Materia : Ciencia de Datos

Alumnas: Amorena Inés, Campasso Delfina y Lauro Renata

Profesores: Anchorena Ignacio y Buscaglia Tomás

Primavera 2025

LINK REPOSITORIO: https://github.com/renatalauro/CC408-Grupo-T3-6

Este trabajo se realizó a partir de una base unificada de la Encuesta Permanente de Hogares para los años 2005 y 2025, y decidimos formalizarla en la región del Gran Buenos Aires. Por lo tanto realizamos un pipeline de limpieza y recodificación que nos permitió comparar las variables en el tiempo. Cuando logramos depurar y estandarizar la base, construimos nuevas variables y realizamos un análisis descriptivo y gráfico de acuerdo con las consignas que se nos asignaron.

Metodología

En primer lugar, preparamos la base de datos. Para ello, identificamos y filtramos la región GBA, implementando funciones como *find_region_col* y *mask_GBA* las cuales nos ayudaron a localizar la columna de región aun cuando tuviera distintos nombres o formatos. De esta forma, se seleccionaron únicamente los registros que correspondían a GBA.
Luego, homogeneizamos los nombres de las columnas a minúsculas y mantuvimos únicamente las variables presentes en ambas bases, con el objetivo de poder concatenarlas en un único dataset.
Posteriormente, procedimos con la normalización y limpieza de los datos. Para ello, estandarizamos la codificación de sexo, estableciendo a Varón como 1 y a Mujer como 2. Además, corregimos los ingresos negativos o fuera de rango. También depuramos edades mayores a 110 y se recodificaron respuestas textuales. Por último, en la variable “ingresos familiares” limpiamos los símbolos para convertirlos a valores numéricos.
De esta forma pudimos seguir con la recodificación, en la cual aplicamos mapeos para variables categóricas, como aglomerado o educación, entre otras. De este modo convertimos textos en códigos numéricos que pudieron ser comparables entre años.
Luego de finalizar con los mencionados pasos previos, pudimos comenzar a resolver el punto uno. Este proceso es indispensable, ya que la preparación de los datos nos permite luego realizar un análisis correcto. Posteriormente realizamos un control de calidad de datos, con el objetivo de asegurarnos de que todas las variables quedaran dentro de sus dominios esperados. Para ello, seleccionamos nuestras variables de interés. Verificamos la consistencia a partir de los tipos de datos y detectamos los posibles valores inesperados con un conteo de los valores más frecuentes. Luego, construimos un diccionario *dom* que estableció los valores correctos para cada variable categórica. Nuestro código se encargó de que, para cada variable, se eliminaran los valores faltantes, se compara lo observado con el conjunto permitido y se reporten los casos de valores fuera del dominio. Luego nos aseguramos de que se imprimiera “Normalizado” en caso de que ninguna variable presenta errores, o “Revisar columnas marcadas” en caso de que alguna variable tuviera un valor fuera del rango.

Resultados

En el punto 1 llevamos adelante la creación de la variable *Edad2* y realizamos un histograma de edad y un KDE de edad, usando un KDE simple sin dependencias. Ver resultados obtenidos en el anexo 1.1.

En el panel A podemos observar un histograma de edad. Muestra una distribución amplia de edades, desde recién nacidos a 100 años. Podemos observar una alta concentración de individuos entre las edades de 10 y 20 años, lo cual denota la presencia de una base etaria joven numerosa. La concentración de individuos desciende con el aumento de edad de forma gradual, lo cual evidencia una menor representación de adultos mayores en la muestra.

Por otro lado, en el panel B (ver anexo 1.2) se puede ver una distribución kernel de edad según pobreza. La curva de población pobre muestra un pico más pronunciado en las edades tempranas, entre los 10 y 15 años, lo cual indica que la pobreza afecta en mayor proporción a hogares con población joven. Mientras tanto, los no pobres muestran una distribución más homogénea, con mayor peso relativo en edades adultas, entre los 30 y 60 años, lo cual denota la influencia de la inserción laboral en la probabilidad de caer en la pobreza.

Para la consigna 2, tuvimos que crear la variable *educ,* la cual expresara los años de educación de los individuos. Para ello, implementamos un proceso de recodificación a partir de las variables originales CH12, la cual mostraba el máximo nivel alcanzado; CH13 (binaria, si finalizo el nivel =1, si no=2); y CH14, la cual expresaba el último año aprobado en el nivel correspondiente.
En primer lugar, convertimos las mencionadas variables a formato numérico, transformando cualquier texto o dato no válido en NaN. Luego creamos las bases *base_anios* la cual definía la cantidad de años acumulados antes de cada nivel educativo; y *Max_en_niv* que expresaba la cantidad máxima de años que se pueden cursar dentro de cada nivel. Posteriormente, se realizaron cálculos dentro del nivel. Es decir, tomamos CH14 como último año aprobado y, si el individuo afirmó haber finalizado en nivel CH13=1, entonces se le asignó el máximo posible del nivel. Por otro lado, si el valor de CH14 supera el máximo, se reemplaza por NaN. Por último, restringimos los valores negativos.
La variable *educ* tiene como objetivo sumar los años totales de educación, por lo tanto es el resultado de la suma de los años acumulados hasta el nivel y los años dentro de dicho nivel. Los resultados se redondearon y convirtieron a número entero.
Finalmente generamos la estadística descriptiva, generando los indicadores de media, desvío estándar, mediana, mínimo y máximo. Revisamos valores faltantes y chequeamos comparando entre 2005 y 2025 para verificar la consistencia. Ver los resultados obtenidos en el anexo 2.1.

En el caso de 2005, se recopilaron 8752 observaciones válidas. Se pudo ver que en promedio, las personas tenían casi nueve años de educación formal (8.79). También se nota una alta dispersión, siendo el desvío estándar 5.11; esto quiere decir que hay gente con muchos años de escolaridad y otros con muy pocos. Que el mínimo sea 0 denota que hay individuos sin años de escolaridad. La mediana es 8, lo que muestra que la mitad de la población tiene ocho años o menos de escolaridad, es decir, no tienen la secundaria completa. Finalmente, el máximo es 22, lo que denota la existencia de individuos con estudios universitarios largos o hasta posgrados realizados.

Por otro lado, se pudieron recopilar 6814 observaciones válidas en la base de 2025.Aunque es más chica que la de 2005, sigue siendo comparable. Pudimos observar que la media ascendió a casi 11 años (10.68) lo cual muestra una suba en el promedio de años de educación formal en GBA. Por otro lado, la dispersión se mantuvo estable, en 5.24. El mínimo continúa siendo 0. La mediana subió en un 50%, a 12 años alcanzados por la mitad de la población, alcanzando a completar el nivel secundario. El máximo se mantuvo estable en 22. Los valores de ambos años sugieren una expansión de la escolaridad obligatoria y finalización del secundario.

Para la consigna número 3, actualizamos la variable *ingreso_total_familiar* con los ingresos de 2005 a precios de 2025. Luego, analizamos la distribución de los ingresos mediante un histograma y un kernel, separando pobres y no pobres. En ambos casos incorporamos la línea de pobreza. En primer lugar, tuvimos que hacer que los ingresos de ambos periodos fueran comparables. Por lo tanto, aplicamos un factor de conversión basado en la Canasta Básica Total de adulto equivalente.

$$\text{factor} = \frac{\text{CBT 2025}}{\text{CBT 2005}} \approx \frac{365,177}{205.07} \approx 1781.2$$

De esta forma, llegamos a que $1 de 2005 equivale aproximadamente a $1781,2 del primer trimestre de 2025. Con estos valores, transformamos los ingresos familiares de 2005 en valores comparables con los de 2025 (anexo 3.1).
En el histograma del panel A, podemos observar que la mayoría de los hogares están concentrados en intervalos de ingresos muy bajos. Existen pocos hogares con ingresos muy altos. La línea punteada representa la mediana de la línea de pobreza en el hogar en 2025. Esta nos permite visualizar a los hogares que quedan por debajo del umbral de pobreza. La gran densidad de hogares acumulada cerca del origen denota un número significativo de hogares que viven con ingresos que rondan o son inferiores a la línea de pobreza.

En el caso de las distribuciones de kernel del panel B (ver 3.2 anexo), podemos ver una clara segmentación en base a la comparación de las densidades de las curvas. Se observa que la curva de pobreza se concentra en ingresos bajos, con una alta cantidad de población a la izquierda de la línea de pobreza. Por otro lado, la línea de los no pobres se desplaza hacia la derecha,mostrando una distribución más extendida y dispersa en torno a ingresos superiores a la línea de pobreza.

Para la consigna 4 construimos la variables *horastrab,* la cual adjunta la suma de horas trabajadas en la ocupación principal y en otras ocupaciones, restringida a jefes del hogar. En primer lugar, verificamos que las variables de origen (*pp3e_tot* y *pp3f_tot*) fueran numericas. Luego aplicamos un rango razonable de 0 a 56 horas semanales con el objetivo de evitar valores extremos que distorsionen la estadística (Ver anexo 4.1). Los resultados descriptivos totales indican una media de horas trabajadas semanales de 24,3, con una

desviación estándar alta (22,5 aprox) lo que refleja heterogeneidad. El valor mínimo es de 0, lo cual indica la existencia de jefes inactivos o desocupados. La mediana se ubica en 24 horas y el máximo de horas es 56.
En el caso de 2005, la media es de 26 horas semanales,mientras que en 2025 se redujo a 22,4. También se puede observar una disminución en la mediana, que era 30 en 2005 y es 24 en el 2025. Esto sugiere una disminución de la carga horaria semanal de los jefes de hogar.

Luego restringimos la muestra únicamente para jefes ocupados (estado=1). Los resultados cambiaron radicalmente, con una media de 33,7 horas semanales y una mediana de 40 horas semanales, lo cual indica que la mayoría de los ocupados trabaja jornada completa. El máximo es 56, lo cual denota presencia de jefes con jornadas de trabajo extendidas.

Para la consigna 5, calculamos el tamaño final de la base unificada, limpia y homogeneizada de 20025 y 2025. Partimos de la base *base_gba1_clean1*, la cual ya contenía las variables depuradas y homogeneizadas. Luego,para la variable definida como *ano4,* se calcularon para cada año y para el total la cantidad de: observaciones; observaciones con NAs en la variable *Pobre*; hogares pobres (*itf_bajo_cbt_hogar = 1*); hogares no pobres (*itf_bajo_cbt_hogar = 0*) y el número de variables limpias y homogeneizadas en la base final.

	2005	**2025**	**Total**
Cantidad de observaciones	9484	7181	16665
Cantidad de observaciones con NAs en la variable “Pobre”	113	2872	2895
Cantidad de pobres	3301	1790	2985
Cantidad de no pobres	6070	2519	8589
Cantidad de variables limpias y homogeneizadas	178	178	356

En base a la tabla, podemos comparar la proporción de NAs en cada año, habiendo en 2005 un 1,19% de NAs versus un 39,99% en 2025, lo cual importa a la hora de comparar ambos períodos. En el total de la muestra, los pobres representan un 37,2% de los hogares, aunque en 2005 se observa una proporción más alta en términos absolutos respecto de 2025. La presencia de 178 variables homogéneas por año nos asegura la consistencia para análisis comparativos.

Para el ejercicio 1 de la parte II, con el objetivo de observar dependencias lineales entre variables, generamos una matriz de correlación de Pearson entre: *edad, edad2, educacion, ingreso total familiar (ITF), cantidad de miembros del hogar (IX_TOT) y horas trabajadas.* Ver anexo II.1.

Podemos observar que las variables Edad y Edad2 correlacionan casi perfecto (0.97), lo cual era esperado, ya que *edad2* es en función de *edad*.Por otro lado, existe una correlación negativa entre *edad* y *miembros del hogar* de (-0,42), lo cual indica que a mayor edad promedio, menos integrante en el hogar. Además, hay una baja correlación positiva entre *educación* y *ITF,* sobre la cual podemos ver que a mayor educación, mayor ingreso. Por último, vemos que las horas trabajadas no se correlacionan prácticamente con ninguna variable.

En segundo lugar, graficamos los ponderadores de PCA para el primer y segundo componente. Para ello, estandarizamos los z-scores y aplicamos un PCA con dos componentes principales. A partir de eso obtuvimos scores y loadings (Ver II.2 del anexo).

Por un lado, el PC1 explica el 39.2% de la varianza, que está muy cargada en edad (0.617), en edad2 (0.599) y en miembros del hogar (0.431). Esto puede interpretarse como un eje que asocia ciclo vital y hogar, es decir, que los hogares grandes suelen tener adultos jóvenes y los pequeños, adultos mayores. Por otro lado, el PC2 explica el 19.3% de la varianza, y cargan fuertemente educación (0.601) e ITF (0.633). Podemos pensarlo como un eje que asocia aspectos socioeconómicos y laborales.En total, las dos PC explican el 58.5% de la variabilidad total.

Luego, alineadas con la consigna 3, creamos un biplot que nos dejó combinar la dispersión de los individuos en el plano de los dos componentes principales; y la representación de las variables originales como vector,que indicaran su peso y dirección. Cada flecha está dibujada con un color distinto, y en la parte superior del gráfico se incluye una leyenda donde ese mismo color se asocia al nombre de la variable correspondiente. De esta forma, la correspondencia entre flecha y variable no se da mediante un texto sobre la flecha, sino a través de la coincidencia de colores entre la flecha en el plano y el nombre en la leyenda (ver anexo II.3).

La nube de puntos azules muestran la distribución de los datos en el espacio PC1-PC2. En cuanto a la dispersión de los individuos, la mayor concentración está alrededor del origen, lo cual refleja que la mayoría de los individuos se encuentran cerca de los valores promedio. En relación con las flechas, podemos ver que las de *edad y edad2* apuntan hacia la derecha, indicando que PC1 captura principalmente la dimensión etaria. Por otro lado, la flecha naranja, que calcula el número de miembros en el hogar se direcciona opuestamente a edad. Esto muestra que pC1 distingue hogares grandes de individuos de mayor edad. Finalmente, las variables *educacion, ITF, y horas trabajadas* tienen una dirección hacia arriba, asociadas al PC2, representando un eje asociado a una dimension socioeconomica-laboral.

En el ejercicio 4 buscamos analizar cuánta variabilidad explican los seis componentes principales extraídos de las variables seleccionadas. Para eso graficamos el scree plot y la varianza explicada acumulada. Obtuvimos una tabla de resultados numéricos y los gráficos (ver anexo II.4).

En el scree plot podemos observar que el PC1 explica la mayor proporción de varianza (39,2%), y que los dos primeros componentes juntos explican más de la mitad de la variabilidad (58.4%). Consecuentemente, si bien PC3 aún aporta información relevante, a partir del tercer componente, la varianza explicada adicional disminuye. Tanto PC5 como PC6 aportan muy poca varianza. En pocas palabras, con los primeros dos o tres componentes se puede retener la may parte de la estructura de datos.

En el caso del gráfico de la varianza acumulada (Ver anexo II.4.2), podemos ver que con los primero tres componentes podemos analizar el 75,3 de la variabilidad total. Aún con cuatro se llega a 88.6, el cual está en un nivel alto de retención de información. Aunque con cinco o seis componentes se explica casi el 100%, el costo de complejidad adicional se eleva demasiado y no se gana tanta interpretabilidad.

Con el objetivo de resolver la consigna 5, aplicamos el algoritmo de k-means utilizando como variables la edad al cuadrado (edad²) y el ingreso total familiar (ITF). Se probaron tres configuraciones con distinto número de clusters: k=2, k=4 y k=10. Los resultados muestran que, en todos los casos, el eje de mayor peso en la separación es el ITF, ya que los centroides y los límites entre los grupos se ubican casi exclusivamente sobre esta variable. Esto indica que el ingreso es la dimensión que efectivamente estructura los agrupamientos, mientras que la edad² apenas aporta matices secundarios (ver anexo II.5). En el caso de k=2, observamos que la partición de los datos se organiza fundamentalmente en torno al ingreso total familiar (ITF), variable que estructura de manera más clara la separación de los grupos. Si bien la edad² se incluyó en el modelo, su efecto resulta secundario frente a las diferencias de ingreso. El valor del silhouette (=0.52) indica una calidad de agrupamiento moderada. En el caso de k=4, la segmentación se vuelve más consistente (silhouette = 0.59) y aparecen subgrupos más interpretables que combinan distintos niveles de ingreso y perfiles etarios. Por ejemplo, es posible distinguir hogares de bajos ingresos, otros de ingresos medios y altos, con algunas diferencias internas según la edad. Este valor de k parece ser el más adecuado, ya que genera clusters más plausibles desde el punto de vista socioeconómico.Por último, con k=10 los clusters se fragmentan en exceso y pierden interpretabilidad. Aunque el silhouette (=0.52) no es bajo, el número elevado de grupos genera superposición y dificulta dar sentido a los perfiles resultantes.

En relación con la consigna sobre la capacidad de k=2 de separar pobres y no pobres, realizamos una matriz de confusión para verlo con claridad. Los resultados fueron los siguientes:

Predicho (k=2)

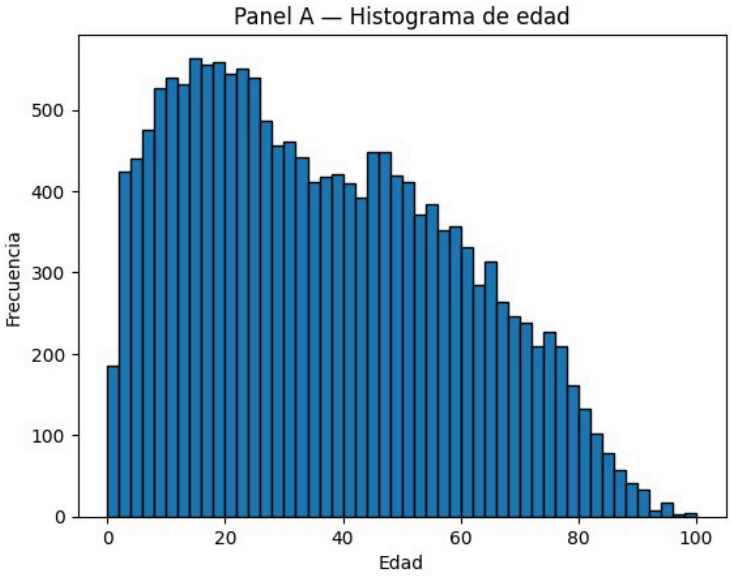
Real (1=pobre)		
	0	8520
	1	4999

En cuanto a la capacidad de k=2 para separar a pobres y no pobres, los resultados muestran limitaciones. La matriz de confusión evidencia que muchos hogares pobres no son correctamente identificados, lo cual se explica porque el algoritmo prioriza la varianza del ITF como criterio de agrupación y no la definición normativa de pobreza (ITF vs. CBT). En consecuencia, si bien el ITF es la variable que estructura los clusters, el modelo no logra reproducir fielmente la división entre pobres y no pobres.

Alineadas con la consigna 5b, realizamos un gráfico del método del codo, aplicado al clustering con K-means (ver anexo II.5.1). En el gráfico podemos observar la evolución de la inercia a medida que aumenta el número de clusters (k). para k=1, vemos una inercia muy alta, ya que todos los datos se encuentran concentrados en un único grupo. A medida que aumenta el número de clusters, la inercia disminuye de marcadamente hasta k=3. A partir de ese punto, la pendiente de la curva se aplana, lo cual denota el poco beneficio marginal de agregar más clusters. En conclusión, el número óptimo de clusters es 3 o 4 como mucho.
En cuanto a la interpretación socioeconómica del gráfico, con k=3 podemos distinguir entre hogares vulnerables o pobres, que cuentan con bajos ingresos, baja educación y un mayor número de integrantes por hogar; hogares intermedios, que cuentan con ingresos medios y niveles educativos moderados; y hogares de mayor nivel socioeconómico, con ingresos y educación más alto y un número de integrantes en el hogar más reducido.

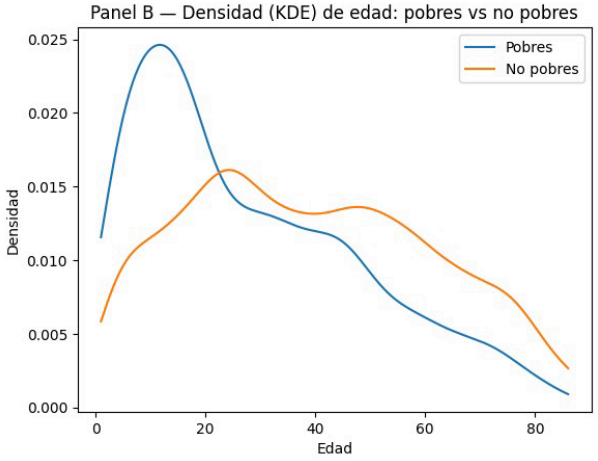
En la consigna 6 elaboramos un dendrograma de clustering jerárquico. Este gráfico en forma de árbol representa cómo se van agrupando progresivamente las observaciones según su similitud. La altura a la que se unen las ramas indica la distancia entre los grupos, de modo que uniones más bajas reflejan mayor semejanza. Al definir un nivel de corte en el árbol, es posible determinar el número de clusters resultante. El eje Y representa la distancia o disimilitud entre grupos: cuanto más baja es la unión, mayor es la semejanza entre los elementos fusionados; cuanto más alta, mayor es la diferencia. En el eje X se ubican los clusters hoja, que en este caso aparecen resumidos en 21 grupos, con el número entre paréntesis indicando la cantidad de observaciones contenidas en cada uno (la suma de esos valores corresponde al total de la base, N = 16.665). A partir de la estructura del dendrograma se observan tres conglomerados principales que se forman a una distancia intermedia (aprox. entre 100 y 150), lo cual sugiere la existencia de tres grupos diferenciados en la muestra. Si se estableciera un punto de corte mayor (por ejemplo, alrededor de 170) los clusters se reducirían a dos grandes grupos, mientras que con un corte más bajo podrían identificarse cuatro o más subgrupos.

Anexo:



1.1

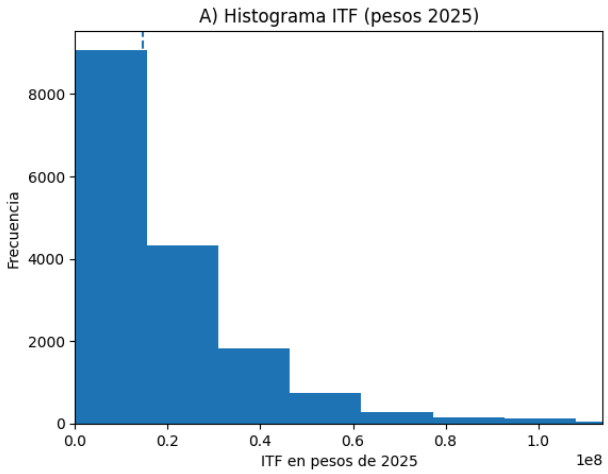
1.2



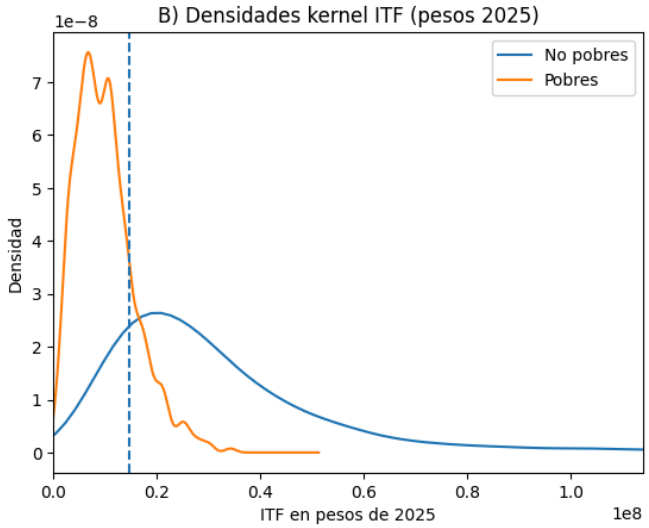
2.1

ano4	count	mean	std	min	50%	max
2005	8752.0	8.796161	5.11453	0.0	8.0	22.0
2025	6814.0	10.680804	5.240058	0.0	12.0	22.0

3.1



3.2



4.1

Estadístico general

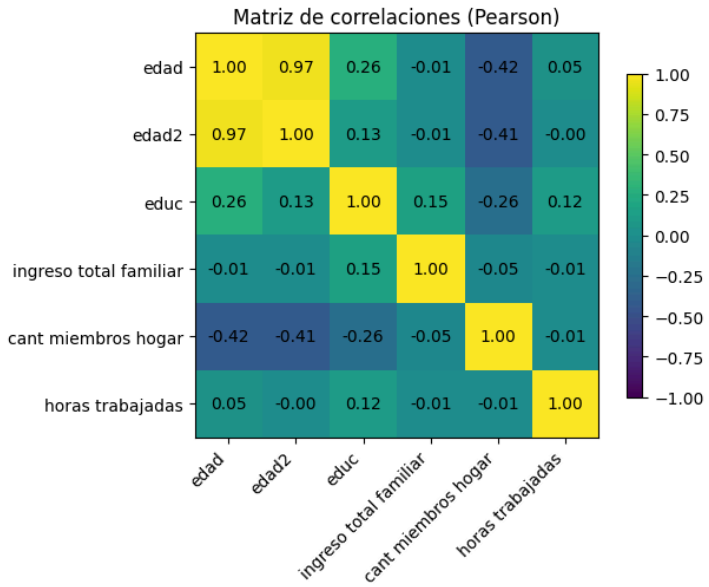
Estadístico	Valor
mean	24.331772
std	22.488368
min	0.0
50% (mediana)	24.0
max	56.0

Descriptivas por año:

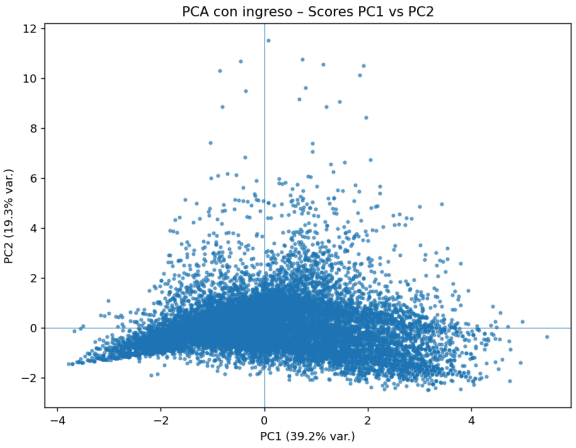
año	mean	std	min	50%	max
2005	26.030921	23.354655	0.0	30.0	56.0
2025	22.415102	21.311697	0.0	24.0	56.0

ano4	mean	std	min	50%	max
2025	33.706628	17.386702	0.0	40.0	56.0

II.1

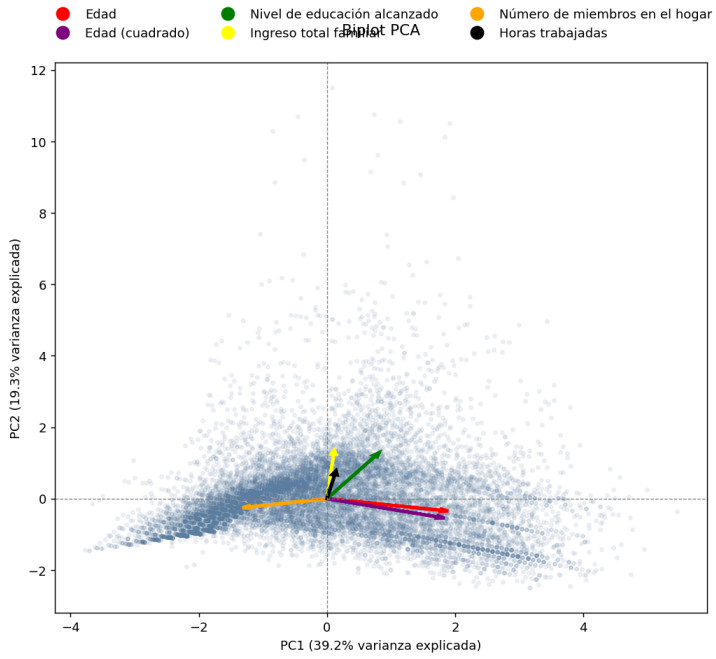


II.2

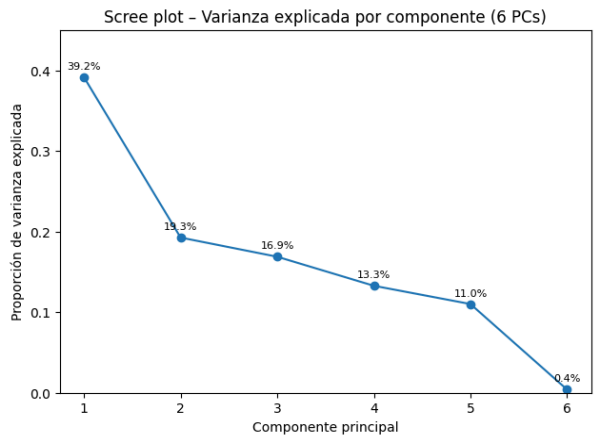


Variable	PC1	PC2
edad	0.617	-0.161
edad^2	0.599	-0.253
educ	0.266	0.601
ITF	0.037	0.633
IX_TOT	-0.431	-0.118
horastrab	0.046	0.366

II.3

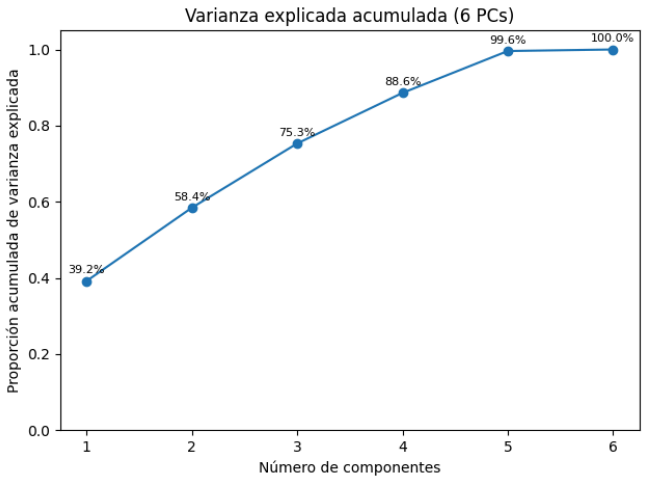


II.4

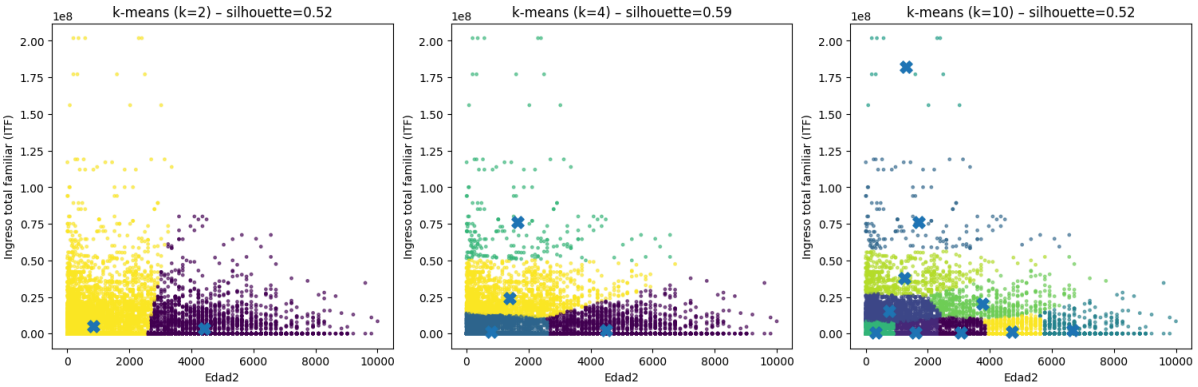


PC	Varianza explicada	Varianza acumulada
PC1	0.3916	0.3916
PC2	0.1926	0.5842
PC3	0.1688	0.7530
PC4	0.1328	0.8859
PC5	0.1101	0.9960
PC6	0.0040	1.0000

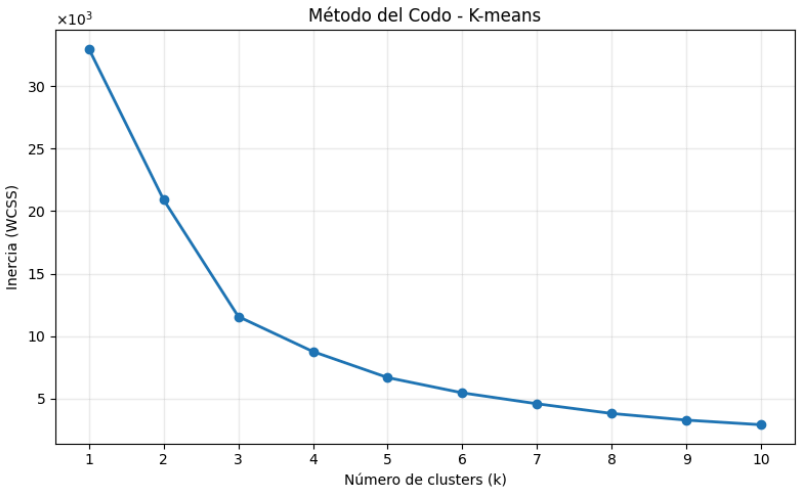
II.4.2



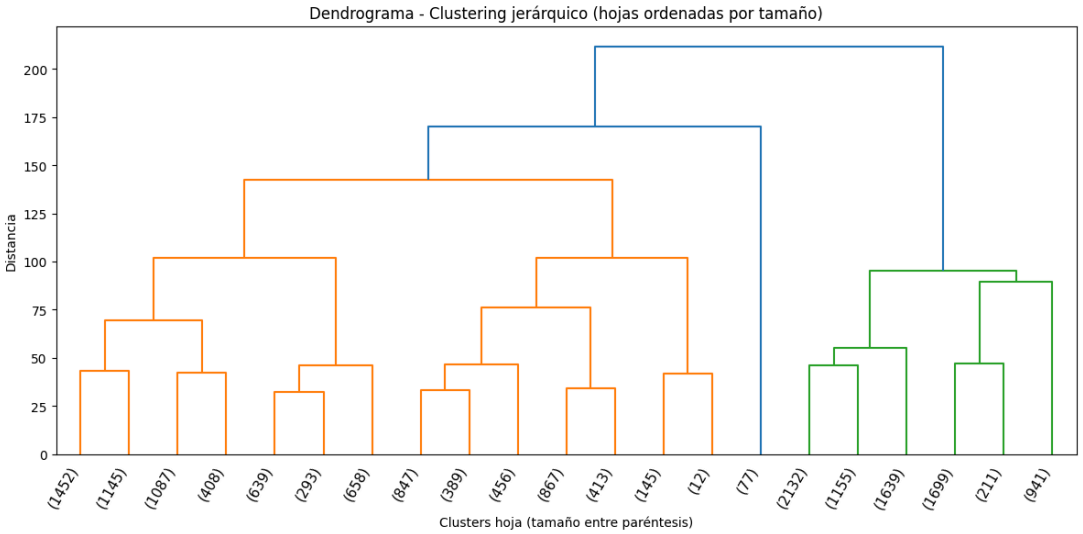
II.5



II.5.1


Método del Codo - K-means
×10³
Inercia (WCSS)
Número de clusters (k)
30
25
20
15
10
5
1
2
3
4
5
6
7
8
9
10

II.6


Dendrograma - Clustering jerárquico (hojas ordenadas por tamaño)
Distancia
Clusters hoja (tamaño entre paréntesis)
200
175
150
125
100
75
50
25
0
(1452)
(1145)
(1087)
(408)
(639)
(293)
(658)
(847)
(389)
(456)
(867)
(413)
(145)
(12)
(77)
(2132)
(1155)
(1639)
(1699)
(211)
(941)