

Introduction to Data Wrangling

Introduction to Data Wrangling

The essential first step in turning raw data into actionable insights.

1. What Is Data Wrangling?

Term	Meaning
Data Wrangling (also called data munging, data cleaning, or data preprocessing)	The process of transforming raw, messy, or unstructured data into a tidy, analysis-ready format.
Goal	Reduce errors, handle inconsistencies, and create a reproducible dataset that can be explored, modeled, and visualized with confidence.

2. Why It Matters

- **Improves model performance** – Clean data leads to more accurate predictions.
 - **Saves time** – Early detection of problems prevents costly rework later.
 - **Ensures reproducibility** – Documented steps make analyses transparent and repeatable.
 - **Facilitates collaboration** – Well-structured data is easier for teammates to understand and use.
-

3. Typical Workflow

1. **Acquire Data** – Load files, query databases, or pull from APIs.
 2. **Inspect & Profile** – Get a quick sense of size, types, missingness, and distribution.
 3. **Clean** – Handle missing values, remove duplicates, correct typos, standardize formats.
 4. **Transform** – Reshape, aggregate, create derived variables, encode categories.
 5. **Validate** – Verify that transformations preserve meaning and meet business rules.
 6. **Export** – Save the tidy dataset in a format suited for downstream analysis (CSV, Parquet, database table, etc.).
-

Introduction to Data Wrangling

4. Core Tasks & Common Techniques

Task	Typical Operations	Tools / Functions (Python)
Missing Data	- Identify (<code>isnull</code>) - Impute (mean, median, mode, K?NN) - Flag or drop rows/columns	<code>pandas.isna()</code> , <code>SimpleImputer</code> , <code>fillna()</code>
Duplicates	- Detect (<code>duplicated</code>) - Remove or consolidate	<code>drop_duplicates()</code>
Data Types	- Cast (<code>astype</code>) - Parse dates (<code>to_datetime</code>) - Convert categorical strings to category	<code>astype()</code> , <code>pd.to_datetime()</code>
String Cleaning	- Trim whitespace, lower?case - Regex replace, extract patterns	<code>str.strip()</code> , <code>str.replace()</code> , <code>str.extract()</code>
Outliers	- Visual inspection (<code>boxplot</code>) - Statistical rules (IQR, Z?score) - Cap or remove	<code>np.percentile()</code> , <code>scipy.stats.zscore</code>
Reshaping	- Pivot / unpivot - Melt, stack, wide ? long - Create ratios, flags, rolling windows	<code>pivot()</code> , <code>melt()</code> , <code>stack()</code>
Feature Engineering	- Encode categorical variables (<code>one?hot</code>, <code>label</code>)	<code>pd.get_dummies()</code> , <code>OrdinalEncoder</code>
Joining Data	- Merge tables on keys, handle many?to?many - Concatenate datasets	<code>merge()</code> , <code>concat()</code>

5. Practical Example (Python)

```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer
```

Introduction to Data Wrangling

```
# 1?? Load raw data
df = pd.read_csv('sales_raw.csv')

# 2?? Profile
print(df.info())
print(df.describe(include='all'))

# 3?? Clean
```