Trabajo Práctico 2 – Taller de programación

Maestría: Economía aplicada

Profesor: Noelia Romero

Alumnos: Shanthal Chavarria, Edgar Julián Pineda Toro, Gustavo Horacio Romero

Enlace del repositorio: https://github.com/shanthalchs/BigDataUBA-Grupo6

1. Introducción

El presente trabajo constituye la segunda etapa del análisis de la Encuesta Permanente de Hogares (EPH) para el Gran Buenos Aires, correspondiente a los primeros trimestres de 2005 y 2025. En esta instancia buscamos profundizar el estudio de las condiciones socioeconómicas de los hogares, con especial atención a las diferencias entre grupos poblacionales en términos de educación, ingresos, pobreza y características laborales.

El trabajo se desarrolla en dos partes complementarias. En la primera, elaboramos variables derivadas que permiten una descripción más precisa de los hogares y las personas —entre ellas, los años de educación, los ingresos ajustados por inflación, las horas totales trabajadas y la edad al cuadrado— y presentamos un análisis descriptivo comparativo entre ambos años. En la segunda parte, aplicamos métodos multivariados, como el Análisis de Componentes Principales (ACP) y la clasificación no supervisada (k-medias), con el propósito de identificar perfiles socioeconómicos y caracterizar la estructura interna de la población bajo estudio.

2. Metodología

El trabajo se realizó a partir de la base de datos depurada en el TP1, garantizando consistencia entre las variables y la comparabilidad entre años. Inicialmente, se construyeron nuevas variables derivadas:

- **edad2**, definida como el cuadrado de la edad;
- **educ**, que representa los años de educación formal completados según el nivel educativo, finalización y último año aprobado;
- **ingreso_total_familiar**, ajustado a precios de 2025 en función de la **Canasta Básica Total (CBT)**;
- **horastrab**, correspondiente al total de horas trabajadas en ocupaciones principales y secundarias;
- y **pobre**, que identifica hogares con ingresos insuficientes respecto a su ingreso necesario.

En la segunda parte, trabajamos con las variables normalizadas y aplicamos el Análisis de Componentes Principales (ACP) para reducir la dimensionalidad del conjunto de datos, identificando los factores que explican mayor varianza. Posteriormente, implementamos un

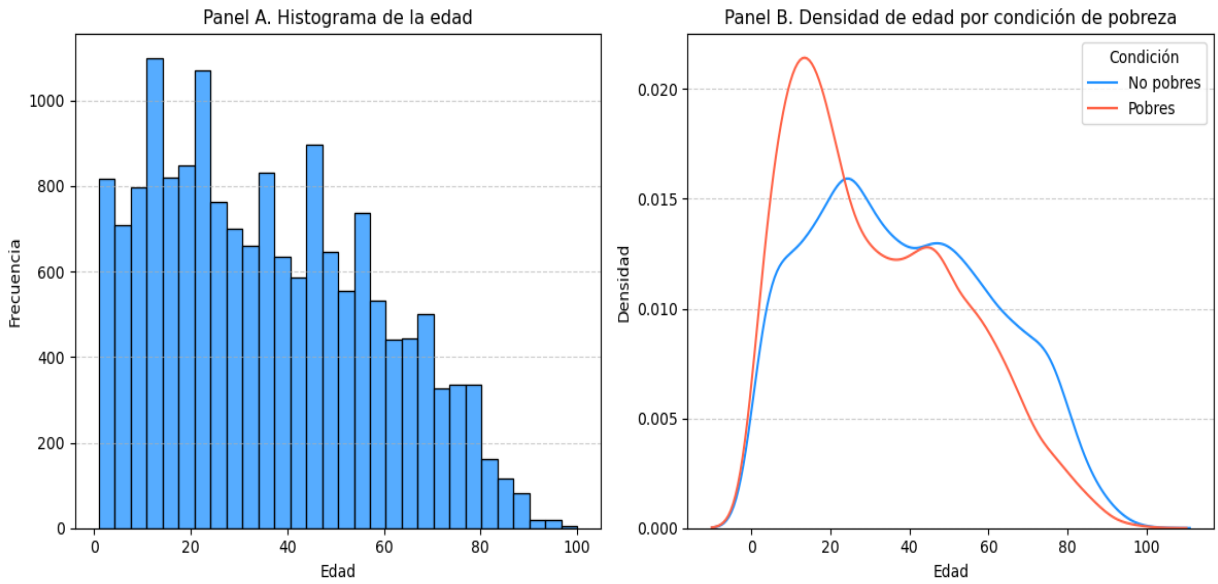
algoritmo de agrupamiento k-medias para clasificar a los individuos según sus características socioeconómicas. Los resultados se presentan mediante gráficos y tablas descriptivas que permiten interpretar la estructura social y económica de la región en ambos períodos.

3. Resultados

El análisis descriptivo permite caracterizar la estructura sociodemográfica y económica de la población del Gran Buenos Aires a partir de las variables construidas en la primera parte del trabajo.

El análisis descriptivo de la edad muestra una estructura poblacional activa en el Gran Buenos Aires, concentrada principalmente entre los 25 y los 55 años. Al comparar la distribución entre pobres y no pobres, se observa que los primeros tienden a ser más jóvenes, lo que sugiere una mayor incidencia de pobreza en hogares encabezados por personas con menor experiencia laboral o menores niveles educativos. En cambio, los no pobres se concentran en edades medias y avanzadas, asociadas a una mayor estabilidad económica y laboral.

Gráfico 1. Distribución de la edad – EPH 2005 y 2025

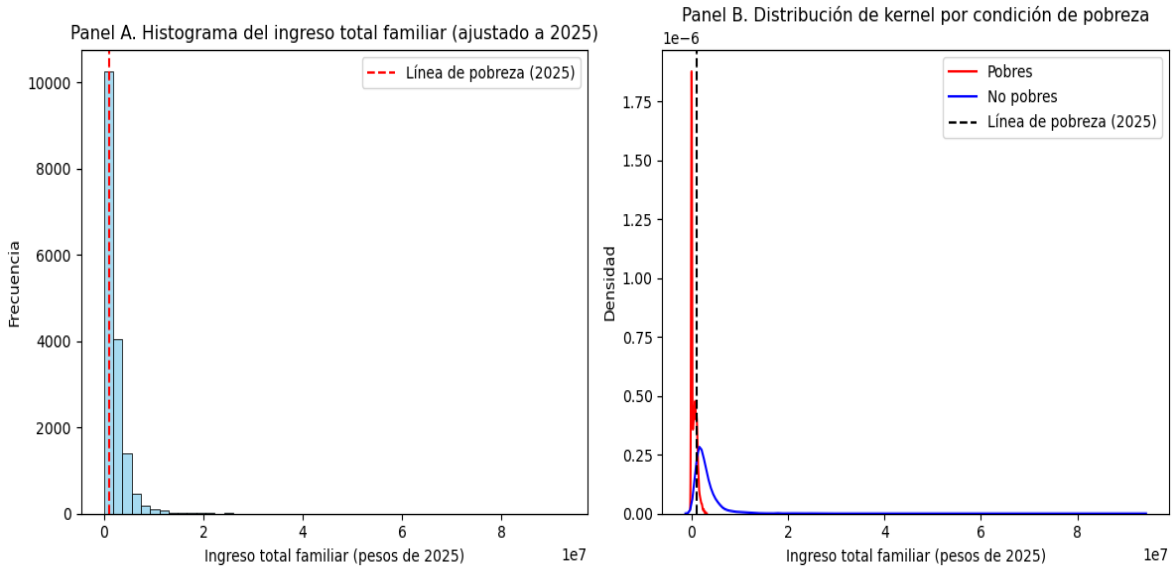


Fuente: Elaboración propia a partir de la EPH 2005 Y 2025.

En cuanto a la variable, *educ* (tabla con estadísticas detalladas se pueden observar en el código anexado) se observa que presenta un promedio de **10,5 años de escolaridad**, con una mediana de **11 años**, lo que corresponde aproximadamente al nivel secundario completo. El rango de valores va de **0 a 23 años**, reflejando tanto a quienes no completaron la educación básica como a aquellos con estudios universitarios y de posgrado. Esta dispersión evidencia la heterogeneidad educativa en el Gran Buenos Aires. En promedio, los hogares pobres registran menos años de educación, lo que confirma la estrecha relación entre nivel educativo y condiciones económicas.

El **Gráfico 2,** en el panel A muestra una distribución del ingreso familiar fuertemente asimétrica, con alta concentración en los tramos bajos. En el panel B, la comparación entre pobres y no pobres evidencia una clara brecha: los hogares pobres se agrupan muy por debajo de la línea de pobreza, mientras los no pobres presentan ingresos más diversificados.

Gráfico 2. Composición por sexo en Gran Buenos Aires – EPH 2005 y 2025



Fuente: Elaboración propia a partir de la EPH 2005 Y 2025.

En cuanto a la variable *horastrab*, que mide el total de horas trabajadas por los jefes de hogar, presenta un promedio de **33,4 horas semanales** y una mediana de **29,5 horas**, lo que sugiere una carga laboral ligeramente inferior a la jornada completa típica. La dispersión es elevada, con una desviación estándar de **83 horas**, lo que refleja una marcada heterogeneidad en la cantidad de horas trabajadas entre los hogares, coherente con la diversidad de situaciones laborales observadas en la región del Gran Buenos Aires.

Por último, en la tabla 1 presenta un resumen del tamaño y la composición de la base final para el Gran Buenos Aires, diferenciando los años 2005 y 2025. En ella se muestra la cantidad total de observaciones, los registros con valores faltantes en la variable *pobre*, y la distribución entre hogares pobres y no pobres, junto con el número de variables limpias y homogéneas en cada período. Esta información permite visualizar la magnitud de la base consolidada y la consistencia del proceso de depuración aplicado antes de los análisis descriptivos y comparativos.

Tabla 1. Resumen de la base final para la región de Gran Buenos Aires

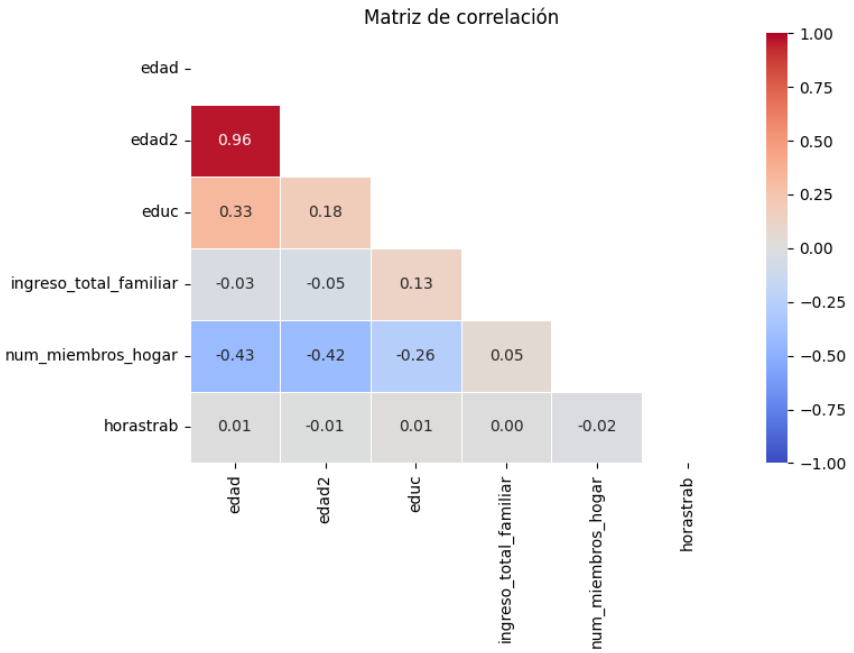
Ítem	2005	2025	Total
Cantidad observaciones	9,484	7,181	16,665
Cantidad de observaciones con NAs en la variable "Pobre"	0	0	0
Cantidad de Pobres	2,506	4,204	6,710
Cantidad de No Pobres	6,978	2,977	9,955
Cantidad de variables limpias y homogeneizadas	173	90	

Fuente: Elaboración propia a partir de la EPH 2005 Y 2025.

3.1 Matriz de correlación

Ahora analizamos las correlaciones entre nuestras variables en al gráfico 3. La matriz muestra una fuerte relación entre edad y edad^2 (0.96), lo cual es esperado dado que una es función de la otra. Se observa una correlación positiva moderada entre educación e ingreso familiar, mientras que el número de miembros del hogar se asocia negativamente con la edad y la educación. En conjunto, las correlaciones son bajas, lo que indica que las variables no están altamente colineadas y son adecuadas para aplicar PCA.

Gráfico 3. Matriz de correlaciones



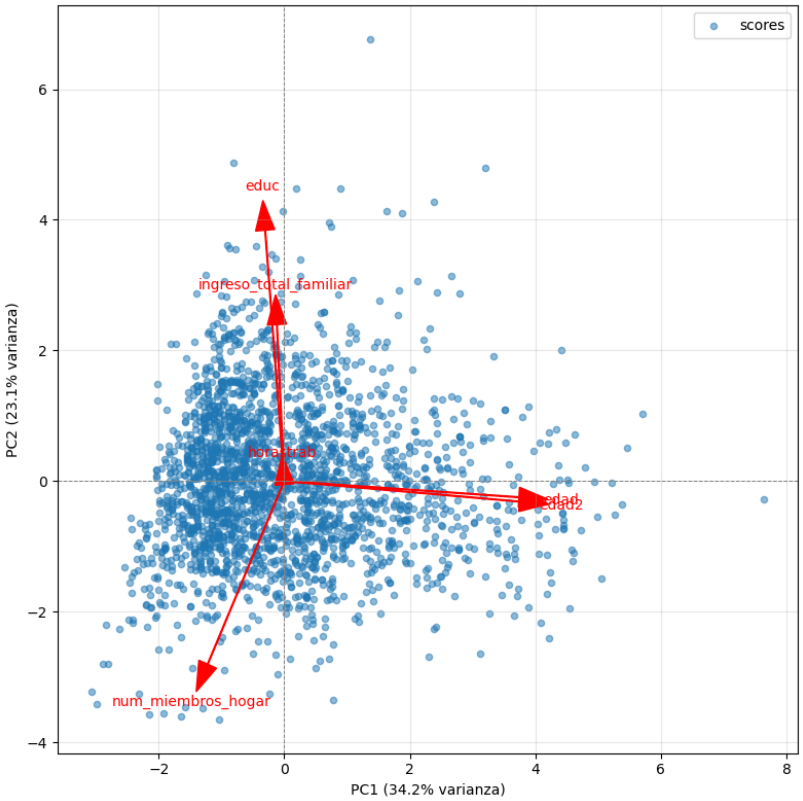
Fuente: Elaboración propia a partir de la EPH 2005 Y 2025.

3.2 PCA

El análisis de componentes principales permite reducir la dimensionalidad conservando la mayor parte de la información. En el gráfico 4 de dispersión de los dos primeros componentes (que explican alrededor del 57% de la varianza total), se observa que las observaciones se agrupan en torno al origen, con mayor dispersión en el eje del primer componente, asociado principalmente a la edad y al ingreso familiar.

También en el gráfico 4 podemos observar las flechas del biplot que muestran que las variables edad y edad^2 contribuyen fuertemente al primer componente, mientras que la educación y el ingreso total familiar tienen mayor peso en el segundo. Esto sugiere que el PCA separa principalmente a los individuos por características etarias y, en segundo término, por nivel educativo y capacidad económica.

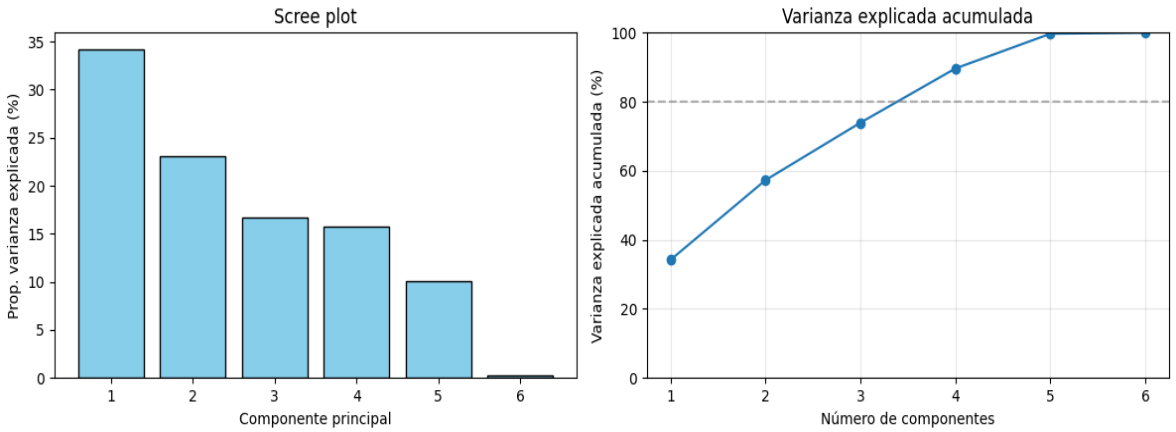
Gráfico 4. PCA y flechas con los ponderados



Fuente: Elaboración propia a partir de la EPH 2005 Y 2025.

Ahora en el gráfico de varianza explicada indica que los dos primeros componentes concentran aproximadamente el 57% de la información total, mientras que con cuatro componentes se alcanza cerca del 90%. Esto implica que la mayor parte de la variabilidad de las variables originales puede resumirse en pocos factores, lo que valida la utilidad del PCA para simplificar el análisis.

Gráfico 5. Varianza explicada

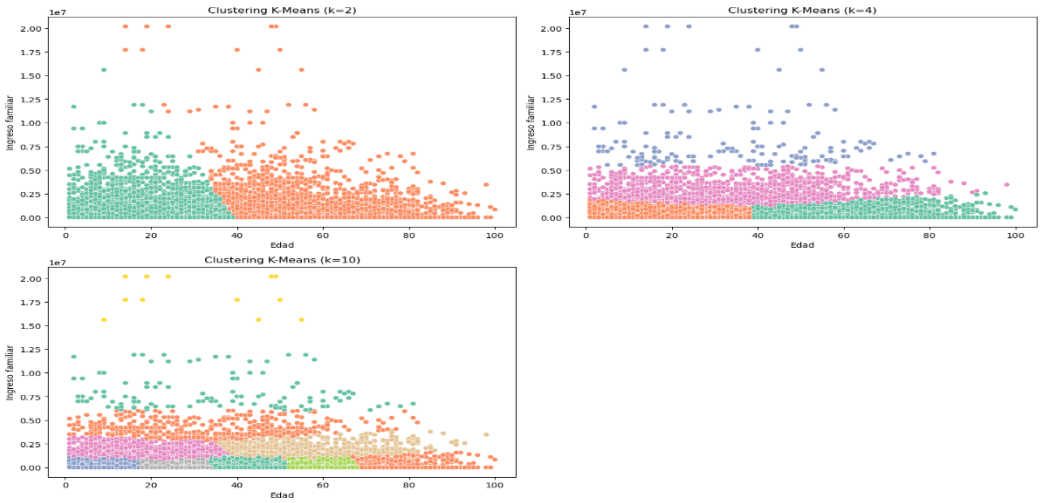


Fuente: Elaboración propia a partir de la EPH 2005 Y 2025.

3.3 Análisis de clusters

El algoritmo de **K-medias** permite segmentar la población del Gran Buenos Aires según edad e ingreso. Con **k = 2**, logra distinguir parcialmente entre pobres y no pobres, aunque existe solapamiento entre ambos grupos. Al aumentar a **k = 4**, los clusters reflejan mejor la heterogeneidad económica, mientras que con **k = 10** la segmentación pierde claridad interpretativa.

Gráfico 6. Segmentación por edad e ingreso familiar en Gran Buenos Aires – EPH 2005 y 2025

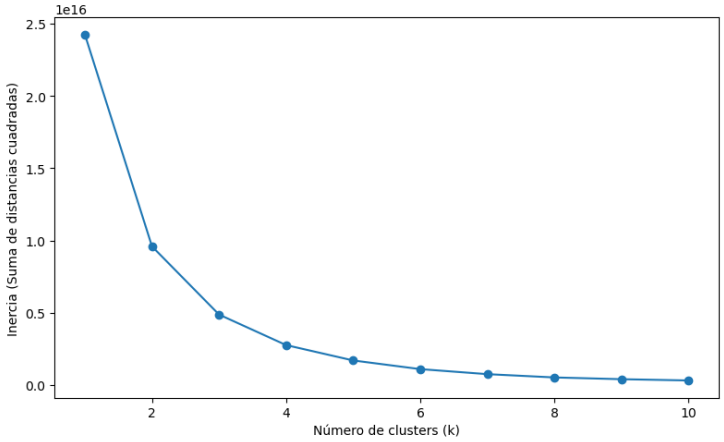


Fuente: Elaboración propia a partir de la EPH 2005 Y 2025.

También aplicamos el método del codo (*Elbow*) con el objetivo de determinar el número óptimo de grupos que puede identificar el algoritmo de K-means a partir de las variables **edad** e

ingreso familiar. El gráfico muestra una reducción pronunciada en la inercia hasta aproximadamente **k = 3**, a partir de donde las mejoras se vuelven marginales. Esto sugiere que el número óptimo de clusters para la región del Gran Buenos Aires se sitúa en torno a **tres grupos**. Esta cantidad de agrupamientos permite captar diferencias socioeconómicas relevantes dentro de la población, distinguiendo no solo entre pobres y no pobres, sino también entre distintos niveles de ingreso o estabilidad económica.

Gráfico 7. Inspección visual de Elbow

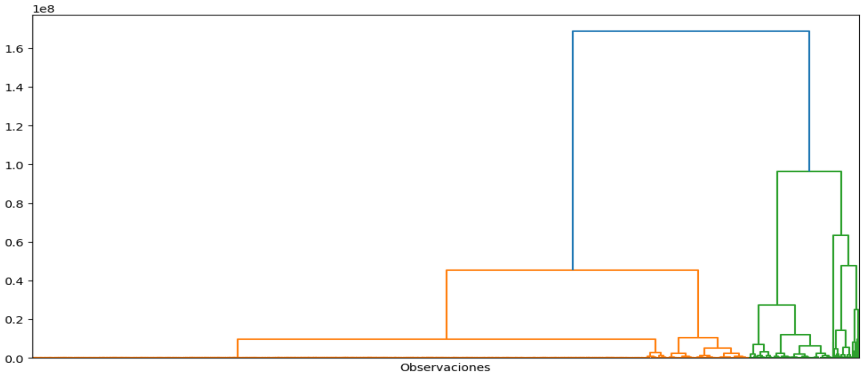


Fuente: Elaboración propia a partir de la EPH 2005 Y 2025.

3.4 Clustering jerárquico

Ahora realizamos el dendrograma, el cual nos permite visualizar la estructura jerárquica de los grupos formados a partir de las variables edad e ingreso familiar. Cada unión entre ramas representa una fusión entre individuos o grupos similares según la distancia euclídea. Se observa que los primeros conglomerados se forman entre observaciones con ingresos y edades cercanas, mientras que las divisiones principales reflejan contrastes más marcados en el nivel de ingreso. En este sentido, el dendrograma evidencia la existencia de grupos socioeconómicos diferenciados, aunque no una separación exacta entre pobres y no pobres.

Gráfico 8. Dendrograma del clustering jerárquico



Fuente: Elaboración propia a partir de la EPH 2005 Y 2025.