Task 1: Data Cleaning and Preprocessing

Objective: Clean and prepare a raw dataset (with nulls, duplicates, inconsistent formats).
Tools: Excel / Python (Pandas)

Deliverables: Cleaned dataset + short summary of changes

Hints / Mini Guide:
Identify and handle missing values using .isnull() in Python or filters in Excel.
Remove duplicate rows using .drop_duplicates() or Excel's "Remove Duplicates".
Standardize text values like gender, country names, etc.
Convert date formats to a consistent type (e.g., dd-mm-yyyy).
Rename column headers to be clean and uniform (e.g., lowercase, no spaces).
Check and fix data types (e.g., age should be int, date as datetime).

Dataset names from Kaggle suitable for Task 1:

- Customer Personality Analysis
- Medical Appointment No Shows
- Mall Customer Segmentation Data
- Netflix Movies and TV Shows
- Sales Data

By completing this task, you will:

- Gain hands-on experience in identifying and fixing common data issues like missing values, duplicates, and inconsistent formatting.
- Learn to use Excel functions or Pandas in Python for real-world data cleaning.
- Improve your understanding of data pre-processing, which is a critical step before data analysis or visualization.
- Build confidence in handling raw datasets independently.
- Create a clean, structured dataset that is ready for analysis or modelling.

Interview Questions Related To Above Task:

1. What are missing values and how do you handle them?
2. How do you treat duplicate records?
3. Difference between dropna() and fillna() in Pandas?
4. What is outlier treatment and why is it important?
5. Explain the process of standardizing data.
6. How do you handle inconsistent data formats (e.g., date/time)?
7. What are common data cleaning challenges?
8. How can you check data quality?

📌 Task Submission Guidelines

- ⏰ **Time Window:**

You can complete the task anytime between 10:00 AM to 10:00 PM on the given day. Submission link closes at 10 :00 PM

- 🔍 **Self-Research Allowed:**

You are free to explore, Google, or refer to tutorials to understand concepts and complete the task effectively.

- 🛠️ **Debug Yourself:**

Try to resolve all errors by yourself. This helps you learn problem-solving and ensures you don't face the same issues in future tasks.

- 💸 **No Paid Tools:**

If the task involves any paid software/tools, do not purchase anything. Just learn the process or find free alternatives.

- 📁 **GitHub Submission:**

Create a new GitHub repository for each task.

Add everything you used for the task — code, datasets, screenshots (if any), and a short README.md explaining what you did.

📤 **Submit Here:**

After completing the task, paste your GitHub repo link and submit it using the link below:

- 👉 [Submission Link]