****

**BUCHAREST UNIVERSITY OF ECONOMIC STUDIES**

**CYBERNETICS, STATISTICS AND ECONOMIC INFORMATICS FACULTY**

SOFTWARE PACKAGES PROJECT

Emigration in the European Union: A Comprehensive Study

**Coordinator:**

**Conf. Univ. Dr. BELCIU Anda**

**Major:**

**Economic Informatics**

**Student: COSTAN CRISTIANA**

**Group: 1103**

**Bucharest**

**2024**

**CONTENTS**

[1. Data overview 3](#_Toc165496541)

[2. Data analysis 3](#_Toc165496542)

[Conclusions 16](#_Toc165496543)

[References 16](#_Toc165496544)

# Data overview

The objective of this project is to analyze migration patterns among European Union (EU) member states. In order to examine emigration within these countries, I chose three data sets: one that includes emigrants, one with various quality of life indicators and one that includes criminal offences. The data selected for this study was obtained from Eurostat[1] and represents values at the end of 2019. Choosing a recent year allowed me to make assumptions about the results easier, since I have some knowledge of the circumstances in different nations. The final model includes 13 variables, excluding the number of emigrants, chosen for studying emigration in various countries from the EU. Each variable represents a quality of life indicator, their abbreviations being displayed in table 1.1.

|  |  |  |
| --- | --- | --- |
| **No.** | **Label of the variable** | **Description of the variable** |
| 1 | AvgIncome | Average Income |
| 2 | PovRate | At-risk-of-poverty rate |
| 3 | GDP | Main GDP aggregates per capita |
| 4 | SVDRate | Severe material deprivation rate |
| 5 | OcwRate | Overcrowding rate |
| 6 | UnpRate | Unemployment rate |
| 7 | LifeExp | Life expectancy |
| 8 | UMN | Unmet medical needs rate |
| 9 | EarlyLeavers | Early leavers from education |
| 10 | Debtors | Arrears |
| 11 | GEmpGap | Gender employment gap |
| 12 | PGEP | Pollution, grime or other environmental problems |
| 13 | Offences | Offences |

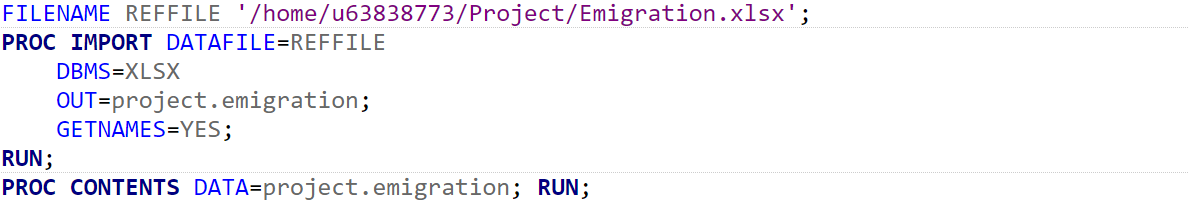
*Table 1.1. Variable labels and descriptions*

# Data analysis

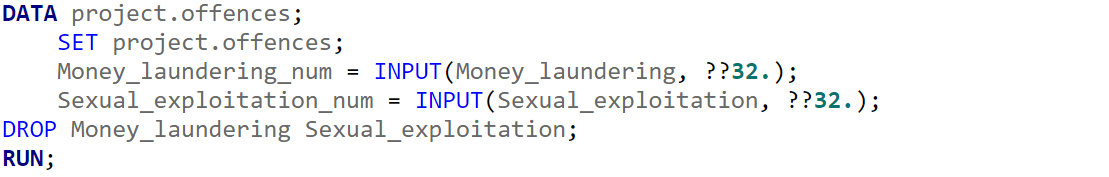
The initial step involved creating a library named "project" to ensure the persistence of data.



To procure a data set suited for analysis, I imported the three Excel files referenced in the preceding chapter into SAS, thereby generating three SAS data sets: *emigration*, *offences*, and *quality\_of\_life*.

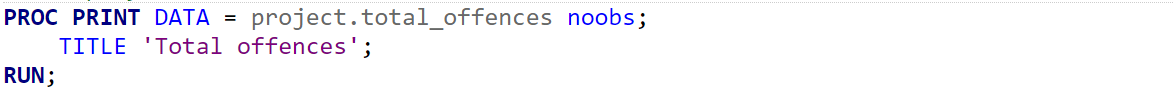


The variables "Money\_laundering" and "Sexual\_exploitation" from the imported XLSX file named *offences* were converted to numeric variables due to their initial import as character variables.



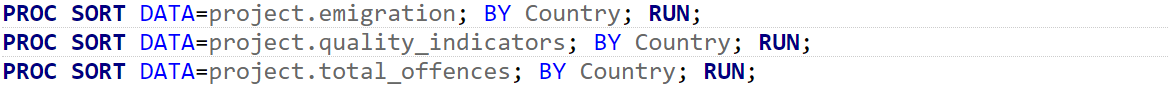
The *total\_offences* data set, which will be further utilized in the analysis, is the result of aggregating all offences from the previously imported and modified XLSX file.

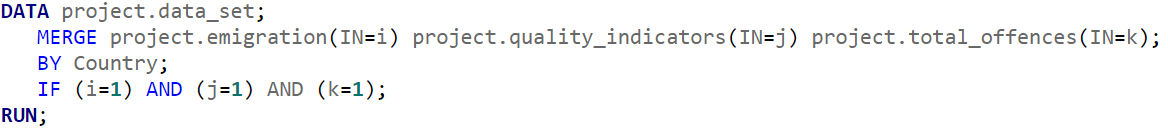




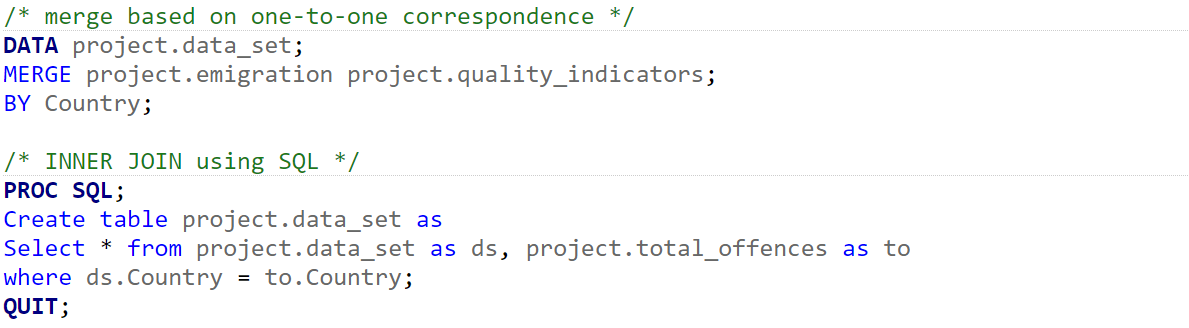


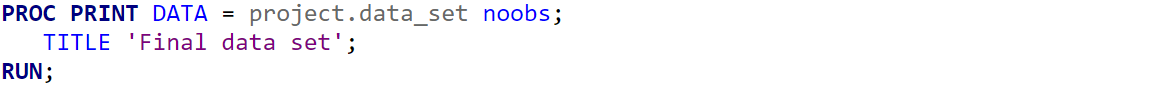
Since I have the same number of countries, hence the same number of observations, I will merge the three data sets into a single one, named *data\_set*. This can be accomplished by sorting the data sets and performing a match-merge, as illustrated in the figure below.

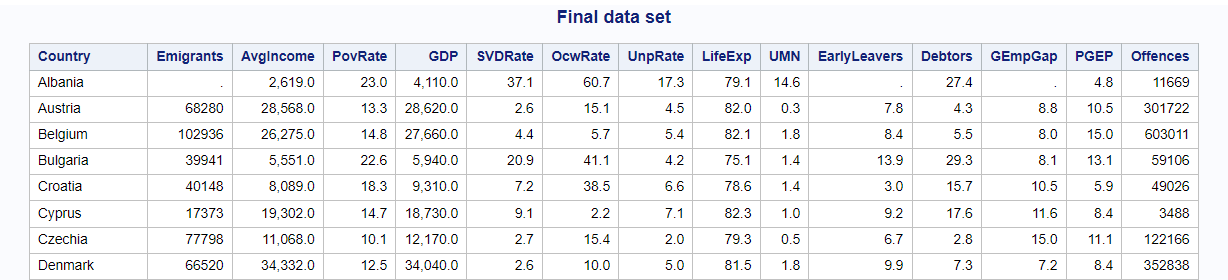




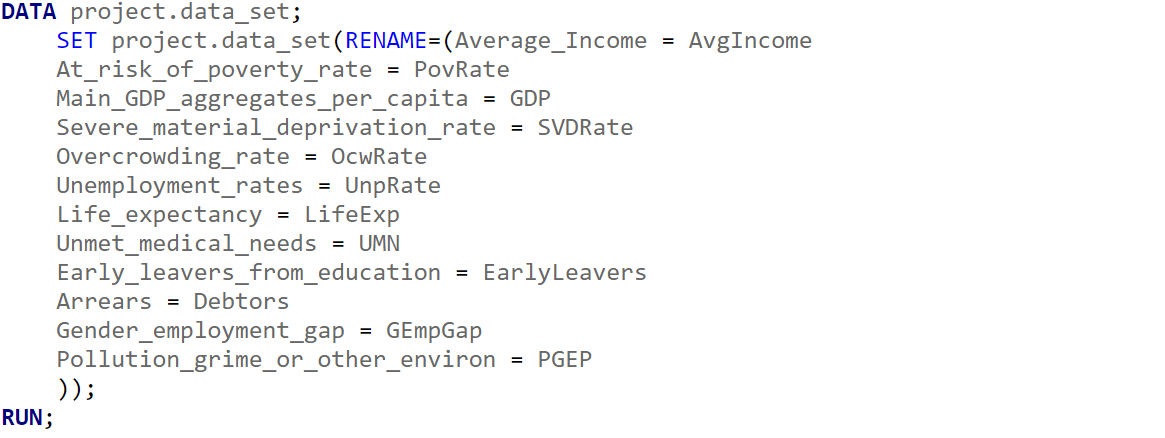
To merge the data sets separately, I can merge "emigration" and "quality\_indicators" based on a one-to-one correspondence. Then, I can use SQL to perform an inner join between the resulting data set and *total\_offences* based on the "Country" column.

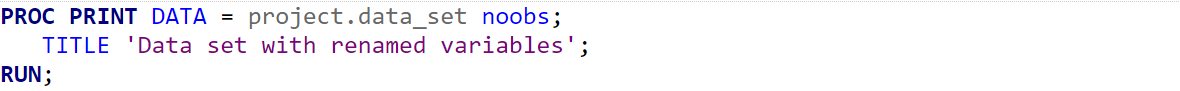


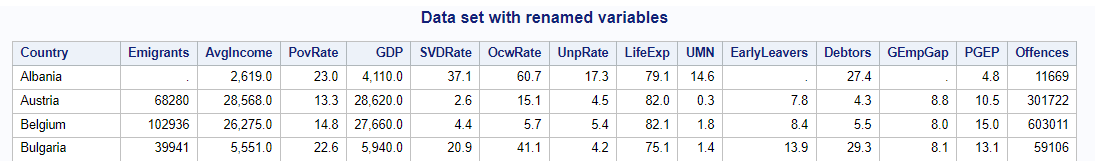




After merging the data sets into one, I renamed the variables to match the table from the previous chapter, facilitating easier usage of variable names.



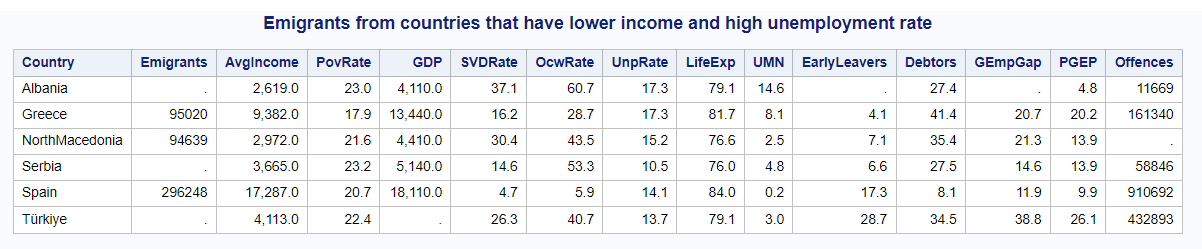
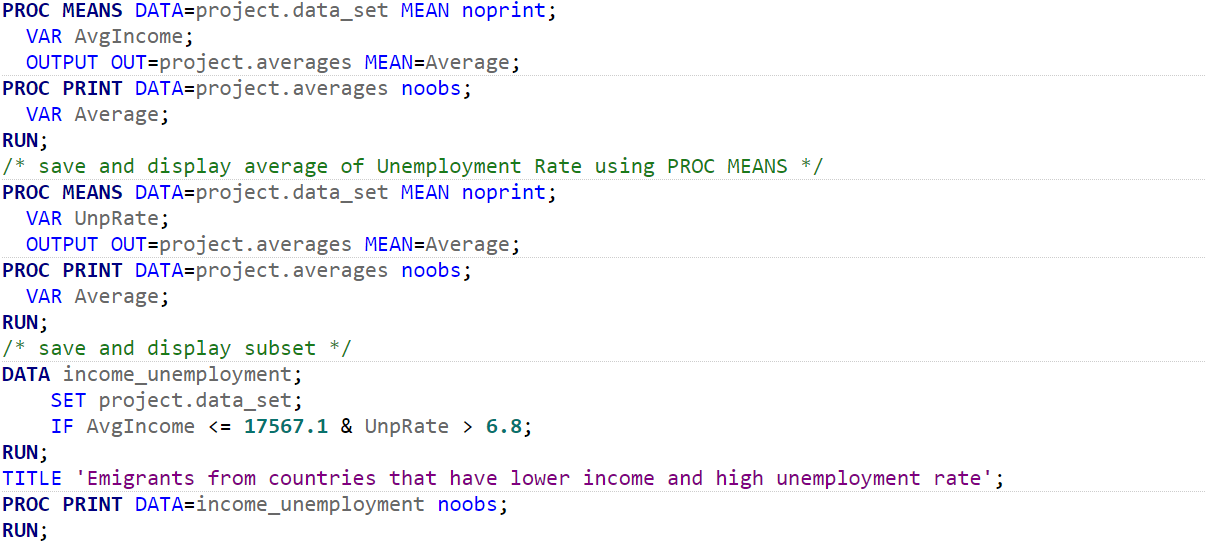




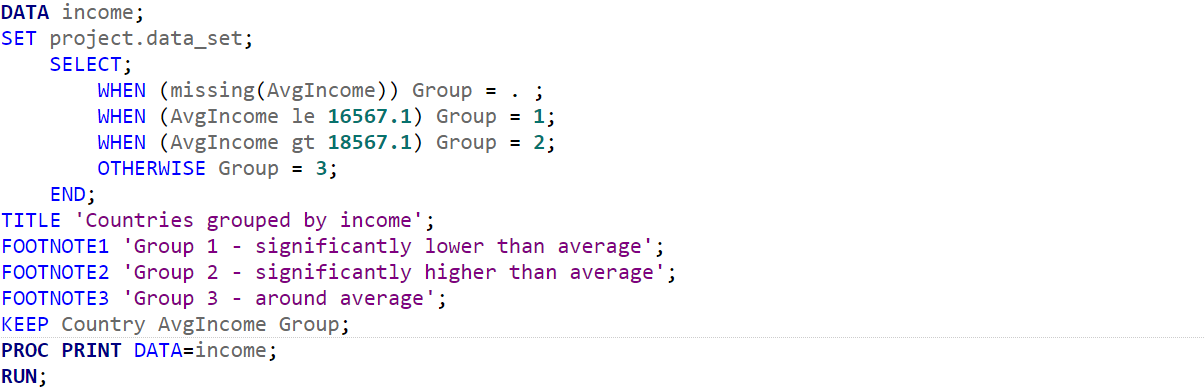
Moreover, I removed leading, between, and trailing spaces in the "Country" column.



The first aspect I aimed to analyze was the number of emigrants from countries characterized by lower income and high unemployment rates. I obtained these data by creating a subset where I specified the income to be lower than the average income and the unemployment rate higher than the average unemployment rate. The averages were calculated using **PROC MEANS**. Upon examining the results, it appears that Albania, Greece, North Macedonia, Serbia, Spain, and Turkey have lower income and high unemployment rates. Furthermore, it can be observed that Greece and North Macedonia exhibit a similar number of emigrants, which could be correlated with their high unemployment rates, substantial levels of debt, and notable gender employment gap.



I continued by grouping countries based on income, with explanations for each group provided in footnotes. Group 3 comprises the fewest number of countries, with only Malta and Spain having income levels around the average, while the remaining countries in the data set exhibit incomes either significantly lower or higher than the average.

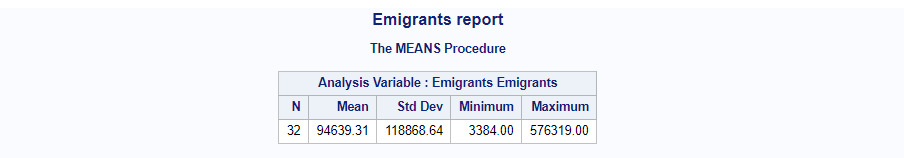




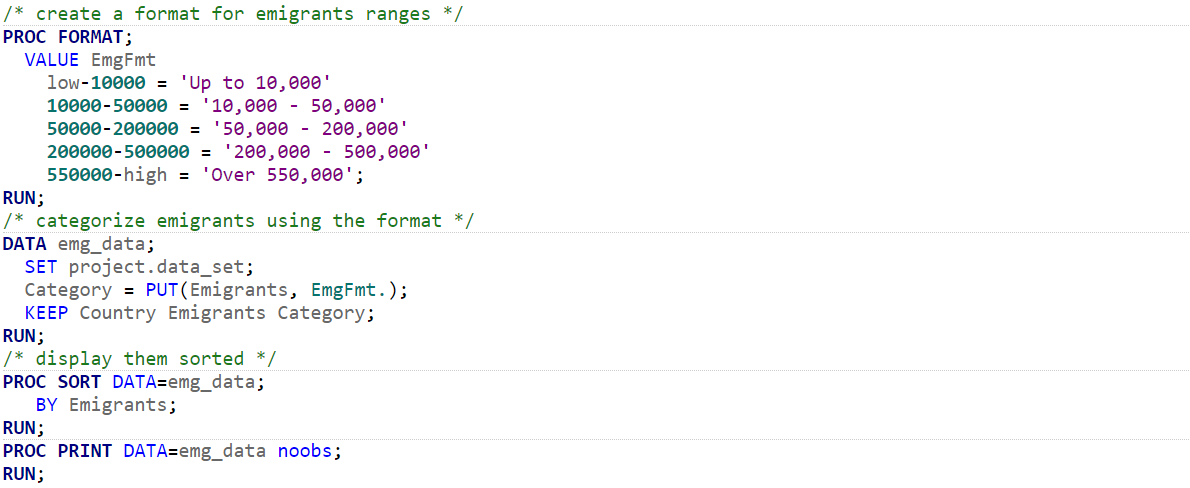


Furthermore, I generated aggregated reports using the **MEANS** procedure, where I analyzed various data about emigrants. The total number of observations is 32, with the minimum number of emigrants being 3,384.00 and the highest being 576,319.00.





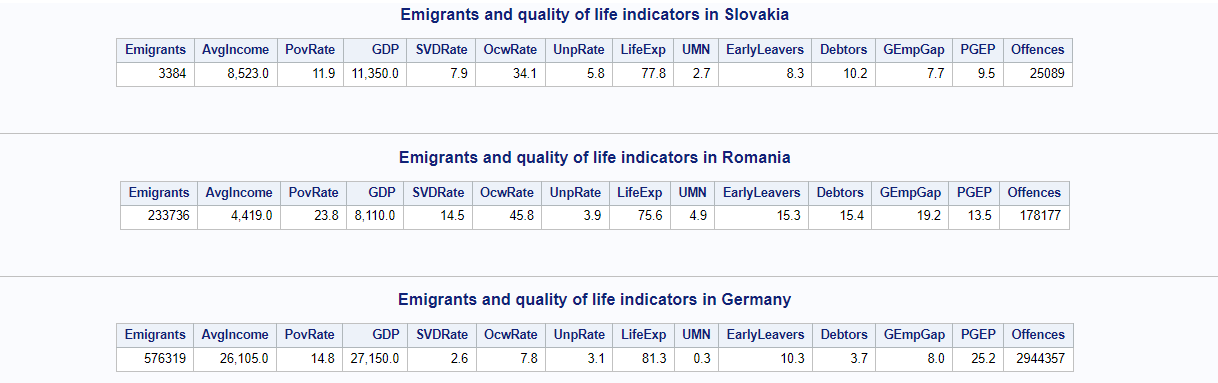
Following that, leveraging the maximum and minimum number of emigrants, I created a format for emigrant ranges and applied it to the *data\_set*, generating a new data set with categorized emigrants. Additionally, I sorted the resulting data set to identify which countries had the lowest and highest number of emigrants. The results indicated that Slovakia has the lowest number of emigrants, as previously identified, while Germany has the highest number of emigrants, being the sole country falling into the category "Over 550,000".



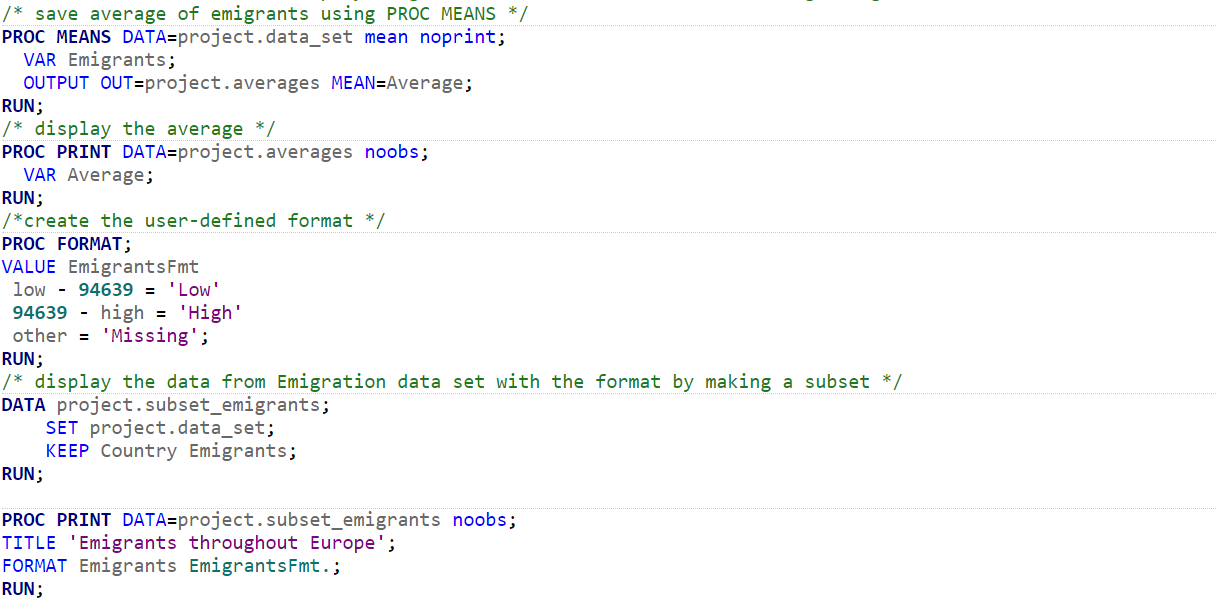


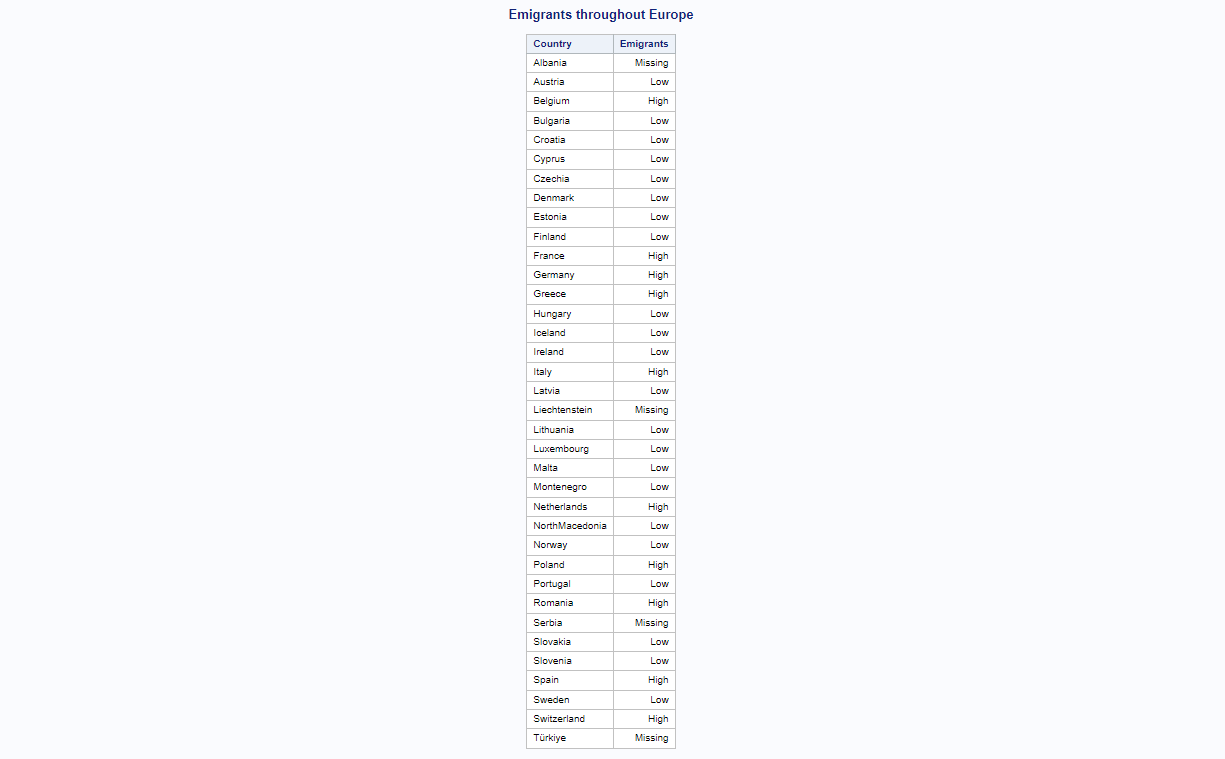
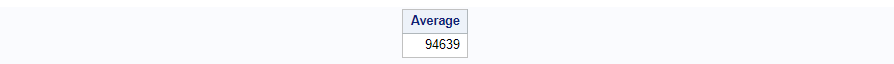


After that, I proceeded by generating three subsets: one containing data for Slovakia, which has the lowest number of emigrants, another for Germany, which has the highest number of emigrants, and a third for Romania, to compare with the identified countries. Compared to the first two countries, Romania has roughly half the number of emigrants, despite its unfavorable material conditions, with high income and poverty rates compared to the other two countries. From these statistics, it can be assumed that the number of offences significantly influences the number of emigrants, as it is a variable that gradually increases in each country with the number of emigrants.

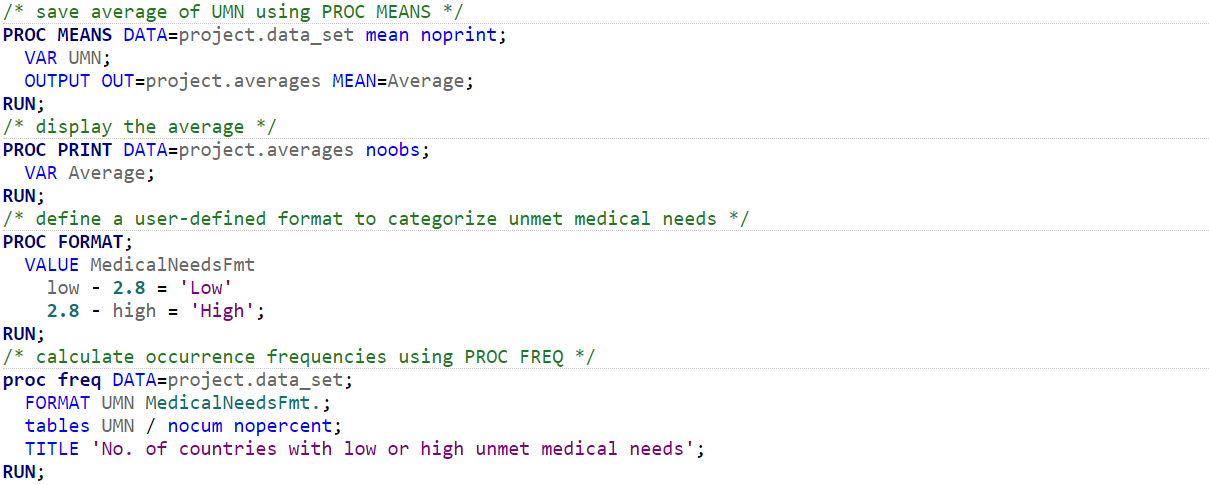


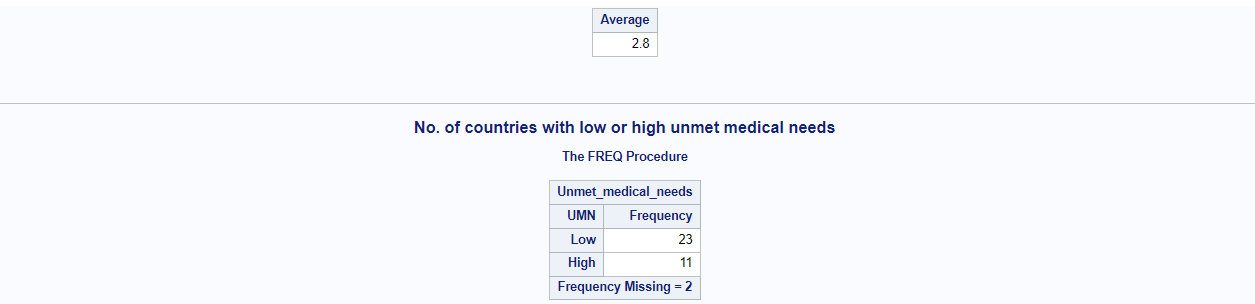
Then, I utilized a user-defined format to present emigrants from countries based on the average number of emigrants. I saved the average of emigrants using **PROC MEANS** to display it individually and then created the format based on the result. The resulting data was displayed in a subset created from the *data\_set*, retaining only the columns "Country" and "Emigrants". The majority of countries exhibited a small number of emigrants compared to the average.



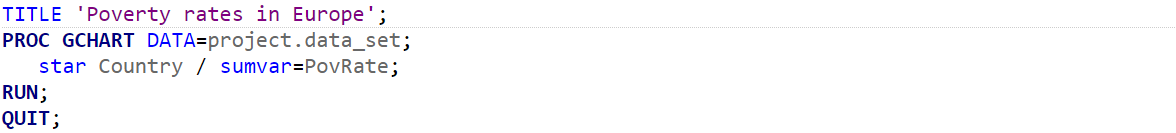


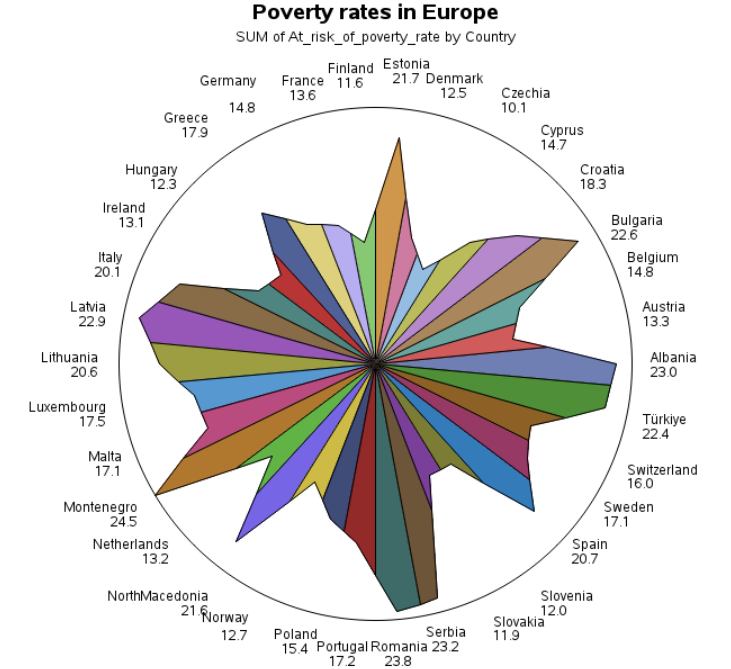
Similarly, I employed the average of "UMN" to determine the number of countries with low or high unmet medical needs. By utilizing the procedure for calculating occurrence frequencies and a user-defined format, I successfully displayed the countries. It was revealed that 23 countries have low unmet medical needs, while 11 countries have high unmet medical needs. Given that countries with low unmet medical needs predominate and, conversely, countries with a small number of emigrants prevail, it can be presumed that medical needs also contribute to emigration.



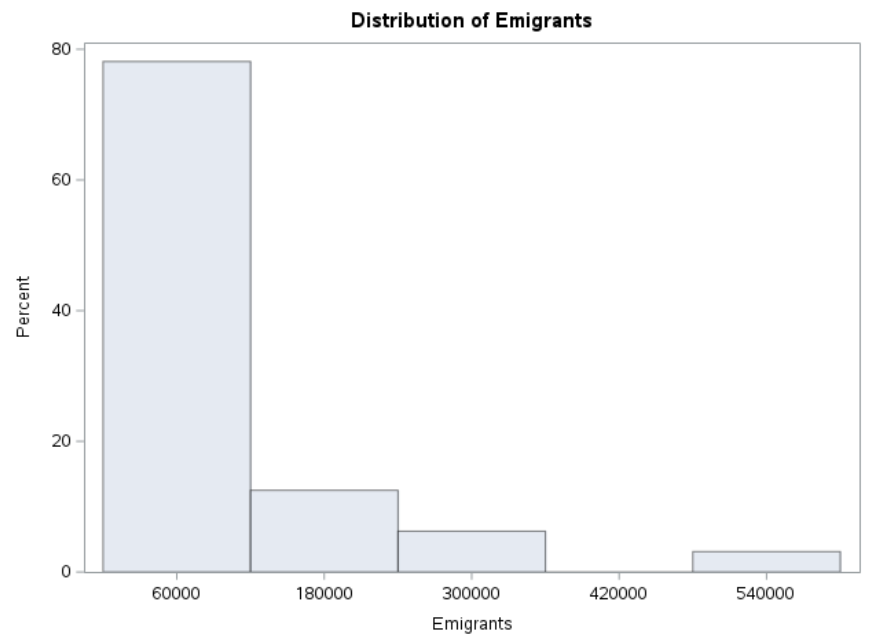
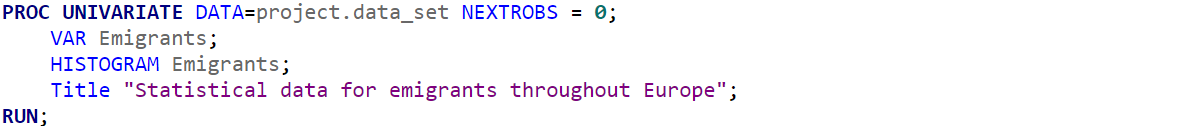


Furthermore, I displayed poverty rates throughout Europe in a star chart. As can be observed, Montenegro had the highest poverty rate (24.5), followed by Romania, Serbia, Albania, and Latvia.

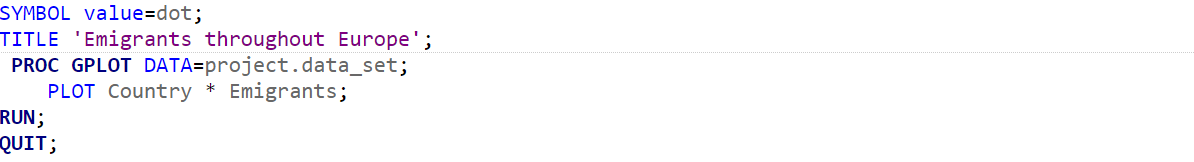


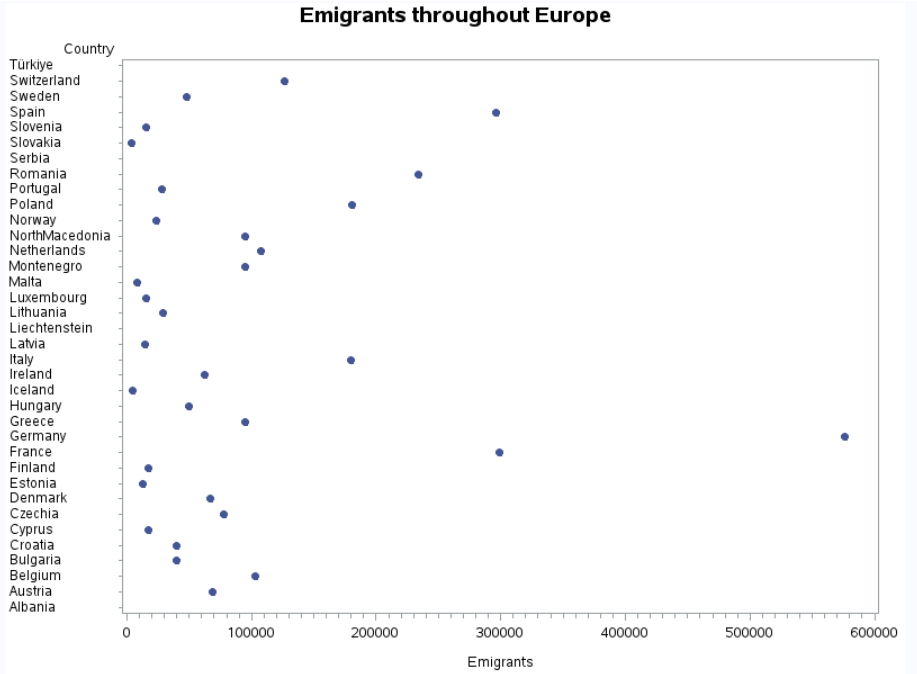


Afterwards, I displayed emigrants in a histogram along with descriptive statistics using **UNIVARIATE**. On average, there are approximately 94,639 emigrants. In 2019, a total of 3,028,458 emigrants were recorded in Europe. Additionally, the statistics included extreme values, comprising the minimum and maximum number of emigrants, alongside quantiles. The resulting histogram illustrates the distribution of emigrants, wherein we can observe that values around 540,000 are outliers. While most of the values fall between 0 and approximately 350,000, one value falls around 540,000, which is the outlier of the data set, lying far away from other data points.

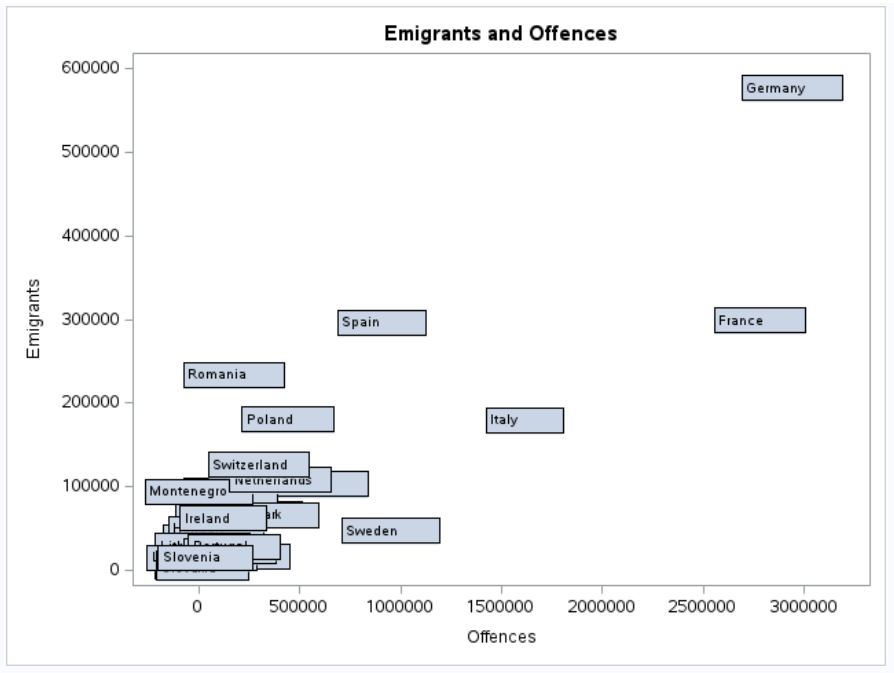
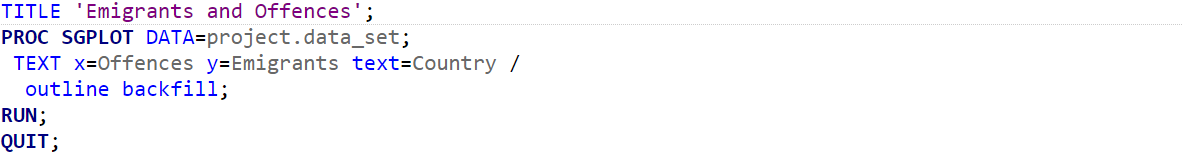


When I displayed emigrants in Europe using a scatter plot with dots, it became evident that only Germany had the highest number of emigrants (close to 600,000), followed by Spain and France.



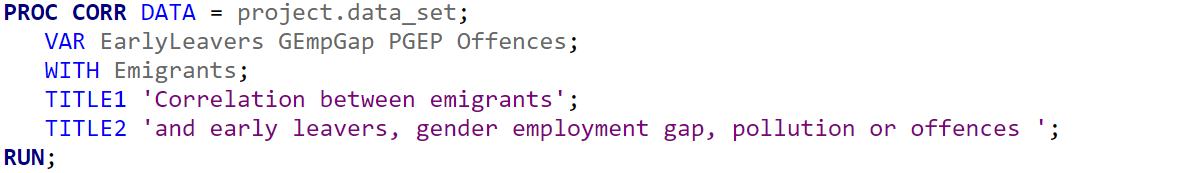


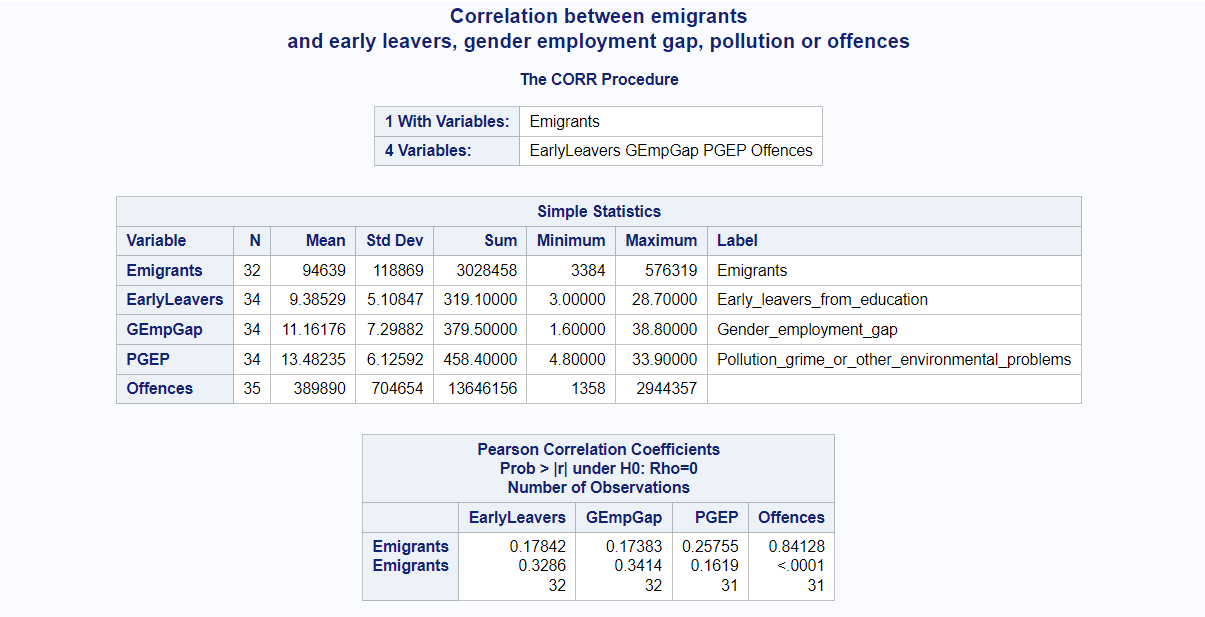
By displaying countries with their emigrants and offences using a text plot, I observed that Germany and France still lead the list, having a large number of emigrants and offences. However, this time Spain is only listed as having a large number of emigrants. Additionally, Italy stands out, being close to having a large number of offences.



From the correlation analysis between all the variables, which displays Pearson's correlation coefficient by default, I extracted "EarlyLeavers", "GEmpGap", "PGEP", and "Offences" to perform an additional, separate correlation, highlighting the high correlation coefficient between "Emigrants" and the selected variables. In the additional statistics, among other things, one can observe the sum, minimum, and maximum of each variable extracted for study. "Offences" exhibit the highest correlation with "Emigrants".

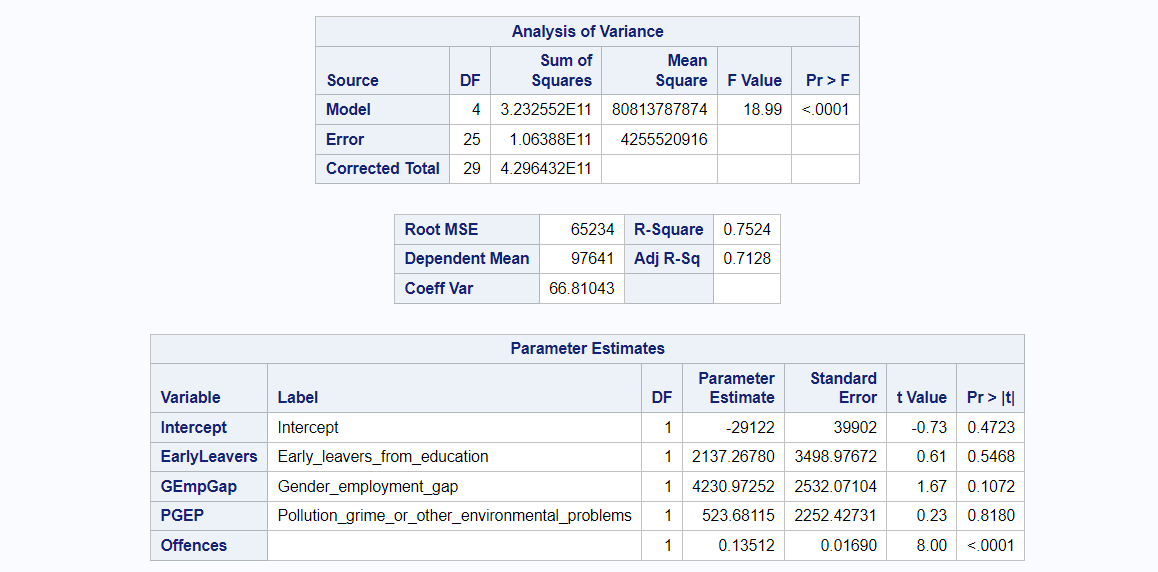






Lastly, I conducted a multiple linear regression (MLR) analysis to demonstrate how emigrants are impacted by "EarlyLeavers", "GEmpGap", "PGEP", and "Offences". With an R-squared value of 0.7524, it indicates that 75.24% of the variation in the number of emigrants is explained by the selected variables. The MLR analysis confirms that offences significantly influence the number of emigrants in the analyzed countries, with an increase in the number of offences by 1 unit corresponding to a rise in the number of emigrants by 8. Thus, high criminality is associated with higher levels of emigration.





# Conclusions

In summary, the comprehensive analysis undertaken in this study sheds light on the intricate relationship between various socio-economic factors and emigration patterns across European countries. Through data exploration, visualization, and statistical modeling, key insights have been uncovered. Moreover, the multiple linear regression analysis highlighted the strong influence of criminal offences on emigration, indicating that higher levels of criminality are associated with increased emigration rates. These findings underscore the multifaceted nature of migration dynamics and emphasize the importance of addressing socio-economic challenges to mitigate emigration pressures. Overall, this study contributes to a deeper understanding of the complex factors driving emigration.

# References

[1] <https://ec.europa.eu/eurostat/web/main/data/database>