**DS1 Final Report**

**Anandi Gupta**

**Word Count: 3027**

**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.

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.

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. 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. 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. 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 case rates between the incarcerated population and the general population in a county; and 2) the difference in Covid-19 death rates between the incarcerated population and the general population in a county.

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 and county-level population estimates by race from the U.S. Census Bureau that I used to calculate the percent of the population in the county that is White. 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. [[ add in population density and capacity]]. I merged these datasets using County FIPS codes (that I converted to be numeric to ensure successful merges).