✅ Task 1: Excel Basics – Data Cleaning & Formatting

Tools:

- Primary: Microsoft Excel / Google Sheets
- Alternatives: LibreOffice Calc, WPS Spreadsheet

Dataset:

- Netflix Movies and TV Shows" dataset
- "Global Superstore" dataset
- Any dataset with 5k–10k rows and messy columns (missing values, duplicates)

Hints / Mini Guide:

1. Download the Kaggle dataset in CSV format and open it in Excel or Google Sheets, ensuring the first row is recognized correctly as column headers and the delimiter is properly separated.
2. Use Freeze Panes to lock header rows and apply filters on all columns so you can explore the dataset efficiently like a real analyst.
3. Identify missing values using Filter → Blank values and highlight them using Conditional Formatting, then decide whether to replace, remove, or leave blanks based on column context.
4. Detect duplicates using Remove Duplicates (based on key columns like ID/Title), but before deleting, create a backup tab to avoid irreversible data loss.
5. Standardize text fields using functions like TRIM, PROPER, UPPER, and remove extra spaces and inconsistent naming patterns.
6. Convert incorrect formats by validating date columns (YYYY-MM-DD), numeric columns (no symbols inside), and categorical values (no spelling variations).
7. Create a "Cleaned_Data" sheet and copy only the cleaned output to maintain separation between raw and processed data like professional pipelines.
8. Add a new column called Data_Quality_Notes and write short notes such as "missing director names present" or "rating values inconsistent" to show analyst thinking.
9. Save the final dataset as Cleaned_dataset.xlsx and export it as cleaned_dataset.csv for further analysis tasks.

Deliverables:

- Raw_Data.xlsx
- Cleaned_dataset.xlsx
- cleaned_dataset.csv

Final Outcome:

Intern learns how to clean and standardize real-world messy data using Excel/Sheets and produces a clean structured dataset ready for analysis.

Interview Questions Related To Above Task:

- What is the difference between deleting missing data vs imputing missing data?
- What risks occur if you remove duplicates incorrectly?
- How do TRIM and CLEAN functions help in data cleaning?
- How do you validate correct data types in a spreadsheet?
- Why should raw data and cleaned data always be kept separately?

📌 **Task Submission Guidelines**

- ⏰ **Time Window:**

Youcan 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:**

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

- 🛠️ **Debug Yourself:**

Tryto 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:**

Ifthe 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]