



Ciencia de Datos

5 de Septiembre de 2025

Trabajo Práctico 1

Docentes: María Noelia Romero, Tomás Buscaglia

Alumnos: Ignacio Matas , Tomás Weissmann y Max Wroclavsky

Link al repositorio: <https://github.com/mwroclavsky/CC408-Grupo-T3-13>

1

A partir de los ingresos de los hogares se establece si éstos tienen capacidad de satisfacer -por medio de la compra de bienes y servicios- un conjunto de necesidades alimentarias y no alimentarias consideradas esenciales. El procedimiento parte de utilizar una canasta básica de alimentos (CBA) y ampliarla con la inclusión de bienes y servicios no alimentarios (vestimenta, transporte, educación, salud, etc.) con el fin de obtener el valor de la canasta básica total (CBT).

Para calcular la incidencia de la pobreza se analiza la proporción de hogares cuyo ingreso no supera el valor de la CBT; para el caso de la indigencia, la proporción cuyo ingreso no supera la CBA.

2A

----- 2005 -----

Filas totales: 47030 | Filas Patagonia: 3229

Columnas totales: 176

Columnas después del filtro: 176

----- 2025 -----

Filas totales: 45425 | Filas Patagonia: 5359

Columnas totales: 235

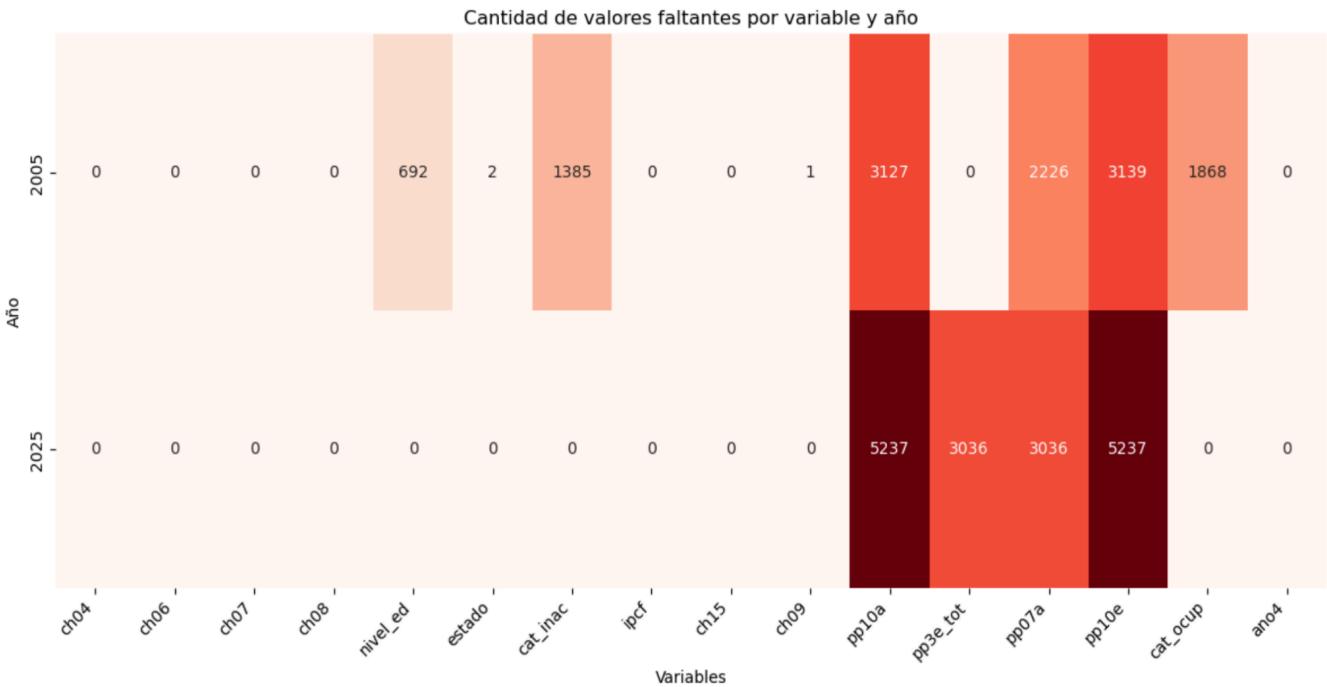
Columnas después del filtro: 235

Se filtraron las bases de 2005 y 2025 para quedarnos solo con la región Patagónica. Así armamos las dos submuestras con las que vamos a trabajar

2B

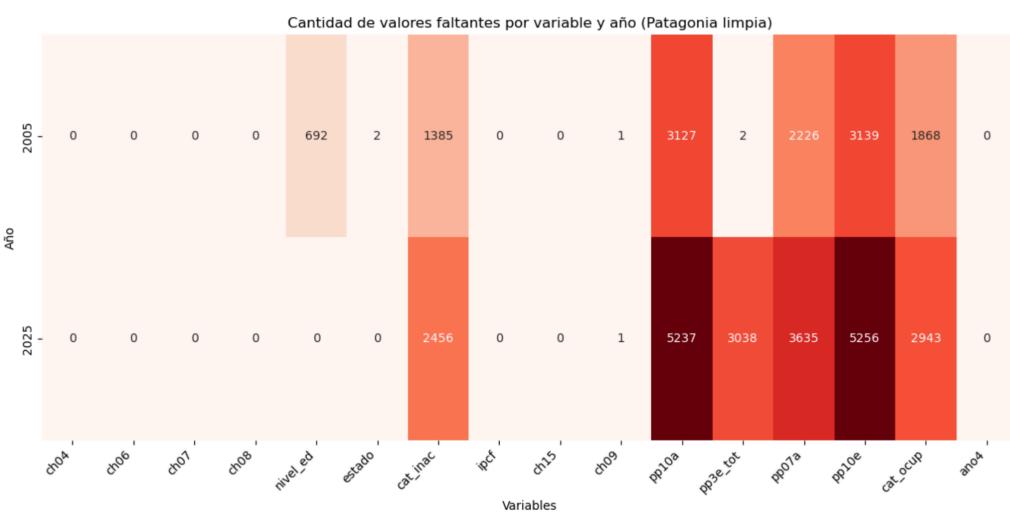
	2005	2025
ch04	0	0
ch06	0	0
ch07	0	0
ch08	0	0
nivel_ed	692	0
estado	2	0
cat_inac	1385	0
ipcf	0	0
ch15	0	0
ch09	1	0
pp10a	3127	5237
pp3e_tot	0	3036
pp07a	2226	3036
pp10e	3139	5237
cat_ocup	1868	0
ano4	0	0

Se analizó la cantidad de valores faltantes en cada variable y se mostró en un heatmap. Esto permite ver en qué variables se concentran los problemas de no respuesta en cada año.



En la comparación de 2005 y 2025, se observa que en 2005 las variables con más valores faltantes son NIVEL_ED (692), CAT_INAC (1385), CAT_OCUP (1868), PP07A (2226), PP10A (3127) y PP10E (3139), sumando un total de 11.440 casos faltantes en toda la base. En 2025, las variables con mayor cantidad de valores perdidos son PP10A (5237), PP10E (5237), PP07A (3036) y PP3E_TOT (3036), junto con ESTADO (2), lo que representa un total de 16.548 casos faltantes. El resto de las variables presentan muy pocos o ningún valor perdido. Esto sugiere que la calidad de los datos es mayor en las variables básicas (como sexo o edad), mientras que la información más sensible o vinculada a condiciones laborales y de búsqueda de empleo presenta mayores problemas de no respuesta.

2C



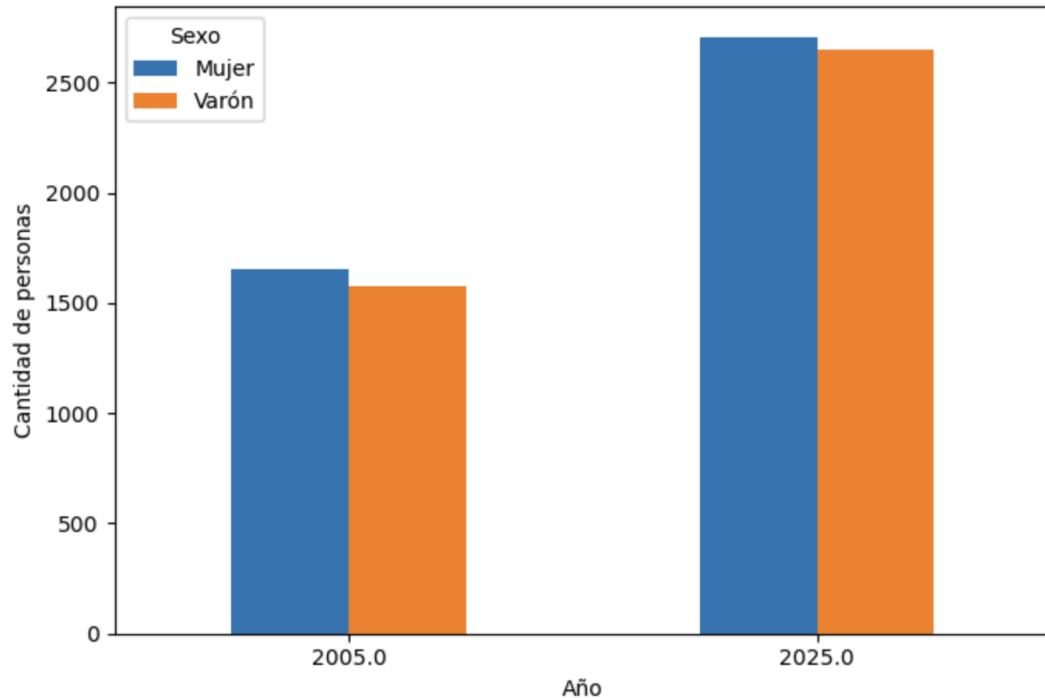
Durante la etapa de limpieza se corrigieron valores fuera de rango en las variables seleccionadas. En particular, se reemplazaron como faltantes los ingresos negativos en ipcfc y las horas semanales de trabajo menores a 0 o mayores a 100 en pp3e_tot. Asimismo, se recodificaron como NaN aquellas categorías que aparecían fuera de los valores válidos definidos por la codificación de la EPH (por ejemplo, códigos no reconocidos en estado civil, cobertura médica o nivel educativo). Este proceso asegura que los análisis posteriores se realicen únicamente con datos consistentes.

3

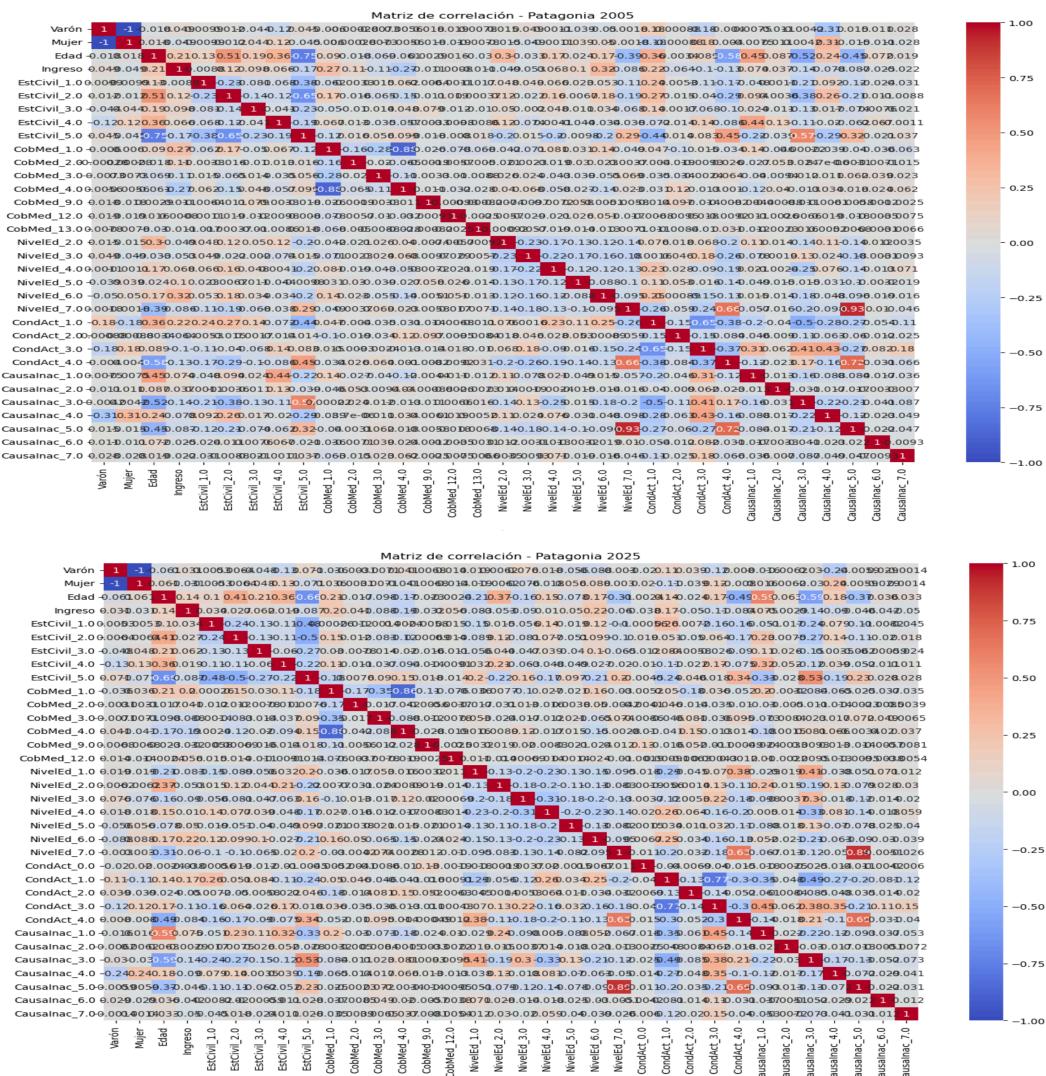
Tabla de composición por sexo en 2005 y 2025:

Sexo	Mujer	Varón
ano4		
2005.0	1654	1575
2025.0	2709	2650

Composición por sexo en 2005 y 2025 - Región 44



4



La matriz de correlación para Patagonia en 2005 y 2025 muestra cómo se relacionan las variables demográficas y laborales. Se observa una correlación positiva entre el nivel educativo y los ingresos, y entre la condición de actividad y el ingreso, lo que indica que personas con mayor educación y en actividad tienden a tener mejores ingresos. El sexo muestra baja correlación con el ingreso, lo que sugiere brecha moderada. Ciertas causas de inactividad y algunas coberturas médicas presentan correlaciones negativas con el empleo y los ingresos, evidenciando exclusión en ciertos grupos. Los patrones se mantienen estables entre 2005 y 2025, aunque puede haber variaciones menores por cambios sociales y económicos. El análisis sintético permite visualizar las conexiones entre educación, mercado laboral y estructura social en la región.

```

No respondieron (total): 13
No respondieron por año:
ano4
2005.0      2
2025.0     11
Name: estado, dtype: int64
respondieron (ITF>0): 7363 | 2005: 3206 | 2025: 4157
no respondieron (ITF=0): 1225 | 2005: 23 | 2025: 1202

```

La tabla diferencia entre quienes respondieron y quienes no informaron sus ingresos familiares. Sirve para ver la magnitud del problema de no respuesta y comparar entre 2005 y 2025

6

	ano4	ch06	ch04	adulto_equiv	ad_equiv_hogar
0	2005.0	46.0	1.0	1.00	937.73
1	2005.0	32.0	2.0	NaN	937.73
2	2005.0	14.0	1.0	0.96	937.73
3	2005.0	9.0	1.0	0.69	937.73
4	2005.0	3.0	2.0	0.51	937.73

```

Totales de adultos equivalentes por año:
ano4
2005.0      937.73
2025.0     1211.10
Name: adulto_equiv, dtype: float64

```

La tabla suma el equivalente adulto según sexo y edad. Así se observa cuántos “adultos equivalentes” hay en total en la región para cada año

7

```

==== CONSIGNA 7: inicio ====
Filas respondieron (antes): 7363
Agrupando patagonia_equiv por (ano4, ipcfc) para sumar adulto_equiv...
Filas en tabla ad_equiv_proxy: 1034
   ano4          ipcfc  ad_equiv_hogar
0  2005.0       0.000000      4.33
1  2005.0      14.400000      2.42
2  2005.0     16.666667      0.74
3  2005.0     18.750000      4.77
4  2005.0    20.000000      0.00

Haciendo merge a respondieron por (ano4, ipcfc)...
Columnas nuevas agregadas: []
Filas respondieron (después del merge): 7363
Primeras filas con ad_equiv_hogar:
   ano4          ipcfc  ad_equiv_hogar
0  2005.0      480.00        7.21
1  2005.0      480.00        7.21
2  2005.0      480.00        7.21
3  2005.0      480.00        7.21
4  2005.0      480.00        7.21

Hogares sin ad_equiv_hogar (NaN): 0

== Resumen ad_equiv_hogar por año ==
ad_equiv_hogar  count  nunique   min    median      mean      max
ano4
2005.0           3206.0      221.0   0.0     3.22    5.108026   22.05
2025.0           4157.0      210.0   0.0     2.79    5.367724   25.03

== Resumen ingreso_necesario por año ==
ingreso_necesario  count  nunique   min    median      mean      max
ano4
2005.0           3206.0      221.0   0.0   6.603254e+02  1.047503e+03
2025.0           4157.0      209.0   0.0  1.018844e+06  1.960169e+06

ingreso_necesario      max
ano4
2005.0      4.521793e+03
2025.0      9.140380e+06

Muestra aleatoria de 5 filas:
   ano4          ipcfc  ad_equiv_hogar  ingreso_necesario
3162  2005.0  2.50000e+02      15.42    3.162179e+03
6270  2025.0  1.466667e+06      0.00    0.000000e+00
4821  2005.0      3.10000e+05      0.93    3.100000e+05
3149  2005.0  2.642857e+02      2.92    5.988044e+02
6528  2025.0  3.50000e+05      3.95    1.442449e+06
==== CONSIGNA 7: fin ====

```

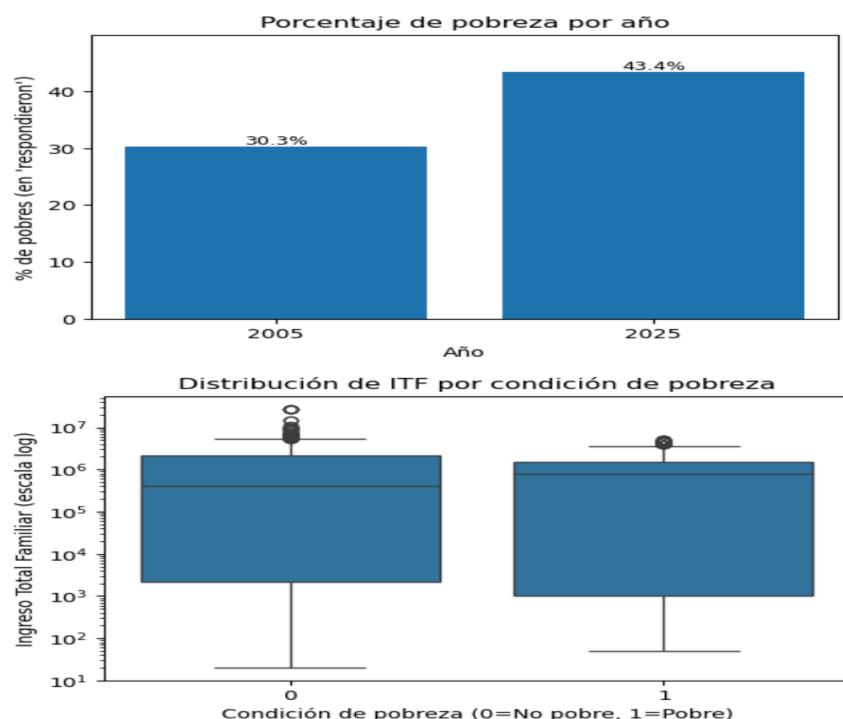
La tabla incorpora la columna ingreso_necesario, que estima cuánto debería tener cada hogar según sus adultos equivalentes para no ser pobre. Se calcula multiplicando la canasta básica de cada año (2005: \$205,07 y 2025: \$365.177) por el total de adultos equivalentes del hogar.

8

	ano4	pobres	n	porcentaje_muestra
0	2005.0	971	3206	30.286962
1	2025.0	1805	4157	43.420736

La tabla agrega la columna pobre, que marca con 1 a los hogares cuyo ingreso total familiar es menor al ingreso necesario y con 0 al resto. Esto permite contar cuántos hogares pobres hay en 2005 y 2025 y qué porcentaje representan dentro de la muestra.

9



Se elaboraron estadísticas descriptivas y dos gráficos exploratorios de la variable “pobre”:

1. Gráfico de barras (% de pobreza por año):
 - Compara visualmente la incidencia de la pobreza en 2005 y 2025.
 - Facilita observar si la tasa de pobreza aumentó o disminuyó entre ambos períodos.
2. Boxplot del ITF según condición de pobreza:
 - Presenta la distribución del ingreso total familiar (ITF) en escala logarítmica, diferenciando entre hogares pobres (1) y no pobres (0).
 - Se observa que los hogares pobres concentran ingresos mucho más bajos y con menor dispersión, mientras que los no pobres tienen ingresos más altos y heterogéneos.