Project Title: Fashion Brand Data Analysis --The *ModaFlick* Case Study

Context:

You are a junior data analyst at ***ModaFlick***, a growing fashion brand that sells stylish wear through physical stores and online platforms. Management wants insight into customer behavior, product performance, inventory management, and ad campaign effectiveness.

Your task is to consolidate data from multiple sources, clean it (*if necessary*), and generate insights that help the business grow.

Objectives:

1. Create a database for ***Modaflick***
2. Extract and load data from PostgreSQL and external sources.
3. Clean and transform datasets to prepare for analysis.
4. Merge and analyze multiple DataFrames.
5. Extract business insights that influence strategic decisions.
6. Develop a local Python package for reusable ETL functions

NOTE: The csv files needed for the database is in the zipped ***ModaflickDB*** file.

Dataset Descriptions:

File Name	Description
customers.csv	Customer profiles (name, age, gender, city)
products.csv	Product catalog (name, category, price, brand)
orders.csv	Order metadata (customer, date, channel)
order_items.csv	Line items per order (product, quantity)
stores.csv	Physical store locations
inventory.csv	Product stock levels by store

Data Cleaning Tasks:

As a first step, perform the following on each dataset:

- Detect and remove duplicates
- Check for missing/null values and handle appropriately
- Standardize text formats (e.g., lowercase categories)
- Convert date strings to datetime objects
- Join foreign keys and validate referential integrity
- Create useful derived columns (e.g., total order value)

Data Integration Tasks:

Using `pandas` and `sqlalchemy`:

1. Load these CSVs as tables or import them into PostgreSQL (your choice).
2. Join `orders` → `order_items` → `products` to compute order values.
3. Join with `customers` to segment users by age or location.
4. Join with `stores` for store-specific analysis.
5. Integrate your inventory data and check for overstock/understock patterns.

Insight Generation (Analysis Questions):

1. What are the **top 5 best-selling products** by revenue and by volume?
2. How do **sales trend over time** (monthly/quarterly)?
3. What is the **average order value** across customer segments (age, city)?
4. Who are the **most frequent buyers** and what do they buy?
5. Are there **inventory mismatches** (e.g., low stock for popular items)?
6. Compare **store vs online** sales: which performs better and why?
7. What **customer behaviors** can you derive from the data?

External Files (These should be pulled directly from the internet. *DO NOT DOWNLOAD*)

1. marketing_budget.csv
2. warehouse_inventory.xlsx

Final Deliverables:

- A Jupyter notebook (`modaflick_analysis.ipynb`) with:
  - Data loading
  - Cleaning
  - Merging
  - Analysis & visualizations
  - Business recommendations
- A local ETL package
- Optional: export cleaned or summarized data to CSV

HOW THE DATASETS FIT INTO THE PROJECT

Component	Data Format	Source	Role in Project
`customers.csv`	CSV	Internal Database	Understand customer demographics and behavior
`products.csv`	CSV	Internal Database	Product catalog and pricing info
`orders.csv`	CSV	Internal Database	Transactional sales data
`order_items.csv`	CSV	Internal Database	Order-level breakdown (units, SKUs)
`stores.csv`	CSV	Internal Database	Store metadata
`inventory.csv`	CSV	Internal Database	Point-in-time stock levels
`marketing_budget.csv`	CSV	External Sheet (Marketing)	Campaign insights and spend-performance correlation
`warehouse_inventory.xlsx`	Excel	External File (Warehouse)	Stock movement by month and warehouse