



Trabajo Práctico 2

Licenciatura en Ciencias del Comportamiento

Alumnas

Mora Palatini

Isabella Martina Brunello

Sofia Chiara Retamal Diniello

Profesores

Maria Noelia Romero

Tomas Enrique Buscaglia

Asignatura

Ciencia de Datos

Tutorial 3 - Lunes 14:00

Semestre y año de presentación

2º semestre 2025

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

1. Tanto en el Panel A como en el Panel B, podemos observar que la distribución de las edades está sesgada hacia la derecha, indicando que la mayoría de la muestra está compuesta por personas jóvenes. En el Panel B podemos ver una función de densidad de las edades al cuadrado desglosada en Pobres y No Pobres. Podemos observar que hay mayor cantidad de personas con edades mayores en el grupo No Pobres que en el grupo Pobres, cuya población se concentra más en edades más jóvenes.

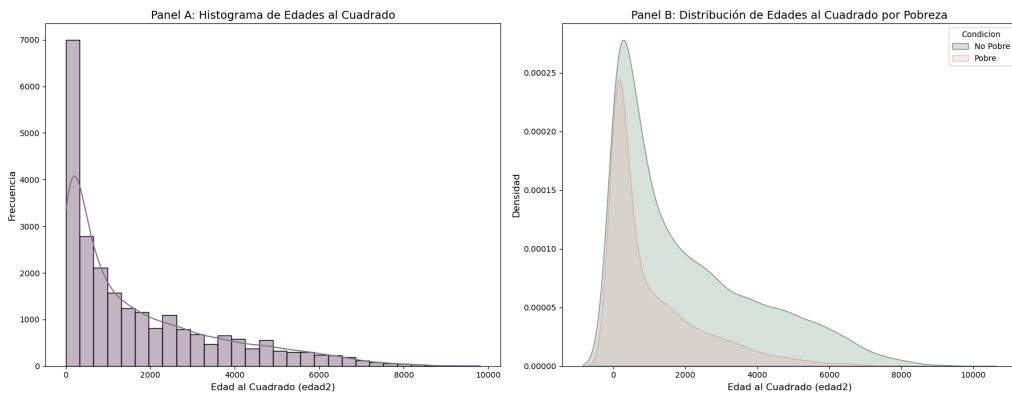


Figura 1. Distribución de edades elevadas al cuadrado

2. En la Tabla 1 podemos observar que el promedio es 9.21 años de educación recibida, lo que significa que la media de la muestra no ha completado el secundario con una variabilidad moderada de 4.84 en torno a la media. El mínimo es de 0.00 años de instrucción formal mientras que el máximo de 17.00 indica que completaron la universidad. El *p50* nos indica que al menos la mitad de la muestra no completó el nivel secundario de educación. La diferencia entre la media (9.21) y la mediana (10) nos indica un sesgo en la distribución de la educación hacia valores más bajos, indicando un menor índice de escolarización.

Promedio (mean)	9.208 años
Desviación Estándar (sd)	4.844 años
Mínimo (min)	0.00 años
Percentil 50 (p50)	10.00 años
Máximo (max)	17.00 años

Tabla 1. Estadística descriptiva de la variable *educ*

3. En la Figura 2 podemos observar la distribución de la variable ITF. Esta se encuentra fuertemente polarizada y separada en dos grupos que se distribuyen por debajo y encima de la línea de pobreza¹ (fijada en el TP1 en \$365.177). La polarización de las familias puede ser

¹ Nótese que en el eje x se observa una escala logarítmica, donde 10^3 indica \$1.000, 10^4 indica \$10.000 y así sucesivamente.

explicado porque una familia numerosa puede tener un ingreso total elevado, pero al repartirlo entre muchos miembros, el monto por persona es insuficiente y por eso se clasifican como pobres. Esto demuestra que la riqueza de un hogar depende tanto de cuánto dinero ingresa como de cuántas personas hay que alimentar con él, es decir entre cuántos miembros del hogar se divide.

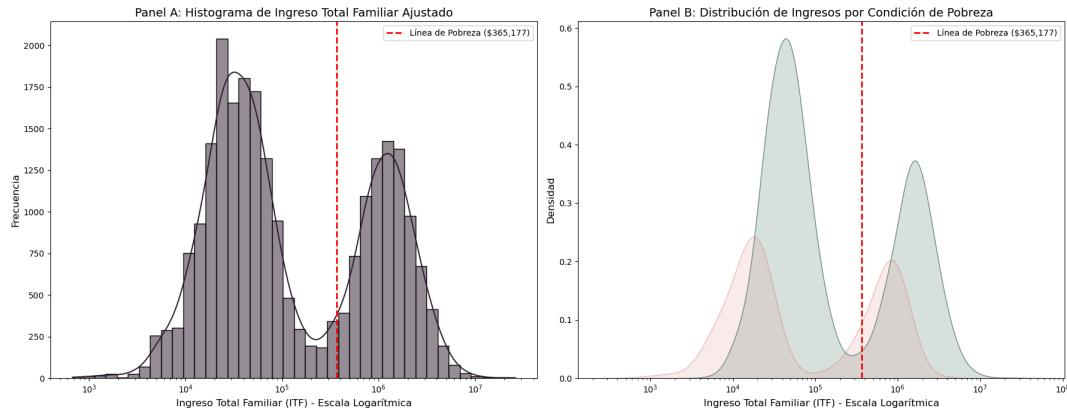


Figura 2. Distribución de la variable ITF

4. A la hora de crear la Tabla 2, notamos valores máximos extraños (como 1998 horas semanales, equivalente a 285 horas por día); es por esto que decidimos sacar los valores mayores a 200 horas semanales. Luego de esta limpieza, la variable *horastrab* muestra un promedio de 24.16 horas semanales y un desvío estándar de 25.51 horas, indicador de una alta variabilidad de las horas de trabajo a la semana. La mediana de 20 horas señala que al menos la mitad de los jefes de hogar trabajan media jornada. La distribución de la muestra en la tabla, demuestra que existen jefes de hogar que no trabajan, otros que cumplen jornada completa y otros con altas cargas de horarios de trabajo.

Promedio (mean)	24.16 hrs.
Desviación Estándar (sd)	25.51 hrs.
Mínimo (min)	0.00 hrs.
Percentil 50 (p50)	20.00 hrs.
Máximo (max)	140.00 hrs

Tabla 2. Estadística descriptiva de la variable *horastrab*

5.

	2005	2025	Total
Cantidad observaciones	14.651	11.698	26.349
Cantidad de observaciones con NAs en la variable “Pobre”	170	2.358	2.528

Cantidad de Pobres	4.264	3.256	7.520
Cantidad de No Pobres	10.217	6.084	16.301
Cantidad de variables limpias y homogeneizadas	19	19 ²	19

Tabla 4. Resumen de la base final para la región Pampeana

Parte II: Métodos No Supervisados

1. En la matriz de correlación Figura 3 se analizan los seis predictores de la región pampeana, es decir *edad*, *edad²*, *educ*, *ingreso_total_familiar (ITF)*, el número de *miembros en el hogar* (2005=IX_TOT y 2025=IX_Tot) y *horastrab*. Primero se observa una correlación muy alta entre *edad* y *edad²* de 0.96, bastante lógico porque son la misma variable con la diferencia de que la segunda está elevada al cuadrado (riesgo de multicolinealidad). Asimismo, *edad* muestra una relación negativa moderada de -0.41, con *miembros del hogar*, lo que refleja que hay cambios en la estructura familiar según el ciclo de vida, es decir, que a medida que una persona es mayor ella suele convivir con menos miembros en el hogar. Por último, el *ingreso familiar total* y *educación* no parecen relacionarse de manera fuerte con ninguna variable del análisis. En conclusión, la matriz sugiere que las variables de estructura demográfica (*edad* y tamaño del hogar) tienen más vínculo entre sí que con el ingreso.

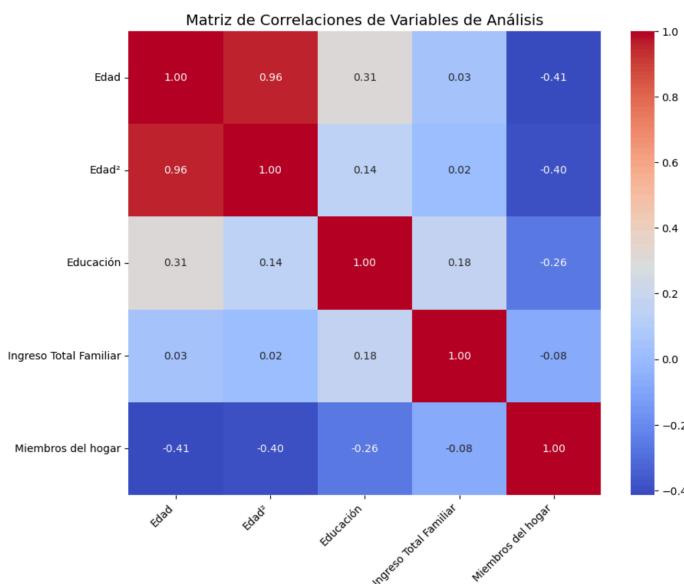


Figura 3. Matriz de Correlaciones de Variables de Análisis

A. PCA

2. Los resultados del gráfico muestran cómo se agrupan los hogares según patrones de edad, educación, ingreso, horas trabajadas y tamaño del hogar. La dispersión indica la variabilidad capturada por los dos primeros componentes. El componente 1 (eje X) capta

² Limpiamos las mismas variables y quedaron unificadas en una misma base de datos, por eso el total de variables en la base unificada es 19

diferencias socioeconómicas (ingreso, educación, horas trabajadas) y el componente 2 (eje Y) refleja en los aspectos del ciclo de vida (edad y tamaño del hogar).

En la Figura 4 (2005) se ve una mayor dispersión vertical que indica como el ciclo de vida (edad y composición del hogar) juega un rol más fuerte en la diferenciación de los hogares. Los hogares pobres (color púrpura) se concentran en la zona baja y media del gráfico y los no pobres (color verde) están más hacia valores altos de componente socioeconómico. Refleja que en 2005 la pobreza estaba más asociada a menores ingresos y años de educación.

En la Figura 5 (2025) su dispersión se concentra más hacia el eje horizontal (componente 1), indicando que el índice socioeconómico es el principal factor de diferenciación entre los hogares. Se ve una separación más clara entre pobres (color rosado) y no pobres (color verde), sobre todo en la parte baja del eje socioeconómico. El componente del ciclo de vida (eje Y) pesa menos en la diferenciación respecto a 2005. Esto sugiere que en 2025 la pobreza está más determinada por las condiciones económicas que por la estructura demográfica.

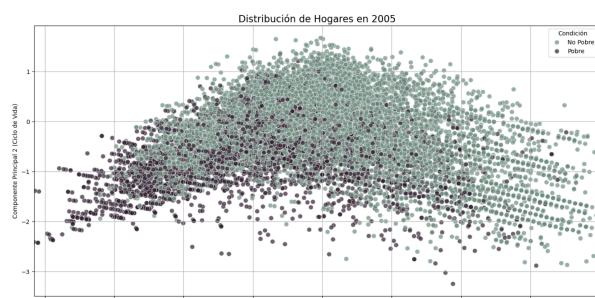


Figura 4. Distribución de Hogares en 2005

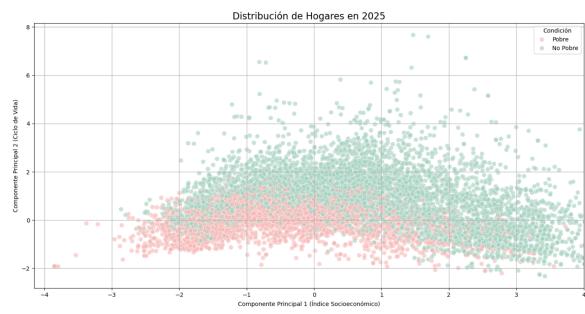


Figura 5. Distribución de Hogares en 2025

3. En la Figura 7 la dirección y longitud de las flechas indican la influencia de cada variable en los componentes y aquellas que tienen flechas largas son las que tienen mayor peso, además la proximidad de flechas indica una correlación positiva entre variables.

El biplot del PCA consiste de un primer componente (*miembros del hogar, ITF, horas trabajadas*) y un segundo componente (*edad* y *edad²*). Es por ello que las dimensiones que mejor diferencian a los hogares en el análisis son las condiciones económicas y laborales y la etapa del ciclo de vida en la que se encuentran.

En cuanto al *ingreso total familiar* y *horas trabajadas* ambas flechas son largas y apuntan hacia la derecha sobre el componente 1, indica que estas variables tienen gran peso en la dimensión socioeconómica. Cuanto más a la derecha está un hogar en el gráfico, mayor es su ingreso y sus horas trabajadas. Además, la cercanía entre estas dos flechas muestra que están positivamente correlacionadas. En cuanto a *miembros del hogar* apunta en dirección opuesta a las variables de ingreso y trabajo, sobre todo en el componente 1. Los hogares más grandes están asociados a menores ingresos y menos horas trabajadas, contribuyendo negativamente a la dimensión socioeconómica.

Las variables *edad* y *edad²* tienen vectores orientados hacia el componente 2, indica que influyen en la dimensión del ciclo de vida y no en la dimensión socioeconómica. La posición cercana entre ambas confirma la alta correlación esperada, ya que *edad²* es una transformación de la variable *edad*.

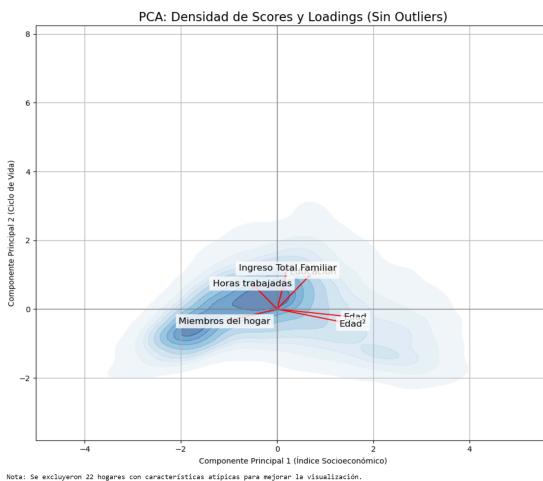


Figura 6. PCA: Densidad de Scores y Loadings (Sin Outliers)

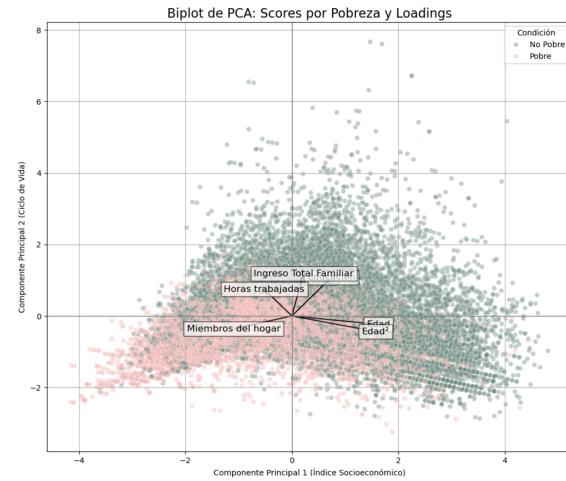


Figura 7. Biplot de PCA: Scores por Pobreza y Loadings

4. El primer componente de la Figura 8, PC1, captura la mayor parte de la variabilidad, en este caso de un 40% y representa la dimensión más importante en la diferenciación de los hogares, la cual está asociada al eje socioeconómico: ingresos y horas trabajadas. El siguiente componente, PC2, captura otra porción importante de la variabilidad de un 20% y refleja la dimensión vinculada principalmente al ciclo de vida (edad y miembros del hogar). El gráfico muestra que los dos primeros componentes (PC1 y PC2) concentran alrededor del 60% de la variabilidad total de los datos, lo cual justifica su uso para representar ejes interpretativos principales (socioeconómico y ciclo de vida).

A partir del tercer componente la proporción de varianza explicada desciende progresivamente hasta el sexto que no aporta información relevante. Esto confirma que el PCA logra una reducción dimensional eficiente, con solo dos componentes se puede resumir buena parte de la heterogeneidad de los hogares. Aquello es notado en que los componentes: PC3 representa solo un 15%, PC4 un 12.5% y PC5 un 11%. Por último, el componente restante PC6, aporta muy poca varianza, casi 0%.

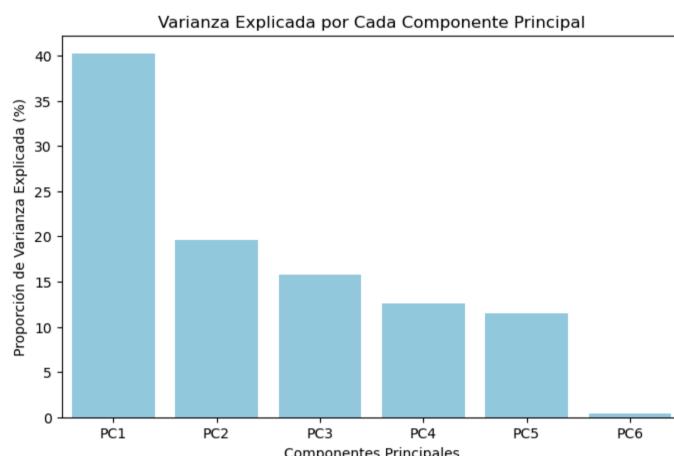


Figura 8. Varianza Explicada por Cada Componente Principal

B. Cluster

5.a. La Figura 9 divide los datos en dos clusters principales: el primero (azul) concentra los ingresos más bajos, es decir, aquellos más cercanos a cero, mientras que el segundo (rojo) agrupa al resto de los ingresos. El inconveniente es que este último cluster resulta demasiado amplio, ya que incluye tanto a personas con ingresos bajos como a aquellas con ingresos muy altos, sin lograr diferenciarlos claramente. En consecuencia, si bien existe una separación inicial entre ‘pobres’ y ‘no pobres’, esta distinción es poco precisa y ambigua.

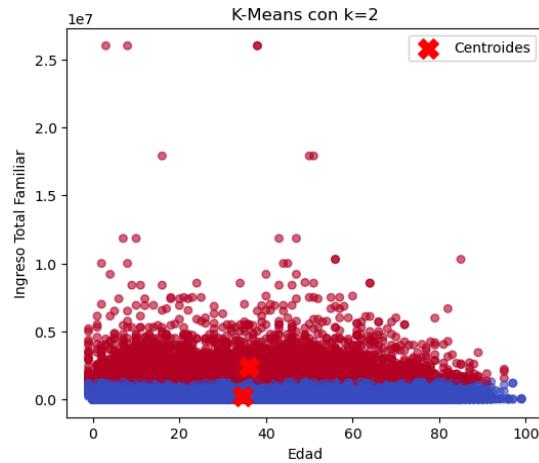


Figura 9. Ingreso Total Familiar con K-Means con $k = 2$

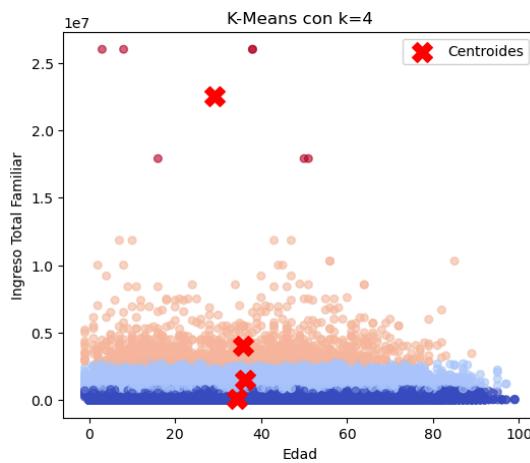


Figura 10. Ingreso Total Familiar con K-Means con $k = 4$

En la Figura 10 se observa una segmentación más detallada respecto de la distribución de ingresos. El algoritmo genera cuatro clusters diferenciados: el primero concentrado en los ingresos más bajos, que podría asociarse a la población en situación de pobreza; el segundo correspondiente a ingresos bajos-medios; un tercer grupo que concentra ingresos medios; y finalmente un cuarto cluster que agrupa a los casos atípicos con ingresos extremadamente altos. Esta mayor cantidad de clusters permite una separación más clara entre los diferentes niveles socioeconómicos y reduce la ambigüedad observada en la Figura 9. No obstante, la clasificación continúa presentando cierto grado de superposición, lo que limita su capacidad de reflejar con precisión la frontera entre pobres y no pobres establecida de manera oficial.

En la Figura 11 se aprecia una segmentación aún más refinada, donde los datos se distribuyen en múltiples clusters que capturan con mayor detalle las distintas escalas de ingreso. El algoritmo genera agrupamientos diferenciados que permiten identificar desde los hogares de ingresos más bajos hasta aquellos con ingresos medios y altos, incluyendo además clusters específicos para los casos atípicos con ingresos extremadamente elevados. Esta mayor granularidad mejora la representación de la heterogeneidad de la población, al reflejar gradientes de ingreso que no eran visibles en particiones con menor número de clusters.

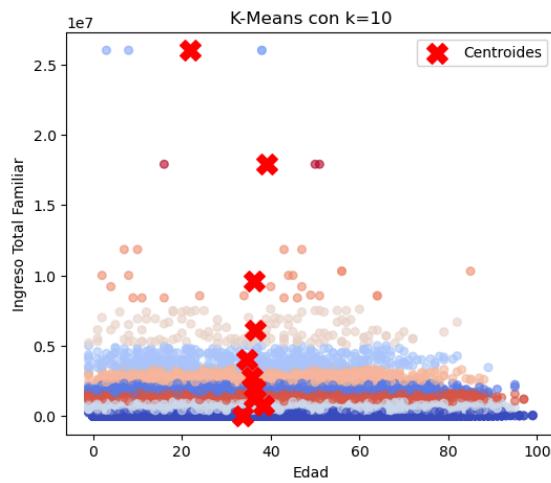


Figura 11. Ingreso Total Familiar con K-Means k = 10

5.b. Al analizar la curva de inercia mediante el método del *Elbow*, en la Figura 12 se observa que la reducción de la varianza intra-cluster es significativa hasta aproximadamente $k = 3$ o $k = 4$, a partir de donde la pendiente de la curva comienza a aplanarse. Esto indica que el número óptimo de clusters para el conjunto de datos se ubica en torno a cuatro, ya que un mayor incremento de k aporta mejoras marginales poco relevantes en términos de compactación de los grupos.

La elección de $k = 4$ resulta adecuada para capturar distintos niveles socioeconómicos, distinguiendo entre hogares de ingresos bajos, medios y altos, además de identificar casos atípicos con ingresos extremos. Sin embargo, esta partición no necesariamente coincide de forma estricta con la clasificación binaria de pobres y no pobres, ya que la línea oficial de pobreza no es considerada explícitamente por el algoritmo. Por lo tanto, $k = 4$ permite una mejor diferenciación de clases socioeconómicas en sentido amplio, pero requiere de un contraste adicional con la línea de pobreza para evaluar su capacidad de discriminar entre pobres y no pobres.

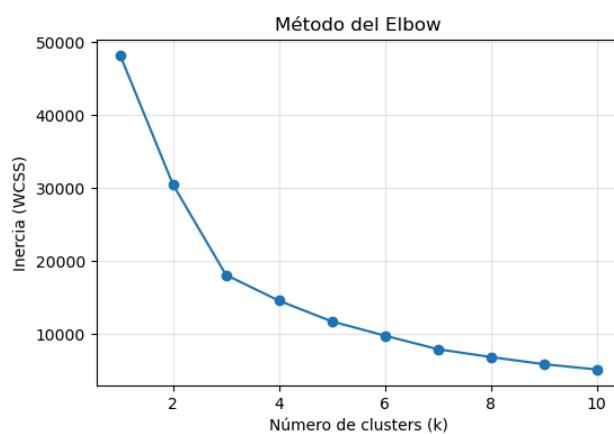


Figura 12. Método del Elbow

6. Un dendrograma es una representación gráfica en forma de árbol que muestra cómo las observaciones se van agrupando jerárquicamente según su nivel de similitud, de manera

que las fusiones en niveles bajos reflejan alta homogeneidad y las uniones a mayor altura evidencian diferencias más marcadas. El dendrograma obtenido mediante el método de Ward (ver Figura 12) permite visualizar cómo las observaciones de la base se dividen en dos grandes grupos a una distancia relativamente alta, lo cual sugiere una partición importante en la población. Asimismo, dentro de cada rama emergen subgrupos más pequeños, lo que refleja heterogeneidad interna asociada a factores socioeconómicos como la edad, el nivel educativo, los ingresos familiares y las horas trabajadas.

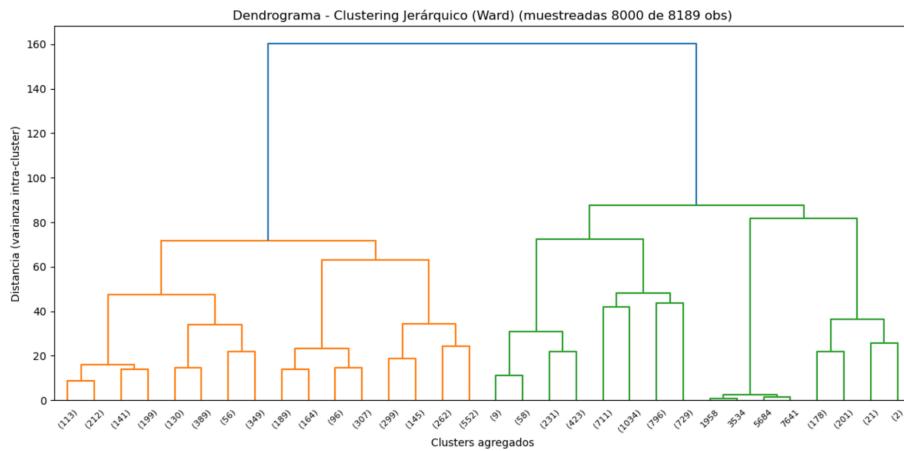


Figura 13. Dendograma

7. La Figura 13 es un diagrama de dispersión $k = 2$ que muestra la asignación de clústeres sobre el plano *edad* e *ingreso total familiar* (ITF). En esta figura se observa una superposición casi completa entre los colores en todo el rango etario y de ingresos: tanto los ingresos muy bajos (concentrados cerca de cero) como los medios/altos aparecen repartidos en ambos grupos. No se aprecia un umbral de ITF ni franjas etarias que separan nítidamente los clústeres; más bien, uno de ellos domina en cantidad de puntos y el otro aparece intercalado sin patrón claro, mientras que los outliers de ingreso alto también quedan mezclados. Esto sugiere que, con $k = 2$, el modelo no captura una división económica equivalente a pobre/no pobre, y que la partición responde a otras dummies sociodemográficas que no se reflejan como fronteras visibles en esta proyección.

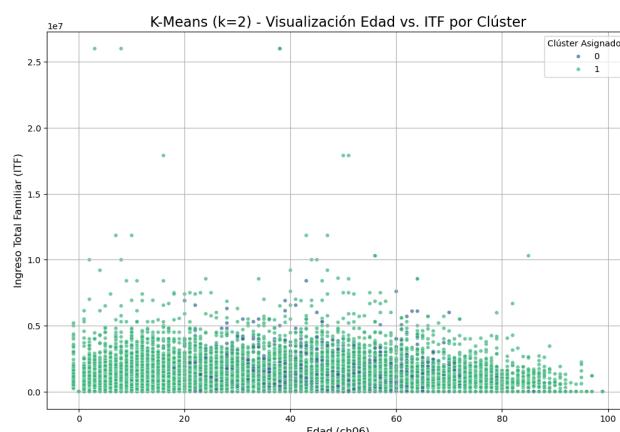


Figura 13. Ingreso Total Familiar con K-Means $k = 2$

Anexo

[Link al repositorio de GitHub](#)

HTTPS: <https://github.com/sofiaretamal/CC408-Grupo-T3-1.git>