DS 4A | EMPOWERMENT

# New York State Socioeconomic Factors & COVID-19 Case Rates

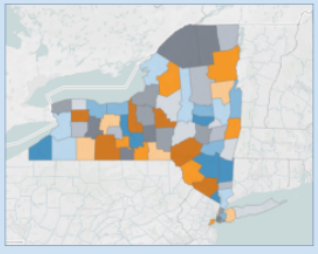
***Correlation One: Data Science for All - Team 88*** *- Idalys Perez, Michael Luevanos, William Johnson, Tolu Olude, Monika Wiggins, Juan Llamas, Sean Kang*

## Highlights

- The linear regression model we used showed how factors like average household size, health uninsured population and Hispanic communities had higher case rates per 100K in NY
- The linear regression results did not depict significant findings for other ethnic communities, median family income, and SNAP recipients
- New York state has 62 counties. Of the top 10 counties with highest COVID-19 case rates, 4 of those counties were located within the New York City area

## Background

The COVID-19 pandemic disproportionately impacted communities of certain racial/ethnic and socioeconomic groups. The following research addresses whether or not specific geographic areas in the state of New York were also disproportionately impacted by the rise in COVID-19 cases.



## Data

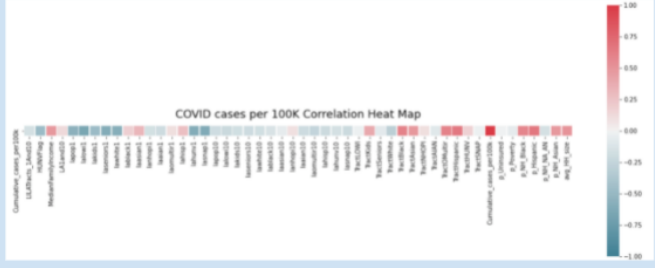
Our research is based on data from the Food Access Research Atlas and the COVID-19 Community Profile Report. After merging and cleaning these data sets, we identified variables on case rates, vehicle access, percent uninsured, average household size, food access, and ethnic population sizes. We then used these variables to explore potential contributing factors that could exacerbate COVID-19 cases in certain geographic areas in New York.

U.S. Department of Agriculture
Food Access Research Atlas

Healthdata.gov
COVID-19 Community Profile Report

## Methods

To evaluate the significance of our variables, we used a correlation heat map to get an overview of our data set.




We then used linear regression to pull out the key variables from our data. Our R-squared value of .622 indicates a fairly strong to strong model. Additionally, the highlighted p-values of .031 (percent uninsured) , .024 (percent hispanic) , and .012 (average household size) show the highlighted variables to be significant.

OLS Regression Results

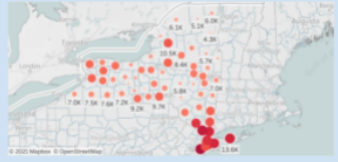
| | | | |
|---|---|---|---|
| Dep. Variable: | Cumulative_cases_per100k | R-squared: | 0.622 |
| Model: | OLS | Adj. R-squared: | 0.556 |
| Method: | Least Squares | F-statistic: | 9.500 |
| Date: | Sat, 07 Aug 2021 | Prob (F-statistic): | 2.21e-08 |
| Time: | 17:12:43 | Log-Likelihood: | -544.87 |
| No. Observations: | 62 | AIC: | 1110. |
| Df Residuals: | 52 | BIC: | 1131. |
| Df Model: | 9 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 72.5575 | 4321.296 | 0.017 | 0.987 | -8598.757 | 8743.872 |
| p_Uninsured | -2.549e+04 | 1.15e+04 | -2.212 | 0.031 | -4.86e+04 | -2367.477 |
| p_Poverty | -9088.5814 | 1.13e+04 | -0.807 | 0.424 | -3.17e+04 | 1.35e+04 |
| p_NH_Black | 7753.0952 | 6479.905 | 1.196 | 0.237 | -5249.786 | 2.08e+04 |
| p_Hispanic | 1.155e+04 | 4960.525 | 2.328 | 0.024 | 1593.670 | 2.15e+04 |
| p_NH_Asian | 6026.5132 | 7906.186 | 0.762 | 0.449 | -9838.410 | 2.19e+04 |
| avg_HH_size | 2949.0872 | 1126.118 | 2.619 | 0.012 | 689.365 | 5208.809 |
| HUNVFlag | 1023.2612 | 1990.069 | 0.514 | 0.609 | -2970.105 | 5016.627 |
| MedianFamilyIncome | 0.0124 | 0.021 | 0.602 | 0.550 | -0.029 | 0.054 |
| TractSNAP | 3.2559 | 5.664 | 0.575 | 0.568 | -8.110 | 14.622 |

## Findings and Results

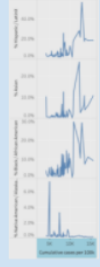
Based on our findings, there is a strong relationship between the average household size, the Hispanic population percentage, and the percentage of uninsured individuals in a given New York county and COVID-19 case rates per 100,000 people in the state of New York.

COVID-19 cases by county



Cumulative COVID-19 Cases and % Uninsured

Cumulative
COVID-19 Cases and
Race/Ethnicity



## Limitations

In gathering and manipulating the data, we found converting census tract to zip code is an industry-wide challenge when working with public census data. To work around this challenge, we relied on county boundaries to join datasets and perform our data analysis. In addition, there may be limitations with how both data sets recorded surveys in racial minority populations as the surveys were voluntary.

## Next Steps

With the Hispanic population percentage showing a positive relationship to COVID-19 case rates, it's imperative to target training, awareness, and vaccination efforts within the Hispanic community. It is also important to collect further data from other minority communities to ensure they are being considered in targeted vaccine efforts.