TP 6 Lab: Pandas Reporting — Merge, Groupby KPIs, Pivot Table, Export + One Chart

---

Goal

Given two tables (clients + transactions), build a small reporting pipeline:

- clean minimal types,
- merge on a key,
- compute KPIs with `groupby().agg()`,
- build a `pivot_table` report,
- export CSV outputs and one chart image.

Constraints

- Use a **virtual environment (venv)**.
- Use **pandas** and **matplotlib** (no seaborn).
- No ML libraries (no scikit-learn, no torch, no tensorflow).
- Organize reusable functions in `src/analysis.py`.

Required Project Structure

```
session6/
  main.py
```

```
  requirements.txt
  README.md
  src/
    analysis.py
  data/
    clients.csv
    transactions.csv
    transactions_dirty.csv
  out/
```

Data Files (copy/paste)

data/clients.csv

```
client_id,age,city,segment
1,23,paris,A
2,45,lyon,B
3,31,lille,A
4,29,paris,A
5,52,marseille,C
6,41,lyon,B
7,36,paris,B
8,27,lille,A
9,19,nantes,C
```

data/transactions.csv

```
tx_id,client_id,date,product,amount
1001,1,2025-12-01,coffee,8.40
1002,2,2025-12-02,tea,12.00
1003,1,2025-12-02,coffee,4.20
1004,3,2025-12-03,chocolate,15.50
1005,4,2025-12-03,coffee,6.30
1006,5,2025-12-04,tea,20.00
1007,6,2025-12-05,coffee,9.10
1008,7,2025-12-05,chocolate,11.00
1009,8,2025-12-06,tea,7.00
```

```
1010,7,2025-12-06,coffee,5.60
1011,9,2025-12-07,coffee,13.00
```

`data/transactions_dirty.csv` (bonus robustness)

```
tx_id,client_id,date,product,amount
1001,1,2025-12-01,coffee,8.40
1002,2,2025-12-02,tea,12.00
1003,1,2025-12-02,coffee,
1004,3,2025-12-03,chocolate,15,50
1005,4,2025-12-03, coffee ,6.30
1006,5,not_a_date,tea,20.00
1007,6,2025-12-05,coffee,9.10
1008,7,2025-12-05,chocolate,11.00
1008,7,2025-12-05,chocolate,11.00
1012,999,2025-12-08,coffee,5.00
```

Tasks

Task A — Load + Minimal Cleaning + Key Checks (25 min)

Create functions in `src/analysis.py`:

- `load_clients(path: str) -> pd.DataFrame`
    - normalize:
        - `city`: strip + lower
        - `segment`: strip + upper
- `load_transactions(path: str) -> pd.DataFrame`
    - normalize:
        - `product`: strip + lower
    - convert:

        - `amount` to numeric (handle `"15,50"` by replacing comma with dot)
        - `date` to datetime (`errors="coerce"`)
- `validate_keys(clients, tx) -> None`
    - print warnings if:
        - `clients["client_id"]` has duplicates
        - `tx["tx_id"]` has duplicates

In `main.py`, call these and print:

- number of rows loaded for each table
- duplicate counts for keys

Task B — Merge + Clean After Merge (25 min)

Implement in `src/analysis.py`:

- `merge_data(clients, tx) -> pd.DataFrame`
    - left join on `client_id`

After merge in `main.py`:

- drop duplicate transactions by `tx_id` (keep first)
- drop invalid transaction rows (drop rows where `amount` is NaN or `date` is NaT)
- fill missing dimension fields after merge:
    - `city` → `"unknown"`
    - `segment` → `"UNKNOWN"`

Print:

- number of rows before/after cleaning

Task C — KPI Tables with `groupby().agg()` (30 min)

Implement in `src/analysis.py`:

- `kpi_by_city(df) -> pd.DataFrame` with:
    - `total_amount` (sum)
    - `n_tx` (count of tx_id)
    - `avg_amount` (mean)
- `kpi_by_city_segment(df) -> pd.DataFrame` with:
    - `total_amount` (sum)
    - `n_tx` (count)

Sort by `total_amount` descending and print both tables in `main.py`.

Task D — Pivot Report + Export (20 min)

Implement in `src/analysis.py`:

- `pivot_city_segment(df) -> pd.DataFrame`
    - index = `city`
    - columns = `segment`
    - values = sum of `amount`

    - fill_value = 0

Export to:

- `out/pivot_city_segment.csv`

Task E — One Chart + Export Image (15 min)

Create a bar chart:

- total amount by city

Export to:

- `out/amount_by_city.png`

Use pandas plotting (matplotlib under the hood). No seaborn.

Bonus Task (Optional, 15 min)

Run the full pipeline on:

- `data/transactions_dirty.csv`

Make sure your cleaning logic handles:

- empty amount → drop
- comma decimals `"15,50"` → convert
- invalid date `"not_a_date"` → drop
- duplicate `tx_id` → drop duplicates
- unknown `client_id` (e.g., 999) → keep but label city/segment as unknown

---

Deliverables

- Working `session6/` project with the required structure
- Exports in `out/`:
    - `pivot_city_segment.csv`
    - `amount_by_city.png`
- `README.md` filled (template below)

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

````
# Session 6 — Pandas Reporting (merge, groupby, pivot, chart)

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

Run

```
python main.py --tx data/transactions.csv
python main.py --tx data/transactions_dirty.csv
```

Output

- Prints KPI tables (city, city+segment)
- Exports `out/pivot_city_segment.csv`
- Exports `out/amount_by_city.png`
```