Trabajo Práctico N°2

HISTOGRAMAS, KERNELS & MÉTODOS NO SUPERVISADOS USANDO LA EPH

link del repositorio: https://github.com/gfeijoo95/BigDataUBA-Grupo7.git

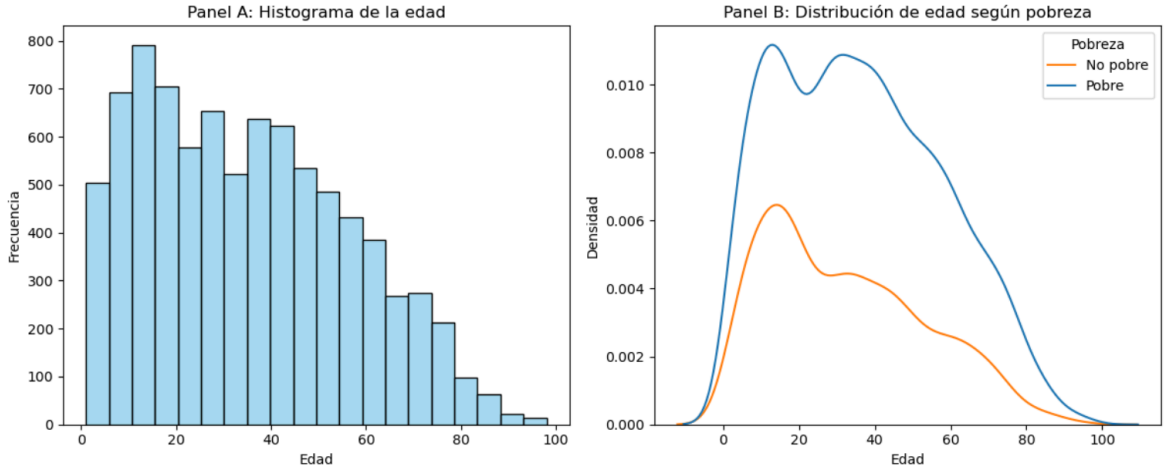
Grupo 7

Dolimpio, Gastón

Feijoo, Guillermo

Rudi, Federico

Parte I:
1)



En el Panel A, la distribución de la edad muestra una mayor concentración de personas en edades adultas jóvenes, con menos frecuencia en los extremos etarios.
En el Panel B, se observa que las personas pobres tienden a concentrarse en edades más jóvenes, mientras que las no pobres presentan una distribución algo más desplazada hacia edades adultas.
Esto sugiere una estructura poblacional más joven entre los hogares pobres, posiblemente asociada a diferencias en la participación laboral o educativa.

2)

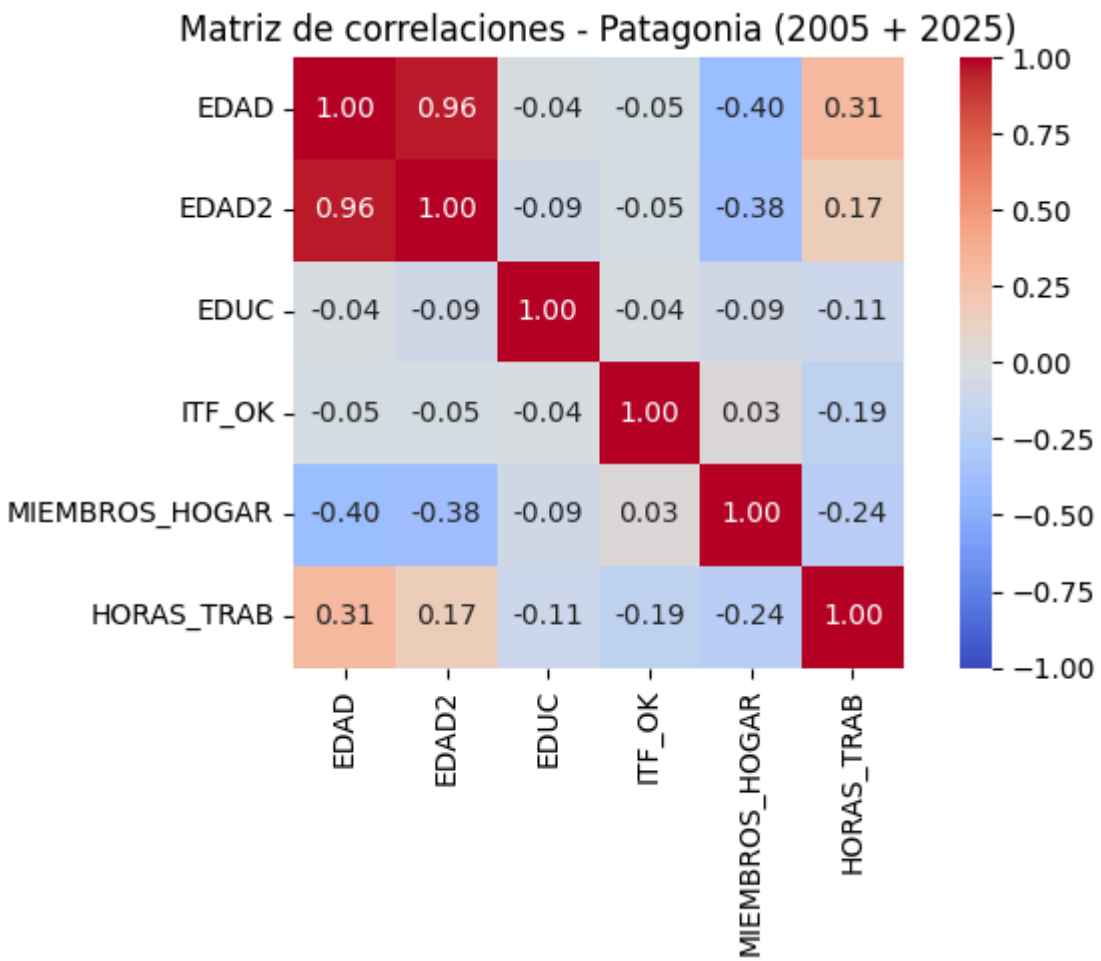
Año	Media	Desvío estándar	Mínimo	Mediana	Máximo	N
2005	8.37	4.50	0.0	8.0	20.0	2,982
2025	9.69	4.67	0.0	11.0	20.0	5,182
Total	9.26	4.66	0.0	9.0	20.0	8,164

En términos generales, se observa una mejora sostenida en el nivel educativo promedio de la población patagónica entre 2005 y 2025. El promedio ponderado de años de educación aumentó de 8.4 a 9.7, y la mediana pasó de 8 a 11 años, lo que implica un corrimiento hacia niveles más altos de escolaridad —principalmente mayor finalización de la secundaria. Sin embargo, la dispersión (aproximadamente 4.6 años) permanece amplia, lo que refleja persistentes desigualdades educativas dentro de la región.

3) A continuación se actualiza el valor del Ingreso Total Familiar del 2005 para expresarlo en pesos del 2025. Para ello se promedia el valor de la CBT para el primer trimestre de cada

año y se obtiene un deflactor de 1400,78. Se multiplican los valores del 2005 por ese índice a fin de que las cifras sean comparables.

Parte II



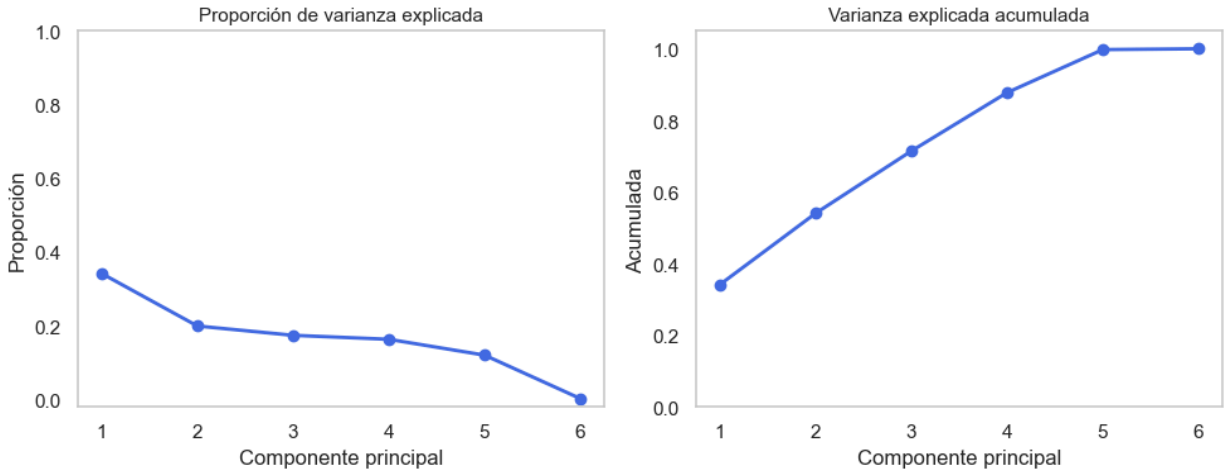
En la matriz de correlaciones para la región Patagónica (2005 y 2025 combinados), se observa una alta relación mecánica entre la edad y edad2 (0,96), mientras que el resto de las asociaciones son en general débiles.

La edad presenta correlación negativa con el tamaño del hogar (-0,40) y positiva con las horas trabajadas (0,31), lo que refleja el ciclo de vida laboral y familiar típico: los jóvenes viven en hogares grandes y trabajan menos horas, mientras que los adultos de mediana edad trabajan más.

La educación presenta correlaciones débiles o nulas con las horas trabajadas (-0,11) y con la condición de ingresos positivos (-0,04), lo cual es coherente dado que la educación influye más en el nivel de ingresos que en su mera existencia.

El tamaño del hogar se asocia negativamente con las horas trabajadas (-0,24), indicando que en hogares con más miembros hay más dependientes y menos personas activas laboralmente.

En conjunto, estas correlaciones ofrecen una primera aproximación coherente con la teoría económica del ciclo de vida y con patrones demográficos y laborales observados en la región.
PCA



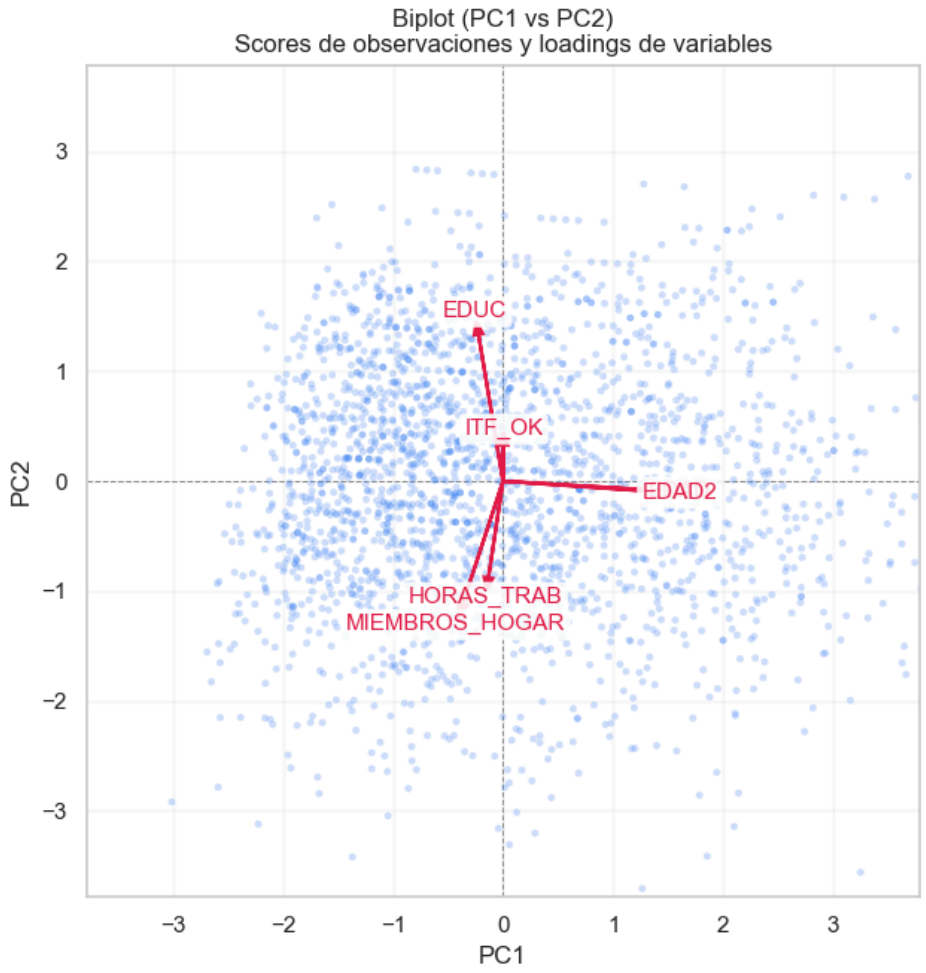
En esta etapa se aplica el Análisis de Componentes Principales (PCA) sobre las variables estandarizadas del conjunto de datos.

El objetivo es reducir la dimensionalidad del espacio original conservando la mayor cantidad posible de información contenida en las variables.

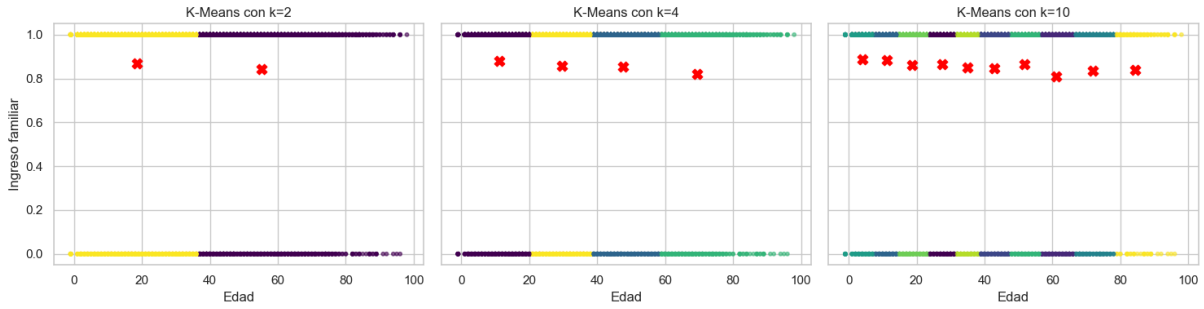
El gráfico de la izquierda muestra la proporción de varianza explicada por cada componente principal, mientras que el de la derecha presenta la varianza acumulada.

Se observa que:

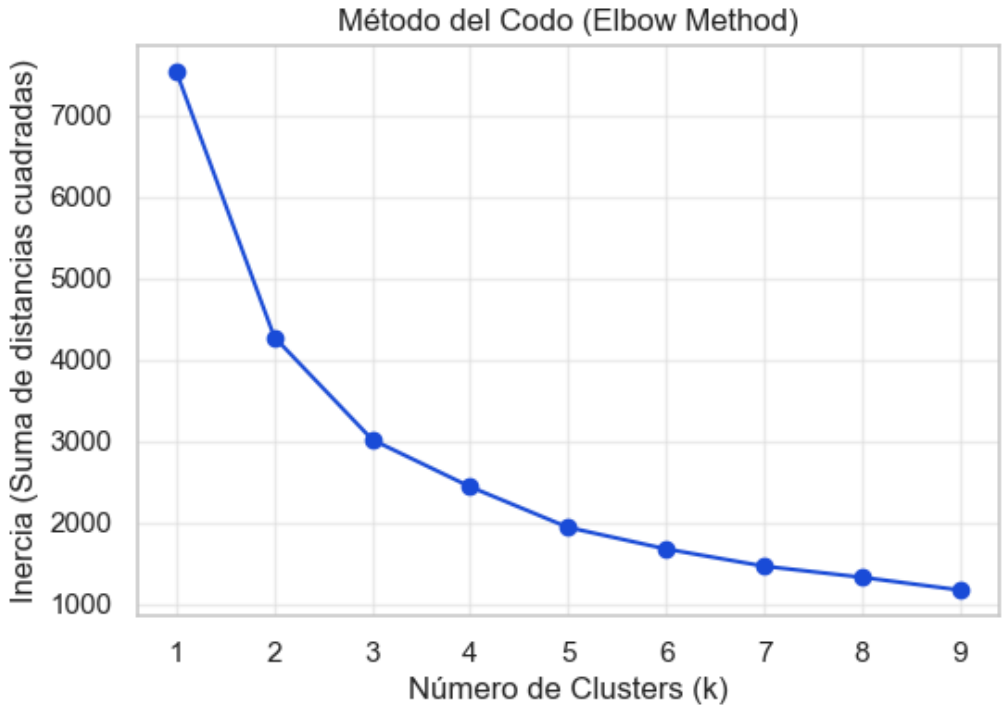
- Los primeros dos componentes explican aproximadamente el 54 % de la varianza total.
- Con cuatro componentes, se alcanza cerca del 88 % de la varianza.
- A partir del quinto componente, el aporte marginal de información es mínimo.



El biplot del PCA muestra la distribución de las observaciones en el plano definido por los dos primeros componentes principales y la contribución de cada variable a su conformación. Las flechas representan las cargas (loadings) de las variables originales sobre los componentes, indicando tanto su dirección de influencia como su intensidad. Se observa que EDAD y EDAD2 se orientan de manera opuesta a HORAS_TRAB y MIEMBROS_HOGAR, lo que sugiere un contraste entre individuos de mayor edad —con menor carga laboral y hogares más reducidos— y aquellos más jóvenes y activos laboralmente. Por su parte, EDUC e ITF_OK se ubican en una dirección similar, reflejando la correlación positiva entre nivel educativo e ingreso familiar per cápita. En conjunto, el gráfico permite visualizar cómo las variables se agrupan en torno a dos dimensiones: una demográfica-laboral, vinculada al ciclo de vida y la participación en el mercado de trabajo, y otra socioeconómica, asociada al capital humano y los ingresos.

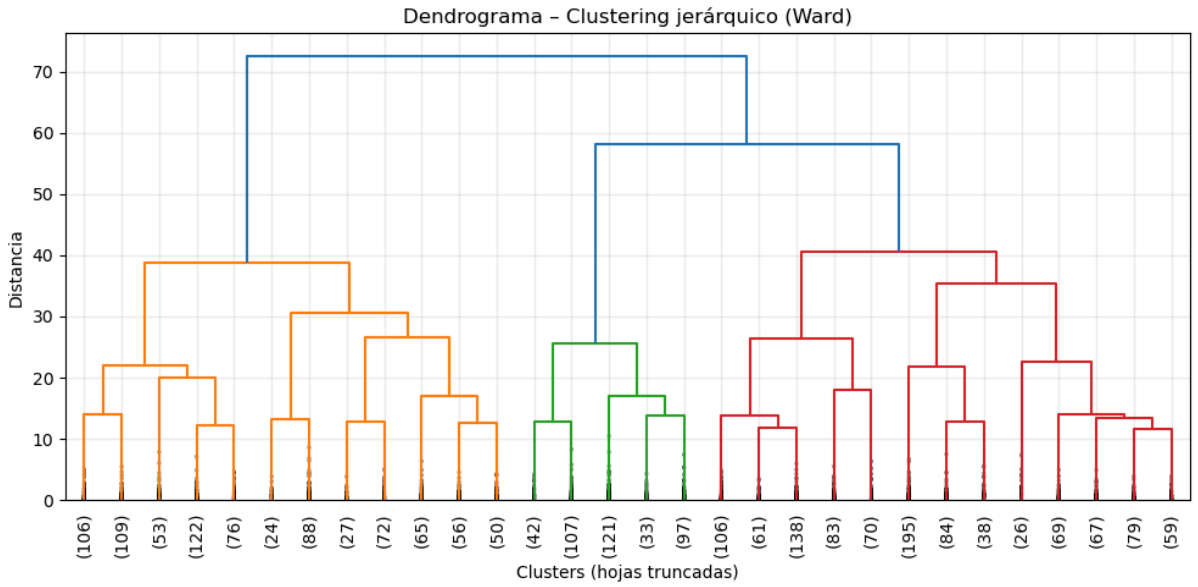


El modelo de K-Means con dos grupos (k = 2) logra una separación simple entre personas con ingresos familiares bajos y altos, identificando de manera clara a los hogares pobres y no pobres en la región. Dado que el ingreso familiar domina la estructura de los datos, la variable edad no aporta una diferenciación sustantiva dentro de cada grupo. En términos socioeconómicos, el algoritmo reproduce una segmentación binaria basada principalmente en el nivel de ingresos, sin capturar heterogeneidad interna ni matices vinculados al ciclo vital o a otras dimensiones del bienestar. Por ello, aunque k = 2 permite distinguir correctamente a los pobres de los no pobres, esta clasificación resulta limitada para analizar la diversidad de situaciones dentro de cada estrato.



El gráfico del método del codo muestra una disminución pronunciada de la inercia entre *k = 1* y *k = 3*, y una reducción más atenuada a partir de *k = 4*. Este patrón sugiere que el punto de inflexión —o "codo"— se encuentra aproximadamente entre tres y cuatro grupos, indicando que ese rango representa una estructura de clusters razonablemente homogénea dentro de cada grupo y suficientemente diferenciada entre ellos. En términos sustantivos, este resultado implica que el número óptimo de clusters en la región no sería *k = 2* (una separación estricta entre pobres y no pobres), sino más bien *k = 3 o k = 4*, que permitiría

captar diferencias adicionales entre distintos niveles socioeconómicos intermedios. Así, el algoritmo no sólo distinguiría la condición de pobreza, sino también gradaciones dentro de la población no pobre, asociadas a variaciones en el ingreso y posiblemente en la edad o composición del hogar.



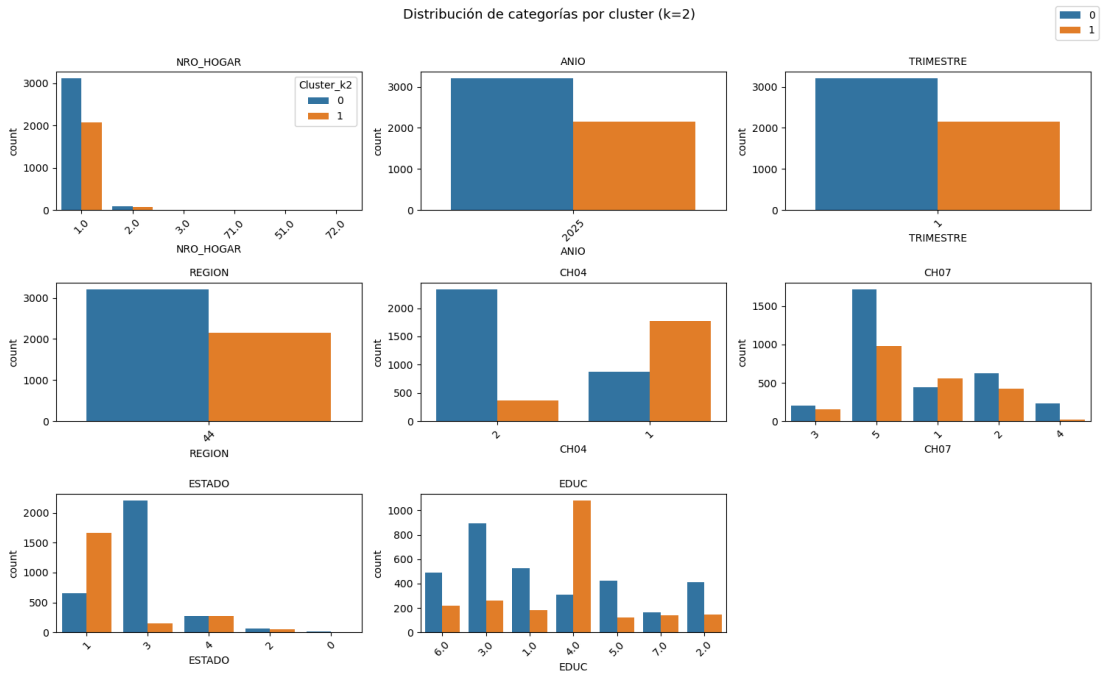
El dendrograma es una herramienta visual utilizada en el análisis de clustering jerárquico que representa la forma en que las observaciones se agrupan progresivamente según su nivel de similitud.

Gráficamente, adopta la forma de un árbol invertido, donde cada hoja en la base corresponde a una observación individual y las ramas muestran las uniones o “fusiones” entre grupos a medida que aumenta la distancia o disimilitud.

Su lectura se realiza de abajo hacia arriba: en los niveles inferiores se encuentran los individuos más similares, que se combinan para formar grupos más amplios conforme se asciende.

El eje vertical indica la distancia o disimilitud entre los grupos fusionados, de modo que los saltos grandes o las ramas largas señalan la existencia de conglomerados más diferenciados entre sí.

En síntesis, el dendrograma funciona como una representación jerárquica y visualmente intuitiva del proceso de agrupamiento, permitiendo identificar cuántos clusters resultan apropiados según los niveles de distancia donde se producen los principales cambios estructurales.

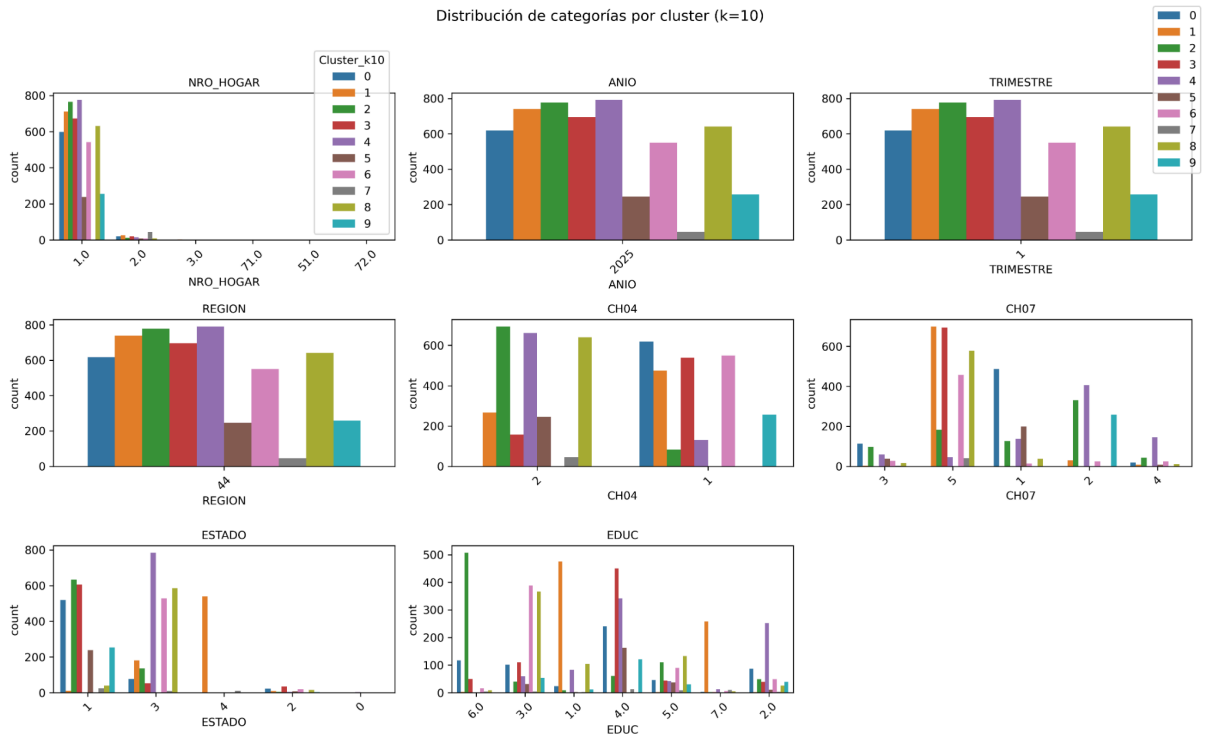


El modelo de K-moda con k = 2 separa la población en dos grupos de características socioeconómicas diferenciadas.

El Cluster 0 agrupa hogares con mayores niveles educativos y probablemente mayor inserción laboral, mientras que el Cluster 1 concentra personas con menor nivel educativo y categorías laborales más precarias.

Asimismo, las diferencias en la composición del hogar (CH04 y CH07) sugieren que el segundo grupo está formado por familias más numerosas o con mayor presencia de dependientes.

En conjunto, la segmentación obtenida reproduce la distinción entre hogares de mayores y menores recursos, análoga a la división entre no pobres y pobres observada en la variable de referencia.



El análisis de *k-modes* permitió explorar la estructura de la población a partir de variables categóricas.

Con k = 4, el algoritmo identifica cuatro grupos socioeconómicos bien definidos, que combinan diferencias en nivel educativo, tamaño del hogar y condición laboral, ofreciendo una visión más matizada que la simple división entre pobres y no pobres.

En cambio, con k = 10, los clusters se vuelven más específicos, reflejando microsegmentos con características particulares (por ejemplo, hogares pequeños con alta educación frente a familias extensas con baja escolaridad).

Si bien esta mayor granularidad aporta detalle, también introduce ruido interpretativo, por lo que el valor de k = 4 se considera un punto de equilibrio entre precisión estadística y claridad analítica.