COVID-19 Case and Death Analysis in New York State

Thomas Pink

COVID-19 has struck the United States particularly hard, with the largest concentration of cases in the New York City area. This paper is an analysis of the New York state daily COVID-19 data in conjunction with a variety of socio-economic predictors to uncover correlations between these predictors and infection rates. The data collected was the daily cumulative total of the number of cases and deaths by county (Source 1). Operations were performed to obtain the day to day cases for each county in New York state. The population of each county was binded to the dataset (Source 2), this allowed for the cases and deaths per capita to be calculated. Datasets were found regarding information on the population density (Source 3), poverty rate (Source 4), ethnic demographics (Source 5) and education statistics (Source 6) for each county. The ethic demographics broke the population down into percent white, black, Asian, and Hispanic. The education statistics broke the population down into percent of the population with no high school diploma, only a high school diploma, some college (incomplete bachelor's, or just associates degree), and having a bachelor's degree or higher. All of this information was added to the NYS COVID-19 dataset.

Before doing any analytics, the dataset was sorted highest to lowest based on rate of cases. At a glance, it seems like a high rate of cases correlates strongly with a high population density, high poverty rate, a high proportion of minorities (non-white) in the population, and having a bachelor's degree or higher. Lower case rates seemed to correlate strongly with low population densities, majority white populations, and education at the high school diploma level. A correlation plot was made as shown in Figure 1. This plot highlights correlations between variable pairs.

The next step before doing the analytics was to create graphs of the cases and deaths per capita versus the number of days since the first reported case in New York. It was decided to only do so for the eight counties with the highest case rates to allow the graphs to be legible and useful. The graph in Figure 2 shows the cases increase drastically at day 20, and they increase overall until around day 45, they then decrease and around day 60 the cases are back down to where they were at day 20. The graph for deaths per capita in Figure 3 shows the deaths begin to increase drastically around day 30, peaking around day 37, then decreasing steadily thereafter. It should be noted that the first graph made for the deaths showed 0 deaths for every county on days 38 and 39, followed by a massive spike at day 40. It was inferred that the deaths for days 38 and 39 were not reported on their respective days, they were then added to the count on day 40. The deaths on day 38, 39, and 40 were then set to be one third the deaths on day 40. The subsequent graph made much more sense, so that correction was carried through in the analysis.

Before running any models, the data was split into 75% train and 25% test datasets. Every model was trained on the train set and tested on the test set. All models trained used the same set of socio-economic predictors described in detail in the first paragraph. After creating each model, the varImp() function was run on the model to determine the variable importance for each predictor. The attempt to use boosting did not work because the dataset is not large enough. The first models created were linear models using the lm() function. Two linear models were made, one for the cases per capita (CPC) and one for the deaths per capita (DPC). The linear model of the CPC had a root mean squared error (RMSE) of 0.0501 and a R^2 of 0.573. Based on the linear model, the strongest predictors for the CPC were population density and poverty rate, with variable importances of 25.1 and 2.58 respectively. The coefficient estimates and t-statistics can be found in Figure 4 and variable importance in Figure 5 for the CPC linear model. Variable importance from Figure 5 is calculated using the values shown in Figure 4. The linear model for the DPC had a RMSE of 0.00462 and a R^2 of 0.433. The linear model of the DPC had a root mean squared error (RMSE) of 0.00462 and a R^2 of 0.433. Based on the linear model, the strongest predictors for the DPC were population density, having some college, having a bachelor's degree or higher, only a high school diploma, and no high school diploma with variable importances of 21.3 , 2.17, 2.09, 1.99, and 1.82 respectively. The coefficient estimates and t-statistics can be found in Figure 6 and variable importance in Figure 7 for the DPC linear model. Variable importance from Figure 6 is calculated using the values shown in Figure 7.

The random forest models were created using the randomForests() function and varImpPlot() was used to plot the variable importance. The models were run with n.tree = 2000, this tree number resulted in an accurate model that did not take very long to run. Random forest models were created for CPC and DPC. The CPC model had an RMSE of 0.0496 and accounted for 56.2% of the variance. The strongest predictors were percent Hispanic, percent black, having a bachelor's degree or higher, and percent Asian with variable importances of 0.826, 0.797, 0.735, and 0.606 respectively. Variable importance numbers are shown in Figure 8 and plotted in Figure 9. The DPC model had an RMSE of 0.00462 and accounted for 38.3% of the variance. The strongest predictors were percent white, having some college, having just a high school diploma, and population density with variable importances of 0.00568, 0.00556, 0.00540, and 0.00335 respectively. Variable importance numbers for DPC are shown in Figure 10 and plotted in Figure 11.

The bagging model was created using the bagging() function. This was done with nbagg = 200, coob = TRUE, minsplit = 2, cp = 0, and n.trees = 2000. These parameters were chosen to get a working model that was accurate and did not take too long to run. Bagging models were created for CPC and DPC. The CPC model had an RMSE of 0.0497. It should be noted that one cannot obtain the % of variance explained from a bagging model. The strongest predictors for the CPC model were population density, percent Hispanic, percent white, and percent black with variable importances of 2.33, 2.89, 2.01, and 1.88 respectively. Variable importance numbers are shown in Figure 12. The DPC model had an RMSE of 0.00461. The strongest predictors were percent white, population density, percent Hispanic, and percent black with variable importances of 1.33, 1.32, 1.10, and 1.03 respectively. Variable importance numbers for DPC are shown in Figure 13.

All of the models had very similarly low errors and similar amounts of variance explained. Because of this fact, it was decided to use a conglomeration of all three model types for the analysis. When predicting the cases per capita, the top predictor by far would be population density. This would make sense, as such an easily transmittable virus is going to spread to more people faster when people are closer together, no matter what demographic they belong to. The other two predictors which seem to be important in predicting CPC are percent Hispanic and percent of people below the poverty line. These predictors likely correlate with high infection rates because if a given individual is Hispanic and/or in poverty, they likely live in a city, that city is likely densely populated. When predicting the deaths per capita, the top predictor by far would be also be population density. This makes sense that the most powerful predictor for cases by far would also predict the most DPC. The other strong predictors of DPC are percent Hispanic and percent white. These findings are important, especially since the average age of a white American is much older than the average age of a Hispanic American (Source 7). The median age for a white American is 44 years, and 30 years for a Hispanic American. The mode for a white American is 58 years old and 11 years old for a Hispanic American. About 80% of COVID-19 deaths have been suffered by people of age 65 or higher (Source 8). Since the white population is much older, it would make sense that their case rate wouldn't be outstanding, but their death rate would be higher than other ethnicities. It is concerning that this data suggests being Hispanic correlates with a high death rate, especially since Hispanics in the US are much younger on average. Many of these deaths are likely preventable, this could be due to lack of education on what one should do if they have symptoms. It is likely that Hispanic people who live in NY and are not US citizens are scared of seeking medical treatment in fear of legal repercussions and are avoiding hospitals. Those in fear of seeking medical treatment are far more likely to spread the virus further in their community and are more likely to die. From a risk analysis perspective, it seems like the biggest risk factor in cases and deaths is obviously population density, this is well known and accounted for in current policies. The risk factors which are most important for new policy implementation would be the high rates of death and infection among those who are Hispanic and/or living in poverty, especially if they're in an urban environment. This implies that the policy which would result in the most harm reduction (which has not been implemented yet) would be one that affects the mentioned demographic. This could materialize in medical teams visiting these areas, setting up triages in which they would offer testing and treatment. Leaders should encourage people who are specifically Hispanic and/or in poverty to reach out for testing and treatment without hesitation.

The data used in this paper is novel and is likely underestimating the number of cases. Testing for COVID-19 must become far more widespread to receive accurate case numbers. It would also be very useful to have accurate data on recovery rates. Early studies suggest the cases rate in New York City is much higher than this data suggests (Source 9). This analysis should be repeated as more data comes out. It would also be useful to have a dataset of the cases and deaths in NYC broken up by district. When used in conjunction with socio-economic data, an analysis of such data could give one far more insight into the details relating to heavily affected populations. These accurate and specific results could help allocate help to the most affected areas in the US.

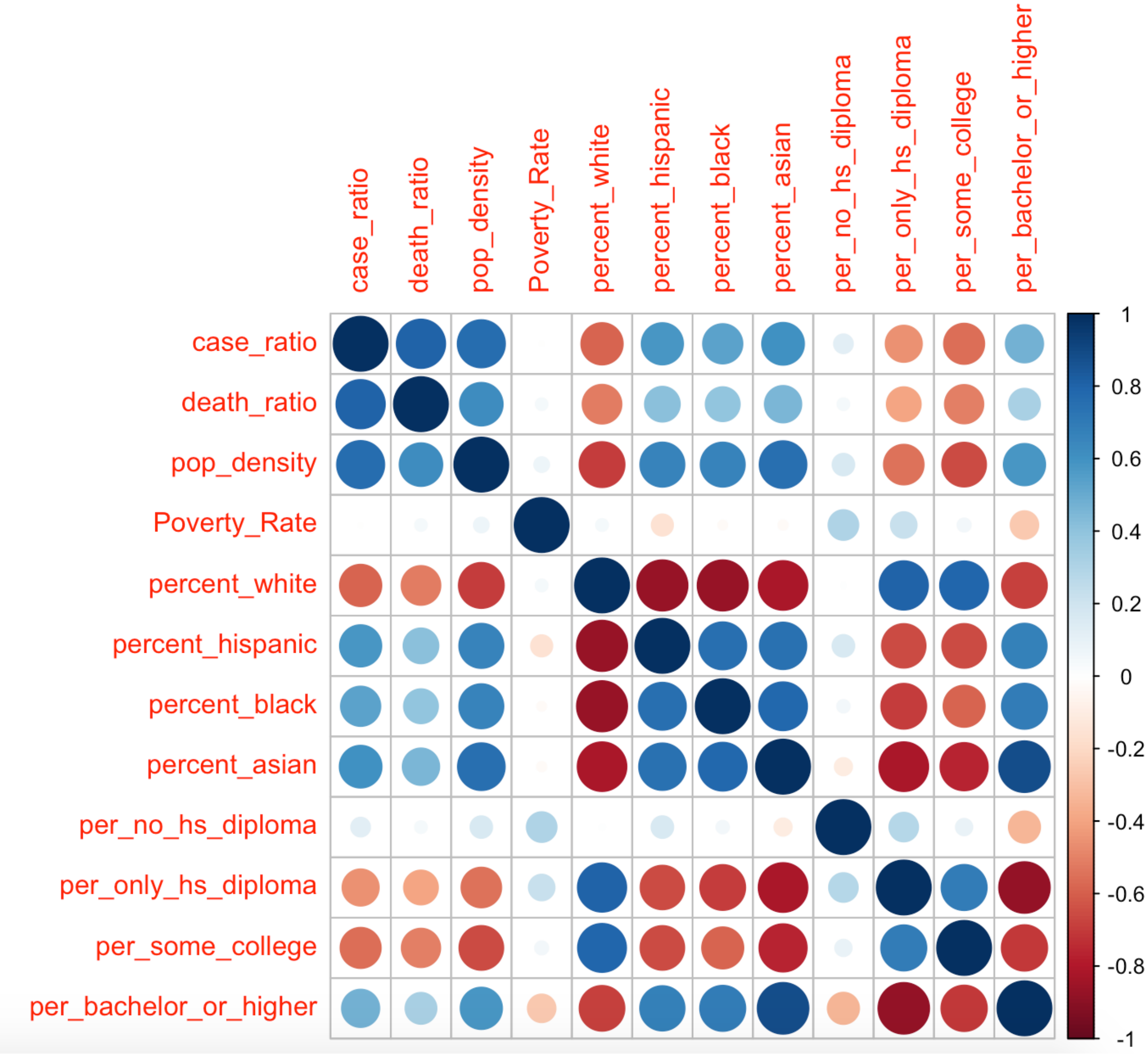


Figure 1. Correlation Plot

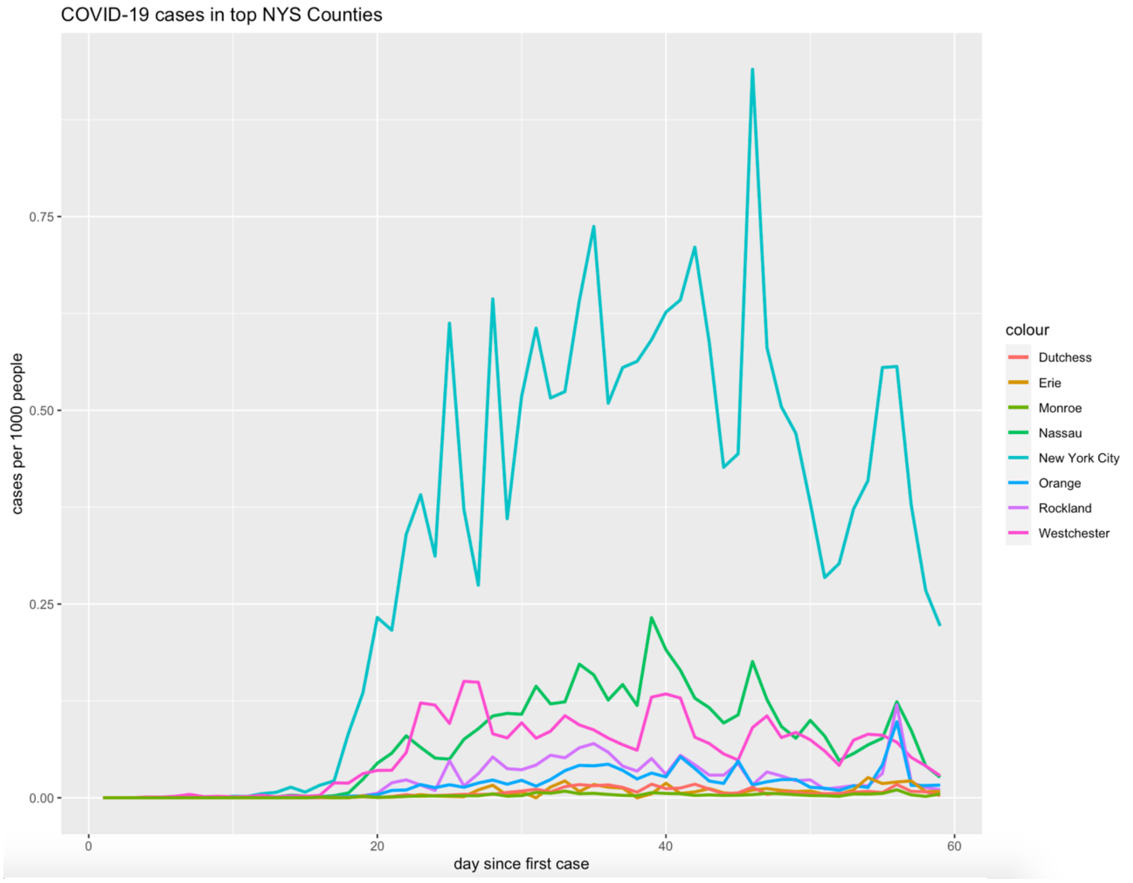


Figure 2. Plot of cases per capita per day since initial case

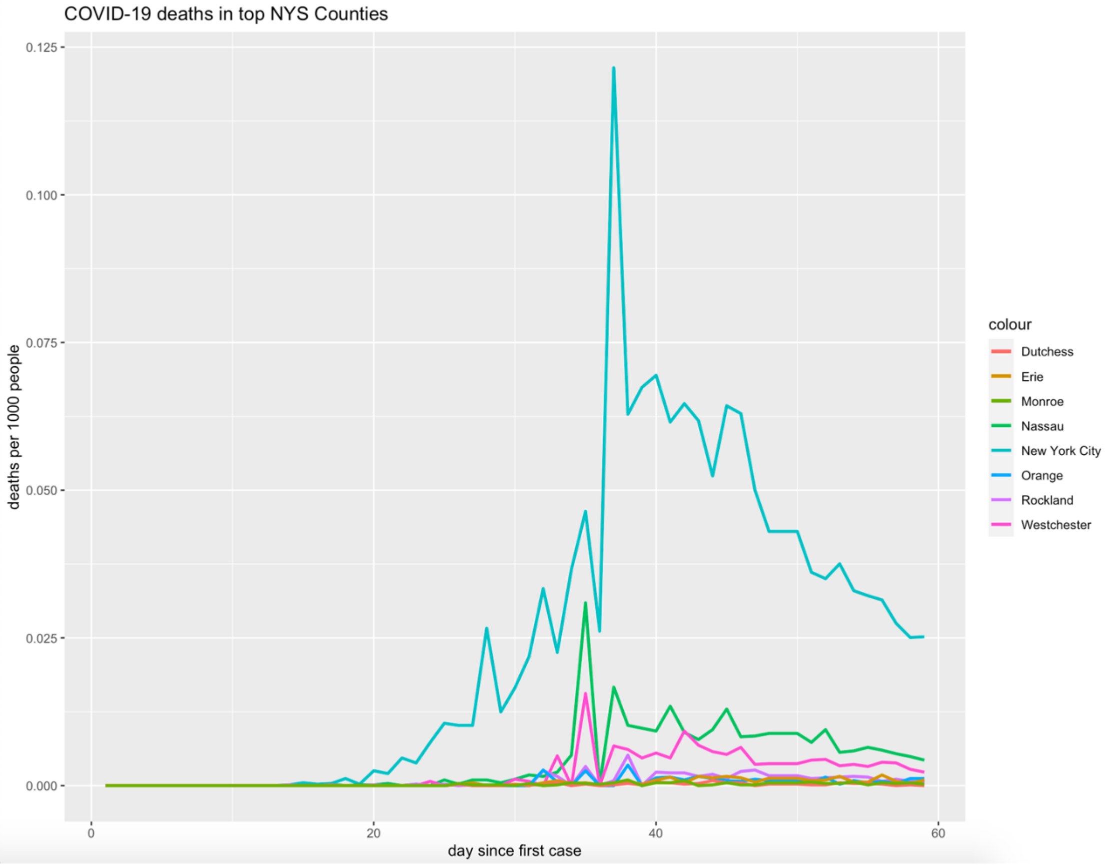
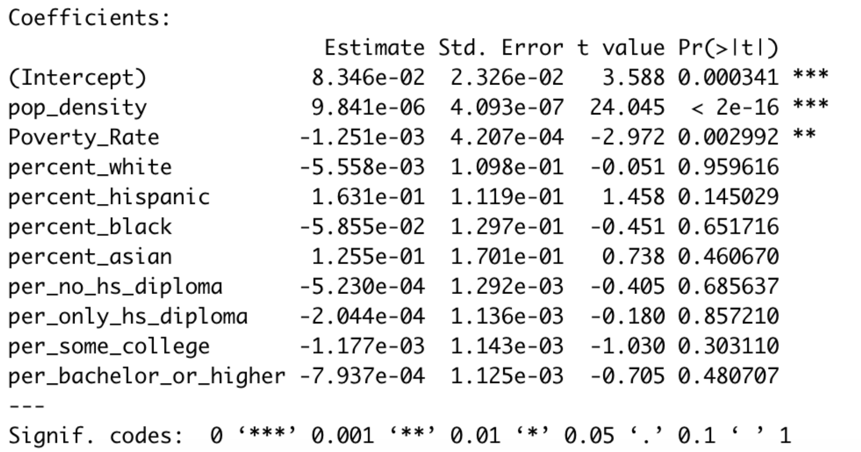
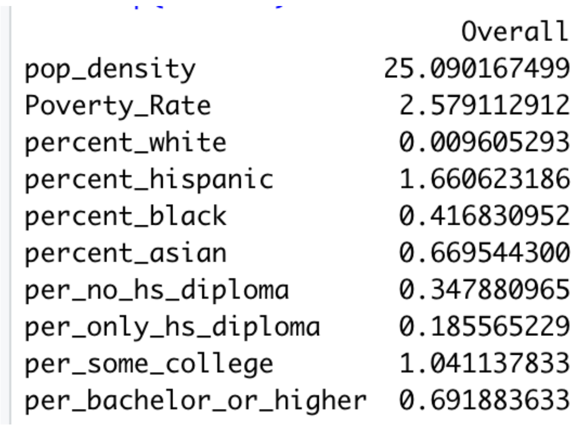


Figure 3. Plot of deaths per capita per day since initial case



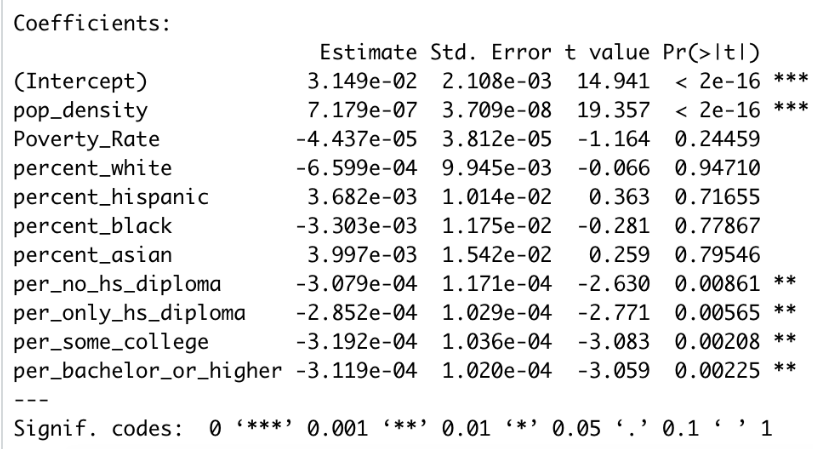
```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             8.346e-02  2.326e-02   3.588 0.000341 ***
pop_density             9.841e-06  4.093e-07  24.045  < 2e-16 ***
Poverty_Rate           -1.251e-03  4.207e-04  -2.972 0.002992 ** 
percent_white          -5.558e-03  1.098e-01  -0.051 0.959616    
percent_hispanic        1.631e-01  1.119e-01   1.458 0.145029    
percent_black          -5.855e-02  1.297e-01  -0.451 0.651716    
percent_asian           1.255e-01  1.701e-01   0.738 0.460670    
per_no_hs_diploma      -5.230e-04  1.292e-03  -0.405 0.685637    
per_only_hs_diploma    -2.044e-04  1.136e-03  -0.180 0.857210    
per_some_college       -1.177e-03  1.143e-03  -1.030 0.303110    
per_bachelor_or_higher -7.937e-04  1.125e-03  -0.705 0.480707    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4. Predictors for Cases Linear Model



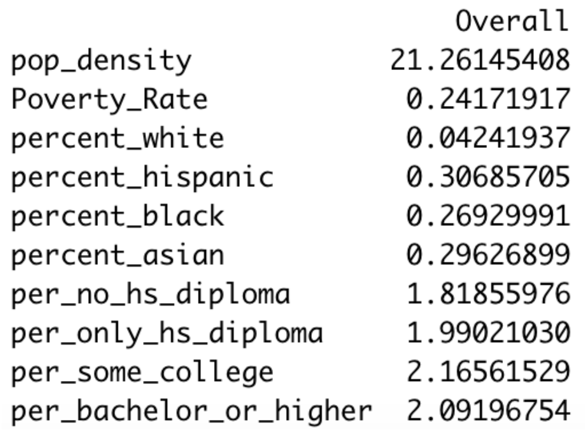
```
                           Overall
pop_density            25.090167499
Poverty_Rate            2.579112912
percent_white           0.009605293
percent_hispanic        1.660623186
percent_black           0.416830952
percent_asian           0.669544300
per_no_hs_diploma       0.347880965
per_only_hs_diploma     0.185565229
per_some_college        1.041137833
per_bachelor_or_higher  0.691883633
```

Figure 5. Variable importance for Cases Linear Model



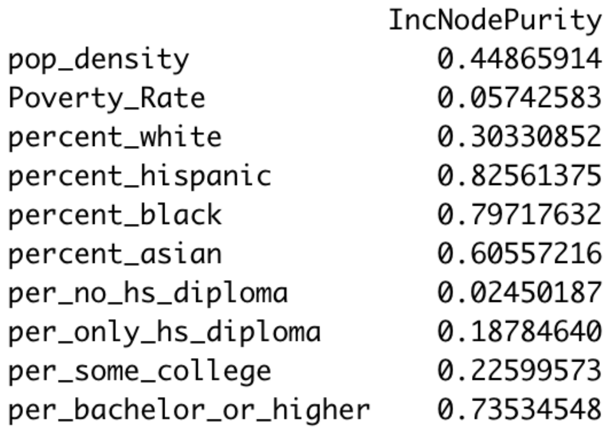
```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             3.149e-02  2.108e-03  14.941  < 2e-16 ***
pop_density             7.179e-07  3.709e-08  19.357  < 2e-16 ***
Poverty_Rate           -4.437e-05  3.812e-05  -1.164  0.24459    
percent_white          -6.599e-04  9.945e-03  -0.066  0.94710    
percent_hispanic        3.682e-03  1.014e-02   0.363  0.71655    
percent_black          -3.303e-03  1.175e-02  -0.281  0.77867    
percent_asian           3.997e-03  1.542e-02   0.259  0.79546    
per_no_hs_diploma      -3.079e-04  1.171e-04  -2.630  0.00861 ** 
per_only_hs_diploma    -2.852e-04  1.029e-04  -2.771  0.00565 ** 
per_some_college       -3.192e-04  1.036e-04  -3.083  0.00208 ** 
per_bachelor_or_higher -3.119e-04  1.020e-04  -3.059  0.00225 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6. Predictors for Deaths Linear Model



```
                          Overall
pop_density            21.26145408
Poverty_Rate            0.24171917
percent_white           0.04241937
percent_hispanic        0.30685705
percent_black           0.26929991
percent_asian           0.29626899
per_no_hs_diploma       1.81855976
per_only_hs_diploma     1.99021030
per_some_college        2.16561529
per_bachelor_or_higher  2.09196754
```

Figure 7. Variable importance for Deaths Linear Model



```
                        IncNodePurity
pop_density               0.44865914
Poverty_Rate              0.05742583
percent_white             0.30330852
percent_hispanic          0.82561375
percent_black             0.79717632
percent_asian             0.60557216
per_no_hs_diploma         0.02450187
per_only_hs_diploma       0.18784640
per_some_college          0.22599573
per_bachelor_or_higher    0.73534548
```

Figure 8. Variable importance for Cases Random Forests

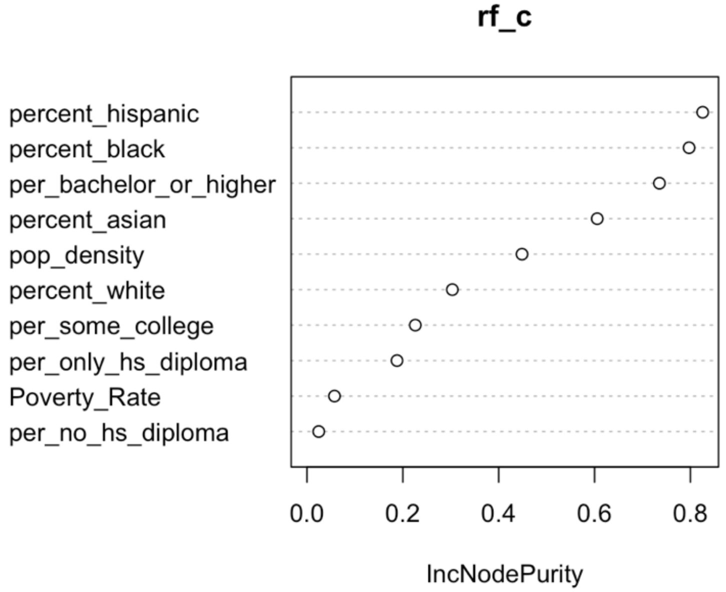
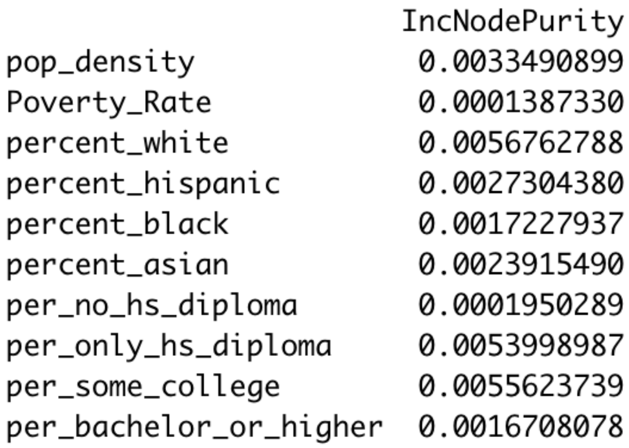


Figure 9. Variable importance Plot for Cases Random Forests



```
                        IncNodePurity
pop_density             0.0033490899
Poverty_Rate            0.0001387330
percent_white           0.0056762788
percent_hispanic        0.0027304380
percent_black           0.0017227937
percent_asian           0.0023915490
per_no_hs_diploma       0.0001950289
per_only_hs_diploma     0.0053998987
per_some_college        0.0055623739
per_bachelor_or_higher  0.0016708078
```

Figure 10. Variable importance for Deaths Random Forests

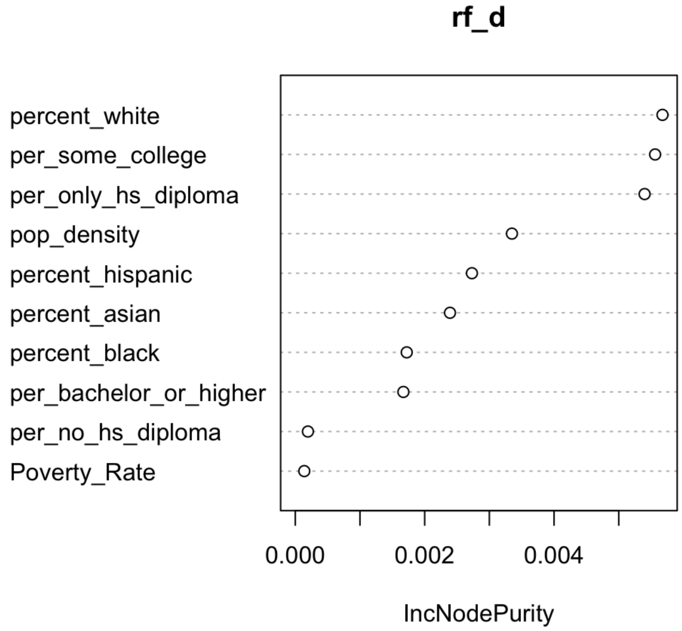
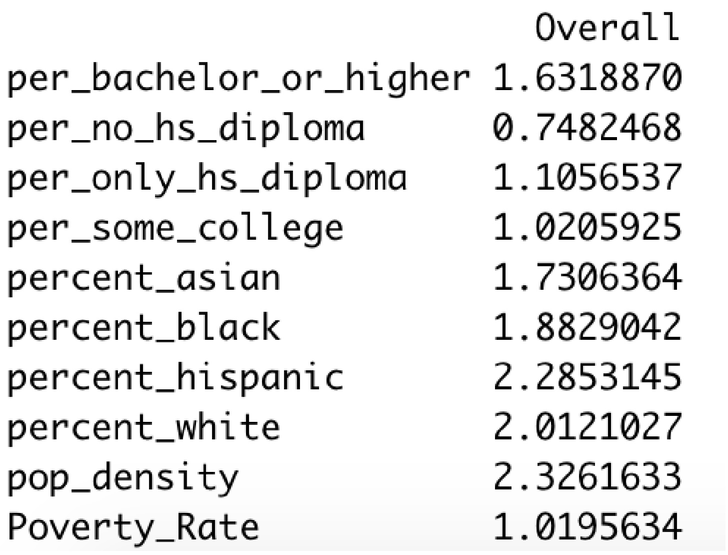
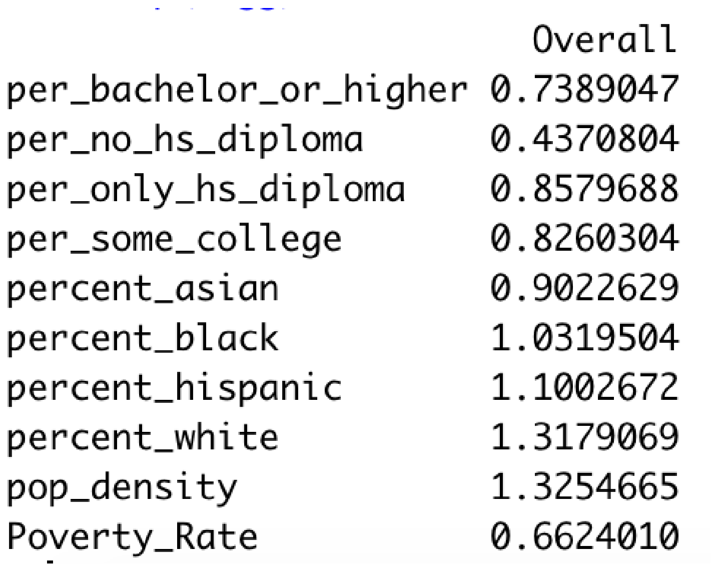


Figure 11. Variable importance Plot for Deaths Random Forests



```
                          Overall
per_bachelor_or_higher 1.6318870
per_no_hs_diploma      0.7482468
per_only_hs_diploma    1.1056537
per_some_college       1.0205925
percent_asian          1.7306364
percent_black          1.8829042
percent_hispanic       2.2853145
percent_white          2.0121027
pop_density            2.3261633
Poverty_Rate           1.0195634
```

Figure 12. Variable importance for Cases Bagging



```
                          Overall
per_bachelor_or_higher 0.7389047
per_no_hs_diploma      0.4370804
per_only_hs_diploma    0.8579688
per_some_college       0.8260304
percent_asian          0.9022629
percent_black          1.0319504
percent_hispanic       1.1002672
percent_white          1.3179069
pop_density            1.3254665
Poverty_Rate           0.6624010
```

Figure 13. Variable importance for Deaths Bagging

Work Cited

1. https://githulkjhg b.com/nytimes/covid-19-data/blob/master/us-counties.csv
2. “Annual Population Estimates for New York State and Counties: Beginning 1970: Open Data NY.” *State of New York*, data.ny.gov/Government-Finance/Annual-Population Estimates-for-New-York-State-and/krt9-ym2k/data.
3. “Department of Health.” *Table 2: Population, Land Area, and Population Density by County, New York State - 2015*, www.health.ny.gov/statistics/vital_statistics/2015/table02.htm.
4. Holshue, Michelle L., et al. “First Case of 2019 Novel Coronavirus in the United States: NEJM.” *New England Journal of Medicine*, 22 Apr. 2020, www.nejm.org/doi/full/10.1056/NEJMoa2001191.
5. “Department of Labor.” *Population Data and Projections - New York State Department of Labor*, labor.ny.gov/stats/nys/statewide-population-data.shtm.
6. “Education.” *Education*, data.ers.usda.gov/reports.aspx?ID=17829.
7. Schaeffer, Katherine. “The Most Common Age among Whites in U.S. Is 58 – More than Double That of Racial and Ethnic Minorities.” *Pew Research Center*, Pew Research Center, 30 July 2019, www.pewresearch.org/fact-tank/2019/07/30/most-common-age among-us-racial-ethnic-groups/.
8. Sonabend, Raphael. “Coronavirus and Probability- The Media Must Learn How to Report Statistics Now.” *Medium*, Towards Data Science, 11 Mar. 2020, towardsdatascience.com/coronovarius-and-probability-the-media-must-learn-how-to report-statistics-now-973ed2d52959.
9. De Avila, Joseph. “New York, New Jersey Lay Out Criteria for Reopening Economics.” *Wall Street Journal*, 28 Apr. 2020.