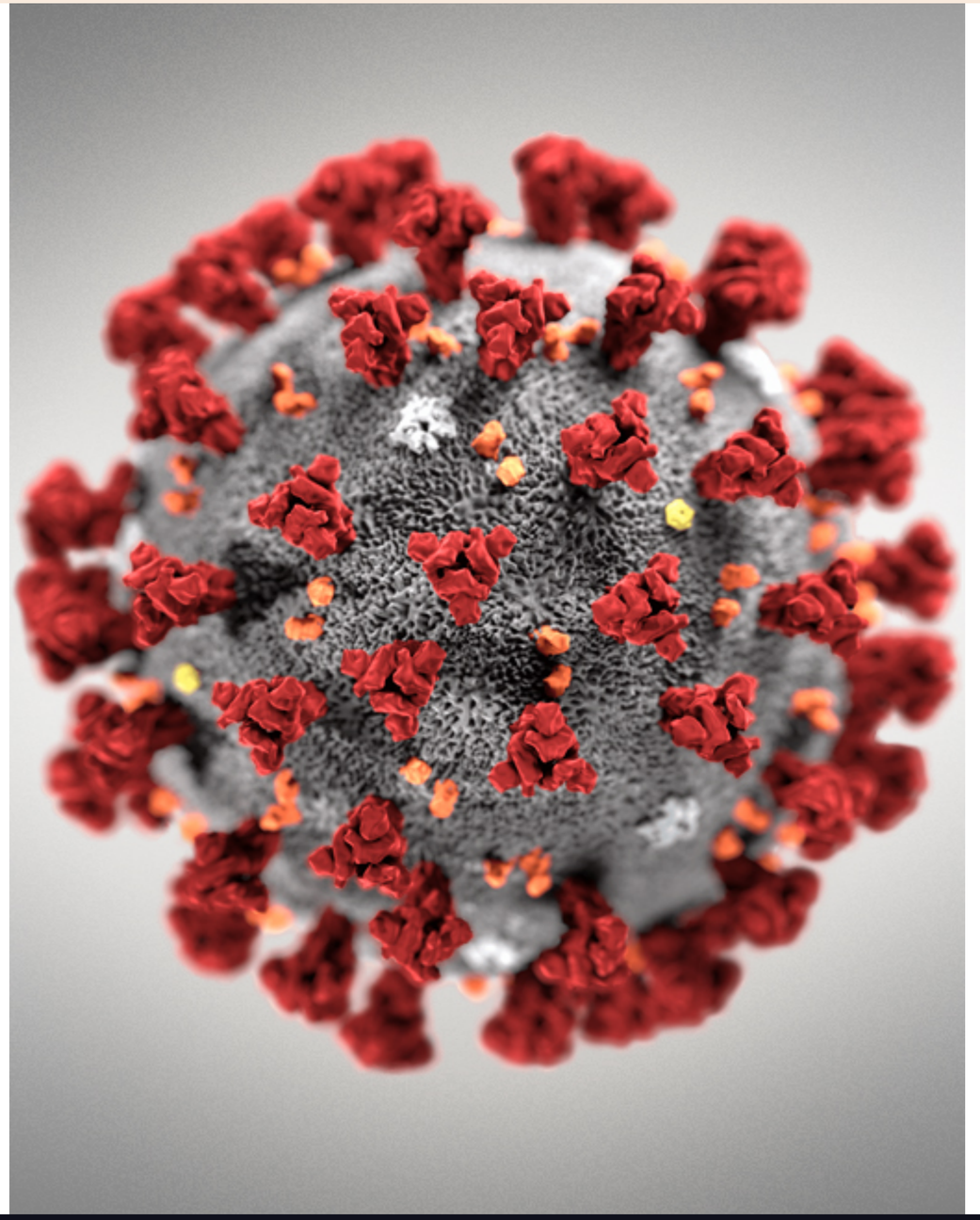
Using location data to understand the dynamics of COVID-19 spread

Giovanne B. Diniz, Ph.D.

The Problem

- The 2020 COVID-19 pandemic has imposed major public health and financial costs
- To contain the spread of the disease, the United States has imposed widespread lockdowns and stay-at-home orders
- With the economy halted, most states now face the challenge of reopening the economy and restarting business
- The question of how to restart the economy is one of the most important decisions states will have to make moving forward
- If data can be leveraged to understand COVID-19 spread, state authorities can make better informed decisions

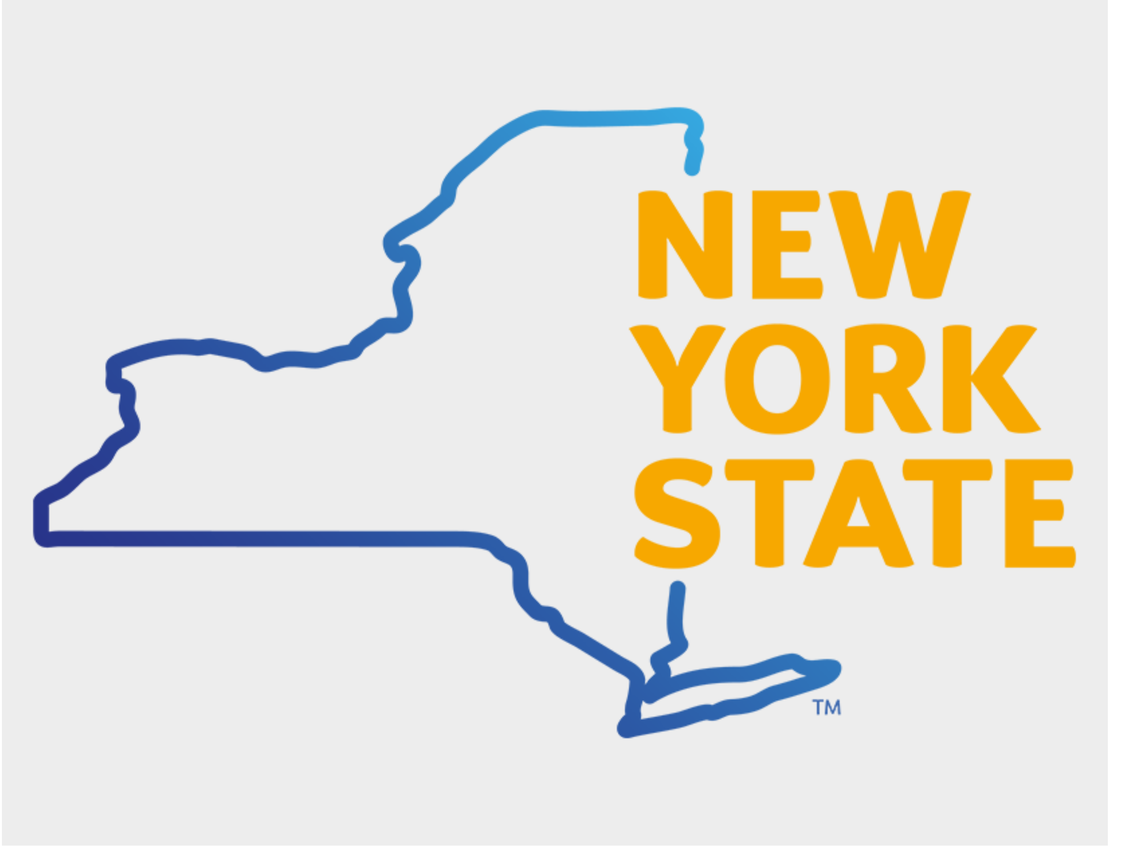


The Data

- In this study, epidemiological and population data were used in addition to the presence of different types of venues in various zip codes distributed throughout the state of New York

- New York State has several characteristics that make it a desirable target of study:
  - NY has seen the most extensive COVID-19 testing effort;
  - NY is an extensive state with numerous and diverse counties
  - NY has a comprehensive and well-polished open data framework, and an easy-to-implement API

- In addition to the state data, the Foursquare API was used to obtain the venues found in different neighborhoods through the state using the explore endpoint



Results - The data

- Datasets xdss-u53e, e9uj-s3sf, and juva-r6g2 were obtained through the Socrata API
  - 62 counties, from Albany to Yates
  - Epidemiological data: total number of positive cases and total number of tests, per county

- Additional metrics: positive cases per capita, tests per capita, and the ratio of positives over tests
  - Ratio of positive cases over test (infection rate) is the most important metric because it accounts both for population and the different rate of tests between counties

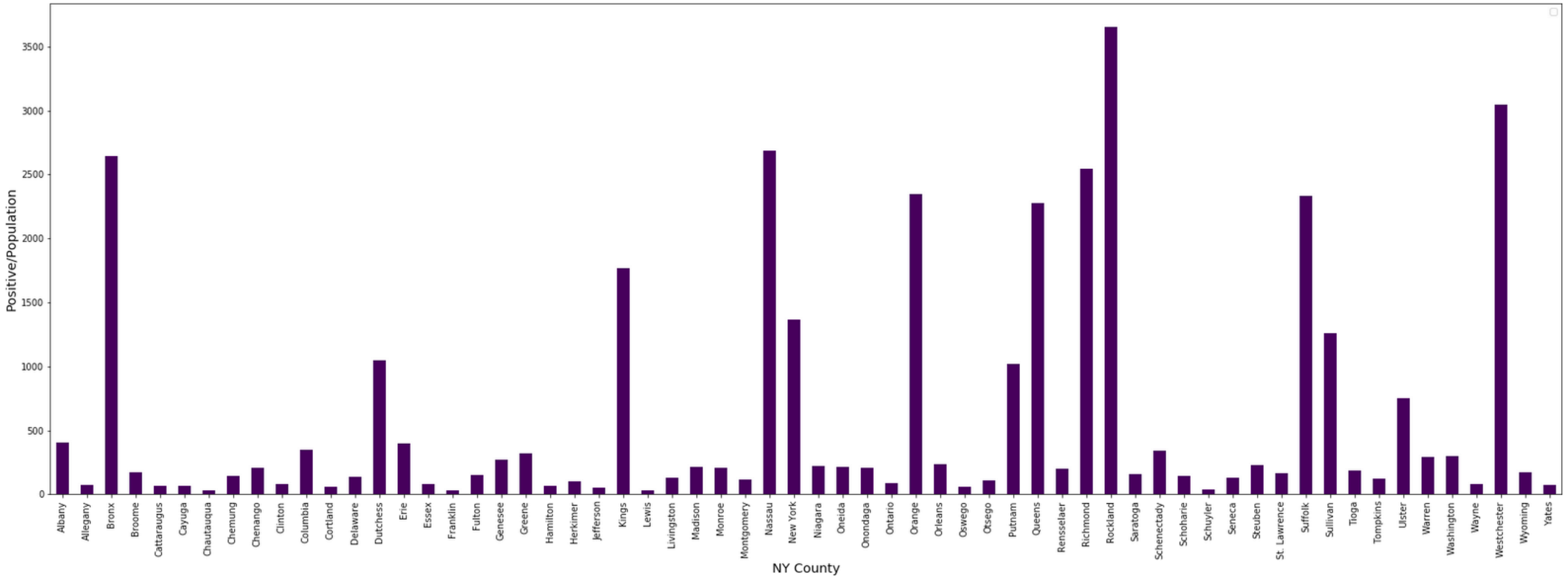
- Data was cleaned up and only the current data was used in the analysis. Duplicated data was removed and county and column names standardized.

	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

	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

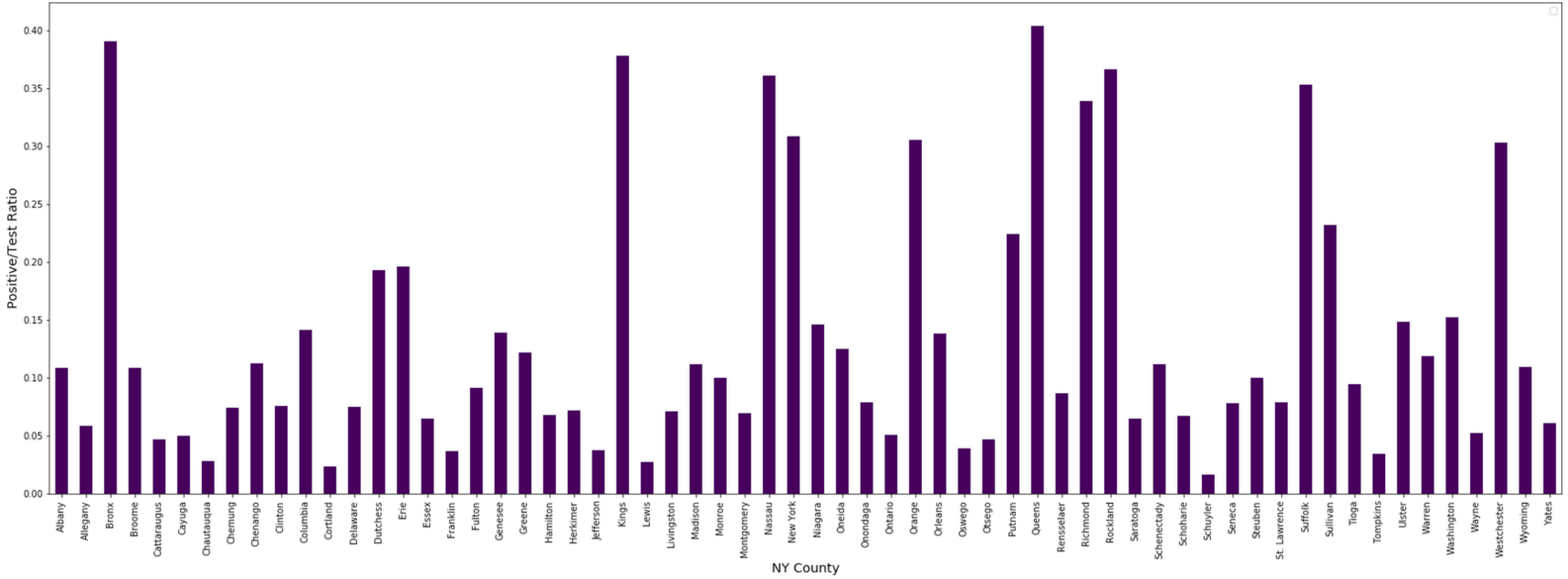
Results - The data

- The number of cases per 100,000 people varied substantially between counties



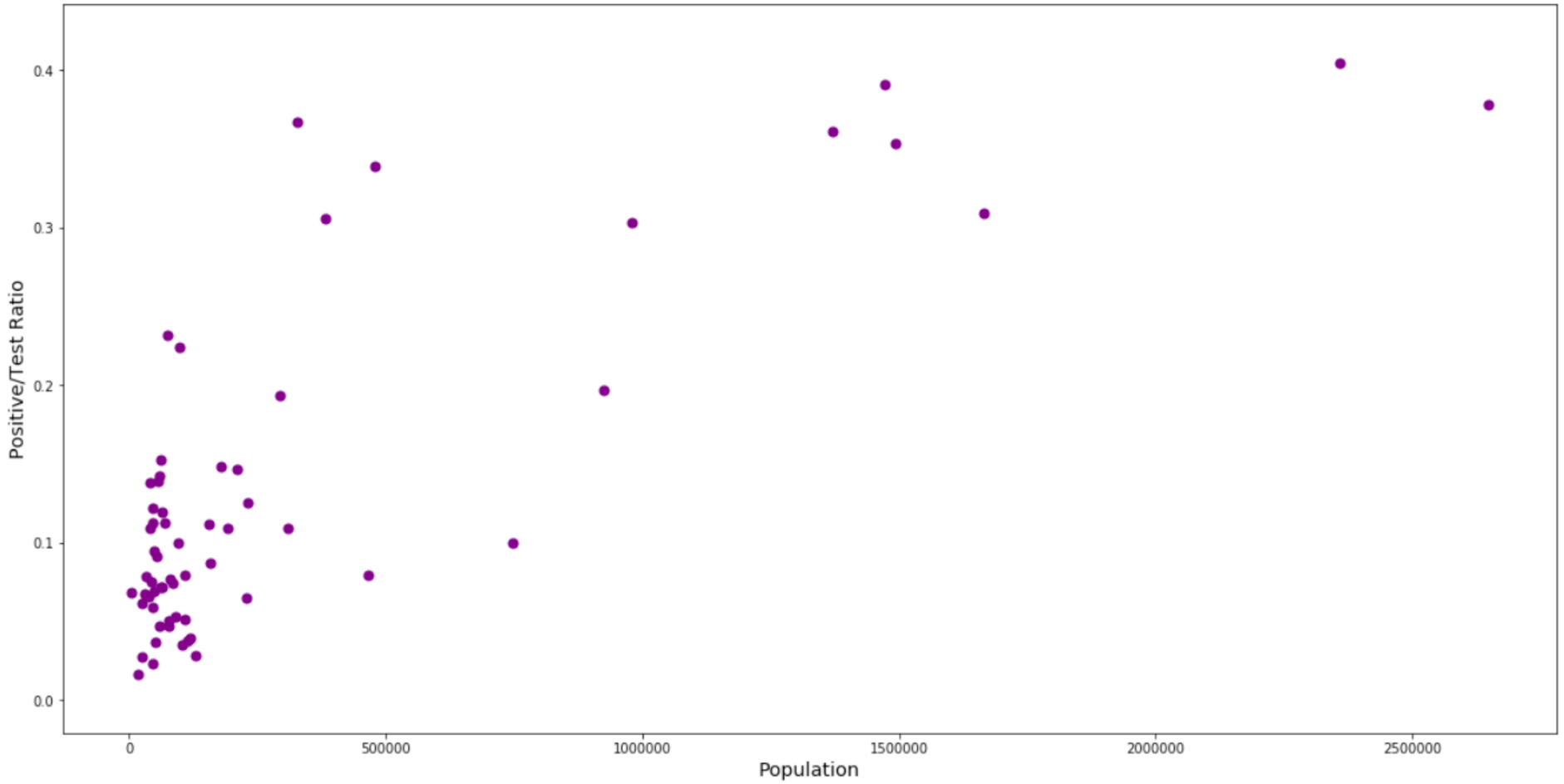
Results - The data

- The number of cases per 100,000 people varied substantially between counties
- The infection rate also varied substantially between counties, with a 24-fold difference between the lowest and the highest rates



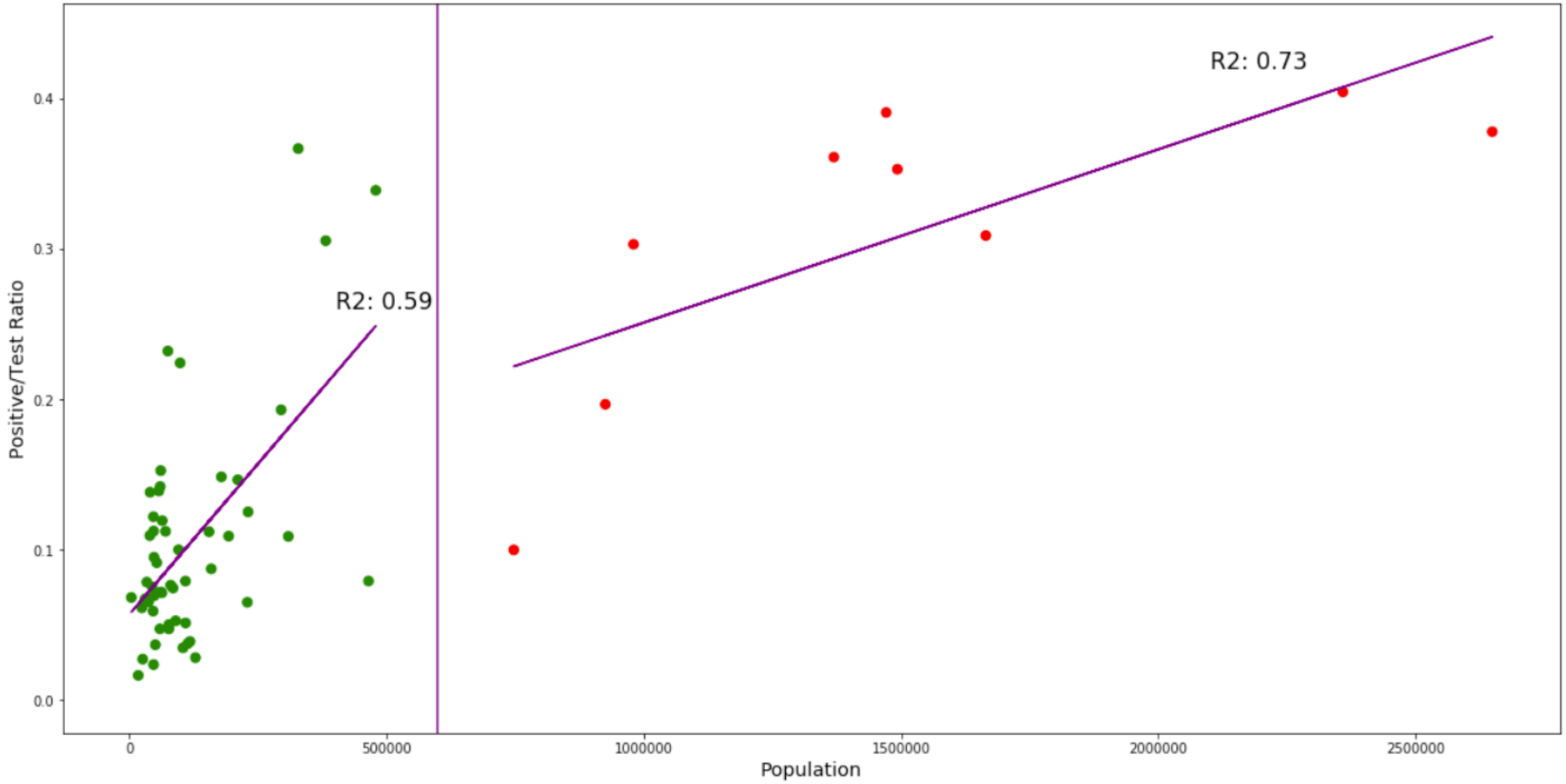
Results - The data

- The rate of infection varies with population, with a Pearson correlation factor of 0.78



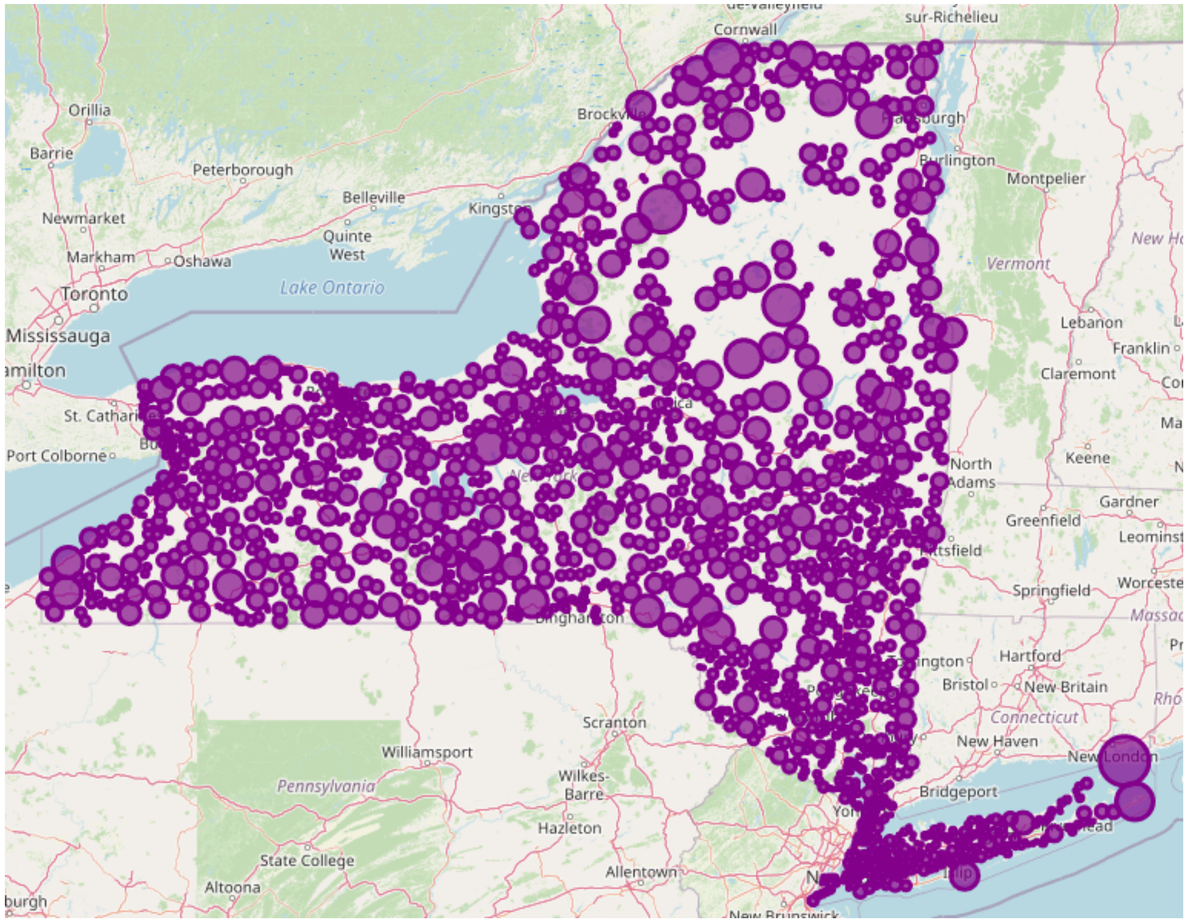
Results - The data

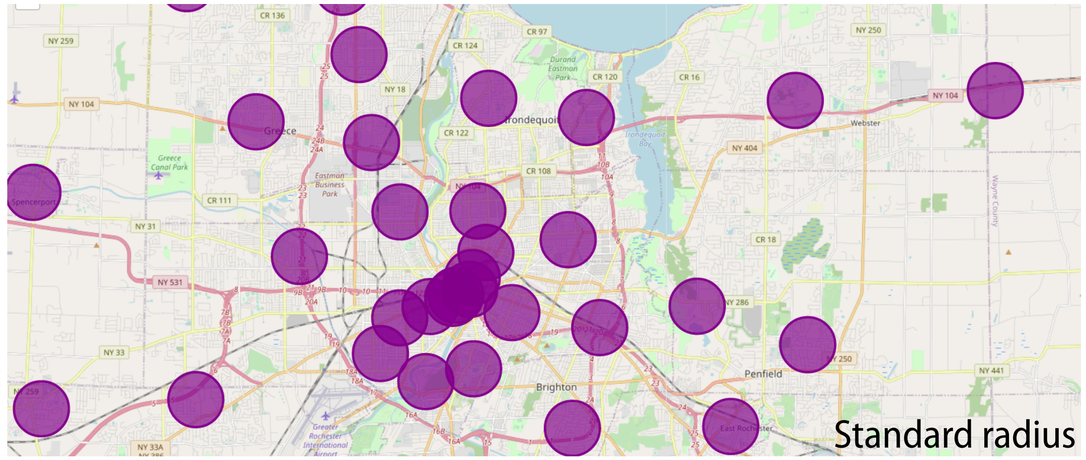
- The rate of infection varies with population, with a Pearson correlation factor of 0.78
- Small counties have a weaker correlation with population when compared to larger counties

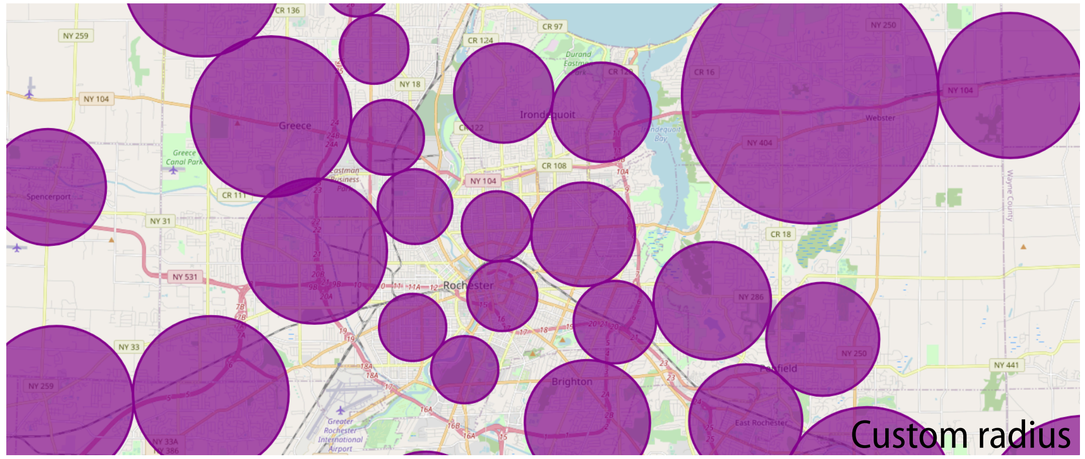


Results - Location data

- A total of 1684 neighborhoods covering a wide area of NY were included in this study
- To avoid duplication of venues, zip codes were merged and custom radius established for each neighborhood

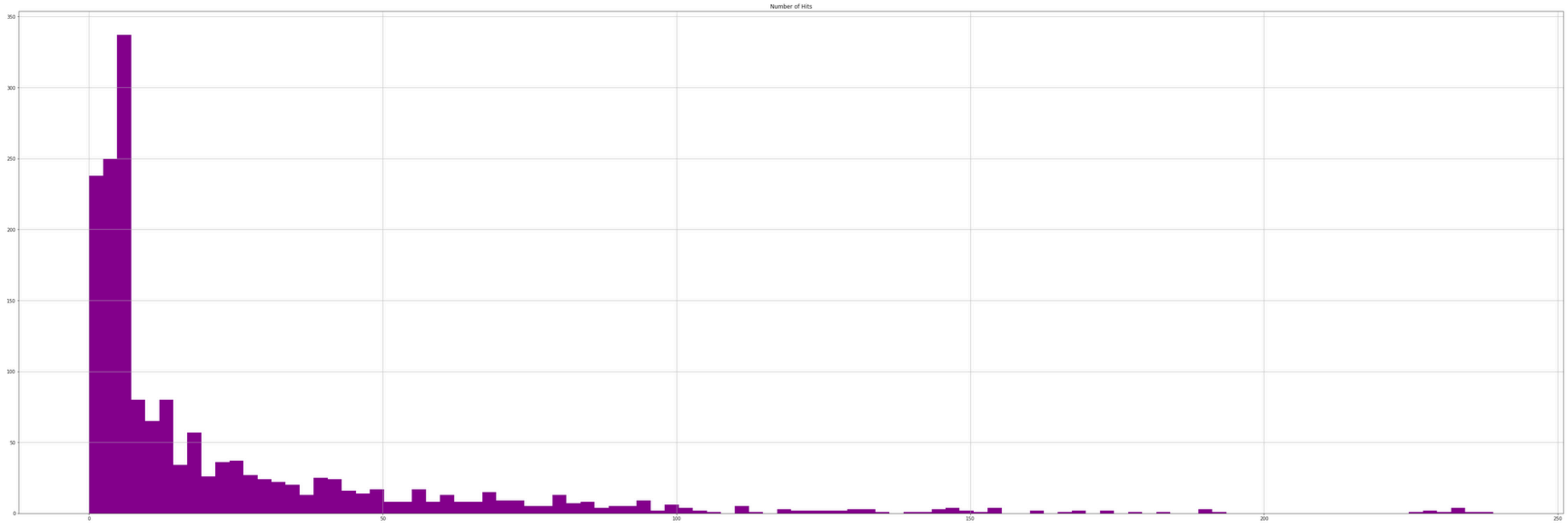






Results - Location data

- A total of 36,151 different venues were obtained through the Foursquare API
- 3.2% of the neighborhoods queried returned no results, 33.7% returned between 1 and 5 venues, and 63.1% returned 6 or more venues



Number of returned venues

Results - Venue types and COVID-19 spread

- A total of 32 venues had positive Pearson correlation factors above 0.5
- Bagel shops, a staple of NY, are the most positively correlated venue at almost 0.8
- Ethnic restaurants comprise half of all the venues with correlation factor above 0.5 (starred venues)
- Granular data will be necessary to understand the relationship between these venues and the higher numbers of COVID-19 spread in their communities

Bagel Shop		0.791398	Deli / Bodega		0.573419	Dumpling Restaurant	★	0.523627
Bank		0.694208	Cosmetics Shop		0.566873	Chinese Restaurant	★	0.522814
Sushi Restaurant	★	0.658005	Mattress Store		0.564517	Health & Beauty Service		0.518143
Spanish Restaurant	★	0.636337	Juice Bar		0.563616	Southern / Soul Food Restaurant	★	0.516078
Latin American Restaurant	★	0.622984	Thai Restaurant	★	0.560069	Boxing Gym		0.514954
Caribbean Restaurant	★	0.614024	Comic Shop		0.547129	Vegetarian / Vegan Restaurant		0.512960
Spa		0.610117	Middle Eastern Restaurant	★	0.544661	Kids Store		0.507896
Mexican Restaurant	★	0.607224	Empanada Restaurant	★	0.538664	Mediterranean Restaurant	★	0.505303
Bakery		0.594169	Seafood Restaurant		0.532133	Italian Restaurant	★	0.504900
Peruvian Restaurant	★	0.587165	Filipino Restaurant	★	0.531120	Metro Station		0.504362
Supplement Shop		0.582200	Portuguese Restaurant	★	0.527683			

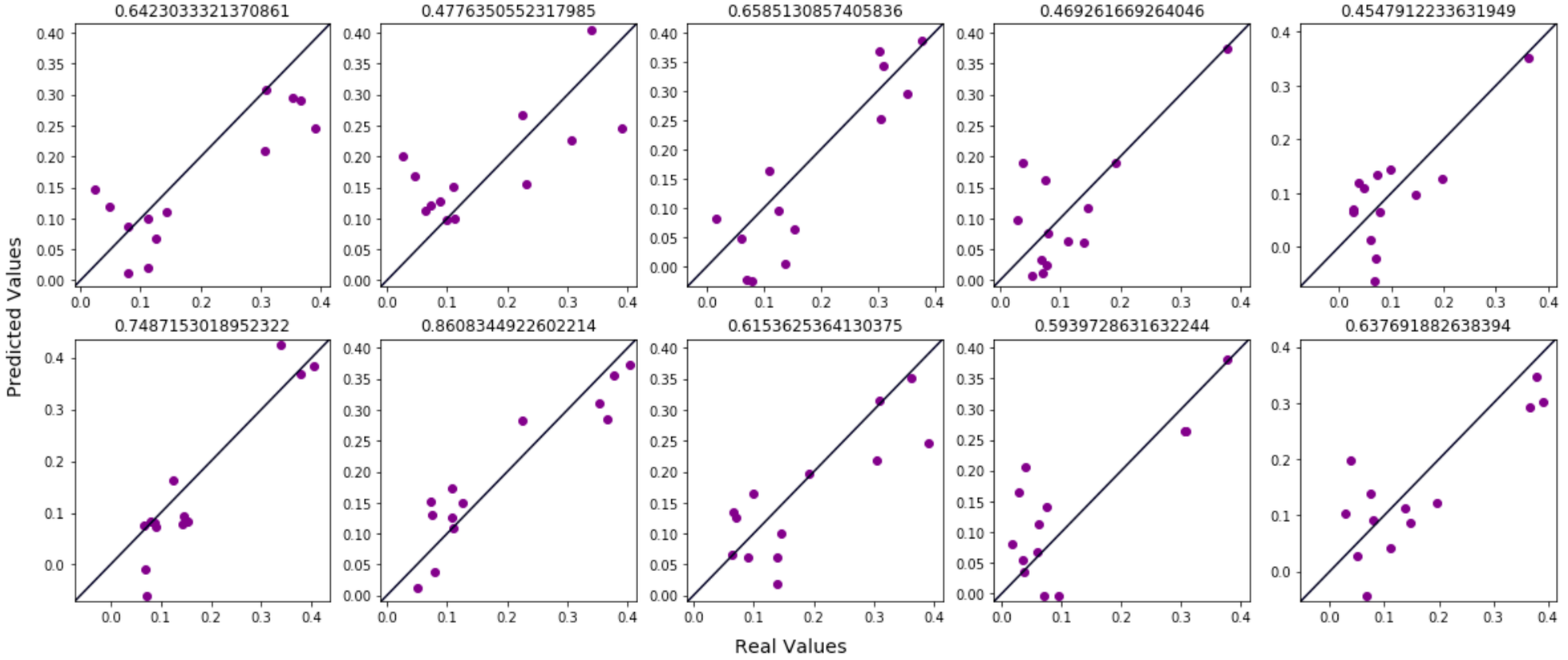
Results - Venue types and COVID-19 spread

- As expected, negative correlation factors were predominantly low
- Outdoor areas were the most predominant type among negatively correlated venues (starred venues)
- Temporary residence venues also displayed negative correlation, such as hotels, motels, and B&Bs
- This data suggests that reopening outdoor areas may be a strategy to alleviate cabin fever and provide entertainment, with low risk of a spike in the number of cases

Airport	-0.187398	Ice Cream Shop	-0.238375	River ★	-0.288216
Motorcycle Shop	-0.192827	Gas Station	-0.242156	Diner	-0.297713
Movie Theater	-0.196132	Gift Shop	-0.244940	Lake ★	-0.316713
Gun Range ★	-0.196889	Bar	-0.248049	Convenience Store	-0.328975
Racetrack ★	-0.202061	Scenic Lookout ★	-0.256180	Business Service	-0.338211
Motel	-0.208667	Ski Area ★	-0.260076	Campground ★	-0.350196
Gun Shop	-0.216796	Farm ★	-0.261986	Brewery	-0.382060
Tourist Information Center ★	-0.216901	Trail ★	-0.264979	Construction & Landscaping	-0.383008
Bed & Breakfast	-0.225144	Home Service	-0.265964	Discount Store	-0.426543
State / Provincial Park ★	-0.237108	Golf Course ★	-0.275048	Hotel	-0.467212

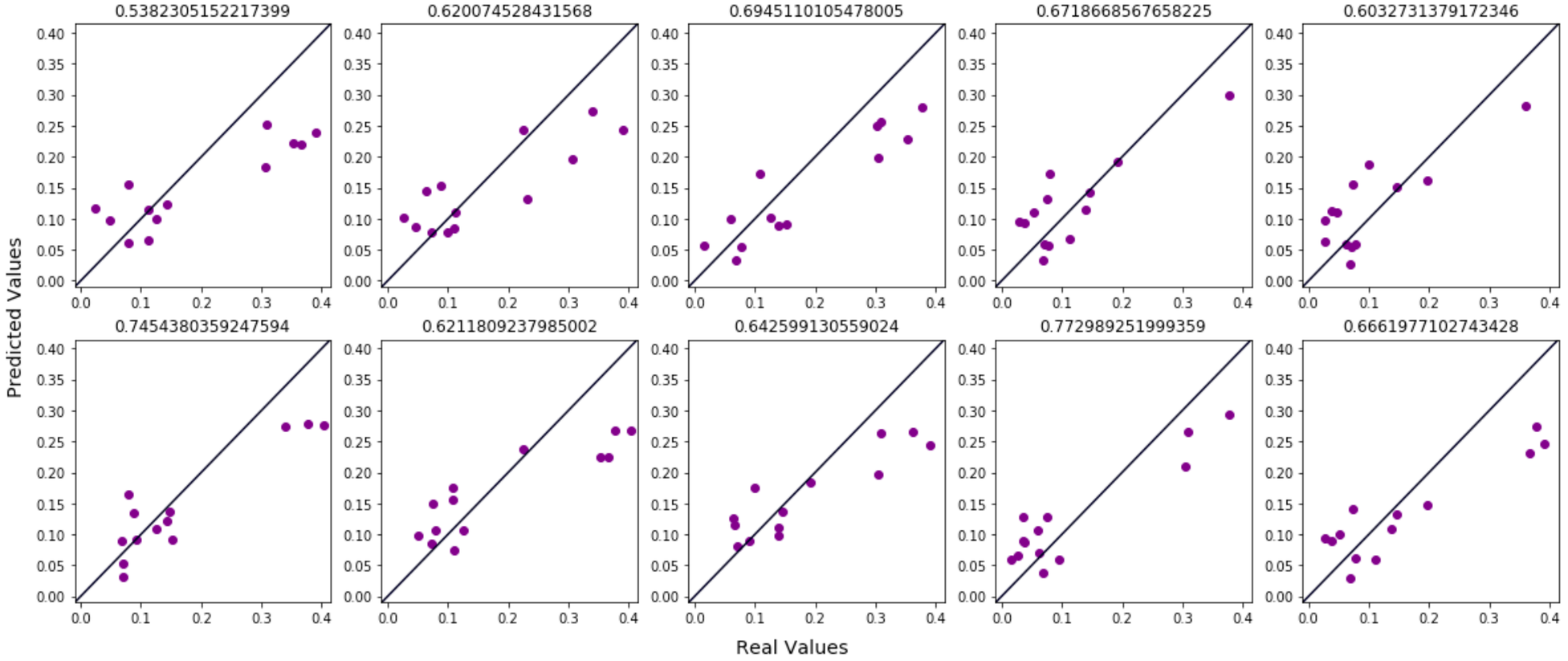
Results - Regression models

- Simple linear regression using the relative frequency of venues for each county resulted in a weak R^2 score of 0.45 and MSE of 0.006 (10 simulations of train-test sets illustrated below)
- When Ridge and Lasso regression were employed to mitigate the effects of collinearity, the R^2 score raised to 0.65, suggesting a moderate effect of venue composition in the rate of COVID infection



Results - Regression models

- Lasso principal component analysis results in the isolation of 160-190 venues. Although the increase in R^2 by using principal component analysis was modest (0.67), the reduced set of features resulted in predictions more balanced around the real values, as opposed to the consistently underestimated values predicted with the full set of features (10 simulations of train-test sets illustrated below)



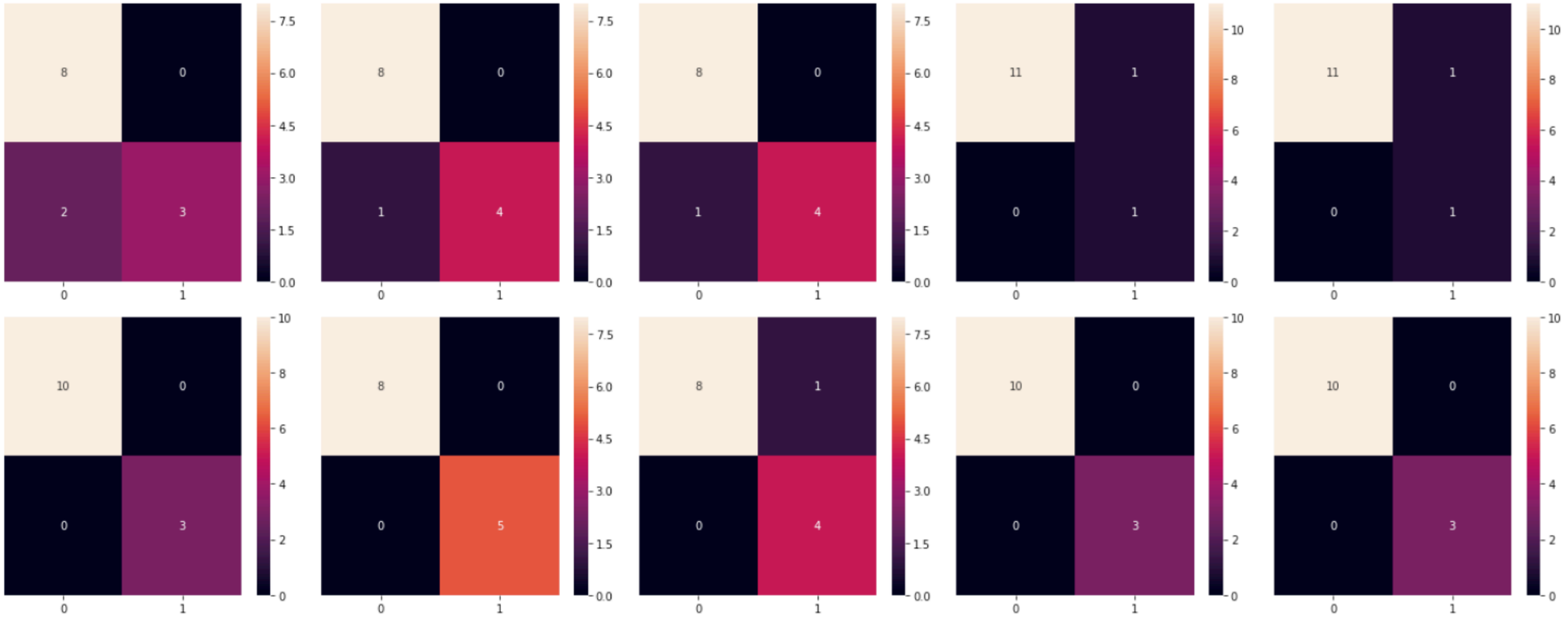
Results - Classifier models

- Instead of predicting the specific infection rate of COVID-19, another strategy is to classify counties as being low or high risk
- In addition to different algorithms, different binning strategies can be employed. Among all attempts, the KNN method for high-risk defined as above 20% infection was the most successful across all metrics

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

Results - Classifier models

- In the best scenario, the KNN method achieved an F1 score of 0.96 and a recall of 0.91, making it very effective in identifying high-risk counties, a trait desirable for a model to be used in public health policies
- Confusion matrices for 10 different train-test splits are presented below, attesting to the high efficacy of the KNN model



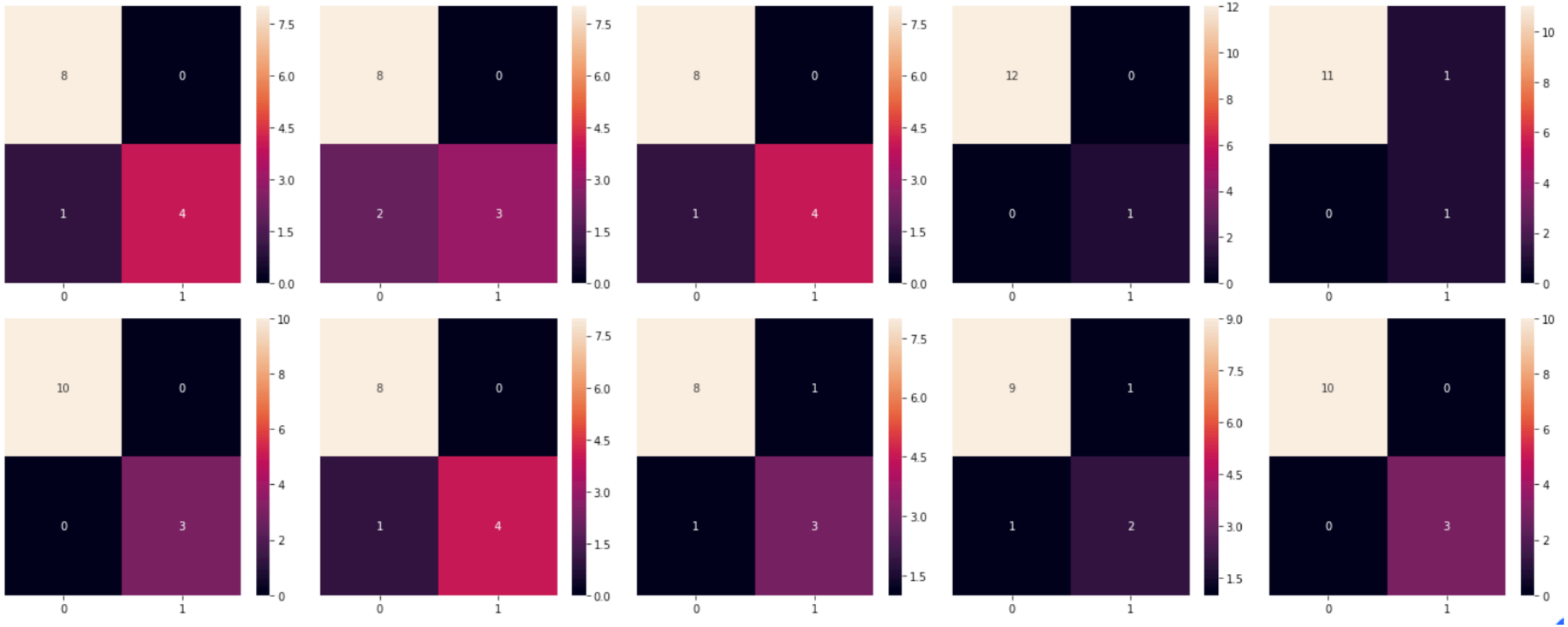
Results - Classifier models

- In addition to classifying the risk of each county by the frequency of venues, neighborhoods can also be clustered in different archetypes using the K-means method and the classification algorithms can then be applied

K Cluster	KNN Acc	KNN Pre	KNN F1	KNN Recall	KNN ROC-AUC	DT Acc	DT Pre	DT F1	DT Recall	DT ROC-AUC	SVM Acc	SVM Pre	SVM F1	SVM Recall	SVM ROC-AUC	LR Acc	LR Pre	LR F1	LR Recall	LR ROC-AUC
1	0.806452	0.193548	0.720046	0.000000	0.500000	0.193548	0.193548	0.062772	1.000000	0.500000	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
2	0.693548	0.217921	0.705835	0.333333	0.556667	0.693548	0.363830	0.725697	0.916667	0.778333	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
3	0.741935	0.261713	0.749228	0.416667	0.618333	0.758065	0.369816	0.779092	0.750000	0.755000	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
4	0.903226	0.610887	0.903226	0.750000	0.845000	0.854839	0.474773	0.857009	0.666667	0.783333	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
5	0.887097	0.567618	0.888785	0.750000	0.835000	0.854839	0.498387	0.860658	0.750000	0.815000	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
6	0.887097	0.567618	0.888785	0.750000	0.835000	0.838710	0.445469	0.843267	0.666667	0.773333	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
7	0.903226	0.627496	0.905960	0.833333	0.876667	0.790323	0.403650	0.806105	0.750000	0.775000	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
8	0.887097	0.567618	0.888785	0.750000	0.835000	0.903226	0.610887	0.903226	0.750000	0.845000	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
9	0.951613	0.789834	0.950801	0.833333	0.906667	0.887097	0.587814	0.891623	0.833333	0.866667	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
10	0.951613	0.789834	0.950801	0.833333	0.906667	0.854839	0.474773	0.857009	0.666667	0.783333	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
11	0.919355	0.673284	0.920560	0.833333	0.886667	0.838710	0.470262	0.846870	0.750000	0.805000	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
12	0.935484	0.726703	0.935484	0.833333	0.896667	0.903226	0.627496	0.905960	0.833333	0.876667	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
13	0.919355	0.673284	0.920560	0.833333	0.886667	0.854839	0.498387	0.860658	0.750000	0.815000	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
14	0.870968	0.530530	0.874614	0.750000	0.825000	0.790323	0.378242	0.802880	0.666667	0.743333	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
15	0.935484	0.726703	0.935484	0.833333	0.896667	0.919355	0.673284	0.920560	0.833333	0.886667	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
16	0.951613	0.789834	0.950801	0.833333	0.906667	0.870968	0.508961	0.870968	0.666667	0.793333	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
17	0.903226	0.610887	0.903226	0.750000	0.845000	0.806452	0.372312	0.811921	0.583333	0.721667	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
18	0.887097	0.567618	0.888785	0.750000	0.835000	0.854839	0.474773	0.857009	0.666667	0.783333	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
19	0.935484	0.731183	0.930273	0.666667	0.833333	0.758065	0.345218	0.776275	0.666667	0.723333	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5
20	0.887097	0.567618	0.888785	0.750000	0.835000	0.903226	0.610887	0.903226	0.750000	0.845000	0.193548	0.193548	0.062772	1.0	0.5	0.193548	0.193548	0.062772	1.0	0.5

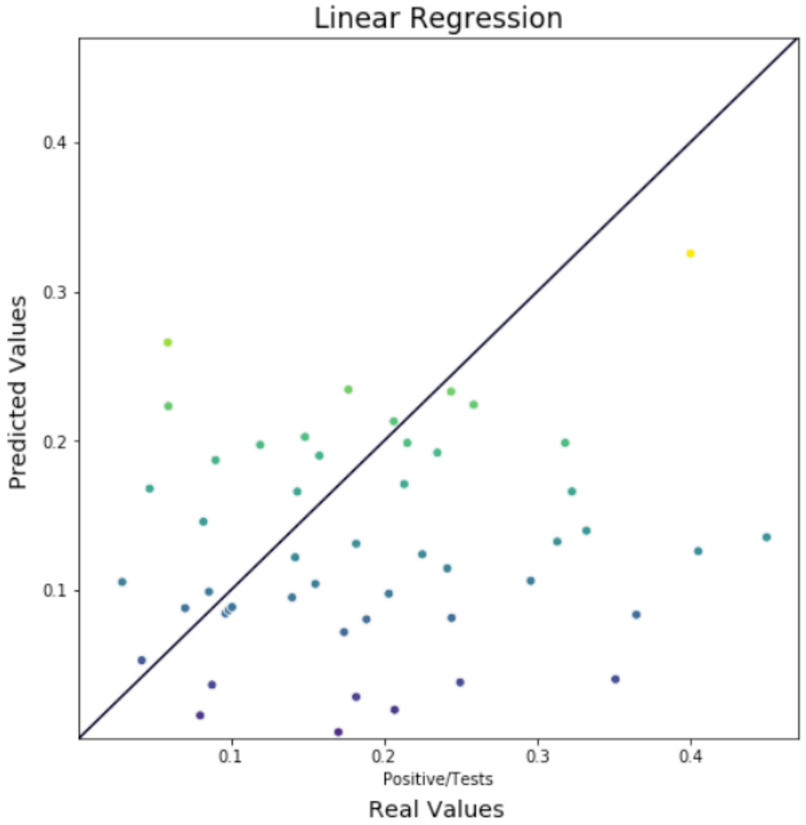
Results - Classifier models

- In the clustering scenario, the best results overall were also obtained with the KNN method, with neighborhoods grouped in 16 different clusters
- This method was slightly less efficient than direct analysis, but it could have applications generalizing the models so they may perform better in other scenarios



Results - Illinois

- The best performing models trained in NY data were tested against an independent dataset: data from Illinois
- Models performed very badly on the independent dataset. Analysis of the data from Illinois suggest the relationship between venues and COVID-19 spread is completely different between the two states, suggesting state- or group-models are necessary



Credit Union	0.410401
Arts & Entertainment	0.378222
Scenic Lookout	0.374564
Light Rail Station	0.374442
Gas Station	0.366522
Mattress Store	0.363426
Cajun / Creole Restaurant	0.363416
Bike Shop	0.362465
Airport Service	0.362465
Platform	0.358304

Food Court	-0.231247
Hot Dog Joint	-0.235331
Music Venue	-0.249392
Noodle House	-0.250356
Warehouse Store	-0.310346
Department Store	-0.333022
Cosmetics Shop	-0.347348
Video Game Store	-0.373520
Burger Joint	-0.395082
Women's Store	-0.410473

Conclusions

- 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;
- 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;
- There is a strong correlation between county population and infection rate that seems to be more prevalent for the larger counties;
- 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;
- Models developed for NY state performed poorly in Illinois, suggesting the relationship between venues and viral spread may be state or region dependents.

Perspective

- Adding more states to this analysis may allow us to identify patterns of similarity between states that could then be used to expand the models beyond single targets
- As the pandemic evolves, more data will become available. In particular, data with high spatial precision can improve the models and help us understand the relationship between particular venue types and the spread of the virus