Trabajo Práctico N°2 - Ciencia de Datos

Histogramas, Kernels & Métodos No Supervisados usando 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

Enlace a la carpeta de Github: https://github.com/franciscacetra/TP2-Grupo10.git

Parte I: Creación de variables, histogramas, kernels y resumen de la base de datos final

1. Histograma de la variable $edad$ y distribución de kernels para pobres y no pobres. (Ver anexo - Figura 1).

   En el Panel A se presenta el histograma de la edad de la población relevada. La distribución exhibe una alta concentración entre los 10 y 20 años, y una disminución progresiva en la frecuencia a medida que aumenta la edad. Esta forma es consistente con la estructura demográfica típica de países con elevada proporción de población infantil y juvenil.

   El Panel B muestra las funciones de densidad estimadas mediante Kernel de la edad según condición de pobreza. Se observa que la población en situación de pobreza se concentra en edades más tempranas, con un pico pronunciado en torno a la niñez y adolescencia, donde se concentra al rededor del 30% de la densidad, mientras que la población no pobre presenta una distribución más homogénea y extendida hacia edades adultas y mayores. Esta diferencia sugiere una mayor incidencia de la pobreza en hogares con niños y adolescentes.

2. La variable $educ$ representa la cantidad de años de educación formal completados por los individuos, calculada considerando el nivel educativo más alto alcanzado, si se completó o no, y los años aprobados en caso de no finalizar un nivel. El promedio de 9,05 años indica que, en general, la población ha completado la educación primaria y cursado parte de la educación secundaria. La desviación estándar de 4,39 años refleja la variabilidad entre los individuos, desde quienes no han recibido educación formal (mínimo 0 años) hasta aquellos con estudios avanzados de nivel terciario o universitario (máximo 19 años). La mediana de 9 años confirma que la mayoría de los individuos ha completado la educación primaria y parte de la secundaria. En conjunto, estos resultados muestran que la población presenta un nivel educativo intermedio, con una proporción significativa de personas que han alcanzado niveles educativos más altos, y algunos que no han recibido educación formal.

3. Teniendo en cuenta que 100 pesos de 2005 son 135,768.99 en 2025, creamos una nueva columna *ingreso_total_familiar* que mantiene los ingresos de 2025 tal cual, y convierte los de 2005 a pesos de 2025 usando el factor de conversión.

   En la Figura 2 (Ver anexo), el Panel A muestra la distribución de los ingresos familiares. La línea roja punteada indica la línea de pobreza. Gran parte de los hogares está concentrada a la derecha de la línea (con ingresos más altos, son "no pobres"). Pero también observamos algunos grupos a la izquierda de la línea (hogares con ingresos por debajo de la línea de pobreza). Con este histograma en escala logarítmica se puede visualizar fácilmente la magnitud relativa de los hogares pobres y no pobres.

   En el Panel B se muestran las densidades del ingreso según la condición de pobreza. En naranja, los hogares clasificados como pobres y en azul, los no pobres. Al observar el Kernel, notamos que la mayoría de las observaciones de los hogares pobres se encuentran del lado derecho de la línea de pobreza, lo que podría indicar que está mal trazada la línea o que si hubiéramos utilizado una función de densidad uniforme o Epanechnikov en lugar de Gaussiana estarían mejor representados los datos.

4. Para el jefe del hogar (CH03=1), calculamos la cantidad de horas trabajadas y las guardamos en una variable llamada $horastrab$ (sumando la ocupación principal y otras ocupaciones). Las estadísticas descriptivas presentan un promedio de aproximadamente 32,6 horas por semana, con una desviación estándar de 25,8 horas, lo que indica una considerable variabilidad entre los individuos. El valor mínimo registrado es 0 horas, correspondiente a jefes de hogar que no trabajaron durante la semana de referencia, mientras que la mediana es de 40 horas, mostrando que la mayoría de los jefes de hogar trabaja jornadas completas. El máximo observado es de 100 horas, límite establecido para descartar valores atípicos poco realistas. En conjunto, estos resultados reflejan que la mayor parte de los jefes de hogar trabaja jornadas habituales, aunque existen casos de alta intensidad laboral.
5. Base final de la región GBA, (Ver anexo - Tabla 1).

Parte II: Métodos No Supervisados

1. Matriz de correlaciones con las variables de Edad, $edad^2$, $educ$ (Nivel educativo), Ingreso familiar total, $TOT\ P12$ (Cantidad de miembros del hogar) y $horastrab$ (Horas trabajadas). (Ver anexo - Figura 3)

En estas matrices de correlaciones, obviamente hay una correlación altísima y positiva entre $edad$ y $edad^2$en ambos años (0.96), porque una es simplemente la transformación de la otra (elevada al cuadrado). Esto es normal, no aporta información relevante. En 2005, la correlación entre *edad* y *nivel educativo* era prácticamente nula (0.09), lo que refleja que la educación estaba relativamente distribuida de manera heterogénea entre las distintas generaciones: tanto jóvenes como adultos mayores podrían tener trayectorias educativas muy variadas. Esto podría explicarse en parte por la expansión previa del sistema educativo, que ya había incorporado a amplios sectores sociales, y por la diversidad socioeconómica del GBA, que diluye la relación entre más años de vida y mayor escolaridad. Por otro lado, en 2025 hay una correlación positiva moderada entre estas dos variables (0.48). Esto lo podemos interpretar como que, en la muestra más nueva, los individuos de mayor edad tienen en promedio más años de educación formal acumulados, tal vez por una expansión educativa en años anteriores. La variable $ingreso\ total\ familiar$ tiene correlaciones muy bajas con el resto de las variables en los dos años. Todos los índices de correlación son menores o iguales a 0.20. Esto sugiere que el ingreso familiar no se explica de manera lineal ni por la edad, ni por la educación, ni por el tamaño del hogar de manera fuerte. La relación parece ser más compleja y estar claramente mediada por otros factores. Las correlaciones entre la variable del número de miembros en el hogar con el ingreso, educación o edad son bajas en ambos años. Esto podría reflejar que el tamaño del hogar tiene mucha variabilidad y no está directamente asociado con estas características incluidas en la matriz. Por otro lado, en 2005 la variable de horas trabajadas correlaciona levemente con edad (0.26) y muy poco con la educación o el ingreso. En 2025, la correlación con la edad sube a 0.32 y con educación a 0.23, sugiriendo que personas con más años de educación y un poco más de edad tienden a trabajar más horas.

A. PCA

2. PCA con ingreso (Ver anexo - Figura 4).

Este gráfico de dispersión muestra las seis variables originales ($edad$, $edad^2$, $educación$, $ingreso\ total\ familiar$, $IX\ TOT$[1] y $horastrab$) pero ahora resumidas en dos dimensiones. El eje X es el primer componente principal (CP1) y el eje Y es el segundo componente principal (CP2). Los componentes son combinaciones lineales de las variables originales que capturan la máxima varianza posible. CP1 es la dirección donde hay más varianza de la información y CP2 concentra la segunda mayor variabilidad, pero en una dirección ortogonal a CP1. La dispersión de estos puntos en el gráfico muestra cómo se distribuyen los scores de las observaciones en este espacio "más reducido". Las observaciones más lejanas podrían ser outliers o casos atípicos (alejados de la nube concentrada de puntos). Se ve que la mayoría de los hogares se concentran más cerca del eje X (es decir, tienen valores bajos en CP2). Algunos hogares se alejan mucho en CP2, esos podrían ser casos con características extremas, como ingresos familiares muy altos o tal vez jornadas laborales muy largas. Además, como las variables fueron estandarizadas, los scores positivos en un componente reflejan observaciones con valores por encima de la media en las variables que más pesan en ese componente, mientras que los scores negativos indican valores por debajo del promedio. Este gráfico no nos da información directa de qué variables (de las originales) están asociadas a cada componente. Nos ayuda a ver patrones generales, aunque no es muy informativo.

3**.** Gráfico con flechas de los ponderadores de PCA, (Ver anexo - Figura 5).

En este biplot de PCA podemos ver combinados los scores (los hogares representados en los puntos azules) en los dos componentes principales junto con los loadings (las flechas rojas) que son los ponderadores que muestran el peso y la dirección de cada variable en esos componentes. El primer componente (CP1) parece estar asociado principalmente con la variable $edad$ (y $edad^2$), mientras que el segundo (CP2) se relaciona con las variables socioeconómicas: educación ($educ$), ingreso familiar total y cantidad de

[1] Esta variable indica la cantidad o el número de miembros del hogar.

miembros del hogar o tamaño del hogar ($TOT\ P12$). Las flechas de educación, cantidad de miembros del hogar e ingreso se orientan en la misma dirección, mostrando que tienen una correlación positiva fuerte. Por otro lado, la variable de horas trabajadas ($horastrab$) aparece en la mitad de la nube de observaciones, está vinculada con la edad ($Edad$ y $edad^2$) y con los factores socioeconómicos. Entonces podemos ver como en este "resumen" de los dos componentes principales se resume la "oposición" entre dimensión etaria y la dimensión educativa y económica de los hogares.

4. Realizamos una tabla con la proporción de varianza explicada por cada componente (Ver anexo - Tabla 2). En la Figura 6 (Ver anexo), el gráfico de la izquierda podemos ver que el primer componente principal (CP1) explica alrededor del 35% de la varianza total, mientras que el segundo componente principal explica cerca del 20%. En conjunto, los primeros dos componentes concentran más del 55% de la variabilidad de los datos. Los componentes que siguen (CP3 y CP4) todavía aportan información relevante (16% y 15%, respectivamente), mientras que el quinto componente apenas contribuye un 13%. El CP6 ya casi no explica varianza (0,005%).

En el gráfico de la derecha, que muestra la varianza acumulada, confirmamos que con los primeros cuatro componentes se logra capturar alrededor de 86% de la variabilidad de los datos, y con cinco componentes prácticamente se alcanza el 100%. Esto significa que, aunque las seis variables originales aportan información, gran parte de la estructura de los datos la podemos resumir en menos dimensiones (2 a 4 componentes) sin perder demasiada información.

B. Cluster

5. a) En la Figura 7 (Ver anexo), se ve que aplicamos el método de clusters con k-medias (utilizando $k = 2$, $k = 4$ y $k = 10$) para las variables normalizadas *Ingreso total familiar* y *edad2* (edad al cuadrado) y eliminamos los outliers de ITF porque estos valores extremos distorsionan la estructura de los datos y afectan la posición de los centroides en el clustering. Además, marcamos en rojo el valor real de la variable *ingreso_necesario* en el eje X para tener un punto de referencia de la línea de pobreza. A

partir de esta herramienta de visualización, observamos que con $k = 2$, el algoritmo divide los datos en dos clusters, pero la separación parece estar más influenciada por la edad que por el nivel de ingreso. Por lo tanto, los clusters obtenidos no coinciden de manera confiable con las categorías de personas pobres y no pobres, y no se puede usar esta clusterización para clasificar correctamente la pobreza en la región. Con $k = 4$, se producen grupos amplios y fáciles de interpretar. Se distinguen dos grupos de hogares de ingresos bajos y medios con edades bajas, un grupo de ingresos altos y un clúster que abarca más niveles de ingresos y edades más altas. Esta solución ofrece perfiles generales pero manejables. Con $k = 10$, los grupos se fragmentan en categorías más específicas. Los hogares se dividen en múltiples subgrupos según ingresos y edad, lo que permite mayor detalle pero complica la interpretación y resta claridad analítica. Si bien sabemos que aumentar la cantidad de clusters aumenta la precisión y mejora el ajuste, reduce la simplicidad y la fácil interpretación para el lector que es un principio importante en la visualización de datos.

b) Por curiosidad hicimos dos gráficos del método Elbow, uno incluyendo y otro excluyendo los outliers de ITF. Cuando graficamos el método Elbow sin outliers (Ver anexo - Figura 8), el k óptimo cambia respecto al gráfico con outliers. Con outliers, el codo se sugiere entre k=4 y k=5, mientras que sin ellos el codo se ubica en k=3. Esto es esperable, porque los outliers incrementan la variabilidad de los datos y el algoritmo requiere más clusters para agruparlos. Al removerlos, la estructura de los datos se simplifica y alcanza con menos clusters para representar adecuadamente la dispersión. Sin embargo, aunque estemos usando ITF, que tiene que ver con la pobreza, los clusters no separan pobres de no pobres porque la división se hace en función de la distancia conjunta entre Edad e ITF, sin usar la etiqueta de pobreza. Como los valores de ITF de pobres y no pobres se superponen, no importa qué k usemos, los clusters nunca coincidirán perfectamente con la condición de pobreza.

6. Cluster jerárquico (Ver anexo - Figura 9)**.** Un dendrograma es un diagrama en forma de árbol que muestra cómo se agrupan progresivamente las observaciones en un análisis de clustering jerárquico, sin necesidad de fijar de antemano el número de clusters. En este

caso, utilizamos el método de Ward con distancia euclídea. El eje vertical representa el nivel de disimilitud: fusiones más bajas indican grupos más similares, mientras que las uniones más altas muestran clusters más heterogéneos. Para facilitar la lectura, se muestran solo los últimos 25 merges en lugar de cada observación individual.

Anexo

Figura 1. Panel A: Histograma de edad y Panel B: Distribución de kernels por condición de pobreza

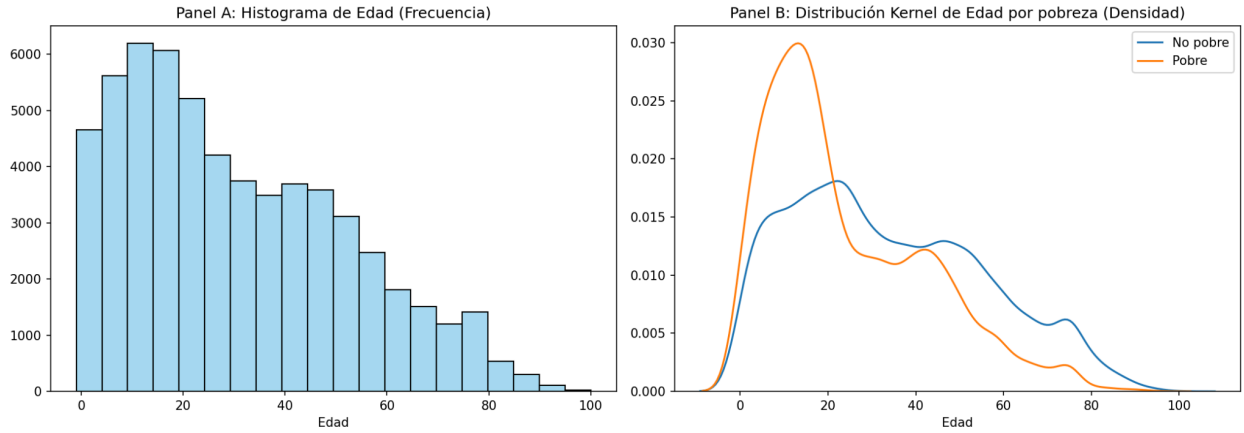

Figura 2. Panel A: Histograma ITF y Panel B: Kernel de ITF para pobres y no pobres

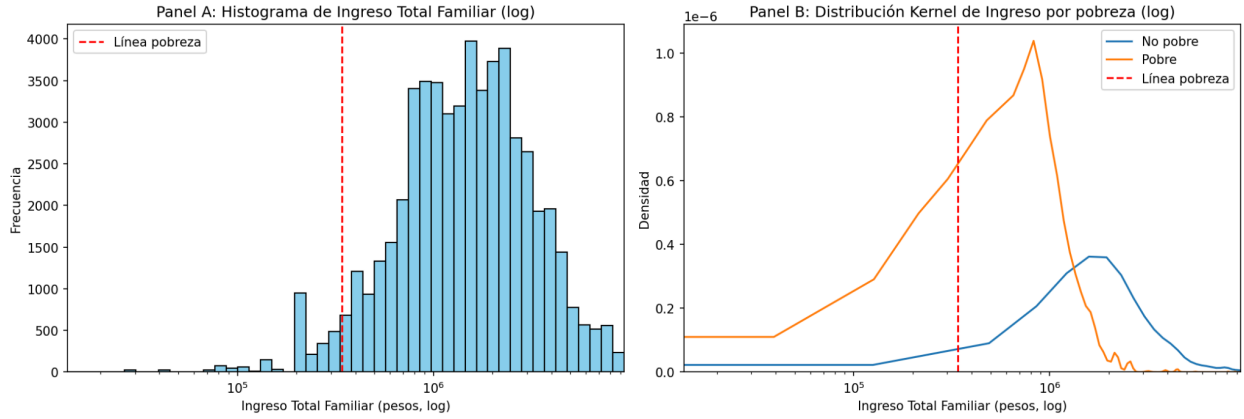

Tabla 1. Resumen de la base final para la región GBA

	2005	2025	Total
Cantidad de observaciones	40969	17968	58937
Cantidad de observaciones con NAs en la variable "Pobre"	112	2890	3002
Cantidad de Pobres	14356.0	5969.0	20325
Cantidad de No pobres	26501	9109	35610
Cantidad de variables limpias y homogeneizadas	175	73	248

Figura 3. Matrices de correlaciones de 2005 y 2025

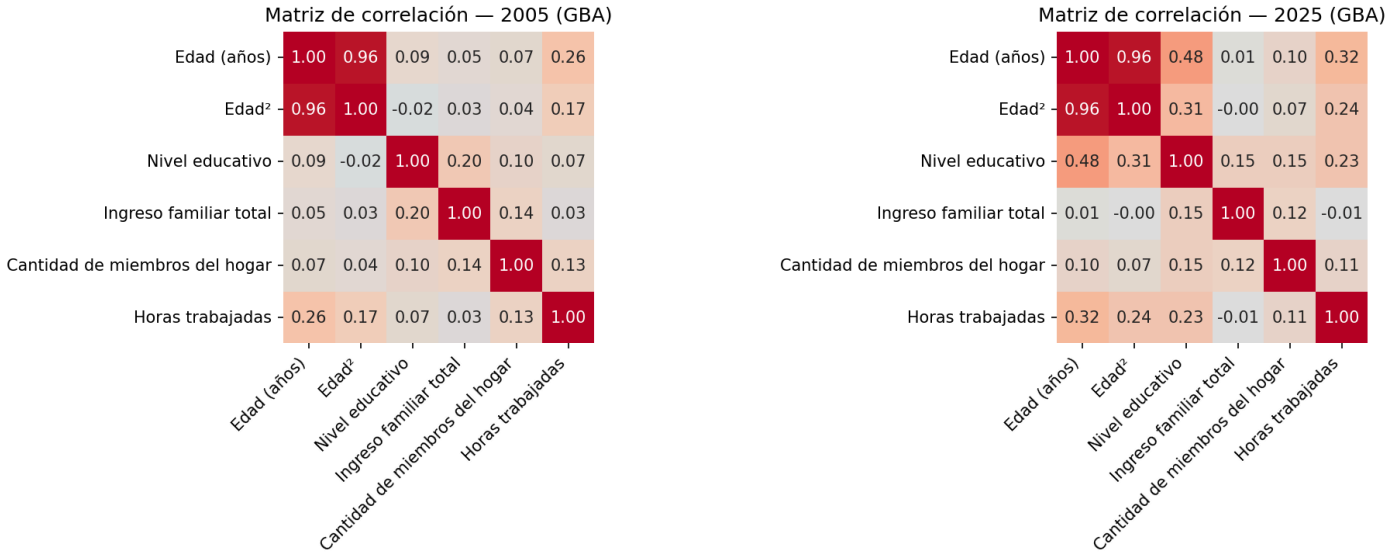

Figura 4. Gráfico de dispersión del primer componente principal (CP1) vs. el segundo componente principal (CP2).

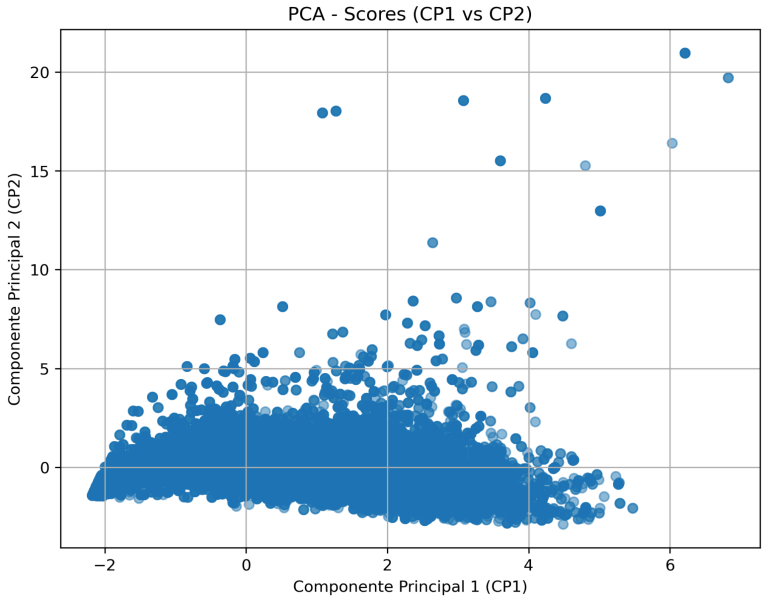

Figura 5. Gráfico de dispersión de PCA con flechas de los ponderadores (loadings)

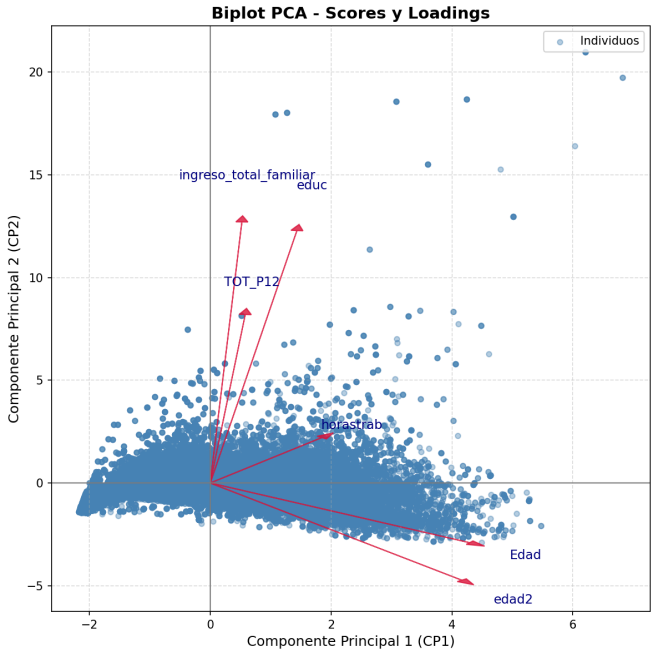

Tabla 2. Proporción de varianza explicada por cada componente.

Componente	Proporción	Acumulada
1	0,354	0,354
2	0,197	0,551
3	0,162	0,713
4	0,149	0,862
5	0,133	0,995
6	0,005	1

Figura 6. Gráficos de la proporción de varianza explicada por cada uno de los 6 componentes y varianza acumulada

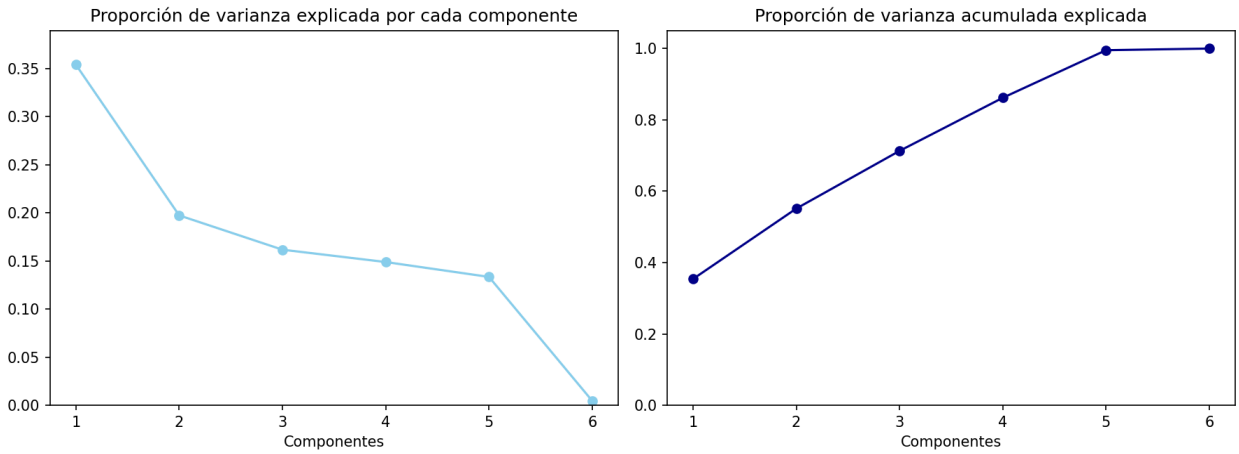

Figura 7. Gráficos Clúster k-media con $k = 2$, $k = 4$ y $k = 10$ sin outliers con vars estandarizadas.

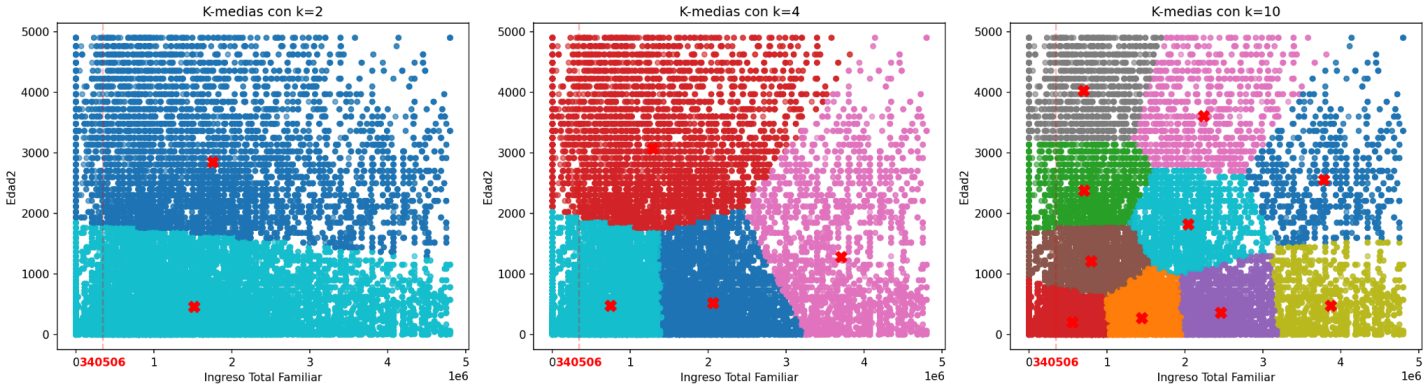

Figura 8 Inspección visual Elbow (quitando outliers).

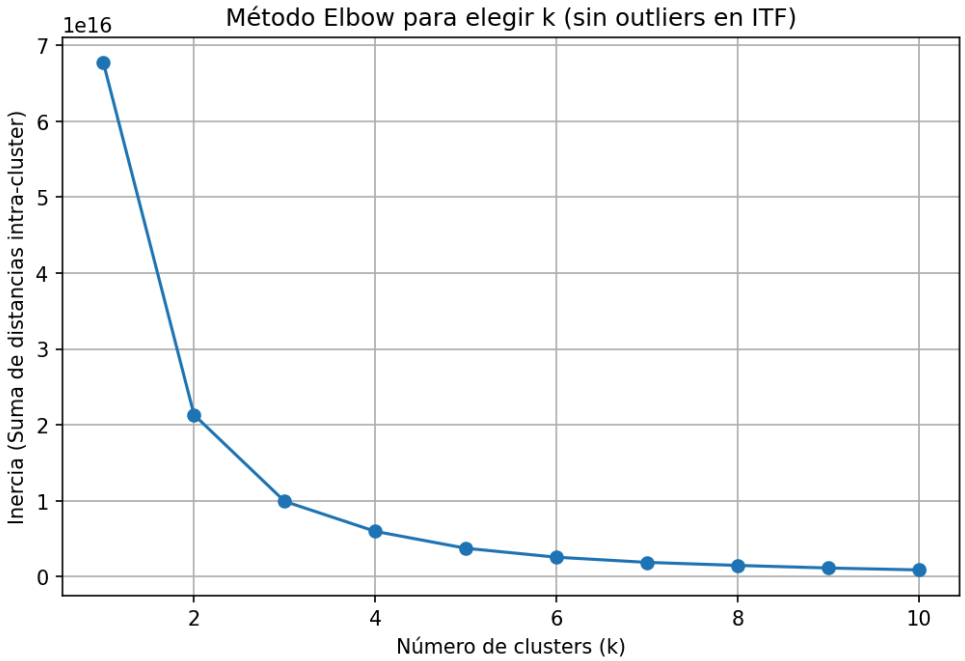

Figura 9. Dendrograma.