Trabajo Práctico 1 - Informe Escrito

Licenciatura en Ciencias del Comportamiento

Alumnas - Grupo 12

Rocío Echavarri (34173)

Martina Sacks (34028)

Malena Salamida (34114)

Profesores

María Noelia Romero

Tomas Enrique Buscaglia

Asignatura

Ciencia de Datos (CC408) - Tutorial 3 - Lunes 14:00

Fecha de presentación

5 de Septiembre de 2025

Parte I

Identificación de personas bajo la línea de pobreza

Para identificar a las personas en situación de pobreza, el INDEC calcula el valor de la Canasta Básica Total (CBT), que incluye una Canasta Básica Alimentaria (CBA) y un conjunto de bienes y servicios tales como vestimenta, transporte, educación, salud, entre otros. El valor de los componentes de estas canastas se determina a partir del Índice de Precios al Consumidor (IPC), que varía según el período de medición (Instituto Nacional de Estadística y Censos [INDEC], 2016).

Aquellos hogares que no cuenten con el ingreso total familiar para costear la CBT son considerados hogares pobres, y dicha clasificación se extiende a cada uno de sus integrantes (INDEC, 2016).

Delimitación de la muestra

Para llevar a cabo el análisis de la Encuesta Permanente de Hogares (EPH) se restringió la muestra a la región noroeste (NOA), puesto que es la tercera con mayor cantidad de datos (n = 19.090), precedida por la región pampeana y Gran Buenos Aires.

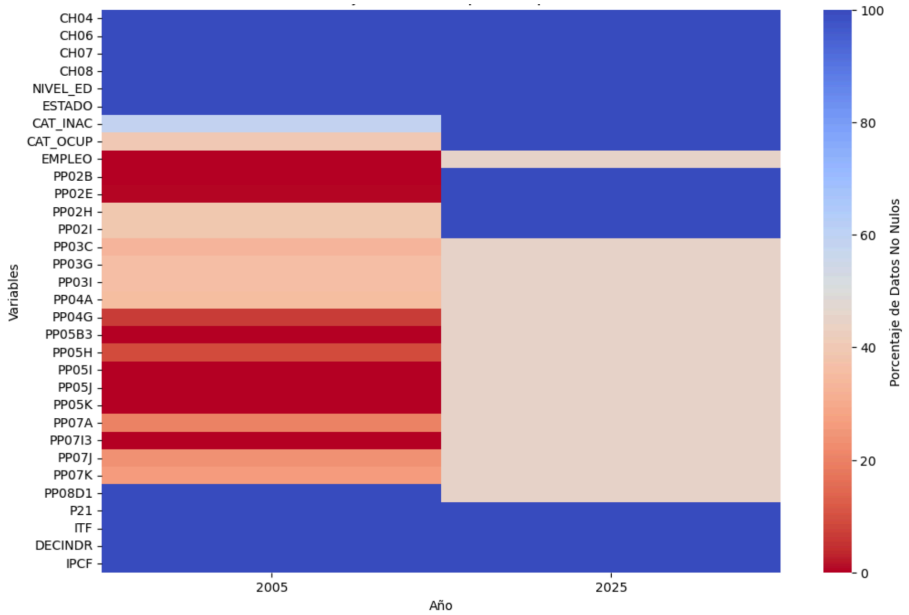
Selección de variables

Se eligieron ocho variables principales que permiten una caracterización adecuada de la muestra estudiada. Se incluyen el sexo (`CH04`), la edad (`CH06`), el estado civil (`CH07`), el tipo de cobertura médica (`CH08`), el nivel educativo (`NIVEL_ED`), la condición de actividad (`ESTADO`), la categoría de inactividad (`CAT_INAC`) y el monto de ingreso familiar per cápita (`IPCF`). De este modo, contaremos con información precisa sobre los miembros de los hogares analizados y así podremos categorizar las diferentes variables elegidas a continuación.

Por otro lado, se seleccionaron veinticuatro variables complementarias para indagar sobre las diferencias de empleabilidad entre las diferentes categorías establecidas por la EPH (económicamente activa, ocupada, desocupada, ocupada demandante de empleo y subocupada). Dichas variables son `CAT_OCUP`, `EMPLEO`, `PP02B`, `PP02E`, `PP02H`, `PP02I`, `PP03C`, `PP03G`, `PP03I`, `PP04A`, `PP04G`, `PP05B3`, `PP05H`, `PP05I`, `PP05J`, `PP05K`, `PP07A`, `PP07I3`, `PP07J`, `PP07K`, `PP08D1`, `P21`, `ITF`, `DECINDR`, `IPCF`.

En el mapa de calor (Figura 1) se observa que las variables con mayor cantidad de datos faltantes en 2005 son `EMPLEO`, `PP02B`, `PP02E`, `PP02H`, `PP05B3`, `PP05I`, `PP05J`, `PP05K`, `PP07I3`. Esto se debe a que, de las variables mencionadas, solo `PP02E` y `PP02H` se encuentran en ambas bases de datos, en tanto que las demás fueron agregadas a la EPH luego del 2005. En cambio, para 2025 se identificó un 50% de valores faltantes en todas las variables desde `PP03C` hasta `PP08D1`, además de para la variable `EMPLEO`. Existe una asimetría considerable en el volumen de valores vacíos entre los dos trimestres analizados, siendo el de 2005 el más afectado.

Figura 1. Porcentaje (%) de valores no nulos en las variables de la EPH consideradas
El color azul indica mayor porcentaje de valores no nulos.



Limpieza de datos

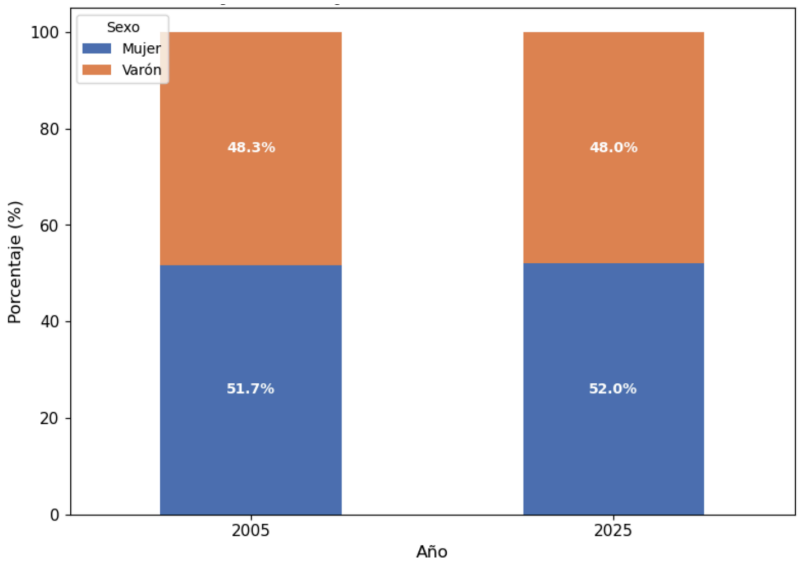
Para la limpieza de datos, se recategorizaron las variables de la EPH 2005, ya que estaban codificadas en formato string, mientras que las de 2025 se encontraban en formato numérico (enteros que corresponden a las categorías establecidas por el INDEC). Asimismo, se unificaron los nombres de las columnas en mayúsculas para facilitar su identificación. Finalmente, dado que algunas variables presentes en 2025 no estaban disponibles en 2005, se crearon dichas columnas en el dataframe de 2005 antes de concatenar ambas bases con la función *pd.concat*, generando un único dataframe denominado `EPH_combinada` (ver Anexo).

Parte II

Análisis exploratorio

En primer lugar, se produjo un aumento en la cantidad de encuestados de 2005 (n = 9348) a 2025 (n = 9742). A continuación se analizó la composición de la muestra por sexo para los dos trimestres de interés. En ambos casos se observó una leve predominancia de mujeres (barras azules) por sobre los varones (barras naranjas) (Figura 2).

Figura 2. Gráfico de barras de la composición de la EPH por sexo para 2005 y 2025



Para construir la matriz de correlación se generaron variables dummies. En **`CH04`**, "Mujer" toma el valor 1 y el resto 0; en **`CH07`**, toma los valores del estado civil, tomando como base cuando están solteros. La variable **`CH08`** se dividió en "Obra Social", "Prepaga" y "Plan Público" según las categorías del INDEC y se tomó como base cuando no contaban con cobertura. En **`NIVEL_ED`**, se simplificaron las variables, se tomó la categoría "Primaria" como base y se codificaron "Secundaria" y "Terciaria/Univ." como dummies. En **`ESTADO`**, la base es "Ocupado" y se crearon las categorías "Desocupado" e "Inactivo". Finalmente, en **`CAT_INAC`** la base es "Otro" y se incluyeron las categorías "Estudiante", "Jubilado" y "Ama de Casa".

La matriz de correlación de la EPH 2005 (Figura 3) muestra cómo la mayoría de las correlaciones son bajas, lo que indica que las variables no se mueven de manera lineal muy fuerte unas con otras. Como se esperaba, surgieron correlaciones con la edad ordinal consistentes con una correlación positiva con estado civil casado (0.50) y con

jubilado/pensionado (0.44), lo que refleja trayectorias vitales previsibles; a su vez, muestra una correlación negativa con la condición de estudiante (-0.54), lo que también es lógico dado que a mayor edad disminuye la probabilidad de estar estudiando. En cuanto a las variables vinculadas al ingreso per cápita familiar (`IPCF`), se observan correlaciones positivas con el acceso a obra social (0.39) y con nivel educativo terciario/universitario (0.31), mientras que aparecen asociaciones negativas con ausencia de instrucción (-0.10) y con la categoría de estudiante (-0.11). Esto sugiere que los hogares con mayor ingreso tienden a tener una mayor inserción formal en el sistema de salud y niveles educativos más altos, lo cual está en línea con lo que se espera en contextos de desigualdad estructural.

En la matriz de 2025 (Figura 4) también se observa que las correlaciones son bajas a moderadas. Sin embargo, las variables con una mayor relación positiva con el IPCF son la educación terciaria/universitaria (0.24) y la cobertura de obra social (0.30), mientras que la falta de instrucción muestra una correlación negativa (-0.06) aunque relativamente baja. Esto indica que un mayor nivel educativo y acceso a cobertura de salud están asociados con mejores resultados socioeconómicos. Además, en las características sociodemográficas, se observan algunas correlaciones internas notables. Por ejemplo, ser mujer se asocia negativamente con no tener instrucción (-0.35) y ser estudiante (-0.61), mientras que la edad tiene una correlación positiva con estar casado (0.38) y jubilado/pensionado (0.43). Asimismo, las variables de cobertura de salud muestran correlaciones moderadas entre sí: obra social y prepaga (0.15), obra social y plan público (-0.03), sugiriendo cierta sustitución entre distintos tipos de cobertura. Finalmente, las relaciones negativas importantes en el ámbito educativo y ocupacional: ser estudiante se correlaciona negativamente con edad (-0.61), casado (-0.26) y jubilado/pensionado (-0.13), reflejando la incompatibilidad temporal de estas categorías.

En conclusión, las asociaciones entre variables se mantienen relativamente estables entre 2005 y 2025. Aunque las correlaciones no sean elevadas y dificulten extraer conclusiones contundentes, reflejan tendencias claras en la relación entre género, educación, edad e ingresos. Por ejemplo, existe una asociación persistente entre “Inactivo” y estudiantes, mujeres y adultos mayores.

Figura 3. Matriz de correlación de variables de la EPH en el primer trimestre de 2005.
Generada a partir de la categorización de las variables `CH04, CH06, CH07, CH08, NIVEL_ED, ESTADO, CAT_INAC, IPCF`

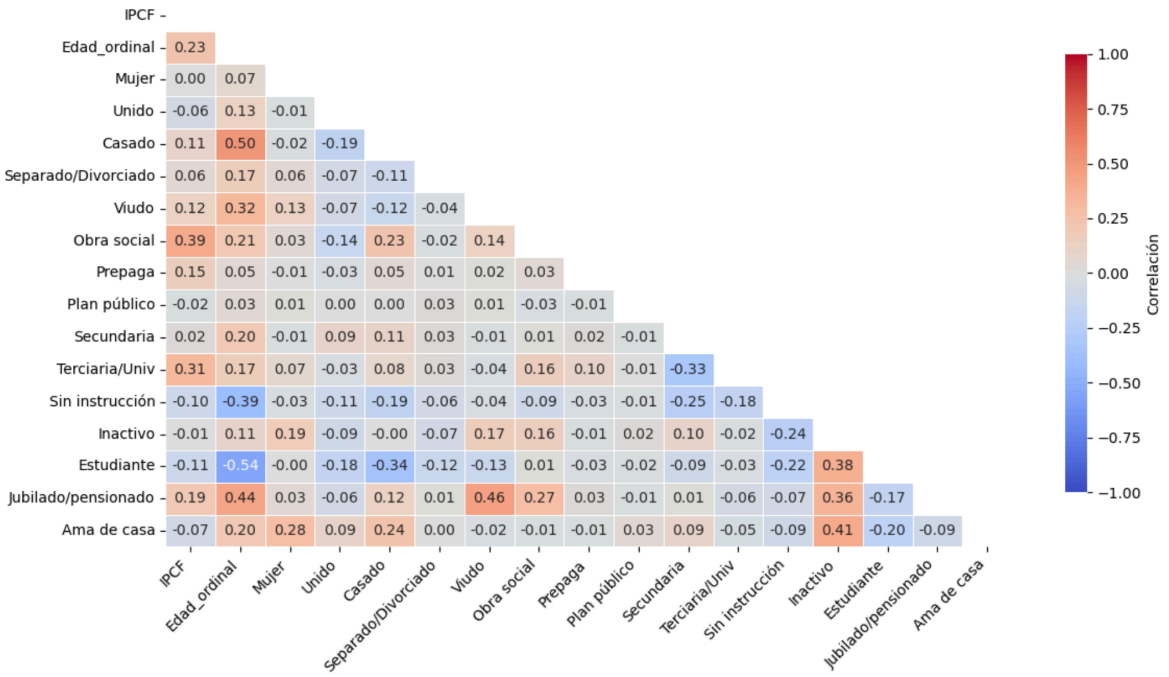
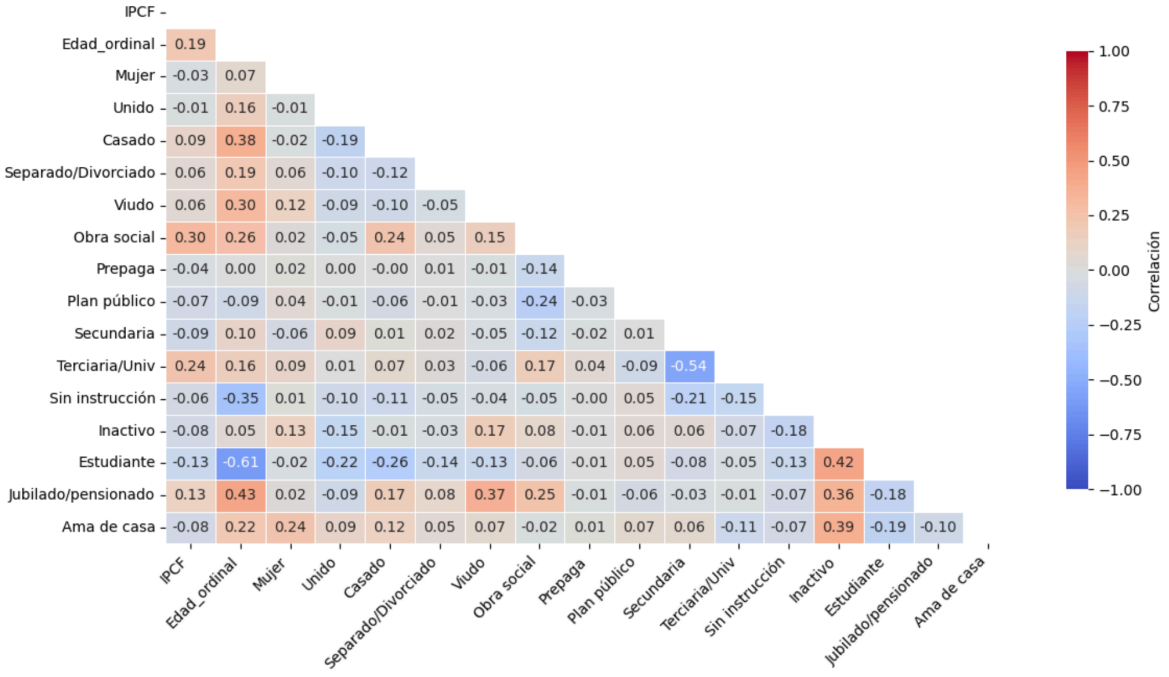


Figura 4. Matriz de correlación de variables de la EPH en el primer trimestre de 2025.
Generada a partir de la categorización de las variables `CH04, CH06, CH07, CH08, NIVEL_ED, ESTADO, CAT_INAC, IPCF`



Parte III

Personas bajo la línea de pobreza en la región noroeste

Solo se hallaron ocho personas que no reportaron su condición de actividad en el primer trimestre de 2005, en tanto que todos la reportaron en el primer trimestre de 2025. Por otra parte, la cantidad de personas que no reportaron sus ingresos fue 1.117, equivalente a un 5.85% de la muestra, mientras que 17.973 personas sí lo reportaron.

En 2005 se identificaron 4.461 pobres y en 2025 se calculó que la cantidad de personas bajo la línea de pobreza es de 3911, representando un 47.56 % y un 44.98% de la muestra total, respectivamente.

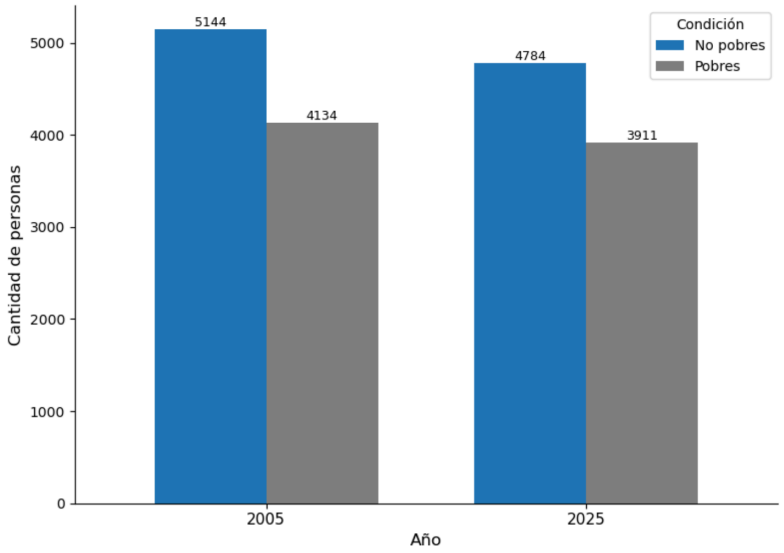
Además, se utilizó estadística descriptiva para observar más claramente las diferencias entre las personas bajo la línea de pobreza para los dos trimestres analizados (Tabla 1).

Tabla 1. Estadística descriptiva de los individuos bajo la línea de pobreza

Año	Porcentaje de pobres	Porcentaje de no pobres	Desviación Estándar	Número de observaciones	Máximo	Mínimo	Media	Mediana
2005	47.56	52.44	0.5	9380	1	0	0.48	0.0
2025	44.98	55.02	0.5	8695	1	0	0.45	0.0

En la Figura 5 se puede observar que en 2005 la cantidad de personas no pobres supera a la de las personas pobres. En cambio en el año 2025, el gráfico muestra que, aunque la cantidad de personas no pobres sigue siendo mayor, la diferencia con el grupo de personas pobres parece haber aumentado un poco. Por lo tanto, a pesar de que la cantidad de personas no pobres sigue siendo superior, la brecha entre ambos grupos es más grande en 2025.

Figura 5. Gráfico de barras comparando a personas bajo y sobre la línea de pobreza en el primer trimestre de 2005 y 2025

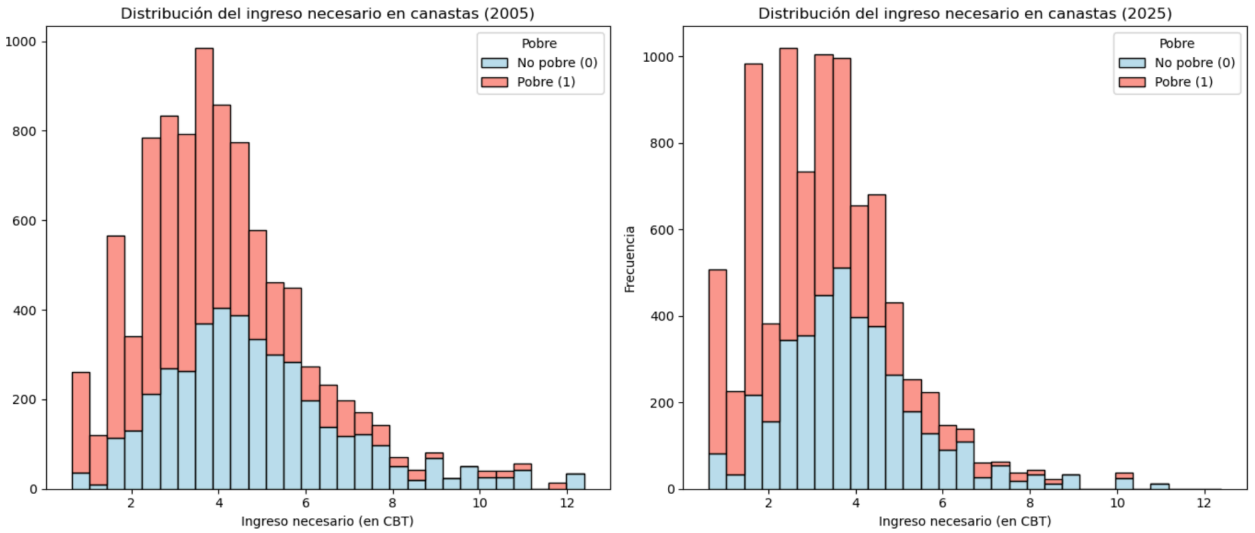


La Figura 6 contiene dos histogramas, uno para el año 2005 y otro para 2025. Se representó el ingreso necesario en función de la cantidad de Canastas Básicas Totales (CBT) que se logran cubrir con dicho ingreso. La comparación de la distribución entre 2005 y 2025 muestra una evolución significativa en la estructura de la pobreza. En 2005, la mayor parte de los hogares se concentraba entre 2 y 5 CBT, con un solapamiento considerable entre pobres y no pobres. Esto implicaba que incluso sectores con ingresos intermedios quedaban clasificados como pobres, reflejando una incidencia elevada y extendida de la pobreza en los niveles medios-bajos de la distribución.

En 2025 la situación presenta un cambio relevante: si bien la concentración en el rango de 2 a 5 CBT se mantiene, se observa un aumento en la proporción de hogares no pobres dentro de este segmento. La pobreza tiende a replegarse hacia los niveles más bajos de ingreso (1–2 CBT), reduciendo su peso relativo en los estratos medios. En términos distributivos, esto indica una mejora relativa, con más hogares que logran superar la línea de pobreza y una mayor segmentación entre quienes se encuentran en condiciones de vulnerabilidad severa y quienes han alcanzado cierto grado de estabilidad económica.

En síntesis, el contraste entre ambos períodos sugiere un proceso de reducción de la pobreza en los tramos medios, acompañado de una concentración del problema en los sectores más vulnerables. Si bien esto representa un avance en la capacidad de muchos hogares para superar la condición de pobreza, también debería focalizarse en los grupos que permanecen en la base de la distribución, donde persiste la situación entre 2005 y 2025.

Figura 6. Histograma de la distribución de canastas básicas (CBT) en 2005 y 2025



Bibliografía

Instituto Nacional de Estadística y Censos (2016). *La medición de la pobreza y la indigencia en la Argentina*. https://www.indec.gob.ar/ftp/cuadros/sociedad/EPH_metodologia_22_pobreza.pdf

Instituto Nacional de Estadística y Censos (2025). INDEC: Instituto Nacional de Estadística y Censos de la República Argentina. https://www.indec.gob.ar/

Anexo

Link al GitHub

Diccionario para base de datos 2005

Diccionario para base de datos 2025