

Spring 2025: IDC 6940 – Data Science Capstone Final Report for Predicting Crime Hotspots in Florida

Prepared By: **Pedro Medina**

Project Mentor: **Dr. Levente Juhasz**

Course Instructor: **Professor Ananda Mondal**

Table of Contents

[Overview 1](#_Toc195962446)

[Goals 1](#_Toc195962447)

[Objectives 1](#_Toc195962448)

[Motivation 2](#_Toc195962449)

[Prior Arts & Challenges 2](#_Toc195962450)

[Prior Arts 2](#_Toc195962451)

[Challenges 2](#_Toc195962452)

[Data Sources & Description 2](#_Toc195962453)

[Methods & Tools 4](#_Toc195962454)

[Tools & Programming Languages Used 4](#_Toc195962455)

[Data Preparation 4](#_Toc195962456)

[Feature Selection 5](#_Toc195962457)

[Predictive Modeling 5](#_Toc195962458)

[Results 6](#_Toc195962459)

[Exploratory Data Analysis (EDA) Results 6](#_Toc195962460)

[Correlation & Feature Selection Results 7](#_Toc195962461)

[Model Performance & Geospatial Visualizations 9](#_Toc195962462)

[Discussion 11](#_Toc195962463)

[Interpreting Findings 11](#_Toc195962464)

[Feature & Model Insights 11](#_Toc195962465)

[Contributions & Conclusions 12](#_Toc195962466)

[Key Contributions 12](#_Toc195962467)

[Overall Conclusions 12](#_Toc195962468)

[References 12](#_Toc195962469)

# Overview

In modern times, Crime continues to be a significant challenge. Impacting public safety, economic stability, and the well-being of communities. Law enforcement agencies and policymakers have grown to rely on data-driven tools to make decisions and efficiently allocate resources. This project focuses on predicting and visualizing crime hotspots across all the counties in the state of Florida. The primary goal of this project is to find ways of enhancing public safety and support proactive policing strategies.

Using Machine Learning techniques and geospatial analysis, this project will analyze trends in Violent and Property Crime rates across Counties in the State of Florida from 2015-2020. The project will also investigate Socio-Economic factors such as Median Household Income and Poverty Rates to determine which factors have the highest influence on crime rates. At the end of the project, Geospatial data will be integrated to provide a visualization of the results and offer actionable insights for policymakers and law enforcement agencies.

Predictive models will be built after all patterns & correlations have been identified and visualized. These models will not only highlight areas with high volumes of crime but also provide actionable insights. The integration of demographic and geographic data providers a deeper understanding of crime trends across the state. The report documents all data sources, methodology process, tools used, and end results of the study. This report is written is such a way that any reader can reproduce this project.

# Goals

The primary goals of this project are:

1. **Predict crime hotspots** across all 67 counties in the state of Florida using crime and socio-economic data from the time range of 2015-2020.
2. **Identify socio-economic factors** that have the strongest correlations with violent and property crime rates.
3. **Develop machine learning models** that are capable of predicting regions of Florida with high crime rates.
4. **Visualize the distribution of crime rates** using geospatial mapping tools provided in coding languages such as R and Python.
5. **Provide data-informed strategies** for crime prevention and resource allocation to law enforcement & policymakers.

# Objectives

To achieve the goals, the project will focus on the following objectives:

1. **Gather and preprocess** datasets on Crime Rates and Socio-Economic Factors for all 67 Florida counties, ranging from the years of 2015-2020.
2. **Merge and normalize** datasets, ensuring that all of the data is compatible and is prepared for a county-level analysis.
3. **Selec**t socio-economic factors with the highest correlations to crime rates using Correlation Analysis.
4. **Train and evaluate** machine learning models to complete the prediction of violent & property crime hotspots.
5. **Visualize** results using geospatial mapping, highlighting counties with high predicted crime rates.

# Motivation

Identifying and understanding the Socio-Economic Factors that influence Crime, and its spatial distribution is vital for decision making. This project aims to reveal these patterns to help to contribute to safer communities and efficient allocation of the proper resources.

# Prior Arts & Challenges

## **Prior Arts**

Previous projects have explored the topic of crime prediction using Machine Learning. Some projects vary in the data that they use. While these projects have provided valuable insights, a notable limitation in many of them is the lack of integration of Geospatial data. This leads to a failure of capturing critical location-based trends that could help to improve prediction accuracy and policy relevance. This project aims to build upon the shortcomings found in previous works by integrating Geospatial analysis with the initial Crime and Socio-Economic Factors data.

## **Challenges**

* **Data integration across sources**
  + Combining crime and socio-economic data required normalization and alignment across different data formats, time periods, and geographic identifiers.
* **County-level geospatial merging**
  + Integrating shapefiles with statistical data required precise joining based on standardized county names and codes.
* **Feature selection & interpretation**
  + Identifying which socio-economic variables had the strongest influence on crime rates required a balance of correlation analysis, domain knowledge, and machine learning feature importance.
* **Modeling tuning & validation**
  + Ensuring reliable model performance across multiple counties and years involved hyperparameter tuning and cross-validation strategies.
* **Maintaining reproducibility**
  + Ensuring that every step taken in this project could be reproduced by outside users involved clear documentation and standardized coding methods.

Despite facing these challenges, this project constructs a pipeline for crime hotspot prediction that integrates spatial, historical crime, and socio-economic data.

# Data Sources & Description

This project utilizes three main datasets that capture the necessary data to fulfill the areas of crime statistics, socio-economic factors, and geographic boundaries. All data was collected and preprocessed for all 67 counties that compile the state of Florida, spanning the time range of the years of 2015 through 2020.

1. **Crime Data**
   * **Source**: [**Florida Department of Law Enforcement**](https://www.fdle.state.fl.us/CJAB/UCR/Annual-Reports/UCR-Offense-Data)
   * **Dataset(s) description**: Data was split into two datasets. One dataset (68 rows x 8 columns) represents the annual counts of reported violent crimes at the county level. The other dataset (68 rows x 7 columns) represents the annual counts of reported property crimes at the county levels.
   * **Data variables and units**:
     + Violent Crimes:
       - **Murder** – Number of reported homicide cases
       - **Rape** – Number of reported rape cases
       - **Robbery** – Number of reported robbery cases
       - **Aggravated Assault** – Number of reported aggravated assault cases
     + Property Crimes:
       - **Burglary** – Number of reported burglary cases
       - **Larceny** – Number of reported theft cases
       - **Motor Vehicle Theft** – Number of reported vehicle thefts
   * **Derived variables:**
     + **Total Violent Crime**
       - Murder + Rape + Robbery + Aggravated Assault
     + **Total Property Crime**
       - Burglary + Larceny + Motor Vehicle Theft
     + All crime counts were normalized to **Incidents per 100,000 people** (done during preprocessing)
   * Data was split into the 3 regions of Florida (North Florida, Central Florida, & South Florida) for regional analysis.
2. **Socio-Economic Factors Data**
   * **Source**: [**U.S. Census Bureau – American Community Survey (ACS) 5-Year Estimates**](https://data.census.gov/profile/Miami-Dade_County,_Florida?g=050XX00US12086)
   * **Dataset(s) description**: Socio-economic indicators collected at the county-year level (68 rows x 8 columns)
   * **Data variables and units**:
     + **Median Household Income** – reported median income (USD) for households per county
     + **Poverty Rate** – percent of total population living below poverty line
     + **Unemployment Rate** – percent of civilian labor force (age 16+) that is unemployed
     + **Educational Attainment Levels** – percent of adults (age 25+) with a **Bachelor’s degree or higher**
     + **Homeownership Rates** – percent of occupied housing units that are owner occupied
     + **Population Density** – **people per square mile** (**people/mi²**) calculated by dividing population by county area
     + **Total Population** – total number of residents per county
3. **Florida County Geospatial Data**
   * **Source**: [**U.S. Census Bureau TIGER/LINE 2020**](https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html#list-tab-790442341)
   * **Dataset(s) description**: Geospatial shapefiles used to map Florida’s 67 counties (67 rows x 13 columns).
   * **Data variables and units**:
     + **STATEFP, COUNTYFP, GEOID** – Federal and geographic codes
     + **NAME** – County name
     + **geomtery** – Polygon coordinates for each county

# Methods & Tools

This project followed a structured process that incorporates data gathering & preparation, feature selection, predictive modeling, and geospatial visualization. All work that was performed throughout the duration of this study was done in a reproducible manner using modern data science tools & methods.

## **Tools & Programming Languages Used**

* **R Studio** (version **4.3.1**)
* **Python** (version **3.10**)
* IDEs used: **RStudio** & Google **Colab**

**R Libraries used**:

* **dplyr & tidyverse** – used for manipulating data and wrangling
* **ggplot2, sf & tmap** – used for visualizations and geospatial plotting
* **readr & janitor** – used for cleaning all data
* **corrplot & psych** – used to generate correlation matrices and perform exploratory analysis

**Python Libraries used**:

* **pandas & numpy** – used for manipulating data
* **scikit-learn** – used for model training, feature selection, and model evaluation (RandomForestClassifier, SVC, train\_test\_split, GridSearchCV)
* **matplotlib & seaborn** – used for creating plots and visualizations
* **geopandas** – used for loading geospatial data and plotting (county shapefiles)
* **shapely, plotly** – advanced spatial joins and interactive maps

## **Data Preparation**

All datasets were loaded in using the .csv and .shp formats to ensure consistency. County crime counts were normalized as **incidents per 100,000 people** using the available population data. All datasets were merged using the county\_name or GEOID using the pandas library within Python and the dplyr library within RStudio. The geospatial shapefiles collected from the TIGER/LINE 2020 dataset were filtered to include only Florida counties (STATEFP = 12), and all geometry was preserved.

## **Feature Selection**

Correlation analysis was performed in R using the corrplot library to visualize the correlation between the socio-economic factors and the various crime types. The factors that had the highest correlation with the crime types were selected. Feature importance was computed from the trained Random Forest model in python to rank predictors based on their influence on crime prediction.

## **Predictive Modeling**

Three machine learning models were developed and trained in Python using the scikit-learn and XGBoost libraries to predict crime levels based on the previously selected socio-economic features. The predictions were measured in terms of **county-level crime rate per 100,000 residents**. Each model was trained to predict violent and property crime rates for all Florida counties.

1. **Decision Tree Regressor**
   * Simple, interpretable baseline model used to establish foundational accuracy.
   * Provides a clear view of decision rules and feature splits.
   * Prone to overfitting on small datasets, but useful for comparing models.
2. **Random Forest Regressor**
   * Ensemble model that builds multiple decision trees and aggregates their predictions to improve generalization.
   * Handles non-linear relationships and reduces overfitting through bootstrapped sampling and averaging.
   * Feature importance values were extracted using .feature\_importances\_.
3. **XGBoost Regressor**
   * High-performance gradient boosting algorithm with built-in regularization.
   * Tuned using GridSearchCV to optimize learning rate (eta), tree depth, and subsample ratio.
   * Especially effective in handling unbalanced data and noisy features.

**Training and Evaluation Strategy**:

* **Train-test split**: 80% training, 20% testing
* **Cross-validation**: 5 fold cross-validation used to ensure model robustness
* **Evaluation metrics**:
  + **RMSE (Root Mean Squared Error)** – measures the prediction error (the lower the better)
  + **Score (Coefficient of Determination)** – indicates how well variance is explained (the higher the better)

Each model was evaluated separately for violent and property crime rates. The best-performing model for each task was determined based on the lowest RMSE and higher score on the test set.

# Results

This section presents the results retrieved following the process that was outlined in the previous section.

## **Exploratory Data Analysis (EDA) Results**

The initial EDA phase provides key insights into crime patterns and socio-economic disparities across Florida’s 67 counties from 2015 to 2020.

**Distribution of Crime types**:

A screenshot of a graph

AI-generated content may be incorrect.

As presented in the pie charts above, **Aggravated Assault** made up the majority of violent crimes (66.9%), followed by Robbery (23.1%), Rape (8.8%), and Murder (1.1%). For property crimes, **Larceny** dominated the total count (76.6%), followed by Burglary (19.6%) and Motor Vehicle Theft (3.7%). The total amount of violent crimes in the dataset is **516,697** and **1,807,105** for property crimes.

A graph of crime and property type

Description automatically generatedThis imbalance between the crime types confirms that property crimes are far more prevalent than violent crimes statewide.

**Regional Crime Patterns**:

Bar charts were generated to visualize the regional comparison of crime totals per each region. **Property crimes consistently exceeded violent crimes** in all regions. **Central and South Florida** reported the highest levels of property crimes. **South Florida** had the highest volume of violent crimes, suggesting a need for a deeper localized analysis.

**Regional Socio-Economic Distributions**:

A diagram of different colored squares

AI-generated content may be incorrect.

**Income & Education**

South Florida reported the **highest median household income** and **educational attainment levels**, suggesting greater access to education and higher-paying jobs. North Florida had the lowest income and education levels across all counties.

**Poverty & Employment**

North Florida showed **higher poverty rates** and **lower employment levels** compared to other regions. Central Florida maintained a more balanced level of socio-economic metrics across all factors.

**Housing & Population Density**

South Florida had the **highest number of housing units** and was the **most densely populated** region by far, followed by Central Florida. North Florida, being more rural, showed lower density and fewer housing units.

## **Correlation & Feature Selection Results**

In order to identify which socio-economic variables were the most relevant for predicting crime rates, a 2-step feature selection process was implemented:

1. **Correlation Analysis**

* 2 correlation matrices were generated within R using the corrplot library. These matrices visualize the relationship between each socio-economic factor and the violent/property crime types.

A screenshot of a graph

AI-generated content may be incorrect.

**Poverty Rate** and **Housing Units** showed the strongest positive correlations with both violent and property crimes. **Educational Attainment levels** and **Employment Rates** had moderate correlations with the crime types. **Poverty Rates** had a negative correlation with property crime types, suggesting counties with higher poverty tend to have higher crime rates. **Median Household Income** showed the weakest correlations with both crime types, suggesting that even wealthier areas are prone to crime.

**Heatmaps of Actual Crime Rates**

To complement the correlation analysis, spatial heatmaps were created to visualize the actual distribution of the crime types across Florida. These maps provide insights into regional crime concentrations.

A map of florida showing the state's population

AI-generated content may be incorrect.

A close-up of several maps

AI-generated content may be incorrect.

1. **Feature Selection**

* Following the completion of the correlation analysis, feature selection was finalized using the features that showcased the highest correlations with the crime types. The goal of this phase was to strictly retain variables that contributed meaningfully to the models’ predictive performance.
* The features that were selected for the modeling phase were **Population Density, Housing Units. Educational Attainment Levels, Employment Rates,** and **Poverty Rates**.



## **Model Performance & Geospatial Visualizations**

**Model Performance Evaluation:**

Three regression models (Decision Tree, Random Forest, and XGBoost) were trained using the previously selected socio-economic featured to predict both violent and property crime rates across Florida.

Robbery was the best-predicted crime type, with XGBoost and Random Forest achieving an above 0.3. **Motor Vehicle Theft** also showed moderate predictability under Random Forest, with lower RMSE. **Murder**, **Rape**, and **Aggravated Assault** had poor scores across all models, indicating that these crimes may be influenced by variables not used in this project. **Larceny and Burglary** also had moderate predictability but high variance.

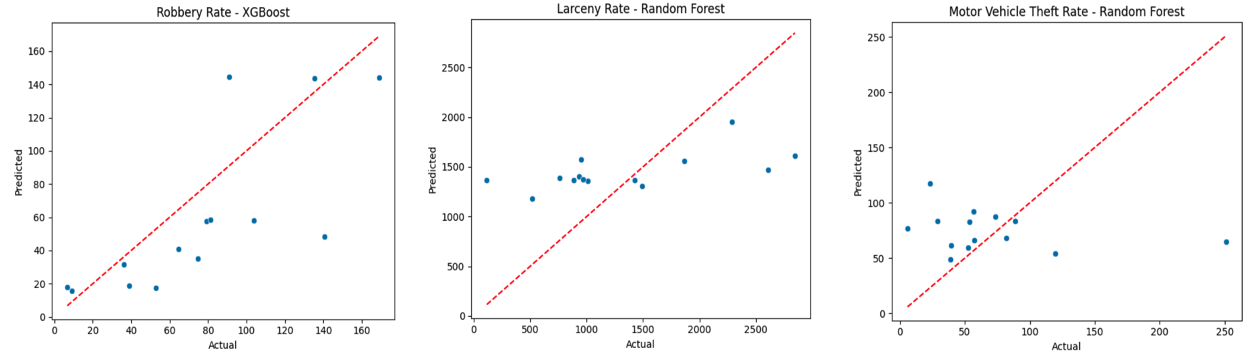
**Importance of Socio-Economic Factors for Predictions:**

A graph with blue and white text

AI-generated content may be incorrect.

To support the feature selection process, feature importance values were extracted from each model during the training phase. These values show the relative importance of each socio-economic factor on predicting specific crime types. These charts were created for the crime type that was predicted the best by each model. Across all three models, factors such as **Population Density**, **Housing Units**, and **Employment Rates** consistently emerged as top key predictors, aligning closely with the results gathered during the correlation analysis phase.

**Actual vs. Predicted Crime Scatterplots:**

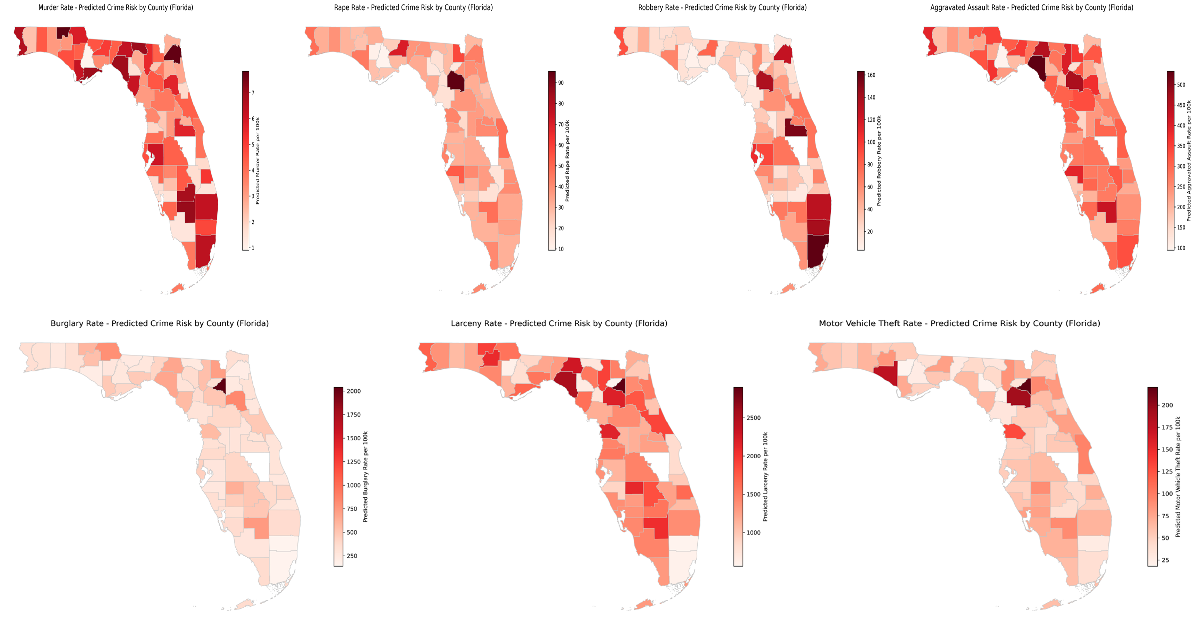


To finalize the evaluation of the performance of the models, scatterplots were generated to compare the **predicted vs. actual crime rates** for the three best predicted crime types: **Robbery**, **Larceny**, and **Motor Vehicle Theft**. These plots provide an assessment of the accuracy of the predictions, with each data point representing a county of Florida. The **red dashed line** represents a perfect 1:1 prediction.

As shown in the plots above, XGBoost has the strongest performance for predicting Robbery rates, as the predicted values where closely aligned within the acutal rates. The Random Forest model for Larceny followed a consistent upward trend with the actual data, indicating moderate predictive reliability. The Random Forest model for Motor Vehicle Theft had the highest variance, with many predictions falling below the 1:1 line.

**Geospatial Visualizations of Predicted Crime:**

To complement the statistical/numerical evaluation of the model performance, predicted county-level crime rates were visualized using heatmaps. These maps reflect **predicted crime risk levels** for all crime types across Florida, where **darker shades indicate higher crime rates per 100,000 residents**.



As shown in the maps above, the spatial distribution of predicted crime risk reveals distinct regional patterns:

* **South Florida** consistently ranked the highest In predicted crime across both violent and property crimes.
* **Central Florida** also emerged as a notable hotspot, especially for crimes such as larceny, burglary, and motor vehicle theft.
* **North Florida** tended to have lower predicted crime rates overall, though a few of its’ counties stood out with elevated risks in specific crime types.
* The maps also highlight disparities in rural counties, where predictions showed more variability. This is likely due to factors such as lower population sizes and smaller incident counts.

# Discussion

## **Interpreting Findings**

The results gathered from this project have confirmed that machine learning models and data science methods, when paired with socio-economic, historical crime, and geospatial data, can provide meaningful insights. Among the various crime types studies, **property crimes** were the most accurately predicted across the 3 models used. The **XGBoost** and **Random Forest** models have the best scores (above 0.3). These crimes are more prevalent and are greatly influenced by the socio-economic factors that were studied in this project.

In contrast, the **violent crimes** had an overall lower accuracy in their predictions. These crime types may be influenced by external factors that were not analyzed in this project, such as policing strategies, accessibility to weapons, etc. This gap highlights the limitations of relying solely on demographic and economic indicators for certain categories of crime.

## **Feature & Model Insights**

The feature importance rankings that were generated in the feature selection phase closely align with the patterns that were discovered during the correlation analysis phase. **Population Density**, **Housing Units**, and **Employment Rates** consistently emerged as the most influential factors across different crime types. **Educational Attainment Levels** had a moderate influence, but the impact varied slightly depending on the crime category.

Out of all of the models, **XGBoost** had a slightly better performance than Random Forest on most tasks. **Decision Trees**, while used as a baseline model, was outperformed in all categories.

# Contributions & Conclusions

This project demonstrates the results gained from the integration of machine learning, socio-economic data, and geospatial analysis to predict & visualize crime trends across the counties in Florida. By analyzing 5 years’ worth of crime and demographic for all 67 counties, the study provided predictive models and spatial insights that can help inform real-world decisions.

## **Key Contributions**

* Developed a **reproducible process** for crime prediction using publicly available tools
* Identified top **socio-economic predictors** of crime
* Trained and evaluated various machine learning models to assess their ability to predict crime
* Generated **predicted crime rates heatmaps** for all studied crime types, offering insights at the county level

## **Overall Conclusions**

* **Property Crimes** were easier predict than **violent crimes** likely due to the more consistent influence of socio-economic factors.
* **XGBoost** emerged as the best performing and most accurate model overall.
* The use of **geospatial data** made the results more interpretable and applicable, transforming raw predictions into **visual data** than can inform real-world decisions.

# References

[Chainey, S., & Ratcliffe, J. (2005). GIS and crime mapping. John Wiley & Sons.](https://www.wiley.com/en-us/GIS+and+Crime+Mapping-p-9781118685198)

[Florida Department of Law Enforcement. (2020). Uniform Crime Reports (UCR) Offense Data.](https://www.fdle.state.fl.us/CJAB/UCR/Annual-Reports/UCR-Offense-Data)

[Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2011). Randomized controlled field trials of predictive policing. Journal of the American Statistical Association, 106(493), 1399–1411.](https://www.tandfonline.com/doi/abs/10.1198/jasa.2011.ap09546)

[U.S. Census Bureau. (2020a). American Community Survey 5-Year Estimates.](https://data.census.gov/)

[U.S. Census Bureau. (2020b). TIGER/Line Shapefiles.](https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html)

[Wang, F., & Brown, D. (2013). Predictive modeling of crime incidents using historical data. International Journal of Forecasting, 29(4), 611–620.](https://www.researchgate.net/publication/309351102_Crime_Rate_Inference_with_Big_Data)

[XGBoost Documentation. (n.d.).](https://xgboost.readthedocs.io/)