Trabajo práctico Nº1

Ciencia de Datos

Grupo Nº4

Alumnos

Juan Augusto Alvarez Simonassi , Camila Gioja , María Delfina González Elosú

Profesores

María Noelia Romero

Ignacio Anchorena

Tomas Enrique Buscaglia

GitHub: https://github.com/mgonzalezelosu/Ciencia_de_Datos_TP1_Grupo4.git

Parte I: Familiarización y limpieza de la base EPH

Siguiendola metodología del INDEC, se considera pobre a un hogar cuyo ingreso total familiar (ITF) resulta inferior al ingreso necesario para cubrir la canasta básica total (CBT). El INDEC no mide la pobreza directamente, sino que establece parámetros objetivos a través del método de la Línea de Pobreza (LP), el cual se basa en los ingresos de los hogares. La medición parte de la construcción de dos canastas:

- Canasta Básica Alimentaria (CBA): incluye un conjunto de alimentos capaces de cubrir un umbral mínimo de necesidades energéticas y proteicas. Cuando los ingresos de un hogar no alcanzan para adquirir esta canasta, este hogar se clasifica dentro de la categoría de indigente.
- Canasta Básica Total (CBT): amplía la CBA al incorporar bienes y servicios no alimentarios considerados esenciales (vestimenta, transporte, educación y salud). Si los ingresos de un hogar no alcanzan para cubrir la CBT, se lo ubica en la categoría de pobre.

El cálculo de estas canastas se realiza a nivel de hogar y no de manera individual, ajustando los valores según la composición familiar mediante el concepto de adulto equivalente. A partir de estas referencias, se elaboran los datos oficiales sobre hogares y personas bajo la LP, utilizando como fuente la Encuesta Permanente de Hogares (EPH). El procedimiento consiste en comparar los ingresos de cada hogar con el costo de la CBA y de la CBT.

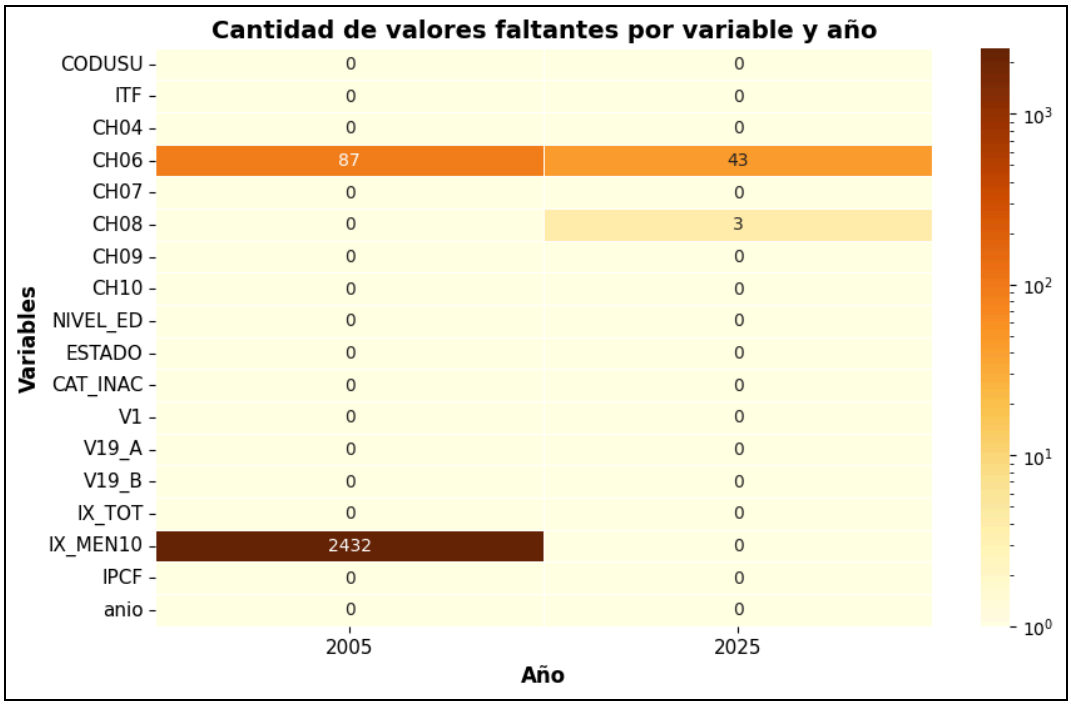
En una primera instancia, se descargaron los archivos de microdatos publicados por el INDEC para el primer trimestre del 2005 en formato .dta, y del primer trimestre del 2025 en formato .xls, junto con los respectivos diccionarios de variables. Posteriormente, se seleccionó el cuyo como región especifica del país trabajar. Cabe destacar que para cada año se trabajó con dos bases de datos distintas: una referida a los hogares y otra a los individuos. La base de hogares contiene información agregada del conjunto familiar, mientras que la base de individuos muestra características particulares de cada miembro.

Siguiendo con la limpieza de datos, se definió un conjunto de quince variables de interés, entre las cuales se incluyeron: CH04 (sexo), CH06 (edad), CH07 (estado civil), CH08 (cobertura médica), NIVEL_ED (nivel educativo), ESTADO (condicion de actividad), CAT_INAC (categoria de inactividad) e IPCF (monto de ingreso per cápita familiar). Además, se decidió incorporar como variables adicionales el ITF (monto de ingreso total familiar), V1 (si el hogar vivió de lo que gana en el trabajo), V19_A (si algún menor de 10 años contribuye con dinero trabajando), V19_B (si algún menor de 10 años contribuye con dinero pidiendo, IX_TOT (cantidad de miembros en el hogar), IX_MEN10 (cantidad de miembros menores de 10 años), CH09 (alfabetización) y CH10 (asistencia escolar). Por último, se utilizó el identificador

CODUSU, que permite relacionar las bases de datos de hogares e individuos con un mismo código. Dado que algunas variables se encontraban en la base de hogares y otras en la de individuos, este identificador resultó fundamental para poder integrar ambas fuentes de información en una única base por año, asegurando la correspondencia entre datos individuales y por hogar. Para garantizar la consistencia entre ambas bases, se verificó la correcta codificación de cada variable y se unificaron los formatos, dado que los archivos de 2005 y 2025 presentaban diferencias en los tipos de dato (string, float, integers, etc.). Además, se almacenaron todas las variables en mayúsculas dado que en la base de 2005 estaban cargadas de manera inconsistente.

En esta etapa también se detectaron valores negativos en la variable CH06 (edad en años cumplidos), que correspondían al código -1, y según la documentación de la EPH indica “no sabe/no responde. Para mantener la coherencia y permitir los análisis válidos, dichos valores se recodificaron como NaN, y posteriormente armamos un heatmap para poder visualizar la cantidad de valores faltantes por variable y año (**Figura 1**). Asimismo, en el dataframe 2005 se observó que la variable IX_MEN10 presentaba 2432 NaNs y ningún caso registrado como cero. Dado que, de acuerdo con el diseño de la EPH, la ausencia de menores de diez años en el hogar debe codificarse con un 0, se corrigió la variable, reemplazando los valores faltantes por ceros. Por otra parte, en la variable CH08 (cobertura médica) del año 2025, los valores faltantes se mantuvieron como NaN, ya que corresponden a casos de no respuesta en una pregunta cerrada. Una vez realizadas estas correcciones, se procedió a concatenar las bases de 2005 y 2025 en un único dataframe, agregando una nueva columna llamada anio, que identifica el período al que corresponde cada observación.

Figura 1



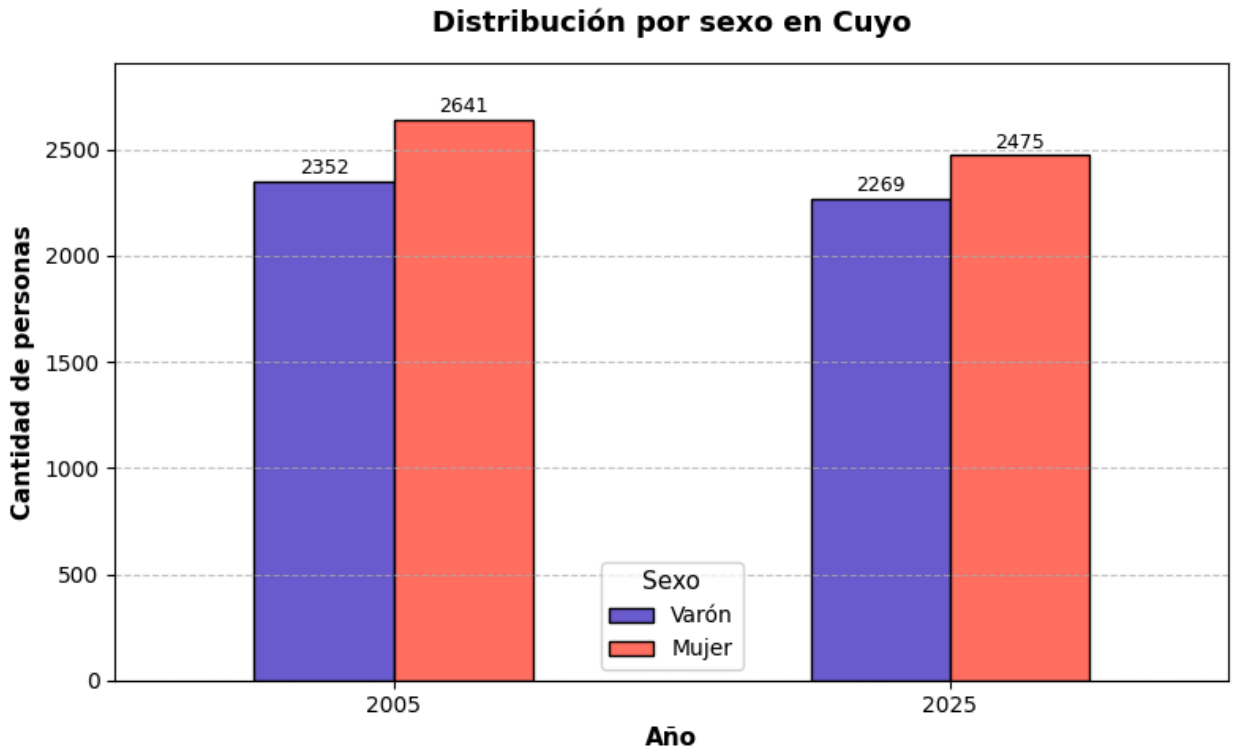
Nota: se utilizó escala logarítmica para una mejor visualización de los datos.

Parte II: Primer análisis exploratorio

En esta etapa se buscó caracterizar de manera descriptiva la composición de la población y las relaciones entre las principales variables sociodemográficas seleccionadas.

En primer lugar, se elaboró un gráfico de barras mostrando la distribución por sexo en 2005 y 2025 dentro de la región analizada. Para el 2005, participaron en la muestra 2352 varones y 2641 mujeres, mientras que en el 2025 participaron 2269 varones y 2475 mujeres. Los resultados indicaron que la proporción de varones y mujeres se mantuvo relativamente estable entre ambos períodos, con variaciones de baja magnitud que no alteran la composición general de la muestra (**Figura 2**). En ambos años, se observó una mayor proporción de mujeres que de varones.

Figura 2



Posteriormente se construyó una matriz de correlación utilizando las variables CH04 (sexo), CH06 (edad), CH07 (estado civil), CH08 (cobertura médica), NIVEL_ED (nivel educativo), ESTADO (condición de actividad), CAT_INAC (categoría de inactividad) e IPCF (ingreso per cápita familiar). Para hacerlo fue necesario generar variables dicotómicas a partir de las categorías originales, de modo que los coeficientes resultaran comparables y pudieran graficarse en un mapa de correlaciones (**Figura 3a y 3b**). El análisis de correlaciones muestra patrones consistentes con lo esperado para ambos años. Por ejemplo, entre edad y ciclo de vida, la edad se asoció positivamente con estar casado (2005 r = 0,48; 2025 r = 0,46), viudo (2005 r = 0,42; 2025 r = 0,37) y con ser jubilado o pensionado (2005 r = 0,50; 2025 r = 0,52), mientras que se correlacionó negativamente con ser soltero (2005 r = -0,73; 2025 r = -0,69) y con categorías como ser estudiante (2005 r = -0,51; 2025 r = -0,55). Asimismo, en 2005 el nivel educativo se vinculaba ligeramente con el ingreso per cápita familiar (r = 0,21), pero en 2025 esta relación ya no apareció tan marcada (el nivel educativo pierde peso como correlato directo del ingreso. En condición de actividad, en 2005 estar ocupado correlacionar positivamente con el ingreso (r = 0,30) y ser inactivo correlacionar negativamente (r = -0,27), pero en 2025 estas correlaciones se debilitan o desaparecen, aunque aún se observan asociaciones internas fuertes, como inactivo con menor de 10 años (r = -0,74) y con jubilados (r = 0,39). Esto nos llevaría a indicar que la variable "inactividad" queda más bien vinculada a grupos etarios específicos que directamente al ingreso. Por último, en 2005 hubo una correlación negativa entre depender de planes y seguros públicos y el ingreso (r = -0,36). En general, no vemos insights particularmente novedosos, sino mas bien correlaciones marcadas por superposición de definiciones (como por ejemplo entre ser menor de 6 años y ser menor de 10 años).

Figura 3a

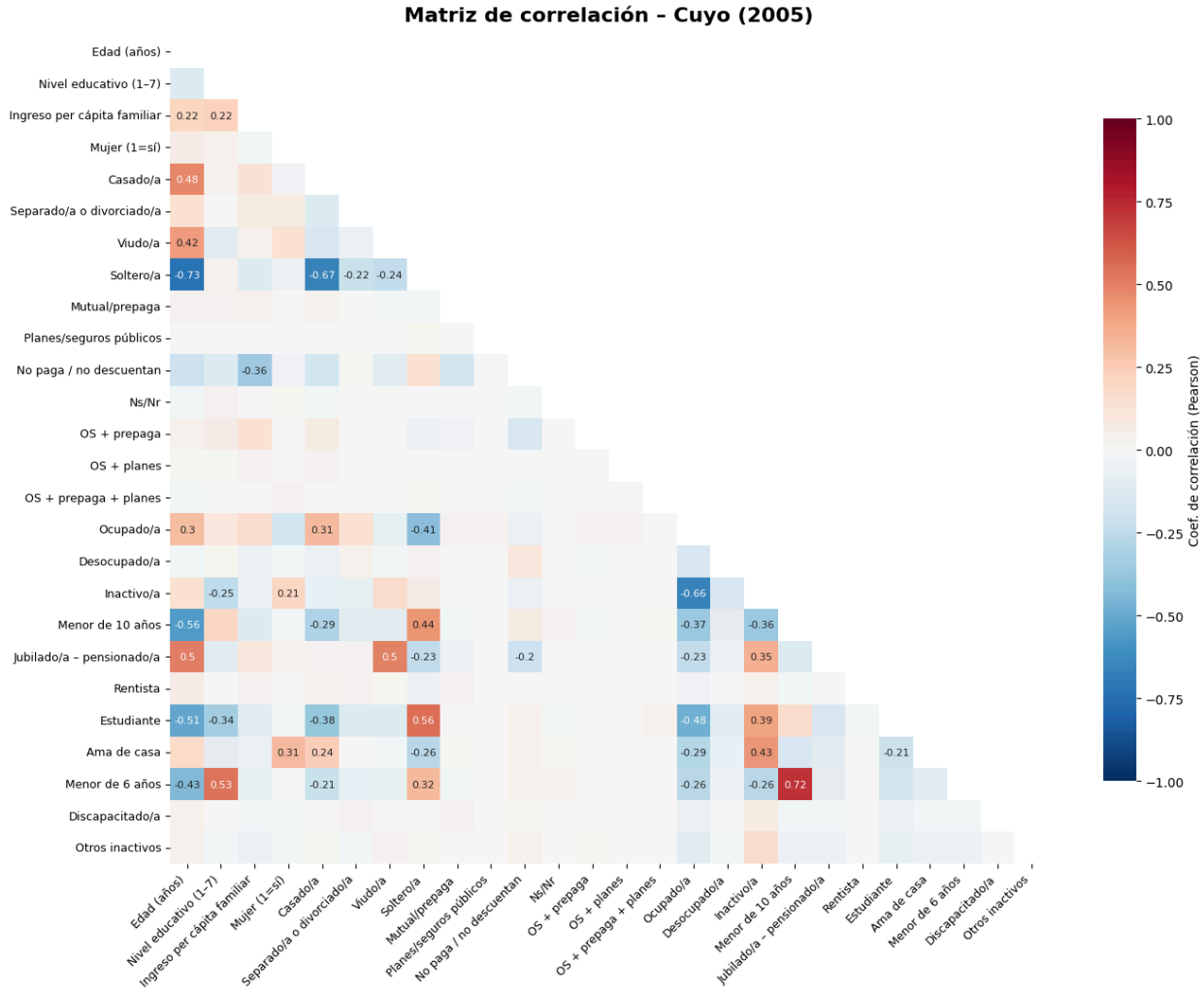
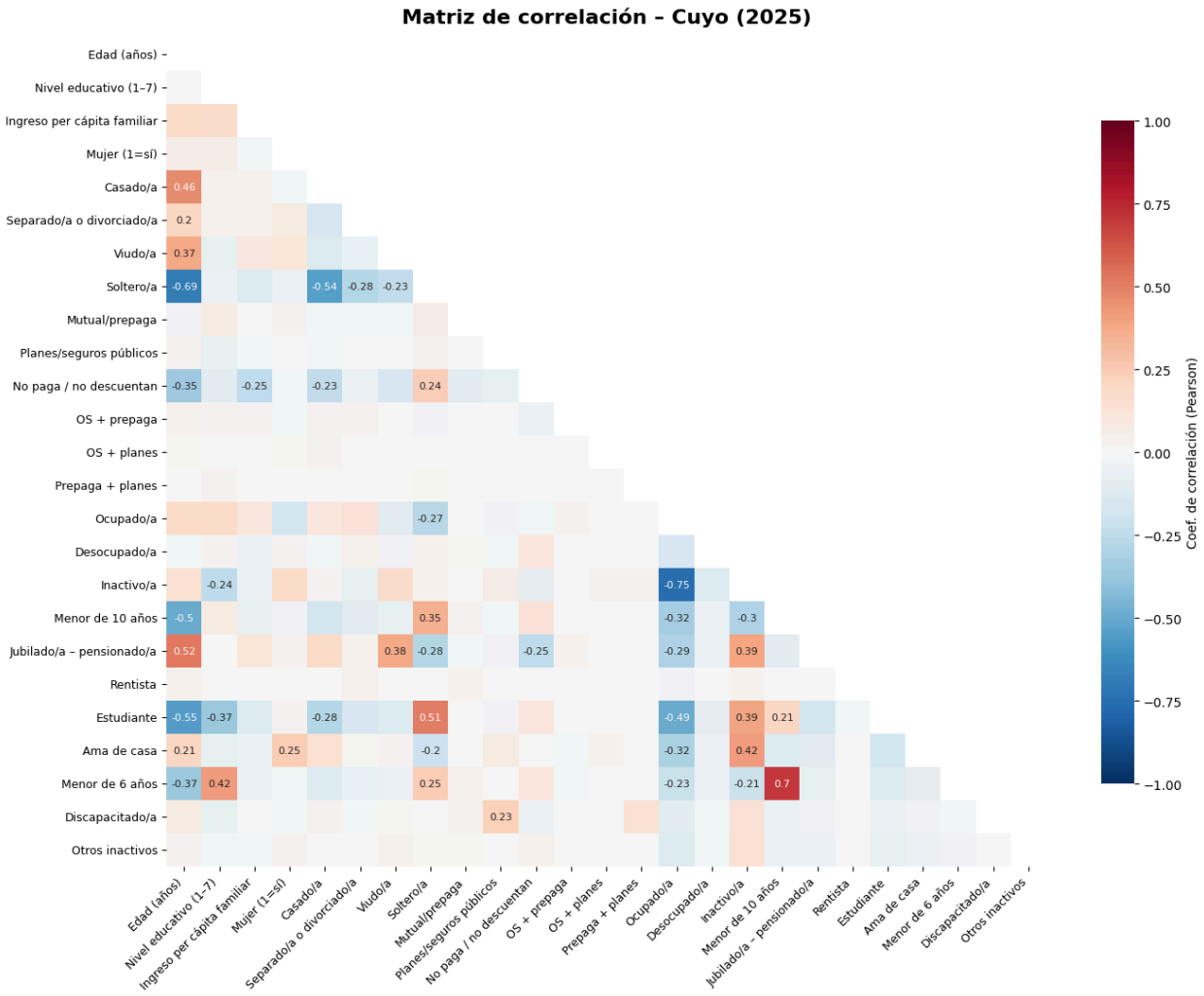


Figura 3b



Parte III: Conociendo a los pobres y no pobres

En esta etapa se buscó delimitar la población de análisis y construir el indicador de pobreza a partir del ingreso total familiar y los adultos equivalentes. Como primer control, se contabilizaron los registros con código de no respuesta en la variable ESTADO (valor igual a 0) en cada año: se observaron cinco casos en 2005 y cuatro casos en 2025. Luego, se hizo el mismo proceso sobre la variable ITF, definiendo subconjuntos y trabajando con bases distintas para los casos donde el ITF sea NaN o cero. Para 2005, la base respondieron2005 (ITF distinto de 0 y no nulo) tuvo 4946 observaciones y norespondieron2005 tuvo 37; para 2025, respondieron2025 registró 3656 observaciones y norespondieron2025, 1088. En total, 8602 personas contaban con ITF válido para el análisis.

A continuación, se incorporó el ajuste por necesidades demográficas utilizando el archivo *tabla_adulto_equiv.xlsx*. Para hacerlo, se implementó una función para mapear la edad y el sexo de cada persona y así asignarle, en cada año, el valor de adulto equivalente indicado en la tabla.

Luego, mediante el identificador CODUSU se agregaron estos valores a nivel de hogar para obtener la cantidad total de adultos equivalentes por vivienda. Con esa información se calculó el ingreso necesario para no ser pobre en cada hogar, multiplicando la canasta básica total por los adultos equivalentes: $205,07 por adulto equivalente para 2005 y $365.177 para 2025.

Para identificar la cantidad de persona debajo de la línea de pobreza se tomó como referencia el ITF (ingreso total familiar). En la base de análisis con ITF válido ("respondieron") se creó la variable pobre, que toma valor verdadero cuando el ITF del hogar es inferior al ingreso necesario del año correspondiente y falso en caso contrario. Con este criterio, en 2005 se identificaron 1.836 personas pobres (37,05% de la muestra con ITF válido) y en 2025 1.637 personas pobres (44,78% de las observaciónes) (**Figura 4**). Aunque el conteo de pobres baja en 2025, es un efecto relativo al menor tamaño muestral; cabe mirar al porcentaje de pobres en 2025 para darnos cuenta que en realidad la situación empeora.

Figura 4

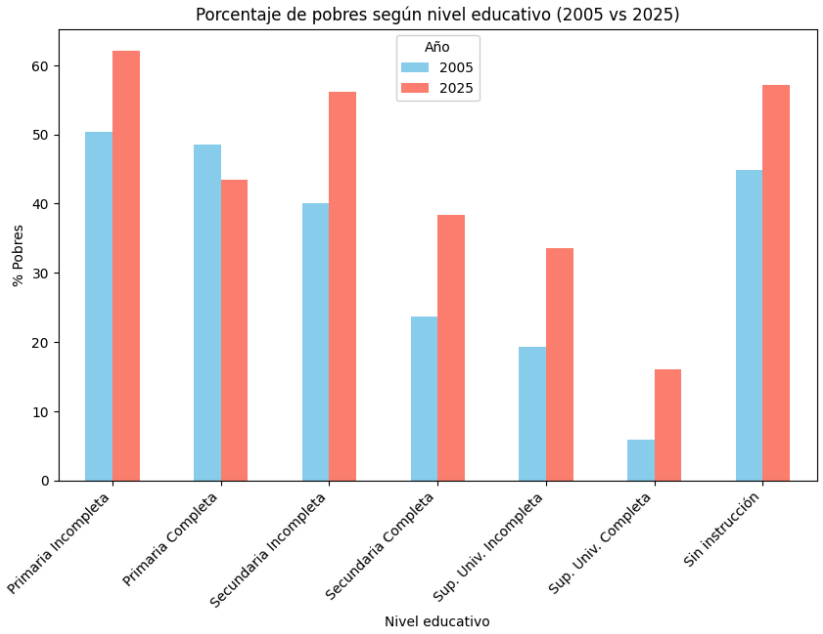
Análisis descriptivo de la pobreza

Año	Observaciones	Pobres	No pobres	% Pobres	% No pobres	DE
2005	4956	1836	3120	37,05	62,95	0,483
2025	3656	1637	2019	44,78	55,22	0,497

Nota: DE = Desvío Estándar

Por último, decidimos elaborar un gráfico de barras comparando niveles educativos y porcentaje de personas pobres entre 2005 y 2025 (**Figura 5**). El gráfico muestra una relación inversa entre nivel educativo y pobreza, donde los porcentajes más altos de pobreza se concentran en los grupos con menor educación, como primaria incompleta y personas sin instrucción, que superan el 50% en ambos años y alcanzan más del 60% en 2025. Asimismo, quienes completaron estudios universitarios presentan niveles de pobreza más bajos, inferiores al 20% en 2025. Al comparar ambos años, se observa un aumento generalizado de la pobreza en casi todos los niveles educativos, siendo más marcado en secundaria incompleta, secundaria completa y educación superior incompleta, con incrementos de entre 10 y 15 puntos porcentuales.

Figura 5



Por otro lado, tambien decidimos comparar el porcentaje de pobres según la alfabetización entre 2005 y 2025 (**Figura 6**). El gráfico evidencia que la pobreza es más elevada entre las personas que no saben leer ni escribir en comparación con quienes sí saben. En 2005, el 44% de las personas analfabetas eran pobres, mientras que en 2025 esta cifra supera el 60%, lo que indica un incremento muy marcado en este grupo. Entre quienes saben leer y escribir, la incidencia de la pobreza es menor, aunque también se puede ver un aumento en el 2025, pasando de aproximadamente 36% en 2005 a más del 40%. Estos resultados refuerzan la idea de que la alfabetización actúa como un factor de protección frente a la pobreza, aunque entre 2005 y 2025 la proporción de pobres aumentó en todos los grupos.

Figura 6