Project Coversheet

Full Name	Andrea Aguirre
Email	andrea.vtag@gmail.com
Contact Number	083 893 5710
Date of Submission	12/08/2025
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.

YOU CAN START YOUR PROJECT FROM HERE

Introduction

Joining *Green Cart Ltd.*, a rapidly growing UK e-commerce company dedicated to eco-friendly household products, I immediately felt welcomed and excited to be part of the team. As a fresh graduate, I've found everyone to be helpful, including my manager on the *Data & Insights* team.

In preparation for the **Q2 performance review**, my first responsibility as a new team member was to **analyse and investigate sales and customer behaviour** across different regions and product categories. This involves *cleaning and merging the datasets, engineering new features* for deeper analysis, and *exploring patterns* in sales and customer behaviour.

The goal of the report is to deliver **clear insights** into revenue trends, customer engagement, and delivery performance over the period covered by the data. Throughout the report, I will be discussing what specific methods I have used in my data analysis.

Data Cleaning Summary

I began my analysis by importing the necessary Python libraries, including **pandas** as **pd** for data manipulation and **matplotlib.pyplot** as **plt** for data visualization. As well, I installed the **seaborn** library in my Jupyter Notebook environment to enhance visual styling and imported it as **sns** to create more insightful and aesthetically pleasing charts. I am familiar with matplotlib.pyplot, but this is my **first time using** seaborn library. I was provided with three CSV files - *sales_data.csv, product_info.csv, customer_info.csv* - which I loaded individually into separate pandas DataFrames. After loading each file, I examined its structure, data types, and checked for missing values to understand the data and identify cleaning needs.

Cleaning was performed individually on each dataset, starting with *sales_data.csv*, then *product_info.csv* and lastly, *customer_info.csv*. The first task was to **standardise text formatting**, as the *delivery_status* column contained misspellings such as "*delyd*" and "*delrd*" instead of "*Delayed*" and "*Delivered*." I corrected these errors by replacing them with the proper spellings, ensuring the first letter was capitalised. To achieve this, I used pandas methods

including .astype(), .lower(), .strip(), .replace(), and .title() for consistent and clean text formatting.

I continued the same process with the **payment_method** and **quantity** columns. This resulted in:

```
0       Delivered
1         Delayed
2       Delivered
3       Cancelled
4         Delayed
          ...
2995    Delivered
2996      Delayed
2997    Delivered
2998    Delivered
2999    Delivered
Name: delivery_status, Length: 3000, dtype: object
0              Paypal
1         Credit Card
2       Bank Transfer
3         Credit Card
4         Credit Card
            ...
2995    Bank Transfer
2996    Bank Transfer
2997      Credit Card
2998      Credit Card
2999      Credit Card
Name: payment_method, Length: 3000, dtype: object
0       3
1       5
2       1
3       1
4       1
       ..
2995    5
2996    4
2997    1
2998    5
2999    3
Name: quantity, Length: 3000, dtype: object
```

After cleaning the text data, I converted the **order_date** column in the sales dataset to **datetime** format using pd.to_datetime(). I applied the same conversion to the **signup_date** column in the *customer_info* dataset and the **launch_date** column in the *product_info* dataset.

```
0       2025-06-07
1       2025-06-07
2       2025-06-07
3       2025-06-07
4       2025-06-07
           ...
2995    2025-06-07
2996    2025-06-07
2997    2025-06-07
2998    2025-06-07
2999    2025-06-07
Name: order_date, Length: 3000, dtype: datetime64[ns]
```

For missing values, I first identified null counts using .isnull().sum(). I then created copies of the data using .copy() before filling missing numerical values (like *discount_applied* and *quantity* with **0.0** and **0**), assuming missing meant none applied or zero quantity. For categorical variables (like *delivery_status* and *region*), missing entries were replaced with "**Unknown**" or "**Other**" to preserve records while indicating incomplete data. Rows missing critical identifiers (*order_id, customer_id, product_id, or order_date*) were dropped, as these are essential for accurate analysis.

```
Missing Values:
 order_id             1
customer_id           2
product_id            5
quantity              0
unit_price            1
order_date            3
delivery_status       0
payment_method        0
region                0
discount_applied    517
dtype: int64
Replaced:
      order_id customer_id product_id quantity  unit_price order_date  \
0     O966977      C00397      P0022        3       39.25 2025-06-07
1     O696648      C00236      P0023        5       18.92 2025-06-07
2     O202644      C00492      P0011        1       29.68 2025-06-07
3     O501803      C00031      P0003        1       32.76 2025-06-07
4     O322242      C00495      P0016        1       47.62 2025-06-07
...       ...         ...         ...      ...         ...        ...
2995  O868860      C00233      P0001        5       43.40 2025-06-07
2996  O949709      C00246      P0029        4       34.04 2025-06-07
2997  O763639      C00182      P0026        1       42.34 2025-06-07
2998  O753958      C00074      P0003        5       35.96 2025-06-07
2999  O929624      C00405      P0004        3       43.23 2025-06-07

     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
...              ...            ...      ...               ...
2995       Delivered  Bank Transfer     West              0.20
2996         Delayed  Bank Transfer     West              0.20
2997       Delivered    Credit Card    South              0.00
2998       Delivered    Credit Card  Central              0.00
2999       Delivered    Credit Card     West              0.10

[3000 rows x 10 columns]
Removed:
      order_id customer_id product_id quantity  unit_price order_date  \
0     O966977      C00397      P0022        3       39.25 2025-06-07
1     O696648      C00236      P0023        5       18.92 2025-06-07
2     O202644      C00492      P0011        1       29.68 2025-06-07
3     O501803      C00031      P0003        1       32.76 2025-06-07
4     O322242      C00495      P0016        1       47.62 2025-06-07
...       ...         ...         ...      ...         ...        ...
2995  O868860      C00233      P0001        5       43.40 2025-06-07
2996  O949709      C00246      P0029        4       34.04 2025-06-07
2997  O763639      C00182      P0026        1       42.34 2025-06-07
2998  O753958      C00074      P0003        5       35.96 2025-06-07
2999  O929624      C00405      P0004        3       43.23 2025-06-07

     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
...              ...            ...      ...               ...
2995       Delivered  Bank Transfer     West              0.20
2996         Delayed  Bank Transfer     West              0.20
2997       Delivered    Credit Card    South              0.00
2998       Delivered    Credit Card  Central              0.00
2999       Delivered    Credit Card     West              0.10

[2989 rows x 10 columns]
```

Duplicates were identified using .duplicated() on order_id and removed with .drop_duplicates(), resulting in two duplicates being deleted.

Finally, I validated the numeric columns (*quantity, unit_price, and discount_applied*) by converting them to numeric types and checking that no negative values were present, ensuring data validity.

The same data cleaning process was applied to all three CSV files to ensure uniformity and accuracy across all datasets.

Feature Engineering Summary

During the analysis, I created several new features:

- **Revenue**: Calculated it by multiplying quantity, unit price, and adjusting for any discount applied. This represents the actual amount earned per order after discounts.
- **Order Week**: Extracted the week number from the *order_date* column to track sales trends every week.
- **Price Band**: Categorised the *unit_price* into three groups, **Low**, **Medium**, and **High**, to analyse sales across different price ranges.
- **Days to Order**: Calculated the number of days between the product's *launch_date* and the *order_date*, indicating how soon customers purchase items after release.
- **Email Domain**: Separated the domain part of customers' *email* to help identify patterns in customer segments. Missing values were replaced with "**Unknown**".
- **Is Late:** Flagged (Boolean flag) orders as late if their *delivery_status* was "Delayed," supporting delivery performance analysis.

Through the analysis, the new features provided clearer insights into sales trends, customer behaviour, and operational performance.

Key Findings & Trends

1. **Revenue Peaked in Week 23:**

Revenue peaked consistently across all regions and sub-regions during Week 23, with the highest total revenue observed in the **West region**, with up to 9,160 units. This suggests a strong sales period likely driven by seasonal factors or promotions.

	revenue
0	117.750
1	94.600
2	25.228
3	26.208
4	38.096
...	...
2993	173.600
2994	108.928
2995	42.340
2996	179.800
2997	116.721

[2998 rows x 1 columns]

2. **Cleaning Enhances Product Category Performance:**

The *Cleaning* category generated the highest total revenue of **£75,484** and quantity sold (**3,589 units**), with an average discount of about 10%. This indicates it is a top-selling category, likely a core part of the product mix, while other categories like Kitchen and Outdoors also contribute significantly, but at lower volumes.

			total_revenue
region_x	region_y	order_week	
Central	Central	23	5618.7830
	East	23	7694.1605
	North	23	8553.8745
	South	23	6885.8005
	West	23	7947.9020
East	Central	23	6406.0925
	East	23	8451.4390
	North	23	6871.6390
	South	23	7919.9075
	West	23	8731.9640
North	Central	23	7334.5375
	East	23	7813.2865
	North	23	8054.9355
	South	23	6902.0275
	West	23	7254.3725
South	Central	23	8451.9320
	East	23	6813.2855
	North	23	8058.1715
	South	23	7252.8005
	West	23	9160.7355
West	Central	23	7010.5420
	East	23	8101.4450
	North	23	9133.0755
	South	23	7762.4095
	West	23	7729.4645
nrth	South	23	19.5120

3. **Delivery Delay Trends by Region and Price Band:**

Higher-priced orders tend to experience more delivery delays, with **delay rates** reaching up to **50%** for ‘High’ price bands in some regions (e.g., Central to Central). Lower-priced orders

generally had fewer delays, suggesting possible logistic prioritisation or challenges with expensive product shipments.

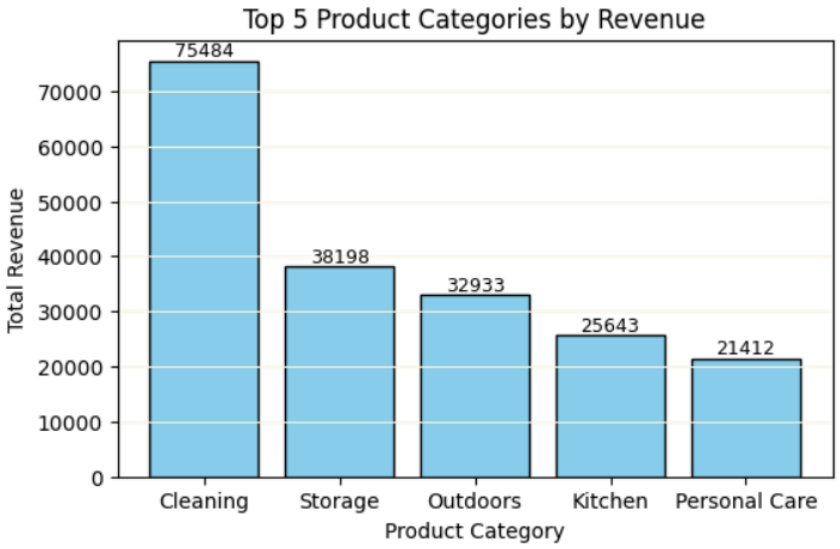
region_x	region_y	price_band	total_orders	delayed_orders	delay_rate
Central	Central	Low	18	7	0.388889
		Medium	45	20	0.444444
		High	40	20	0.500000
	East	Low	21	7	0.333333
		Medium	37	9	0.243243
...			...	...	...
nrth	South	Medium	1	0	0.000000
		High	0	0	NaN
	West	Low	0	0	NaN
		Medium	0	0	NaN
		High	0	0	NaN

[90 rows x 3 columns]

Business Question Answers

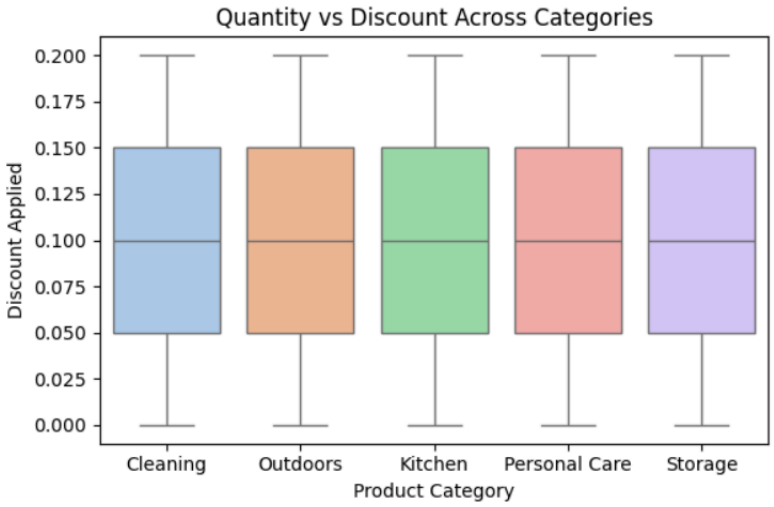
1. The *Cleaning* category made about **£75,484** and **sold 3,589 units**, leading all categories. Storage and Outdoors followed with revenues of around £38,198 and £32,933. Most sales come from the West and South regions, which show the **highest weekly revenue totals**.

   This data can be visualized using the bar chart shown below, with *Cleaning* leading.



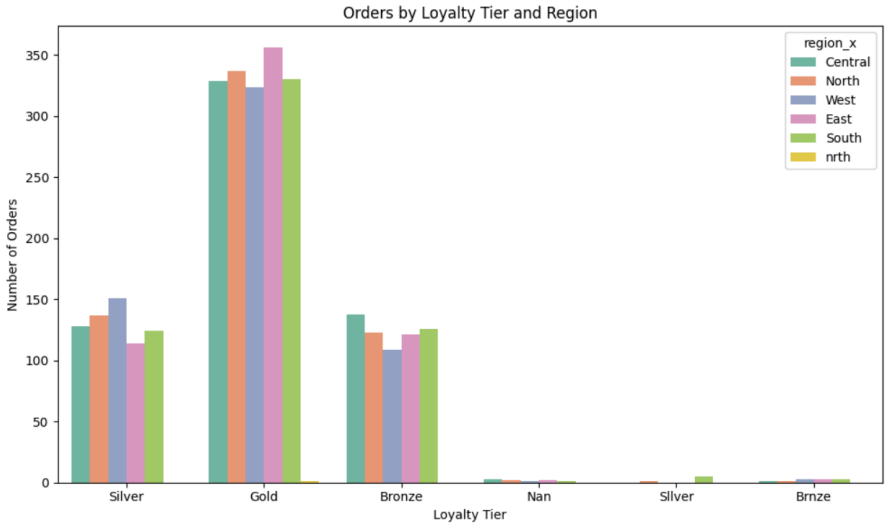
2. Discounts vary by category, but they don't always mean more sales. For example, categories with an average discount of around 10% don't always sell more than those with lower discounts. So, giving bigger discounts doesn't guarantee higher quantity sold.

The information is represented in the boxplot below.



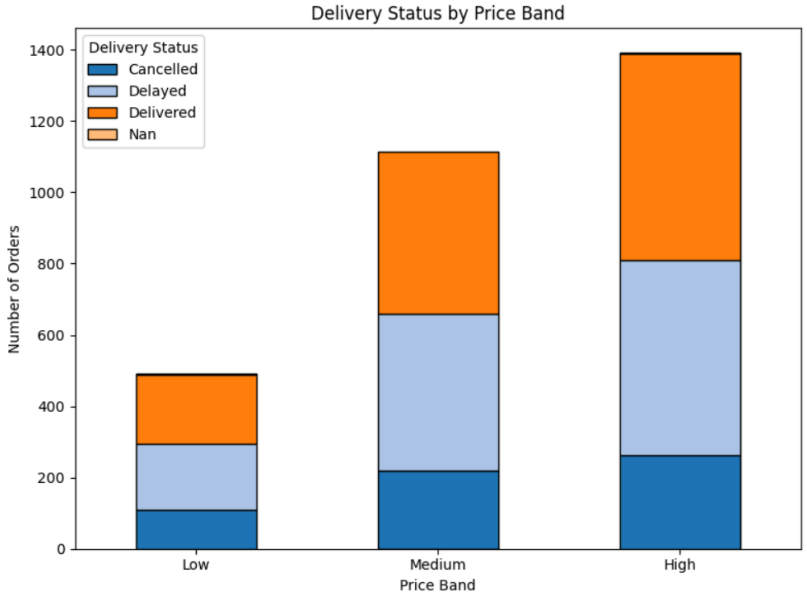
3. **Gold members** place the most orders and generate the highest revenue, followed by Silver and Bronze. Gold tier customers make up the majority of high-value purchases.

   The following countplot depicts the data, which shows Gold being the highest.



4. Yes. The **Central region** has a delay rate of up to 50% for high-priced orders, while other regions like North and East have lower delay rates (around 20-30%). This shows that delays are mostly a problem in certain areas.

The data is summarized in the stacked bar chart shown below, with the number of high-priced orders that are delayed.



5. Customers who signed up earlier tend to order more and spend more. For example, customers signed up in early 2024 have higher average order counts and revenue compared to those signed up later in 2025.

Recommendations

1. **Boost Marketing for Top-Selling Categories in Key Regions**
   Focus promotions on the **Cleaning** and **Storage** categories, especially in the **West** and **South** regions, where revenue is already high. This could maximise sales where demand is strongest.
2. **Target Loyalty Tier Upselling**
   Since **Gold** members spend the most, introduce exclusive bundles or early-access deals for **Silver** members to encourage upgrades to **Gold**, increasing their lifetime value.
3. **Improve Delivery Reliability in Delay-Prone Regions**
   Prioritise logistics improvements in the **Central** region, where delivery delays are highest. Offering faster shipping or delay compensation could improve customer satisfaction and retention.

Data Issues or Risks

Several issues were identified in the dataset, including blank values in **order_id, customer_id, product_id,** and **signup_date**, which indicate incomplete or inaccurate data collection. These fields should be marked as mandatory (star/required) and automatically generated where possible (e.g., signup_date should be system-generated rather than user-entered).

Additionally, there are misspellings in fields like *gender* and *loyalty_tier* due to manual typing (e.g., "**Brnze**" instead of "**Bronze**"), which could be prevented by replacing free-text entry with dropdown/select options. Implementing these fixes at the data entry source would ensure completeness, consistency, and accuracy for future analyses.