Trabajo práctico Nº2

Ciencia de Datos

Grupo Nº4

Alumnos

Juan Augusto Alvarez Simonassi , Camila Gioja , María Delfina González Elosú

Profesores

María Noelia Romero

Ignacio Anchorena

Tomas Enrique Buscaglia

GitHub: Link

Parte 0: Limpieza de datos I

En una primera instancia, se comenzó el trabajo filtrando la base de la EPH, de los archivos de individual y hogar de ambos años, 2005 y 2025. Se utilizó una función llamada filtrar_region_42 para extraer únicamente los registros correspondientes a la región identificada: Cuyo, con el código 42. Luego se convierten las columnas a formato numérico y las guarda en la variable ***indiv_2005*** y ***indiv_2025***. Luego se unificaron los dos conjuntos de datos, tanto ***data_2005*** y ***data_2025***, filtrando las variables únicamente por las que tienen en común ambas bases.

Para continuar, se buscó explorar los tipos de variables. Se construyó una lista de *ingresos* de las columnas de ***data_2005*** con todas las variables monetarias y los módulos de ocupación con ingresos como: monto de ingreso total familiar ("ITF"); monto de ingreso per cápita familiar (*"IPCF"*); monto de ingreso total individual (*"P47T"*); ingreso de ocupación principal (*"P21"*), etc.

Se revisaron las variables numéricas de las bases de 2005 y 2025 para detectar valores negativos o inconsistentes. Los códigos de no respuesta (9, 99, 9999) fueron reemplazados por NaN, lo que permitió contar los faltantes por variable y comparar ambos años en una tabla resumen. También se verificaron casos sospechosos (por ejemplo, cuando en un año los faltantes aparecían como ceros y en otro como NaN). En particular, para la variable *"ITF"*, se consideró como "no respondió" cuando era NaN o igual a cero. Finalmente, se hizo un chequeo de tamaños muestrales para confirmar la consistencia entre las bases de 2005 y 2025.

Continuamos utilizando la base de datos ***tabla_adulto_equiv***. Convertimos la información individual de edad y sexo en un valor de adulto equivalente utilizando una tabla de factores. Esto ajusta las diferencias de consumo según edad y sexo, para ponderar valores y que sean comparables los dos años. Siguiente, agrupamos por "CODOSU" y "NRO_HOGAR" y seleccionamos la columna "*adulto_equiv"*. Sumamos los valores de adulto_equiv dentro de cada grupo de hogar y asigna ese valor a cada fila del grupo para que todos los miembros de mismo hogar tengan el mismo valor de adulto equivalente del hogar. Se hizo lo mismo para 2025.

A continuación se identificó quien es pobre y quien no clasificando el nivel de pobreza utilizando la variable *"ITF"* (Ingreso Total Familiar) y el tamaño de hogar en "*adultos_equiv"*. Se utilizaron solamente los individuos que respondieron. Sin embargo, luego se asignó NaN para todas las observaciones, y luego se asignaron los valores de los que sí respondieron. Luego se multiplicó 205.07 y 355177 de 2005 y 2025 respectivamente por la columna de adulto equivalente para obtener el "*ingreso_necesario"*. Luego se armó una tabla para ambos años para comparar por ejemplo: cantidad de pobres, cantidad de no pobres, porcentaje de pobres, porcentaje de no pobres y porcentaje de NaNs. Finalmente se concatenó la base de 2005 con la de 2025 que se llama ***data_final***.

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

Para comenzar con el ejercicio 1 creamos la edad al cuadrado y eliminamos los valores faltantes y negativos de *"EDAD"*. Se realizó un Kernel Density Estimate que demuestra la densidad de edad para cada grupo; pobre, no pobre y faltante. El histograma de edad en el Panel A **(Figura 1)** muestra que la distribución de edades es amplia, con mayor concentración en edades jóvenes y jóvenes adultos.
En el Panel B (**Figura 1**), la curva de pobres y no pobres parece tener una forma dentro de todo similar, aunque desplazadas. En los pobres la concentración de las edades es en edades tempranas, mientras que los no pobres presentan una distribución más pareja, aunque el máximo sigue siendo en adultos jóvenes. Se puede decir que la mayoría de la población para pobres y no pobres se encuentra entre los 5 y los 40 años.

A continuación, creamos una función llamada calcular_educ que tomaba una fila y devuelve los años de educación. Para hacer las reglas de construcción del índice, se utilizó un *if* para asignar el número 6 si terminó primaria, un 12 si terminó secundaria y en caso contrario se le suman los años que haya aprobado de secundaria. Con el nivel terciario, universitario y posgrado se hizo lo mismo. La función resultante llamada *"EDUC"* nos permitió convertir variables categóricas en una variable continua que transformaba los años de educación de los individuos. En cuanto a las estadísticas descriptivas de *"EDUC"*, la media es de 10.84 años. En 20 años la media aumentó 31 años. Esto indica que hubo una mejora en la escolaridad promedio a lo largo del tiempo. En cuanto a la mediana, la mitad de la población tiene al menos una secundaria completa (12 años). La desviación estándar creció de 3.88 a 4.04 de 2005 a 2025. Esto quiere decir que aumentó la heterogeneidad en los años de educación, es decir que hay más diferencia entre los que estudiaron pocos años y los que estudiaron más años.

Ejercicio 3

Antes de realizar los gráficos, se guardaron los parámetros de los precios en dos variables distintas y se calculó un deflactor. Luego nos aseguramos con una función que la variable ingreso (*"ITF"*) sea numérica y definimos el equivalente adulto del hogar. Luego se calcularon los ingresos familiares en pesos del 2005 para que equivalgan los pesos de 2025. Calculamos la línea de pobreza de hogar en 2025 ajustando por tamaño (*LINEA_2025*) y composición (*EQ_HOGAR*).
Armamos una base unificada y se eliminaron los ceros, negativos y NaNs. Calculamos la línea de pobreza mediana para después graficar la línea de referencia para los dos gráficos. En el histograma de Panel A (**Figura 2**) la distribución de ingresos familiares, que fue expresada en términos de pesos de 2025, presenta mayor concentración cerca de la línea de pobreza. La distribución es asimétrica y desplazada significativamente hacia la izquierda, es decir hay menos hogares a medida que el ingreso aumenta. Se puede observar una cola larga hacia la derecha, lo que indica que hay una proporción importante de hogares con ingresos más altos. En el panel B (**Figura 2**) al separar la densidad por condición de pobreza, los hogares pobres se encuentran con sus ingresos a la izquierda mientras que los no pobres a la derecha, y se superponen. Esto denota

la desigualdad en la distribución de ingreso y que es relevante segmentar la población según el umbral de pobreza.

A continuación, para el ejercicio 4, sumamos ambas variables *"PP3E_TOT"* y creamos una variable llamada "*horas"* para ambos años con las "*horastrab"*. En las dos bases, la variable *horastrab* muestra que la media es de 15 horas semanales trabajadas, con desviación estándar de 23 y 22 horas para 2005 y 2025 respectivamente. Como la mediana para ambos años dio cero, se puede afirmar que por lo menos la mitad de horas trabajadas durante una semana fue de cero horas, incluyendo a ocupados ausentes, población inactiva y desempleada. El contraste entre la media (15hs) y la mediana (0hs) muestra la heterogeneidad de la muestra: muy poco porcentaje de la población trabaja muchas horas pero la mayoría de la población no trabaja ninguna.

Posteriormente, para el ejercicio 5, creamos un resumen de la base final para la región del Cuyo (**Tabla 1**) incluyendo cantidad de observaciones en total (9459), pobres (3156) y cantidad de variables limpias y homogeneizadas (186). Podemos ver como la base de datos de 2005 tiene aproximadamente 300 observaciones más que la de 2025. Al mismo tiempo, hay menos cantidad de observaciones de NAs con la variable pobre en este año. Sin embargo, hay solamente 200 observaciones de diferencia entre la cantidad de pobres entre los dos años. Las variables para ambos años fueron las mismas ya que se unieron las bases para poder compararlas posteriormente.

Parte II: Métodos No Supervisados

Para comenzar con el análisis con métodos no supervisados, primero creamos una matriz de correlaciones de las variables *"EDAD", "EDAD2", "EDUC", "ingreso_total_familiar", "miembros_hogar"* y *"horastrab"* (**Figura 3**). Los resultados muestran una correlación positiva fuerte entre *edad* y *edad2*, como es esperable dado que una es función de la otra. Se ve una correlación positiva moderada entre *"EDUC"* e "*ingreso_total_familiar"*, lo que refleja que mayores niveles educativos tienden a asociarse con ingresos familiares más altos. Correlaciones bajas entre "*horastrab"* y las demás variables, indicando que la cantidad de horas trabajadas no se relaciona fuertemente con edad, educación o ingreso.

Parte A: PCA

Para comenzar con nuestro análisis de componentes principales, primero estandarizamos las variables y luego aplicamos PCA utilizando la librería sklearn.decomposition. A su vez, agregamos los *loadings* y los *scores*; estos últimos corresponden a las nuevas coordenadas de cada observación en el espacio definido por los PCA. Realizamos gráficos de dispersión con los dos primeros componentes por año y vimos que los datos se superponen mucho, aunque se puede ver que en 2025 hay una leve mayor dispersión hacia valores positivos de PC1 (**Figura 4**).

Posteriormente, realizamos gráficos de dispersión de los primeros dos componentes con las condiciones de pobreza (**Figura 5**) donde la separación es un poco más clara. Los valores negativos de PC1 se distribuyen con mayor amplitud hacia valores negativos. Los pobres se distribuyen con mayor amplitud hacia la derecha del gráfico (PC1 positivos). Con estos gráficos se puede afirmar que PC1 está fuertemente asociado a diferencias en ingresos y condiciones de vida. Se visualiza la desigualdad económica entre grupos más que diferencias por año. En la tabla de loadings (**Tabla 2**), los dos primeros componentes concentran el 60% de la información. PC1 representa ejes demográficos de edad y tamaño de hogar y el PC2 un eje socioeconómico de educación trabajo e ingreso.

Siguiendo con este punto, en el gráfico de loadings (**Figura 6**) podemos ver los componentes principales que toman dos ejes claros de variación. El primer componente está asociado a *"EDAD"* y *"EDAD2"* (edad al cuadrado) lo que refleja un patrón demográfico. El componente 2 está asociado con educación, ingreso familiar y horas trabajadas, de allí proviene el componente socioeconómico. El *"IX_TOT"* (tamaño de hogar) presenta un loading negativo y más pequeño que el resto, esto explica el peso marginal para la explicación de la varianza.

Para el ejercicio 4 hicimos un gráfico de barras (**Figura 7**), donde se puede visualizar el porcentaje de la varianza de cada componente. Se puede decir que el componente 1 explica aproximadamente el 37,5% de la varianza de los datos. En segundo lugar tenemos al componente 2 con el 22,5%, y posteriormente tenemos al componente 3 y 4 que explican alrededor del 15%. Así desciende cada vez más hasta llegar al componente 6 que ya explica menos de 2% de la varianza de los datos.

Parte B: Cluster

Para la primera parte del ejercicio 5 utilizamos la biblioteca `sklearn-preprocessing` y `sklearn.cluster`. Estandarizamos la base sólo con *"EDAD"*, *"ITF_2025"* y *"POBRE"*. A su vez, quitamos los outliers, ya que para visualizar los gráficos se podía notar que había demasiados datos dispersos y no se lograba comprender la información. Posteriormente, creamos un modelo de k-medidas con k = 2 **(Figura 8)**, k = 4 (**Figura 9**) y k = 10 (**Figura 10**) clusters para EDAD al cuadrado e Ingreso Total Familiar. En la **Figura 8**; k=2 se puede ver la división clara de edad que se separa entre menos y mayores de 40 años. En la **Figura 9**; k=4, se pueden ver más grupos, jóvenes de bajos ingresos (naranja), adultos con ingresos bajos/medios (azul), adultos con ingresos medios (rojo) y personas de todas las edades con ingresos altos (verde). Aquí se ve claramente como el algoritmo separa en niveles socioeconómicos. La **Figura 10**; k=10, segmenta en forma más específica los rangos etarios y sus ingresos. Hay una diferencia entre niños, jóvenes adultos, y adultos y adultos mayores para ingresos bajos. Luego se va disipando esta diferencia de rangos etarios a medida que aumenta el ingreso familiar hasta quedar en ingresos altos un grupo que contiene todos los rangos etarios.

Por último, se muestra la clasificación real de pobres y no pobres (**Figura 11**). Esto nos permite comparar por ejemplo el caso de k=2 y concluir que el algoritmo capta parcialmente la brecha económica, segmenta más por edad que por ingreso. En cambio se observó que con más clusters se puede visualizar mejor que la variable ingreso sí tiene mayor peso de segmentación. Este método sirve para determinar cuál es el número óptimo de clusters para usar en el análisis de K-means. La inercia mide que tan bien se agrupa un conjunto de datos mediante K-Means. Mientras más chica es la inercia, mejor es la comparación de los clusters. Según este gráfico (**Figura 12**), habría que quedarse con 3 clusters, ya que el método del codo dice que cuando se produce el punto de inflexión, ya agregar clusters deja de aportar información valiosa. Estos 3 grupos nos podrían ayudar a distinguir entre pobres, clase media y clase alta.

En la **Figura 13,** podemos ver el cluster con con k = 3. Se parece un poco al de k=4 pero no distingue entre por ejemplo: "clase media" y "clase alta". Sin embargo se divide mayormente en 2 grandes grupos económicos; baja y alta. Se visualiza mejor los 3 clusters en relación a la edad, niños, jóvenes y jóvenes adultos con ingresos hasta la mitad del total de ingresos de la muestra (4), adultos y adultos mayores con ingresos también hasta 4 y por último ingresos altos donde hay mayor variabilidad de edades.

Posteriormente, para el ejercicio 6 utilizamos las 6 variables indicadas: *"EDAD"*, *"EDAD2"*, *"EDUC"*, *"ITF_2025"*, *"IX_TOT"*, *"horastrab"*. Como cada variable tiene una escala distinta primero se estandarizaron con *StandardScaler* para ponderar todas. Luego se realizó un cluster jerárquico (**Figura 14**) con el método de Ward que agrupa las observaciones y minimiza la varianza. Un dendrograma es un diagrama de árbol que permite visualizar la relación jerárquica de cómo se van uniendo los individuos. El eje X representa a los individuos y el eje Y representa la distancia que hay entre individuos. Cuanto menor es la altura, menor es la diferencia entre individuos que se unen.

Anexo

Figura 1

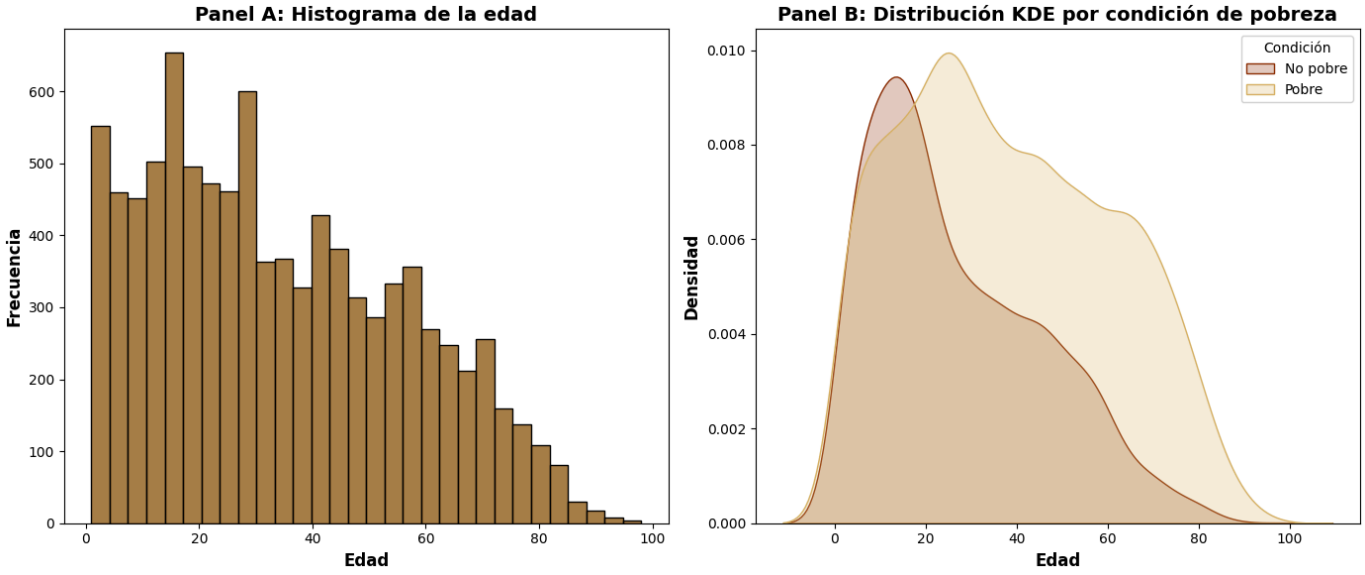
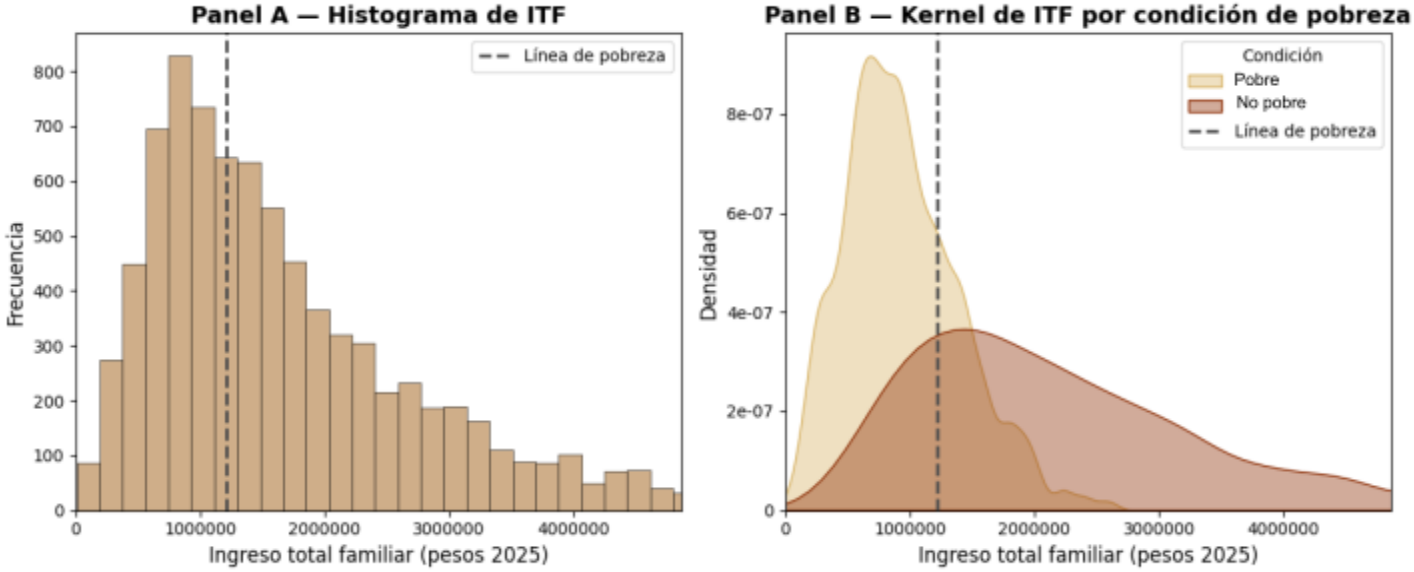


Figura 2



Nota: En ambos paneles, el ingreso total familiar está estandarizado a los valores de 2025

Tabla 1

Resumen de la base final para la región del Cuyo

	2005	2025	Total
Cantidad de observaciones	4865	4594	9459
Cantidad de observaciones con NANs en la variable “Pobre”	35	1063	1098
Cantidad de Pobres	1675	1481	3156
Cantidad de No Pobres	3155	2050	5205
Cantidad de variables limpias y homogeneizadas	186	186	186

Nota: La variable “Pobre” no tiene NaNs ya que fue previamente limpiada, en esta tabla se cuenta como NaNs a las personas que no respondieron ITF.

Figura 3

Matriz de correlaciones para la región de Cuyo (2005 + 2025)

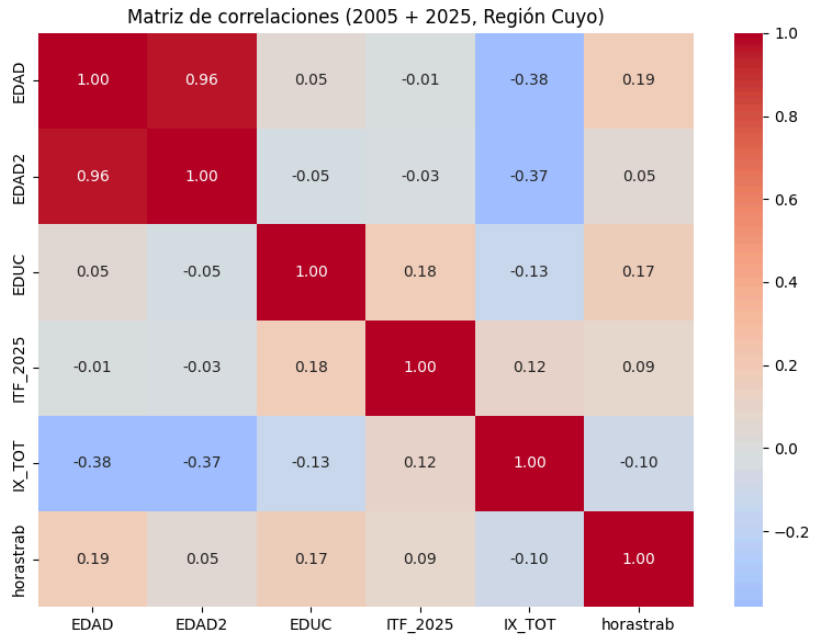


Figura 4

Gráfico de dispersión por año

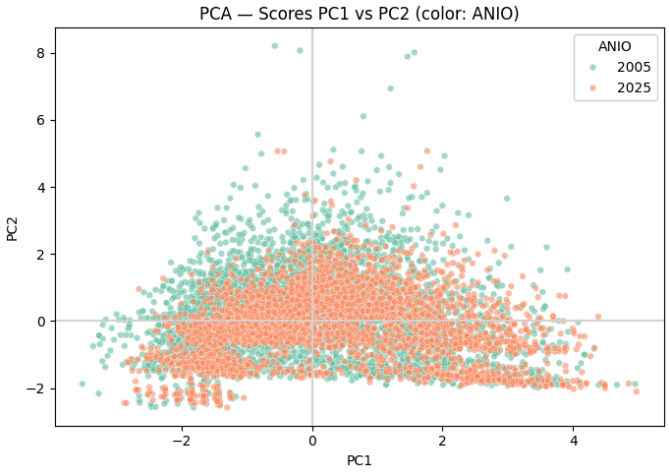


Figura 5

Gráfico de dispersión según pobreza

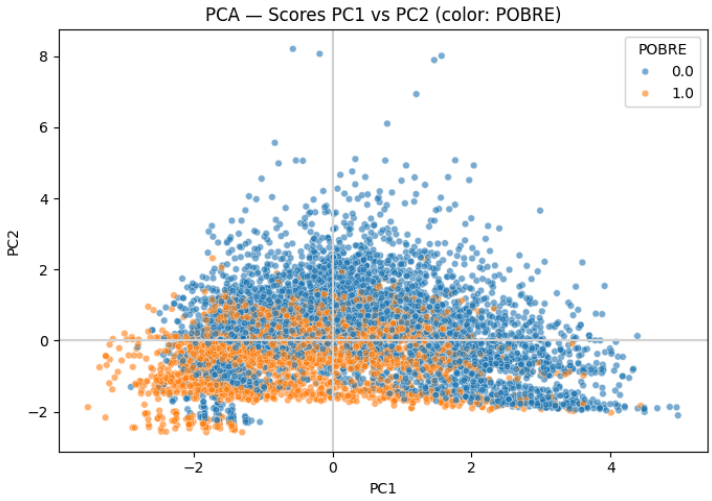


Tabla 2 *Tabla de loadings*

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
EDAD	0.643	-0.004	0.195	0.056	0.221	0.705
$EDAD^2$	0.633	-0.110	0.261	0.004	0.162	-0.702
EDUCACIÓN	0.037	0.637	-0.291	-0.510	0.494	-0.064
ITF 2025	-0.029	0.583	0.645	-0.110	-0.481	0.011
Miembros del hogar	-0.410	-0.011	0.543	0.297	0.670	-0.010
horas de trabajo	0.124	0.493	-0.314	0.798	-0.017	-0.081
Varianza explicada por componente	0.371	0.224	0.156	0.143	0.102	0.004

Varianza explicada acumulada	0.371	0.595	0.752	0.895	0.996	1.000

Figura 6

PCA - Loadings en el plano PC1-PC2

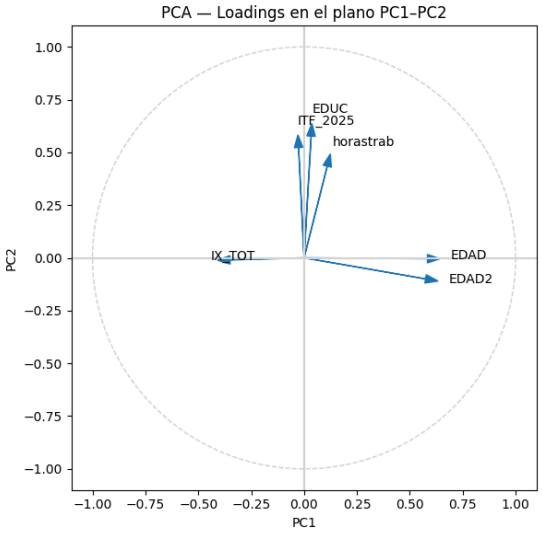


Figura 7

Gráfico de barras de PCA de varianza explicada por componente

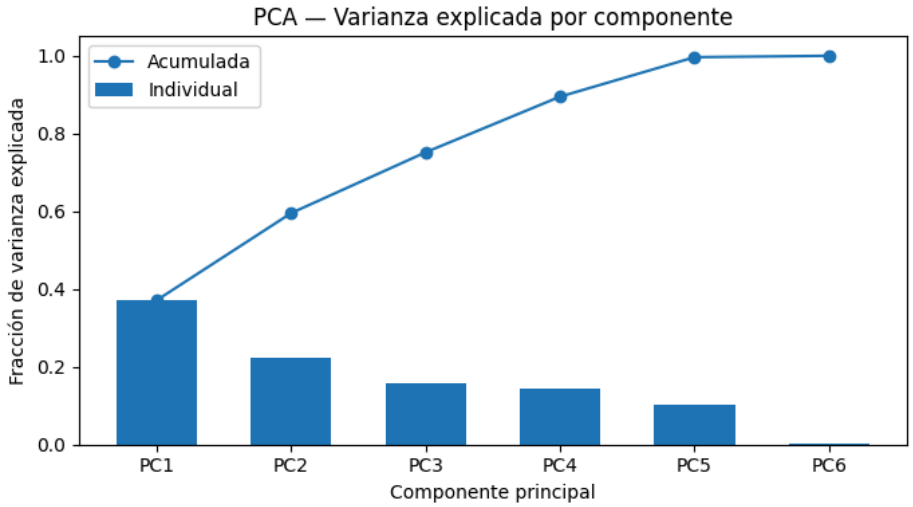


Figura 8

Gráfico de dispersión k=2

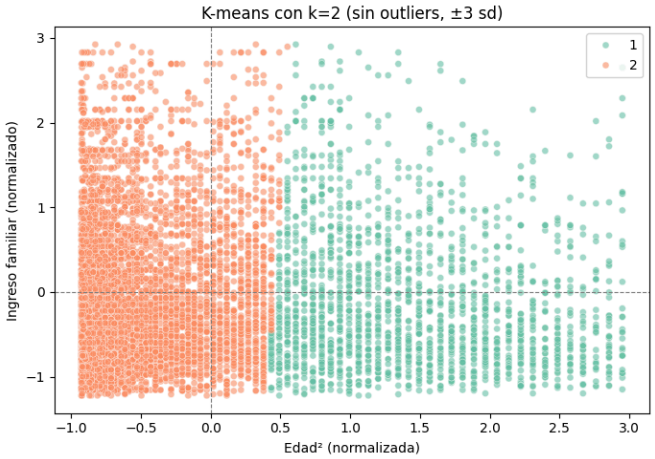


Figura 9

Gráfico de dispersión k=4

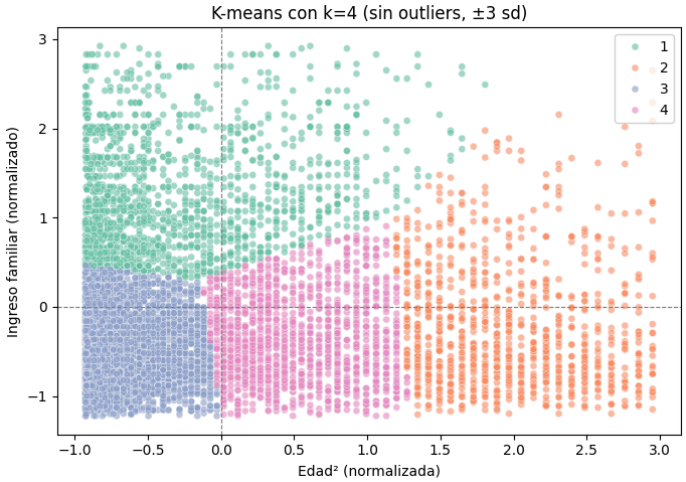


Figura 10

Gráfico de dispersión k=10

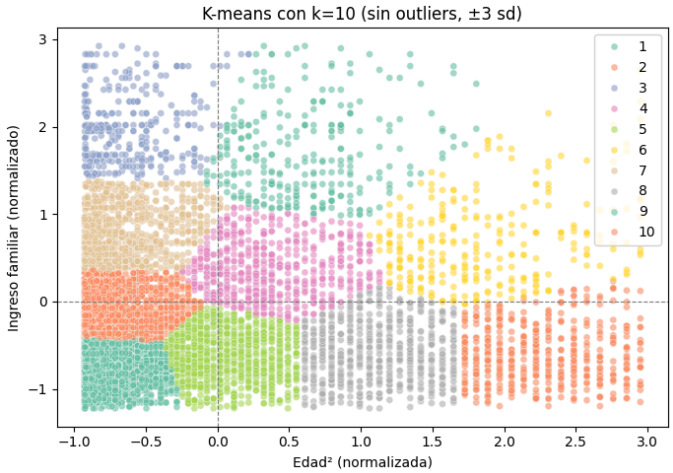


Figura 11

Gráfico de dispersión de clasificación real de pobres y no pobres

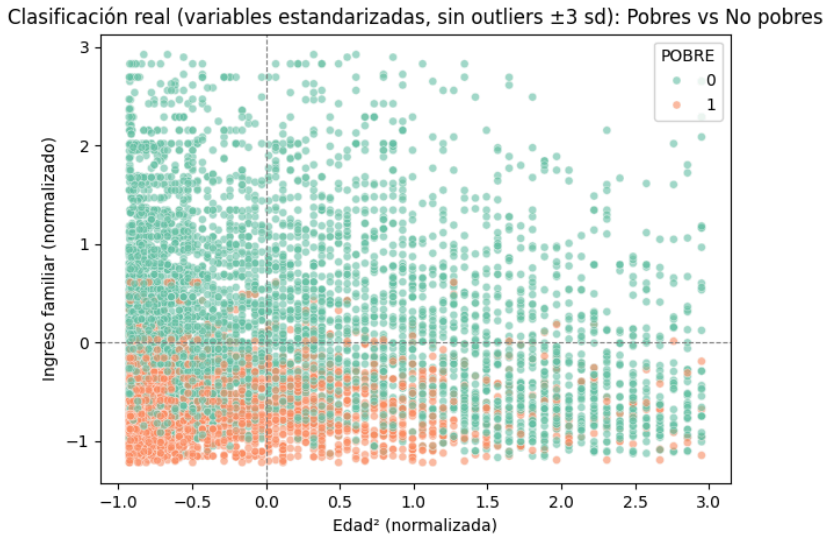


Figura 12

Método de codo - K-Means

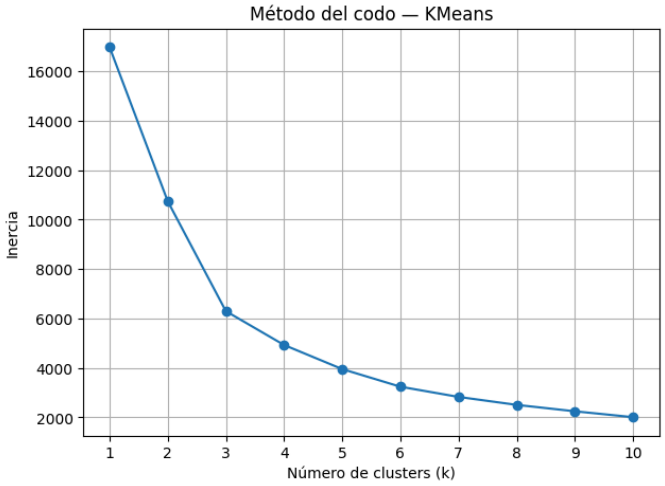


Figura 13

Gráfico de dispersión con k=3 según método del codo

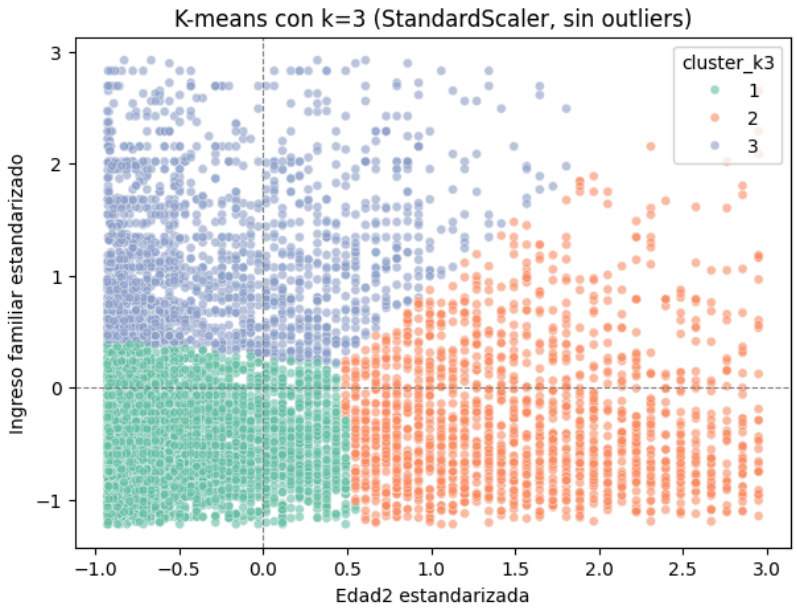


Figura 14

Cluster jerárquico - Dendograma