✅ Task 5: Python Basics – Reading Data + Simple Cleaning (Pandas)

Tools:

- Primary: Google Colab (free)
- Alternatives: Jupyter Notebook, Kaggle Notebooks
- Libraries: pandas, numpy

Dataset:

- "Titanic Dataset"
- "House Prices Dataset"
- "Students Performance Dataset"

Hints / Mini Guide:

1. Open Google Colab and upload dataset or load directly through Kaggle notebook to ensure environment is ready and reproducible.
2. Import pandas and read the dataset using pd.read_csv(), then print .head() and .info() to understand structure and missing values.
3. Identify missing values using .isnull().sum() and analyze which columns require cleaning before modeling or reporting.
4. Clean missing values using proper approaches: drop if irrelevant, fill using mean/median if numeric, mode if categorical.
5. Remove duplicates using .drop_duplicates() and verify before/after row count to confirm duplicates were removed properly.
6. Convert datatypes using .astype() (example: date string into datetime) to allow correct calculations.
7. Create new columns using logic, such as profit margin, age category, or price band, demonstrating transformation ability.
8. Save cleaned dataset using .to_csv() and confirm by downloading output file.
9. Write short markdown notes in the notebook explaining what cleaning steps were applied and why.

Deliverables:

- Task5_Cleaning.ipynb
- cleaned_data.csv
- 5–10 markdown notes inside notebook

Final Outcome:

✅ Intern gains hands-on Pandas cleaning skills and learns how Python replaces manual Excel cleaning in large datasets.

Interview Questions Related To Above Task:

- How do you handle missing values in pandas?
- Difference between fillna() and dropna()?
- How do you detect duplicates?
- Why datatype conversion is important in analytics?
- What are the benefits of Python over Excel in cleaning?

📌 **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]