# Information Visualization

# CHECKPOINT II: Data cleaning and processing

G15 - A

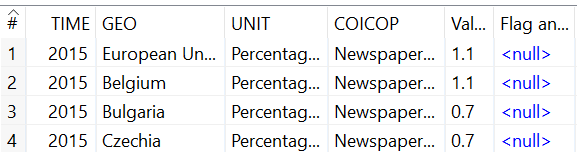
**1. Initial Dataset**

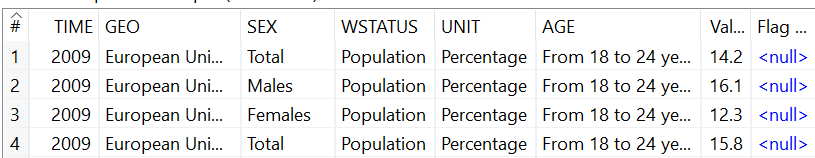
Falta pôr aqui o tamanho do dataset inicial e também o final

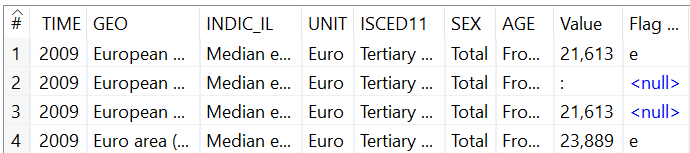
The aim of our information visualization is to correlate the reading habits of EU members and some demographics indicators such as average income per level of education, student performance in mathematics, science and reading.

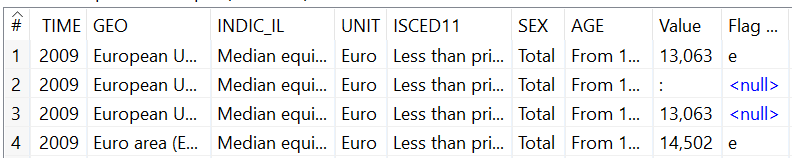
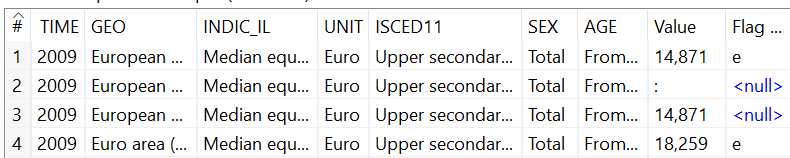
The datasets necessary to use in our visualization were obtained in the EUROSTAT database.

**Household Expenditure in Books**

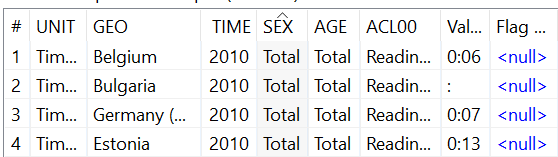


**Early leavers from education and training**

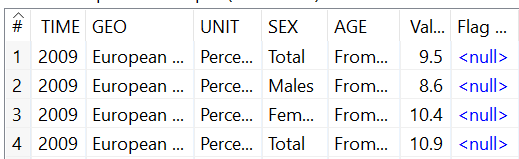
**Avg income by level of education**

****

**Time Spent Reading**



**Participation Rate in Education**

****

**2. Selected/Derived Data**

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Attributes | Selected Attributes | Derived Attributes |
| Time spent reading books | UNIT, GEO, TIME, SEX, AGE, ACL00, VALUE, Flag and Footnotes | Country, Value | by Country |
| Participation rate in education and training | TIME, GEO, UNIT, SEX, AGE, Value, Flag and Footnotes | Year, SEX, Value, Country | by Country and/or SEX |
| Early leavers from education | TIME, GEO, SEX, WSTATUS, UNIT, AGE, Value, Flag and Footnotes | Year, SEX, Value, Country | by Country and/or SEX |
| Household expenditure in books | TIME, GEO, UNIT, COICOP, Value, Flag and Footnotes | Year, Value, Country | by Country |

**3. Data abstraction**

|  |  |  |
| --- | --- | --- |
|  | **Attribute** | **Semantics** |
| “Value” – isto tem de ser discriminado, foi o que o professor avisou na aula | Quantitative | Percentage, time, Salary |
| “Year” | Quantitative | Stand for evaluate the evolution of data over time |
| “Country” | Nominal |  |
| “Sex” | Nominal | All, Males, Females |

* + - 1. **4. Dataset processing**

Devemos mudar aqui e pôr os passos mesmo? Como no power point mais ou menos?

To ensure the quality of the data needed to answer raised questions, it is necessary to attend a process of cleaning and structure them. Most of data acquired for it were in csv format, in order to clean, eliminate unnecessary attributes, and to keep the coherency of the data, we use Pentaho Data Integration as the main tool.

Using transformations and other resources available on Pentaho Data Integration, we perform some operations, such as fill the empty cells in the data sources and assign values. since we were working with historical data related with countries, that changes over years and for the case of countries that were in a conflict and got divided, we assign value “0” to cells that were empty or filled with “-”. However, some cases, we had to eliminate entire column or row.

The entire csv data processed, were transformed to .json files, for later use in D3.

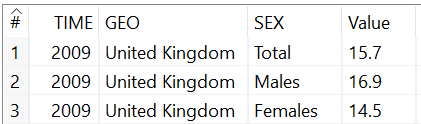
* + - 1. **5. Mapping (Data sample / Questions)**

1. **How many hours, in average, do the countries in EU spend reading?**

Time spent reading

1. **Given a country´s reading habits, what is the rate of dropout?**

Falta o time spent reading

****

1. **What is the average percentage of household expenditure in reading material by country?**

Book Expenditure

1. **Given the reading habits of each country, what is the average income for education level comparing to other EU countries?**

Falta o time spent reading

Education level 1

1. **What is the adult participation in learning after leaving the formal education?**

Participation Rate in Education and Training