Intermediate Assessment 1

🔹 Part 1: Data Loading

The **California Housing Dataset** is included inside **scikit-learn**. You can **load it directly** without even needing a .csv file.

Here's how you can load it in Python:
python

```
from sklearn.datasets import fetch_california_housing
import pandas as pd

housing = fetch_california_housing(as_frame=True)
df = housing.frame

# Now df is your full California Housing dataset!
print(df.head())
```

1. Load the california_housing.csv file using pandas.
2. Display the first 5 rows.
3. Print the column names and their data types.
4. Check for missing values in the dataset.
5. Get basic statistical summaries using .describe().

🔹 Part 2: Data Cleaning (30-40 min)

6. Check for and remove any duplicated rows.
7. Handle missing values if any:
    - Fill missing numerical features with the **mean** value.
8. Create a new column PricePerRoom = median_house_value / total_rooms.
9. Create a column HighPopulationArea:
    - 1 if population > 500, else 0.
10. Bin the median_income into 5 equal-sized bins and label them as Very Low, Low, Medium, High, Very High. (Hint: pd.cut)
11. Drop columns that seem redundant after feature creation (if any).

🔹 Part 3: Data Visualization (40-50 min)

12. Plot the distribution of median_house_value with a histogram.
13. Create a scatter plot of longitude vs latitude, colored by median_house_value.

14. Plot a boxplot of median_house_value grouped by the new income categories.
15. Plot the correlation matrix heatmap between numerical features.
16. Create a bar plot showing average median_house_value for high population vs low population areas.
    Create a pairplot (sns.pairplot) for selected features: median_income, housing_median_age, median_house_value.

✅ Deliverables

- Jupyter Notebook (.ipynb)
- Cleaned dataset file
- At least 5 clean, labeled plots
- Add all submission files to a GitHub repository and share the link of the same via Paatshala

---