IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Data Preparation

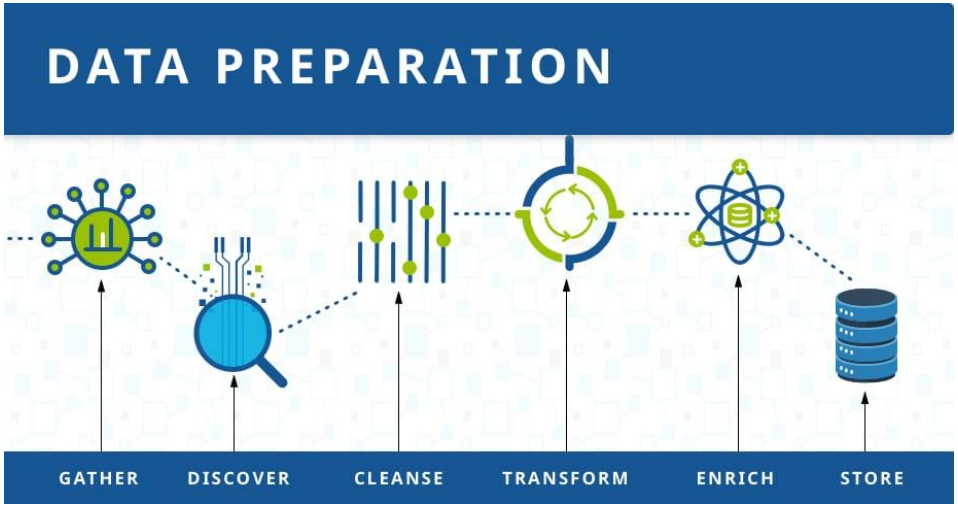
Data Science Toolkit & Applications
Lina Fahed & Laurent Brisson

Data Preparation: Transforming Raw Data into Gold

The Art of **Cleaning**, **Transforming**, and **Structuring** Data for Modeling

The Hidden Work of Data Science

- 80% of a data scientist's time is spent on data preparation (source: Forbes, 2023).
- Example: A dataset with missing values, inconsistent formats, or outliers will lead to poor model performance.
- "Data preparation is like cooking: the **quality** of the ingredients determines the **quality** of the dish."

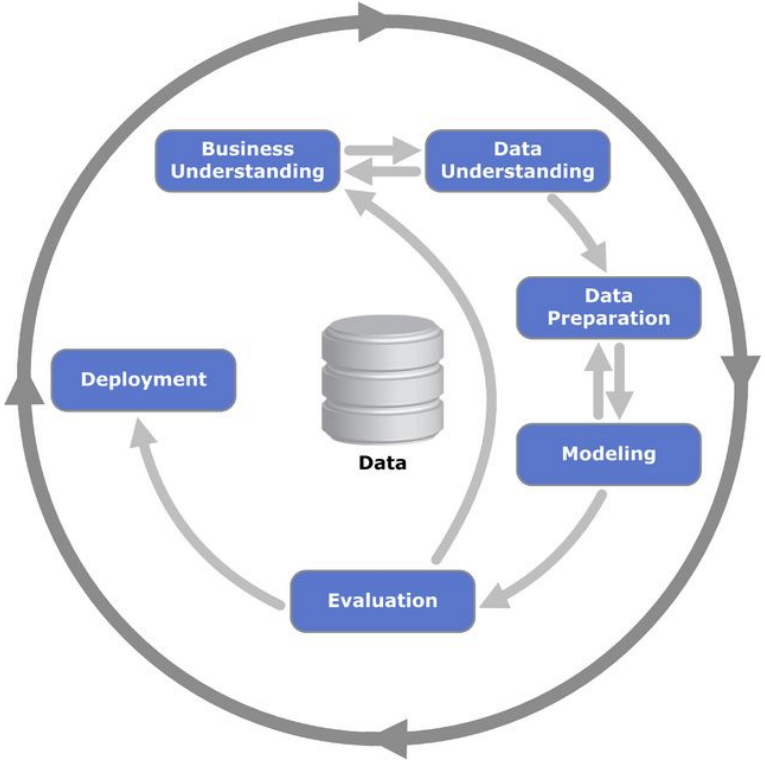


Where Does Data Preparation Fit?

CRISP-DM Steps:

1. Business Understanding ✅
2. Data Understanding ✅
3. **Data Preparation** (Current focus)
4. Modeling
5. Evaluation
6. Deployment

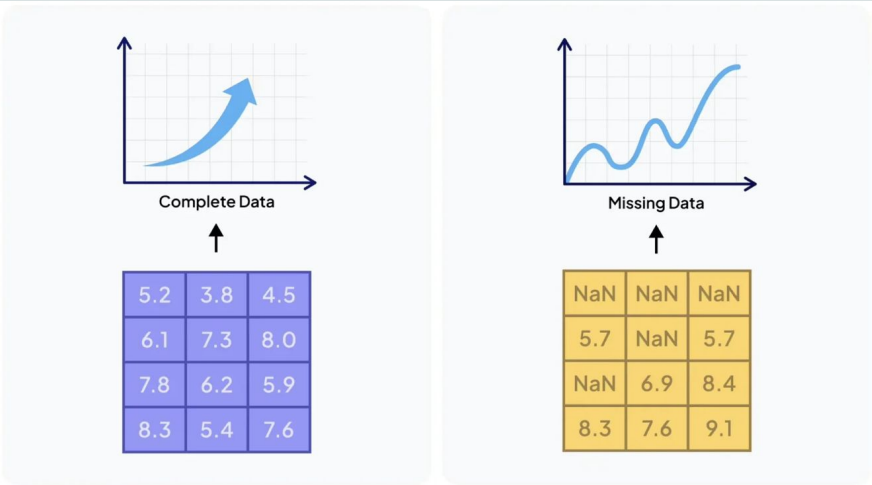
Data preparation **Goal**: Clean, transform, and structure data to make it model-ready.



What Does Data Preparation Involve?

1. **Data Cleaning**: Handle missing values, remove duplicates, correct errors.
2. **Data Transformation**: Normalize, scale, encode categorical variables.
3. **Data Integration**: Merge datasets, join tables.
4. **Data Reduction**: Reduce dimensionality (e.g., PCA), select features.

Data Cleaning – Handling Missing Data



Why Missing Values Occur:

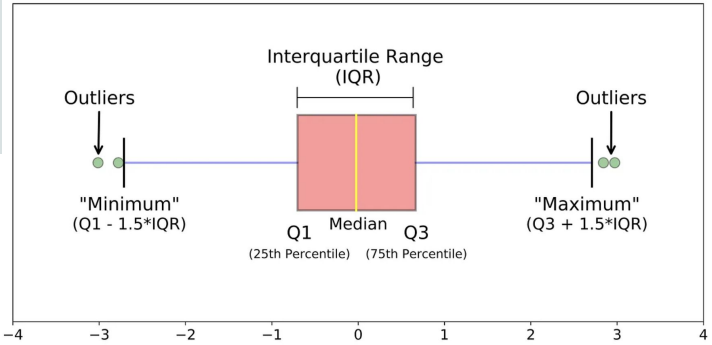
- Data not collected.
- Human error.
- Sensor failures.

Strategies:

- **Deletion**: Remove rows/columns with missing values (if <5% of data).
- **Imputation**: Fill with mean/median (numerical) or mode (categorical).
- **Advanced** Methods: Use algorithms (e.g., KNN imputation).

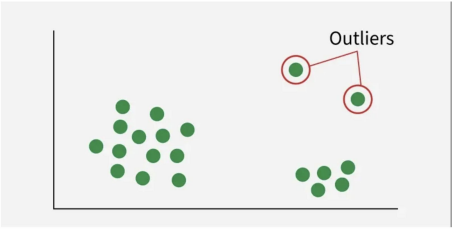
Example: If 10% of 'Age' values are missing ⇒ Impute with median age.

Data Cleaning – Dealing with Outliers



What is an Outlier?

- A data point **significantly different** from others
- e.g., salary of $1M in a dataset of $50K

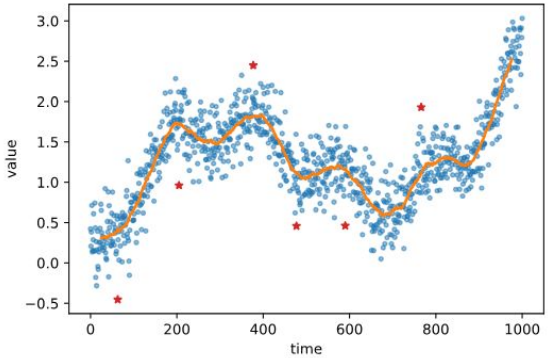


Detection Methods:

- Visual: Box plots, scatter plots.
- Statistical: Z-score, IQR (Interquartile Range).

Strategies:

- Remove: If it's an error (e.g., data entry mistake).
- Cap: Replace with a threshold (e.g., 95th percentile).
- Transform: Use log transformation to reduce skew.



Data Transformation – Scaling and Normalizing Data

Why Scale Data?
- Algorithms like SVM, KNN, or neural networks **require comparable scales** ⇒ **distance** based algorithms.

Common **Techniques**:

- Normalizing / Min-Max Scaling:
  - rescale to [0, 1].
  - formula: (x - min) / (max - min)
- Standardization (Z-score):
  - rescale to mean=0, std=1.
  - formula: (x - mean) / std
- Log Transformation:
  - for highly skewed data (e.g., income).

Example:
- Age (20-80) vs. Income ($20K-$200K)
- → Standardize both.

Data Preparation: Transforming Raw Data into Gold

Why Encode?
- Most ML algorithms require **numerical inputs**.

Techniques:

One-Hot Encoding:

- Create binary columns for each category.
- Example: Color (Red, Blue) → Color_Red, Color_Blue.

Label Encoding:

- Assign a unique integer to each category.
- Example: Red=0, Blue=1.

Ordinal Encoding:

- For ordered categories (e.g., Low=0, Medium=1, High=2).

Pitfall: Avoid the dummy variable trap ⇒ perfect multicollinearity.

- The dummy variable trap occurs because the dummy variables become perfectly predictable from one another, creating perfect multicollinearity

Example

Suppose the categorical variable **Color** has three categories:

Color	Red	Blue	Green
Red	1	0	0
Blue	0	1	0
Green	0	0	1

Notice that for **every observation**:

$$\text{Red} + \text{Blue} + \text{Green} = 1$$

This means one dummy variable can always be computed from the other two:

$$\text{Green} = 1 - \text{Red} - \text{Blue}$$

The three variables are therefore **perfectly linearly dependent**.

Data Integration – Merging and Joining Data

Why Integrate?

- Data is often spread across multiple tables/files.

Methods:

- **Concatenation**: Stack datasets vertically (same columns).
- **Merge** (Join): Combine datasets horizontally using a key (e.g., customer_id).
- Types: Inner, Outer, Left, Right.

Example: orders.csv + customers.csv → Merge on customer_id.

Data Reduction – Dimensionality & Feature Selection
Simplifying Without Losing Information

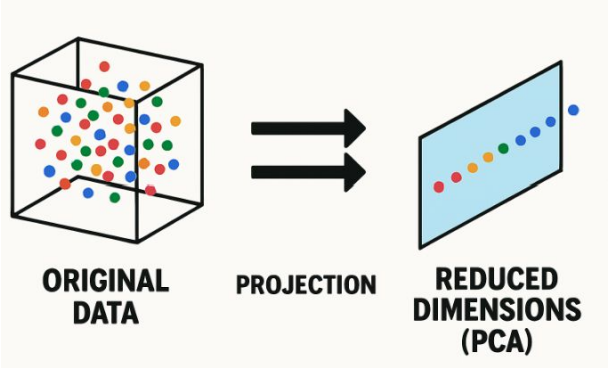
Why Reduce Data?

- Reduce **noise**, improve model **performance**, and **speed up** training.

Techniques:

- **Feature Selection**: Remove irrelevant features (e.g., using correlation analysis).
- **Dimensionality Reduction**: PCA (Principal Component Analysis).

Example: 100 features → PCA reduces to 10 principal components.



Essential Tools and Libraries

Python:

- Pandas (data manipulation)
- NumPy (numerical operations)
- Scikit-learn (scaling, encoding, PCA)

What to Remember

- **Data Preparation** is 80% of the Work – But it’s what makes or breaks your model.
- **Clean First, Model Later** – Always address missing values, outliers, and inconsistencies.
- **Automate Where Possible** – Use libraries like Pandas and Scikit-learn to save time.