Lic. en Ciencias del Comportamiento

Trabajo Práctico N° 1

Alumnas

Isabela Nicola

Agustina Kiessner

Clara Mollón

Profesores

María Noelia Romero

Tomas Enrique Buscaglia

Ignacio Anchorena

Asignatura

Ciencia de datos - Tutorial 3

Fecha de presentación

05/09/2025

Parte I: Familiarizandonos con la base EPH y limpieza

1.

Según el INDEC, las personas pobres se identifican basándose en el umbral de la línea de pobreza, que incluye no solo los insumos alimentarios mínimos, sino también otros consumos básicos no alimentarios, como por ejemplo transporte, vestimenta y salud, que conforman la Canasta Básica Total (CBT). Esta canasta se compara con los ingresos de los hogares relevados por la Encuesta Permanente de Hogares (EPH) y, cuando dichos ingresos resultan inferiores al valor de la CBT, el hogar es considerado en situación de pobreza (INDEC, 2025).

2.

A.

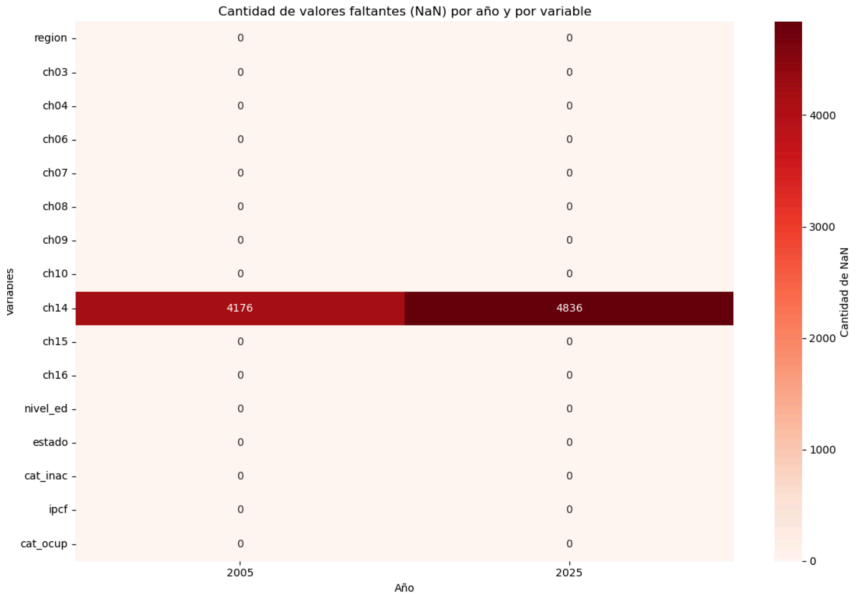
La región elegida fue NOA (Noroeste Argentino), luego de la filtración, la base de datos tiene 19.090 observaciones totales. En primer lugar, pusimos en minúscula los nombres de las variables de cada base de datos (2005 y 2025) para que todas tengan el mismo nombre y de esa manera evitamos inconsistencias al momento de integrarlas. Luego, al intentar filtrar la base de datos por la región seleccionada, el código saltaba con un error debido a que las variables de cada base de datos no estaban tipificadas de la misma manera. En la de 2005 se encontraba en formato de texto, mientras que en la de 2025 aparecía en formato numérico. Para solucionar este problema, cambiamos los valores de región en la base de datos de 2005 a número, para que sea igual a la de 2025 mediante una función que llamamos "reemplazar_valores". Después, unimos ambas bases de datos con la función *concat*. Es importante destacar dos particularidades. Por un lado la variable "ch14" en la base de datos de 2005, mediante la función *unique* notamos que entre sus valores figuraba uno que correspondía únicamente a un espacio en blanco. Para evitar que este tipo de registros distorsionara los resultados dentro de la función "reemplazar_valores" los convertimos en NaN para simplificar los análisis posteriores. Por otro lado, la variable "ch06" que corresponde a la pregunta "¿cuántos años cumplidos tiene?" la respuesta "Menos de 1 año" la transformamos en el valor 0. Por último, dentro de esta función convertimos los valores a tipo *float*, con el objetivo de garantizar la homogeneidad en el tipo de datos de toda la base.

B.

Las 15 variables seleccionadas fueron: "ch03" (Relación de parentesco), "ch04" (Sexo), "ch06" (Años cumplidos), "ch07" (Estado civil), "ch08" (Cobertura médica), "ch09"

(¿Sabe leer y escribir?), "ch10" (¿Asiste o asistió a algún establecimiento educativo?), "ch14" (¿Cuál fue el último año que aprobó?), "ch15" (Lugar de nacimiento), "ch16" (¿Dónde vivía hace 5 años?), "nivel_ed" (Nivel educativo), "estado" (Cuestion de actividad), "cat_inac" (Categoría de inactividad), "ipcf" (Monto de ingreso per cápita familiar) y "cat_ocup" (Categoría ocupacional). Además, agregamos las variables "región" y "ano4" para poder simplificar análisis posteriores.

Figura 1



Nota. La Figura 1 muestra la cantidad de valores faltantes (NaN) por variable y por año en las bases de datos analizadas (2005 y 2025).

Luego de realizar el heatmap (Figura 1) encontramos que simplemente la variable “ch14” era la única que presentó valores faltantes *NaN*, donde 4836 de datos corresponden a la base de datos de 2025 y 4176 a la de 2005. El resto de las variables no presenta valores faltantes en ninguno de los dos años. Dado que estos resultados nos resultaban sospechosos, procedimos a realizar estadísticas descriptivas con el fin de asegurarnos que los resultados obtenidos fueran correctos. Además, calculamos los valores faltantes antes de unificar las bases de datos para comprobar que no sea un problema de la unificación de bases a lo que obtuvimos los mismos resultados que antes.

C.

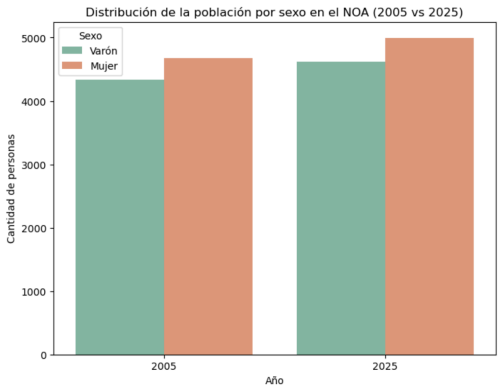
Para identificar los valores sin sentido, es decir, los extraños, primero fuimos chequeando con el código *unique* qué variables tienen valores que no corresponden al diccionario. Luego de realizar este análisis, concluimos que solo 3 variables tienen valores extraños: ch10, edad e ipcf. Durante la revisión se identificó que algunos participantes respondieron con el valor "0" en la variable "ch10". Este valor fue considerado inválido, dado que no corresponde a ninguna de las categorías especificadas en el diccionario de la encuesta. Además, en los comentarios generales del cuestionario se aclara que el "0" se utiliza únicamente para indicar casos en los que no corresponde responder la secuencia de preguntas asociadas. Sin embargo, en este caso en particular la pregunta "ch10" debía ser respondida por la totalidad de los participantes, por lo cual los registros con "0" fueron clasificados como extraños. También consideramos a los valores de las variables "edad" y "ipcf" que eran negativos como extraños.

Los valores considerados extraños fueron recodificados con el valor -1, con el objetivo de identificarlos fácilmente para después contarlos. Elegimos este valor dado que dicho no corresponde a ninguna categoría válida en la base de datos. Por lo tanto, para "ipcf" encontramos 0 valores extraños, para "ch06" hay 51 valores extraños y para ch10 hay 466 variables extraños. Posteriormente, estos registros fueron eliminados para garantizar la consistencia y validez de los análisis. Luego de esta eliminación mediante el uso del comando .loc, la base de datos quedó conformada por un total de 18.624 observaciones.

Parte II: Primer Análisis Exploratorio

3.

Figura 2

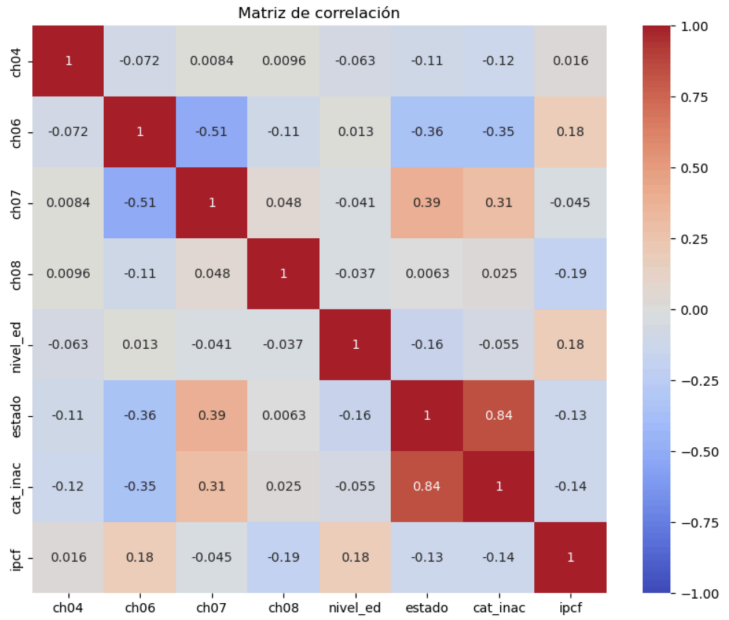


Nota. El gráfico muestra la composición por sexo en la región seleccionada para los años 2005 y 2025

En la Figura 2 podemos observar la distirbucion de la población por sexo en los años 2005 y 2025 en la región seleccionada, es decir el NOA. Como muestra la figura en ambos años hay levemente más mujeres que varones. En la base de datos de 2005 hay 4.332 varones y 4.677 mujeres, mientras que en la de 2025 hay 4.618 varones y 4.997. Asimismo, se evidencia un aumento en el número de personas de ambos sexos entre 2005 y 2025, manteniéndose la brecha relativa a favor de la población femenina.

4.

Figura 3



Nota. La matriz de correlación de ambos años muestra la relación entre las principales variables sociodemográficas seleccionadas.

En primer lugar, creamos la variable dicotómicas binarias para la columna “ch04”, es decir la sexo, ya que es la única dicotimica dentro de la selcción de vairables. En esta recodificación se mantuvo el valor “1” para identificar a los varones, mientras que el valor “2” fue transformado en “0” para representar a las mujeres.

Como podemos observar en la Figura 3, la correlación más fuerte y positiva (0.84) es entre la variable “estado” (condición de inactividad) y “cat_inac” (categoría de inactividad) lo cual es esperable, ya que ambas variables capturan de manera complementaria la condición de actividad laboral de las personas. Otra correlación fuerte y negativa es entre edad (CH06) y estado civil (CH07) con un r=-0,51, lo cual indica que a menor edad, es más probable estar soltero/a. A su vez, se presentan correlación moderados, como por ejemplo entre “ch06” se

correlaciona de manera negativa tanto con “cat_inac” con un r=-35 y “estado” con un r=-36. Esto nos dice que a medida que aumenta la edad se incrementa la probabilidad de ser inactivo.

Parte III: Conociendo a los pobres y no pobres

5.

En la base de 2005, 8 personas no respondieron su condición de actividad, y en la base de 2025, 16 personas no respondieron esta pregunta. Luego, para la variable “itf” contestaron 17.973 participantes, por lo que en la base “no respondieron” quedaron 1.117 observaciones.

6.

Para esta consigna agregamos la columna “adulto_equiv” a la base de datos “respondieron”, que contiene los valores de adulto equivalente a cada persona segun sexo y edad. Después, usamos el comando *groupby* para sumar esta columna para las personas que pertenecen a un mismo hogar, y la guardamos bajo el nombre de “ad_equiv_hogar”, como se muestra en la Tabla 1 a continuación.

Tabla 1

	ano4	codusu	itf	ch04	ch06	grupo_edad	adulto_equiv	\	ad_equiv_hogar
0	2005.0	125666	700.0	1	36.0	30 a 45 años	1.00		4.18
1	2005.0	125666	700.0	2	35.0	30 a 45 años	0.77		4.18
2	2005.0	125666	700.0	2	15.0	15 años	0.77		4.18
3	2005.0	125666	700.0	1	12.0	12 años	0.85		4.18
4	2005.0	125666	700.0	1	10.0	10 años	0.79		4.18
5	2005.0	126344	3800.0	1	53.0	46 a 60 años	1.00		2.79
6	2005.0	126344	3800.0	2	52.0	46 a 60 años	0.76		2.79
7	2005.0	126344	3800.0	1	16.0	16 años	1.03		2.79

Nota. Esta tabla muestra las primeras ocho observaciones de la base “respondieron.”

7.

Con el objetivo de determinar el ingreso mínimo requerido por cada hogar para no ser considerado pobre, agregamos a la base “respondieron” la columna llamada “ingreso_necesario”. Para esto, se utilizaron los valores de la CBT por adulto equivalente publicados por el INDEC tanto para el 2005 como el 2025. Posteriormente, el cálculo del ingreso necesario de cada hogar se realizó multiplicando el valor de la CBT por adulto equivalente según el año por la variable “ad equiv hogar”, algunos de estos resultados puede observar se en la Tabla 2.

Tabla 2

	ano4	ad_equiv_hogar	ingreso_necesario
0	2005.0	4.18	857.1926
1	2005.0	4.18	857.1926
2	2005.0	4.18	857.1926
3	2005.0	4.18	857.1926
4	2005.0	4.18	857.1926
...	...	...	...
17958	2025.0	1.78	650015.0600
17959	2025.0	0.76	277534.5200
17960	2025.0	1.99	726702.2300
17961	2025.0	1.99	726702.2300
17962	2025.0	1.99	726702.2300

Nota. La tabla muestra las primeras y últimas cinco observaciones de la base de datos.

8.

Agregamos a la base de datos "respondieron" la columna "pobre" que toma el valor 1 si el ITF es menor al ingreso necesario y 0 en caso contrario. El total de pobres para el año 2005 en el NOA fue de 4.196, que representa un 45.23% de la muestra, y para el año 2025 fue de 3.967, que representa un 45.62% de la muestra. Esto puede observarse en la tabla 3.

Tabla 3

ano4	pobres	total de observaciones	porcentaje
2005.0	4196	9278	45.23
2025.0	3967	8695	45.62

Nota. Tabla donde se muestra la cantidad total de "pobres" encontrada en cada año, el total de observaciones y el porcentaje de muestra que representa.

9.

En primer lugar, realizamos estadísticas descriptivas relevantes de "pobre" comparando 2005 y 2025. En la Tabla 4, podemos ver la cantidad total de observaciones en cada año (9278 en 2005 y 8695 en 2025), la tasa de pobreza estimada (media de la variable binaria), que es de 45,2 % en 2005 y de 45,6 % en 2025, lo que muestra estabilidad en el tiempo. El desvío estándar cercano a 0,5 en ambos años confirma una distribución equilibrada entre pobres y no pobres, sin grandes variaciones temporales. La mediana y los cuartiles indican que la distribución de la variable está fuertemente concentrada en los valores extremos (0 o 1), lo cual es esperable dado su carácter dicotómico. Por ejemplo, la mediana toma valor 0 en ambos años, lo cual refleja que la mayoría relativa de la población no se encuentra en situación de pobreza (55%). Por último, en las filas "min" y "max", podemos ver los valores mínimos y máximos (0 y 1) que confirman la coexistencia de hogares pobres y

no pobres. En síntesis, los resultados sugieren que, pese a transformaciones en la estructura demográfica y económica entre 2005 y 2025, la incidencia de la pobreza sobre el total de la población analizada se mantuvo relativamente constante.

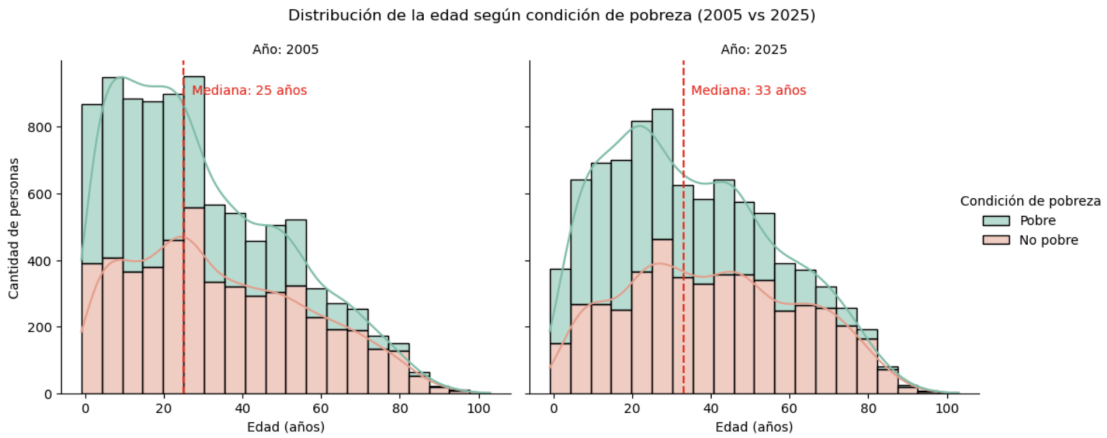
Tabla 4

ano4	2005.0	2025.0
count	9278.000000	8695.000000
mean	0.452253	0.456239
std	0.497742	0.498110
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	1.000000	1.000000
max	1.000000	1.000000

Nota. Tabla de estadísticas descriptiva de la variable “pobre” por año.

Luego, realizamos dos gráficos exploratorios usando la variable “pobre”. Por un lado, en la Figura 5 podemos ver la distribución de la edad según condición de pobreza donde se evidencia un cambio en la estructura etaria de la población entre 2005 y 2025. En 2005, la mediana de edad se ubicaba en 25 años, reflejando que la pobreza afectaba en mayor medida a personas jóvenes, en particular a quienes se encontraban en edades tempranas de inserción laboral. Mientras que para 2025, la mediana ascendió a 33 años, lo que indica un envejecimiento relativo de la población pobre y una mayor presencia de adultos en situación de vulnerabilidad. Este desplazamiento sugiere que, si bien la pobreza continúa siendo un fenómeno transversal a distintas edades, con el tiempo ha tendido a consolidarse también en grupos de mayor edad, lo que puede estar asociado a trayectorias laborales inestables y a limitaciones en la movilidad social. Asimismo, la superposición entre pobres y no pobres es más marcada en 2025, lo que evidencia una atenuación de las diferencias etarias.

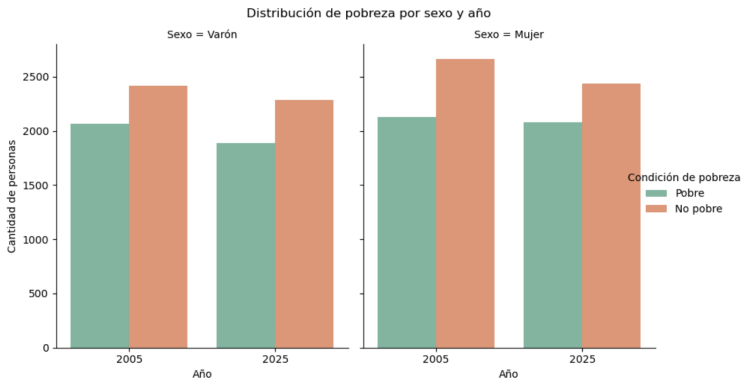
Figura 5



Nota. La figura muestra un análisis de la distribución de la edad según condición de pobreza comprando los años 2005 y 2025.

Por otro lado, la Figura 6 muestra un grafico de barras que representa la distribución de pobres y no pobres en los años 2005 y 2025, diferenciados según sexo. En el caso de las mujeres, la mayor parte no son pobres en ambos años, aunque hay una proporción considerable de pobreza. En los hombres, podemos ver que en ambos años ocurre la misma tendencia, la cantidad de no pobres supera a la de los pobres. Comparando ambos años, los valores son relativamente estables para ambos sexos, con una pequeña disminución en la cantidad de hogares pobres de varones. Sin embargo, la brecha entre ambos grupos se reduce en 2025, lo que indica un incremento relativo de la pobreza. Por lo tanto, independientemente del sexo, la proporción de pobreza aumenta en términos relativos en 2025 respecto de 2005, reflejando un deterioro en las condiciones socioeconómicas en el país.

Figura 6



Nota. El gráfico presenta la distribución de pobreza según el sexo de la población en los años 2005 y 2025.

Bibliografía

Instituto Nacional de Estadística y Censos de la República Argentina. (s. f.). *Página principal*. INDEC. https://www.indec.gob.ar/

Anexo

Link al GitHub: https://github.com/isanicola/CC408-Grupo-T3-8.git

Diccionario para la base de datos 2005
https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_disenoreg_T1_2005.pdf

Diccionario para la base de datos 2025
https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_registro_1T2025.pdf