Data Handover Report: Verulam Blue E-Commerce Analytics

Document Version: 1.0
To: Analytics Engineering Team
From: Data Acquisition & Migration Team, Verulam Blue

1. Introduction & Background

Following the acquisition of DataFlow Inc., you are now the primary owner of the historical e-commerce transaction dataset generated by their legacy order management system. This dataset is critical for our initial performance assessment.

> **Important:** The legacy DataFlow team is no longer available for support; this document contains all known information regarding the data's structure, quality, and the business requirements for its use.

Your role is to validate, clean, and transform this raw data into an analysis-ready state to produce a standard set of performance indicators.

2. Business Context & Objectives

The Leadership Team requires a clear, accurate view of sales performance, customer value, and operational efficiency. Your primary deliverable is a report containing the following Key Performance Indicators (KPIs):

KPI 1: Average Order Value (AOV)

The average value of all orders.

KPI 2: Overall Gross Margin %

Total profit (revenue minus cost) as a percentage of total revenue.

KPI 3: Return Rate

The percentage of all orders that were returned.

KPI 4: Median Order Amount

The middle value of all order amounts.

KPI 5: Return Rate by Payment Method

The return rate, broken down by each payment method.

KPI 6: High-Value Customer GMV Share

The percentage of total Gross Merchandise Volume (GMV) attributed to the 'Premium' and 'Platinum' customer segments.

KPI 7: Below-Target Margin Rate

The percentage of orders that do not meet the segment’s minimum

KPI 8: Top-GMV Month in 2024

The month with the highest total sales (format YYYY-MM).

profit margin requirement.

KPI 9: Latest Month-over-Month (MoM) GMV Growth %

The GMV growth rate from the second-to-last month to the final month in the dataset.

KPI 10: Max Month-to-Month Payment-Method Share Shift (pp)

Largest percentage-point change in any single payment method's order share from one month to the next.

Note: The target profit margin floors for KPI 7 are business-critical and are as follows:

Standard customers	Premium customers	Platinum customers
≥ 40%	≥ 30%	>25%

3. Dataset Description

You will receive a single dataset with the following columns. Note that the data types are not enforced and the column is stored as provided.

Column Name	Description	Notes
`row_id`	A unique identifier for each row.	Should correspond to a single order.
`date`	The date of the order.	Stored as a string with format inconsistencies.
`customer_segment`	The customer's tier.	Contains typographical errors.
`order_amount_old`	The value of the order in USD.	Contains missing values.
`cost`	The cost of goods sold for the order in USD.	
`is_return`	A flag indicating if the order was returned.	1 for yes, 0 for no.
`payment_method`	The method used to pay for the order.	
`hour_of_day`	The hour of the day the order was placed.	24-hour format.

4. Known Data Quality Issues

The following issues were identified during the final export and transfer process from DataFlow Inc.'s systems and must be addressed before analysis:

1. Duplicate Records

A small number of duplicate rows were introduced during the data extraction process. All duplicate records must be identified and removed.

2. Date Format Inconsistency

The date column suffers from a critical formatting issue. A portion of the records were exported in DD-MM-YYYY format, while the remainder use a non-standard YYYY.MM.DD format. The column is stored as a string and must be parsed into a standard date type.

3. Missing Values

The `order_amount_old` field contains null values for a small subset of records. Per business rules, all orders must have a value greater than or equal to $5.00. Records with null or otherwise invalid order amounts must be removed.

4. Typographical Errors

The `customer_segment` column contains minor spelling errors (e.g., `standrad`, `premuim`, `platnum`) for a very small fraction of records. These must be mapped to the canonical values: `standard`, `premium`, and `platinum`.

5. Schema Drift

The column for order value was recently renamed in the source system from `order_amount` to `order_amount_old`.
Feel free to use the current name but be aware of its original purpose.

5. Validation Rules & Business Logic

Please apply the following rules during your data cleaning and validation process:

> **Order Amount (order_amount_old):** Must be a positive number with a minimum value of $5.00.

> **Cost:** Must be a positive number.

> **Gross Margin:** Calculated as `(order_amount_old - cost) / order_amount_old`.

> **Customer Segments:** Must be one of: `standard`, `premium`, `platinum`. Map any typos to these canonical values.

> **Returns:** The `is_return` flag should only contain the values 0 or 1.

> **Duplicates:** All duplicate records must be removed.

> **Null Values:** Any records containing null values in key fields must be removed.

6. Next Steps & Deliverables

Cleaned Dataset

Analysis-ready dataset in standardised format

KPI Calculations

Complete set of 10 calculated KPIs in specified structure

Data Cleaning

Create a cleaned, analysis-ready dataset by addressing all known data quality issues:

1. Removing duplicate records
2. Standardising date formats to a consistent DD-MM-YYYY format
3. Handling null values according to business rules
4. Ensuring data type consistency

KPI Calculation

Using the cleaned and joined data, calculate the 10 KPIs listed in Section 2

7. Closing and Way Forward

Reference: This document will serve as your foundational reference as you move forward.

Handover of Responsibility: With the background and context provided in this report, we are confident in your ability to take the project forward and complete the assignment.