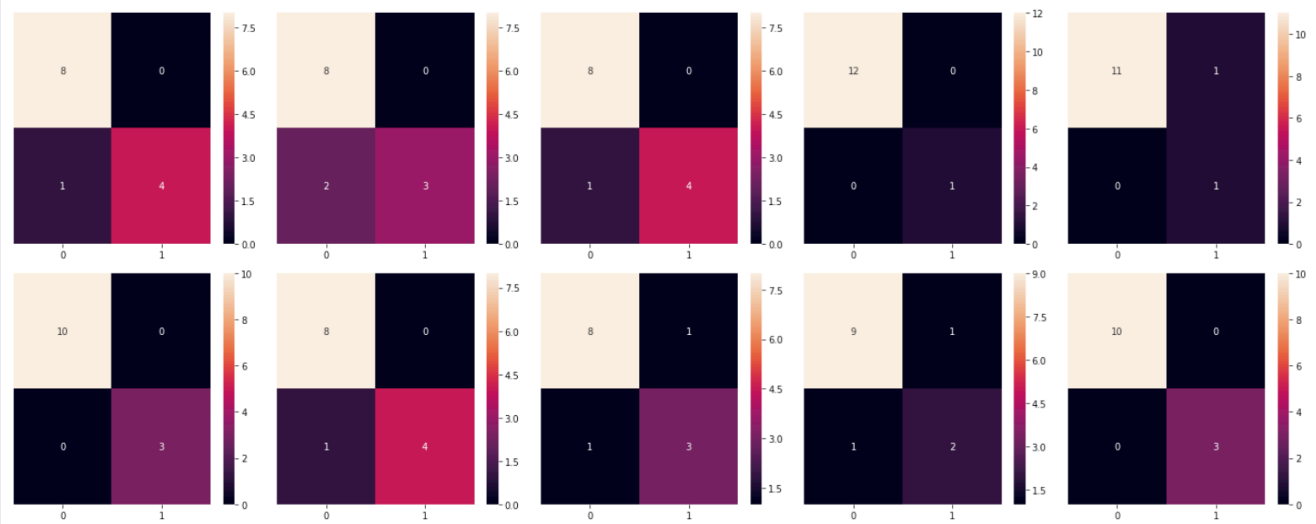
Table 7

K Cluster	KNN Acc	KNN Pre	KNN F1	KNN Recall	KNN ROC-AUC	DT Acc	DT Pre	DT F1	DT Recall	DT ROC-AUC
1	0.806452	0.193548	0.720046	0.000000	0.500000	0.193548	0.193548	0.062772	1.000000	0.500000
2	0.693548	0.217921	0.705835	0.333333	0.556667	0.693548	0.363830	0.725697	0.916667	0.778333
3	0.741935	0.261713	0.749228	0.416667	0.618333	0.758065	0.369816	0.779092	0.750000	0.755000
4	0.903226	0.610887	0.903226	0.750000	0.845000	0.854839	0.474773	0.857009	0.666667	0.783333
5	0.887097	0.567618	0.888785	0.750000	0.835000	0.854839	0.498387	0.860658	0.750000	0.815000
6	0.887097	0.567618	0.888785	0.750000	0.835000	0.838710	0.445469	0.843267	0.666667	0.773333
7	0.903226	0.627496	0.905960	0.833333	0.876667	0.790323	0.403650	0.806105	0.750000	0.775000
8	0.887097	0.567618	0.888785	0.750000	0.835000	0.903226	0.610887	0.903226	0.750000	0.845000
9	0.951613	0.789834	0.950801	0.833333	0.906667	0.887097	0.587814	0.891623	0.833333	0.866667
10	0.951613	0.789834	0.950801	0.833333	0.906667	0.854839	0.474773	0.857009	0.666667	0.783333
11	0.919355	0.673284	0.920560	0.833333	0.886667	0.838710	0.470262	0.846870	0.750000	0.805000
12	0.935484	0.726703	0.935484	0.833333	0.896667	0.903226	0.627496	0.905960	0.833333	0.876667
13	0.919355	0.673284	0.920560	0.833333	0.886667	0.854839	0.498387	0.860658	0.750000	0.815000
14	0.870968	0.530530	0.874614	0.750000	0.825000	0.790323	0.378242	0.802880	0.666667	0.743333
15	0.935484	0.726703	0.935484	0.833333	0.896667	0.919355	0.673284	0.920560	0.833333	0.886667
16	0.951613	0.789834	0.950801	0.833333	0.906667	0.870968	0.508961	0.870968	0.666667	0.793333
17	0.903226	0.610887	0.903226	0.750000	0.845000	0.806452	0.372312	0.811921	0.583333	0.721667
18	0.887097	0.567618	0.888785	0.750000	0.835000	0.854839	0.474773	0.857009	0.666667	0.783333
19	0.935484	0.731183	0.930273	0.666667	0.833333	0.758065	0.345218	0.776275	0.666667	0.723333
20	0.887097	0.567618	0.888785	0.750000	0.835000	0.903226	0.610887	0.903226	0.750000	0.845000

K Cluster	SVM Acc	SVM Pre	SVM F1	SVM Recall	SVM ROC-AUC	LR Acc	LR Pre	LR F1	LR Recall	LR ROC-AUC
1	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
2	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
3	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
4	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
5	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
6	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
7	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
8	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
9	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
10	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
11	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
12	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
13	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
14	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
15	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
16	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
17	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
18	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
19	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
20	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5

Metrics for different classifier algorithms when neighborhoods were clustered based on the relative presence of venues. KNN – K-nearest neighbor; DT – decision tree; SVM – support vector machine; LR – logistic regression. Acc – accuracy; Pre – precision; F1 – f1-score; ROC-AUC – receiver operating characteristic area under the curve.

Figure 6



Confusion matrices for 10 different train-test splits using the KNN method with automatic selection of algorithm, 5 neighbors, a p of 2, and uniform weights. Neighborhoods were clustered in 16 different neighborhood categories before the analysis.

Table 8

Credit Union	0.410401	**Food Court**	-0.231247
Arts & Entertainment	0.378222	**Hot Dog Joint**	-0.235331
Scenic Lookout	0.374564	**Music Venue**	-0.249392
Light Rail Station	0.374442	**Noodle House**	-0.250356
Gas Station	0.366522	**Warehouse Store**	-0.310346
Mattress Store	0.363426	**Department Store**	-0.333022
Cajun / Creole Restaurant	0.363416	**Cosmetics Shop**	-0.347348
Bike Shop	0.362465	**Video Game Store**	-0.373520
Airport Service	0.362465	**Burger Joint**	-0.395082
Platform	0.358304	**Women's Store**	-0.410473

Venues with the highest positive (left) and negative (right) Pearson correlation factor with respect to infection rate.

5. Discussion

In this study, the possibility that location profiles and COVID-19 spread was evaluated. Models with moderate predictive power were built using location data, suggesting there is a certain degree of correlation between that data modality and the disease spread dynamics, even if a causality relationship cannot be established between any one type of venue and disease spread. Regardless, it shows that location data is a valid source of information and can be used as an auxiliary input when studying the epidemiology of infectious diseases.

Readily discernable from the data is the fact that states are fairly heterogeneous concerning virus spread and disease prevalence. There is a 135-fold difference between the counties with the least and the higher number of cases per capita, and a 12-fold difference in the number of tests per capita. That reveals that both diseases spread, and testing efforts vary substantially. Even when both population and testing capability are taken into consideration in the form of a positive/tests ratio, the difference can still be as large as 25 times.

An interesting aspect about the infection spread with respect to population is that it seems to follow two distinct patterns. Highly populated counties displayed higher Pearson correlation factors between infection and population, similar or higher to that obtained for the combined data. On the other hand, the correlation factor for smaller counties remained relatively low, even when data was split to maximize the correlation factor for both groups. Separating large and small counties, however, did not result in better model performance, suggesting the relationship between population and spread may not be as simple, which may be explored in a further work.

Correlation between individual venue categories and infection rate also varied substantially. Only 32 out of the 556 venues had a Pearson correlation factor above 0.5, and only 8 above 0.6. Perhaps surprisingly so, bagel shop is the venue category most strongly correlated with infection spread, with a factor of 0.79. This is somewhat hard to explain, given the overall similarity of bagel shops to cafes and other establishments that are not featured among those venues with high correlation factors. A possible explanation, however, is the intrinsic connection between bagels and New York state and, in particular, New York city[15]. Bagels are considered a staple of the Big Apple, and numerous shops can be found in the state. Although an interesting observation, it's unlikely that other states will have equitable numbers of bagel shops to make that a decisive factor in COVID-19 spread.

A more generalizable observation, however, is the remarkable presence of ethnic restaurants among the most positively correlated venues. Six out of the top 10 venues are ethnic restaurants, and half of all venues with correlation factor above 0.5 are in that category. Other types of venues, such as American food, diners, and fast food joints, are not present in the list, suggesting there are particularities to neighborhoods containing high numbers of ethnic restaurants that make those communities more susceptible to viral spread. It's also conceivable that the restaurants themselves may act as vectors, given that ethnic restaurants tend to be smaller, host larger groups, and for longer periods of time. More granular data will be necessary to dissect the specific role of restaurants in the disease spread. Regardless, restaurants are an important predictor of spread, making them good targets for public health initiatives and social distancing enforcement.

[15] Source: https://www.6sqft.com/the-new-york-bagel-the-hole-story-from-history-and-chemistry-to-where-youll-find-the-good-ones/

In addition to restaurants, venues such as banks, spas, supplement shops, juice bars, and boxing gyms also figured preeminently among the most strongly correlated venues. Notably, not figuring in this list, are offices and office-based business, suggesting workplace transmission may play a smaller role in the disease spread. These kinds of jobs, therefore, could potentially be resumed first to start the economy before restaurants and spas are allowed to reopen. One could raise the possibility that area positively or negatively correlated merely reflect a dichotomy between urban (restaurant- and service-rich) and rural areas. However, when the large counties are removed from the analysis, the broad strokes of correlation remain the same, suggesting the venues indicated here work as predictors for small counties as well independent of NYC.

On the flip side, no venue has a correlation factor larger than 0.5 while being negatively correlated with disease spread. This is good agreement with the nature with the data, as very few venues can potentially actually slow the spread of the disease, except for a moderate effect of some poorly conductive venues taking the place of more conductive ones. Perhaps that's the reason outdoor areas are numerously featured in the list of negatively correlated features, including campgrounds, lakes, rivers, rest areas, golf courses, trails, farms, ski areas, scenic lookouts, state parks, tourist information centers, racetracks, and gun ranges. Reopening those areas to the public, therefore, is a strategy to alleviate cabin fever and provide entertainment while keeping transmission under surveillance. Curiously, hotels, motels and bed and breakfast were also negatively correlated with spread, suggesting these venues, or the neighborhoods that typically host those business, are less conductive to disease transmission than other areas.

Regarding the prediction power of the regression models developed in this work, linear regression achieved a modest 0.45 R^2 score, and techniques to mitigate collinearity were able to improve the models to a moderate 0.65 R^2 score. This indicates a reasonable degree of correlation between venue data and COVID-19 spread. Predicted vs. real value plots reveal that, under these training circumstances, the models tend to slightly overestimate small infection rates, while more strongly underestimating large infection rates. That is likely a result of the larger number of less infected counties in the analysis. Using LASSO to perform feature selection resulted in the selection of anywhere between 160-190 venues (the best R^2 was selected from anywhere between 20 and all the venues). Although using a subset of features resulted in a minimal improvement of R^2 score, its main benefit was better distributing predictions between under and overestimates, with less high-infection counties being subdetected. These results show there is substantial, even if not large, prediction potential in location data towards COVID-19 infection spread.

Another strategy for the use of location data in this study was to classify counties as being low or high risk, instead of defining specific infection numbers. This strategy has the added benefit of being more easily interpreted by policy makers, as they can focus their efforts in the flagged counties. Since infection rate is a continuous variable in nature, it is necessary to employ binning strategies to determine what are high-risk counties. In this study, a backward approach was taken, using model metrics to find a binning strategy that facilitates prediction and makes biological sense. After running four different classifying algorithms, a cut-off around 0.20 was the most effective, and has the added benefit of being a simple number that can be easily interpreted. In this scenario, the KNN method was able to accurately and precisely identify most counties that are considered high-risk. This indicates that risk classification can be an easier and more effective way to approach COVID-19 spread between different counties.

Finally, the models developed in this study were tested against an independent dataset: Illinois. In all cases, the models performed very poorly, performing worse than the average of the data. This was somewhat surprising given the good performance of the models with NY data. One possibility is that the two states do not share enough venues in common, but the data does not support this hypothesis. The two states share 427 venue types, corresponding to 93.4% of all Illinois venues and 76.7% of New York state venues. Another explanation is a different dynamic between location data and COVID-19 spread between the two states. This seems to be supported by the data. No venue type from Illinois had a correlation factor larger than 0.5, and only one above 0.4. Furthermore, the venue profile was substantially different between the two states. This indicates that the relationship between location data and infection spread is not a one-fit-all relationship, but it follows dynamics that are specific for each state. Illinois was chosen due to several characteristics in common with NY state, but perhaps these similarities were too superficial. Although out of the scope of this study, it would be interesting to include more states in this analysis to see if states can be grouped in different categories. Perhaps models are not state-specific, but they are applicable to certain types of state.

Regardless, a relationship between location data and COVID-19 was successfully established in this work. State by state analysis can be performed to identify those venues that are most or least conductive to the virus spread. Hopefully, as the pandemic develops, more granular and specific data will be generated, and that data will help refine these models and build a more powerful picture of how diseases spread. Such study is fundamental to enact data-driven public policy to mitigate the health and economic impact of this, and future, diseases.

6. Conclusion

In this study, we were able to draw several conclusions about the relationship between location data and COVID-19 spread. These include:

a. There is substantial variability between counties within a state with respect to per capita number of cases, per capita number of tests, and infection ratio;
b. There is a strong correlation between county population and infection rate that seems to be more prevalent for the larger counties;
c. Some individual venue types seem to be moderately correlated with infection rate (Pearson score 0.5-0.75), what can inform public policy for targeted actions;
d. Location data can be used both for regression and classification of COVID-19 risk, further reinforcing the idea that there is an underlying relationship between venue composition and viral spread;
e. Models developed for NY state performed poorly in Illinois, suggesting the relationship between venues and viral spread may be state or region dependents.

This information may help policymakers and governmental bodies make better informed decisions to mitigate the negative impact of COVID-19.