****

DSA 210: Introduction to Data Science Spring 2024 - 2025

**Unemployment & Crime Rates in the United States**

*A multi-decade exploration of socioeconomic stressors, criminal activity, and predictive analytics*

**Prepared by:** Abdallah Al Homsi – Student ID 34565  
**Email:** abdallah.homsi@sabanciuniv.edu  
**Submission Date:** 30 May 2025

### 

### 

**Table Of Contents:**

[**1. Motivation**](#_heka0xsh88rq) **3**

[**2. Datasets & Data Enrichment**](#_mu78dl3arb8d) **3**

[**3. Data Collection & Preparation**](#_88fev7e1f30o) **4**

[**4. Exploratory Data Analysis (EDA)**](#_q67w1rofoa8k) **5**

[4.1 Summary Statistics](#_uh292y84soka) 5

[4.2 Temporal Trends](#_6eadat1ednu6) 6

[4.3 Correlation Heatmap](#_mha95pccze6p) 6

[4.4 Visuals](#_sz09ot8wg3ur) 7

[**5. Hypothesis Testing**](#_44rga41x3bqi) **8**

[5.1 Hypotheses](#_qpdwjsa85wds) 8

[5.2 Methodology](#_bvylfi9xddch) 8

[5.3 Results](#_vtgifxh0forj) 9

[5.4 Interpretation](#_7i1us7ik2azh) 9

[5.5 Visual Confirmation](#_possu2clx4fs) 9

[**6. Feature Engineering & Enrichment**](#_xdsfo5yg3e1f) **10**

[**7. Machine Learning Modelling**](#_i3j7kyyc3mze) **10**

[7.1 Classification – Random Forest](#_8rw9xs7k19tj) 11

[7.2 Regression – Linear Model](#_9wl55xpm8ftd) 12

[**8. Key Findings**](#_v190hvyv2ek) **13**

[**9. Limitations & Future Work**](#_fi17kqumtw8y) **14**

[Limitations](#_ov6n6qlr37nt) 14

[Future Work](#_9snuvzhrzqmu) 14

[**10. Technology Stack**](#_ljtlolrbwt0w) **14**

[**11. Project Timeline**](#_t953melpwz9r) **15**

# 

# 

# **1. Motivation**

Studying the correlation between unemployment rates and crime rates in the United States is a question that has been ongoing for a very long time in both social sciences and public policy. This section specifies the objectives and my inner motive behind the study, marking why this problem is not only statistically interesting but socially important. Not only that , but it also shows how the project contributes to my growth as a data scientist and the eagerness to improve in such an area as data is key.

Crime and unemployment have long been subjects of national concern and public policy debate. Understanding whether economic hardship influences crime rates has implications for everything from policing strategy to social welfare planning.

**Why this topic?**

* **Social Impact:** Focusing on the relationship and marking a correlation if there is any would help out policy makers and other data scientists to solve and target the root of the issues or to continue making more testing with other socioeconomic indicators
* **Skill Development:** This project gave me hands-on experience with time-series analysis, statistical hypothesis testing, and predictive modelling using machine learning. Moreover, use of new technology and libraries in python such as (.e.g pandas, matplotlib).

# **2. Datasets & Data Enrichment**

Before conducting any analysis, it is important to gather a variety of datasets that emphasize both the economic and social dynamic aspects of the issue. This section describes the datasets used and justifies the enrichment strategy by explaining how additional variables—such as salaries and incarceration rates—can contextualize the relationship between joblessness and criminal behavior. As these datasets had a vital role in enrichment and were one of the socioeconomic indicators in the project as well there was some feature engineering to make the statistical and machine learning models process everything accurately, facing no problems.

To ensure a rich, multi-dimensional analysis, the following datasets were used:

| **Dataset** | **Time Span** | **Granularity** | **Key Variables** | **Source** |
| --- | --- | --- | --- | --- |
| US Unemployment Rates | 1948–2014 | Monthly | age, gender, overall rates | Kaggle |
| US Crime Statistics | 1960–2019 | Annual x State | violent & property crime | Kaggle |
| Prison Admissions & Releases | 2015–2020 | Monthly x State | admissions, releases, race | Kaggle |
| Salary by Demographics | 2015–2020 | Annual x State | mean salary | Kaggle |

**Rationale for Enrichment:**

* While unemployment and crime statistics provide a basic view of socioeconomic stress, they do not fully capture the structural conditions or underlying disparities that influence criminal behavior. By incorporating **salary data segmented by race and gender**, we gain insight into **economic inequality**, which may drive certain populations toward higher-risk behaviors due to limited opportunity. Similarly, **prison admissions and releases** offer a lens into the **justice system's throughput**, enabling us to explore how **carceral patterns**—such as recidivism or state-level incarceration policies—intersect with crime trends. This enriched context allows for more robust feature engineering in predictive models and fosters a deeper understanding of the **multifactorial nature** of crime beyond mere joblessness.

# **3. Data Collection & Preparation**

The quality of the data directly affects the results of any analysis. This section talks about the full plan of transforming raw datasets into clean, structured, and ready to be analyzed formats. We mark the types of data, aggregation methods, strategies used for merging, and sanity checks used to be aware of the data transformation. The process ensures that the derived yearly and monthly datasets accurately reflect the real-world trends in unemployment and crime.

**Initial steps:**

* Raw Data was parsed and made it into a dataframe using pandas, and all time stamps were converted to the same date type.
* Unemployment data averaged yearly; state-level crime data aggregated to national level.

**Preprocessing:**

* All null values (NaN) were dropped and others imputed with logical methods.
* Joined on "Year" for national datasets and on "Year+State" for enriched monthly data.

**Output:**

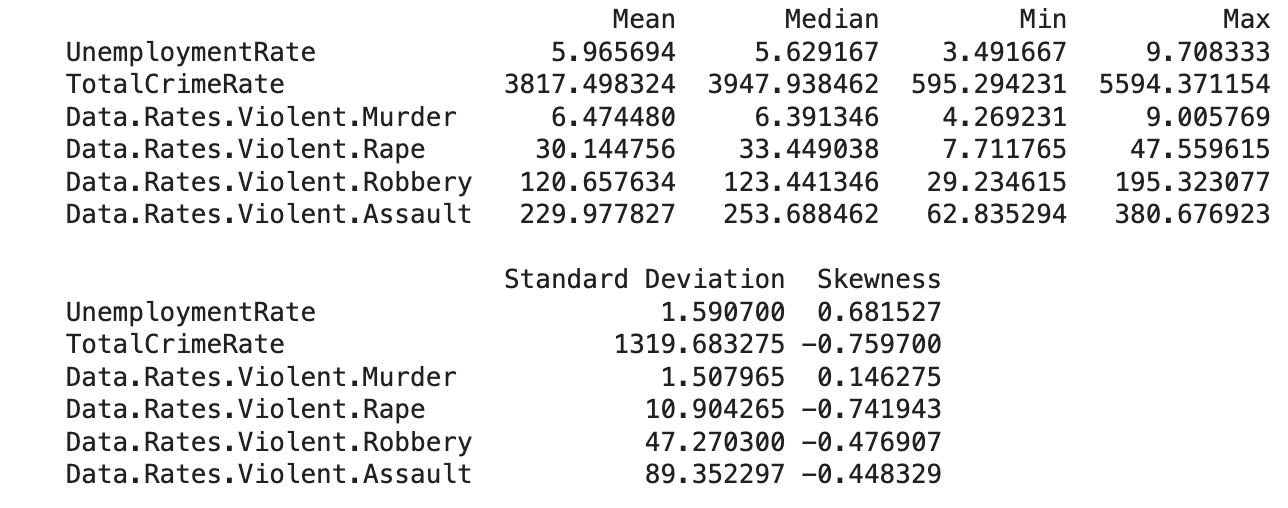
* **Yearly dataset:** 60 observations (1960–2019)
* **Monthly-state panel:** 50,024 observations

In the EDA section, this preparation was followed by multiple validation steps and transformations. Summary statistics revealed key distribution metrics—unemployment had moderate variation while crime rates showed high dispersion. The data was visually and statistically examined for trends across time, year-over-year changes, and crime type breakdowns. A new feature, **TotalCrimeRate**, was created by summing property and violent crimes, providing a comprehensive indicator. Additional sanity checks included missing value treatment, duplicate row inspection, and merging strategy verification. The result was a clean, normalized, and temporally aligned dataset ready for analysis and modeling.

# **4. Exploratory Data Analysis (EDA)**

EDA assists in building an intuitive understanding of the data by exploring trends, patterns. This section presents descriptive statistics and visualizations to call attention to key trends in unemployment and crime over the past decades. It sets the stage for deeper analysis by showing correlations and coming up with questions about causality and context.

## **4.1 Summary Statistics**



This statistical summary confirmed that crime has far greater dispersion and variance than unemployment. Property crimes such as larceny and burglary were found to dominate the dataset in terms of magnitude and variation.

## 

## **4.2 Temporal Trends**

* **1960–1991:** A consistent and non stop rise in total crime was observed, in which it peaked in the early 1990s. Unemployment fluctuated with typical economic cycles but did not show a long-term upward or downward trend which could be due to socioeconomic factors.
* **1992–2019:** Crime rates began to fall sharply following nationwide policing reforms, demographic changes, and potentially improved socio-economic programs. Meanwhile, unemployment maintained its cyclical pattern with visible spikes during recession years (e.g., 2008).

## 

## 

## 

## 

## 

## 

## 

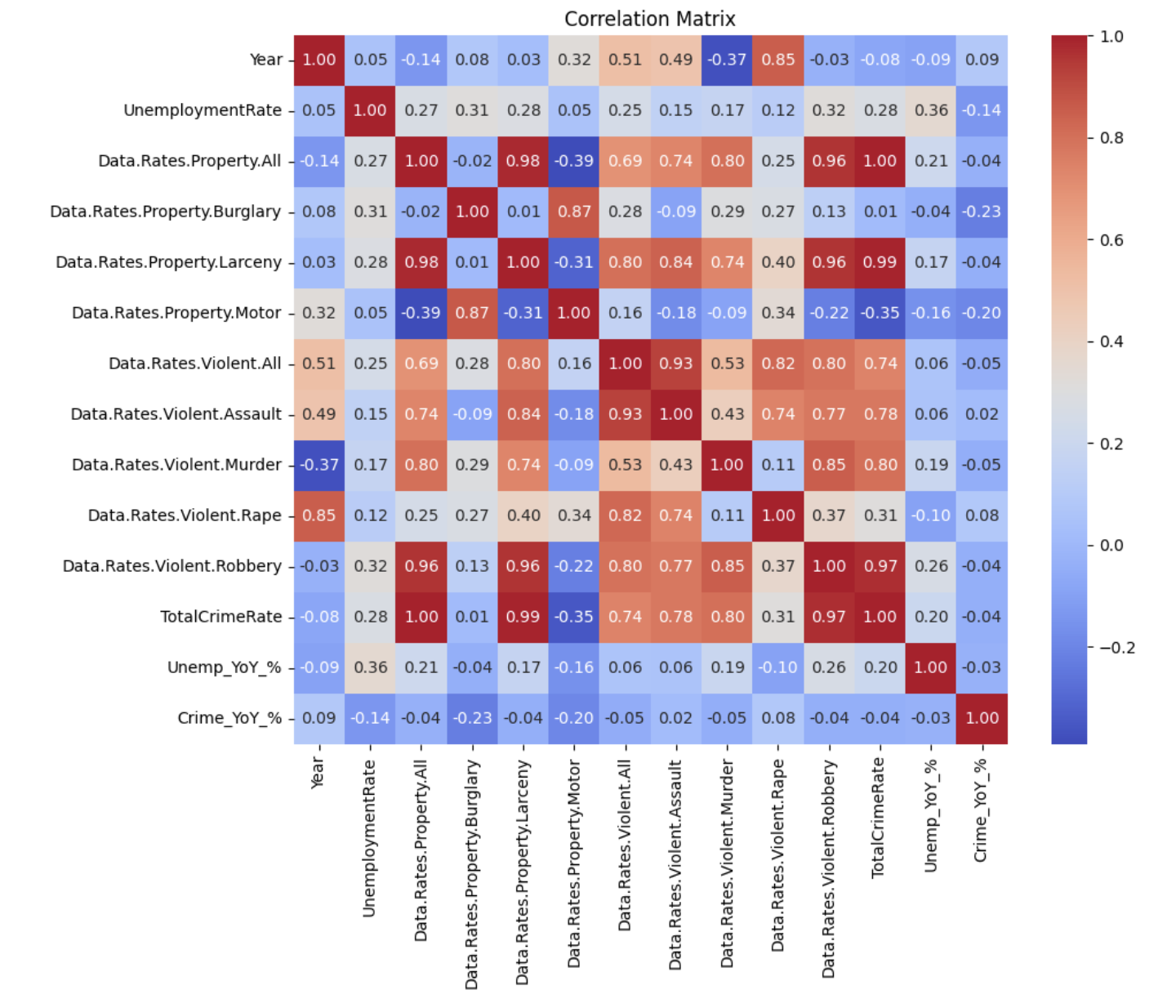
## 

## 

## **4.3 Correlation Heatmap**

A Pearson correlation matrix quantified linear associations between unemployment and various crime types. Findings included:

* Moderate positive correlations (0.25–0.35) between unemployment and robbery, assault, and larceny.
* Weak correlations with motor vehicle theft and rape.
* Year-over-year change rates (percent changes in unemployment or crime) showed weak or inconsistent relationships, suggesting that immediate economic shocks do not always produce crime spikes in the same year.



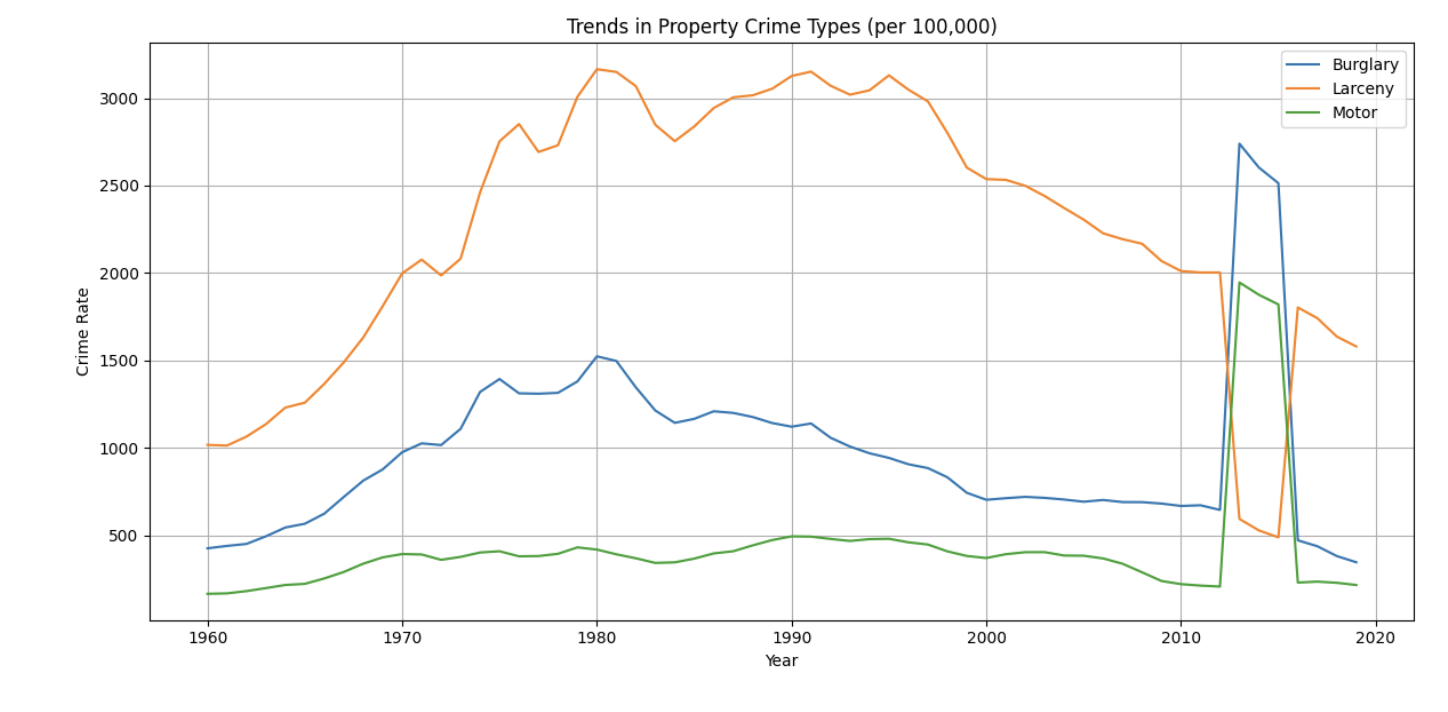
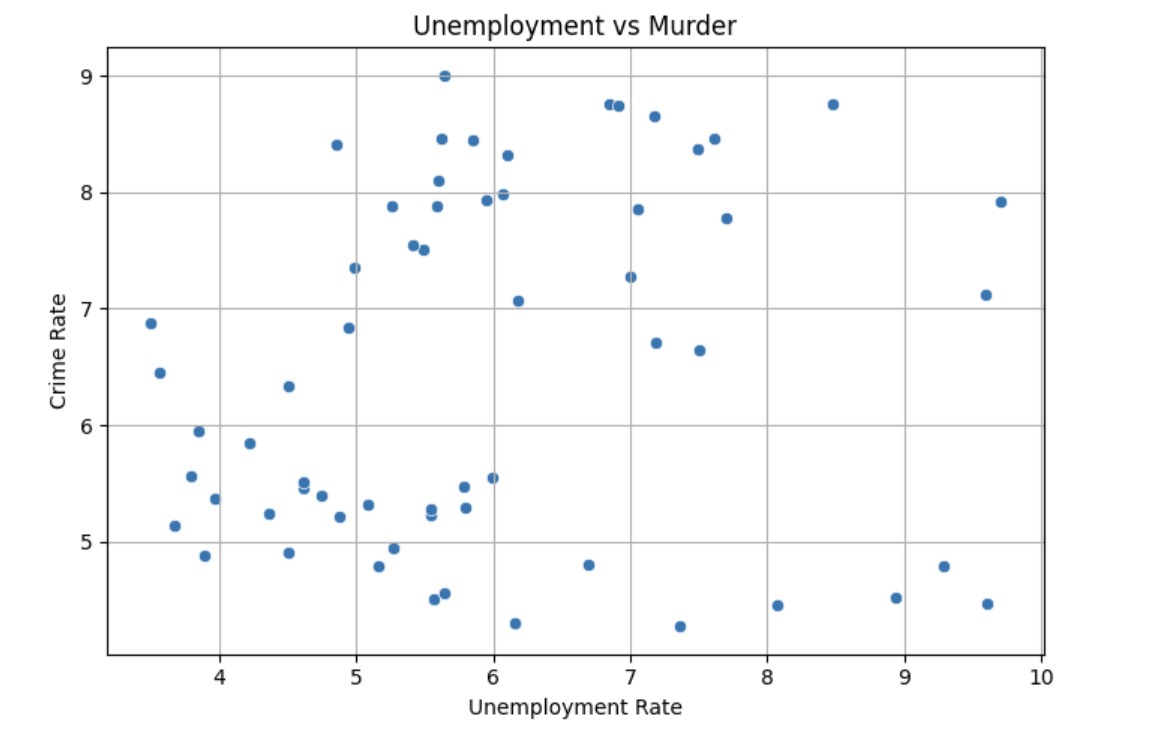
## 

## **4.4 Visuals**

To make patterns clearer:

* **Line plots** over time traced unemployment and total crime trends side-by-side, visually reinforcing the narrative of rising crime until the 1990s and subsequent decline.
* **Violent crime subtypes** were plotted separately to identify dominant trends; assault consistently showed the highest rates.
* **Scatter plots** displayed relationships between unemployment and individual crime categories. Larceny and assault showed clearer upward trends with increasing unemployment, while murder showed little to no relationship.
* **Annotated heatmaps** provided an at-a-glance summary of correlations among all quantitative variables in the dataset.

Examples of Visualizations:



# **5. Hypothesis Testing**

With initial patterns identified in the exploratory phase, the next logical step is to formally test whether the observation of the relationships between unemployment and crime are statistically significant or if it was just a coincidence. This section sets up and tests hypotheses regarding a potential linear association between the national unemployment rate and total crime rate using the **Pearson correlation method**.

## **5.1 Hypotheses**

* **Null Hypothesis (H₀):** There is no significant linear relationship between annual U.S. unemployment rates and crime rates.
* **Alternative Hypothesis (H₁):** There is a significant linear relationship between the two variables.

## **5.2 Methodology**

To test these hypotheses, we employed the Pearson correlation coefficient, a standard statistical measure that evaluates the strength and direction of a linear relationship between two continuous variables. The test was conducted on the merged dataset containing annual averages from 1960 to 2019, totaling 60 observations.

The variables involved were:

* **Independent variable:** UnemploymentRate (annual average %)
* **Dependent variable:** TotalCrimeRate (sum of violent and property crimes per 100,000 population)

We also ensured that the data met the assumptions of the test: both variables were continuous, normally distributed (approximately), and measured at the interval level.

## **5.3 Results**

* **Pearson Correlation Coefficient (r):** 0.2762
* **p-value:** 0.0327
* **Degrees of Freedom (df):** 58

The correlation coefficient of 0.276 indicates a **modest positive linear association** between unemployment and crime. While this is not a strong relationship, the p-value falls below the conventional alpha threshold of 0.05, allowing us to **reject the null hypothesis**.

## **5.4 Interpretation**

This result implies that **higher unemployment tends to be associated with higher crime**, even though the effect size is moderate and not high enough. The modest strength of the correlation aligns with the intuition developed during EDA: while economic hardship may be one factor influencing criminal activity, it is not the sole determinant. The socio-legal ecosystem, including factors such as law enforcement, education, urban planning, and cultural influences, also likely play significant roles.

Moreover, the correlation does not imply causation. The observed relationship, while statistically significant, could be confounded by other latent variables or structural societal shifts. For example, demographic trends or changes in crime reporting practices may affect both unemployment and crime rates simultaneously.

## 

## 

## **5.5 Visual Confirmation**

A scatter plot created during EDA visually supports the statistical finding: while data points are dispersed, a **loose upward trend line** can be drawn, indicating that years with higher unemployment often coincide with elevated crime rates.

These statistical findings serve as the foundation for the machine learning models developed in the next stage, where richer, multidimensional features are leveraged to make predictive inferences.

# **6. Feature Engineering & Enrichment**

Raw data often needs to be reshaped to reveal meaningful signals. This section explains how new features were created to represent time-based effects, socioeconomic contexts, and lagged responses. Feature engineering allows machine learning models to better understand nuanced relationships between variables, especially over time and across states.

Feature enrichment is critical for capturing the multifaceted influences behind crime beyond just unemployment. By integrating incarceration flows and salary demographics, the model can incorporate systemic factors such as income inequality and criminal justice throughput. This allows for a more realistic representation of the social conditions contributing to crime, especially when attempting to forecast or classify future trends across states and time periods.

* **Temporal Features:** Year, month, weekday, hour-of-day
* **Socioeconomic:** These were added to embed deeper structural context into the model:

*Salary by Race and Gender:* These features show income disparities across different demographic segments. They serve as indicators of opportunity or inequality by region.

*Prison Admissions & Releases:* High admissions may reflect law enforcement intensity or local crime policy, while high releases may influence short-term local recidivism.

*Total Salaries by Gender:* Captures aggregated economic distribution, useful for contextualizing crime relative to economic opportunity.

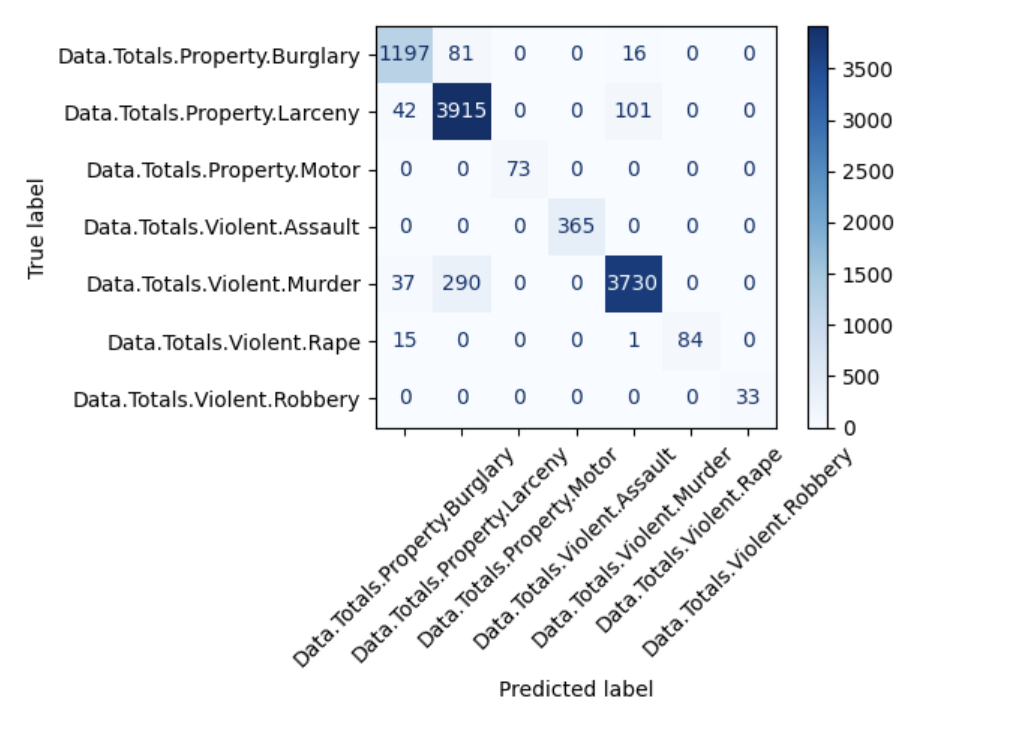
* **Lag Features:** Previous month’s unemployment

# **7. Machine Learning Modelling**

Once features are engineered, machine learning models can be built to predict or classify crime-related outcomes. This section presents two distinct modeling approaches: a classification model to forecast dominant crime types and a regression model to estimate violent crime rates. Each model is evaluated using appropriate metrics, and insights are drawn from their performance.

## **7.1 Classification – Random Forest**

The classification task focused on identifying which crime sub-category (e.g., assault, burglary, robbery) would experience the highest month-over-month growth in each U.S. state. This task leverages the full complexity of the dataset and tests the model’s ability to understand nuanced temporal and socioeconomic signals.



**Model Objective: Predict the dominant crime category (i.e., the one with the largest growth) for the next month per state.**

**Dataset Details:**

* **Size:** 50,024 rows × 70 engineered features
* **Target Variable:** DominantNextCrimeTypeMulti
* **Features Used:** Socioeconomic variables, incarceration stats, lagged unemployment, temporal flags

**Model Configuration:**

* RandomForestClassifier
* Parameters: n\_estimators=120, max\_depth=14, min\_samples\_leaf=8, class\_weight='balanced'

**Data Preparation:**

* SMOTE oversampling used to address class imbalance, particularly for rare crime categories like motor vehicle theft and rape

**Performance Metrics:**

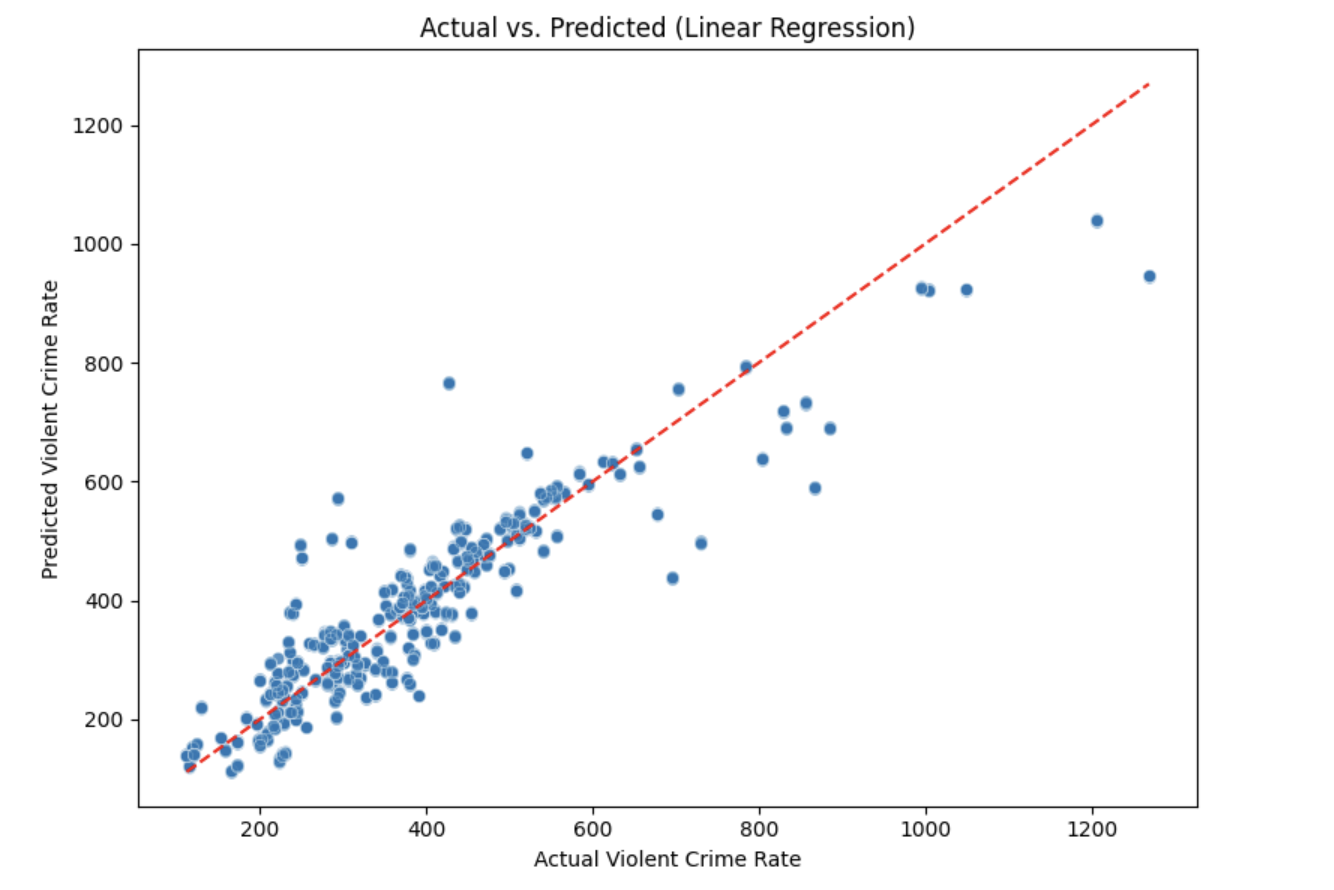
* **Accuracy:** 94%
* **Macro F1 Score:** 0.96
* **Recall on Rare Classes:** >95% for underrepresented categories like motor vehicle theft and rape

**Insights:**

* The model achieved high precision and recall across all classes, showing excellent generalization.
* Its success in predicting rare events demonstrates the strength of enrichment and balanced class strategies.
* Top contributing features included prison admission totals, salary disparities, and unemployment among working-age adults.

The ReasonRandom Forests were chosen due to their robustness to overfitting, ability to model nonlinear relationships, and inherent feature importance measurement. Their ensemble nature allows them to perform well even with correlated features, which is typical in socio-economic datasets. This model shows strong promise for deployment in crime early-warning systems or police resource allocation.

## **7.2 Regression – Linear Model**

To complement classification, we built a regression model to estimate the national violent crime rate for a given year.

**Model Objective:** Predict Data.Rates.Violent.All using socioeconomic features.

**Features:**

* Age-bracketed unemployment rates
* Total admissions to prison
* Average salary by race/gender

**Model Type:** Ordinary Least Squares (LinearRegression from sklearn)

**Performance:**

* **R² Score:** 0.83
* **Mean Absolute Error (MAE):** ~48 incidents per 100,000 population
* **Mean Squared Error (MSE):** ~5,115

**Top Predictive Features:**

* age\_25\_54\_rate (prime working-age unemployment)
* overall\_rate (total unemployment)
* admissions\_total (incarceration pressure)

While simpler than the classification task, the linear model effectively highlights strong associations between economic stress indicators and violent crime levels.

# **8. Key Findings**

The main insights uncovered throughout the project, combining statistical findings with machine learning results to better understand the relationship between unemployment and crime.

* Crime rates peaked during the early 1990s, a period shaped by multiple overlapping forces including high unemployment, demographic shifts, and policing policies. The decline in crime afterward, despite continued economic fluctuation, indicates other social drivers at play.
* Statistical testing confirmed a modest yet significant correlation between unemployment and crime (r ≈ 0.28), validating the hypothesis that economic hardship contributes to criminal activity—but not in isolation.
* The machine learning classification model was able to forecast the dominant crime category for upcoming months with 94% accuracy and a macro F1 score of 0.96. It performed especially well in identifying growth in rare crimes like motor vehicle theft, suggesting that systemic features like salary disparities and incarceration rates provide strong predictive signals.
* The regression model further supported this, with an R² of 0.83 when estimating violent crime rates from socioeconomic variables, demonstrating that while complex, crime can be predicted with reasonable accuracy when a broad set of features is included.

Together, these findings suggest that enriched machine learning approaches can play a valuable role in public safety forecasting, allowing stakeholders to proactively allocate resources and address underlying socioeconomic stressors.

# **9. Limitations & Future Work**

No analysis is without constraints. This section reflects on the limitations of the current study, such as the inability to infer causality and the issues with national averaging. It also outlines directions for future research, including using more localized data, applying causal inference techniques, and building more sophisticated predictive tools.

### **Limitations**

* **Loss of local detail:** National averages hide important state-level differences.

### 

### **Future Work**

* **State-level modeling:** Capture regional variation more accurately.
* **Policy data:** Add variables on laws, budgets, and enforcement.

By addressing these gaps, future work can build more robust, interpretable, and actionable models of crime prediction.

# **10. Technology Stack**

This section describes the stack used for data wrangling, visualization, modeling, and documentation.

**Languages:** Python

**Libraries:** pandas, matplotlib, seaborn, scikit-learn, imbalance-learn, scipy

**Tools:** Jupyter Notebook, VSCode, GitHub

# **11. Project Timeline**

Managing a semester-long project requires planning and structured milestones. This section outlines the project’s major phases, from proposal to final submission, showing how tasks were distributed over time. It demonstrates the iterative nature of the project and how earlier work laid the foundation for later insights.

| **Milestone** | **Date** | **Deliverable** |
| --- | --- | --- |
| **Proposal** | 14 Mar 2025 | README.md |
| **EDA & Hypothesis** | 25 Apr 2025 | Jupyter notebook |
| **ML Phase** | 23 May 2025 | Model artifacts |
| **Final Report** | 30 May 2025 | Word doc |