Data Cleaning & Engineering Report

Objective: Transform raw, messy data into a consistent, reliable dataset suitable for high-level business analysis and SQL querying.

1- Missing Values Strategy (Handling Nulls)

We categorized missing data into three specific actions based on business context:

- **Action A: Drop (Remove Data)**
    - **Columns (notes, internal_flag in Sessions):** Dropped entirely. These contained 100% nulls or random garbage data with no analytical value.
    - **Rows (campaign_id in Ad Spend):** Dropped rows where the campaign ID was missing (<0.5%). We cannot attribute spend to a non-existent campaign.
- **Action B: Impute (Fill Data)**
    - **utm_source & utm_medium:** Imputed with **'unknown'**.
        - *Rationale:* Preserves total session counts. Dropping these rows would artificially deflate site traffic metrics.
    - **gender:** Imputed with **'Unknown'**.
        - *Rationale:* Ensures these customers remain visible as a distinct demographic segment rather than disappearing from analysis.
    - **session_id (in Orders):** Imputed with **'offline_or_missing'**.
        - *Rationale:* These represent Call Center orders or tracking failures. We must retain the revenue data, even if the digital source is missing.
- **Action C: Intentional Ignore (Keep as Null)**
    - **end_date (in Campaigns):** Kept as Null.
        - *Meaning:* Indicates an **Active/Ongoing** campaign.
    - **customer_id (in Sessions):** Kept as Null.
        - *Meaning:* Represents **Guest/Anonymous Visitors** who have not logged in.
    - **csat_score (in Support):** Kept as Null.
        - *Meaning:* The customer declined to rate the service (**No Rating**).

2- Standardization (Categorical Cleanup)

Addressed inconsistent naming conventions to ensure accurate aggregation:

- **Whitespace Removal (Trimming):** Fixed "ghost" spaces in column headers (e.g., campaign_id) that cause code errors.

- **Case Normalization:** Converted all categorical text to **lowercase** (e.g., MOBILE $\rightarrow$ mobile) to eliminate case-sensitivity duplicates.
- **Entity Mapping (Dictionaries):**
  - **Traffic Sources:** Mapped variations like fb, Meta, facebook ads $\rightarrow$ **facebook**. (Applied similarly to Google and TikTok).
  - **Payment Methods:** Mapped cod, Cash, cash $\rightarrow$ **cash on delivery**.

3- Business Logic & Sanity Checks

Fixed data points that were technically valid but logically impossible:

- **Session "Time Travel" & Bots:**
  - Identified sessions with **negative duration** (End Time < Start Time) and **extreme duration** (> 6 hours).
  - *Action:* Removed these rows to prevent skewing the "Average Session Duration" metric.
- **Invalid Campaign Dates:**
  - Identified campaigns where end_date was earlier than start_date.
  - *Action:* Removed these illogical records.
- **Negative Financials:**
  - Identified orders with negative prices or shipping costs.
  - *Action:* Converted values to absolute (positive) numbers, assuming data entry sign errors.

4- Financial Accuracy

- **Calculation Audit:** Discovered that ~1% of orders had incorrect total_amount values due to system errors.
  - *Action:* Overwrote total_amount using the correct formula: $(Subtotal + Shipping + Tax) - Discount$.
- **Rounding Noise:** Differentiated between tiny floating-point differences (< 1.00 EGP) and real calculation errors, ensuring we fixed the significant discrepancies without over-correcting noise.

5- Referential Integrity

- **Orphan Orders:** Identified orders linked to a customer_id that does not exist in the Customers table.
    - *Action:* Set the invalid customer_id to **Null** instead of deleting the order.
    - *Rationale:* This preserves the **Revenue** in financial reports, attributing it to an "Unknown Customer" rather than losing the sale entirely.

6- Data Quality & PII

- **Email Validation:**
    - Removed invalid email formats (missing @ or .).
    - Retained **Duplicate Emails** (multiple accounts per email) as this reflects valid customer behavior useful for marketing analysis (e.g., Loyalty checks).