# Task 1: Data Cleaning and Preprocessing

## ✅ Objective

Clean and prepare a raw dataset (with nulls, duplicates, inconsistent formats).

## 📂 Dataset Used

- [Medical Appointment No Shows](https://www.kaggle.com/datasets/joniarroba/noshowappointments)

- Original: `KaggleV2-May-2016.csv` (inside `archive.zip`)

- Cleaned: `cleaned\_kaggle\_appointments.csv`

## 🔧 Tools Used

- Python

- Pandas

## 🔍 Data Cleaning Steps Performed

1. \*\*Removed Missing Values\*\*

→ No missing values found.

2. \*\*Removed Duplicates\*\*

→ 0 duplicate rows removed.

3. \*\*Standardized Column Names\*\*

→ Converted to lowercase and replaced spaces/hyphens with underscores.

4. \*\*Fixed Invalid Age Values\*\*

→ Removed rows with negative age (only 1 row affected).

5. \*\*Standardized Text Values\*\*

→ Gender: Ensured only 'M' and 'F' values.

6. \*\*Converted Date Columns\*\*

→ `scheduledday` and `appointmentday` changed to format `dd-mm-yyyy`.

7. \*\*Checked & Fixed Data Types\*\*

→ `age` as integer, dates as `datetime`.

## 📊 Final Result

- Shape: 110526 rows × 14 columns

- No nulls or duplicates

- All columns standardized and date formats unified

## 📝 Files Included

- `archive.zip` – Original dataset from Kaggle

- `cleaned\_kaggle\_appointments.csv` – Final cleaned data

- `data\_cleaning.ipynb` – Python code used

- (Optional) Screenshots if any Excel work done