



## **Trabajo Práctico 2**

### **Ciencia de Datos**

#### **Alumnas**

Juana Cisneros, Francesca Lo Tartaro , Lucía Martín

#### **Profesores**

María Noelia Romero, Ignacio Anchorena

#### **Asignatura**

Ciencia de Datos

Teórica 1 - Tutorial 2.

#### **Semestre y año de presentación**

2do Semestre 2025

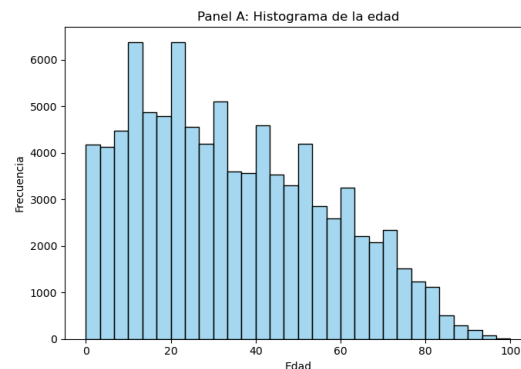
Link al repositorio: <https://github.com/lucia1martin/Trabajo-Practico-2>

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

1)

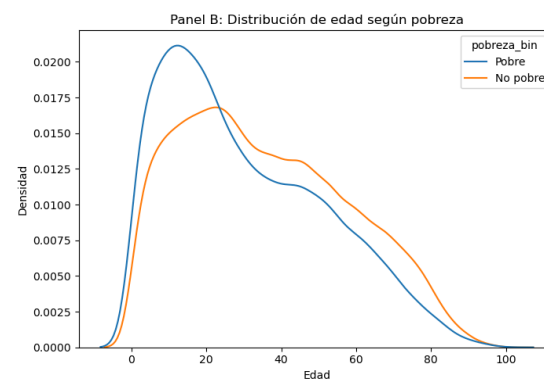
**Figura 1**

*Histograma de la variable edad*



**Figura 2**

*Distribución de edad según pobreza*



El histograma (**Figura 1**) muestra mayor concentración de jóvenes, sobre todo entre 0 y 20 años, y una caída progresiva desde los 30, típica de una pirámide poblacional. En la **Figura 2**, la curva de pobreza alcanza su pico en niñez y adolescencia, indicando más menores en hogares pobres; con la edad la brecha se reduce y, desde los 60, la población total desciende por efecto de la mortalidad.

2)

**Figura 3**

Estadística de educación (Años de educación)	
Count	64431.00

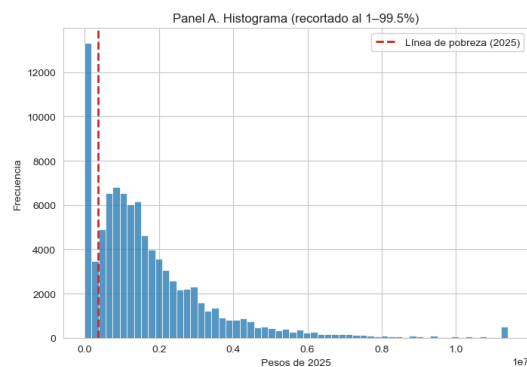
<b>Promedio</b>	11.12
<b>Desviación estándar (sd)</b>	9.22
<b>Mínimo</b>	0.00
<b>Mediana (p50)</b>	13.00
<b>Máximo</b>	111.00

La variable de educación muestra un promedio de 11,1 años y una mediana de 13, lo que indica que la mayoría completó al menos el secundario. El mínimo de 0 refleja casos sin escolaridad y el máximo de 111 posibles errores o atípicos. La alta dispersión (desvío estándar de 9,2) se relaciona con la heterogeneidad etaria y problemas de registro.

3)

**Figura 4**

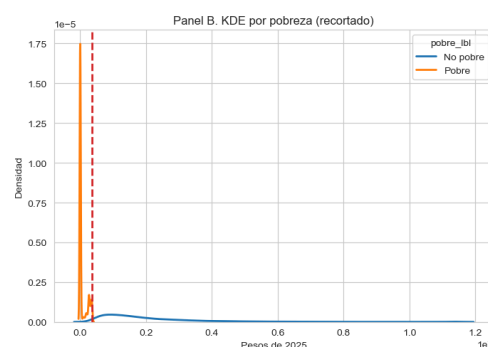
*Histograma del ingreso total familiar (a precios de 2025).*



El histograma revela una fuerte concentración de hogares en ingresos bajos y una distribución asimétrica a la derecha. La mayoría se ubica cerca de la línea de pobreza (marcada en rojo), lo que evidencia vulnerabilidad, mientras que solo unos pocos alcanzan ingresos muy altos.

**Figura 5**

*Distribución Kernel del ingreso total familiar según condición de pobreza.*



La distribución muestra que los hogares pobres se concentran en ingresos muy bajos y por debajo de la línea de pobreza, mientras que los no pobres exhiben mayor dispersión y presencia en niveles altos. La comparación evidencia una marcada brecha y fuerte desigualdad en los ingresos.

4)

**Figura 6***Estadística de 'horastrab' por año (sólo jefes/as):*

Año	2005	2025
N	13597.00	9904.00
Promedio	29.52	37.38
Desviación estándar (sd)	50.54	39.90
Mínimo	0.00	0.00
Mediana (p50)	28.00	40.00
Máximo	1998.00	999.00

Entre 2005 y 2025 se observa un aumento en las horas trabajadas por los jefes/as de hogar (promedio de 29,5 a 37,4 y mediana de 28 a 40). No obstante, los valores extremos y la presencia de ceros sugieren posibles errores de registro.

**Figura 7***Estadística de 'horastrab' (solo jefes/as, total):*

2005 & 2025	
N	23501.00
Promedio	32.83
Desviación estándar (sd)	46.52
Mínimo	0.00
Mediana (p50)	35.00
Máximo	1998.00

El promedio de horas trabajadas por los jefes de hogar es de 37,4 semanales y la mediana de 40 coincide con la jornada estándar. La alta dispersión (39,9) indica casos extremos: desde 0 horas de desocupados hasta un outlier de 999 horas por probable error de carga. En general, la mayoría se concentra en torno a la jornada típica, aunque con registros atípicos que requieren tratamiento.

5)

**Figura 8***Resumen de la base final para la región y año (2005–2025).*

	2005	2025	Total
Cantidad de Observaciones	47030	45425	92455
Cantidad de observaciones con NAs en la variable "Pobre"	0	0	0

<b>Cantidad de Pobres</b>	2662	13954	16616
<b>Cantidad de No Pobres</b>	44368	31471	75938
<b>Cantidad de variables limpias y homogeneizadas</b>	151	151	151

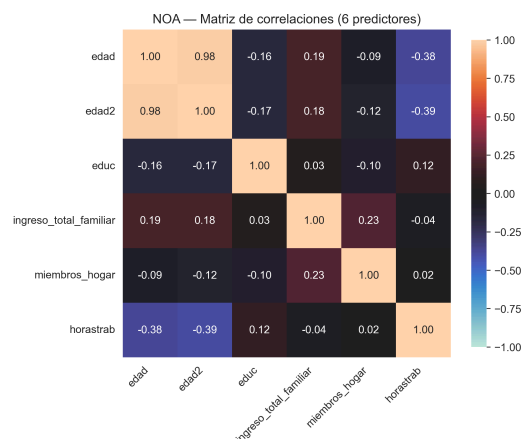
La base unificada tiene 92.455 observaciones (2005: 47.030; 2025: 45.425) y 0 NAs en la variable *Pobre*, señal de buena calidad de limpieza. La cantidad de pobres pasa de 2.662 en 2005 a 13.954 en 2025; eso implica una tasa de pobreza aproximada de 5,7% en 2005 y 30,7% en 2025 (total período: 18,0%). Los no pobres suman 75.938 en el total del panel. Se mantuvieron 167 variables limpias y homogeneizadas en ambos años, lo que permite comparaciones consistentes entre 2005 y 2025.

## Parte II: Métodos No Supervisados

1)

### Figura 9

*Matriz de correlaciones (NOA)*



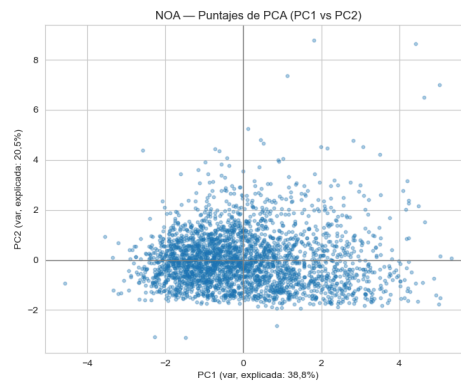
La edad y edad<sup>2</sup> muestran alta correlación por construcción. Con mayor edad disminuyen las horas trabajadas, mientras que el ingreso familiar se asocia débilmente de forma positiva al tamaño del hogar. La educación presenta correlaciones bajas con el resto, indicando asociaciones limitadas.

### A. PCA

2)

### Figura 10

*PCA: puntajes PC1 vs PC2 (NOA)*

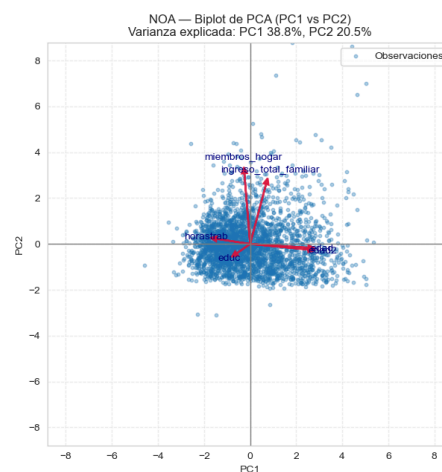


El primer componente principal explica el 38,8% de la varianza y el segundo un 20,5%, sumando cerca del 60% total. Los puntos se concentran alrededor del origen, sin agrupamientos claros, aunque con algunos valores atípicos en ambas direcciones.

3)

**Figura 11**

*Biplot PCA: puntajes y loadings (NOA)*

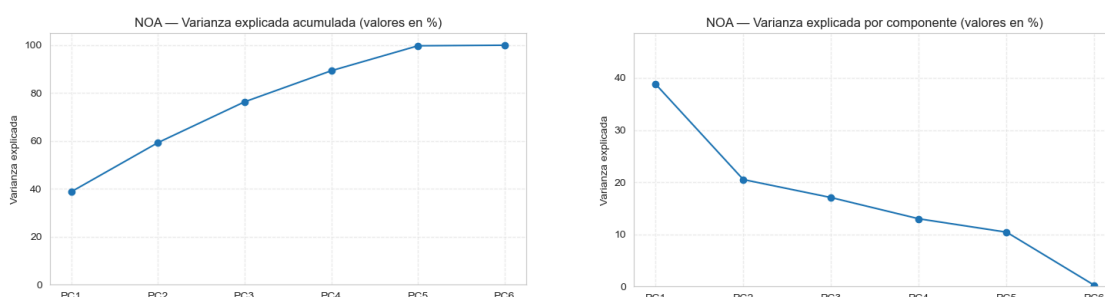


El PC1 se asocia sobre todo con edad y educación, mientras que el PC2 con el tamaño del hogar. El ingreso familiar aparece en el cuadrante superior derecho, en relación positiva con ambas dimensiones, y las horas de trabajo explican menos variabilidad

4)

**Figura 12**

*PCA: varianza explicada (acumulada y por componente)*



El PC1 explica el 38,8% de la varianza y el PC2 un 20,5%, sumando cerca del 60%. Con tres componentes se alcanza el 77% y con cinco se supera el 99%, lo que indica que la mayor parte de la información puede resumirse en los dos o tres primeros sin necesidad de usar los seis.

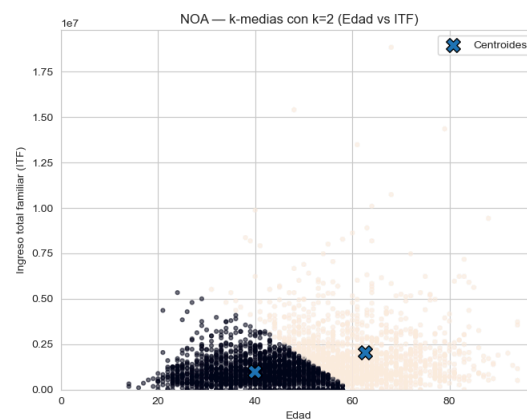
### B. Cluster

5)

a.

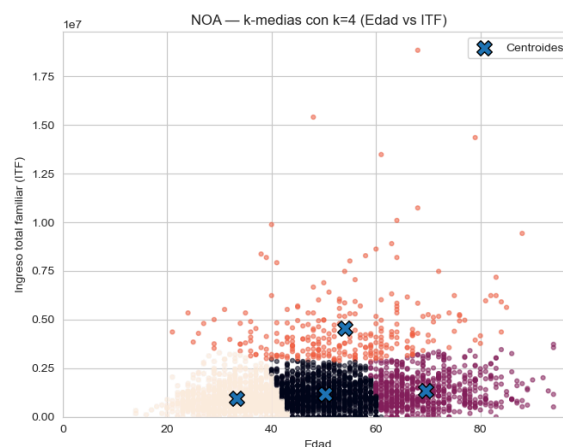
**Figura 13**

*k-medias (k=2): Edad vs ITF (NOA)*



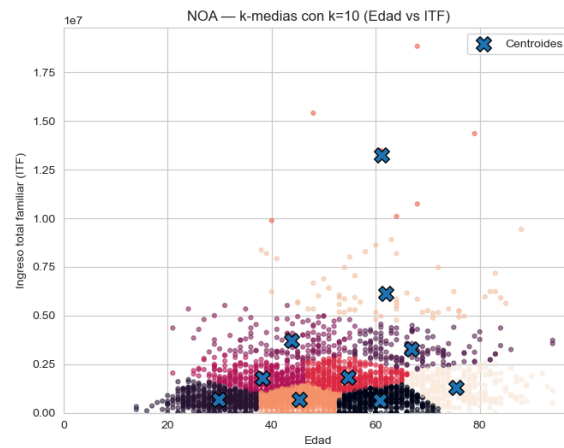
**Figura 14**

*k-medias (k=4): Edad vs ITF (NOA)*

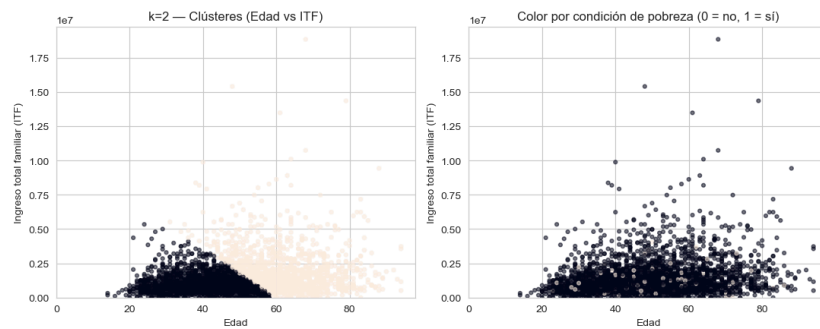


**Figura 15**

*k-medias (k=10): Edad vs ITF (NOA)*

**Figura 16**

*k=2: clústeres vs pobreza (NOA)*

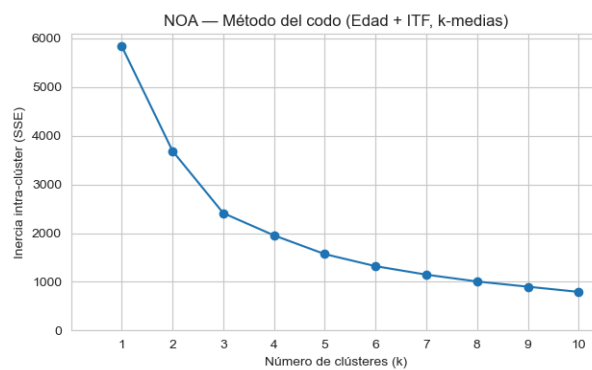


Con  $k=2$  el algoritmo separa de forma general por edad e ingreso, pero no distingue bien la pobreza. Con  $k=4$  los grupos capturan mejor la heterogeneidad por rangos de edad e ingreso. Con  $k=10$  la partición se fragmenta en subgrupos pequeños sin interpretación clara.

**b.**

**Figura 17**

*Método del codo ( $k$ -medias, NOA)*

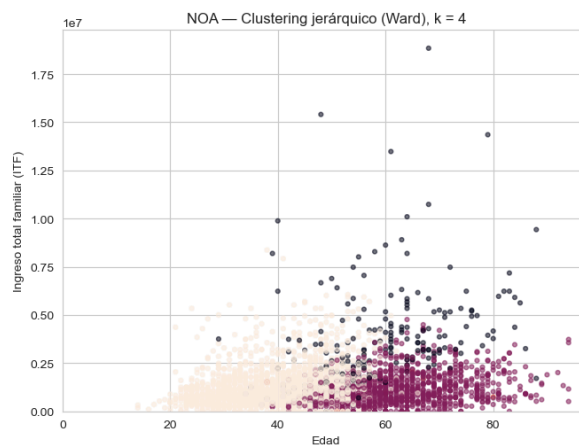


**6)**

**Figura 18**



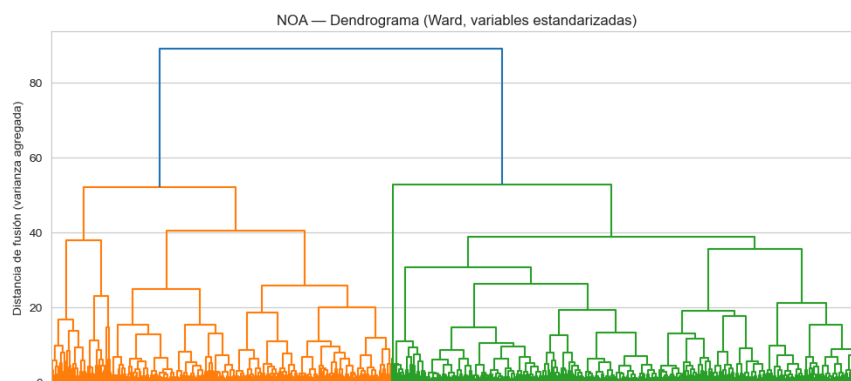
*Jerárquico (Ward)  $k=4$ : Edad vs ITF (NOA)*



El clustering jerárquico con edad e ingreso identificó 4 grupos: uno concentra a la población pobre y los demás reflejan niveles intermedios y altos, mostrando estratos socioeconómicos más allá de la simple división pobre/no pobre.

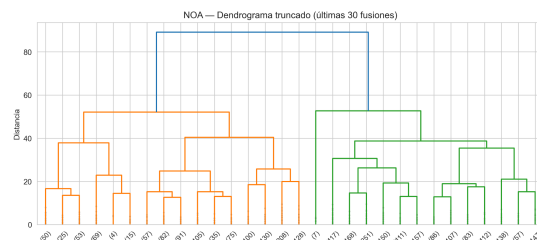
**Figura 19**

*Dendrograma (Ward, NOA)*



**Figura 20**

*Dendrograma truncado (últimas 30 fusiones, NOA)*



Un dendrograma es un gráfico en forma de árbol que muestra cómo se agrupan jerárquicamente las observaciones, reflejando la similitud entre ellas y permitiendo identificar distintos niveles de agrupamiento y el número óptimo de clusters.