Trabajo Práctico N°1 - Ciencia de Datos

Un primer encuentro con la EPH

Grupo 10

Francisca Cetra, Mariano Ripoll y Justina Rivero Ayerza

Profesora Magistral

María Noelia Romero

Profesor Tutorial

Tomas Enrique Buscaglia

Semestre y año de presentación

2do Semestre 2025

Link al repositorio de Github: https://github.com/franciscacetra/TP1-Grupo10.git

Parte I: Familiarizándonos con la base EPH y limpieza.

1. Con respecto a la metodología del INDEC; usa el método de la línea de pobreza: compara el ingreso total del hogar con el costo de una canasta básica total (CBT) que incorpora bienes y servicios esenciales. La CBT se obtiene a partir de la canasta básica alimentaria (CBA), que cubre requerimientos kilocalóricos y proteicos según sexo y edad, ampliada mediante un coeficiente de Engel; ambas se valorizan con precios del IPC y se ajustan por composición del hogar mediante adultos equivalentes (AE). La clasificación es a nivel de hogar: es indigente el hogar cuyo ingreso no alcanza su CBA; es pobre (no indigente) el hogar cuyo ingreso no alcanza su CBT; las personas heredan la condición de su hogar (INDEC, 2016; INDEC, 2018).

2. **Proceso de Limpieza (a,b,c):**

   Lo primero que hicimos fue acceder a la web del INDEC y efectuar la descarga de las bases de microdatos de la encuesta permanente de hogares para los primeros trimestres de 2005 y 2025. Asimismo, descargamos sus respectivos diseños de registro y estructura, asegurándonos de entender adecuadamente la codificación de cada variable, que medía y como operaba. Seleccionamos los datos de la región de Gran Buenos Aires porque sentíamos que era la que más nos interesaba y lo que nos afecta más directamente. Agregamos una columna referenciando cada dato a su año correspondiente para poder apilar las bases de datos sin perder el sentido de los distintos indicadores y la relación entre un hogar y una persona.

   Homogeneizamos los nombres y tipos de variables de un año respecto del otro y tomamos como 15 variables de interés: CH04 (sexo), CH06 (edad), CH07 (estado conyugal), CH08 (cobertura de salud), NIVEL_ED (nivel educativo), ESTADO (condición de actividad), CAT_INAC (categoría de inactividad), IPCF (ingreso per cápita familiar), ITF (ingreso total familiar), PONDERA (factor de expansión), CAT_OCUP (categoría ocupacional), AGLOMERADO (aglomerado), CODUSU (identificador de cada hogar), NRO_HOGAR (número de hogar), PP3E_TOT (horas trabajadas)

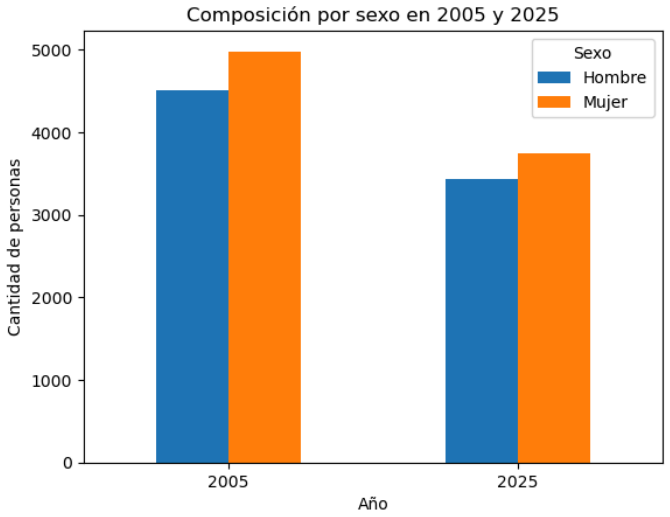
   Previo a llevar a cabo cualquier operación, cálculo o representación de las variables; corrimos un heatmap de faltantes respecto a las mismas, buscando así identificar qué

huecos había en la información. Acordemente, lo que hallamos fue que la variable referida a horas trabajadas (PP3E_TOT) concentraba las no-respuestas.

En última instancia limpiamos los valores que no tenían sentido en concordancia con la configuración del diccionario. Esto es, ingresos negativos a faltantes, edades fuera de rango, horas semanales imposibles, y códigos de "Ns/Nr". En este sentido también alineamos y corregimos, cuando 2005 y 2025 asignaban nombres diferentes para una misma categoría. De este modo posibilitamos efectuar el resto del análisis.

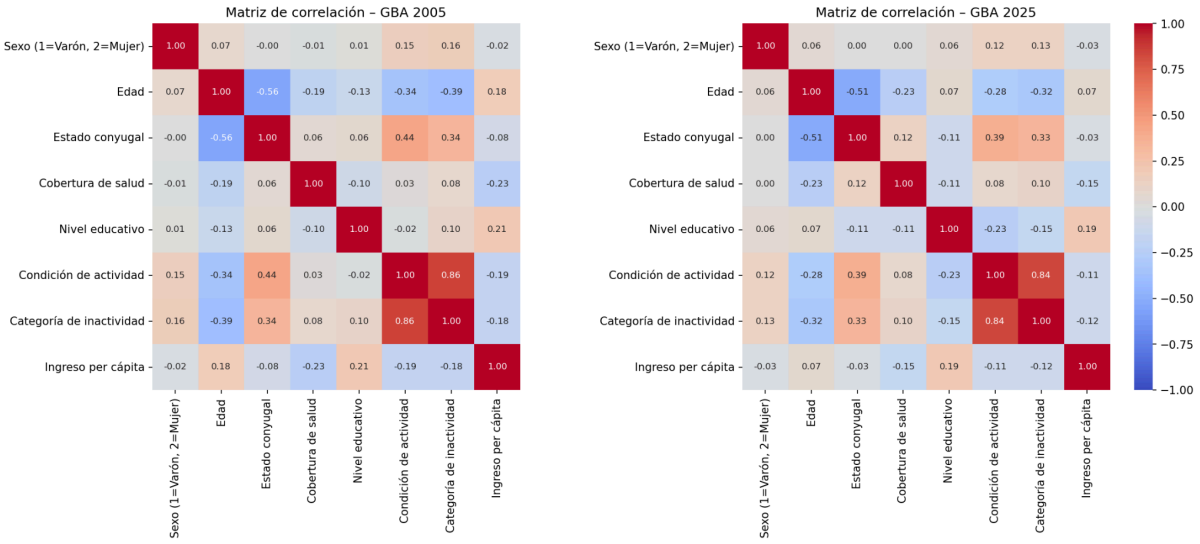
Parte II: Primer Análisis Exploratorio.

3.



Respecto a un breve comentar de la composición por sexo en 2005 y 2025 en Gran Buenos Aires: podemos ver que proporcionalmente hablando parecería corresponderse la proporción entre sexos en ambos años. No obstante, aparentemente se redujo la muestra tomada en 2025. Pero creemos que la representatividad está intacta.

4. Matriz de correlación (para 2005 y 2025).



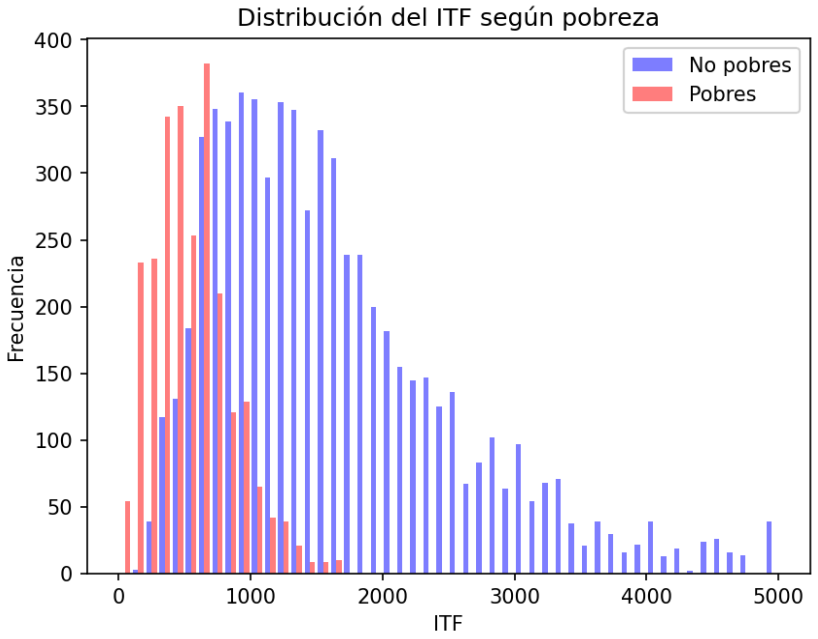
En las matrices de correlación se observa que algunas relaciones estructurales se mantienen estables a lo largo del tiempo. Por ejemplo, la condición de actividad y la categoría de inactividad presentan una correlación muy alta, lo cual resulta esperable dado que ambas variables reflejan la situación laboral de las personas. De manera similar, la edad mantiene una relación negativa moderada con el estado conyugal, indicando que a menor edad es menos probable estar en pareja o casado, mientras que esta probabilidad aumenta en edades mayores. No obstante, también se identifican cambios entre los años analizados. En 2005, la edad mostraba una correlación negativa con el nivel educativo, lo que reflejaba que los más jóvenes tendían a alcanzar mayores niveles de instrucción que las generaciones mayores; en 2025, esta relación prácticamente desaparece e incluso invierte su signo, sugiriendo un proceso de convergencia intergeneracional en el acceso a la educación. Asimismo, la asociación negativa entre edad y cobertura de salud se intensifica levemente en 2025, mientras que la relación positiva entre nivel educativo e ingreso per cápita, aunque persiste, disminuye ligeramente en magnitud. Por último, el sexo no presenta correlaciones relevantes con las demás variables, y el ingreso per cápita muestra asociaciones significativas únicamente con el nivel educativo.

Parte III: Conociendo a los pobres y no pobres.

Para poder explorar coherentemente el ámbito de la pobreza, primero, hicimos un chequeo de control sobre las respuestas; contamos cuanta gente no dijo su condición de actividad y, con respecto al ITF armamos dos subconjuntos que permitieran más prolijidad: en respondieron

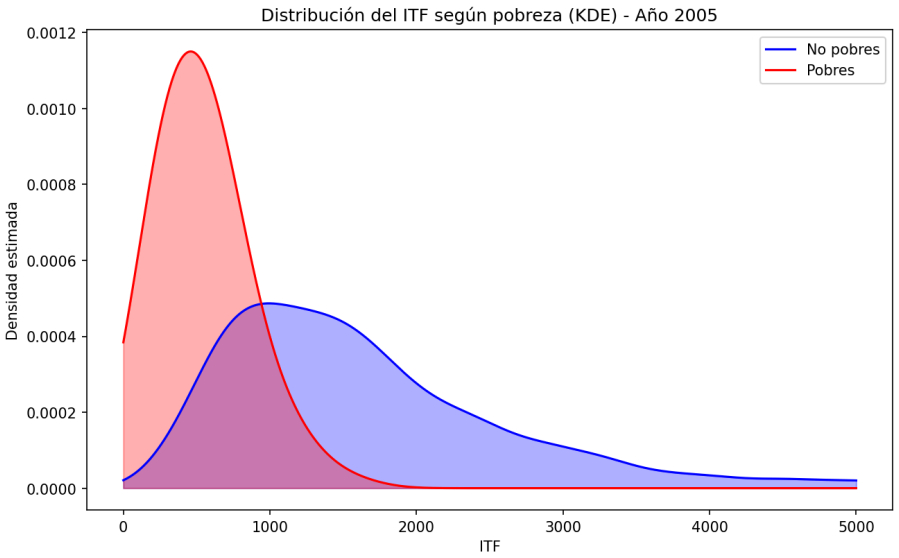
dejamos a quienes declararon ingresos positivos y en “no respondieron” guardamos los casos con ITF=0. A partir de este filtrado avanzamos a construir la variable pobre: calculamos adultos equivalentes con edad y sexo, los sumamos por hogar y comparamos el itf correspondiente del hoghar con el umbral correspondiente del período (la canasta por adulto equivalente). Si el ingreso no alcanza, ese hogar se marca como pobre; caso contrario, como no pobre. Con la variable de pobreza ya definida, armamos una tabla de descriptivas que resume el panorama general por año y, para analizarlo mejor, construimos gráficos exploratorios con la variable:

Figura 1. Histograma de distribución del ITF según pobreza



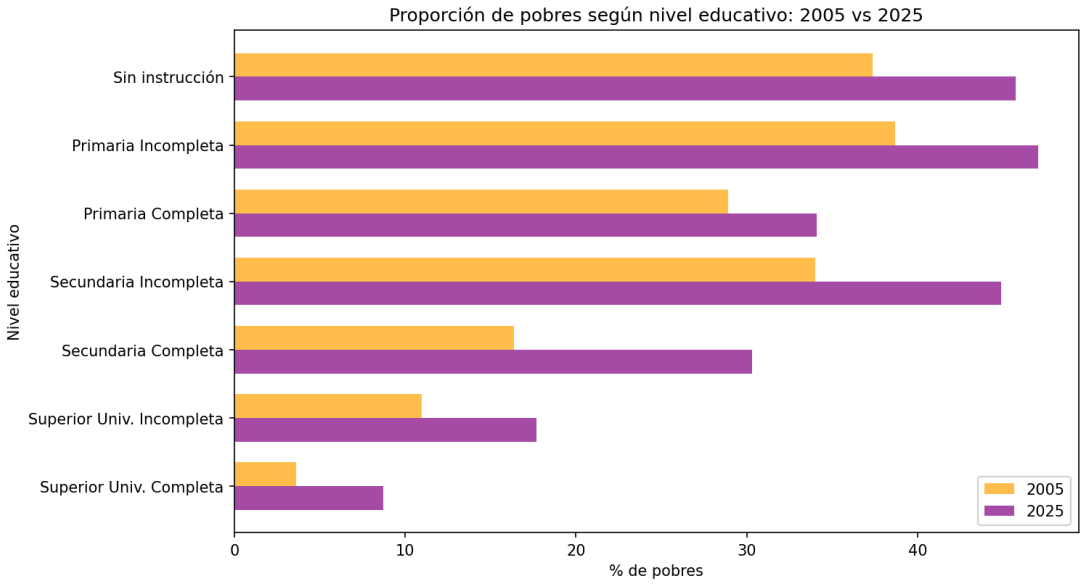
El gráfico permite ver que la pobreza tiende a conglomerarse en torno a los ingresos más bajos y caen rápidamente a medida que el ITF aumenta. Caso distinto es aquel de las barras azules (no pobres) que se extienden hacia la derecha con mucho más recorrido y una cola larga de ingresos altos. Es interesante notar la existencia de la superposición entre ambos grupos cerca del umbral: existen ahí los hogares “al límite”; se entiende que con una suba o baja chica de ingresos, podrían pasar de un grupo a otro.

Figura 2. Kernels con función Gaussiana del histograma anterior



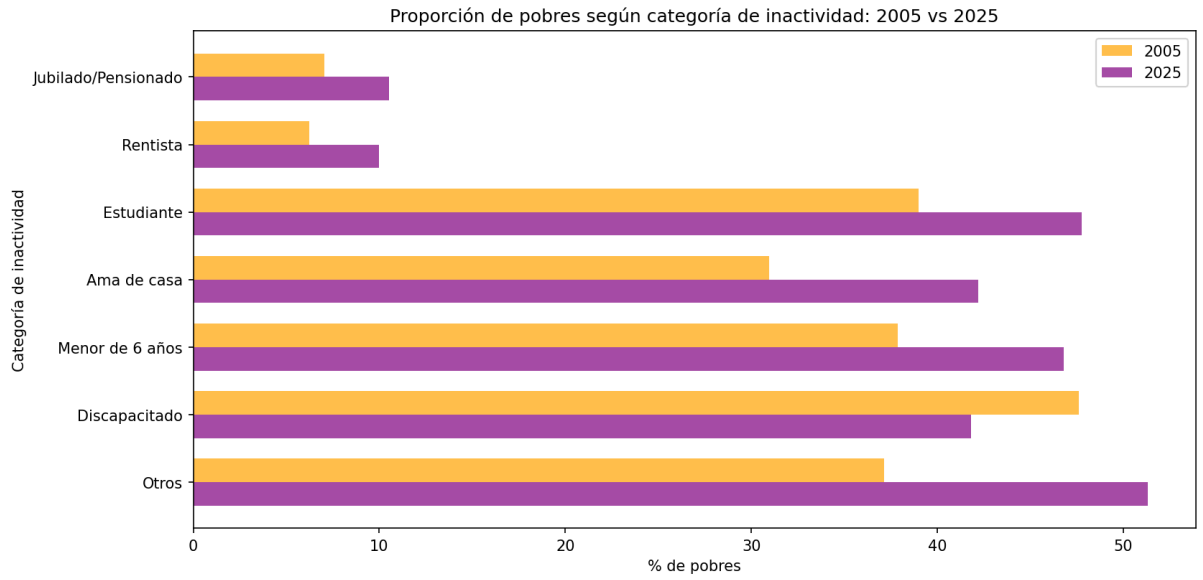
Como los histogramas le asignan igual valor de densidad a todos los puntos dentro de un bin aunque no necesariamente todas las observaciones de ese bin tengan exactamente igual fx de densidad, usamos Kernels para aumentar la precisión. En este caso, decidimos usar una función normal estandarizada. Graficamos la densidad También retoma la información del anterior, pero construye una curva suavizada que, en vez de las barras, muestra la forma continua de cada distribución. Así no se depende del tamaño muestral ni del número (ni ubicación) de bins y puede verse mejor el solapamiento cerca del umbral.

Figura 3. Gráfico de barras de proporción de pobres según Nivel educativo por año



Distinguiendo la proporción de pobres acorde al nivel educativo, se facilita la comprensión de cómo influye la educación en esta ecuación: a menor nivel educativo, mayor proporción de pobres; a medida que sube la educación, la barra cae. Esta tendencia es consistente en ambos años, pero en 2025 todas las barras quedan más altas que en 2005: sube en los niveles bajos, en los intermedios y en los altos. Esto puede entenderse simplemente como que la educación protege, reduce las probabilidades de pobreza, pero no es un factor que blinde contra ella; en 2025, incluso con secundaria o estudios superiores hay más pobreza que en 2005. Claramente, la pobreza, es una variable contingente a fenómenos macroeconómicos y tendencias generales más amplias.

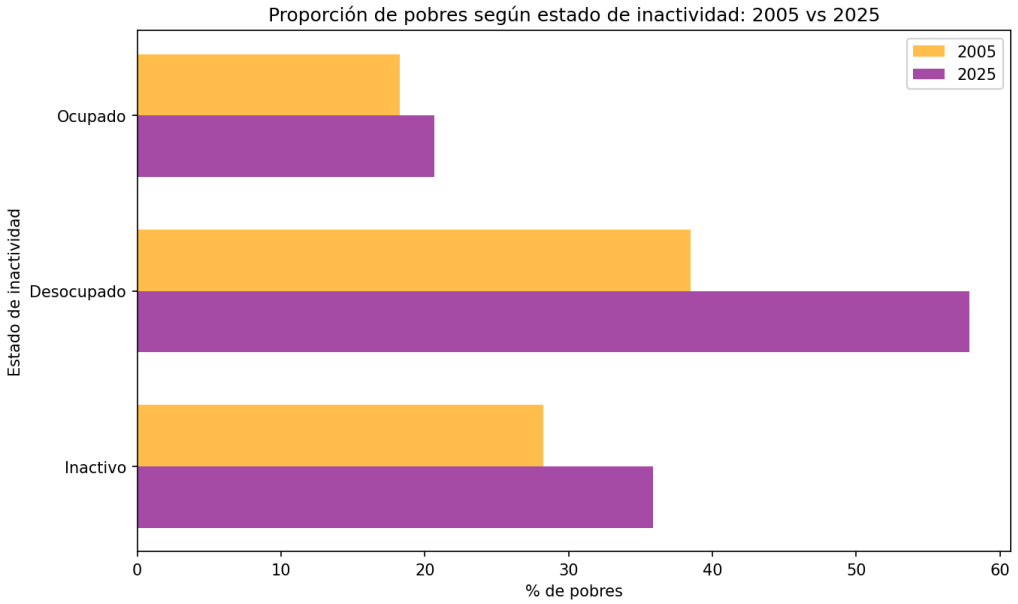
Figura 4. Gráfico de barras de proporción de pobres según Categoría de inactividad por año



El presente gráfico muestra los inactivos por subcategoría y compara 2005 respecto de 2025, si bien sí, todas las barras de 2025 quedan por encima de 2005, el patrón general interanual se mantiene. Las incidencias más altas están en *menores de 6 años*, *personas con discapacidad*, *amas de casa* y *estudiantes*: todos estos son perfiles que dependen del ingreso del hogar; si ese ingreso no alcanza, quedan expuestos. Caso contrario sucede con *jubilados/pensionados* y *rentistas*, que muestran las tasas más bajas, lo que es coherente con ingresos más estables. Ósea sí, la pobreza sube en 2025 dentro de la inactividad, pero la "geografía" es la misma, los

grupos sin ingreso propio o con menor autonomía económica siguen siendo los más vulnerables.

Figura 5. Gráfico de barras de proporción de pobres según Estado de inactividad por año



Otra perspectiva que creímos relevante para considerar la pobreza fue mediante su categorización por estado de actividad. En este caso, también, el orden se repite entre ambos años: desocupados, inactivos, ocupados. También es el caso que las tres barras suben para el año 2025, pero el salto más grande está dado por los desocupados; similar a la educación, uno puede inferir que tener trabajo, protege, pero no hace inmune a la pobreza. Entre inactivos la incidencia es intermedia porque dependen del ingreso del resto del hogar. El gráfico nos deja pensar que la inserción con el mercado laboral podría ser la línea que mejor separa los riesgos de pobreza en la región.

Referencias

INDEC (2016). La medición de la pobreza y la indigencia en la Argentina (Metodología INDEC N.º 22). (ver sección de Metodologías en indec.gob.ar). https://sitioanterior.indec.gob.ar/ftp/cuadros/sociedad/EPH_metodologia_22_pobreza.pdf

INDEC (2018). Incidencia de la pobreza y la indigencia. Perspectivas metodológicas (Informes Técnicos). (véase Información de archivo y series de CBA/CBT).