Q2 Performance Insights: Product Sales, Customer Loyalty, and Regional Trends

Table of Contents

Introduction

This report investigates sales and customer behavior across regions and product lines during the second quarter (Q2) of the business year at Green Cart Ltd. The analysis aims to uncover performance patterns, operational inefficiencies, and customer engagement trends by addressing the following key questions:

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

Three datasets were used in this investigation: customer data, sales transaction records, and product catalog information. All datasets were provided by the Data & Insights team at Green Cart Ltd. The analysis was carried out using Python (Pandas, Matplotlib, and Seaborn), with supporting visualizations and statistical summaries to assist interpretation.

Data Cleaning Summary

Customers Data

The customer dataset presented two main issues: missing values and inconsistent formatting.

1.Text Standardization
Several categorical columns contained inconsistencies in spelling. These were standardized to ensure consistency across analyses:

- **Gender column**: Values like "FEMALE" and "femle" were corrected to "female".
- **Loyalty Tier column**: Misspellings and inconsistent capitalizations such as "GOLD", "gld", "sllver", and "brnze" were corrected to "gold", "silver", and "bronze" respectively. The strip() method was also applied to remove any leading or trailing whitespace.

2.Handling Missing Values
Different techniques were applied based on the variable type and its importance to the analysis:

- **Email column**: Missing values were filled with the placeholder "unknown" to retain records while clearly identifying unavailable data.
- **Region column**: Missing values were imputed using the mode (most frequent value), assuming it reflects the most likely assignment.
- **Gender column**: Missing values were imputed using the mode to ensure consistency in demographic analyses.
- **Loyalty Tier column**: Missing values were imputed using the mode to avoid bias.
- **Customer ID**: Observations with missing customer IDs were dropped, as this field serves as the primary key and is essential for merging datasets.

Product Data

The product dataset did not require any data cleaning

Sales Data

The Sales dataset presented two main issues: missing values and inconsistent formatting similar to the customer dataset.

1.Text Standardization
Several categorical columns contained inconsistencies in spelling. These were standardized to ensure consistency across analyses:

- **Delivery status column**:
    - " delrd " → " delivered "
    - " delyd " → " delayed "
- **Payment method column**:
    - " bank transfr" → " Bank Transfer "
    - " credit card " → " Credit Card "
- **Region column**:

    - " nrth " → " north "
- **Quantity column**:
    - " three" → "3"
    - " five" → "5"
- **2.Handling Missing Values**
  Different techniques were applied based on the variable type and its importance to the analysis:
- **Unit Price column**: Missing values were filled with the mean (28.95) to retain unbiasedness.
- **Discount Applied column**: Missing values were imputed using the mean (0.10).
- **Delivery status, Payment method, and Product ID columns**: Missing values were filled with “unknown”.
- **Quantity column**: Missing values were imputed using the mode to avoid bias.

Feature Engineering Summary

Merging Datasets: The datasets were merged using a left join to ensure all entries from the sales dataset were preserved, prioritizing sales data.

- **Revenue Column**: A new column named revenue was created by multiplying the values in the quantity, unit_price, and discount_applied columns, representing the actual revenue after discounts.
- **Price Band Column**: A price_band column was created by categorizing the unit_price into three bands:
    - **Low**: Prices less than 15
    - **Medium**: Prices between 15 and 30
    - **High**: Prices greater than 30
- **Days to Order Column**: A days_to_order column was created to calculate the number of days between the product_launch_date and the order_date, measuring the time taken to place the order after launch.

- **Email Domain Column**: A new email_domain column was created by extracting the domain part from the email addresses in the customer data, allowing for domain-level analysis.
- **Is Late Column**: An is_late column was created to identify whether an order was delivered or delayed, providing insight into delivery performance.

Key Findings & Trends

Signup Month	Month Name	Revenue
1	January	17,378.6
2	February	23,209.3
3	March	19,461.4
4	April	15,030.9
5	May	15,666.0
6	June	16,718.2
7	July	19,179.7
8	August	23,477.7
9	September	18,333.3
10	October	22,036.2
11	November	22,193.7
12	December	19,864.3

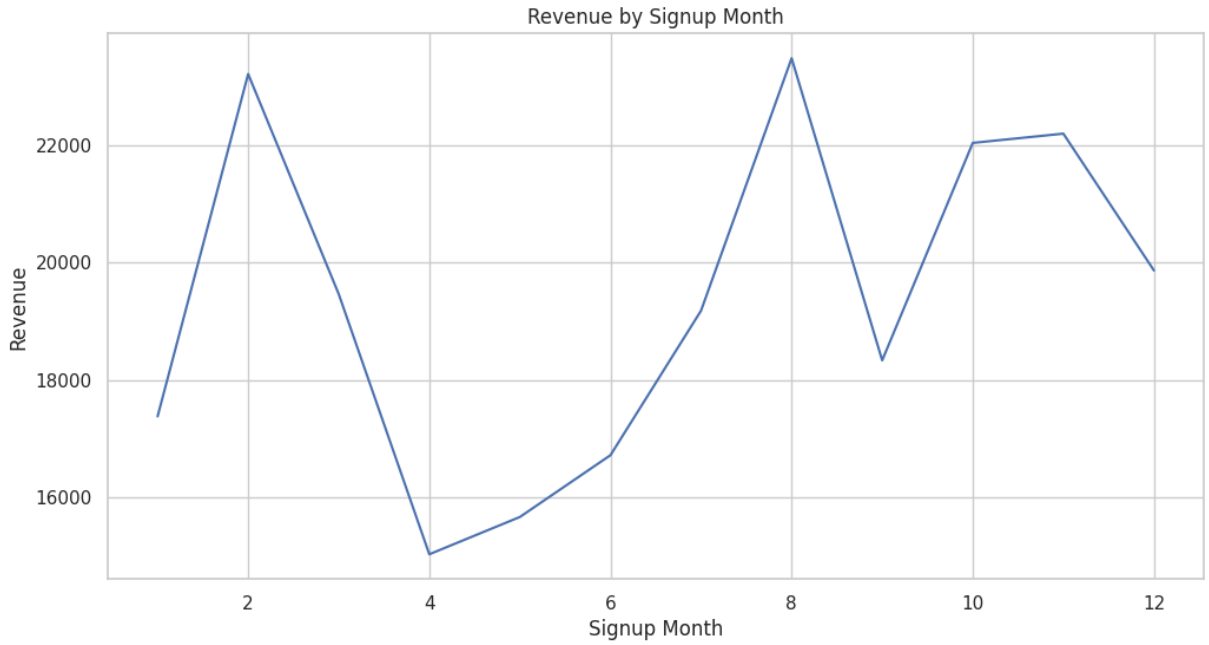

Figure 1: line plot of Revenue by Signup Month

Both the line plot and the table of total revenue by signup month reveals two significant peaks in February and August, indicating higher revenue from customers who signed up during these months. In contrast, April had the lowest revenue, marking it as the month with the least contribution from new signups.

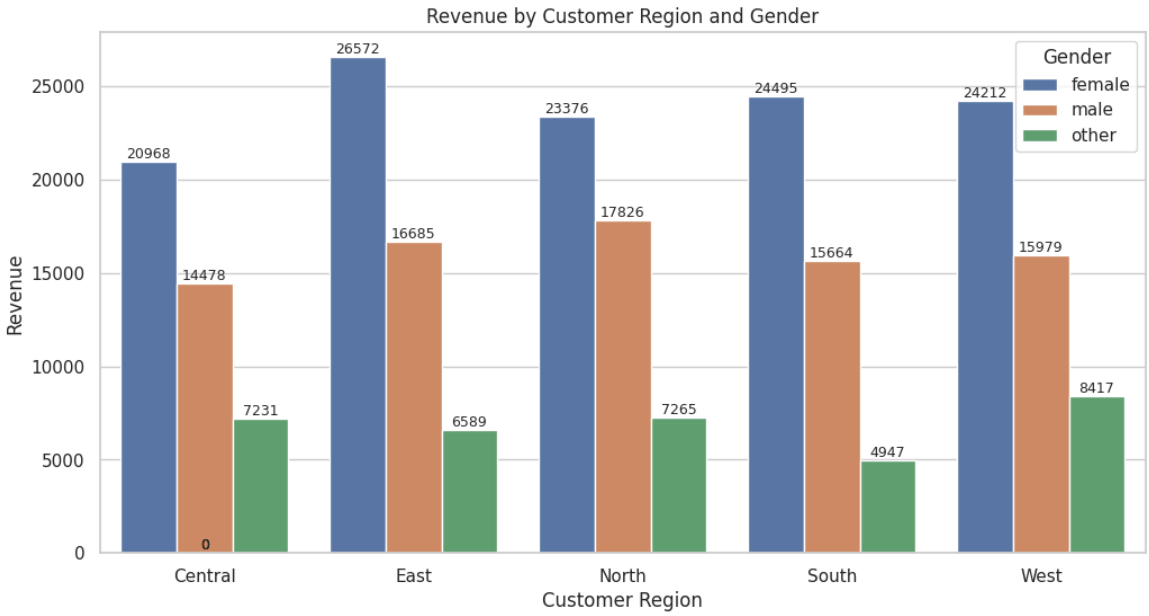

Figure 2: Bar plot of Revenue by customer region and gender

Female customers generated the highest revenue across all regions, followed by male customers and then others. The highest revenue for female customers was observed in the East region, while for male customers it was in the North region, and for the “other” category, the West region contributed the most. The Central region recorded the lowest revenue for both female and male customers, whereas the South region had the lowest revenue for the “other” category.

Business Question Answers

Which product categories drive the most revenue, and in which regions?

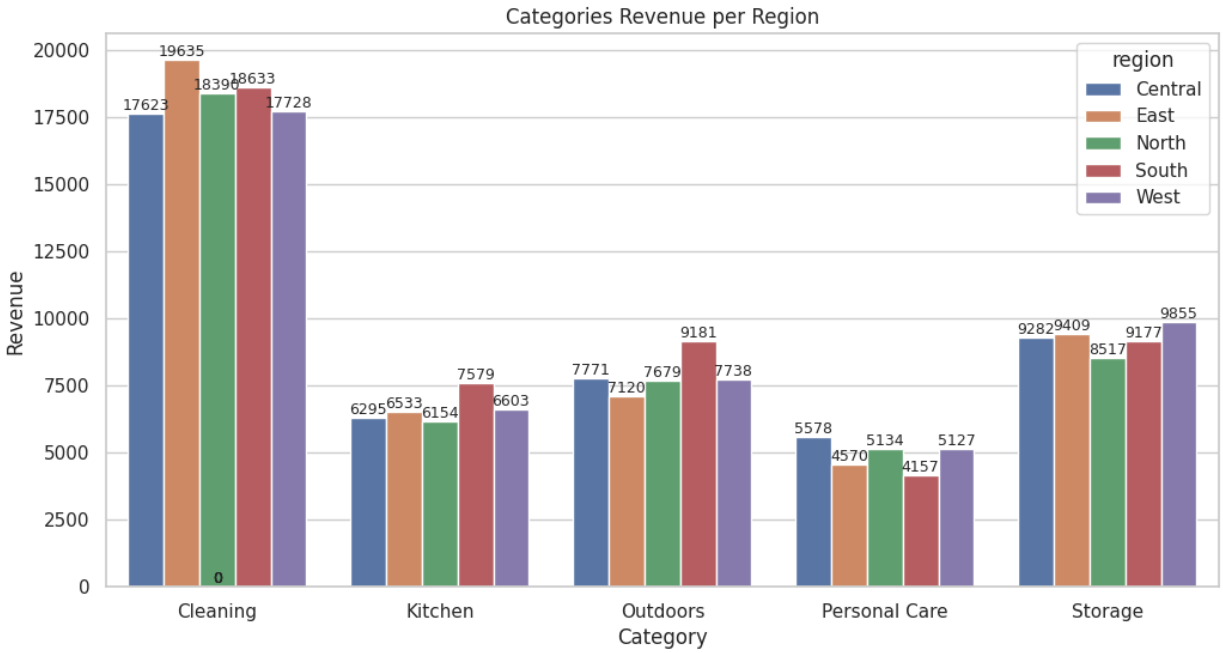

Figure 3: Bar plot of Categories Revenue Per Region

Based on Figure 3, cleaning products generated the highest sales revenue across all regions. Their revenue was approximately twice as high as that of the next highest category (Storage), indicating a strong and consistent demand for cleaning items regardless of region.

On the other hand, personal care generated the least sales revenue across all regions.

Do discounts lead to more items being sold?

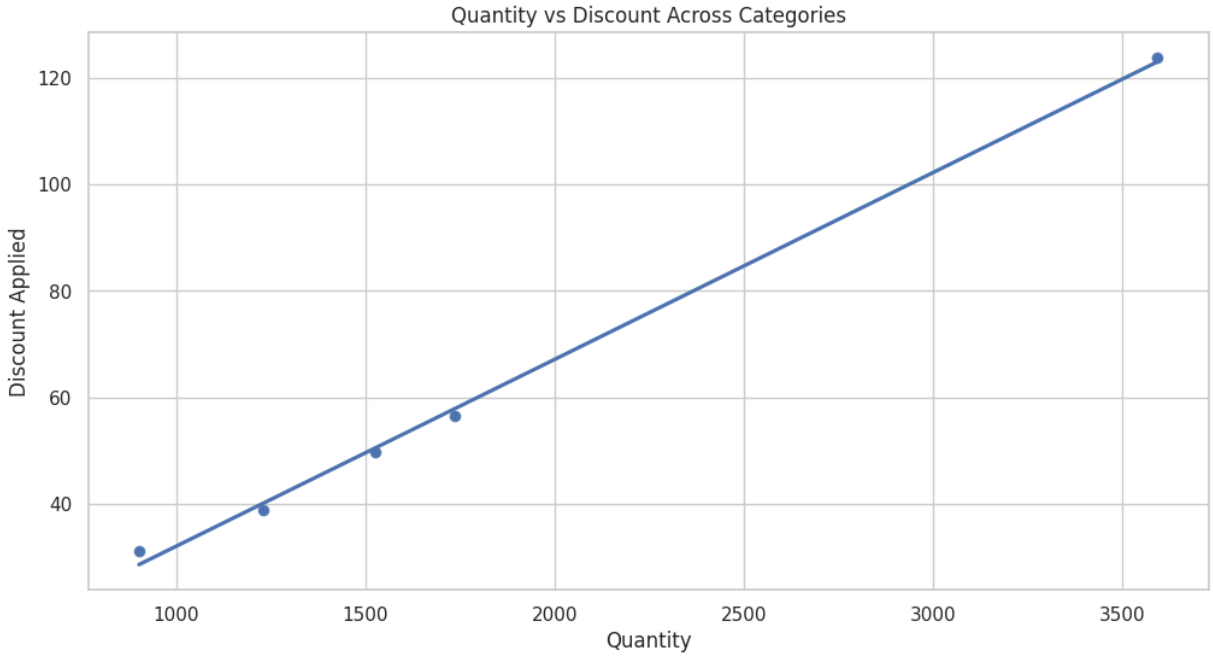

Figure 4: regression plot of Quantity to Discount Applied

As shown in Figure 4, there is a clear trend indicating that higher discounts are positively associated with increased quantities sold across all product categories. This suggests that applying discounts can be an effective strategy to boost sales volume

Which loyalty tier generates the most value?

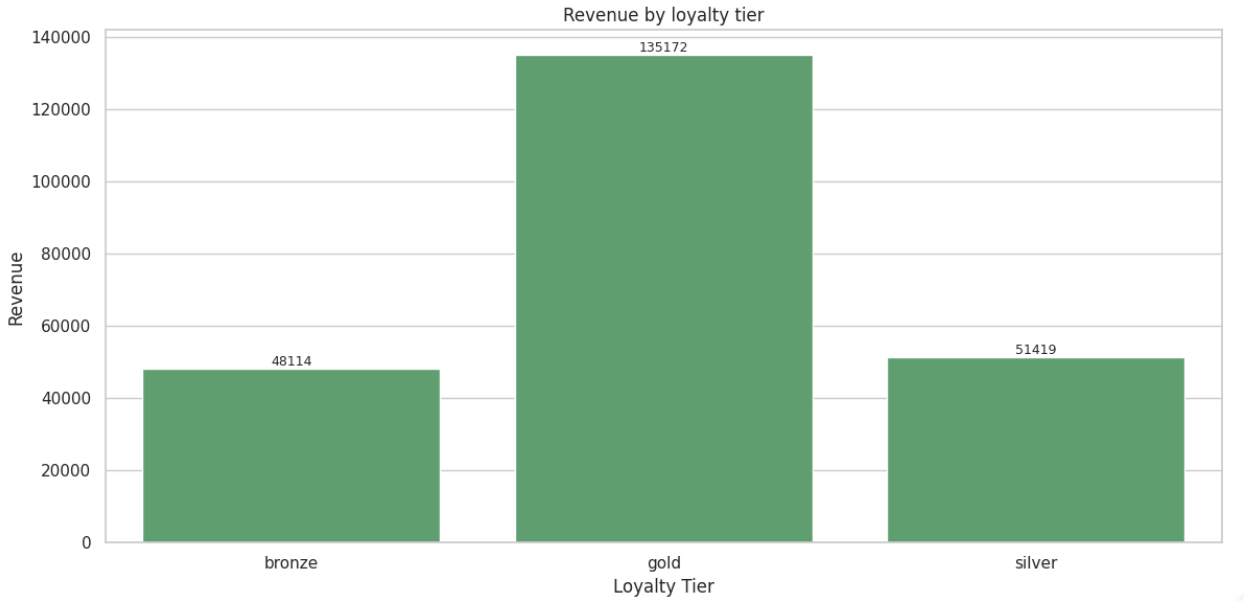

Figure 5: bar plot of Revenue by loyalty tier

The bar chart above shows that the gold loyalty tier generated the highest revenue during the second quarter.

Are certain regions struggling with delivery delays?

Figure 6: Bar plot of Delivery Status Per Region

While Figure 6 does not indicate a specific region struggling with delivery delays, it is noticeable that the East region has a higher number of delayed orders compared to other regions.

Do customer signup patterns influence purchasing activity?

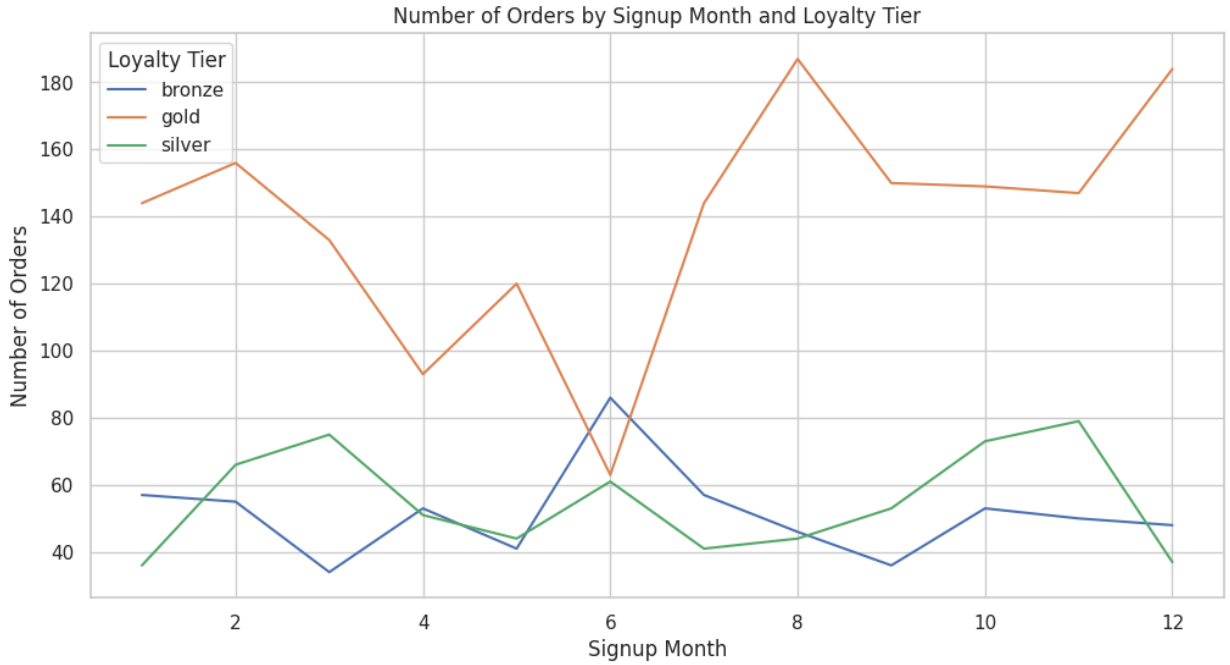

Figure 7: Line plot of the number of orders by signup month and loyalty tier

The line plot shows that customer signup patterns, reflected by both loyalty tier membership and signup month, influence purchasing activity. Specifically, customers in the gold loyalty tier generally make more orders compared to those in the silver and bronze tiers, who exhibit similar ordering behavior. However, customers who signed up in June tend to place approximately the same number of orders regardless of their loyalty tier.

Recommendations

1.Strengthen Engagement During Peak Signup Months (February & August):

Given the high revenue from customers who signed up in February and August, we should consider aligning major promotions, referral incentives, and onboarding campaigns with these months to capitalize on naturally higher engagement from the customers.

2.Target Female Customers in the East Region with Personalized Offers:

Since female customers in the East region generate the highest revenue overall, the business should focus targeted marketing campaigns and loyalty programs on this segment to boost revenue and encourage repeat purchases.

Data Issues or Risks

One of the key data quality problems identified was the presence of missing values across several fields. Missing data can affect the accuracy of analysis and limit the ability to draw reliable conclusions.

Suggested Fix:
To reduce missing values in future datasets, it is recommended to store the data in a structured SQL database and to enforce the following measure:

Enforce NOT NULL constraints on essential fields such as region, age, and plan to prevent incomplete entries

Implementing this constraint at the database level will help ensure higher data integrity and reduce the amount of cleaning needed in future analyses.