1. Investigación de las variables y dominio

El dataset contiene información de individuos del censo estadounidense. La variable objetivo es income, que indica si una persona gana más de 50K o no. Algunas variables clave:

- age, education-num, hours-per-week: numéricas.
- workclass, education, marital-status, occupation, relationship, race, sex, native-country: categóricas.
- capital-gain y capital-loss: variables monetarias, numéricas.

Esto permite explorar cómo estas variables se relacionan con el nivel de ingreso.

---

2. Análisis descriptivo (estadístico)

Observamos

- Estadísticas generales (media, mediana, etc.).
- Distribución de clases.
- Detección de outliers.

Análisis Descriptivo: Resultados

- **Distribución de Clases (income):**
    - <=50K: 76%
    - >50K: 24%
    - El dataset está desbalanceado, lo que puede afectar el modelo
- **Variables Numéricas:**
    - age: distribución normal, sin valores imposibles.
    - capital-gain y capital-loss: presentan **muchos ceros** y valores extremos (outliers fuertes, como 99999).

- hours-per-week: tiene valores mínimos de 1 y máximos de 99. No parecen errores, pero son extremos.

3. Verificación y ajuste de tipos de variables

Ahora chequeamos que las variables tengan el tipo correcto (numérico vs categórico).

Tipos de Variables

- Se corrigieron los tipos de datos para las variables categóricas, que inicialmente estaban como object.

- Ahora workclass, education, sex, income, etc., están correctamente como category, optimizando uso de memoria y permitiendo análisis más adecuado.

4. Datos ausentes y atípicos

No hay valores faltantes tras procesar los " ?" como NaN. Todo está limpio.

Valores Atípicos

- **capital-gain = 99999**: es un outlier claro.
  Para capital-gain y capital-loss: podemos aplicar transformación logarítmica o binarizar (tiene ganancia/sí o no).
  Para outliers extremos: podemos recortar o usar modelos robustos.

5. Correlaciones

Se intentó calcular la correlación entre las variables numéricas y el ingreso (income binarizado), pero hubo un problema técnico con la inclusión de la variable en la matriz de correlación.