Abstract

The question of which European Union country offers the best opportunities for immigrants is as important as it is complex. While the EU is seen as a region of opportunity and diversity, the conditions for immigrants differ widely between member states, influenced by economic, social, and institutional factors.

To better understand these differences, we conducted a thorough analysis to explore which countries provide the most supportive environments for immigrants. By examining a wide range of relevant factors, we identified patterns that reveal not only where opportunities are most abundant but also which nations foster conditions for long-term integration and growth.

This study seeks to shed light on the EU countries that stand out for their welcoming and supportive environments, offering valuable insights for individuals considering immigration. By drawing attention to these examples, we hope to inspire more inclusive and forward-thinking approaches across the region

KEY WORDS: *Immigration; EU Countries; Opportunities; Inclusivity*

Introduction

Immigration significantly influences the social, economic, and cultural dynamics of European Union (EU) countries. People migrate to the EU seeking better opportunities in areas such as employment, healthcare, and overall quality of life. However, the conditions and support systems provided to immigrants vary widely across EU countries. While some nations have robust structures to help immigrants integrate and thrive, others face challenges in addressing their needs effectively.

Modern immigration is no longer just about seeking economic stability. Immigrants also prioritize access to quality healthcare, financial security, and fair opportunities for education and social integration. Factors like corruption levels, healthcare expenditure, financial security, and median income play critical roles in shaping the experience of immigrants. Moreover, the ability to acquire citizenship and build a sustainable livelihood are essential for long-term success and inclusion.

The findings will provide valuable insights for immigrants exploring their options and for policymakers striving to create more inclusive and equitable societies. By understanding which countries excel in supporting immigrants, we hope to inspire the adoption of policies that promote opportunity, inclusion, and a higher quality of life for all.

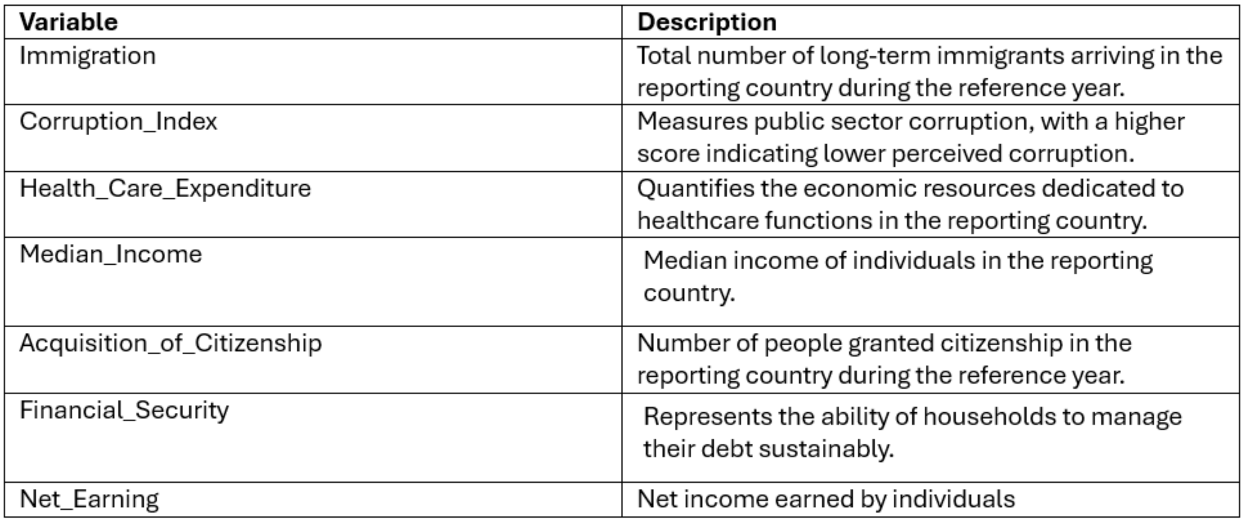
METHODOLOGY

DATA DESCRIPTION

The data used in this study was gathered from **Eurostat**, the statistical office of the European Union, which offers reliable and comprehensive statistics on various aspects of life in EU member states. In our analysis, we decided to focus exclusively on **EU member states** because they share common policies and standards that influence economic conditions, social services, and immigration processes. This made them an ideal group for comparison, as they operate within a unified legal and regulatory framework while still displaying variations in how they approach immigration and integration.

Finding reliable data from EU member states was crucial for the success of this study. The accuracy and consistency of the data were key in ensuring that our analysis was based on trustworthy and comparable indicators across all countries. By using data from a reputable source like Eurostat, we were able to ensure the validity of our findings, which ultimately provide a clearer understanding of the opportunities available to immigrants in different EU countries.

The dataset is organized into the following categories:



Variable	Description
Immigration	Total number of long-term immigrants arriving in the reporting country during the reference year.
Corruption_Index	Measures public sector corruption, with a higher score indicating lower perceived corruption.
Health_Care_Expenditure	Quantifies the economic resources dedicated to healthcare functions in the reporting country.
Median_Income	Median income of individuals in the reporting country.
Acquisition_of_Citizenship	Number of people granted citizenship in the reporting country during the reference year.
Financial_Security	Represents the ability of households to manage their debt sustainably.
Net_Earning	Net income earned by individuals

Figure 1 – Data

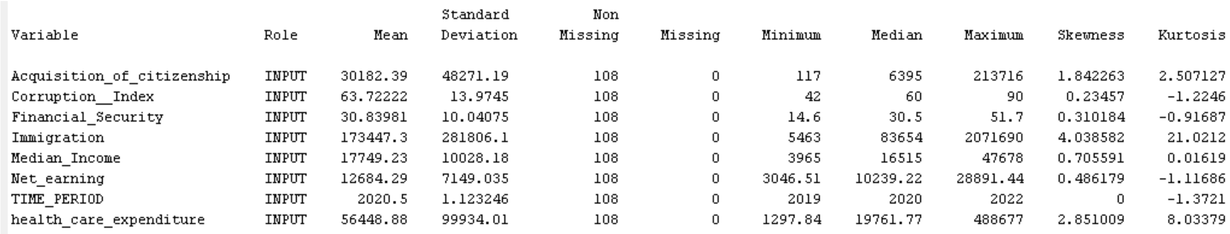
VARIABLE SELECTION

The variables selected for this study were based on our research, which identifies them as crucial factors influencing immigrant integration in EU member states. Economic indicators, such as median income, financial security, and net earnings, were chosen because they directly reflect the material conditions immigrants face, which are key to understanding their economic integration and opportunities.

Institutional factors, including the corruption index and acquisition of citizenship, were selected based on our findings that the quality of governance and access to legal pathways are essential for immigrants' successful integration. These variables provide insight into how countries create inclusive environments for immigrants.

Data Treatment

In our dataset of EU member states, no missing values were identified, allowing for the use of the complete dataset without the need for imputation or exclusion.



```
                                                 Standard        Non
Variable                     Role        Mean    Deviation    Missing     Missing     Minimum      Median     Maximum    Skewness    Kurtosis

Acquisition_of_citizenship   INPUT   30182.39    48271.19         108           0         117        6395      213716    1.842263    2.507127
Corruption__Index            INPUT   63.72222     13.9745         108           0          42          60          90     0.23457     -1.2246
Financial_Security           INPUT   30.83981    10.04075         108           0        14.6        30.5        51.7    0.310184    -0.91687
Immigration                  INPUT   173447.3    281806.1         108           0        5463       83654     2071690    4.038582     21.0212
Median_Income                INPUT   17749.23    10028.18         108           0        3965       16515       47678    0.705591     0.01619
Net_earning                  INPUT   12684.29    7149.035         108           0     3046.51    10239.22    28891.44    0.486179    -1.11686
TIME_PERIOD                  INPUT     2020.5    1.123246         108           0        2019        2020        2022           0     -1.3721
health_care_expenditure      INPUT   56448.88    99934.01         108           0     1297.84    19761.77      488677    2.851009     8.03379
```

Figure 2 – Summary Statistics

The summary statistics (Figure 2) show considerable variation in the means and standard deviations across variables due to differences in their measurement scales. For example, health care expenditure has a maximum value of 488,677 with a high standard deviation, while net earnings range between 3,046.51 and 22,891.44, reflecting a much smaller spread. These differences in scale could influence the relative weight of variables in the analysis.

To account for these variations, the data was standardized by transforming all variables into z-scores. This step adjusted for the differences in scale, allowing for comparability across variables.

A review of the dataset did not reveal significant outliers or anomalies that would affect the analysis. All data points were retained, and no inconsistencies were observed across the variables.

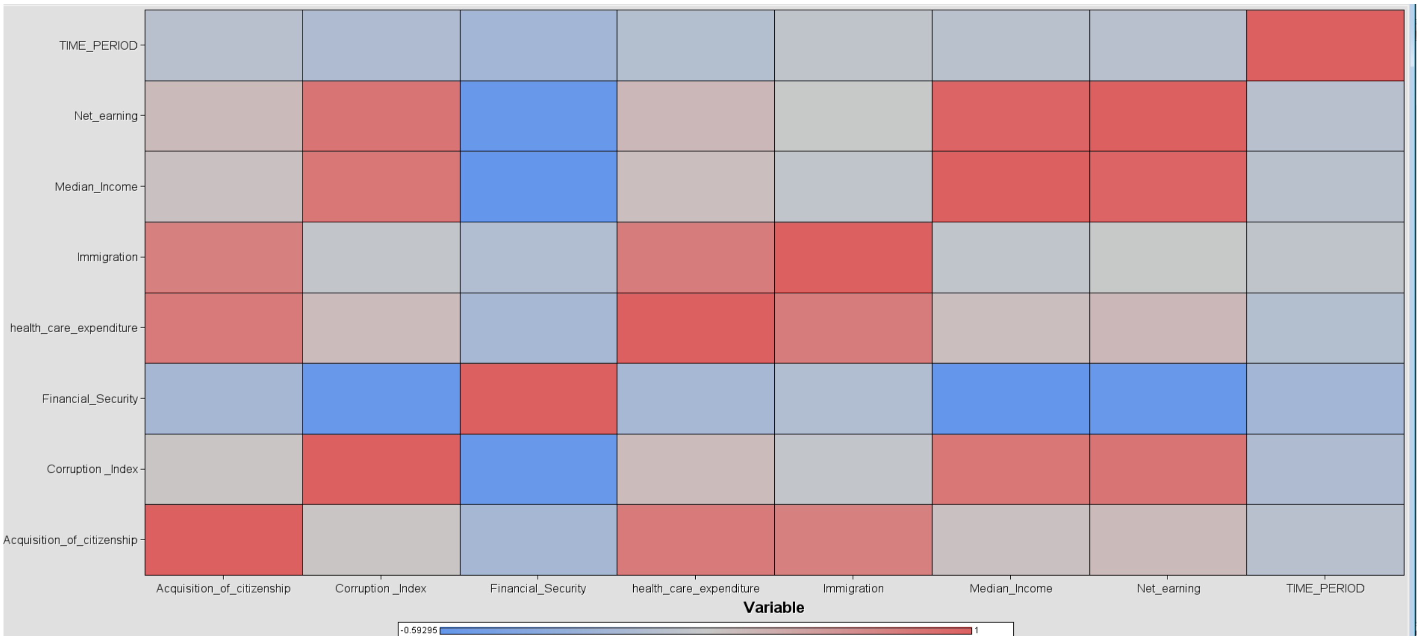


Figure 3 – correlation matrix

The correlation matrix demonstrates that most variables are strongly correlated, indicating that they are likely to capture related aspects of the phenomena under study. Variables such as *Net_Earning*, *Median_Income*, and *Financial_Security* show positive associations, suggesting that they jointly represent dimensions of economic well-being.

Similarly, *Health_Care_Expenditure* and *Immigration* are also correlated, reflecting potential links between public spending and migration dynamics

METHODS

Principal Components Factoring (PCF)

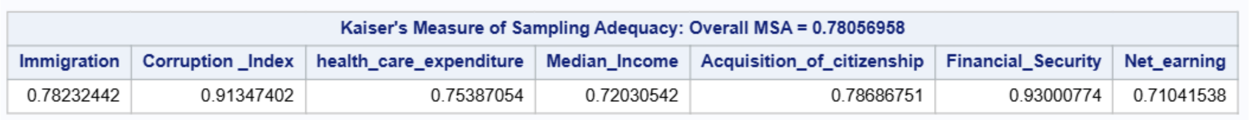
For our analysis, we decided to use Principal Components Factoring (PCF) rather than Principal Axis Factoring (PAF). The primary reason for this choice lies in the nature of our dataset, which is sourced from trustworthy, objective, and factual platform like Eurostat. Since the data is reliable and robust, Principal Components Factoring allows us to efficiently reduce the dimensionality of the dataset while retaining the maximum amount of variance across the variables.

The correlation matrix has already demonstrated that the majority of our variables are strongly correlated, which is a key requirement for conducting a meaningful factor analysis. Strong correlations among variables indicate shared underlying dimensions.

Is Our Data Suitable for Factor Analysis?

To determine the suitability of our dataset for factor analysis, we evaluated several key statistical measures. The **Kaiser-Meyer-Olkin (KMO)** measure of sampling adequacy was calculated as 0.78, indicating a

good level of adequacy as per the commonly accepted scale. This result suggests that the data is well-suited for factor analysis, as it exceeds the threshold of 0.5. Furthermore, all individual KMO values for the variables were above 0.5, confirming that each variable contributes sufficiently to the shared variance structure.

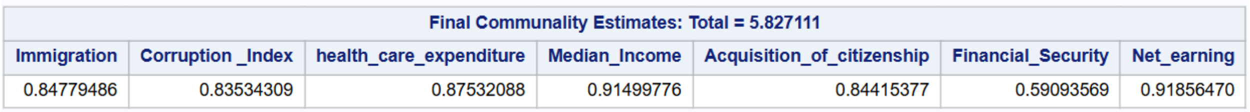


Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.78056958

Immigration	Corruption _Index	health_care_expenditure	Median_Income	Acquisition_of_citizenship	Financial_Security	Net_earning
0.78232442	0.91347402	0.75387054	0.72030542	0.78686751	0.93000774	0.71041538

Figure 4 – KMO

The communalities, which represent the proportion of variance each variable shares with the extracted factors, were also examined. The final communality estimate was 5.8, with all variables showing values above the minimum threshold of 0.5. This indicates that all variables are well-represented by the factors, reinforcing the appropriateness of the dataset for this analysis.

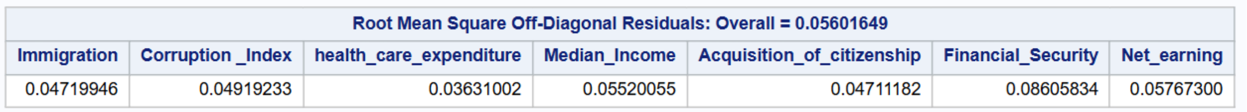


Final Communality Estimates: Total = 5.827111

Immigration	Corruption _Index	health_care_expenditure	Median_Income	Acquisition_of_citizenship	Financial_Security	Net_earning
0.84779486	0.83534309	0.87532088	0.91499776	0.84415377	0.59093569	0.91856470

Figure 5 – Final Communality Estimates

The residuals of the reproduced correlation matrix were evaluated using the root mean square off-diagonal residual (RMSR). The overall RMSR value was 0.05, with none of the residuals exceeding 0.1. This result signifies a good fit between the observed and reproduced correlations, suggesting that the factor model is effective in capturing the structure of the data.



Root Mean Square Off-Diagonal Residuals: Overall = 0.05601649

Immigration	Corruption _Index	health_care_expenditure	Median_Income	Acquisition_of_citizenship	Financial_Security	Net_earning
0.04719946	0.04919233	0.03631002	0.05520055	0.04711182	0.08605834	0.05767300

Figure 6 – RMSR

Based on these findings, we conclude that the dataset meets the necessary criteria for conducting **PCF**. The high communalities, acceptable KMO value, and low residuals collectively demonstrate the suitability of the data.

Number Of Factors

Number of Factors

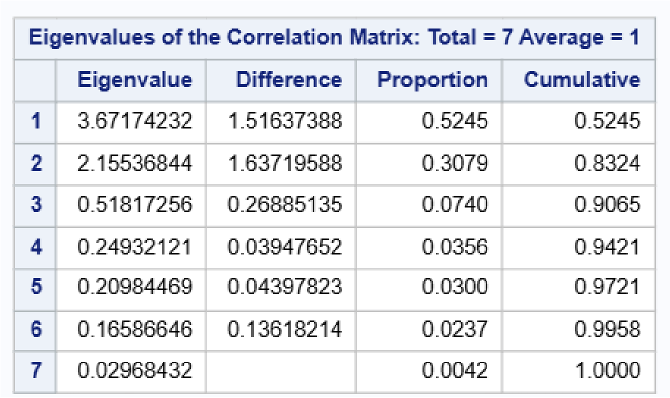
Determining the appropriate number of factors to retain is a critical step in factor analysis, as it directly impacts the interpretability and utility of the results. Several criteria were applied to ensure the selection of an optimal factor structure:

Kaiser’s Criteria

According to Kaiser’s criterion, factors with eigenvalues greater than 1 should be retained. This method assumes that a factor with an eigenvalue below 1 contributes less to explaining the variance than a single variable and is, therefore, not meaningful for inclusion in the analysis. In our results, two factors were identified with eigenvalues exceeding 1, suggesting that these two factors are significant contributors to the overall variance in the dataset.

Pearson’s Criteria

Pearson’s criterion considers the cumulative percentage of explained variance to determine the number of factors. A commonly accepted threshold is retaining enough factors to explain at least 80% of the total variance. The results indicated that the first two factors together accounted for approximately 80% of the total variance, confirming the retention of two factors as optimal.



Eigenvalues of the Correlation Matrix: Total = 7 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.67174232	1.51637388	0.5245	0.5245
2	2.15536844	1.63719588	0.3079	0.8324
3	0.51817256	0.26885135	0.0740	0.9065
4	0.24932121	0.03947652	0.0356	0.9421
5	0.20984469	0.04397823	0.0300	0.9721
6	0.16586646	0.13618214	0.0237	0.9958
7	0.02968432		0.0042	1.0000

Figure 7 – EigenValue, Cumulative

Scree Plot's Criteria

The scree plot provides a visual method for determining the number of factors by identifying the "elbow," where the curve of eigenvalues flattens. In this analysis, the scree plot suggested ambiguity between two and three factors, as the inflection point occurred between the second and third eigenvalues. However, the sharp decline after the second factor aligned more closely with the two-factor solution.

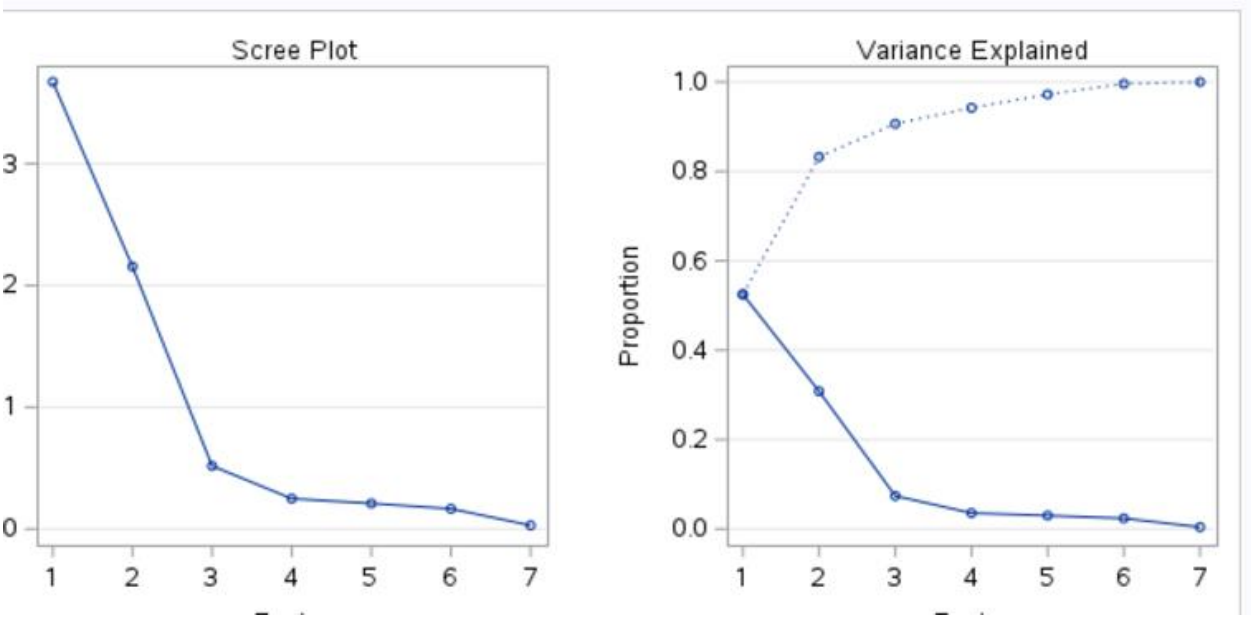


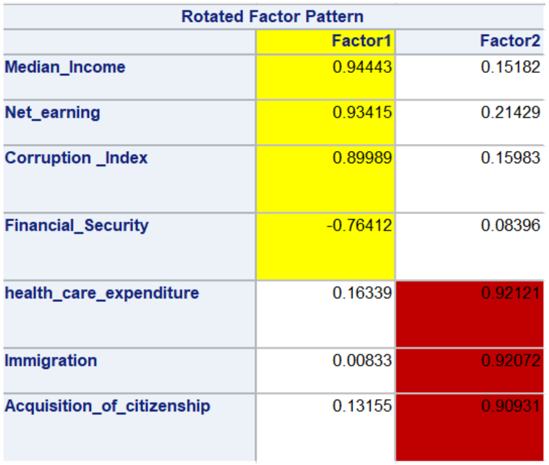
Figure 8 – Scree Plot's

Considering the agreement between Kaiser's criterion and Pearson's criterion, and noting the slight ambiguity in the scree plot, we decided to proceed with a two-factor model. This approach balances simplicity and the explanatory power of the factors.

Factor Rotation

Varimax Rotation

Varimax rotation produced factors with slightly higher loadings, which resulted in a clearer differentiation between the two factors identified in our analysis—Economic Development and Social Inclusion. The higher loadings helped strengthen the relationship between variables and their respective factors, making the factor structure easier to interpret in the context of our study.

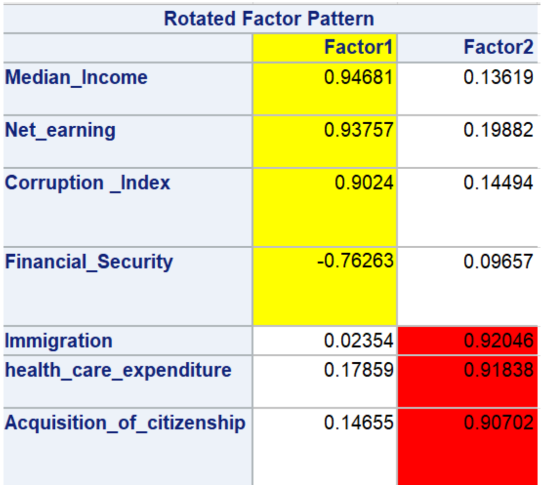


Rotated Factor Pattern		
	Factor1	Factor2
Median_Income	0.94443	0.15182
Net_earning	0.93415	0.21429
Corruption _Index	0.89989	0.15983
Financial_Security	-0.76412	0.08396
health_care_expenditure	0.16339	0.92121
Immigration	0.00833	0.92072
Acquisition_of_citizenship	0.13155	0.90931

Figure 8 – Varimax Factor Rotation

Quartimax Rotation

Quartimax rotation, on the other hand, provided the same factor structure, but the loadings were ever more slightly different across the factors. While this didn't alter the interpretation of the factors, the choice between Varimax and Quartimax wouldn't have made a major difference in our analysis, but comparing the two helped highlight the subtle differences in how the factors were represented.



Rotated Factor Pattern		
	Factor1	Factor2
Median_Income	0.94681	0.13619
Net_earning	0.93757	0.19882
Corruption _Index	0.9024	0.14494
Financial_Security	-0.76263	0.09657
Immigration	0.02354	0.92046
health_care_expenditure	0.17859	0.91838
Acquisition_of_citizenship	0.14655	0.90702

Figure 9 – Quartimax Factor Rotation

Comparison

Both rotation methods led to the same interpretable factor structure with

Factor1: Labeled as **Economic Development**

Factor 2: Labeled as **Social Inclusion**

Since the interpretation of the factors was consistent across both methods, we proceeded with **Varimax** as we believed it was the best option for our analysis.

Factor visualization

Factor 2: Social Inclusion

The map illustrates the distribution of Factor 2, which we have labeled **Social Inclusion**, across EU member states. This factor is influenced by variables such as **immigration**, **health care expenditure**, and **acquisition of citizenship**.

The color gradient ranges from darker shades (representing lower values, around -0.7) to lighter shades (indicating higher values, approaching 1).

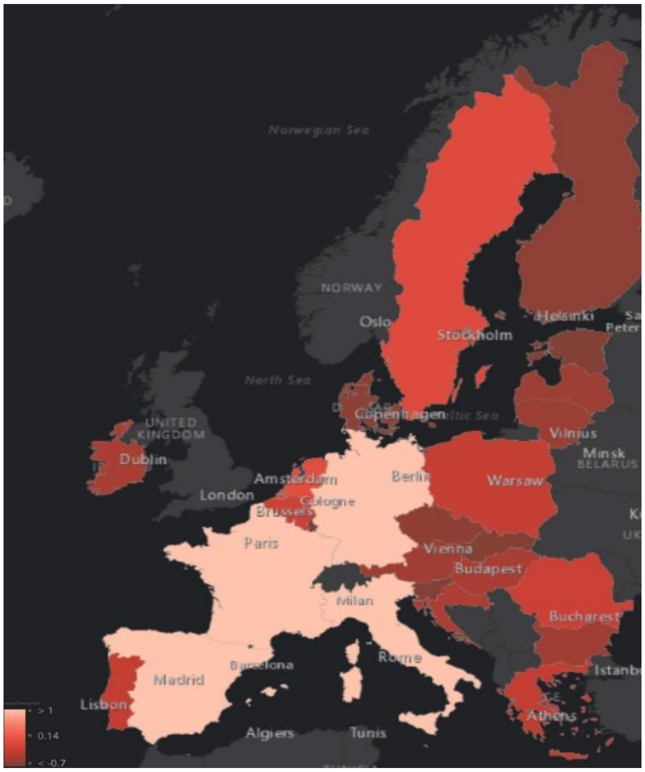


Figure 10 – Social Inclusion Map

- Countries **with Lighter Shades**

higher Social Inclusion scores, marked by lighter shades. These countries demonstrate significant investment in public health care, higher immigration levels, and relatively accessible pathways for acquiring citizenship. Their policies often reflect a commitment to integrating immigrants and ensuring access to public services.

- Countries **with Darker Shades**

Eastern and some Northern European countries, including **Poland, Hungary, and the Baltic States**, appear in darker shades. These regions are characterized by lower scores, indicating less emphasis on immigration, limited health care expenditure per capita, and more restrictive pathways to citizenship. This could reflect economic, political, or historical factors influencing their approach to inclusivity.

- Intermediate **Scores**

Countries such as **Belgium, Sweeden, and the Netherlands** exhibit medium levels of Social Inclusion. These countries have established systems to accommodate immigrants and allocate resources to public health care but may vary in their accessibility to citizenship or the proportion of public spending dedicated to these factors.

Factor 1: Economic Development

The map illustrates the distribution of **Economic Development** (Factor 1) across European countries, using a gradient where lighter colors represent higher scores (approaching 1) and darker colors correspond to lower scores (closer to -0.7). This factor is primarily defined by variables such as **Median Income**, **Net Earnings**, **Corruption Index**, and **Financial Security**, which collectively reflect the economic well-being of the population and institutional quality.

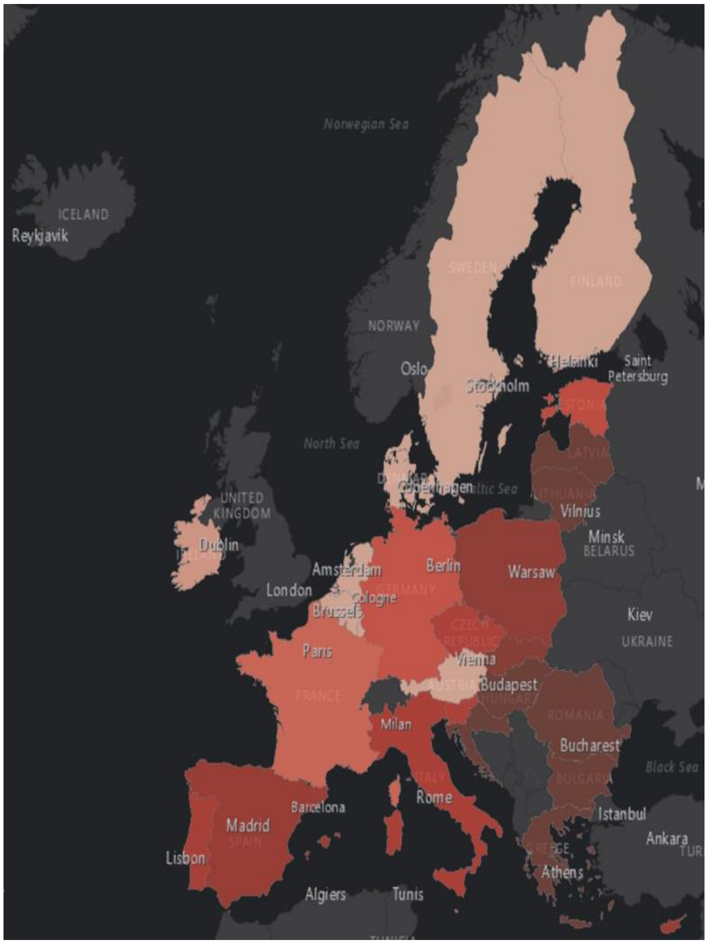


Figure 11 – Economic Development

Regions with Higher Scores

Countries in **Northern Europe**, such as **Sweden, Finland, and Denmark**, exhibit the highest scores for economic development, as indicated by the lightest shades. These nations are characterized by robust financial systems, higher median incomes, low perceived corruption, and significant financial security, all of which contribute to their favorable positioning.

Regions with Lower Scores

Countries in **Southern Europe**, including **Portugal and Spain**, as well as parts of **Eastern Europe**, are depicted with darker shades, reflecting lower levels of economic development. These regions face challenges such as lower average incomes, less financial stability, and higher perceptions of corruption, which are consistent with their relatively weaker performance on this factor.

Regions with Moderate Scores

Countries such as **Germany, France, and the Netherlands** show intermediate scores, reflected by moderate shades on the map. These nations exhibit balanced economic conditions, combining relatively high median incomes and financial security with moderate levels of corruption and institutional challenges.

The factors reveals significant disparities in Social Inclusion and Economic Development across EU member states.

- **Social Inclusion** is highest in countries with progressive immigration policies, substantial healthcare investment, and accessible citizenship pathways, while lower scores reflect limited inclusivity, often due to economic or historical factors. Intermediate levels of inclusion highlight varying degrees of commitment to integration and public resource allocation.
- **Economic Development** follows a similar gradient, with Northern European countries excelling due to strong financial systems and institutional quality. In contrast, Southern and Eastern Europe face economic and systemic challenges. Intermediate performers balance economic stability with areas for improvement.

Cluster Analysis

Cluster analysis is a statistical technique used to group similar observations into clusters, ensuring that observations within the same cluster are highly similar (internally homogeneous) while observations in different clusters are distinctly dissimilar (externally heterogeneous). This method is particularly useful for uncovering patterns and structures in the data, which may not be immediately evident.

Clustering Methods

To perform cluster analysis, a decision must first be made on the type of clustering method to employ. Two primary types are commonly used:

1. **Hierarchical Clustering**: This method builds a hierarchy of clusters based on the pairwise distances or similarities between data points.
2. **Non-Hierarchical Clustering**: This approach partitions the dataset into a predetermined number of clusters without forming a hierarchical structure.

Hierarchical Clustering

Hierarchical clustering iteratively merges or splits clusters based on specific criteria. In this approach, the distance between clusters can be calculated using various linkage methods, which influence how clusters are formed:

- **Single Linkage (Nearest-Neighbor)**: The shortest distance between two points in different clusters determines cluster proximity. This method tends to form elongated clusters.
- **Complete Linkage (Farthest-Neighbor)**: The largest distance between any two points in different clusters is used. It tends to produce compact and evenly sized clusters.

- **Average Linkage**: The average distance between all points in one cluster and all points in another determines proximity, balancing between single and complete linkage.
- **Ward's Method**: This approach forms clusters by minimizing the increase in within-cluster variance. It emphasizes creating clusters that are internally homogeneous.
- **Centroid Linkage**: The distance between cluster centroids is used to determine proximity. This method assumes that clusters are spherical in shape.

For our analysis, hierarchical clustering was employed as an exploratory tool to identify the optimal number of clusters by testing these linkage methods. The method that provided the clearest and most interpretable results was selected for further analysis.

Non-Hierarchical Clustering

Non-hierarchical methods, such as k-means clustering, involve assigning observations to a predefined number of clusters. This method requires the number of clusters to be determined beforehand, making hierarchical clustering a valuable preliminary step for determining this optimal number.

In this project, after determining the number of clusters using hierarchical clustering, non-hierarchical clustering was applied to refine the solution and interpret the results more clearly.

Clustering Solutions of standardized original variables.

After standardizing the original variables, we applied Cluster Analysis to group countries based on their similarities. This method was chosen to improve interpretability and derive meaningful conclusions from the dataset.

Analysis of Clustering Methods

The SAS output provided insights into five Hierarchical Clustering methods:

- **Single Linkage**
- **Complete Linkage**
- **Average Linkage**
- **Ward's Method**
- **Centroid Linkage**

To determine the most optimal clustering approach, we analyzed the Dendrogram, and the R-Square values generated by these methods. Ward's method emerged as the most suitable solution for our analysis, this method proved to be the most effective among the five options in our analysis, as it achieved the highest R-Square value for the same number of clusters, solidifying it as the optimal choice.

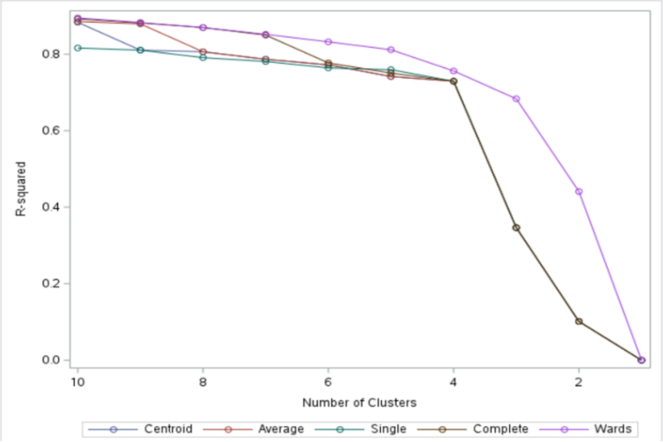


Figure 12 – Clustering methods 1

After examining the dendrogram, we observed the "first big jump," which represented a substantial increase in the distance between clusters. This jump marked the point where distinct groupings began to emerge, a crucial aspect in hierarchical clustering for determining the optimal number of clusters. We concluded that three clusters provided the most suitable solution. This choice allowed for a more nuanced interpretation of the data, maintaining a higher level of detail compared to simpler configurations.

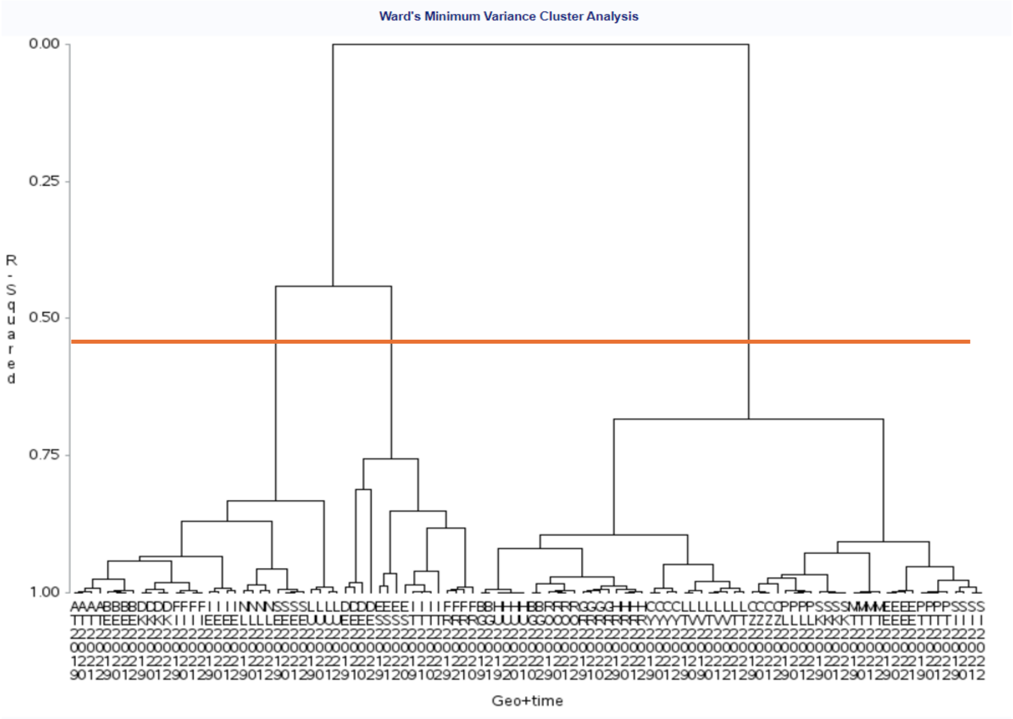


Figure 13 – Dendrogram 1

HIERARCHICAL CLUSTER ANALYSIS WITH THE FACTORS

When conducting the cluster analysis using the factors, we observed that the results aligned closely with those obtained from the original variables. Both approaches confirmed Ward's method as the most optimal algorithm to use. This alignment highlights that the extracted factors effectively represent the structure of the original.

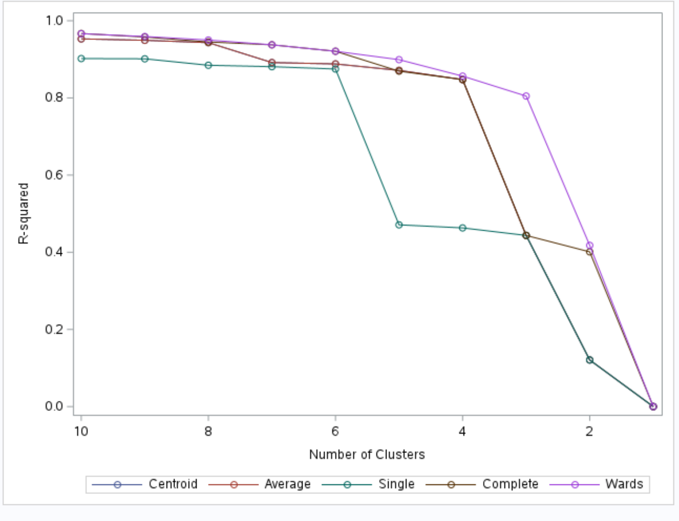


Figure 14 – Clustering methods 2

Examining the dendrogram generated with the original variables reveals the same results to the factor-based approach. the dendrogram for factors pointed toward retaining three clusters which is aligned with the analysis of original variables which showed three clusters too. To confirm,, the report also shown that the factors effectively represent the original variables and provide a solid foundation for further analysis.

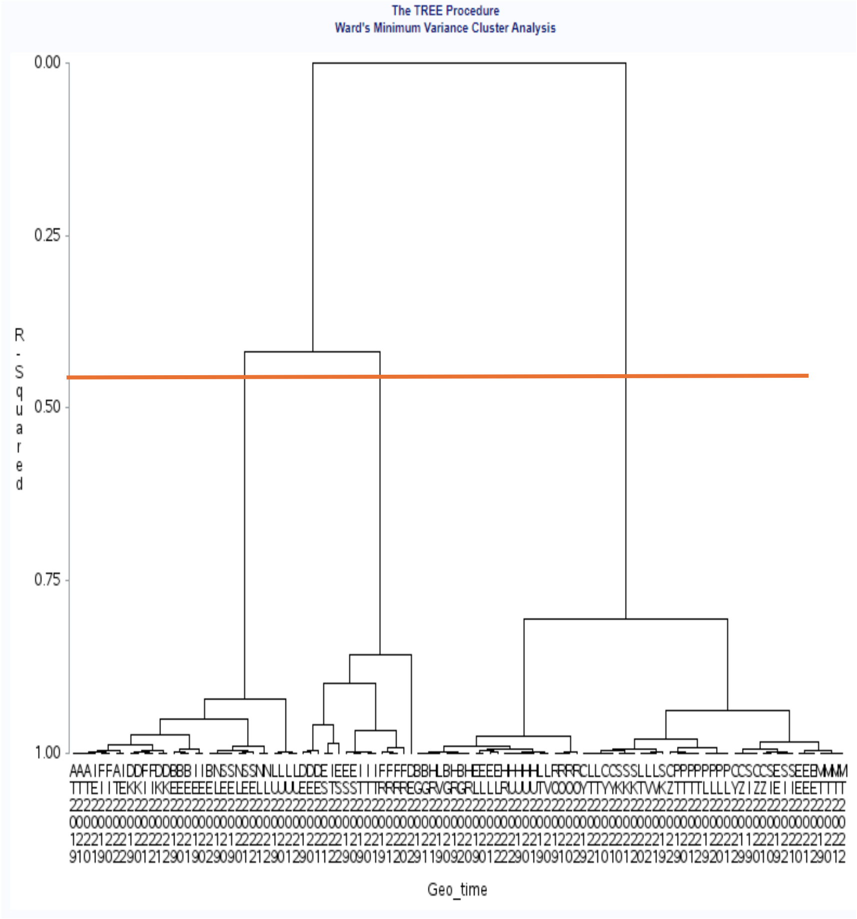


Figure 15 – Dendrogram 2

To refine our results, we will start the K-means clustering with the three clusters identified through the factors and original variables. This approach will allow us to further assess the consistency and validity of the factor-based clustering outcomes and potentially uncover more precise groupings in the data.

NON-HIERARCHICAL CLUSTERING (K-Means)

K-means clustering is a popular non-hierarchical method used to partition data into a specified number of clusters. Unlike hierarchical clustering, which creates a tree-like structure, K-means directly assigns each data point to a cluster, aiming to minimize the variance within each cluster and maximize the distance between clusters. This approach makes K-means particularly useful for large datasets and when the desired number of clusters is predefined.

For our analysis, we decided to start with three clusters based on the results from factor analysis. This initial decision was supported by the interpretation of the dendrogram and the alignment between the factors and the original variables, as we previously discussed.

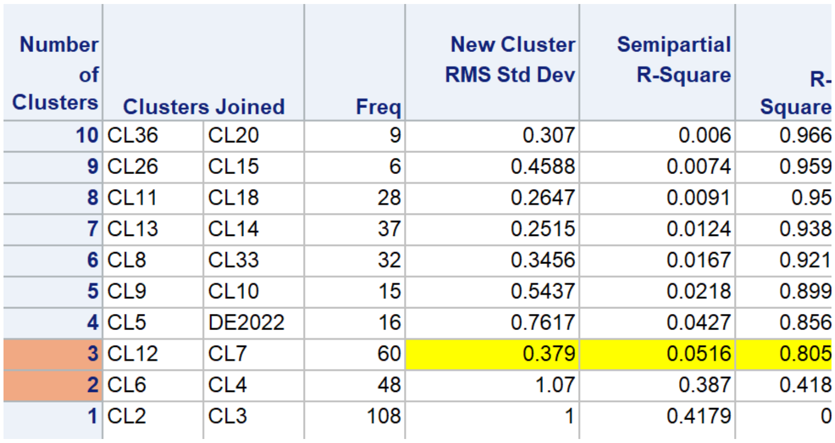
***K-means** works as follows:*

1. Assign each data point to the nearest centroid based on the Euclidean distance.
2. Update the centroid by calculating the mean of the points assigned to it.
3. Repeat steps 1 and 2 until the assignments no longer change, or a predefined number of iterations is reached.

By starting with three clusters, we aim to further evaluate the stability and validity of the factor-based clustering results. K-means provides a more direct, computationally efficient approach to clustering, and the results can be compared with those from hierarchical clustering to assess whether they lead to similar groupings.

K-Means with Factors Scores

The evaluation of the optimal number of clusters began with an analysis of the **Root Mean Square Standard Deviation (RMSSTD)**, a metric that measures the variability within clusters. A smaller RMSSTD indicates more compact and homogenous clusters, which is desirable. As clusters are merged, the RMSSTD typically increases due to the introduction of greater variability within the groups.



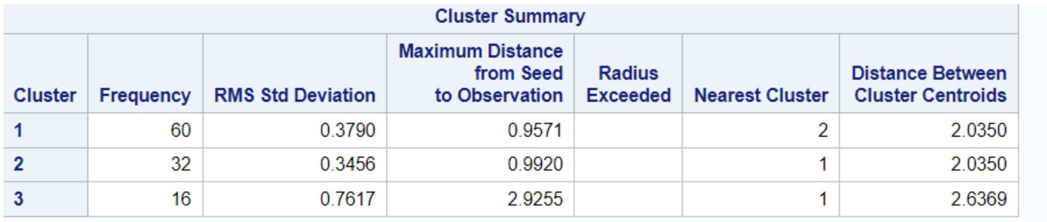
Number of Clusters	Clusters Joined		Freq	New Cluster RMS Std Dev	Semipartial R-Square	R-Square
10	CL36	CL20	9	0.307	0.006	0.966
9	CL26	CL15	6	0.4588	0.0074	0.959
8	CL11	CL18	28	0.2647	0.0091	0.95
7	CL13	CL14	37	0.2515	0.0124	0.938
6	CL8	CL33	32	0.3456	0.0167	0.921
5	CL9	CL10	15	0.5437	0.0218	0.899
4	CL5	DE2022	16	0.7617	0.0427	0.856
3	CL12	CL7	60	0.379	0.0516	0.805
2	CL6	CL4	48	1.07	0.387	0.418
1	CL2	CL3	108	1	0.4179	0

Figure 16 – Cluster History

From the **Cluster History** table, the RMSSTD (Root Mean Square Standard Deviation) displayed a steady increase as clusters were merged, with the most significant change observed between 3 clusters (0.379) and 2 clusters (1.07). This marked increase indicates that merging clusters beyond 3 introduces substantial heterogeneity, reducing the cohesion of the configuration. Furthermore, the semipartial R-Square also highlights a notable decrease in explained variance when moving from 3 clusters (0.0516) to 2 clusters (0.387), further supporting the decision to maintain 3 clusters. Therefore, we opted to proceed with a 3-cluster solution.

Cluster Summary

To further validate the decision to retain 3 clusters, we looked at the metrics from the Cluster Summary table were analyzed



Cluster Summary

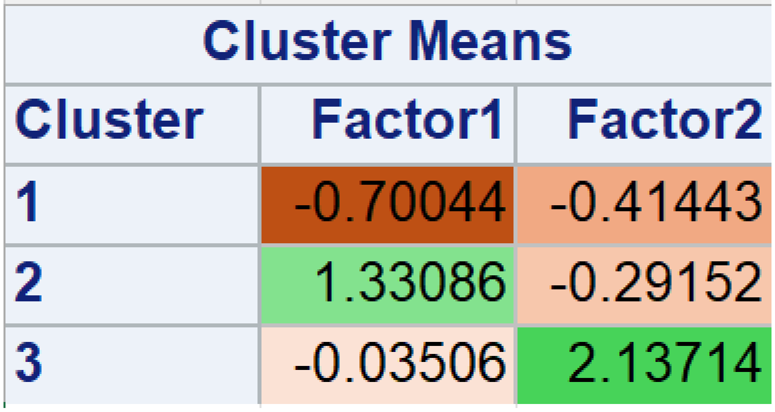
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	60	0.3790	0.9571		2	2.0350
2	32	0.3456	0.9920		1	2.0350
3	16	0.7617	2.9255		1	2.6369

Figure 17 – Cluster Summary

The variability within clusters remains consistent with the 3-cluster solution. Clusters 1 and 2 exhibit relatively low RMS deviations (0.3790 and 0.3456, respectively), confirming the homogeneity of these clusters. Cluster 3 has a higher RMS deviation (0.7617), which aligns with its smaller frequency and inclusion of more diverse data points.

Analysis of Cluster Means

The **Cluster Means** table provides insight into the distinguishing characteristics of each cluster by summarizing the average values for the variables (factors) within each group. In this case, the variables **Factor 1** and **Factor 2** serve as key dimensions for differentiating the clusters.



Cluster Means

Cluster	Factor1	Factor2
1	-0.70044	-0.41443
2	1.33086	-0.29152
3	-0.03506	2.13714

Figure 18 – Cluster Means

Cluster 1

- **Economic Development (Factor 1)**: -0.70044
- **Social Inclusion (Factor 2)**: -0.41443

Cluster 1 is characterized by negative values for both factors, indicating lower levels of both economic development and social inclusion compared to the other clusters. This cluster likely represents a group of observations associated with underperformance or challenges in both areas.

Cluster 2

- **Socio-Economic Development (Factor 1)**: 1.33086
- **Social Inclusion (Factor 2)**: -0.29152

Cluster 2 is distinguished by a high positive mean for economic development, indicating strong performance in this dimension. However, the slightly negative mean for social inclusion suggests that while these observations excel economically, they face some deficits or challenges in fostering social inclusion.

Cluster 3

- **Socio-Economic Development (Factor 1)**: -0.03506
- **Social Inclusion (Factor 2)**: 2.13714

Cluster 3 stands out for its significantly high positive mean in social inclusion, indicating strong performance in this area. The near-zero value for economic development suggests a neutral or balanced position in this dimension, with no distinct trend relative to the other clusters.

*To better understand the differences among the clusters, we visualized the cluster means for the two factors—**Economic Development** (Factor 1) and **Social Inclusion** (Factor 2).*

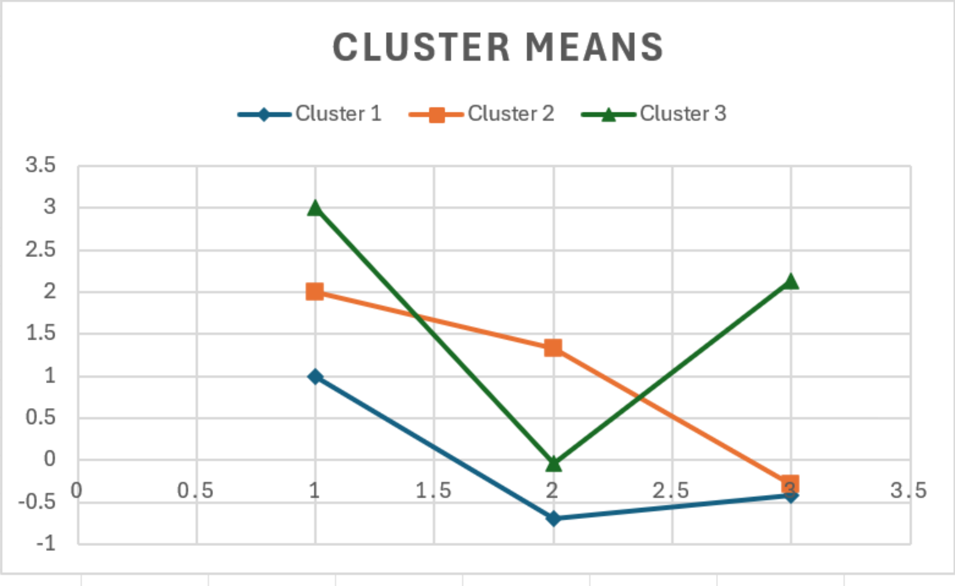


Figure 19 – Cluster Means graph

Observations from the Visualization

1. **Cluster 1**:
   - ○ The negative values for both factors are evident in the visualization, showing that this cluster is consistently below average in both economic development and social inclusion. This supports the conclusion that Cluster 1 represents entities with challenges in both dimensions.
2. **Cluster 2**:
   - ○ The high positive value for economic development (Factor 1) is clearly visible, indicating strong performance in this area. However, the negative value for social inclusion (Factor 2) is also prominent, highlighting the relative gap in social inclusion for this cluster.

3. **Cluster 3**:

   - The significant positive value for social inclusion (Factor 2) dominates this cluster, while economic development (Factor 1) remains neutral. The visualization reinforces that Cluster 3 is defined by its strong performance in fostering social inclusion.

INTERPRETATION

Each cluster consists of:

None of the countries moved between clusters during the analysis period

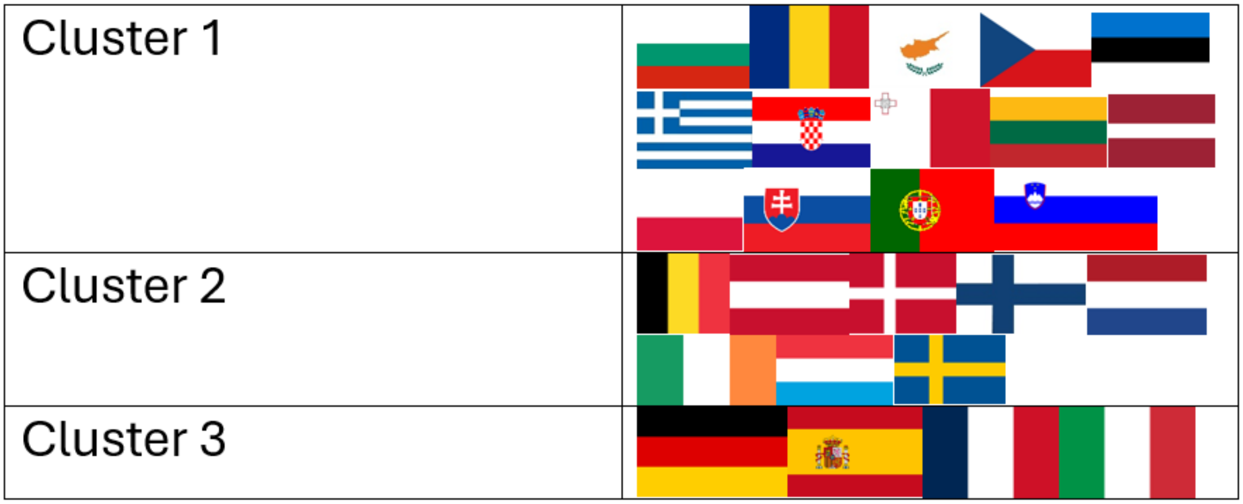


Figure 20 – Clusters

Cluster 1 (Challenged Economies and Inclusion) includes countries like Greece, Croatia, Portugal, and Slovakia, which have below average values for both economic growth and inclusive societies. These Countries often face challenges such as high unemployment, lower living standards, and limited public investment in essential services like healthcare and social programs. Additionally, these countries tend to have low immigration rates, indicating a lack of established immigrant communities and fewer opportunities for cultural integration

Cluster 2 (Economic Powerhouses with Social Gaps) comprises countries like Luxembourg, Sweden, Denmark, and Ireland, which countries in this cluster have the highest score in economic development. These countries offer high income levels, financial stability, and robust job markets, making them ideal destinations for skilled professionals seeking economic success. However, their social inclusion policies lag behind their economic performance, with barriers to integration and lower citizenship acquisition rates.

Cluster 3 (Inclusive Societies with Balanced Economies) includes countries such as Germany, Spain, Italy, and France, which were reported the highest immigration rates during the analyzed period. These countries excel in fostering inclusivity and offering progressive immigration policies, with smoother integration and access to citizenship for newcomers. Economically, they are more balanced, and they have an average value compared to Cluster 2, providing stable but less lucrative opportunities.

To illustrate the analysis, the scatter plot clearly highlights the distribution of countries within the three clusters. Cluster 1 countries, concentrated on the lower left, reflect weaker economic performance and limited social inclusion. Cluster 2 countries, positioned towards the right, demonstrate the highest levels of economic development but moderate social inclusivity. Meanwhile, Cluster 3 countries, such as Germany, are located in the upper right quadrant, combining high immigration rates and robust social inclusion with balanced but average economic development. Germany's distinct placement with high scores on both factors further reinforces its role as a leader in achieving inclusivity while maintaining economic stability. This visualization effectively captures the defining characteristics of each cluster and emphasizes the relationships between economic and social dimensions across countries.

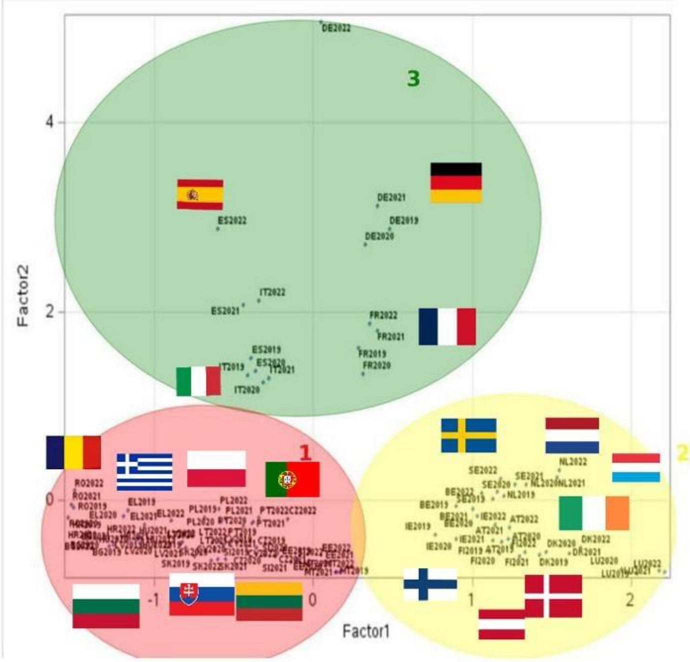


Figure 21 – Scatter Plot

Geographical Distribution of Countries

Interestingly, throughout the analysis period from 2019 to 2022, none of the countries shifted between clusters. This consistency highlights how deeply rooted economic development and social inclusion are within the structural and policy frameworks of each nation. It suggests that significant changes in these factors are slow to occur and often require long-term reforms or external influences, such as shifts in governance, global economic trends, or demographic transitions. The stability of the clusters also reflects the persistent disparities in economic opportunities and social integration policies across different regions in Europe. This observation further emphasizes the importance of sustained efforts to address systemic issues in underperforming countries and to improve social inclusion in even the most economically developed nations.

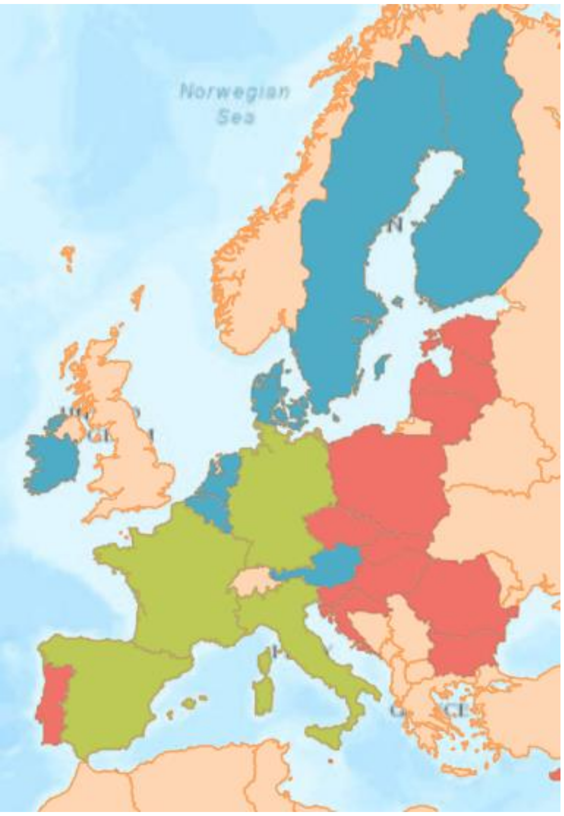


Figure 22 – Clusters map

The map also reveals a clear regional pattern, with countries in **Cluster 1 (Challenged Economies and Inclusion)**, highlighted in red, predominantly located in Southern and Eastern Europe. These nations struggle with both economic and social challenges, such as lower living standards, high unemployment, and limited immigration, which reflect systemic barriers to integration and development. Conversely, countries in **Cluster 2 (Economic Powerhouses with Social Gaps)**, marked in blue, are concentrated in Northern and Western Europe. These nations lead in economic development but exhibit gaps in social inclusion, suggesting that economic success does not always translate into societal cohesion.

Cluster 3 (Inclusive Societies with Balanced Economies), shown in green, covers parts of Southern and Western Europe, including countries like Germany, France, and Spain, which are notable for their progressive immigration policies and high levels of inclusivity. While their economic development is moderate compared to Cluster 2, their balance between stable economic opportunities and a strong focus on integration makes them appealing destinations for immigrants.

This map-driven perspective reinforces the idea that geographical and historical factors heavily influence a country's economic and social outcomes. Moreover, the absence of movement between clusters underscores the challenge of achieving rapid transformation in these areas, as economic growth and social inclusivity require sustained and deliberate efforts.

Conclusion

The primary aim of this study was to determine which European Union countries provide the most supportive environments for immigrants by examining critical economic and social factors such as median income, healthcare expenditure, corruption levels, and inclusivity policies. This analysis was designed to offer insights for individuals seeking opportunities and for policymakers working to foster integration.

Our findings revealed substantial variations among EU member states. Countries like Germany, France, and Spain stand out for their balanced approach, combining progressive immigration policies with a focus on social integration and moderate economic stability. These nations provide a welcoming environment for immigrants, ensuring access to public services, citizenship pathways, and community integration.

However, economically advanced countries such as Sweden, Denmark, and Luxembourg demonstrate strong financial systems and job markets but fall short in social inclusion due to barriers in integration policies and limited pathways to citizenship. Meanwhile, nations such as Greece and Portugal face systemic challenges, including lower economic performance and underinvestment in public services, which hinder their ability to provide opportunities for immigrants.

The study underscores that there is no single "right answer" to which country is best for immigrants, as each nation offers a unique mix of advantages and challenges. The choice depends on individual priorities, whether they emphasize economic opportunities, healthcare access, or community integration.

Ultimately, the findings highlight that economic prosperity does not always translate into inclusivity, and social inclusion requires deliberate and sustained policy efforts. For immigrants, this means weighing trade-offs between financial stability and social cohesion. For policymakers, it calls for long-term reforms aimed at addressing systemic barriers and fostering equitable opportunities for all.

This report has be done by: Ishak Soltani 20221999, Mohamed Neil 20230009, Wael meziane 20231497, Youssef nachi 20230014, Pablo Rojas 20222003

References

- Ajdin Kamber, European Parliament, & Duch, J., Guillot. (2024). Exploring migration causes: why people migrate. *European Parliament*, 4. https://www.europarl.europa.eu/pdfs/news/expert/2020/7/story/20200624STO81906/20200624STO81906_en.pdf
- Overview - Eurostat. (n.d.). Eurostat. https://ec.europa.eu/eurostat/web/migration-asylum

- Galão, M., & Jesus, F. (2024). Chapter 3: Factor Analysis [Slides]. E-Learning Nova IMS. Retrieved December 7, 2024, from https://elearning.novaims.unl.pt
- Galão, M., & Jesus, F. (2024). Chapter 3: Cluster Analysis [Slides]. E-Learning Nova IMS. Retrieved December 7, 2024, from https://elearning.novaims.unl.pt

Attachments

```
%LET LIB_DATA = res; /* Input SAS dataset library name */
%LET LIB_RES = res; /* Output SAS dataset library name */
%LET DATASET = Factors23; /* Input SAS dataset name */
%LET VARIABLES = Factor1 Factor2; /* Variables to use in the cluster analysis */
%LET ID = "Geo_time"; /* Column name for unique identifiers */
%LET ALGORITHM = WARDS; /* Hierarchical clustering method (wards, centroid, average, single, complete) */
%LET NCLUS_LIMIT = 10; /* Maximum number of clusters to form */
/* Perform hierarchical clustering using the specified algorithm
  and output the hierarchical clustering tree */
PROC CLUSTER
    DATA = &LIB_DATA..&DATASET /* Input dataset */
    METHOD = &ALGORITHM /* Specify the clustering algorithm to use */
    NOTIE /* Do not allow tied distances when clustering */
    STANDARD /* Standardize variables */
    SIMPLE /* Include simple statistics */
    NOEIGEN /* Suppress eigenvalue computation */
    RMSSTD /* Include Root Mean Square Standard Deviation */
    OUT = HC_Tree_&ALGORITHM; /* Output dataset containing the hierarchical clustering tree (dendrogram) */
ID &ID; /* Specify the ID */
VAR &VARIABLES; /* Specify the variables to be used for clustering */
RUN;
/* Create a consolidated table with RSQ values for different clustering methods,
  by joining the output from various hierarchical clustering trees */
PROC SQL;
CREATE TABLE &LIB_RES..&DATASET._RSQ_ALL_HC_METHODS AS
    SELECT DISTINCT
        A._NCL_, /* Cluster number */
        A._RSQ_ AS RSQ_CENTROID, /* R-squared for centroid linkage method */
        B._RSQ_ AS RSQ_AVERAGE, /* R-squared for average linkage method */
        C._RSQ_ AS RSQ_SINGLE, /* R-squared for single linkage method */
        D._RSQ_ AS RSQ_COMPLETE, /* R-squared for complete method method */
        E._RSQ_ AS RSQ_WARDS /* R-squared for Ward's method */
    FROM
```

```
        HC_TREE_CENTROID A /* Centroid linkage output */
    LEFT OUTER JOIN
        HC_TREE_AVERAGE B /* Average linkage output */
        ON (A._NCL_ = B._NCL_) /* Matching cluster numbers */
    LEFT OUTER JOIN
        HC_TREE_SINGLE C /* Single linkage output */
        ON (A._NCL_ = C._NCL_) /* Matching cluster numbers */
    LEFT OUTER JOIN
        HC_TREE_COMPLETE D /* Complete linkage output */
        ON (A._NCL_ = D._NCL_) /* Matching cluster numbers */
                                                                                LEFT OUTER JOIN
        HC_TREE_WARDS E /* Ward's output */
        ON (A._NCL_ = E._NCL_) /* Matching cluster numbers */
                                                                                WHERE A._NCL_ <=
&NCLUS_LIMIT; /* Filter for clusters with a number less than or equal to the specified limit */
RUN;
/* Create a plot to assess the clustering methods by plotting R-squared values */
PROC SGPLOT DATA = &LIB_RES..&DATASET._RSQ_ALL_HC_METHODS;
    /* Reverse the X-axis for a descending cluster count */
    XAXIS REVERSE LABEL = "Number of Clusters";

    /* Set up the Y-axis to represent the R-squared statistics */
    YAXIS LABEL = "R-squared";

    /* Plot the R-squared values for each method using different lines */
    SERIES X = _NCL_ Y = RSQ_CENTROID / MARKERS LEGENDLABEL = "Centroid"; /* Centroid linkage method */
    SERIES X = _NCL_ Y = RSQ_AVERAGE / MARKERS LEGENDLABEL = "Average"; /* Average linkage method */
    SERIES X = _NCL_ Y = RSQ_SINGLE / MARKERS LEGENDLABEL = "Single"; /* Single linkage method */
    SERIES X = _NCL_ Y = RSQ_COMPLETE / MARKERS LEGENDLABEL = "Complete"; /* Complete linkage method */
    SERIES X = _NCL_ Y = RSQ_WARDS / MARKERS LEGENDLABEL = "Wards"; /* Ward's method */
RUN;
```

```
%LET ALGORITHM = WARDS; /* Final hierarchical clustering method */
%LET NCLUS = 3; /* Final number of clusters to form */

PROC TREE
DATA = HC_Tree_&ALGORITHM /* Input the output dataset from PROC CLUSTER */
  NCLUSTERS = &NCLUS /* Specify the number of clusters to form */
    HEIGHT = RSQ /* Display R-squared values on the dendrogram"s vertical axis */
    OUT = &LIB_RES..&DATASET._HC_&ALGORITHM; /* Output dataset with cluster assignments at different levels */
ID &ID; /* Use the ID variable to label observations in the dendrogram */
COPY &VARIABLES; /* Copy the clustering variables to the output dataset for reference */
RUN;

/* Sort the hierarchical clustering assignment dataset by the 'CLUSTER' variable */
PROC SORT DATA = &LIB_RES..&DATASET._HC_&ALGORITHM; /* Input the output dataset from PROC TREE */
    BY CLUSTER;
RUN;

/* Standardize the hierarchical clustering assignment dataset
   to allow for standardized initial seeds */
PROC STANDARD DATA = &LIB_RES..&DATASET._HC_&ALGORITHM
    MEAN = 0 /* Set the mean to 0 */
    STD = 1 /* Set the standard deviation to 1 */
    OUT = HC_STD; /* Output the standardized dataset */
    VAR &VARIABLES; /* Specify the variables to standardize */
RUN;

/* Calculate the cluster centroids to
   define initial seeds for k-means clustering */
PROC MEANS DATA = HC_STD /* Use &LIB_RES..&DATASET._HC_&ALGORITHM to get non-standardized cluster means */
    MEAN /* Calculate the mean for each cluster */
    NWAY /* Only output the mean for non-missing data */
    NOPRINT; /* Suppress the printing of results */
    VAR &VARIABLES; /* Specify the clustering variables */
    BY CLUSTER; /* Group the data by cluster */
    OUTPUT OUT = INITIAL_SEEDS MEAN = ; /* Output the means (centroids) for each cluster as initial seeds */
RUN;
/* Standardize the original dataset for input into k-means */
PROC STANDARD DATA = &LIB_DATA..&DATASET
    MEAN = 0 /* Set the mean to 0 */
    STD = 1 /* Set the standard deviation to 1 */
    OUT = DATASET_STD; /* Output the standardized dataset */
    VAR &VARIABLES; /* Specify the variables to standardize */
RUN;

/* Perform k-means clustering using the initial seeds (centroids)
   from hierarchical clustering and the standardized dataset */
PROC FASTCLUS DATA = DATASET_STD /* Input standardized dataset */
    SEED = INITIAL_SEEDS /* Use centroids from hierarchical clustering as initial seeds */
    MAXCLUSTERS = &NCLUS /* Define the number of clusters */
    OUT = &LIB_RES..&DATASET._KMEANS /* Output dataset with k-means results */
    MAXITER = 30; /* Limit iterations to 30 */
    ID &ID; /* Use ID variable to label observations */
    VAR &VARIABLES; /* Specify clustering variables */
RUN;
/* Sort the k-means results by the cluster variable */
PROC SORT DATA = &LIB_RES..&DATASET._KMEANS;
    BY CLUSTER;
RUN;
/* Print the observations of each cluster */
PROC PRINT DATA = &LIB_RES..&DATASET._KMEANS;
    BY CLUSTER;
    VAR &ID &VARIABLES; /* Display the ID and clustering variables */
RUN;
/* Print the statistics (and generate dataset) for each cluster */
PROC MEANS DATA = &LIB_RES..&DATASET._KMEANS
NWAY /* Output only statistics grouped by clusters */
N /* Count of observations */
MEAN /* Mean */
MEDIAN /* Median */
```

```
STD /* Standard deviation */

MIN /* Minimum value */

MAX /* Maximum value */

P10 /* 10th percentile */

P90; /* 90th percentile */

    VAR &VARIABLES; /* Variables for which to calculate statistics */

    BY CLUSTER; /* Group statistics by cluster */

    OUTPUT OUT = KMEANS_STATISTICS MEAN = ; /* Save the calculated cluster statistics */

RUN;
```