# Michigan Covid Analysis

Tyler Zender

6/16/2021

## Introduction and Background

Throughout the global COVID-19 pandemic, many eyes had been on the growth trends and occurrences of cases in areas of different backgrounds and demographics. I personally had been interested in the spread and growth of the virus since its initial discovery in early 2020, but became much more invested in tracking its growth when it arrived in the USA as well as my home state of Michigan. As the pandemic progressed, more and more data had been amassed by the state government which detailed the nature of this growth and the details surrounding each case. Particularly notable is the fact that this collection of data had been noted with the county of residence for the infected individual. Because of this, we can effectively perform an analysis of the counties in which cases occur in and draw conclusions about the tendency of the virus to propagate in each population. Inspecting the characteristics of each county and comparing these characteristics to the number of total cases that had occurred in each county can give us insight on where COVID-19 spread most effectively.

## Purposes and Limitations

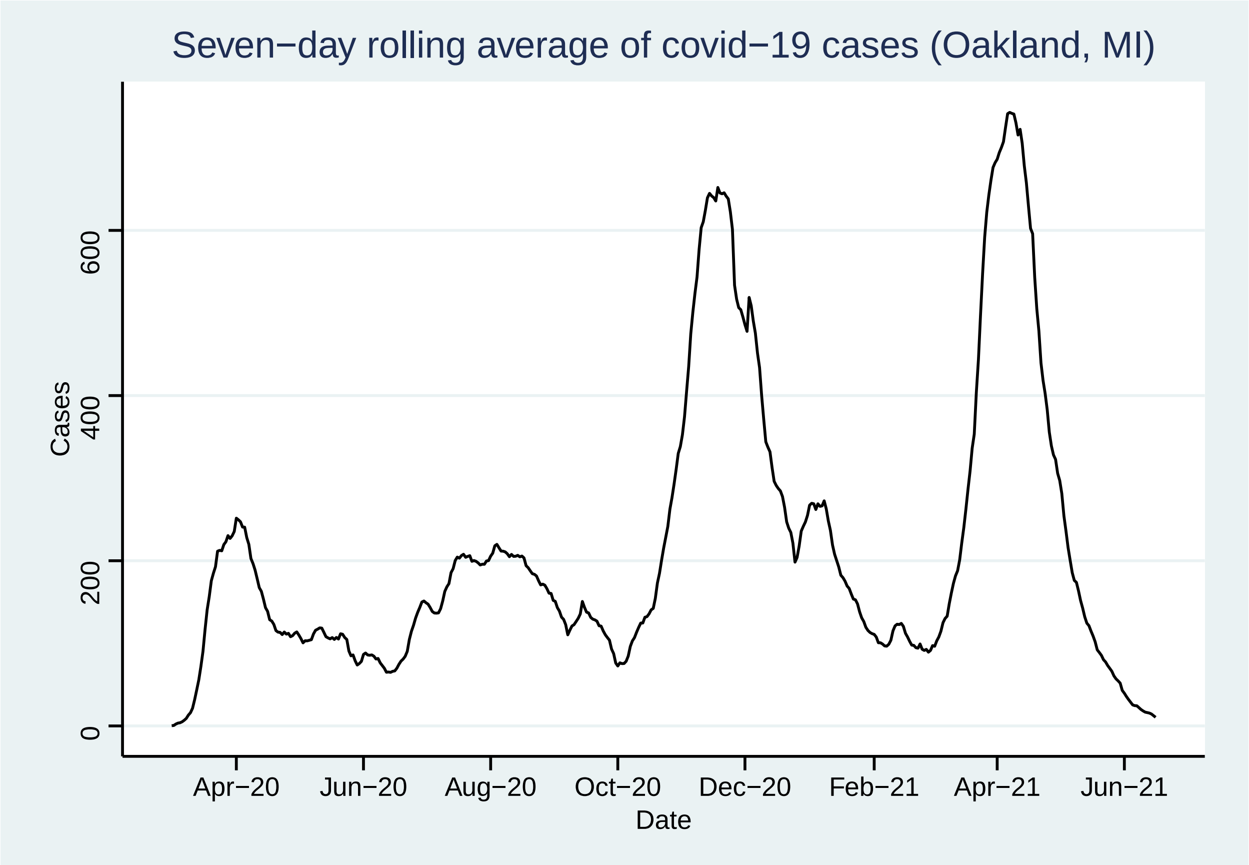
Using data from the "michigan.gov/coronavirus" website, as well as manually-collected information from "census.gov" for a handful of counties, an analysis of county characteristics can help indicate under what conditions the COVID-19 virus reproduces rapidly. Specific characteristics being investigated to derive a conclusion include population, percent of population over 65 years of age, housing units in the county, and per capita income. These characteristics may all be related to the cumulative cases in the county for computing a mathematical relation. It should be noted that the number of data points that represent individual counties are limited (n=19) due to the laborious process of manual collection. Because of this, these results should be taken with caution and would need to be investigated further.

## Methods:

Data is extracted from .csv and .xlsx files for COVID-19 cases and demographics for 19 counties in the state of Michigan. An initial plot is made of the seven-day rolling average of cases in Oakland county to get a background on the general trend of cases. Afterwards, a number of simple linear regression models are used to evaluate which criteria most affect the cases per capita. Plots are generated to visualize data and models. Finally, a multiple linear regression model is generated to improve upon the simple linear regression models.

## Results:

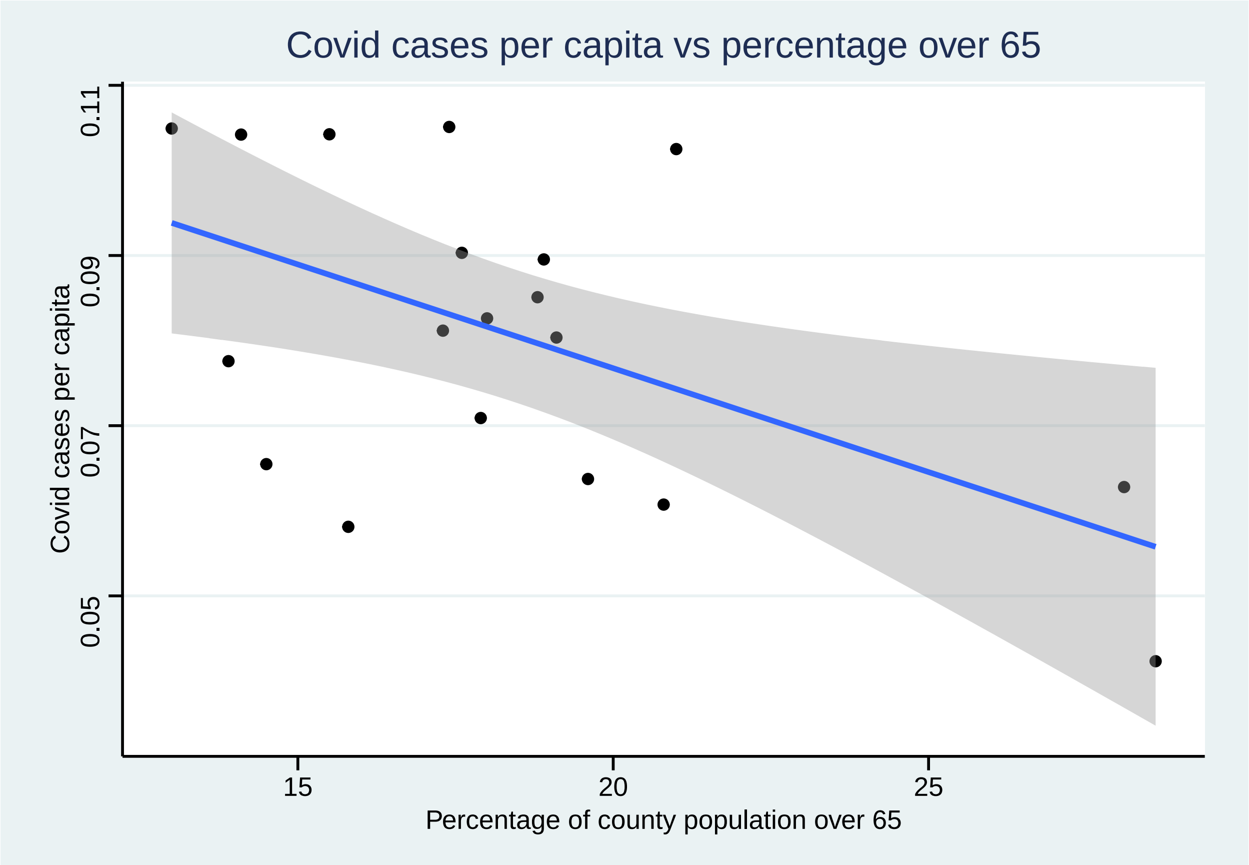
```
## Loading required package: carData
```



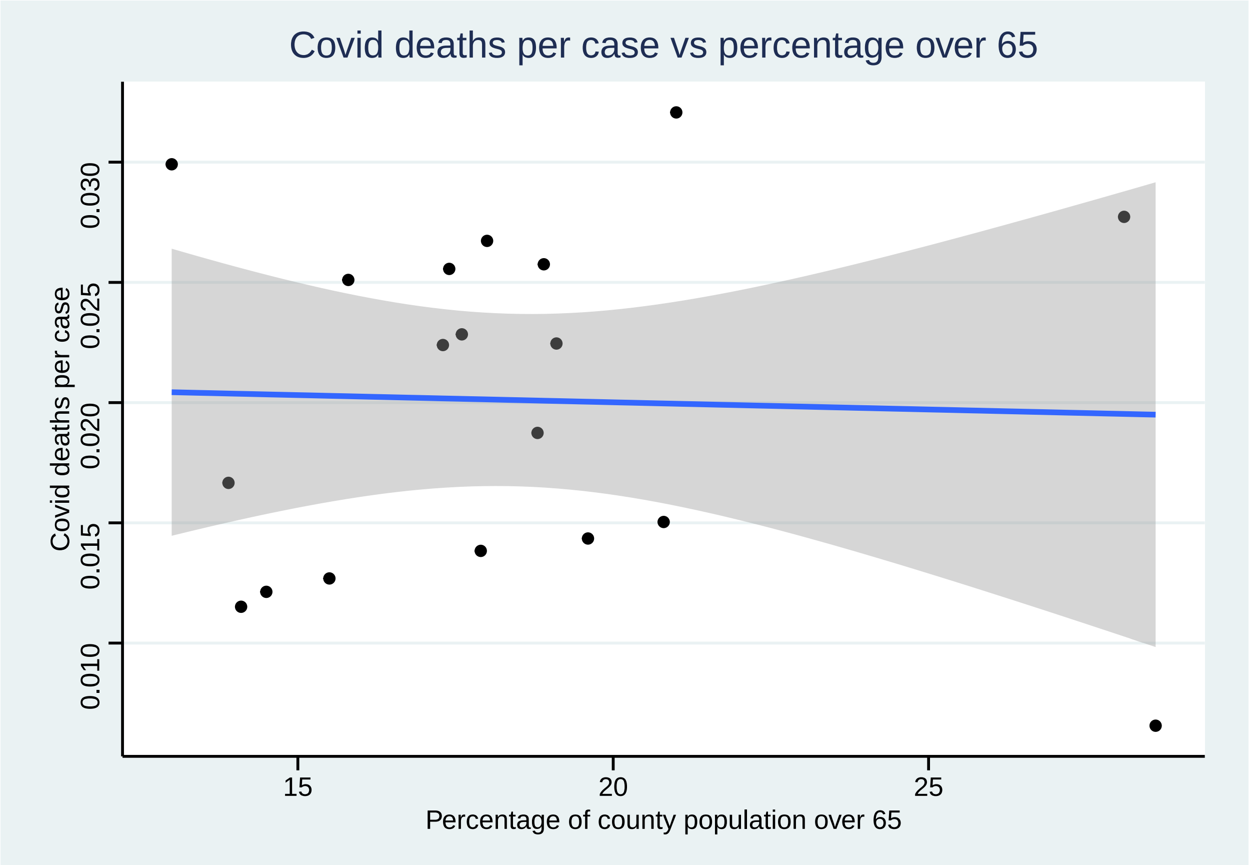
Using our first predictor, the percent of the population over 65, we get our first linear model with a low p-value for the t-test, being .01. At an alpha = .05 level, we would reject the null hypothesis that the coefficient for the relationship between percent over 65 and cases per capita is 0. The same data and model was plotted with a .95 confidence interval.

```
countysOfInterest = c("Oakland", "Muskegon", "Ottawa", "Bay", "Cheboygan", "Macomb", "Wayne", "Washtenaw
```

```
##
## Call:
## lm(formula = countyData$Cases.Per.Capita ~ Over.65, data = countyData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.028884 -0.013721  0.001412  0.010597  0.028195
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1255490  0.0171361   7.327 1.18e-06 ***
## Over.65     -0.0024399  0.0009086  -2.685   0.0156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01611 on 17 degrees of freedom
## Multiple R-squared:  0.2978, Adjusted R-squared:  0.2565
## F-statistic: 7.211 on 1 and 17 DF,  p-value: 0.01565
```



In the graph visualization, we can see the clearly defined trend of cases per capita in relation to percent over 65. Though one may have suspected a strong linear relationship between these two variables, it may not have been clear that the trend would have been negative. With the value of the linear coefficient coming out to -0.0024, we see that as the population becomes dominated by the elderly the cases per capita tend to decrease. One may speculate that this may have resulted from caution taken by these older folks knowing they are at risk. Another cause may have been due to less social exposure due to a lack of employment due to large portions of the population having been retired, thus reducing cases.



To investigate further into the effect of an elderly-dominated population on COVID-19 outcomes, the probability of each case resulting in a death was plotted against the percent of the population over 65. Surprisingly, there does not appear to be a strong positive relationship between these two variables as one might expect.

The next three sets of coefficients look at the relationship between cases per capita vs housing per capita, per capita income, and population, respectively.

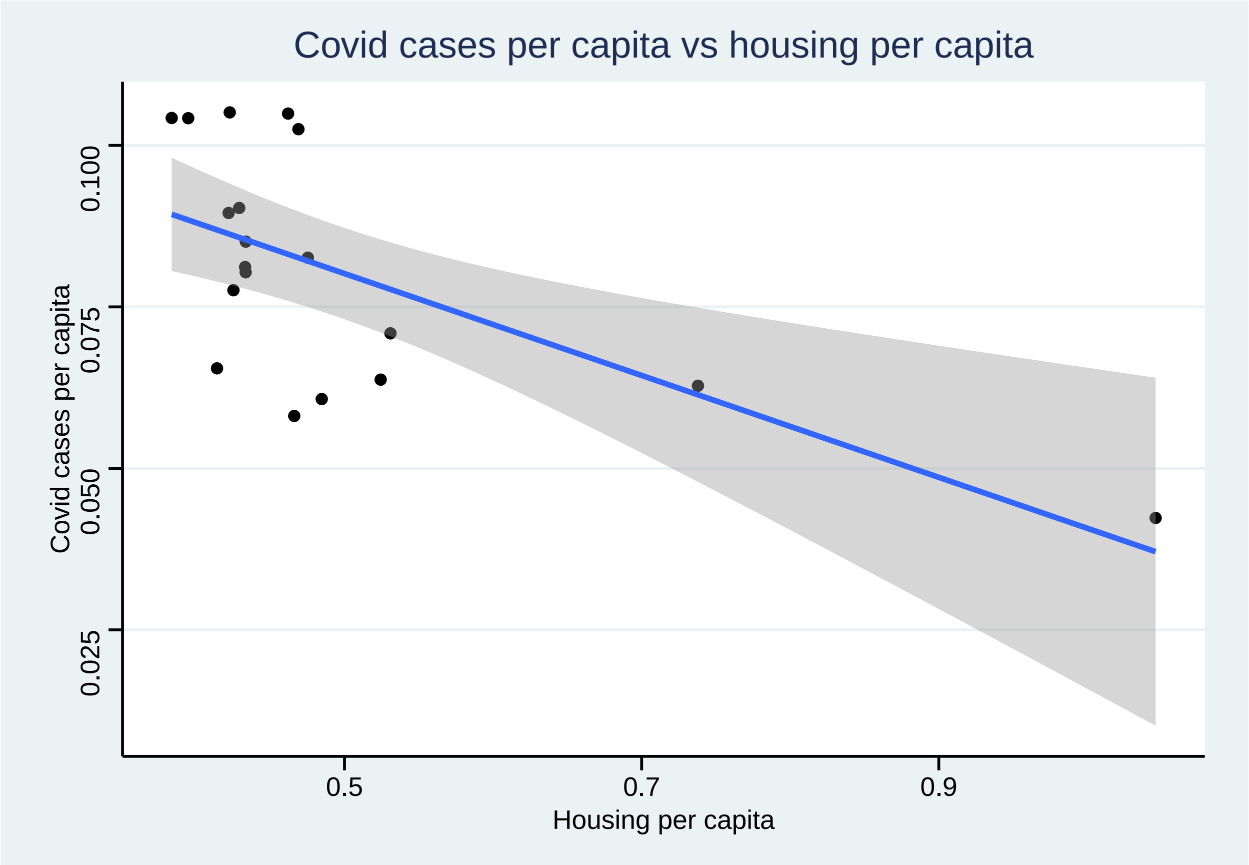
```
##
## Call:
## lm(formula = Cases.Per.Capita ~ Housing.Per.Capita, data = countyData)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0247058 -0.0076405  0.0005141  0.0100690  0.0217758
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.11959    0.01153  10.372 9.04e-09 ***
## Housing.Per.Capita -0.07887    0.02233  -3.532  0.00256 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0146 on 17 degrees of freedom
## Multiple R-squared:  0.4233, Adjusted R-squared:  0.3894
## F-statistic: 12.48 on 1 and 17 DF,  p-value: 0.002558

##
```

```
## Call:
## lm(formula = Cases.Per.Capita ~ Per.Capita.Income, data = countyData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.038270 -0.016067  0.000343  0.015857  0.024467
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.015e-02  2.690e-02   2.980   0.0084 **
## Per.Capita.Income 1.506e-08  8.617e-07   0.017   0.9863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01923 on 17 degrees of freedom
## Multiple R-squared:  1.797e-05,  Adjusted R-squared:  -0.0588
## F-statistic: 0.0003056 on 1 and 17 DF,  p-value: 0.9863
```

```
##
## Call:
## lm(formula = Cases.Per.Capita ~ Population, data = countyData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.037817 -0.015808  0.000088  0.016100  0.024544
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.012e-02  5.663e-03  14.148  7.8e-11 ***
## Population  1.356e-09  9.725e-09   0.139    0.891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01922 on 17 degrees of freedom
## Multiple R-squared:  0.001142,   Adjusted R-squared:  -0.05761
## F-statistic: 0.01944 on 1 and 17 DF,  p-value: 0.8907
```

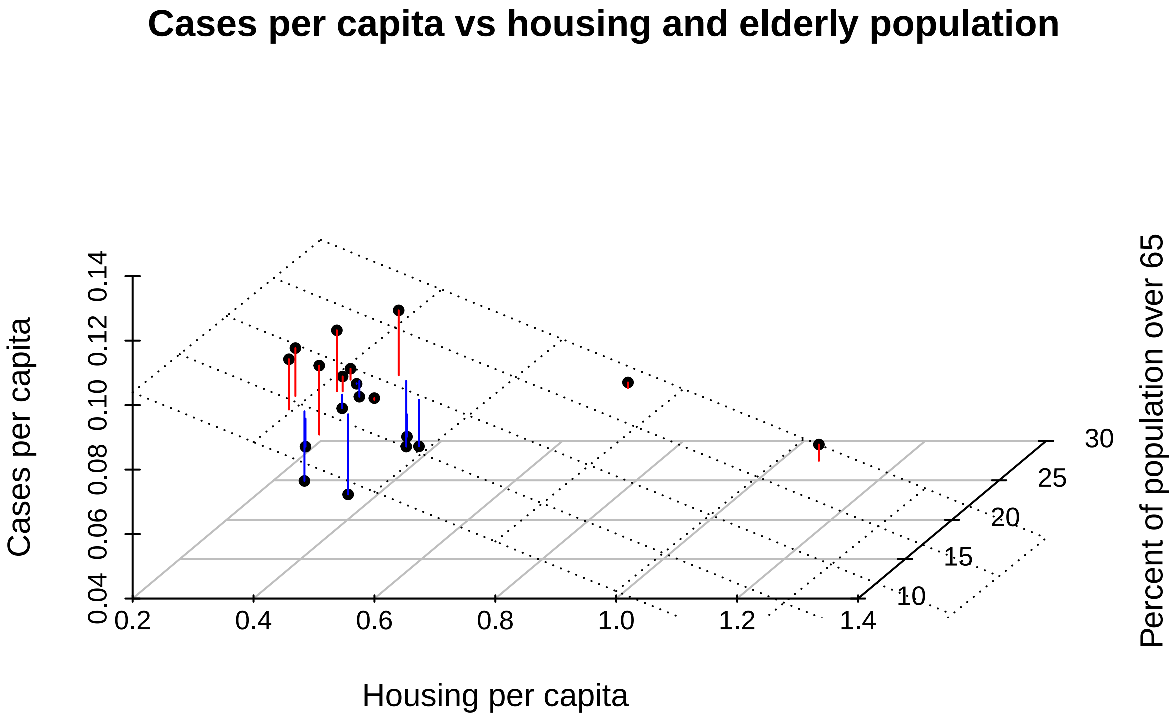
We find that two of these three models provide generally little useful information. Using population of the county and per capita income of the county as predictors gives an a strong hint that neither of these independent variables has an impact on our variable of interest.

However, we do see that the only the model utilizing housing per capita as the predictor displays a low p-value for the t-test. The produced linear coefficient is -0.0788, indicating that as housing per capita increases, cases decrease. This is as one might expect, where it may be expected that as less and less individuals are in a common household, there is are less individuals at risk if one were to contract the disease. At an alpha level of .05, we would reject the null hypothesis that there is no linear relationship between housing per capita and cases per capita. Below is a plot which visualizes these findings.



After noticing how well the model fit the data at the far right end of the plot, I quickly calculated the RMSE for the trained data and ended up with a value of .0138. This is in comparison to the residual standard error displayed in the earlier model summary of 0.0146. Though it is difficult to comprehend the meaningfulness of these values in relative isolation and with little frame of reference, I would believe them to both demonstrate relative efficiency and wellness of fit on the regression model on the data.

Finally, one last multiple regression model was made with the two best predictors found across the previous models - housing per capita, and percentage of the population over 65.



In this increased complexity model, we generate a much lower p-value for the relationship between the predictors and the variable of interest. However, even in this more expanded model, we are able to reject the null hypothesis that there is no linear relationship between housing per capita and cases per capita, but only at an alpha = .1 level.

## Conclusion

Throughout this analysis, an evaluation of COVID-19 cumulative cases per capita for a handful of counties was evaluated in conjunction with demographics of each county. Specific demographic characteristics included information on percent of the population over 65, houses per capita, income per capita, and overall population. In isolated models, it was demonstrated that there is strong evidence to suggest a linear relationship between percent of the population over 65 and cumulative covid cases per capita, as well as housing per capita and cumulative covid cases per capita. In a combined model, more evidence was found to suggest that there is a linear relationship between housing per capita and cumulative covid cases per capita.

## Future Work

Though this analysis provided significant insights on the nature of COVID-19 for counties in the state of Michigan, the extent of the analysis is limited. Future work could included more data points, such as using counties across the USA, which would allow increased examination of COVID cases in response to factors such as state-wide lockdowns or restrictions put in place to limit the spread of the disease. Future analysis might also include the investigation of additional demographic variables available from sites such as census.gov such as sex makeup which would allow us to investigate differences between COVID-19 growth for male vs female populations.