Data Science Assignment Report

1. Data Generation Methodology

The relational database ECommerce_DataScience_DB.sqlite was constructed using a Python script employing the sqlite3 library for database management and Faker for generating realistic synthetic records. The process involved schema definition, parent table population, and transaction table generation, ensuring full referential integrity and realistic data variation.

Intentional data imperfections were introduced, including missing values and a deliberate duplicate email insertion to demonstrate constraint enforcement within the database.

2. Database Schema and Rationale

The database uses a normalized, star-like structure consisting of three tables (Products, Customers, Order_Items). The schema adheres to Third Normal Form, minimizing redundancy and enforcing consistency through primary keys, foreign keys, CHECK constraints, and UNIQUE constraints.

3. Data Types and Measurement Levels

The Order_Items table includes variables across all four measurement levels: nominal (warehouse_location), ordinal (discount_level), interval (shipping_time_days), and ratio (quantity).

4. Ethical and Data Privacy Considerations

Although all data is synthetic, the project applies ethical principles relevant to real-world data handling, including anonymization practices, data minimisation, transparency, bias awareness, and adherence to data protection standards such as GDPR.