Final Project: Retail Store Data Pipeline & Analysis

Technologies: Python, Pandas, SQL, SQL Server,

SQLAlchemy **Project Overview**

In this project, students will build a mini end-to-end data pipeline using real retail data. They will extract data from multiple CSV files, clean and transform it using Python and Pandas, then load it into SQL Server, and finally perform analytical SQL queries.

This project combines everything learned throughout the course in Python + SQL.

Project Requirements

1. Data Loading (Python + Pandas)

Students must:

- Load all CSV files using pandas.read_csv() ====> Data Link
- Handle encoding issues and separators if needed
- Print rows/columns count for each dataset
- Standardize column names:
- all lowercase
- replace spaces with _

2. Data Cleaning Requirements

Each dataset must be properly cleaned.

✔ Missing Values

- Detect and handle missing values
- Drop rows if necessary
- Fill values when logical (e.g., missing phone → "Unknown")

✔ Data Types

- Convert IDs to integers
- Convert order dates to datetime
- Convert prices to float
- Ensure foreign key columns have matching data types

✔ Outliers & Format Issues

- Fix incorrect or inconsistent values
- Remove negative quantities

- Clean multi-valued phone numbers, e.g.: “3389745, 3389744, 5123445” → extract the first or keep as list

✔ Duplicates

- Detect duplicate rows
- Remove or fix them

3. Data Transformation Requirements

- Mandatory transformations:
- Merge products with brands and categories
- Calculate total price per order item:
- total_price = quantity * list_price
- Calculate order total amount (sum of all items grouped by order_id)
- Create a full_name column for customers (if applicable)
- Clean and standardize customer phone numbers:
- remove spaces
- keep digits only

4. SQL Server Database Requirements

✔ Create a database:

- **RetailDB**

✔ Create the following tables:

- **Brands**
- **Categories**
- **Products**
- **Customers**
- **Orders**
- **OrderItems**
- **Staffs**
- **Stores**
- **Stocks**

✔ Include:

- **Primary keys**
- **Foreign keys**
- **Correct data types**
- **Proper normalization (up to 3NF)**

✔ Load the cleaned data to SQL Server using:

- Python (SQLAlchemy)

5. Required SQL Analysis Queries

A) Sales Analysis

- Top 10 best-selling products
- Top 5 customers by spending
- Revenue per store
- Revenue per category
- Monthly sales trend

B) Inventory Analysis

1. **Products with low stock**
2. **Stores with the highest inventory levels**

C) Staff Performance

1. **Number of orders handled by each staff member**
2. **Best-performing staff member by total sales**

D) Customer Insights

1. **Customers with no orders**
2. **Average spending per customer**

6. Python Final Report (Optional but Recommended) -> Bonus

- Students may create a Python script that:
- Retrieves analysis results
- Generates simple plots using Pandas
- Saves a final report (CSV)

Deliverables

Students must submit:

- Python script (main.py)
- Loading
- Cleaning
- Transforming
- SQL loading
- Cleaned datasets (CSV format)
- SQL Server database

- All tables created
- Data loaded
- SQL script (analysis_queries.sql)
- Contains all required analytical queries
- README file
- **Documentation of the pipeline, steps, and insights in a GitHub repo.**