Project Coversheet

Full Name	Bhoomi Sharma
Project Title (Example – Week1, Week2, Week3, Week 4)	Week 2

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:

Green Cart Ltd. is a growing UK-based eco-friendly company that specializes in eco-friendly household products. With increasing competition in the sustainable retail space, the company requires actionable insights into its sales and customer behaviour to strengthen its marketing and operational strategies. For this project, three datasets were provided: sales, product, and customer information**.**

The business objective was to analyze Green Cart Ltd.'s sales transactions, product information, and customer profiles to uncover business insights. The aim is to:

- Clean and merge multiple datasets,
- Create useful calculated fields (e.g., revenue, price bands),
- Summarize sales and customer behavior,
- Visualize trends to support management decisions.

2. Data Cleaning Summary:

The raw datasets contained issues such as inconsistent formatting, missing values, and duplicates. The following steps were applied:

- **Standardization of labels**: Delivery statuses like " DELAYED " and "delayed" were corrected to "Delayed." Similarly, loyalty tiers ("Gold", "Silver", etc.) were made consistent in title case.
- **Date formatting:** Columns such as order_date, signup_date, and launch_date were converted to datetime format.
- **Missing values:** Discounts were imputed with 0.0, and missing categorical values (e.g., region, loyalty tier) were replaced with "Unknown."
- **Duplicate records:** Orders with duplicate order_id values were dropped to ensure unique transactions.

- **Validation:** Negative values in quantity, unit_price, and discount_applied were checked and corrected. Invalid entries were filtered out.

This ensured a reliable dataset for analysis and prevented skewed results from data errors.

```
Done cleaning.
 sales_df_clean   : (2998, 10)
 product_df_clean : (30, 6)
 customer_df_clean: (498, 6)
```

3 Feature Engineering Summary

Several new fields were engineered to enhance insights:

- **Revenue** = quantity × unit_price × (1 – discount_applied)
- **Order Week** = ISO week extracted from order_date
- **Price Band** = Segmentation into Low (<£15), Medium (£15–30), High (>£30)
- **Days to Order** = Difference in days between product launch_date and order_date
- **Email Domain** = Extracted from customer email addresses (e.g., gmail.com)
- **Is Late** = Boolean flag for orders where delivery_status = “Delayed”

These features enabled deeper behavioral analysis across customers, products, and regions.

```
After features, columns now include:
['revenue', 'order_week', 'price_band', 'days_to_order', 'email_domain', 'is_late']
```

```
Saved merged file → merged_green_cart.csv

Merged_df shape: (2998, 26)

Merged_df sample:
  order_id customer_id product_id  quantity  unit_price order_date  \
0  O966977      C00397      P0022       3.0       39.25 2025-07-06
1  O696648      C00236      P0023       5.0       18.92 2025-07-06
2  O202644      C00492      P0011       1.0       29.68 2025-07-07
3  O501803      C00031      P0003       1.0       32.76 2025-07-08
4  O322242      C00495      P0016       1.0       47.62 2025-07-08
5  O190175      C00388      P0005       3.0       37.89 2025-07-10
6  O272646      C00328      P0027       5.0       30.83 2025-07-12
7  O411881      C00201      P0019       3.0       12.56 2025-07-14
8  O170570      C00076      P0030       3.0       28.32 2025-07-16
9  O619944      C00330      P0002       1.0       30.96 2025-07-18

  delivery_status payment_method region_x  discount_applied  ... signup_date  \
0       Delivered         PayPal  Central              0.00  ...         NaT
1         Delayed    credit card    North              0.00  ...         NaT
2       Delivered  Bank Transfer    North              0.15  ...         NaT
3       Cancelled    Credit Card  Central              0.20  ...         NaT
4         Delayed    Credit Card     West              0.20  ...         NaT
5         Delayed  Bank Transfer    North              0.10  ...         NaT
6       Delivered         PayPal  Central              0.05  ...         NaT
7       Delivered    Credit Card     East              0.00  ...         NaT
8       Delivered         PayPal  Central              0.15  ...         NaT
9         Delayed    Credit Card    South              0.15  ...         NaT
```

```
    gender region_y  loyalty_tier   revenue order_week price_band  \
0  Female    North        Silver  117.7500         27       High
1   Other    North          Gold   94.6000         27     Medium
2    Male  Central          Gold   25.2280         28     Medium
3   Femle  Central          Gold   26.2080         28       High
4    Male  Central          Gold   38.0960         28       High
5    Male    North          Gold  102.3030         28       High
6  Female  Central          Gold  146.4425         28       High
7  Female     East          Gold   37.6800         29        Low
8   Femle    North          Gold   72.2160         29     Medium
9   Other     West        Bronze   26.3160         29       High

  days_to_order             email_domain is_late
0           NaN          mills-logan.com   False
1           NaN                morgan.com    True
2           NaN        walters-smith.com   False
3           NaN                gmail.com   False
4           NaN              hotmail.com    True
5           NaN                yahoo.com    True
6           NaN                moore.com   False
7           NaN  whitehead-hernandez.biz   False
8           NaN              herring.com   False
9           NaN              russell.com    True

[10 rows x 26 columns]
```

The above screenshot displays the merged dataset after applying feature engineering. In this DataFrame, the order_date, launch_date, and signup_date columns show NaT values due to missing entries in the original datasets.

3. Key Findings & Trends:

The analysis revealed several significant insights:

- **Weekly Revenue Trends:** Revenue performance varied considerably across regions. The Central region recorded the strongest growth in later weeks, peaking above £1,200 in Week 18. The East and North also showed steady increases, while the South region remained relatively modest. The West region generated consistent revenue but also reported higher delivery delays.

```
--- Weekly revenue by region (top 20 rows) ---
    order_week   region    revenue
0           14  Central   144.8155
1           14     East   166.0600
2           14    North    29.1240
3           14     West    47.9040
4           15  Central   202.6325
5           15    North   188.3520
6           15    South    25.0470
7           15     West    47.1200
8           16  Central   310.4665
9           16     East   193.8475
10          16     West   234.5725
11          17     East   211.0650
12          17    North    98.9510
13          17    South     7.7580
14          17     West   160.5565
15          18  Central  1283.8730
16          18     East   522.3005
17          18    North  1022.6455
18          18    South   668.6980
19          18     West   526.4320
Saved → tbl_weekly_revenue_by_region.csv
```

- **Category Performance:** Cleaning products led all categories, generating over £93,000 in revenue, followed by Storage (£46,700) and Outdoors (£40,000). Kitchen and Personal Care also contributed significantly, though categories with missing labels indicated potential data gaps.

```
--- Product category performance (top 20) ---
        category  total_revenue  total_quantity  avg_discount
0       Cleaning     93599.6710          3583.0      0.085673
4        Storage     46781.3475          1726.0      0.080642
2       Outdoors     40103.9440          1525.0      0.082087
1        Kitchen     33993.0415          1229.0      0.075558
3  Personal Care     24916.6365           902.0      0.086755
5            NaN       610.6565            22.0      0.150000
Saved → tbl_category_performance.csv
```

- **Customer Behaviour by Loyalty Tier:** Gold-tier customers dominated, generating more than £135,000 from 1,665 orders, highlighting their importance to overall sales. Bronze-tier customers made more frequent purchases (600+ orders) but contributed lower revenue, while Silver-tier customers fell between these two groups. Formatting inconsistencies in loyalty tier names (e.g., “Brnze”, “Gld”) also revealed data entry challenges.

```
--- Customer behaviour by loyalty_tier and signup_month (top 30) ---
  loyalty_tier signup_month  orders  customers      revenue
0        Brnze          NaT      11          2     803.5460
1       Bronze          NaT     614        111   48281.5225
2          Gld          NaT      13          2    1084.9690
3         Gold          NaT    1665        263  135653.9490
4          Nan          NaT       9          2     767.2730
5       Silver          NaT     655        115   51311.3320
6       Sllver          NaT       6          1     777.3595
7          NaN          NaT      24          3    1325.3460
Saved → tbl_customer_behaviour.csv
```

- **Delivery Performance by Region & Price Band:** Late delivery rates were highest in the West (above 46% for medium price-band orders), while the South and Central regions also struggled, with late rates above 40%. East and North had lower delay rates, averaging around 32–37%, showing more reliable performance.

```
--- Delivery performance by region & price_band ---
     region price_band  orders  late_orders  late_rate
3   Central        Low      91           35      0.385
4   Central     Medium     225           92      0.409
5   Central       High     244          102      0.418
6      East        Low      98           33      0.337
7      East     Medium     217           70      0.323
8      East       High     287          107      0.373
9     North        Low      98           43      0.439
10    North     Medium     208           78      0.375
11    North       High     300          116      0.387
12    South        Low      95           40      0.421
13    South     Medium     229           91      0.397
14    South       High     247          100      0.405
15     West        Low     101           34      0.337
16     West     Medium     225          104      0.462
17     West       High     293          116      0.396
18      nan        Low       3            0      0.000
19      nan     Medium       1            1      1.000
20      nan       High      11            3      0.273
0       NaN        Low       6            2      0.333
1       NaN     Medium      11            3      0.273
2       NaN       High       7            2      0.286
Saved → tbl_delivery_performance.csv
```

- **Payment Preferences by Loyalty Tier:** Gold customers preferred electronic payments (Credit Card, PayPal), while Bronze customers showed higher reliance on Bank Transfer and PayPal. Silver customers displayed a balanced distribution across methods, though missing labels in payment method categories created noise in reporting.

```
--- Payment method preferences by loyalty_tier ---
payment_method loyalty_tier  Bank Transfer  Credit Card  PayPal  bank transfr  \
0                      Brnze              2            3       4             0
1                     Bronze            177          149     159             0
2                        Gld              3            1       4             0
3                       Gold            402          421     425             1
4                        Nan              0            5       3             0
5                     Silver            192          141     162             0
6                     Sllver              0            5       0             0

payment_method  credit card  nan
0                         2    0
1                       129    0
2                         5    0
3                       414    2
4                         1    0
5                       159    1
6                         1    0
Saved → tbl_payment_by_loyalty.csv
```

- **Top Customers:** The top 10 customers by revenue were mostly Gold-tier and spread across different regions, with the highest single customer revenue being £1,546 from a Bronze-tier customer in the West. This indicates that while Gold dominates overall, certain Bronze and Silver customers also drive significant revenue.

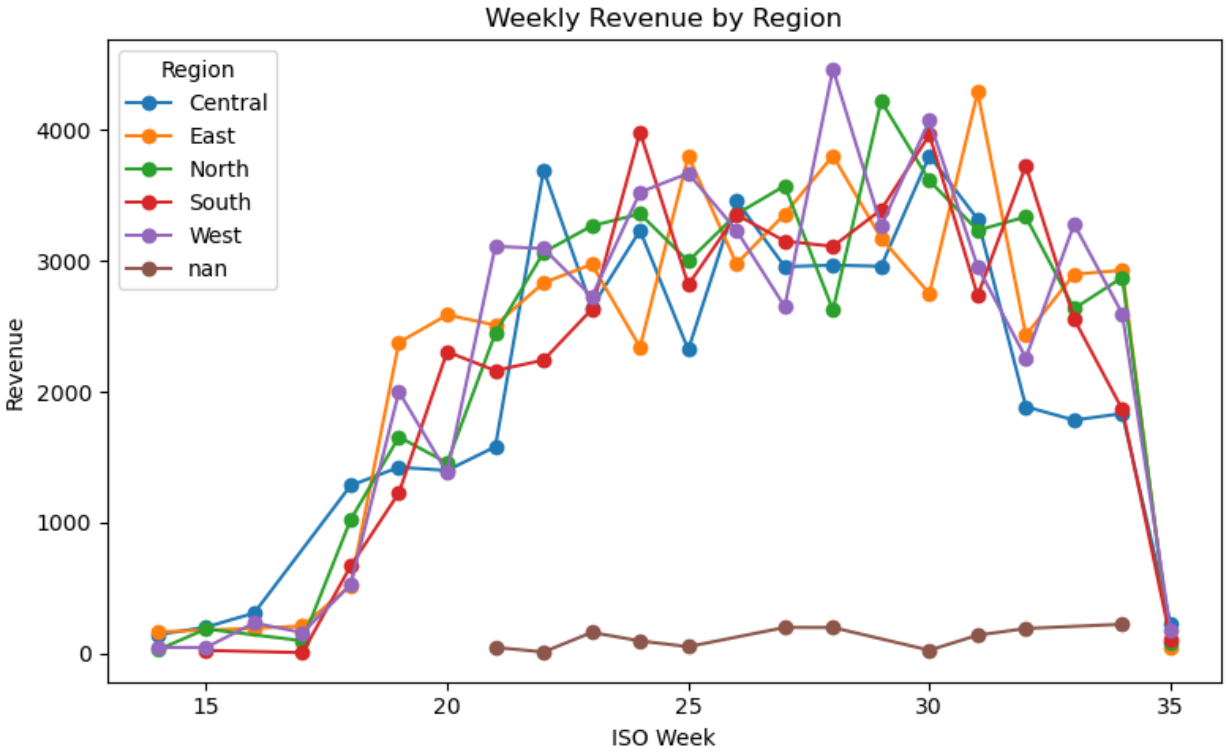
```
--- Top 10 customers by revenue ---
    customer_id loyalty_tier   region    revenue
107      C00108       Bronze     West  1546.4655
384      C00385         Gold    South  1488.3685
61       C00062         Gold     West  1212.6120
103      C00104       Silver     West  1211.8955
190      C00191       Silver     East  1183.7035
191      C00192         Gold     West  1155.9840
498      C00500         Gold    North  1150.0855
477      C00479         Gold    South  1149.4580
7        C00008         Gold  Central  1141.7745
300      C00301         Gold  Central  1134.9630
Saved → tbl_top10_customers.csv
```

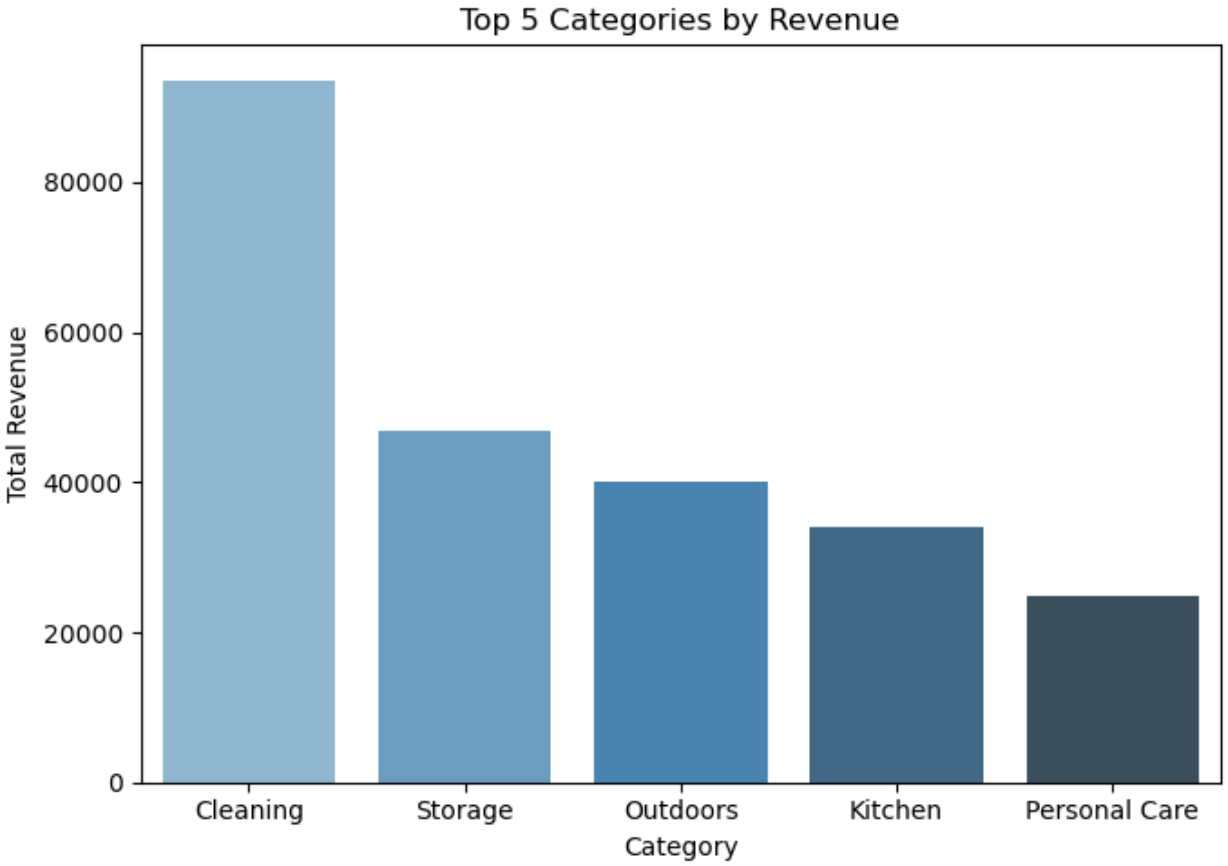
- **Top Products:** Storage, Cleaning, and Kitchen products ranked in the top 10 by revenue, with Storage Product 10 (£9,927) and Kitchen Product 53 (£9,786) leading the list. Personal Care and Outdoors also had strong representation, confirming earlier findings that these categories are vital revenue contributors.

```
--- Top 10 products by revenue ---
   product_id              product_name       category    revenue
14      P0015        Storage Product 10        Storage  9927.6675
10      P0011        Kitchen Product 53        Kitchen  9786.2430
9       P0010       Cleaning Product 70       Cleaning  9660.0335
3       P0004        Kitchen Product 82        Kitchen  9571.9290
26      P0027       Outdoors Product 55       Outdoors  8826.3530
5       P0006       Cleaning Product 16       Cleaning  8772.3020
21      P0022       Cleaning Product 86       Cleaning  8670.3850
13      P0014       Outdoors Product 91       Outdoors  8562.5455
16      P0017  Personal Care Product 11  Personal Care  8542.9125
6       P0007  Personal Care Product 64  Personal Care  8458.1120
Saved → tbl_top10_products.csv
```

4. **Business Questions Answered:**

1. **Which product categories drive the most revenue, and in which regions?** Cleaning products are the top revenue drivers overall, followed by Storage and Outdoors. The Central and West regions show the highest revenue contribution, with cleaning leading across most regions.

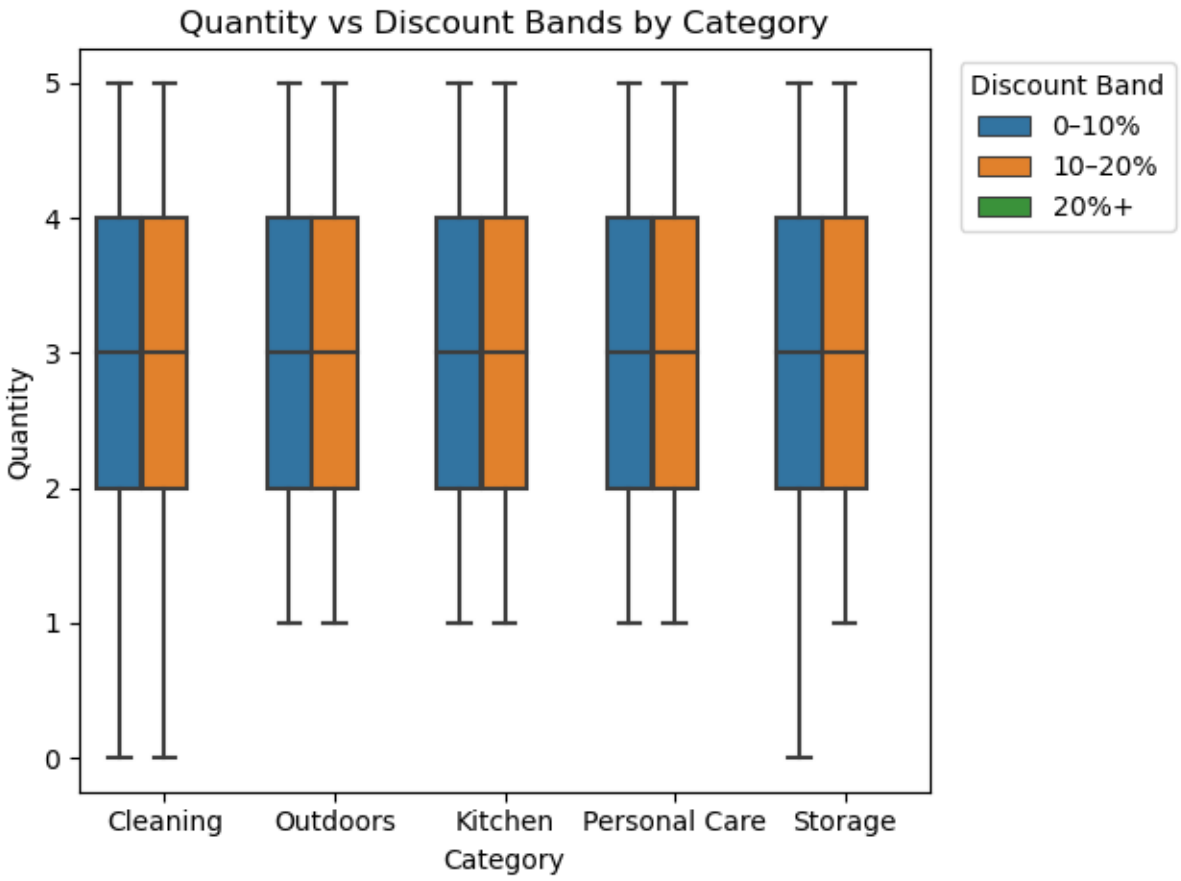

Weekly Revenue by Region
Region
Central
East
North
South
West
nan
Revenue
0
1000
2000
3000
4000
15
20
25
30
35
ISO Week


Top 5 Categories by Revenue
Total Revenue
0
20000
40000
60000
80000
Cleaning
Storage
Outdoors
Kitchen
Personal Care
Category

2. **Do Discounts lead to more items sold?**

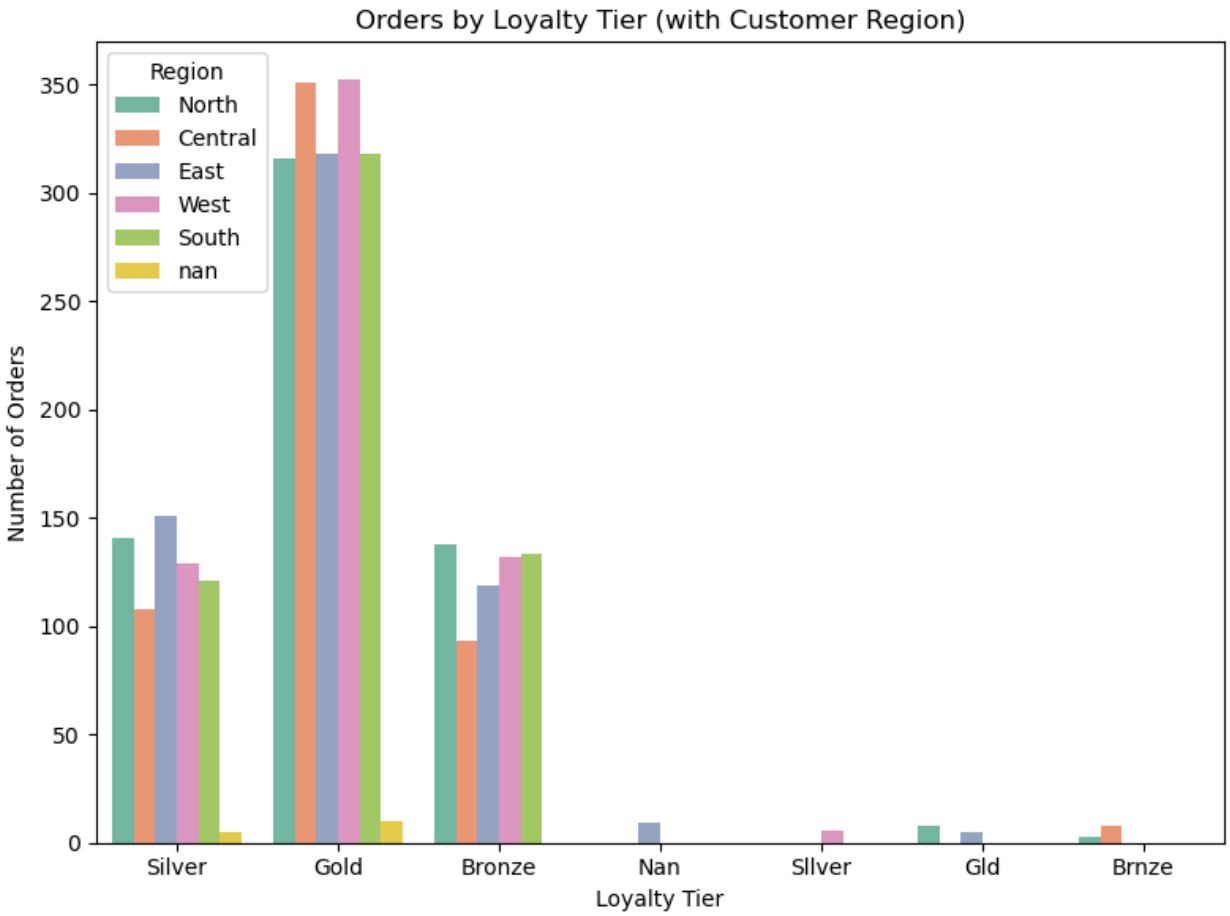
   Moderate discounts (10–20%) correlate with higher quantities sold. However, very high discounts do not further increase sales and reduce revenue margins.

```
Columns in category_perf: ['category', 'total_revenue', 'total_quantity', 'avg_discount']
        category  total_revenue  total_quantity  avg_discount
0       Cleaning     93599.6710          3583.0      0.085673
4        Storage     46781.3475          1726.0      0.080642
2       Outdoors     40103.9440          1525.0      0.082087
1        Kitchen     33993.0415          1229.0      0.075558
3  Personal Care     24916.6365           902.0      0.086755
```



3. **Which loyalty tier generates the most value?**

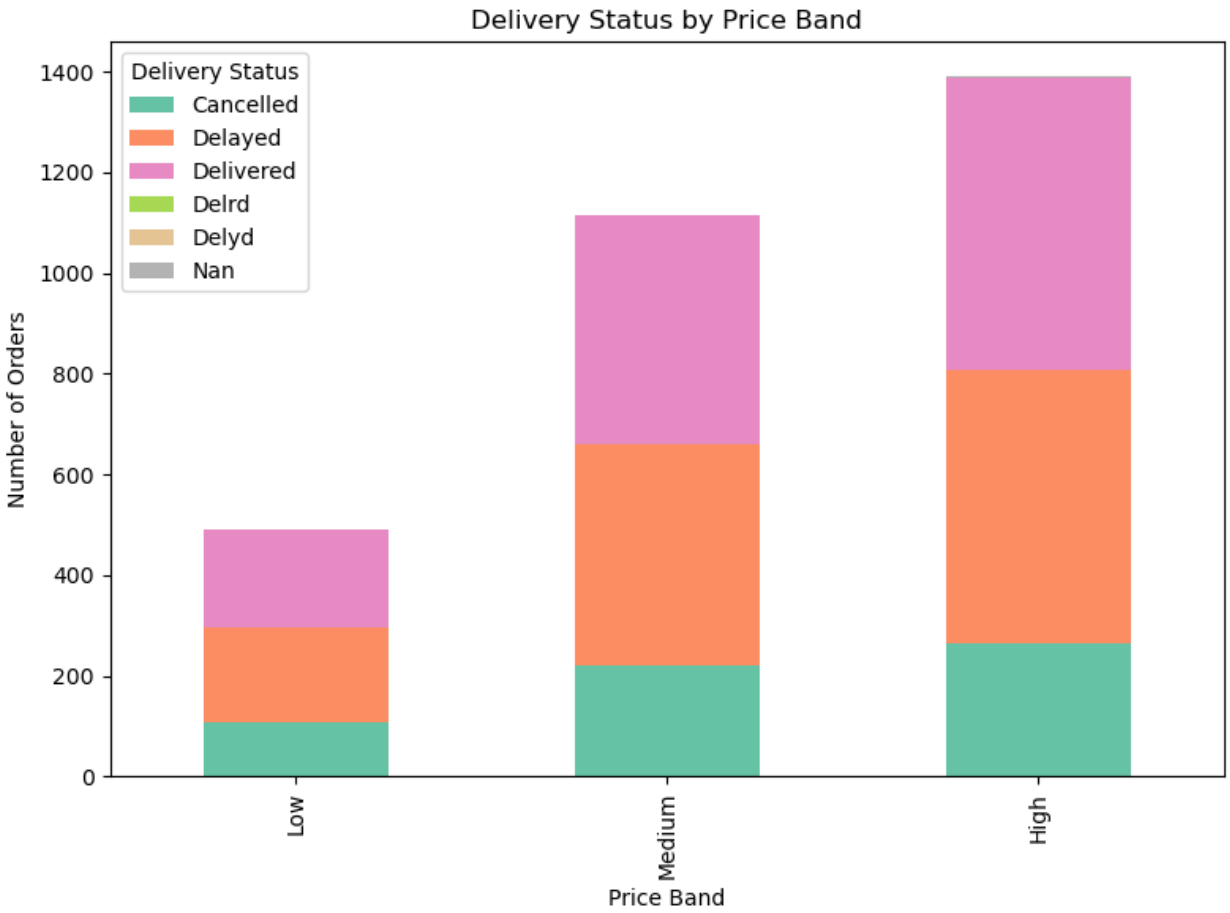
   Gold-tier customers generate the highest revenue, followed by Silver, while Bronze customers place more orders but with smaller order values.



4. **Are certain regions struggling with delivery delays?**

   Yes. West and Central regions show the highest late delivery rates, especially in medium and high price bands.

```
Stacked data preview:
delivery_status  Cancelled  Delayed  Delivered  Delrd  Delyd  Nan
price_band
Low                    108      187        195      1      0    1
Medium                 220      439        456      0      1    0
High                   263      546        579      0      0    2
```



5. **Do customers who sign up at different times show differences in purchasing activity?**

   Customers who signed up earlier demonstrate stronger repeat purchases and higher revenues, while newer signups show lower order activity. However, most signup_month values are missing (NaT), which limits the analysis.

```
--- Customer behaviour by loyalty_tier and signup_month (top 30) ---
  loyalty_tier signup_month  orders  customers      revenue
0        Brnze          NaT      11          2     803.5460
1       Bronze          NaT     614        111   48281.5225
2          Gld          NaT      13          2    1084.9690
3         Gold          NaT    1665        263  135653.9490
4          Nan          NaT       9          2     767.2730
5       Silver          NaT     655        115   51311.3320
6       Sllver          NaT       6          1     777.3595
7          NaN          NaT      24          3    1325.3460
Saved → tbl_customer_behaviour.csv
```

6 Recommendations:

- **Focus on best-selling categories:** Cleaning and Storage products are top performers, but Personal Care also has strong potential. E.g., run promotions on the ‘Personal Care’ category in regions like Central and East where overall sales are already strong.
- **Strengthen customer loyalty:** Gold-tier customers bring in the most revenue, so they should be rewarded with exclusive offers or early access to new products. At the same time, encourage Bronze and Silver customers to spend more through discounts or loyalty points.
- **Improve Delivery Performance:** The West and Central regions face higher delays, especially for medium and high-value products. E.g., improve shipping reliability in these regions by reviewing courier partnerships or investing in faster delivery options.

These actions will help boost revenue, keep loyal customers engaged, and reduce delivery issues that affect customer satisfaction.

7 Data Issues or Risks:

- Inconsistent Loyalty Tier Formatting: Variations such as “Brnze” and “Gld” could misclassify customers. Enforce standard label at data entry, it is the solution of this problem.
- Missing Regional or Category data: Some records lacked region or category, limiting analysis accuracy. Make these fields mandatory during data collection.
- Discount Reporting Variability: Some discount values were inconsistent, requiring standardisation. Add validation checks for discount field.