Student Homework Sheet — Stage 04: Data Acquisition and Ingestion

Due: Next class session

Assignment

In the lecture, we learned **API acquisition, scraping with BeautifulSoup, secrets via `.env`, validation, and saving to `data/raw/`**.
Now, you will adapt these to build a reproducible ingestion workflow for one market-related dataset.

Tasks

1. **API Pull (required):**
    - Choose one ticker or endpoint.
    - Load API key from `.env` (if required).
    - Request data with `requests` *(or `yfinance` as fallback)*.
    - Convert to DataFrame; parse dtypes (dates, floats).
    - Validate (required columns, NA counts, shape).
    - Save raw CSV to `data/raw/`.
2. **Scrape a Small Table (required):**
    - Public, permitted page with a simple table.
    - Parse with `BeautifulSoup`; build DataFrame.
    - Validate numeric/text columns.
    - Save raw CSV to `data/raw/`.
3. **Documentation (required):**
    - In the notebook: list data sources/URLs, params, and validation logic.
    - Confirm `.env` is **not committed**.
    - Include a short "assumptions & risks" cell.

Step-by-Step

- Start from `stage04_data-acquisition-and-ingestion_homework-starter.ipynb`.
- Fill in the TODO cells (API, scrape, validate, save).
- Run all cells; push notebook + saved CSVs to GitHub.

Grading Rubric (100 pts)

- **API Ingestion (30)** — works, handles errors, parses types
- **Scraping Ingestion (30)** — correct parse, resilient selectors
- **Validation (20)** — required cols, NA counts, basic rules
- **Reproducibility & Docs (20)** — `.env`, filenames, sources, explanations

Example Deliverables

- Notebook with executed cells
- `data/raw/api_<SOURCE>_<TICKER>_<YYYYMMDD-HHMM>.csv`
- `data/raw/scrape_<SITE>_<TABLE>_<YYYYMMDD-HHMM>.csv`
- `.env` present locally; `.env.example` in repo

> **Chain:** In the lecture, we learned how to **pull data via APIs, scrape tables, validate, and save raw files**.
> **Now, you will adapt those patterns** to **ingest one API dataset and one scraped table**, with validations and reproducible filenames.
