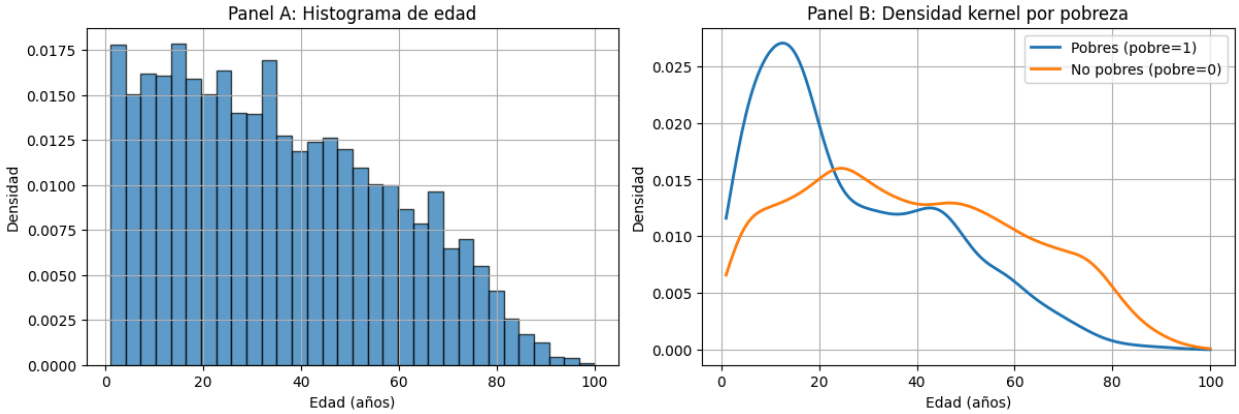
Mora Barrientos, Máxima Dinger Marciale y Florencia Falaq

TRABAJO PRÁCTICO N° 2

Link al repositorio de GitHub: https://github.com/morabarrientos/TP2
Link a carpeta con las tres bases usadas (también subidas en GitHub):
https://drive.google.com/drive/folders/18dxczgb1mmMdeMzhHZPY52ZovVjsW3Yy?usp=sharing

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

1) Creación de variable "*edad2*", histograma de la variable edad en un panel A, y a la par una distribución de kernels para los pobres y no pobres en un panel B



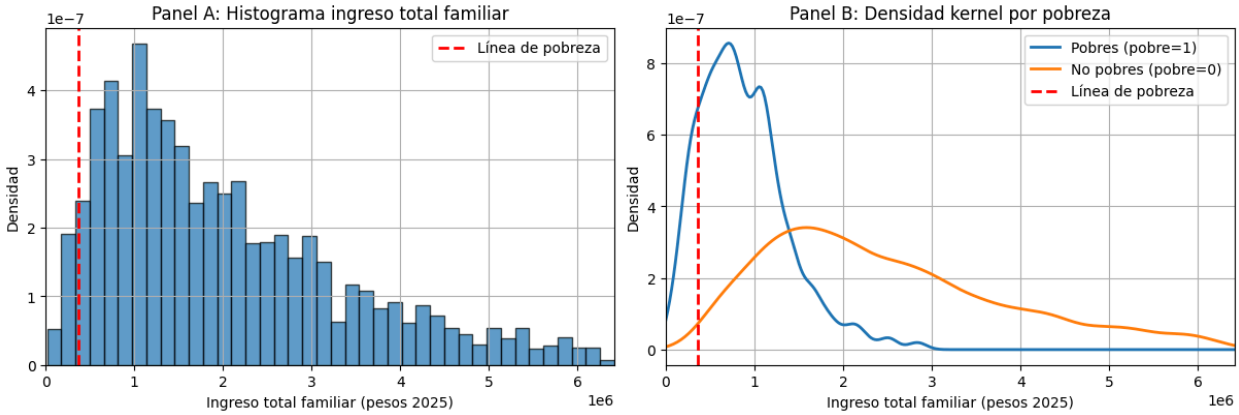
En el Panel A, la distribución etaria muestra una alta concentración en edades tempranas, que desciende de manera sostenida hacia las edades más avanzadas. Esto refleja una población con predominio de personas jóvenes y menor presencia relativa de adultos mayores. En el Panel B, la densidad de los hogares pobres se concentra especialmente en la infancia y adolescencia, mientras que entre los no pobres la distribución se desplaza hacia edades adultas. Esta diferencia sugiere que la pobreza se asocia de forma más marcada a estructuras familiares jóvenes, mientras que los no pobres concentran mayormente individuos en edad productiva.

2) Creación de variable *educ* & estadística descriptiva (promedio, sd, min, p50, max)

	Cantidad	Promedio	Desviación Estándar	Valor Mínimo	25%	Mediana	75%	Valor Máximo
educ	7096.000000	10.454622	5.222780	0.000000	4.000000	12.000000	14.000000	21.000000

La variable *educ*, medida en años de educación formal, presenta un promedio de aproximadamente 10,5 años y una mediana de 12, lo que refleja que gran parte de la población alcanzó el nivel secundario completo. El rango intercuartílico se ubica entre 4 y 14 años, mostrando la heterogeneidad entre quienes tienen primaria incompleta y quienes avanzaron hacia estudios terciarios o universitarios. Se observan casos sin escolaridad (0 años), mientras que los valores superiores a 30 años de educación fueron considerados implausibles y recodificados como NA, de modo que el valor máximo real es 21. En conjunto, la distribución es consistente con la diversidad educativa de la población y más robusta tras la depuración de outliers.

3) Actualización de variable *ingreso_total_familiar* con el total de ingresos habituales (ITF), histograma de la variable *ingreso_total_familiar* y las distribuciones de kernels para pobres y no pobres.



En los gráficos se muestra la distribución del ingreso total familiar expresado en millones de pesos de 2025. Para mejorar la visualización se optó por recortar el eje en el percentil 95 (≈ 6,4 millones), ya que los valores del 5% superior distorsionaban la escala y dificultaban la interpretación de la estructura central. Con este recorte puede observarse con mayor claridad que los hogares pobres se concentran en niveles de ingreso cercanos a la línea de pobreza, mientras que los no pobres presentan una distribución más amplia, con mayor densidad entre 1 y 3 millones de pesos y una cola extendida hacia ingresos más altos. Esta decisión permite resaltar las diferencias entre ambos grupos y dar una idea más precisa de la desigualdad sin que los valores extremos opaquen el comportamiento de la mayoría de los hogares.

4) Variable *horastrab* + estadística descriptiva (promedio, sd, min, p50, max) de dicha variable.

	Cantidad	Promedio	Desviación Estándar	Mínimo	Mediana (p50)	Máximo
horastrab	4527.000000	29.061630	23.106431	0.000000	35.000000	84.000000

En este inciso se ajustaron las variables *PP3E_TOT* y *PP3F_TOT* para excluir valores no plausibles: todos aquellos menores a 0 o mayores a 84 horas semanales fueron recodificados como NA. Consideramos que trabajar 84 horas semanales implica un promedio de 12 horas diarias o 16,8 si solo se contemplan los días hábiles, por lo que superar este umbral debe deberse a errores de carga. De esta manera, la variable total de horas trabajadas (*horastrab*) descarta automáticamente cualquier caso inválido, evitando que distorsione las medidas estadísticas. Los resultados muestran un promedio de 29 horas semanales y una desviación estándar alta (23 horas), lo que evidencia gran dispersión en la cantidad de horas declaradas. La mediana (35 horas) supera al promedio, lo que refleja la influencia de casos con 0 horas trabajadas. El valor máximo, fijado en 84 horas, representa el tope plausible considerado en el análisis.

5) Tamaño de la base de datos para nuestra región con las variables originales unificadas.

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

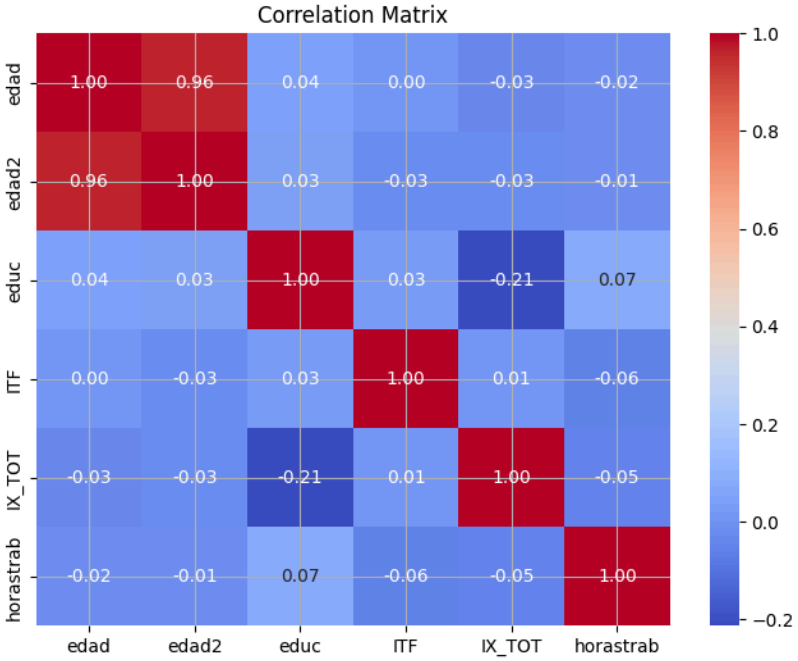
	2005	2025	Total
Cantidad de observaciones	9484	7181	16665
Cantidad de observaciones con NAs en la variable "Pobre"	0	0	0
Cantidad de Pobres	2438	1334	3772
Cantidad de No Pobres	6933	2975	9908

Nota: la cantidad de pobres y no pobres fue calculada a partir de la variable de **Pobre** que creamos en el trabajo práctico 1.

La tabla muestra que en la región 1 se dispone de 16.665 observaciones en total, distribuidas entre 9.484 en 2005 y 7.181 en 2025. En ambos años no hay valores faltantes en la variable pobre. En cuanto a la condición de pobreza, se registran 3.772 casos pobres frente a 9.908 no pobres, lo que indica una mayoría relativa de hogares no pobres. La comparación entre años refleja una reducción en el número absoluto de observaciones de pobres hacia 2025.

Parte II: Métodos No Supervisados

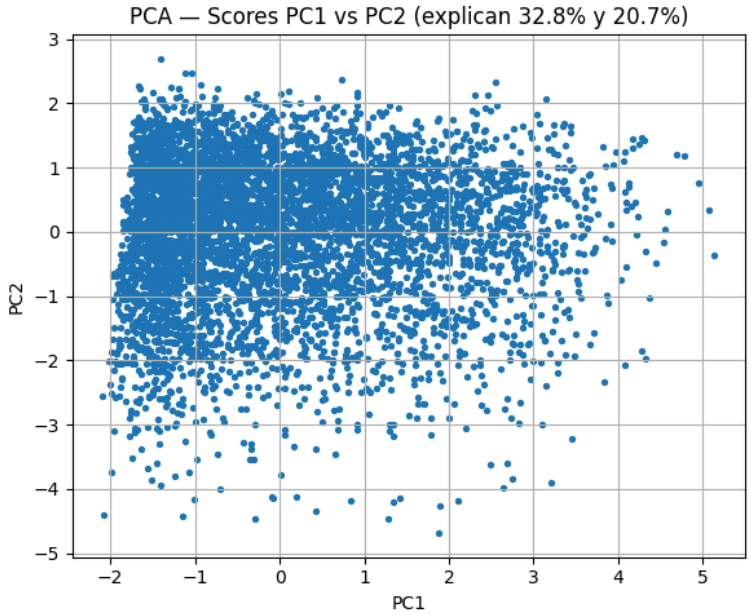
1. Matriz de correlaciones con los seis predictores (edad, edad2, educ, ingreso_total_familiar (ITF), el número de miembros en el hogar (2005=IX_TOT y 2025=IX_Tot) y horastrab) para nuestra región.



El gráfico confirma que la relación más fuerte se da entre *edad* y *edad2*, con una correlación cercana a 1, lo que es esperable porque una es una transformación de la otra. El resto de las variables muestran correlaciones muy bajas entre sí, lo que sugiere que aportan información relativamente independiente. Ni el nivel educativo, ni las horas trabajadas, ni el tamaño del hogar presentan asociaciones lineales fuertes con el ingreso total familiar, al menos en este corte simple. En conjunto, la matriz refleja que no hay multicolinealidad alta entre los predictores, salvo la obvia entre *edad* y *edad2*.

A. PCA

2. <u>PCA con ingreso</u>: Apliquen PCA a las seis variables seleccionadas para esta parte. Recuerde primero estandarizar las variables como vimos en la tutorial. En un gráfico de dispersión muestren los índices (scores) calculados del primer y segundo componente de PCA y comente los resultados.



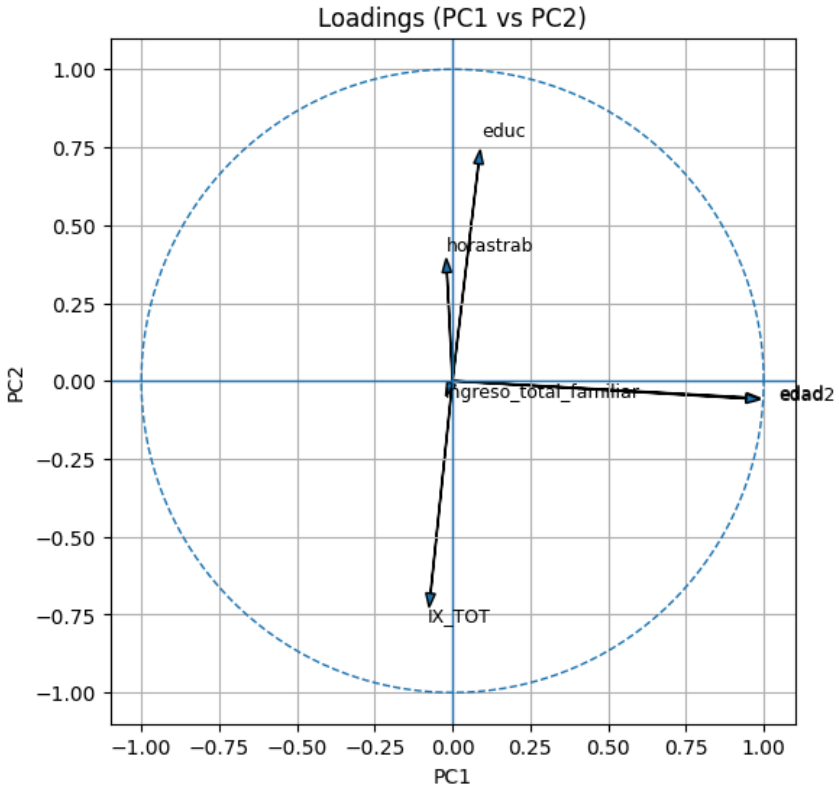
El análisis de componentes principales (PCA) realizado sobre las seis variables seleccionadas (*edad*, *edad2*, *educación*, *ingreso total familiar*, *número de miembros en el hogar* y *horas trabajadas*) permitió reducir la dimensionalidad de la información conservando una proporción considerable de la varianza total. En particular, el primer componente explicó un 32,8 % de la variabilidad de los datos y el segundo un 20,7 %, lo que en conjunto representa más de la mitad de la información contenida en las variables originales.

La representación gráfica de los scores en el plano definido por el primer y segundo componente muestra una dispersión amplia de los casos, sin una estructura de clústeres claramente definidos. No obstante, se observan valores atípicos hacia la derecha del eje del primer componente y hacia los valores negativos del segundo, lo que refleja la presencia de observaciones con características extremas en términos de ingreso, educación u horas trabajadas.

En términos interpretativos, el primer componente se asocia principalmente con un eje socioeconómico, dado que las variables de educación e ingreso tienden a concentrar la mayor parte de la varianza. El segundo componente, en cambio, parece vincularse con la estructura del hogar y la carga laboral, ya que incorpora la variación relacionada con el número de miembros en la vivienda y las horas trabajadas.

En síntesis, los resultados evidencian que las principales fuentes de heterogeneidad en la población analizada responden a diferencias en el nivel socioeconómico y en la organización demográfica-laboral de los hogares. La ausencia de agrupamientos bien definidos sugiere, además, que estas dimensiones atraviesan a los casos de manera continua más que segmentada.

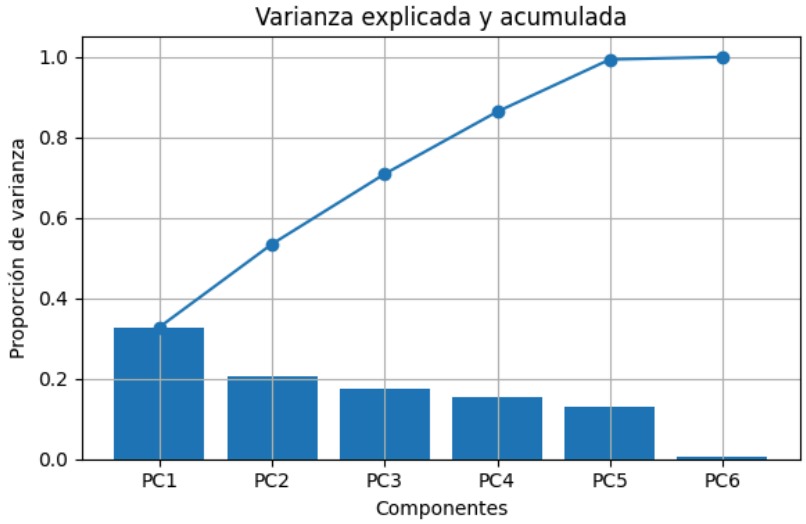
3. Gráfico con flechas de los ponderadores (loading) de PCA para el primer y segundo componente. Pesos que le dan a cada variable utilizada.



El gráfico de *loadings* evidencia las variables que más influyen en la construcción de los dos primeros componentes principales. En el PC1, la mayor carga corresponde a *edad2*, acompañada en menor medida por el *ingreso total familiar*, lo que indica que este eje representa principalmente diferencias de tipo demográfico y socioeconómico. En el caso del PC2, se destacan las cargas positivas de *educación* y *horas trabajadas,* mientras que el tamaño del hogar (*IX_TOT*) aparece con una contribución negativa. Esto sugiere que el segundo componente refleja una dimensión asociada al capital humano y la inserción laboral en tensión con la estructura familiar. En conjunto, ambos componentes permiten distinguir dos ejes de variación relevantes: uno ligado a edad e ingresos y otro a la combinación de

educación, empleo y tamaño del hogar, capturando de manera complementaria las principales fuentes de heterogeneidad en la muestra.

4. Proporción de la varianza explicada para cada uno de los seis componentes.

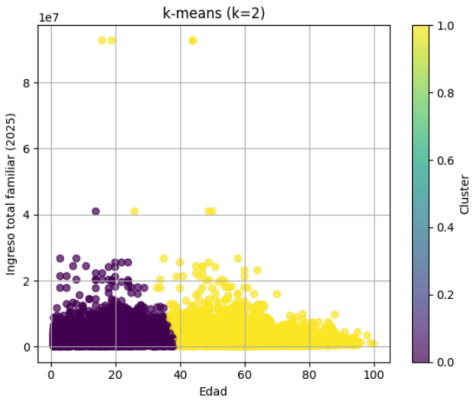


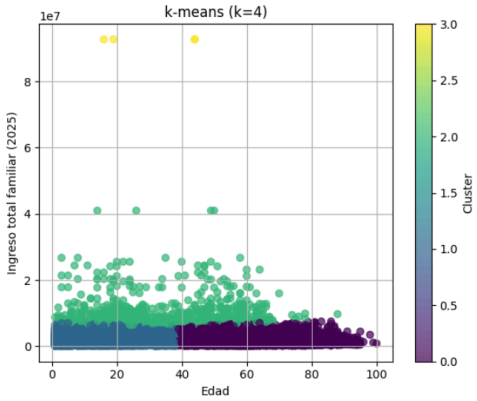
El gráfico muestra que el primer componente principal (PC1) explica un 32,81% de la varianza y el segundo (PC2) un 20,67%, alcanzando juntos más del 50% de la variabilidad total. A partir del tercer componente, la varianza explicada es mucho menor y se distribuye de forma más pareja. Esto indica que gran parte de la información de las seis variables originales puede resumirse en los dos primeros componentes.

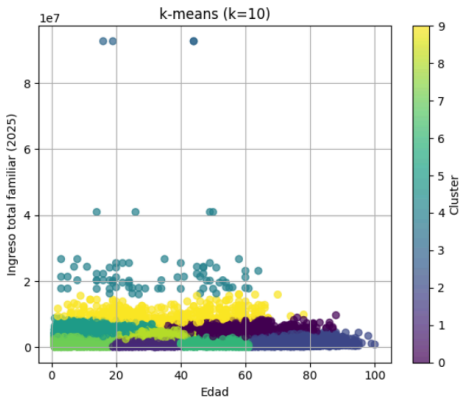
B. Cluster

5. <u>Cluster k-medias</u>:

a. Algoritmo con k = 2, k = 4 y k = 10 usando n_init =20. Gráfico de los resultados usando edad e ingreso familiar. Interpretación y análisis de si e el algoritmo con k = 2 puede separar correctamente a las personas pobres y no pobres en nuestra región.





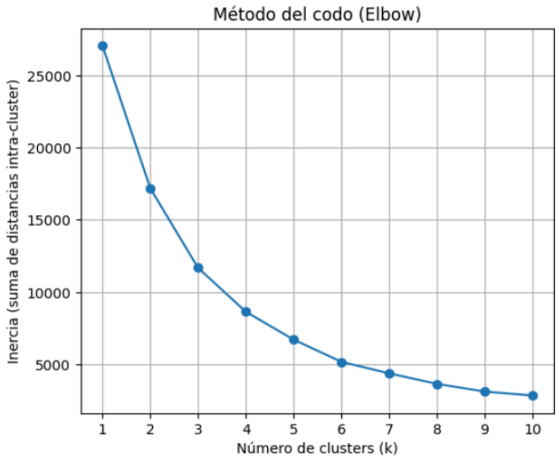


pobre	0	1
kmeans_k2		
0	5107	2681
1	4700	1031

	no_pobre(0)	pobre(1)
kmeans_k2		
0	0.656	0.344
1	0.820	0.180

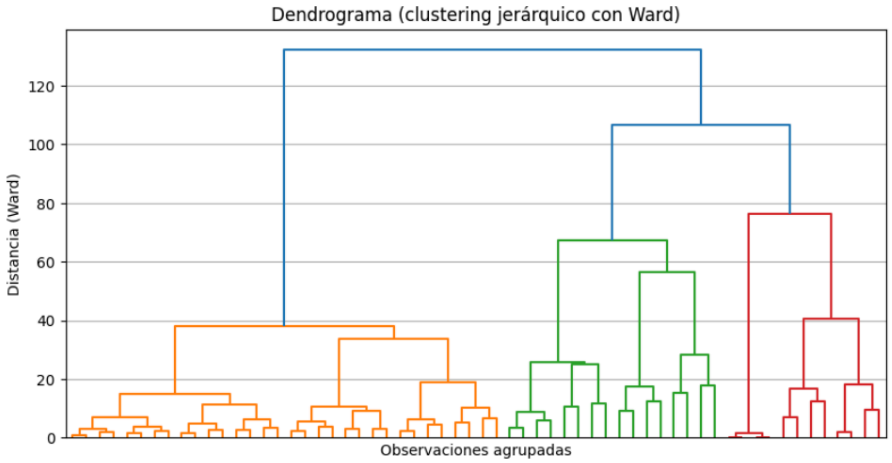
En los dispersogramas K-means segmenta prácticamente solo por ingreso: los cortes entre colores forman bandas horizontales y la edad apenas incide (mayor dispersión en jóvenes y leve caída del ingreso medio a edades altas). Con k=2 la nube queda partida en un estrato bajo y otro medio/alto. Al comparar esa partición con la etiqueta externa pobre/no pobre se observa solapamiento: en el cluster 0 hay 65,6% de no pobres y 34,4% de pobres; en el cluster 1, 82,0% de no pobres y 18,0% de pobres. En consecuencia, con k=2 K-means no separa correctamente a pobres y no pobres; lo que aprende son niveles de ingreso, no la frontera normativa de pobreza. Con k=4 la estructura se subdivide en franjas (baja, media-baja, media y un grupo pequeño de altos/outliers), lo que mejora la descripción pero mantiene el mismo criterio (ingreso). Con k=10 la segmentación se vuelve muy fina y pierde interpretabilidad social. Consistente con el "codo", la mayor ganancia aparece al pasar de 1→2 y, en menor medida, 2→4; luego los retornos son decrecientes. En síntesis, K-means es útil para estratificar por ingreso, pero no para clasificar pobreza/no pobreza.

b. Gráfico de alguna medida de disimilitud para k = 1 hasta k = 10. Usando la inspección visual de Elbow para identificar el número óptimo de clusters en nuestra región. Identificación sobre si dicha cantidad de grupos nos ayudaría a distinguir entre pobres y no pobres o entre distintas clases socioeconómicas.



La inercia (suma de distancias intra-cluster) cae muy fuerte al pasar de k=1 a k=2 y sigue bajando de forma marcada en k=3. Desde k=4 la pendiente se aplana y aparecen rendimientos decrecientes. Con el criterio del "Elbow", el quiebre está entre k=2 y k=3 , por lo que k=3 es un punto razonable: reduce bastante la inercia sin perder interpretabilidad. Con ese k, los grupos tienden a reflejar estratos de ingreso (bajo–medio–alto) condicionados por la edad, más que la frontera normativa de pobre/no pobre. Podemos decir que el número de clusters sugerido por el Elbow es útil para describir clases socioeconómicas dentro de la región, pero no separa de manera exacta a pobres y no pobres (el algoritmo es no supervisado y sólo ve edad+ingreso).

6. Cluster jerárquico: Utilizando las variables mencionadas arriba, realizamos un análisis de clustering jerárquico. Generamos un dendograma y explicamos brevemente qué es un dendograma.



El dendrograma es el "árbol" que produce el clustering jerárquico: cada hoja es una observación, y las uniones verticales muestran en qué orden se van fusionando los grupos. La altura de cada unión (eje "Distancia – Ward") refleja cuán diferentes eran los grupos que se juntan: cuanto más alta la barra, más disímiles. Un "corte" horizontal del árbol define cuántos clusters quedan. En nuestro dendrograma, las fusiones al principio son bajas (grupos muy parecidos) y aparecen dos saltos grandes alrededor de las alturas ~75 y ~105. Si cortamos a la altura intermedia (≈80–90), quedan tres ramas bien separadas; si cortamos más abajo (≈40) aparecen 4–5 subgrupos, y si cortamos muy arriba (≈110) todo se reduce a 2. Por tanto, la estructura sugiere 3 clusters como partición razonable, consistente con lo que vimos con el codo (k-means). Estos grupos muestran perfiles socioeconómicos distintos en las variables consideradas, sirven para describir clases dentro de la región, pero no equivalen a "pobres vs. no pobres".