Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Taller de Programación

Maestría en Economía Aplicada

Tercer trimestre 2025

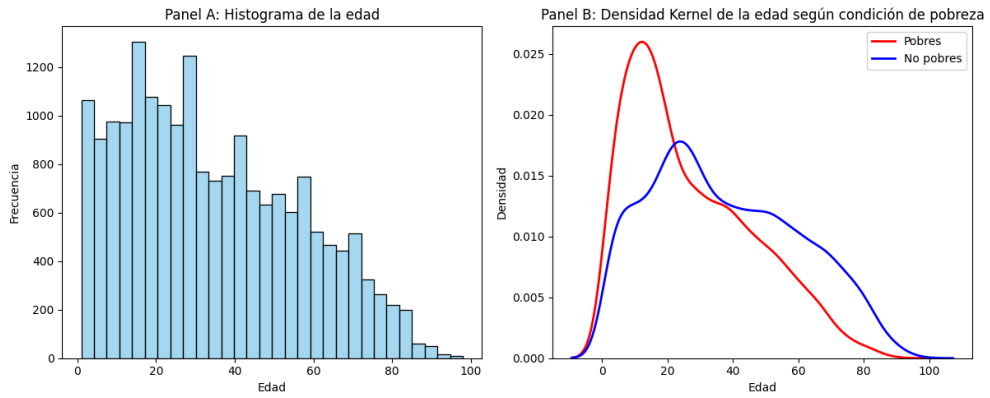
Trabajo Práctico N° 2

Link GitHub https://github.com/jorge-ux272/Big-Data-UBA--Grupo-8

Docente: Noelia Romero

Alumnos:
Jorge Eduardo Bolaños Gamarra
Mario Antonio Valdivia Reyes
Héctor Sebastián San Martin

Parte I: Creación de variables, histogramas, kernels y resumen de la base de datos final1.
Edad al cuadrado y distribución



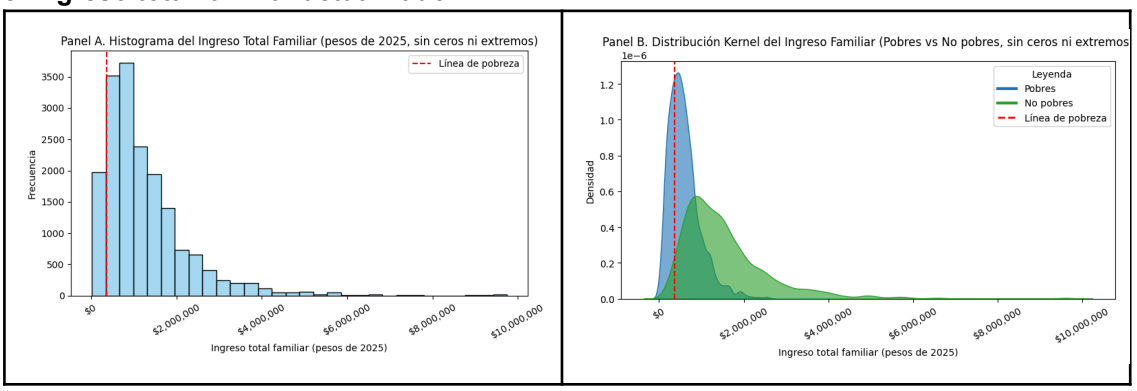
La población relevada es predominantemente joven, con mayor concentración entre los 10 y 40 años. No obstante, se observa que los individuos pobres tienden a ser más jóvenes, en comparación a los no pobres. Esto sugiere que la pobreza está asociada a una mayor densidad de población joven, lo cual puede tener implicancias en términos de vulnerabilidad económica, acceso a oportunidades y diseño de políticas sociales focalizadas.

2. Años de educación

Estadístico	Valor
Cantidad de casos válidos	18,622
Casos faltantes	775
Media	9.94
Mediana	10.00
Moda	13.00
Desvío estándar	5.06
Varianza	25.62

La población de la región Pampeana se concentra en niveles medios de educación, con valores modales en torno al secundario completo. Esto sugiere una población con acceso educativo relativamente extendido, aunque con dispersión significativa. La amplitud del rango y los valores extremos permiten explorar desigualdades educativas por edad, pobreza o región.

3. Ingreso total familiar actualizado



El ingreso familiar muestra una distribución desigual y sesgada a la derecha, con la línea de pobreza indicando que 25-30% de la población está por debajo del umbral. Las densidades de kernel muestran por separado a pobres y no pobres, resaltando que los ingresos bajos se concentran entre los pobres, aunque hay cierta superposición.

4. Horas trabajadas totales

Estadístico	Valor
Media	40.46
Desvío estándar	18.24
Mínimo	1
Mediana	40
Máximo	99

La media indica una carga laboral relativamente baja en términos generales para la población de esta región, pero el desvío estándar elevado (que incluso supera a la media) y el máximo de 99 horas reflejan una alta dispersión, con presencia de casos extremos. La mediana en cero sugiere que al menos la mitad de los individuos no trabajaron en la semana relevada, lo que puede incluir personas desocupadas, inactivas o con trabajo informal no captado.

5. Tamaño de la muestra

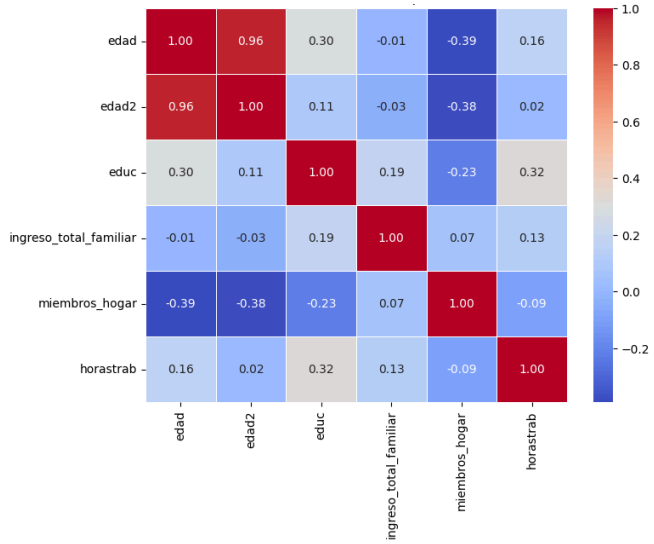
Tabla 1. Resumen de la base final para la región Pampeana

	2005	2025	Total
Cantidad observaciones	14.651	4.746	19.397
Cantidad de observaciones con NAs en la variable “Pobre”	170	1.406	1.576
Cantidad de Pobres	4.195	1.652	5.847

Cantidad de No Pobres	10.286	1.688	11.974
Cantidad de variables limpias y homogeneizadas	26	23	49

Parte II: Métodos No Supervisados

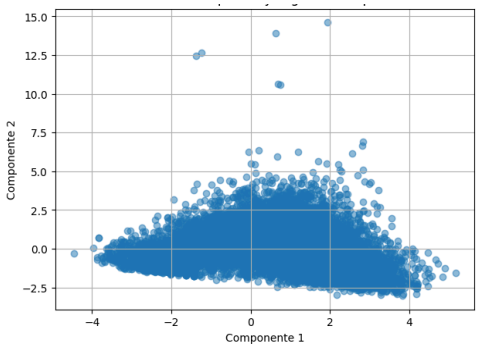
1. Matriz de correlaciones con los seis predictores para la región



Se observa una correlación muy alta entre "edad" y "edad2", dado que "edad2" es una transformación de la anterior. "Educación" tiene correlaciones moderadas con "edad" y "horas de trabajo". Por su parte, "Ingreso total familiar" muestra correlaciones débiles con las demás variables, excepto con "miembros del hogar" y "horas de trabajo". Las variables "miembros del hogar" y "horas trabajadas" presentan una correlación baja y negativa.

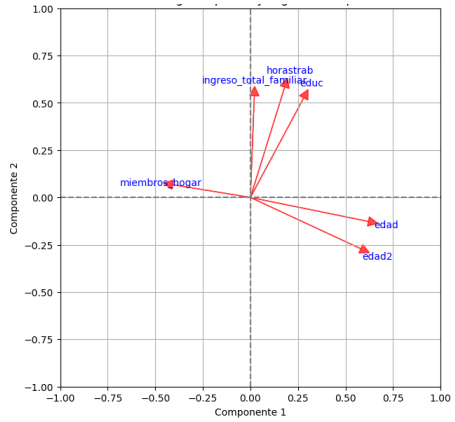
2. PCA con ingreso

Dispersión de los índices (scores) del primer y segundo componente PCA



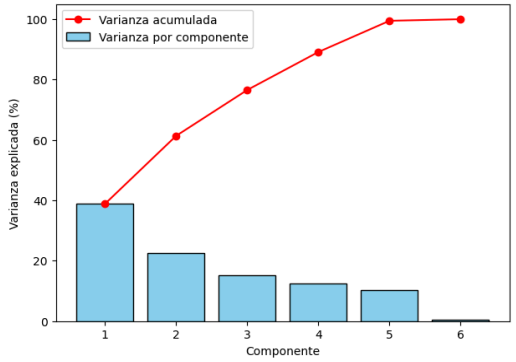
La mayoría de los datos se concentran alrededor de los valores cercanos a 0 en ambos componentes, por lo cual se forma una distribución densa y simétrica con valores similares en estos componentes. Esto sugiere que el primer y segundo componente explican una porción significativa de la varianza total; donde el segundo componente exhibe una amplitud menor al primero.

3. Ponderadores loading



Hay diferencias en las direcciones y magnitudes de las contribuciones de las variables a los componentes 1 y 2. El ingreso total familiar, las horas trabajadas y la cantidad de años de educación contribuyen de forma positiva en ambos componentes, sobre todo al segundo (más de 0.50). Por su parte, las variables “edad” y “edad2” contribuyen de forma positiva e importante al componente 1 (más de 0.50). Finalmente, “miembros_hogar” también contribuye más a este componente, aunque negativamente.

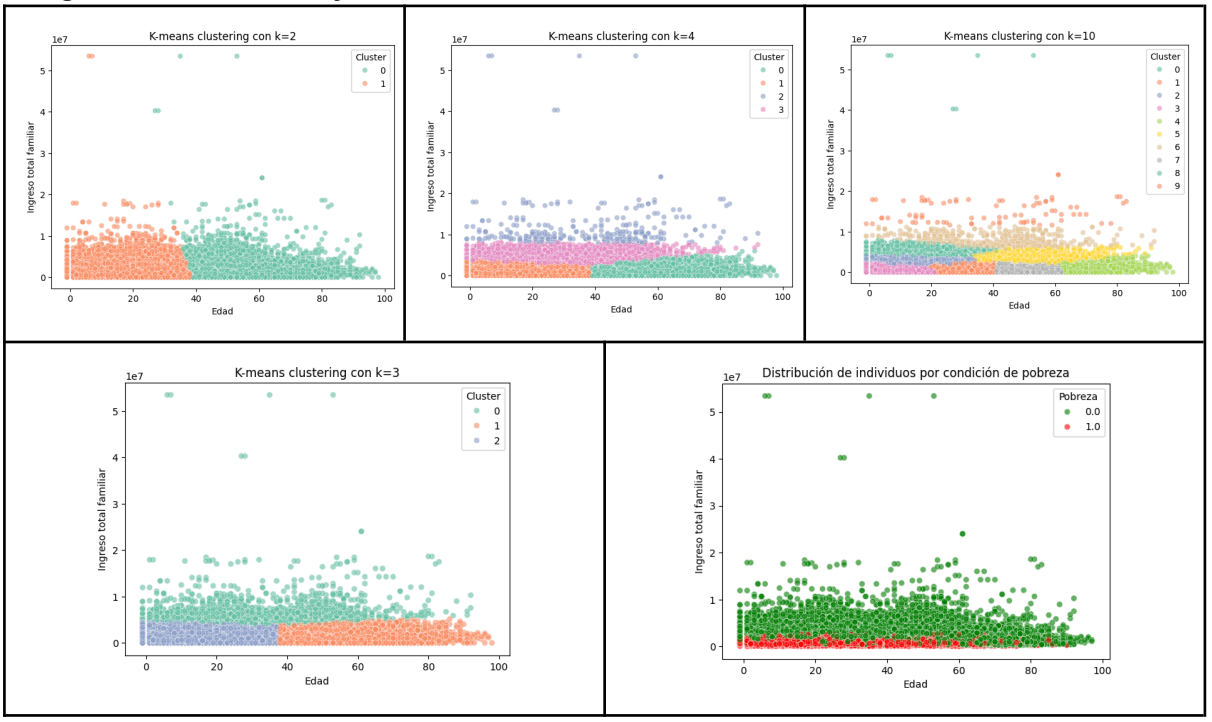
4. Proporción de varianza explicada



El primer componente explica aproximadamente el 40% de la varianza total, por lo cual se lo debe considerar como el más significativo, mientras que el segundo componente contribuye un 20% de la varianza, sumándose hasta ese punto un 60% acumulado. Dos componentes son suficientes para capturar la mayor parte de la variabilidad de los datos analizados en este informe.

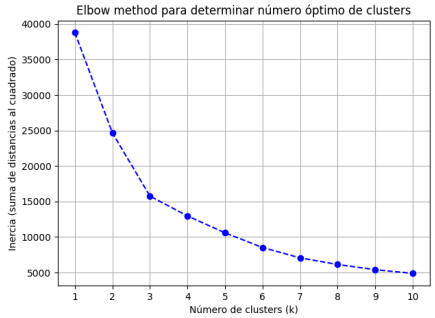
5. Cluster

a. Algoritmo con k=2, k=4 y k=10



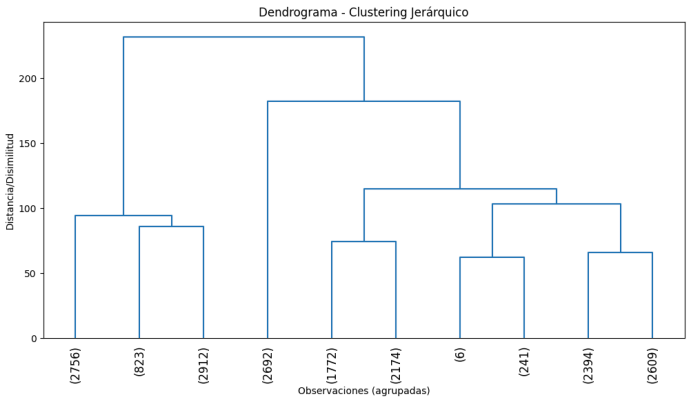
Al aplicar el modelo K-medias, se obtiene como resultado que con k=2 los clusters no logran separar de manera perfecta a los pobres y no pobres, aunque sí se refleja parcialmente la condición socioeconómica. Por otro lado, al aumentar a k=4 o k=10, surgen subgrupos que combinan ingresos y edades, evidenciando de manera más clara distintos niveles socioeconómicos en esta región.

b. Medida de disimilitud



La inercia disminuye rápidamente al aumentar el número de clusters de 1 a 3; a partir de ese punto se observa un cambio notable, sugiriendo que el "codo" óptimo está alrededor de k=3, ya que a partir de ahí la reducción de la inercia se vuelve más gradual. Con k=3, se logra clasificar de manera más o menos realista los casos de pobreza, pero se sobrestima el resultado, en comparación a la distribución observada.

6. Cluster jerárquico



Un dendrograma es un diagrama en forma de árbol que representa la estructura de agrupamiento de un conjunto de observaciones en un análisis de clustering jerárquico. Las alturas en el eje vertical representan la distancia o disimilitud entre las observaciones, donde una mayor altura indica mayor diferencia.

7. Cluster k-moda

Clúster	k=2 (POBRE=0)	k=2 (POBRE=1)	k=4 (POBRE=0)	k=4 (POBRE=1)	k=10 (POBRE=0)	k=10 (POBRE=1)
0	6785	2485	921	761	1550	1507
1	5189	3362	5510	1774	2499	713
2	-	-	3467	2400	563	479
3	-	-	2076	912	735	565
4	-	-	-	-	1478	127
5	-	-	-	-	1084	540
6	-	-	-	-	1149	166
7	-	-	-	-	1423	517
8	-	-	-	-	749	574
9	-	-	-	-	744	659

Con K-moda, los clusters contienen mezclas significativas de pobres y no pobres, reflejando una separación poco efectiva o imperfecta. Aumentar el número de clusters permite identificar subgrupos socioeconómicos más finos, aunque todavía no se logra una clasificación totalmente precisa.