

## SQL Analysis – Query-by-Query Explanation

This document explains the purpose of each SQL query used during the analysis phase, and how each step logically led to the next.

### Query 1: Initial Data Exploration

**SQL Logic:** `SELECT * FROM bigquery-public-data.thelook_ecommerce.order_items`

- Purpose: Take an initial exploratory look at the raw data.
- Reason: Before asking any business questions, it is critical to understand available columns, data types, and general structure.
- Outcome: Identified revenue-related fields and grain (order-item level).

### Query 2: Date Range Validation

**SQL Logic:** `MIN(created_at), MAX(created_at) FROM orders`

- Purpose: Determine the temporal coverage of the dataset.
- Reason: Understanding the start and end dates is necessary to decide whether yearly or monthly analysis is appropriate.
- Outcome: Confirmed multi-year coverage suitable for trend analysis.

### Query 3: Orders and Users Volume Check

**SQL Logic:** `COUNT(DISTINCT order_id), COUNT(DISTINCT user_id)`

- Purpose: Measure dataset scale in terms of customers and orders.
- Reason: Validates whether the dataset is large enough to support retention analysis.
- Outcome: Confirmed sufficient volume for customer segmentation.

### Query 4: Total Revenue Calculation

**SQL Logic:** `SUM(sale_price) FROM order_items`

- Purpose: Establish baseline total revenue.
- Reason: Serves as a reference point for later revenue splits (new vs returning).
- Outcome: Used as a control number for validation in Power BI.

### Query 5: Yearly Performance Metrics

**SQL Logic:** Revenue, Orders, AOV grouped by year

- Purpose: Understand high-level business performance trends over time.
- Reason: Identifies whether growth is volume-driven or value-driven.

- Outcome: Observed relatively stable AOV, suggesting growth depends on user behavior.

## Query 6: New vs Returning Users Analysis

**SQL Logic:** First order CTE + yearly user classification

- Purpose: Identify proportion of new users per year.
- Reason: Hypothesis emerged that revenue relies heavily on new customers.
- Outcome: Confirmed high dependency on new users, leading directly to the business question.