APPLIED MULTIVARIATE DATA ANALYSIS PROJECT

What are the most significant dimensions shaping innovation and economic performance in EU countries?

Gabriel Calero 20240357
Hamza Tahboub 20241553
Sebastian Biolley 20241378
Mohamed Etman 20240770
Zeineb Boughattas 20241586

DECEMBER 2024

Index

Abstract

Despite the availability of various datasets exploring innovation and economic performance across European Union (EU) countries, the relationships among these dimensions and their role in driving competitiveness and development remain underexplored. Our goal is to identify the **key latent dimensions** that shape national innovation and economic performance by analyzing real-world numeric variables that reflect technological capacity, human capital development, and economic strength. To better understand the dataset, we performed a series of linear transformations and descriptive analyses, including **Factor Analysis** and **Cluster Analysis**, supported by effective visualization tools.

Among our results, we were able to identify the most significant dimensions that drive innovation and economic performance and reveal clear patterns differentiating EU countries in terms of competitiveness and development. Our findings highlight actionable strategies for fostering innovation ecosystems and provide policymakers with insights to bridge developmental disparities and enhance sustainable growth.

Keywords: Innovation, EU Countries, Technology, Human Capital, Economic Performance

Introduction

What sets countries apart in their pursuit of innovation and economic growth? Is it their capacity to embrace emerging technologies, invest in human capital, or sustain economic performance—or is it a delicate balance of all three? In today's fast-paced, interconnected world, innovation is no longer just an aspiration; it is the foundation for long-term competitiveness, economic resilience, and societal progress. However, the factors driving innovation remain nuanced, particularly within the European Union (EU), a region characterized by diversity in policy priorities, resources, and development levels.

Innovation, at its core, relies on three interconnected dimensions: technological capacity, human capital, and economic strength. Technology drives advancements in productivity and competitiveness. Human capital—nurtured through education, workforce policies, and research—acts as the engine that transforms ideas into tangible progress. Economic performance and stability, meanwhile, provide the environment where innovation can take root and thrive. Understanding how these dimensions interact within the EU is essential to addressing developmental disparities and strengthening the region's collective ability to compete on a global scale.

This study seeks to analyze the innovation and economic performance of EU countries by examining key variables that represent technological readiness, human capital development, and economic performance. Variables such as R&D personnel, government expenditure on education, AI preparedness, productivity, and GDP per capita reflect critical aspects of national innovation systems. Through Factor Analysis, we identify the latent dimensions that underlie these relationships, revealing structural patterns that drive innovation performance. Additionally, Cluster Analysis is employed to group EU countries into distinct categories based on their performance, offering insights into their relative strengths and weaknesses.

By uncovering these dimensions and identifying country clusters, this research aims to address two key questions:

1. What are the most significant dimensions shaping innovation and economic performance in EU countries?

2. How do these dimensions differentiate countries in terms of competitiveness and development?

The findings of this study provide valuable insights for policymakers, helping them understand the critical drivers of innovation and economic progress. By identifying areas of strength and opportunities for improvement, EU countries can implement targeted strategies to foster innovation, bridge developmental gaps, and maintain their competitive edge in the global economy.

1. Data

1.1. Data identification

This study aims to evaluate the innovation and economic performance of European Union (EU) countries by analyzing a set of 11 carefully selected variables for the year 2022. These variables reflect key dimensions such as infrastructure, research intensity, human capital, technological readiness, and economic performance. Data was sourced from trusted platforms, including Eurostat and the World Bank, ensuring reliability and comprehensive coverage. The dataset was merged, organized, and prepared using Microsoft Excel and analyzed through SAS Enterprise Guide to identify patterns and relationships among EU countries.

To organize and analyze the data, we utilized Microsoft Excel for preprocessing and merging, followed by statistical analyses using SAS Enterprise Guide. The table below presents the selected variables, along with their descriptions, highlighting their relevance to our research.

Data description

Variables	Description
ElecCons	Electricity consumption (% of population * 1000K): Measures energy accessibility and infrastructure
ResRD	Researchers in R&D (per million people): Indicates research capacity and innovation potential
GovEduExp	Government expenditure on education (% of GDP): Reflects public investment in workforce development
AIPrepIdx	AI Preparedness Index: Measures readiness for adopting AI technologies
InnovEconInt	Innovation and Economic Integration Index: Assesses innovation's role in economic systems
HCandLMIdx	Human Capital and Labor Market Policies Index: Evaluates workforce policies and labor efficiency
RegEthIdx	Regulation and Ethics Index: Captures governance quality and regulatory standards
GDPpc	GDP per capita (current US$): Represents economic output and standard of living
FDI	Direct investment (% of GDP): Measures foreign investment inflows and economic attractiveness
EmpRate	Employment rate: Reflects the share of the working-age population that is employed
Prod	Productivity (GDP per hour worked): Measures economic efficiency and output per worker

Figure 1 *- Data description*

1.2. Data treatment

To finalize the dataset, the remaining step was to identify and analyze outliers. Using the SAS program and the Summary Statistics task, we generated Box-and-Whisker Plots for each variable to pinpoint any outliers. While outliers often result from database errors, this was not the case here, and since they did not represent extreme values, we opted to proceed without making any modifications.

Outlier analysis revealed notable extremes: GDP per capita (GDPpc) had a maximum of 128,259.40, and FDI showed a minimum of -596.90, both reflecting economic variability rather than errors. Similarly, Productivity (Prod) had a high maximum of 132.97, while EmpRate included an unusual minimum of 0. These values were retained as they were valid observations.

Finally, we noticed substantial differences in variance across the variables. For example, GDPpc exhibited a much higher variance compared to AIPrepIdx and other more stable indices. To account for these varying scales and ensure comparability, we standardized all variables before moving forward with the analysis.

Descriptive Statistics of Variables

Variable	Mean	Std Dev	Minimum	Maximum	N
ElecCons	5878.17	2518.04	2429.17	14077.85	26
ResRD	4547.78	1915.14	985.4933	8130.79	26
GovEduExp	5.043756	1.040327	3.00598	7.5722	26
AIPrepIdx	0.668074	0.067599	0.562962	0.778522	26
InnovEconInt	0.15867	0.014935	0.135067	0.183759	26
HCandLMIdx	0.157317	0.015893	0.12634	0.184953	26
RegEthIdx	0.178652	0.029152	0.137404	0.23044	26
GDPpc	43047.9	26201.24	15797.6	128259.4	26
FDI	-17.3962	120.0975	-596.9	108.4	26
EmpRate	37.06538	10.26095	0	58.9	26
Prod	57.22407	22.35838	28.76859	132.9687	26

Figure 2 *- Descriptive Statistics of Variables*

2. Factor analysis

Factor Analysis aims to identify the underlying factors responsible for the correlations between indicators (Sharma & Mukherjee, 1996). In our case, we chose Principal Component Factoring (PCF) because our data is derived from reliable, objective sources with minimal error. To perform the analysis, we used SAS Enterprise Guide. We began by analyzing the Correlation Matrix subjectively, identifying pairs of indicators with high correlations and others with weaker correlations, which informed our decision-making process for factor extraction. Throughout this process, we ran multiple trials, removing variables based on multicollinearity, weak associations in the correlation matrix, and other rules. This iterative approach ensured that the dataset was suitable for factor analysis. Additionally, we considered the Kaiser-Meyer-Olkin (KMO) measure for sampling adequacy, which confirmed that the data was homogeneous and appropriate for Factor Analysis.

Correlations

	1	2	3	4	5	6	7	8	9	10	11
1 - ElecCons	1.000	0.735	0.552	0.613	0.518	0.449	0.662	0.523	-0.290	0.478	0.430
2 - ResRD	0.735	1.000	0.653	0.697	0.725	0.561	0.636	0.489	-0.086	0.447	0.490
3 - GovEduExp	0.552	0.653	1.000	0.538	0.635	0.390	0.467	0.014	0.269	0.228	0.001
4 - AIPrepIdx	0.613	0.697	0.538	1.000	0.797	0.884	0.961	0.638	-0.211	0.551	0.575
5 - InnovEconInt	0.518	0.725	0.635	0.797	1.000	0.569	0.686	0.415	0.101	0.478	0.445
6 - HCandLMIdx	0.449	0.561	0.390	0.884	0.569	1.000	0.816	0.565	-0.195	0.451	0.548
7 - RegEthIdx	0.662	0.636	0.467	0.961	0.686	0.816	1.000	0.659	-0.297	0.572	0.546
8 - GDPpc	0.523	0.489	0.014	0.638	0.415	0.565	0.659	1.000	-0.666	0.645	0.927
9 - FDI	-0.290	-0.086	0.269	-0.211	0.101	-0.195	-0.297	-0.666	1.000	-0.407	-0.407
10 - EmpRate	0.478	0.447	0.228	0.551	0.478	0.451	0.572	0.645	-0.407	1.000	0.533
11 - Prod	0.430	0.490	0.001	0.575	0.445	0.548	0.546	0.927	-0.407	0.533	1.000

***Figure 3**. Correlation Matrix*

We reviewed the correlation matrix to identify issues such as multicollinearity or weak associations among variables. The analysis highlights strong correlations, such as between Electricity Consumption and Researchers in R&D (r=0.73), which, despite their relationship, represent distinct aspects of societal infrastructure and innovation capacity. Similarly, the high correlation between the AI Preparedness Index and Regulation and Ethics (r=0.96) reflects their interconnected role in advancing technological readiness, while each variable emphasizes unique contributions—technological infrastructure in one and ethical oversight in the other.

Weaker correlations, such as between Government Expenditure on Education and GDP per capita (r=0.01), illustrate that some variables operate independently yet remain critical to the broader context, like societal investments in education versus direct economic output. Foreign Direct Investment, though weakly correlated with most variables, captures external economic influences distinct from internal indicators like Employment Rate or Productivity.

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.66522146										
ElecCons	ResRD	GovEduExp	AIPrepIdx	InnovEconInt	HCandLMIdx	RegEthIdx	GDPpc	FDI	EmpRate	Prod
0.7818	0.9170	0.7038	0.6483	0.6348	0.6800	0.6460	0.6328	0.3479	0.8417	0.5768

Figure 4 *- KMO's values before and after removing variables*

After that, we checked the KMO value for our dataset. As we can see in **Figure 4** we got 0.665, which is tolerable for factor analysis. ResRD had a high KMO, indicating strong suitability, while FDI had a low KMO, suggesting weaker suitability. Despite this, the overall MSA value was acceptable for proceeding with the analysis.

The next step was to determine how many factors to retain. Based on the Eigenvalues from the Correlation Matrix, we found that the Kaiser's Criteria suggested retaining 2 factors, as only the first 2 factors had eigenvalues greater than 1. According to Pearson's method, we would also retain 2 factors, and the Scree Plot method further confirmed this decision.

Eigenvalues of the Correlation Matrix: Total = 11 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	6.23990827	4.20263276	0.5673	0.5673
2	2.03727551	1.24027422	0.1852	0.7525
3	0.79700129	0.18854115	0.0725	0.8249
4	0.60846014	0.08463065	0.0553	0.8802
5	0.5238295	0.26206419	0.0476	0.9279
6	0.26176531	0.02361098	0.0238	0.9517
7	0.23815433	0.05498409	0.0217	0.9733
8	0.18317024	0.08633805	0.0167	0.99
9	0.09683219	0.08862217	0.0088	0.9988
10	0.00821002	0.00281682	0.0007	0.9995
11	0.00539321		0.0005	1

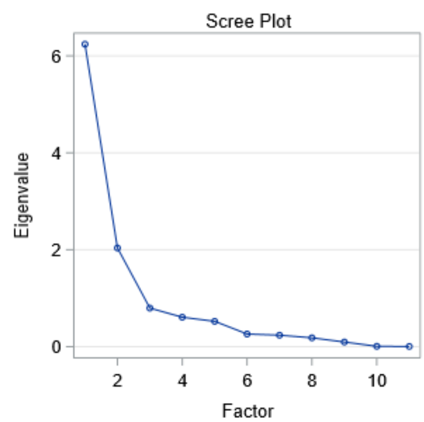

Figure 5. *Eigenvalue Table and Scree Plot for Factor Retention*

After selecting the number of factors to retain, we assessed the factor solution using the **Residual Correlation Matrix and the Root Mean Square Off-Residuals**, confirming a good fit.

Root Mean Square Off-Diagonal Residuals: Overall = 0.07557392										
ElecCons	**ResRD**	**GovEduE**	**AIPrepIdx**	**InnovEco**	**HCandLM**	**RegEthId**	**GDPpc**	**FDI**	**EmpRate**	**Prod**
0.0949	0.0695	0.0620	0.0764	0.0631	0.0942	0.0694	0.0437	0.0897	0.0599	0.0899

__Figure 6__ - Root mean square off diagonal residuals

To improve interpretability, we applied both Varimax and Quartimax rotations and found the factor solution consistent across methods. We proceeded with Varimax, ensuring the analysis captured key relationships while aligning with the study's framework.

Rotated Factor Pattern		
	Factor1	**Factor2**
InnovEconInt	**0.87174**	0.11281
GovEduExp	**0.85302**	-0.30702
AIPrepIdx	**0.84689**	0.42145
ResRD	**0.83699**	0.21147
RegEthIdx	**0.77446**	0.48323
ElecCons	**0.69218**	0.33714
HCandLMIdx	**0.69094**	0.42446
GDPpc	0.32562	**0.9172**
Prod	0.33255	**0.79366**
EmpRate	0.4294	**0.60661**
FDI	0.15712	**-0.83312**

***Figure 7**. Varimax rotated factor Pattern*

Based on the rotated factor pattern, we grouped variables according to their highest loadings to label the factors. **Factor 1** was labeled **"Innovation and Knowledge"** due to its high loadings from variables like the Innovation and Economic Integration Index, Government Expenditure on Education, AI Preparedness Index, Researchers in R&D, Regulation and Ethics Index, Electricity Consumption, and Human Capital and Labor Market Policies. These variables collectively reflect aspects of a country's readiness and capacity in innovation, knowledge, and technological development.

Factor 2 was labeled **"Economic Performance and Market Dynamics"** as it predominantly consisted of variables GDP per capita, Productivity, Employment Rate, and Direct Investment, which highlight a country's economic performance and market dynamics.

Variance Explained by Each Factor	
Factor1	**Factor2**
4.8873	3.3899

Final Community Estimates: Total = 8.277184										
ElecCons	**ResRD**	**GovEduExp**	**AIPrepIdx**	**InnovEconInt**	**HCandLMIdx**	**RegEthIdx**	**GDPpc**	**FDI**	**EmpRate**	**Prod**
0.5928	0.7453	0.8219	0.8948	0.7727	0.6576	0.8333	0.9473	0.7188	0.5524	0.7405

Figure 8. *Final Communality Estimates*

The variance explained by the two factors further supports the adequacy of this labeling, with Factor 1 accounting for **44.41%** and Factor 2 for **30.80%** of the variance. Together, these factors explain **75.21%** of the total variance.

Figure 8 shows that the variables are generally well explained by the factors. Key variables, such as GDP per capita and AI preparedness, have strong associations with the factors, highlighting their importance in innovation and economic performance. In contrast, variables like electricity consumption and employment rate have weaker associations. This indicates that the factors capture most of the variance, though with varying strength across the variables. The use of Varimax rotation ensures a balanced distribution of variance between the factors, enhancing interpretability.

3. Cluster Analysis

In the cluster analysis, we aimed to group countries based on their factor scores. Initially, we employed a hierarchical clustering approach to determine the optimal number of clusters. Once the appropriate number of clusters was identified, we used this value as a parameter to perform clustering with a non-hierarchical method, specifically k-means.

The process was applied to clustering based on both the factor scores alone and all 11 variables combined.

3.1. Cluster analysis using factor scores

For the hierarchical clustering, we evaluated various linkage methods and determined that Ward's Method was the most suitable for forming distinct clusters, based on the R-squared chart, which measures cluster heterogeneity. Furthermore, the dendrogram was used to determine the optimal number of clusters, resulting in four distinct clusters.

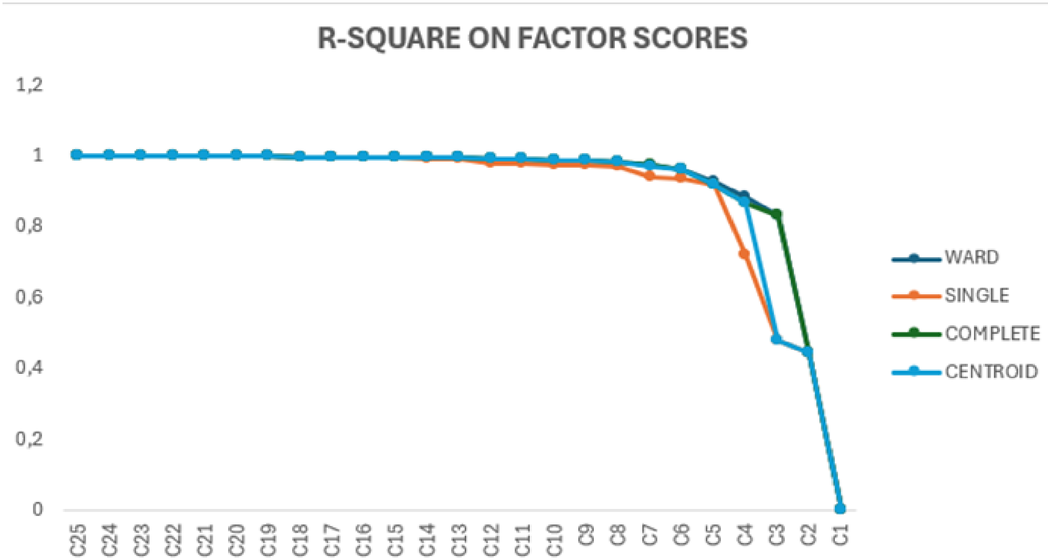

Figure 9.. R-Square on factor scores

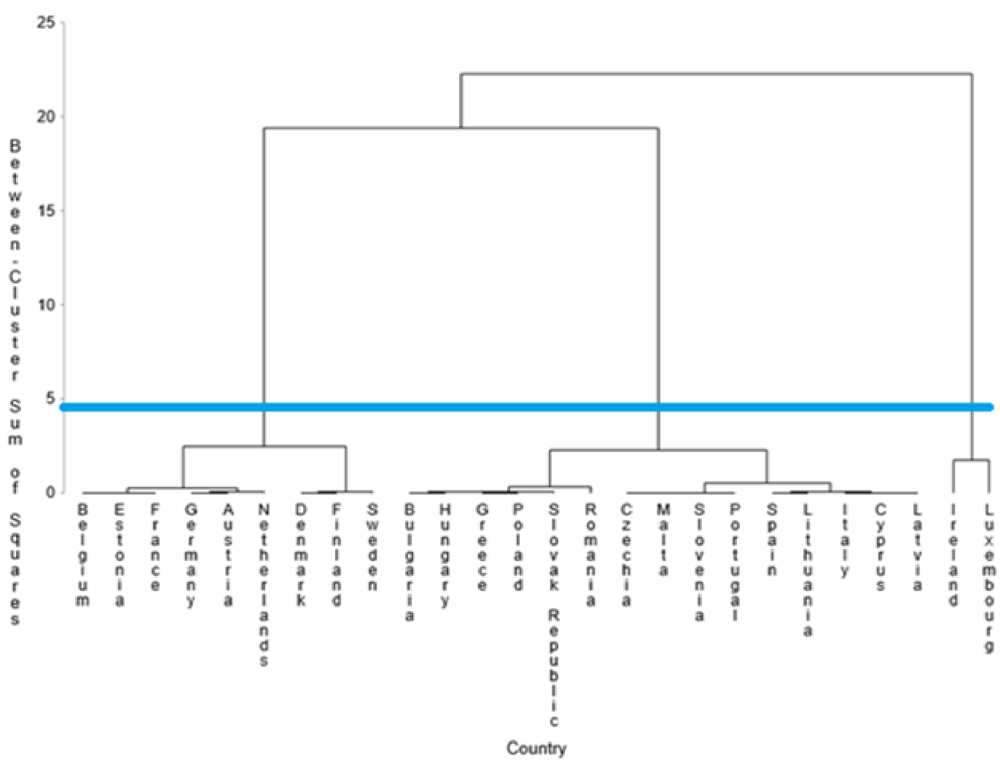

Figure 10.. Ward's dendogram on factor scores

Using the number of clusters determined from the hierarchical analysis, we applied a non-hierarchical method, specifically k-means, to group the countries into the following three clusters:

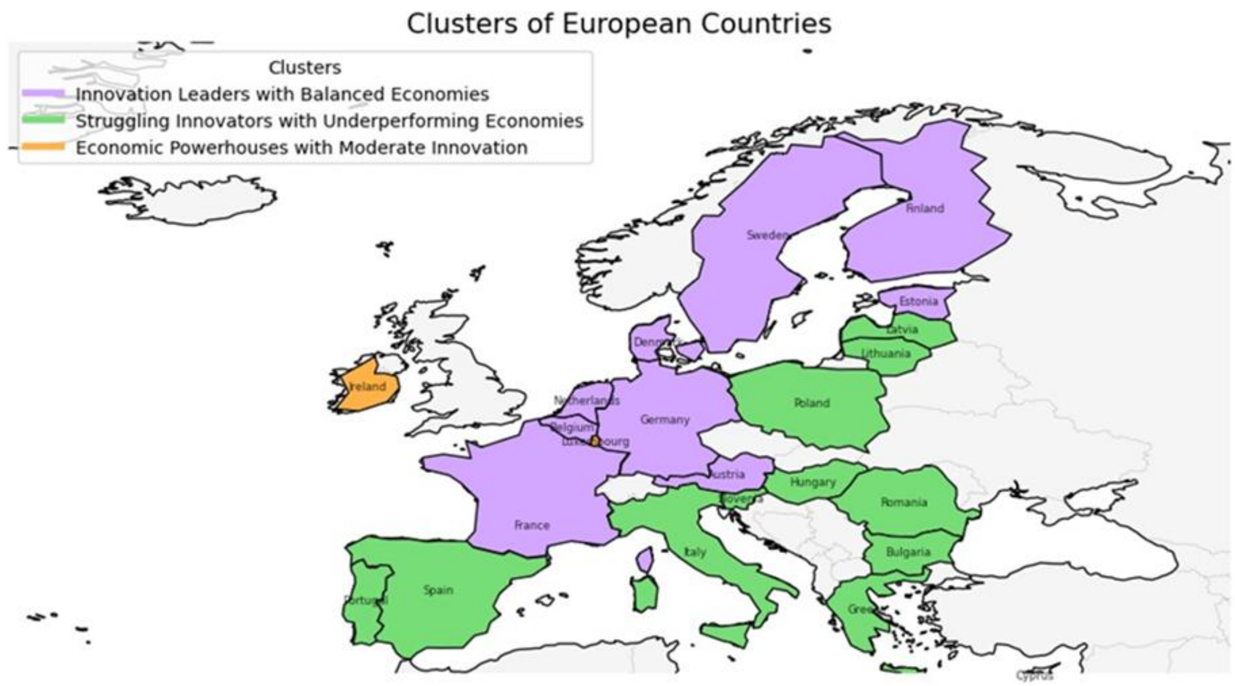

Figure 11.. Map illustrating resulting clusters

Cluster 1: Innovation Leaders with Balanced Economies:

- High performance in Factor 1 (Innovation and Knowledge Infrastructure).

- Neutral performance in Factor 2 (Economic Performance and Market Dynamics).

- These countries focus heavily on innovation and education but maintain moderate economic outcomes.

Cluster 2: Struggling Innovators with Underperforming Economies:

- Below-average performance in both factors.

- Economies in this group show limited innovation capacity and weak market competitiveness.

Cluster 3: Economic Powerhouses with Moderate Innovation:

- High performance in Factor 2 (Economic Performance and Market Dynamics).

- Moderate or below-average performance in Factor 1.
- These countries excel economically but rely less on innovation for growth.

Our findings reveal that EU countries exhibit distinct strengths and weaknesses when it comes to innovation and economic performance:

- Countries in **Cluster 1** should focus on translating their innovation efforts into stronger economic outcomes.
- Countries in **Cluster 2** need substantial policy interventions to improve both innovation capacity and economic competitiveness.
- Countries in **Cluster 3** have strong economies but may benefit from increasing investment in innovation to sustain long-term growth.

These insights can guide policymakers in crafting targeted strategies to enhance innovation capacity, improve economic performance, and reduce disparities across the EU.

3.2. Cluster analysis using the original 11 variables

On the other hand, using the original 11 variables, we first applied hierarchical clustering to determine the optimal number of clusters. Based on the dendrogram and R-squared chart, three clusters were identified as the most suitable solution. Subsequently, we used this number of clusters as a parameter for a non-hierarchical method, specifically k-means, to group the countries into three clusters.

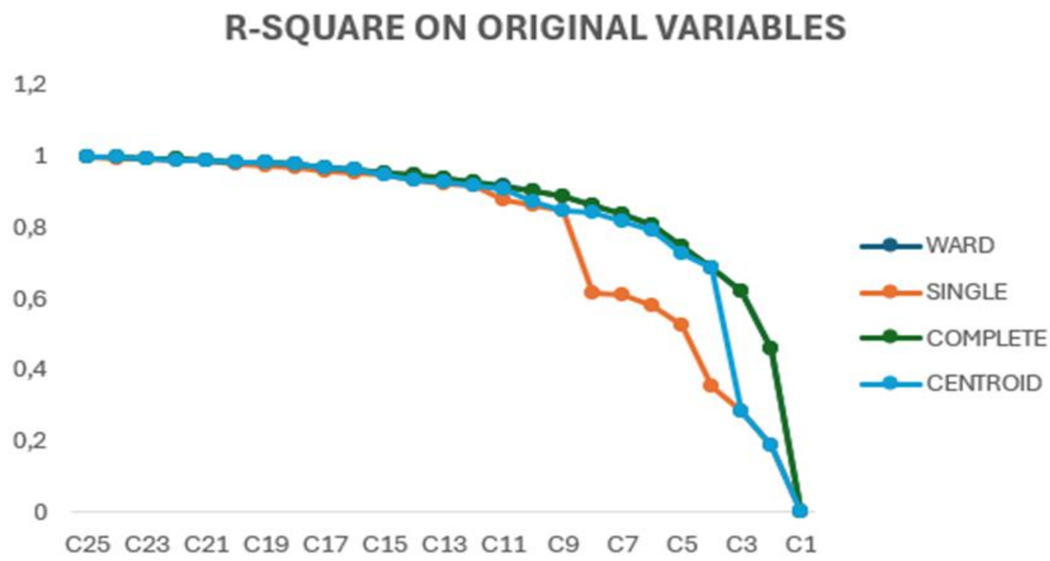

Figure 12.. R-Square plot using the original variables

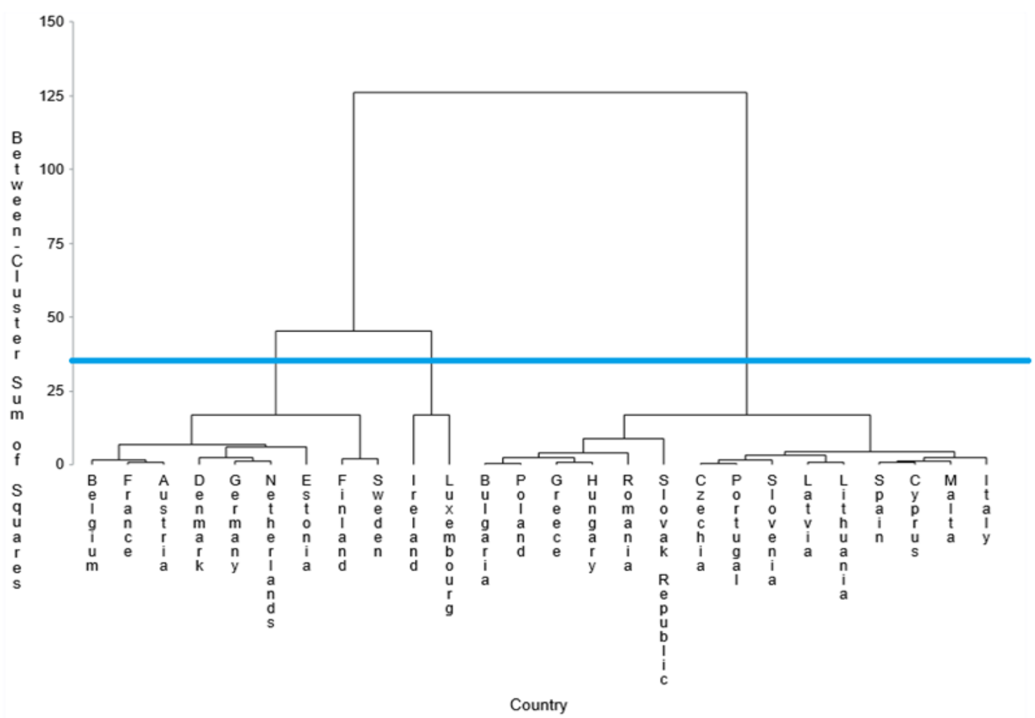

Figure 13.. Ward's dendogram using the original variables

The clustering results based on factor scores and the original 11 variables were identical. Comparing the dendrograms from both approaches revealed the same three clusters, with each cluster comprising the exact same composition of countries. This consistency confirms that the factors are a strong and accurate representation of the original data.

4. Result and discussion

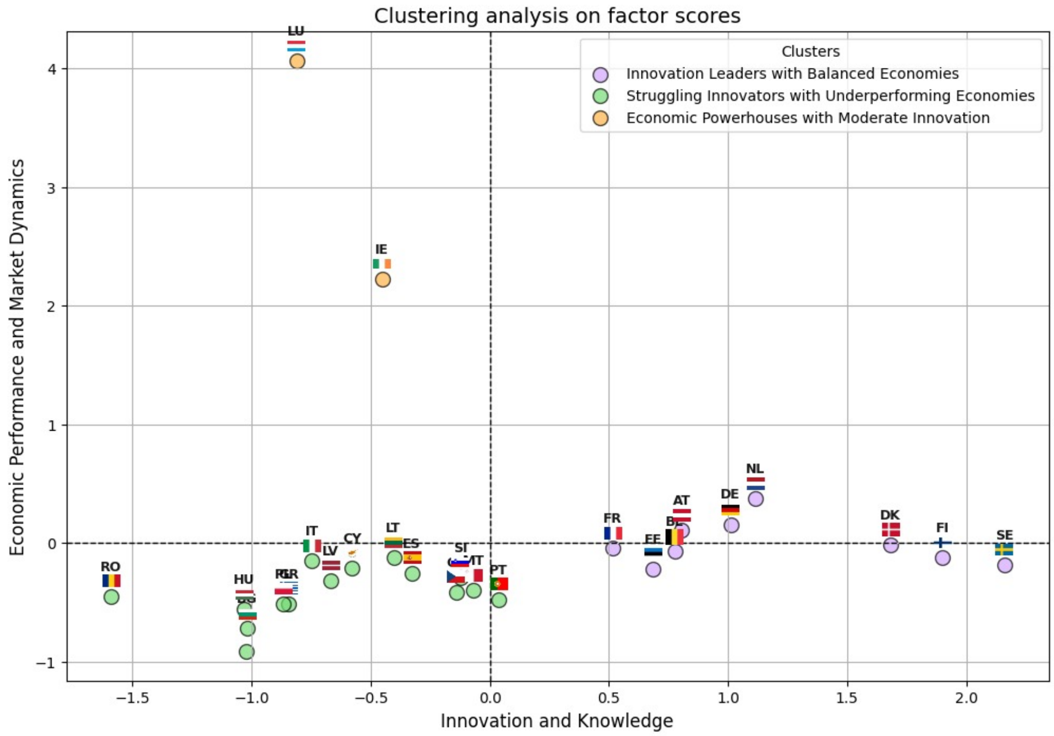

Figure 14. Economic Performance Market Dynamics vs Innovation and Knowledge

The figure above represents the correlation between **Factor 1 – Innovation and Knowledge** and **Factor 2 – Economic Performance and Market Dynamics**. By analyzing the clustering of countries within this framework, distinct patterns and relationships have been identified.

The clustering analysis presented in **Figure 12** identifies three distinct groups of EU countries based on their performance in **Innovation and Knowledge** and **Economic Performance and Market Dynamics**. The **Purple Cluster** includes countries like Sweden, Finland, Denmark, and the Netherlands, which achieve high scores in both dimensions, demonstrating balanced competitiveness through advanced innovation systems and strong economies. The **Orange Cluster**, represented by countries such as Ireland and Luxembourg, excels in economic performance, with high GDP and strong investment levels, but their moderate innovation scores indicate a focus on economic growth over innovation-driven strategies. In contrast, the **Green Cluster** consists of countries like Romania, Hungary, and Bulgaria, which lag in both dimensions due to weaker economic performance and underdeveloped innovation ecosystems.

This analysis highlights the complex relationship between economic performance and innovation, showing that high economic performance does not always correlate with strong innovation capacity, and vice versa. Countries like Sweden, Finland, and Denmark exemplify a balanced approach, excelling in both areas, while others, such as Ireland and Luxembourg, demonstrate the need for greater investment in innovation to complement their economic success. Conversely, countries in the Green Cluster face significant challenges in both dimensions, emphasizing the need for tailored policies to foster innovation and economic development simultaneously.

5. Conclusion

This study has offered a thorough examination of the economic performance and innovation of EU member states, uncovering important hidden factors that propel growth and competitiveness. We discovered two main dimensions using factor analysis: Innovation and Knowledge Infrastructure and Economic Performance and Market Dynamics. These two dimensions together explain 75.21% of the dataset's variance. These factors show how economic power, human capital development, and technology readiness interact to shape national innovation systems. EU nations were further divided into three groups by the clustering analysis: Economic Powerhouses with Moderate Innovation, Struggling Innovators with Underperforming Economies, and Innovation Leaders with Balanced Economies. These clusters highlight the various paths taken by EU countries in their quest for development and innovation.

The following are the main conclusions: 1. Sweden, Finland, and Denmark are excellent examples of balanced approaches, scoring highly in both economic performance and innovation.
2. Countries with strong economies, like Ireland and Luxembourg, have strong economic performance but a limited capability for innovation, indicating a chance to improve long-term competitiveness through investments in innovation ecosystems.

3. Low economic performance and weak innovation systems are two issues that countries like Romania, Hungary, and Bulgaria must address with major policy changes.

The study's conclusions offer practical suggestions for decision makers. The goal of nations in the Innovation Leaders cluster should be to leverage their strengths in innovation to generate long-term economic growth. Targeted tactics are necessary for struggling innovators to increase their capacity for invention and competitiveness in the market. However, in order to ensure long-term sustainability, Economic Powerhouses can gain from a greater focus on innovation. EU officials may create customized plans to close gaps, improve cooperation, and promote sustainable growth throughout the region by knowing the key forces behind innovation and economic success. These initiatives will improve the EU's overall standing in the world economy in addition to strengthening individual countries.

References

Sharma, S., & Mukherjee, S. (1996). Applied multivariate techniques.

Eurostat. (2022). Research and development statistics. European Commission. Retrieved from https://ec.europa.eu/eurostat

World Bank. (2022). World Development Indicators: Innovation and Economic Indicators. World Bank Group. Retrieved from https://data.worldbank.org

Organisation for Economic Co-operation and Development (OECD). (2022). Education, labor market, and economic indicators. Retrieved from https://www.oecd.org