# Information Visualization

# CHECKPOINT II: Data cleaning and processing

G15 - A

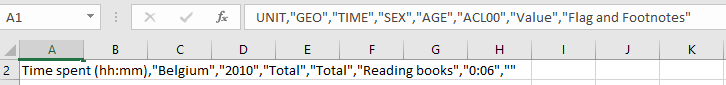
**1. Initial Dataset**

Initial dataset description: a high-level description of the initial dataset (with a data sample).

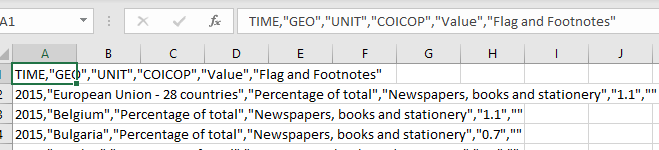
The aim of our information visualization is to correlate the reading habits of EU members and some demographics indicators such as average income per level of education, student performance in mathematics, science and reading.

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)

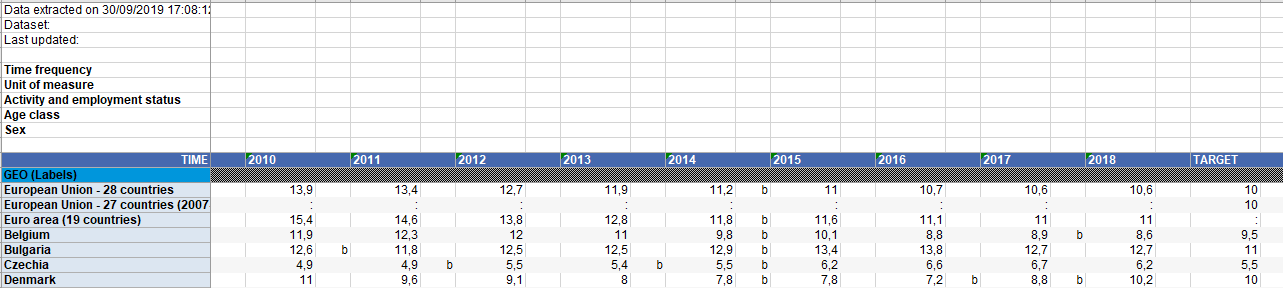


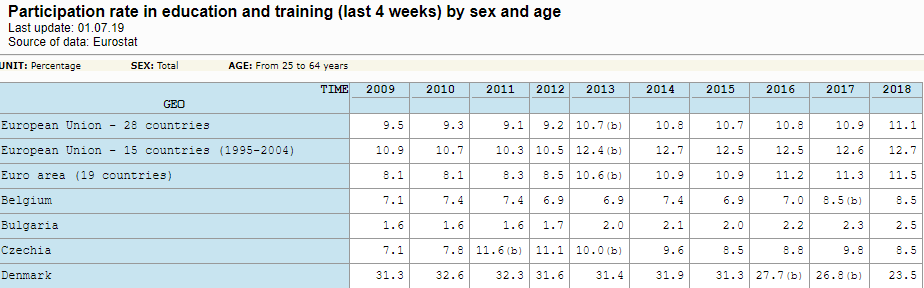
Book spending: Consumption expenditure for household by consumption purpose (2015-2018)



Early Leavers: Education and training (in percentage) from 1992-2018

File Structure: Summary, Sheet1(population), Sheet2(Males), Sheet3(Females)



Participation rate in education(non-traditional) and training (html file, group by all)

**2. Selected/Derived Data**

Description of:

* Which were selected from your dataset?
* Which derived measures did you consider and why (based on your tasks and questions)?

|  |  |  |  |
| --- | --- | --- | --- |
| 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 | (4) |
|  |  |  | (5) |

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**
4. Data abstraction description:

* Description of the dataset type (spatial, table, field, etc.);
* Description of each item and attribute (nominal/ordinal/etc., diverging/sequential scale, etc.);
* Semantics (what does each attribute and item stand for).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| “UNIT” |  |  |  |  |  |
| “Value” |  |  |  |  |  |
| “TIME” |  |  |  |  |  |
| “Flag and Footnotes” |  |  |  |  |  |
| “GEO” |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

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

Description of how the dataset was processed (cleaned, problems found and solutions, how did you fix missing values, cross-referenced different tables/datasets, etc.).

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

Some examples that show that with your data sample you will be able to provide the answers to the questions you formulated.