MEMORIA PRÁCTICA 1:

EXTRACCIÓN DE INFORMACIÓN DE LA ESTADÍSTICA "MERCADO DE TRABAJO Y PENSIONES" DE LA AGENCIA TRIBUTARIA (AEAT)

1. Contexto:

La información se ha recolectado en un contexto de investigación estadística sobre las diferentes fuentes de datos que hay en España acerca de salarios, con el objetivo de poder realizar estudios comparativos a nivel de CCAA, distinguiendo principalmente las variables sexo y edad.

Aparte del Instituto Nacional de Estadística, que difunde datos salariales a nivel CCAA fácilmente descargables, recolectados a partir de encuestas (Encuestas de Coste Laboral y Encuestas de Estructura Salarial). Otra fuente muy interesante que me encontré fue la Agencia Tributaria, que publica una estadística denominada "Mercado de trabajo y pensiones en las fuentes tributarias", una investigación censal realizada a partir de la información que este organismo recolecta a partir de la Declaración Anual de Retenciones e Ingresos a Cuenta sobre Rendimientos del Trabajo (Modelo 190), que presentan anualmente los empleadores.

https://sede.agenciatributaria.gob.es/AEAT/Contenidos_Comunes/La_Agencia_Tributaria/Estadisticas/Publicaciones/sites/mercado/2021/jrubikf556e330817d393a43709315680b99c80da2c25b9.html

Me pareció una estadística muy útil y muy completa, ya que se analizaban todas las variables interesantes para el estudio con un nivel de desagregación mayor que las encuestas del INE, pero me encontré con que la Agencia Tributaria publica las tablas en su web en formato HTML sin enlaces para poder descargarlas y tampoco dispone de API para poder acceder a los datos cómodamente. Con lo que la única opción para un usuario no experto en temas informáticos sería hacer un copia y pega de las tablas que se necesiten, un método muy rudimentario. Por lo que vi la ocasión perfecta de aplicar el uso del web scraping y así mecanizar la extracción y rápidamente extraer todos los cruces posibles de la tabla que necesito.

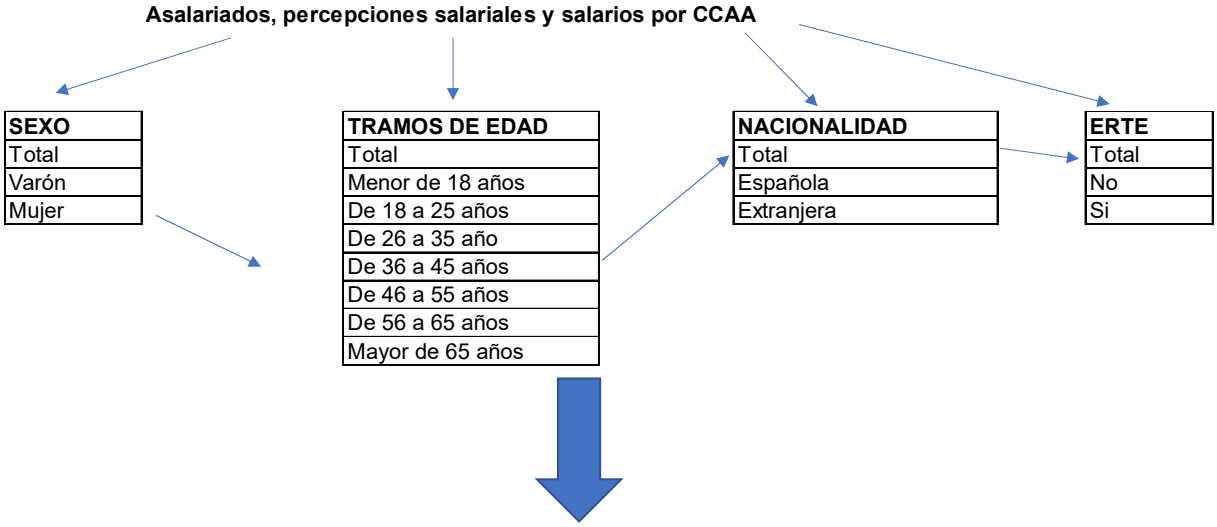
2. Título

"Asalariados, percepciones salariales y salarios por CCAA", ya que esta sería la información principal de todas las tablas extraídas, independientemente de que se hagan cruces por sexo, grupos de edad, nacionalidad y situación de ERTE.

3. Descripción del dataset

El dataset extraído está constituido por un conjunto de tablas, una para cada uno de los posibles cruces de las variables Sexo (Total, Varón, Mujer), Tramos de edad (Total, Menor de 18 años, De 18 a 25 años, De 26 a 35 años, De 36 a 45 años, De 46 a 55 años, De 56 a 65 años, Mayor de 65 años), Nacionalidad (Total, Española, Extranjera) y ERTE (Total, No, Si). Por lo que hay 216 cruces y tablas posibles, y en cada una de ellas se mostraría la información correspondiente sobre asalariados, percepciones salariales y salarios por CCAA.

4. Representación gráfica



	Asalariados	Percepciones por persona	Salarios	Salario Medio Anual
Total	2.394.456	1,3	48.693.595.162	20.336
Andalucía	440.814	1,38	7.194.739.551	16.321
Aragón	71.561	1,28	1.417.880.037	19.814
Asturias, Principado	50.258	1,24	966.447.502	19.230
Balears,Illes	66.359	1,24	1.300.679.542	19.601
Canarias	115.505	1,24	2.026.549.767	17.545
Cantabria	30.929	1,26	588.540.429	19.029
Castilla y León	117.613	1,29	2.229.124.728	18.953
Castilla - La Mancha	107.070	1,3	1.982.379.921	18.515
Cataluña	430.553	1,28	10.224.171.419	23.747
Comunitat Valenciana	266.470	1,31	5.002.952.212	18.775
Extremadura	51.860	1,41	831.359.589	16.031
Galicia	143.477	1,26	2.728.100.608	19.014
Madrid, Comunidad de	398.661	1,26	10.297.343.246	25.830
Murcia, Región de	79.976	1,33	1.436.814.433	17.966
Rioja, La	16.446	1,28	317.304.754	19.294
Ceuta	3.462	1,22	73.317.019	21.178
Melilla	3.442	1,25	75.890.406	22.048

Combinando entre sí las distintas categorías correspondientes a cada una de las variables, se obtendrían todas las tablas del dataset. En el diagrama se pone el ejemplo de una, la

correspondiente a los datos del nº de personas asalariadas, percepciones salariales y salarios por CCAA, de la población de mujeres que tienen entre 36 a 45 años, de todas las nacionalidades y que están o no en situación de ERTE.

5. CONTENIDO

El data set incluiría entonces a todas las posibles tablas ordenadas en un dataframe constituído por 3.600 filas (no hay 216x18 = 3.888 observaciones porque para algunos cruces no se publican datos de Ceuta y Melilla) con las siguientes variables:

- Espacio
- Nº de personas asalariadas: nº de personas empleadas por cuenta ajena en un determinado año contadas de forma única, independientemente de que hayan trabajado para una o varias empresas o entidades.
- Percepciones salariales por persona: nº medio de retribuciones salariales percibidas por una persona y pagadas por distintas empresas en ese año.
- Salarios: suma de los salarios anuales recibidos por cada una de esas personas.
- Salario medio anual: salarios/asalariados
- Sexo
- Edad
- Nacionalidad
- Erte

El período de referencia de los datos es el año 2021.

6. PROPIETARIO

El propietario de estos datos es la Agencia Tributaria dependiente del Ministerio de Hacienda y Función Pública del Gobierno de España.

En cuanto a los principios éticos y legales de este proyecto, al ser información estadística difundida por un organismo público, no le aplica ningún tipo de protección especial, ya que no le afecta la LOPD al ser datos agregados y totalmente anonimizados. Tan solo habría que citar que la fuente es la AEAT al realizar cualquier tipo de estudio.

Algunas citas relacionadas con estos datos las encontramos en recientes noticias de prensa:

https://www.20minutos.es/noticia/5076824/0/la-mitad-de-los-espanoles-ganan-21-347-euros-o-menos-de-sueldo-al-ano-segun-la-agencia-tributaria/

7. INSPIRACIÓN

Este conjunto de datos resultaría muy interesante para realizar estudios sobre la brecha salarial entre hombres y mujeres, comparándola por CCAA, grupos de edad y nacionalidad. En la noticia anterior se señala que la brecha en 2021 alcanza el 10%.

Sacando este mismo dataset para años anteriores podríamos establecer a su vez una comparativa temporal.

8. LICENCIA

La licencia que le aplicaría sería Open Database License (ODbL), al tratarse de información pública. Esta permite a los usuarios hacer uso libre de los datos contenidos en el repositorio sin temor a la infracción de derechos de autor, para compartir, para crear y para adaptar; bajo las condiciones de Atribución, Compartir Igual y Mantener Abierta.

9. CÓDIGO

El código se ha implementado usando el lenguaje de programación R. El script correspondiente se ha subido a la carpeta /source del repositorio.

Se han usado los paquetes:

- rvest: para capturar páginas estáticas, lee el html de la página
- htttr: para trabajar con páginas web
- glue: para combinar textos
- openxlsx: para manipular libros excel
- tidyverse, tidyselect: para manipular datos

Para realizar el proceso de extracción de los datos se partió de la página web de una tabla inicial (en la que todas las variables son igual a Total, que se llamó pag_url) y se observó que las páginas de las distintas tablas que se podían crear, cruzando variables, tenían todas una estructura similar, en la que cada página contenía su vez en su html 17 enlaces.

Por ejemplo, en la página inicial, los 17 enlaces vincularían a las tablas (1.Total Total Total Total; 2. Varón Total Total Total; 3. Mujer Total Total Total; 4. Total Total Total Total; 5. Total Menor de 18 años Total Total, ….)

Por lo que si se iban recorriendo las páginas que había dentro de otras de forma recursiva, al final se obtendrían todos los cruces posibles.

Se crearon dos funciones:

Una para extraer todos los enlaces que estaban contenidos en una determinada página.

Otra que extraía la tabla correspondiente a un determinado enlace.

Se trabajó con la herramienta Selector Gadget para construir las expresiones css de los elementos que se querían seleccionar. Dar con el selector adecuado fue la parte que me resultó más dificultosa.

10. Enlace zenodo

https://doi.org/10.5281/zenodo.7343586

11. Enlace vídeo

https://drive.google.com/file/d/1wqRflKi3yOlNrjieOplASfVu9tZOqs07/view?usp=share_link

Esta práctica se ha realizado individualmente por Lucía Fernández González.