Data Pipeline

Three-stage workflow for data integrity and reproducibility.

Overview

1. **Input** - Raw data storage
2. **Cleaned** - 3NF data storage
3. **Output** - Analysis results

Input Folder

- Raw, unprocessed data
- Original sources (surveys, indicators, market data)
- Single source of truth
- Preserve original state (may have inconsistencies)

Cleaned Folder

- Data in Third Normal Form (3NF)
- Intermediary between raw and analysis
- Standardized, structured
- Consistent access across analyses

Output Folder

- Organized by analysis type (e.g., simulate)
- Final results, models, visualizations
- Formats: .pkl (Python), .rds (R)
- Reproducibility by isolating results

Third Normal Form (3NF)

1. **1NF**: Primary key, atomic values, no repeating groups
2. **2NF**: 1NF + all non-key attributes fully dependent on primary key
3. **3NF**: 2NF + no transitive dependencies (non-key depends only on primary key)

Benefits of 3NF

- Eliminates redundancy
- Reduces inconsistencies
- Improves data integrity
- Easier updates and maintenance
- Reliable foundation for analysis