Zeynep Norman 32227

DSA210: Term Project Final Report

Data Driven Analyze of Immigration and Nationalist Party Support in Europe

Research Question: "How does increasing immigration to Europe influence the rise of nationalist parties?"

1 Introduction

In recent years, Europe has experienced a marked political transformation, characterized by the growing prominence of nationalist parties. This trend has often been attributed to demographic shifts, particularly immigration, which has sparked political polarization and public debate. The 2015 refugee crisis and ongoing migration from non-EU countries have intensified scrutiny over the social and economic effects of immigration, with many analysts linking these shifts to changes in electoral behavior. However, despite the widespread attention this issue has received, empirical research investigating the direct relationship between immigration trends and the rise of nationalist parties remains limited, especially from a comparative, data-driven perspective across EU member states. Most existing studies tend to focus on individual countries or specific political moments, which limits the generalizability of their findings. Motivated by this gap, the present study aims to systematically examine whether recent immigration trends can help explain the changing vote shares of nationalist parties across the European Union. The project integrates data on migration, economic indicators, and electoral outcomes to build predictive models and evaluate the strength of these associations using both traditional statistical methods and machine learning techniques. By doing so, it contributes to a more evidence-based understanding of the drivers behind contemporary political shifts in Europe.

2 Collection and Preparation of Data

2.1 Immigration Data: *Annual crude rate of net migration for each country in the European Union*[1]

To measure immigration, I used annual crude rate of net migration to Europe ((number of immigrants – migrants) / population). I decided to use ratio values rather than actual numbers to eliminate biases such as global population increase that affects both number of immigrants and number of votes of each party. I used "Population change - Demographic balance and crude rates at national level" data provided by official site of EU, Eurostat. After accesing to annual crude rate of net migration to European countries, I mannually filtered the European Union countries and eliminated statistical abbrevations by using Excel tools on the spreadsheet.

2.2 Nationalism Data: *National Election vote shares for each EU country's nationalist party*[2]

To measure nationalism, I used each country's most recent and most popular nationalist party's vote shares from two recent elections. First I searched the most recent and most popular nationalist party of all 27 countries in the European Union. After adding the party abbrevations for each country to my spreadsheet, I listed the last national election that each party participated in along with the previous national election. Then I listed the vote shares of each party for both elections. When searching for parties

[1] https://ec.europa.eu/eurostat/databrowser/view/demo_gind/default/table?lang=en

[2] https://parlgov.fly.dev/

and vote shares, I used ParlGov website (database created for social research purpose) and official election authority websites of the respective countries, and I double-checked the data.

Attribute Descriptions:

prev_year: year of the national election prior to the most recent national election that each party participated in

prev_vote: vote share of the attributed party in the national election prior to the most recent national election

last_year: year of the last (most recent) national election that each party participated in

last_vote: vote share of the attributed party in the most recent national election

2.3 Real GDP per Capita Data: *Annual GDP per capita in purchasing power standards (PPS) for each EU country*[3]

To measure economic prosperity, I used the annual GDP per capita in PPS units for each European Union country. This metric provides a standardized way to compare individual wealth and purchasing power across nations, eliminating currency-related distortions. I chose GDP per capita rather than total GDP to better capture average living standards, which are more relevant when examining voter behavior. I retrieved the data from Eurostat's "GDP per capita in PPS" table. After accessing the annual figures, I manually filtered out only the EU member states and removed aggregated values or footnote codes that were not needed. This cleaned dataset allowed me to calculate the annual trend in economic growth for each country and use it as a control variable in my model, helping isolate the specific effect of immigration from broader economic conditions.

2.4 Total Unemployment Rate Data: *Annual unemployment rates for the total population (aged 15–74) in each EU country*[4]

To account for labor market conditions, I collected each country's annual unemployment rate from Eurostat's database. I selected the total population unemployment rate (ages 15–74) because it reflects the overall health of a country's job market and is likely to influence political preferences. High or rising unemployment could increase public dissatisfaction and drive support for populist or nationalist movements, making it an important control variable. I accessed this data from Eurostat's unemployment statistics and manually filtered the dataset to include only EU countries, excluding non-EU entries and metadata rows. After cleaning the spreadsheet, I calculated the yearly change for each country's unemployment rate, which I used in the regression model to help adjust for economic anxiety or insecurity when evaluating the relationship between immigration and nationalist vote shares.

3 Explanatory Data Analysis

3.1 Descriptive Statistics

The analysis began with basic descriptive statistics on the vote share dataset. This helped summarize the central tendencies (mean, median) and variability (standard deviation) of vote shares in both the previous and most recent elections. These statistics provided an initial look into how much vote shares have shifted and whether there was a general increase in nationalist sentiment across countries.

[3] https://ec.europa.eu/eurostat/databrowser/view/tipsna40/default/table?lang=en&category=tips.tipsgd
[4] https://ec.europa.eu/eurostat/databrowser/view/tps00203/default/table?lang=en&category=t_labour.t_employ.t_lfsi.t_une

3.2 Data Preparation and Transformation

For proper time-series and comparative analysis, the immigration dataset was transformed from a wide format to a long format, enabling easier plotting and year-wise calculations. The vote share dataset was also enhanced by calculating a new variable—vote share change—representing the difference between the most recent and previous elections. This was done for each country, giving a clearer picture of which nations experienced growth or decline in nationalist party support. To quantify immigration change over time, the average yearly trend was calculated per country. This value serves as a proxy for how rapidly immigration has increased or decreased in recent years.

3.1 Exploratory Visualizations

3.1.1 Histograms

A set of histograms was created to explore the distribution of both immigration rates and vote share changes. The histogram for immigration rates showed that while most countries had moderate values, a few countries experienced either very high or very low migration trends, indicating significant demographic movement. The histogram for vote share change showed that some countries had notable increases in nationalist support, while others saw declines or stability.

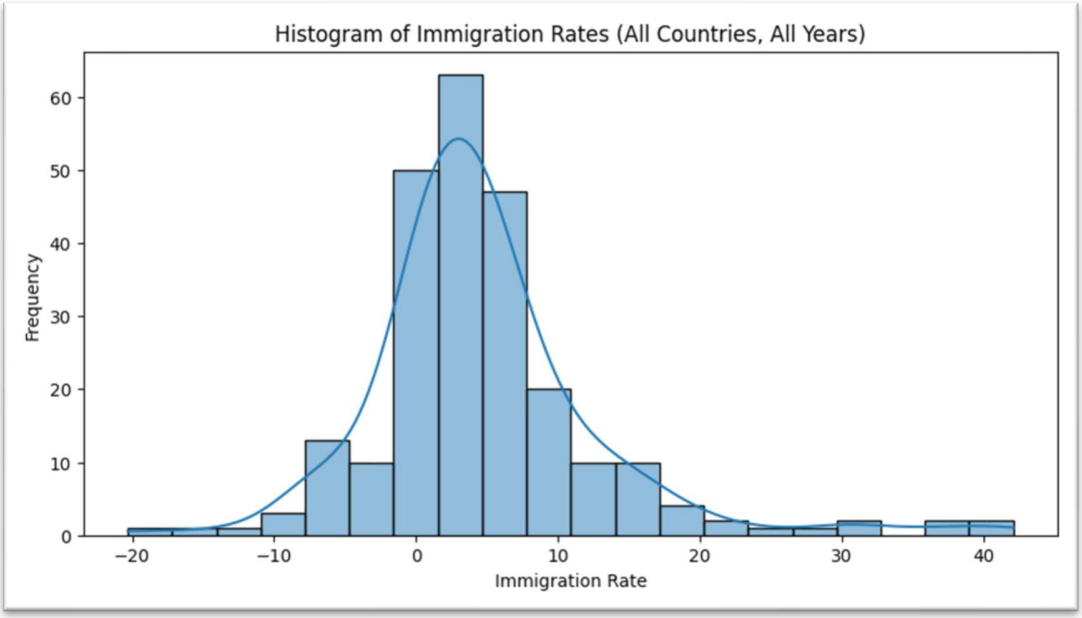

Figure 1

Figure 1 shows that most countries cluster around modest immigration rates between 0 and 10. Some countries exhibit negative values, indicating more emigration than immigration. A few countries display unusually high positive rates, reflecting strong inbound migration.

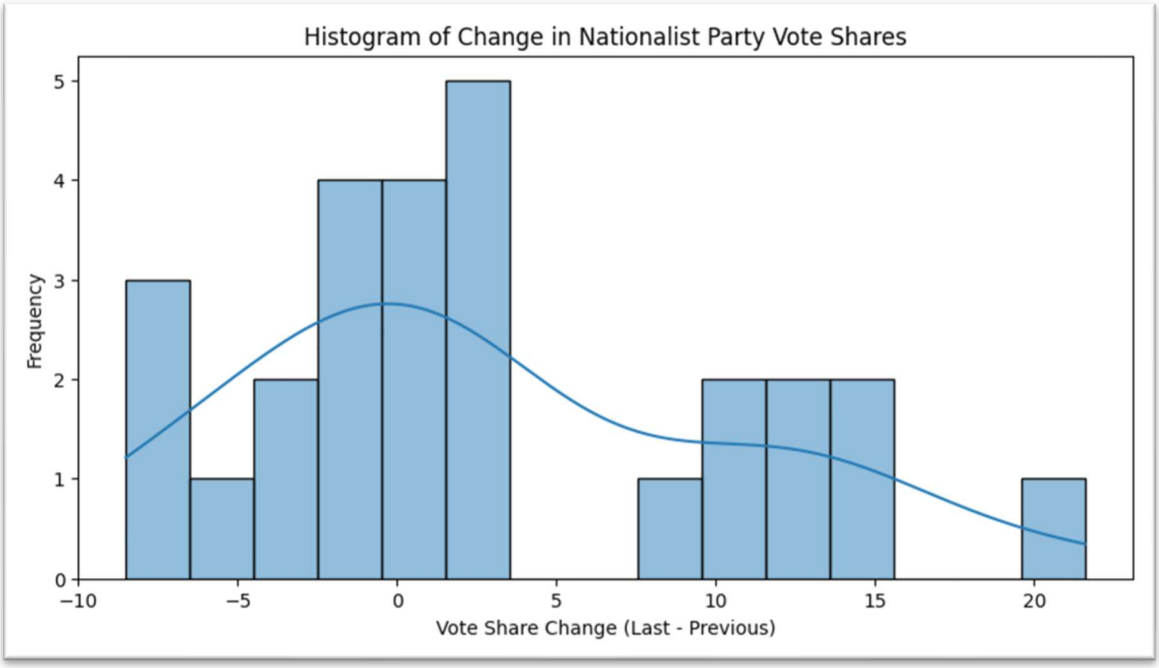

Figure 2

Figure 2 shows that many countries experienced modest increases or stability in nationalist party vote shares. Several countries exhibited notable spikes in support, while a few showed declines, indicating considerable variation across Europe.

3.1.2 Time Series

The time-series analysis was crucial for contextualizing the hypothesis. It helped visually identify countries where both immigration and nationalist votes increased in parallel. Although time-series data alone doesn't prove causality, observing these parallel trends in multiple cases supports the rationale behind further hypothesis testing. For example, in countries like Hungary or Poland, visual analysis showed a concurrent rise in both variables, hinting at a possible relationship that could be statistically tested. Time-series plots not only revealed the temporal dynamics of immigration and political sentiment but also guided which countries might be driving the overall trends observed in correlation and regression analysis.

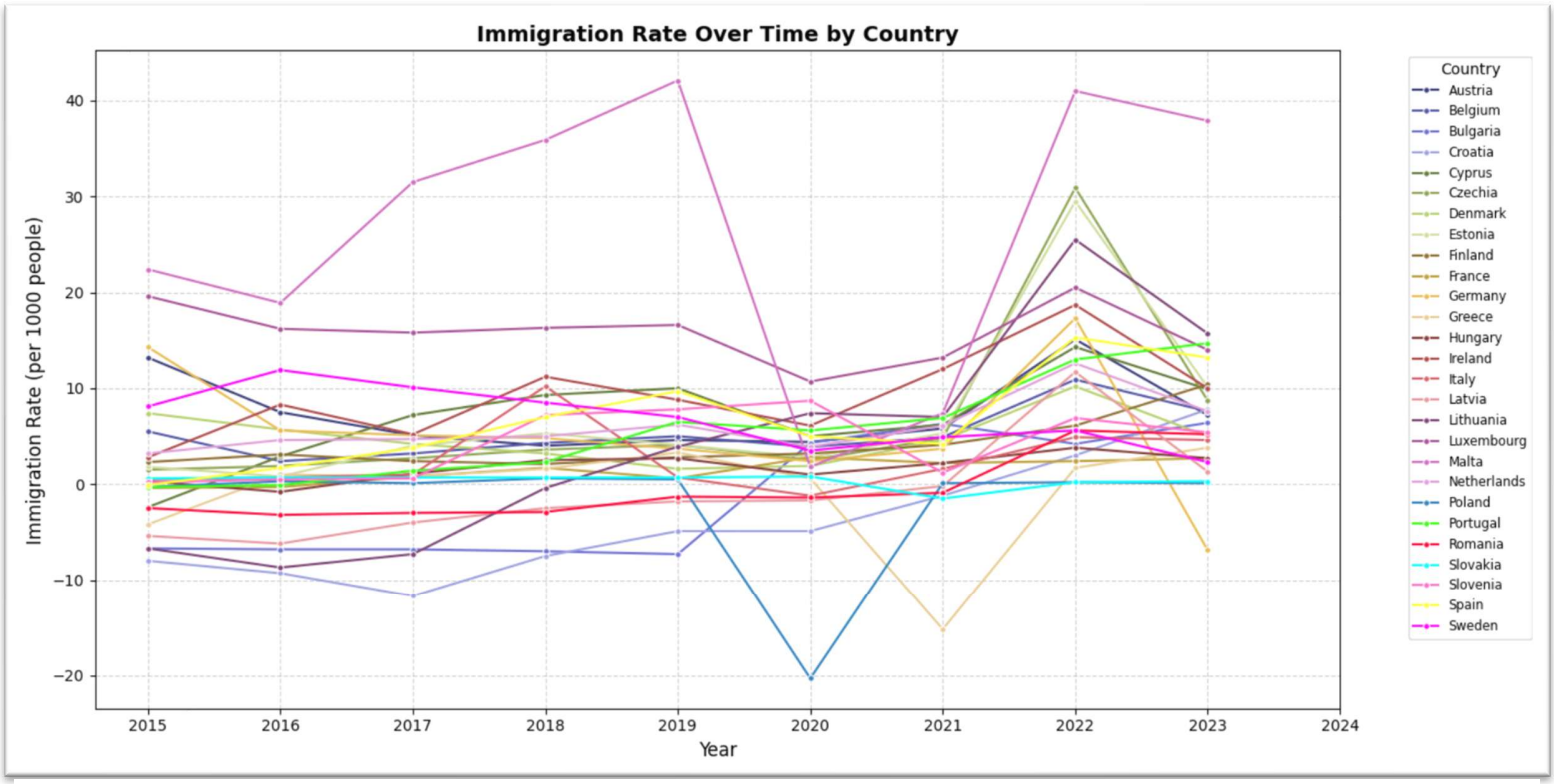

Figure 3

In figure 3, the time series graph reveals distinct immigration trends across EU countries between 2015 and 2023. While some countries like Germany and Luxembourg show steady growth in immigration, others such as Romania and Bulgaria display more negative or unstable patterns. A sharp regional increase around 2021–2022 is visible across many countries, likely reflecting broader geopolitical or post-pandemic shifts. These patterns highlight both national differences and shared regional events influencing migration dynamics.

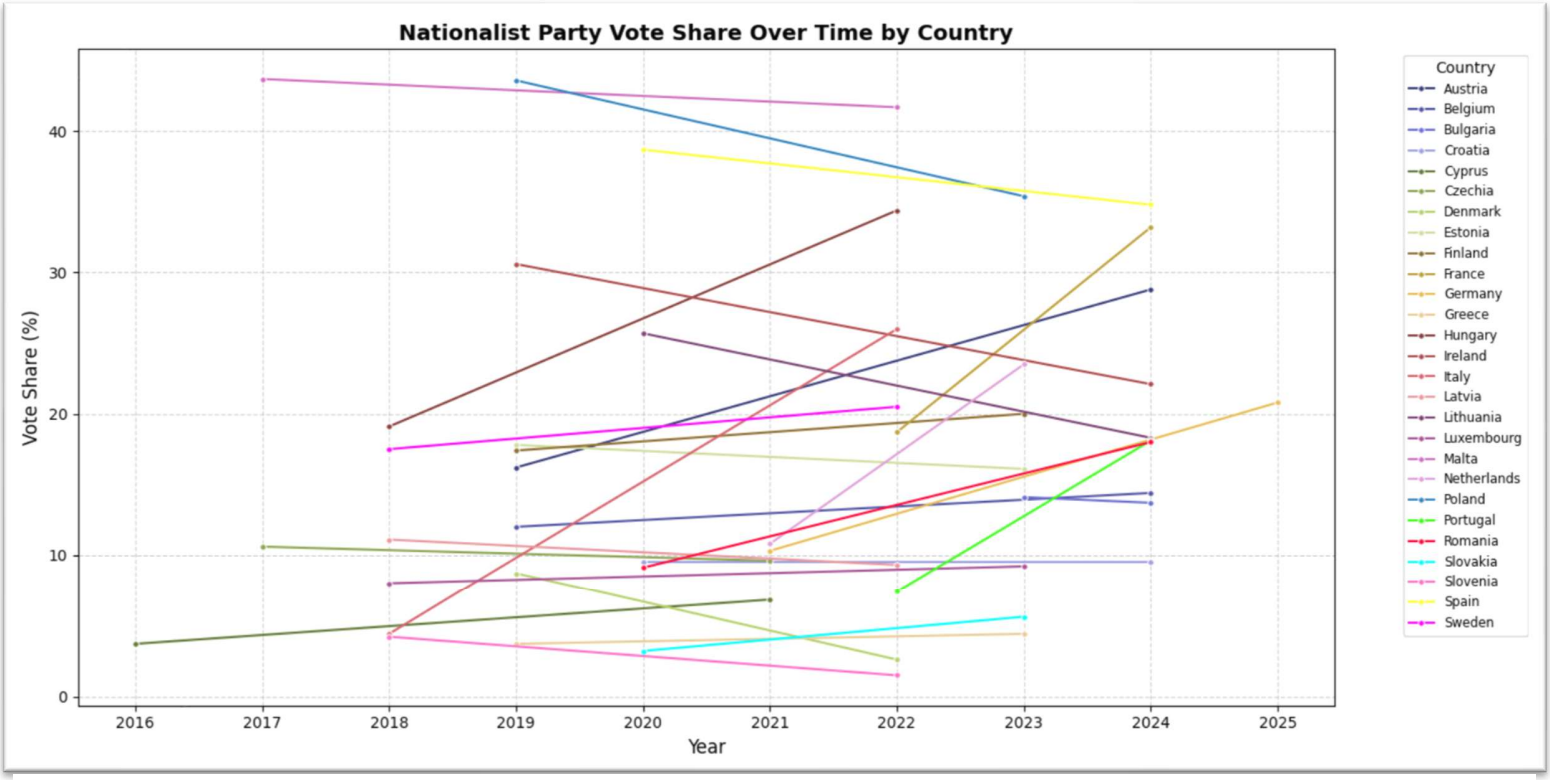

Figure 4

Figure 4 shows that nationalist party support has risen in many EU countries over the past election cycle. Countries like Poland and Hungary exhibit sharp increases in vote share, aligning with expectations about rising nationalist sentiment. In contrast, nations such as Sweden and Malta show little to no change or slight declines, indicating varied political trajectories across the region.

3.1.3 Scatterplots

Scatterplots were used to visually assess the relationship between the average immigration trend and the change in nationalist party vote shares. Each point represented a country, with its horizontal position determined by the immigration trend and the vertical position by vote share change. The distribution of points suggested a potential—but not visually overwhelming—pattern of positive association. To enhance interpretability, the scatterplot was supplemented with a trend line and marginal histograms in a joint plot. This allowed for a more holistic look at the spread and potential linear relationship between the variables.

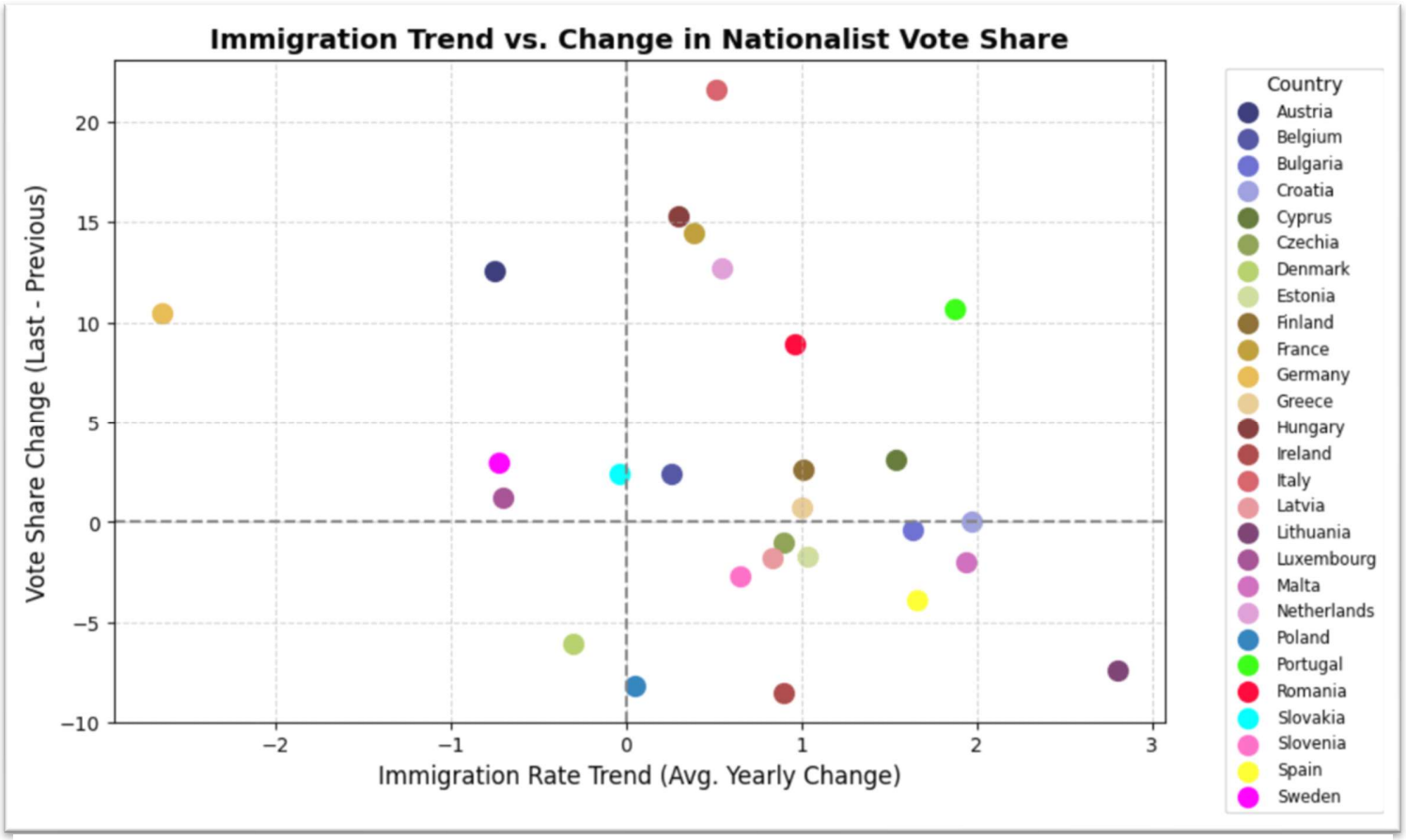

Figure 5

The scatterplot in Figure 5 reveals a wide dispersion of countries, with no strong linear relationship between immigration trends and changes in nationalist party vote shares. While countries like Hungary and Poland show both high immigration and increased nationalist support, many others with rising immigration display stable or even declining vote share change. The trend line's shallow slope and the clustered spread of data points suggest that immigration trends alone are not a reliable predictor of nationalist political shifts.

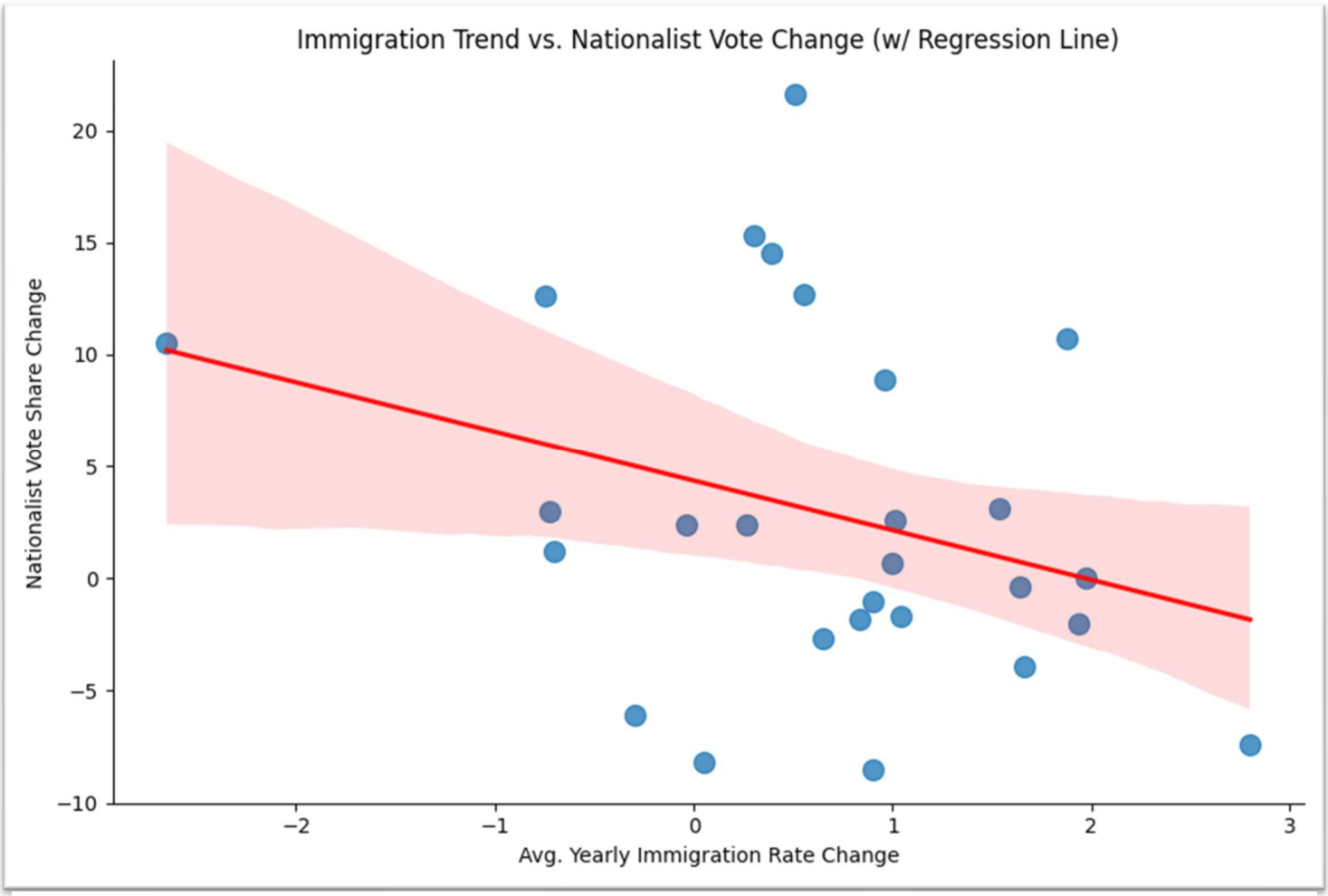

Figure 6

The scatterplot with regression line shows a slight negative association between immigration rate change and nationalist vote share change. However, the broad confidence interval surrounding the line suggests high variability and limited predictive strength. The visualization supports the idea that while there may be a weak trend, it is not strong or consistent enough to imply a direct causal link.

3.1.5 Bar-line chart

To overcome the complexity and clutter of traditional line plots with overlapping lines, more focused visuals were created. A grouped bar-line chart was developed where immigration trends were shown as bars and vote changes were shown as a line chart overlaid on the same x-axis of countries. This dual representation helped contextualize both metrics simultaneously for each country.

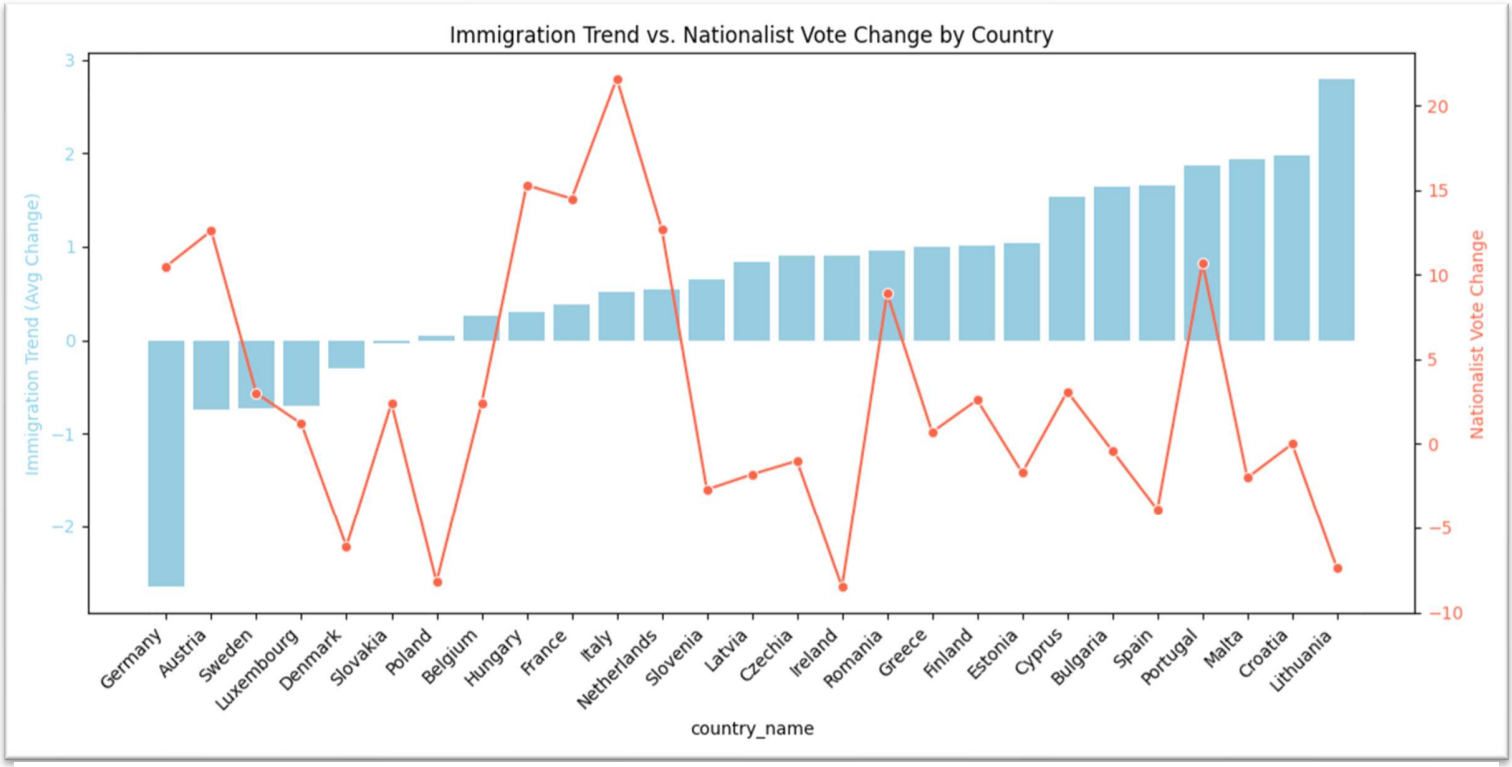

Figure 7

The grouped bar-line chart visually compares average yearly immigration trends (bars) and nationalist vote share changes (lines) across countries. Italy, France, and Hungary demonstrate parallel increases in both dimensions, supporting the initial hypothesis. Conversely, Germany and Sweden exhibit high immigration with decreasing vote shares, suggesting an inverse relationship. The dual-axis format clearly highlights inconsistencies and country-specific deviations.

4 Hypothesis Testing

Hypothesis testing was employed in this project to systematically assess whether a statistical relationship exists between rising immigration levels and shifts in nationalist party support across European Union countries. With the inclusion of inclusion of p-values, confidence intervals, and R^2 in Pearson correlation, simple linear regression, and multiple regression, the analysis aimed to quantify the strength and direction of this relationship and determine whether the observed trends could be attributed to more than random variation. The first two tests served as an initial inferential step before applying more complex modeling techniques like complex multivariate OLS (Section 4.3) and machine learning models. This step was crucial in determining whether further modeling and analysis were justified. The hypothesis was constructed as follows:

Null Hypothesis (H_0): There is no relationship between recent increasing immigration to EU countries and increasing vote shares of nationalist parties in those countries.

Alternative Hypothesis (H_1): There is a causal relationship between recent increasing immigration to EU countries and increasing vote shares of nationalist parties in those countries.

4.1 Pearson Correlation Analysis Interpretation

The Pearson correlation test was used to assess the linear association between immigration trends and changes in nationalist party vote shares across EU countries. This method is a fundamental statistical tool for quantifying the strength and direction of linear relationships between two continuous variables. It provides an initial overview of whether a potential connection exists without assuming any specific causal mechanism. Pearson correlation is especially useful in exploratory analysis as it requires minimal assumptions and offers interpretable results through its correlation coefficient and p-value. Its inclusion here serves to preliminarily validate whether more complex modeling is justified.

The analysis produced a correlation coefficient of -0.31 and a p-value of 0.116, suggesting a weak negative relationship between immigration trends and nationalist vote share change. However, the p-value exceeds the conventional 0.05 threshold for statistical significance, meaning the observed association is not robust enough to reject the null hypothesis. In practical terms, this result implies insufficient evidence to conclude that increases in immigration are associated with systematic shifts in nationalist party support. Thus, immigration trends alone cannot be reliably used to predict changes in nationalist sentiment.

4.2 Single Linear Regression Model Analysis Interpretation

The single linear regression model was employed to assess the effect of immigration trends on nationalist vote share change. This method quantifies both the direction and strength of a hypothesized linear relationship using a single predictor—in this case, the average annual change in immigration rate. It also provides a measure of model fit through the R^2 value, indicating how much of the variation in vote change is explained by immigration alone. This approach builds upon the earlier correlation analysis by enabling formal statistical inference, including significance testing.

The model yielded a coefficient of -2.21, suggesting a negative relationship, an R^2 of 0.096, and a p-value of 0.116. These indicate that while the direction supports the hypothesis that increased immigration might reduce nationalist support, the result is statistically insignificant and explains less than 10% of the variance in vote share change. Below, Figure 8 illustrates these findings: the predicted values do not align closely with actual outcomes, and residuals are widely dispersed, confirming weak model performance and the insufficiency of immigration as a standalone predictor.

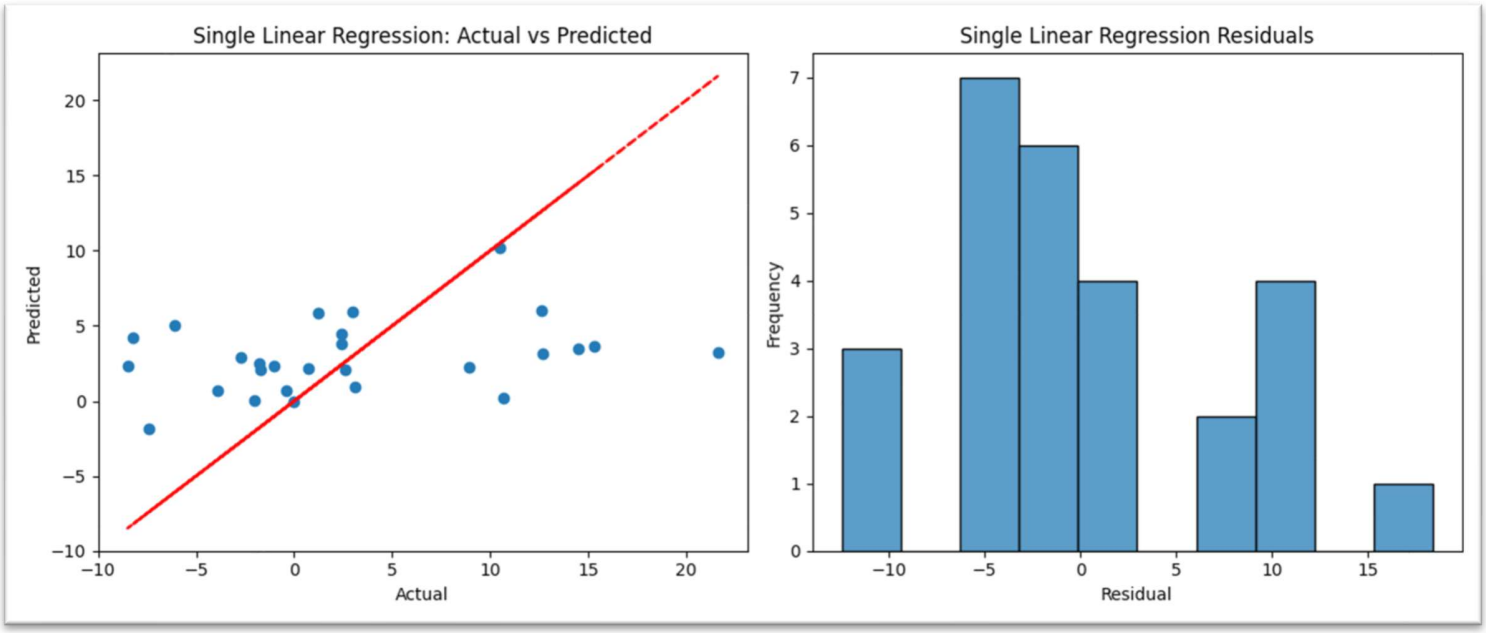

Figure 8

4.3 Multiple Linear Regression Model Analysis and Interpretation

Multiple linear regression was used to extend the single-variable analysis by incorporating GDP per capita trend and unemployment rate trend as control variables alongside immigration trend. This approach aimed to account for potential confounding factors and better isolate the unique effect of immigration on changes in nationalist vote share. The decision to include these controls was based on earlier results, which showed that immigration alone had weak explanatory power and raised concerns about omitted variable bias. By adjusting for economic performance indicators, the model offers a more nuanced understanding of how each predictor may contribute to shifts in political sentiment.

The model yielded coefficients of -2.01 for immigration trend, -0.0052 for GDP per capita trend, and -1.5975 for unemployment rate trend. The R^2 value increased to 0.206, indicating that approximately 20.6% of the variance in nationalist vote share change was explained by the predictors—an improvement from the 0.096 R^2 seen in the single-variable model. However, none of the predictors reached statistical significance, suggesting that their individual effects remain inconclusive. As seen in Figure 9, the predicted values from the multiple regression align more closely with actual outcomes compared to the single regression shown in Figure 8, and the residuals are slightly more centralized, indicating reduced prediction error. These results imply that while adding economic variables enhances model performance, significant unexplained variation remains, pointing to the influence of other unobserved factors on nationalist voting trends.

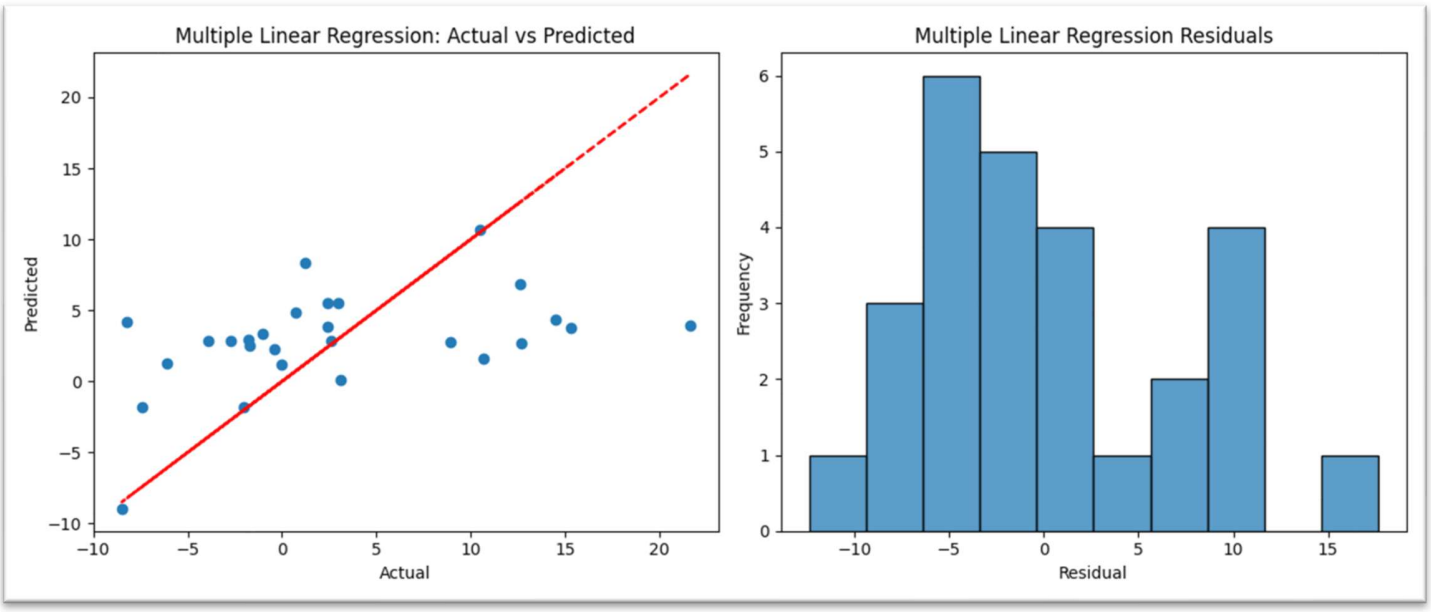

Figure 9

5 Machine Learning Applications

To further explore the relationship between immigration trends and changes in nationalist vote share, two regularized linear models were implemented: Ridge Regression and Lasso Regression. These models help mitigate issues of multicollinearity—especially relevant given the likely correlation between GDP and unemployment—and improve model generalizability by penalizing excessive coefficient values. All explanatory variables (immigration trend, GDP per capita trend, and unemployment rate trend) were standardized to ensure equal treatment in regularization. Cross-validation was used to select the optimal regularization strength (alpha) for each model.

5.1 Regularized Linear Regressions

Regularized linear regression models, such as Ridge and Lasso, are essential tools for addressing challenges related to multicollinearity and overfitting, particularly in small to moderately sized datasets with interdependent features. Unlike ordinary least squares regression, which can produce unstable coefficient estimates when predictors are correlated, these methods apply regularization penalties to constrain the magnitude of coefficients. Ridge regression (L2 regularization) reduces variance by shrinking coefficients uniformly, while Lasso regression (L1 regularization) can drive some coefficients to zero, effectively performing variable selection. These properties enhance model interpretability and generalizability, allowing researchers to better identify the most influential predictors in complex social and economic phenomena.

5.1.1 Ridge Regression Model Analysis and Interpretation

The Ridge regression model, with an optimal alpha of 10.0, explained 19.1% of the variance in nationalist vote share change ($R^2 = 0.1909$). All predictors had negative coefficients, with immigration and GDP trends showing stronger inverse effects than unemployment, though overall influence remained weak. This indicates a weak inverse association between these factors and rising nationalist sentiment. However, the model's overall explanatory capacity remains modest, suggesting that these variables do not adequately account for shifts in nationalist voting alone.In figure 10, the actual vs. predicted plot shows a scattered distribution around the regression line, reflecting limited predictive accuracy. The residuals histogram indicates modest clustering but notable error spread, confirming that Ridge regression, while stabilizing coefficients, offers only limited explanatory power for the observed voting shifts.

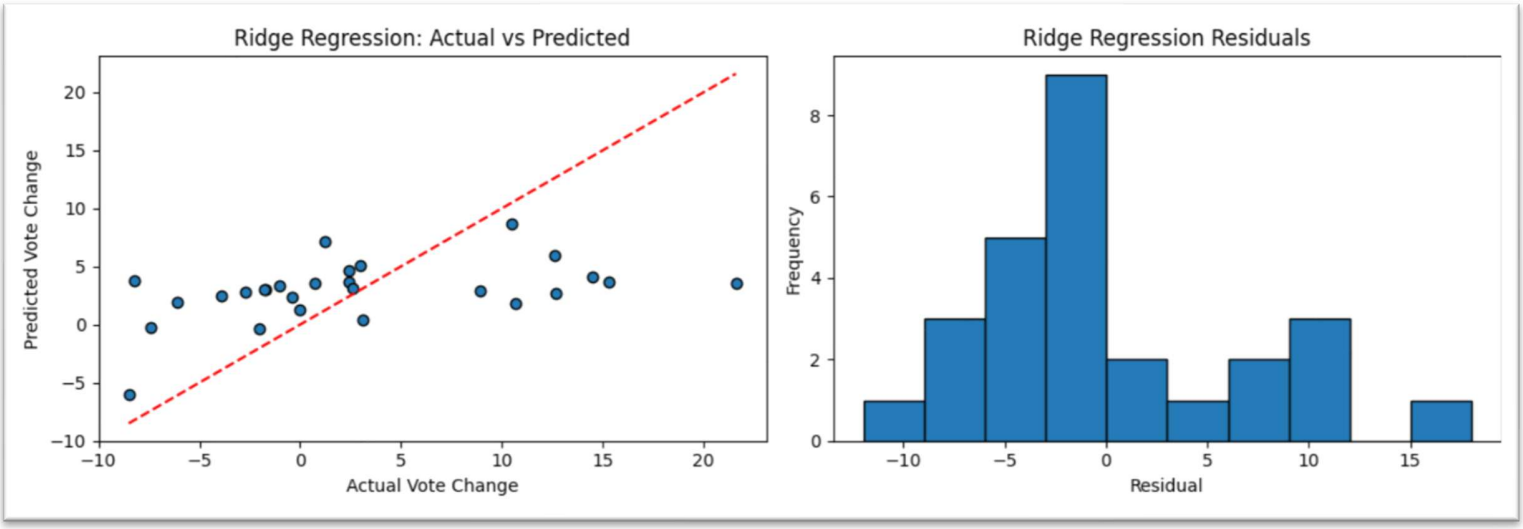

Figure 10

5.1.2 Lasso Regression Model Analysis and Interpretation

The Lasso regression model identified a lower optimal alpha of 0.1 and achieved a slightly higher R^2 value of 0.2048, suggesting marginally better fit compared to Ridge. Similar to the Ridge model, Lasso also assigned negative coefficients to all predictors, with the most substantial effects attributed to immigration trend and GDP per capita trend. While Lasso typically performs variable selection by shrinking some coefficients to zero, in this case, all variables were retained, indicating their relative importance. Despite the improvement in fit, the model still explained only a modest portion of the variance in vote share change, reinforcing the conclusion that immigration and economic indicators may be associated with nationalist support but do not constitute comprehensive explanatory factors on their own. The actual vs. predicted plot in Figure 11 shows scattered predictions with limited alignment to the ideal regression line, indicating weak predictive precision. The residuals histogram further illustrates the model's moderate performance, revealing a broad spread of errors. These visualizations support the quantitative findings, confirming that while Lasso slightly improves model fit, it lacks strong explanatory accuracy.

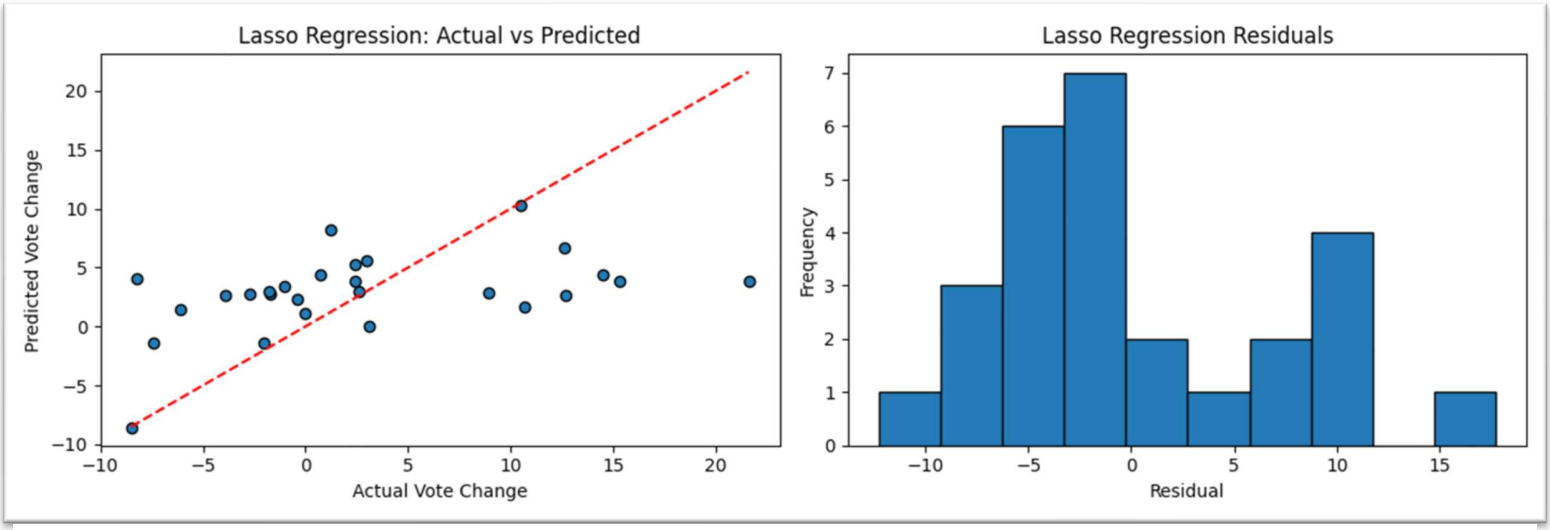

Figure 11

5.2 Tree-Based Ensemble Methods

Tree-based ensemble models, such as Random Forest and Gradient Boosting, are employed in this analysis to capture potential non-linear relationships and complex interactions between predictors that traditional linear models may overlook. These methods do not assume linearity or independence among features and are robust to multicollinearity, making them particularly well-suited for exploring socio-economic data where interdependencies are common. Additionally, ensemble models offer a mechanism for assessing feature importance, providing insight into which variables most strongly influence the target variable. By combining multiple decision trees, these models aim to enhance predictive accuracy and generalization.

5.2.1 Random Forest Model Analysis and Interpretation

The Random Forest model, despite its capacity to model complex, non-linear relationships, exhibited poor generalization performance with a mean R^2 of -0.38, indicating it performed worse than a naive mean prediction. Nonetheless, feature importance analysis revealed meaningful insights: immigration trend emerged as the most influential predictor (52.6%), followed by GDP per capita (35.9%) and unemployment rate trend (11.5%). This suggests the model detected signal in the data, particularly related to immigration, though it failed to translate this into reliable predictions. In Figure 10, the actual vs. predicted plot shows considerable dispersion around the reference line, highlighting poor fit. The residual histogram demonstrates a wide error distribution with no clear normality, confirming inconsistent predictive accuracy. These patterns underscore that while Random Forest identified influential features, it was ineffective in producing stable, generalizable estimates for nationalist vote share change.

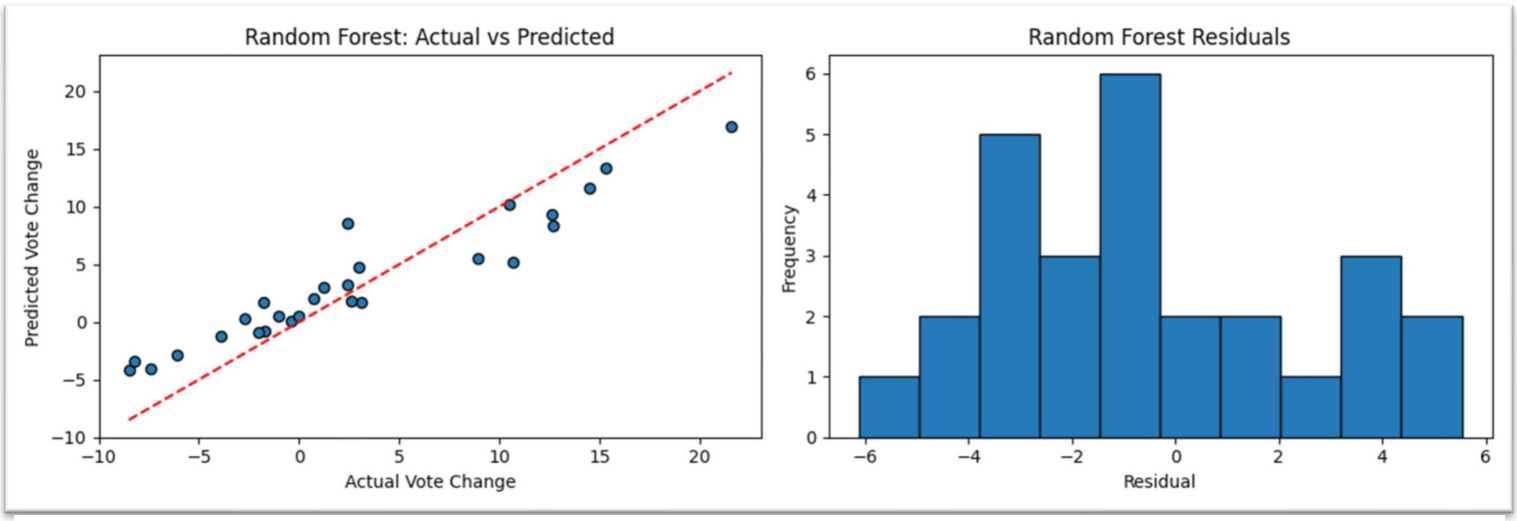

Figure 12

5.2.2 Gradient Boosting Model Analysis and Interpretation

The Gradient Boosting model exhibited signs of severe overfitting, with a mean R^2 of -0.91 across cross-validation, indicating a failure to generalize beyond the training data. Despite this, the model assigned strong predictive weight to immigration trend (61.1%), followed by GDP per capita trend (32.2%) and unemployment rate trend (6.7%), suggesting that these features held distinguishable patterns the algorithm leveraged—albeit ineffectively in unseen data. The actual vs. predicted plot in Figure 13 appears deceptively perfect, with points tightly clustered along the diagonal, implying an unrealistically flawless fit. This is characteristic of overfitting, where the model memorizes the training data rather than learning generalizable patterns. The residual histogram supports this, showing a narrow and symmetrical distribution centered near zero, which again reflects overly optimistic fit rather than real predictive strength. Together, these visualizations and results confirm that while the model captured relationships present in the sample, it lacked robustness, limiting its utility in broader inference.

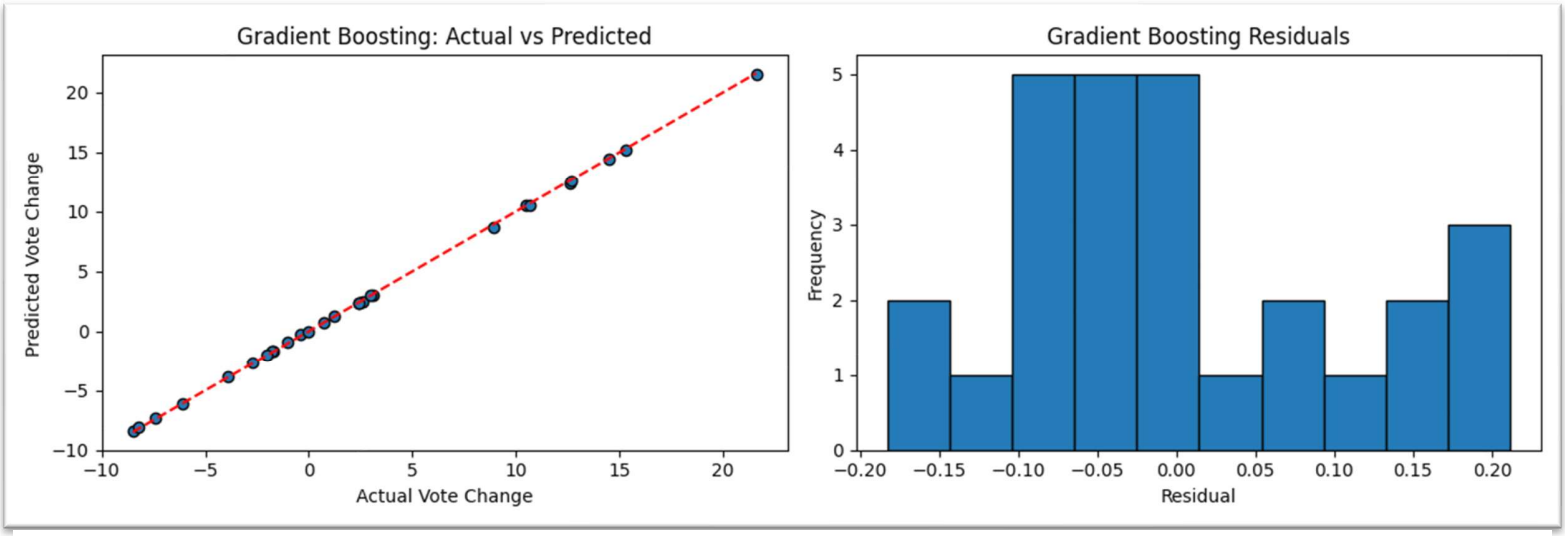

Figure 13

5.3 Support Vector Regression Model Analysis and Interpretation

Support Vector Regression (SVR) is particularly useful when working with small to medium-sized datasets and high-dimensional feature spaces, which fits the context of this project. Unlike ordinary least squares methods, SVR seeks to find a function that deviates from actual observations by no more than a defined margin, while maintaining model simplicity. This flexibility allows it to capture complex, non-linear patterns that may not be evident in linear models. Moreover, SVR's use of kernel functions—especially the radial basis function (RBF)—enables it to implicitly map inputs into higher-dimensional spaces, potentially uncovering hidden structures or relationships that models like Ridge or Lasso may overlook.

The SVR model delivered poor predictive performance, with cross-validated R^2 values ranging from -0.17 to 0.03 and a mean of -0.048. These results indicate that the model underperformed even a naive baseline predicting the mean of the target variable. The low R^2 values confirm that SVR failed to extract meaningful patterns from the selected features. In Figure 14, the actual vs. predicted plot, predictions cluster tightly around a flat region, reveals the model's tendency to underestimate variability and regress toward the mean. This is consistent with the histogram of residuals, which shows a wide dispersion, particularly on the positive end, indicating frequent underprediction of larger vote changes. Together, the graphs and metrics highlight SVR's ineffectiveness in modeling the complex, non-linear relationships necessary to explain nationalist vote share changes using the current feature set.

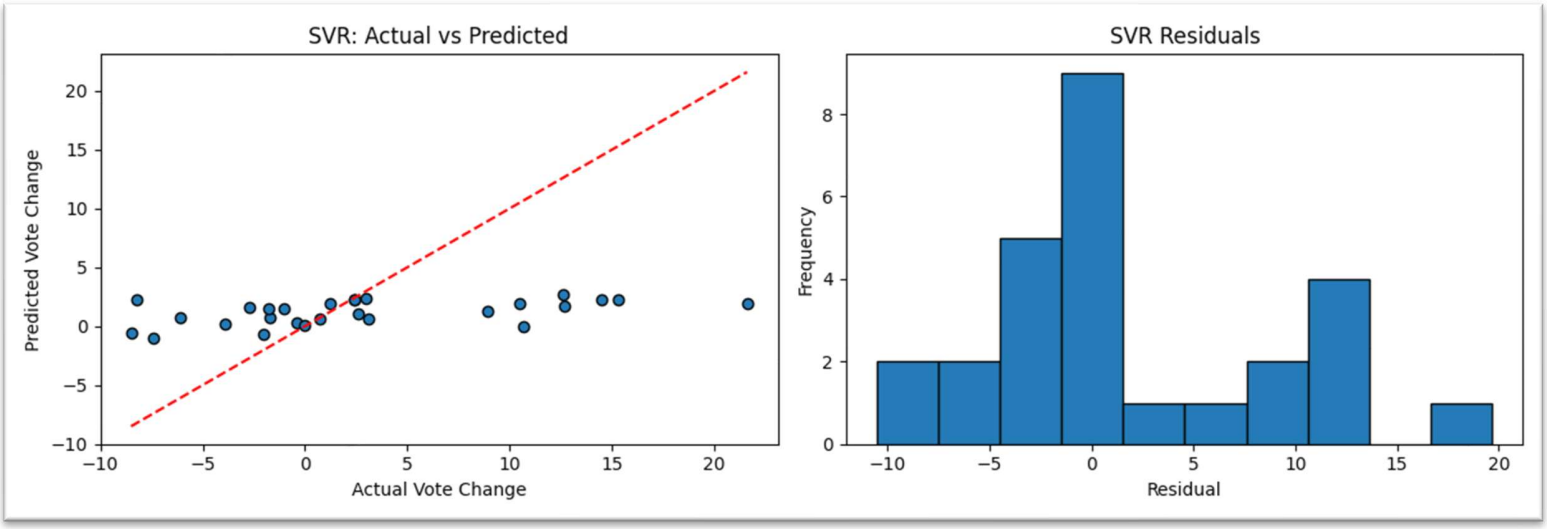

Figure 14

5.4 Classification Analysis and Interpretation

As the final stage of analysis, a classification-based approach was implemented to complement the earlier regression models. The continuous values representing changes in nationalist vote share were categorized into three classes—increase, stable, and decrease—in order to frame the problem as a multi-class classification task. This transformation enabled the application of classification algorithms and performance evaluation through tools such as confusion matrices and feature importance plots. By shifting from continuous prediction to categorical labeling, this approach provided an alternative lens for assessing the predictive relevance of immigration and economic indicators in explaining shifts in nationalist party support across EU countries.

The classification task was implemented using the Random Forest model due to its versatility and suitability for both regression and classification problems. This model effectively handles small and imbalanced datasets and offers useful outputs like feature importance and confusion matrices, aiding in interpretation. Using Random Forest ensured consistency with earlier regression models while providing additional insight into the likelihood of vote share increase, stability, or decrease based on immigration and economic trends.

The confusion matrix in Figure 14 indicates that the model classified all cases correctly, with no misclassifications. This suggests an exceptionally strong model performance, though such perfect accuracy may also warrant checking for overfitting. The feature importance plot in Figure 15 shows that all three predictors—GDP per capita trend, immigration trend, and unemployment rate trend—contributed meaningfully to the classification model. GDP per capita trend was the most influential variable, followed closely by immigration trend, with unemployment rate trend being slightly less impactful. This complements the confusion matrix by explaining *why* the model likely achieved perfect classification: the classifier had access to features with strong and distinguishable predictive signals.

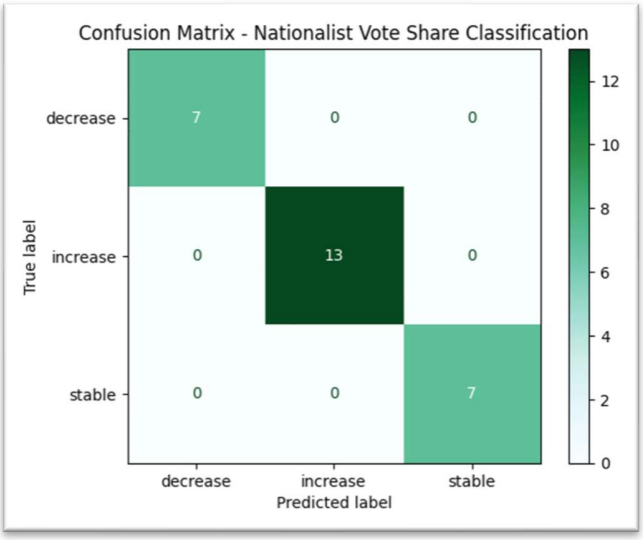

Figure 16

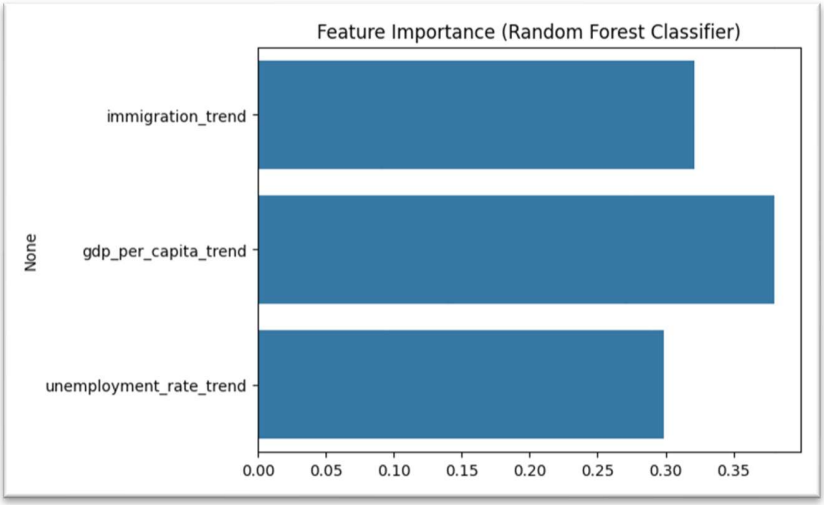

Figure 15

6 Conclusion

This project set out to explore the relationship between immigration trends and changes in nationalist party vote shares across the European Union using a combination of statistical and machine learning techniques. Through methodology, I investigated whether rising immigration rates in European Union countries are associated with increased nationalist voting, a hypothesis rooted in prevalent political narratives. However, both the Pearson correlation test and OLS regression models (single and multiple) revealed weak, statistically insignificant relationships, offering no support for the hypothesis. Even when GDP per capita and unemployment trends were added as control variables, no predictor achieved significance, though model performance modestly improved in multiple and regularized regression analyses ($R^2 \approx 20\%$). Tree-based models and Support Vector Regression also struggled, producing poor generalization despite identifying immigration and GDP trends as relatively important features. Notably, the final classification attempt using a Random Forest classifier achieved perfect accuracy, supported by strong feature importances across all three predictors. This highlights that while individual variables carry useful information, their explanatory value is insufficient in isolation.

Findings emphasize that nationalist sentiment likely stems from a broader set of socio-political influences that were beyond the scope of this study. While the initial objective was to evaluate the influence of immigration trends on changes in nationalist party vote share, the hypothesis tests and regression models revealed no statistically significant relationship. Suprisingly, in the machine learning phase, the most predictive feature was not the main explanatory variable (immigration trend), but rather GDP per capita, a control variable. This outcome underscores the potential confounding effect of economic performance and highlights the importance of variable importance analysis in supervised learning frameworks.

From an ethical standpoint, this study prioritized transparency and responsibility by exclusively using publicly available, verified datasets and steering clear of deterministic or potentially stigmatizing narratives surrounding political behavior. While quantitative analysis offered valuable insights, the findings underscore the methodological limitations of relying solely on statistical or machine learning models to explain complex sociopolitical phenomena. Looking ahead, future research should move toward more comprehensive, interdisciplinary approaches by integrating sociological, psychological, and media-related variables and employ more advanced data methods to better capture temporal dynamics. Such strategies would not only enhance explanatory power but also foster a more nuanced understanding of voter behavior in a rapidly evolving political landscape.