Full Name	Zahra Hayati
Project Title	Week 2: Sales & Customer Behaviour Insights

1. Introduction

Green Cart Ltd., a UK-based e-commerce company specialising in eco-friendly household products, aims to evaluate Q2 sales performance, customer behaviour, and delivery efficiency. This report analyses sales, product, and customer datasets to identify trends, highlight high-performing product categories, evaluate customer loyalty, and assess delivery performance. The insights will support marketing strategies, operational improvements, and targeted promotions for growth.

2. Data Cleaning Summary

The datasets required extensive cleaning before analysis:

1. Text Standardisation
    - delivery_status, Delivered / Delayed / Cancelled / Other
    - loyalty_tier, Gold / Silver / Bronze / Other
    - region, consistent capitalization (e.g., "North")
    - payment_method, consistent titles (e.g., "Credit Card", "Bank Transfer")
2. Missing Values
    - discount_applied, missing values set to 0.0 with imputation flag
    - quantity, missing values replaced with 1 and flagged
    - email_domain, missing values set to "unknown"
    - Missing product or customer IDs were filled as "UNKNOWN"
3. Date Parsing
    - Converted order_date, launch_date, signup_date to datetime format with error coercion.
4. Duplicate
    - Duplicate order lines and customer records flagged and resolved.
5. Numeric Validation
    - quantity, unit_price, discount_applied verified as numeric and non-negative

3. Feature Engineering Summary

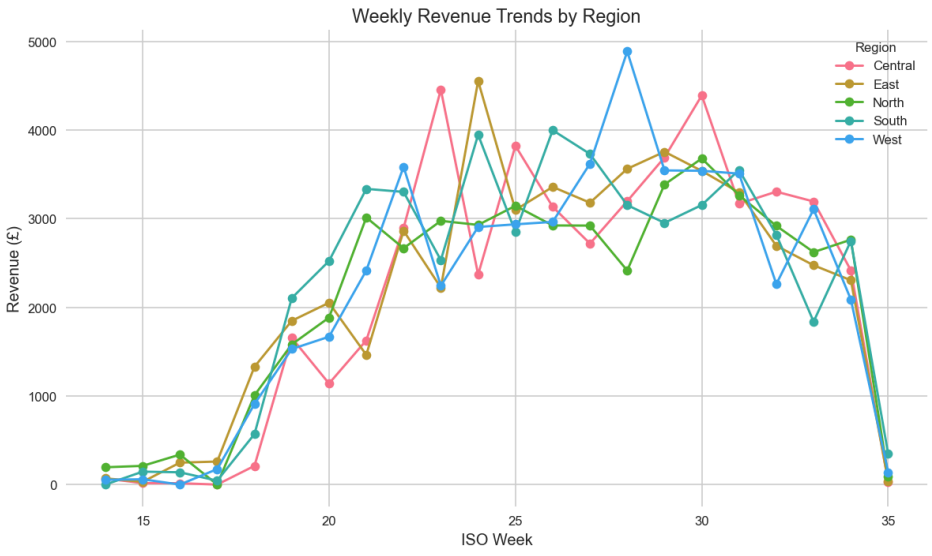
- revenue, Calculates the actual value of each order after applying discounts, serving as a key performance metric for sales analysis.

- order_week, ISO week extracted from order_date. Facilitates trend analysis over time and helps identify weekly sales patterns or seasonal fluctuations.
- price_band, Categorises products into Low (<£15), Medium (£15–30), and High (>£30) price ranges. Enables segmentation and comparison of sales performance across price tiers.
- days_to_order = Difference between order_date and launch_date. Measures the product's adoption speed and lifecycle effectiveness.
- email_domain = Extracted domain from customer email addresses. Useful for analysing customer source types or detecting organisational purchase behaviour.
- is_late = Boolean flag set to True if delivery_status = "Delayed". Assesses delivery reliability and supports logistic performance evaluation.

4. Key Findings & Trends

1. Weekly Revenue Trends by Region

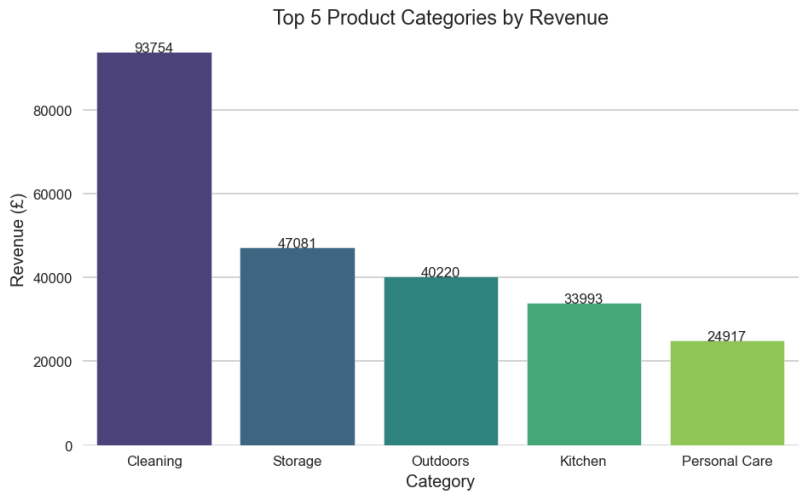
- The North and South regions show high and generally consistent revenue, especially after Week 20.
- The West region demonstrates a late-stage surge, reaching the highest single-week revenue around Week 28.
- The overall trend shows a sharp ramp-up in revenue starting around Week 19, peaking between Weeks 23-31, and then slightly declining toward the final weeks.



2. Top Product Categories

- The top two categories, Cleaning ($93,754.26) and Storage ($47,081.34), significantly outperform Outdoors ($40,220.50) and Kitchen ($33,993.04), confirming your key finding.

- Cleaning stands out as the dominant category, generating nearly double the revenue of the next highest category, Storage.

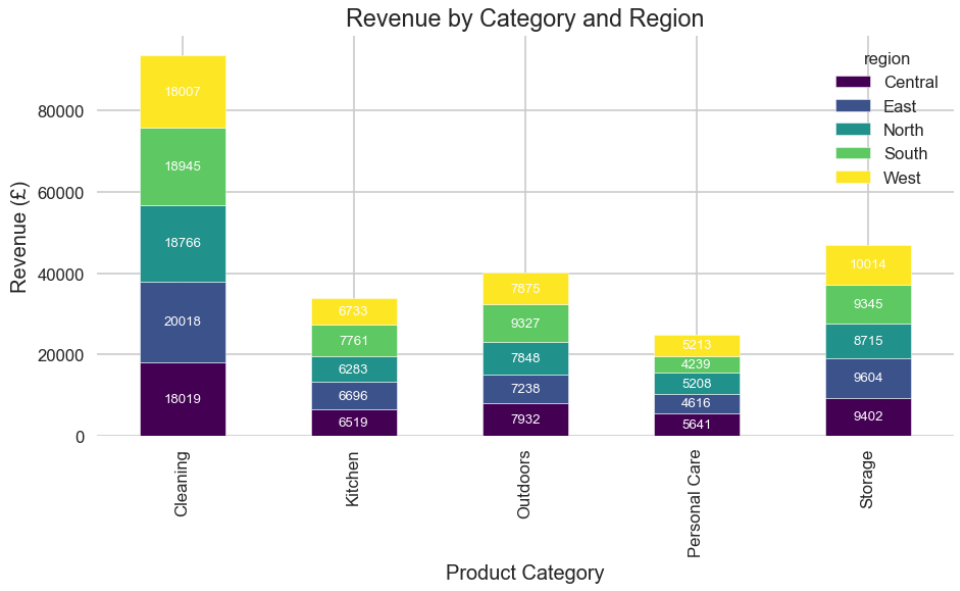


3. Quantity Distribution Across Categories
   - The mean quantity for all categories is very consistent, hovering right around 3 units (≈2.98 to 3.05), supporting the finding that the median is similar across categories.
4. Customer Behaviour by Loyalty Tier
   - The Gold tier represents the highest volume of orders in every region.
   - The total order count is relatively similar across all regions (around 580 to 600 orders per region).
   - The Bronze tier has a lower total order count than Gold or Silver in all regions, which might seem to contradict the finding of "highest order count in low-value segments." However, this chart represents the count of unique customers, not the total order count.

5. Delivery Performance
    - Delayed and Delivered are the two dominant statuses across all price bands.
    - The absolute count of Delayed orders is highest in the High-Price band (545), followed by Medium (441), which supports the finding that delayed orders are concentrated in the medium/high price bands.
6. Correlation Analysis
    - The correlation is -0.12, confirming the weak negative correlation, meaning that as discounts increase, revenue slightly tends to decrease (or vice-versa).
    - A strong positive correlation of 0.72 is observed, which is expected, higher quantity orders generally lead to higher revenue.
    - The correlation is near zero at -0.007, indicating that the discount level has almost no relationship with the number of items purchased in an order, which supports the narrative of quantity being consistent across categories regardless of discount.

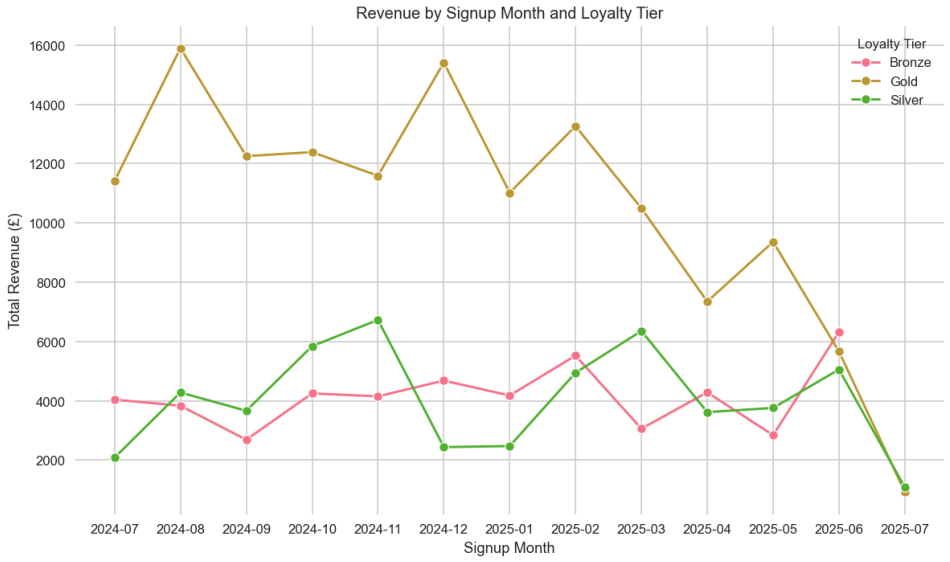
5. Business Question Answers

1. **Which product categories drive the most revenue, and in which regions?**
   The Cleaning category drives the highest total revenue (£93,754), followed by Storage (£47,081) and Outdoors (£40,220).
   Regionally, revenue trends show that the South and East regions consistently outperform others across weeks 20–35, suggesting strong sales concentration there.



2. **Do discounts lead to more items sold?**
   The correlation between discount and quantity is -0.0074, which is effectively zero. This suggests no meaningful relationship, offering discounts did not lead to higher quantities sold.
   Sales volumes are consistent (mostly 2–4 units per order) across all categories regardless of discount.

3. **Which loyalty tier generates the most value?**
   The Gold tier generates the most orders across all regions (300–350 orders per region). This tier likely represents your most valuable customer group in both frequency and revenue contribution.
   Bronze and Silver segments lag significantly behind in activity.
4. **Are certain regions struggling with delivery delays?**
   Certain regions show higher delivery struggles. The East region consistently shows the highest percentage of delayed orders overall. Furthermore, this struggle is concentrated in the Medium and High-Price Band orders across most regions, pointing to a bottleneck in handling logistics for higher-value shipments.
5. **Do customer signup patterns influence purchasing activity?**
   Customer signup timing does appear to influence purchasing activity.
   Customers who joined earlier (mid to late 2024) generated higher revenues than those who joined in 2025, suggesting that newer signups are less active or have not yet built strong purchasing habits.



6. Recommendations

Based on the analysis of sales, customer behaviour, and delivery performance, the following strategic recommendations are proposed:

1. **Revise Discount Strategy and Focus on Gold Tier**

   - Action: Immediately cease deep, general discounts aimed at increasing cart quantity, as the correlation is near zero.
   - Focus: Instead, implement value-add bundling strategies (e.g., "Buy one, get complimentary product") to lift AOV organically. Design exclusive, high-value

promotions and early access programs specifically for the “Gold” loyalty tier to reward and retain this critical high-volume customer base.

2. **Address High-Value Delivery Bottlenecks**

- Action: Investigate the logistics route and partner serving the East region to resolve the identified local delay bottleneck.
- Focus: Introduce a Premium Shipping option for all High-Price Band (>£30) orders. This involves using a more reliable carrier for high-value shipments to reduce the high delay rate in this segment, improving customer satisfaction for your most valuable transactions.

7. Data Issues or Risks

1. Duplicate and Conflicting Records:
    - Critical data integrity violations were found where single order_id was linked to two or more different customer_id, order_date and region values.
    - Risk: An order_id should uniquely identify a single transaction by a single customer at a single point in time. This conflict means the core sales data is unreliable, making accurate attribution of revenue, delivery performance, and customer lifetime value (CLV) impossible for the affected orders. These conflicting orders could lead to misreported regional sales figures.
2. Text and Category Inconsistencies:
    - Categorical variables like region, delivery_status, and loyalty_tier included inconsistent or misspelled entries, requiring string standardisation.
    - Risk: Minor misclassifications could influence frequency counts or category-based plots.
3. Customer Data Completeness:
    - Some customer records were missing fields such as region, signup_date, or loyalty_tier, which were inferred where possible.
    - Risk: May bias loyalty or geographic segmentation result.