TRABAJO PRÁCTICO N° 1

UN PRIMER ENCUENTRO CON LA EPH

Martin Cargnel

Link al repositorio: https://github.com/mcargnel/Big-Data-UBA-Grupo-001/tree/main/TP1

1 Familiarizandonos con la base EPH y limpieza

1.1 Breve introducción a la base

En este trabajo se utilizará la encuesta permanente de hogares para hacer una limpieza de datos y sentar las bases para un estudio sobre las personas pobres. La identificación de personas pobres en la EPH se realiza comparando el ingreso total familiar (ITF) del hogar con el valor de la Canasta Básica Total (CBT) correspondiente al período y ajustada por el número de adultos equivalentes en el hogar. Si el ITF es menor al ingreso necesario para cubrir la CBT según la composición del hogar, se considera que las personas de ese hogar son pobres. Este método permite tener en cuenta tanto el tamaño como la estructura etaria y de género del hogar, ajustando el umbral de pobreza a las necesidades reales de consumo.

Para ello, se utilizarán los primeros trimestres de 2005 y 2025 para la región de Gran Buenos Aires.

Uniendo estos períodos se obtiene un dataset de 16.665 filas, donde cada una corresponde a un individuo. Sin embargo, a continuación se procederá a realizar una limpieza de estos datos.

Las variables utilizadas en el análisis son las siguientes: **CODUSU**: Código que permite identificar de manera única cada vivienda, **nro_hogar**: Código que distingue a los diferentes hogares dentro de una misma vivienda, **componente**: Número de orden asignado a cada persona dentro del hogar, **ch04**: Sexo de la persona, **ch06**: Edad de la persona, **ch07**: Estado civil, **ch08**: Tipo de cobertura de salud, **nivel_ed**: Nivel educativo alcanzado, **estado**: Condición de actividad (ocupado, desocupado, inactivo, etc.), **cat_inac**: Categoría de inactividad, en caso de no estar ocupado, **ipcf**: Ingreso per cápita familiar, **pp04a**: Rama de actividad principal en la que trabaja la persona, **pp05b2_ano**: Años de antigüedad en la ocupación actual, **pp07c**: Indica si el empleo tiene fecha de finalización, **pp07g_59**: Señala si el empleo carece de beneficios como vacaciones pagas, aguinaldo, días pagos por enfermedad u obra social, **itf**:

Monto del ingreso total familiar, **pp07j**: Turno de trabajo (día, noche, otros) y **pp11o**: Motivo por el cual la persona dejó su trabajo anterior.

Cabe aclarar que, adicionalmente, se realizó una limpieza de las variables para que tengan nombres consistentes entre los dos períodos. La principal diferencia fue que en la tabla de 2025 las respuestas están encodeadas como números, mientras que en la de 2005 las respuestas están tipeadas. Por lo tanto, para mantener la legibilidad, se optó por reemplazar los números por respuestas tipeadas.

Luego de limpiar y filtrar los datos, se puede ver en la Tabla 1 y en la Figura 1 cuantos datos faltantes hay en cada una de las variables incluidas.

Tabla 1: Tabla con valores nulos por variable

ano4	2005	2025
codusu	0	0
nro_hogar	0	0
componente	0	0
ch04	0	0
ch06	0	0
ch07	0	0
ch08	0	0
nivel_ed	0	0
estado	0	0
cat_inac	0	0
ipcf	0	0
pp04a	0	3741
pp05b2_ano	0	3741
pp07c	0	3741
pp07g_59	0	3741
itf	0	0
pp07j	0	3741
pp11o	0	6818

2 Primer Análisis Exploratorio

En la Figura 2 se puede ver la composición de observaciones por género y año. Si bien la proporción se mantiene similar entre los dos años (mujeres ~52% y varones ~48%), la cantidad total de observaciones es menor en 2025 que en 2005.

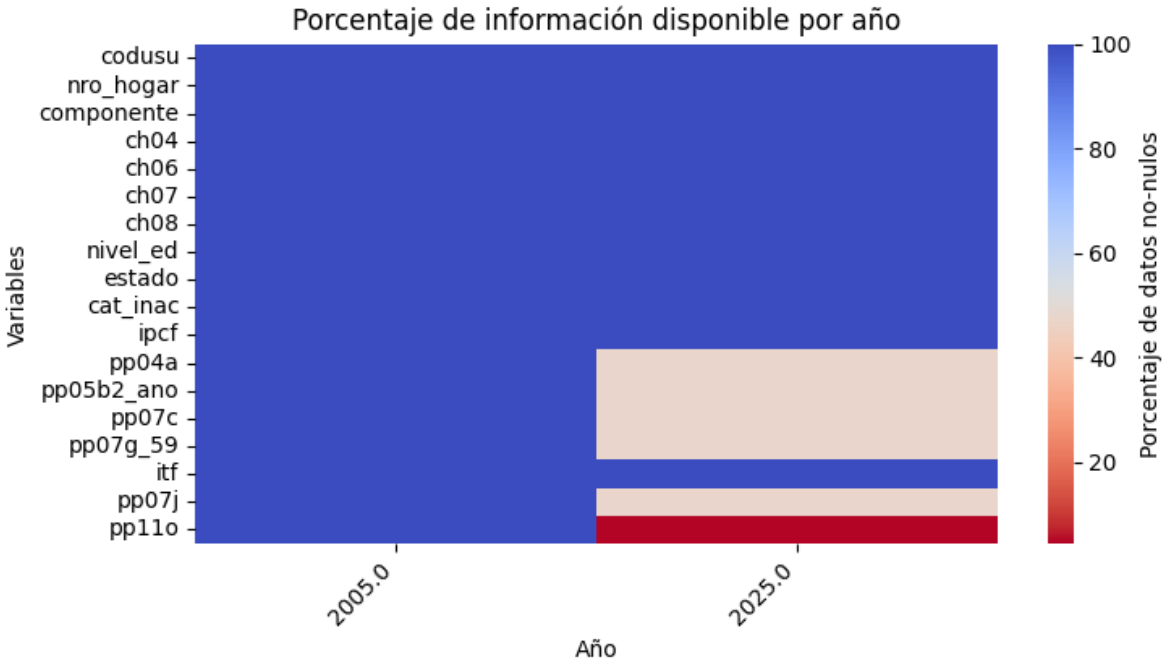

Figura 1

2.1 Matriz de correlación

En la Figura 3 se pueden ver las correlaciones entre algunas de las variables incluidas. En particular, se destaca:

- Alta correlación positiva entre "sin instrucción" y "menor de 6 años".
- Alta correlación negativa entre estar soltero y la edad.
- Alta correlación negativa entre no pagar una cobertura de salud y tener obra social (que incluye PAMI).

El único punto que no parece intuitivo sería el último, pero puede explicarse porque PAMI está incluido. El resto de las variables presenta una correlación relativamente baja, lo cual es positivo.

3 Conociendo a los pobres y no pobres

Según lo que se vio en esta base, en 2005, 30 personas no respondieron el estado, mientras que en 2025 ese número disminuyó a 21. Sin embargo, si se tiene en cuenta la respuesta a los ingresos, este número aumenta a 2.857 en 2025 y a 113 en 2005.

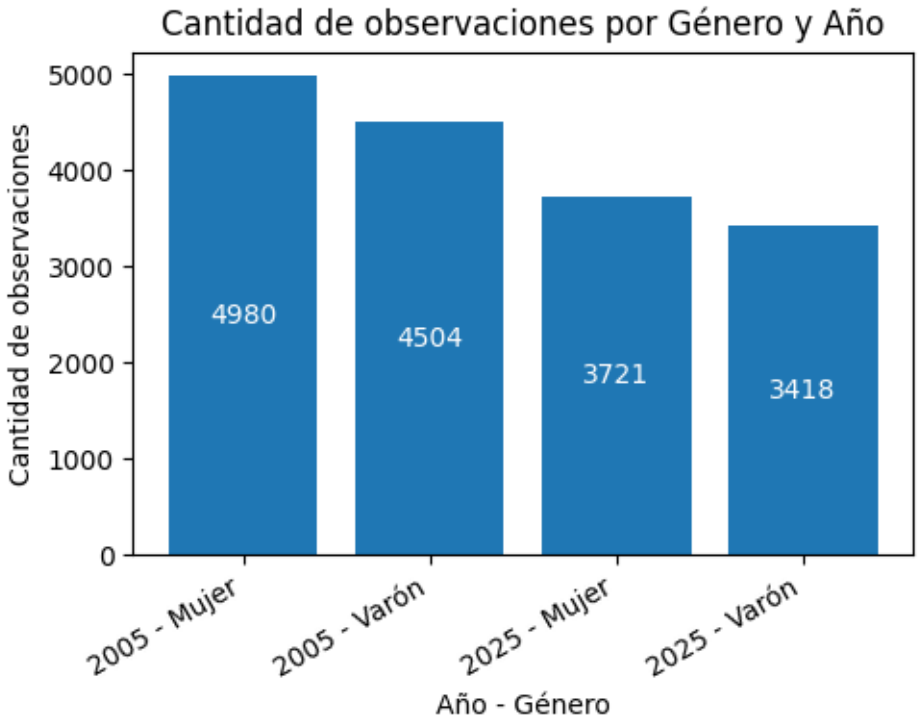
Cantidad de observaciones por Género y Año
Cantidad de observaciones
5000
4000
3000
2000
1000
0
4980
4504
3721
3418
2005 - Mujer
2005 - Varón
2025 - Mujer
2025 - Varón
Año - Género

Figura 2

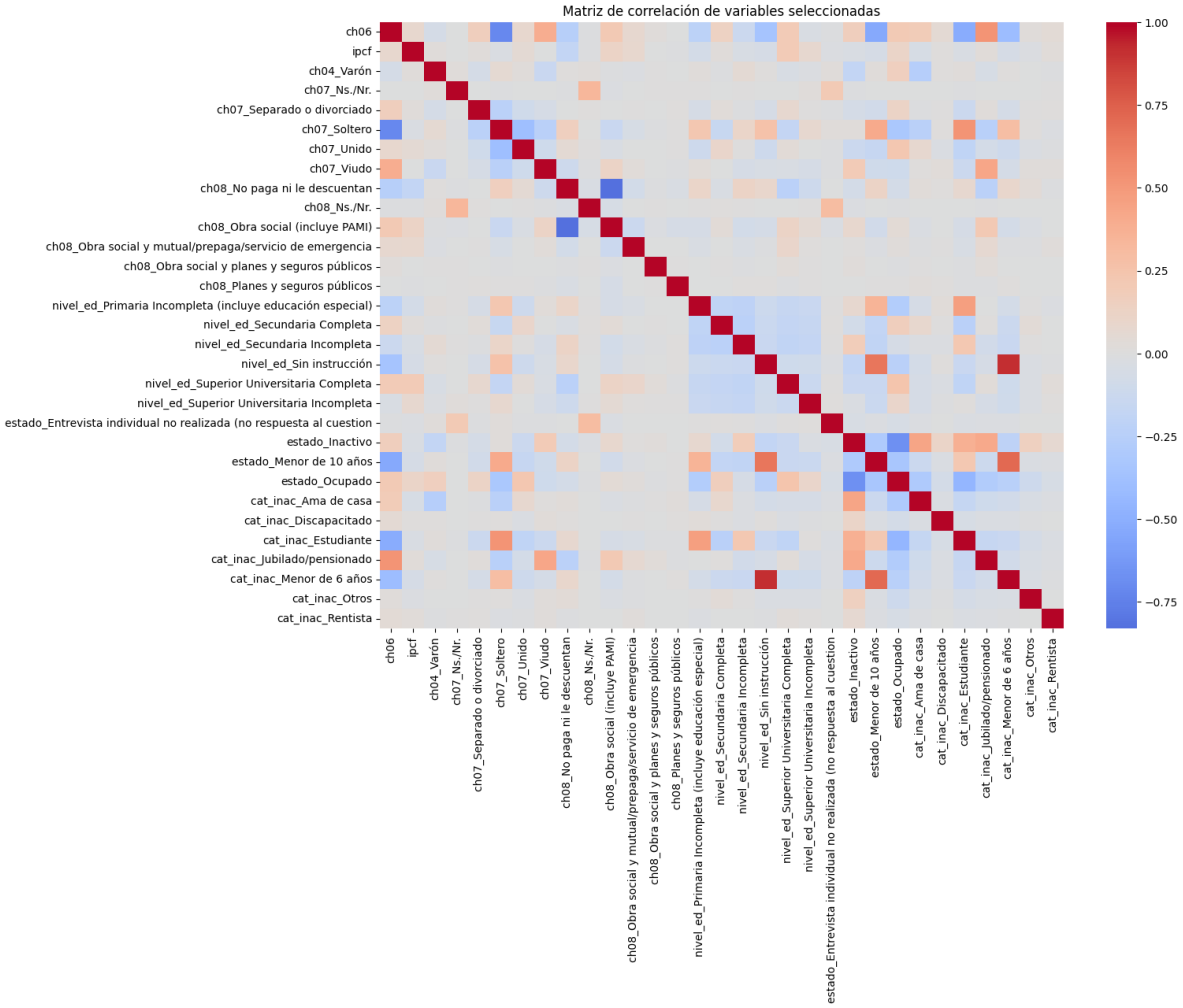
Matriz de correlación de variables seleccionadas
ch06
ipcf
ch04_Varón
ch07_Ns./Nr.
ch07_Separado o divorciado
ch07_Soltero
ch07_Unido
ch07_Viudo
ch08_No paga ni le descuentan
ch08_Ns./Nr.
ch08_Obra social (incluye PAMI)
ch08_Obra social y mutual/prepaga/servicio de emergencia
ch08_Obra social y planes y seguros públicos
ch08_Planes y seguros públicos
nivel_ed_Primaria Incompleta (incluye educación especial)
nivel_ed_Secundaria Completa
nivel_ed_Secundaria Incompleta
nivel_ed_Sin instrucción
nivel_ed_Superior Universitaria Completa
nivel_ed_Superior Universitaria Incompleta
estado_Entrevista individual no realizada (no respuesta al cuestion
estado_Inactivo
estado_Menor de 10 años
estado_Ocupado
cat_inac_Ama de casa
cat_inac_Discapacitado
cat_inac_Estudiante
cat_inac_Jubilado/pensionado
cat_inac_Menor de 6 años
cat_inac_Otros
cat_inac_Rentista
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Figura 3

Para continuar con el análisis se incluyó a la base una variable que tiene los valores de adulto equivalente de cada persona según su sexo y edad. Dicha variable se agregó a nivel hogar y se guardó en una columna llamada ad_equiv_hogar. Luego, se multiplicó por la canasta básica para cada año (2005: $205,07 y 2025: $365.177)y comparó contra el ITF con el objectivo de clasificar como pobres a quienes tengan un ingreso menor a esta métrica.

Por último, puede verse en la @tab-pobres la cantidad y proporción de pobres en ambos períodos. En 2005, el 26,6% de las personas se encontraba bajo la línea de pobreza, mientras que en 2025 este porcentaje aumenta al 31,1%. Esto indica un incremento en la proporción de personas pobres en el Gran Buenos Aires entre ambos períodos. Además, la cantidad total de observaciones es menor en 2025, lo que puede estar relacionado con la mayor cantidad de personas que no respondieron el ingreso total familiar en ese año.

Tabla 2: Tabla con la cantidad y proporción de pobres por año

	Año	Pobre	Cantidad	Proporción
0	2005.0	0	6878	0.7340
2	2005.0	1	2493	0.2660
1	2025.0	0	2950	0.6889
3	2025.0	1	1332	0.3111