

Big Data y Machine Learning (UBA) - 2025

Trabajo Práctico 2: Un primer encuentro con la EPH

Grupo 1

Francisco Ariel Gonzalez Häberle 903012

David Jimenez Jaldin 909323

Lourdes Sofia Beltramo 888198

Link github: <https://github.com/davidjjj14082000-png/BigDataUBA-Grupo-1>

Parte 1

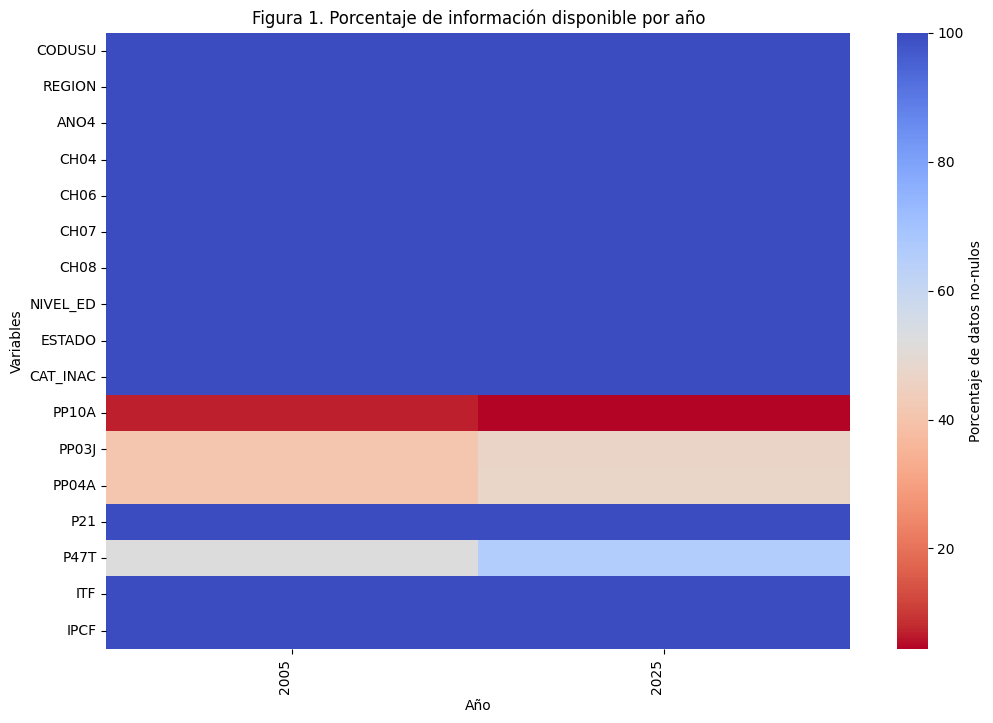
Ejercicio 1

Las nociones de pobreza empleadas por el INDEC se corresponden con el método de medición indirecta, denominado también “línea”[[1]](#footnote-1).

La “Línea de Pobreza” (LP) incluye no sólo los consumos alimentarios mínimos (como lo hace la línea de indigencia) y también otros consumos básicos no alimentarios. La suma de ambos conforma la Canasta Básica Total (CBT), la cual es también contrastada con los ingresos de los hogares relevados por la Encuesta Permanente de Hogares (EPH).

Por lo tanto, es necesario contar con el valor de la Canasta básica total y ampliarlo 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), es decir se construye en base a la evidencia empírica que refleja los hábitos de consumo alimentario y no alimentario de la población de referencia.

Ejercicio 2



La figura muestra el porcentaje de información disponible por año (2005 y 2025) para las variables seleccionadas.

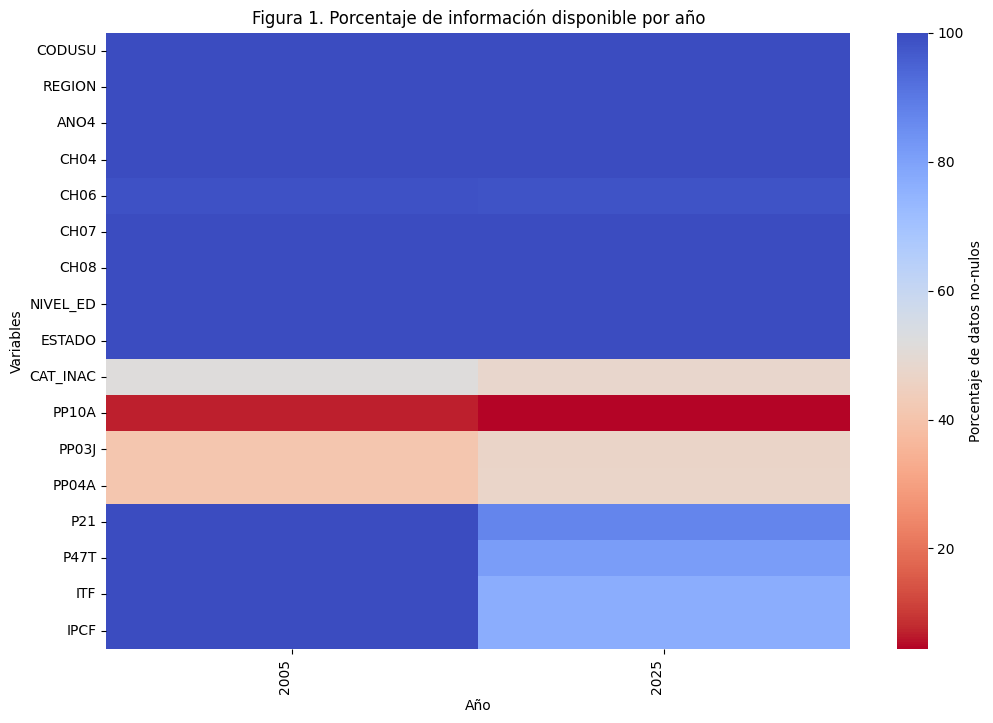
La mayoría de las variables, como CODUSU, REGION, ANO4, CH04, CH06, CH07, CH08, NIVEL\_ED, ESTADO, CAT\_INAC, P21, ITF e IPCF, presentan 100% de datos disponibles en ambos años, sin valores faltantes relevantes.

En contraste, tres variables presentan altos niveles de datos faltantes corresponden con:

* PP10A (desocupados: Cuánto hace que está buscando trabajo): es la más incompleta, con apenas un porcentaje mínimo de valores no nulos en ambos años (menos del 5%).
* PP03J (Aparte de este/os trabajo/s, ¿estuvo buscando algún empleo/ ocupación/actividad?) y PP04A (refiere al tipo de institución en la que trabaja, estatal, privada u otra): muestran una proporción de datos faltantes significativa, cercana al 50%, aunque con una leve mejora en 2025 respecto de 2005.
* P47T (Monto de ingreso total individual): presenta también datos faltantes, pero en menor medida que PP03J y PP04A, con un aumento de disponibilidad en 2025.

Conclusión: las variables con más valores faltantes son PP10A, PP03J, PP04A y P47T, siendo PP10A la más incompleta en ambos períodos. El resto de las variables mantiene cobertura total de información en 2005 y 2025.

A continuación, corregimos los valores que no tienen sentido o que cumplen algunas condiciones para poder observar la esas variables incoherentes ya corregidas y dar así una mejor claridad de los datos.



Se observa en la gráfica que, después de corregir las variables, más de la mitad de los indicadores muestra niveles altos de completitud (alrededor del 100%) en los años 2005 y 2025.

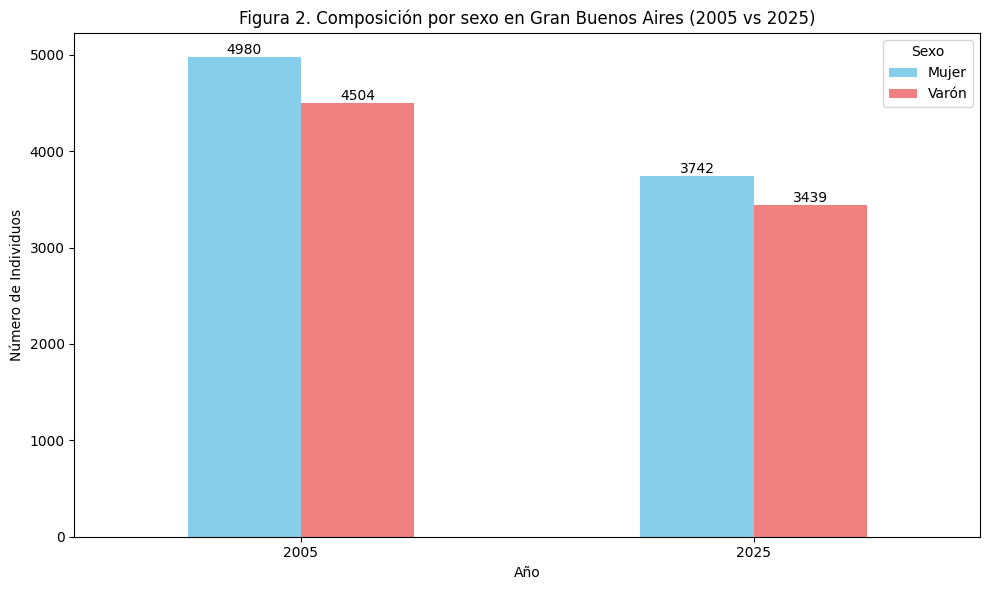
Solo algunas variables presentan proporciones más bajas de datos disponibles:

CAT\_INAC (categoría de inactividad), PP10A (desocupados: Cuánto hace que está buscando trabajo), PP03J (Aparte de este/os trabajo/s, ¿estuvo buscando algún empleo/ ocupación/actividad?) y PP04A con valores intermedios (refiere al tipo de institución en la que trabaja, estatal, privada u otra), y en menor medida P47T (monto de ingreso total individual), ITF (monto de ingreso total familiar); e IPCF (monto de ingreso per cápita familiar) en 2025.

En general, la corrección posibilitó que se disminuyeran las inconsistencias y se mejorara la calidad de los datos, lo cual asegura una mayor solidez para los análisis futuros.

Parte 2

Ejercicio 3



El gráfico expuesto muestra la composición por género en Gran Buenos Aires en dos periodos de tiempo: 2005 y 2025.

En 2005, la población de mujeres (4980) es más alta que la de hombres (4504), con una diferencia aproximada de 476 personas.

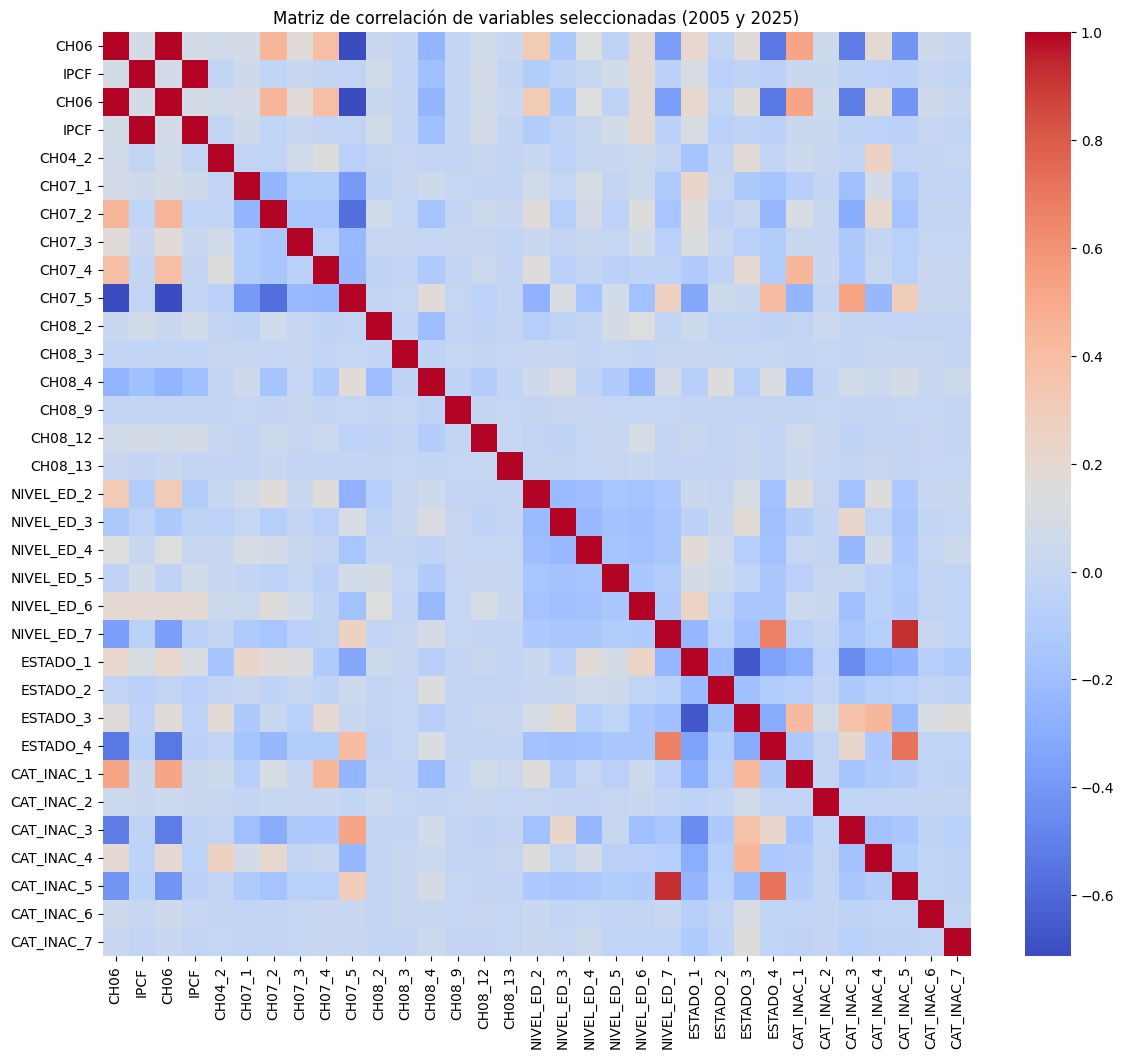
Para el 2025, ambas cifras de población se reducen: la de las mujeres baja a 3742 y la de los hombres se reduce a 3439. Pese a ello, continúa la tendencia de contar con más mujeres que hombres, aunque con una diferencia menor (303 individuos).

En conclusión, el gráfico muestra:

una disminución de la población total, tanto masculina como femenina, entre 2005 y 2025;

y que la disminución de la población afecta a hombres y mujeres, aunque el descenso es más significativo en términos absolutos para las mujeres (1238 menos) que para los hombres (1065 menos).

Ejercicio 4



La figura muestra la matriz de correlación de variables sociodemográficas, educativas y laborales en Gran Buenos Aires (2005 y 2025).

Las diferentes categorías educativas presentan correlaciones con las de ocupación y condición de actividad, lo que evidencia el vínculo entre la participación en el mercado laboral y el capital educativo.

Las variables de cobertura médica (ch08) tienen correlaciones distintas con la condición de actividad y la condición de inactividad, lo que indica una relación entre la situación ocupacional y el acceso a sistemas de protección.

La edad (ch06) muestra correlaciones positivas con categorías de inactividad como jubilación o pensión, lo cual refleja comportamientos demográficos esperados.

En conclusión, se proporcionó una amplia perspectiva sobre las relaciones entre variables sociodemográficas en el Gran Buenos Aires. Entre los descubrimientos más importantes sobresale la veracidad de las conexiones entre el acceso a servicios sanitarios, la educación y la condición laboral, además de que se han verificado patrones demográficos comunes (sexo-estado civil, edad-jubilación).

1. La medición de la pobreza y la indigencia en la Argentina. Metodología INDEC Nº 22 https://www.indec.gob.ar/ftp/cuadros/sociedad/EPH\_metodologia\_22\_pobreza.pdf [↑](#footnote-ref-1)