Task 2: Data Cleaning & Missing Value Handling

Tools:

- Python (Pandas, NumPy)
- Alternatives: R (tidyverse)

Dataset:

- “House Prices Dataset”
- “Medical Appointment No Shows”

Hints / Mini Guide:

1. Load dataset and identify missing values using .isnull().sum().
2. Visualize missing data patterns using simple bar charts.
3. Apply mean/median imputation for numerical columns.
4. Apply mode imputation for categorical columns.
5. Remove columns with extremely high missing values.
6. Validate dataset after cleaning.
7. Compare before vs after dataset size and quality.

Deliverables:

- Cleaned dataset file
- Notebook with cleaning steps

Final Outcome:

Intern gains hands-on data preprocessing skills.

Interview Questions Related To Above Task:

- Mean vs median imputation?
- When should rows be dropped?
- Why missing data is harmful?
- What is data leakage?
- What is data quality?

📌 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]

⭐⭐⭐⭐⭐