Análisis de la EPH: T1 2005 y 2025

Ciencia de Datos

Grupo: 16

Maia Findor - Mora Dal Lago Carabajal

TP1

Entrega: 5/9/2025

Parte 1:

1) Según el INDEC, las personas consideradas pobres son las personas que están por debajo de la línea de pobreza (LP). Por esta razón, los datos sobre la pobreza se basan en la EPH, ya que para delimitar si alguien está o no por debajo de la lìnea de pobreza, se calculan los ingresos de cada hogar y se establece si es que los hogares tienen o no la capacidad para poder comprar bienes y servicios esenciales. Para esto se utiliza la canasta básica de alimentos (CBA) la cual establece un umbral mínimo a cubrir de necesidades energéticas y proteicas, y se amplía 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). Los hogares que no tengan los ingresos suficientes para cubrir la canasta básica total serán considerados pobres.

2) Reducción y limpieza de la base:

   a) Para el análisis de la EPH se utilizaron datos correspondientes a la región Patagónica. En este sentido, se unificaron los formatos en los que se presentaban los datos de ambas bases y se eliminaron los registros pertenecientes a otras regiones.

   b) La variable con más datos faltantes es la variable PP03G en el año 2025

Figura 1

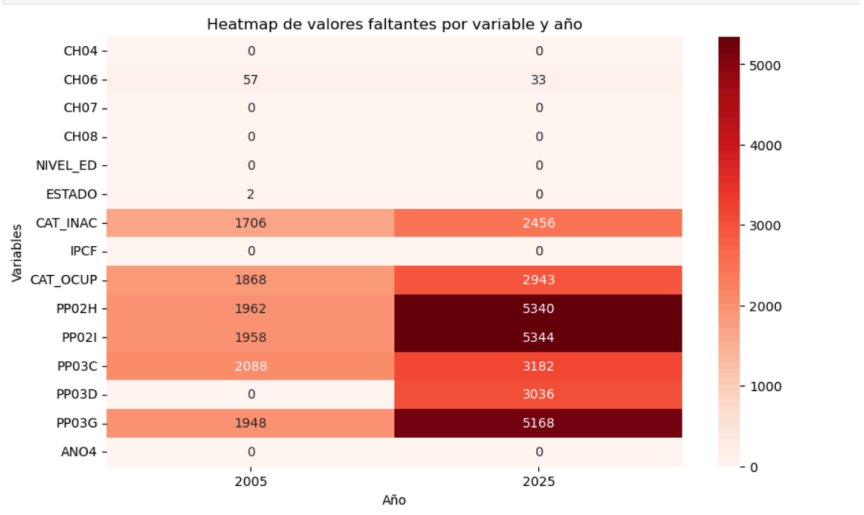


Figura 2

Variable	2005	2025	Total
CH04	0	0	0
CH06	57	33	90
CH07	0	0	0
CH08	0	0	0
NIVEL_ED	0	0	0
ESTADO	2	0	0
CAT_INAC	1706	2456	4162
IPFC	0	0	0
CAT_OCUP	1868	2943	4811
PP02H	1962	5340	7302
PP02I	1958	5344	7302
PP03C	2088	3182	5270
PP03D	0	3036	3036
PP03G	1948	5168	7116
ANO4	0	0	0
TOTAL	11589	27502	39091

Figura 1: El heatmap muestra la cantidad de valores faltantes por variable y año en las bases de la EPH correspondientes a 2005 y 2025. Se observa que la mayoría de las

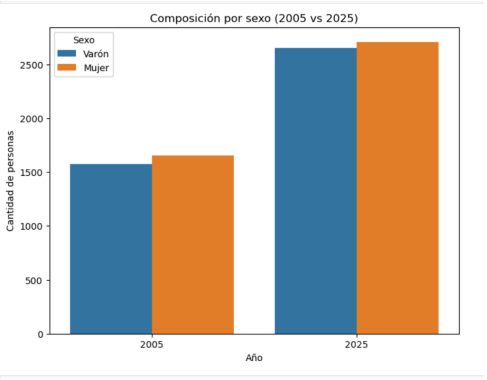
variables no presentan datos perdidos (como CH04, CH07, CH08, NIVEL_ED, IPCF o ANO4), mientras que en 2005 los faltantes son mínimos en variables puntuales como CH06 (57 casos) o ESTADO (2 casos). Sin embargo, tanto en 2005 como en 2025 los valores faltantes se concentran en las variables vinculadas a la condición de actividad y ocupación (CAT_INAC, CAT_OCUP, PP02H, PP02I, PP03C, PP03D y PP03G). En particular, en 2025 se registra un incremento considerable en la magnitud de los faltantes, destacándose PP02H, PP02I y PP03G con más de 5.000 casos cada una. En síntesis, el gráfico evidencia que las pérdidas de información se concentran en variables laborales y que su magnitud es notablemente mayor en 2025 respecto de 2005, lo cual constituye un aspecto a considerar en los análisis posteriores.

Figura 2: Tabla con los datos faltantes de cada variable por año. La variable con más datos faltantes es la variable PP03G en el año 2025. Esta segunda figura muestra la misma información que la Figura 1 pero puesta en formato de tabla.

c) Durante el proceso de limpieza de la base de datos se identificaron y codificaron valores raros o códigos especiales utilizados por la EPH para indicar no respuesta. Valores como 99, 999 o negativos (que corresponden a respuestas “no sabe/no responde”, o categorías sin sentido en ciertas variables) fueron transformados en valores faltantes (NaN). Asimismo, en variables binarias (las cuales inicialmente estaban codificadas como 1 = afirmativo, 2 =negativo) se reemplazaron los 0 (valores inválidos para la variable) por valores de no respuesta NA. Además, se estandarizaron los códigos para que tomaran el valor 1 en caso afirmativo y 0 en caso negativo. De esta manera, se homogeneizó la codificación y se evitó confundir respuestas válidas con valores que en realidad corresponden a no respuestas

Parte 2:

Figura 3:



3) Siguiendo por la región que filtramos, en 2005, tanto varones como mujeres presentan cantidades relativamente similares. En 2025, la cantidad de personas aumenta

considerablemente en ambos grupos, reflejando un mayor tamaño muestral. La composición por sexo sigue siendo equilibrada, con un ligero predominio femenino.

4)

Figura 4:

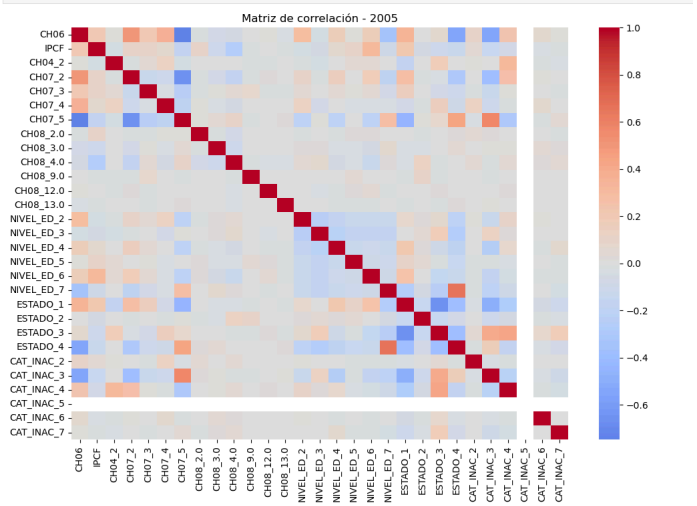


Figura 5:

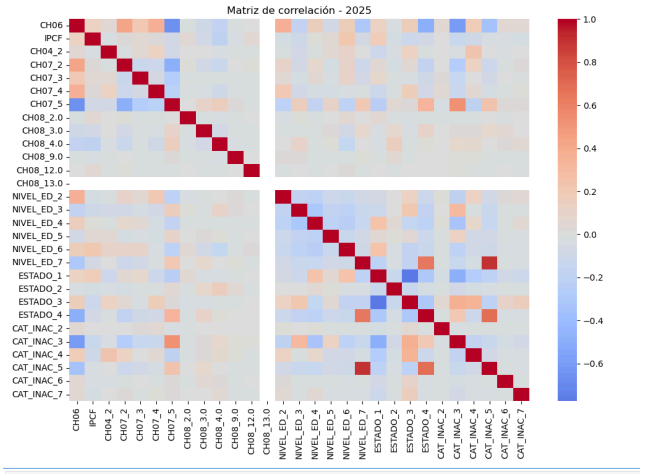


Figura 4: Se ven algunas relaciones esperables entre las variables. Por ejemplo, el estado civil (CH07 / ESTADO, donde 1 = Soltero/a, 2 = Casado/a, 3 = Separado/a o Divorciado/a, 4 = Viudo/a, 5 = Unión libre/conviviente) aparece relacionado con la inactividad (CAT_INAC, donde 1 = Jubilado/a o pensionado/a, 2 = Rentista, 3 = Estudiante, 4 = Ama de casa, 5 = Discapacitado/a permanente, 6 = Menor de 6 años, 7 = Otros inactivos), ya que ciertas situaciones de inactividad se asocian más a determinados estados civiles. A mayor nivel educativo (CH08 / NIVEL_ED, con categorías: 0 = Sin instrucción, 1 = Primaria incompleta, 2 = Primaria completa, 3 = Secundaria incompleta, 4 = Secundaria completa, 5 = Terciario/universitario incompleto, 6 = Terciario/universitario completo) hay menos chances de estar inactivo. Además, la edad (CH06, medida en años) se vincula con el estado civil, como la mayor probabilidad de estar casado en edades más avanzadas.

Figura 5: Se mantiene la relación negativa entre educación (CH08 / NIVEL_ED) e inactividad (CAT_INAC), lo que confirma que la educación funciona como un factor que reduce la probabilidad de inactividad laboral. Además, el ingreso per cápita familiar (IPCF, variable continua que surge de dividir el ingreso total del hogar por la cantidad de integrantes

equivalentes) se asocia de forma positiva con la educación y de forma negativa con la inactividad, lo cual es lógico: más educación suele estar vinculada a mejores ingresos y menor inactividad.

Ambas matrices muestran que los vínculos entre estas variables se mantienen estables en el tiempo. Las relaciones entre educación, inactividad e ingresos se repiten en los dos años, y lo mismo pasa con los vínculos entre edad y estado civil. Esto sugiere que, a pesar de los veinte años de diferencia, la estructura de estas relaciones no cambia demasiado.

Parte 3:

5)
Uno de los principales problemas de la EPH es la falta de respuesta en torno a los ingresos. Al separar las bases según la variable Ingreso Total Familiar (ITF), se observa que en 2005 46.592 personas (99,1%) respondieron su ingreso, mientras que sólo 438 personas (0,9%) no lo hicieron (casos con ITF = 0). En cambio, en 2025 la diferencia es mucho más marcada: 33.372 personas (73,5%) respondieron, pero 12.053 personas (26,5%) no reportaron sus ingresos.
De un año a otro, aumentó de manera considerable la falta de respuestas con respecto al ingreso. Mientras que en 2005 aproximadamente 1 de cada 100 no respondía, 20 años más tarde se convirtió en un 1 cada 4 que no responde (dentro de la región considerada)

7) En 2005, sobre un total de 47.030 personas, se clasificaron como pobres 16.820 personas, lo que representa el 35,8% de la muestra. En 2025, sobre un total de 45.425 personas, se clasificaron como pobres 31.515 personas, lo que equivale al 69,4% de la muestra.

8)

Figura 6:

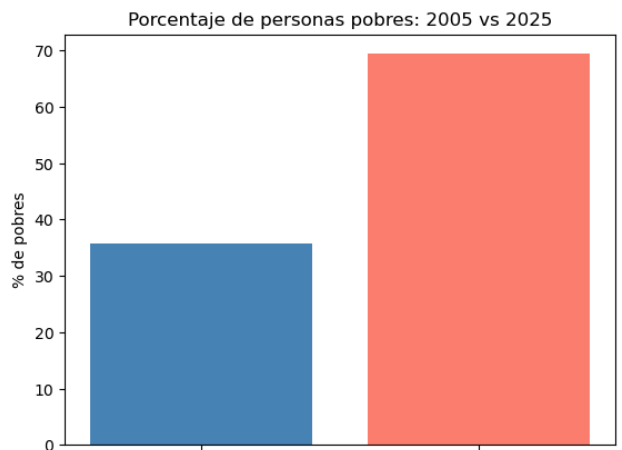


Figura 7:

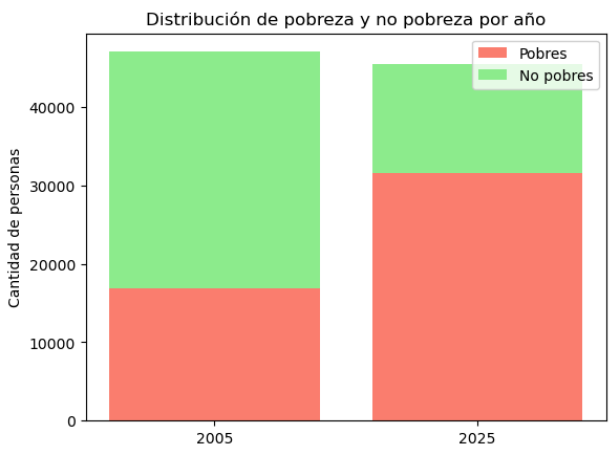


Figura 6: Muestra la comparación del porcentaje de personas pobres en la región analizada entre 2005 y 2025. En 2005, aproximadamente el 37% de la población se encontraba por debajo de la línea de pobreza, mientras que en 2025 este valor asciende a alrededor del 70%. Esto refleja un aumento muy significativo en la incidencia de la pobreza a lo largo de las dos décadas, lo que evidencia un deterioro en las condiciones socioeconómicas de la población.

Figura 7: Muestra la distribución de la población entre pobres y no pobres en 2005 y 2025. En 2005 predominaban los no pobres, que representaban la mayor parte de la población, mientras que los pobres eran una proporción menor. En cambio, en 2025 la cantidad de personas pobres supera a la de no pobres, reflejando un aumento considerable de la pobreza en comparación a 2005.