Universidad de San Andrés

Escuela de Educación

Licenciatura en Ciencias del Comportamiento

Ciencias de Datos - 1er Informe

Gastón Perez Raffo y Nicolás De la Rosa-Soiza

Profesora: María Noelia Romero

Buenos Aires, Argentina

5 de septiembre del 2025

https://github.com/tonga86/Ciencia-Datos---TP-1---Grupo-14

Parte I: Familiarización con la base EPH y limpieza.

1. El INDEC clasifica a las personas como pobres o no pobres, a partir de que tienen la capacidad de cubrir sus necesidades elementales. Esto se establece a medida que su hogar cuenta con suficiente ingreso para cubrir una Canasta Básica Total (CBT). Esta es, la canasta básica alimentaria, más algunos bienes y servicios no alimentarios (educación, salud, transporte, vestimenta, etc.)

2. Este informe se va a centrar en los resultados de la encuesta permanente de hogares (EPH) del Gran Buenos Aires (GBA) del 2005 y 2025, tanto del cuestionario individual como el de hogar. Las variables tratadas van a ser 12 sobre características de los individuos y 3 sobre características de la vivienda.

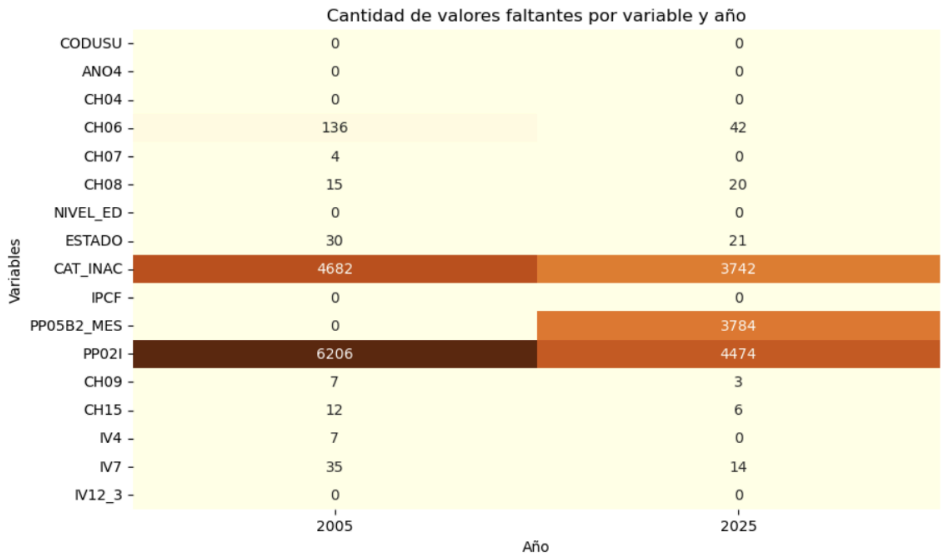
   Las variables individuales son: el sexo (CH04), edad (CH06), estado civil actual (CH07), cobertura medica (CH08), nivel educativo (NIVEL_ED), condicion de actividad (ESTADO), categoria de inactividad (CAT_INAC), monto del ingreso per capita familiar (IPCF), cantidad ed meses eguidos de trabajo en el empleo actual (PP05B2_MES), estado de actividad en el ultimo año (PP02I), analfabetismo (CH09), y lugar de nacimiento (CH015)

   Por otro lado, se van a tener en cuenta las siguientes variables del hogar: material de la cubierta exterior del techo (IV4), qué fuente de agua tiene el hogar (IV7), y si la vivienda está ubicada en una villa de emergencia (IV12_3).

   Por último, a fin de unificar las bases del cuestionario de hogar y el individual, se va a mantener el código para distinguir viviendas (CODUSU), y para unificar las bases del 2005 y 2025 se va a mantener la variable año (ANO4) para poder distinguir cuando fue tomado cada dato.

   Ante un análisis inicial, se puede ver que hay 3784 datos faltantes solamente en la variables *PP05B2_MES* del año 2025. Sin embargo, la documentación de la EPH indica que las respuestas no sabe/no responde fueron codificadas con los números 9, 99, 999 y 9999. Así mismo, cuando las preguntas no aplicaban para el individuo encuestado se codificó con un 0. Tomando estas consideraciones, y recodificando estos valores como respuestas faltantes (NaN en python), la variable con mayores datos faltantes es *PP02I* con 6206 en el año 2005.

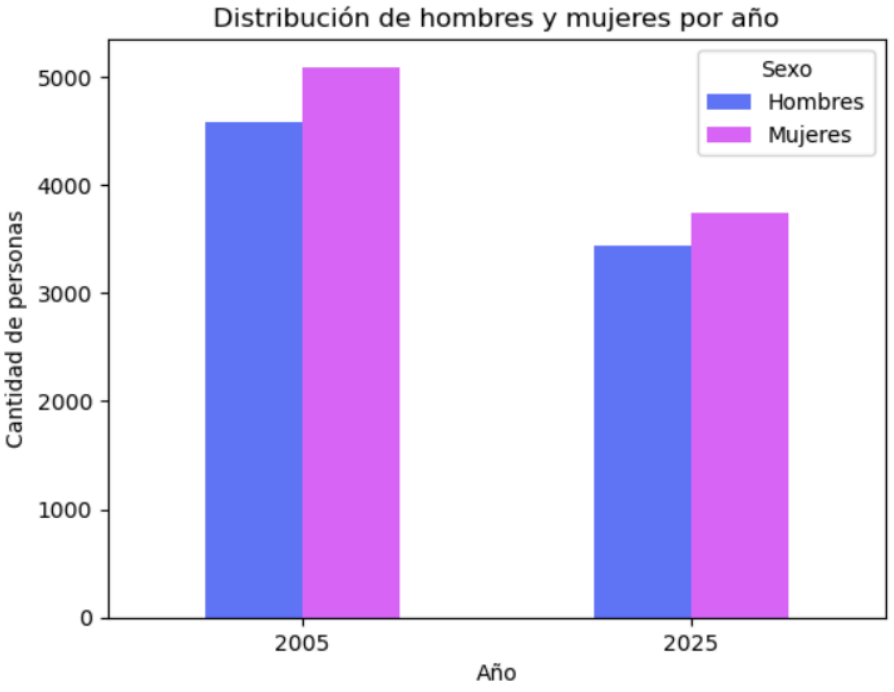
   Figura 1



Parte II: Análisis Exploratorio

3. Entre el 2005 y el 2025 se puede observar una proporción similar en la diferencia entre hombres y mujeres. En 2005 hubieron 4587 (47.42%) hombres y 5027 (52.58%) mujeres. En cambio, en 2025 hubieron 2055 (47.66%) de hombres y 5027 (52.34%) de mujeres.

Figura 2.1



4. A fin de estudiar la correlación entre variables, se realizaron 3 matrices de correlación en formato de heatmap, una por año y otra en conjunto. Se renombraron las variables para que puedan ser interpretadas mejor.

Figura 3.1

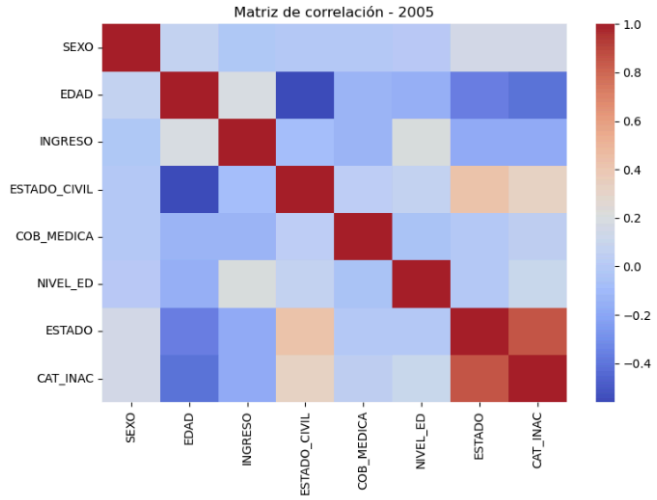


Figura 3.2

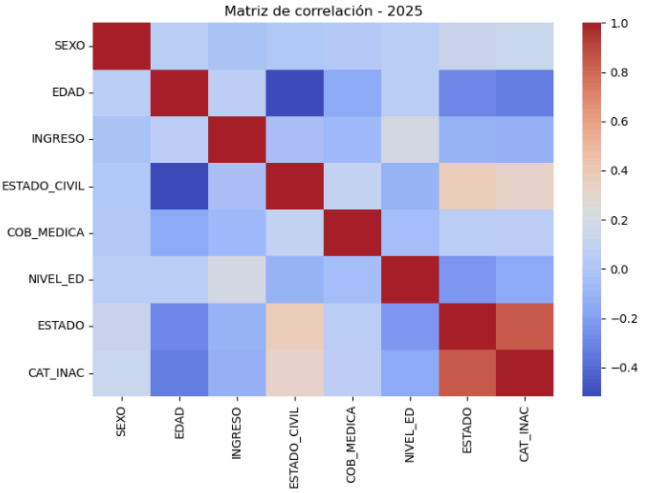
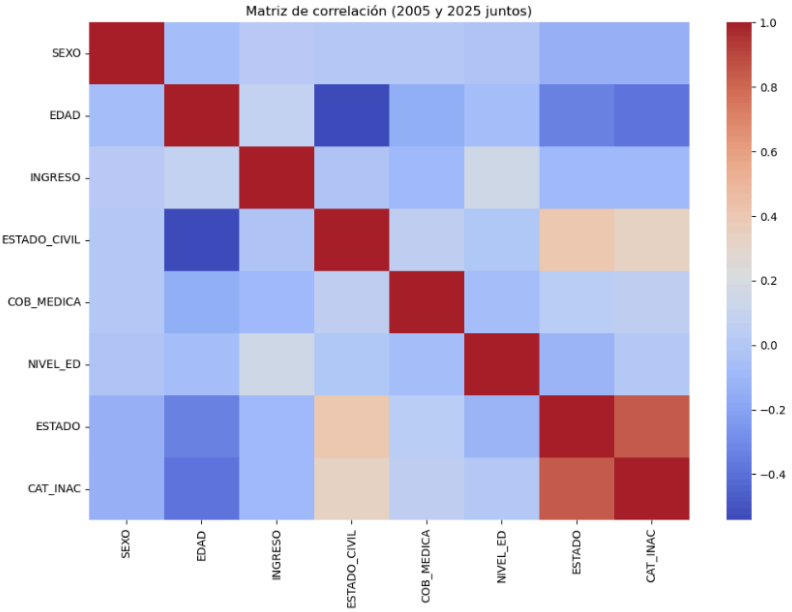


Figura 3.3

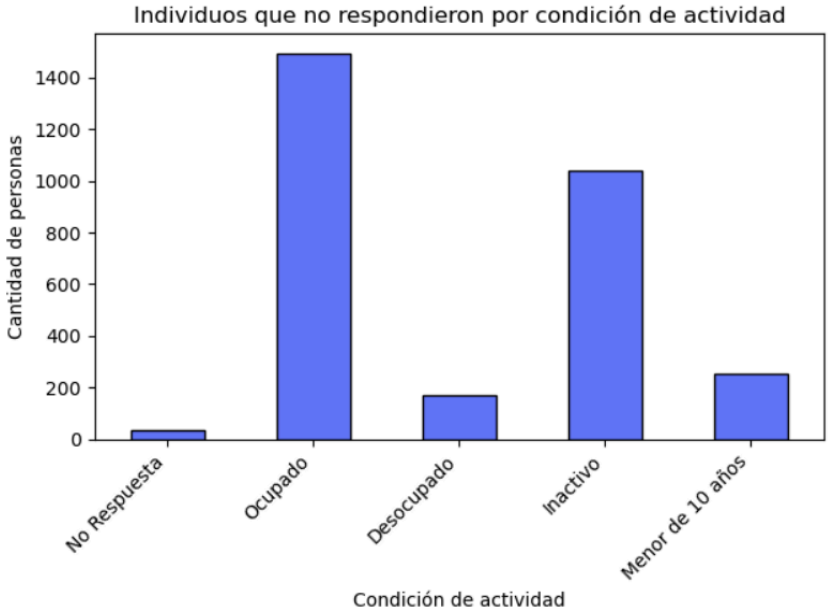


Parte III: Conociendo a los pobres y no pobres

5. Incluyendo la variable ingreso total familiar (ITF) en el análisis, se armaron dos bases de datos en base a los encuestados que respondieron y no respondieron esta pregunta; los que no respondieron fueron codificados con el valor 0 en la variable ITF. 2989 individuos no respondieron la pregunta sobre su ITF, mientras que 13873 sí.

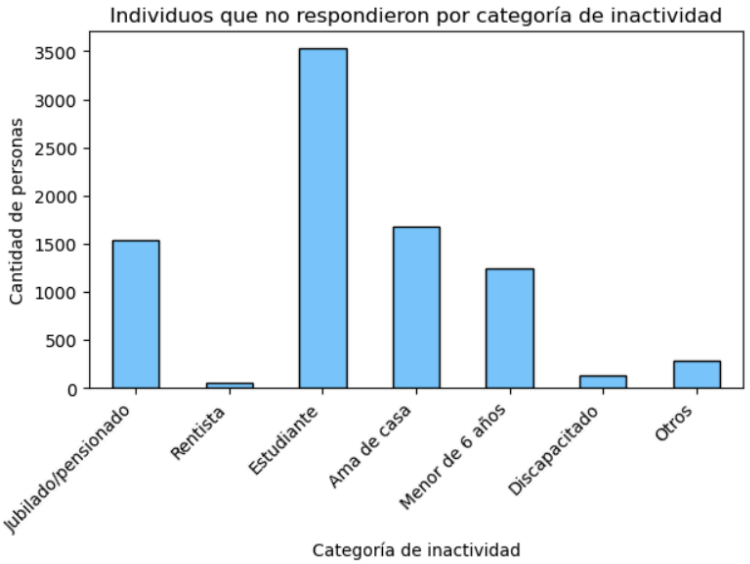
   Posteriormente, se realizó un histograma para visualizar la distribución de aquellos que no respondieron en las distintas condiciones de actividad. Puntualmente, se pudo observar que un 43.83% (7390 individuos) estaban ocupados.

   Figura 4.1



   A partir de los 6182 individuos (36.66%) que reportaron estar inactivos, se realizó otro histograma para analizar sus categorías de inactividad. En este caso, el 41.83% (3530 individuos) reportaron ser estudiantes.

   Figura 4.2



6. De manera subsecuente, en la base de datos de personas que respondieron la pregunta sobre su ITF, se creo una nueva variable con las necesidades energetias de cada individuo a partir de su edad y sexo llamada *respondieron_equiv*. Esto se logró gracias a la Tabla de equivalencias de necesidades energéticas del INDEC. Hecho esto, se agregó otra variable más para las necesidades energéticas de cada hogar, llamada *ad_equiv_hogar*. Para hacer esto, se usó la función group.by() a partir del código de vivienda, habiendo juntado los individuos de cada casa, se hizo una sumatoria de todas sus necesidades energéticas.

7. Seguido a esto, se creó otra variable más llamada *ingreso_necesario*. Esta variable representa el ingreso que una familia tiene que tener para no ser considerada pobre. Para obtener este dato se multiplicó la necesidad energética del hogar por la Canasta Básica Total (CBT) para un adulto equivalente en el primer trimestre de 2025 ($365177 aprox).

8. Habiendo hecho esto, se clasificó a cada individuo como pobre (1) o no pobre (0) en una nueva variable dummy llamada *pobre*. El criterio de clasificación fue si el ITF superaba o no el ingreso necesario del hogar. De esta manera, se encontró que en el 2005 de las 9561 personas encuestadas en el GBA, que reportaron su ITF, 2660 personas son pobres (27.82%). En el caso del 2025, de 4312 personas, 1340 son pobres (31.08%). En definitiva, hubo un aumento del 3.26% de pobreza entre el 2005 y el 2025.

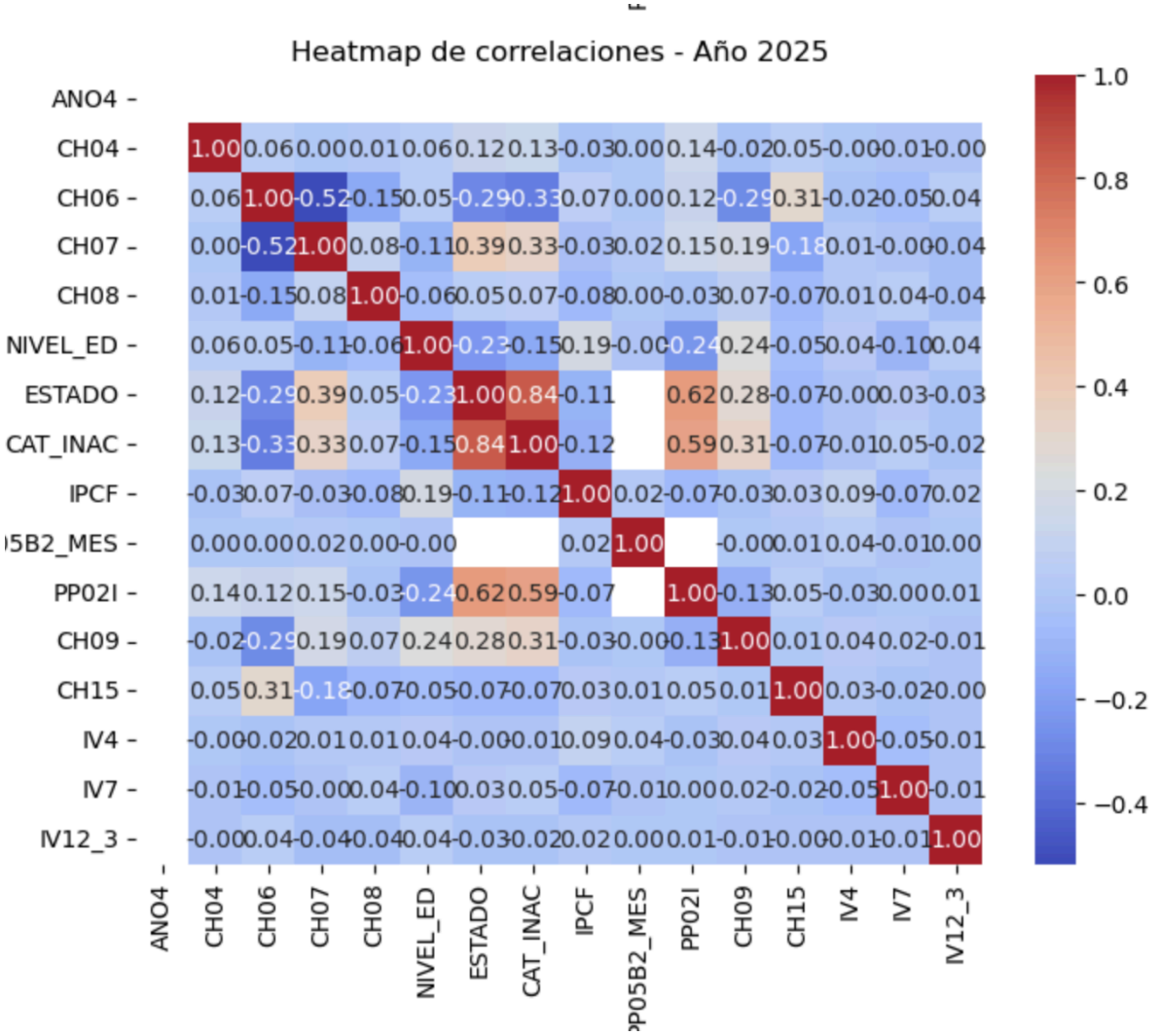
9. Como análisis exploratorio, realizamos una matriz de correlación entre la variable pobre y variables seleccionadas.

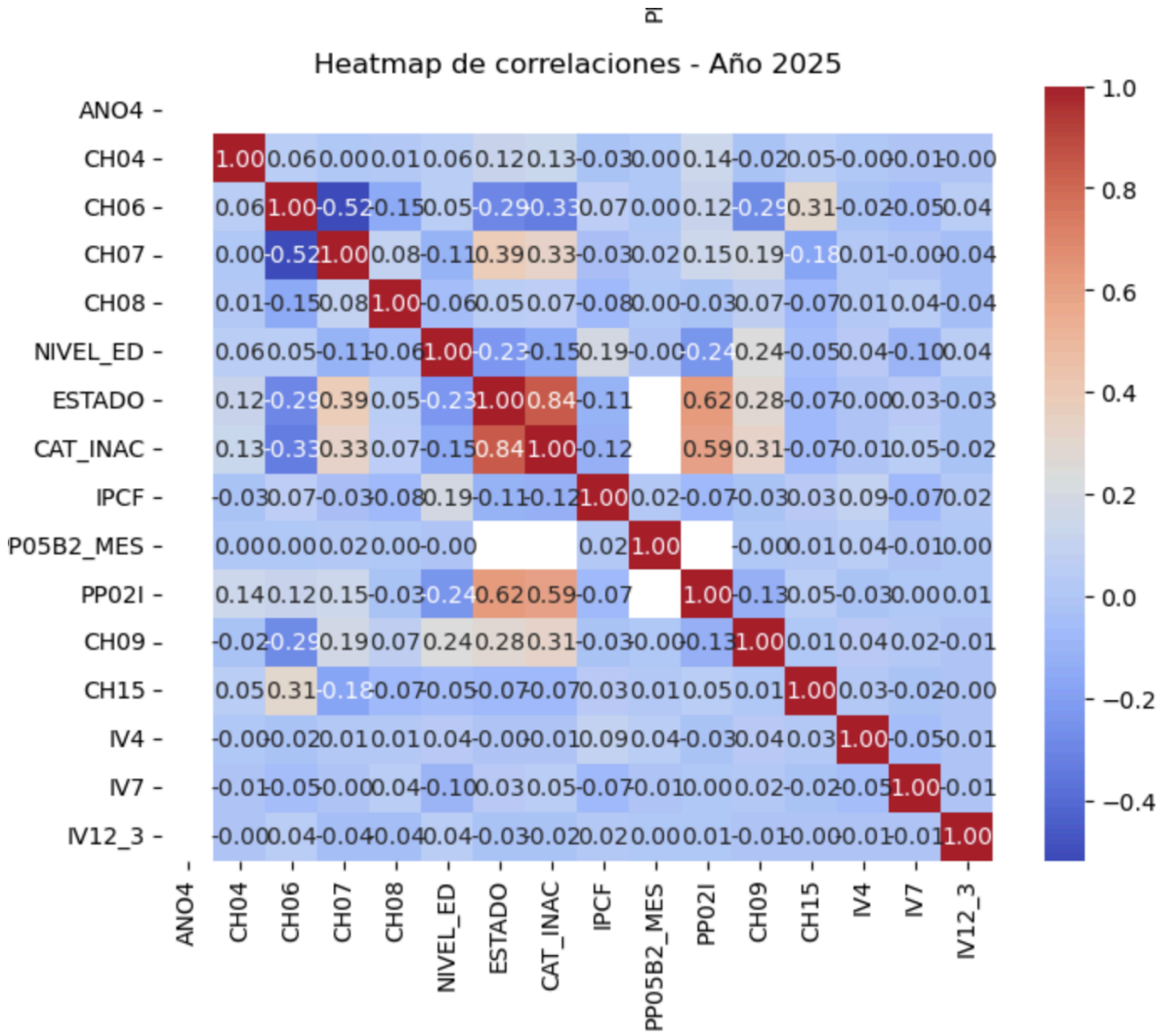
PORCENTAJE

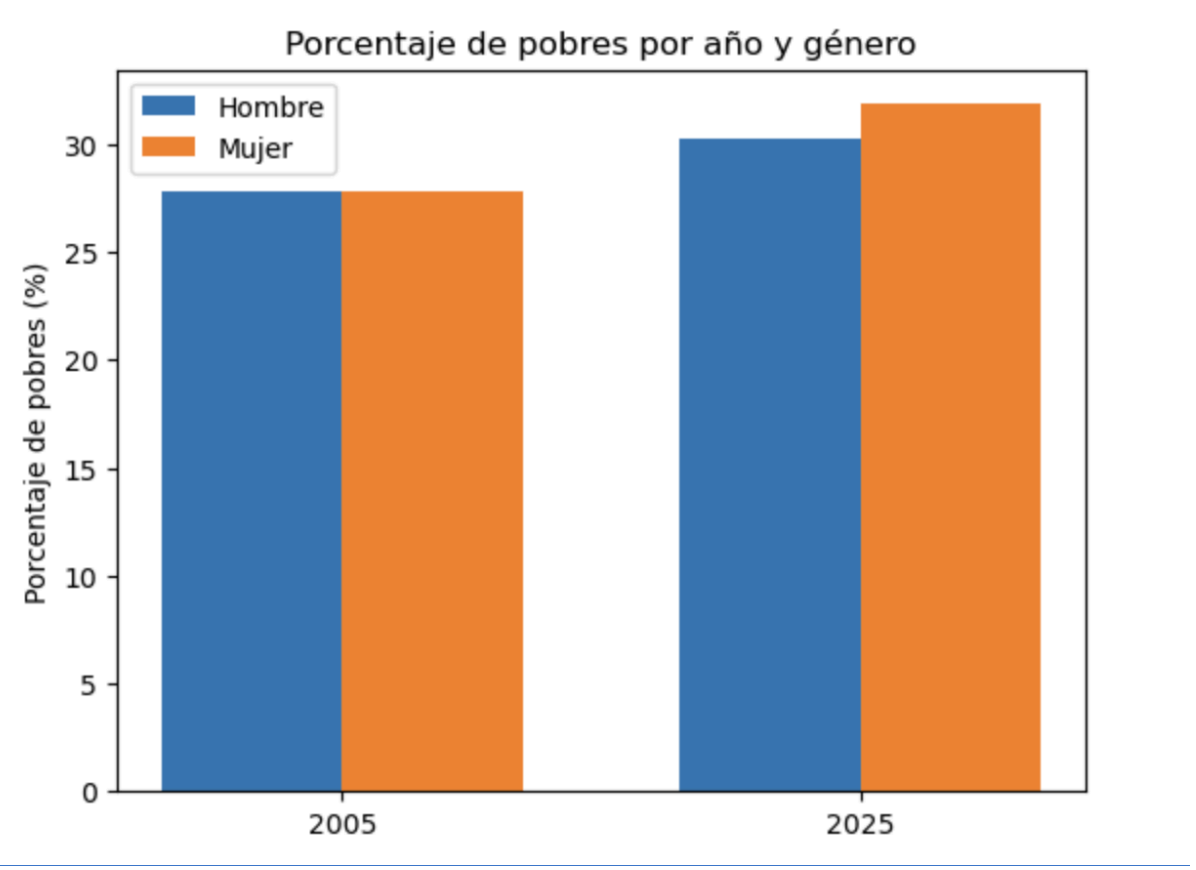
ANO4	
2005	27.82
2025	31.08

CANTIDAD

ANO4	
2005	9561
2025	4312

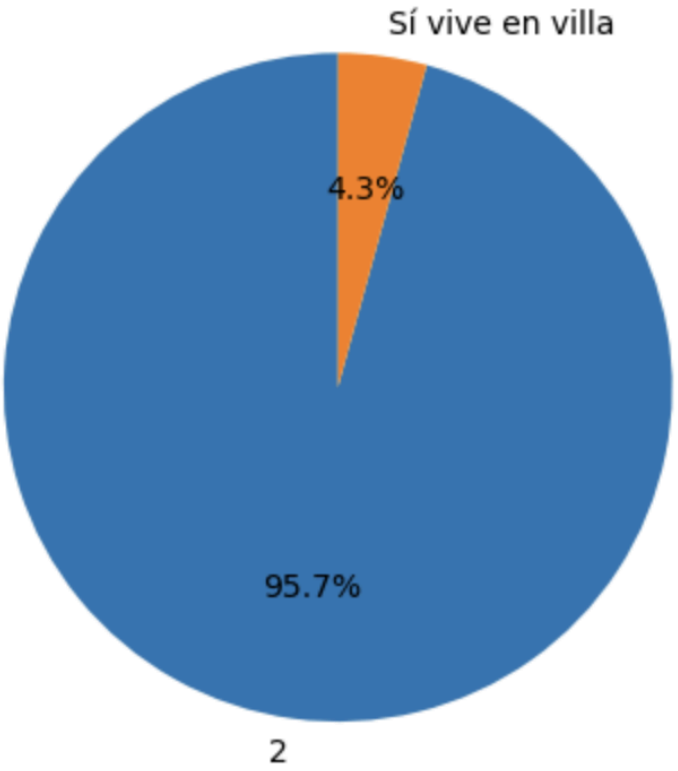



Heatmap de correlaciones - Año 2025
ANO4
CH04
CH06
CH07
CH08
NIVEL_ED
ESTADO
CAT_INAC
IPCF
P05B2_MES
PP02I
CH09
CH15
IV4
IV7
IV12_3
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

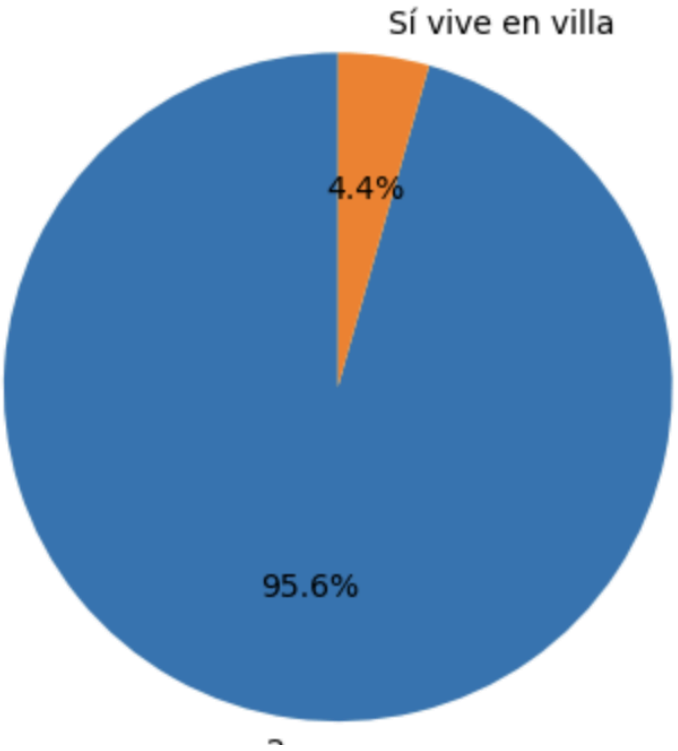


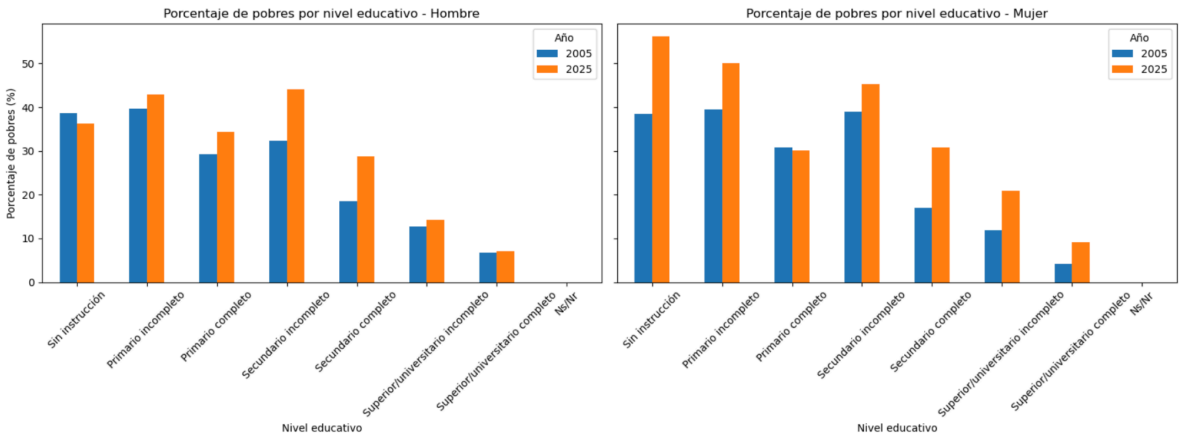
Aca observamos como se comporta la variable de pobreza con respecto a la educación seccionado por genero

Pobres que viven en villas de emergencia - Año 2005



Pobres que viven en villas de emergencia - Año 2025





De forma adicional, se hicieron dos análisis exploratorios sobre la pobreza

En primer lugar, se encontró una correlación negativa entre el nivel educativo alcanzado y la probabilidad de ser pobre; si bien no es muy alta, resulta estadísticamente significativa.

Logit Regression Results

Dep. Variable:	pobre	No. Observations:	13873
Model:	Logit	Df Residuals:	13871
Method:	MLE	Df Model:	1
Date:	Fri, 05 Sep 2025	Pseudo R-squ.:	0.01846
Time:	10:41:00	Log-Likelihood:	-8179.0
converged:	True	LL-Null:	-8332.8
Covariance Type:	nonrobust	LLR p-value:	7.254e-69

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2829	0.040	-7.153	0.000	-0.360	-0.205
NIVEL_ED	-0.1807	0.011	-17.171	0.000	-0.201	-0.160

En segundo lugar, se realizó un modelo de regresión logística teniendo en cuenta varias variables ('CH06', 'CAT_INAC', 'CH08_Obra_Social', 'CH08_No_Paga', 'NIVEL _ED', 'ESTADO OCUPADO') para predecir la condición de pobreza. Fue necesario crear variables dummy ya que había ciertas respuestas en algunos ítems que correlacionaban más individualmente que entre todas las respuestas posibles al ítem.

El modelo estimado se empleó para evaluar la existencia de sesgo de no respuesta en la variable de ingreso total familiar, mediante la comparación de las predicciones de pobreza en los casos sin información declarada. Se encontró que efectivamente existe un sesgo de no respuesta. El AUC del modelo fue de 0.796, por lo cual se considera que tiene un poder discriminador aceptable. Los resultados del modelo realizado presentaron un accuracy de 0.7607.

Posteriormente, se aplicó el modelo de regresión logística entrenado sobre la base con información completa de ingreso a la base de datos de no respondedores. Los resultados indican la presencia de un sesgo de no respuesta en la variable de ingreso total familiar: el modelo predijo un 32,41% de individuos en situación de pobreza en la muestra de no respondedores, comparado con un 28,83% de pobres en la base de respondedores.

Al desagregar por año, en 2005 la predicción del modelo fue de 36%, mientras que la incidencia observada en los respondedores fue de 27,82%; en 2025, la predicción fue de 31%, frente a un 31,08% en los respondedores. Estos resultados sugieren una mejoría en la aleatorización de la muestra a lo largo del tiempo, evidenciándose un sesgo de no respuesta menos pronunciado en 2025 en comparación con 2005.