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

Sıla Horozoğlu

CONTENT

PROJECT STRUCTURE

This project was structured in two phases to enable both a broad overview and a more focused, in-depth examination of the relationship between immigration and crime. The first analysis provided a general understanding by investigating overall crime rates in relation to immigration levels across European countries. This approach helped identify high-level patterns and assess the existence of any statistically significant associations using aggregated crime data.

However, relying solely on a single overall crime metric risks oversimplifying the complex nature of criminal activity. Different types of crime may have different social drivers and relationships with immigration. Therefore, a second, more detailed analysis was conducted using disaggregated data on specific crime categories—such as theft, rape, and homicide—and included the Migrant Integration Policy Index (MIPEX) to contextualize findings within national policy environments.

By dividing the project into two analytical stages, the study balances comprehensiveness with depth, ensuring that both general trends and specific nuances are captured in evaluating whether and how immigration relates to crime in Europe.

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.

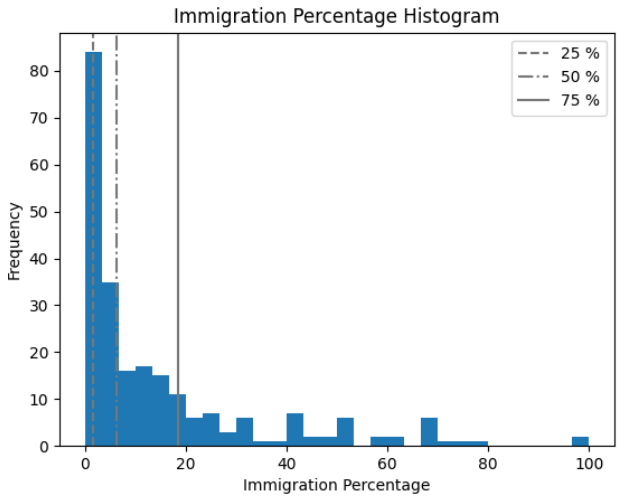
3. Data Pre-Processing

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

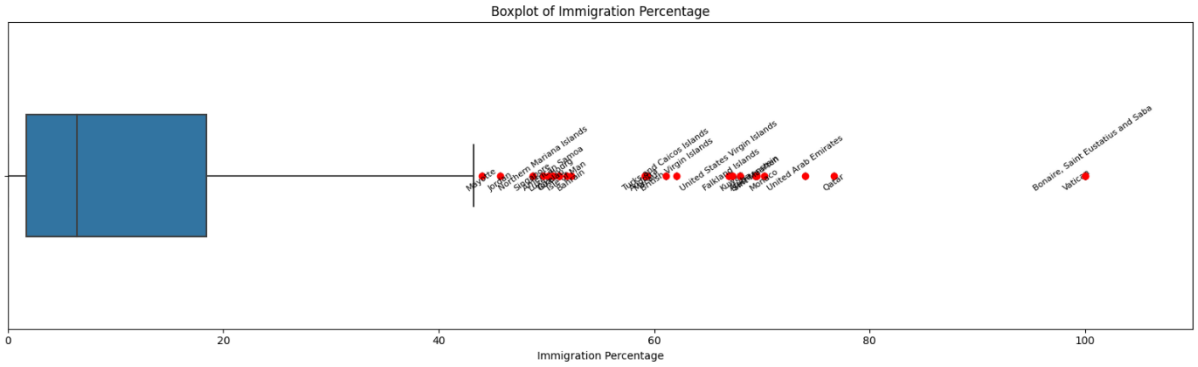


Observations:

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.

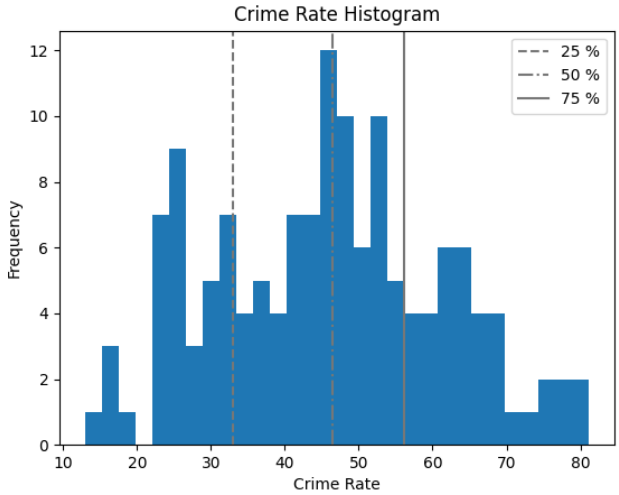
Bonaire, Saint Eustatius and Saba and Vatican show 100% immigration percentage. It is due to how immigration percentage is calculated. It refers to the ratio of number of immigrants to the total population. Bonaire, Saint Eustatius and Saba and Vatican have small population but high proportion of foreign-born 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.

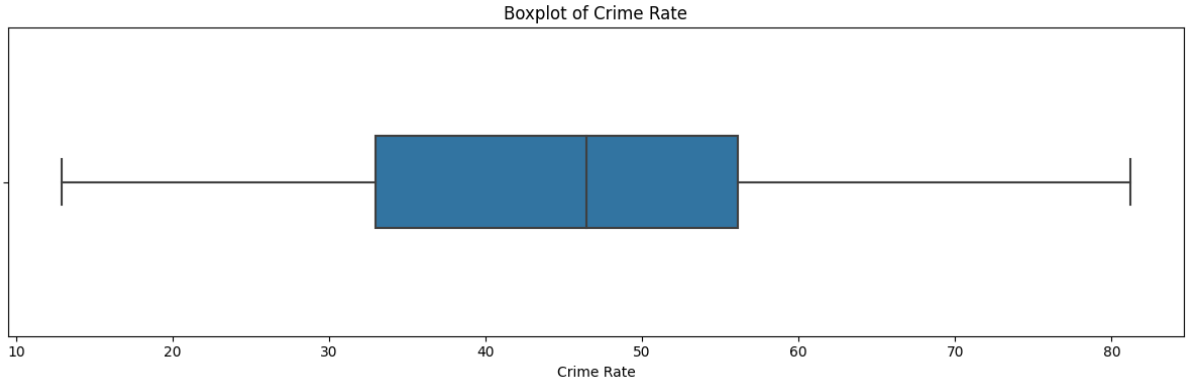
4.2 Crime Rate



Observation:

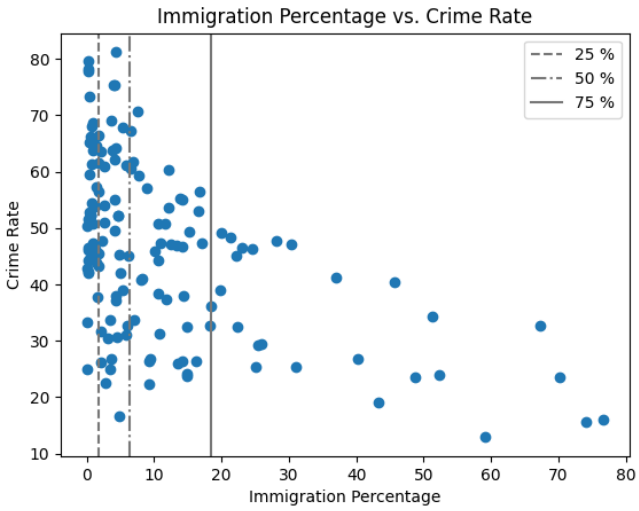
1. Roughly symmetrical distribution: The data appears more balanced with less skewness.

Outliers:



No outliers are detected for crime rate data.

4.3 Immigration Percentage vs. Crime



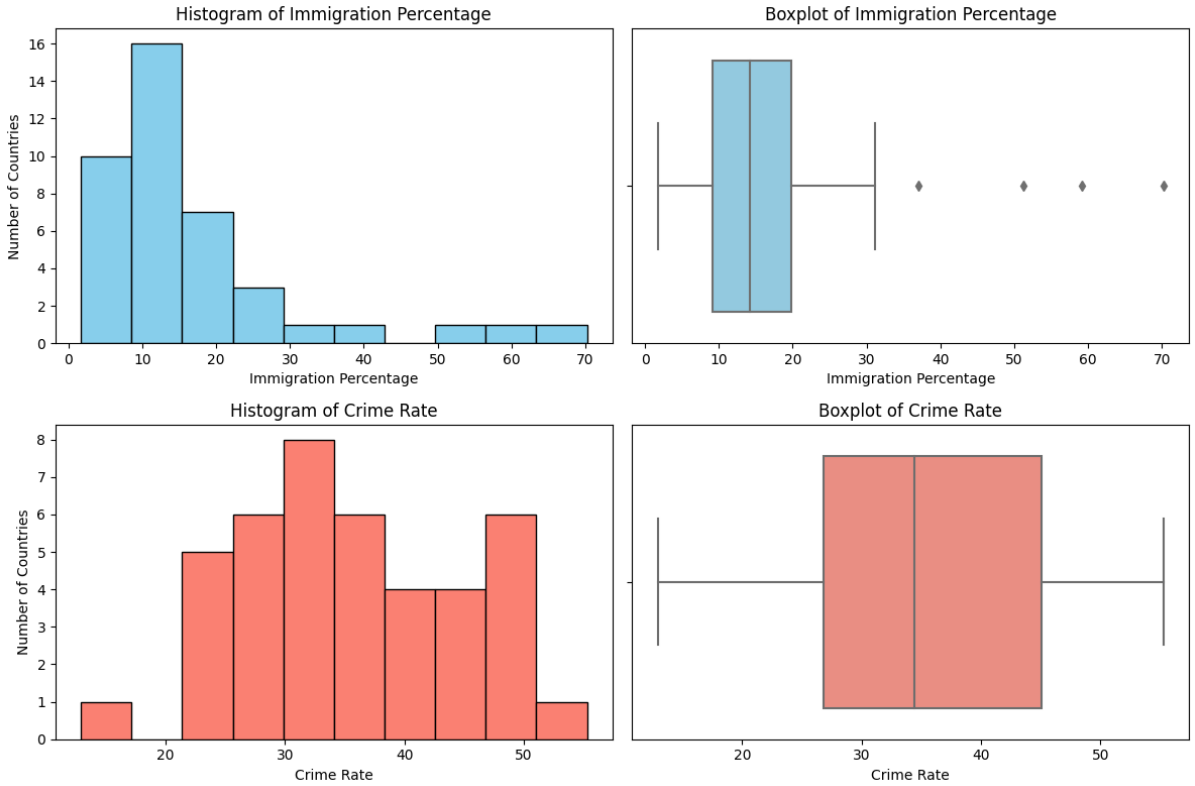
Observations:

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:

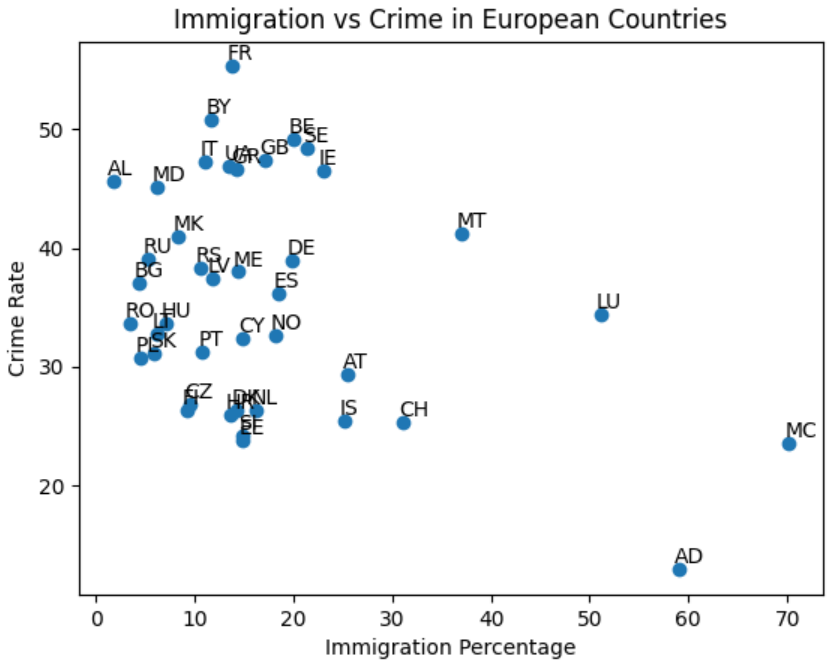
Immigration Outliers (based on IQR method):

	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.

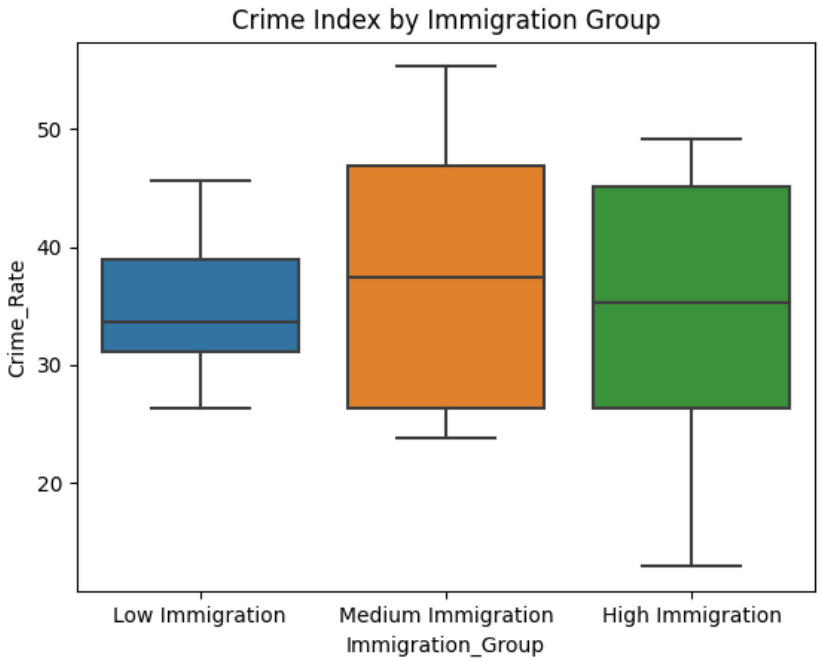
Spearman's rank correlation is measured as -0.174 indicating a weak negative monotonic relationship. P-values is measured as 0.276 which is above the significance level. Therefore, the correlation is not statistically significant, we cannot reject the null hypothesis.

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.

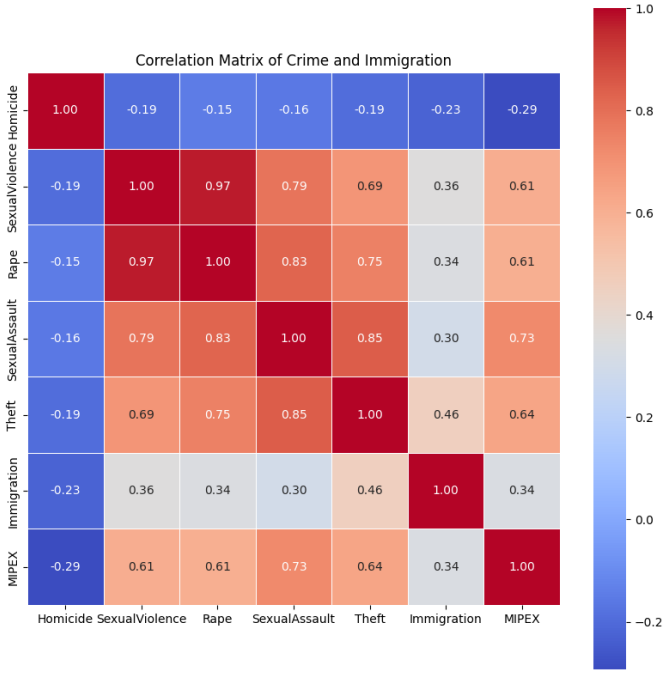
Since the most recent MIPEX data is from 2019, I aligned all other features—such as crime rates and immigration percentages—to that year to maintain consistency. Using 2019 data also avoids potential distortions caused by the COVID-19 pandemic, which significantly disrupted migration patterns and crime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

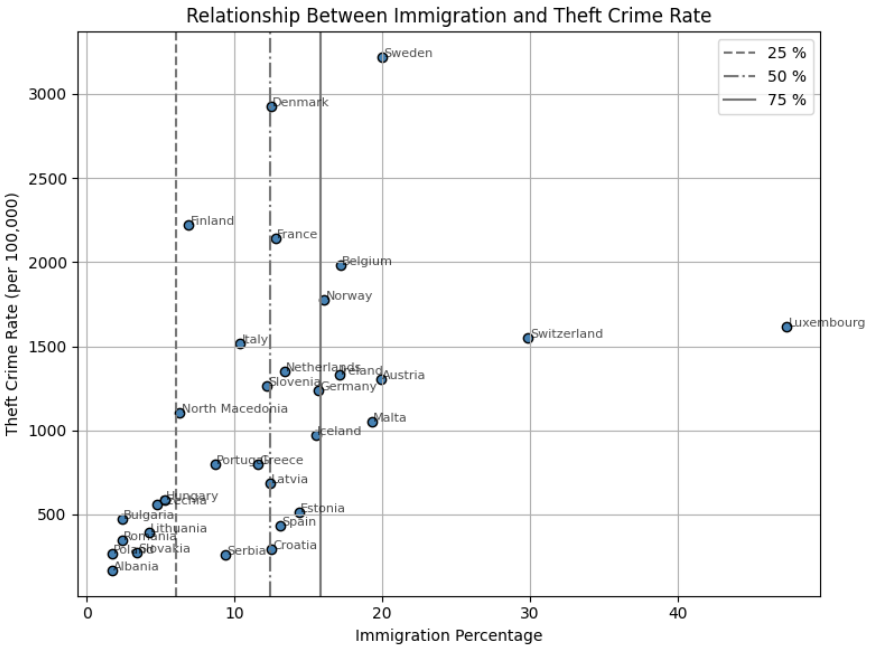
3. Exploratory Data Analysis



Observations:

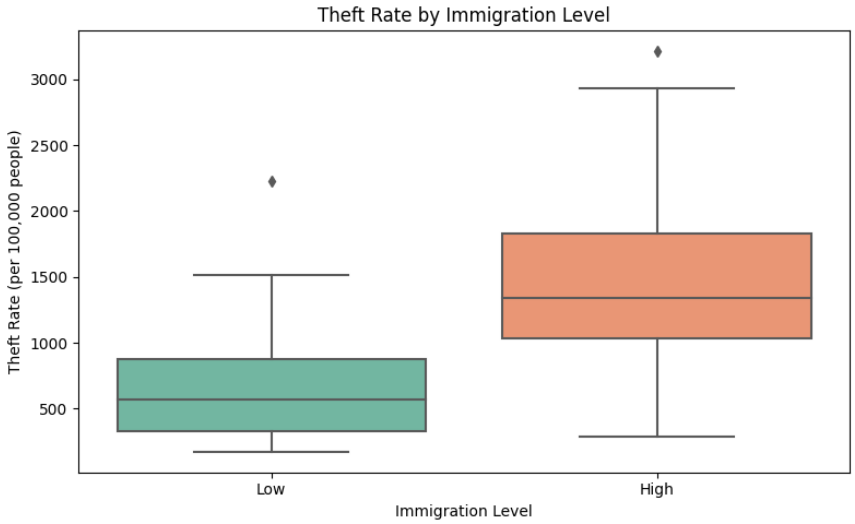
1. Rape, sexual assault and sexual violence show very strong positive correlations with each other. Rape and sexual assault are both subcategories of sexual violence under the International Classification of Crime for Statistical Purposes (ICCS) — they were included individually to preserve the level 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.