



Grado de diseño y creación digitales.

Infografía y visualización

Práctica: parte 1, selección de un conjunto de datos.

Autor: **Juan Rafael Reina Valle**

Curso 2023-2024

Contenido

- 1. Elección del conjunto de datos.....3
- 2. Conversión de datos.5
- 3. Bibliografía.....7
 - 3.1 Herramientas usadas.....7

1. Elección del conjunto de datos.

Para elegir el conjunto de datos lo primero que tuve en cuenta es la temática del mismo, me interesan los temas sociales así que accedí al portal de datos abiertos Eurostat <https://ec.europa.eu/eurostat> y busque todos los conjuntos de datos relacionados con asuntos sociales y condiciones de vida.

En el he encontrado un conjunto de datos que me ha interesado especialmente, “Severe material and social deprivation rate by age and sex” (Porcentaje de privación material y social grave por edad y sexo)

https://ec.europa.eu/eurostat/databrowser/view/ilc_mdspd11/default/table?lang=en

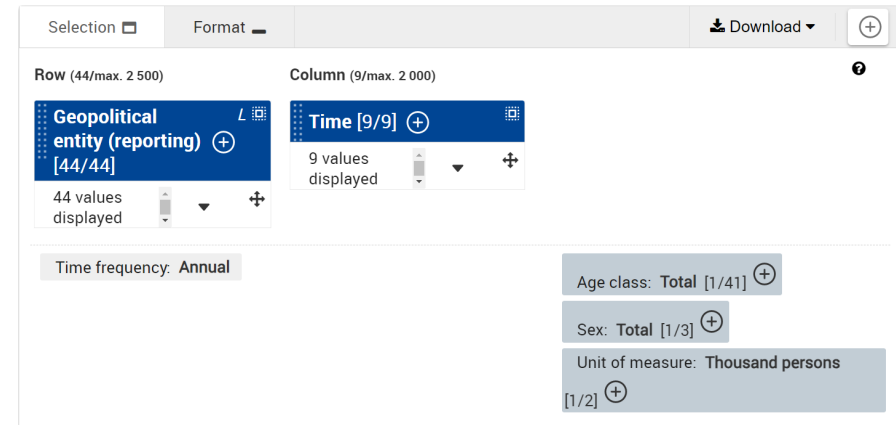
Según la propia explicación del conjunto de datos “Las Estadísticas de la Unión Europea sobre la Renta y las Condiciones de Vida (EU-SILC) recopilan microdatos multidimensionales actualizados y comparables sobre la renta, la pobreza, la exclusión social y las condiciones de vida.” (Eurostat, s. f.)

Aunque inicialmente el conjunto de datos parece sencillo, dado que solo podemos ver una tabla con el porcentaje de personas en riesgo de privación material y social en función del país y el año, la página de Eurostat permite modificarlo para añadir datos que nos interesen.

En mi caso me interesaban la edad y el sexo de los participantes en los datos para poder tener, entre otros, perspectiva de género.

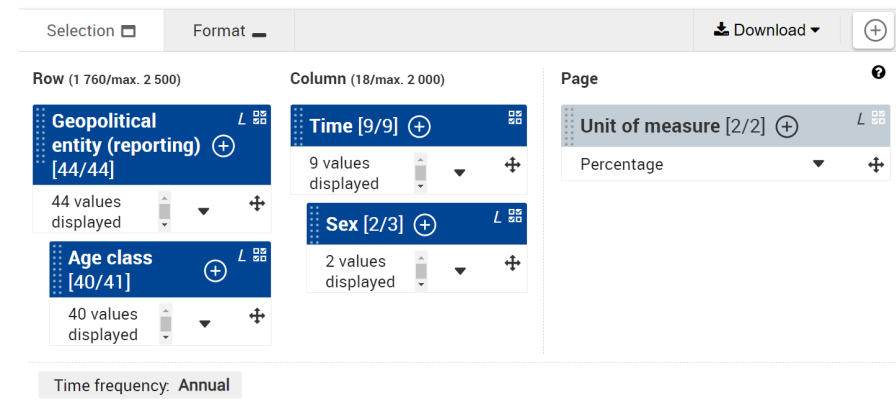
De esta manera pasé de este conjunto de datos:

https://ec.europa.eu/eurostat/databrowser/view/ilc_mdspd11/default/table?lang=en



a este

https://ec.europa.eu/eurostat/databrowser/view/ilc_mdspd11_custom_8195786/default/table?lang=en



Esto hace que el conjunto de datos sea mucho más completo e interesante, ya que incluye estadísticas de edad y perspectiva de género. De tal manera que nos podía proporcionar datos que nos permitieran comprender con mucha más exactitud que corpúsculos sociales son más afectados por esta situación y como ha avanzado esta situación a lo largo de los años.

Sin embargo, al pasar de dos dimensiones, país y año, a cuatro dimensiones la tabla descargada se complica de manera sustancial, mientras que la tabla original tiene este aspecto:

| | TIME | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|------|------|-------|-------|-------|-------|--------|--------|--------|--------|
| GEO (Labels) | | | | | | | | | | |
| European Union (EU6-1958, EU9-1973) | | | | | | | 28 025 | | | |
| European Union - 27 countries (from 2007) | | | | | | | 28 025 | | | |
| European Union - 28 countries (2013-2022) | | | | | | | 28 025 | | | |
| European Union - 27 countries (2007-2013) | | | | | | | | | | |
| Euro area (EA11-1999, EA12-2001, EA12-2002) | | | | | | | 18 852 | 19 998 | 18 798 | |
| Euro area - 20 countries (from 2023) | | | | | | | 19 035 | 20 109 | 19 024 | 20 061 |
| Euro area - 19 countries (2015-2022) | | | | | | | 18 852 | 19 998 | 18 798 | |
| Euro area - 18 countries (2014) | | | | | | | 18 586 | 19 777 | | |
| Belgium | | | 795 | 932 | 794 | 718 | 700 | 758 | 710 | 662 |
| Bulgaria | | | 2 636 | 2 387 | 2 124 | 1 568 | 1 549 | 1 538 | 1 320 | 1 282 |
| Czechia | | | 464 | 360 | 348 | 251 | 215 | 194 | 186 | 211 |
| Denmark | | | 181 | 146 | 205 | 196 | 215 | 201 | 178 | 185 |
| Germany | | | 4 356 | 3 632 | 2 887 | 2 675 | 2 283 | 3 585 | 3 563 | 5 019 |
| Estonia | | | 45 | 29 | 51 | 39 | 34 | 30 | 24 | 40 |
| Ireland | | | 329 | 281 | 247 | 220 | 263 | 205 | 252 | 284 |
| Greece | | | 1 868 | 1 952 | 1 937 | 1 688 | 1 651 | 1 562 | 1 449 | 1 424 |
| Spain | | | 3 356 | 3 878 | 3 813 | 3 999 | 3 537 | 3 930 | 3 801 | 3 554 |
| France | | | 4 080 | 4 101 | 3 880 | 4 079 | 4 384 | 4 305 | 3 719 | 4 855 |
| Croatia | | | 345 | 299 | 289 | 243 | 183 | 171 | 137 | 150 |
| Italy | | | 7 386 | 6 085 | 3 960 | 3 925 | 3 827 | 3 602 | 3 483 | 2 613 |
| Cyprus | | | 67 | 56 | 52 | 33 | 28 | 28 | 23 | 24 |
| Latvia | | | 295 | 240 | 239 | 191 | 137 | 126 | 95 | 135 |
| Lithuania | | | 420 | 438 | 401 | 332 | 266 | 222 | 175 | 163 |
| Luxembourg | | | 13 | 12 | 11 | 9 | 8 | 10 | 13 | 10 |

La tabla con los datos añadidos tiene una complejidad añadida al disponer de datos de sexo y edad.

| | | TIME | | 2014 | | 2015 | | 2016 | |
|--------------|---------------------|--------------|--|-------|--|---------|------|-------|------|
| | | SEX (Labels) | | Males | | Females | | Males | |
| GEO (Labels) | AGE (Labels) | | | | | | | | |
| Belgium | Less than 6 years | | | | | 12,3 | 10,1 | 10,6 | 13,4 |
| Belgium | From 6 to 10 years | | | | | 10,8 | 11,0 | 11,6 | 12,8 |
| Belgium | From 6 to 11 years | | | | | 10,9 | 11,0 | 10,9 | 12,5 |
| Belgium | From 11 to 15 years | | | | | 8,1 | 9,3 | 10,1 | 11,4 |
| Belgium | From 12 to 17 years | | | | | 6,9 | 8,2 | 8,8 | 10,1 |
| Belgium | From 15 to 19 years | | | | | 7,0 | 7,8 | 6,8 | 10,7 |
| Belgium | From 15 to 24 years | | | | | 6,0 | 7,5 | 8,7 | 9,0 |
| Belgium | From 15 to 29 years | | | | | 6,1 | 7,8 | 8,1 | 9,1 |
| Belgium | Less than 16 years | | | | | 10,5 | 10,1 | 10,8 | 12,6 |
| Belgium | From 16 to 19 years | | | | | 5,8 | 6,8 | 6,4 | 9,6 |
| Belgium | From 16 to 24 years | | | | | 5,4 | 7,1 | 8,8 | 8,3 |
| Belgium | From 16 to 29 years | | | | | 5,7 | 7,5 | 8,1 | 8,7 |
| Belgium | From 16 to 64 years | | | | | 6,5 | 8,0 | 7,5 | 9,7 |
| Belgium | 16 years or over | | | | | 5,8 | 7,2 | 6,6 | 8,7 |
| Belgium | Less than 18 years | | | | | 10,1 | 9,8 | 10,1 | 12,1 |
| Belgium | From 18 to 24 years | | | | | 5,1 | 7,1 | 9,9 | 8,6 |
| Belgium | From 18 to 59 years | | | | | 6,6 | 8,0 | 7,7 | 9,6 |

Como podemos observar, ahora las columnas tienen datos combinados y no es tan sencillo extraer los datos para su análisis.

El conjunto de datos seleccionado según esta variación dispone de las siguientes características:

1441 registros.

20 variables.

Variables categóricas, como los grupos de edad.

Datos combinados, en una variable disponemos del porcentaje de personas en riesgo de exclusión tanto para un año como para un sexo en concreto. Por tanto deberemos de desnormalizar dichas variables para separarlas en variables nominales, como el sexo, y variables cuantitativas como el porcentaje.

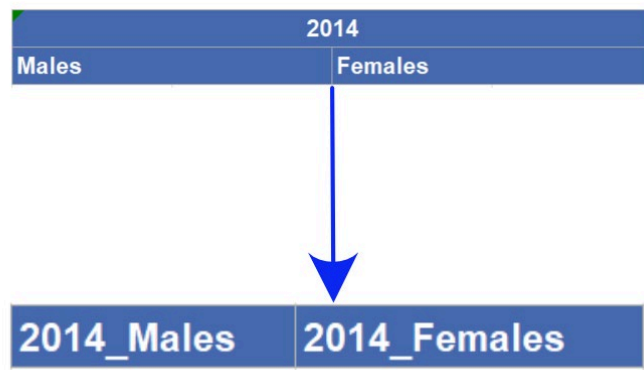
Formato de datos Wide(Minguillón, 2023), lo cual complica su tratamiento con Tableau o Flourish

2. Conversión de datos.

Para ello podemos usar muchas variantes, desde copiar y pegar en el propio Excel hasta programas de cálculos estadísticos como R.

He elegido para hacerlo la plataforma Google Collab dado que me permite usar la potencia de Python, el paquete estadístico Pandas y las hojas de trabajo de Jupyter, todo ello sin necesidad de instalar nada en el propio ordenador y con la capacidad de poder compartirlo en Internet.

El primer paso será limpiar los datos del Excel, para ellos dividiremos la celdas combinadas del año y sexos en dos celdas independientes y crearemos un csv que subiremos a GitHub



El objetivo de esto es facilitar el poder hacer pivote(Minguillón, 2023) de formato wide a long mediante la función [melt de pandas](#) en Google Collab.

The diagram illustrates the transformation of a wide table into a long table. At the top, a table with four columns, 'GEO', 'AGE', '2015_Males', and '2015_Females', is shown. A green arrow points down to a new table with five columns, 'GEO', 'AGE', 'year_sex', and 'PovertyRate'. A red arrow points from the '2015_Males' column of the top table to the 'year_sex' column of the bottom table. A blue arrow points from the '2015_Females' column of the top table to the 'year_sex' column of the bottom table.

| | GEO | AGE | 2015_Males | 2015_Females |
|---|---------|---------------------|------------|--------------|
| 0 | Belgium | Less than 6 years | 12,3 | 10,1 |
| 1 | Belgium | From 6 to 10 years | 10,8 | 11,0 |
| 2 | Belgium | From 6 to 11 years | 10,9 | 11,0 |
| 3 | Belgium | From 11 to 15 years | 8,1 | 9,3 |
| 4 | Belgium | From 12 to 17 years | 6,9 | 8,2 |

| | GEO | AGE | year_sex | PovertyRate |
|---|---------|---------------------|--------------|-------------|
| 0 | Belgium | Less than 6 years | 2015_Males | 12,3 |
| 1 | Belgium | From 6 to 10 years | 2015_Males | 10,8 |
| 2 | Belgium | From 6 to 11 years | 2015_Males | 10,9 |
| 3 | Belgium | From 11 to 15 years | 2015_Males | 8,1 |
| 4 | Belgium | From 12 to 17 years | 2015_Males | 6,9 |
| 5 | Belgium | Less than 6 years | 2015_Females | 10,1 |
| 6 | Belgium | From 6 to 10 years | 2015_Females | 11,0 |
| 7 | Belgium | From 6 to 11 years | 2015_Females | 11,0 |
| 8 | Belgium | From 11 to 15 years | 2015_Females | 9,3 |
| 9 | Belgium | From 12 to 17 years | 2015_Females | 8,2 |

Una vez tenemos los datos en dicho formato podemos dividir la columna year_sex en dos columnas mediante la función [str.split de pandas](#).

| | GEO | AGE | year_sex | PovertyRate |
|---|---------|---------------------|--------------|-------------|
| 0 | Belgium | Less than 6 years | 2015_Males | 12,3 |
| 1 | Belgium | From 6 to 10 years | 2015_Males | 10,8 |
| 2 | Belgium | From 6 to 11 years | 2015_Males | 10,9 |
| 3 | Belgium | From 11 to 15 years | 2015_Males | 8,1 |
| 4 | Belgium | From 12 to 17 years | 2015_Males | 6,9 |
| 5 | Belgium | Less than 6 years | 2015_Females | 10,1 |
| 6 | Belgium | From 6 to 10 years | 2015_Females | 11,0 |
| 7 | Belgium | From 6 to 11 years | 2015_Females | 11,0 |
| 8 | Belgium | From 11 to 15 years | 2015_Females | 9,3 |
| 9 | Belgium | From 12 to 17 years | 2015_Females | 8,2 |

| | GEO | AGE | Year | Sex | PovertyRate |
|---|---------|---------------------|------|---------|-------------|
| 0 | Belgium | Less than 6 years | 2015 | Males | 12,3 |
| 1 | Belgium | From 6 to 10 years | 2015 | Males | 10,8 |
| 2 | Belgium | From 6 to 11 years | 2015 | Males | 10,9 |
| 3 | Belgium | From 11 to 15 years | 2015 | Males | 8,1 |
| 4 | Belgium | From 12 to 17 years | 2015 | Males | 6,9 |
| 5 | Belgium | Less than 6 years | 2015 | Females | 10,1 |
| 6 | Belgium | From 6 to 10 years | 2015 | Females | 11,0 |
| 7 | Belgium | From 6 to 11 years | 2015 | Females | 11,0 |
| 8 | Belgium | From 11 to 15 years | 2015 | Females | 9,3 |
| 9 | Belgium | From 12 to 17 years | 2015 | Females | 8,2 |

En el siguiente [github](#) público se pueden ver:

[El fichero de datos original.](#)

[El csv que he subido a Google collab](#)

[La hoja de trabajo en Google Collab con la que he hecho la conversión de la tabla.](#)

[El csv final que usaremos en Tableau.](#)

Este csv final es mucho más facil de tratar en Tableau y será el que usaremos para crear las infografías de la práctica. Dado que tenemos tanto información de sexo, de las edades y de la evolución a través de los años podremos hacer varias graficas que respondan a múltiples preguntas:

- ¿Influye el sexo en el índice de pobreza?
- ¿Qué diferencias hay entre los países de la Unión Europea?
- ¿Como han evolucionado los índices de pobreza a través de los años?
- ¿Como influye la edad de los habitantes en el índice de pobreza?
- ¿Cómo se relacionan todos los anteriores conceptos entre ellos, ¿con el paso de los años las posibles diferencias entre sexos se igualan o se incrementan?¿como se relacionan la edad y el sexo?

3. Bibliografía.

Alcalde, I., & Minguillón, J. (s. f.). *Introducción a la visualización de la información*. Recuperado 20 de noviembre de 2023, de https://materials.campus.uoc.edu/daisy/Materials/PID_00272000/html5/PID_00272000.html?utm_source=meus_materials_app&utm_medium=campus&utm_campaign=multiformat

Minguillón, J. (2023, noviembre 16). *Introducción a la preparación de datos*. Recuperado 28 de noviembre de 2023, de https://aula.uoc.edu/courses/9391/pages/recursos-de-aprendizaje-de-la-practica?module_item_id=852344

Eurostat. (s. f.). *Income and living conditions (ilc)*. Recuperado 20 de noviembre de 2023, de https://ec.europa.eu/eurostat/cache/metadata/en/ilc_sieusilc.htm

Pascual Cid, V. (s. f.). *Buenas prácticas en visualización de datos*. Recuperado 20 de noviembre de 2023, de https://materials.campus.uoc.edu/daisy/Materials/PID_00272015/html5/PID_00272015.html?utm_source=meus_materials_app&utm_medium=campus&utm_campaign=multiformat

3.1 Herramientas usadas.

<https://jupyter.org/>

<https://github.com/>

<https://colab.research.google.com/>

<https://pandas.pydata.org/>