Project Coversheet

Full Name	Rory Scott
Email	Rdscott40ksx@hotmail.co.uk
Contact Number	+447590235745
Date of Submission	18/07/2025
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.

YOU CAN START YOUR PROJECT FROM HERE

Uptrail Internship Program Week 2 Project

Sales and Customer Behaviour Insights

By Rory Scott

Introduction

As part of the Data & Insights team at Green Cart Ltd., I was tasked with analysing sales and customer behaviour to support the company's Q2 performance review. Green Cart is a UK-based e-commerce business specialising in eco-friendly household products.

This report explores trends in revenue, product performance, customer loyalty, and delivery reliability across regions. The analysis is based on three datasets:

- sales_data.csv
- customer_info.csv
- product_info.csv

To complete the analysis, I used Python along with key data libraries: NumPy and Pandas for data cleaning, wrangling and manipulation; along with Seaborn and Matplotlib for data visualisation. The goal is to present findings in a clear, visual, and business-friendly format to support decision-making across marketing and operations.

Section 1: Data Cleaning Summary

📌 Part 1: Identifying Data Frame Structures

1. **sales_data**
   - 3000 rows and 10 columns
   - Primary Key – order_id
   - Foreign Key – customer_id
   - Foreign Key – product_id

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   order_id          2999 non-null   object
 1   customer_id       2998 non-null   object
 2   product_id        2995 non-null   object
 3   quantity          2997 non-null   object
 4   unit_price        2999 non-null   float64
 5   order_date        2997 non-null   object
 6   delivery_status   2997 non-null   object
 7   payment_method    2997 non-null   object
 8   region            3000 non-null   object
 9   discount_applied  2483 non-null   float64
dtypes: float64(2), object(8)
memory usage: 234.5+ KB
```

2. **customer_info**
   - 500 rows and 6 columns
   - Primary Key – customer_id

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 6 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   customer_id   497 non-null    object
 1   email         494 non-null    object
 2   signup_date   496 non-null    object
 3   gender        496 non-null    object
 4   region        497 non-null    object
 5   loyalty_tier  498 non-null    object
dtypes: object(6)
memory usage: 23.6+ KB
```

3. **product_info**
   - 30 rows and 6 columns
   - Primary Key – product_id

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   product_id     30 non-null     object
 1   product_name   30 non-null     object
 2   category       30 non-null     object
 3   launch_date    30 non-null     object
 4   base_price     30 non-null     float64
 5   supplier_code  30 non-null     object
dtypes: float64(1), object(5)
memory usage: 1.5+ KB
```

📌 Part 2: Handling Inconsistent Categorical Values

Several columns contained inconsistent categorical values due to variations in spelling, casing, and formatting (e.g., 'DELAYED', 'delrd', 'delyd'). These inconsistencies were identified across fields such as gender, category, region, loyalty_tier, payment_method, and delivery_status.

To address this, we applied:

- **Basic standardisation** using lowercase conversion and whitespace stripping.

```
🧼 'delivery_status': 1190 values changed via basic standardisation.
🔄 'delivery_status': 2 values changed via fuzzy matching
--------------------------------------------------------------------------------
🧼 'payment_method': 1485 values changed via basic standardisation.
🔄 'payment_method': 764 values changed via fuzzy matching
--------------------------------------------------------------------------------
🧼 'region': 1 values changed via basic standardisation.
🔄 'region': 1 values changed via fuzzy matching
--------------------------------------------------------------------------------
🧼 'gender': 269 values changed via basic standardisation.
🔄 'gender': 92 values changed via fuzzy matching
--------------------------------------------------------------------------------
🧼 'loyalty_tier': 382 values changed via basic standardisation.
🔄 'loyalty_tier': 5 values changed via fuzzy matching
--------------------------------------------------------------------------------
🧼 'region': 3 values changed via basic standardisation.
🔄 'region': 0 values changed via fuzzy matching
--------------------------------------------------------------------------------
🧼 'category': 0 values changed via basic standardisation.
🔄 'category': 0 values changed via fuzzy matching
--------------------------------------------------------------------------------
```

- **Fuzzy matching** techniques to group and unify similar but non-identical entries.

This process improved data consistency and prepared categorical fields for accurate downstream analysis.

📌 Part 3: Handling Datatypes

The following columns were successfully converted to appropriate data types:

- **quantity**: Converted to float64 — enabling accurate numerical analysis.

- **Date Fields (order_date, signup_date, launch_date)**:
  All converted to datetime64[ns], supporting reliable date-based operations.

Invalid or unparseable entries were safely coerced to NaN/NaT, maintaining the integrity of the dataset during transformation.

```
📊 'quantity' data type format converted to: float64

📊 'order_date' data type format converted to: datetime64[ns]

📊 'signup_date' data type format converted to: datetime64[ns]

📊 'launch_date' data type format converted to: datetime64[ns]
```

📌 Part 5: Handling Missing/Duplicate/Invalid Fields

A range of strategies were applied to handle missing and duplicate values across datasets:

```
🗑️ Dropped 1 rows with missing 'order_id'.
🗑️ Dropped 2 rows with missing 'customer_id'.
🗑️ Dropped 5 rows with missing 'product_id'.
⏭️ Skipped handling missing values in 'quantity'.
⏭️ Skipped handling missing values in 'unit_price'.
⏭️ Skipped handling missing values in 'order_date'.
🖍️ Filled 516 missing 'discount_applied' with '0.0'.
🗑️ Dropped 3 rows with missing 'customer_id'.
🖍️ Filled 6 missing 'email' with 'Unknown'.
⏭️ Skipped handling missing values in 'signup_date'.
⏭️ Skipped handling missing values in 'launch_date'.
⏭️ Skipped handling missing values in 'base_price'.
🧹 Sales Data: 2 duplicate rows dropped based on 'order_id'.
🧹 Customer Info: 0 duplicate rows dropped based on 'customer_id'.
🧹 Product Info: 0 duplicate rows dropped based on 'product_id'.
```

1. **Missing Values**

- **Identifier Columns** (order_id, customer_id, product_id):
  Rows with missing critical IDs were **dropped** to maintain referential integrity.
- **Categorical Fields** (loyalty_tier, gender, delivery_status, email, etc):
  Missing values in email were **filled with 'Unknown'** to retain record completeness without introducing bias.
- **Discount Field**:
  Missing values in discount_applied were **filled with 0.0**, assuming no discount was given where nothing was recorded.
- **Date Fields (**order_date, signup_date, launch_date):

  Rows with missing dates were dropped to ensure accuracy in time-based analysis and avoid misleading results.
- **Other Fields** (quantity, unit_price, base_price):
  These were **left unchanged** at this stage - to avoid introducing inaccurate assumptions.

2. **Duplicate Rows**

- **Sales Data**: 2 duplicate rows dropped based on order_id.
- **Customer & Product Info**: No exact duplicates found based on customer_id or product_id.

3. **Negative Numeric Values**

- Checked for negative values in numeric columns, 0 were found.

```
Number of corrected negative values:
sales_data.quantity: 0
sales_data.unit_price: 0
sales_data.discount_applied: 0
product_info.base_price: 0
```

These steps ensured that only clean, consistent data was retained for further analysis while avoiding overly aggressive imputation that could distort insights.

Section 2: Feature Engineering Summary

📌 New features Created / Why

- **revenue** - Tracking total income per order or customer. Enables analysis of high-revenue customers, best-selling products, and financial performance over time.
- **order_week** - Analysing weekly trends, seasonality, and campaign performance. Useful for understanding peak ordering periods and planning resources or marketing.
- **price_band** - Segmenting products or customers based on price tiers (e.g., low, mid, high). Helps analyse buying behavior, customer preferences, and value perception.
- **days_to_order** - Measuring time between signup and first order. Useful for analysing onboarding effectiveness, conversion delays, and identifying quick vs. slow converters.
- **email_domain** - Identifying B2B vs. B2C customers (e.g., Gmail vs. company domains). Enables segmentation for targeted marketing and communication strategies.
- **is_late** - Tracking delivery or order fulfillment delays. Supports operational efficiency analysis, customer satisfaction metrics, and identifying process bottlenecks.
- **sale_region / customer_region** – The merged dataset contained duplicate region columns from sales and customer data, so we renamed them for clarity and consistency.

Section 3: Key Findings & Trends

Part 1: Key Findings

Total Revenue Generated: $ 239224.63

Average Revenue per Order: $ 80.2

Total Number of Products sold: 30

Total Quantity of Products sold: 8956

Total Number of Orders: 2987

Average Quantity per Order: 3.0

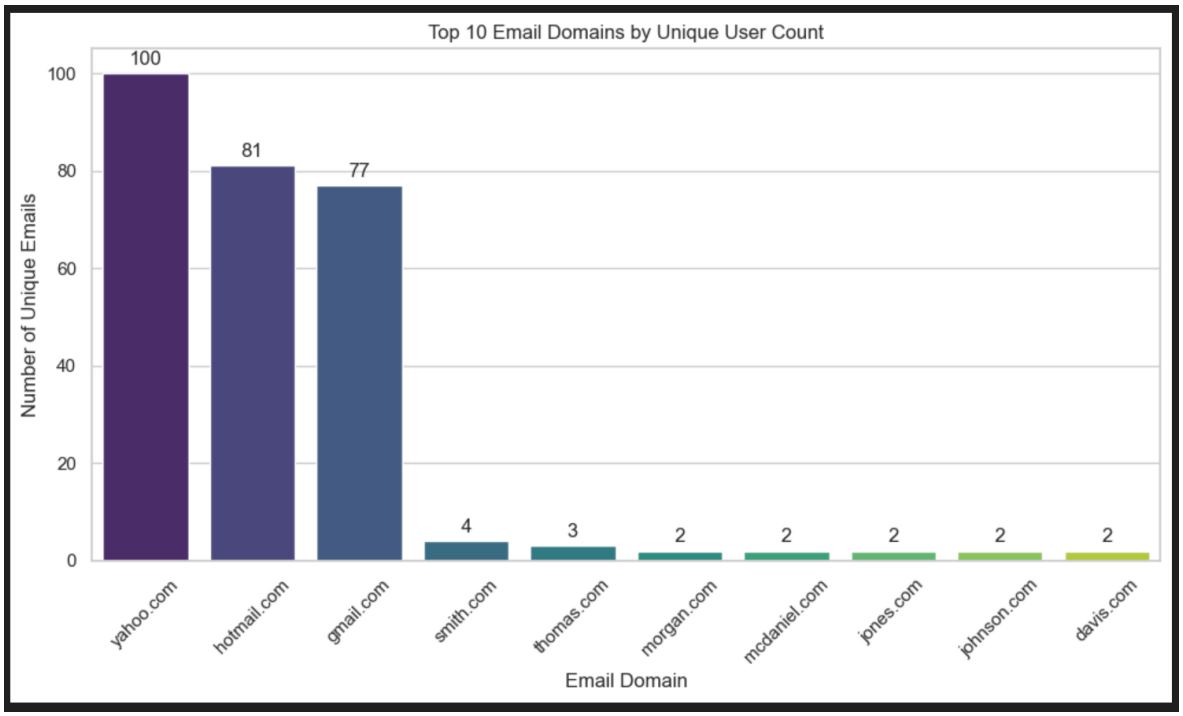
Total Number of Unique customers: 499

Total number of unique payment methods: 4

Part 2: Identified Trends

1. **Majority of Users Use Mainstream Email Providers**

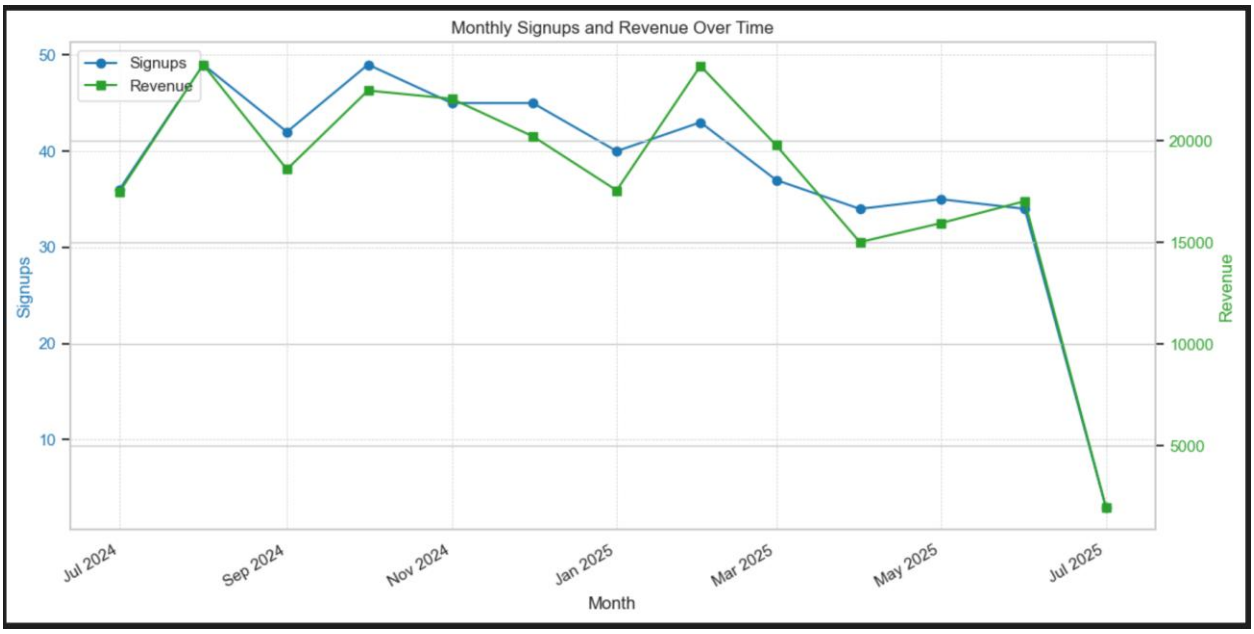
This shows that most users sign up with well-known email services, which could help



when planning email campaigns.

2. More Signups Lead to More Revenue

- A Pearson correlation analysis revealed a very strong, statistically significant relationship between user signups and revenue ($r = 0.967$, $p < 0.001$).
- This suggests that increases in customer acquisition are closely tied to increases in revenue, supporting a direct relationship between user growth and business performance.



Section 4: Stretch Task (Optional

📌 Part 1: Querying

```
Dataframe of customers that signed up in Q2 and made and order within 14 days with a discount of 0.2 or greater:  Empty DataFrame
Columns: [customer_id, signup_date, order_date, days_to_order, discount_applied]
Index: []
```

- Within 14 days of signup of users in Q2, there were no orders with a discount of 20% or greater.

📌 Part 2: MinMaxScaler

- The normalised_revenue column rescales raw revenue values to a 0–1 range, where 117.75 maps to **0.457** relative to the min and max revenue in the dataset.

	revenue	normalised_revenue
0	117.75	0.457
1	94.6	0.362
2	25.228	0.076
3	26.208	0.080
4	38.096	0.129

📌 Part 3: Under-Performing Product Flagging

Method:

- I grouped the dataset by product_id and calculated average revenue, quantity sold, average discount, and base price. I then computed z-scores on the average revenue and flagged any product with a z-score below -1.5 as a statistical underperformer.

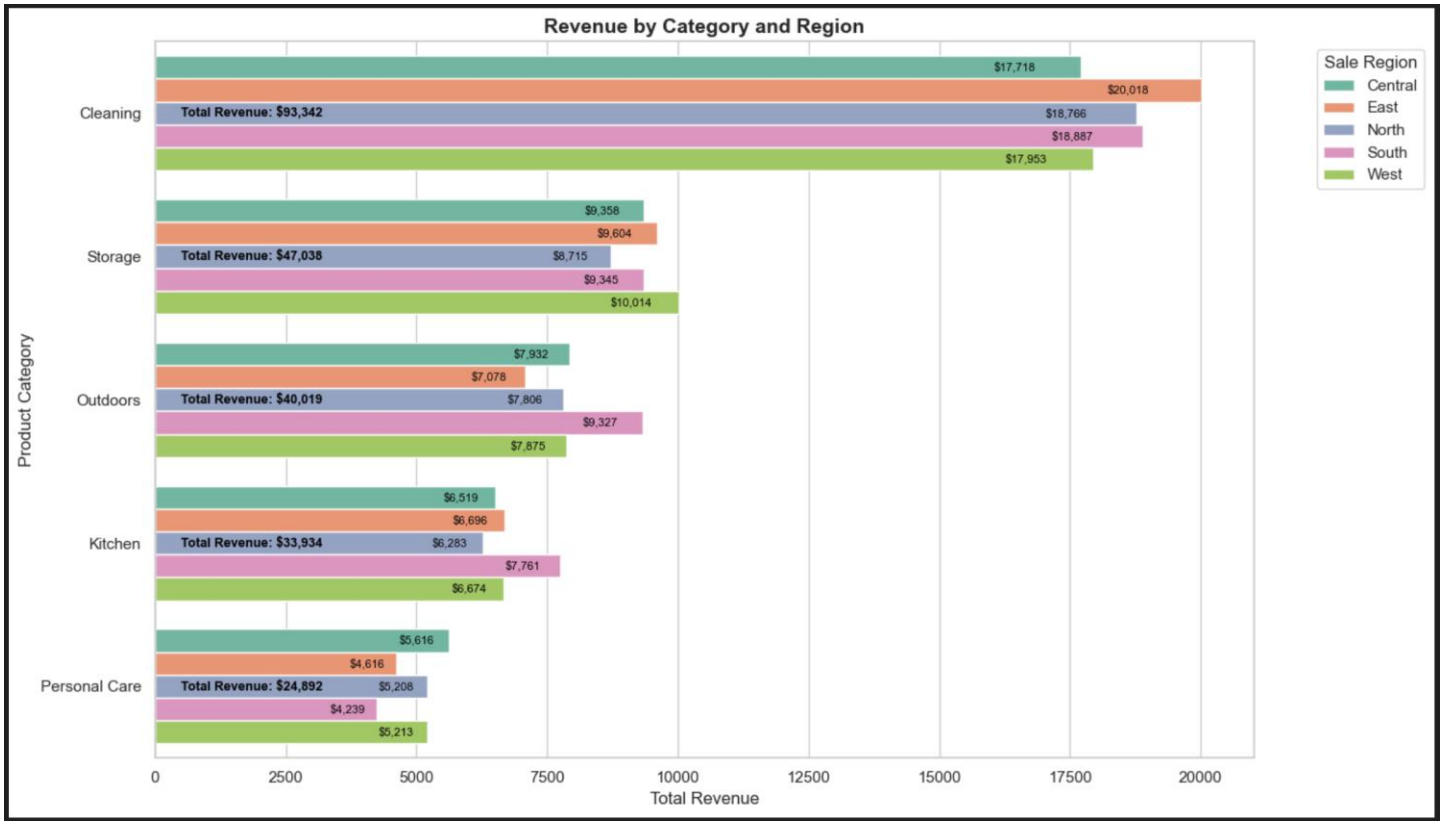
Conclusion:

- Products **P0008** and **P0012** were flagged as underperformers, with average revenues more than **2 standard deviations below the mean**, despite moderate discounts - indicating low perceived value or market demand.

	product_id	avg_revenue	total_revenue	total_quantity	avg_discount	base_price	revenue_z	statistical_underperformer
7	P0008	67.392172	6671.825	252	0.087879	31.96	-2.051538	True
11	P0012	66.592423	6459.465	265	0.084021	14.67	-2.179753	True

Section 5: Business Insights and Answers

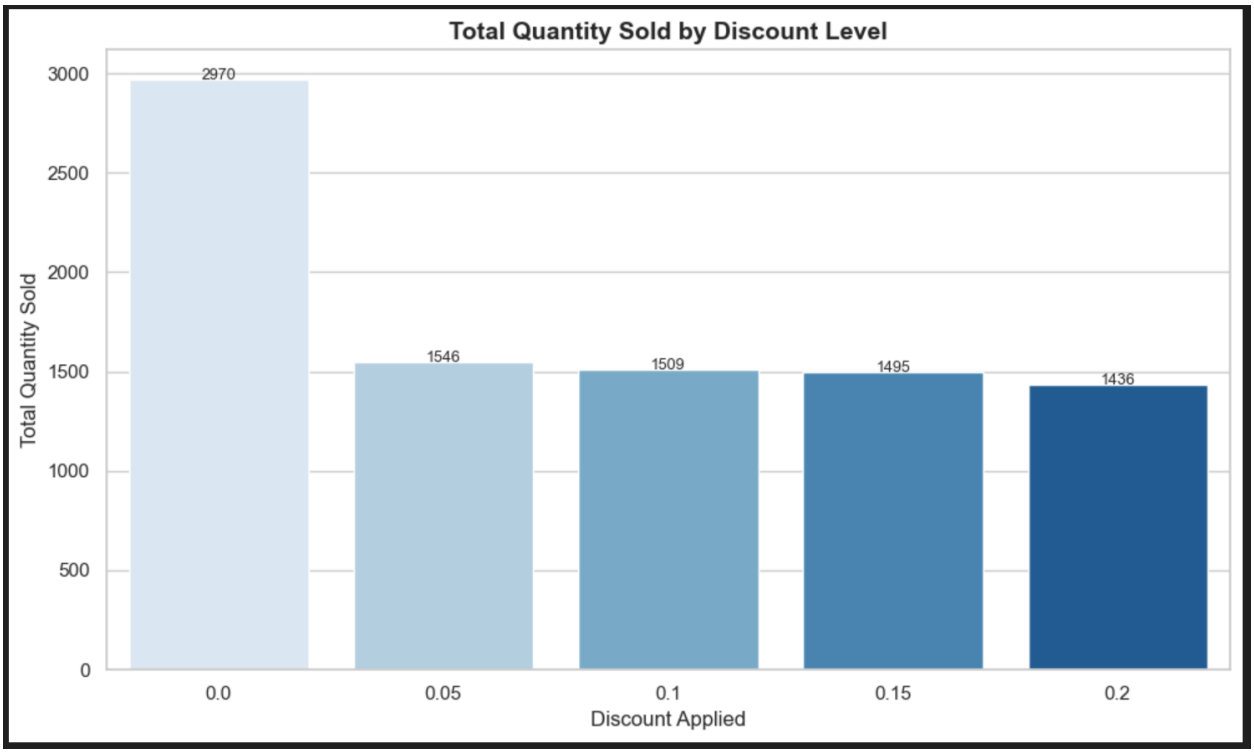
📌 Question 1: Which product categories drive the most revenue, and in which regions?



Insight:

Cleaning products dominate revenue across all regions, while regional differences are minimal, suggesting product category is a stronger driver of revenue than geography.

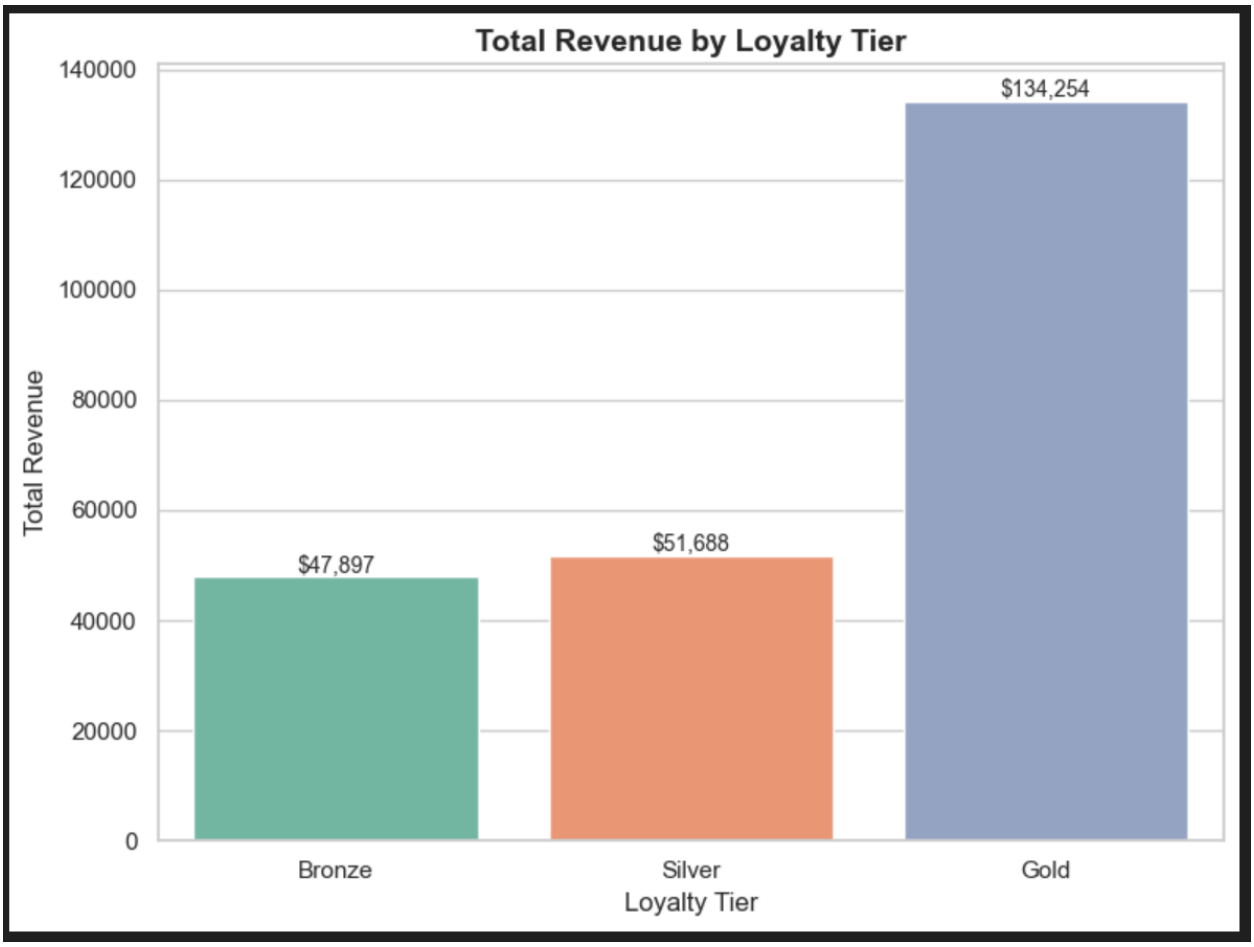
📌 Question 2: Do discounts lead to more items sold?



Insight:

- Discounts did not lead to more items **being sold. I**n fact, sales volume declined as discounts increased. This suggests discounts may be applied to lower-demand products rather than **being used** effectively to drive higher sales.

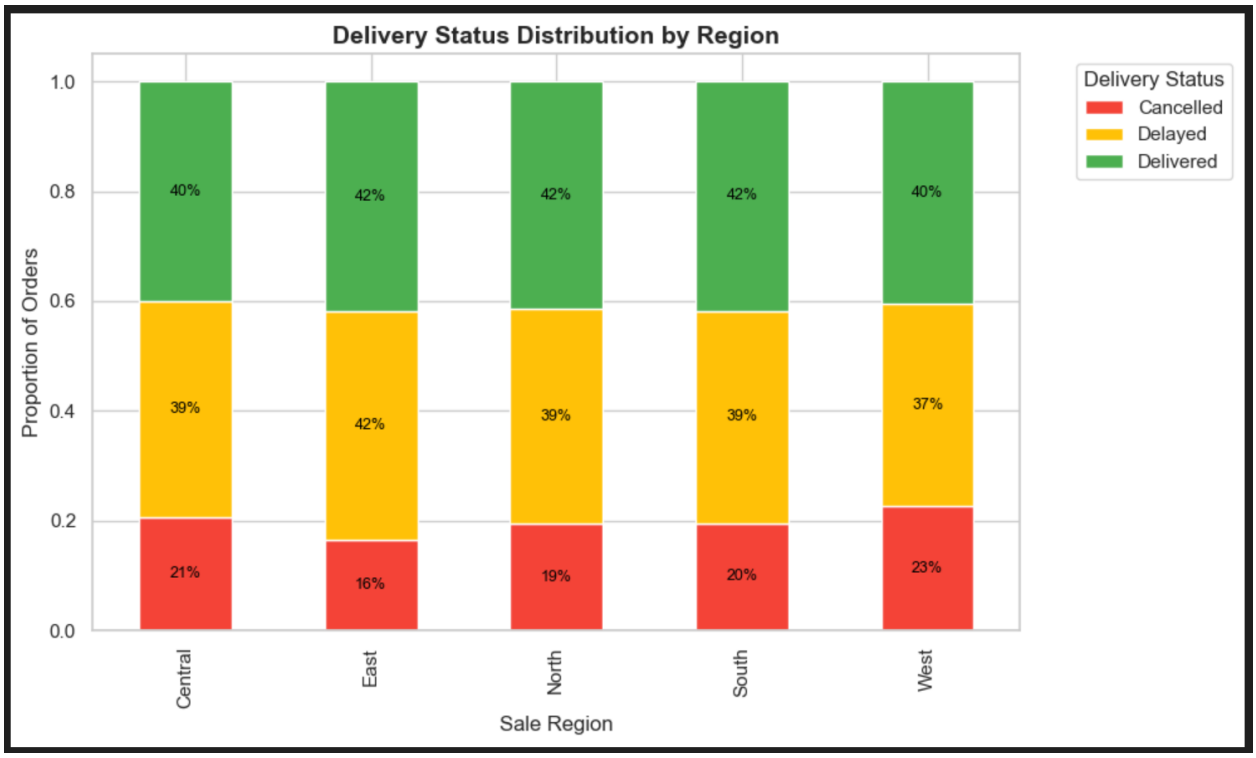
📌 Question 3: Which loyalty tier generates the most value?



Insight:

- Gold-tier customers are the most valuable segment - efforts to retain and grow this tier could yield the highest returns.

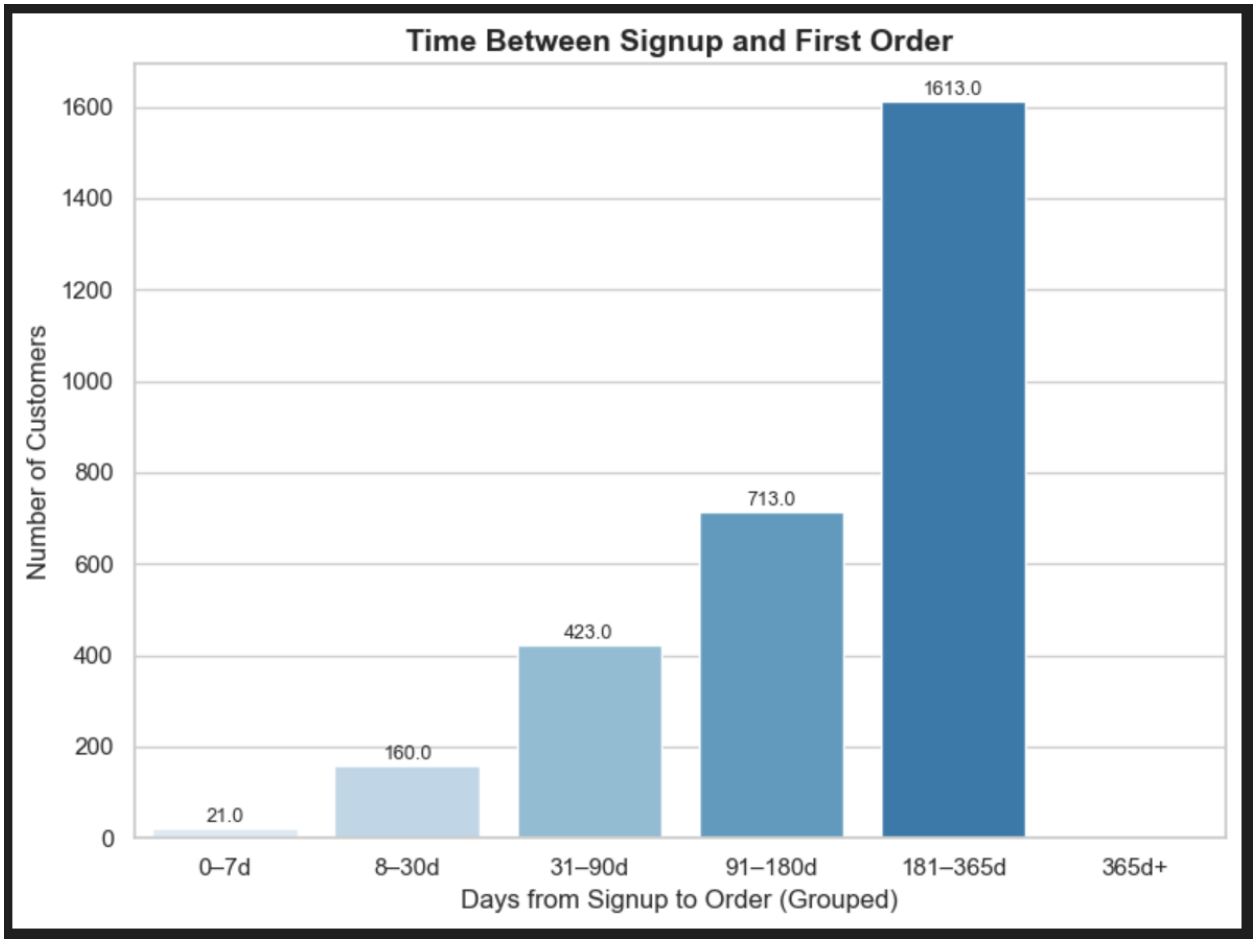
📌 Question 4: Are certain regions struggling with delivery delays?



Insight:

While there are **minor regional variations**, the **overall delivery performance is consistent**, suggesting that **delivery delays are not region-specific** but rather a system-wide behavior.

📌 Question 5: Do customer signup patterns influence purchasing activity?



Insight:

Most customers take over 6 months to place their first order, with very few converting early. This suggests delayed engagement and a potential opportunity to drive faster conversions.

Section 6: Business Recommendations, Data Issues and Risks

📌 Part 1: Business Recommendations

- **Prioritize High-Value Categories and Customers**
  Focus marketing and stock on top-performing categories like cleaning and invest in retaining Gold-tier customers who generate the most revenue.
- **Revamp Discounting and Onboarding Strategies**
  Current discounts don't drive volume - test strategic promotions on popular products. Also, improve early engagement to reduce long signup-to-purchase delays.
- **Fix Regional Fulfilment Gaps**
  The West region has elevated cancellation rates - review and enhance delivery operations to improve reliability and customer satisfaction.

📌 Part 2: Data Issues and Risks

1. **Time-Based Analysis Is Unreliable**
   All order dates are identical, and the overall time range is unclear, preventing valid trend, seasonality, or forecasting analysis.
2. **User Identity & Behaviour May Be Inaccurate**
   Missing email validation and unusual signup-to-order gaps suggest potential duplicate users or test data, which can distort user-level insights.
3. **Imputation May Skew Insights**
   Filling 516 missing discount values with 0 may have overstated the ineffectiveness of discounts, biasing conclusions around promotional impact.

[END OF PROJECT]