Project Coversheet

Full Name	Shabnam Shaik
Email	shabnamsmiles@gmail.com
Contact Number	+44 7880803036
Date of Submission	12-08-2025
Project Week	Week 2

1. Introduction:

- ### Project Overview

Green Cart Ltd., a UK-based e-commerce company specializing in eco-friendly household products, is preparing for its Q2 performance review. To support strategic decision-making in marketing and operations, the Data & Insights team has initiated a project to analyse customer behaviour and sales performance across different regions and product lines.
The primary objective is to explore and extract actionable insights from the company's datasets. The analysis will focus on identifying trends, performance metrics, and behavioural patterns that can drive informed business decisions for the upcoming quarter.

Objective:
Integrate and analyse sales, product, and customer data to uncover actionable business insights.

Key Steps:

1. **Data Preparation** – Clean and standardize datasets, resolve missing values, and merge sources into a unified dataset.
2. **Feature Engineering** – Create meaningful variables (e.g., revenue, price bands, customer tenure) to enable deeper analysis.

3. **Exploratory Analysis** – Identify trends, patterns, and performance drivers across products, customers, and regions.
4. **Insights & Visualization** – Use charts, tables, and KPIs to present findings clearly for decision-making.

Outcome:
A comprehensive analytical view of business performance, highlighting opportunities for growth, efficiency, and customer engagement.

2. Dataset Overview and Cleaning Summary

This project leverages three primary datasets that together provide a comprehensive view of sales transactions, product attributes, and customer profiles.

1. **sales_data.csv**

Purpose: The sales data file records transactional details for each order, including order and customer identifiers, product references, purchase quantity, unit price, order date, delivery status, payment method, sales region, and any discount applied.

2. **product_info.csv**

Purpose: The product information file contains key details for each item, including its Product ID, name, category, launch date, base price, and supplier Code.

3. **customer_info.csv**

Purpose: The customer information file captures essential customer details, including unique customer ID, contact email, signup date, gender, region, and loyalty tier (Bronze, Silver, Gold)

Data Cleaning Summary

1. Standardized text formatting , stripped and used title case

- delivery status were corrected
  "delrd" → "Delivered"
  "delyd" → "Delayed"

- payment method was corrected
  "bank transfr" → "Bank Transfer"

- The region field was corrected
  "nrth" → "North"

- Misspelled gender entries were corrected
  "femle" → "Female"

- loyalty tier values were **corrected**
  “gld” → “Gold”

- For quantity, text numbers (e.g., "three", "five") were **converted to digits**, and values converted to **integer type**

2. Converted date columns:

order_date, signup_date, launch_date are converted to datetime using pd.to_datetime()

3. Handled missing values:

- Replaced missing values of all the categorical fields to “**Unknown"** and **“Unspecified”** for gender
- missing order dates were filled with the **mode date** in the dataset.
- Any **missing signup dates** were filled with the **median signup date** to ensure completeness.
- Missing emails were replaced with **“Unknown@example.com”.**
- Discount values with nulls were set to 0.0. as it is assumed that no discount was given.
- Imputed delivery status of “**unknown”** to “**Delivered”** as there are significantly small.

- Missing values in **quantity**, **unit_price**, **order_id**, **product_id**, and **customer_id** were removed from the dataset, as these incomplete records do not contribute meaningfully to the sales analysis.

4. **Removed duplicates:**

   Duplicate orders (based on order_id) were dropped.

5. **Validate numeric columns:**

   No negative values found for the numeric columns: quantity, unit_price, and discount_applied.

Post cleaning, sales_data is merged with product_info using product_id which is further merged with customer_info using customer_id

3. Feature Engineering Summary :

Several new features were engineered in the merged dataset to enrich analysis:

- **revenue** was calculated as the product of quantity, unit price, and discount adjustment using the below calculation.
  revenue = quantity × unit_price × (1 - discount_applied)
- **order_week** extracted the ISO week number from the order date for time-based grouping
- **price_band** categorized unit prices into Low (<£15), Medium (£15–30), and High (>£30) bands
- **days_to_order** measured the days elapsed between product launch and order date
- **email_domain** extracted the domain part of customer emails for segmentation
- **is_late =True ,** flagged orders with delivery status marked as “Delayed” to identify late deliveries.

Dropped the columns product_name and supplier_code as they do not contribute to analysis

4. Key Findings & Trends:

- **South** region leads in sales revenue with approximately £49,561, followed closely by East (£48,056) and West (£47,730). Central and North regions trail slightly behind.

sales_region	revenue
South	49560.5725
East	48055.8520
West	47729.8220
Central	47444.2915
North	46778.4895

- Among product categories, **Cleaning** generates the highest revenue (£93,643) and quantity sold (3,585 units), despite an average discount of 8.6%. Other notable categories include Storage and Outdoors, both contributing significant revenue with similar discount levels around 8%. Kitchen and Personal Care have lower revenues and quantities but maintain comparable discount rates.

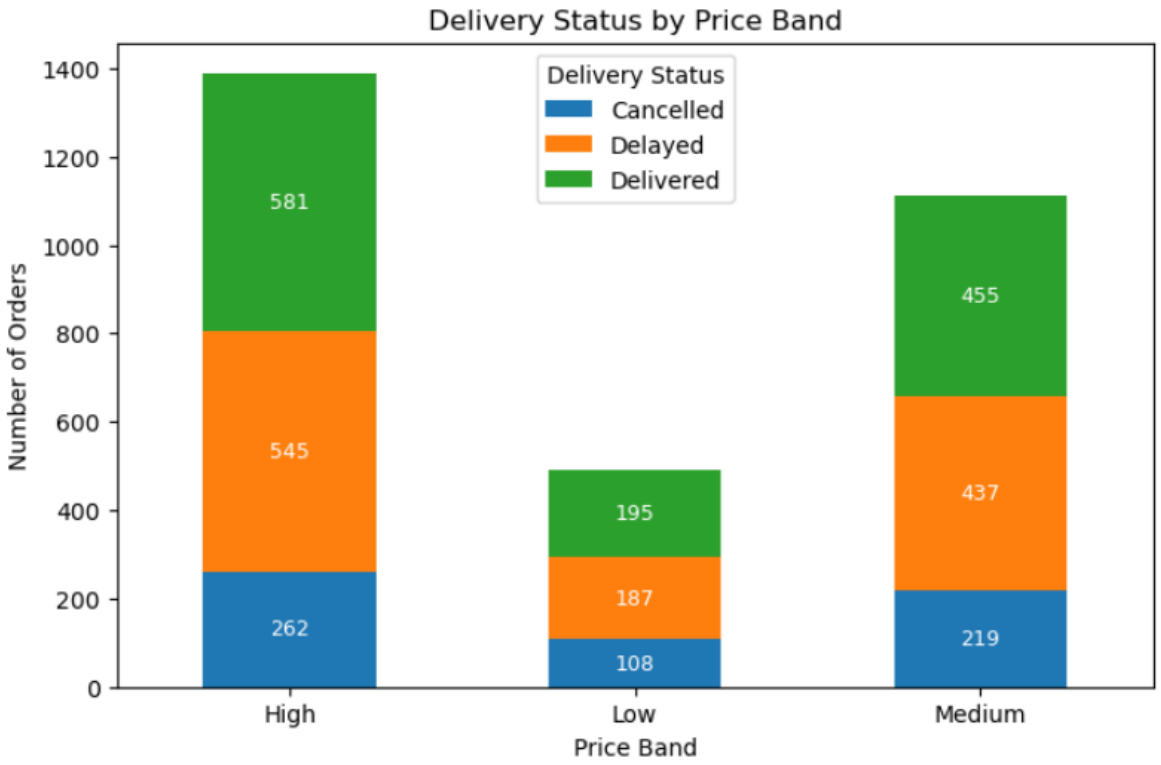
category	revenue	quantity	discount_applied
Cleaning	93643.26	3585	8.6
Storage	47037.75	1733	8.1
Outdoors	40062.07	1519	8.2
Kitchen	33933.68	1226	7.6
Personal Care	24892.28	900	8.7

Overall, the South region drives the strongest sales, while Cleaning stands out as the top-performing category by revenue and volume, with moderate discounts applied across categories.

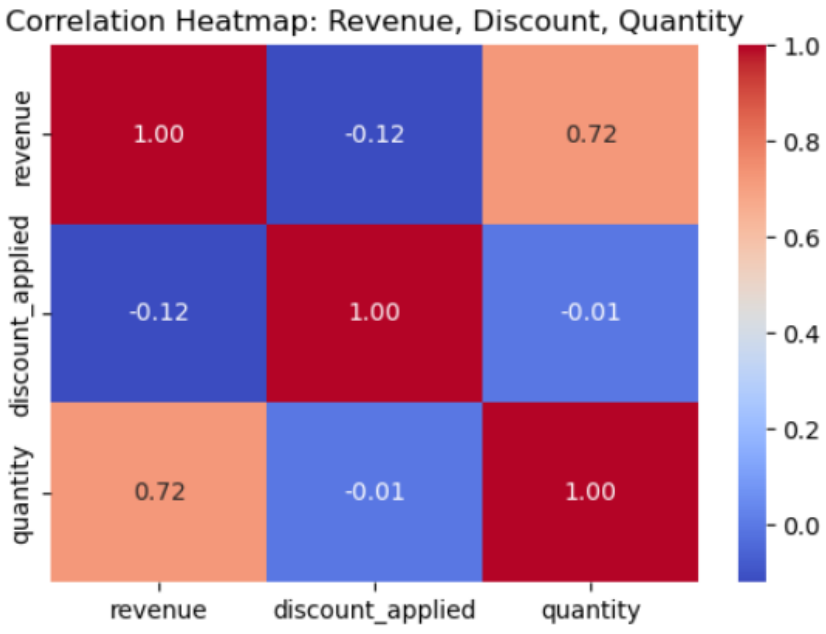
- Delivery status by price band

Delivery outcomes are fairly consistent across all price bands, with roughly 40% of orders delivered and around 38–39% delayed. However, low-priced products

experience a slightly higher cancellation rate (~22%) compared to medium (~20%) and high-priced (~19%) items.



- correlation between revenue, discount, and quantity



- **Revenue and Quantity** show a strong positive correlation (0.72), indicating that higher quantities sold generally lead to higher revenue.
- **Discount Applied** has a weak negative correlation with revenue (-0.12) and virtually no correlation with quantity (-0.01), suggesting discounts do not significantly impact sales volume or revenue in this dataset.

5. Business Question Answers

5.1) Which product categories drive the most revenue, and in which regions?

From Fig. 5.1, it is clear that, the 'Cleaning' product category generates the highest revenue, with a slightly stronger contribution from the East region, followed by the South. The remaining regions contribute nearly equal amounts of revenue.

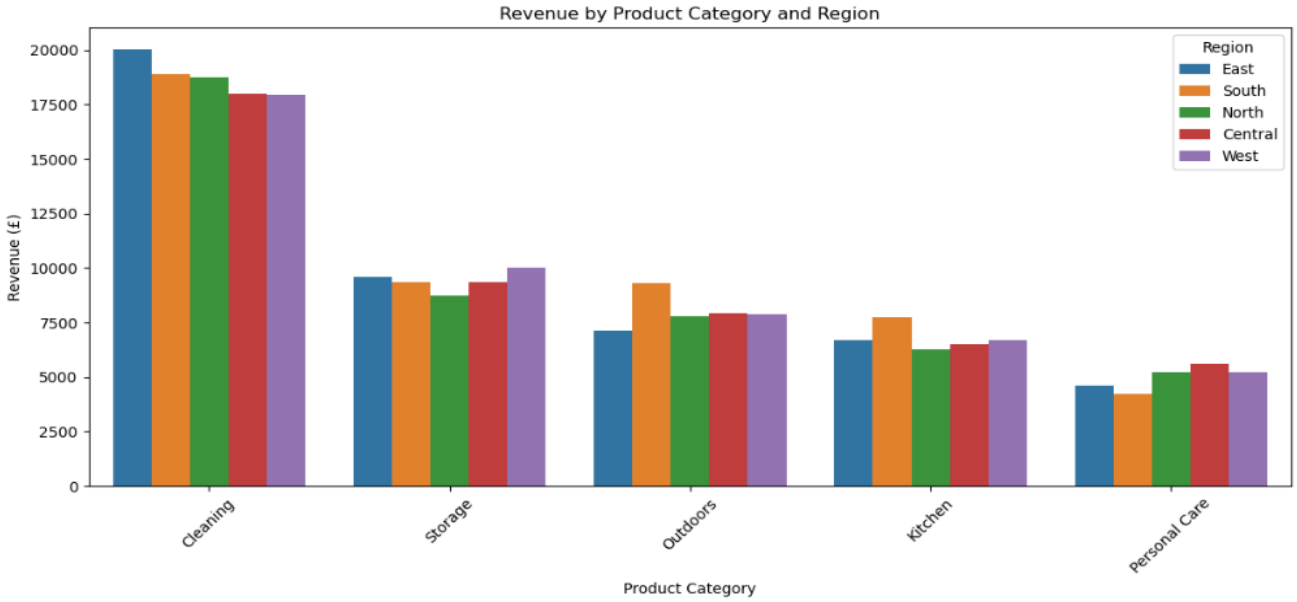


Fig: 5.1 Revenue by product Category and Region

5.2) Do discounts lead to more items sold?

From Fig. 5.2, the average quantity sold across all discount bands is identical (3 units), indicating no observable relationship between discount levels and quantity sold.

	discount_band	quantity
0	Low Discount	3
1	Medium Discount	3
2	High Discount	3

Fig: 5.2 Average quantity sold in each Discount band

5.3) Which loyalty tier generates the most value?

The Gold loyalty tier generates the highest revenue at **£136,283.85 (57%)**, significantly outperforming Silver **£52,138.90 (22%)** and Bronze **£49,053.66 (21%)**, while customers with an Unknown tier contribute negligibly **£767.27 (0.3%)**.

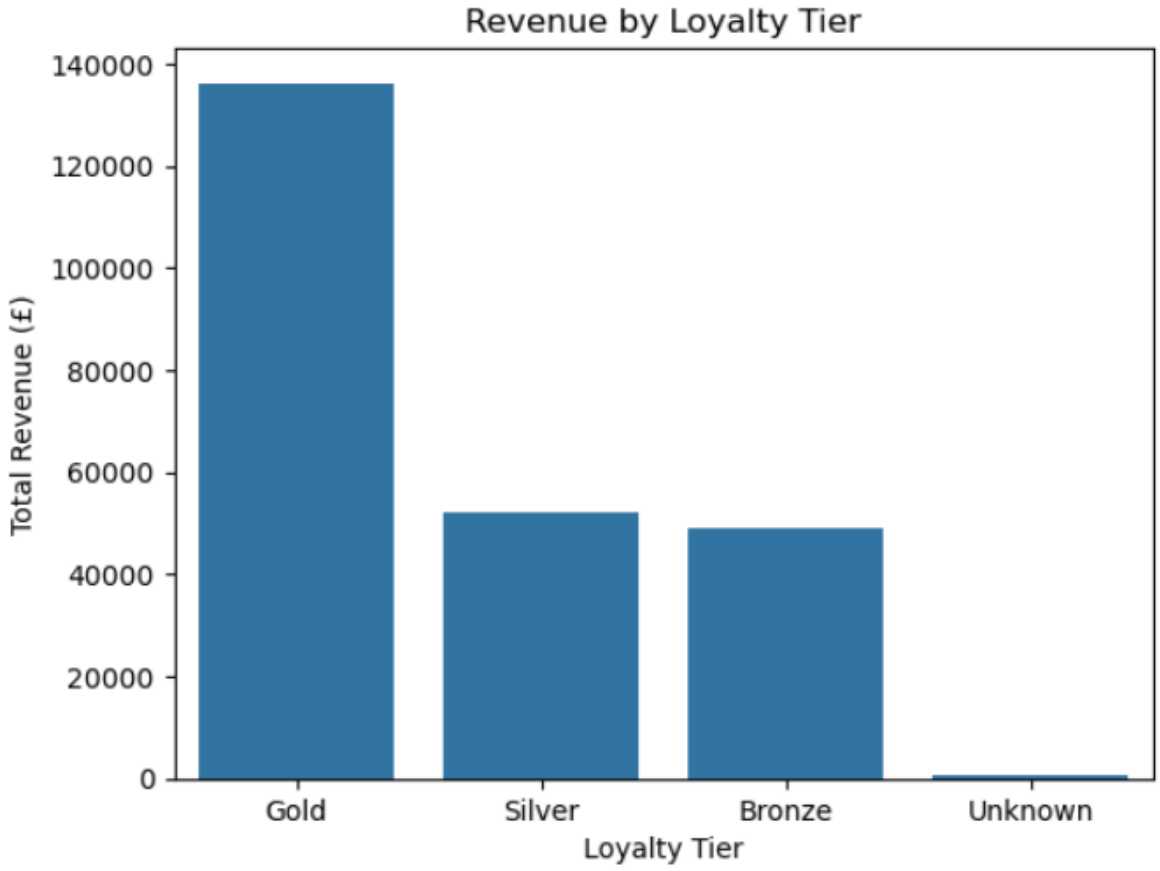


Fig: 5.3 Revenue by Loyalty Tier

5.4) Are certain regions struggling with delivery delays?

All regions have a similar distribution of delivery statuses with most orders being delivered, followed by delayed and then cancelled. The East region has the

highest number of delayed orders (251), closely followed by North (236) and Central (235)

delivery_status	Cancelled	Delayed	Delivered
sales_region			
Central	125.0	235.0	242.0
East	98.0	251.0	251.0
North	117.0	236.0	251.0
South	116.0	230.0	249.0
West	133.0	217.0	238.0

Fig: 5.4a Region wise delivery status counts

delivery_status	Delayed	Total	Delayed_pct
sales_region			
Central	235.0	602.0	39.04
East	251.0	600.0	41.83
North	236.0	604.0	39.07
South	230.0	595.0	38.66
West	217.0	588.0	36.90

Fig: 5.4b Percentage of Delayed orders in each Region

When looking at the percentage of delayed deliveries (fig 5.4b) relative to total orders, East also leads with 41.8% delayed, while West has the lowest delay rate at 36.9%. Overall, delivery delays affect roughly 37% to 42% of orders across regions, indicating consistent challenges with timely deliveries throughout.

5.5) Do customer signup patterns influence purchasing activity?

From July 2024 to June 2025, revenue, quantity sold, and unique orders generally fluctuate within a moderate range, with peaks around August 2024 and February 2025, indicating stronger purchasing activity in these months. There is a noticeable dip in April and May 2025, suggesting a slowdown in sales and orders. The sharp decline in all metrics in July 2025 is likely due to incomplete data for the latest month, rather than a true drop in the customer activity. Overall, the data reflects some seasonal variation, with possible factors including marketing campaigns, seasonal demand.

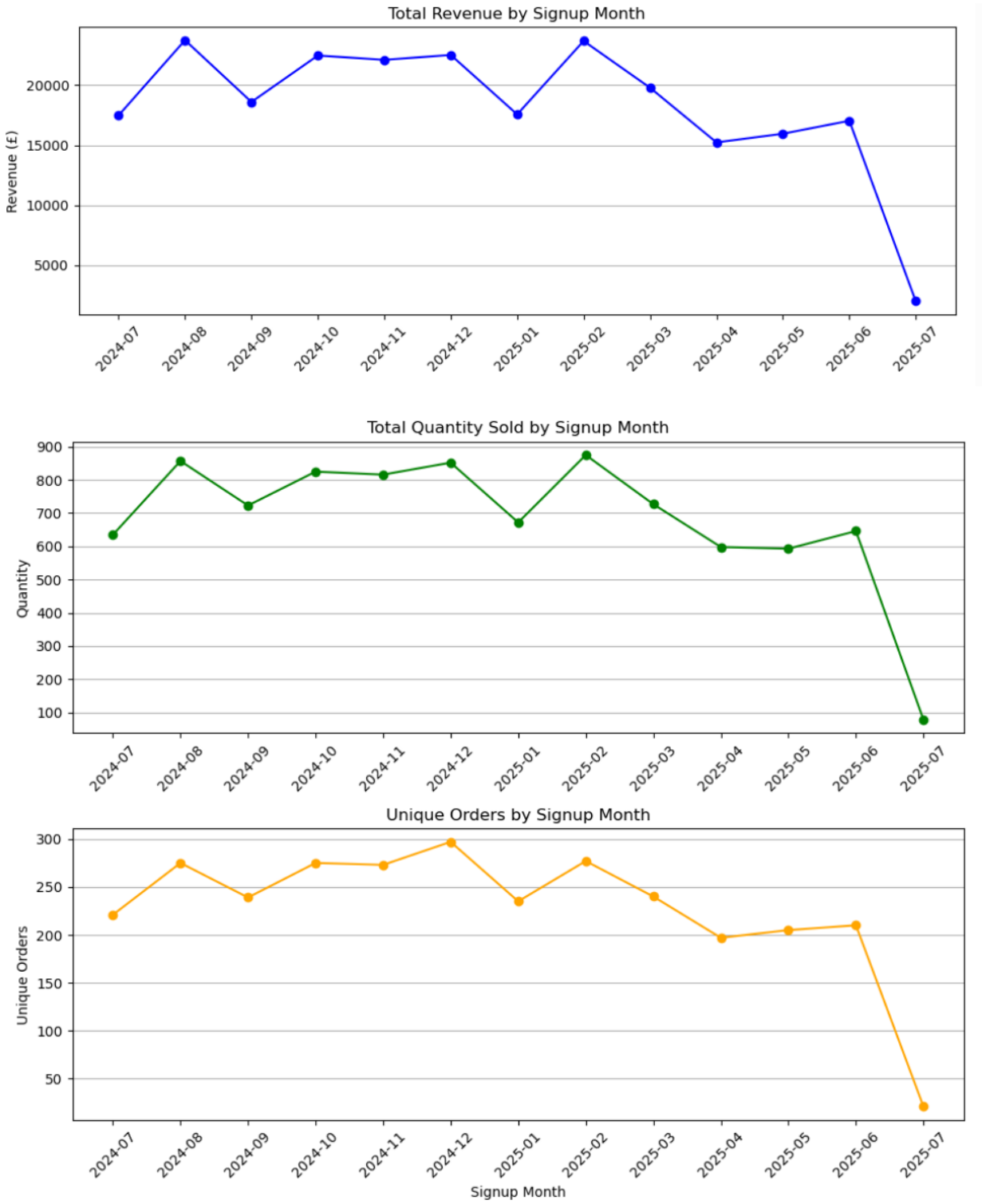


Fig: 5.5 Total Revenue, Quantity and Unique Orders for each signup month

6. Recommendations

1. **Target Promotions in High-Value Categories:** Focus marketing and discount efforts on the **'Personal Care'** and **'Cleaning'** categories, which show strong revenue performance, especially in the regions **Central** and **East** respectively.
2. **Enhance Delivery Performance:** Improve logistics and shipping processes in regions like **East**, North and **Central**, where delivery delays are notably higher (over 39%), to boost customer satisfaction and reduce cancellations.
3. **Loyalty Program Optimization:** Invest in strategies to encourage customers to move into the **Gold loyalty tier**, as this segment generates the highest revenue share (~57%).

7. Data Issues or Risks

- **Issue:** Inconsistent and overlapping **region** data between sales and customer datasets led to 2,410 mismatched entries.
  Possible Fix: Implement stricter data entry validation and automated cross-checks between datasets at the point of data capture to ensure consistent region naming and reduce manual errors.
- **Issue:** The entire sales dataset contains information for the same order date for all transactions, which prevents any meaningful analysis of trends over time (e.g., weekly or monthly sales patterns).
  Possible Fix: Ensure that future data captures accurate and varied order dates to enable time-series and trend analysis.