Materia : Ciencia de Datos

Alumnas: Amorena Ines, Campasso Delfina y Lauro Renata

Profesores: Anchorena Ignacio y Buscaglia Tomás

Primavera 2025

LINK REPOSITORIO: https://github.com/iamorena/TP1

1.Introducción y marco teórico

Consideramos que en el presente trabajo es necesario realizar un paréntesis teórico que introduzca al lector en la definición que el INDEC da sobre lo que es la pobreza y la indigencia.

"Las nociones de pobreza e indigencia empleadas por el INDEC para el cálculo de incidencia se corresponden con el método de medición indirecta, denominado también "línea"..."
Para establecer una correcta definición de pobreza, hay que establecer el concepto de línea de indigencia. El INDEC define la línea de indigencia como "si los hogares cuentan con ingresos suficientes para cubrir una canasta de alimentos capaz de satisfacer el umbral mínimo de necesidades energéticas y proteicas, denominada Canasta Básica Alimentaria (CBA)." Los hogares que no superen esta línea básica son considerados indigentes. La referencia para evaluar la CBA es el Índice de Precios al Consumidor (IPC). La CBA se calcula mediante el Coeficiente de Engel, que es Gasto Alimentario / Gasto Total.

$$\textbf{Coeficiente de Engel} = \frac{\textbf{Gasto Alimentario}}{\textbf{Gasto Total}}$$

Por otro lado, la línea de pobreza (LP), no solo incluye insumos alimentarios mínimos sino que también agrega consumos básicos no alimentarios, es decir, la Canasta Básica Total (CBT). Como consecuencia, se le suma a la CBA bienes y servicios no alimentarios, como vestimenta, transporte, salud, educación, entre otras. Esta canasta se basa en los hábitos de consumo generales de la población, tanto alimentarios como no alimentarios. El calculo de la CBT consiste en multiplicar el valor de la CBA por la inversa del Coeficiente de Engel.

$$\textbf{CBT} = \textbf{CBA} * \textbf{ICE}$$

Las líneas de hogar se construyen de acuerdo a su tamaño y disposición. Además, existen medidas de referencia en cuanto a necesidades nutricionales de un "adulto equivalente" a partir de la cual se establecen las relaciones en necesidades energéticas según edad y sexo. A partir de las medidas de referencia y de la construcción particular de cada hogar, se comparan los ingresos totales del hogar en cuestión con las necesidades correspondientes.

2. Metodología

La recolección de datos se obtuvo mediante la página web del Instituto de Estadísticas y Censos (INDEC). Descargamos los microdatos de la Encuesta Permanente de Hogares (EPH) del primer trimestre de 2005 y de 2025. A partir de la mencionada base, decidimos trabajar con

la región del Gran Buenos Aires (GBA). Consecuentemente, eliminamos los datos que no correspondiera a nuestra región, obteniendo una nueva base únicamente con los datos que nos resultaban relevantes. Trasladando esto a *python,* para realizar una única base de datos que englobe a las dos previamente descargadas, tuvimos que hacer un filtro de las columnas que tenían distintas. Luego, a las columnas que eran iguales en ambas bases las normalizamos, es decir, las convertimos todas a minúsculas. Para realizar el correcto filtrado, también tuvimos que normalizar la región ya que en una de las dos bases estaba indexada con un "1", mientras que en la otra como "Gran Buenos Aires".

Luego, seleccionamos 15 variables de interés del dataset original y calculamos los valores no nulos y los valores faltantes por cada año. Visualizamos los resultados mediante dos heatmaps (Figura 1), herramienta que sirve para comparar rápidamente diversas variables entre categorías. Quisimos identificar qué variables presentaban mayor proporción de valores faltantes y en que años ocurría (ver figura 1 en el anexo).

De acuerdo con el gráfico obtenido, podemos ver que la mayoría de las variables presentan una cercanía al 100% de disponibilidad de datos, lo cual indica consistencia en el relevamiento. De todas formas, pudimos observar excepciones en las variables "sector_trabajo_prinicpal" y "tiene_socios_familiares", específicamente en años más recientes. Para su correcta lectura y fácil entendimiento, cambiamos los nombres de las variables que aparecían en la base de datos ("cho04, cho06…) por sus nombres reales. Por otro lado, en el código, el gráfico de "Missing Data" muestra exactamente lo opuesto.

Posteriormente corregimos variables que tuvieran valores carentes de sentido, de acuerdo a la documentación de la EPH, como por ejemplo la edad <0 0 >110, ingresos negativos etc... El proceso de limpieza consistió, en primer lugar, en normalizar la variable "sexo". Luego, pasamos todos los datos a valor string, trim, en minúscula y sin acentos. De esta forma logramos unificar codificaciones heterogéneas de sexo y asegurarnos de que solo existan tres posibilidades; 1, 2 o NaN. Luego definimos grupos de variables y pruebas. Para ello, armamos una lista por tipo de control para poder aplicar reglas homogéneas sin modificar columnas inexistentes.

Para los valores negativos de la variable "Ingresos", se forzó a que los valores deberían tomar obligatoriamente un formato numérico, de lo contrario serían NaN. También se detectaron negativos y se redecodificaron y se tomó un registro de los casos afectados por la limpieza. Para la variable "edad", se realizó algo similar: se limpiaron valores biológicamente imposibles o errores en la carga. En los casos de variables binarias, se aseguró que en los casos en que no hubiera {1,2} se indicará no respuesta. Posteriormente, imprimimos una bitácora de recodificaciones por regla o variable, que muestra el tamaño final de la base y que nos asegurara que la normalización no hubiera vaciado años enteros. De esta forma, obtuvimos una nueva base, la cual preserva la mayor cantidad posible de información válida, sin introducir imputaciones.

Por último, gracias a un *print* notamos que la variable “sexo” estaba indexada de manera distinta en ambas bases iniciales. En *base_gba2025* se indicaba el género a través de una variable dicotómica (1: varón, 2:mujer) mientras que en *base_gba2005* estaba escrito en string “Varón”/”Mujer”. Para evitar posibles NaN en ejercicios posteriores, normalizamos aquel detalle.

3.Resultados

3.1 Análisis exploratorios

Luego de limpiar nuestra base de datos, encontramos la composición por sexo (ver Figura 2) de 2005 y 2025 en la región del Gran Buenos Aires.

En primer lugar, podemos observar que en 2005 encontramos una mayor cantidad de mujeres que de varones, a razón de 54% contra 46% respectivamente (aproximadamente). En el caso de 2025, persiste la mayoría femenina, sin embargo se puede observar que la brecha se redujo en alguna pequeña medida (algo así como 52% contra 48%). Consecuentemente vemos que la composición por sexo en el Gran Buenos Aires es relativamente estable, con un predominio femenino.

A continuación, generamos esta matriz, de 7 variables. En el caso del 2005 (ver Figura 3), podemos ver que sexo sigue sin correlacionar de manera significativa con variaciones en educación, ingreso o cobertura. La edad se correlaciona negativamente con “estado civil”, lo cual es esperable. Existe una correlación positiva entre “edad” y “cobertura médica”: las personas con más edad tienden a tener algo más de cobertura formal y mejores ingresos. Además, hay una correlación negativa entre edad y nivel de instrucción. En cuanto a la “cobertura médica”, correlaciona más con ingreso y con edad que con nivel educativo. Nivel educativo se correlaciona positivamente con “ingreso per cápita familiar”.

En la matriz de 7 variables del año 2025 (Figura 4) podemos observar correlaciones mucho mas fuertes que en las de 2005, entre casi todas las variables. En primer lugar, se mantiene la correlación negativa entre edad y estado civil. Sin embargo, la correlación entre edad e ingreso o cobertura desaparecen o se vuelven negativas, a diferencia de la matriz de 2005.

Por otro lado, sexo en la matriz anterior prácticamente no tenia correlaciones; sin embargo, en la de 2025 (ver Figura 4) aparece fuertemente correlacionada con estado civil, cobertura medica, nivel educativo y condición de actividad. Esto sugiere que el sexo se volvió un valor esctructuralmente significativo de las desigualdades sociales. En el caso de nivel educativo, aumenta su relación con casi todas las variables, con excepción al ingreso. Correlaciona fuertemente con cobertura, actividad y sexo. La variable “cobertura medica” pierde correlación con ingreso pero gana con sexo, estado civil, educación y empleo. La “condición de actividad” se volvió altamente dependiente de variables sociales como sexo, estado civil y

cobertura. El ingreso per cápita, que en la matriz anterior contaba con correlaciones mas consistentes, deja de ser un eje articulador.

Sin embargo, una de las razones por las cuales la variable “Ingreso” podría haber perdido centralidad se debe a que menos personas reportan sus ingresos o cuentan cuál es su condición de actividad. Pudimos observar que, en 2005, de las 9484 personas encuestadas, solo 108 se negaron a responder a la pregunta “condición de actividad”. Sin embargo, en el caso de 2025, de 7181 encuestados, 2872 se negaron a responder cuando se les preguntó por su condición de actividad. Creemos que esto puede haber incidido en la matriz de correlación de 2025. Las diferencias entre años también puede deberse a errores en esta celda de código.

3.2 Reconociendo a los pobres y no pobres

Una vez realizados los análisis exploratorios, notamos que uno de los principales problemas de la EPH es la creciente cantidad de hogares que **NO** reportan sus ingresos. Para saber cuántas personas respondieron y no respondieron, establecimos que los hogares que tenían un ITF>0 fueron los que respondieron, mientras que los que tenían un ITF = 0. Eso se guardó en una base de datos nueva que guardamos como .csv para poder manipular en los ejercicios posteriores (ver Figura 5). Al analizar el ITF, observamos que en 2005 casi todos respondieron (98.8%) mientras que en 2025 la no respuesta subió al 40% reduciendo la tasa de respuesta al 60%. Algunas de las razones posibles que pensamos para que se dé este fenómenos son la sensibilidad de la pregunta sobre ingresos, es decir, muchas personas se incomodan al declarara cuánto gana su hogar, mayor informalidad y precarización laboral (en contextos económicos más inestables, los hogares no siempre tienen ingresos fijos”, aumento de la fatiga de encuesta entre otras. El filtro que hicimos en este punto, permite poder continuar con otros análisis posteriores. Además, a este csv también sumamos las columnas de “codosu” y “nro_hogar” para poder seguir con los pasos siguientes.

Al abordar la consigna 6, trabajamos, por un lado con el archivo csv de respondieron y con la tabla de equivalencias de adultos según sexo y edad. Dado que el excel original contenía formados de texto que dificulta la lectura en Python y generaba valores faltantes (Nan), resolvimos el problema reconstruyendo un nuevo archivo de excel con los mismos datos originales, pero en un formato limpio y numérico llamado “ADULTO_EQUIVALENTE_xlsx”. Este cambio no modificó la información utilizada, sino que permitió procesarla correctamente y evitar errores. A partir de la tabla Adulto Equivalente mapeamos la variable sexo a números EPH y normalizamos la edad. Luego, prepáranos una base que pudiera cruzar los valores numéricos de edad y sexo e hicimos el merge. Usamos groupby para tener el número de adultos equivalentes dentro de cada hogar; y transform para calcular el total del hogar. Entonces, la fila “adulto_equiv” nos mostraba valores individuales, mientras que “ad_equiv_hogar3” nos mostraba la suma en el hogar.

Por otro lado, en la consigna 7 se incorporó la variable “ingreso_necesario”, definida como el producto de la Canasta Básica Total por adulto equivalente del año de referencia ($205,07 en 2005 y $365.177 en 2025) y la cantidad de adultos equivalentes del hogar (ad_equiv_hogar3).. Por lo tanto, el objetivo del ejercicio era sumar a la base “respondieron” una nueva columna llamada “ingreso_necesario” que refleja este cálculo y permitiera luego clasificar a los hogares en pobres y no pobres al comparar contra su ingreso total familiar. Para resolver este ejercicio, partimos de la base enriquecida construida en el ejercicio anterior, en la cual ya contábamos con la variable ad_equiv_hogar3. Esta variable representa la suma de los adultos equivalentes de cada hogar, obtenida a partir del sexo y la edad de cada persona y los coeficientes de equivalencia correspondientes. Una vez cargada esta base, nos aseguramos de que la variable de año estuviera correctamente tipificada como numérica, de modo de poder asignar sin problemas el valor de la canasta básica correspondiente a 2005 o 2025. Después definimos en el código un diccionario con el valor de la CBT por adulto equivalente en cada año y aplicamos una función fila por fila que multiplicó este valor por la cantidad de adultos equivalentes del hogar. De esta manera se generó la nueva columna ingreso_necesario, que asigna a cada hogar el umbral mínimo de ingresos que necesita para no ser considerado pobre. Además, corregimos los casos donde ad_equiv_hogar3 aparecía como cero, reemplazándolos por valores faltantes, ya que no representan hogares válidos. Finalmente, la base fue exportada como respondieron_p7.csv y se realizaron chequeos con estadísticas descriptivas y vistas previas para confirmar la coherencia de los cálculos.

En 2005 los 9.371 hogares analizados necesitaban en promedio $656 para no ser pobres, con un rango que iba de $129 a $1.921. En 2025, los 4.309 hogares requerían más de $942.000 en promedio, con un mínimo de $230.061 y un máximo de $3.399.798 (ver Tabla 1). Estos resultados reflejan el fuerte aumento de la Canasta Básica Total y la incidencia de la composición de los hogares. La variable “ingreso_necesario” marca el umbral mínimo de ingresos de cada hogar y resulta clave para clasificar la pobreza, mostrando un crecimiento muy significativo de dicho umbral entre 2005 y 2025 (ver Tabla 1).

Respondiendo al punto 8, se incorporó a la base respondieron una nueva columna denominada pobre, que toma el valor 1 cuando el ingreso total familiar (ITF) resulta menor al ingreso necesario estimado para el hogar, y 0 en caso contrario. Esta variable permitió clasificar a los hogares según si alcanzaban o no el umbral mínimo de ingresos establecido a partir de la Canasta Básica Total por adulto equivalente.
Para resolverlo, cargamos la base *respondieron_p7.csv* que ya contenía el cálculo del ingreso necesario por hogar. A continuación nos aseguramos de que tanto el ITF como la variable *ingreso_necesario* estuvieran en formato numérico. Luego generamos la variable pobre con un criterio lógico simple: se asigna 1 si el ITF del hogar es inferior a su ingreso necesario y 0 si lo supera. Con esta nueva columna pudimos agrupar por año y elaborar un resumen con el número de pobres identificados (n_pobres), el total de hogares analizados (n_total) y el porcentaje que representan dentro de la muestra (pct_pobres).

Los resultados muestran que en 2005 se identificaron 2.438 hogares pobres sobre un total de 9.371, lo que equivale al 26,02% de la muestra (ver Tabla 2). En 2025 se clasificaron como pobres 1.334 hogares de un total de 4.309, representando el 30,96% de la muestra. Este aumento en la proporción de hogares pobres refleja que, aunque cambian los valores absolutos de los ingresos y las canastas, la presión sobre los hogares para alcanzar el umbral mínimo de subsistencia se incrementa entre ambos períodos (ver Tabla 2).

Por último, en la consigna 9 trabajamos con la base *respondieron_p8.csv*, que ya contenía la variable pobre. El primer paso del código fue calcular estadísticas descriptivas de pobre por año, mostrando medidas como el número de observaciones, la media, la desviación estándar, los cuartiles y los valores mínimos y máximos. Dado que pobre es una variable dicotómica, su media puede interpretarse directamente como la proporción de pobres en cada año. El resumen confirmó lo que ya habíamos visto en la consigna anterior: en 2005 alrededor del 26% de los hogares fueron clasificados como pobres, mientras que en 2025 la proporción ascendió al 31% (ver Tabla 3). Esto marcó un aumento en la incidencia de la pobreza en la muestra a lo largo del tiempo.

Uno de los gráficos elaborados fue de barras comparando los porcentajes de hogares pobres entre 2005 y 2025 (ver Figura 6). En él se observa con claridad cómo la proporción de pobreza crece entre ambos períodos, pasando de aproximadamente una cuarta parte de los hogares en 2005 a casi un tercio en 2025.

El segundo gráfico fue un diagrama de dispersión que muestra la relación entre el ingreso total familiar declarado (ITF) y el ingreso necesario calculado para cada hogar (ver Figura 7). Los puntos aparecen coloreados en verde si el hogar no es pobre y en rojo si es pobre. La línea de corte diagonal representa la igualdad entre ITF e ingreso necesario: los hogares ubicados por debajo de la línea (en rojo) son aquellos que su ingreso declarado no alcanza para cubrir sus necesidades básicas. Este gráfico permite ver que la mayoría de los hogares pobres se concentran en la parte baja de la distribución, mientras que los hogares no pobres se ubican claramente por encima de la línea de corte, con ingresos que superan su umbral de subsistencia.

Finalmente, se construyó un tercer gráfico exploratorio adicional que analizó la pobreza desagregada por sexo y año (ver Figura 8). Para ello se estandarizó la variable ch04 a etiquetas de texto (“Varón” y “Mujer”) y se calculó la proporción de hogares pobres en cada grupo. El gráfico de barras comparativo muestra que tanto en 2005 como en 2025 las tasas de pobreza son similares entre varones y mujeres, aunque en 2025 las mujeres presentan un porcentaje ligeramente mayor (31,8% frente a 30,1% en varones). Este resultado señala que el incremento de la pobreza a lo largo del tiempo afecta a ambos sexos, aunque se observa un impacto un poco más elevado en los hogares con jefa de hogar femenina.

Anexo

Figura 1

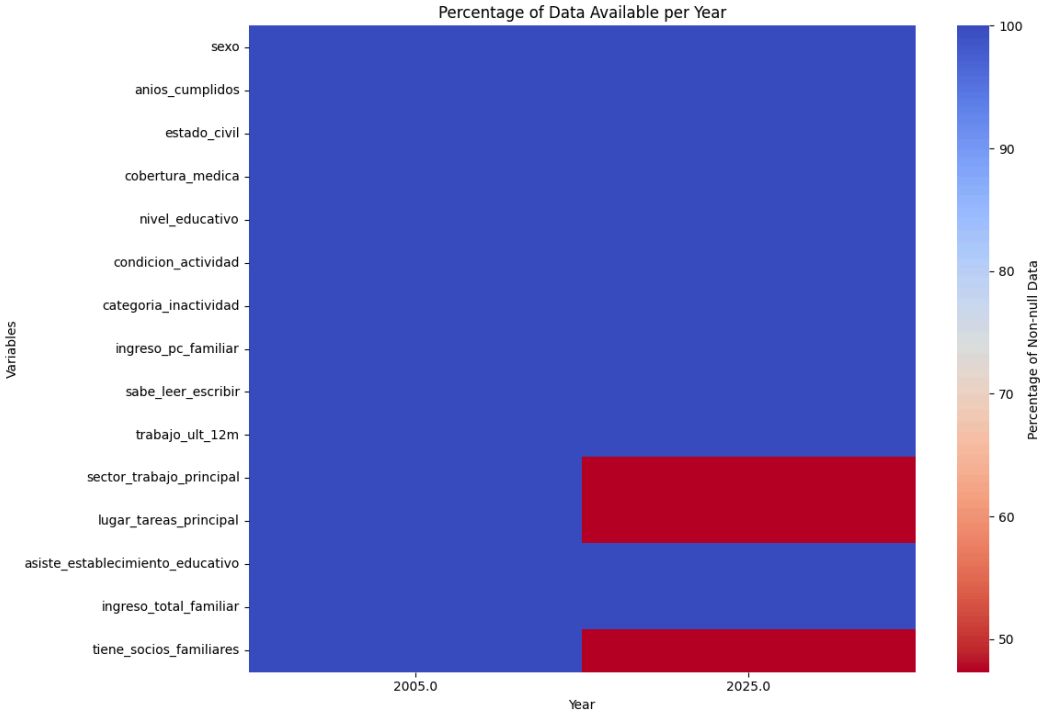

Figura 2

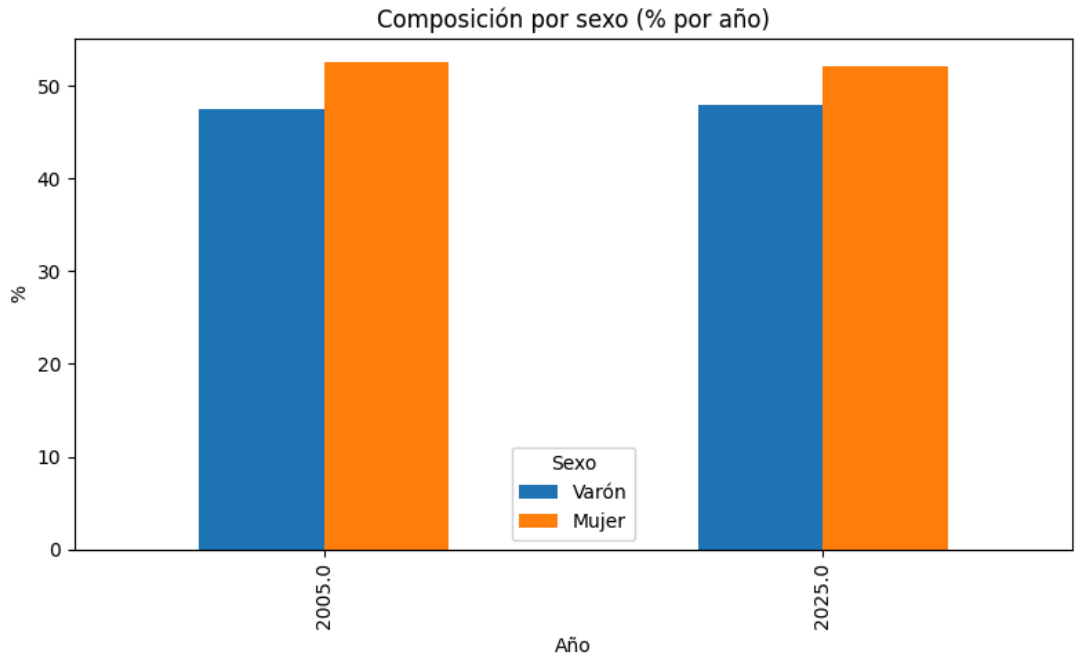

Figura 3

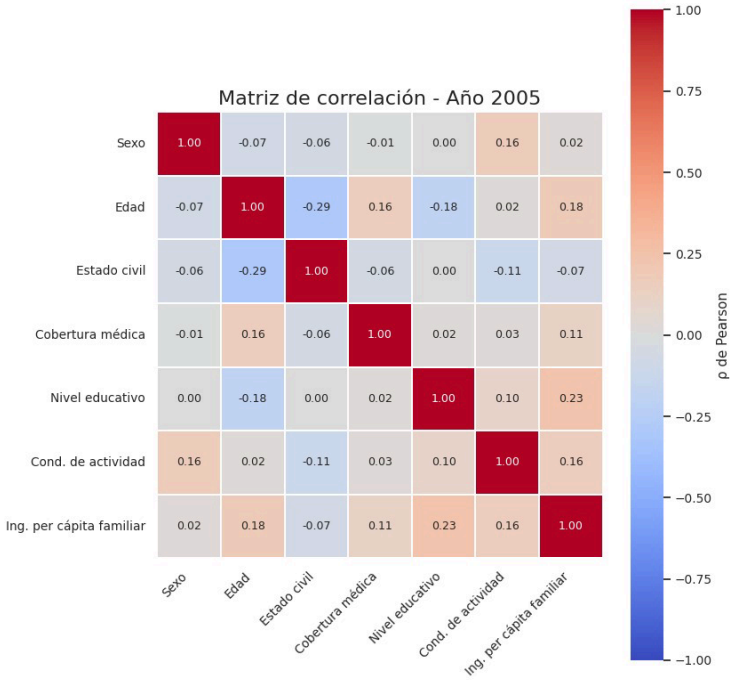

Figura 4

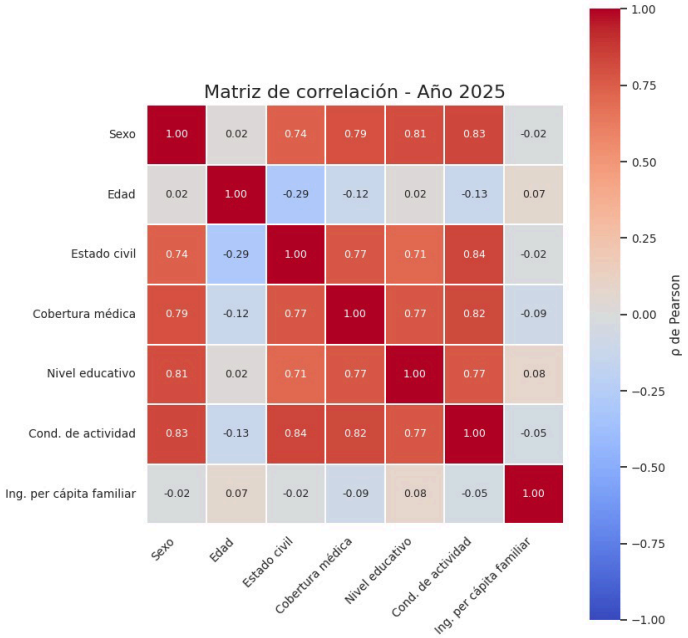

Figura 5

```
=== Condición de actividad ===
Personas que NO respondieron condición de actividad: 9484

=== Resumen por año — ITF (criterio: >0 / =0) ===
   ano4  respondieron  no_respondieron  otros  total  % respondieron  % no_respondieron
 2005.0          9371              113      0   9484           98.81               1.19
 2025.0          4309             2872      0   7181           60.01              39.99
  TOTAL         13680             2985      0  16665           82.09              17.91

=== Contingencia año × respuesta — ITF ===
ano4               2005.0  2025.0
respuesta_ITF
No respondió (=0)     113    2872
Respondió (>0)       9371    4309
```

Tabla 1:

```
Distribución ingreso_necesario por año:
        count          mean            std          min          25%  \
ano4
2005.0  9371.0  6.913332e+02     333.553564     129.1941     465.5089
2025.0  4309.0  1.016889e+06  503884.900600  230061.5100  642711.5200

               50%           75%           max
ano4
2005.0     656.224  8.797503e+02  1.921506e+03
2025.0  942156.660  1.300030e+06  3.399798e+06

=== Vista previa de la tabla corregida ===
      ano4  ad_equiv_hogar3  ingreso_necesario
0   2005.0             0.74           151.7518
1   2005.0             1.78           365.0246
2   2005.0             1.78           365.0246
3   2005.0             1.78           365.0246
4   2005.0             0.67           137.3969
5   2005.0             4.30           881.8010
6   2005.0             4.30           881.8010
7   2005.0             4.30           881.8010
8   2005.0             4.30           881.8010
9   2005.0             4.30           881.8010
10  2005.0             2.24           459.3568
11  2005.0             2.24           459.3568
12  2005.0             2.24           459.3568
13  2005.0             0.63           129.1941
14  2005.0             3.28           672.6296
15  2005.0             3.28           672.6296
16  2005.0             3.28           672.6296
17  2005.0             3.28           672.6296
18  2005.0             4.44           910.5108
19  2005.0             4.44           910.5108
```

Tabla 2:

```
Resumen de pobreza por año:
        n_pobres  n_total  pct_pobres
ano4
2005.0      2438     9371       26.02
2025.0      1334     4309       30.96
```

Tabla 3:

```
=== Estadísticas descriptivas de pobreza por año ===
        count  mean    std  min  25%  50%  75%  max
ano4
2005.0  9371.0  0.26  0.439  0.0  0.0  0.0  1.0  1.0
2025.0  4309.0  0.31  0.462  0.0  0.0  0.0  1.0  1.0
/tmp/ipython-input-2710264195.py:19: FutureWarning:
```

Figura 6:

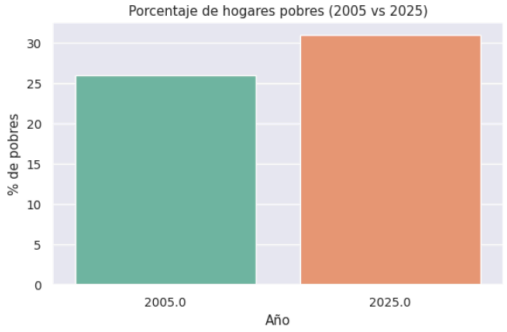

Figura 7:

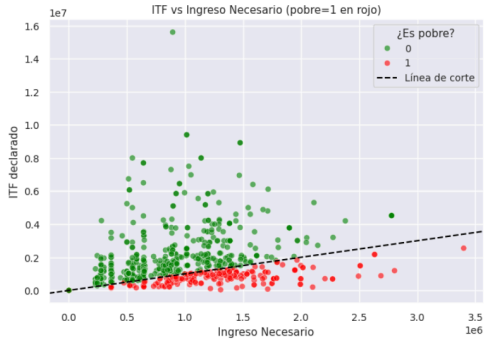

Figura 8:

```
   ano4   sexo  % pobres
0  2005.0  Mujer      25.8
1  2005.0  Varón      26.2
2  2025.0  Mujer      31.8
3  2025.0  Varón      30.1
```