SQL for Data Analysis – Online Retail Dataset

Objective

The objective of this project is to use SQL queries to extract, analyze, and gain insights from a structured Ecommerce dataset. The dataset used is the Online Retail dataset from Kaggle, containing transaction-level information such as customer IDs, invoice numbers, quantities, and product details.

Learning Outcomes

- Learned to connect a dataset with a SQL database using Python (SQLite).
- Performed data extraction, aggregation, and analysis using SQL queries.
- Executed SQL queries in Google Colab.
- Saved results, took screenshots, and submitted the project on GitHub.

Tools & Technologies Used

• Google Colab – Run SQL and Python code online • SQLite3 – Lightweight SQL database • Pandas – Data handling and analysis • Kaggle Dataset (Online Retail) • GitHub – Version control and submission

Step-by-Step Process

1. Open Google Colab and create a new notebook named SQL_Task4_Ecommerce. 2. Upload dataset using files.upload(). 3. Import required libraries (pandas and sqlite3). 4. Load dataset and view columns. 5. Create SQLite database and load data. 6. Verify table using SELECT * query. 7. Execute the required SQL queries (total revenue, top products, etc.). 8. Save query results as CSV. 9. Download CSVs and screenshots. 10. Upload everything to GitHub.

Key SQL Queries

1. Display first 10 rows. 2. Total revenue per customer. 3. Top 5 countries by revenue. 4. Average order value per customer. 5. Monthly sales trend. 6. Highest selling products.

Interview Questions

1. WHERE vs HAVING – WHERE filters before grouping, HAVING filters after. 2. Joins – INNER, LEFT, RIGHT, FULL OUTER. 3. Average revenue per user – Use AVG() with GROUP BY. 4. Subquery – Query inside another query. 5. Optimization – Use indexes, avoid SELECT *. 6. View – Virtual table based on query. 7. Handle NULL – Use IFNULL() or COALESCE().

Final Outcome

Successfully created an SQL analysis workflow using Google Colab and SQLite. Extracted valuable insights from the ecommerce dataset, including top customers, countries, and product trends. Results were exported and documented on GitHub.