Data Cleaning (Missing Values, Duplicates, Types)

🔹 Reminder

- Yesterday: we learned how to **import & export data** (CSV, Excel, JSON).
- Today: imported datasets are often **messy** → we must **clean before analysis**.

🔹 Why Cleaning Matters

Notes

- **Garbage in = garbage out** → ML models and visualizations are only as good as the data.
- Real-world issues:
  - Missing values → `NaN` (Not a Number).
  - Duplicates → repeated rows.
  - Wrong data types → e.g., numbers stored as text.

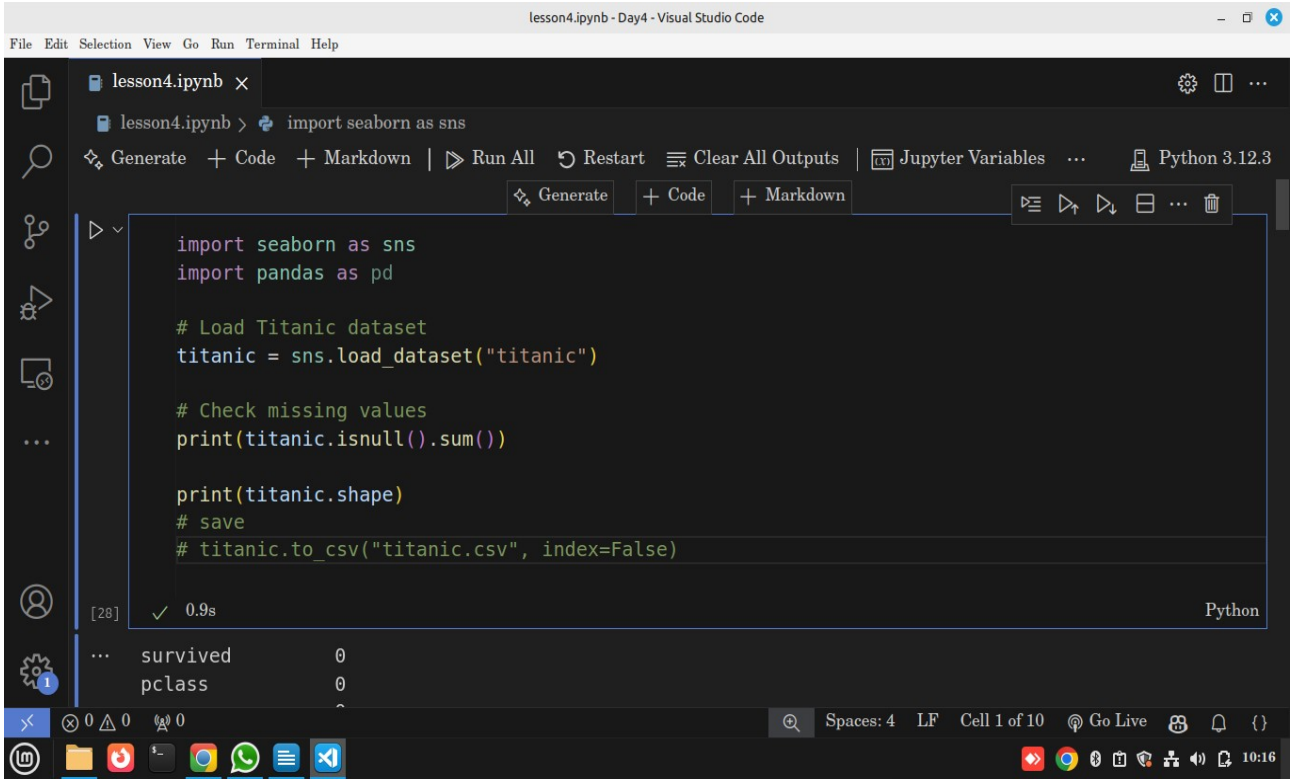
Use in Data Science

- Data cleaning can take **50–80% of project time**.
- A well-cleaned dataset improves **accuracy, reliability, and reproducibility**.

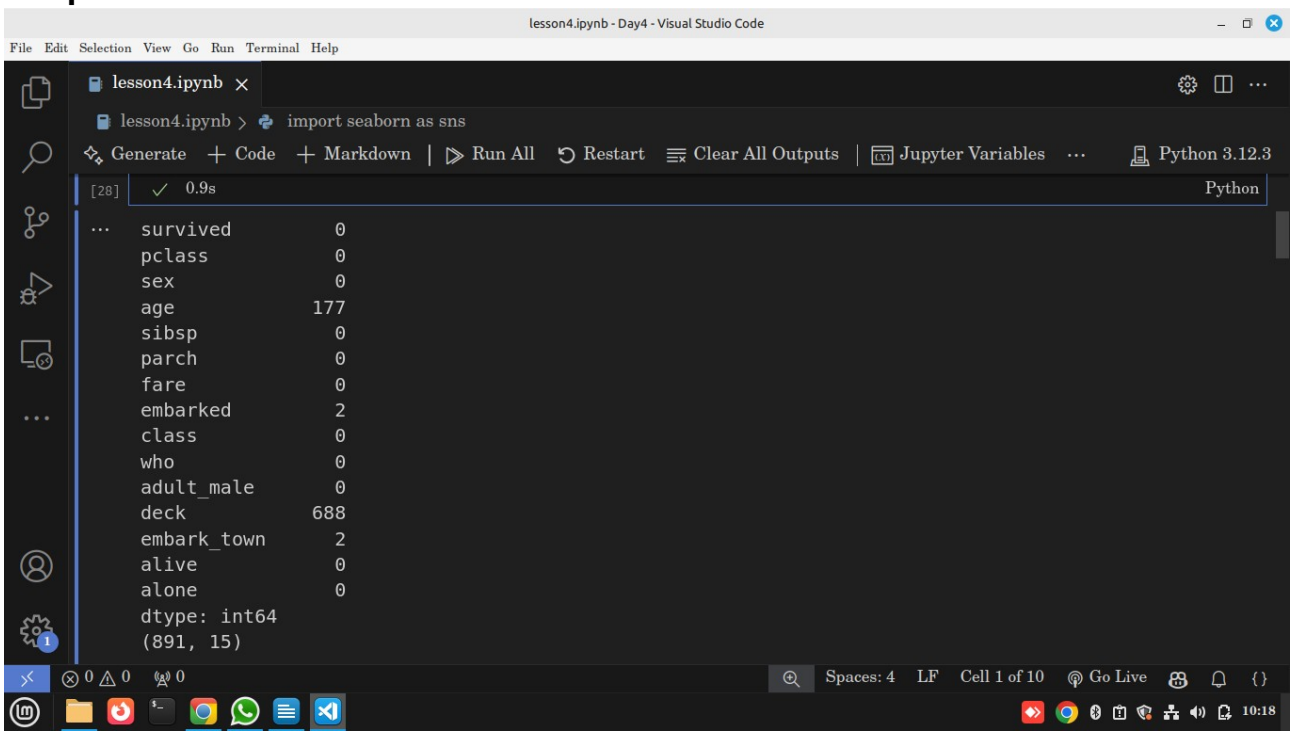
🔹 Handling Missing Values

Practical

Check for missing values



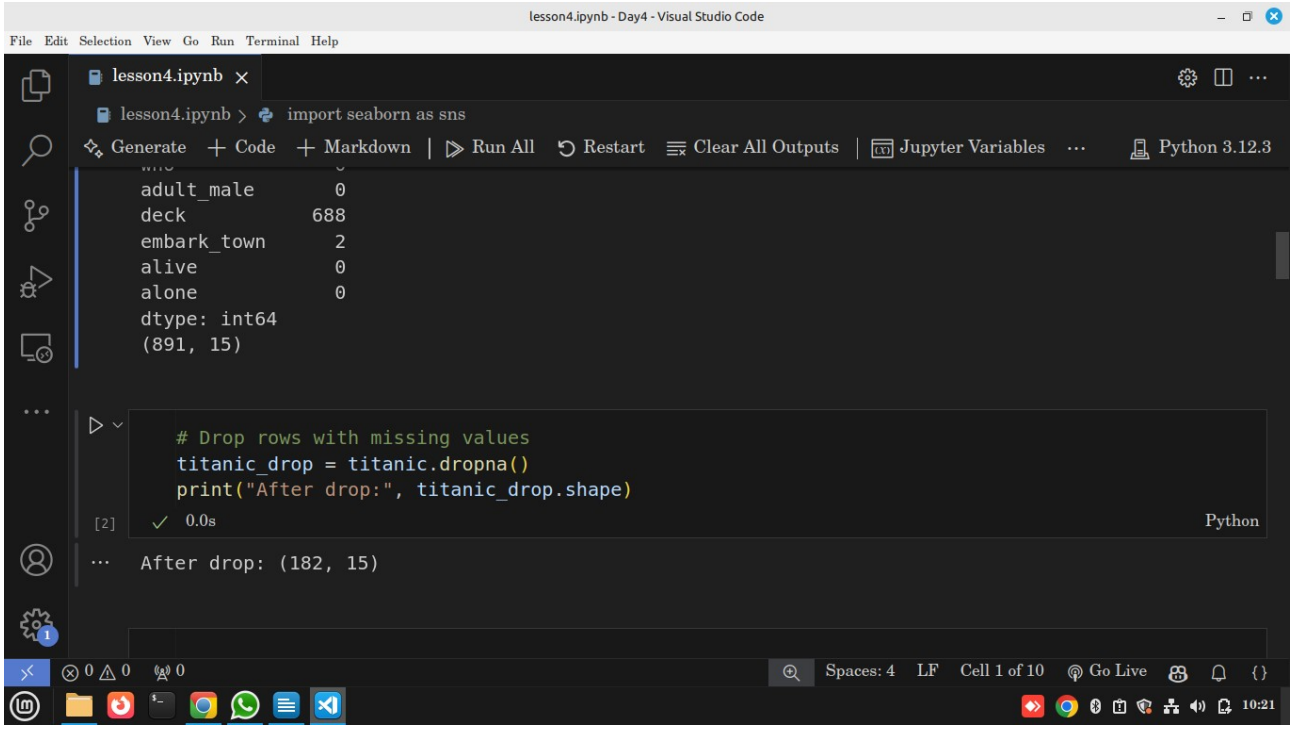
Output



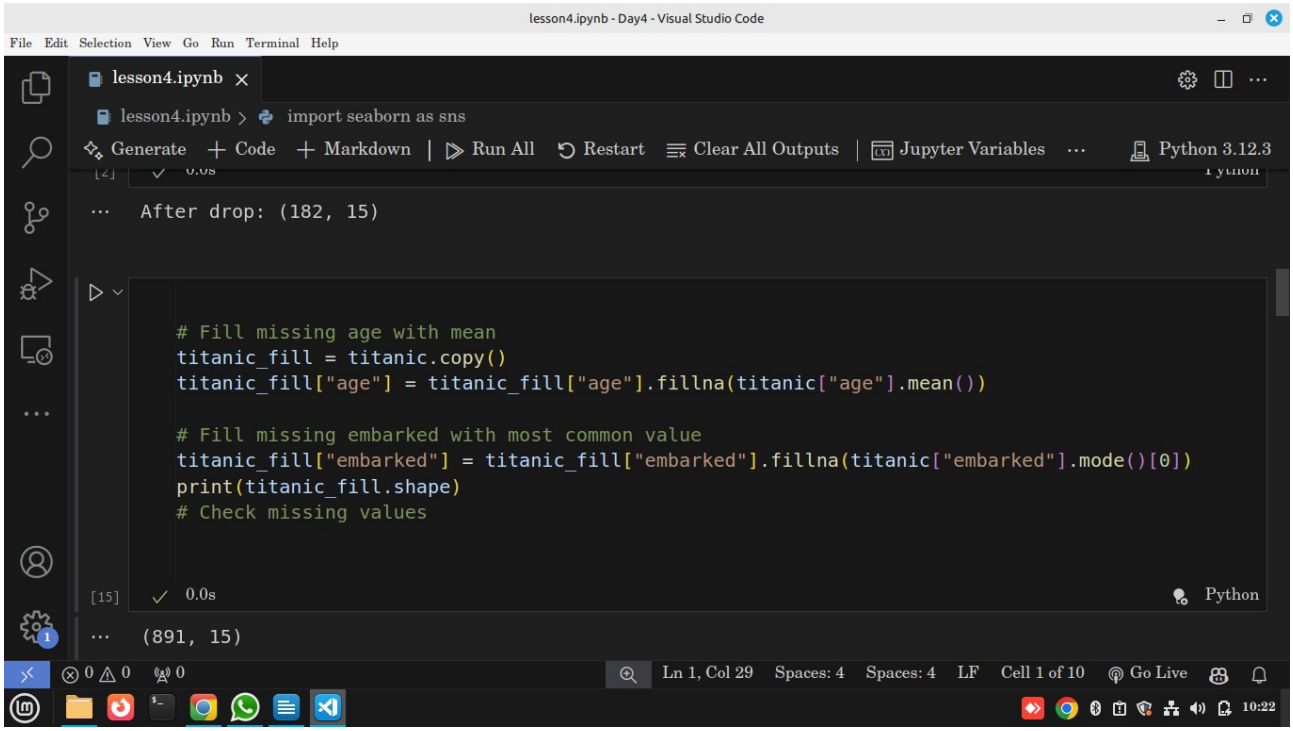
- We have some missing values in age ,deck, and embark_town

- we have a total of 891 rows and 15 columns

Option 1: Drop the empty rows



Option2 : Fill the missing values with Mean or Most common used value (mode)



Explanation

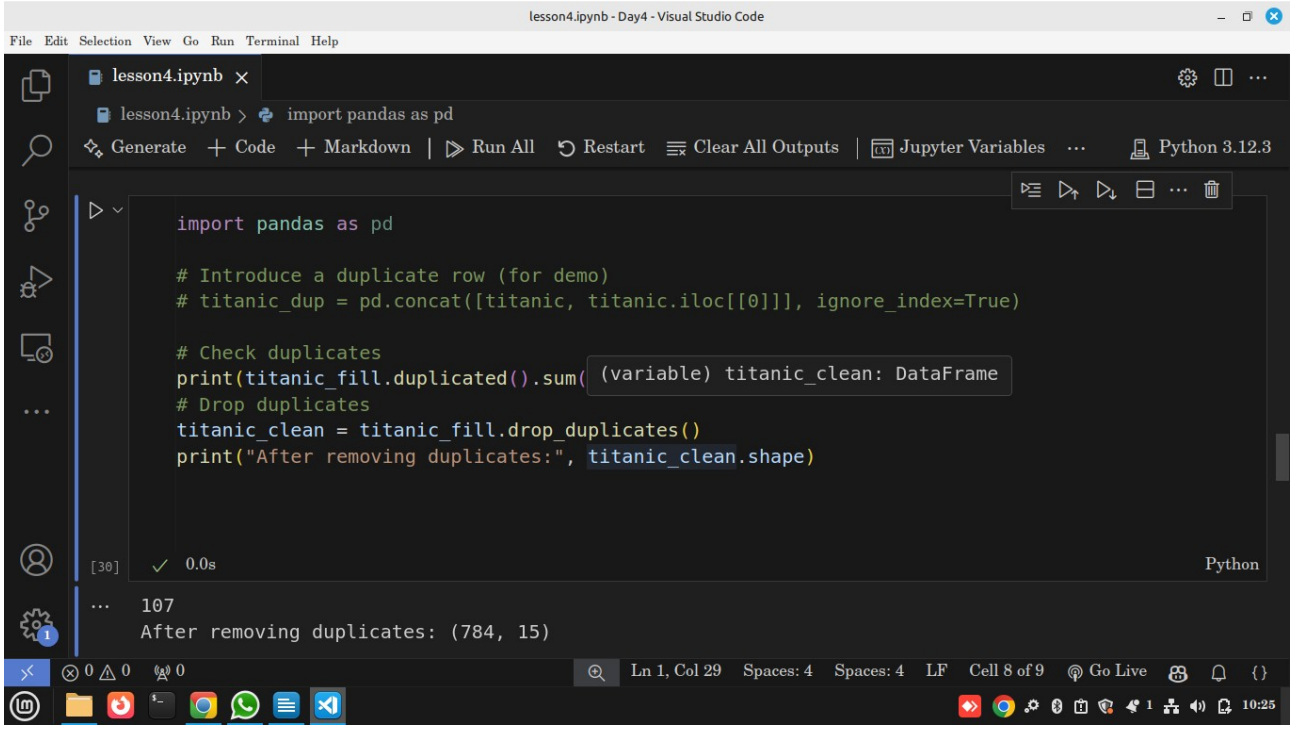
- `.isnull().sum()` → counts missing values per column.
- `.dropna()` → removes rows with `NaN`.
- `.fillna()` → replaces missing values.
- `.mode()[0]` → fills with the most frequent category.

Use in Data Science

- Dropping vs filling depends on dataset size.
- Filling prevents data loss and makes models work.

🔹 Handling Duplicates

Practical



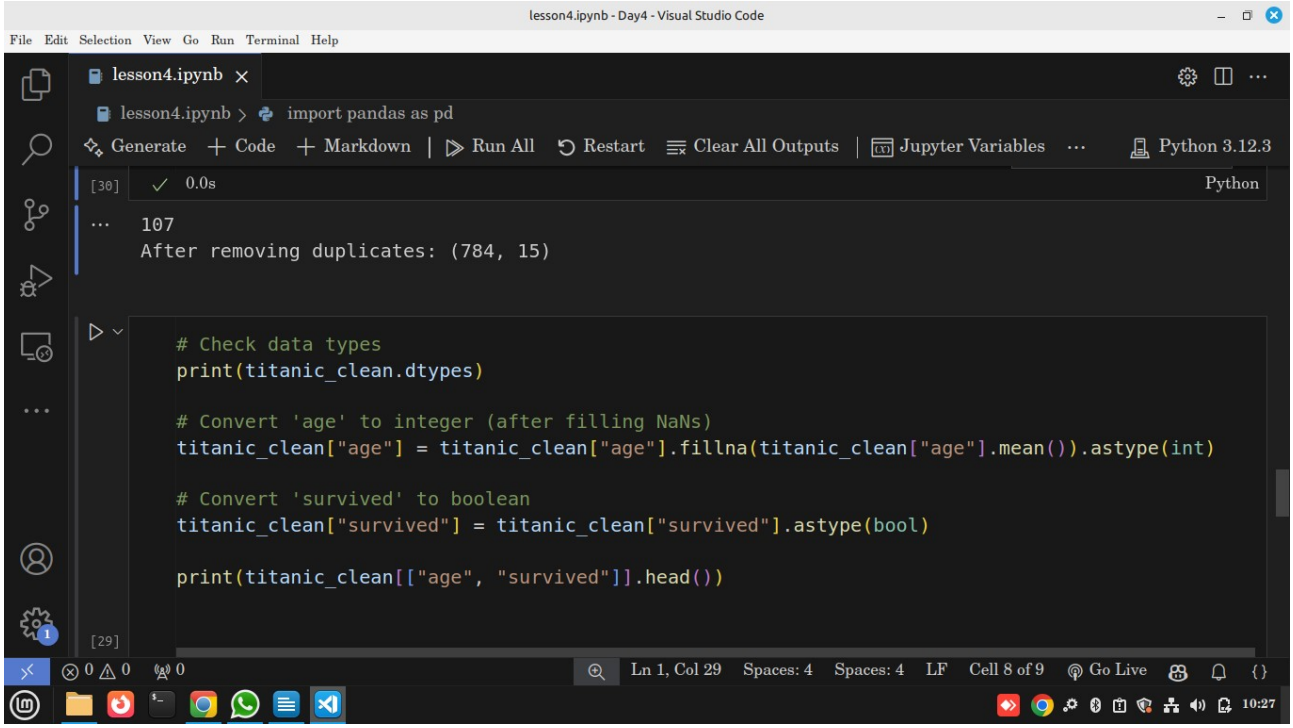
- `.duplicated()` → marks duplicates as True/False.
- `.drop_duplicates()` → removes them.

Use in Data Science

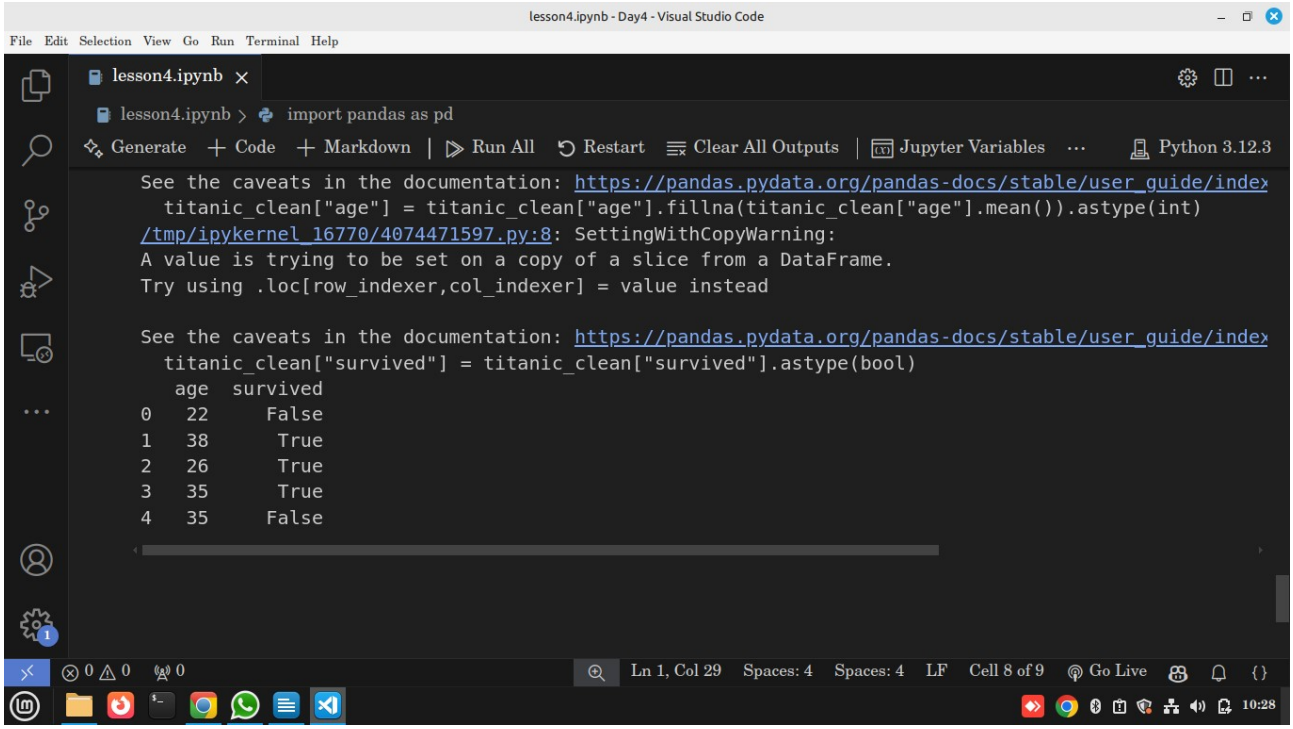
- Duplicates can bias results (e.g., same passenger counted twice).

🔹 Fixing Data Types

Practical



Output



Explanation

- `.dtypes` → shows column data types.
- `.astype(int)` → ensures numerical consistency.
- `.astype(bool)` → makes survival easier to understand (True/False).

Use in Data Science

- Correct data types = fewer bugs + more efficient memory usage.

🔹 Mini Challenge

- On Titanic dataset:
  1. Fill missing `age` with median instead of mean.
  2. Drop duplicate rows.
  3. Convert `class` column to category type.
  4. Save cleaned dataset as `titanic_clean.csv`.

🔍 Reflection

- Why does cleaning take 80% of time?
- Would you always drop missing values?
- How does fixing data types help later in **visualization & ML**?