Identifying Variation in Jail Stays Across Counties in California

University of California, Irvine

Antonio Rodriguez | Linh Trinh | Brenda Van Riper

Abstract

This comprehensive report presents a detailed analysis of daily rosters derived from the Jail Data Initiative (JDI), offering valuable insights into the complex dynamics of incarceration. The study categorizes data into four core domains to provide a holistic understanding of the incarceration landscape. Our analysis begins by examining both the aggregate and average incarceration durations at the county level, shedding light on the temporal aspects of confinement. Additionally, we investigate the average rates of re-incarceration, a crucial metric that underscores the recurring nature of involvement with the criminal justice system. Furthermore, the study delves into the intervals between successive jail terms, revealing patterns and trends that influence the frequency of incarcerations. Cross-county incarcerations are scrutinized, unveiling the interconnected nature of incarceration events across different jurisdictions. A significant focus of this analysis is the exploration of repeated incarcerations for individuals over time. We uncover patterns and driving factors behind the recurrent engagement of individuals with the criminal justice system, providing critical insights for policy considerations and decision-making processes. Incorporating extensive census data, our findings offer a comprehensive overview of how demographic, socio-economic, and educational factors intersect with incarceration trends, durations, re-entry dynamics, and patterns. This enriched understanding of the multifaceted nature of the criminal justice system serves as a valuable resource for policymakers, researchers, and stakeholders, facilitating informed decision-making and the formulation of effective strategies to address the challenges associated with incarceration.

1 Introduction

In an era where data-driven decision-making is pivotal, this paper explores the complex landscape of incarceration, drawing from a rich repository of Jail Data Initiative (JDI)(3) daily rosters. We embark on a multifaceted analytical journey, organizing voluminous data into four distinct categories to unravel the intricate dynamics of jail stays and factors influencing recidivism. Employing advanced statistical methods and machine learning, our study delves into the temporal and personal-level nuances of incarceration patterns across counties. Integrating Census data adds a vital dimension, enriching our understanding of the socio-economic underpinnings and demographic factors at play. This comprehensive approach transcends traditional analyses, offering novel insights into the criminal justice system, aiding in the development of informed policies and strategic interventions aimed at addressing the complexities of incarceration and its societal impact (5).

1.1 Analysis Goals and Objectives

The significance of this project lies in its endeavor to unravel the complexities embedded within the JDI daily rosters, a repository of data crucial for understanding the dynamics of incarceration. As the project unfolds, it addresses several key objectives:

1. **Data Organization and Linkage:** The first step involves meticulously organizing JDI daily rosters into four distinct analysis categories. This organizational effort is pivotal in ensuring the correct linkage across datasets, laying the groundwork for subsequent analyses.

2. **County-Level Incarceration Analysis:** By analyzing total and average incarceration periods for each county, we aim to identify influencing factors using the person-stay file. This county-level perspective offers insights into regional variations and disparities in the justice system.

3. **Person-Level Insights:** Leveraging the person-level file, the project explores metrics such as the average number of incarcerations per person and the time spent outside jail between incarcerations. This facet provides a more nuanced understanding of individual experiences within the criminal justice system.

4. **Cross-County Patterns and Repeated Incarcerations:** The study delves into the patterns of individuals being incarcerated across multiple counties, uncovering influencing factors behind such occurrences. Additionally, the project quantifies instances of repeated incarcerations involving the same individuals over different periods, shedding light on long-term trends and potential systemic issues.

Overall, this project is pivotal in enhancing our comprehension of incarceration dynamics, offering a holistic view that goes beyond individual counties. The findings are anticipated to inform policy considerations, contribute to evidence-based decision-making, and foster a more equitable and informed criminal justice system.

2 Data Description

2.1 Jail Data Initiative

The data used for this analysis was provided by the Jail Data Initiative team. The dataset is comprised of information gathered from 35 different counties, each collecting data uniquely, particularly focusing on distinct person-related features that varied in naming conventions. It consisted of over 600,000 inmate records, spanning the period between 2020 and 2023. However, across multiple counties, there were discrepancies in data availability, with some commencing data recordings at later periods or experiencing gaps in updates. For instance, Los Angeles County provided data solely for the years 2021 and 2022. As stipulated by our sponsors, a list of 26 required features was provided, directing us to extract these specific attributes from the dataset for both their and our future use.

2.1.1 Data Pre-processing

The initiation of this analytical journey transpired with the introduction of a diverse 1GB+ JSON dataset sourced from 35 counties. Leveraging the capabilities of MongoDB, this dataset underwent a comprehensive transformation, resulting in the creation of 35 distinct CSV files. This transformational process serves as the bedrock for subsequent analytical endeavors, laying a robust foundation for in-depth investigation.

2.1.2 MongoDB Transformation

To mitigate the challenges presented by the original dataset, MongoDB was employed to perform a strategic transformation. The following steps outline the transformation process:

- **County-Level CSV Extraction:** MongoDB's capabilities facilitated the extraction of county-specific data, resulting in the creation of 35 individual CSV files—one for each county. This segmentation set the stage for focused preprocessing on a county-by-county basis.
- **Facilitating Parallel Processing:** The decision to split the dataset into individual CSV files enabled parallel processing. This approach optimized computational efficiency, making it feasible to address county-specific inconsistencies simultaneously.

2.1.3 Preprocessing Steps

The initial step involves selecting and renaming columns to ensure uniformity across counties. The list of essential columns provides a standardized framework for subsequent analysis. Here is our data cleaning - general approach:

- *Housing Information:* Housing information, spread across multiple columns ('housing1', 'housing1a', 'housing2', 'housing3'), is processed to provide a consistent representation. This involves mapping and filling missing values where applicable.
- *Name Information:* Full names ('nameFull') are retained, and individual components ('nameLast', 'nameFirst', 'nameMiddle') are standardized to ensure consistency.
- *Date Handling:* Dates ('bookingDate', 'dob', 'date') undergo conversion to a uniform 'YYYY-MM-DD' format, enabling consistent temporal analysis.
- *Gender, Race, Eye Color, Hair Color:* Missing values in categorical attributes like 'sex,' 'race,' 'eyeColor,' and 'hairColor' are addressed by filling them with appropriate placeholders ('Unknown'). Existing values are mapped using predefined dictionaries to standardize representations.

- *Height and Weight:* Entries in 'height' are converted to inches for uniformity, and 'weight' is standardized by replacing 'Unknown' entries and converting to numeric values.
- *Bond and Bail:* The 'bond' and 'bail' columns undergo cleaning by removing symbols and converting to numeric values for consistency in financial representations.
- *Charge Information:* The 'topcharge' column is treated to handle missing values, filling them with 'Unknown' and mapping existing values for a consistent representation.
- *Age Calculation:* Age is computed based on the provided date of birth ('dob') and booking date ('bookingDate').
- *Incarceration Duration:* The duration of incarceration is computed by subtracting relevant dates, providing insights into the length of individuals' interactions with the legal system.
- *Time Spend Outside of Jail:* The dataset is sorted based on 'nameFull' and 'firstappearance' columns. Subsequently, the 'time_gap' column is computed, representing the temporal duration individuals spend outside of incarceration between successive legal detainments.
- *Incarceration Count:* The 'incarceration_counts' are computed by systematically tracking occurrences of each individual in the dataset using the 'nameFull' column. This quantifies the number of times each person experiences incarceration, offering a comprehensive representation of their involvement with the criminal justice system.
- *Re-Incarcerations:* The 'reincarceration' column is determined with a binary value (0 or 1), indicating whether an individual has more than one incarceration instance. If the incarceration count surpasses 1, the column value is set to 1; otherwise, it is assigned 0, distinguishing between multiple (1) and single or no instances (0).
- *Re-Incarceration in Different Counties:* To analyze re-incarceration patterns, records with 'reincarceration' equal to 1 are filtered. The dataset is grouped by 'nameFull' and 'county,' counting occurrences for each unique pairing. A new column indicates if an individual experienced incarceration in different counties, seamlessly integrating this information using the 'nameFull' column.

2.1.4 Algorithm for Repeat Incarceration with Same Individual

In our investigation focusing on shared accommodations within a subset of 35 counties, we specifically directed our attention to six counties that provide detailed information about cell blocks. This analysis was aimed at quantifying instances of repeated incarcerations involving the same individuals over different periods. This strategic selection aims to enhance analysis effectiveness, concentrating on locales with available cell block data to uncover meaningful patterns among inmates sharing a cell. The algorithm for identifying overlapping stays is executed independently for each of these six counties, and the results are later merged into a consolidated dataset.

- *Grouping Data:* Individuals sharing the same housing facility or cell are grouped based on the 'housing3' column.
- *Identify Overlapping Stays:* The algorithm examines combinations of inmates within these groups to identify potential overlaps in their stays, focusing on temporal aspects. This scrutiny involves utilizing the maximum start date value and minimum end date value for each inmate, ensuring precise determination of overlapping stays based on the earliest commencement and latest conclusion of their housing tenure.
- *Calculate Overlap Duration:* Considering a minimum overlap duration parameter ('min_overlap_duration') of 2 days for precision.
- *Create DataFrame:* A set of unique inmate groups meeting specific criteria forms the basis for creating a DataFrame for each county, including columns for 'housing3,' inmate names, and the durations of overlapping stays.

2.1.5 Census Data

In our study, we strategically enriched the JDI dataset by integrating it with Census data(1), focusing on three critical factors: population, poverty, and unemployment, due to their significant influence on criminal justice outcomes. This integration provided a richer analysis on a county level. Population metrics such as demographic composition, household dynamics, and political trends offered insights into community characteristics correlating with crime patterns. Meanwhile, the exploration of poverty, encompassing both overall and race-specific rates, along with housing costs, shed light on the socio-economic conditions tied to criminal involvement. Additionally, investigating unemployment rates across various demographics helped us understand its relationship with criminal activity. By merging these complex societal and economic factors, our analysis aimed to capture the nuanced interplay influencing criminal justice system trajectories.

3 Statistical Methods

3.1 Data Visualization and Analysis

3.1.1 JDI Data

Our dataset, predominantly populated by data from Los Angeles, Orange, and Riverside counties (Figure 3), reflects their prominence in California. However, notable limitations exist, such as Los Angeles County contributing data only for 2021-2022 (Figure 4), a limitation shared by Orange and Riverside counties. This limitation creates a significant decrease in data volume after 2022, requiring careful navigation for accurate interpretation of incarceration trends. The absence of data for entire years poses challenges in effectively tracking longitudinal reincarceration rates and population dynamics.

Transitioning from our analysis of the population composition within the JDI dataset, we pivot our focus to the gender dynamics among the incarcerated over the years (Figure 5). Female incarceration rates remain consistent over the observed time period, yet a noticeable shift occurs in male population trends, decreasing while the 'Unknown' category sees a significant rise from 2020 to 2023. Although technology allows gender inference from names, our current focus doesn't include its implementation.

Examining age distribution (Figure 6), the dataset presents a concentration of individuals in the mid-thirties, aligning with national incarceration trends. A closer look reveals a slight decrease in the average age from 2020 to 2022, followed by an increase in 2023. This fluctuation warrants deeper investigation, potentially linked to the COVID-19 pandemic, prompting questions about its wider socio-economic impact and policy responses affecting incarceration rates and demographics.

As we shift from examining age trends in our dataset, we delve into the racial dynamics of incarceration (Figure 7). The data portrays a nuanced racial landscape: a rise and subsequent fall in the White population in 2023, a steady increase among Black individuals over the years, and a noticeable peak in the Hispanic population in 2021, followed by a decline. These patterns, while revealing, must be approached with caution due to the incomplete nature of our dataset.

In our analysis of incarceration data, we examined three crucial metrics: average counts of incarceration (Figure 8), time gaps between incarcerations (Figure 9), and durations of incarceration across various charges (Figure 10). We found that 'Property' and 'Drug' charges had higher incarceration frequencies, while 'Unknown' charges indicated recurrent involvement with the justice system. 'DUI Offenses' had the lowest average count. Time gaps varied among charges, with 'Property' showing longer intervals between incarcerations. 'Violent' offenses had the longest average incarceration duration, while 'DUI Offenses' suggested shorter periods. This analysis offers insights into the landscape of incarceration critical for policymakers and legal professionals navigating the criminal justice system's complexities. (Figures 8, 9, 10)

In comparing different offense types, notable distinctions emerge in incarceration patterns. 'Violent' offenses, while resulting in longer incarceration durations, do not consistently yield the highest average incarcerated counts, implying potentially fewer re-offenses but more severe legal consequences. Conversely, 'Property' and 'Drug' charges, with higher counts, exhibit varied time gaps that might indicate different legal and rehabilitative approaches. 'DUI Offenses', distinct for their shortest duration and lowest counts, suggest a different legal treatment compared to more severe crimes. The intricate dynamics within the 'Unknown' category, characterized by high counts, lengthy durations, and short time gaps, imply complex interactions within the justice system due to uncertainties in case processing. This comparative analysis illuminates how offense nature significantly shapes incarceration patterns, offering essential insights for tailored policy interventions, legal reforms, and targeted support programs aimed at addressing specific challenges within each offense category.

3.1.2 Census Data

In our analysis of Census data(1) from California's counties, comparing those covered and not covered by the Jail Data Initiative, our focus was on key demographic, socio-economic, and educational indicators. We observed significant variations in racial populations in both JDI and non-JDI counties, revealing lower proportions of Hispanic, Black, and Asian communities, and higher White populations compared to the state average (Figure 11). These variations prompt important inquiries into racial and ethnic influences on criminal justice involvement. Additionally, our poverty analysis (Figure 12) highlighted marked disparities. Non-JDI areas exhibited higher Black poverty rates, diverging from JDI counties and state averages, indicating unique socio-economic challenges. We also noted distinct differences in poverty rates for Black and Asian communities between JDI and non-JDI counties, suggesting racial and ethnic economic disparities. Unemployment trends (Figure 13) revealed distinctive patterns among JDI and non-JDI counties compared to California overall. Non-JDI areas exhibited higher Black unemployment rates exceeding the state average, while JDI counties generally had lower rates, except in specific demographics such as Asians. In terms of gender, non-JDI regions showed higher male unemployment, contrary to the state pattern. Across educational levels, non-JDI counties faced higher unemployment, except for those with higher education degrees.

Further examination of the census data against JDI and non-JDI counties, as well as California as a whole, reveals critical socio-economic insights. Differences in household size, homeownership rates, health coverage, and median housing costs across these regions highlight the diverse socio-economic pressures they face. These findings suggest that while JDI counties might have better healthcare access, they also grapple with higher living costs and lower home ownership rates, informing the need for tailored policy interventions to address these community-specific challenges.

3.1.3 Correlation Analysis - JDI and Census

Our analysis initially focuses on identifying key predictors of jail incarceration spells and related features. The correlation heatmaps, placed side by side, reveal intriguing connections within our dataset (Figure 14). Notably, there's a significant correlation between bond and bail, suggesting a link between legal financial obligations and the duration of incarceration. Other correlations, such as those between race, hair color, and eye color, indicate nuanced intersections of these variables. However, the overall low correlation levels across many variables highlight the complex dynamics of incarceration.

Our exploratory journey involves formulating hypotheses to guide our investigation into jail incarceration patterns, such as potential relationships between demographic factors and the length of jail stays, as well as anticipated regional disparities. The incorporation of census data into our analysis, reflected in the second heatmap (Figure 15), markedly enhances our understanding. Seven out of the top ten highly correlated features are from the census data, emphasizing the significant role of factors like Hispanic Poverty Percentage and Female Unemployment in predicting jail stays. The strong intercorrelations among these features indicate a nuanced interplay, which might influence their use in model building due to potential redundancy.

3.2 Principle Component Analysis

In our pursuit of a comprehensive understanding of the dataset dynamics, Principal Component Analysis (PCA) stands out as a powerful tool, revealing hidden patterns, reducing dimensionality, and extracting insights. This section details the application of PCA to uncover the dataset's structure.

- **Enhanced Representation:** Categorical variables are preprocessed through One-Hot Encoding, crucial for effective PCA representation. This step, while increasing dimensionality, allows nuanced variance exploration in categorical attributes.

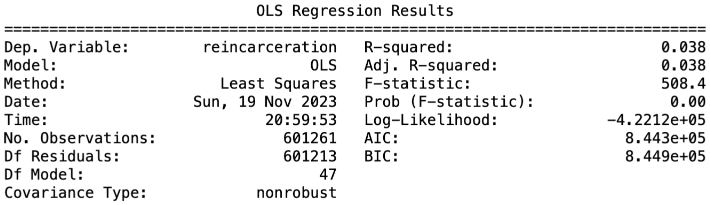
- **Multifaceted Dimensions:** PCA is applied to four distinct features—incarcerated counts, time gap, reincarceration, and incarcerated days. This approach aims to unravel the dataset's multifaceted aspects and gain deeper insights into underlying patterns.

- **Interpreting Principal Components:** Focusing on the two prominent principal components,**PC1 captures the variance associated with socio-economic status**, employment-related aspects, and various demographic features, while **PC2 delves into demographic attributes**, encompassing age, sex, race, physical characteristics, and other relevant variables.

Through PCA, we navigate data complexity, uncover core factors behind trends, and draw insights from this rich and multifaceted dataset. PCA proves invaluable in exploring, interpreting, and understanding the dataset's intricate nature.

3.3 Modeling Process and Methodology

In this report, we delve into the findings of a simple regression model designed to understand the factors influencing the duration of incarceration, time spend outside of jails, incarceration counts and re-incarcerations. The analysis begins with a basic linear regression, progresses to cluster-robust standard errors, and concludes with the consideration of a multi-level model.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:        reincarceration   R-squared:                       0.038
Model:                            OLS   Adj. R-squared:                  0.038
Method:                 Least Squares   F-statistic:                     508.4
Date:                Sun, 19 Nov 2023   Prob (F-statistic):               0.00
Time:                        20:59:53   Log-Likelihood:            -4.2212e+05
No. Observations:              601261   AIC:                         8.443e+05
Df Residuals:                  601213   BIC:                         8.449e+05
Df Model:                          47
Covariance Type:            nonrobust
```

Figure 1: Simple linear regression model

The R-squared value of 0.038 indicates that approximately 3.8% of the variability in the number of incarcerated days can be explained by the independent variables in the regression model. While this percentage is modest, the log-likelihood suggests a good fit to the data. AIC and BIC, both indicating model fit and complexity, are observed to be the same. Lower values are indicative of a better-fitting model. Among the predictors, those with p-values less than our significance level of 0.05 are considered statistically significant.

3.3.1 Cluster-Robust Standard Errors

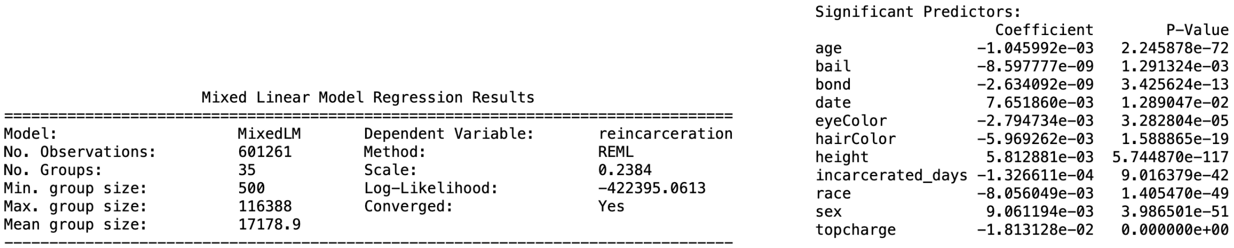
To address potential correlations or heteroscedasticity within clustered data, the 'cluster' option in the 'fit' method for the OLS model is utilized. This method adjusts standard errors at the county level. (Figure 16)

The F-statistic in the clustering method is reported as 'nan', which can occur with cluster-robust standard errors. However, the R-squared value remains the same between the non-clustered and clustering methods (Figure 16).

3.3.2 Multi-Level Model

Considering a multi-level model becomes pertinent when dealing with nested or clustered data, such as individuals within counties. Fixed effects, including demographic and socioeconomic predictors, are included in the model. 'County' is considered a random effect to account for county-level variability.

- *Fixed Effects:* These are variables for which we want to estimate the average effect. In this case, features like 'age', 'sex', 'race', 'eyeColor', 'hairColor', etc., could be considered fixed effects.
- *Random Effects:* These are variables for which we want to account for variation at a higher level, such as individual or county-level variability. In this case, 'county' could be a good candidate for a random effect.



```
                  Mixed Linear Model Regression Results
=============================================================================
Model:                    MixedLM      Dependent Variable:      reincarceration
No. Observations:         601261       Method:                  REML
No. Groups:               35           Scale:                   0.2384
Min. group size:          500          Log-Likelihood:          -422395.0613
Max. group size:          116388       Converged:               Yes
Mean group size:          17178.9
-----------------------------------------------------------------------------
```

```
Significant Predictors:
                    Coefficient        P-Value
age               -1.045992e-03   2.245878e-72
bail              -8.597777e-09   1.291324e-03
bond              -2.634092e-09   3.425624e-13
date               7.651860e-03   1.289047e-02
eyeColor          -2.794734e-03   3.282804e-05
hairColor         -5.969262e-03   1.588865e-19
height             5.812881e-03  5.744870e-117
incarcerated_days -1.326611e-04   9.016379e-42
race              -8.056049e-03   1.405470e-49
sex                9.061194e-03   3.986501e-51
topcharge         -1.813128e-02   0.000000e+00
```

Figure 2: Output for the multi-level model with random effect 'county'

To assess model fit, we compare log-likelihood values, particularly between the current model and a null model (with only an intercept). The Likelihood Ratio Test Statistic, measuring the difference in fit between the full model (with predictors and random effects) and the null model, yielded a significant result of 4732.82 (Figure 2). Larger values indicate a substantial difference in fit. The degrees of freedom (54 in this case) represent additional parameters in the full model compared to the null model. The very low p-value indicates strong evidence against the null hypothesis, signifying that the full model fits the data significantly better than the null model.

The statistical analysis journey, from simple regression to multi-level modeling, provides valuable insights into the factors influencing incarceration duration. The inclusion of demographic and socioeconomic predictors, along with consideration of county-level variability, enhances the model's ability to explain the observed variation in the data. This comprehensive approach lays the groundwork for nuanced and informed interpretations, guiding future analyses and policy considerations related to incarceration dynamics.

3.4 Time Series Analysis

3.4.1 Model Selection

SARIMAX(4), a robust forecasting model, effectively predicts reincarceration counts in daily roster data from 35 California counties (Figure 17). Its strength lies in capturing intricate time series patterns and adapting to diverse dataset characteristics. The model incorporates exogenous variables, accommodating external factors like policy changes and events such as the COVID-19 pandemic. This feature enhances adaptability to unforeseen influences, demonstrating SARIMAX's versatility in addressing real-world complexities.

Future improvements can refine the model's sensitivity, particularly in uncertain scenarios like the ongoing pandemic, and exploring additional features, such as census data, holds potential for enhancing predictive accuracy and robustness (Figure 17).

3.4.2 Evaluation of Time Series Model

- **Insights from 2023 SARIMAX Predictions:** The Root Mean Squared Error (RMSE) scores for SARIMAX predictions in 2023 provide valuable accuracy insights. The Average RMSE for all counties is 246.20, reflecting the average difference between actual and predicted values.
- **Exclusion of Counties with Zero Counts:** Certain counties, like San Luis Obispo, lacked records for 2023, resulting in all counts being zero. To address potential data issues and avoid artificially inflated RMSE scores, counties with more than half the year having zero counts were excluded. This strategic decision focuses the evaluation on regions with reliable and complete data, preventing disproportionate influence from areas with missing or incomplete records.
- **Improved Accuracy with Exclusion:** The Average RMSE for 2023 (excluding specified counties) significantly improves to 117.73. This highlights the impact of excluding regions with peculiar characteristics or missing records, resulting in a more accurate representation of the model's predictive performance on reliable data.

This strategic exclusion ensures a fair evaluation, considering the challenges posed by incomplete or unavailable data in specific regions. It also highlights the importance of adapting modeling approaches based on the data characteristics of each region to derive meaningful insights into the reincarceration forecasting process.

4 Machine Learning Models

The incorporation of machine learning methodologies into our analytical framework was driven by the overarching goal of extracting nuanced insights from complex incarceration datasets. Traditional statistical approaches are often constrained by assumptions and may struggle to capture intricate non-linear relationships present in the data. For this type of analysis in our model:

1. **Model Categories:** Employing four primary machine learning categories, we construct prediction models to anticipate incarceration days, duration out of jail, incarceration numbers, and the likelihood of re-incarceration.
2. **Metric Selection:** To assess prediction accuracy, we utilize the Root Mean Squared Error (RMSE) metric for the first three values, which involve regression tasks. Conversely, for the binary nature of re-incarceration prediction, we employ accuracy as the metric.

4.1 Model Selection

In our machine learning approach for this dataset, we opted for three models—decision tree, neural network, and XGBoost— with the intention to identify the most effective one for subsequent analysis. We chose these models due to their effectiveness in handling diverse data types within our dataset, addressing tasks from predicting incarcerated days to assessing the likelihood of re-incarceration. Their simplicity and interpretability make them optimal for transparently discerning feature impacts on predictions. Proficiency in managing various data types, modeling non-linear relationships, and notable strengths such as seamless handling of missing data and insensitivity to numerical feature scales, position these models well for datasets with intricate patterns. However, rigorous performance validation against research objectives and exploration of alternative models are crucial steps for ensuring optimal fit in our research endeavors.

4.1.1 Decision Tree Exploration

The machine learning phase of our analysis began with the deployment of a simple yet insightful decision tree model. This initial model aimed to predict incarcerated spells or stays, yielding an RMSE (Root Mean Squared Error) of 56.4 for predicting incarcerated days.

4.1.2 Neural Network Exploration

To probe the capabilities of neural networks in our predictive modeling, we introduced a neural network model. The initial iteration focused on incarcerated spells or stays, yielding a notably high RMSE of 4173.7. Acknowledging the necessity for optimization, we undertook meticulous tuning, leading to a refined yet significant improvement, resulting in an RMSE of 4453.5.

4.1.3 XGBoost and Feature Importance

In pursuit of heightened predictive performance, we transitioned to the formidable XGBoost (2) algorithm for forecasting incarcerated spells or stays. The XGBoost model exhibited substantial improvements, achieving an RMSE of 52.69. Further fine-tuning enhanced the model, resulting in a slightly improved RMSE of 53.686. The strategic choice to implement XGBoost for all other metrics, including duration out of jails, incarcerated counts, and re-incarcerations, stemmed from its effectiveness in handling complex, non-linear relationships within the dataset.

4.1.4 XGBoost: Elevating Incarceration Analysis

XGBoost was chosen as our primary machine learning model due to its proven effectiveness in handling non-linear relationships and complex interactions within our incarceration dataset. Its iterative boosting algorithms and ensemble learning approach enable the model to sequentially correct errors and adapt through tuning, enhancing predictive accuracy. This capability, combined with XGBoost's interpretability, made it particularly suited for a nuanced analysis of incarceration patterns, allowing for a deeper understanding of the factors influencing incarceration durations.

4.2 Improvement on Machine Learning Model with Census Data

The integration of external census data into our analysis has yielded substantial enhancements in the predictive performance of our XGBoost model, particularly in the context of re-incarceration prediction. The model's accuracy has demonstrated a noteworthy improvement, escalating from an initial 64.6% to a significantly elevated 71%. This uptick in accuracy is indicative of the model's heightened ability to correctly classify instances of re-incarceration.

Furthermore, the incorporation of additional socio-economic features has notably bolstered the model's discriminatory power. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC), a crucial metric for evaluating classification model performance, has surged from an initial value of 0.7 to an impressive 0.8. This substantial increase in AUC-ROC signifies a more robust ability to distinguish between individuals with varying re-incarceration outcomes.

The observed improvements underscore the critical role played by the enriched dataset derived from the integration of census data. The model now benefits from a more comprehensive understanding of the contextual socio-economic factors, allowing it to capture and leverage intricate patterns that were previously beyond its reach. The nuanced relationships uncovered through this richer dataset have translated into heightened accuracy and discriminative power, marking a significant stride in our quest to comprehensively model and predict re-incarceration dynamics. This progress reaffirms the value of augmenting our original dataset with external sources, reinforcing the depth and efficacy of our analytical insights.

5 Results

In our exploration of incarceration data, we have calculated several key averages to provide valuable insights into the experiences of individuals within our dataset.

- **Average Number of Incarceration Spells:** On average, individuals within our dataset experience approximately 4.94 incarceration spells.
- **Average Duration out of Jails:** The average duration spent out of jails between incarceration spells is approximately 94.82 days.
- **Average Duration of Incarceration:** Across all counties, the average duration of incarceration is approximately 42.73 days.
- **Average Rate of Re-Incarceration:** The average rate of re-incarceration stands at approximately 1.17, the average rate per month is 0.0313 when taking into account the length of exposure of each county. This rate signifies the likelihood of individuals returning to incarceration after their previous spell, highlighting the recurring nature of involvement with the criminal justice system.

These averages serve as fundamental indicators of the patterns and trends within our dataset, offering a comprehensive overview of the experiences of individuals as they navigate the complex landscape of incarceration and re-entry into society.

5.1 Feature Importance in XGBoost Modeling of Incarceration Metrics

The application of XGBoost to diverse incarceration-related metrics, spanning incarcerated counts, duration out of jails, and re-incarceration, has yielded profound insights into the hierarchy of feature importance within the predictive modeling framework.

Primarily, **the model identifies age, incarcerated days, county, bond, topcharge, weight, height, and race as pivotal contributors to predictive accuracy.** These features emerge as highly influential, signifying their substantive impact on the model's predictive outcomes concerning incarceration metrics. In contrast, attributes such as eye color, hair color, and sex are deemed less influential within the model.

Upon integration of census data into the XGBoost model, the analysis upholds the significance of previously identified key features. Noteworthy is the introduction of socioeconomic features such as population, poverty percentage, and unemployment percentage, which assume heightened importance. This underscores the augmented relevance of broader socio-economic factors in shaping predictive outcomes related to incarceration metrics. Conversely, eye color, hair color, and sex exhibit a relative decrease in importance within the model when considering the broader context provided by census data. This implies a moderated influence of these attributes on predictive power, emphasizing their less pronounced role in forecasting incarceration outcomes within the expanded analytical framework.

5.2 Repeat-Incarceration Analysis with Same Individual

The examination of the dataset focused on the co-incarceration patterns among pairs of individuals, particularly those who have been incarcerated together multiple times. The following summarizes the findings for each county:

Table 1: Co-Incarceration Counts by County

County	**Pairs Incarcerated Twice**	**Pairs Incarcerated Three Times**
Fresno	2	0
Lake	1015	18
Monterey	4	0
Placer	4266	75
Riverside	250	0
Shasta	153	0

These results shed light on the frequency of co-incarceration occurrences within different counties. Notably, Lake County, for instance, shows a substantial number of pairs who have been incarcerated together twice, and a smaller number who have experienced three instances of co-incarceration. These patterns provide valuable insights into the dynamics of co-incarceration, offering potential avenues for further investigation into the social and institutional factors contributing to these occurrences.

6 Conclusion

6.1 Discussion: Unraveling the Dynamics of Feature Importance in Incarceration Modeling

The exploration of feature importance within the XGBoost predictive modeling framework for incarceration metrics offers valuable insights into the intricate dynamics governing the predictive accuracy of the model. The identified key features highlight the profound impact of individual-level attributes on forecasting incarceration-related outcomes.

The significance of age and incarcerated days suggests a direct correlation between an individual's temporal engagement with the legal system and the predictability of their incarceration patterns. County emerges as a critical factor, underscoring the localized nature of legal dynamics and the distinct influences that jurisdiction-specific factors may exert on incarceration metrics.

The importance assigned to demographic attributes such as weight, height, and race suggests nuanced associations between individual characteristics and their interactions within the criminal justice system. Specifically, the significance of height and weight may indicate how physical attributes contribute to various aspects of the legal process, including law enforcement interactions, court proceedings, or experiences during incarceration. These factors could be linked to issues of physical restraint, accommodation in correctional facilities, or potential biases within the criminal justice system related to body size. The correlation of height and weight with critical variables influencing incarceration metrics may further contribute to their prominence in predictive models, highlighting the need for further investigation and contextual analysis to understand the nuanced associations comprehensively. Conversely, the downgraded importance of attributes like eye color, hair color, and sex within the model emphasizes their limited role in predicting incarceration outcomes. It might be influenced by their limited variability, missing data points, or their perceived lesser impact on the specific outcomes of interest in the dataset.

The incorporation of census data enriches the predictive landscape, introducing population, poverty percentage, and unemployment percentage as influential features. These variables offer insights into regional structural and economic conditions, influencing law enforcement activities and crime rates. Higher population density may correlate with increased law enforcement engagement, while elevated poverty and unemployment percentages often contribute to higher crime rates. Integrating these socio-economic features enhances the model's ability to capture the complex interplay between demographic and economic factors, providing a nuanced understanding of the determinants influencing incarceration patterns.

The nuanced adjustments in feature importance when transitioning from individual-level attributes to broader socio-economic indicators underscore the model's adaptability and responsiveness to varying data landscapes. This adaptability reinforces the need for comprehensive, multi-dimensional approaches in understanding and predicting incarceration trends.

6.2 Limitations and Considerations

In our examination of incarceration data, while we've gained valuable insights, it's important to acknowledge certain limitations that may impact the generalizability and interpretation of our findings. Firstly, our analysis doesn't explicitly consider the potential influence of the COVID-19 pandemic on incarceration patterns, given the unprecedented disruptions it introduced. Secondly, the dataset may have instances of missing or incomplete information, which could introduce bias despite our use of imputation methods. Thirdly, our focus on 35 counties in California might not fully represent the entire state's incarceration landscape, cautioning against broad extrapolation. Moreover, integrating census and jail data, although enriching our analysis, poses challenges due to differences in data collection methodologies. Lastly, our examination of co-incarceration patterns is confined to specific pairs within the dataset, emphasizing the need for caution when generalizing these patterns to broader societal trends that may vary across regions and populations.

References

[1] U.S. Census Bureau. U.S. Census Bureau Data, 2023.

[2] XGBoost Developers. Xgboost documentation, 2023.

[3] Jail Data Initiative. Jail data initiative, 2023.

[4] Statsmodels Developers. Seasonal autoregressive integrated moving average with exogenous regressors (sarimax) - statsmodels documentation, 2023.

[5] Kristin Turney and Emma Conner. Jail incarceration: A common and consequential form of criminal justice contact. *Annual Review of Criminology*, pages 1, 29, 2019.

Appendix

A Jail Data Initiative Visualization and Analysis

A.1 JDI Population Graph

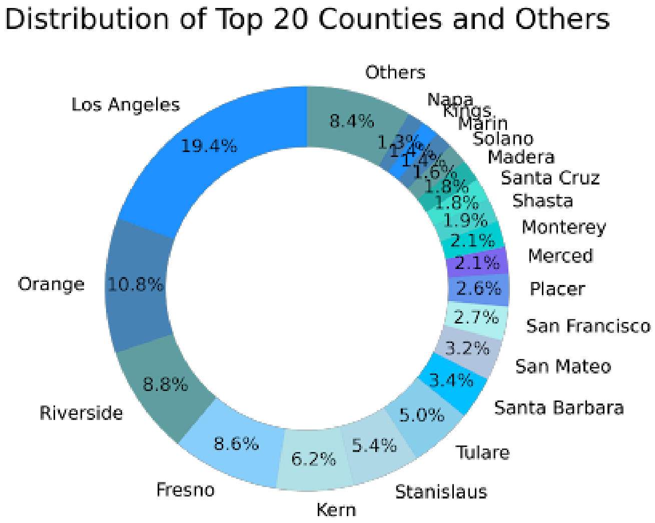
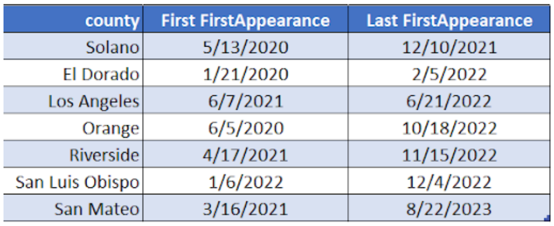


Figure 3: Distribution of all of JDI data

A.2 JDI Missing Data Table



county	First FirstAppearance	Last FirstAppearance
Solano	5/13/2020	12/10/2021
El Dorado	1/21/2020	2/5/2022
Los Angeles	6/7/2021	6/21/2022
Orange	6/5/2020	10/18/2022
Riverside	4/17/2021	11/15/2022
San Luis Obispo	1/6/2022	12/4/2022
San Mateo	3/16/2021	8/22/2023

Figure 4: Table of which counties have missing data

A.3 JDI Sex by Year

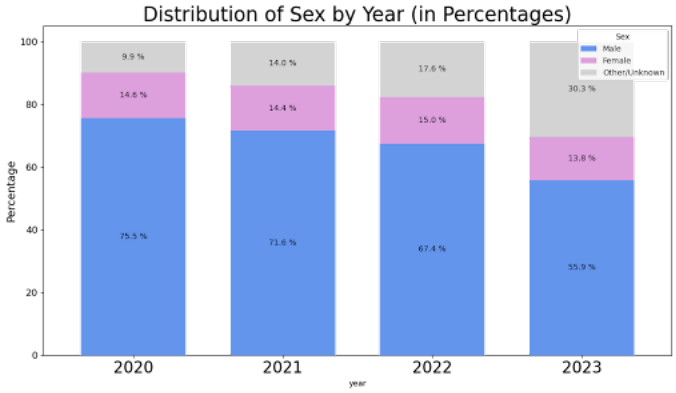


Figure 5: Sex stacked bar graph with male, female, and 'Unknown'

A.4 JDI Age by Year with average lines

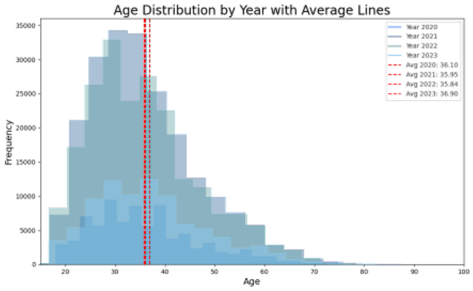


Figure 6: Age distribution with average lines and average year old of incarceration

A.5 JDI Race by Year

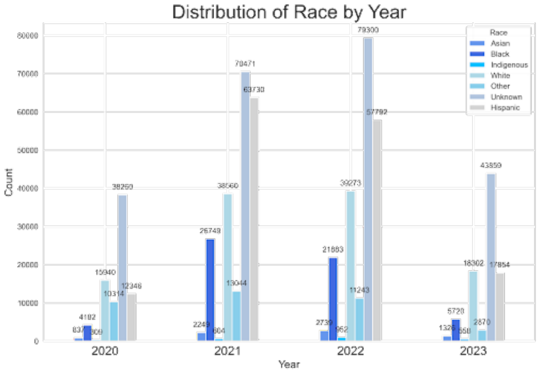


Figure 7: Bar graph of race by year with all 7 categories

A.6 Average Incarceration Count based on Top Charge categories

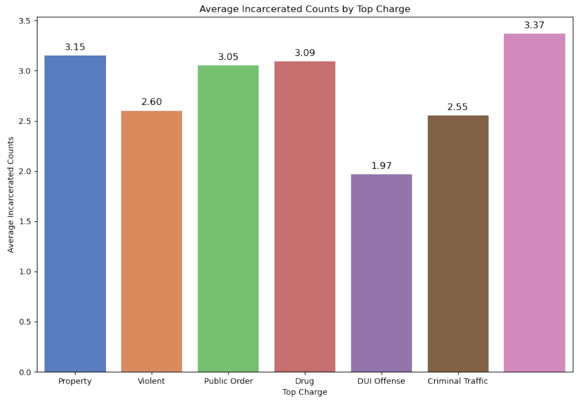


Figure 8: Bar graphs based on top charge for incarceration count

A.7 Average Duration out of Jail(time gap) based on Top Charge categories

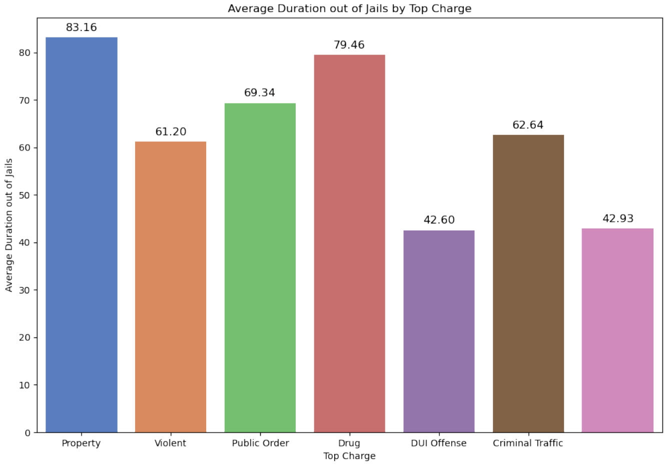


Figure 9: Bar graphs based on top charge for duration out of jail(time gap)

A.8 Average Incarceration Days based on Top Charge categories

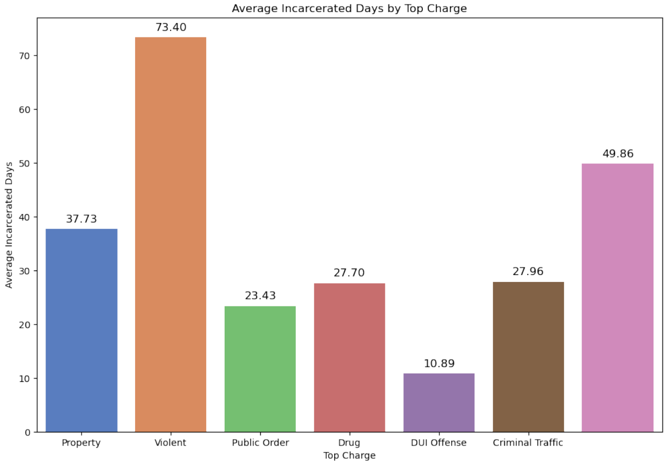


Figure 10: Bar graphs based on top charge for incarcerated days

B Census Data

B.1 Census Data for JDI, nonJDI, and CA averages for Race %

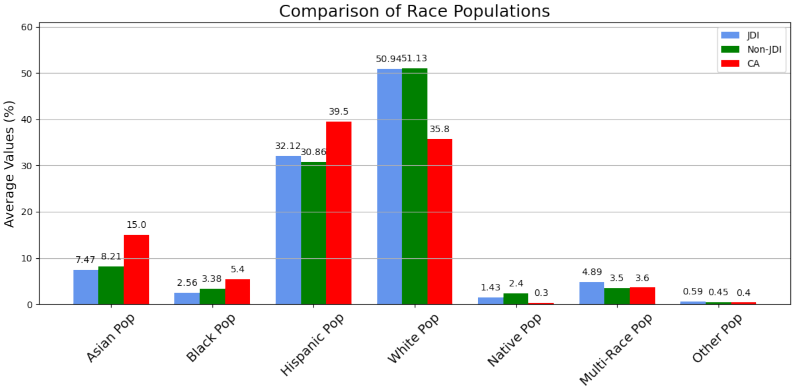


Figure 11: Race Census comparison for JDI, nonJDI and CA

B.2 Census Data for JDI, nonJDI, and CA averages for Poverty %

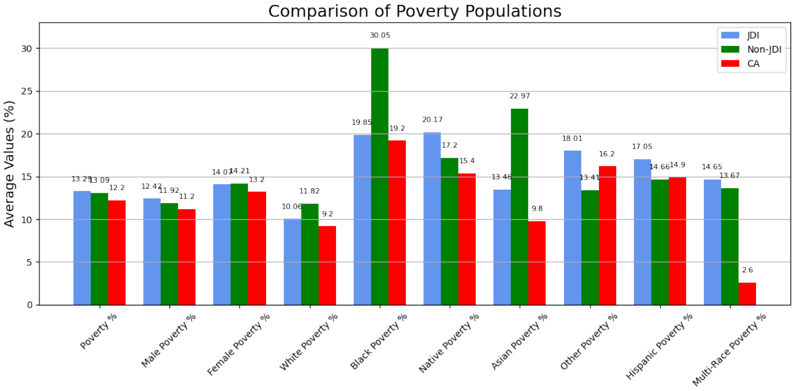


Figure 12: Poverty % Census comparison for JDI, nonJDI and CA

B.3 Census Data for JDI, nonJDI, and CA averages for Unemployment %

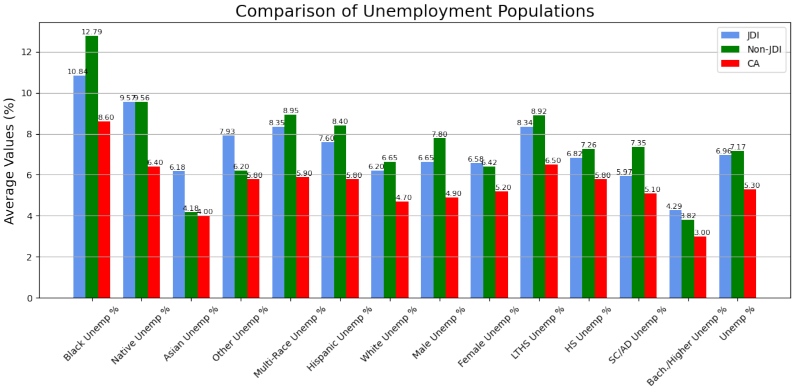


Figure 13: Unemployment % Census comparison for JDI, nonJDI and CA

B.4 Heatmap with JDI Data

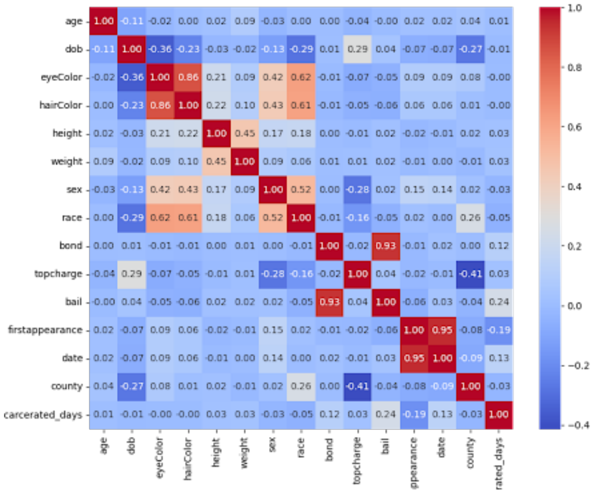


Figure 14: Correlation heatmap for jail incarceration stays with only JDI features

B.5 Heatmap with JDI Data

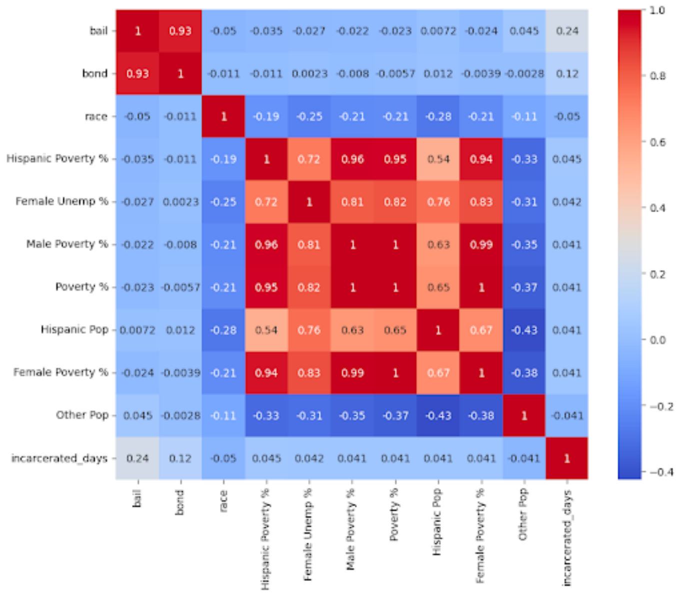
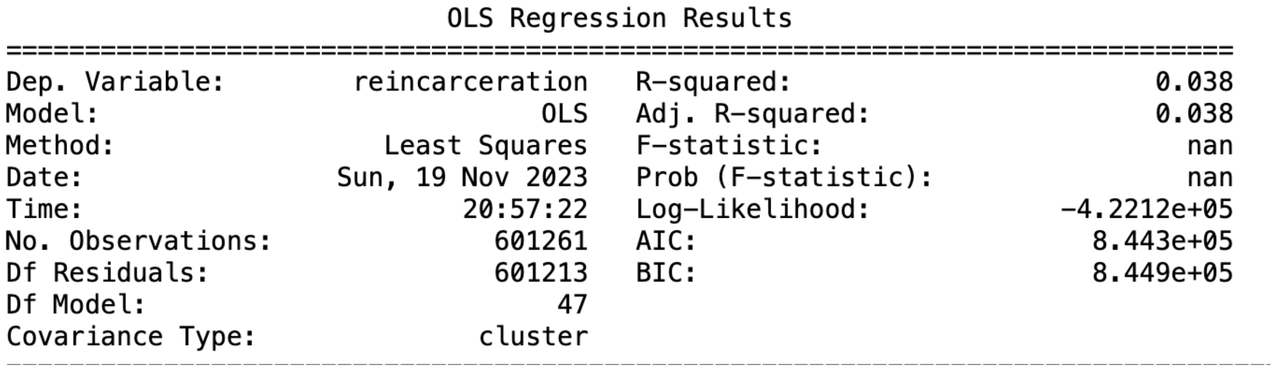


Figure 15: Correlation heatmap for jail incarceration stays after Census data was added

C Statistical Methods

C.1 Cluster Robust Statistical Model



```
                            OLS Regression Results
==============================================================================
Dep. Variable:        reincarceration   R-squared:                       0.038
Model:                            OLS   Adj. R-squared:                  0.038
Method:                 Least Squares   F-statistic:                       nan
Date:                Sun, 19 Nov 2023   Prob (F-statistic):                nan
Time:                        20:57:22   Log-Likelihood:            -4.2212e+05
No. Observations:              601261   AIC:                         8.443e+05
Df Residuals:                  601213   BIC:                         8.449e+05
Df Model:                          47
Covariance Type:              cluster
------------------------------------------------------------------------------
```

Figure 16: Cluster Stats Model

C.2 SARIMAX Time Series

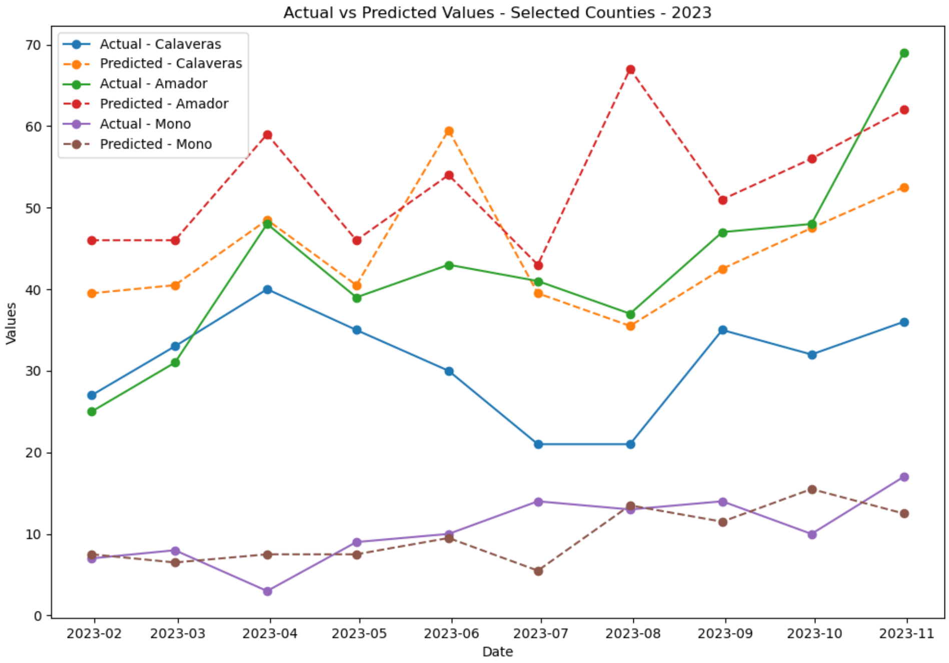


Figure 17: SARIMAX Time Series Graph