# Information Visualization

# CHECKPOINT II: Data cleaning and processing

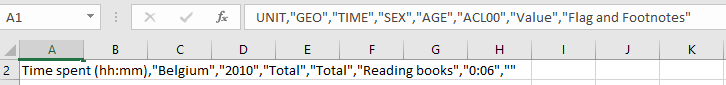
G15 - A

**1. Initial Dataset**

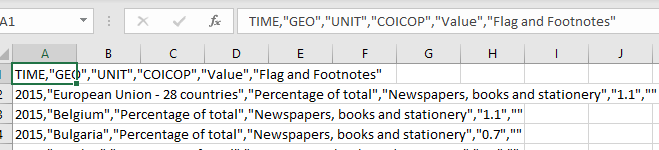
The aim of our information visualization is to correlate the reading habits of EU members and some demographics indicators such as average income per level of education, student performance in mathematics, science and reading.

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

**Time Spent Reading Books (By countries members of EU) (in the year 2010)**



**Book spending: Consumption expenditure for household by consumption purpose (2015-2018)**



**2. Selected/Derived Data**

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Attributes | Selected Attributes | Measures |
| Time spent reading books | UNIT, GEO, TIME, SEX, AGE, ACL00, VALUE, Flag and Footnotes | Country, Value | (1) |
| Participation rate in education and training | TIME, GEO, UNIT, SEX, AGE, Value, Flag and Footnotes | Year, SEX, Value, Country | (2) |
| Early leavers from education | TIME, GEO, SEX, WSTATUS, UNIT, AGE, Value, Flag and Footnotes | Year, SEX, Value, Country | (3) |
| Household expenditure in books | TIME, GEO, UNIT, COICOP, Value, Flag and Footnotes | Year, Value, Country |  |

1. The only selected attributes were the country and value, since for our questions these are the only ones we need. Although we had interest in the attributes sex and age, these were not discriminated in the dataset, there was only a value which was “TOTAL”.
2. See the questions related to the dataset
3. This dataset is going to be related with the first one, and answer to our question of “Given a country’s reading habits, what is the rate of dropout?”, for this reason, the only selected values consist in the year (2009-2018), sex (totals, male and female), value (percentage) and country.
   * + 1. **3. Data abstraction**

|  |  |  |
| --- | --- | --- |
|  | **Attribute** | **Semantics** |
| “Value” | 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**
      2. 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.
      3. Through the use of transformations and other more 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 also 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.
      4. The entire csv data processed, were transformed to .json files, for later use in D3.
      5. **5. Mapping (Data sample / Questions)**
* **How many hours, in average, do the countries in EU spend reading?**

{"time\_spend\_reading": [{"Value":"0:06","Country":"Belgium"},..

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

{"early\_leaver": [{"Year":2009,"SEX":"Total","Value":"15.7","Country":"UnitedKingdom"},

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

{"book\_expenditure": [{"Year":2015,"Value":"0.9","Country":"United Kingdom"},…

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

{"edu\_level\_1":[{"Year":2009,"Value":"13,411","Country":"United Kingdom"}

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

{"part\_rate\_edu\_training": [{"Year":2009,"SEX":"Total","Value":"20.7","Country":"United Kingdom"}