FCE-UBA
UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS ECONÓMICAS

## BIG DATA Y MACHINE LEARNING

# TRABAJO PRÁCTICO N° 2

*"Un Primer Encuentro con la EPH"*

(Encuesta Permanente de Hogares)

---

**GRUPO 3**

**Rafael Pablo Pinto Chambi** - Reg. 908586

**Javier Rodolfo Aguirre** - Reg. 819698

**Lautaro Manuel Bogado** – Reg. 899596

---

**Enlace de repositorio:** https://github.com/RafaelPCh/BigDataUBA-Grupo3.git

**Docente a cargo:** María Noelia Romero

Segundo Cuatrimestre de 2025

## Parte 1:Familiarizandonos con la base EPH y limpieza.

### Punto 1:

En Argentina, el Instituto Nacional de Estadística y Censos (INDEC) identifica a las personas pobres mediante un método indirecto o monetario basado en la Encuesta Permanente de Hogares (EPH). Este procedimiento consiste en comparar los ingresos totales de cada hogar con el valor monetario de una canasta de bienes y servicios básicos. Para ello se construyen dos umbrales: la Canasta Básica Alimentaria (CBA), que incluye los alimentos necesarios para cubrir requerimientos calóricos y proteicos mínimos, y la Canasta Básica Total (CBT), que amplía la CBA para incorporar otros bienes y servicios esenciales como vestimenta, transporte, salud y educación.

Cada canasta se valora mensualmente y se ajusta por la composición del hogar (edad y sexo de sus integrantes) para reflejar las necesidades específicas de cada familia. Si el ingreso total del hogar no alcanza para adquirir la CBT correspondiente, el hogar y todas las personas que lo integran se clasifican como pobres. Si el ingreso es insuficiente incluso para cubrir la CBA, se considera al hogar indigente. Este enfoque permite obtener, con periodicidad trimestral, tasas de pobreza e indigencia comparables entre regiones y a lo largo del tiempo.

Además, en los censos el INDEC utiliza un método complementario denominado Necesidades Básicas Insatisfechas (NBI), que identifica carencias directas en aspectos como vivienda, saneamiento, hacinamiento, escolaridad o capacidad de subsistencia. Así, se dispone de dos formas de medición: la indirecta, centrada en ingresos, y la directa, basada en carencias efectivas.

## Punto 2.c: (gráficos 1.1 y 1.2)

Los gráficos presentan los valores faltantes en las bases de 2005 (t0105) y 2025 (t0125). Se observa que la gran mayoría de las variables no registran datos faltantes en ninguno de los dos años; en particular, CODUSU, NRO_HOGAR, AGLOMERADO, PONDERA, CH03, CH04, CH06, CH07, CH08, NIVEL_ED, ESTADO, IPCF e ITF están completas en ambas bases.

En contraste, sólo cuatro variables concentran los valores perdidos en ambos períodos:CH14 , CAT_OCUP,  CAT_INAC y PP04B_COD

Comparando ambos años, se aprecia que la magnitud de los faltantes varía. En 2005, las variables con más datos perdidos fueron PP04B_COD (5.767 casos) y CAT_OCUP (5.420 casos), seguidas por CH14 (4.018 casos) y CAT_INAC (3.875 casos). En 2025, se mantiene el mismo patrón, aunque con cifras algo distintas: PP04B_COD (5.337 casos) y CAT_OCUP (5.167 casos) continúan encabezando, mientras que CH14 (4.784 casos) y CAT_INAC (4.592 casos) presentan un ligero aumento en comparación con 2005.

En síntesis, los resultados muestran que la problemática de los datos faltantes se concentra exclusivamente en las variables vinculadas con la edad y la condición laboral de la persona. Además, entre 2005 y 2025 se mantiene la consistencia del patrón, aunque con variaciones en el número absoluto de registros nulos: aumentan los faltantes en CH14 y CAT_INAC, mientras que disminuyen levemente en CAT_OCUP y PP04B_COD.

Este comportamiento refleja que, a pesar de la estabilidad en la mayoría de las variables, las relacionadas con la caracterización laboral y etaria siguen siendo las más problemáticas para el análisis estadístico, lo que obliga a considerar estrategias de imputación o tratamiento de valores faltantes en futuras etapas del procesamiento.

## Punto 2.d:

Para poder trabajar de manera conjunta con las bases de 2005 y 2025, se llevó adelante un proceso de limpieza y homogeneización de variables. En primer lugar, se seleccionaron las 15 variables de interés, entre las cuales se incluyeron aquellas exigidas en el enunciado (CH04, CH06, CH07, CH08, NIVEL_ED, ESTADO, CAT_INAC e IPCF), más un conjunto adicional que resulta relevante para el análisis.

Dado que la codificación de respuestas difiere entre los años, se implementó una función (mapear_columna) que permitió transformar categorías textuales en códigos

numéricos homogéneos, de acuerdo con los diccionarios oficiales de la EPH. De esta manera se recodificaron variables como aglomerado, parentesco (CH03), sexo (CH04), estado civil (CH07), cobertura médica (CH08), nivel educativo (NIVEL_ED), condición de actividad (ESTADO), categoría ocupacional (CAT_OCUP) y condición de inactividad (CAT_INAC).

Se realizaron también ajustes para depurar valores fuera de rango. En particular, en la variable edad (CH06) se eliminaron registros con respuestas no válidas como “Menos de 1 año” o “98 y más años”, así como valores atípicos presentes en 2025 (“-1”, “103”).

Asimismo, se corrigieron valores que figuraban como 0 pero no estaban tipificados en los cuadros de referencia (por ejemplo, en CAT_INAC, CAT_OCUP y PP04B_COD), reemplazándolos por NaN para que fueran tratados como datos faltantes.

Finalmente, se unificaron los nombres de columnas en ambas bases para garantizar compatibilidad y permitir concatenar los datasets en un solo marco de análisis.

## Parte 2: Análisis exploratorio.

### Punto 3:

Los gráficos permiten visualizar la evolución de la estructura por sexo y edad de la población del NOA entre 2005 y 2025.

En primer lugar, en el grafico 2, la composición por sexo se mantiene relativamente estable en ambos años. En 2005 los hombres representaban el 48,2 % de la población y las mujeres el 51,8 %, mientras que en 2025 las cifras son 48 % y 52 %, respectivamente. Esto indica que la región continúa presentando una ligera mayoría femenina, fenómeno habitual en la mayoría de los contextos demográficos por la mayor esperanza de vida de las mujeres. El hecho de que la proporción por sexo no varíe de forma apreciable sugiere que, pese a otros cambios estructurales, no se han producido transformaciones profundas en la relación hombres-mujeres.

En segundo lugar,gráfico 3, el desglose por rangos de edad revela cambios mucho más marcados. Se observa una disminución clara del peso relativo de la población menor de 18 años: en 2005 representaban más de un tercio del total, mientras que en 2025 apenas superan una quinta parte. Paralelamente, crecen los grupos adultos y de mayores, lo que refleja un proceso de envejecimiento poblacional propio de las etapas avanzadas de la transición demográfica. Este envejecimiento también explica por qué en los rangos de edad más avanzados predominan las mujeres, dado que tienen mayores tasas de supervivencia.

En conjunto, los gráficos muestran un NOA con estructura por sexo estable pero con un cambio significativo en la estructura etaria hacia una población más adulta y envejecida. Este proceso puede tener efectos relevantes en la planificación de políticas públicas y sociales, especialmente en materia de salud, pensiones, educación y mercado laboral, ya que implica una disminución relativa de la población infantil y juvenil y un aumento de las necesidades de servicios orientados a la población adulta y mayor.

### Punto 4 : (Gráficos 4 y 5)

En primer lugar, en ambos períodos se observa una correlación positiva entre edad y estar casado, y una correlación negativa entre edad y ser soltero. Esto refleja patrones esperables: a mayor edad aumenta la probabilidad de estar casado y disminuye la de estar soltero. También se aprecia que la correlación negativa entre edad y condición de estudiante se mantiene fuerte en los dos años, aunque algo menos intensa en 2025, mostrando que la escolarización sigue concentrándose en edades jóvenes.

En cuanto a las condiciones económicas y de seguridad social, la correlación entre edad y cobertura de obra social es levemente positiva en ambos años, indicando que las personas de mayor edad tienden a estar más cubiertas. Destaca que la variable “No paga ni descuenta” mantiene una correlación negativa fuerte con la obra social y con la condición de ocupado, lo que es consistente con la informalidad laboral: quienes no aportan tampoco cuentan con cobertura y suelen tener empleos más precarios. En 2025 esta relación negativa se intensifica, lo que podría sugerir un aumento de la informalidad o una mayor concentración de la falta de aportes en ciertos grupos.

Respecto de la actividad económica, en ambos años se observa que estar ocupado se relaciona negativamente con ser estudiante (trade-off natural entre trabajo y estudio). En 2025 la correlación negativa entre “Inactivo” y “Ocupado” se vuelve más pronunciada (-0,76 frente a -0,62 en 2005), reflejando una segmentación más clara entre quienes trabajan y quienes no.

En relación con la variable Mujer, las correlaciones con otras variables son en general bajas, lo que indica que el sexo por sí solo no determina fuertemente las otras condiciones, aunque se observan patrones interesantes. En ambos años existe una correlación positiva con la categoría “Ama de casa” y con “Casado”, lo que refleja que las mujeres siguen representando la mayor parte de quienes declaran tareas domésticas y que el matrimonio se asocia más a mujeres en términos de rol social. En 2025 estos vínculos se mantienen aunque con leves variaciones, lo que puede interpretarse como una persistencia de la división sexual del trabajo, a pesar de ciertos avances en la participación laboral femenina.

Por último, la variable IPCF (ingreso per cápita familiar) muestra correlaciones moderadas y consistentes con otras variables. En ambos años presenta una correlación positiva con la edad y con la cobertura de obra social, lo que indica que a medida que aumenta la edad y se formaliza la situación laboral del hogar, mejora el ingreso per cápita familiar y también la cobertura. También se observan correlaciones negativas con “No paga ni descuenta” y con “Primaria incompleta”, coherentes con la idea de que hogares más educados y con aportes tienden a tener mayores ingresos per cápita. Entre 2005 y 2025 estas relaciones no desaparecen pero sí se suavizan levemente, sugiriendo cambios en la composición de los hogares y en la distribución del ingreso en la región.

## Parte III: Conociendo a los pobres y no pobres

### Punto 5:

Para la base de datos del 2005 la cantidad de gente que no respondieron acerca de su ingreso total familiar son 8. En cambio para el 2025 ascendieron a 25.

### Punto 8:

El análisis de la muestra revela un aumento tanto en el número absoluto como en el porcentaje de personas en situación de pobreza entre 2005 y 2025. En 2005, se identificaron 4,098 personas pobres, lo que representaba el 44.63% de la muestra total. Para el año 2025, esta cifra aumentó a 4,909 personas pobres, lo que significa que el 50.66% de la muestra se encontraba en hogares pobres.

A nivel de hogares, la muestra también muestra un incremento en la pobreza. En 2005, se identificaron 843 hogares pobres, lo que representaba el 36.13% de la muestra. Para 2025, esta cifra ascendió a 1,288 hogares pobres, lo que constituye el 42.22% de la muestra total.

### Punto 9:

Los gráficos presentados permiten analizar la evolución de la pobreza en Argentina en dos momentos del tiempo, 2005 y 2025, a partir de dos dimensiones: la incidencia a nivel de hogares y la distribución por grupos etarios.

En el gráfico 6 se observa que la pobreza medida por hogares pasó del 36,1% en 2005 al 42,2% en 2025. Este aumento de más de seis puntos porcentuales refleja un deterioro en las condiciones de vida y muestra que, lejos de resolverse, el problema de la pobreza se ha intensificado. La magnitud de esta cifra implica que, en la actualidad, prácticamente 4 de cada 10 hogares se encuentran en situación de pobreza, lo cual plantea un desafío estructural para las políticas públicas.

El gráfico 7 desagrega la información según grupos etarios, destacando la marcada vulnerabilidad de niños y adolescentes. En 2005, el 57,6% de la población de 0 a 14 años residía en hogares pobres, mientras que en 2025 esta cifra asciende a 62,7%. Esto indica que la niñez continúa siendo el sector más afectado, con consecuencias a largo plazo en términos de nutrición, educación y oportunidades de desarrollo. Por su parte, los adultos mayores (65 años o más) también experimentan un empeoramiento: la pobreza pasó de 21,4% a 26,7%, lo que evidencia limitaciones en la capacidad de los sistemas de protección social para garantizar ingresos adecuados.

En conjunto, ambos gráficos subrayan que la pobreza en Argentina sigue siendo un fenómeno persistente y que afecta de manera desproporcionada a los grupos más vulnerables, especialmente a los niños, consolidando un patrón de transmisión intergeneracional que profundiza las desigualdades sociales.

# Apéndice:

Gráfico 1.1:

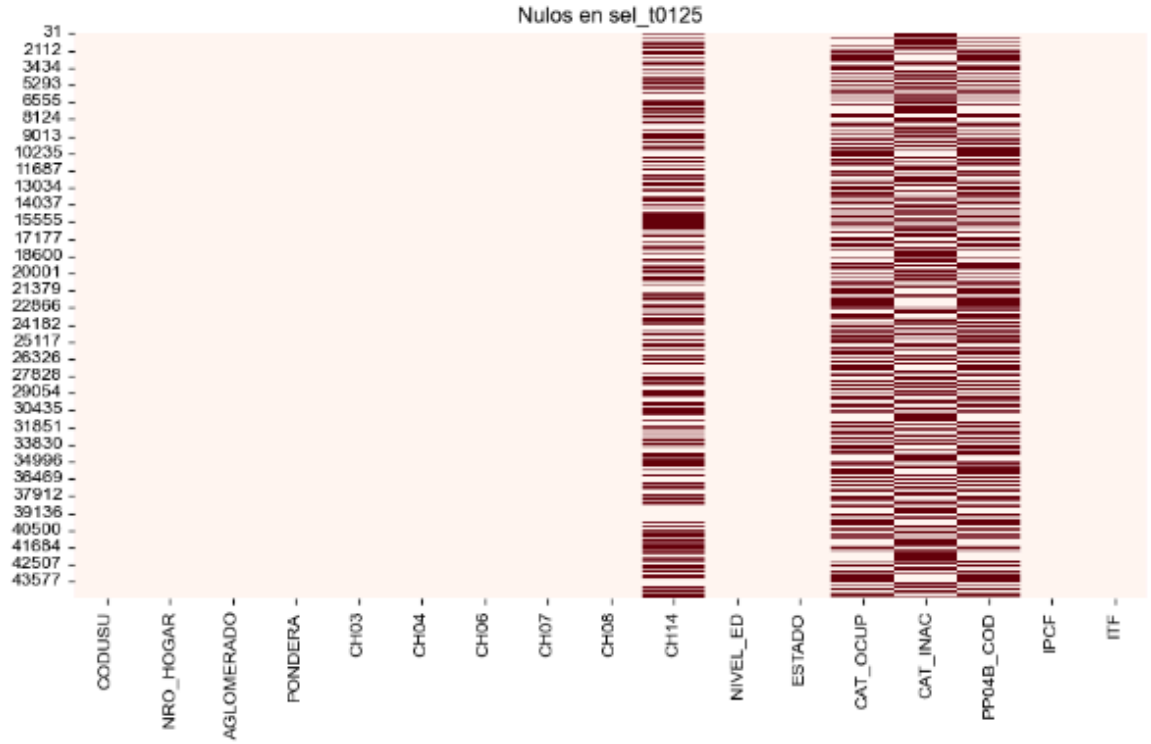


Gráfico 1.2:

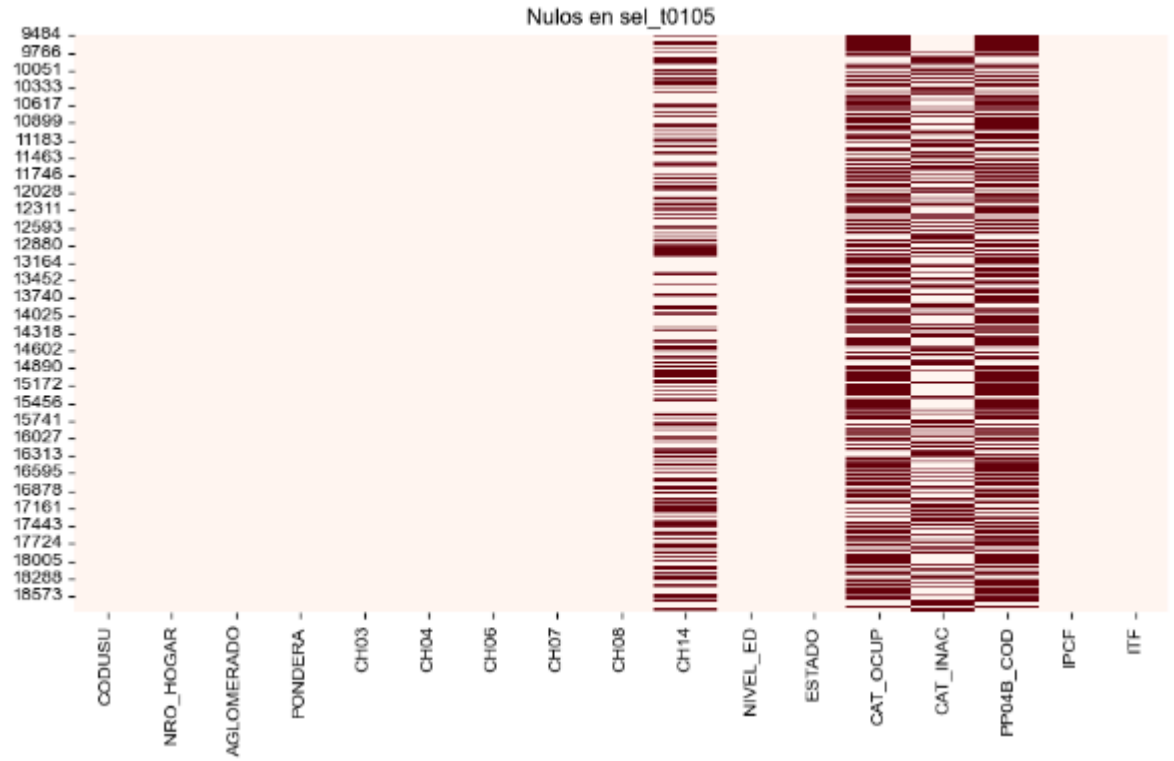

Gráfico 2

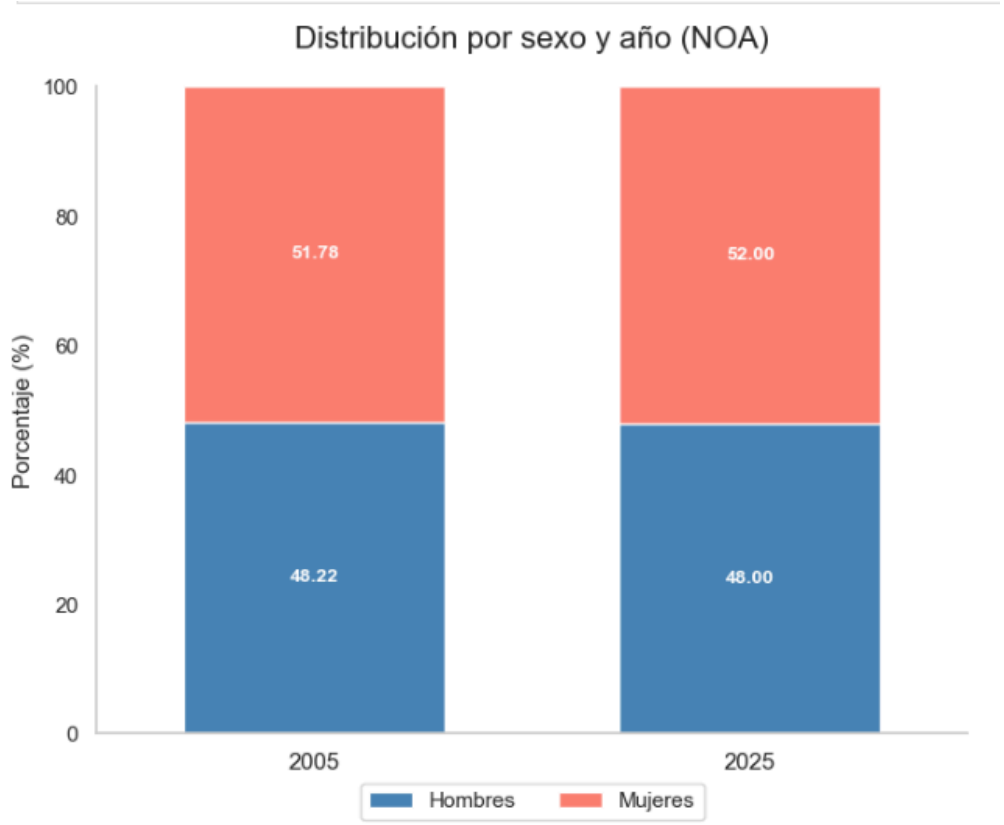


Gráfico 3:

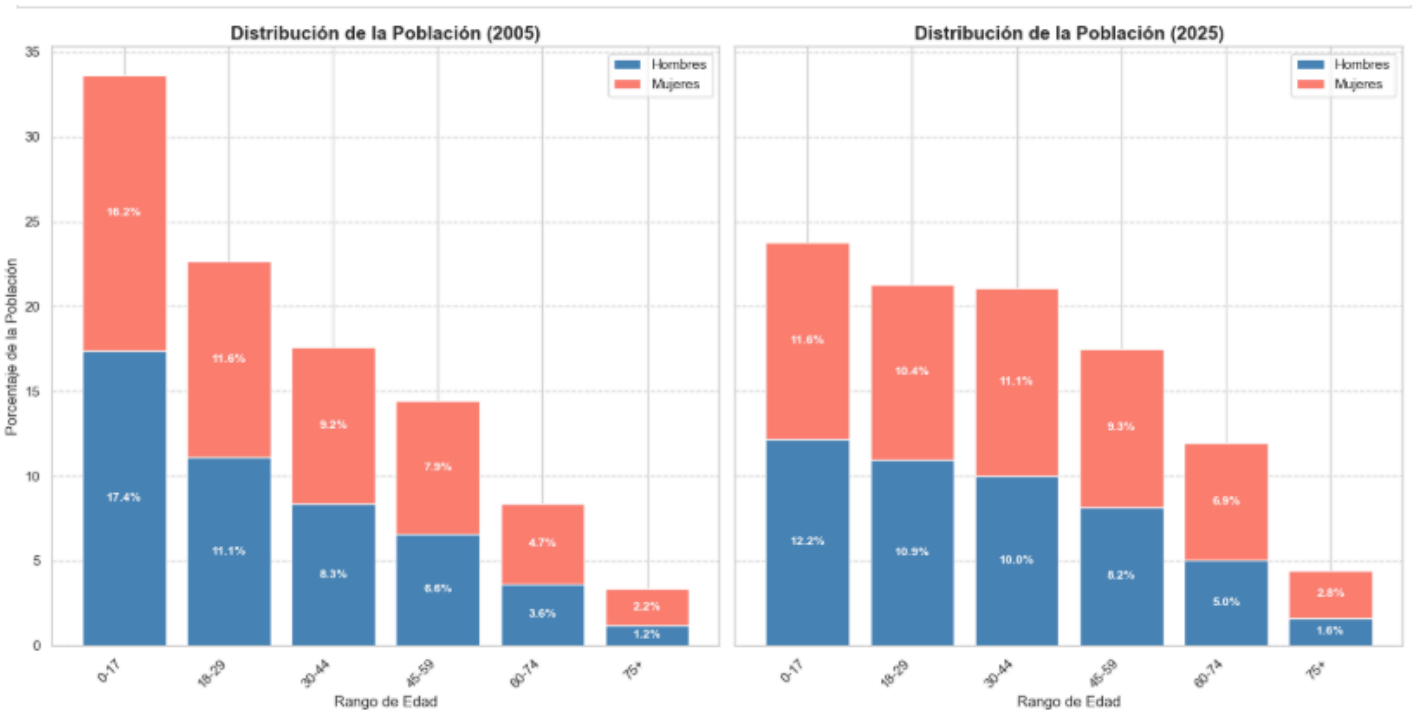

Gráfico 4:

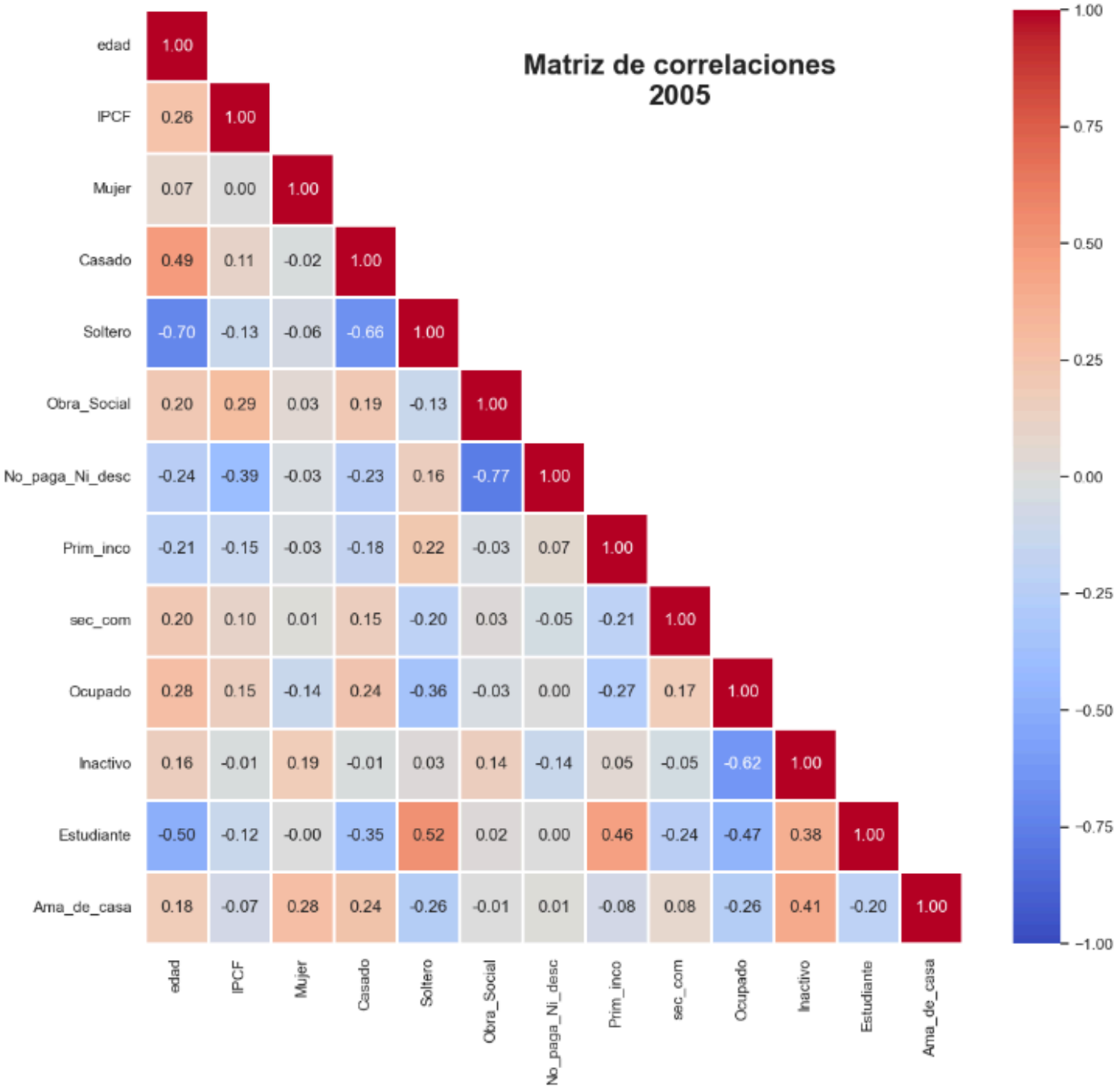
Matriz de correlaciones
2005
edad 1.00
IPCF 0.26 1.00
Mujer 0.07 0.00 1.00
Casado 0.49 0.11 -0.02 1.00
Soltero -0.70 -0.13 -0.06 -0.66 1.00
Obra_Social 0.20 0.29 0.03 0.19 -0.13 1.00
No_paga_Ni_desc -0.24 -0.39 -0.03 -0.23 0.16 -0.77 1.00
Prim_inco -0.21 -0.15 -0.03 -0.18 0.22 -0.03 0.07 1.00
sec_com 0.20 0.10 0.01 0.15 -0.20 0.03 -0.05 -0.21 1.00
Ocupado 0.28 0.15 -0.14 0.24 -0.36 -0.03 0.00 -0.27 0.17 1.00
Inactivo 0.16 -0.01 0.19 -0.01 0.03 0.14 -0.14 0.05 -0.05 -0.62 1.00
Estudiante -0.50 -0.12 -0.00 -0.35 0.52 0.02 0.00 0.46 -0.24 -0.47 0.38 1.00
Ama_de_casa 0.18 -0.07 0.28 0.24 -0.26 -0.01 0.01 -0.08 0.08 -0.26 0.41 -0.20 1.00
edad
IPCF
Mujer
Casado
Soltero
Obra_Social
No_paga_Ni_desc
Prim_inco
sec_com
Ocupado
Inactivo
Estudiante
Ama_de_casa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Gráfico 5:

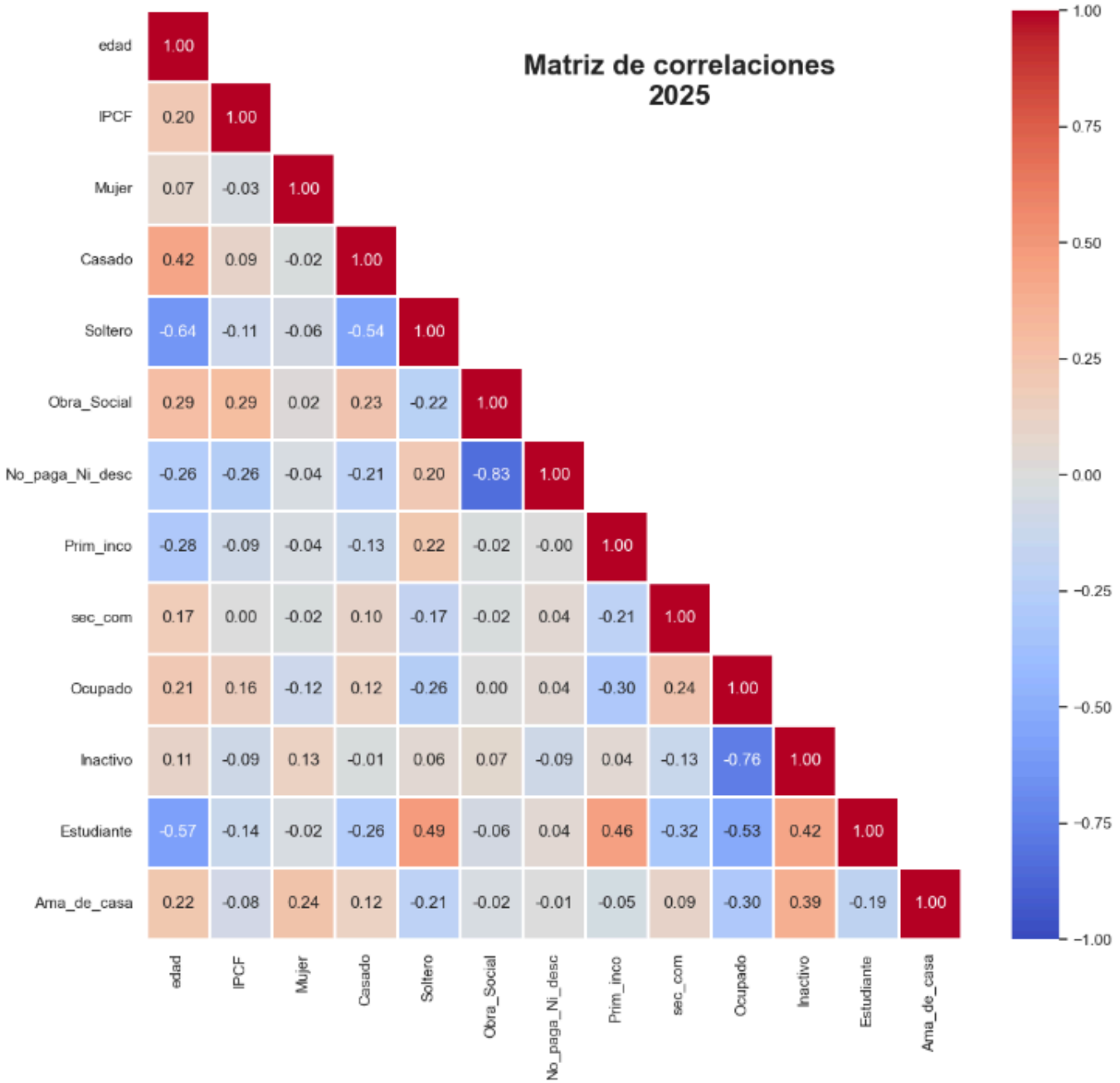
Matriz de correlaciones
2025
edad 1.00
IPCF 0.20 1.00
Mujer 0.07 -0.03 1.00
Casado 0.42 0.09 -0.02 1.00
Soltero -0.64 -0.11 -0.06 -0.54 1.00
Obra_Social 0.29 0.29 0.02 0.23 -0.22 1.00
No_paga_Ni_desc -0.26 -0.26 -0.04 -0.21 0.20 -0.83 1.00
Prim_inco -0.28 -0.09 -0.04 -0.13 0.22 -0.02 -0.00 1.00
sec_com 0.17 0.00 -0.02 0.10 -0.17 -0.02 0.04 -0.21 1.00
Ocupado 0.21 0.16 -0.12 0.12 -0.26 0.00 0.04 -0.30 0.24 1.00
Inactivo 0.11 -0.09 0.13 -0.01 0.06 0.07 -0.09 0.04 -0.13 -0.76 1.00
Estudiante -0.57 -0.14 -0.02 -0.26 0.49 -0.06 0.04 0.46 -0.32 -0.53 0.42 1.00
Ama_de_casa 0.22 -0.08 0.24 0.12 -0.21 -0.02 -0.01 -0.05 0.09 -0.30 0.39 -0.19 1.00
edad
IPCF
Mujer
Casado
Soltero
Obra_Social
No_paga_Ni_desc
Prim_inco
sec_com
Ocupado
Inactivo
Estudiante
Ama_de_casa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Gráfico 6:

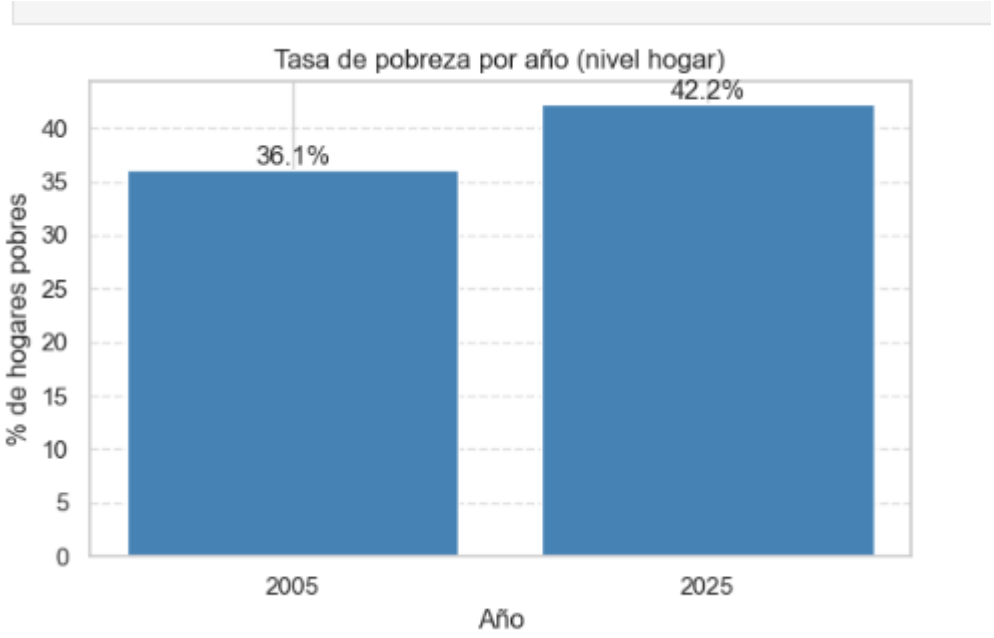


Gráfico 7: