Software Development

for Data Analysis

Team project

Vlad Liviu-Alexandru

Simion Cristina-Florentina

Sandu Victor-Iulian

# Data:

The entirety of the data was gathered from the [eurostat database](https://ec.europa.eu/eurostat/data/database). There are 41 observations from the EU countries in the year 2021, where only 36 made it in the final, apart from one variable detailed below.

There are 5 variables in the final dataset:

*- Life satisfaction (index):* This is the only variable to be from 2020.

*- Hours Worked (no/month)*

*- Unemployment (rate)*

*- Poverty risk (index)*

*- Graduation rate (computed rate):* It is computed as Students enrolled (ED6 + ED7) / First-time graduates (ED6 + ED7)

We took 5 variables that are somewhat related to the general work-life of individuals but far apart enough as to minimize collinearity. With the data at hand, we had multiple paths we could take for the analysis, the one we chose is to try to find a way to group them, visualize them and conclude what the underlying correlation between them is, if there is any.

Firstly, to accomplish this goal, we chose the PCA analysis in order to reduce the complexity and make it possible to visualize, paint a better picture of what the dataset represents and the collinearity present in the underlying data. PCA allows you to actually plot a 5-dimensional dataset and also understand how distributed the data is along those dimensions by looking at the PC percentages.

Secondly, we chose to use EFA as it builds on PCA and tells us if a division into factors or categories is feasible for the dataset as we initially assumed.

# Results:

### PCA:

PC1: The countries with higher positive values on PC1 include Bulgaria, Greece, and Turkey, while countries with lower values include Denmark, Belgium, and Switzerland. The first principal component shows a distinction in economic factors. Countries like Bulgaria, Greece, and Turkey have higher positive values, indicating potential economic challenges, while Denmark, Belgium, and Switzerland have lower values, suggesting higher economic stability.

PC2: This component appears to represent a contrast between Southern European countries (e.g., Greece, Italy, Spain) and Northern European countries (e.g., Sweden, Finland, Denmark). This could be associated with underlying socio-economic differences or regional trends.

PC3: This component seems to be associated with variations in poverty risk and graduation rates, countries with higher values contribute positively to this component.

PC4: Shows variations related to unemployment and life satisfaction. Countries like Greece and Spain have higher values, while Denmark and Switzerland have lower values. Greece and Spain, with higher values on this component, may face challenges in life satisfaction compared to countries like Denmark and Switzerland.

PC5: Is showing variations in hours worked. Countries with higher values include Turkey, Greece, and Bulgaria. Which have higher values on this component, suggesting potential challenges in these areas.

In conclusion, the PCA results suggest that the European countries can be characterized and differentiated based on the patterns of variation captured by the principal components. Different countries contribute differently to these components, allowing for a nuanced understanding of their positions in the multidimensional space defined by these variables.

### EFA:

The estimated inverse covariance determinant had a value of 0.9725. A value close to 1 suggests that there is no multicollinearity issue, and the variables are not highly correlated. This is a favorable condition for factor analysis.

The Bartlett sphericity test shows whether the variables are correlated, indicating if factor analysis is appropriate. The result was a fail to reject the null hypothesis (P-Value = 0.9999999275, chi-sq 0.197) which suggests that there is significant correlation among the variables, supporting the suitability of factor analysis.

Keiser-Meyer-Olkin Measure results in a KMO index is 0.2634, indicating an *unacceptable* level for factor analysis. The KMO index assesses the adequacy of the sample size for factor analysis, with values closer to 1 indicating better suitability. The low index value suggests that the variables may not be well-suited for factor analysis, and other techniques might be more appropriate. Individual KMO indices for each variable fall between 0.0673 and 0.2626, further indicating that some variables might not contribute sufficiently to the factor analysis due to low sampling adequacy.

In summary, the factor loadings indicate potential relationships among variables, but the low KMO suggests caution in the interpretation of the results. The Bartlett Sphericity Test supports the presence of correlation among variables, but for a factor analysis some consideration in picking the data would be needed if the analysis was to go further.

# Conclusion:

The analysis was relatively successful in providing the insight we were looking for and managed to better show the division of the countries by looking at the PCA results. On the other hand, EFA was less successful, as it reinforced the presence of correlation but failed on the KMO test, thus, a factor analysis and division is not very suitable on this dataset in its current form.