PROJECT SPECIFICATION: RETAIL SALES INSIGHT DASHBOARD

Role: Junior Data Analyst

Status: OPEN

Priority: HIGH

Date: October 26, 2023

1. Business Objective

The Goal:

The executive team needs to identify high-performing regions and product lines to inform the Q4 marketing strategy. We currently have raw, unstructured monthly sales data that provides no actionable insights.

** The Deliverable:**

A Python-based analysis pipeline that cleans raw data, stores it in a structured SQL database, and generates visual reports answering: *"Which products should we prioritize for the Q4 marketing campaign?"*

2. Technical Requirements

- **Language:** Python 3.10+
- **Core Libraries:**
  - pandas (Data manipulation)
  - matplotlib or seaborn (Data visualization)
  - sqlalchemy (Database ORM/Connection)
- **Database:** SQL (SQLite for local portability).
- **Input Source:** sales_data_raw.csv
  - *Columns:* TransactionID, Date, Product, Region, Price, Quantity.

3. Functional Specifications (Scope of Work)

Phase 1: ETL Pipeline (Extract, Transform, Load)

The script must perform the following operations automatically:

1. **Ingest:** Read the raw CSV file.
2. **Clean & Transform:**

    - Identify and drop rows with NULL values in critical columns ('Price', 'Quantity').
    - Cast the Date column to standard ISO format (YYYY-MM-DD).
    - **Calculated Field:** Create a new column Total_Revenue = (Price * Quantity).
3. **Load:**
    - Initialize a local SQLite database (sales_data.db).
    - Write the cleaned dataframe to a SQL table named clean_sales_records.

Phase 2: SQL Analysis

The Python script must execute SQL queries against the database to extract:

- Total revenue grouped by Region.
- Total units sold grouped by Product.

Phase 3: Visualization Output

The system must generate and save the following assets to an /output folder:

1. **Bar Chart:** Top 5 selling products by revenue.
2. **Line Chart:** Monthly revenue trends over the last fiscal year.
3. **Heatmap (Optional):** Sales performance by Region vs. Product.

4. Acceptance Criteria (Definition of Done)

- [ ] Script runs from start to finish without syntax errors.
- [ ] A local file named sales_data.db is created and populated.
- [ ] SQL queries return accurate aggregations (verified against manual CSV check).
- [ ] Two distinct, legible PNG image files are generated.
- [ ] Code is PEP8 compliant and includes docstrings explaining the ETL logic.