



FCE-UBA
UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS ECONÓMICAS

BIG DATA Y MACHINE LEARNING

TRABAJO PRÁCTICO N° 3

*HISTOGRAMAS, KERNELS & MÉTODOS NO
SUPERVISADOS USANDO LA EPH*

(Encuesta Permanente de Hogares)

GRUPO 3

Rafael Pablo Pinto Chambi - Reg. 908586

Javier Rodolfo Aguirre - Reg. 819698

Lautaro Manuel Bogado – Reg. 899596

Enlace de repositorio: <https://github.com/RafaelPCh/BigDataUBA-Grupo3.git>

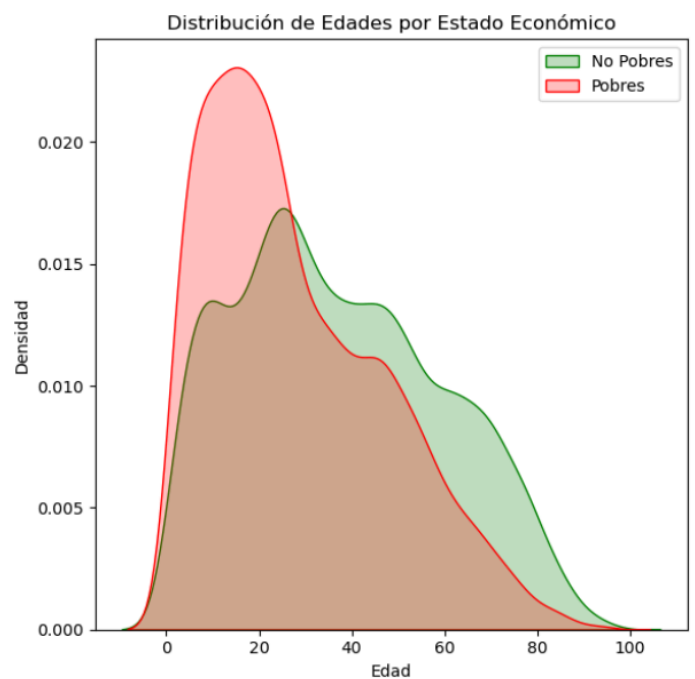
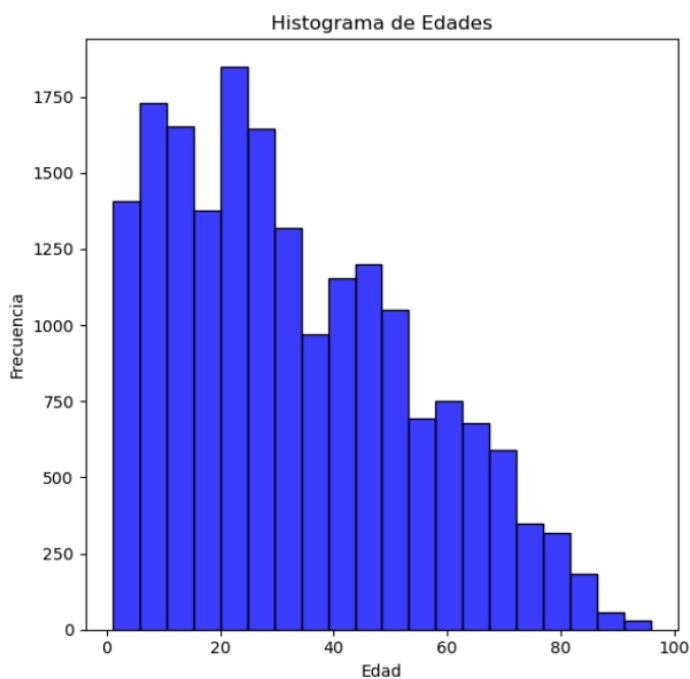
Docente a cargo: María Noelia Romero

Segundo Cuatrimestre de 2025

Parte uno:

Punto 1:

La distribución de edades en el histograma muestra que existe una frecuencia estable de personas hasta los 45 años, luego surge una caída constante en las frecuencias hacia los extremos mayores. En el panel de distribución de kernels, se observa que los pobres tienden a concentrarse en un rango de edad más joven en comparación con los no pobres, quienes tienen una distribución más amplia. Esto podría indicar que hay menos oportunidades económicas para los jóvenes, lo que podría limitar su desarrollo personal y profesional.



Punto 2:

Se observa en el boxplot que la mayor frecuencia se encuentra alrededor de los 7-12 años de educación, lo que sugiere que esta es la cantidad de años de educación de primaria y secundaria es lo más común en la población.

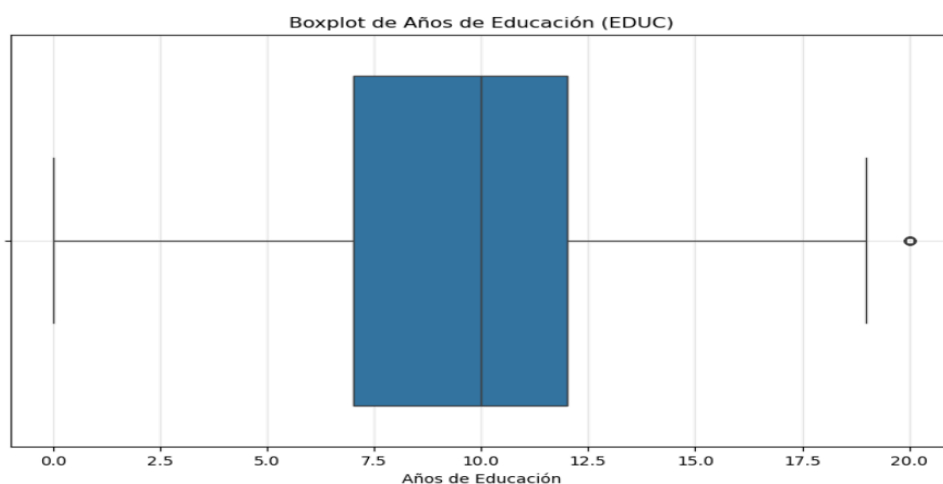
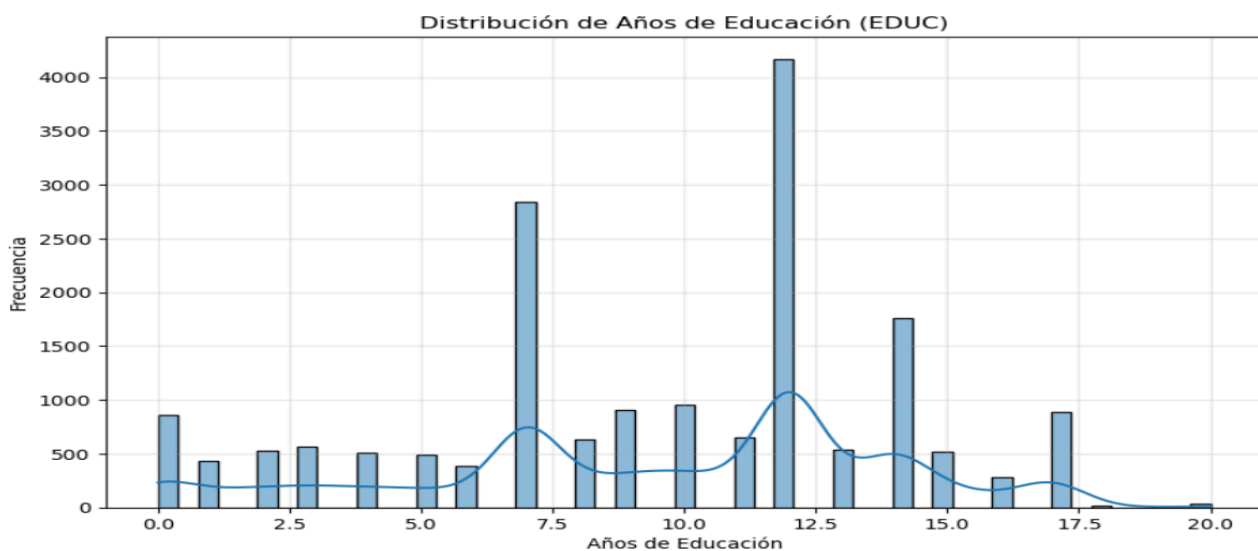
Así mismo se observan picos en los:

7 años (por la primaria completa)

12 años (el mayor pico, por la secundaria completa)

14 años (por la terciario completo)

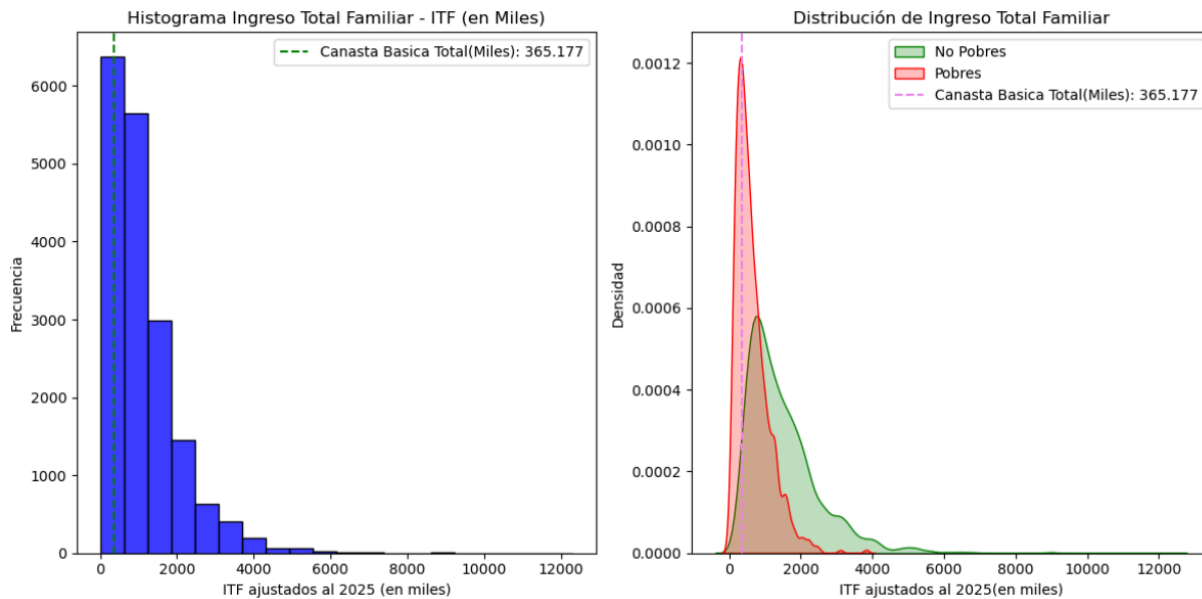
17 años (por la universidad completa).



Punto 3:

En el Histograma la mayor frecuencia se encuentra alrededor de los \$400.000 de ITF. Siendo muy cercana a la línea verde representa la Canasta Básica Total, que se ubica en 365.177 miles.

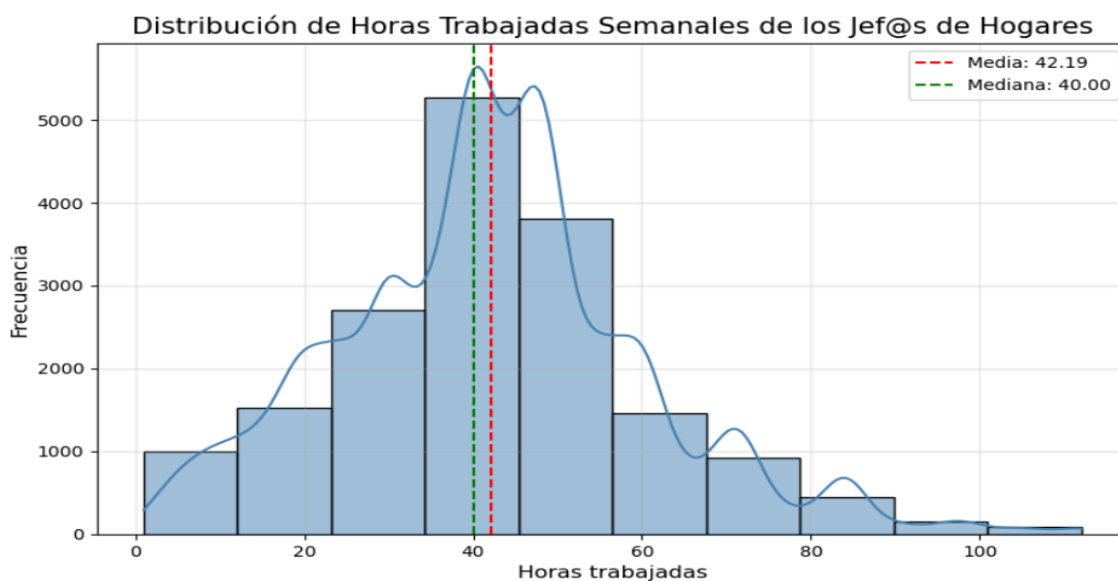
En el gráfico de densidad se visualiza nuevamente la gran relevancia de la pobreza en los mas jóvenes y adicionalmente se puede observar la mayor amplitud en los niveles de ingresos de los no pobres en relación la mayor frecuencia en los niveles de ingreso menores.



Punto 4:

El gráfico nos habla de que la mayoría de los jefes de hogar trabajan entre 20 y 80 horas por semana, con un pico alrededor de las 50 horas.

Como muestra de ello, vemos la media en 42,19 (muy cerca de la mediana de 40) y un desvío standar cercano a la 20 hs



Punto 5:

La región que se había elegido a desarrollar era la del NOA (Noroeste Argentino) que comprende las provincias de Jujuy, Salta, Tucumán, Catamarca, La Rioja y Santiago del Estero. En ambas encuestas, nos encontramos con 19006 registros. Con los pobres calculados en el TP2, tenemos un 47% de porcentaje promedio del periodo en el NOA. Registrándose un aumento del 6% entre el 1er trimestre de 2005 y el 1er trimestre de 2025, llegando al 50.86 en el 2025. Las variables originales, tomadas en cuenta de las encuestas, son 28. Requirieron limpieza de datos 22 variables de 2005 y 15 de 2025 para poder homogeneizarlas y unificarlas en una sola base de datos como se requería.

	2005	2025	Total
Cantidad observaciones	9244	9762	19006
Cantidad de observaciones con NAs en la variable “Pobre”	0	0	0
Cantidad de Pobres	4142	4965	9107
Cantidad de No Pobres	9107	4797	9899
Cantidad de variables limpias y homogeneizadas	50	43	50

Parte Dos:

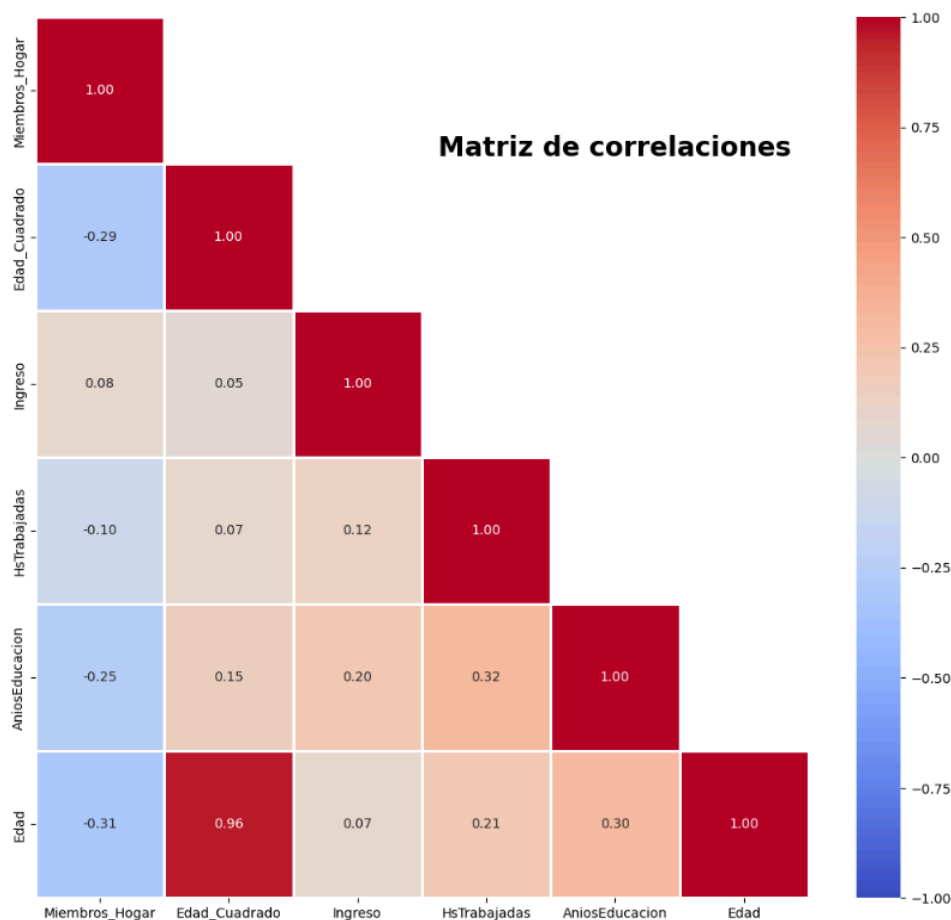
Pca:

Punto 1:

En la matriz de correlaciones (EPH) predominan relaciones bajas a moderadas, lo que indica que las variables aportan información complementaria.

Sobresalen patrones sociodemográficos conocidos: la correlación negativa entre Miembros del Hogar y Años de educación ($\rho \approx -0,25$) sugiere que los hogares más numerosos se concentran en estratos con menor capital educativo. En la misma línea, la relación negativa entre Miembros del Hogar y Edad podría interpretarse como una tendencia a que los hogares con mayor número de integrantes corresponden a núcleos familiares más jóvenes.

Desde el plano económico, el Ingreso familiar se asocia de manera positiva pero moderada con Años de educación ($\rho \approx 0,20$) y con Horas trabajadas ($\rho \approx 0,12$), lo que sugiere que mayor capital educativo y mayor participación laboral se traducen, en promedio, en más ingresos. Además, la correlación positiva entre Horas trabajadas y Años de educación es consistente con mejores oportunidades de empleo estable para quienes poseen mayor formación, y la relación moderada y positiva entre Edad y Horas trabajadas refleja trayectorias laborales más prolongadas a medida que avanza el ciclo de vida.



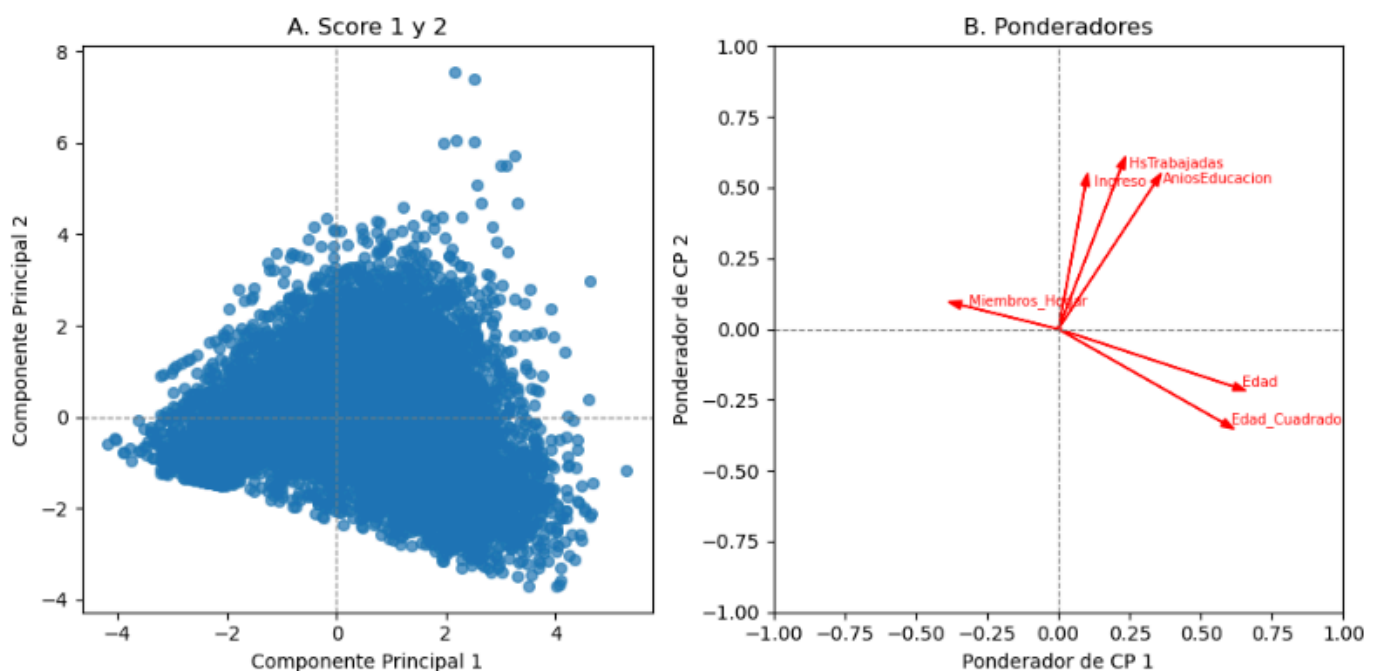
Punto 2:

En el gráfico de dispersión, cada punto representa el score o índice de un individuo en el nuevo espacio definido por los dos primeros componentes principales. La distribución de las observaciones adopta una forma aproximadamente triangular, lo que evidencia una considerable heterogeneidad en la muestra de individuos.

El plano puede dividirse en cuatro cuadrantes según el signo de los componentes. Se observa una menor densidad de puntos en el cuadrante inferior izquierdo.

El análisis del gráfico de ponderadores permite interpretar la composición de cada componente principal. Para el Componente Principal 1, se observa que las variables Edad y Edad al Cuadrado ejercen la influencia más significativa, ambas con un gran peso positivo. En contraste, la variable Miembros del Hogar contribuye de forma moderada y negativa, mientras que los ponderadores restantes tienen una influencia relativamente menor. Esta configuración de pesos indica claramente que el primer componente principal captura principalmente la dimensión etaria del individuo, funcionando como un índice de su etapa en el ciclo de vida.

Por otro lado, la estructura del Componente Principal 2 está dominada por las variables Ingreso Familiar, Horas Trabajadas y Años de Educación, las cuales presentan los mayores pesos positivos. La variable Miembros del Hogar, nuevamente, tiene una influencia menor en la conformación de este segundo componente. En consecuencia, el Componente Principal 2 puede ser interpretado como un indicador sintético que resume las características del capital socioeconómico y laboral del individuo.



Punto 3 y 4:

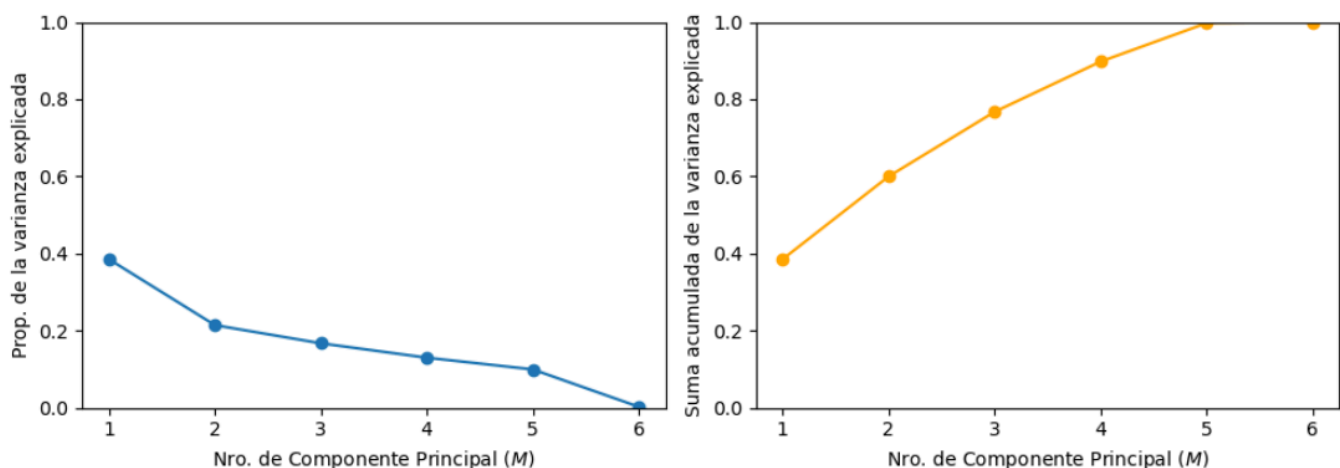
El análisis de la varianza explicada permite cuantificar la cantidad de información de los datos originales que es capturada por los componentes principales, siendo un paso crucial para determinar el número óptimo de componentes a retener.

El gráfico de proporción de varianza individual (gráfico de la izquierda) indica que el Componente Principal 1 captura aproximadamente el 38% de la variabilidad total, mientras que el Componente Principal 2 explica un 22% adicional y el Componente Principal 3, un 18%. A partir de este último, la contribución de cada nuevo componente disminuye considerablemente, sugiriendo que los tres primeros son los más significativos para caracterizar la estructura de los datos.

El gráfico de varianza acumulada confirma esta apreciación. Con los dos primeros componentes se logra explicar el 60% de la varianza total. Al incluir el tercer componente, esta cifra asciende a un 78%, representando una porción sustancial de la información original. Para capturar el 90% de la variabilidad, sería necesario retener cuatro componentes.

En conclusión, la utilización de 2 o 3 componentes principales constituye un balance eficiente entre la simplificación de la dimensionalidad y la pérdida de información.

Gráfico de varianza explicada

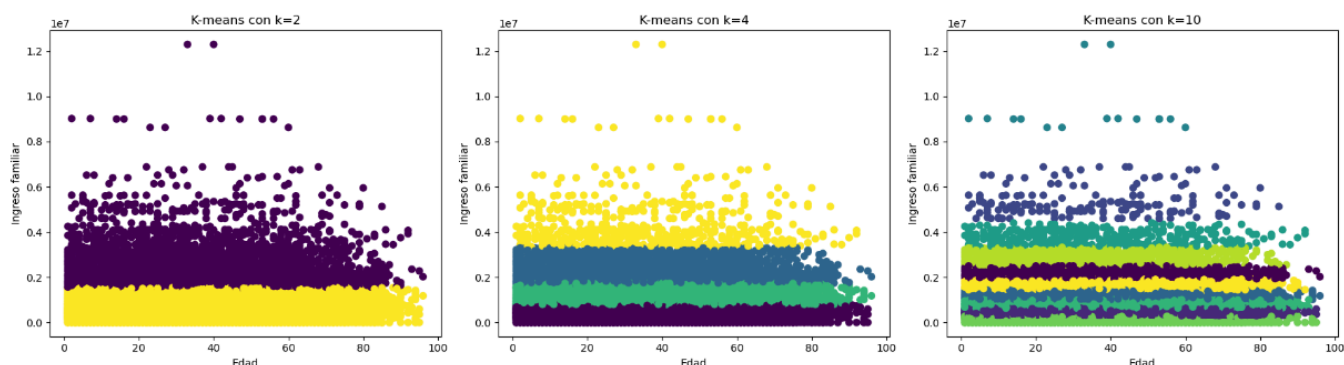


Clusters:

Punto 5 a:

Con $k=2$, $k=4$ y $k=10$ (con $n_{\text{init}}=20$) los gráficos muestran que K-means segmenta casi exclusivamente por ingreso familiar: aparecen bandas horizontales y la edad no aporta separación visible (los colores se mezclan a lo largo del eje x). Con $k=2$ se forman dos franjas amplias; con $k=4$ y $k=10$ esas franjas se refinan en cortes más finos del ingreso, pero la estructura sigue siendo “capas” paralelas y sensibles a outliers.

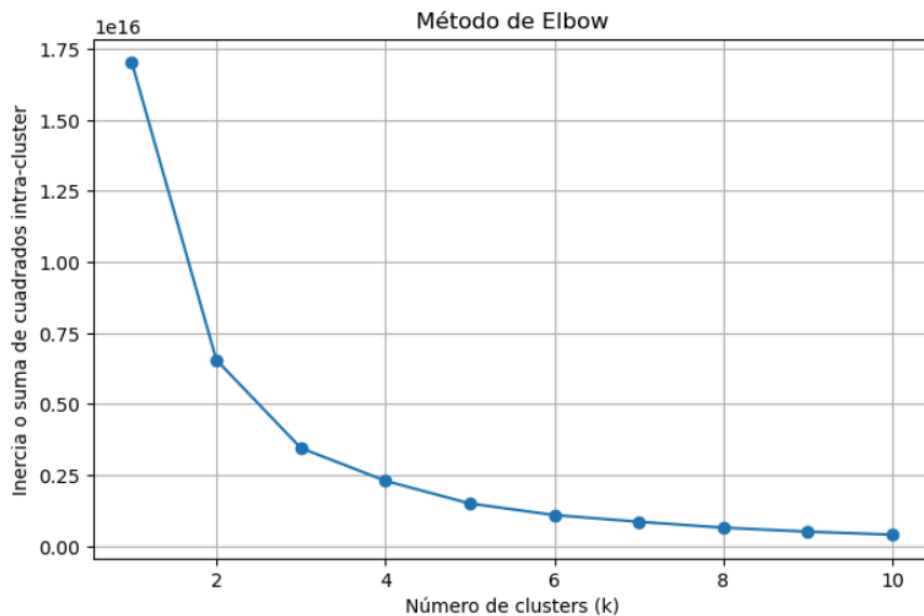
K-means no usa un umbral normativo de pobreza ni ajusta por tamaño/composición del hogar; solo minimiza distancia a centroides y asume clusters aproximadamente esféricos con varianza similar. Dada la asimetría y dispersión del ingreso, muchos casos cercanos a la frontera quedan mal asignados. En suma, $k=2$ produce un corte grueso por ingreso, no una clasificación válida de pobreza para la región.



Punto 5 b:

La curva de inercia desciende con mucha fuerza al pasar de $k = 1$ a $k = 2$ y aún muestra una mejora importante entre $k = 2$ y $k = 3$. Entre $k = 3$ y $k = 4$ la reducción sigue siendo apreciable pero claramente menor, y desde $k \geq 5$ la pendiente se aplana con caídas pequeñas y casi lineales, evidenciando rendimientos marginales decrecientes. El punto de mayor curvatura (donde se pasa de mejoras grandes a incrementos marginales) se ubica en torno a $k=4$, que ofrece un compromiso razonable entre parsimonia y compacidad de los conglomerados (con $n_{\text{init}}=20$, la solución es además estable).

Con $k \approx 4$, los clusters se alinean sobre todo con niveles de ingreso, por lo que es razonable asignarlos a categorías socioeconómicas más que a una dicotomía pobre/no pobre. Ahora bien, una clasificación socioeconómica robusta no se determina solo por ingreso: idealmente incorpora ingreso per cápita equivalente y otras dimensiones (educación, ocupación, calidad de la vivienda, activos, acceso a servicios, etc.). Aun así, el ingreso suele ser el proxy más operativo y eficaz por su disponibilidad y facilidad de medición; por eso $k=4$ ayuda a aproximar estratos, pero no garantiza por sí mismo una identificación normativa de pobreza.



Punto 6:

Aplicamos clustering jerárquico sobre edad e ingreso familiar. Bajo este método, cada observación inicia como su propio grupo y, en pasos sucesivos, se fusionan los pares más similares según una métrica de distancia (euclídea, salvo indicación en contrario) y un criterio de enlace (completo, único, promedio, centroide, Ward, etc.). Metodológicamente, para mejorar la estabilidad de las distancias es recomendable estandarizar las variables y considerar $\log(\text{ingreso})$ dada su asimetría antes de fijar definitivamente el corte.

El resultado se visualiza mediante un dendrograma, que es un “árbol” donde cada unión horizontal representa la fusión de dos clusters y la altura en el eje “y” indica la disimilitud al momento de unirse (a mayor altura, mayor diferencia entre grupos). Para determinar el número de clusters se “corta” el árbol a una altura fija; los saltos verticales largos sugieren una mayor diferencia entre observaciones. Un único dendrograma permite, además, inspeccionar distintas particiones al variar la altura de corte. Además, no podemos asegurar similitud entre dos observaciones en base al eje horizontal.

En el dendrograma obtenido para nuestra región se distinguen tres ramas: dos de ellas se fusionan entre sí a una altura intermedia y recién después se unen con la tercera a una altura claramente superior, lo que sugiere tres grupos principales. Por lo tanto, una partición inicial en tres clusters resulta razonable y puede refinarse en más subgrupos si se corta a una altura inferior.

