|  |  |
| --- | --- |
| **Ex No: 1**  **Date: 06-08-25** | **Exploring the Data Engineering Lifecycle and Stakeholder Roles** |

**Objective:**

This lab provides hands-on experience exploring the data engineering lifecycle and understanding the roles of key stakeholders. Participants will simulate responsibilities of data engineers, data scientists, and business analysts while examining raw data sources and planning a data-driven solution.

**Outcomes:**

1. Identify and describe each stage of the data engineering lifecycle.
2. Explain the specific responsibilities of stakeholders across the lifecycle.
3. Collaborate to define a business problem using raw data sources.
4. Draft a requirements document based on the business use case.

**Materials:**

* Raw sales data CSV file (sales\_data\_raw.csv)
* Customer feedback JSON file (customer\_feedback.json)
* Folder structure representing a mock data warehouse or data lake

**Lab Procedure:**

**Stage 1: Problem Definition and Requirements Gathering (Business Analyst)**

1. Review both datasets provided (sales\_data\_raw.csv and customer\_feedback.json).
2. Formulate a business question, e.g., “What are the top 5 products by revenue, and how does customer sentiment vary for them?”
3. Identify required data points (e.g., product\_id, sale\_price, customer\_id, sentiment\_score).
4. Create a short requirements document outlining the problem, key metrics, and desired insights.

**Stage 2: Role-Based Collaboration Simulation**

1. Discuss and map out how the Data Engineer will ingest and clean the data.
2. Identify how the Data Scientist will analyze and model insights based on the cleaned data.
3. Define how the Business Analyst will interpret and report results.
4. Define how each stakeholder contributes to the overall data solution.
5. Document the flow of responsibilities and dependencies between roles.

**GitHub Link:** [**https://github.com/tanishasn/FDE\_LAB1\_1RVU23CSE500**](https://github.com/tanishasn/FDE_LAB1_1RVU23CSE500)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Stage 1: Problem Definition and Requirements Gathering**

**Role:** Business Analyst

**Dataset Review**

We have been provided with two datasets:

1. **sales\_data\_raw.csv** — contains transactional sales data with columns such as:
   * product\_id — Unique identifier for products
   * sale\_price — Price at which the product was sold
   * quantity — Number of units sold in a transaction
   * customer\_id — ID of the purchasing customer
   * sale\_date — Date of the sale
2. **customer\_feedback.json** — contains customer sentiment and review data with columns such as:
   * customer\_id — Links feedback to sales records
   * product\_id — Identifies the product reviewed
   * sentiment\_score — Customer satisfaction score (scale of 1–5)
   * review\_date — Date when the feedback was given

Using Python (pandas), both datasets can be loaded, previewed, and analyzed for completeness, column types, and potential data quality issues.

**Business Question**

“What are the top 5 products by total revenue, and how does customer sentiment vary for them?”

**Required Data Points**

| **Data Point** | **Source File** | **Purpose** |
| --- | --- | --- |
| product\_id | sales\_data\_raw.csv & customer\_feedback.json | Links sales data with customer feedback |
| sale\_price | sales\_data\_raw.csv | Calculates revenue per sale |
| quantity | sales\_data\_raw.csv | Calculates total units sold |
| customer\_id | Both datasets | Connects purchases with feedback |
| sentiment\_score | customer\_feedback.json | Measures customer satisfaction |
| sale\_date | sales\_data\_raw.csv | Tracks sales trends |
| review\_date | customer\_feedback.json | Tracks sentiment trends |

**Problem Statement**

The business wants to identify the top-performing products based on **total revenue** and examine how customer sentiment scores vary for these products. By integrating sales and feedback data, we can uncover products that are high in revenue but low in satisfaction (potential improvement areas) and track performance trends over time.

**Key Metrics**

**From Sales Data**

* Total Revenue = sum(sale\_price × quantity) per product
* Units Sold per product
* Monthly Revenue Trends

**From Feedback Data**

* Average Sentiment Score per product
* Monthly Sentiment Trends

**Combined Metrics**

* Revenue vs. Sentiment correlation
* High Revenue / Low Sentiment products

**Desired Insights**

1. Top 5 products ranked by total revenue.
2. Average sentiment score for each top product.
3. Products with high sales but low sentiment (potential quality/service issues).
4. Trends in sales and sentiment over time for top products.

**Code Alignment**

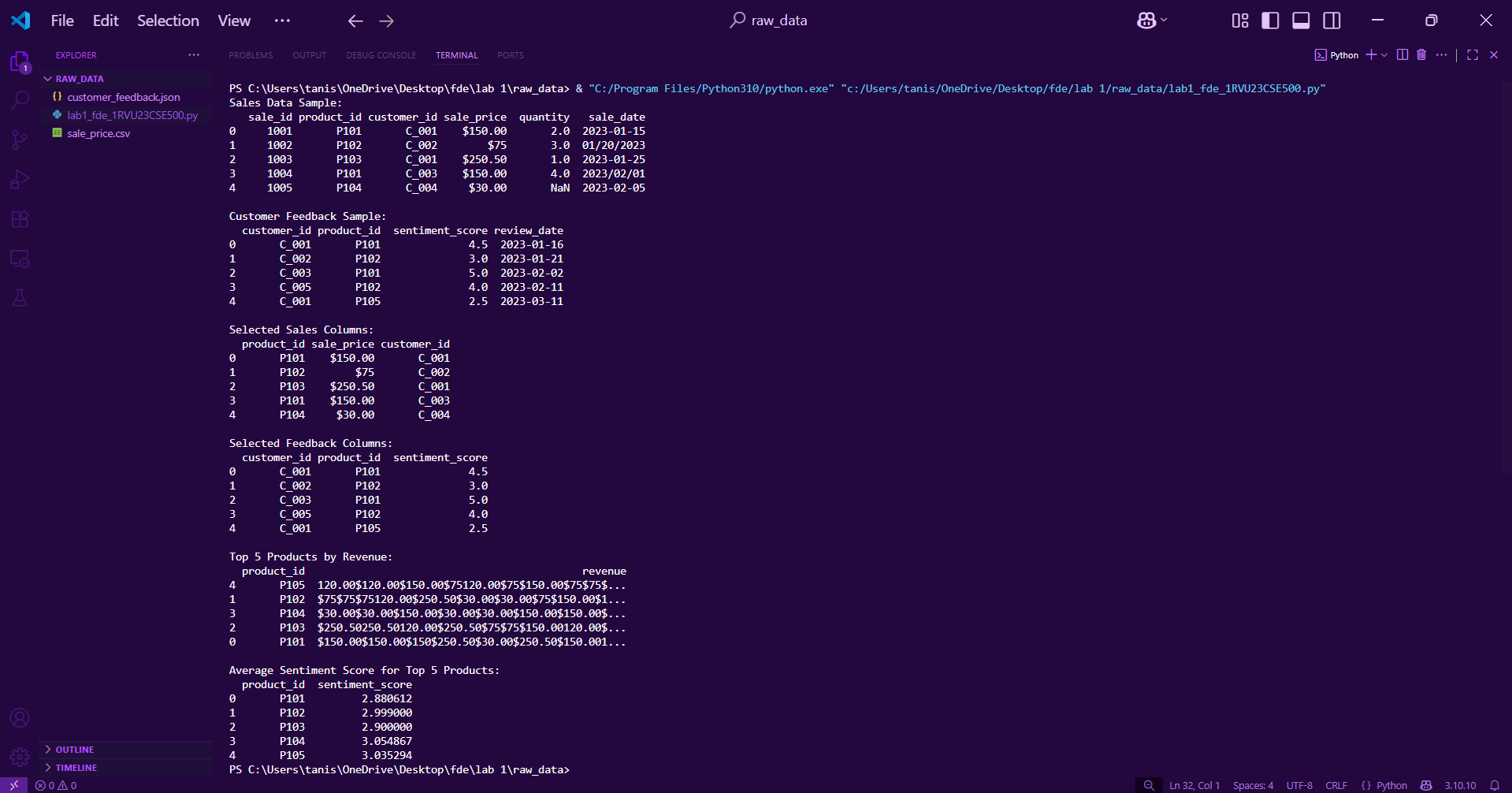
The accompanying Python code:

* Loads both CSV and JSON datasets using pandas.
* Selects relevant columns for analysis.
* Calculates total revenue per product.
* Identifies the top 5 products by revenue.
* Merges top products with customer feedback data.
* Calculates the average sentiment score for each top product.

LAB1 CODE:-



LAB1 OUTPUT:-



**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Stage 2:-

**1. Discuss and map out how the Data Engineer will ingest and clean the data.**

* **Ingest raw data:** Import raw sales data from sales\_data\_raw.csv and customer feedback from customer\_feedback.json into the data warehouse or data lake, ensuring correct schema mapping for compatibility.
* **Schema validation:** Verify that data types, column names, and structures match the defined schema before ingestion to avoid mismatches during processing.
* **Data cleaning:** Remove duplicate records, fix inconsistent formatting (e.g., date formats, currency units), and ensure all product IDs follow a standardized format across both datasets.
* **Data enrichment:** Fill missing values where possible (e.g., sentiment scores) using statistical imputation or default values.
* **Merging datasets:** Join sales data with customer feedback on common keys like Product\_ID to enable combined analysis.
* **Optimization:** Create partitioned and indexed tables/views in the data warehouse for faster querying and analysis.

**2. Identify how the Data Scientist will analyze and model insights based on the cleaned data.**

* **Exploratory Data Analysis (EDA):** Generate statistical summaries, visualizations, and correlation maps to identify trends, relationships, and potential anomalies in the dataset.
* **Trend detection:** Determine seasonal sales patterns, high-demand product categories, and emerging customer preferences.
* **Sentiment analysis:** Analyze sentiment distribution for each product using text mining/NLP techniques on feedback data.
* **Model development:** Build scripts or models to identify top revenue-generating products and measure the relationship between sentiment scores and sales performance.
* **Validation:** Apply statistical tests (e.g., t-tests, chi-square) to ensure findings are statistically significant and not due to random chance.
* **Reproducibility:** Ensure all scripts and workflows are version-controlled, parameterized, and documented for future reuse.

**3. Define how the Business Analyst will interpret and report results.**

* **Review analytical results:** Evaluate the list of top-performing products and sentiment trends against the company’s strategic goals.
* **Identify business opportunities:** Highlight which products require marketing boosts, quality improvements, or stock adjustments.
* **Translate insights:** Convert technical analytical findings into business-friendly recommendations, avoiding excessive jargon.
* **Decision support:** Suggest targeted marketing campaigns, bundling strategies, or product development ideas based on insights.
* **Reporting:** Create easy-to-understand dashboards and reports (e.g., in Power BI/Tableau) tailored for stakeholders and executives.

**4. Define how each stakeholder contributes to the overall data solution.**

* **Data Engineer:** Handles data ingestion, transformation, and integration from multiple sources into a clean, structured format.
* **Data Scientist:** Applies analytical techniques, develops predictive/statistical models, and extracts actionable patterns from data.
* **Business Analyst:** Aligns technical outputs with business needs, making data-driven recommendations for strategy and operations.

**5. Document the flow of responsibilities and dependencies between roles.**

* **Data Engineer → Data Scientist:** Supplies well-structured, optimized datasets that are free from errors and ready for advanced analysis.
* **Data Scientist → Business Analyst:** Delivers processed results, predictive models, and visualizations for interpretation.
* **Business Analyst → Stakeholders:** Presents findings in a clear, actionable format to aid decision-making and strategy formulation.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**