DATA CLEANING AND PREPROCESSING
(Medical Appointment No Shows)

For this project, I used the **Medical Appointment No-Shows** dataset from Kaggle, which I renamed as task1. Two methods were applied for data cleaning and preprocessing: **Microsoft Excel** and **Python**.

1) Microsoft Excel

To work with the dataset in Excel, I first converted it from CSV to XLSX format. After reviewing the data, I performed several cleaning and preprocessing steps:

Column Names and Data Standardization:

- Column names had mixed casing and spaces. I standardized them by converting to lowercase and replacing spaces with underscores.
    - Example: Patient ID → patient_id, Appointment Date → appointment_date
- The gender column used abbreviations M and F instead of full names.
- scheduled_day and appointment_day columns combined both date and time. The appointment_day time was always 00:00:00.
- The age column contained a value of -1.

Handling Missing Values:

- Applied filters to all columns to identify missing values.

Removing Duplicates:

- Used the **Remove Duplicates** tool from the Data tab to eliminate duplicate rows.

Standardizing Data Values:

- Changed column headings to a consistent format (lowercase, no spaces).
- Used **Find and Replace** to update F → Female and M → Male. (Initially, I mistakenly replaced all occurrences of M in the sheet, which required correction.)

Date and Time Formatting:

- Used the **Text to Columns** tool to separate dates and times.
- Original format: 2016-07-21T08:39:56Z
- Steps:
    1. Split by T → 2016-07-21 (date) and 08:39:56Z (time)
    2. Split by Z → 08:39:56
    3. Reformatted date from YYYY-MM-DD to DD-MM-YYYY
- Deleted appointment_time column since it was always 00:00:00.

Correcting Age Values:

- Changed invalid age -1 to 1.

Mistake I have done so far:

I used find and replace for whole sheet so, every M change into Male. And I clean the error once again.

2) Python

For automated and reproducible cleaning, I used Python with **Jupyter Notebook** in Visual Studio.

Steps:

Load Dataset:

```
import pandas as pd

df = pd.read_csv("task 1 csv.csv")
```

Handle Missing Values:

```
print("Missing values:\n", df.isnull().sum())
```

Remove Duplicates:

```
df.drop_duplicates(inplace=True)
```

Standardize Text Columns:

```
for col in df.select_dtypes(include='object').columns:
    df[col] = df[col].str.strip().str.lower()
```

```
df['gender'] = df['gender'].astype(str).str.strip().str.upper().map({
    'F': 'female',
    'M': 'male'
})
```

Format Date and Time Columns:

```
if 'scheduledday' in df.columns:
    df['scheduledday'] = pd.to_datetime(df['scheduledday'], errors='coerce')
    df['scheduled_date'] = df['scheduledday'].dt.strftime('%d-%m-%Y')
    df['scheduled_time'] = df['scheduledday'].dt.strftime('%H:%M:%S')
    df.drop(columns=['scheduledday'], inplace=True)

if 'appointmentday' in df.columns:
    df['appointmentday'] = pd.to_datetime(df['appointmentday'], errors='coerce')
    df['appointment_date'] = df['appointmentday'].dt.strftime('%d-%m-%Y')
    df['appointment_time'] = df['appointmentday'].dt.strftime('%H:%M:%S')
    df.drop(columns=['appointmentday'], inplace=True)
```

Rename Columns:

```
df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')
df.drop(columns=['appointment_time'], inplace=True)
```

Fix Data Types:

```
if 'age' in df.columns:
    df['age'] = df['age'].astype(int, errors='ignore')
if 'date' in df.columns:
    df['date'] = pd.to_datetime(df['date'], errors='coerce')
```

Export Cleaned Data:

```
df.to_csv("cleaned_data.csv", index=False)
print("\nData cleaning completed! Cleaned file saved as 'cleaned_data.csv'")
```

Summary of Python Cleaning:

- Missing values identified and handled.
- Duplicate rows removed.
- Text columns standardized.
- Gender values mapped correctly to Male and Female.
- Dates and times extracted and formatted properly.
- Unnecessary columns removed.
- Data types corrected.
- Cleaned dataset saved as cleaned_data.csv.