Data Science Final Project

Anandi Gupta

11/10/2021

Introduction

The Covid-19 pandemic has devastated the incarcerated population in the U.S., infecting over 400,000 prisoners and killing over 2,600 prisoners to date.[1] It is estimated that one in three inmates in state prisons were infected by the virus as of March 2021.[2] Overcrowding, low testing rates, poor sanitation, and limited access to personal protective equipment have made prisoners particularly vulnerable to the disease.

For my project, I compare infection rates and death rates among incarcerated populations in state and federal prisons and the general population by county to highlight the disproportionate impact the pandemic has had on prisoners. I further examine the extent to which county level demographic and economic characteristics can predict the gap in infection rates between the incarcerated population and the general population.

In the first section of this report, I provide a problem statement and background to contextualize why this research question is important. In the next two sections, I describe the data I obtained and the machine learning techniques I used to wrangle, visualize, and model the relevant data answer my research question. In the last two sections, I discuss the results of the analysis and my recommendations for future research.

Background and Problem Statement

The U.S. has the highest incarceration rate in the world, with over 2 million people behind bars. Research suggests that the incarcerated population faced disproportionate consequences of the pandemic due to poor conditions in these facilities such as high prisoner density, lack of room for social distancing, limited medical supplies, and inadequate sanitation, which were conducive to rapid community transmission of Covid-19. In the one-year period leading up to March 2021, these facilities recorded over 1,400 cases and 7 deaths per day on average.[3]

State and local authorities undertook various policies to curb the spread of the disease among this vulnerable population including providing testing and protective equipment such as masks to incarcerated people and prison staff, reducing admissions and releasing prisoners, eliminating medical co-pays, and prioritizing the incarcerated population in early stages of the vaccine rollout. Policy responses in states, counties, and localities varied widely. For example, at the state level, the New Jersey legislature passed a bill that allowed for people with less than a year left on their sentences to be released up to eight months early starting October 19th. Similarly, at the county level, the Hays County sheriff announced a new "Cite and Divert" program in an effort to reduce arrests, jail time, and criminal charges.[4]

[1] National Overview. (2021, September 21). COVID Prison Project. https://covidprisonproject.com/data/national-overview/

[2] Burkhalter, E., Colón, I., Derr, B., Gamio, L., Griesbach, R., Klein, A. H., Issawi, D., Mensah, K., Norman, D. M., Redl, S., Reynolds, C., Schwing, E., Seline, L., Sherman, R., Turcotte, M., & Williams, T. (2021, August 12). Incarcerated and Infected: How the Virus Tore Through the U.S. Prison System. The New York Times. Available at https://www.nytimes.com/interactive/2021/04/10/us/covid-prison-outbreak.html.

[3] Burkhalter, E., Colón, I., Derr, B., Gamio, L., Griesbach, R., Klein, A. H., Issawi, D., Mensah, K., Norman, D. M., Redl, S., Reynolds, C., Schwing, E., Seline, L., Sherman, R., Turcotte, M., & Williams, T. (2021, August 12). Incarcerated and Infected: How the Virus Tore Through the U.S. Prison System. The New York Times. Available at https://www.nytimes.com/interactive/2021/04/10/us/covid-prison-outbreak.html.

[4] Criminal justice responses to the coronavirus pandemic. Prison Policy Initiative. 2021. https://www.prisonpolicy.org/virus/virusresponse.html

In this project, I examine the extent to which the effects of the pandemic were concentrated among the incarcerated, by comparing their infection rates and death rates to those of their surrounding communities. I further explore whether there exists a relationship between regional economic and demographic characteristics and the performance of prisons or jails during the pandemic? Although several policies related to reducing the burden of the pandemic on the incarceration population were implemented at the state level, I choose to conduct my analysis at the county level to expand my dataset beyond 50 observations.

In order to develop the framework for my models, I reviewed literature investigating county level predictors of 1) Covid-19 infection rates in the general population and 2) incarceration rates. For example, McLaughlin et al. (2021) found that rates of COVID-19 cases and deaths were higher in US counties that were more urban or densely populated or that had more crowded housing, air pollution, women, persons aged 20–49 years, racial/ethnic minorities, residential housing segregation, income inequality, uninsured persons, diabetics, or mobility outside the home during the pandemic. [5]

For incarceration rates, Riley et al. (2018) found that county-level poverty, police expenditures, and spillover effects from other county and state authorities are significant predictors of local jail rates. [6] Additionally, Durante (2017) found that the presence of large shares of African Americans and of Republican voters were indicative of the total prison admission rates in a region. [7]

Thus, for my model I choose economic variables such as poverty rates, unemployment rates, and median household income, demographic variables such as race (percent of the population that is White) and education, health indicators such as the percent of the population that is uninsured, and political ideology measured by the share of vote for President Trump in 2016 as my predictors of the performance gap in infection rates between the incarcerated population and the general population in the county.

Data

For my project, I utilized data from numerous public sources. For Covid-19 data in prisons, I utilized a novel dataset published by the New York Times that tracked Covid-19 cases in prisons and jails at the facility level through March 2021, and aggregated the data to the county level, which is the unit of analysis I chose for the project. Specifically, I focus on state and federal prisons for my analysis. I converted cases and death counts into rates by dividing by the inmate population provided in the same source. To be conservative, I used the maximum 2020 inmate population as my denominator where available. If this variable was missing for a facility, I used the latest inmate population. There were 6 instances where the rate of Covid-19 infections was greater than 100%. that appear to be data errors. Given the small number of instances, and my hesitancy in imputing my outcome variables (particularly given the relatively small size of the dataset), I dropped these cases from my dataset. Thus, my final dataset consisted of Covid-19 counts from prisons in 751 counties.

In order to maintain consistency between my sources, I scraped county level data from the New York Times that tracked counts of cases and deaths by county using Pandas. To ensure consistency in the timeframe used for this analysis, I limit this data to cumulative counts of cases and deaths on March 31, 2021. I manually cleaned the data (for example, the FIPS code for "New York City" was missing and had to be manually entered.) I then converted case and death counts to rates by dividing by county populations that I merged in from data published by the U.S. Department of Agriculture.

I then merged these datasets together and constructed two variables that are my primary outcomes of interests: 1) the difference in Covid-19 cases per 100 population between the incarcerated population and the general population in a county; and 2) the difference in Covid-19 death rates per 1000 population (for ease of visualization and interpretation) between the incarcerated population and the general population in a county.

[5] McLaughlin, J. M., Khan, F., Pugh, S., Angulo, F. J., Schmitt, H. J., Isturiz, R. E., ... & Swerdlow, D. L. (2021). County-level Predictors of Coronavirus Disease 2019 (COVID-19) Cases and Deaths in the United States: What Happened, and Where Do We Go from Here?. Clinical Infectious Diseases, 73(7), e1814-e1821.

[6] Rachael Weiss Riley, Jacob Kang-Brown, Chris Mulligan, Vinod Valsalam, Soumyo Chakraborty & Christian Henrichson (2018) Exploring the Urban–Rural Incarceration Divide: Drivers of Local Jail Incarceration Rates in the United States, Journal of Technology in Human Services, 36:1, 76-88, DOI: 10.1080/15228835.2017.1417955

[7] Durante, K. A. (2020). Racial and Ethnic Disparities in Prison Admissions Across Counties: An Evaluation of Racial/Ethnic Threat, Socioeconomic Inequality, and Political Climate Explanations. Race and Justice, 10(2), 176–202. https://doi.org/10.1177/2153368717738038

For my prediction variables, I obtained data on county-level unemployment, poverty, and graduation rates from the U.S. Department of Agriculture. I additionally obtained the percent of the population under the age of 65 that was uninsured in 2019, county-level population density statistics, and county-level race demographics that I used to calculate the percent of the population in the county that is White from the U.S. Census Bureau. I obtained county-level incarceration rates for 2020 published by the Marshall Project. I also included a score indicating how rural or urban a county is obtained from U.S. Department of agriculture, and Presidential election results by constituency from the MIT Election Data Science Lab that I used to calculate the share of votes for Donald Trump in the 2016 election. I merged these datasets using County FIPS codes (that I converted to be numeric to ensure successful merges).

Finally, I obtained estimates of the operating level of prisons (as a percent of capacity) at the state level. Ideally, I would have been able to identify these numbers at the county level, but chose to include the values using a state-level merge, as policies to release prisoners/reduce admissions during the pandemic were often implemented state-wide. Note, Connecticut and Ohio did not report capacity data dropped from my analysis.

Analysis

Beyond the initial data wrangling and cleaning described in the data description above, my analysis consisted of two main components: 1) visualization and exploration, and 2) modeling. For my visualizations in the initial data wrangling stage, I used the Geopandas library to create spatial depictions of the difference in case rates and death rates at the county level. See Figures 1 and 2 below.

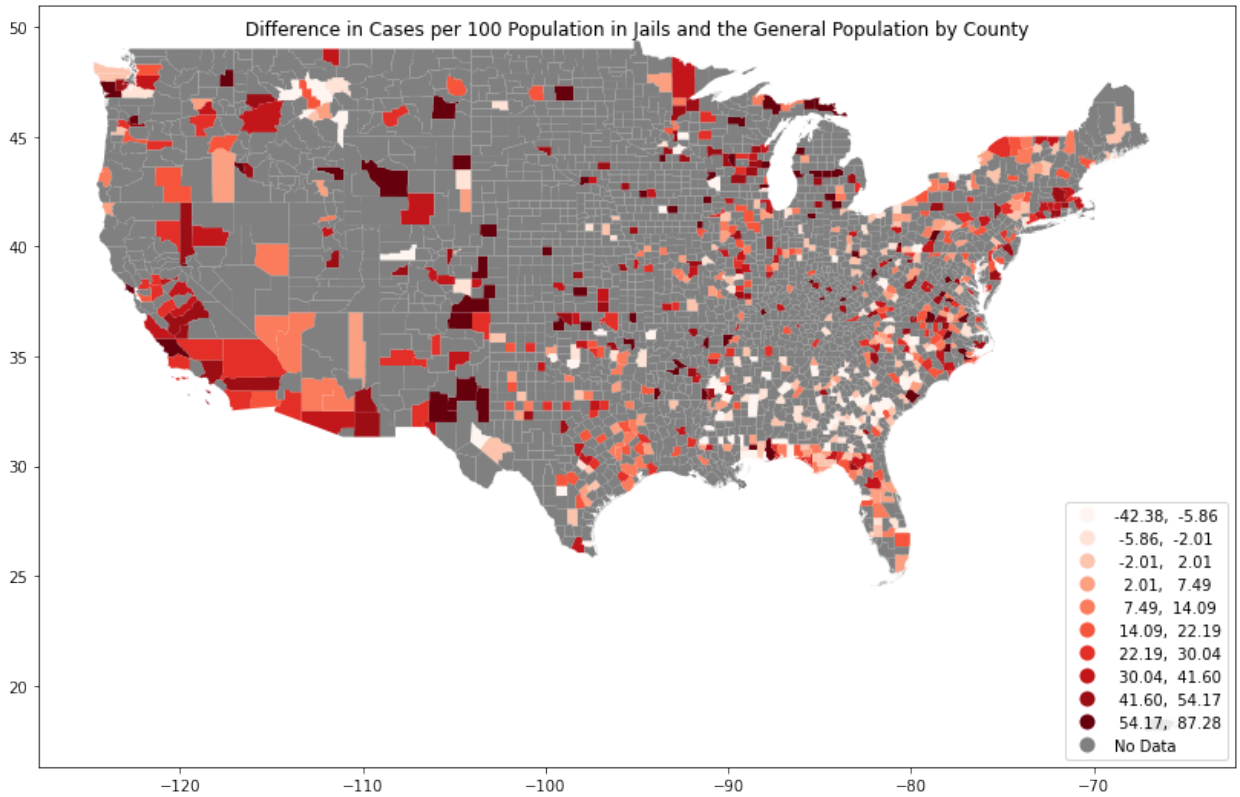

Figure 1: Comparison of Case Rates in Prisons vs the General Population by County

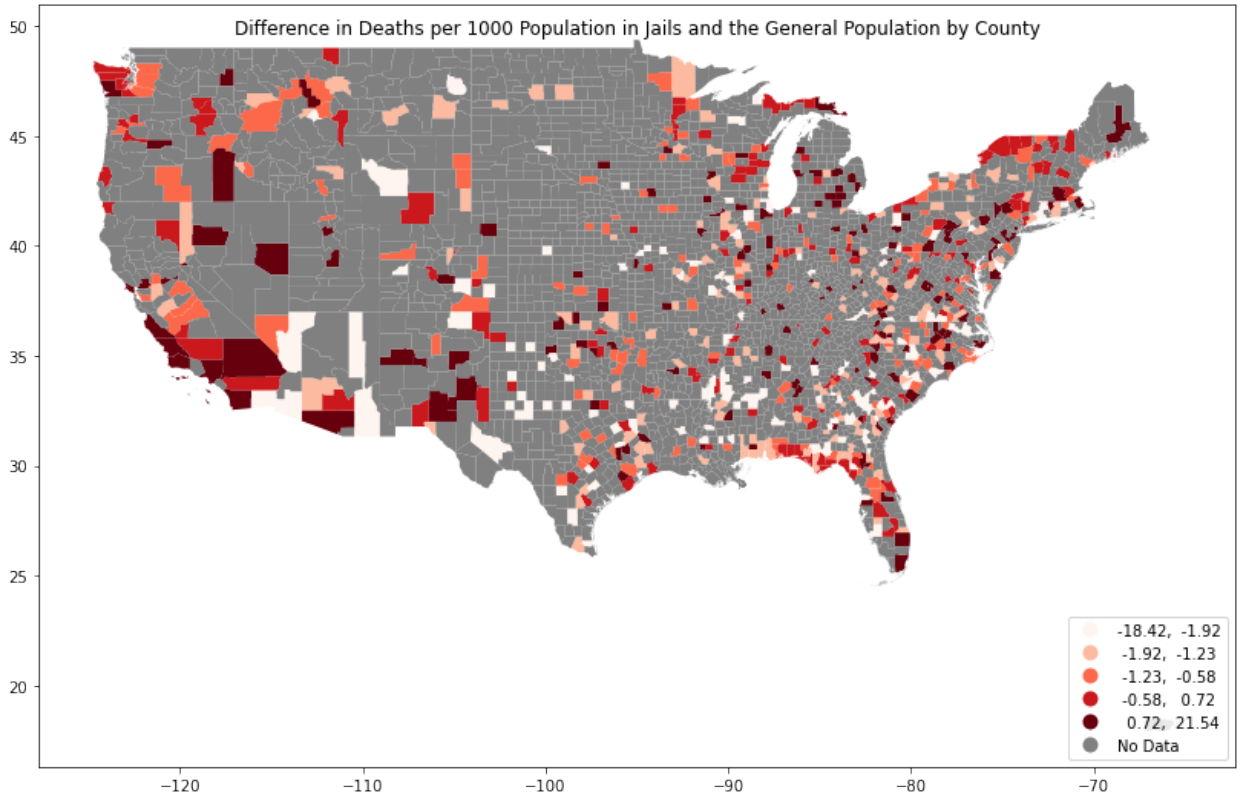

Figure 2: Comparison of Death Rates in Prisons vs the General Population by County

From these visualizations it is evident that there was substantial variation in the performance of prisons versus the general population for Covid-19 case rates (such that in the majority of counties, prisons performed worse than the general population). The visualization shows that in several counties, case rates in prisons exceeded rates in the general population by more than 20 cases per 100 population and in some counties, the performance gap is over 50 cases per 100 population. For death rates, the difference between rates in the incarcerated and general population was substantially lower, and in many counties death rates were lower in prisons as compared to the general population. I hypothesize that these discrepant patterns are driven by the higher mortality rates among older people (as I suspect the percentage of the population that is over the age of 65 and particularly vulnerable to dying from the disease is much lower in prisons.) Thus, for the purposes of modeling, I chose to focus on the difference in cases rates between the incarcerated and general population as my outcome variable.

I then examined the correlation between my predictor variables by creating a correlation matrix using the Seaborn package.[8] Through this process, I found that poverty rates, the percent the population that is uninsured, race (measured by the percent of the population that is White), and education levels (specifically, the percent of adults in the county with less than a high school diploma) were most correlated with my dependent variable (see results in Figure 3 below).

[8]Michael Waskom and team. mwaskom/seaborn, Zenodo, (2020). DOI: 10.5281/zenodo.592845

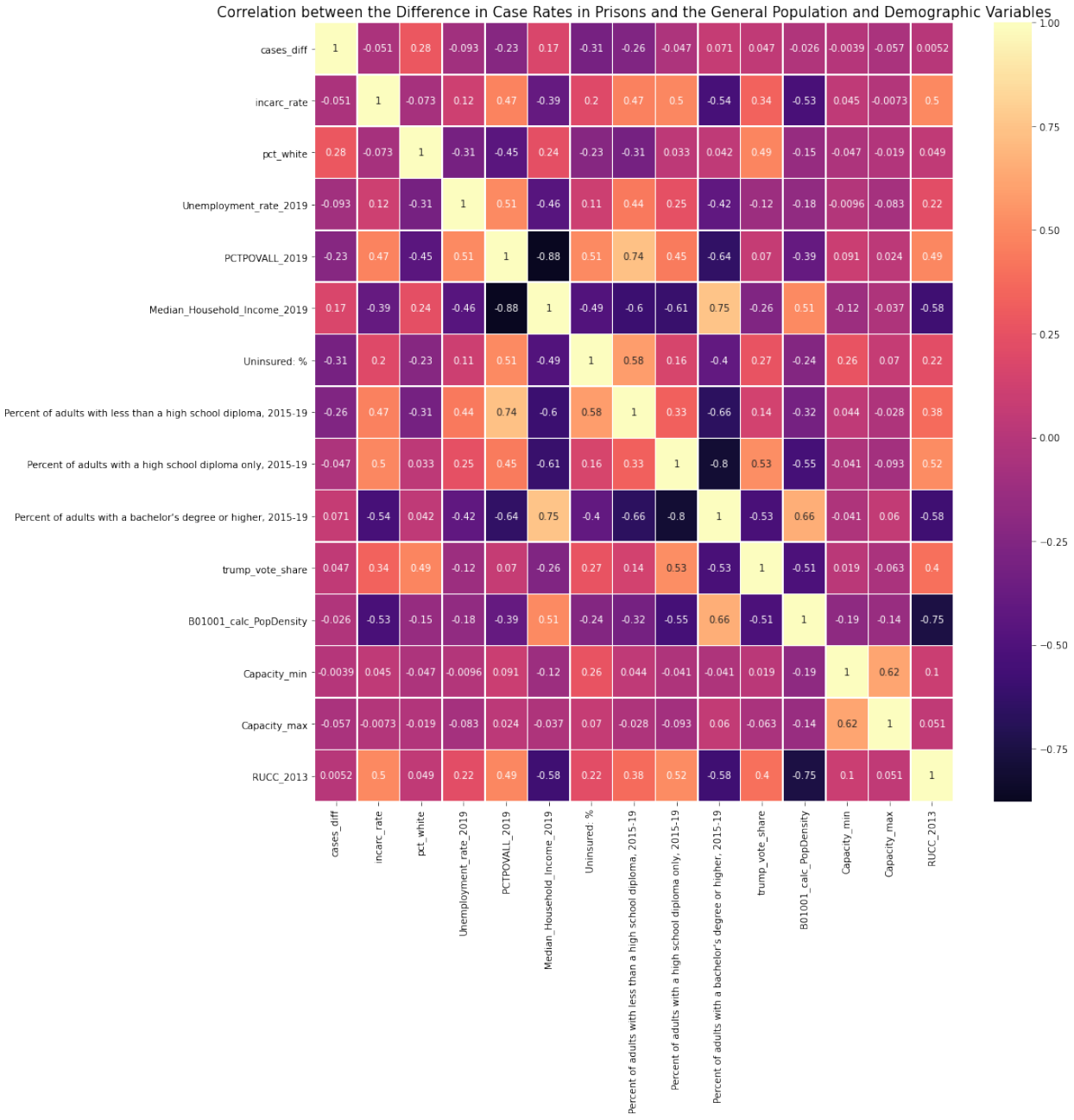

Figure 3: Relationships between Independent Variables and Dependent Variable

Lastly, I modelled my data using Scikit-learn machine learning techniques applicable to regression problems.[9] Specifically, I ran 4 models – a simple linear model (aka OLS regression), K-Nearest Neighbors (KNN), Decision Trees and Random Forests. I then split my data into a test and training dataset using random sampling (specifically I used a split 75% of the total observations for my training dataset, and 25% for my test dataset). I then examined the distribution of my variables in the training dataset to determine

[9]Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011).

the transformations I would need to make. I found that several of my continuous variables such as poverty rates, insurance rates, unemployment rates, population density, and incarceration rates were highly skewed. To address this, I included a logarithmic transformation of these variables in my pre-processing pipeline. I additionally scaled each of these variables in my pre-processing pipeline. I did this separately for my test and training datasets as this involved learning information such as the mean, minimum, and maximum values.

I chose negative mean squared error as my scoring metric for measuring the performance of each of my models. I tuned the hyper-parameters used in my models and ran my training data through my modeling pipeline. I then used GridSearchCV to identify the estimators that resulted in the highest predictive accuracy, as measured by Mean Squared Error (MSE) and the R-squared (R2) score. For the KNN model, this involved choosing the optimal number of "nearest neighbors" in my training data to make predictions in my test data. For the decision tree, this involved identifying the maximum branch depth to use. For the random forest model, I hyper-tuned both the maximum branch depth and the maximum number of predictors to use as inputs. Once I identified the model that resulted in the best score for my training data, I then fit the test data to my model by running it through my pipeline and generated MSE and R2 scores to evaluate the predictive accuracy of the model. I further visualized the predictive accuracy using a scatter plot. Finally, I used 40 permutations of the predictor variables within the model to determine the importance of each feature (measured by the amount the inclusion of each feature reduced the MSE) and depicted feature importance graphically.

Results and Conclusions

To arrive at my final specifications, I tested several models in which I included additional variables/dropped existing variables, applied transformations to additional variables to mitigate skewness, and different methodologies for addressing missing values (i.e., simply dropping rows with missing observations or imputing with the mean using the Sci-Kit Learn Simple Imputer package.

Several of these experiments yield extremely low (or even negative) R2 squares, and thus were excluded from my final specifications to increase the predictive power of my model.

The model with the best score (i.e., the lowest MSE) was the simple OLS linear regression model, which yielded an R2 score of 0.19 and an MSE of 409.7. The R2 indicates that the model was only able to explain 19 percent of the variation in the dependent variable, indicating that the model is not particularly useful for predicting whether prisons in certain counties performed worse compared to the general population in terms of Covid-19 case rates. Taking the square root of the MSE makes this metric easier to interpret, and indicates that on average, my model's predictions for the difference in case rates between the incarcerated population and the general population had an average error of approximately 20 cases per 100 population. Thus, overall, my model performs relatively poorly, and county-level characteristics do not appear to be good predictors of the performance of prisons relative to the general population in terms of Covid-19 case rates during the pandemic.

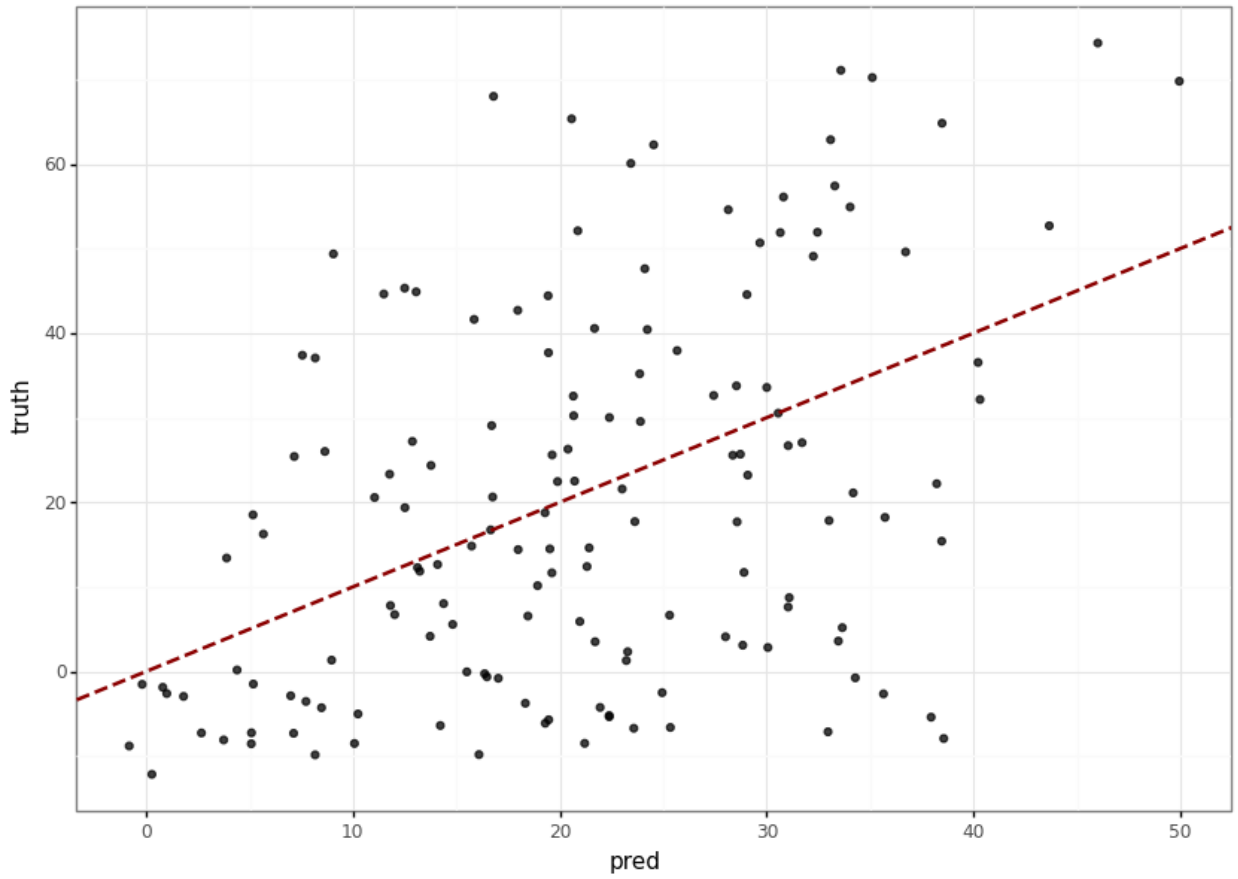

Figure 4: Goodness of Fit

Further, the permutation exercise indicates that of the county-level indicators I chose to include in my model, education levels in a county were most predictive of my outcome variable. The percent of adults with a bachelor's degree or higher, in particular, was most helpful in lowering the MSE. The percent of the population that is uninsured, race, and median household income were also important in lowering the MSE. On the other hand, county-level incarceration rates, population density, and unemployment rates did very little in lowering the MSE.

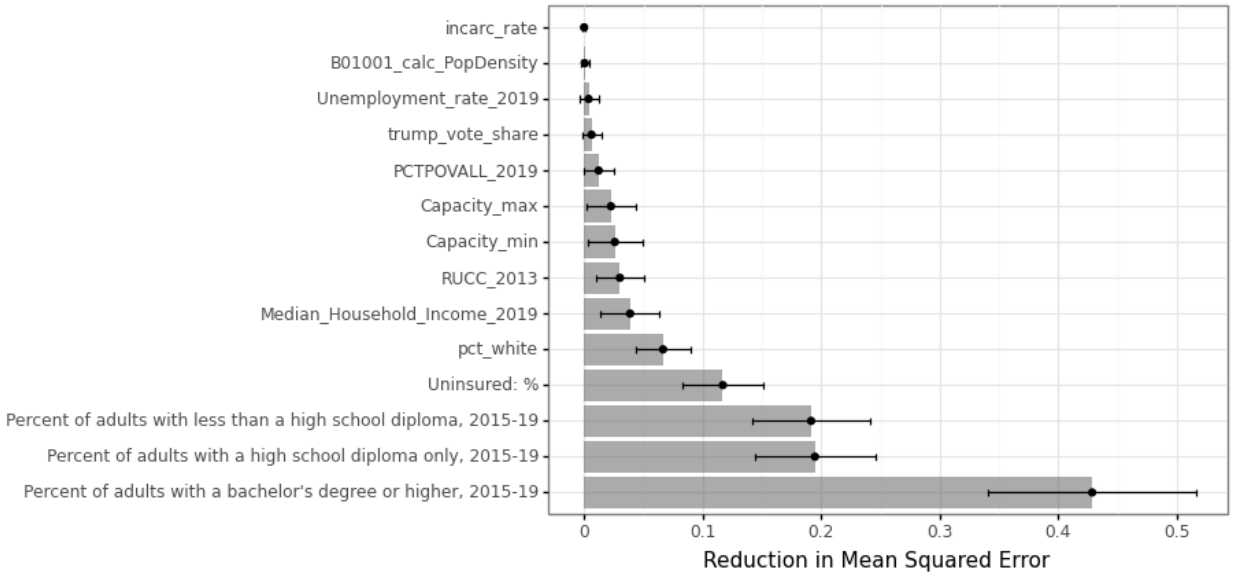

Figure 5: Permutation Exercise: Variables by Importance

Discussion

Upon reflecting more on my research question, I realize that is it not surprising that my model was unable to shed much light on the performance gap between prisons and the general population during the pandemic. The variables I chose to include in my model such as poverty rates, the percent of the population that is uninsured, unemployment rates, and race likely influenced both Covid-19 rates in the general population and within prisons, and thus would have low explanatory power for explaining the difference in performance between the two populations.

Further, research shows that Covid-19 outbreaks in prisons and jails were also associated with widespread transmission of the disease in the surrounding communities as well. For example, a study conducted by the Vera institute shows that at both the county level and broader community (areas that share a local economy) level, larger incarcerated populations were associated with earlier reported cases of COVID-19 in the spring of 2020 and with a spike in confirmed cases over the summer of 2020.[10] Thus, factors that led to higher case rates within prisons, likely also led to higher case rates in the surrounding community, indicating that the predictor variables I chose for my model, by construction, would not be able to explain the difference in case rates between the incarcerated and general population in a county.

It is plausible that factors such as correctional budgets and variations in policies such as the provision of masks and PPE, the release of prisoners or the lowering of prison admission rates during the pandemic, or the prioritization of the incarcerated population during the early stages of the vaccine rollout would have been better indicators of the performance gap. However, such data was hard to obtain, either because reporting from facilities was incomplete or because such policies were implemented at the state level and recommend that future research incorporate such factors when making predictions about the poor performance of prisons relative to the general population during the pandemic.

Overall, I believe I have met the goals I had set in my proposal for a "successful project." One of my priorities was to improve my data visualization skills. In both my project presentation and this report I included several publishable-quality visualizations including geospatial visualizations which included learning new libraries such as Geopandas. I also generally used efficient code, thought carefully about how the decisions I made during the data wrangling portion of the project would impact my results, and minimized iterating through unnecessary loops.

[10]Mass Incarceration, COVID-19, and Community Spread. Prison Policy Initiative. 2021. https://www.prisonpolicy.org/reports/covidspread.html

While I had hoped to build a model with higher predictive power, I have developed a better understanding of how the models work through this project and have developed clear hypotheses for why the predictor variables I selected for my model may not have been the best choice for answering my research question. I believe that developing this sort of thinking will help me design stronger models in my future research.

Word Count: 2803

Github: https://github.com/anandigupta/Data-Science-I-Final-Project