Full Name	Mermoul Badis
Email	sabermermoul50@gmail.com
Contact Number	+4915168850754
Date of Submission	2025-07-18
Project Week	Week 2

Sales & Customer Behavior Insights- Green Cart Ltd.

Introduction:

Green Cart Ltd., a rapidly growing e-commerce company specializing in eco-friendly household products, is preparing for its Q2 performance review.

This report aims to analyze sales trends and customer behavior across various regions and product categories. The insights derived will support the marketing team in developing more targeted campaigns and help inform data-driven operational strategies for the upcoming quarter.

The business Questions:

To guide our analysis and inform strategic decision-making, we aim to answer the following key business questions:

1. Which product categories derive the most revenue,and in which regions?
2. Do discounts lead to more items sold ?
3. Which loyalty tier generates the most value?
4. Are certain regions struggling with delivery delays?
5. Do customers signup patterns influence purchasing activity?

Our analysis was structured into five key steps to ensure data accuracy, depth of insight, and actionable outcomes:

- Loading and data cleaning data.
  We began by importing the raw datasets and addressing missing values, duplicates, and formatting inconsistencies to ensure data quality and integrity.

- Dataset Integration
  Multiple data sources were merged into a unified dataset to enable holistic analysis across customers, orders, products, and regions.

- Feature Engineering
  New variables were created to enrich the dataset and capture key behavioral or operational indicators.
- Summary Table Generation
  We utilized pivot tables and groupby() operations to produce aggregated views that highlight trends by category, region, customer tier, and other dimensions.

- Visual Analysis & Insight Discovery
  Visualizations were developed to reveal patterns and correlations not easily observed through raw numbers alone. These insights support data-driven recommendations for marketing and operations.

Loading and data cleaning data:

We began our analysis by importing and reviewing three key datasets provided in CSV format:

1. Customer_info.csv

    - Contains customer details including gender, signup date, and region.
    - Initial size: 500 records × 6 columns.

2. Sales_date.csv

    - Includes transaction-level sales information for Q2, specifically dated "2025-06-07."
    - Contains 3,000 records × 10 columns, with fields such as quantity sold, price, and region.

3. Product_info.csv

    - Describes 30 unique products, including product name, category, and additional descriptive attributes.
    - This dataset supports the classification and enrichment of sales data.

Cleaning Steps Performed:

- Data Type Corrections
  To ensure accurate analysis and enable time-based operations, we verified and corrected data types across the datasets:

    1. signup_date (Customer dataset) , order_date (Sales dataset) and launch_date (product dataset) were initially stored as object types (strings). These were converted to datetime format to support chronological sorting, filtering, and time-based calculations.

    2. quantity (Sales dataset) was also stored as an object. It was converted to an integer type to enable accurate aggregation and arithmetic operations.

- Handling Missing Values:

  We addressed missing or null values using context-aware strategies:

    1. signup_date: Missing values were imputed using the most recent valid signup date in the dataset, assuming late data entry for active customers.

    2. order_date: Missing values were filled with the transaction date "2025-06-07," as all sales were known to occur on this specific Q2 date.

    3. discount_applied: Nulls were replaced with 0, under the assumption that a missing entry indicates no discount was used.

    4. Other missing values: Records with nulls in non-critical fields were dropped. These omissions were minimal and did not pose a risk of data loss or introduce bias, and no reliable imputation method was applicable.

Data Consistency Verification:

As part of our data cleaning process, we reviewed categorical (object-type) variables to ensure consistency across all datasets. This step was crucial for avoiding mismatches during grouping, filtering, and merging operations.

- Gender: Inconsistencies such as mixed casing (e.g., "Male", "male", "FEMALE") and trailing spaces were corrected. All entries were standardized to lower case (e.g., “male”, “female”)
- Quantity: Although this field should be numeric, some records had it stored as strings (e.g., " three ", "five"). These were cleaned by trimming spaces, correcting obvious textual entries, and converting values to integers.
- Loyalty Tier: Entries in the loyalty_tier column (e.g., “Gold”, “gold ”, “ SILVER”,”gld”,”brnze”) were cleaned by removing leading/trailing whitespace ,correcting misspelling and standardizing to proper case formatting (e.g., "gold", "silver", "bronze").

Handling Duplicate Records

During data quality checks, we identified instances of duplicated orders—specifically, rows with the same order_id (e.g., “O515400”, “O916245”) but associated with different customer_id values. This raised concerns about potential data entry errors or duplicate transactions.

- We analyzed the delivery_status field for each duplicated order.
- Only records with a status of "Delivered" were retained, under the assumption that successfully fulfilled orders represent valid transactions.
- Duplicates with undelivered or unclear statuses were removed to maintain data accuracy and avoid inflating sales metrics.

This approach ensured the integrity of our order-level analysis and prevented double-counting of transactions.

Handling illogical signup dates:

During data validation, we identified customer records where the signup_date occurred after the order_date—an illogical sequence indicating data entry errors.
To correct these anomalies, we replaced the invalid signup_date values with the most frequently occurring (mode) signup date in the dataset. This approach ensured data consistency while minimizing distortion of customer behavior patterns.

Feature Engineering

To enhance the dataset and enable deeper analysis, we created several derived features that add context and analytical value:

- Dataset Merging : All three source datasets — customer_info, sales_data, and product_info — were merged into a single unified dataset using appropriate join keys (e.g., customer_id, product_id). This consolidation simplified downstream analysis and ensured a holistic view of transactions.

- Net Revenue Calculation: A new column, revenue, was created to represent net revenue per order line, calculated as: revenue = quantity × price (1 − discount_applied),,where discount is present .

- Price Band Categorization
  The product price column was used to segment products into three price bands:
  - Low
  - Medium
  - High

- Time-to-Market Metric : A column called days_to_order was created by calculating the number of days between a product's launch_date and the order_date. This helps evaluate product maturity and its effect on sales performance.

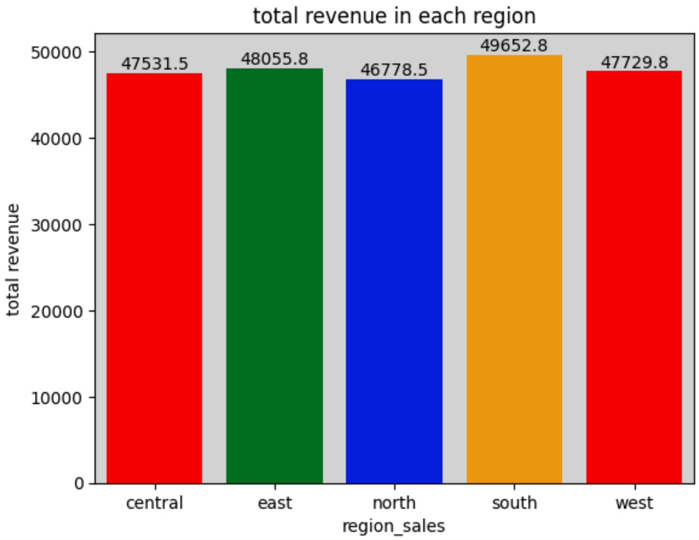
- Email Domain Extraction : From the customer email address, we extracted the domain portion (e.g., gmail.com, yahoo.com) into a separate column. This feature can be used for marketing segmentation and email performance analysis.

- Delivery Delay Indicator : A boolean column, is_delayed, was added to flag whether a given order experienced a delivery delay. This simplifies aggregation and filtering when analyzing logistical performance.

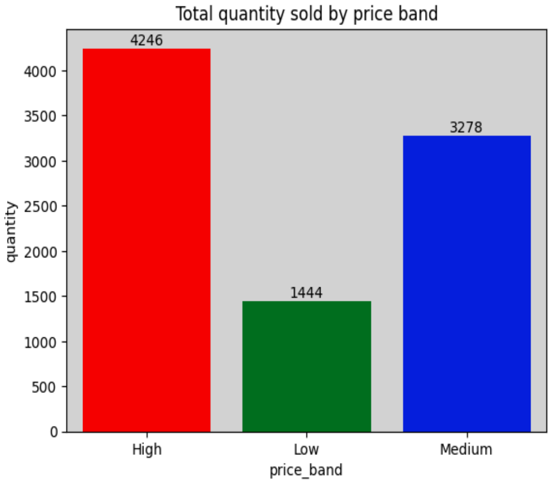
  Since all sales occurred on the same date (2025-06-07), extracting features like day-of-week or week-of-quarter was deemed unnecessary.

Key findings & Trends:

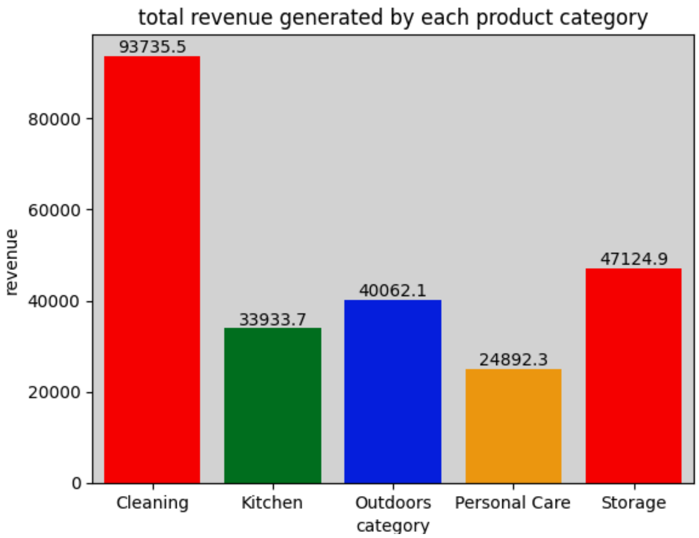
- The South region generated the highest revenue among all regions, with a total of $49,652.00 during the analysis period.

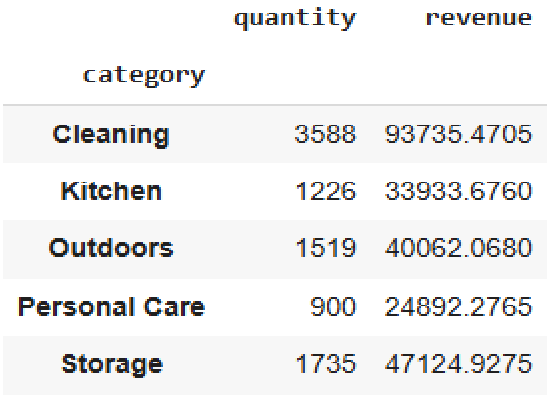


- “High” priced products had the highest quantity sold with a total of 4246 units.



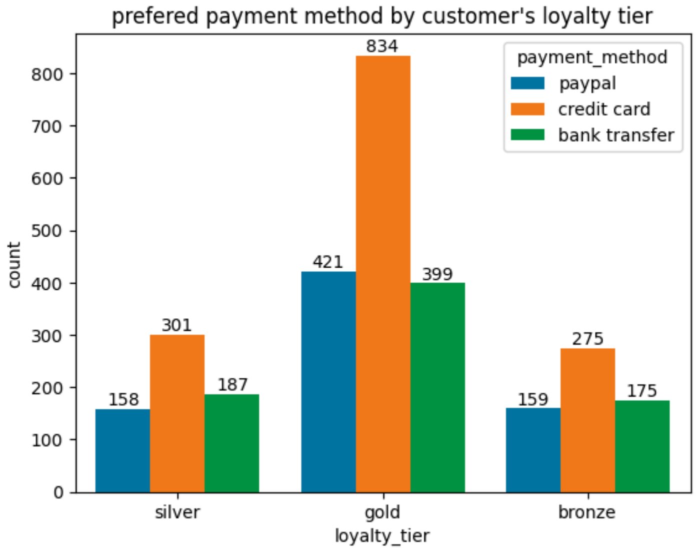
- The Cleaning Products category generated the highest revenue, totaling over $93,537.50 and 3588 units sold during the analysis period. In contrast, the Personal Care category contributed the least, with a total revenue of only $24,892.30 and only 900 units sold.





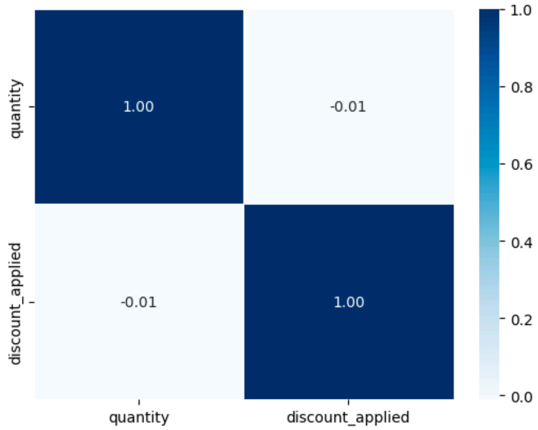
	quantity	revenue
category		
Cleaning	3588	93735.4705
Kitchen	1226	33933.6760
Outdoors	1519	40062.0680
Personal Care	900	24892.2765
Storage	1735	47124.9275

- Credit Card is the preferred payment method among customers



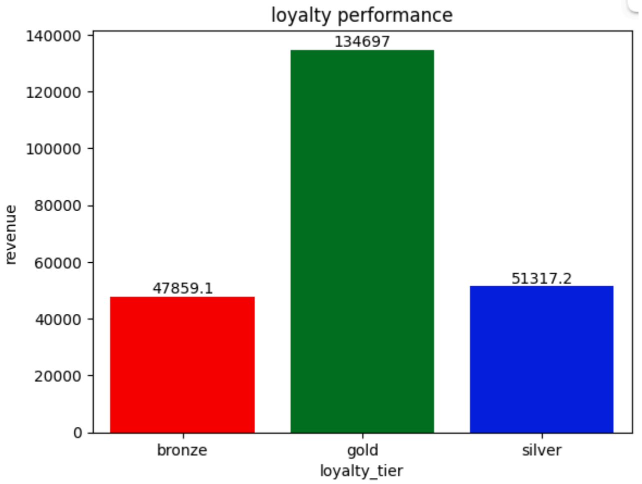
Answering business Questions:

- The Cleaning Products category generated the highest revenue, totaling over $93,537.50 and 3588 units sold during the analysis period. In contrast, the Personal Care category contributed the least, with a total revenue of only $24,892.30 and only 900 units sold.
- The correlation between discount applied and quantity sold is -0.01, indicating virtually no linear relationship between the two variables.This suggests that changes in discount levels do not significantly affect the quantity purchased.

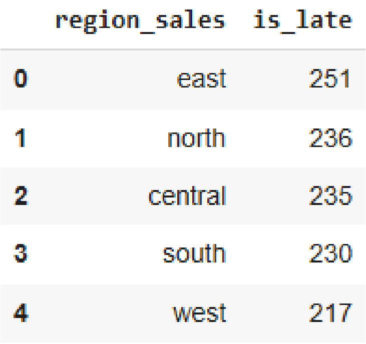


- Gold-tier customers generated the highest overall revenue, contributing $134,697 during the analysis period.

  This highlights the strong value of high-tier loyalty members and underscores the importance of retention strategies and exclusive offers to further engage this segment.

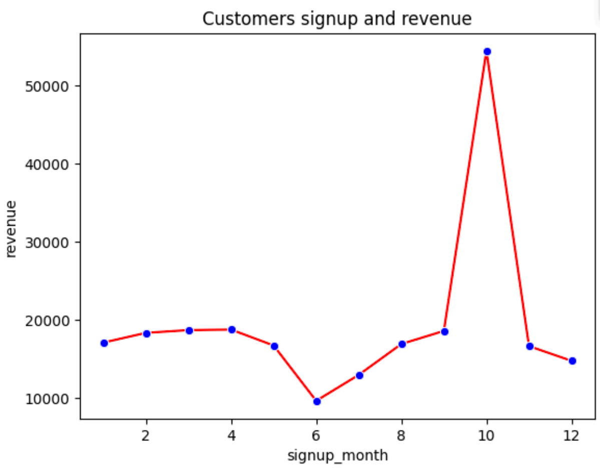


- The East region recorded the highest incidence of delivery delays compared to other regions.



	region_sales	is_late
0	east	251
1	north	236
2	central	235
3	south	230
4	west	217

- There is an influence of signup date on customer purchasing behavior, especially in October,which is an outlier.



Recommendations:

- Given that Cleaning Products are the top-performing category in terms of revenue — particularly within the High price band — we recommend focusing upcoming marketing campaigns on these premium items,while reassessing the positioning and pricing of Personal Care items to improve their sales performance.
- The South region leads in revenue generation. Consider targeted marketing campaigns, localized promotions, and inventory optimization in this region to capitalize on its strong performance.
- Investigate the root causes of frequent delivery delays in the East and take corrective action—such as optimizing courier selection, reviewing warehouse logistics, or improving delivery scheduling.

- With Gold customers contributing the most revenue, enhance retention strategies by offering exclusive perks, early product access, or tailored loyalty incentives to increase customer lifetime value.
- Address data integrity issues such as illogical signup dates and duplicate order IDs by implementing validation checks during data entry and enhancing ETL (Extract, Transform, Load) processes.