

Universidad de San Andrés

Ciencias del comportamiento



Ciencias de Datos

Docentes:

Maria Noelia Romero

Ignacio Anchorena

Trabajo Práctico 1

Alumnos:

Delfina Borrescio

Catalina Cricco

Candelaria Gilles

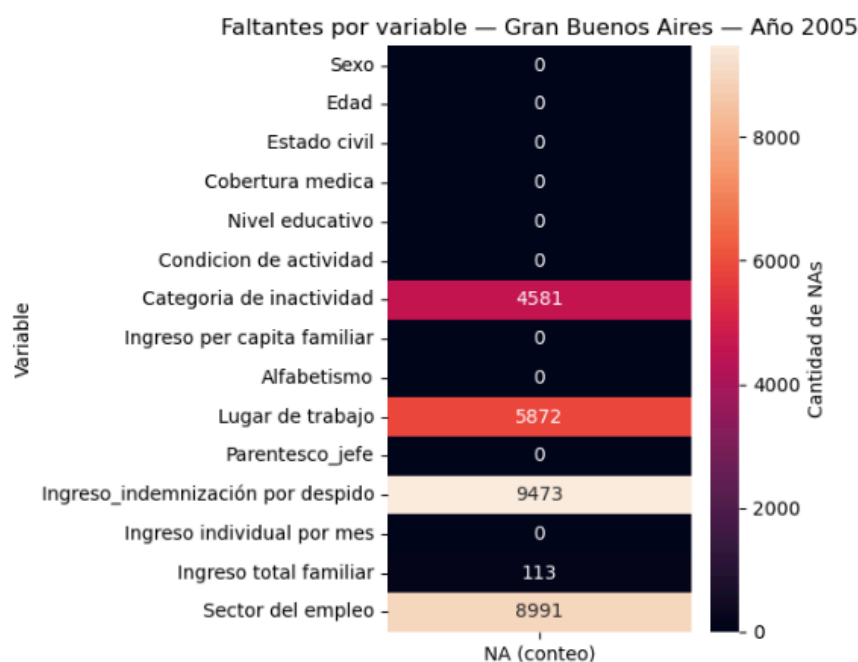
Link al repositorio: <https://github.com/ccricco/CC408-Grupo-T2-3>

Parte I: Familiarizandonos con la base EPH y limpieza

1. Según la página oficial del INDEC (INDEC, 2016), una persona se identifica como pobre si integra un hogar cuyo ingreso total revelado en la EPH (Encuesta Permanente de Hogares del INDEC) es inferior a la CBT (Canasta básica total), que incluye el mínimo alimentario y otros gastos básicos como la vivienda, transporte, salud, etc. Es decir, si el ingreso es menor al valor de la CBT, los individuos de ese hogar se clasifican como pobres.

2. b. La figura 1 presenta un heatmap que muestra la cantidad de datos faltantes (NA) en las variables seleccionadas de la base de datos del año 2005 para la región Gran Buenos Aires. Se observa que la mayoría de las variables, como sexo, edad, estado civil, cobertura médica y nivel educativo, no presentan valores faltantes. En cambio, otras variables registran una cantidad considerable de NA, especialmente sector de empleo (8.991), ingreso por indemnización por despido (9.473), lugar de trabajo (5.872) y categoría de inactividad (4.581). Esto sugiere que las dimensiones vinculadas al mercado laboral y a los ingresos presentan mayores dificultades de registro o respuesta en comparación con las sociodemográficas.

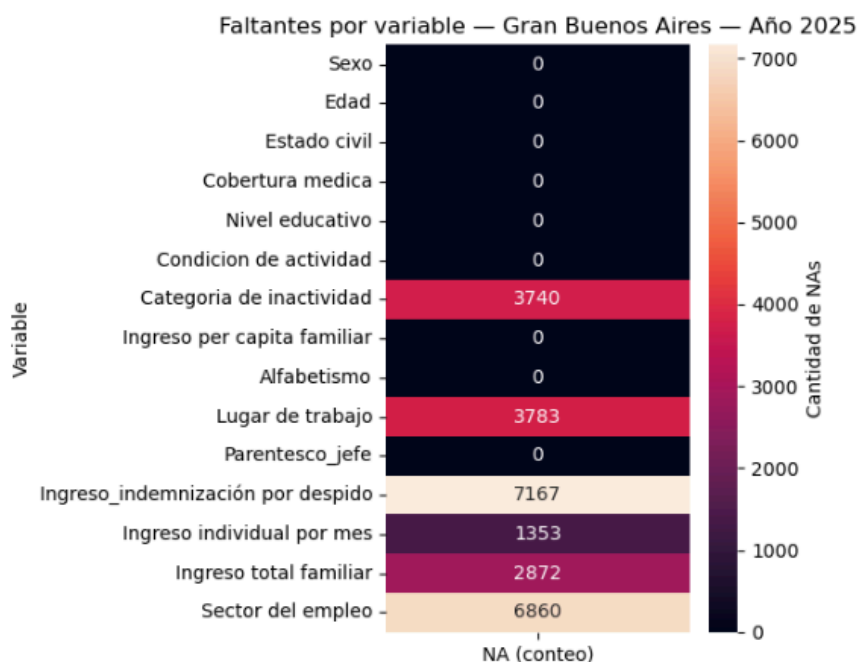
Figura 1.



La Figura 2 presenta un heatmap que muestra la cantidad de datos faltantes (NA) en las variables seleccionadas de la base de datos del año 2025 para la región Gran Buenos Aires. Al igual que en 2005, la mayoría de las variables sociodemográficas, como sexo, edad, estado civil, cobertura médica, nivel educativo y condición de actividad, no presentan valores faltantes. Sin embargo, se observan vacíos de información relevantes en variables asociadas

al mercado laboral y los ingresos. Entre ellas, se destacan el sector de empleo (6.860 NA), ingreso por indemnización por despido (7.167 NA), lugar de trabajo (3.783 NA), categoría de inactividad (3.740 NA) e ingreso total familiar (2.872 NA). También se registran valores faltantes, aunque en menor medida, en ingreso individual por mes (1.353 NA). Estos resultados muestran que, en comparación con las variables sociodemográficas, los mayores problemas de completitud de datos se concentran en las dimensiones económicas y laborales.

Figura 2.



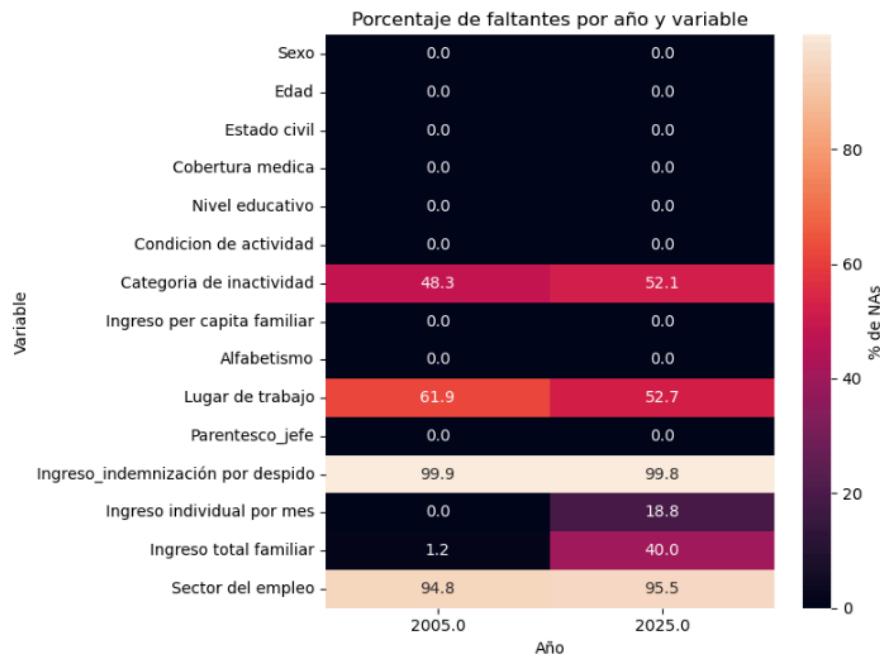
La figura 3 presenta un heatmap que muestra el porcentaje de datos faltantes (NA) por variable y por año en la región Gran Buenos Aires, comparando 2005 y 2025. Se observa que las variables sociodemográficas básicas (sexo, edad, estado civil, cobertura médica, nivel educativo y condición de actividad) no presentan valores faltantes en ninguno de los dos años, lo que indica consistencia en la calidad del registro para estas dimensiones.

En contraste, las variables relacionadas con el mercado laboral y los ingresos exhiben porcentajes elevados de faltantes en ambos períodos. En particular, el ingreso por indemnización por despido alcanza valores cercanos al 100% en 2005 (99,9%) y 2025 (99,8%), mientras que el sector de empleo también presenta una proporción muy alta de faltantes (94,8% y 95,5%, respectivamente). Asimismo, las variables lugar de trabajo (61,9% en 2005 y 52,7% en 2025) y categoría de inactividad (48,3% y 52,1%) muestran niveles intermedios pero igualmente significativos de incompletitud.

Finalmente, se destaca un incremento notable en los faltantes de ingreso total familiar, que pasa de 1,2% en 2005 a 40% en 2025, así como la aparición de un 18,8% de faltantes en la variable ingreso individual por mes en 2025, ausente en 2005. Esto evidencia que, si bien la

cobertura de las variables sociodemográficas se mantiene robusta, las variables económicas y laborales continúan siendo las más problemáticas en términos de calidad y completitud de los datos.

Figura 3.

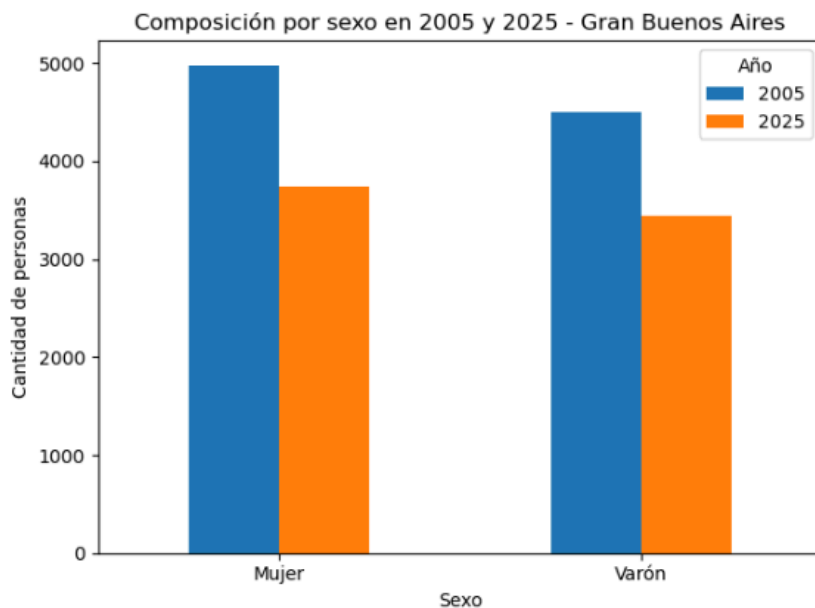


Parte II: Primer Análisis Exploratorio

3. La figura 4 muestra la composición por sexo en la región Gran Buenos Aires en los años 2005 y 2025. En ambos períodos se observa una distribución relativamente equilibrada entre mujeres y varones, aunque con una leve mayoría femenina.

Según los datos, en 2005, la muestra incluía 4.980 mujeres y 4.504 varones, mientras que en 2025 las cifras descienden a 3.742 y 3.439 respectivamente. Esto refleja una reducción en la cantidad total de casos relevados en el tiempo, pero mantiene la proporción similar entre sexos.

Figura 4.



4. La Figura 5 muestra la matriz de correlación para el año 2005 muestra las relaciones lineales entre las principales variables sociodemográficas y económicas de la base de datos del Gran Buenos Aires. Se destaca la correlación negativa entre edad y estado civil (-0,57), así como entre edad y nivel educativo (-0,38) y condición de actividad (-0,37). También se observa una relación positiva entre nivel educativo y categoría de inactividad (0,47). En general, las variables laborales y de ingresos presentan vínculos coherentes pero de menor magnitud.

Figura 5.

Matriz de correlación 2005:

	Sexo	Edad	Estado civil	Cobertura médica	Nivel educativo	Condición de actividad	Categoría de inactividad	Ingreso per cápita
Sexo	1.00	0.07	0.00	-0.01	0.00	0.15	0.02	-0.02
Edad	0.07	1.00	-0.57	-0.14	-0.38	-0.37	-0.50	0.18
Estado civil	0.00	-0.57	1.00	0.04	0.28	0.44	0.11	-0.08
Cobertura médica	-0.01	-0.14	0.04	1.00	-0.05	0.01	0.21	-0.15
Nivel educativo	0.00	-0.38	0.28	-0.05	1.00	0.26	0.47	0.16
Condición de actividad	0.15	-0.37	0.44	0.01	0.26	1.00	0.40	-0.19
Categoría de inactividad	0.02	-0.50	0.11	0.21	0.47	0.40	1.00	-0.16
Ingreso per cápita	-0.02	0.18	-0.08	-0.15	0.16	-0.19	-0.16	1.00

La figura 6 muestra la matriz de correlación para 2025 muestra relaciones en general débiles entre las variables sociodemográficas y laborales. Se destaca la correlación negativa entre edad y estado civil (-0,52) y entre edad y categoría de inactividad (-0,39). A su vez, estado civil se asocia positivamente con condición de actividad (0,39), mientras que nivel educativo presenta una correlación moderada y positiva con ingreso per cápita (0,19). En síntesis, las

asociaciones son consistentes con las tendencias observadas en 2005, aunque con magnitudes algo menores.

Figura 6.

Matriz de correlación 2025:

	Sexo	Edad	Estado civil	Cobertura médica	Nivel educativo	Condición de actividad	Categoría de inactividad	Ingreso per cápita
Sexo	1.00	0.06	0.00	0.01	0.06	0.12	0.03	-0.03
Edad	0.06	1.00	-0.52	-0.15	0.05	-0.30	-0.39	0.07
Estado civil	0.00	-0.52	1.00	0.08	-0.11	0.39	0.16	-0.03
Cobertura médica	0.01	-0.15	0.08	1.00	-0.06	0.06	0.13	-0.08
Nivel educativo	0.06	0.05	-0.11	-0.06	1.00	-0.23	0.14	0.19
Condición de actividad	0.12	-0.30	0.39	0.06	-0.23	1.00	0.26	-0.11
Categoría de inactividad	0.03	-0.39	0.16	0.13	0.14	0.26	1.00	-0.12
Ingreso per cápita	-0.03	0.07	-0.03	-0.08	0.19	-0.11	-0.12	1.00

Las Figuras 7 y 8 presentan las matrices de correlación para el Gran Buenos Aires en 2005 y 2025 en formato de *heatmap*. A diferencia de la tabla numérica, este recurso permite visualizar de manera inmediata la intensidad y dirección de las asociaciones entre variables sociodemográficas, educativas y económicas, ya que las correlaciones negativas se representan en tonos azules y las positivas en tonos rojizos. De este modo, patrones que en la tabla requieren una lectura más detallada se vuelven más evidentes visualmente, ya que el color resalta los contrastes y magnitudes de las relaciones. El heatmap no agrega información nueva respecto de los valores, pero sí facilita la comparación entre los dos años analizados, destacando que en 2025 las correlaciones tienden a ser más débiles que en 2005. En conjunto, la representación gráfica funciona como un complemento a las tablas, ofreciendo una lectura más clara e intuitiva de los vínculos entre las variables.

Figura 7.

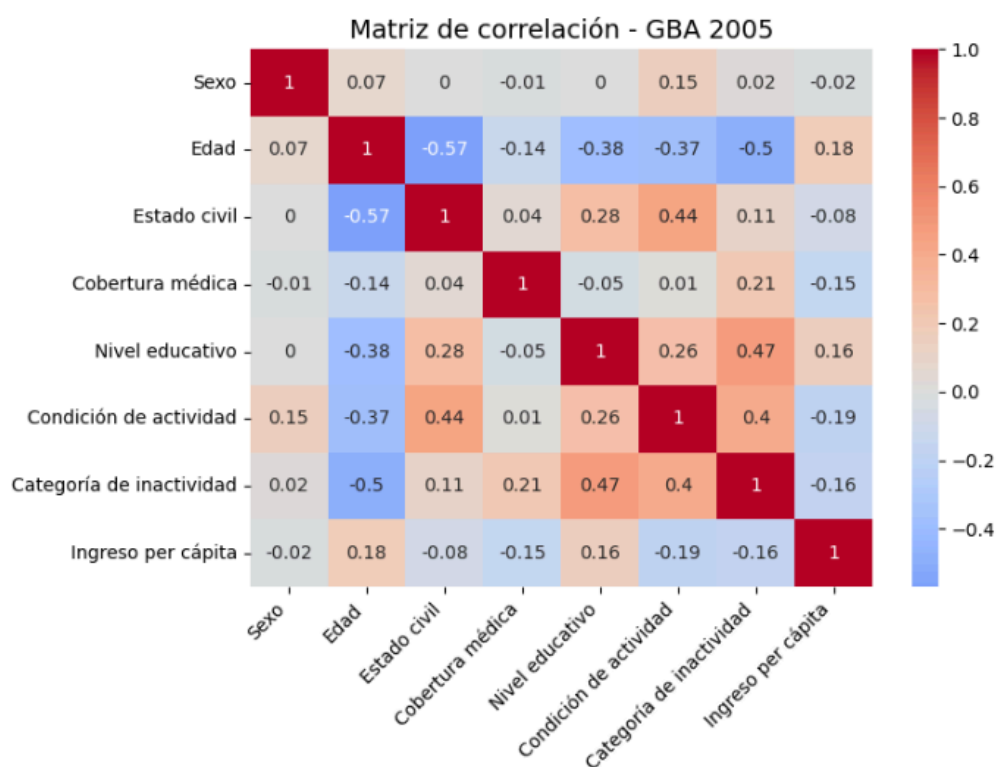
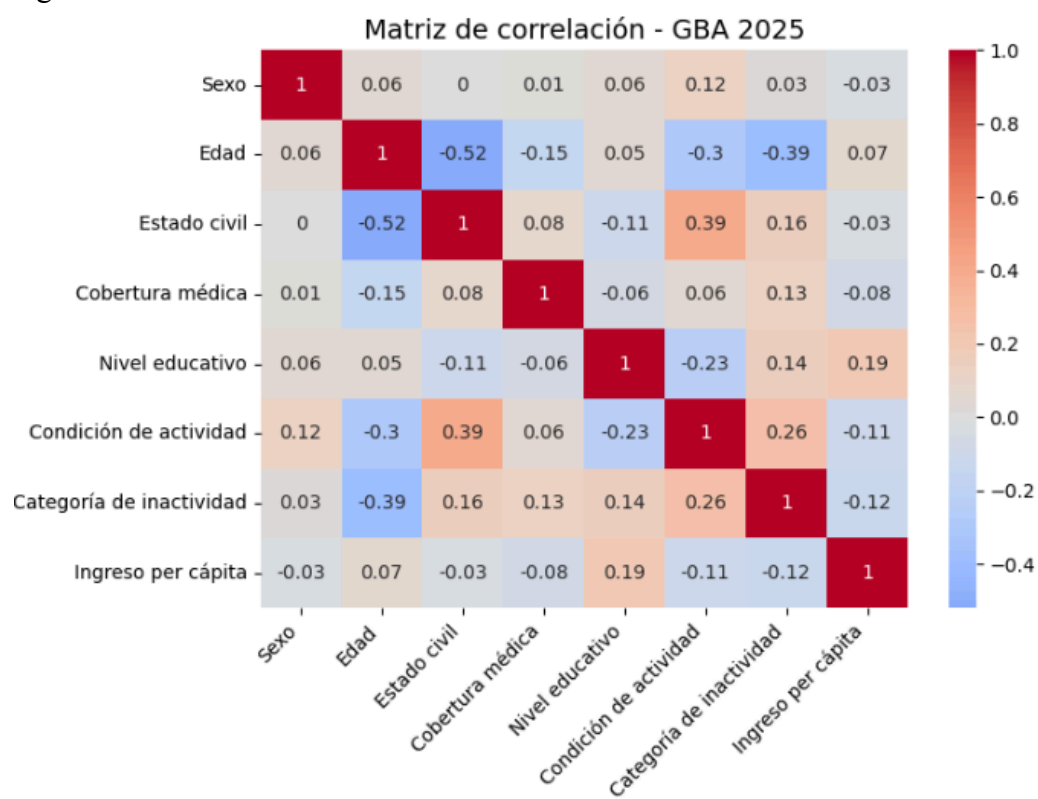


Figura 8.



Parte III: Conociendo a los pobres y no pobres


Para el análisis de pobreza en el Gran Buenos Aires se siguieron una serie de pasos metodológicos que permitieron identificar a los hogares pobres y no pobres en los años 2005 y 2025.

En primer lugar, se trabajó con la variable de ingreso total familiar (ITF), distinguiendo entre los hogares que reportan sus ingresos y aquellos que no lo hicieron. Los casos con ITF igual a 0 fueron clasificados como no respondientes, mientras que los ingresos distintos de cero se consideraron observaciones válidas. De esta manera, se construyó una nueva base denominada *respondieron*. En total, 13.680 hogares reportaron su ingreso, mientras que 2.985 no lo hicieron. Este primer paso resulta relevante, dado que la no respuesta en la variable de ingresos constituye un problema creciente en la encuesta permanente de Hogares (EPH).

A continuación, se incorporó a la base la variable de adultos equivalente, que pondera a cada miembro del hogar según sexo y edad para obtener una medida ajustada de las necesidades del grupo familiar. Luego, se calculó el ingreso necesario para cada hogar a partir de la canasta básica total (CBT) en cada año, multiplicando por la cantidad de adultos equivalentes. Para el primer trimestre de 2005, el valor de la CBT por adulto equivalente era \$205,07, mientras que en 2025 ascendía a \$365.177.

8. Con esta información se generó la variable *pobre*, que toma valor 1 cuando el ITF de un hogar resulta menor al ingreso necesario y 0 en caso contrario. Este procedimiento permitió identificar la proporción de hogares pobres en cada año. La Figura 9 muestra que en 2005, de un total de 9.371 hogares, el 26% se encontraba por debajo de la línea de pobreza. En 2025, con una muestra de 4.309 hogares, la incidencia de la pobreza aumentó al 31,1%. Estos resultados evidencian un incremento en la pobreza a lo largo del periodo analizado.

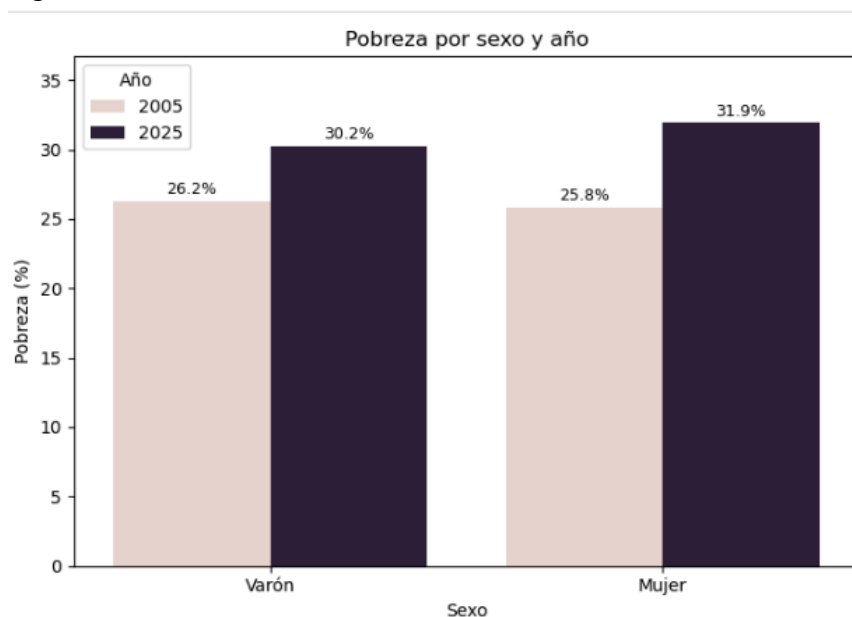
Figura 9.

 Pobreza por año en Gran Buenos Aires

Año	Total_muestra	Pobres_n	Pobres_pct
2005	9371	2438	26.0%
2025	4309	1341	31.1%

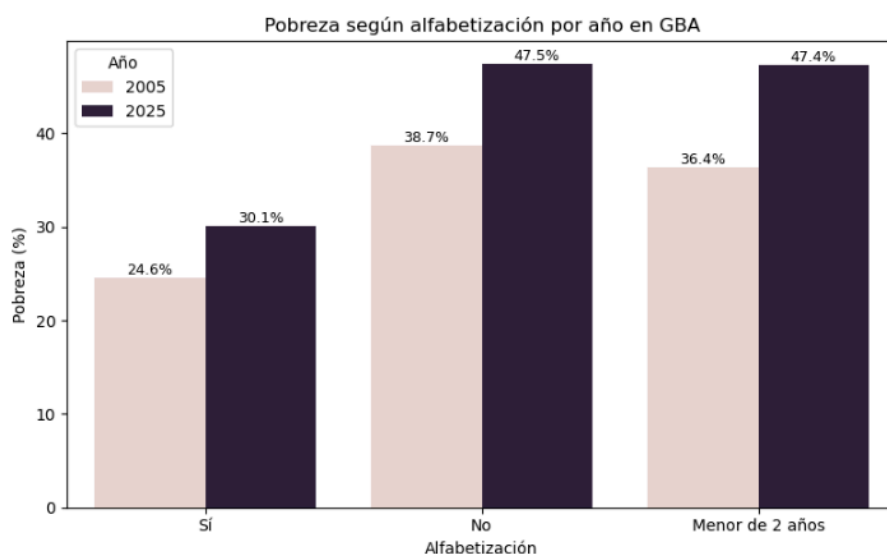
9. La Figura 10 desgrega los datos por sexo y por año. Se observa que en 2005 las diferencias de género eran reducidas (26,2% de pobreza en varones frente a 25,8% en mujeres). Sin embargo, en 2025 la brecha se amplía: la pobreza alcanza al 30,2% de los varones y al 31,9% de las mujeres. Esto refleja un mayor impacto de la pobreza sobre la población femenina en el periodo más reciente.

Figura 10.



Finalmente la Figura 11 muestra la pobreza según condición de alfabetización, osea si saben leer y escribir. Los resultados evidencian una fuerte relación entre nivel educativo y situación de pobreza. En 2005, las personas alfabetizadas presentaban una tasa de pobreza del 24,6%, frente a valores considerables más altos entre quienes no sabían leer ni escribir 38,7%. En 2025 estas brechas se mantienen, aunque en un contexto de aumento generalizado: la pobreza en alfabetizados se eleva al 30,1%, mientras que entre los no alfabetizados asciende a 47,5%. Además, se tomó en consideración los niños menores a dos años de edad como una tercera categoría ya que todavía no era posible decretar si eran alfabetos. En 2005, se registraron 36.4% niños dentro de la categoría mientras que en 2025, creció a un 47.4%.

Figura 11.



Referencias:

Instituto Nacional de Estadística y Censos (INDEC). (2016). La medición de la pobreza y la indigencia en la Argentina (Metodología INDEC, N.º 22). INDEC. ISBN 978-950-896-487-8.