Project Coversheet

Full Name	Maria Schiza
Email	mariaschiza005@gmail.com
Contact Number	07899940691
Date of Submission	19/07/2025
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:
  - Submit all four weekly projects, and
  - 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.

1. Introduction

This report presents an analysis of sales, product and customer data for Green Cart Ltd., a local e-commerce company specialising in eco-friendly household products. As a member of their Data and Insights team, I conducted the following analysis using Python programming language to support the upcoming Q2 performance review. The primary objective is to deliver a clear business overview focusing on sales trends, customer behaviour, and product category insights across different UK regions.

The datasets provided to support the review include:

- sales_data.csv: Contains order-level sales records.
- customer_info.csv: Includes customer profile information.
- product_info.csv: Provides product-level details including categories and pricing.

The report aims to find which product categories contribute the most to overall revenue, evaluate customer loyalty behaviours, uncover issues regarding delivery performance, and investigate how customer sign-up patterns relate to purchasing activity.

The insights gained will help inform strategic decisions across marketing, inventory planning, and customer engagement to enhance business operations.

2. Data Cleaning Summary

The 3 datasets include important user information, and it is therefore necessary, to ensure data accuracy and reliability using the following cleaning steps before conducting meaningful analysis.

Standardised Text Fields

Used .str.strip(), .str.lower(), .str.title() and .replace() to remove whitespace and unify inconsistent entries were found due to capitalisation differences or typos (eg. ‘gld’, ‘DELAYED’). This command was applied to most columns across all datasets including delivery_status, loyalty_tier, and region.

Datetime Conversion

Converted signup_date, launch_date, and order_date columns using pd.to_datetime() to the same datetime format.

Missing Values

- Filled missing discount_applied with 0.0.
- Replaced missing categorical values like loyalty_tier with "Unknown".

- Dropped rows when critical identifiers like order_id, customer_id or product_id were missing, as well as for order_date or signup_date since such columns contain important information for time-based analysis.

Removed Duplicates

Checked for duplicate customer_ids and removed exact duplicate rows using .drop_duplicates() in customer_info dataset.

Similar approach was applied to the other 2 datasets based on product_id and order_id in product_info and sales_data datasets respectively.

Numeric Validation

Converted columns like quantity, unit_price, and discount_applied to numeric. Replaced worded numbers (e.g., "five") in quantity column with their numeric counterparts and removed any records with negative values.

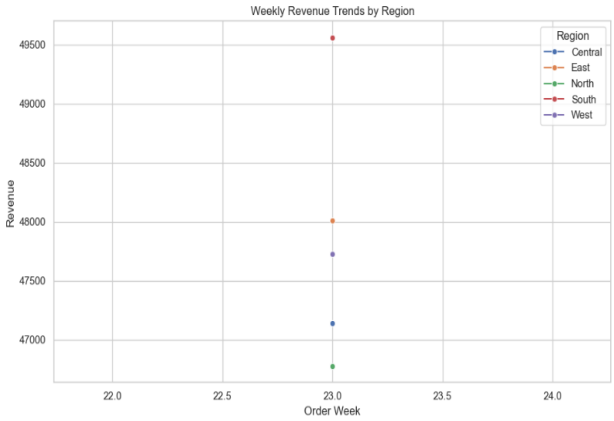
3. Feature Engineering Summary

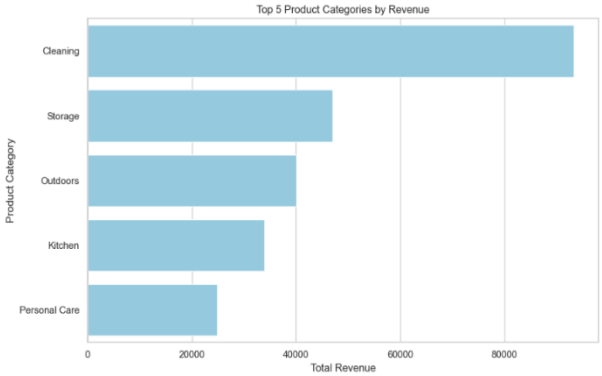
The following features were created to enhance analysis:

- **Revenue**: Calculated by multiplying the quantity sold by the price of each unit and then reduce it be any discount applied
- **Order_week**: The week number of the year, found using the command dt.isocalendar().week
- **Price_band**: Grouping product prices into 'Low' (<£15), 'Medium' (£15-£30), 'High' (>£30).
- **Days_to_order**: Number of days between the launched and ordered date of a product.
- **Email_domain**: Part of the email after the @ symbol. Can be extracted using .str.split('@').str[-1]
- **Is_late**: Boolean flag indicating whether an order was delivered late.
- **Signup_month**: The month a customer signed up.

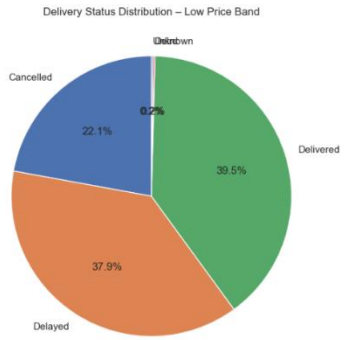
4. Key Findings & Trends

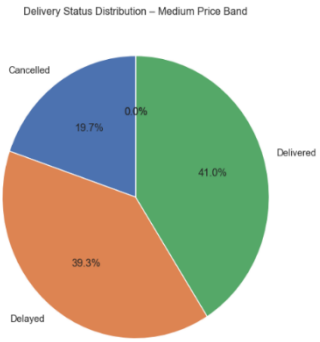
- Revenue reached its peak in week 23 within the South region, recording a value of approximately £50,000 as shown in the line graph to the left. The Cleaning category has contributed the most to this figure, generating substantially higher revenue compared to the other categories as shown in the bar chart.

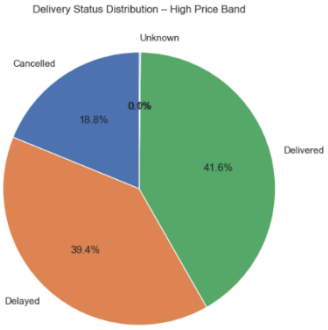




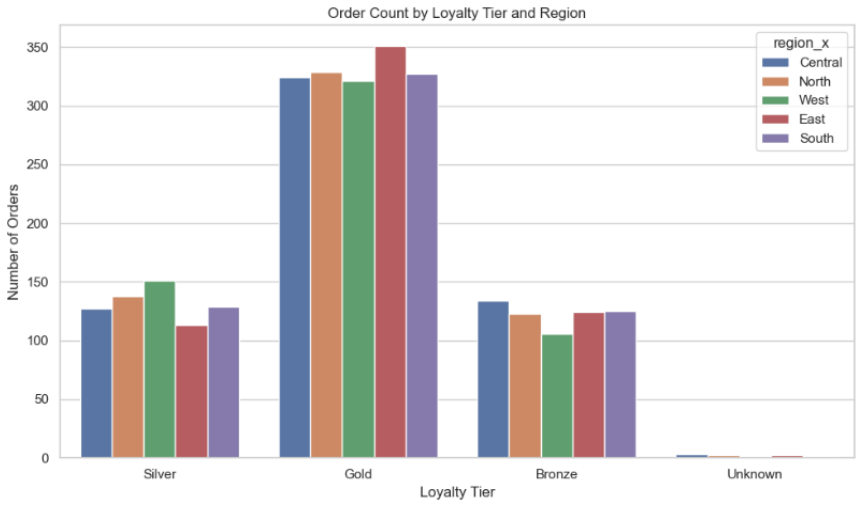
- Across all price bands—Low, Medium, and High—the delivery status distribution remained relatively consistent. Approximately 40% of orders were delivered on time, just under 40% were delayed, and around 20% were cancelled, as we can see from the pie charts below.







- Gold-tier customers, particularly in the East region, placed the highest number of orders at approximately 350. On the contrary, both silver and bronze customers demonstrated similar ordering behaviour, each contributing around 125 orders. This order count relationship by loyalty tier and region is presented below in a count plot.



5. Business Question Answers

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

- The Cleaning category generated the highest total revenue, with a strong performance in the East region. Other regions also exhibited high revenue figures for this category, each exceeding £17,500. The Storage category ranked second in overall revenue, with the West region being the most profitable, contributing approximately £10,000.

2. Do discounts lead to more items sold?

- The correlation number between discount and quantity is -0.01. This tells us that there is almost no correlation between the two variables, thus as discount increases, quantity sold doesn't change in any meaningful way.

3. Which loyalty tier generates the most value?

- The most revenue is generated by gold tier customers with a value of around £130,000.

4. Are certain regions struggling with delivery delays?

- The East region has the highest proportion of delayed deliveries at 41.7%. This indicates potential logistical or operational challenges in that area. The rest of the regions show moderately high delay rates as well, ranging from around 35-40%.

5. Do customer signup patterns influence purchasing activity?

- Customers who signed up in late 2024 and early 2025 contributed the most to overall purchasing activity, with consistently high order volumes and total revenue.

- For example, October 2024 to April 2025 saw an average of over 200 orders per month, peaking in October 2024 (261 orders) and April 2025 (214 orders).
- Revenue was also highest in these periods, with October 2024 reaching over £20,000.
- Customers who signed up in early 2024 showed lower purchasing activity, with fewer than 50 orders per month and total revenue below £4,500.

- On the contrary, signup months including October and November 2025 have lower order volumes, but higher average revenue per order—£93.10 in October 2025 and £100.76 in November 2025—suggesting newer customers may be making larger or higher-value purchases despite placing fewer orders.

6. Recommendations

- Focus on the Cleaning category**,** especially in the East region, where revenue is the highest by allocating more inventory, promotions, and marketing efforts to these areas to sustain strong performance.

- The negligible correlation between discounts and quantity sold suggests a different approach to be taken in targeting discount campaigns. Consider personalised promotions that are more likely to influence customer purchasing behaviour.

- Prioritise retention and rewards for gold customers as they drive the highest revenue and order volumes. Additionally, develop strategies for silver and bronze customers such as exclusive offers and rewards to encourage purchasing activity.

- Conduct an operational audit in the East region, which has the highest delay rate, to identify causes of delay. Actions like route optimisation or vendor performance reviews should be implemented to improve delivery reliability.

- To enhance on time delivery and minimise cancellations it is important to strengthen end-to-end supply chain visibility and enforce predictive delivery tracking or customer alerts to manage expectations and reduce cancellations.

7. Data Issues or Risks

- A frequent issue faced was inconsistent text entries such as varying capitalisations or misspellings in categorical columns which led to inaccurate aggregations during analysis. Examples include, ‘gld’ instead of ‘Gold’ or ‘DELAYED’ instead of ‘Delayed’. To address the issue there should be implemented dropdown menus or controlled vocabulary in the user sign-up interface.

- The dataset contains missing values in essential columns like customer_id, product_id and email, which are crucial for user tracking and communication. Dropping these rows results in data loss, reducing dataset completeness. Best way to address this is as mentioned before, by making these fields mandatory for users in providing valid information as well as real-time validation, such as email verifications.