Herramienta para visualización de la Encuesta de Población Activa (EPA)

Objetivos

Desarrollar una herramienta que permita visualizar de varias formas los datos de la Encuesta de Población Activa para permitir un análisis cuantitativo y cualitativo sencillo y automatizado de varios aspectos de la sociedad española. Para ello, se diseñan 3 herramientas:

1. Visualizador 2D: Una *heatmap* de dos valores de la encuesta para buscar relaciones entre esas dos variables.
2. Gráficos de barras automáticos de una variable o una serie de variables a modo de comparación relativa entre ellas.
3. Evolución temporal y estudio de la asimetría a lo largo de un determinado período de tiempo.

Instrucciones Previas

1. Ejecuta la celda de las librerías.
2. Procesado:
    a. Escoge los años que quieres y los trimestres que quieres de cada año

    ```
    trimestres = [1,2,3,4] #D
    anyos = [2019,2020,2021]
    ```

    En este caso cogeríamos los trimestres 1 a 4 (todos) de los años 2019 a 2021 y combinaríamos los resultados en un solo DataFrame de forma automática. Podemos escoger datos desde 2005 hasta 2021. En caso de no estar disponibles unos datos, como por ejemplo los del año 2022 (en la fecha que escribo esto, Diciembre 2021) saltará un mensaje diciendo que esos datos no están disponibles y el programa continuará sin esos datos.
    b. Escoger si quieres que iguale el número de hombres y mujeres de la encuesta. Si hay por ejemplo 500000 hombres y 490000 mujeres en los datos escogidos, el algoritmo elimina 500000 - 490000 = 10000 hombres aleatorios de la encuesta. Esto a veces resulta útil si quieres comparar asimetrías y cosas así pero puede cambiarse.
    c. Escoger si quieres cambiar los valores numéricos de algunas variables por sus valores en texto, cuyo significado puede se puede encontrar en los ficheros de “Diseño de registro y valores válidos” en la página web del INE, en la sección de EPA en la pestaña de Microdatos https://bit.ly/3om1abK. De momento están

disponibles únicamente algunas de las variables. Por ejemplo, si pones sust_regna = True los números de esta columna que indican la región de nacimiento de las personas se cambian por el lugar en sí, como España, Norteamérica, etc. Por defecto se cambian todos los posibles.

Heatmap Histogram 2D

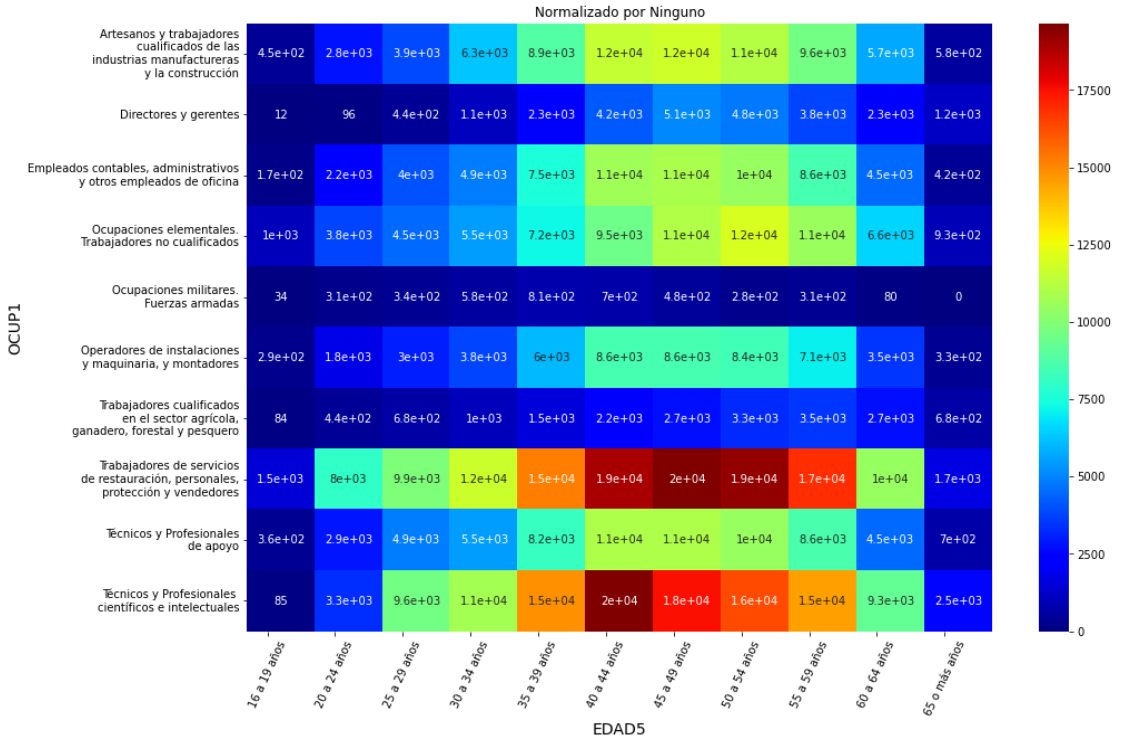
Esta sección contiene un código que dadas dos variables de la EPA, te dibuja una especie de histograma 2D, con el que se puede visualizar relaciones entre esas dos variables. En cada celdilla hay un número de personas o una fracción de un total. Para aclarar el funcionamiento, veamos un ejemplo.

Digamos que queremos estudiar la relación entre la edad y el sector donde trabaja la gente. El orden no es muy importante, pero nosotros elegiremos en vertical, como *feature* 1 (feat1) el sector, OCUP1, y en horizontal (feat2) la edad, EDAD5 [1]. Podemos imponer una limitación a la cantidad de valores que cogerá el programa para hacer la gráfica, limi y limi2. De este modo, si ponemos por ejemplo limi = 3 y limi2 = 5 solo estaremos representando las 3 categorías de ocupación profesional más populares (por ser la primera variable) y los 5 grupos de edad más poblados. Si ponemos un número muy alto, como 99 por ejemplo, selecciona todas las categorías. Esta es la opción que yo suelo elegir, porque me gusta ver la imagen completa, aunque las gráficas resultantes pueden ser algo grandes.

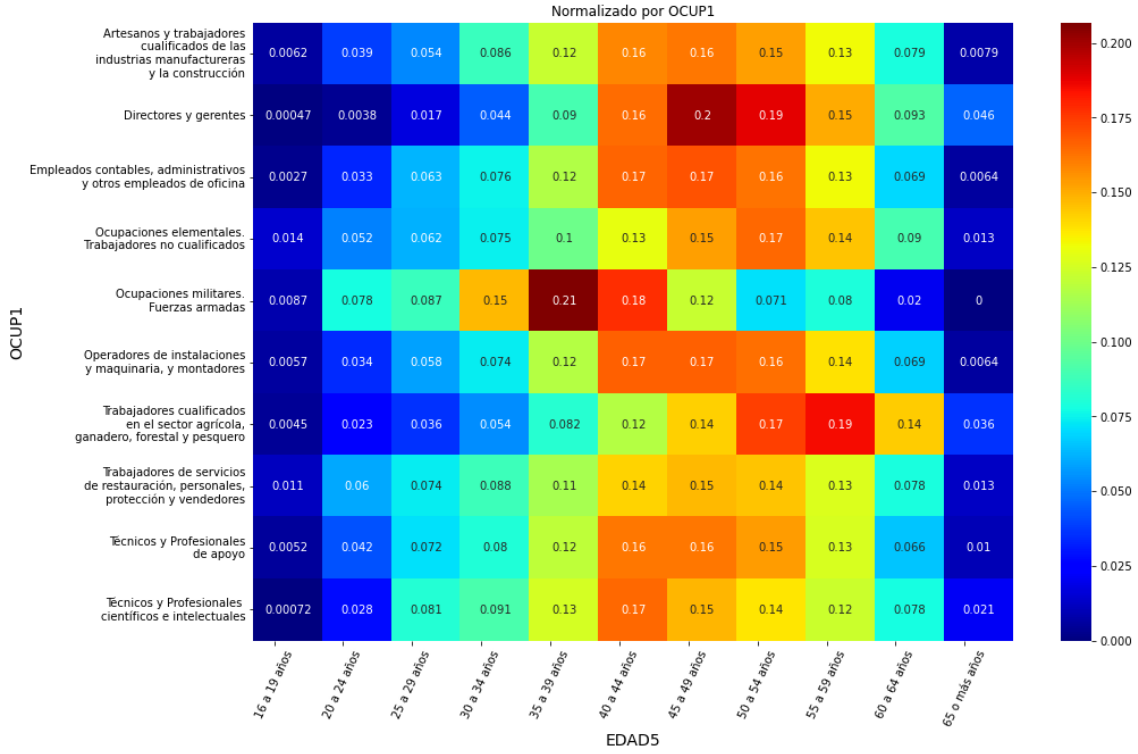
```
feat1 = 'OCUP1'
feat2 = 'EDAD5'
limi = 99
limi2 = 99
```

Ahora tenemos que elegir si normalizar por alguna de las variables o no. Al normalizar, estamos dividiendo entre el total de una categoría. Digamos que hay muy poca gente en el grupo de los militares. Si no normalizamos (normalize = False) , la imagen resultante sería algo como la figura 1a, donde en las ocupaciones militares vemos una franja azul, y no tenemos información acerca de la distribución de edades. Solo vemos las combinaciones con más gente, que ahora mismo no nos es muy útil. Si normalizamos, sin embargo, por ocupación (normalize = 1) la imagen nos estaría mostrando cómo se distribuyen las edades dentro de una misma ocupación. La suma de todas las celdillas de una ocupación entonces da 1. Esta es la figura 1b, en la que claramente vemos que los puestos militares los ocupan principalmente jóvenes, mientras que el sector de la agricultura y la pesca es un sector más envejecido. Podríamos también normalizar por edades, para ver qué sector prefiere cada grupo de edad, poniendo normalize = 2. Este resultado no aparece en la figura 1.

[1] En 2021, la edad cambia de EDAD1 a EDAD5, pero eso ya lo tenemos en cuenta en el procesado y le cambiamos el nombre a EDAD5 porque así es como lo pone el resto de años.



(a)



(b)

Figura 1. Heatmaps 2D de Ocupación y Edad usando datos de la EPA del INE. (a) Sin normalizar. (b) Normalizados por ocupación.

Este proceso se puede llevar a cabo con todas las variables de la encuesta de población activa, aunque es interesante hacerlo primero con las que están sustituidas por su significado. Algunos ejemplos interesantes son ocupación y región de nacimiento, o región de nacimiento y nivel de estudios.

Por último, puedes poner un filtro de edad por si quieres mirar datos relativos a menores o mayores de cierto límite. Por ejemplo se puede mirar la relación entre el sector de trabajo y el nivel de estudios pero solo en menores de 40 años.

Las opciones
```
sort_1 = True
sort_2 = True
```
sirven para ordenar alfabéticamente (o de menor a mayor) las categorías dentro de las variables 1 y 2.

Bar Plots

Esta sección te da la opción de hacer gráficos de barras de cualquier variable que quieras de forma automática, y con diversas opciones se explican en el notebook. Un ejemplo es el de la figura 2.

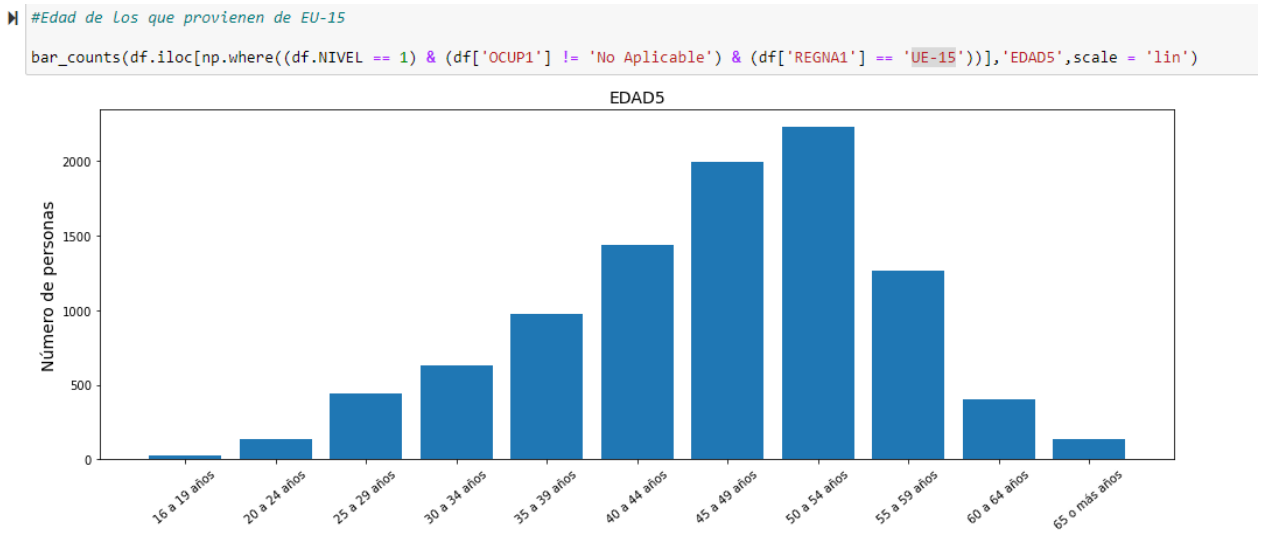


Figura 2. Ejemplo de gráfico de barras, en el que se muestra la distribución de edades de las personas encuestadas provenientes de países de la categoría UE-15 (ver Diseño de Registro EPA).

También es posible hacer diagramas de barras de una variable pero para varios valores de otra variable. Un ejemplo de esto sería un gráfico de la distribución de edades, pero comparando varios países de origen. Así vemos, en la figura 3 que en España, la población de origen africano y sudamericano es en general más jóven que la población de origen UE-15.

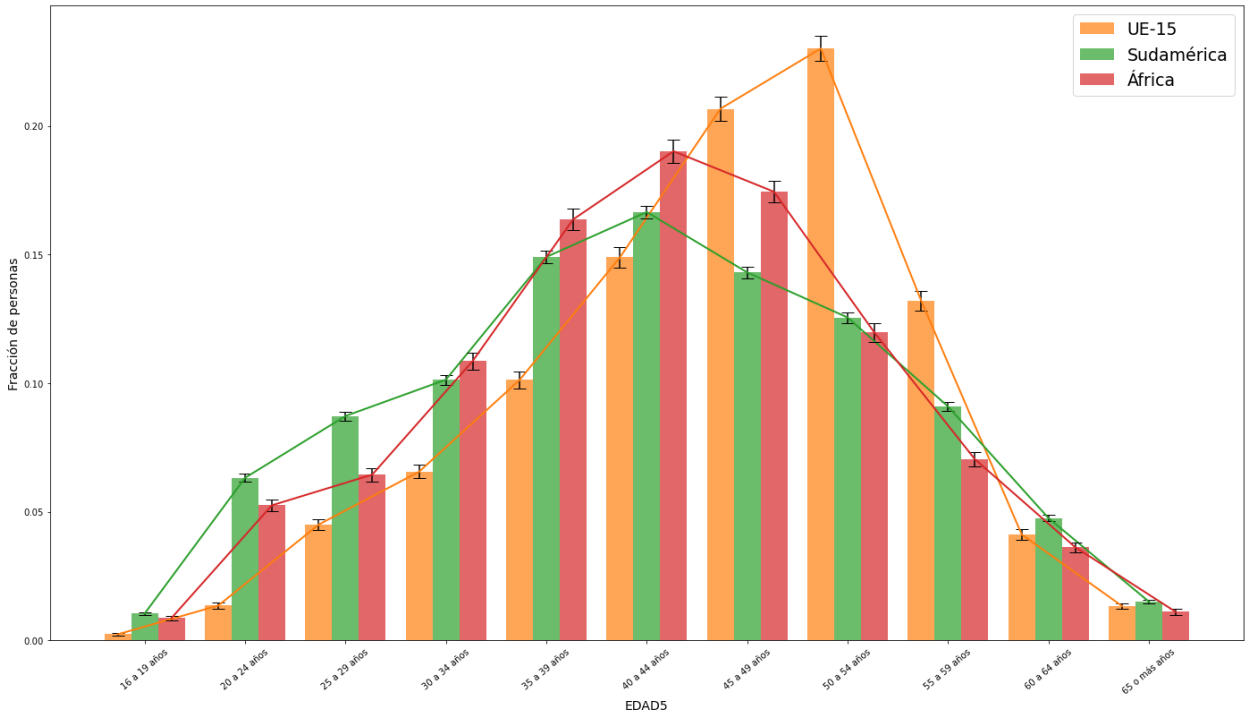


Figura 3. Ejemplo de gráfico de barras múltiple, en el que se muestra la distribución de edades de las personas encuestadas provenientes de países de la categoría UE-15, Sudamérica y África (ver Diseño de Registro EPA). Al comparar las tres distribuciones, la población más envejecida es la de UE-15.

Evolución Temporal

Seleccionando las opciones de forma muy parecida a lo que se explica en las instrucciones previas, podemos recopilar los datos de muchos años y estudiar la serie temporal. Podemos coger por ejemplo el primer y tercer trimestre de cada año desde 2005 hasta la actualidad. A partir de ahí, se puede hacer infinidad de análisis, que todavía están en desarrollo. Un ejemplo es el de la asimetría hombre-mujer o asimetría de género a lo largo del tiempo.

La asimetría de género la definimos como el número de hombres menos el número de mujeres entre el número de personas. Por ejemplo, podemos hacer eso para puestos de trabajo, y tendríamos cómo se van masculinizando, feminizando o igualando los distintos sectores. Actualmente vemos que hay el doble de mujeres que de hombres en puestos de administración y oficina y el doble de hombres que de mujeres en puestos directivos. La evolución temporal de estas cifras se muestra, calculada como la asimetría, en la figura 4. La asimetría de género se puede estudiar para cualquier variable que se escoja.

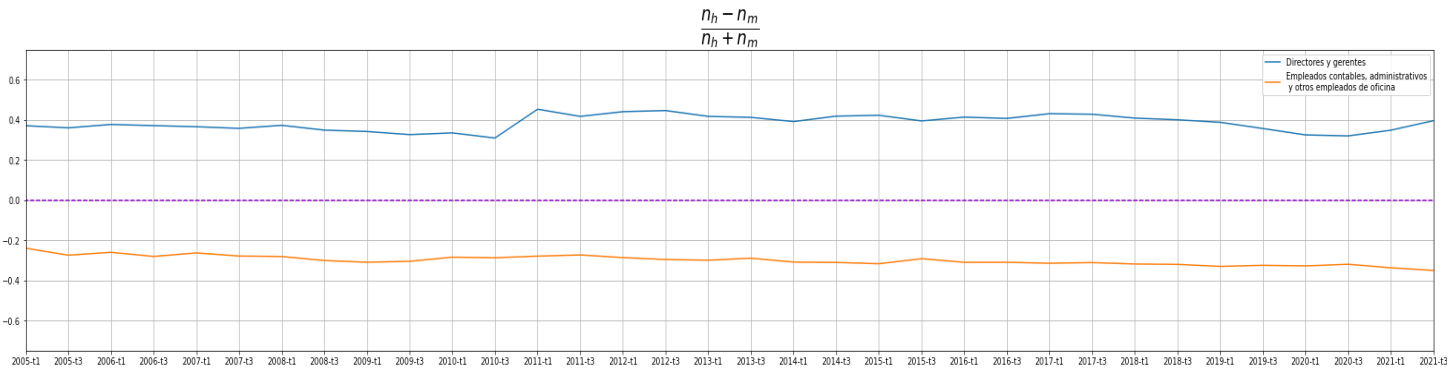


Figura 4. Evolución temporal de la asimetría en puestos directivos (azul) y en puestos de oficina (amarillo). Cuando la asimetría es 0, en dicho sector hay el mismo número de hombres que de mujeres (línea punteada). Cuando la asimetría está por encima de esta línea, hay más hombres y por debajo hay más mujeres. En 2011 hubo un cambio en la definición CNACE de lo que se considera puestos directivos, cosa que explica el salto de la gráfica.