Milestone 2 (Data Acquisition & Ingestion Pipeline)

1. Objective

The objective of Milestone 2 is to design and implement a **robust data ingestion and preprocessing pipeline** that can reliably load, validate, clean, and transform the loan dataset. The pipeline ensures that the data is consistent, high-quality, and ready for downstream machine learning modelling (Milestone 3).

2. Purpose

The purpose of this milestone is to:

- Establish a standardized workflow for importing and preparing loan data
- Improve the overall data quality by addressing missing values, duplicates, and invalid records
- Transform categorical attributes into machine-readable formats
- Generate a clean, structured dataset for training ML models
- Measure data ingestion and data quality KPIs

This milestone ensures the dataset is **accurate**, **efficiently processed**, and **ML ready**.

3. Work Performed

3.1 Data Loading

- Loaded the raw dataset using pandas
- Measured ingestion time as a KPI
- Validated dataset shape: **249,999 rows × 18 columns**

3.2 Data Cleaning

Actions performed:

- Removed duplicate records
- Filled missing numerical values using **median**
- Filled missing categorical values using **mode**

- Removed invalid entries:
    - Age < 18
    - Income ≤ 0
    - LoanAmount ≤ 0

Result:
Dataset remained consistent with **249,999 valid rows**.

3.3 Data Transformation

- Performed safe One-Hot Encoding on selected categorical fields:
    - Education
    - EmploymentType
    - MaritalStatus
    - LoanPurpose
- Ensured no high-cardinality columns were one-hot encoded to avoid memory issues
- Final dataset shape after encoding: **249,999 rows × 26 columns**

3.4 Data Quality KPIs Generated

- Total Rows After Cleaning: **249,999**
- Missing Values After Cleaning: **0 across all columns**
- Memory Usage: **~29.56 MB**
- Duplicate Rows Removed: **0**
- Data Ingestion Time: **~0.46 seconds**

3.5 Saved Cleaned Dataset

The final processed dataset is saved as:

data/loan_data_clean.csv

4. Key Outputs

Files Generated

- scripts/data_ingestion.py — ingestion + cleaning + transformation pipeline
- data/loan_data_clean.csv — cleaned ML-ready dataset

KPIs Produced

- Data ingestion speed
- Missing value summary
- Duplicate count
- Final dataset shape
- Memory usage

5. Conclusion

Milestone 2 successfully establishes a **scalable, efficient, and reliable data ingestion pipeline**. The dataset is now:

- Fully cleaned
- Validated
- Transformed
- Encoded
- Free from missing values
- Memory-optimized
- Ready for Machine Learning model training in **Milestone 3**

This milestone ensures a strong foundation for model development by delivering a high-quality dataset and a reusable ingestion workflow.