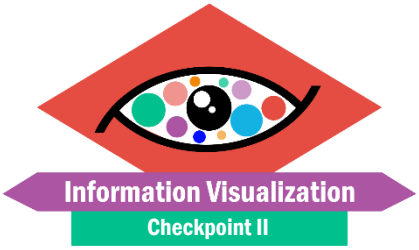

Checkpoint II: Data Cleaning & Processing

Group: <G29>
Date: <2020/10/16>

Initial Dataset

The initial dataset chosen for this project is a set of tables divided by Year, Country and a certain value (which, for each situation, contains information about: gross domestic product, the rate of population at risk for poverty, mean income by educational level, population by educational level, employment rate by sex and % of women in senior management positions). Some tables had other attributes, like Sex (M/F) or ISCED11. Some tables had columns for each year. Some tables had missing values and/or extra rows.

ISCED11	Sex	Country	Year	value
ED0-2	F	AT	2005	16254
ED0-2	F	AT	2006	16359

Datasets we'll be using: **(A)** Gross domestic product at market prices, 2008-2019; **(B)** At-risk-of-poverty rate by sex, 2005-2019; **(C)** Mean and median income by educational attainment level, 2003-2019; **(D)** Population by educational attainment level, sex, age and country of birth, 2005-2019; **(E)** Employment rate by sex, 2005-2019; **(F)** Positions held by women in senior management positions.

Selected/Derived Data

We used all the initial tables considered. Some of the tables had columns that were not significant for our project (we didn't need them for any of the derivations or correlations), so we decided to delete them, such as "Flags". From the initial tables, we selected the attributes "country", "year", "ISCED11", "genre", "employment rate", "percentage by genre and country for educational level", "salary of women per year", "salary of men per year" and "gross domestic product of a country per year".

- **AVG_%** - which represents an average of the level of poverty of men and women (with ages between 15-64), per level of education in a country for a certain year. This gives us information about how poverty-educational levels are related. (Derived)
- **Income Differences** - the differences between men's and women's salary, for the same educational level, for a certain country and year; (Derived)
- **Gender wage gap (GWG)** – in percentage, for a given country, for a certain year. They allow us to understand the contrast existing in the total income, for people with different gender but same educational level, by looking at the GWG and considering only the entries bigger than 5% (the relevant ones to study). (Derived)
- **Growth Rate of Women in High Positions** - that represents the growth rate of women in senior management positions, between years, for a certain country. (Derived)
- **Genre** - which is the sex of the person (female or male, for statistical simplicity). (Selected)
- **Country** - represents the country of birth for each case. (Selected)
- **Year** - Represents at what time the data is referred to. (Selected)
- **Employment rate** - represents the percentage of employment per country and year. (Selected)
- **Percentage by genre and country** - It's the percentage for each measure of ISCED11, per year, country and genre. (Selected)
- **Salary of women/man** - the income of women or man, per year and country. (Selected)
- **ISCED11** - which is the International Standard Classification of Education to measure internationally the level of education of a person.(Selected)
- **GDP** - in euros, represents the market value of all the final goods and services produced.(Selected)

Data Abstraction

All datasets (7) are static tables, with the attributes like Country (nominal attribute that represents the country of origin), Year (continuous and sequential attribute, with hierarchy (because it's time-based) that represents the time we are considering) and:

Genre - It is a nominal attribute ('F' or 'M'), without hierarchy.
ISCED11 - It is a nominal and hierarchical attribute, once there are levels of education, from 0-9.
GDP - It's a sequential and continuous attribute, without hierarchy.
AVG% - It's a sequential and continuous attribute, without hierarchy.
Salary of Women/Men - It's a ratio attribute, without hierarchy.
Income Differences – It's a ratio attribute, without hierarchy.
GWG – It's a ratio attribute, without hierarchy.
Employment rate - It is a ratio attribute, with hierarchy, because it depends on the previous year.
Grow Rate of Women in High Positions – It is a ratio attribute, with hierarchy, because it depends on the previous year.

Data Processing

We used Pentaho with csv files to transform the datasets. We erased the columns that were not needed (Flags and foots, etc.) and ordered the data by Year and then Country, with the help of "select values - Select&Alter". Then, we erased all the values that were from before 2005, the lines EU (27 countries), EU (28 countries) and the countries that do not belong to the EU with the help of "select values - remove". Plus, we used "replace string" for erasing flags (for example "7009.2 b" we erased "b"). Finally, for missing values we input sentinel values. Some of the tables, instead of having the Year column, they had one column for each year. In those cases, we had to rotate the tables using "row normalizer" before doing the process above. For the questions we need to establish correlations between datasets, we used the function of 'Merge Join' after the datasets were treated to our likings. We merged based mostly on the 'Year' and 'Country' parameters, but there were cases where we also used 'ISCED11'. Finally, we removed the repeated columns, created by that tool.

Mapping (Data sample/Questions)

- Q1((Can the poverty level affect the level of education a person can achieve?) Create and AVG_% by relating (B and D)
- Q3 (The level of education per gender has an influence in people's future jobs?) By observing the education level and the employment rate (of a country, for a year, and for genre), we will understand if there is any correlation between them. (D and E)
- Q4 (Is there a contrast in income based on gender with the same education level?) Create Income Differences and Gender wage gap (GWG) by relating (C)
- Q5 (What is the metric between the richness of a country and its gender wage gap?), and looking at the data sample, we can make a parallel between the gender wage gap and the gross domestic product. (A and Q4)
- Q6 (How has the number of females in positions of power changed over the years?), we use the parameter "growth rate of female in high positions" (E and F)
- Q7 (Is inequality on the employment rate related to the richness of a country?), we answer using the variables "employment rate" and "gross domestic product" (A and E)