# Utilisation of EU Employment Data in Lecturing Data Mining Course

Jan Masopust, Zdena Dobesova(✉), and Karel Macků

Department of Geoinformatics, Faculty of Science, Palacky
University, 17. listopadu 50, 779 00 Olomouc, Czech Republic
{jan.masopust01,zdena.dobesova,karel.macku}@upol.cz

**Abstract.** This article describes the utilisation of Eurostat employment data in the Data Mining course. The course is the obligatory course for a master degree Geoinformatics and Cartograhy study program at Palacký University in Olomouc. The article shows an example of the implementation of several methods like correlation, principal components analysis, k-means and hierarchical clustering on the same dataset in the course's teaching. The processing data in the Orange software and following interpretation of results gained by these methods are explained to students. Moreover, students create the MS PowerBI dashboard based on the same data. Teacher final finding is that the use of the current European data is for students more illustrative and increases their awareness of the status of employment in European countries within economic activities categorised by NACE. Practical processing of real data brings a deeper understanding of the lectured topics. Presented outputs, such as clustering, discover similar countries according to the same sector of employment.

**Keywords:** Education · Data Mining · European Union · Eurostat · NACE · GIS · MS PowerBI · UrbanDM

## 1 Introduction

The educational aspect of teaching data science-oriented courses faces many challenges. One of them lies in selecting relevant sample data that are sufficiently illustrative and contribute to an easier understanding of the presented methods. The authors of the article lecture the course Data Mining in the first grade of master study program Geoinformatics and Cartography at Palacký University in Olomouc, Czech Republic. The Data Mining course syllabus consists of basic topics like data pre-processing, multi-dimensional statistics methods (such as clustering or principal component analysis), decision trees, association rules and time series analysis. The student's evaluation of previous courses brought criticism of repetitive using of the sample dataset Iris. Iris flower dataset is probably one of the best-known multivariate sample datasets for many statistical techniques such as classification or clustering. The dataset consists of 150 samples from three species of Iris with four measured characteristics. Many manuals and teaching materials for data mining software frequently use this dataset as a transparent example.

R. Silhavy (Ed.): CSOC 2021, LNNS 229, pp. 601–616, 2021.
https://doi.org/10.1007/978-3-030-77445-5_55

Nevertheless, in the field of geoinformatics, students would prefer more geographically oriented data. As a response to student evaluation, the new practical examples were prepared using the EU (European Union) data. There are dozens of interesting themes in the Eurostat database [1], so we decided to employ this data source to bring a more European perspective into the Data Mining course. The data about employment in various sectors in the EU states were selected for practical exercises to make taught data mining methods easier to understand. Simultaneously, students' subconsciousness about the structure of employment on a European scale is raising.

The finding of new ways of education is a topical theme for pedagogy at universities. Real data and a combination of new methods are mentioned in the literature [2, 3]. The article's authors have a positive experience using a practical example to boost students' knowledge in other courses. Contributive experience is mentioned in the article about the relational database design for the botanical garden plant database under the BotanGIS project [4]. Another example is from the same Data Ming course about constructing a decision tree by self-gathered data [5].

In the practical example demonstrated in this paper, the main goal was to confirm differences between west and middle/east European countries using employment data. From a historical perspective, the separation after the Second World War has influenced industries' development in two parts of Europe for near 45 years in the 20th century. East Europe has been oriented more to agriculture and heavy industry and (primary and secondary sector); west Europe has emphasised the public services, science and research (tertiary and quaternary sector). Based on this historical development, the research question arises: how have the following 30 years (since the 90s of the 20th century and the beginning of the 21st century) changed employment structure in economy sectors? Do the disparities remain or have they disappeared?

This article presents practical examples using data about employment in different sectors of economic activities in Europe. Related to the Data Mining course, several sample tasks were presented as they are taught in the course, including interpreting results.

## 2 Data and Methods

For successful lecturing, it is crucial to choose software and training data wisely. The next chapters briefly explain used data source Eurostat database, Statistical Classification of Economic Activities NACE and practical examples in ORANGE software.

### 2.1 Eurostat Database

Eurostat is the main statistical body of the European Union. Its main task is to provide harmonised statistics at the EU level at both national and regional levels to compare different areas in selected topics. All data published by Eurostat are obtained from the statistical authorities of individual member states (national statistical offices). Eurostat itself is primarily concerned with the collection of these data and their harmonisation into a comparable form. Unfortunately, this approach is also the main weakness, as member states are only required to supply a limited amount of statistical data. The remaining part is not legally enforceable, so their availability depends only on the member state's willingness.

For this reason, much of the data is regionally or temporally unavailable. Most of the European statistics is freely disseminated in the Eurostat Database. Eurostat also provides microdata, such as data from sample survey EU-SILC (European Union Survey on Income and Living Condition). This data is only available on request, consisting of verification of the applicant entity and submission of a research plan in which the data is planned to be used.

### 2.2 Statistical Classification of Economic Activities - NACE

The Statistical Classification of Economic Activities in the European Community, commonly referred to as NACE, is the industry-standard classification system used in the EU. The Regulation establishing NACE Rev. 2 was adopted in December 2006 [6].

NACE uses four hierarchical levels:

- Level 1: 21 sections identified by alphabetical letters A to U;
- Level 2: 88 divisions identified by two-digit numerical codes (01 to 99);
- Level 3: 272 groups identified by three-digit numerical codes (01.1 to 99.0);
- Level 4: 629 classes identified by four-digit numerical codes (01.11 to 99.00).

Eurostat offers data under *National accounts employment data by industry* section [7], online data code NAMA_10_A64_E. Two levels of detail - Level 1 and Level 2 are reported. For a practical example, Level 1 with 21 categories was selected. This level was downloaded in the format for MS Excel spreadsheet. The reported unit of measure is a thousand persons. This unit is incomparable for searching for similar countries and unsuitable for particular methods like clustering. The purpose was to find similar countries with a similar or equal structure of employee by economic activities. Students recalculate the original values from a thousand persons to the relative value - the percentage of each country's total number of employed persons in lectures.

Some statistics are available for countries outside the EU, but they suffer from a certain level of incompleteness. In total, countries Montenegro, Serbia, Liechtenstein, and Malta were excluded from the processed dataset, and 30 countries remain for father processing. The reported year 2018 was selected as a reference point for a practical example. The data for 2019 is also accessible, but many values are labelled as *(p) provisional*. The detection of the incompleteness of data is the first practical experience for students. Recalculation data and solving incompleteness belongs to the introductory lesson about pre-processing data in the Data Mining course.

Table1 shows the classification and description of economic activities at Level 1.

**Table 1.** NACE classification at Level 1

| Code | Description |
|---|---|
| A | AGRICULTURE, FORESTRY AND FISHING |
| B | MINING AND QUARRYING |
| C | MANUFACTURING |
| D | ELECTRICITY, GAS, STEAM AND AIR CONDITIONING SUPPLY |
| E | WATER SUPPLY; SEWERAGE, WASTE MANAGEMENT AND REMEDIATION ACTIVITIES |
| F | CONSTRUCTION |
| G | WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES |
| H | TRANSPORTATION AND STORAGE |
| I | ACCOMMODATION AND FOOD SERVICE ACTIVITIES |
| J | INFORMATION AND COMMUNICATION |
| K | FINANCIAL AND INSURANCE ACTIVITIES |
| L | REAL ESTATE ACTIVITIES |
| M | PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES |
| N | ADMINISTRATIVE AND SUPPORT SERVICE ACTIVITIES |
| O | PUBLIC ADMINISTRATION AND DEFENCE; COMPULSORY SOCIAL SECURITY |
| P | EDUCATION |
| Q | HUMAN HEALTH AND SOCIAL WORK ACTIVITIES |
| R | ARTS, ENTERTAINMENT AND RECREATION |
| S | OTHER SERVICE ACTIVITIES |
| T | ACTIVITIES OF HOUSEHOLDS AS EMPLOYERS; UNDIFFERENTIATED GOODS- AND SERVICES-PRODUCING ACTIVITIES OF HOUSEHOLDS FOR OWN USE |
| U | ACTIVITIES OF EXTRATERRITORIAL ORGANISATIONS AND BODIES |

## 3 Practical Example

The introductory chapters and theoretical presentation for the Data Ming course's lecturing are based on books by P. Pavel and J. Šarmanová [8, 9]. The examples presented in these lecture-books are mostly oriented on medical care, insurance, and the financial sector. Therefore, the theme of employment by economic activities in the EU, which belongs to economic geography, is more suitable for Geoinformatics students.

### 3.1 Example EU Employment in Software ORANGE

The WEKA software has been used in the Data Mining course for five years due to its free access. Alternatively, software R is practised. Starting at last academic year, 2019/2020, Orange software has been implemented into practical lessons. Orange is freely available [10, 11]. The interface is user friendly and easy to handle using the drag and drop approach. Figure 1 shows an example of a workflow containing several data mining methods. This workflow presents all practised methods that are described in the following section of this article.

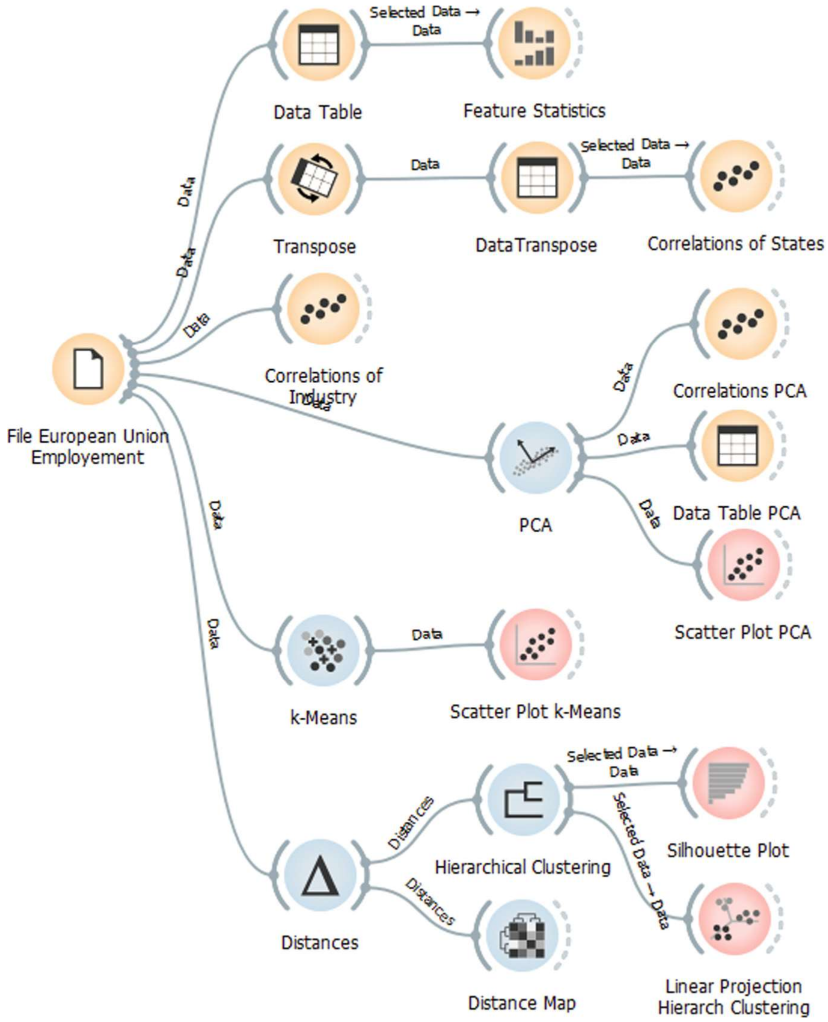
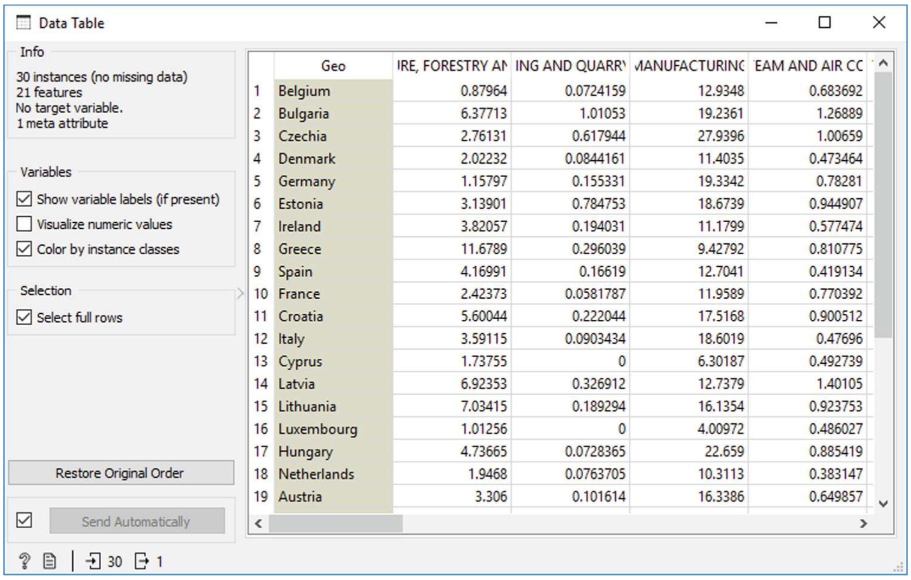


**Fig. 1.** The workflow in Orange software for processing data by several methods. Methods are widgets represented by colour circular nodes.

The first widget, *File European Union Employment* at the left, connects the source XLSX data in the workflow. The first top branch from the *File* widget connects the *Data Table* widget. This widget shows the input table content about employments in sectors. Each row represents one country as an instance. Columns contain the type of economic activity in percentage (Fig. 2).



Info

30 instances (no missing data)
21 features
No target variable.
1 meta attribute

Variables

☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

☑ Send Automatically

| | Geo | IRE, FORESTRY AN | ING AND QUARRY | MANUFACTURING | EAM AND AIR CC |
|---|---|---|---|---|---|
| 1 | Belgium | 0.87964 | 0.0724159 | 12.9348 | 0.683692 |
| 2 | Bulgaria | 6.37713 | 1.01053 | 19.2361 | 1.26889 |
| 3 | Czechia | 2.76131 | 0.617944 | 27.9396 | 1.00659 |
| 4 | Denmark | 2.02232 | 0.0844161 | 11.4035 | 0.473464 |
| 5 | Germany | 1.15797 | 0.155331 | 19.3342 | 0.78281 |
| 6 | Estonia | 3.13901 | 0.784753 | 18.6739 | 0.944907 |
| 7 | Ireland | 3.82057 | 0.194031 | 11.1799 | 0.577474 |
| 8 | Greece | 11.6789 | 0.296039 | 9.42792 | 0.810775 |
| 9 | Spain | 4.16991 | 0.16619 | 12.7041 | 0.419134 |
| 10 | France | 2.42373 | 0.0581787 | 11.9589 | 0.770392 |
| 11 | Croatia | 5.60044 | 0.222044 | 17.5168 | 0.900512 |
| 12 | Italy | 3.59115 | 0.0903434 | 18.6019 | 0.47696 |
| 13 | Cyprus | 1.73755 | 0 | 6.30187 | 0.492739 |
| 14 | Latvia | 6.92353 | 0.326912 | 12.7379 | 1.40105 |
| 15 | Lithuania | 7.03415 | 0.189294 | 16.1354 | 0.923753 |
| 16 | Luxembourg | 1.01256 | 0 | 4.00972 | 0.486027 |
| 17 | Hungary | 4.73665 | 0.0728365 | 22.659 | 0.885419 |
| 18 | Netherlands | 1.9468 | 0.0763705 | 10.3113 | 0.383147 |
| 19 | Austria | 3.306 | 0.101614 | 16.3386 | 0.649857 |

**Fig. 2.** Source data with EU states as records with the percentage of employment in specific sectors of economic activities NACE. Widget Data Table depicts it in Orange software.

### Feature Statistics

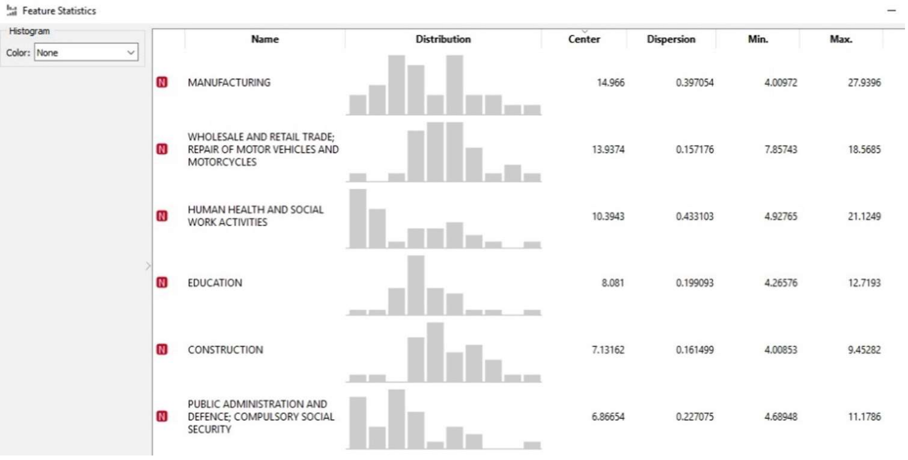
Students are taught that exploratory data analysis (EDA) should always be the first step in the data processing process. EDA includes, for example, basic descriptive statistics of centre and variability, which provide an elementary overview about the nature of data. Basic feature statistics (minimum, maximum and mean values) has been calculated in the Orange software by widget *Feature Statistics* (Fig. 3). This widget follows the widget *Data Table* in the workflow (Fig. 1). It is possible to order rows interactively by maximum or minimum values in columns. Data is ordered by the maximum value in Fig. 3. The maximum value in Manufacturing (near 28% employee) was observed in Czechia, followed by Slovenia and Slovakia. The simple graph with the distribution of values (histogram) is also depicted in the *Feature Statistics* window. The distribution is very variable in different economic activities in Europe. The widget brings a quick and complete first overview of data.

### Correlation Between Sector and Correlation Between Countries

The next basic tasks is the examination of correlation using the Pearson correlation coefficient. The widget node *Correlation* shows the list of Pearson correlation coefficients between pairs sorted form the highest absolute value to zero (Fig. 4a). The values are accompanied by colour lines (green positive, blue negative correlation) where length expresses the value. The interesting thing is the investigation of the correlation between partial economic activities. Very high correlation (above + 0.8) are between *Activities of extraterritorial organisations* and *Financial and insurance activities; Electricity, gas and steam supply* correlates with *transportation and storage.* Surprisingly, negative correlation (-0.7) of *Manufacturing* with *Professional, scientific and technical activities.* The second-highest negative correlation is between *Professional, scientific and technical activities* and *Water supply and sewerage* (0.652).

After transposing the data table (widget *Transpose* in Fig. 1), it is possible to calculate the correlation coefficient for countries (Fig. 4b). The strongest correlation is between *Hungary* and *Slovakia* (+0.989*), Czechia* and *Slovenia* (+0.988) and *Belgium* and *France* (0.982). Significant correlations are evident among the east (former socialist) countries, and the western countries also correlate.

No negative correlation was observed, but several combinations report low, insignificant correlations: *Luxembourg – Romania* (0.07), *Czechia* (0.252), *Bulgaria* (0.277). Especially *Luxembourg* can be marked as a very specific country form the point of employment by economic activities.



Feature Statistics

Histogram

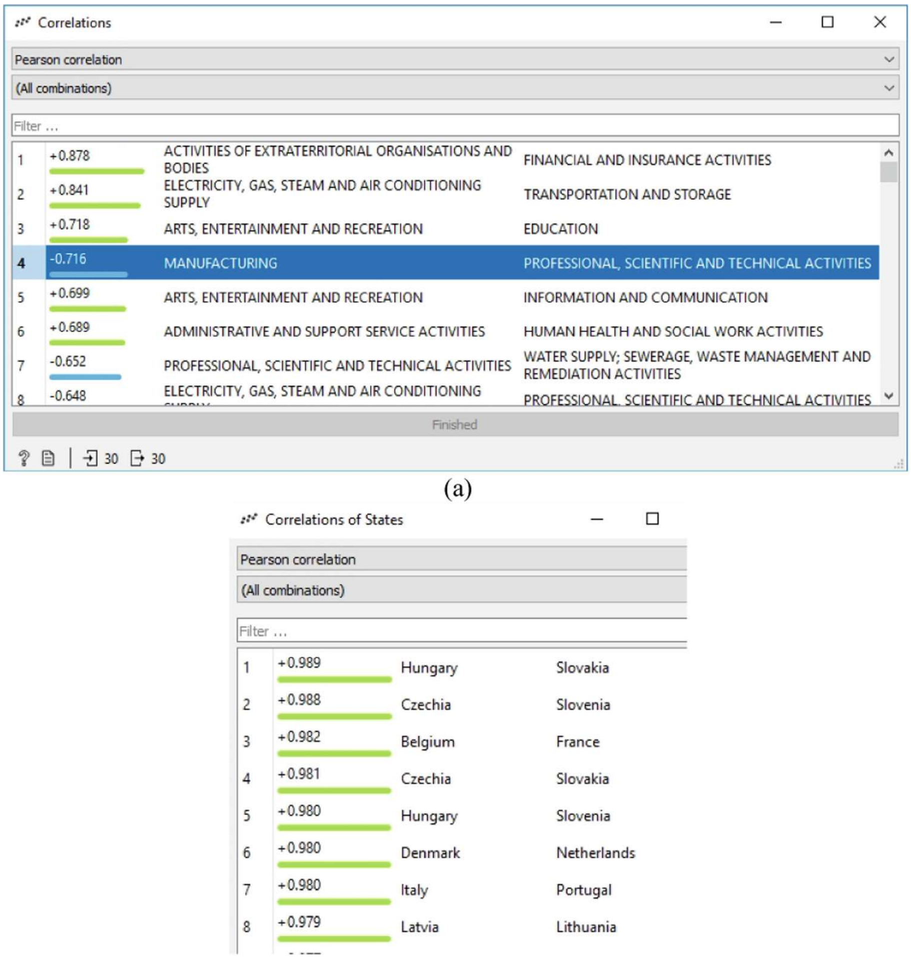
Color: None

| | Name | Distribution | Center | Dispersion | Min. | Max. |
|---|---|---|---|---|---|---|
| N | MANUFACTURING | | 14.966 | 0.397054 | 4.00972 | 27.9396 |
| N | WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | | 13.9374 | 0.157176 | 7.85743 | 18.5685 |
| N | HUMAN HEALTH AND SOCIAL WORK ACTIVITIES | | 10.3943 | 0.433103 | 4.92765 | 21.1249 |
| N | EDUCATION | | 8.081 | 0.199093 | 4.26576 | 12.7193 |
| N | CONSTRUCTION | | 7.13162 | 0.161499 | 4.00853 | 9.45282 |
| N | PUBLIC ADMINISTRATION AND DEFENCE; COMPULSORY SOCIAL SECURITY | | 6.86654 | 0.227075 | 4.68948 | 11.1786 |

**Fig. 3.** Basic description by Feature Statistics widget.

## Principal Component Analysis and Scatter Plot in PC1 a PC2

A principal component analysis (PCA) is a useful technique for revealing patterns when dealing with multi-dimensional correlated attributes. PCA can reduce the processed file's dimension and reveal new, hidden properties - components [12]. The main components are linear combinations of the original attributes, which in descending order capture as much of the input data variance as possible. PCA can be used widely - not only for simple data size reduction but also for checking the data file's consistency or calculating indexes (see example in [13]).



| | | | |
|---|---|---|---|
| 1 | +0.878 | ACTIVITIES OF EXTRATERRITORIAL ORGANISATIONS AND BODIES | FINANCIAL AND INSURANCE ACTIVITIES |
| 2 | +0.841 | ELECTRICITY, GAS, STEAM AND AIR CONDITIONING SUPPLY | TRANSPORTATION AND STORAGE |
| 3 | +0.718 | ARTS, ENTERTAINMENT AND RECREATION | EDUCATION |
| 4 | -0.716 | MANUFACTURING | PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES |
| 5 | +0.699 | ARTS, ENTERTAINMENT AND RECREATION | INFORMATION AND COMMUNICATION |
| 6 | +0.689 | ADMINISTRATIVE AND SUPPORT SERVICE ACTIVITIES | HUMAN HEALTH AND SOCIAL WORK ACTIVITIES |
| 7 | -0.652 | PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES | WATER SUPPLY; SEWERAGE, WASTE MANAGEMENT AND REMEDIATION ACTIVITIES |
| 8 | -0.648 | ELECTRICITY, GAS, STEAM AND AIR CONDITIONING | PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES |

(a)

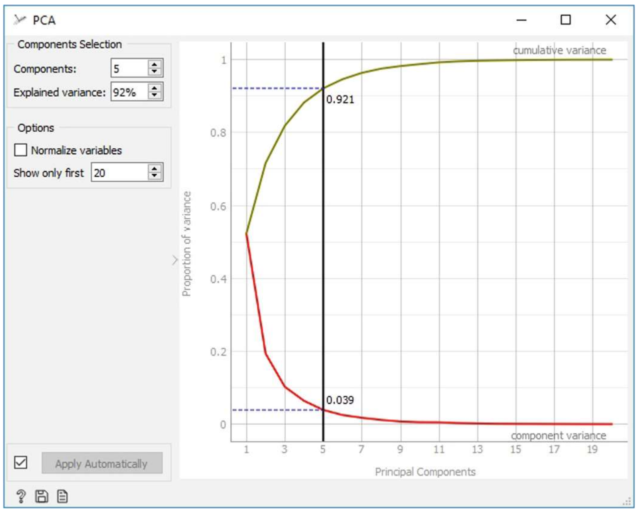
| | | | |
|---|---|---|---|
| 1 | +0.989 | Hungary | Slovakia |
| 2 | +0.988 | Czechia | Slovenia |
| 3 | +0.982 | Belgium | France |
| 4 | +0.981 | Czechia | Slovakia |
| 5 | +0.980 | Hungary | Slovenia |
| 6 | +0.980 | Denmark | Netherlands |
| 7 | +0.980 | Italy | Portugal |
| 8 | +0.979 | Latvia | Lithuania |

(b)

**Fig. 4.** (a) Correlation of economic sectors, (b) Correlation of countries.

First, the number of new significant principal components is determined using the Kaiser criterion. If the component's eigenvalue is greater than one, the component brings new information, and it is advisable to include it for analysis. In our case, five components were labelled as significant, and they together describe 92% of the variance in original input data (Fig. 5). The widget PCA shows the red scree graph where the user can interactively move the black vertical line to select the final number of components. The graph also shows a green curve of cumulative variance labelled by explained variance [14]. The interactive control is very user–friendly and helps student in the selection of the final count of a new component.

Several graphical outputs can analyse the results of PCA. The graph of instances has the form of the two-dimensional scatterplot, showing the position of individual inputs projected into two-dimensional space defined by the scores of two selected components (one graph for each pairwise combination of components is possible). In Fig. 6, the input instances (countries) are displayed in the context of the first and second component, PC1 and PC2. The similarities among countries can be then observed concerning the newly created components. The mutual position of individual states describes their similarity in the context of the components of the score. For example, *Hungary, Slovakia, Slovenia* and *Czechia* are close to each other (yellow group), which means they have quite similar components scores. The green group in the middle consists of Lithuania, Croatia, Latvia, Austria, Portugal, and Italy. There is another significant group of similar countries consisting of the *United Kingdom, Netherlands, Sweden, Denmark, France, Switzerland* and *Belgium* (blue group).



PCA

Components Selection

Components: 5

Explained variance: 92%

Options

☐ Normalize variables

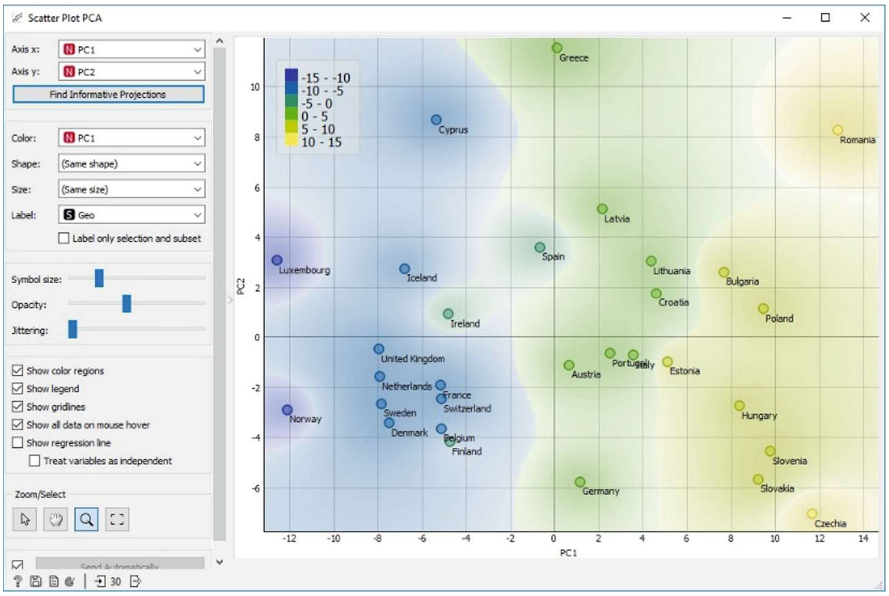
Show only first 20

☑ Apply Automatically

**Fig. 5.** The widget PCA with interactive determining of number of components in the scree graph.

Similarities within only one component can be observed as well (Fig. 6).: *Luxembourg and Norway* have an almost identical score on the first component, on the second component *Greece* has the highest component score, while the first component has a score of around zero, which points to the typical profile of *Greece* on the second component.

Depending on the PCA's efficiency to the data size reduction, all of the components might be deeper interpreted and named. This task is often extremely challenging when there are many correlated indicators, which form a particular component. The similarity of countries can be explained by detecting the correlation between the original attributes and new components PC1 and PC2. The following node *Correlation PCA* represents this step after node *PCA* (Fig. 1). In our example, attributes *Manufacturing* (0.91), *Water supply* (0.77), *Electricity supply* (0.62), and *Agriculture* (0.59) have the highest contribution for PC1. This contribution value of original attributes helps with the labelling and interpretation of components. With this finding, we can label the PC1 as *Industry and production. Agriculture* (correlation 0.62), *Whole retail trade* (0.57) and *Accommodation and food service* (0.48) have the highest contribution for PC2. We can assign the



Scatter Plot PCA

Axis x: PC1

Axis y: PC2

Color: PC1

Shape: (Same shape)

Size: (Same size)

Label: Geo

☐ Label only selection and subset

☑ Show color regions

☑ Show legend

☑ Show gridlines

☑ Show all data on mouse hover

☐ Show regression line

☐ Treat variables as independent

**Fig. 6.** The result PCA scatter plot of instances in PC1 and PC2 coordinate system.

label PC2 as *Trading and services*. *Financial and insurance activities* have the highest contribution to PC3.

There are several graphical tools, which can provide deeper insight into PCA results: namely *correlation circle*, which visualise the relation between input indicators in the coordinates of PCA; *scree plot* showing the importance of calculated components; *bar plots* presenting the contribution of particular attributes on the component (or the quality of their representation on the component) or *graph of individuals* showing the component score of every record included in the analysis. All of these outputs are the subject of detective work, aiming at a better interpretation of the PCA and understanding the revealed links.

One of the benefits of a *Scatter plot* implemented for PCA is the possibility of changing pairs of the axis (PCs), forming the plot's coordinates system and distinguishing the potential clusters (based on PC score) by colour. The colouring depends on the selected component (PC1 in case Fig. 6), so the different results and cluster arrangement could be deduced. E.g., Greece creates a single element cluster due to many people working at accommodation and food services. Luxembourg and Cyprus have many employees in financial and insurance activities.

### Clustering Methods

Another group of typical data mining method is cluster analysis. In general, it is an objective numerical approach aiming at searching for the similarities in n-dimensional space. Meloun & Militký [15] describe cluster analysis as a method to investigate the similarity of multi-dimensional objects and their classification into classes. The essence of cluster analysis is calculating distances between individual objects in n-dimensional space, which represents the input data. Subsequently, clusters (groups of objects similar to each other and simultaneously differ as much as possible from the objects of other clusters) are defined using various algorithms.

In the literature, many authors point out the importance of selecting a suitable clustering algorithm, but not to the selection of a suitable degree of similarity [16]. According to Hastie, Tibshirani, & Friedman [17], the specification of an appropriate degree of similarity (distance) is often more important than a particular algorithm's choice. At least the elementary examples of the distances, such as Euclidean, Mahalanobis, Manhattan or Cosine measures, are introduced to students in the Data Mining course. It is always emphasised that it is necessary to play and experiment with the metrics and choose how they make sense for the given data (e.g. use the Mahalanobis distance for strongly correlated data). The Orange offer several metrics in the widget *Distances*. The visualisation of the used metrics in the form of a heatmap helps to understand the mutual relations between the objects analysed. The widget *Distance Map* shows a matrix of cosine distances with a dendrogram (Fig. 7). In Fig. 7, two significant blue rectangles (couples close to each other, i.e. potential clusters) can be distinguished. Luxembourg and Romania are both at the edges of the matrix, which determines they are dissimilar to other countries. The finding about Luxembourg correspond to the previous findings outputting from correlation analysis; the separation of Romania was also visible in the

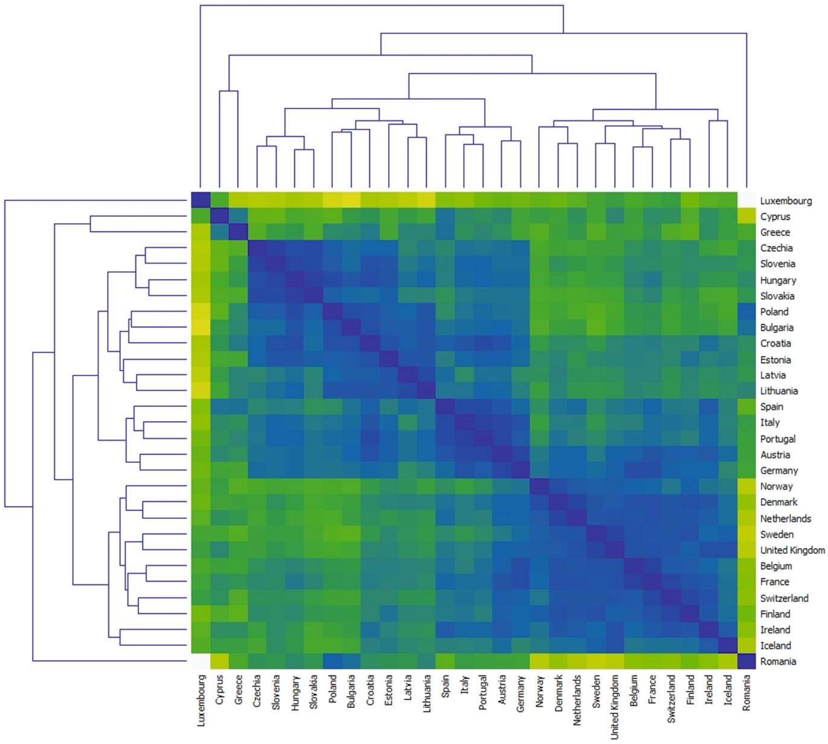


**Fig. 7.** Visualisation of distance matrix based on cosine distance for economic activities

result of PCA. The explanation lies in the great portion of employees in *agriculture, forestry and fishing* (19.8%). That belongs to the primary sector. The second country with a high amount of employees in agriculture is Greece (11.7%).

### Hierarchical Clustering with Ward's Method and Silhouette Plot

There are many possibilities of clustering approaches and algorithms – see, e.g. [18–20]. In the course, the main difference between the hierarchical and non-hierarchical are introduced and demonstrated in the examples of Ward's method (hierarchical group) and K-means (non-hierarchical). These methods were also applied to find similar enterprises from the C sector – Manufacturing in previous research. The source data was about various types of innovations [21]. Ward's method was selected due to the best results in the presented example of economic activities.

The widget *Hierarchical Clustering* computes dendrogram from a matrix of distances. The level of selection could be stated manually, by height ratio or by top N nodes. The big advantage is the interactive manual placing of the vertical dot line of selection and automatic colouring clusters (Fig. 8a). Clustering was processed on a cosine distance matrix, using Ward's method in the presented dendrogram in Fig. 8a.

The widget *Silhouette Plot* offers a graphical representation of consistency within clusters. User can visually assess cluster quality. The silhouette score close to 1 indicates that data instances are close to the centre of the cluster. Instances possessing a score close to 0 are on the border between two clusters. Figure 8b verifies the consistency of the selected level of clustering. Students can experiment with selecting the number of clusters and assessing the consistency of clusters via the widget *Silhouette Plot*. The colours are assigned automatically, and they correspond between the dendrogram and silhouette plot.

They are seven result clusters in the dendrogram. The arrangement of clusters corresponds to the previous results. Luxembourg creates a single-element entity - cluster C1, since it is very dissimilar to other European countries. The next widget, *Linear Projection,* helps with an examination of other clusters and similarities. Clusters of countries are placed in multiple axes of source attributes. Cluster C2 consists of the western countries of Europe. These have more employees in *Human health and social activities, Educations, Professional, scientific and technical activities, Arts, entertainment and recreation.* Four countries from cluster C3: Czechia, Slovakia, Slovenia and Hungary, are similar due high portion of employees in manufacturing and water supply.

Moreover, they are neighbouring countries (except Slovenia). Cluster C4 is a single-element entity of Romania. Cluster C5 is consists of Lithuania, Latvia, Bulgaria, Croatia, Estonia and Poland. They have a similar portion of employees in electricity, gas, steam supply and transportation and agriculture. An interesting cluster is C6 which consists of Austria, Portugal, Italy, Spain and Germany. They are similar in *Accommodation and food services, trading* and *transportation.* Finally, cluster C7 consists of Cyprus and Greece that were also found similar by PCA (sectors like accommodation and trading).

### K-Means Clustering

The last presented method is k-Means clustering. It belongs to unsupervised machine learning as well as PCA and hierarchical clustering [8]. The disadvantage of k-Means is the need of choosing the number of clusters at the start. This widget helps to solve this

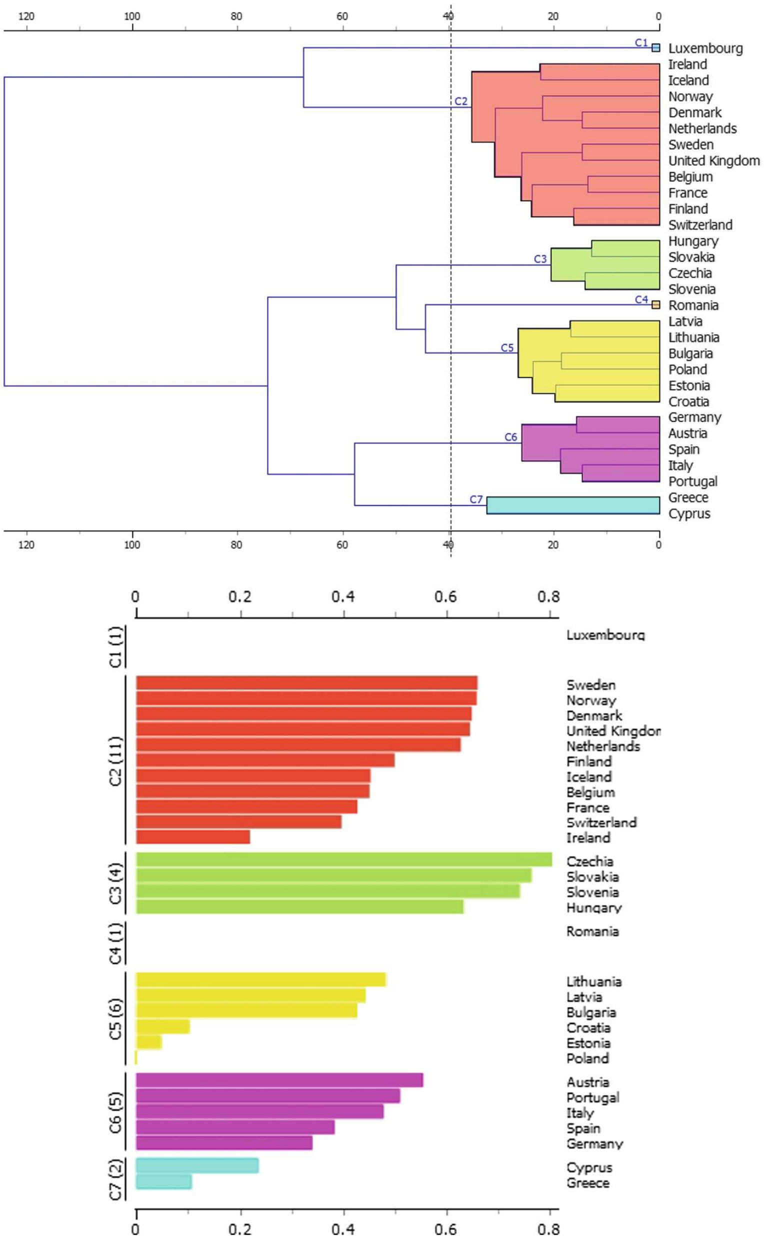


**Fig. 8.** Hierarchical clustering (a) dendrogram, (b) corresponding silhouette plot.

problem by finding the number of clusters using silhouette scores. In the case of NACE economic activities, the suggested number of clusters is two.

Moreover, there are two options for determining the starting centroids of clusters: *Random initialisation* or *Initialisation with KMeans++* in Orange. The followed widget

*Scatter Plot* shows the result of K-Means in automatically selected the best projection (Fig. 9). For economic activities, the suggested projection is for the X-axis *Human health and social activities*, for the Y-axis *Manufacturing*. The result red and blue clusters confirm previous findings. Blue cluster with Norway, Denmark, Netherlands and others have a higher percentage of employees in *Human health and social activities*. Red clusters are countries with higher manufacturing numbers (Czechia, Slovenia, Slovakia, Hungary, Poland, Romania, Italy, Bulgaria, Germany). Germany and Austria belong to the red cluster, but they are near the clusters' border because they also have many employees in *Human health and social activities*.

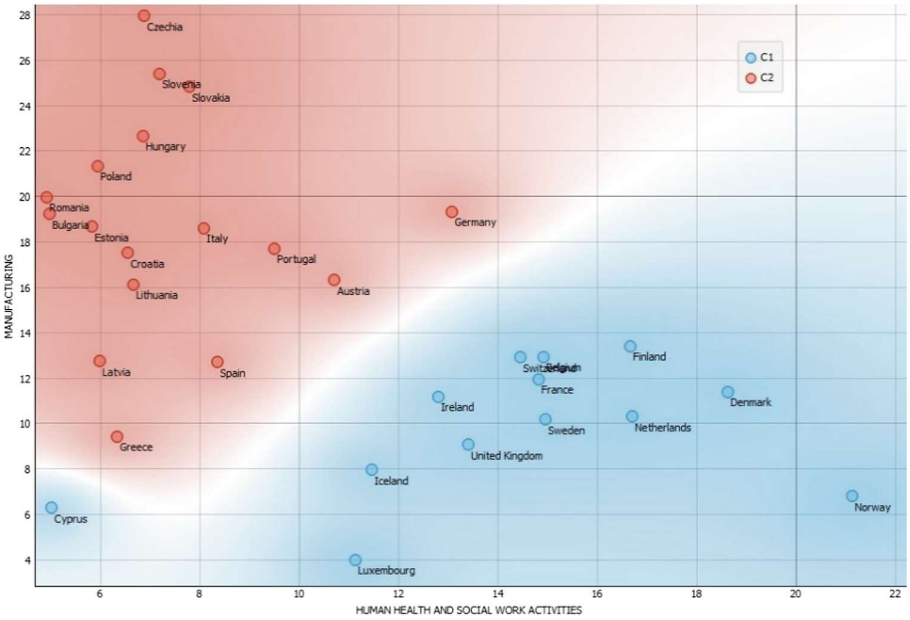
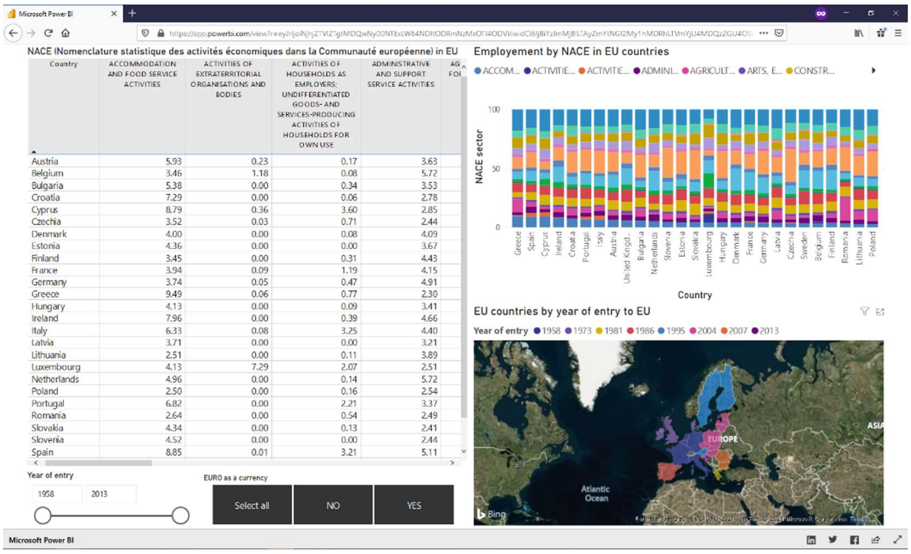


**Fig. 9.** Clusters determined by the K-Means method in two coordinates (economic sectors).

### PowerBI Dashboard

After statistical analysing of the dataset in the Orange software, we had visualised data using online interactive business intelligence (BI) dashboards (Fig. 10). A dashboard is a basic tool of the visual analytics technique. They are one of the most common use cases for data visualisation, and their design and contexts of use are considerably different from exploratory visualisation tools [22]. We had chosen the Microsoft Power BI Desktop. It is free for use. Students had uploaded data into the application. Then they used some visuals suitable for the dataset. We used the matrix, bar chart, choropleth map, filters, or slicers in this tutorial. The dashboard allows interactive selecting and filtering of data by clicking into any visual. Students can click on the YES or NO button in the EURO section to filter out only countries within the Eurozone. They can click anywhere in the table, map, or chart also. The given country will be highlighted in every visual. The final dashboard from our course is available online on https://bit.ly/3dpSV9w.


Microsoft Power BI
https://app.powerbi.com/view?r=eyJrIjoiNjhjZTVlZTgtMDQwNy00NTExLWE4NDItODRmNzMxOTI4ODViIiwidCI6IjBiYzJmMjBiLTAyZmYtNGI2My1hMDRhLTVmYjU4MDQzZGU4OS
Employement by NACE in EU countries
● ACCOM… ● ACTIVITIE… ● ACTIVITIE… ● ADMINI… ● AGRICULT… ● ARTS, E… ● CONSTR…
100
50
0
NACE sector
Greece
Spain
Cyprus
Ireland
Croatia
Portugal
Italy
Austria
United Kingd…
Bulgaria
Netherlands
Slovenia
Estonia
Slovakia
Luxembourg
Hungary
Denmark
France
Germany
Latvia
Czechia
Sweden
Belgium
Finland
Romania
Lithuania
Poland
Country
EU countries by year of entry to EU
Year of entry ● 1958 ● 1973 ● 1981 ● 1986 ● 1995 ● 2004 ● 2007 ● 2013
EUROPE
ASIA
Atlantic Ocean
Bing
Year of entry
1958
2013
EURO as a currency
Select all
NO
YES
Microsoft Power BI

NACE (Nomenclature statistique des activités économiques dans la Communauté européenne) in EU

| Country | ACCOMMODATION AND FOOD SERVICE ACTIVITIES | ACTIVITIES OF EXTRATERRITORIAL ORGANISATIONS AND BODIES | ACTIVITIES OF HOUSEHOLDS AS EMPLOYERS; UNDIFFERENTIATED GOODS- AND SERVICES-PRODUCING ACTIVITIES OF HOUSEHOLDS FOR OWN USE | ADMINISTRATIVE AND SUPPORT SERVICE ACTIVITIES |
|---|---|---|---|---|
| Austria | 5.93 | 0.23 | 0.17 | 3.63 |
| Belgium | 3.46 | 1.18 | 0.08 | 5.72 |
| Bulgaria | 5.38 | 0.00 | 0.34 | 3.53 |
| Croatia | 7.29 | 0.00 | 0.06 | 2.78 |
| Cyprus | 8.79 | 0.36 | 3.60 | 2.85 |
| Czechia | 3.52 | 0.03 | 0.71 | 2.44 |
| Denmark | 4.00 | 0.00 | 0.08 | 4.09 |
| Estonia | 4.36 | 0.00 | 0.00 | 3.67 |
| Finland | 3.45 | 0.00 | 0.31 | 4.43 |
| France | 3.94 | 0.09 | 1.19 | 4.15 |
| Germany | 3.74 | 0.05 | 0.47 | 4.91 |
| Greece | 9.49 | 0.06 | 0.77 | 2.30 |
| Hungary | 4.13 | 0.00 | 0.09 | 3.41 |
| Ireland | 7.96 | 0.00 | 0.39 | 4.66 |
| Italy | 6.33 | 0.08 | 3.25 | 4.40 |
| Latvia | 3.71 | 0.00 | 0.00 | 3.21 |
| Lithuania | 2.51 | 0.00 | 0.11 | 3.89 |
| Luxembourg | 4.13 | 7.29 | 2.07 | 2.51 |
| Netherlands | 4.96 | 0.00 | 0.14 | 5.72 |
| Poland | 2.50 | 0.00 | 0.16 | 2.54 |
| Portugal | 6.82 | 0.00 | 2.21 | 3.37 |
| Romania | 2.64 | 0.00 | 0.54 | 2.49 |
| Slovakia | 4.34 | 0.00 | 0.13 | 2.41 |
| Slovenia | 4.52 | 0.00 | 0.00 | 2.44 |
| Spain | 8.85 | 0.01 | 3.21 | 5.11 |

**Fig. 10.** MS PowerBI Dashboard built with EU NACE employment dataset.

## 4 Conclusion

The use of real Eurostat data is very valuable for students. The best way is a combination of several data mining methods. Some methods bring new findings, and also the combination of a method brings confirmation of partial finding. This article discussed the correlation, PCA, clustering, and interactive visualisation using Orange and MS PowerBI software.

The result is that some differences in the west and middle/east European countries exist. West countries are more oriented to a tertiary and quaternary economic sector (human health, science, services), and east countries are oriented to agriculture, manufacturing and industry (primary and secondary sector). Some outliers are identified, like Romania, Luxembourg and Greece.

Moreover, there is a very good experience with Orange software for the Data Mining course. Orange software is very intuitive and allows the quick design of workflow, especially for novice students in analysing multi-dimensional data. The advantage is interactive exploration results in multiple windows, compare, combine and verify the results in the interpretation phase.

**Acknowledgement.** This article was created with the support of the Erasmus+ Programme of the European Union**,** Jean Monnet Module (Project No. 620791-EPP-1-2020-1-CZ-EPPJMO-MODULE, UrbanDM - Data mining and analysing of urban structures as a contribution to European Union studies). The second supporting project is the Application of geospatial technologies for spatial analysis, modelling, and visualisation of spatial phenomena (IGA_PrF_2021_020) of the Internal Grant Agency of Palacký University Olomouc.

## References

1. Eurostat. https://ec.europa.eu/eurostat/data/database. Accessed 31 Dec 2021
2. Pánek, J., Pászto, V., Perkins, C.: Flying a kite: playful mapping in a multidisciplinary field-course. J. Geogr. High. Educ. **42**, 317–336 (2018). https://doi.org/10.1080/03098265.2018.1463975
3. Egiebor, E.E., Foster, E.J.: Students' perceptions of their engagement using GIS-story maps. J. Geogr. **118**, 51–65 (2019). https://doi.org/10.1080/00221341.2018.1515975
4. Dobesova, Z.: Teaching database systems using a practical example. Earth Sci. Inf. **9**(2), 215–224 (2015). https://doi.org/10.1007/s12145-015-0241-3
5. Dobesova, Z.: Teaching decision tree using a practical example. In: Silhavy, R. (ed.) CSOC 2020. AISC, vol. 1226, pp. 247–256. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51974-2_23
6. NACE Rev. 2 - Statistical Classification of Economic Activities. https://ec.europa.eu/eurostat/web/nace-rev2/overview. Accessed 10 Feb 2021
7. National accounts employment data by industry. https://ec.europa.eu/eurostat/databrowser/view/NAMA_10_A64_E__custom_221911/default/table?lang=en. Accessed 12 Feb 2021
8. Šarmanová, J.: Metody analýzy dat. VŠB – Technická univerzita, Ostrava (2012)
9. Pavel, P.: Metody Data Miningu, část I. Univerzita Pardubice, Fakulta ekonomicko-správní, Pardubice (2014)
10. University of Ljubljana: Orange. https://orange.biolab.si/. Accessed 04 May 2019
11. Orange visual programming documentation, Release 3. https://buildmedia.readthedocs.org/media/pdf/orange-visual-programming/latest/orange-visual-programming.pdf. Accessed 24 Sep 2020
12. Jolliffe, I.: Principal Component Analysis. Springer, New York (2002)
13. Macků, K., Voženílek, V.: Statistická syntéza indikátorů kvality života – návrh tvorby indexu v evropských regionech. Geographia Cassoviensis 13 (2019). https://doi.org/10.33542/GC2019-2-06
14. Getting Started with Orange 09: Principal Component Analysis. https://www.youtube.com/watch?v=OmaAC8a52YI&t=2s. Accessed 10 Dec 2019
15. Meloun, M., Militký, J.: Statistical Data Analysis. Woodhead Publishing India (2011)
16. Mimmack, G.M., Mason, S.J., Galpin, J.S.: Choice of distance matrices in cluster analysis: defining regions. J. Clim. **14**, 2790 (2001). https://doi.org/10.1175/1520-0442(2001)014%3c2790:codmic%3e2.0.co;2
17. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer New York Inc., New York (2001)
18. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: a K-means clustering algorithm. J. Roy. Statist. Soc. Ser. C (Appl. Stat.) **28**, 100–108 (1979). https://doi.org/10.2307/2346830
19. Ward, J.H.: Hierarchical grouping to optimise an objective function. J. Am. Stat. Assoc. **58**, 236–244 (1963). https://doi.org/10.1080/01621459.1963.10500845
20. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley (1990)
21. Dobesova, Z., Paszto, V., Macku, K.: Analysis of similarities in context of enterprise innovations. In: Slavíčková, P. (ed.) Knowledge for Market Use, p. 8, Olomouc (2017)
22. Sarikaya, A., Correll, M., Bartram, L., Tory, M., Fisher, D.: What do we talk about when we talk about dashboards? IEEE Trans. Visual Comput. Graphics **25**, 682–692 (2019). https://doi.org/10.1109/TVCG.2018.2864903