DSA 210 Project
Breaking The Myth: A Data-Driven Perspective of Immigration's Impact on Criminal Activity

Sıla Horozoğlu

PROJECT STRUCTURE

This project investigates the often-politicized question of whether immigration influences crime rates in Europe. Motivated by public discourse that frequently links immigration to rising crime, this study applies a rigorous, data-driven approach to assess the validity of such claims. To ensure both breadth and depth, the analysis is conducted in two structured phases, each designed to progressively refine the research scope and methods.

In Part I, the project begins with a high-level analysis that examines the relationship between overall crime rates and immigration percentages across European countries. This initial phase uses traditional statistical tools—such as correlation analysis and ANOVA—to identify whether broad trends exist between immigration levels and crime incidence. While the results show no statistically significant positive relationship—and even suggest a weak negative association—this phase highlights the limitations of treating crime as a singular, aggregated measure. Different types of crime may have distinct social and institutional drivers, and collapsing them into a general index risks masking these nuances.

Part II addresses this limitation by introducing disaggregated crime data, enabling a more detailed examination of specific categories such as theft, intentional homicide, rape, and sexual assault. This phase also incorporates the Migrant Integration Policy Index (MIPEX), which provides essential policy context by capturing how inclusive or restrictive each country's immigration policies are. The refined analysis reveals that among various crime types, only theft demonstrates a statistically significant relationship with immigration levels, warranting further investigation.

Building on these insights, the final component of the project introduces machine learning models to enhance predictive accuracy and uncover complex, non-linear relationships that conventional methods may overlook. Theft—identified as the most immigration-associated crime—is used as the target variable. A variety of algorithms, including decision trees, random forests, XGBoost, and k-nearest neighbors, are applied to evaluate the predictive contribution of features such as refugee and asylum seeker percentages, socioeconomic indicators, and MIPEX scores. Dimensionality reduction techniques like Principal Component Analysis (PCA) are used to resolve multicollinearity, particularly among highly correlated sexual crime indicators. XGBoost emerges as the most effective model, revealing that economic conditions and crime context features significantly influence theft rates in ways that linear models may fail to detect. By combining traditional statistical reasoning with modern computational tools, this project not only tests prevailing assumptions about immigration and crime but also demonstrates the value of machine learning in social science research.

CONTENT

Part I: Preliminary Analysis

1.Introduction

This study investigates the widely debated relationship between immigration levels and crime rates, a topic of public and policy interest, particularly in the context of global migration trends. While immigration is often framed as a potential driver of criminal activity, opposing narratives highlight that immigrants may be less likely than native populations to engage in crime. This project aims to explore this relationship through data-driven methods, focusing on a cross-national comparison and narrowing the scope to European countries to improve data reliability and consistency. By applying statistical and visualization tools, the project seeks to assess whether immigration levels meaningfully affect crime rates and to challenge commonly held assumptions in public discourse.

2. Data Import

Covering the period from 1990 to 2024, the immigration dataset includes estimates of the total number of international migrants by sex, as well as their places of origin and destination, for 233 countries and areas. The data used form this dataset is international migrant stock as a percentage of the total population in 2024.

Immigration data is downloaded from undesa_pd_2024_ims_stock_by_sex_and_destination.xlsx

The overall crime rate is calculated by dividing the total number of reported crimes of any kind by the total population, then multiplying the result by 100,000 (because crime rate is typically reported as X number of crimes per 100,000 people). The data belongs to the year 2024, which is in accordance with the immigration data.

Crime rate data is downloaded from Crime Rate by Country.

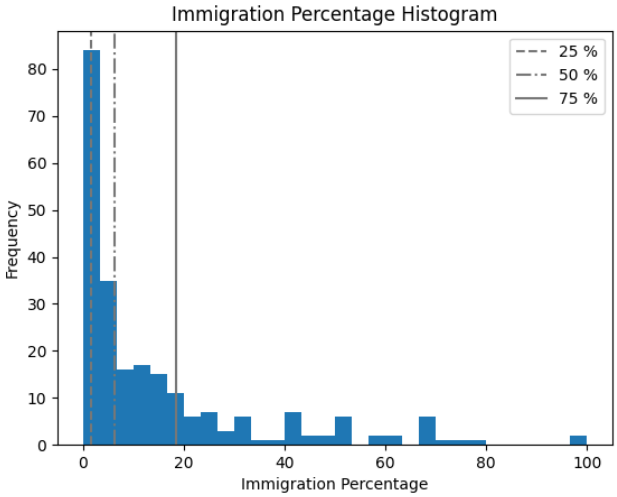
3. Data Pre-Processing

To ensure consistency across datasets from different sources, several preprocessing steps were performed. Country names were first standardized by removing non-alphabetical characters and aligning naming conventions (e.g., ensuring “United States of America” and “USA” matched across datasets). This step was critical for merging the datasets accurately.

An outer merge is used to combine all rows from two datasets, regardless of whether they have matching keys. This means that if a record exists in only one of the dataframes, it will still appear in the merged result, with missing values (NaN) filled in for the columns from the other dataframe. This approach is especially useful when working with datasets that may have only partial overlap—for example, when some countries appear in one dataset but not in another. By using an outer merge, we ensure that no potentially valuable data is lost during the merge process, which is important for comprehensive data exploration and analysis.

4. Exploratory Data Analysis

4.1 EDA of Immigration Percentage

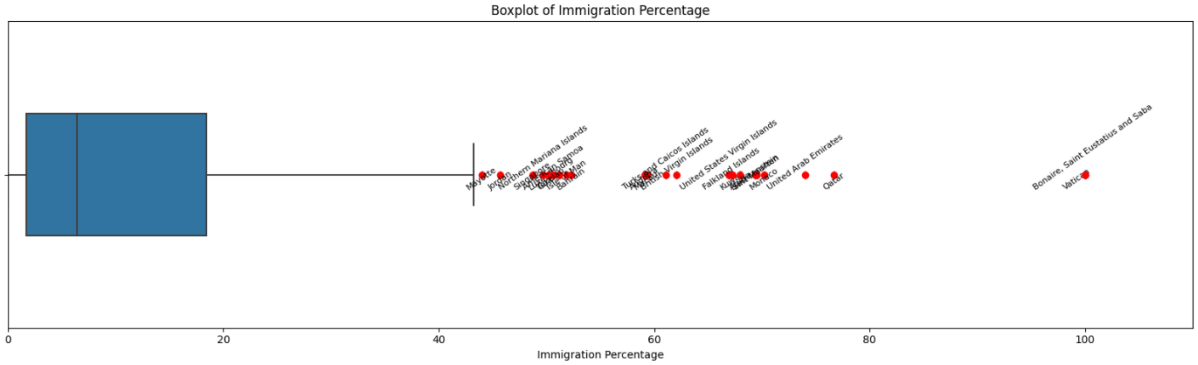


Observations:

1. Highly skewed distribution: The data is heavily right-skewed, indicating that most countries have low immigration percentages.

2. Long tail: There are some countries with very high immigration percentages, which are potential outliers.

Outliers:



When we look at the countries and regions that stand out with exceptionally high immigration percentages, a pattern quickly emerges. These aren't typical large, mainland nations — they're mostly tiny places with unique characteristics that make them statistical outliers.

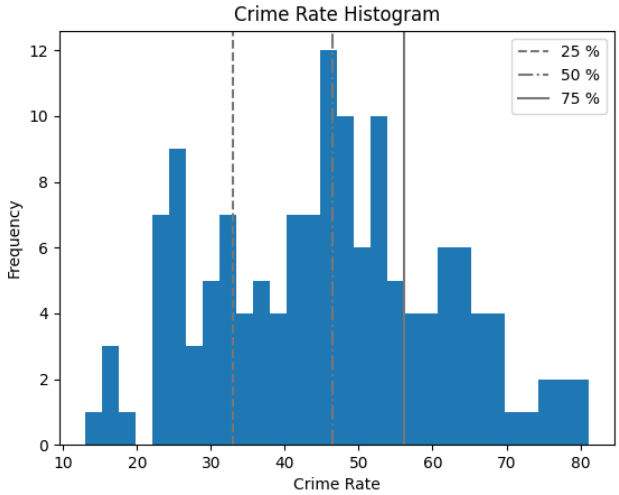
Bonaire, Saint Eustatius and Saba and Vatican show 100% immigration percentage. It is due to how immigration percentage is calculated. It refers to the ratio of number of immigrants to the total population. Bonaire, Saint Eustatius and Saba and Vatican have small population but high proportion of foreign-born residents, creating this statistical effect. So, they are statistical extremes due to their incredibly small populations. In such cases, just a few hundred people moving in can dramatically shift the numbers.

For instance, countries like Qatar (76.7%), Kuwait (67.3%), and the United Arab Emirates (74%) have booming economies but rely heavily on foreign labor. In these countries, the majority of the workforce consists of expatriates, often outnumbering native citizens.

Then there are microstates like Monaco (70.2%), Liechtenstein (69.4%), and Luxembourg (51.2%). These are small, wealthy European countries that attract high-skilled workers and businesspeople from across the continent. Because their populations are so small, even modest immigration numbers can create a very high immigration percentage.

Island territories make up another big chunk of the list — places like Aruba, Guam, American Samoa, and the British Virgin Islands. These often serve as tourist hotspots, offshore financial centers, or overseas dependencies of larger nations like the U.S., UK, or the Netherlands. Many residents are foreign-born workers, and migration from the mainland is counted as international immigration in the statistics.

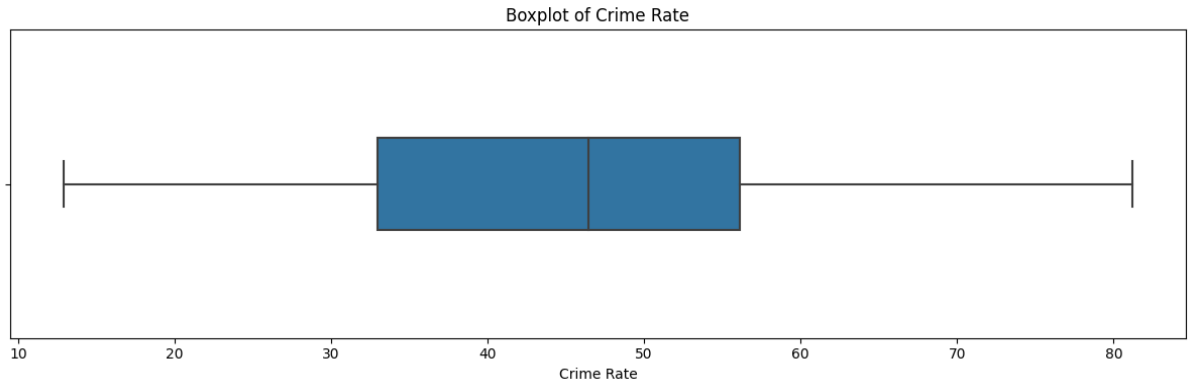
4.2 Crime Rate



Observation:

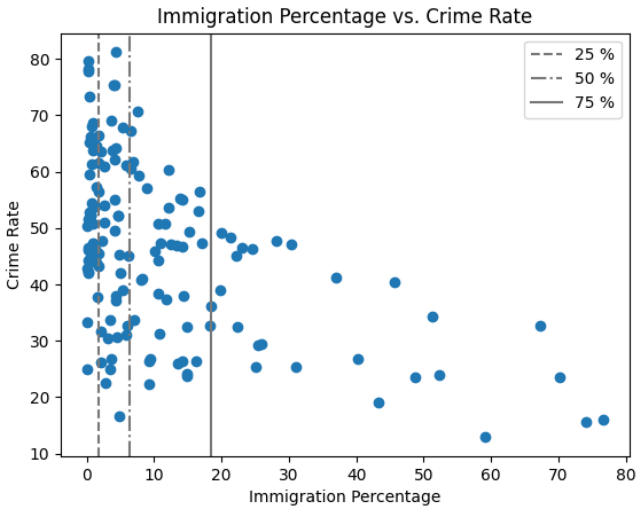
1. Roughly symmetrical distribution: The data appears more balanced with less skewness.

Outliers:



No outliers are detected for crime rate data.

4.3 Immigration Percentage vs. Crime



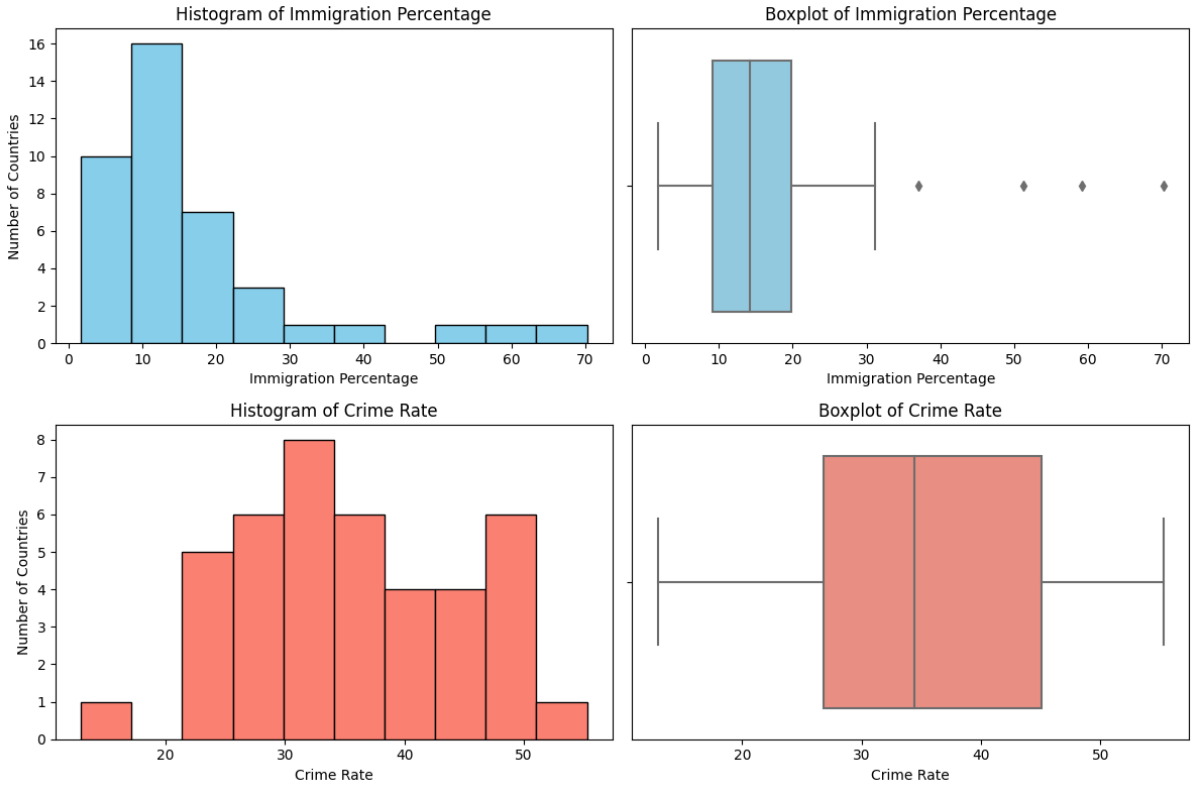
Observations:

1. Overall Trend: There appears to be no clear linear correlation between immigration percentage and crime rate.

2. Clustering: Most countries cluster in the low immigration(0-10%) and moderate crime rate zone.

While the initial goal of this project was to conduct a global analysis, data exploration revealed that certain regions—particularly in South America and parts of the Global South—had extremely high crime rates despite low immigration levels. These extreme cases were often driven by deep-rooted structural issues such as political instability, organized crime, poverty, and fragile institutions, which could obscure the specific relationship between immigration and crime. To ensure a more meaningful and controlled analysis, the study was narrowed to include only European countries, which generally have more consistent crime reporting standards, stable institutions, and comparable socioeconomic conditions. During exploratory analysis, several outliers were also identified—countries with crime or immigration values that skewed statistical results disproportionately. Narrowing the study to European countries also deals with these missing values by excluding outlier countries which are all outside Europe. This targeted and cleaned dataset allowed for more valid comparisons and increased the likelihood of uncovering patterns specific to the immigration-crime dynamic, rather than broader structural issues that dominate in less stable regions.

4.4 Europe Statistics



Observations:

1. The histogram of immigration percentage in Europe shows a right-skewed distribution. Compared to histogram of global data, the European immigration distribution is more compact.
2. The crime rate histogram shows a more uniform distribution.

Outliers:

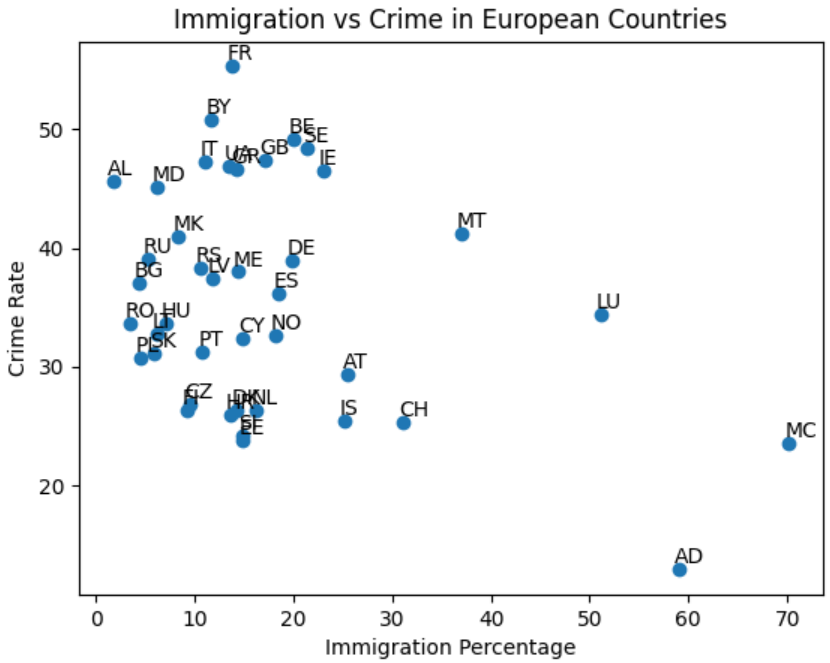
Immigration Outliers (based on IQR method):

	Country	Immigration Percentage
133	Andorra	59.1
140	Malta	37.0
153	Luxembourg	51.2
154	Monaco	70.2

The European countries identified as outliers in immigration percentage — Andorra, Malta, Luxembourg, and Monaco — share notable characteristics. They are all geographically small, economically strong, and maintain policies or conditions favorable to foreign residents. These factors make them highly attractive destinations for immigrants, and their small populations magnify the statistical impact of even modest migration levels.

No outliers are detected for crime rate, just as global data.

4.5 Immigration vs Crime in Europe



As we observe in above figure, there is no clear linear correlation. Further statistical analysis is required to confirm or reject potential relationships.

5. Hypothesis Testing

5.1 Correlation Coefficients

Immigration percentage and crime rate exhibits a Pearson correlation coefficient of -0.33. This suggests that, within Europe, higher immigration percentages tend to be associated with lower crime rates — or at the very least, there is no evidence of a positive relationship.

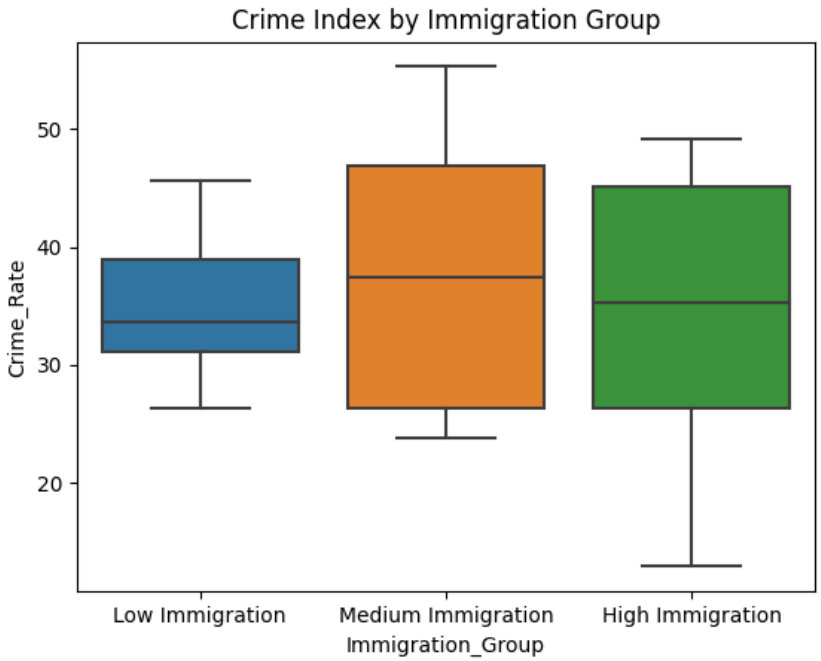
Spearman's rank correlation is measured as -0.174 indicating a weak negative monotonic relationship. P-values is measured as 0.276 which is above the significance level. Therefore, the correlation is not statistically significant, we cannot reject the null hypothesis.

5.2 ANOVA test

To evaluate whether there is a relationship between immigration levels and crime rates, I divided the countries into distinct groups based on their immigration percentages as low, medium, and high immigration using quantiles (tertiles for 3 groups). This method allows us to search for potential non-linear patterns beyond linear relationships we can observe with Pearson or Spearman correlation coefficients. Using tertiles also ensures that each group contains roughly equal number of observations, which improves the comparability of statistical analyses. By applying ANOVA, I can statistically test whether the differences in crime indices across these groups are significant, thereby providing a clearer understanding of the impact of immigration on crime.

Null Hypothesis: Immigration percentage does not have an effect on the crime rate.

Alternative Hypothesis: There is at least one group whose mean crime rate is significantly different from others.



A boxplot of crime rate across three immigration-level groups reveals a slight increase in median crime from low to medium immigration groups, with the high immigration group showing the widest variability. While the trend may suggest a weak association, the overlapping interquartile ranges indicate that immigration level alone does not strongly predict crime rate. These findings align with the overall pattern observed in the correlation analysis.

The results showed an F-statistic of 0.17 and a p-value of 0.845. The p-values is far greater than the significance level of 0.05. Therefore, we **fail to reject the null hypothesis** concluding **no statistically significant difference** in mean crime rates between the groups.
This supports the previous correlation and visual analysis, suggesting that immigration percentage does not have a meaningful impact on crime rate among European countries.

6. Findings

Contrary to popular belief, the analysis revealed no evidence that higher immigration leads to higher crime. Instead, a weak negative association was observed, though not statistically significant. This suggests that countries with high immigration may have other structural or governance-related attributes—such as better reporting practices, inclusive policies, or stronger institutions—that help maintain lower crime rates.

PART II: DETAILED ANALYSIS

To strengthen the validity and depth of my analysis, I decided to enhance my dataset by including multiple types of crime rather than relying solely on a general crime index. While my initial hypothesis testing did not reveal a clear relationship between immigration levels and overall crime, I recognized that aggregating all crimes into a single measure may obscure more nuanced patterns. Different types of crimes—such as violent crimes, property crimes, and white-collar crimes—can have distinct social drivers and may be differently affected by immigration dynamics. By expanding the dataset to include specific crime categories like intentional homicide, sexual violence, sexual assault, theft, and rape, I aim to conduct a more targeted and meaningful analysis. This allows me to explore

whether immigration correlates more strongly with certain types of crime and perform more robust hypothesis testing that accounts for these distinctions.

To enrich the analysis, I included the Migrant Integration Policy Index (MIPEX) as it provides valuable insight into how well countries support the integration of immigrants through 8 policy areas. The policy areas of integration covered by MIPEX are the following: labor market mobility, family reunification, education, political participation, permanent residence, access to nationality, anti-discrimination and health. This index helps contextualize crime and immigration data by highlighting whether inclusive policies may influence societal outcomes.

1. Data Import

MIPEX score is based on a set of indicators covering eight policy areas that has been designed to benchmark current laws and policies. A policy indicator is a question relating to a specific policy component of one of the 8 policy areas. For each answer, there are a set of options with associated values (from 0 to 100, e.g., 0-50-100). The maximum of 100 is awarded when policies meet the highest standards for equal treatment. Within each of the 8 policy areas, the indicator scores are averaged together to give the policy area score for each of the 8 policy areas per country which, averaged together one more time, lead to the overall scores for each country.

Migrant Integration Policy Index is from MIPEX 2020 - Data Analysis Tool.

Intentional homicide, sexual violence, rape, sexual assault and theft data is collected from the Eurostat data in police-recorded offences. The Eurostat crime data are collected from police, prosecution service, courts and prison departments. The recorded values are in units of per hundred thousand inhabitants. In the International Classification of Crime for Statistical Purposes (ICCS) framework, sexual violence encompasses both rape and sexual assault. Rape is a more specific subset involving penetrative acts, whereas sexual assault includes a wider range of non-consensual sexual behaviors. Importantly, intentional homicide and theft are distinct categories that do not overlap with sexual violence or with each other.

Police-recorded offences data is from [crim_off_cat] Police-recorded offences by offence category.

International migrant stock as a percentage of the total population is taken from undesa_pd_2019_migrant_stock_total_dataset.xlsx.

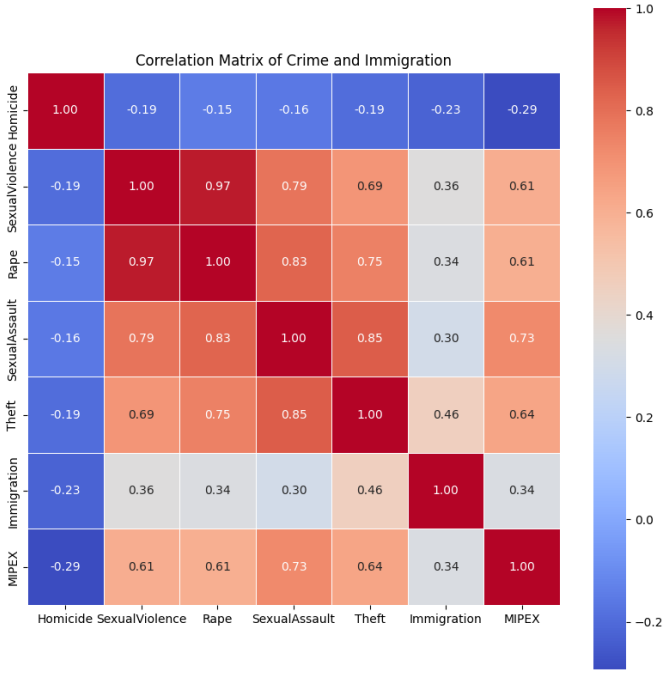
Since the most recent MIPEX data is from 2019, I aligned all other features—such as crime rates and immigration percentages—to that year to maintain consistency. Using 2019 data also avoids potential distortions caused by the COVID-19 pandemic, which significantly disrupted migration patterns and crime statistics.

2. Data Pre-Processing

As in the first analysis, country names were standardized using the country_converter Python package to address inconsistencies in naming conventions across different datasets. After standardization, only European countries were selected to ensure regional comparability. The dataframes varied in shape and structure, so an inner merge was used to retain only countries present in all datasets, ensuring consistency across features. Unnecessary columns were dropped to streamline the analysis. Upon checking the data types, it was noted that several numeric columns were incorrectly stored as object types. Initial attempts to convert them using standard numeric converters failed, which led to further investigation. It was discovered that missing values were represented using the character ':', rather than standard null types. These were manually replaced

with NaN, and then mean imputation was applied to fill in the missing values. A final check confirmed that all missing values were successfully handled, resulting in a clean and analysis-ready dataset.

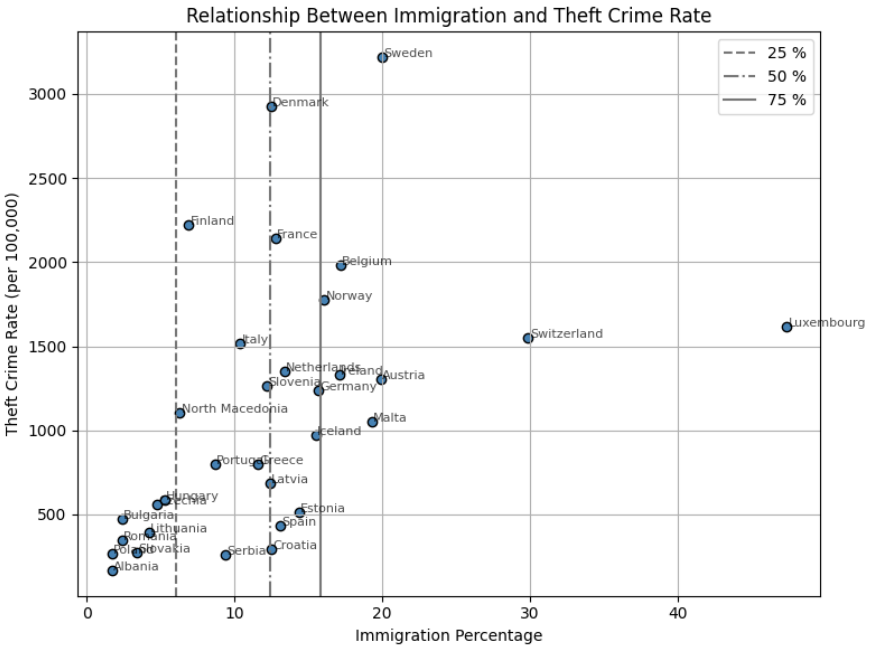
3. Exploratory Data Analysis



Observations:

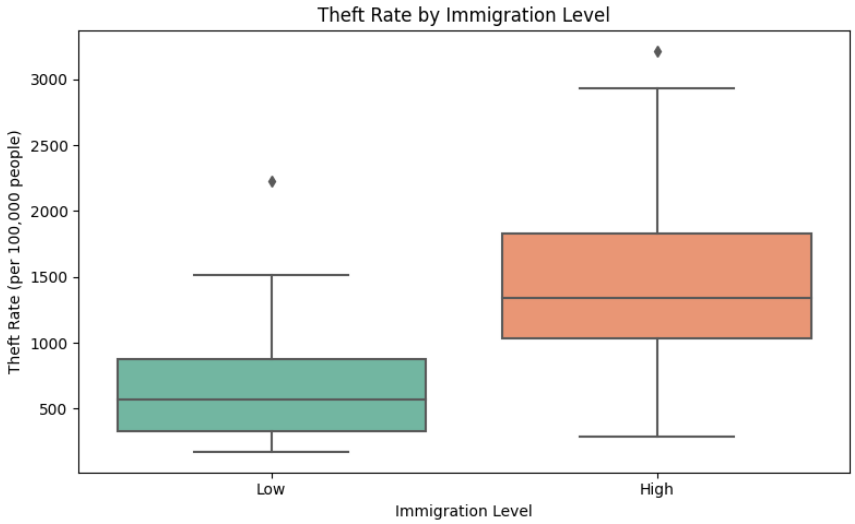
1. Rape, sexual assault and sexual violence show very strong positive correlations with each other. Rape and sexual assault are both subcategories of sexual violence under the International Classification of Crime for Statistical Purposes (ICCS) — they were included individually to preserve the level of detail provided by the source datasets.
2. Theft also shows positive correlations with sexual violence. This observation is not in the scope of this project.
3. Immigration shows weak positive correlation with most crimes, which will be the topic of hypothesis testing.
4. MIPEX and immigration are moderately correlated. This reflects logical consistency since countries with open immigration policies are likely to attract more immigrants and score higher on MIPEX.

No strong correlation is observed between immigration and violent crime. Highest correlation between immigration and any crime is with theft with 0.46 correlation coefficient, which will be the focus of the hypothesis testing.



A scatter plot comparing immigration percentage and theft crime rate across European countries shows **no clear linear trend**, despite a moderate positive correlation in the overall data. High theft rates are observed in countries with both low and moderate immigration levels, and some high-immigration countries (e.g., Luxembourg) exhibit only moderate theft.

Northern European countries with high theft rates (e.g., Sweden, Denmark) also have strong institutional transparency and reporting systems. So higher theft statistics might reflect better reporting, not necessarily higher actual crime. Mediterranean and Eastern European countries generally show lower reported theft rates, even with varying immigration percentages.



The boxplot shows that countries with high immigration levels tend to have higher median theft rates and greater variability compared to countries with low immigration. The interquartile range and upper whisker for the high immigration group are both elevated, suggesting that higher immigration is associated with higher reported theft in some countries.

4. Hypothesis testing

To statistically assess whether immigration levels have a significant effect on theft crime rates, a one-tailed independent t-test was conducted.
Countries were divided into two groups based on the median immigration percentage:

- Low Immigration Group: Countries below the median
- High Immigration Group: Countries at or above the median

The test compared the mean theft rates between these groups under the following hypotheses:

- Null Hypothesis (H_o): Immigration has no effect on theft rates.
- Alternative Hypothesis (H_1): Immigration has an increasing effect on theft rates — i.e., countries with higher immigration have higher theft rates.

The analysis produced a t-statistic of 3.04 and a one-tailed p-value of 0.0024, which is below the 0.05 significance level.
We **reject the null hypothesis**, concluding that high immigration countries have significantly higher theft rates.

5. Findings

The detailed analysis reveals that among various crime categories, only theft shows a statistically significant association with immigration levels. A one-tailed independent t-test comparing theft rates between high and low immigration countries yielded a p-value of 0.0024, leading to the rejection of the null hypothesis. This suggests that countries with higher immigration percentages tend to show higher theft rates.

However, this does not imply a causal relationship. Higher theft figures may reflect better reporting systems rather than higher actual crime incidence. Notably, countries with strong institutions and transparency, such as Sweden and Denmark, report high theft rates alongside high immigration, while some lower-immigration countries may underreport crime.

The inclusion of MIPEX scores provided contextual support, showing a moderate correlation with immigration levels but no direct link to crime rates. Overall, the analysis indicates that immigration is not broadly associated with crime increases, though theft appears as an exception worth further investigation.

6. Limitations

As with the earlier analysis, this phase is limited by the reliability and consistency of officially recorded crime data. The data refer only to crimes recorded by authorities and consequently reported to the police by victims and witnesses, among other things. Inferring crime occurrence from official crime figures can therefore be misleading. Furthermore, definitions and counting of official crime vary between countries, and comparisons between countries can therefore be misleading. The MIPEX index, while informative, is also subject to normative judgments in policy assessment. Most critically, correlation and group comparisons do not provide causal insight. Additional data on demographics, socioeconomic conditions, or internal migration could further strengthen future analyses.

PART 3

Data Enrichment

I introduced new data to create a better understanding of immigration's effect on crime. An international migrant is defined as someone living in a country other than the one in which they were born, usually for one year or more. The UN includes refugees in this definition, so I added information on different immigrant types such as refugees and asylum seekers. Refugees are people who flee their country due to war, violence, or persecution and are recognized under international protection. Asylum seekers are individuals who apply for refugee status but have not yet received a decision. The data is selected to be from 2019 to be in accordance with MIPEX and already existing crime data. Refugee and asylum seeker numbers are provided as raw counts, to make them comparable across countries with different population sizes, I divide them by total population and multiply by 100.

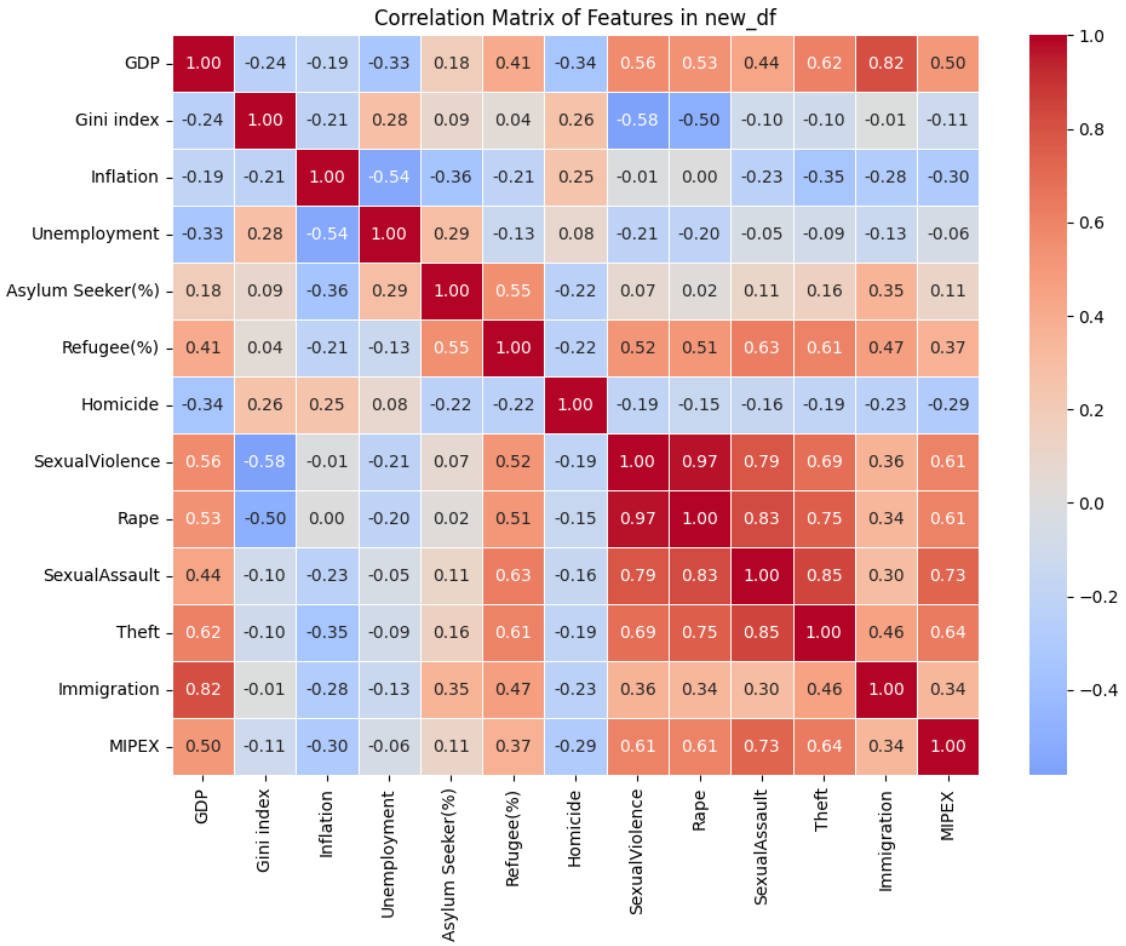
I also incorporated economic indicators such as GDP per capita, unemployment, inflation, and the Gini index. These variables serve two important purposes. First, they provide critical context for interpreting crime trends, as socioeconomic instability can independently influence criminal behavior regardless of immigration levels. Second, including these features helps control for potential confounding effects—ensuring that observed relationships between immigration and crime are not simply proxies for underlying economic conditions.

Unemployment is expressed as the percentage of the total labor force that is unemployed. GDP is reported as per capita in U.S. dollars. Inflation is measured as the annual percentage change in consumer prices. The Gini index ranges from 0 to 100, where 0 represents perfect income equality and 100 indicates maximum inequality.

Refugee, asylum seeker, GDP per capita, inflation, unemployment and Gini index data are downloaded from World Development Indicators | DataBank

Multicollinearity

To assess potential multicollinearity among the independent variables, a correlation matrix was visualized as a heatmap. As shown in the figure below, several strong pairwise correlations are observed, especially among the sexual crime indicators. Notably, sexual violence and rape ($r = 0.97$), sexual violence and sexual assault ($r = 0.79$), and rape and sexual assault ($r = 0.83$) exhibit extremely high correlation, indicating a high degree of redundancy. Additionally, theft shows strong correlations with sexual assault ($r = 0.85$), rape ($r = 0.75$), and sexual violence ($r = 0.69$), suggesting that these features may also interact in a multicollinear fashion. Immigration is strongly correlated with GDP ($r = 0.82$), which may also introduce overlap in their explanatory power. These patterns underscore the need for formal multicollinearity diagnostics, such as the Variance Inflation Factor (VIF), to ensure model stability and interpretability.



Variance Inflation Factor (VIF) is a metric used to detect multicollinearity among the independent variables in a regression model. Multicollinearity occurs when two or more predictor variables are highly correlated, which can distort the estimated effects of each variable and reduce the reliability of the model.

- A **VIF of 1** indicates no multicollinearity.
- **VIFs between 1 and 5** are generally acceptable.
- **VIFs above 10** suggest high multicollinearity and potential issues with coefficient stability.

In this project, VIF was used to assess the independence of predictors in regression models and identify any redundant or overlapping features that may affect interpretability and predictive performance.

Sexual Violence (VIF ≈ 28.83) and Rape (VIF ≈ 23.54) exhibited very high multicollinearity, indicating a strong correlation and potential redundancy when included together in the model. Sexual Assault also showed a high VIF (≈ 9.44), suggesting additional overlap with the other two variables — an expected outcome given the conceptual similarity among these features.
To address this issue and enhance model stability, the three variables were combined using Principal Component Analysis (PCA), which reduced dimensionality while preserving the shared variance. The data is scaled before applying PCA.

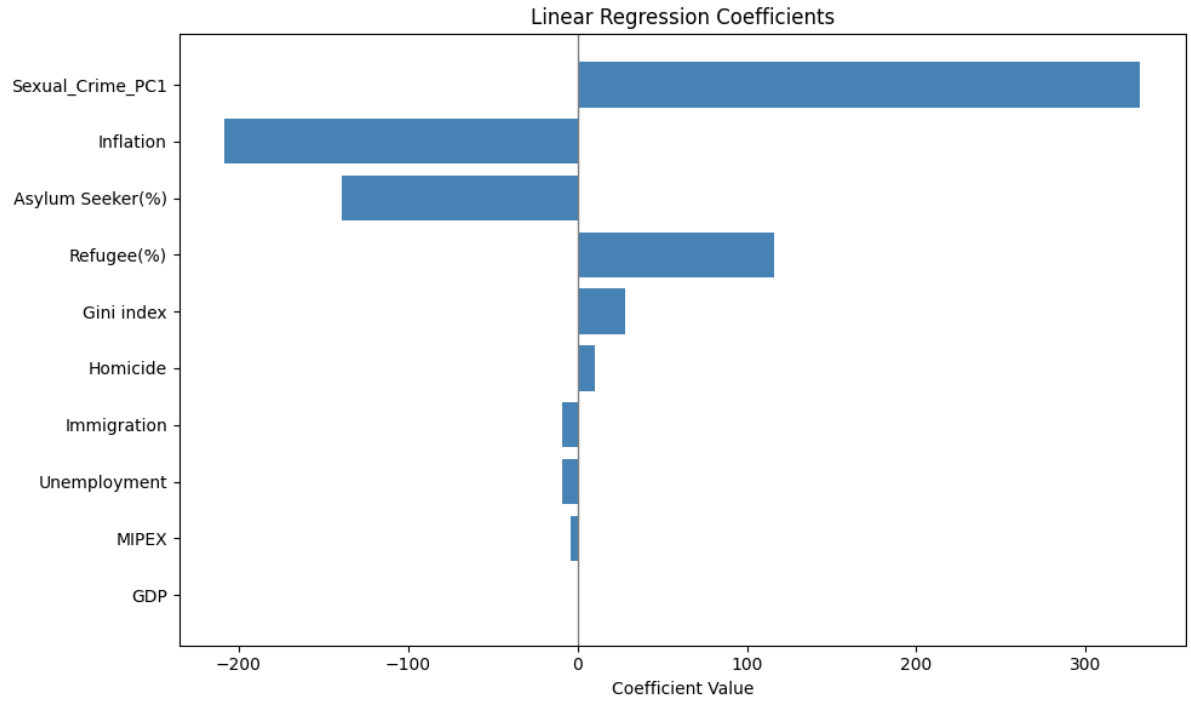
Machine Learning Models

Machine learning models were employed to predict theft crime rates, which showed the strongest correlation with immigration in the dataset. A variety of models were explored to compare performance, including decision tree, random forest, XGBoost, and k-nearest neighbors (k-NN) regression. Although linear regression is not counted as a valid machine learning model under the project requirements, it was included to establish a baseline error for comparison. Due to the limited number of European countries, the dataset was extremely small (32 rows), which constrained model complexity and the extent of hyperparameter tuning that could be applied.

Mean Squared Error (MSE) and R-squared (R^2) are standard metrics used for evaluating regression models. MSE quantifies the average squared difference between predicted and actual values. Root Mean Squared Error (RMSE) is the square root of MSE which is used for easier interpretation. R^2 measures how well the model explains the variance in the target variable. An R^2 of 1 indicates perfect predictions, while values close to 0 (or negative) indicate poor model performance.

Linear Regression

Linear regression was applied both to the original dataset (with multicollinearity) and to the PCA-transformed dataset. The **R^2 score improved from 0.75 to 0.85** after applying PCA, demonstrating the value of thoughtful feature engineering in improving model performance.



Key observations from coefficient figure:

- Sexual_Crime_PC1 has the largest positive coefficient by far, indicating that the composite measure of sexual crimes is the strongest linear predictor of theft.
- Refugee(%) also has a sizable positive coefficient suggesting a potential association between refugee concentration and theft—though this should be interpreted cautiously due to potential confounding factors.
- Inflation and Asylum Seeker(%) have strong negative coefficients, indicating that, within this dataset, increases in these variables are associated with decreased predicted theft.

Coefficient difference between refugees and asylum seekers could reflect differences in legal status, mobility, or settlement patterns compared to registered refugees.

- Socioeconomic indicators such as Gini index, Unemployment, Immigration, MIPEX, and GDP have relatively small coefficients, suggesting they play a minor role in the linear prediction of theft when more predictive crime-related and demographic variables are included.

There is research suggesting that inflation is a crime driver which contradicts with the result of this study. [1] One possible explanation is the presence of multicollinearity in the model. Having correlated predictors can make it difficult to interpret the sign and value of regression coefficient. The inflation is correlated with unemployment (-0.54) as observed in correlation matrix, which complicates their individual effects on theft. Negative correlation between inflation and unemployment is in accordance with economic theory (Phillips Curve).

Decision Tree

Decision tree resulted in R^2 of 0.78 without PCA data. Applying PCA to the dataset reduced the Decision Tree model's R^2 score from **0.78 to 0.56**, indicating a significant drop in predictive performance. This suggests that the Decision Tree model is capable of handling feature redundancy on its own, and that PCA may remove information useful for tree-based models. This highlights the importance of aligning feature engineering techniques with the characteristics of the model — in this case, PCA may be more appropriate for models sensitive to multicollinearity, such as linear regression, rather than for decision trees.

Some key hyperparameters of the DecisionTreeRegressor include:

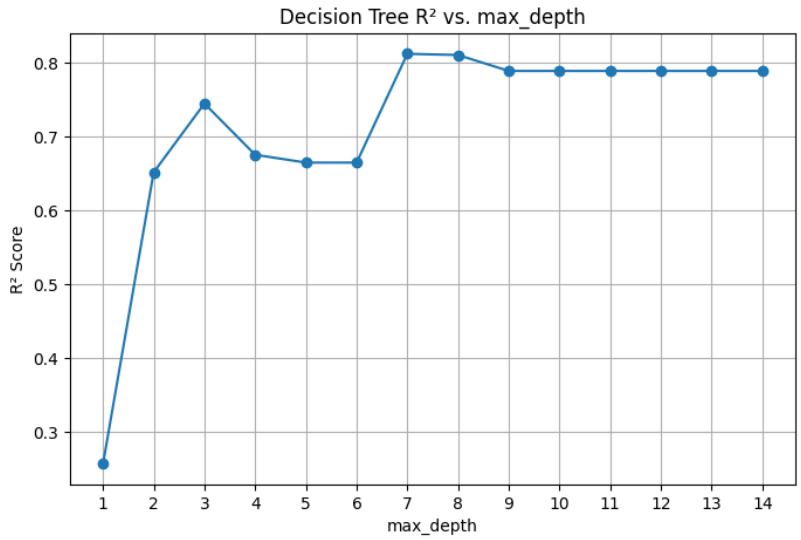
- criterion: *{"squared_error", "friedman_mse", "absolute_error", "poisson"}, default="squared_error"*
- splitter: *{"best", "random"}, default="best"*
- max_depth: *int, default=None*
- min_samples_split: *int or float, default=2*
- min_samples_leaf: *int or float, default=1*
- max_features: int, float or {"sqrt", "log2"}, default=None

Given the small dataset size (32 rows), there is a limited range for numerical hyperparameters. Based on this, I applied GridSearchCV with the following parameter grid:

```
param_grid = {
    'max_depth': [2, 3, 4, 5, None],
    'criterion': ['squared_error', 'friedman_mse'],
    'splitter': ['best', 'random'],
    'max_features': [None, 'sqrt', 'log2']
}
```

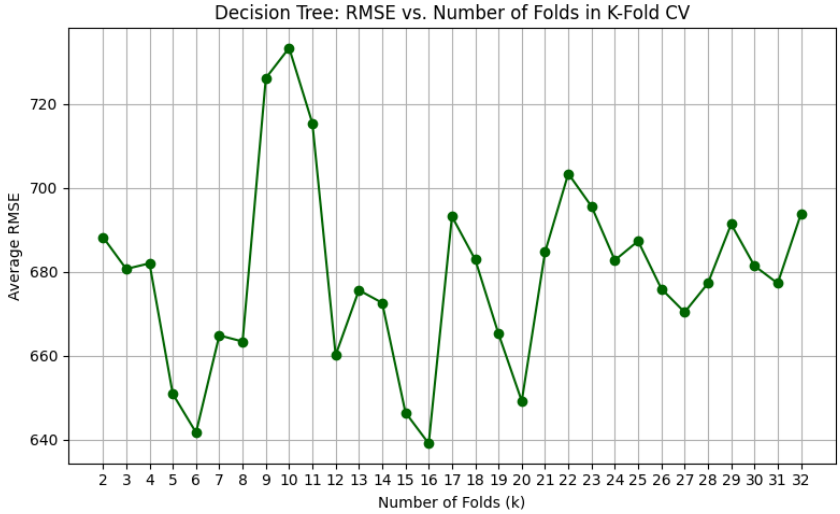
The tuned model, selected based on cross-validation performance, achieved an R^2score of **0.527**, which is **lower than the default model's** R^2 **of 0.78**. There are two key differences between the tuned and default models that explain this performance gap:

1. Max Depth: The tuned model selected max_depth=4, which likely led to underfitting — the tree was too shallow to capture relevant patterns. In contrast, the default tree grew to an actual depth of 9, providing enough capacity to model the dataset's complexity.



2. Splitter: The tuned model used splitter='random', which adds randomness to split selection and is generally less optimal for single decision trees. The default model used splitter='best', which selects the best split based on the criterion and typically leads to better performance.

3. Cross-Validation: As shown in the figure below, the average RMSE varies substantially with different numbers of folds in k-fold cross-validation. This instability highlights how sensitive model evaluation can be when the dataset is small, potentially leading GridSearchCV to favor underperforming configurations due to noisy validation scores.



Random Forest

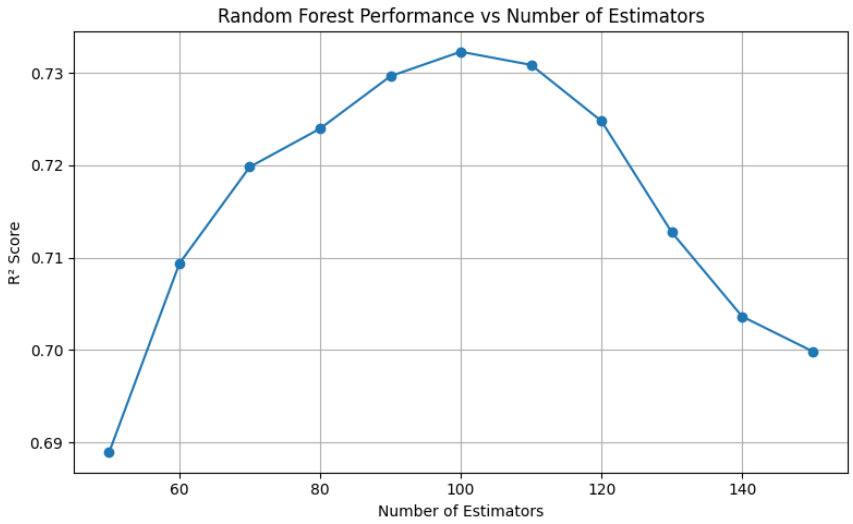
The RandomForestRegressor was evaluated as an ensemble model capable of reducing variance and improving generalization over individual decision trees. Using the default hyperparameters, the model achieved an R^2score of **0.73**, outperforming the tuned version, which yielded an R^2 of **0.71**. Additionaly, using PCA data decreased R^2 to 0.64.

Despite applying hyperparameter tuning with GridSearchCV, performance slightly declined. The best parameters selected were:

{'criterion': 'squared_error', 'max_depth': 5, 'max_features': None, 'n_estimators': 50}

The key differences between tuned and default versions:

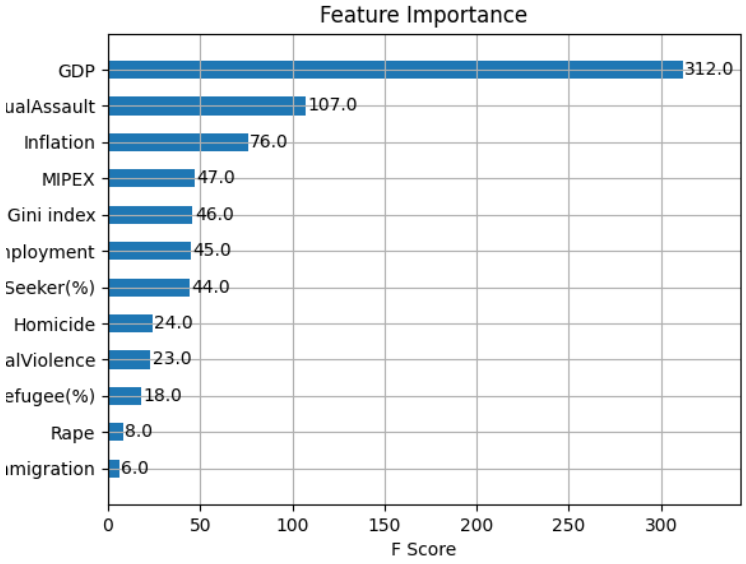
1. Max depth: The tuned model selected max_depth=5, which may have led to underfitting as observed in single decision tree model. The default random forest average tree depth is 6.44, which is in accordance with decision tree's optimal max_depth of 7 as shown in the figure of R^2 vs max depth figure before. Therefore, max_depth of 5 leads to underfitting.
2. Number of estimators: The tuned model selected n_estimators=50 compared to default of 100. R^2 vs number of estimators figure is given below. Increasing the number of estimators from 50 to 100 reduces variance by averaging over more bootstrap samples and feature subsets, smoothing out random fluctuations and yielding a measurable boost in test R^2. The most optimal values is 100 as used in default version.



XGBoost

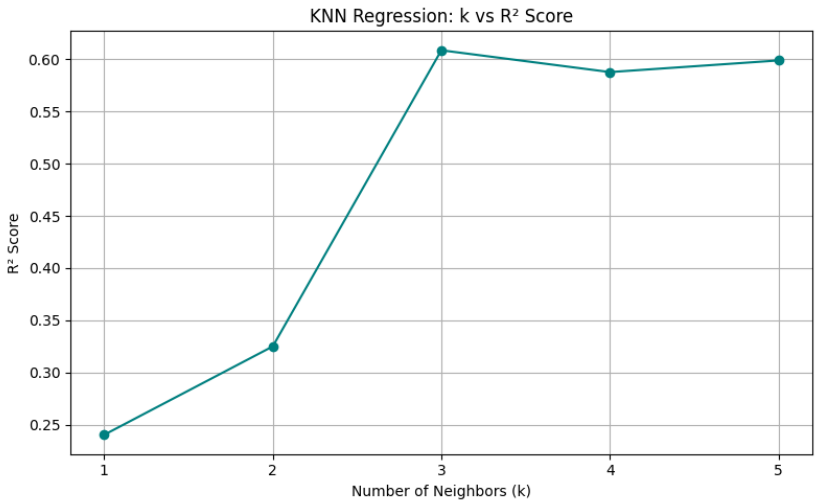
XGBoost is an optimized gradient-boosting library that builds trees sequentially with regularization and system-level enhancement. XGBoost proved to be the strongest model with R^2 of 0.78. As with tree-based learners, applying PCA decreased performance, R^2 fell to 0.71. RandomizedSearchCV was applied over n_estimators, max_depth and learning_rate, which lifted R^2 further to 0.80.

XGBoost's gain scores show that GDP is by far the strongest splitter, meaning differences in national income explain the largest swings in theft rates. Sexual assault and inflation follow, suggesting that higher violent-crime levels and economic instability create conditions where property crime also rises. In contrast, the linear model placed most weight on the first principal component of sexual crimes and on inflation, but downplayed GDP—highlighting that XGBoost uncovers nonlinear interactions that a straight-line fit cannot.



K-Nearest Neighbours (kNN)

KNN regression predicts the target value for new data point by averaging the target values of the k nearest neighbours in feature space. Since it is a distance-based model, we must do scaling to ensure every feature contributes proportionally. kNN regressor resulted in the lowest R^2 which is 0.61. We can observe the effect of number of neighbours with the figure given below. The optimal value is 3 which gave R^2 of 0.61.



Conclusion

The machine learning analysis provided a valuable extension to the statistical methods employed in earlier phases, offering predictive insights and highlighting interactions that linear models alone may overlook. Among all tested models, XGBoost achieved the highest R^2 score of 0.80 after hyperparameter tuning, demonstrating its ability to capture complex, non-linear relationships between immigration, socioeconomic indicators, and theft crime rates.

Interestingly, despite its simplicity, linear regression performed remarkably well—achieving an R^2 of 0.85 after applying PCA to address multicollinearity, particularly among highly correlated sexual crime

variables. This result emphasizes the importance of effective feature engineering and suggests that even straightforward models can deliver strong predictive performance when provided with clean, structured inputs. PCA proved especially useful for stabilizing the model by reducing redundancy in the dataset without sacrificing explanatory power.

The inclusion of economic and demographic variables—such as refugee percentages, inflation, GDP per capita, and the Gini index—further enhanced model interpretability by accounting for broader structural factors influencing crime. These findings reinforce the notion that crime trends cannot be meaningfully explained by immigration levels in isolation; rather, they emerge from the intersection of migration, policy, and socioeconomic context.