Visual Analytics Project
Data quality on ETER database

Alessandro Migliore
Gavrila George Daniel

Abstract

Purpose: This paper is related to the project of the Visual Analytics exam of Masters in Engineering in computer science of Sapienza University.

The main purpose is to present a data visualization system to analyze the data quality of the *European Tertiary Education Register* **(ETER)[4]** database. We will illustrate below the main functionality of the system and its possible usage.

Design: the visualization system was designed primarily to make comparisons between institutions and countries, below we will illustrate the main design choices and the reasons for these choices.

Analysis: The analysis is based on three main data quality indicators: percentage of missing values, consistency of the data present evolution of the lack of data over the years.

1 - Introduction

The European Tertiary Education Register (ETER) is the first comprehensive database on European higher education institutions (HEIs). It is a publicly available data source constructed mostly from statistical data in the participating countries, which covers most of the Higher Education Institutions (HEIs) contributing at the bachelor, master and PhD level in the European Union member countries, EEA/EFTA countries and candidate countries.

The main goal of this paper is to describe a flexible approach developed to monitor the data quality of ETER. We are going to analyze the quality of the pure data of individual universities and countries showing a graphical analysis on them.

The paper is organized as follows :

The next section provides an outline of the ETER data collection.

Data quality may be easy to recognize but it is difficult to determine precisely. You can consider multiple attributes of data to get the correct context and measurement approach to data quality.

Through a study we obtained a percentage of **missing values** and **consistency** on the entire dataset. These values allow us to analyze data quality in an extremely efficient way.

Missing values - In most datasets the missing values are marked as Nan, while in our case their identifier is *m*. For this reason a very precise distinction is needed in the study of individual values, as we will see in the next section.

Consistency - This dimension represents if the same information stored and used at multiple instances matches (considering only values that do not contain missing values). It is expressed as the percent of matched values across various records. Data consistency ensures that analytics correctly capture and leverage the value of data.

Then a visualization of the entire dataset, through numerous graphs that interact with each other and a section regarding the use of analytics.

2 - Data and precomputation

First of all we went to get the data of interest on the ETER site, in fact their site provides a large amount of information of various kinds on tertiary education institutions. Of these we have selected the most important ones, divided by macro categories, they are:

- Basic institutional descriptors
- Financial data
- Students data
- Staff data
- Other data of interest

Basic institutional descriptors
Eter id
Institution name
English institution name
Reference Year
Country code
Latitude
Longitude

Financial
Personal Expenditure (PPP)[1]
Non-Personnel Expenditure (PPP)[1]
Expenditure Unclassified (PPP)[1]
Capital Expenditure (PPP)[1]
Total Current Expenditure (PPP)[1]
Total core budget (PPP)[1]
Total third party Funding (PPP)[1]
Student Fees Funding (PPP)[1]
Revenue Unclassified (PPP)[1]
Total Current Revenues (PPP)[1]

Students data
Total students enrolled for every ISCED[2]
Total graduates students for every ISCED[2]
Share of women students ISCED[2] 5-7
Share of foreign students ISCED[2] 5-7

Staff data
Total staff
Total academic staff
Share of women academic staff

Other data
PhD intensity
Lowest degree delivered
Highest degree delivered

[1] PPP - Purchasing power parity is a measurement of prices in different countries that uses the prices of specific goods to compare the absolute purchasing power of the countries' currencies.

[2] ISCED - The International Standard Classification of Education (ISCED) is a statistical framework for organizing information on education maintained by the United Nations Educational, Scientific and Cultural Organization (UNESCO)

Pre Computation

We initially spent some time analyzing the data at our disposal, making preliminary analyzes. As a result, the entire dataset had a missing value equal to 53%, and a consistency percentage of 60%. This meant that we were dealing with an extremely sparse database. In addition to the missing data, several acronyms of ETER indicated confidential data or data in the process of being added.

This made the job not easy, in fact the first thing we had to do was adapt the data to the visualization and to the precomputation, finding a solution to their heterogeneity. For this we have made a lot of use of python in particular of the powerful numpy and pandas libraries.

Moreover when data are not available for any variable, in order to avoid blank cells, a specific level of metadata should be inserted substituting the missing figure.

In our case we have selected the main ones for our study.

m : refers to the fact that the data in question is missing.

c : is used in the public database only f
or data with restricted access.

nc : should be used for data that have not been collected in the reference year

s : is used in the public database only for data below 3 to keep anonymity of individuals.

The essential premise is that data points that are identified as outliers are highly likely to be invalid. **Outliers are data points that are significantly different from most of the data**. We will be using statistical model based outlier detection techniques. This will require computation of mean and standard deviation. Based on mean and standard deviation, we will use statistical quantities called Z score to detect outliers.

To calculate the Z-score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation. Mathematically, the formula for that process is the following:
z-score equation :

$$Z = \frac{X - \mu}{\sigma}$$

Extract data by country

One of the main pre-calculations was to calculate the aggregate of the main pure data and quality metrics for countries and for the entire db, which we would later need for the visualizations.This was done by making an average. Calculating for the percentage of missing value the number of cells in which there was either "m" or a null value.

Instead for the consistency we have referred to ten parameters, illustrated below, calculated for each row.

Consistency indicators

1	Total Expenditure=SUM(personnel expenditure, non-personnel expenditure, capital expenditure, unclassified expenditures)
2	Total expenditure>0
3	Total Income=SUM(core budget, third party funding, tuition fees, revenues unclassified)
4	Total Income>0
5	Staff Total (HC and FTE)=SUM(academic staff, non-academic staff)
6	Staff Total>0
7	If highest degree delivered=ISCED

	5 then Enrolled Students, Graduates ISCED 6-8 =”a” If highest degree delivered=ISCED 6 then Enrolled Students, Graduates ISCED 7-8 =”a” If highest degree delivered=ISCED 7 then Enrolled Students, Graduates ISCED 8 =”a”
8	If lowest degree delivered=ISCED 8 then Enrolled Students, Graduates ISCED 5 - 7 =”a” If lowest degree delivered=ISCED 7 then Enrolled Students, Graduates ISCED 5 - 6 =”a” If lowest degree delivered=ISCED 6 then Enrolled Students, Graduates ISCED 5 =”a”
9	SUM(Total students enrolled ISCED 5 - 7, Total students ISCED 8)>0
10	SUM(Total graduates ISCED 5 - 7, Graduates ISCED 8)>0

3 - Visual Environment[3]

The entire system was designed primarily to allow a comparison of data quality between various universities and countries, in particular by analyzing the three main parameters: the lack of data, the consistency of the data present and the presence in the period 2011-2017 of these data and their evolution.

The system can also be used to analyze the entire db in order to understand how to improve the quality of the data present in it.

Map

The map was designed to be able to see in the first place how the universities are positioned, see which ones are in the database, and where they are geographically located with respect to their home countries. Secondly, it acts as a selection display, as it is much more natural for any user to go and look for the university by country. It was built to be interactive via zoom,selections and overview.

For the creation of the map we first used a database of the countries extracted in geojson to form the countries, and secondly we placed the universities on top of them by latitude and longitude.

The colors have been chosen to represent waters and lands in a natural way, and the universities in black so that there is a good contrast and is visible.

Precisely also during a selection of a university and a country, colors were chosen that contrast appropriately.

PCA

The initial idea was to cluster the universities using a TruncatedSVD and a MiniBatchKmeans given the very sparse nature of the data. This is to categorize the various universities based on their pure data in the database, having tried to implement this solution and having not had satisfactory results, as the database presented a very high number of missing data and other types of incorrect data, we opted for a simple reduction of dimensionality through PCA to see how institutions were positioned through this reduction. A solution could have been to carry out clustering only on institutions that presented correct data, but in our opinion it would have been a solution that is not very consistent and coherent with the system.

We then made a selection of the most interesting data for each institution between educational financial and other interesting ones and we reduced them to 2 dimensions.

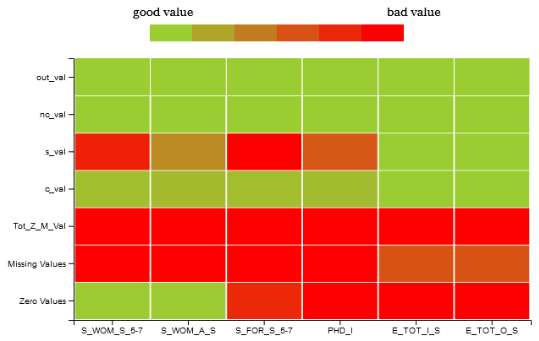
Table lens

The table to first display the data quality parameters for each institution and country. Each parameter is represented by a percentage and a colored rectangle based on the quality of the data. The color scale is a natural scale recognized at Western level, ranging from red for the worst values to green for the best values. The table can be ordered by each value so that it is more usable and secondly to have an account of the data trend.

The average of the selected values always appears as the last line.

This allows you to see how the selected values rank relative to the average.

Quality metrics matrix



Boxplot allow to quickly check distribution and eventual violation.Through a scale of colors, green for good value and red for bad value (shown in a legend), we can visualize our study on the dataset carried out previously, also through a filter it is possible to have an optic on the entire database ETER, in fact it is possible to filter the data for financial, educational and others. As you can immediately deduce the percentage of missing value is extremely high , while we have fewer outliers and null values.

Radar Chart

A radar web chart is a two-dimensional chart type designed to plot one or more series of values over multiple quantitative variables. Each variable has its own axis, all axes are joined in the center of the figure.

The user can carry out a study on the pure data of the dataset, divided by categories (financial , educational , others). We can decide to visualize the values of the single universities and nations compared with the total average of the dataset and the average of the selected elements, in such a way as to be able to analyse the evolution of the individual, compared to the total. Another option allows us to compare individual universities and nations to each other, this comparison is important to understand in which situations we have greater accuracy in entering data. The user can also filter data for consistency and missing value, allowing him to understand in which range he will have greater availability of data.

Linear time chart

To see the trend in data quality for each institution or country, a linear chart was designed with the years on the x axis and the percentages on the y axis.

It shows both the trend of these qualitative data over time for each institution and their presence or absence.

For each selected institution, colors are associated that dynamically appear in the legend so that the display is more usable and legible.

Furthermore there are the data quality parameters for the entire db to be able to compare the selected values with respect to the whole db, this if not of interest can be removed using a button and made to reappear at a later time.

Other aspects

The top bar allows you to filter the data in various ways, analyzing the data for the years in which the data is present, there is a filter by categories that allows you to select the pure data of interest.

Two buttons to clean the selections.

Two filters based on data quality parameters to be able to filter the data and the various views.

4 - Conclusions

Finally the development of the system and using it to have a general overview of the database, we can see that the entire database for the 2011-2017 time horizon settles on a percentage of approximately 50% of missing values and 60% of the consistency of the data we can note, however, that in 2017 there is a collapse with a percentage of missing value that rises to 60% and an important drop in the consistency of the data present to 20%, this could be caused by an abandonment of the project or of the collaborations between ETER and institutions, especially in 2017 there are only the basic description data of France.

More specifically we can see how the countries with the best data quality are in general those of Western Europe and in particular the Netherlands, Finland and the Czech Republic. Among the worst we can find in general those of Eastern Europe in particular Hungary, Romania, Lithuania.

5 - Related works

In the literature there are many scientific articles related to data quality, with different types of analysis and methodology, in particular the closest to our work is "A tailor-made Data Quality Approach for Higher Educational Data"[1], of which however there is no visual analysis system. Many similarities with our graphic environment can be found instead in different papers[2] that deal with different analyzes on European universities.

Below we will illustrate different design choices shared with these projects and other ideas of difference with them.

Similarity in data and analysis

Data quality is a relevant interdisciplinary issue, studied in statistics, management and computer science. Poor data quality greatly reduces data value: inaccuracy, incompleteness, out-of-dateness may cause data to become useless.

For a deep analysis on the data quality of the ETER database we refer to a study carried out.

[1]In the first section, they describe the data quality checks developed to identify outliers, extreme observations and to detect ontological inconsistencies not described in the available meta-data.

the methodology used is developed in two main phases:

1. Multiannual checks,
2. Cross-sectional checks

The multiannual checks beside the identification of individual outlier cases and mistakes in the reporting that were revised with data providers and corrected, allowed highlighting problems of comparability across waves of data collection.

For the Cross-Sectional checks a set of eight ratios was defined mainly considering financial and staff data, thresholds have been defined through an expert based approach in order to spot data inconsistencies both within each country and between different countries.

In our study we chose a consistency and missing value analysis. For consistency we used 10 indicators taken from the previous study[1]. Through them we quantify the value of the institution.We referred to the data flags , which allow us to understand why a data has not been entered. It was useful to note that some values are not present for restrictions on access to the data.

In conclusion we can see a result that is close to the previous one, the change in the result was also due to the fact that their study is carried out in a shorter time range (2012-2016), with a higher data flow, while in our case the percentage of missing values especially in 2017 is high.

Similarity in visualization[2]

In all the literature papers concerning studies conducted on universities and on various aspects of it, there is often always a geographical map with the positioning of the institutions on it, this in fact very often allows you to select the institutions of interest or the scale of the study , for example in our project there are two scales of study, at the institutional level or at the national level.

In relation to Scientometric work, the geographic map is shared but treated differently, specifically in Scientometric there are several selectable analysis scales (NUTS level from 0 Nations to 3 Provinces).

In particular, always with regard to the data exploration environment, we share a use of the same types of graphs.

In particular, the radar chart is used for both visual systems to examine the dimensionality of the tuples, drawing on it the data of interest contained in the tuples, these allow the user to make a comparison at different levels for the data present.

We also share the use of a line chart for a trend analysis of different data but which share a comparison to see their evolution over time.

There is also a sharing of some types of filters such as the choice of the time horizon, or filters used as thresholds.

Conclusion

In conclusion, this data quality project can be both a starting point for its further development and improvement and a useful system that can be adapted for other types of databases. In fact, the consistency indicators can be adapted to other types of data while remaining with the same visualization schemes that allow both a comparison between pure data and a comparison between data quality parameters, also analyzing the time trend.

Furthermore, the system in and of itself can prove useful to study, in this case, the ETER database to understand how to best integrate the database and improve its quality.

References

[1]Cinzia Daraio, Renato Bruni,Giuseppe Catalano, Alessandro Daraio, Giorgio Matteucc i, Monica Scannapieco , Daniel Wagner-Schuster, Benedetto Lepori, A tailor-made Data Quality Approach for Higher Educational Data https://www.researchgate.net/publication/342860207_A_Tailor-made_Data_Quality_Approach_for_Higher_Educational_Data

[2] Angelini, M., Daraio, C., Lenzerini, M. et al. Performance model's development: a novel approach encompassing ontology-based data access and visual analytics. Scientometrics 125, 865–892 (2020). https://doi.org/10.1007/s11192-020-03689-x

[3] Data-Driven Documents, https://d3js.org/

[4] European Tertiary Education Register , https://www.eter-project.com/