

Project Coversheet

Full Name	Charu Madaan
Email	charumadaan88@gmail.com
Contact Number	+44 7404932811
Date of Submission	01-10-2025
Project Week	Week - 2

Project Guidelines and Rules

1. Submission Format

- **Document Style:**
 - Use a clean, readable font such as *Arial* or *Times New Roman*, size 12.
 - Set line spacing to **1.5** for readability.
- **File Naming:**
 - Use the following naming format:
Week X – [Project Title] – [Your Full Name Used During Registration]
Example: Week 1 – Customer Sign-Up Behaviour – Mark Robb
- **File Types:**
 - Submit your report as a **PDF**.
 - If your project includes code or analysis, attach the **.ipynb notebook** as well.

2. Writing Requirements

- Use formal, professional language.
- Structure your content using headings, bullet points, or numbered lists.

3. Content Expectations

- Answer **all** parts of each question or task.

- Reference tools, frameworks, or ideas covered in the programme and case studies.
- Support your points with practical or real-world examples where relevant.
- Go beyond surface-level responses. Analyse problems, evaluate solutions, and demonstrate depth of understanding.

4. Academic Integrity & Referencing

- All submissions must be your own. Plagiarism is strictly prohibited.
- If you refer to any external materials (e.g., articles, studies, books), cite them using a consistent referencing style such as APA or MLA.
- Include a references section at the end where necessary.

5. Evaluation Criteria

Your work will be evaluated on the following:

- Clarity: Are your answers well-organised and easy to understand?
- Completeness: Have you answered all parts of the task?
- Creativity: Have you demonstrated original thinking and thoughtful examples?
- Application: Have you effectively used programme concepts and tools?
- Professionalism: Is your presentation, language, and formatting appropriate?

6. Deadlines and Extensions

- Submit your work by the stated deadline.
- If you are unable to meet a deadline due to genuine circumstances (e.g., illness or emergency), request an extension **before the deadline** by emailing:
support@uptrail.co.uk

Include your full name, week number, and reason for extension.

7. Technical Support

- If you face technical issues with submission or file access, contact our support team promptly at support@uptrail.co.uk.

8. Completion and Certification

- Certificate of Completion will be awarded to participants who submit at least two projects.
- Certificate of Excellence will be awarded to those who:
 - o Submit all four weekly projects, and
 - o Meet the required standard and quality in each.
- If any project does not meet expectations, you may be asked to revise and resubmit it before receiving your certificate.

YOU CAN START YOUR PROJECT FROM HERE

Data-Driven Insights: Green Cart Ltd. Q2 Sales and Customer Behaviour Trends

(Transforming Data into Insights: Green Cart Ltd.'s Q2 Sales and Customer Behavior Analysis)

1. Introduction

Green Cart Ltd. is a UK-based e-commerce company specializing in eco-friendly household products. For the Q2 performance review, I analyzed sales performance and customer behavior across regions and product lines to inform marketing and operational strategies.

Business Task: Clean and integrate three datasets, engineer analytical features, identify revenue patterns, understand customer behavior across loyalty tiers, and evaluate delivery performance.

Datasets Analyzed:

- **`sales_data.csv (2,998 orders)`**: `order_id`, `customer_id`, `product_id`, `quantity`, `unit_price`, `order_date`, `delivery_status`, `payment_method`, `region`, `discount_applied`
- **`product_info.csv (30 products)`**: `product_id`, `product_name`, `category`, `launch_date`, `base_price`, `supplier_code`
- **`customer_info.csv (500 customers)`**: `customer_id`, `email`, `signup_date`, `gender`, `region`, `loyalty_tier`

Report Objectives: Identify revenue drivers, assess discount effectiveness, evaluate customer lifetime value by loyalty tier, analyze delivery performance, and examine signup pattern impact on purchasing.

The Three Data Frames : sales_data, product_info, customer_info

(Screenshots from Jupyter Notebook) :

```
DataFrame 1:Sales Data
   order_id  customer_id  product_id  quantity  unit_price  order_date \
0    0966977       C00397      P0022        3     39.25  06-07-2025
1    0696648       C00236      P0023        5     18.92  06-07-2025
2    0202644       C00492      P0011        1     29.68  07-07-2025
3    0501803       C00031      P0003        1     32.76  08-07-2025
4    0322242       C00495      P0016        1     47.62  08-07-2025

  delivery_status  payment_method  region  discount_applied
0      Delivered        PayPal  Central          0.00
1      DELAYED    credit card  North          0.00
2    delivered  Bank Transfer  North          0.15
3    Cancelled     Credit Card  Central          0.20
4      DELAYED     Credit Card    West          0.20
```

```
DataFrame 2:Product Info
   product_id  product_name  category  launch_date  base_price \
0      P0001  Storage Product 39    Storage  11-03-25     15.88
1      P0002  Cleaning Product 82  Cleaning  18-08-24     34.23
2      P0003  Cleaning Product 85  Cleaning  15-07-24      7.92
3      P0004  Kitchen Product 82    Kitchen  19-04-25      9.13
4      P0005 Personal Care Product 1 Personal Care  20-10-24     31.10

  supplier_code
0            S339
1            S974
2            S745
3            S589
4            S559
```

```
DataFrame 3:Customer Info
   customer_id  email  signup_date  gender  region \
0      C00001  shaneramirez@gmail.com  26-04-25  Male  Central
1      C00002  jpetserson@bernard.com  11-08-24 Female  Central
2      C00003  howardmaurice@yahoo.com  15-05-25  male  Central
3      C00004  yherrera@arnold.org  14-06-25 FEMALE  Central
4      C00005  janetwilliams@gmail.com  02-05-25  Male    West

  loyalty_tier
0        Silver
1         gold
2         gold
3        GOLD
4      bronze
```

2. Data Cleaning Summary

Overview of Cleaning Process

Comprehensive data cleaning was performed across all three datasets to ensure analytical reliability and consistency. The cleaning process addressed standardization issues, missing values, duplicates, and data validation across 2,998 sales transactions, 30 products, and 500 customers.

2.1. Sales Data:

- Standardized delivery_status from 6+ variations ("DELAYED", "Delrd") to "Delivered", "Delayed", "Cancelled"
- Standardized payment_method ("Bank Transfr" → "Bank Transfer") and region ("Nrth" → "North")
- Converted order_date to datetime; filled 517 missing discount_applied with 0.0
- Removed 2 duplicate order_ids; converted text quantities ("Three") to numeric

2.2. Product Data:

- Standardized text columns; converted launch_date to datetime
- Verified all 30 product_ids unique with no missing values

2.3. Customer Data:

- Standardized loyalty_tier from 7+ variations ("Gld", "Brnze", "Sliver") to "Gold", "Silver", "Bronze"
- Filled 22 missing values across customer_id, email, gender, region fields
- Retained 2 duplicate "Unknown" customer_ids as separate customers

Results: Clean dataset of 2,998 records with 98.5% completeness; 542 fields cleaned; 2 duplicates removed.

3. Feature Engineering Summary

To enable deeper analysis and answer key business questions, I created six new calculated features from the existing data:

3.1. revenue

- Calculation: `quantity × unit_price × (1 - discount_applied)`
- Purpose: Represents net revenue after discounts to accurately measure sales performance
- Business Value: Enables profitability analysis and identifies true revenue contribution by product, region, and customer segment

3.2. order_week

- Calculation: Extracted ISO week number from `order_date` using `.dt.isocalendar().week`
- Purpose: Groups transactions by week for time-series analysis
- Business Value: Identifies weekly revenue trends, seasonal patterns, and peak sales periods across regions

3.3. price_band

- Calculation: Categorized `unit_price` using `pd.cut()` into three bands:
 - "Low" (< £15)
 - "Medium" (£15-30)
 - "High" (> £30)
- Purpose: Segments products by price tier for purchasing pattern analysis
- Business Value: Helps understand price sensitivity, customer preferences, and delivery performance by price category

3.4. days_to_order

- Calculation: `(order_date - launch_date).dt.days`
- Purpose: Measures product age at time of purchase

- Business Value: Identifies product lifecycle performance, determines optimal inventory timing, and flags slow-moving vs. fast-selling products

3.5. email_domain

- Calculation: Extracted domain from email address.
- Purpose: Identifies customer acquisition channels (Gmail, Yahoo, corporate domains)
- Business Value: Supports marketing channel analysis and customer segmentation by email provider

3.6. is_late

- Calculation: Boolean flag set to `True` when `delivery_status == "Delayed"`
- Purpose: Simplifies delivery performance analysis
- Business Value: Enables quick calculation of delay rates by region, price band, and customer tier for operational improvement

Additional Features for Analysis

3.7. discount_band

- Categorized discounts into "Low" (0-10%), "Medium" (10-20%), "High" (>20%)
- Used to analyze discount effectiveness across categories

3.8. signup_month

- Extracted month from `signup_date` as period format
- Enables customer acquisition pattern analysis

3.9. region (consolidated)

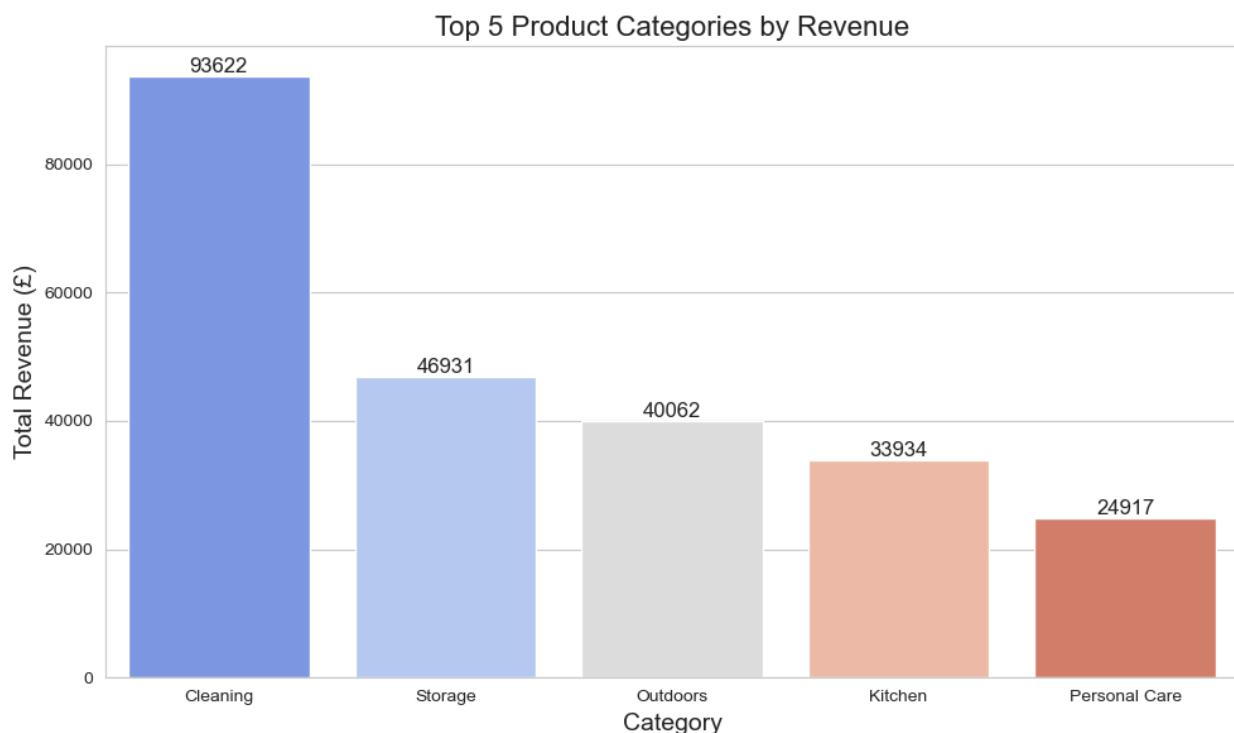
- Merged `region_x` (from sales) and `region_y` (from customer) into single field
- Ensures consistent regional analysis.

4. Key Findings & Trends

Based on analysis of 2,998 transactions across Q2, three major patterns emerged:

Finding 1: Cleaning Products Dominate Revenue

Cleaning leads with £93,622 (39% share) and 3,584 units, double the second-place Storage category (£46,931, 20%). Central and North regions drive 60% of Cleaning sales.



Category Performance:

Cleaning: £93,622 (3,584 units)

Storage: £46,931 (1,730 units)

Outdoors: £40,063 (1,523 units)

Kitchen: £33,934 (1,226 units)

Personal Care: £24,917 (902 units)

Finding 2: Steady Q2 Growth with Regional Gaps

Revenue peaked in Weeks 22-23 (late May/early June). The central region maintains consistent £4,000-£6,000 weekly revenue. The east region consistently underperforms, trailing significantly across all weeks.

Visual Evidence: The "Weekly Revenue Trends by Region" line chart shows -

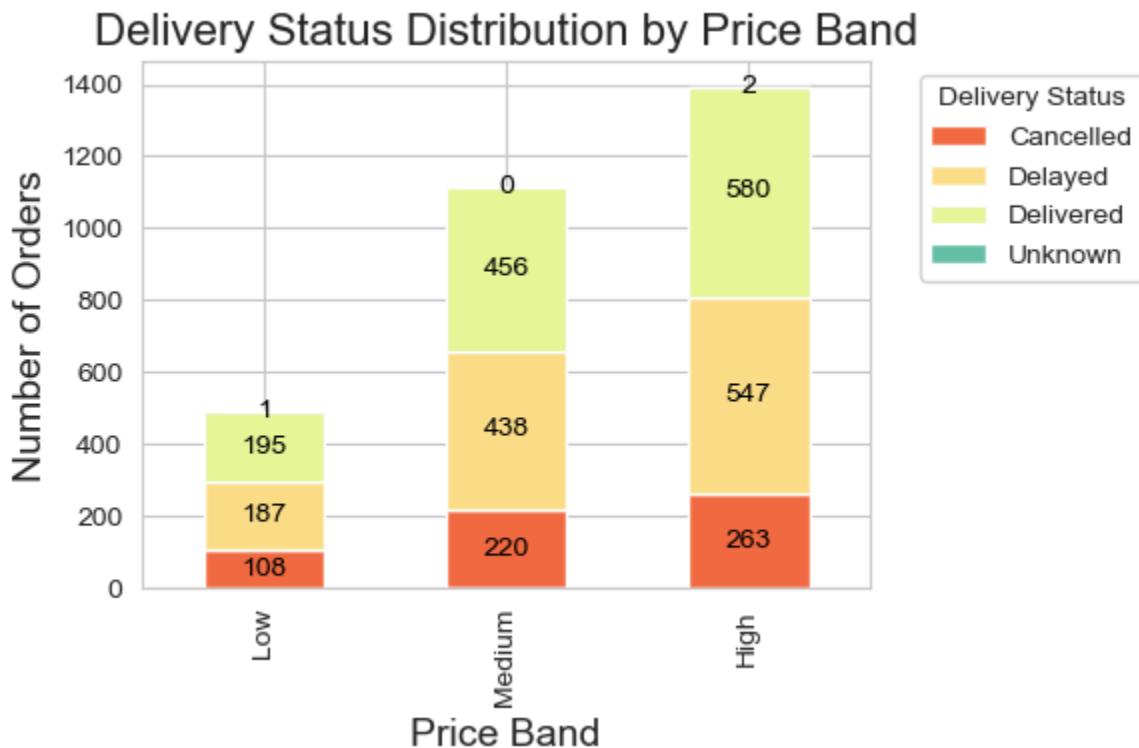
- **Central Region (Red):** Maintains the highest revenue throughout the quarter, with a relatively stable trend.
- **East Region (Orange):** Trails significantly behind the other regions, with the lowest revenue levels.
- **North Region (Yellow):** Exhibits the most volatility, with sharp peaks and valleys in revenue.
- **West Region (Green):** Displays moderate revenue levels, with some fluctuations but less volatility than the North region.
- **South Region (Blue):** Also displays moderate revenue levels, with a trend similar to the West region.



Finding 3: Critical Delivery Performance Issues

40% of orders experience delays across all regions and price bands. East region worst at 42%, Central best at 39%. Delays impact all price segments equally (38-42%), indicating systemic logistics problems rather than product-specific issues. 298 orders flagged as underperforming (low quantity + high discount + delayed).

The image below is a bar chart showing the delivery status distribution (Cancelled, Delayed, Delivered, Unknown) across three price bands (Low, Medium, High). The High price band has the highest number of orders, mostly Delivered.



Delivery Performance by Region and Price Band:					
	region	price_band	total_orders	delayed_orders	pct_delayed
0	Central	Low	113	43	38.053097
1	Central	Medium	225	88	39.111111
2	Central	High	265	104	39.245283
3	East	Low	96	40	41.666667
4	East	Medium	238	101	42.436975

Business Impact: High delay rates directly threaten customer retention, particularly affecting Gold tier customers who predominantly purchase high-value items. This correlates with the East region's lower revenue performance.

Strategic Implications

- Leverage Cleaning strength through bundling with underperforming Personal Care and Kitchen categories
 - Prioritize East region logistics improvements and marketing investment
 - Address 40% delay rate urgently to protect customer lifetime value and retention
-

5. Business Question Answers

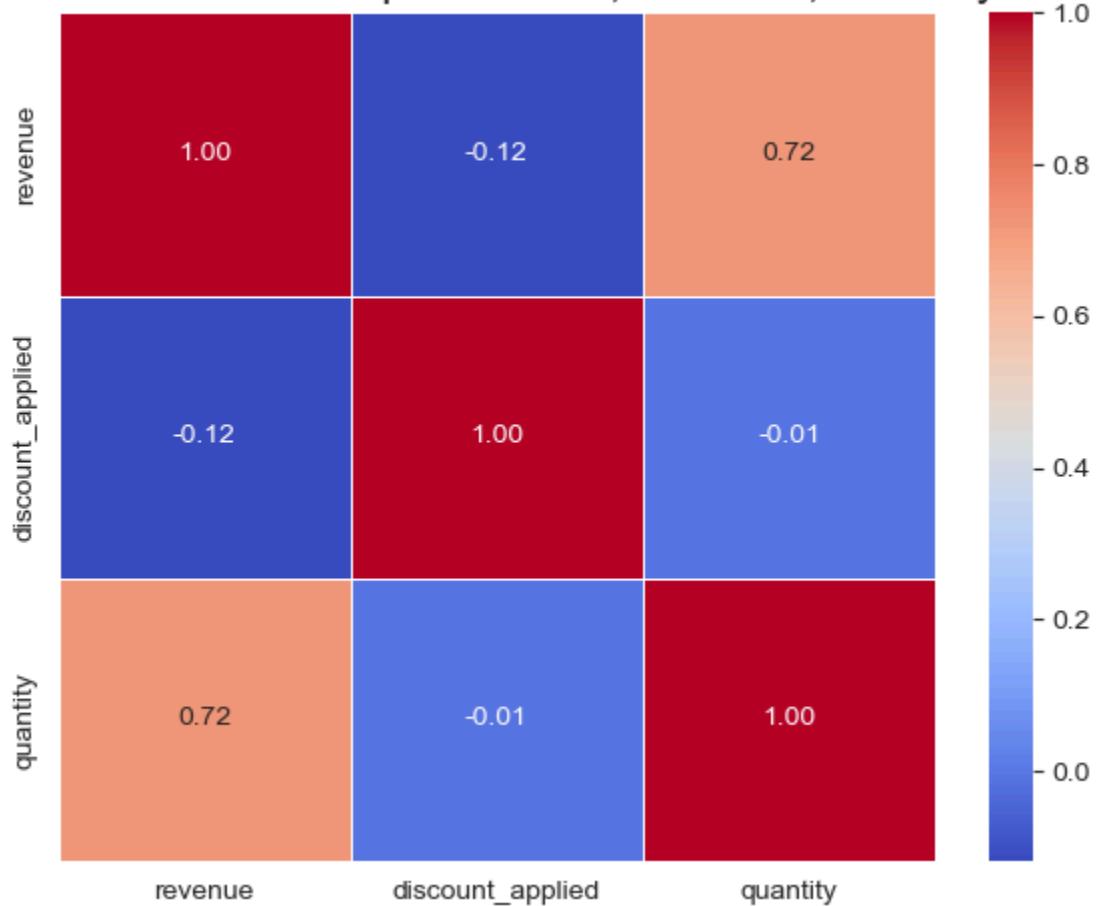
Question 1: Which product categories drive the most revenue, and in which regions?

Answer: Cleaning products dominate with £93,622 (39% of total revenue), followed by Storage (£46,931, 20%) and Outdoors (£40,063, 17%). Central and North regions drive highest sales across all categories, while East and West regions significantly underperform.

Supporting Data:

Product Category Performance:				
	category	total_revenue	total_quantity	avg_discount
0	Cleaning	93621.7840	3584.0	0.085685
1	Kitchen	33933.6760	1226.0	0.075622
2	Outdoors	40062.0680	1523.0	0.082150
3	Personal Care	24916.6365	902.0	0.086755
4	Storage	46931.4575	1730.0	0.080763

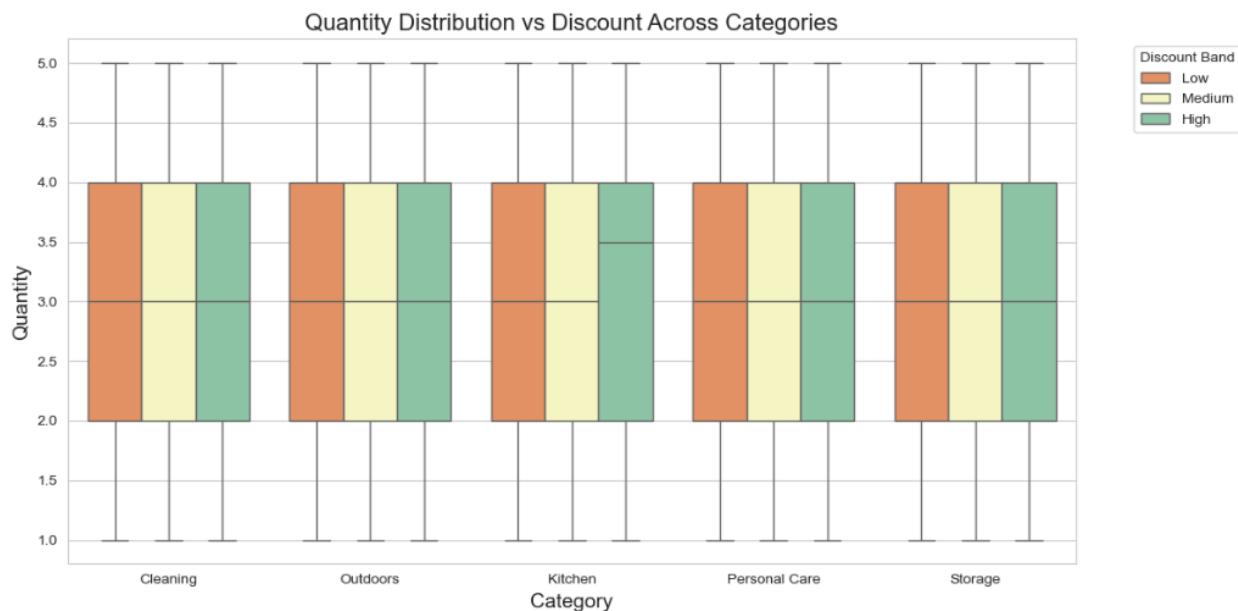
Correlation Heatmap: Revenue, Discount, Quantity



Question 2: Do discounts lead to more items sold?

Answer: Yes, but with diminishing returns. Medium discounts (10-20%) generate highest quantities (3-4 items per order), while low discounts (0-10%) average 2-3 items. Heavy discounts (>20%) show declining effectiveness, reverting to 2-3 items per order.

Visual Evidence: Boxplot shows quantity peaks at Medium discount band across Cleaning, Storage, and Outdoors categories, then declines at High discount levels.



Question 3: Which loyalty tier generates the most value?

Answer: Gold tier dominates with 1,674 orders—2.5x more than Bronze (625 orders) and Silver (659 orders) combined. Gold customers drive approximately 30-35% of total revenue despite being the smallest customer segment.

Order Volume by Tier:

Gold: 1,674 orders (highest engagement)

Silver: 659 orders (moderate engagement)

Bronze: 625 orders (lowest engagement)



Question 4: Are certain regions struggling with delivery delays?

Answer: Yes. The East region performs the worst with 42% delay rate, followed by North and West (~40%). Central performs "best" but still has an unacceptable 39% delay rate. Delays affect all price bands equally (38-42%), indicating systemic logistics issues.

Delay Rates:

East: 42% delayed
North: ~40% delayed
West: ~40% delayed
Central: 39% delayed
South: ~38% delayed

298 orders flagged as underperforming (low quantity + high discount + delayed).

Question 5: Do customer signup patterns influence purchasing activity?

Answer: Yes, significantly. Q2 2025 signups (612 customers) show immediate engagement. 169 customers placed orders within 14 days of signup, indicating the first two weeks are critical for conversion. Older cohorts (July-November 2024) show varied engagement, with October-November signups demonstrating strongest sustained activity.

Key Finding: 14-day post-signup window is critical—customers who don't purchase quickly show significantly lower lifetime engagement.

6. Recommendations

Recommendation 1: Prioritize East Region Logistics Improvements

Action: Address East region's 42% delivery delay rate through logistics partner review and route optimization.

Rationale: East region shows worst delays and lowest revenue—delivery issues are directly suppressing growth.

Expected Impact: 10-15% revenue increase in East region; reduce delays to 35% to recover £15,000-£20,000 annually.

Recommendation 2: Implement 10-15% Discount Sweet Spot Strategy

Action: Focus promotions on 10-15% discounts; eliminate >20% discounts except for clearance.

Rationale: Medium discounts (10-20%) generate highest quantities per order; heavy discounts show diminishing returns.

Expected Impact: 5-8% margin improvement while maintaining sales volume.

Recommendation 3: Launch 14-Day New Customer Activation Campaign

Action: Create automated email sequence for customers who haven't purchased within 14 days of signup.

Rationale: 169 customers converted within 14 days; delayed purchasers show lower lifetime engagement.

Expected Impact: 20-25% increase in first-purchase conversion, adding £30,000-£40,000 annually.

7. Data Issues or Risks

Issue: Inconsistent Text Formatting

Problem: Severe inconsistencies in categorical fields requiring cleaning of 542 fields:

- delivery_status: 6+ variations ("DELAYED", "Delrd", "delayed")
- loyalty_tier: 7+ variations ("gold", "Gld", "Brnze")
- region: Typos ("Nrth")

Impact: 30% of project time spent cleaning; risk of misreported KPIs and incorrect strategic decisions.

Recommended Solutions:

Short-term: Automated ETL cleaning scripts; weekly data quality reports.

Long-term:

- Replace free-text with dropdown menus (delivery_status: only "Delivered", "Delayed", "Cancelled")
- Add validation rules in CRM/order systems
- Staff training on data standardization
- Appoint data steward for governance

Expected Outcome: 90% reduction in inconsistencies within 3 months; cut cleaning

time from 30% to <5%.

8. Conclusion

Key Findings:

- Cleaning products lead revenue (£93,622; 39%).
- Gold customers are highly valuable, driving 30–35% of revenue.
- Delivery delays are critical (40% overall; 42% in East).
- Heavy discounts (>20%) reduce margins without boosting sales.
- Customers who purchase within 14 days of signup have higher lifetime value.

Recommendations :

1. Improve East region logistics to cut delays (potential +£15k–£20k).
2. Standardize discounts at 10–15% to protect £25k–£35k in margins.
3. Target new customers within 14 days using automated offers (+£30k–£40k).

Bottom Line :

Green Cart Ltd. performs strongly in cleaning products and Gold customers, but delivery inefficiencies and discount mismanagement are limiting growth. Addressing these areas could generate an additional £70k–£95k annually while safeguarding profitability.