Fragrantica Data Project

# 1. Data Pipeline

**Stage 1 — Raw Data**

* Collect CSVs (fragrance\_raw.csv, brand\_raw.csv).
* Store **exactly as received** in /data/raw/.
* Do not touch (acts as single source of truth).

**Stage 2 — Light Cleaning (Excel)**

* Fix column names, remove obvious duplicates.
* Handle missing values where easy (replace N/A with NULL, fix decimal separators).
* Save into /data/staging/ (e.g., fragrance\_excelcleaned.csv).

**Stage 3 — Import into PostgreSQL (Staging)**

* Create **staging tables** mirroring CSVs (staging\_fragrance, staging\_brand).
* Import with COPY or \copy in psql.

**Stage 4 — Heavy Cleaning / Transformation (SQL)**

* Parse and normalize:
  + Split accords into separate rows (limit to first 5).
  + Split top/middle/base notes into separate tables.
  + Handle NULL values properly.
  + Extract brand name from URL if missing.
* Insert into **normalized schema** (fragrance, brand, accord, fragrance\_accord, notes, perfumers).

**Stage 5 — Processed Exports**

* Export clean tables back into /data/processed/ with \copy:
  + fragrance\_cleaned.csv
  + brand\_cleaned.csv
  + fragrance\_accord.csv

**Stage 6 — Analysis & Visualization**

* Use SQL queries (/sql/queries/) to answer questions:
  + What are the most common accords?
  + Which brand has the highest average rating?
  + Rating distribution by gender?
* Create visualizations in Tableau, Power BI, or Python (visualisation/).

# 2. Naming Conventions

* **CSV files**: <table>\_<stage>.csv
  + fragrance\_raw.csv
  + fragrance\_excelcleaned.csv
  + fragrance\_cleaned.csv
* **SQL files**: <stage>\_<table>.sql
  + schema\_fragrance.sql
  + etl\_load\_fragrance.sql
  + query\_top\_accords.sql

# 3. Workflow

1. Drop new CSV into data/raw/.
2. Light clean → save to data/staging/.
3. Run sql/staging/\*.sql to create staging tables.
4. Import CSVs into staging tables.
5. Run sql/etl/\*.sql to load into main schema.
6. Export cleaned tables to data/processed/.
7. Run queries from sql/queries/ or build dashboards.