Project Coversheet

Full Name	RAMIL KHALILLI
Project Title (Example – Week1, Week2, Week3, Week 4)	WEEK 2 - SALES & CUSTOMER BEHAVIOUR INSIGHTS - GREEN CART Ltd.

Instructions:

Students must download this cover sheet**,** use it as the first page of their project, and then save the entire document as a PDF before submission**.**

Project Guidelines and Rules

1. Formatting and Submission

- Format: Use a readable font (e.g., Arial/Times New Roman), size 12, 1.5 line spacing.
- Title: Include Week and Title (Example - Week 1: Travel Ease Case Study.)
- File Format: Submit as PDF or Word file
- Page Limit: 4–5 pages, including the title and references.

2. Answer Requirements

- Word Count: Each answer should be within 100–150 words; Maximum 800–1,200 words.
- Clarity: Write concise, structured answers with key points.
- Tone: Use formal, professional language.

3. Content Rules

- Answer all questions thoroughly, referencing case study concepts.

- Use examples where possible (e.g., risk assessment techniques).
- Break complex answers into bullet points or lists.

4. Plagiarism Policy

- Submit original work; no copy-pasting.
- Cite external material in a consistent format (e.g., APA, MLA).

5. Evaluation Criteria

- Understanding: Clear grasp of business analysis principles.
- Application: Effective use of concepts like cost-benefit analysis and Agile/Waterfall.
- Clarity: Logical, well-structured responses.
- Creativity: Innovative problem-solving and examples.
- Completeness: Answer all questions within the word limit.

6. Deadlines and Late Submissions

- Deadline: Submit on time; trainees who fail to submit the project will miss the “Certificate of Excellence”

7. Additional Resources

- Refer to lecture notes and recommended readings.
- Contact the instructor or peers for clarifications before the deadline.

1. Introduction

This project analyses **sales performance and customer behaviour** for *Green Cart Ltd.*, a UK-based e-commerce company selling eco-friendly household products. The goal is to support the company's **Q2 performance review** by identifying key patterns in revenue, customer activity, product performance, discounts, and delivery outcomes across regions.

The work follows the project brief in structured stages: first, the three provided datasets (**sales_data**, **product_info**, **customer_info**) are **loaded, cleaned, and merged** into a single analytical dataset. Next, several **new features** are engineered (such as revenue, order week, price band, delivery delay flag, and customer email domain) to enable deeper analysis. Using these features, the project then produces **summary tables and visualisations** that reveal trends in weekly revenue, category performance, loyalty-tier behaviour, delivery reliability, and payment preferences.

Finally, the findings are used to **answer the required business questions** and produce clear, actionable recommendations for marketing and operational decision-making.

2. Data Cleaning Summary

As the tasks wanted me to I standardized text formatting inside some columns and made some cleaning such as converting different entries into a single entry that means the same:
'DELAYED' -> 'Delayed'
'gold' -> 'Gold'
'gld' -> 'Gold'
'brnze' -> 'Bronze'
'sllver' -> 'Silver'
'femle' -> 'Female'

Then I converted the date columns (order_date, signup_date, launch_date) into a datetime data type using pd.to_datetime(). Then came handling missing values. Depending on the case I treated missing values in different columns differently. In primary columns of the tables such as 'customer_id' in customers table, 'order_id' in sales table and 'product_id' in products table, I dropped rows where there was a missing value since primary keys should be non-null. I also dropped rows where columns like 'unit_price' and 'quantity' had missing values since these are key business columns and we want them non-null and imputing them artificially will add only noise to the data. Ignoring them is not the best idea either since these are very important metrics for us and we should get rid of null values. I treated null values in categorical columns like 'gender', 'region', 'loyalty_tier', 'delivery_status' and 'payment_method' as their own category ('Unknown'). And I ignored null values in 'order_date' column as droping valid transactons becuse of missing date is not right and logically imputing these missing values is not possible either. We just need to exclude them from Time-Series Analysis. I filled missing values inside 'discount_applied' column with 0.0 since missing value in this column mostly means that there is a 0 discount.

Then I removed duplicate values according to the primary columns of the tables. Finally for this part I made sure that columns like 'unit_price' and 'discount_applied' were non-negative.

3. Feature Engineering Summary

To enable deeper analysis of sales performance and customer behaviour, several new features were created from the original datasets:

Revenue: Calculated as *quantity × unit_price × (1 − discount_applied)* to reflect the actual monetary value generated per order after discounts.

Order Week: Extracted as the ISO week number from the order date, allowing weekly trend analysis of sales and revenue.

Price Band: Products were categorised into **Low (£<15)**, **Medium (£15–30)**, and **High (£>30)** price ranges to support price-based performance and delivery analysis.

Days to Order: Represents the number of days between a product’s launch date and the order date, helping assess how quickly products generate sales after launch.

Email Domain: Extracted from customer email addresses (e.g., *gmail.com*) to identify potential patterns in customer sign-up behaviour.

Is Late: A boolean flag indicating whether an order’s delivery status was marked as *Delayed*, enabling focused analysis of delivery performance issues.

These engineered features form the foundation for the summary tables, visualisations, and business insights presented in the report.

4. Key Findings & Trends

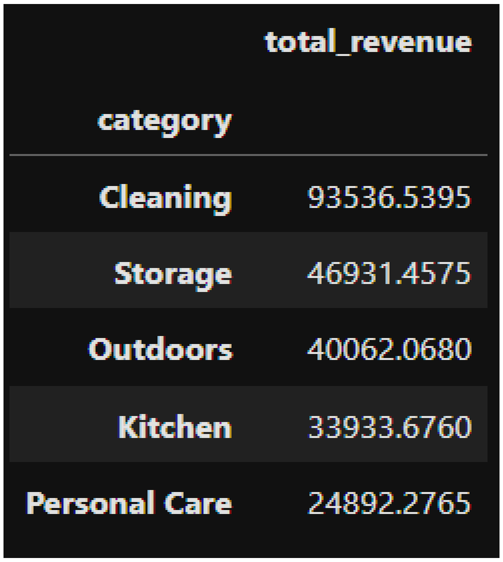
Cleaning products were sold the most but on average kitchen products brings the most revenue.
For bronze loyalty tier the most orders were made during June and the most number of customers signed up during June and October. For silver loyalty tier the month of March is the most fruitful with the most number of customers signing up and orders made. For gold loyalty tier however the month of September was the most fruitful in terms of number of orders made and customers signing up.

Also in the delivery of high price band products there was less delay than other price bands. All customers belonging to different loyalty tier prefers credit card as a means of payment.

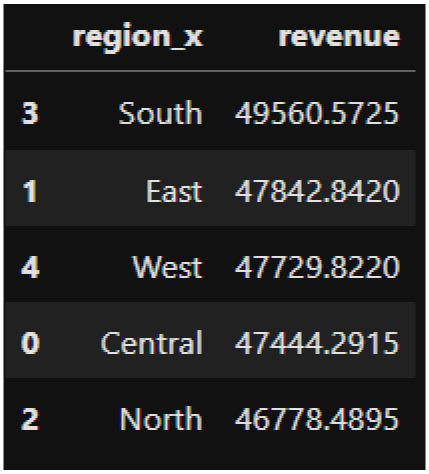
5. Business Question Answers

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

- Cleaning products drive the most revenue overall and South region drives the most revenue.



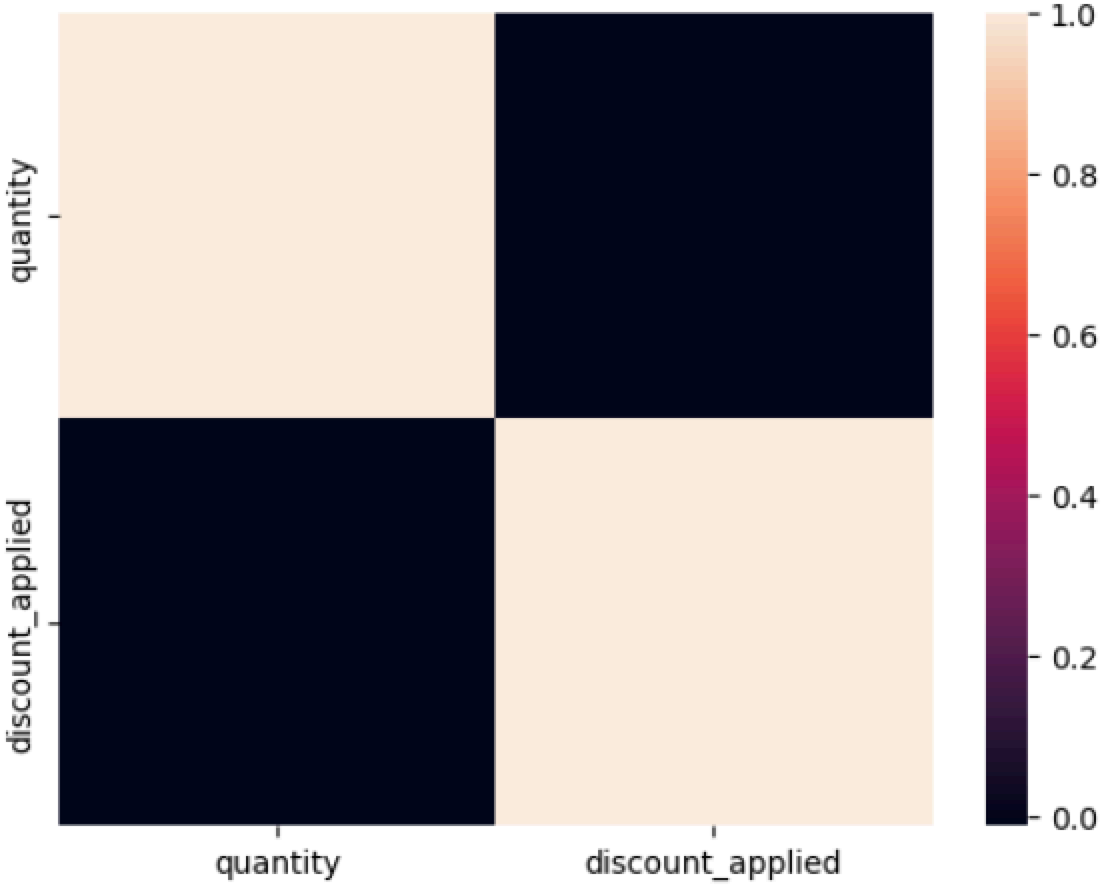
	total_revenue
category	
Cleaning	93536.5395
Storage	46931.4575
Outdoors	40062.0680
Kitchen	33933.6760
Personal Care	24892.2765



	region_x	revenue
3	South	49560.5725
1	East	47842.8420
4	West	47729.8220
0	Central	47444.2915
2	North	46778.4895

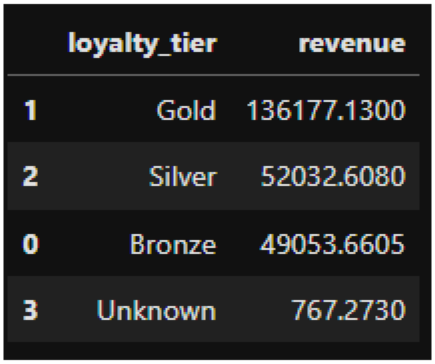
2. Do discounts lead to more items sold?

- Yes, there is strong positive relationship between discount applied and the quantity of items sold.



3. Which loyalty tier generates the most value?

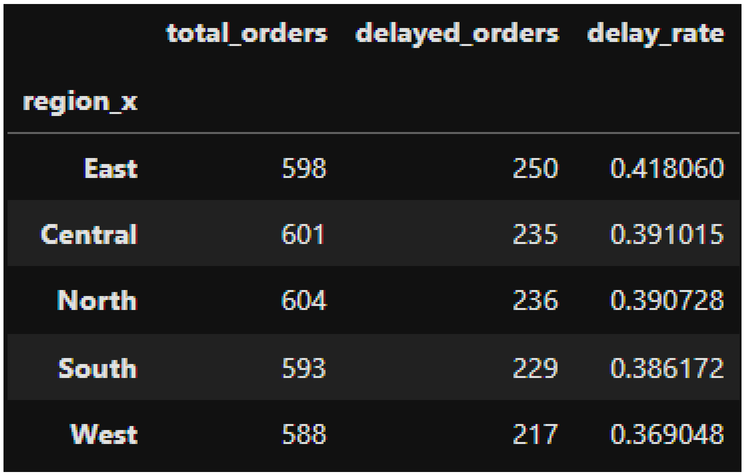
- Gold tier generates the most revenue.



	loyalty_tier	revenue
1	Gold	136177.1300
2	Silver	52032.6080
0	Bronze	49053.6605
3	Unknown	767.2730

4. Are certain regions struggling with delivery delays?

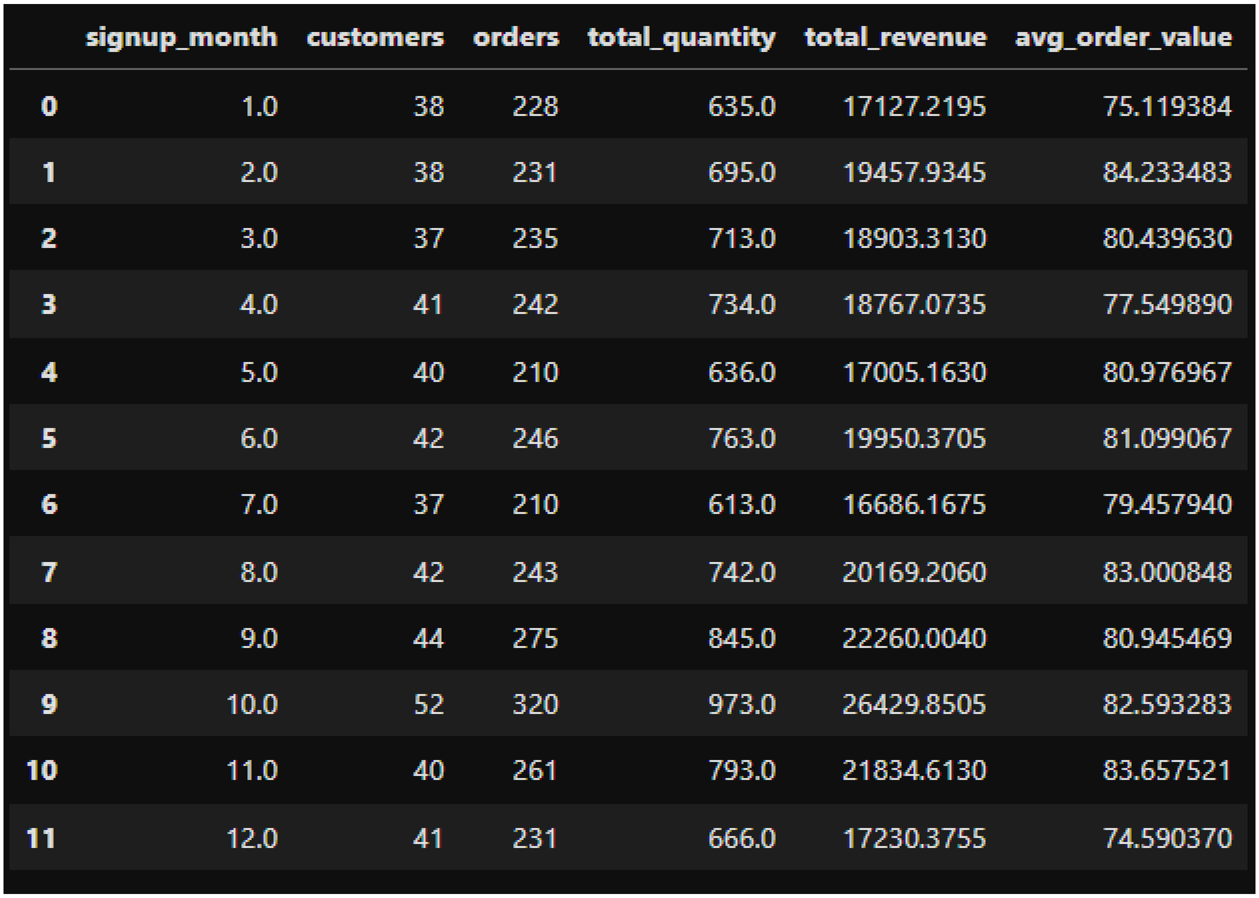
- The East region has the highest delivery delay rate; however, delay rates across regions are relatively similar, suggesting no major regional disparity.



	total_orders	delayed_orders	delay_rate
region_x			
East	598	250	0.418060
Central	601	235	0.391015
North	604	236	0.390728
South	593	229	0.386172
West	588	217	0.369048

5. Do customer signup patterns influence purchasing activity?

- The 11th month (November) shows great numbers across the metrics such as order quantity sold, the number of total customers, the number of total orders and total revenue. For other months there is no any significant difference.



	signup_month	customers	orders	total_quantity	total_revenue	avg_order_value
0	1.0	38	228	635.0	17127.2195	75.119384
1	2.0	38	231	695.0	19457.9345	84.233483
2	3.0	37	235	713.0	18903.3130	80.439630
3	4.0	41	242	734.0	18767.0735	77.549890
4	5.0	40	210	636.0	17005.1630	80.976967
5	6.0	42	246	763.0	19950.3705	81.099067
6	7.0	37	210	613.0	16686.1675	79.457940
7	8.0	42	243	742.0	20169.2060	83.000848
8	9.0	44	275	845.0	22260.0040	80.945469
9	10.0	52	320	973.0	26429.8505	82.593283
10	11.0	40	261	793.0	21834.6130	83.657521
11	12.0	41	231	666.0	17230.3755	74.590370

6. Recommendations

1. There should be focus on promotions on the 'cleaning' product category in high performing regions.
2. East region has the most delay rates we should be taken an action against.

7. Data Issues or Risks

Data had some missing values which should be looked upon to see what is the source of this problem. The most likely answer is upstream where we obtain the data. This should be prevented to achieve completeness of data via automated checks or in other way.