# Information Visualization

# CHECKPOINT II: Data cleaning and processing

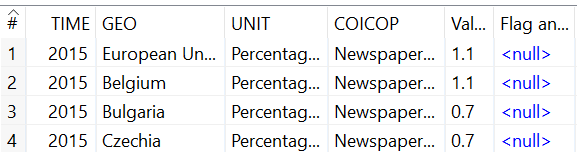
G15 - A

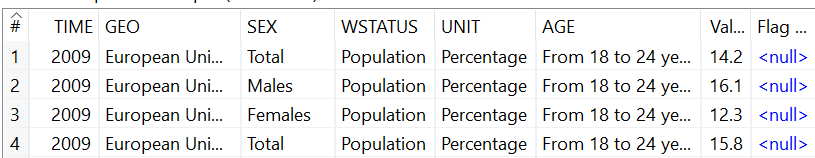
**1. Initial Dataset**

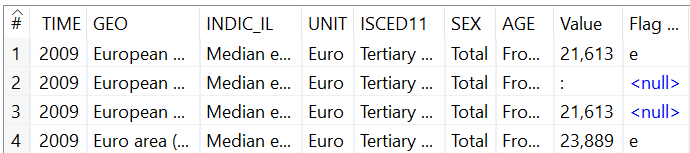
The aim of our information visualization is to correlate the reading habits of EU members and some demographics indicators such as average income per level of education, student performance in mathematics, science and reading.

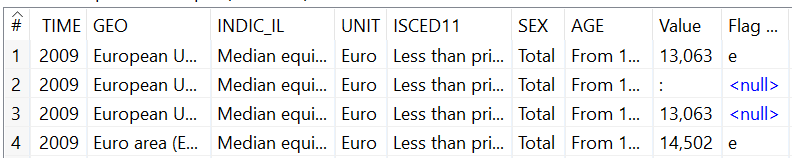
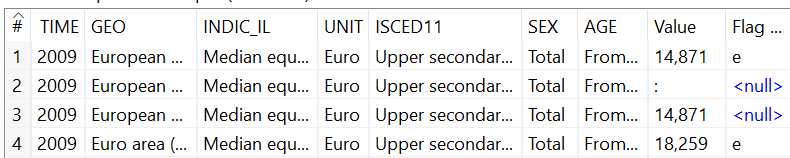
The datasets necessary to use in our visualization were obtained in the EUROSTAT database.

**Household Expenditure in Books**

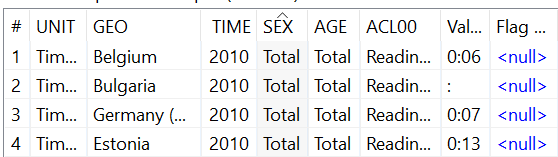


**Early leavers from education and training**

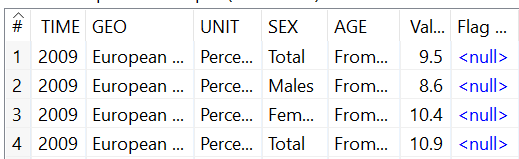
**Avg income by level of education**

****

**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(GEO), Value\* | \*avg\_reading\_minutes |
| Participation rate in education and training | TIME, GEO, UNIT, SEX, AGE, Value, Flag and Footnotes | Year(TIME),SEX,Value\*\*,Country  (GEO) | \*\*percentage\_pop |
| Early leavers from education | TIME,GEO,SEX, WSTATUS, UNIT, AGE, Value, Flag and Footnotes | Year(TIME),SEX, Value\*, Country  (GEO) | \*percentage\_pop |
| Household expenditure in books | TIME, GEO, UNIT, COICOP, Value, Flag and Footnotes | Year(TIME),Value\*,Country(GEO) | \*percentage\_total\_expend |
| Income level by education level | TIME,GEO,SEX,INDIC\_IL, UNIT, AGE, Value, Flag and Footnotes, ISCED11 | Year(Time), Country(Geo),Value\* | avg\_year\_income |

**3. Data abstraction**

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Attribute** | **Semantics** |
| “avg\_year\_income” | Quantitative | Average year income level, in euros |
| “percentage\_total\_expend” | Quantitative | Percentage of total expenditure in books and newspapers by household income |
| “avg\_reading\_minutes” | Quantitative | Average reading time in minutes (\* dataset doesn’t specify if by day,hours,month or year) |
| “percentage\_pop” | Quantitative | Percentage of population |
| “Year” | Quantitative | Stand for evaluate the evolution of data over time |
| “Country” | Nominal |  |
| “Sex” | Nominal | All, Males, Females |

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

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.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Initial number of instances** | **Final number of instances** |
| Time spent reading books | 66 | 22 |
| Participation rate in education and training | 1140 | 1080 |
| Early leavers from education | 1140 | 1080 |
| Household expenditure in books | 149 | 149 |
| Income level by education level | 400 + 400 + 400 | 1050 |

* + - 1. **5. Mapping (Data sample / Questions)**
      2. Here are some of the questions that we are going to base our visualization on.

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

{"Value":"0:06","Country":"Belgium"}

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

{"Value":"0:06","Country":"Belgium"}

{"Year":2009,"SEX":"Total","Value":"11.1","Country":"Belgium"},{"Year":2009,"SEX":"Males","Value":"12.8","Country":"Belgium"},{"Year":2009,"SEX":"Females","Value":"9.3","Country":"Belgium"}

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

{"Year":2015,"Value":"0.9","Country":"United Kingdom"}