Data Cleaning in Python

* Data cleaning:

(1) Data collection:

Multiple sources like

(1) web scraping
(2) Databases
(3) API
(4) Surveys

} Data Storage

(2) Raw Data :

- Issues in Raw data are below:

(1) Missing values
(2) Duplicate values
(3) Spelling errors
(4) Outliers
(5) Incorrect datatypes
(6) Inconsistent formatting

③ Data cleaning Techniques:

① Summary statistics
② Distribution plots
③ info (data types/Null count
④ frequency → categorical data
⑤ Visual inspect

④ Handling Missing data:

① Delete rows / columns:
→ More than 50 % of data is missing.

② Imputation: Filling Missing values
① For Numerical columns:
→ Mean / Median / Mode

② For categorical values:
→ Mode / unknown

③ Datetime columns (Time Series)
→ forward fill / backward fill

④ Predictive Imputation using ML Models.
(for Most Accurate information using ML model to fill the data)

⑤ Inconsistent formatting:

→ Replace / lowercase / uppercase
→ Extra spaces Removal.

⑥ Handling duplicate Rows:

① Remove duplicate Records.

② Check for primary key.

③ Remove near duplicate Records.

⑦ Correcting Datatypes:

① Numeric → int / float

② Categorical → object

③ Datetime → datetime

⑧ Handling outliers:

① Genuine outliers

② Errors

↳ Remove outliers using Z-score, IQR

→ Domain knowledge

→ Transforming Data