**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**](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.

1. **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:

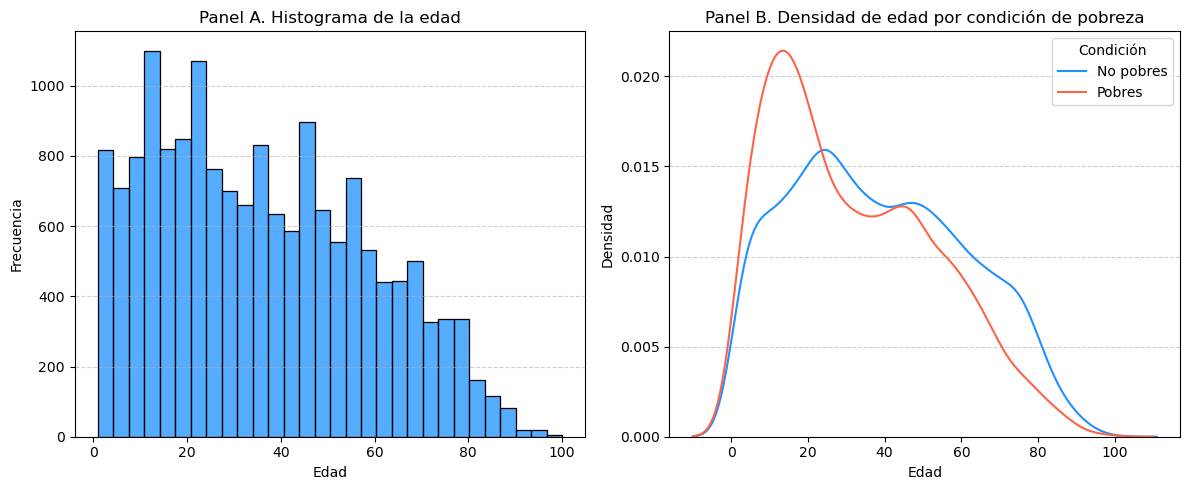
* **edad2**, definida como el cuadrado de la edad;
* **educ**, que representa los años de educación formal completados según el nivel educativo, finalización y último año aprobado;
* **ingreso\_total\_familiar**, ajustado a precios de 2025 en función de la **Canasta Básica Total (CBT)**;
* **horastrab**, 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.

1. **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 la tabla 1, se observa que la variable *educ* 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.

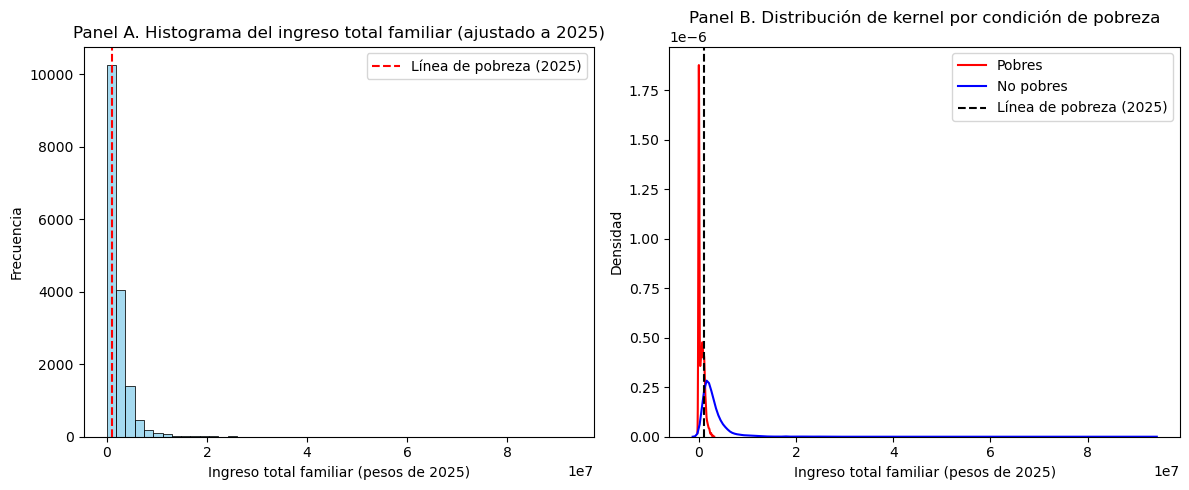
**Tabla 1. Estadísticas de educación – Gran Buenos Aires (2005 y 2025)**

|  |  |
| --- | --- |
| **Estadístico** | **Total** |
| Cuenta | 15,828 |
| Media | 10.48 |
| Desviación estandar | 5.01 |
| Mediana | 11.00 |
| Valor máximo | 23.00 |

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

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.

**Tabla 2. Estadísticas horas trabajadas – Gran Buenos Aires (2005 y 2025)**

|  |  |
| --- | --- |
| **Estadístico** | **Total** |
| Media | 33.35 |
| Desviación estandar | 83.06 |
| Mediana | 29.50 |
| Valor máximo | 1,047 |

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

Por último, en la tabla 3 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 3. 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.**