Bucharest University of Economic Studies

Faculty of Cybernetics, Statistics and Economic Informatics

Software Development for Data Analysis Project

Coordinating teacher

Conf. univ. dr. VINȚE Claudiu

Student

Carp Cosmin

Bucharest

2023

The main aim of the project is to analyze the evolution of European countries by studying several variables. The variables taken into consideration are:

- The employment (thousands of persons)
- The unemployment rate (%)
- The gross average salary (euro/year)
- Persons with a bachelor's degree or equivalent (number of persons)
- Passengers carried by air (number of persons)
- Consumption on cultural goods (PPS – Purchasing Power Standard)
- Tourist arrivals (number of persons)
- Internet usage per individual (%)
- Mean consumption expenditure (euro/inhabitant)
- Crime index
- Literacy rate (%)

The main source of data is the Eurostat Database. The data gathered from Eurostat is saved as an Excel spreadsheet, which was then converted into a .CSV file.

In order to classify the countries, we want to reduce the dimensionality, i.e., finding some relevant indicators for analysis, synthetic indicators reduced from the initial ones, based on which the data will be easier to interpret. To achieve this, it is used the **Principal Component Analysis** (PCA) technique based on the correlation matrix. New components will express new attributes of the countries and are built in a way so that they are uncorrelated with each other, each of these new variables being a linear combination of the original variables.

The analysis of the data gathered is done using Python. The project is divided into directories, packages and different functions which fulfill different roles, such as data analysis and the creation of graphs which help visualize the data. The directory *"dataIN"* contains the Excel spreadsheet and the .CSV file, the directory *"dataOUT"* contains the data which is processed using Python's data analysis tools which is then saved as .CSV files. The package *"pca"* is made of a class which implements the Principal Components Analysis. Some libraries which were used are *numpy* and *pandas*.

The first step is to compute the descriptive analysis of the variables. For this, we calculate the central tendency indicators, such as mean, minimum value and value maximum, standard deviation.

	MIN	MAX	MEAN	STD. DEVIATION
Employment (thousand persons)	258	39148	6722	9203,85
Unemployment rate (%)	1,9	9,6	4,37	1,79
Gross average salary (euro/year)	10272	74076	31221	17431,09
No. of persons with a bachelor's degree	3056	2002583	383637	481757,21
Passengers carried by air	419346	91898241	17818837	24201178,06
Consumption on cultural goods (PPS)	11422	49838	25635	9285,66
Tourist arrivals	853436	117442342	21119666	30690194,73
Internet usage per individual (%)	75	99	90	6,62
Mean consumption expenditure (euro/inhabitant)	3447	38424	15547	9786,54
Crime index	22,65	52,41	34,50	8,34
Literacy rate	94,07	100	98,74	1,45

Figure 1. Descriptive statistics

From the table displayed above, we can conclude that the literacy rate has, on average, a high value, as well as the internet usage per individual. The difference between the mean gross salary and the extreme values is rather large, meaning that there are substantial differences in the salaries from country to country. The standard deviation of number of tourists ais substantial as well, meaning that there are countries which have far more visitors than others. It can be observed that the standard deviations of the variables have quite different values. Data standardization is required.

Next, the data gathered in the .CSV file is read using the *read_csv* function from the *pandas* library and stored as a matrix. Because the data is expressed in different units of measure, it needs to be standardized. The aim of this step is to standardize the range of the initial variables so that each one of them contributes equally to the analysis. This is done by subtracting the mean and dividing by the standard deviation for each value of each variable. The main properties of standardized data are: the standard deviation is equal with 1, the mean is equal with 0, and the covariance matrix is equal with the correlation matrix. Once the standardization is done, all the variables will be transformed to the same scale. After the data is standardized, it will be saved as a pandas DataFrame called "*table*". The standardized DataFrame is exported as an .CSV file which will be found in the "*dataOUT*" directory.

Following standardization, the previously expressed variables in different measurement units, are now perfectly comparable between them, and country rankings can be made depending on each individual variable. For example, the country with the greatest number of employees is Germany, while the country with the least number of employees is Malta. As for the gross average salary, Denmark holds the greatest value, while Serbia the lowest.

The next step is to compute the variance-covariance matrix. The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to

each other, or in other words, to see if there is any relationship between them. The matrix is also stored in the "*dataOUT*" directory and then the correlogram of the variance-covariance matrix will be created and shown.

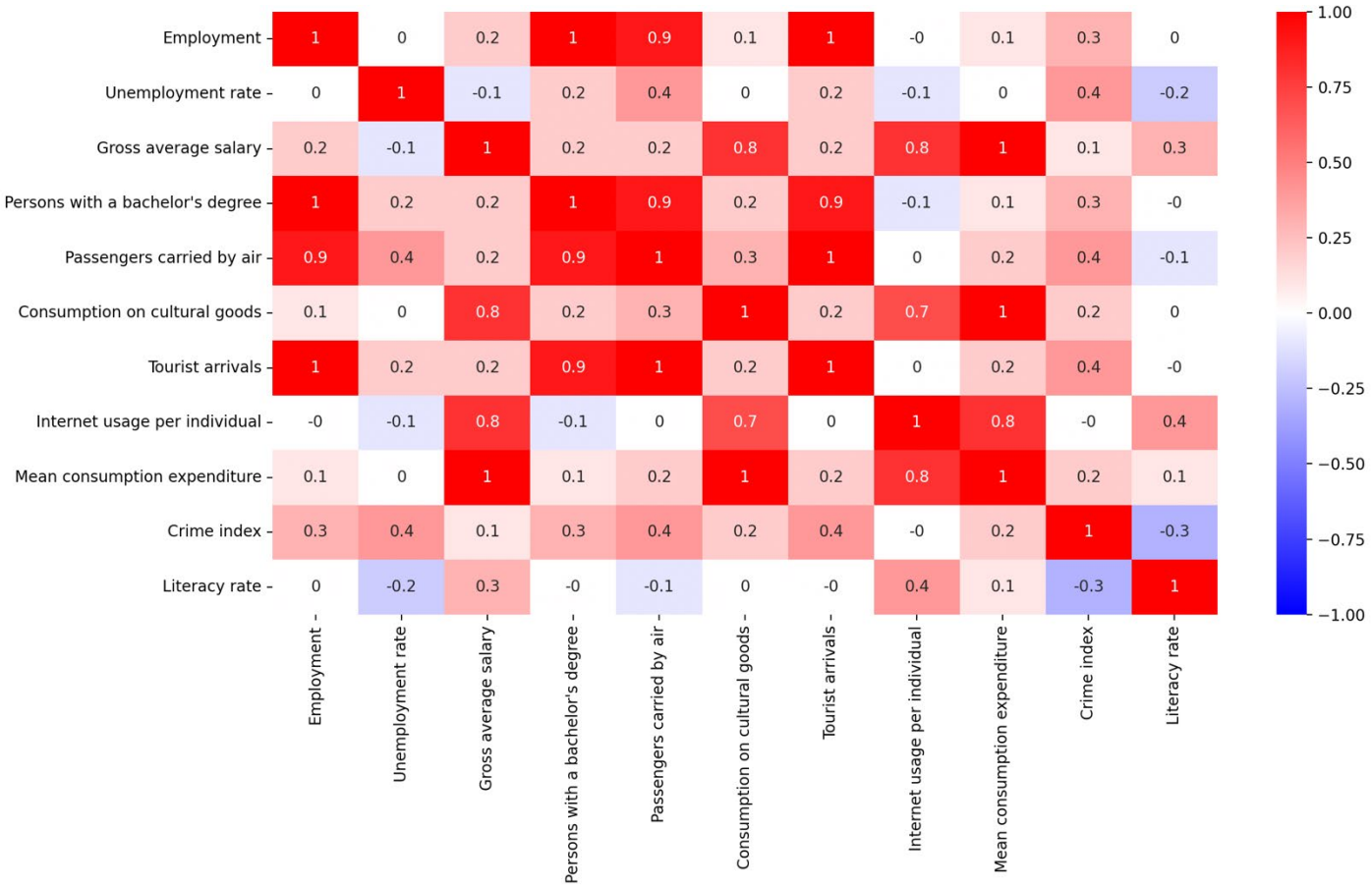

Figure 2. Variance-covariance matrix

From the analysis of the variance-covariance matrix, we can conclude that some of the highest correlations are between the gross average salary and mean consumption expenditure, persons with a bachelor's degree and employment (meaning that the number of employees is strongly related to the number of persons with a superior studies), persons carried by air and tourist arrivals (meaning that most tourist travel by plane), mean consumption expenditure and the consumption on cultural goods (meaning that a change in the mean consumption of cultural goods will have a significant impact on the mean consumption expenditure). In addition, we can conclude that the unemployment rate, literacy rate and the crime index have little to no effect on the other variables.

Moving forward, the next step is to compute the eigenvectors and eigenvalues of the variance-covariance matrix to identify the principal components. Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables.

These combinations are done in such a way that the new variables are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components. As there are as many principal components as there are variables in the data, principal components are constructed in such a manner that the first principal component accounts for the largest possible variance in the data set. A scree-plot is used to determine a suitable number of principal components. Eigenvalues are displayed on the y-axis, while the number of principal components is displayed on the x-axis. In order to decide which principal components will be kept in our analysis, we will use the Kaiser criterion. The Kaiser criterion states that components based on eigenvalues greater than 1 should be retained (their variance is greater than 1). In our case, according to the scree-plot displayed below, the number of retained principal components is 3.

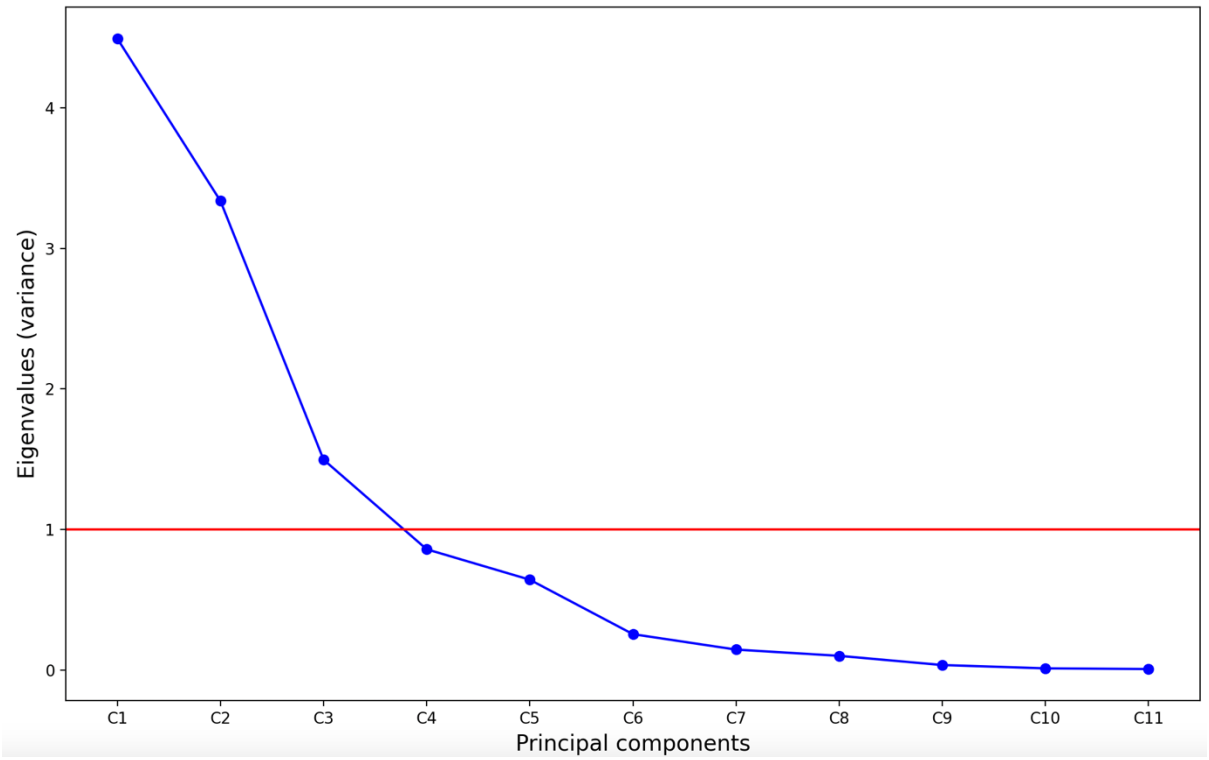

Figure 3. Scree-plot of the principal components

Then, the correlogram of scores is created. Scores are the standardized values of the principal components.

Lastly, the factor loadings are computed. These represent the correlation between the observed variables and the principal components. Based on this matrix and its graphical representations it can be determined the information contained in each principal component. The bigger the value of the correlation coefficient between an original variable and a principal component, the more adequate and complete is the informational expression of the original

variable through the respective. The matrix of factor loadings is stored in the "*dataOUT*" directory.

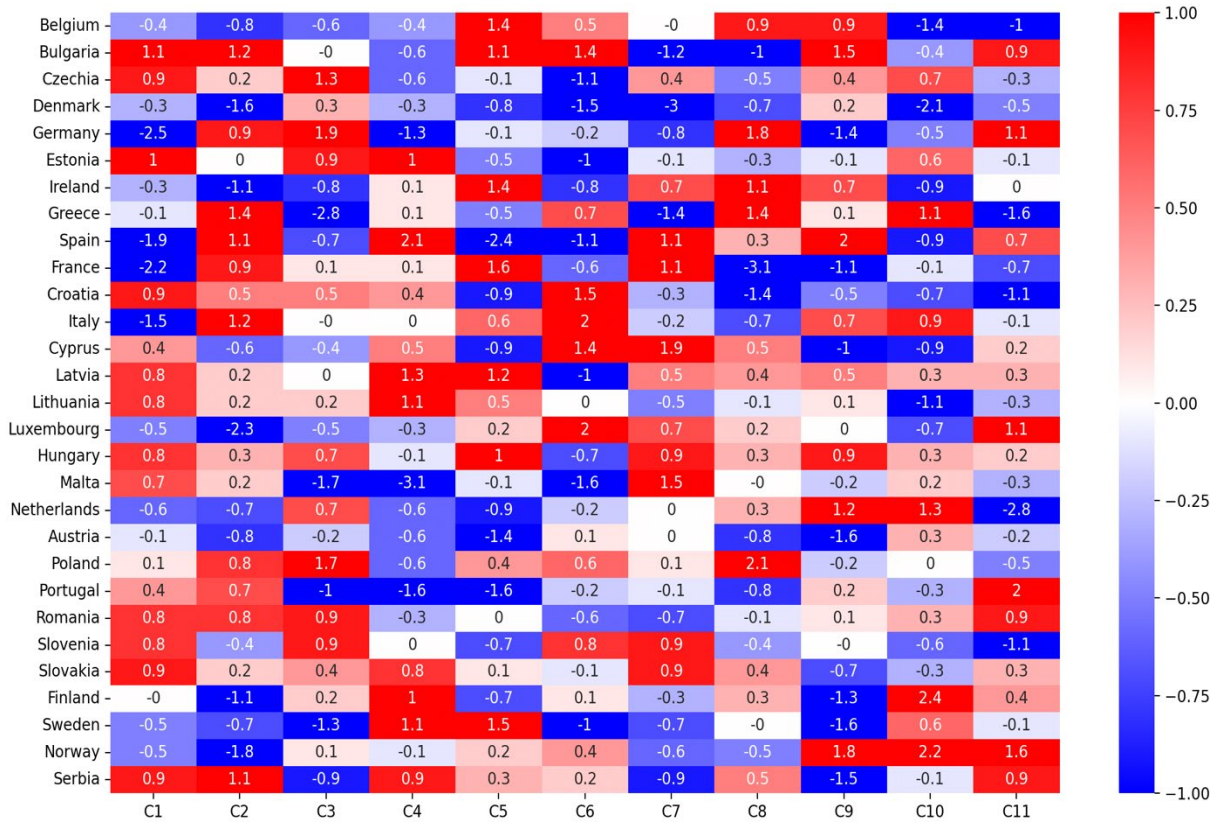

Figure 4. Scores

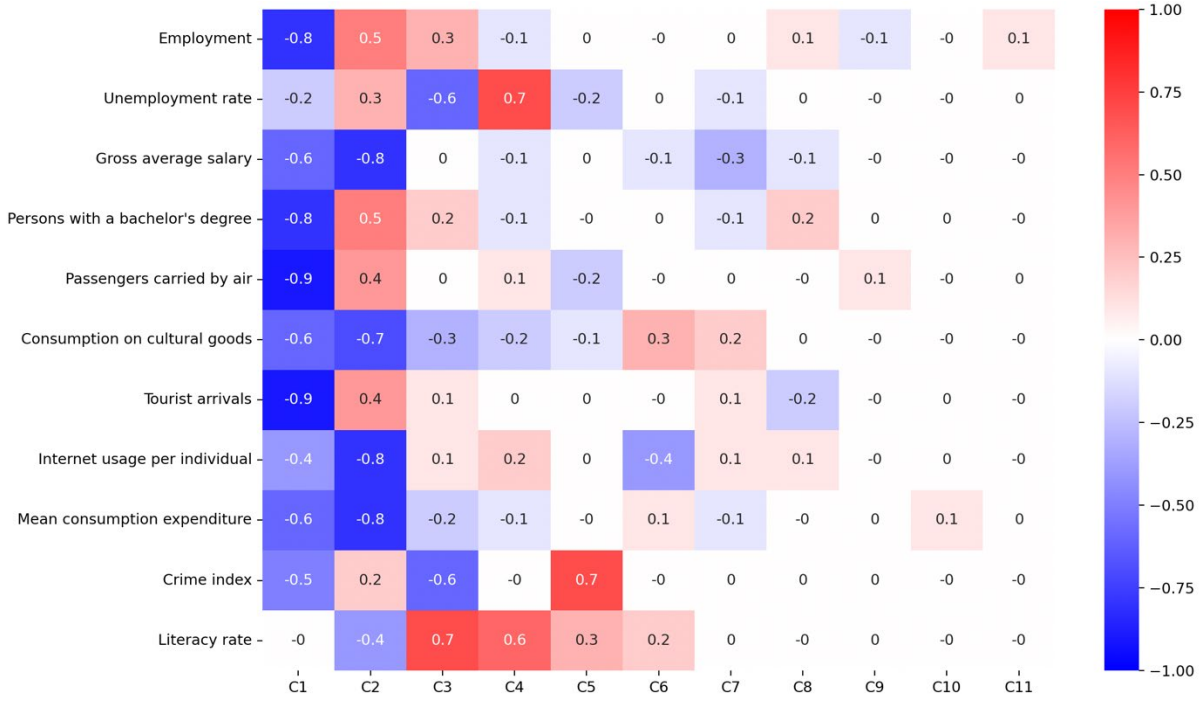

Figure 5. Factor loadings matrix

The first principal component is composed of employment, persons with a bachelor's degree, passengers carried by air and tourist arrivals. The second principal component is made up of the gross average salary, mean consumption expenditure, internet usage per individual and consumption on cultural goods. Lastly, the third and final principal component is comprised of the unemployment rate, literacy rate and crime index.

In conclusion, based on the analysis of the data carried out on 29 countries and on the 11 influencing factors, we were able to establish which are the most developed countries and what characteristics each country has. In order to prepare the data for the analysis of the principal components, the data was standardized and centered. In the analysis carried out on this sample, it could be observed that some variables were strongly correlated with each other, so that they did not bring more relevant information. By using the principal components analysis, we were able to reduce the system to only 3 principal components which describe the data, so that the informational losses are minimum.