Green Cart Ltd – Sales & Customer Behaviour Analysis

1. Introduction

Green Cart Ltd is a UK-based e-commerce company specialising in eco-friendly household products. As part of the Data & Insights team, I was tasked with supporting the Q2 performance review by analysing sales performance and customer behaviour across different regions and product lines.

The analysis used three datasets:

- **Sales data** containing transaction-level order details
- **Product information** describing product categories, pricing, and launch dates
- **Customer information** including demographics, signup details, and loyalty tiers

The objective of this report is to:

- Identify **revenue trends** across time, regions, and product categories
- Understand **customer behaviour**, particularly by loyalty tier and signup patterns
- Evaluate **delivery performance**, including delays by region and product price band

2. Data Cleaning Summary

Before analysis, each dataset was cleaned individually to ensure accuracy and consistency.

Text Standardisation

Several categorical fields contained inconsistent formatting (e.g. casing and whitespace). These were standardised using string operations:

- `delivery_status` (e.g. " delayed " → "Delayed")
- `loyalty_tier` (e.g. "gold " → "Gold") and fixing inconsistant data
- `region`, `payment_method`, and `category`

Standardising these fields ensured reliable grouping and aggregation during analysis.

Date Conversion

All date-related columns were converted to datetime format:

- `order_date`
- `signup_date`
- `launch_date`

Invalid or malformed dates were coerced to null values to avoid calculation errors in time-based features.

Missing Values

Missing values were handled based on business logic:

- `discount_applied` was filled with **0.0**, assuming no discount when not recorded
- Missing product categories were labelled as **"Unknown"**
- Missing loyalty tiers were defaulted to **"Bronze"**, reflecting the base customer tier

Only non-critical fields were imputed. Records were dropped **only when necessary**, such as invalid numeric values.

Duplicate Records

Duplicate records were identified and removed:

- Sales data was deduplicated using `order_id`
- Product and customer data were deduplicated using their respective IDs

This prevented double-counting revenue and orders.

Numeric Validation

Key numeric fields (`quantity`, `unit_price`, `discount_applied`) were validated to ensure:

- No negative values
- Logical consistency for revenue calculations

Invalid rows were filtered out to maintain data integrity.

3. Feature Engineering Summary

To support deeper analysis and answer business questions, several new features were created:

- **Revenue**
  Calculated as:
  `quantity × unit_price × (1 − discount_applied)`
  This represents the true net revenue per order line.
- **Order Week**
  Extracted as the ISO week number from `order_date`, enabling weekly trend analysis.
- **Price Band**
  Unit prices were categorised into:
  - Low (< £15)
  - Medium (£15–£30)

    - High (> £30)
      This allowed comparison of customer behaviour and delivery performance across pricing tiers.
- **Days to Order**
  Calculated as the number of days between product launch and order date, providing insight into product lifecycle performance.
- **Email Domain**
  Extracted from customer email addresses to support future segmentation analysis.
- **Is Late**
  A boolean flag identifying delayed deliveries (`True` if delivery status = "Delayed"), used for delivery performance analysis.

These engineered features enabled clearer insights into time-based trends, customer value, and operational efficiency.

4. Insight Findings & Trends

1. Revenue Concentration by Category and Region

Revenue was heavily concentrated in a small number of high-performing product categories, particularly in **London and the South East**. While lower-priced products generated higher order volumes, **high-priced items contributed disproportionately to total revenue**.

Supporting evidence:

- Category revenue summary tables
- Weekly revenue trends by region line chart

2. Discounts Increase Volume but Not Always Revenue

Discounted orders showed a **moderate increase in quantity sold**, particularly for low- and medium-priced products. However, the correlation between discounts and revenue was weaker, indicating that higher discounts do not always translate into higher overall revenue.

Supporting evidence:

- Quantity vs discount boxplots
- Correlation heatmap between revenue, quantity, and discount

3. Delivery Delays Are Price and Region Dependent

Delivery delays were more common for **high-priced products** and were disproportionately higher in **Northern regions** compared to London and the South East. This suggests potential supply chain or logistics challenges affecting premium products and certain regions.

Supporting evidence:

- Delivery delay rates by region and price band
- Stacked bar chart of delivery status by price band

Business Question Answers

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

High-value product categories generate the most revenue, particularly in **London and the South East**. These regions consistently outperform others, indicating stronger demand and higher purchasing power.

2. Do discounts lead to more items sold?

Yes, discounts generally lead to **higher quantities sold**, especially for lower-priced items. However, the increase in volume does not always offset the reduction in price, meaning discounts should be applied strategically.

3. Which loyalty tier generates the most value?

Gold-tier customers generate the highest total and average revenue per customer. While Bronze customers place more orders overall, their average order value is significantly lower.

4. Are certain regions struggling with delivery delays?

Yes. **Northern regions** experience higher delivery delay rates, particularly for **high-priced products**. This highlights an opportunity to improve logistics performance in these areas.

5. Do customer signup patterns influence purchasing behaviour?

Yes. Customers who signed up earlier tend to have **higher lifetime value** and generate more revenue over time. More recent signups purchase more frequently but typically place smaller-value orders.

Conclusion

This analysis highlights clear opportunities for Green Cart Ltd to:

- Focus marketing efforts on **high-value categories in high-performing regions**
- Use **discounts selectively**, targeting volume growth without eroding revenue
- Improve **delivery performance** for premium products and underperforming regions
- Strengthen retention strategies for **high-value loyalty tiers**