THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

THE FACULTY OF CYBERNETICS, STATISTICS AND ECONOMIC INFORMATICS

ECONOMIC INFORMATICS SPECIALIZATION

Software Development for Data Analysis Project

COORDINATOR:

Conf. Univ. Dr. Claudiu VINȚE

STUDENT:

Burada Andreea-Daniela

Group 1096, G Series

BUCHAREST

2022

# Description of the topic and selected data

The main purpose of the analysis is to identify correlation between different educational levels, risk of poverty and unemployment rate at European level. The data has been collected from EUROSTAT. It includes 7 statistical variables:

* No. of primary education enrolments – number of students enrolled
* No. of lower secondary education enrolments – number of students enrolled
* No. of upper secondary education enrolments – number of students enrolled
* No. of Bachelor’s enrolments – number of students enrolled
* No. of Master’s enrolments- number of students enrolled
* Percentage of persons at risk of poverty or social exclusion – percentage of total population
* Unemployment rate – percentage of total labour force

The data was collected in 2020 and the observations are represented by the current 27 countries members of the European Union.

Additionally, the classification of all countries by European region is provided in the “regions.xlxs” file in order to compute the data grouped by regions.

# Explained Variance by the Principal Components

Chart, line chart

Description automatically generated

Given that not all variables are the same unit, the matrix of values extracted from the Excel file needs to be standardized. The variance explained by each principal component constitutes the eigenvalues and the largest variance is explained by the first principal component denoted as C1. It is important to note that all principal components are independent and have no correlation, whereas the original variables may be correlated.

# The correlogram of Kaiser-Meyer-Olkin indices

Table

Description automatically generated

The Kaiser–Meyer–Olkin (KMO) test is a statistical measure **to determine how suited data is for factor analysis**. The test measures sampling adequacy for each variable in the model and the complete model. The statistic is a measure of the proportion of variance among variables that might be common variance.

Following the computation of the indices, we have the following interpretation: only one index is of average factorability, the rest being mediocre or below. This may indicate that the sampling may not be adequate.

# Representation Quality of Observations on the Factors Axis

Graphical user interface, application

Description automatically generated with medium confidence

For the first three factors we have values different than 0 for the quality representation. In the first column we can observe that countries like Latvia, Lithuania and Poland have the highest values, which may indicate that they have a higher number of students at primary education level.