TP 5 Lab: Pandas Cleaning Pipeline — Load, Inspect, Clean, KPI, Export

---

Goal

Build a simple, reusable pandas pipeline:

- load a CSV (clean and dirty versions),
- inspect quality (types, missing values, duplicates),
- clean and standardize fields,
- compute a few KPIs,
- export a cleaned CSV.

Constraints

- Use a **virtual environment (venv)**.
- Use **pandas** (NumPy is allowed indirectly).
- No ML libraries (no scikit-learn, no torch, no tensorflow).
- Your cleaning logic must be reusable: implement `clean_df(df)`.

Required Project Structure

```
session5/
  main.py
  requirements.txt
  README.md
```

```
src/
  cleaning.py
data/
  clients_clean.csv
  clients_dirty.csv
out/
```

Data Files (copy/paste)

data/clients_clean.csv

```
client_id,age,city,income,spend,segment,signup_date
1,23,paris,2100,1200,A,2025-10-02
2,45,lyon,4200,2300,B,2025-09-15
3,31,lille,3200,1800,A,2025-10-20
4,29,paris,2900,1600,A,2025-08-11
5,52,marseille,5100,2600,C,2025-07-19
6,41,lyon,3900,2100,B,2025-11-01
7,36,paris,3600,1950,B,2025-12-05
8,27,lille,2500,1400,A,2025-10-09
```

data/clients_dirty.csv

```
client_id,age,city,income,spend,segment,signup_date
1,23, Paris ,2100,1200,a,2025-10-02
2,45,LYON,4200,2300,B,2025-09-15
3,31,,3200,1800,A,2025-10-20
4, 29 ,paris,2900,1600,A,2025-08-11
5,52,marseille,5100,2600,C,2025-07-19
6,41,lyon,3900,2100,B,not_a_date
7,36,paris,3600,1950,B,2025-12-05
8,27,lille,2500,1400,A,2025-10-09
8,27,lille,2500,1400,A,2025-10-09
9,not_an_int,nantes,9999,9999,C,2025-10-31
10,19,PARIS,2,500,1500,A,2025-11-12
```

Tasks

Task A — Load + Inspect (20 min)

In `main.py`:

1. Load `clients_dirty.csv`
2. Print:

- `df.shape`
- `df.columns`
- `df.head()`
- `df.info()`
- missing values per column: `df.isna().sum()`

Task B — Cleaning Function (40 min)

In `src/cleaning.py`, implement:

- `clean_df(df: pd.DataFrame) -> pd.DataFrame`

Cleaning rules

- Convert numeric columns with coercion:
    - `age`, `income`, `spend` → numeric (`errors="coerce"`)
- Convert dates:
    - `signup_date` → datetime (`errors="coerce"`)
- Normalize text:

    - `city` → strip + lower, fill missing with `"unknown"`
    - `segment` → strip + upper
- Remove duplicates:
    - drop duplicates by `client_id` (keep first)
- Drop invalid rows:
    - drop rows where `age` or `income` or `spend` is missing
- Create derived columns:
    - `margin = income - spend`
    - `spend_ratio = spend / income`

Task C — KPI Computation (25 min)

Using the cleaned DataFrame:

- print final number of rows
- print mean `income` and mean `spend`
- print top 3 cities by number of clients (`value_counts().head(3)`)
- print mean `margin`

Bonus (optional): mean margin by city (`groupby("city")["margin"].mean()`).

Task D — Export (15 min)

Export the cleaned dataset to:

- `out/clients_cleaned.csv` (no index)

Deliverables

- `src/cleaning.py` with a correct `clean_df`
- `main.py` runnable from project root
- exported file: `out/clients_cleaned.csv`
- `README.md` filled (template below)

README Template (copy/paste into `README.md`)

````
# Session 5 — Pandas Cleaning Pipeline

## Setup
```bash
python -m venv venv
# activate venv (OS-specific)
pip install -r requirements.txt
````

Run

```
python main.py --input data/clients_clean.csv
python main.py --input data/clients_dirty.csv
```

Output

- Prints inspection info and KPIs
- Exports `out/clients_cleaned.csv`
```