# Project Coversheet

| | |
|---|---|
| Full Name | Robert May |
| 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.

## YOU CAN START YOUR PROJECT FROM HERE

### 1. Introduction

The task was to investigate sales and customer behaviour across regions and product lines for Green Cart Ltd. There were sales, product and customer datasets, and this project aims to analyse patterns and performance, and present insights.

### 2. Data Cleaning Summary

The three datasets were loaded into dataframes in Python. Inconsistent data columns such as delivery_status and payment_method were standardised, and dates were converted to ‘datetime’. In the customer dataset, loyalty_tier and gender values were standardised.

Missing values in order_id and customer_id were initially set to Not known. However, as these identifiers could potentially be found using other available data, they were changed to Missing to avoid introducing incorrect assumptions. Empty date fields were left unchanged, as this is considered acceptable for analysis.

Most other null values were standardised to Missing. An exception was discount_applied, which was set to 0 where missing. The discount_applied field was also converted to a numeric data type to ensure consistency and enable calculations. No missing values were identified in the product dataset.

In the customer dataset, missing customer_id and email values were filled with Missing. Empty date fields were left unchanged. Within the gender field, four entries labelled as “Nan” were identified as valid string values rather than true nulls; these were recoded as Unknown for clarity. All other missing values in the customer dataset were also updated to Unknown, in line with the task specification.

The data summary is shown below. There were major issues with some columns as mentioned earlier.

| **sales** | Missing values | % missing | duplicates | inconsistent data | % |
|---|---:|---:|---:|---:|---:|
| order_id | 1 | 0.03% | 2 | 0 | |
| customer_id | 2 | 0.07% | | 0 | |
| product_id | 5 | 0.17% | | 0 | |
| quantity | 3 | 0.10% | | 0 | |
| unit_price | 1 | 0.03% | | 6 | 0.20% |
| **order_date** | 1820 | 60.67% | | 30 | 1.00% |
| **delivery_status** | 0 | 0.00% | | 1187 | 39.57% |

| | | | | | |
|---|---|---|---|---|---|
| **payment_method** | 3 | 0.10% | | 719 | 23.97% |
| region | 0 | 0.00% | | 2 | 0.07% |
| | | | | | |
| **customer** | | | | | |
| signup_date | 4 | 0.80% | | | |
| **loyalty_tier** | 9 | 1.80% | | 119 | 23.80% |
| **gender** | 4 | 0.80% | | 265 | 53.00% |
| **payment_method** | 3 | 0.60% | | 719 | 143.80% |

### Duplicates

In the sales dataset, two duplicate order_id values were identified. As these rows differed in other fields, they were not removed. Instead, they were flagged as Duplicate or Unique to allow for further investigation.

Duplicate customer_id values were expected and represent repeat customers. Duplicate values in other columns were also considered valid. No duplicate rows were found in the product dataset.

Missing values in the quantity field were investigated by comparing sales data with product data. However, the product dataset contains a base_price field, while the sales dataset uses unit_price. Attempts to find quantity by comparing with base_price resulted in no whole values. As a result, these missing quantities were flagged for further investigation.

The sales and product datasets were successfully merged using a left join. When the resulting dataset was merged with the customer dataset, four additional rows were created. This was traced back to three rows with missing customer_id values that had previously been flagged. To preserve join integrity, these rows were removed prior to merging. Following this step, the merged dataset contained the expected 3,000 rows. The final table was inspected using .head() and .info(), and no further issues were identified.

## 3. Feature Engineering Summary

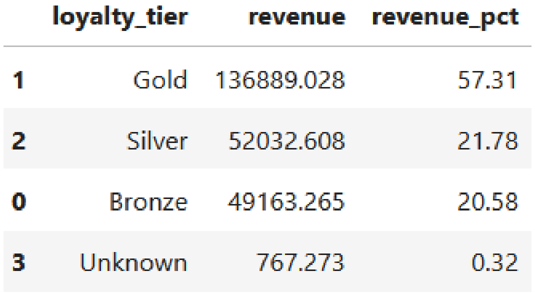
The following new features/columns were created:

- Revenue: quantity * unit_price * 1-discount_applied.
- order_week, using the week number from the ISO.
- Price_band, categorizing unit price as Low, (<£15), Medium (£15–30), High (>£30).
- Days_to_order = order date – launch date
- email_domain

- is_late: true if delivery status is delayed
- Missing emails were changed to unknown@unknown.com to extract domains.

4. Key findings and trends

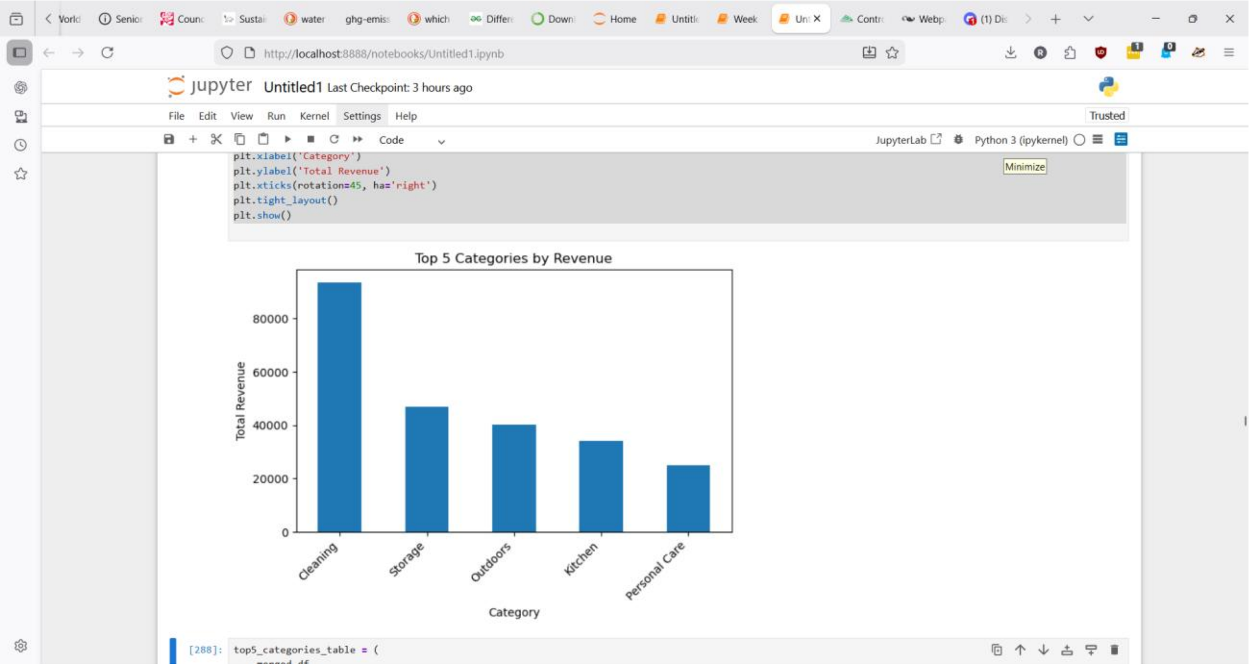
- Gold tier brought in more than twice as much revenue as any other tier. 57.3% compared to 21.8 for Silver.



| | loyalty_tier | revenue | revenue_pct |
|---|---|---|---|
| **1** | Gold | 136889.028 | 57.31 |
| **2** | Silver | 52032.608 | 21.78 |
| **0** | Bronze | 49163.265 | 20.58 |
| **3** | Unknown | 767.273 | 0.32 |

- Storage Product 10 brought in the most revenue, with Kitchen Product 53 next. In general, Cleaning products brought in more revenue than any other category. 39%, more than double the next highest which was Storage.

Figure 1

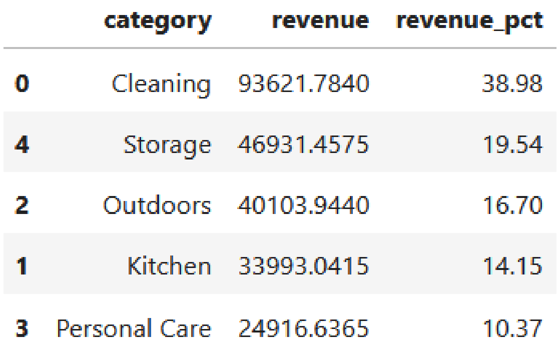


http://localhost:8888/notebooks/Untitled1.ipynb
jupyter Untitled1 Last Checkpoint: 3 hours ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Minimize

```
plt.xlabel('Category')
plt.ylabel('Total Revenue')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

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

```
[288]: top5_categories_table = (
```

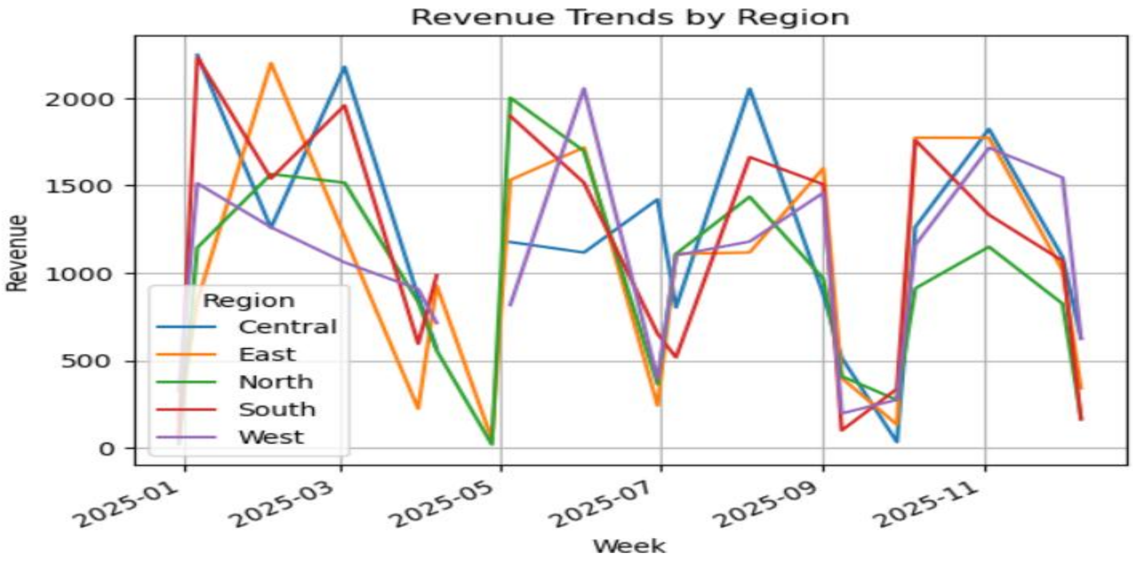


| | category | revenue | revenue_pct |
|---|---|---|---|
| **0** | Cleaning | 93621.7840 | 38.98 |
| **4** | Storage | 46931.4575 | 19.54 |
| **2** | Outdoors | 40103.9440 | 16.70 |
| **1** | Kitchen | 33993.0415 | 14.15 |
| **3** | Personal Care | 24916.6365 | 10.37 |

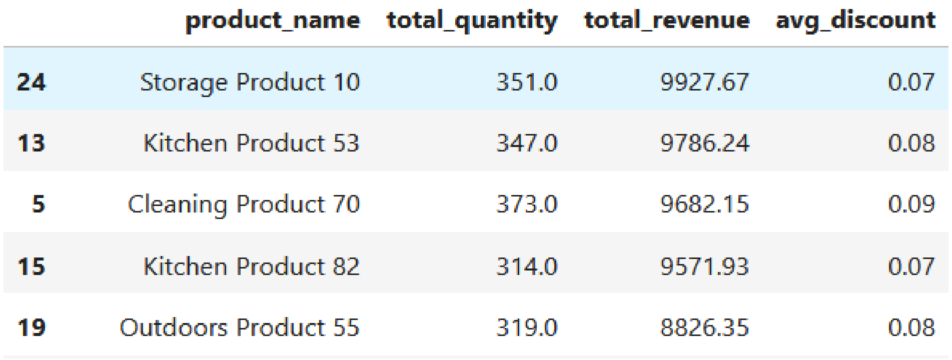
Summary tables: Pivot tables were found to give more clear results, so these were used in each instance.

Weekly revenue trends by region: You can see that for two weeks in the year, almost nothing is ordered. Perhaps this is a data issue or the reason for this is already known. It's also true that the labelling on the x axis is for monthly strangely, however the data is correct apart from this.

Figure 2



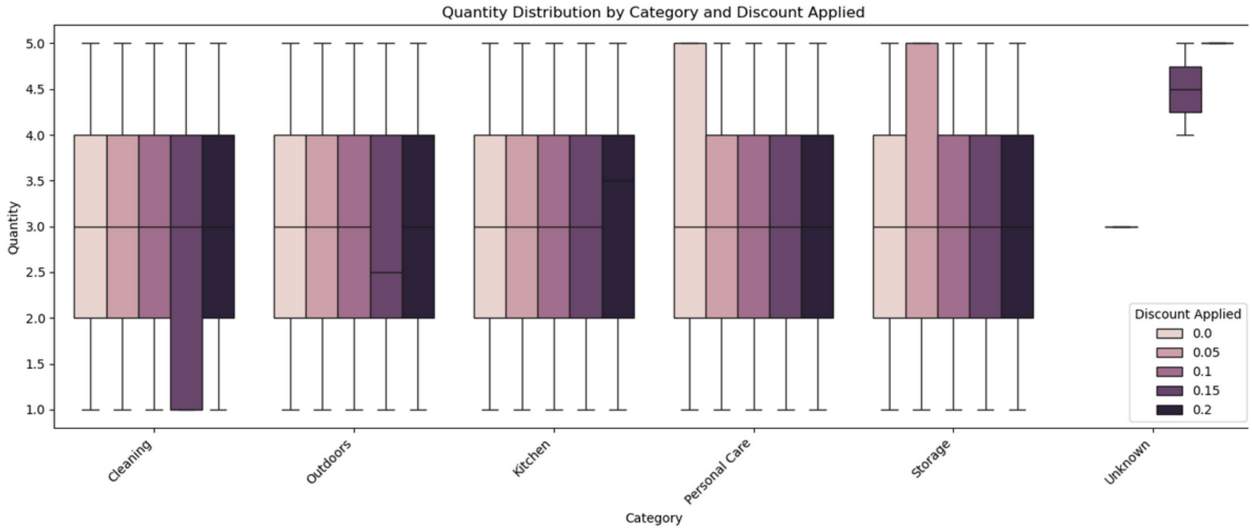
Product category performance:



|  | product_name | total_quantity | total_revenue | avg_discount |
|---|---|---|---|---|
| **24** | Storage Product 10 | 351.0 | 9927.67 | 0.07 |
| **13** | Kitchen Product 53 | 347.0 | 9786.24 | 0.08 |
| **5** | Cleaning Product 70 | 373.0 | 9682.15 | 0.09 |
| **15** | Kitchen Product 82 | 314.0 | 9571.93 | 0.07 |
| **19** | Outdoors Product 55 | 319.0 | 8826.35 | 0.08 |

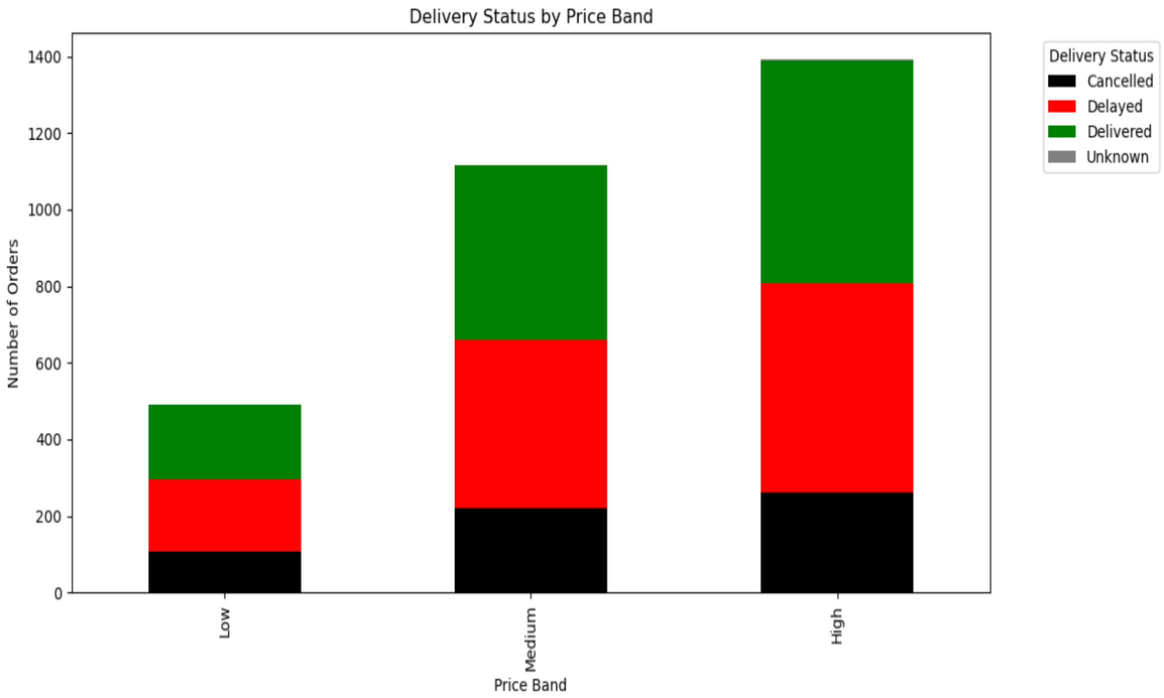
Boxplot showing statistical distribution of quantity vs discount for each category:

Figure 3

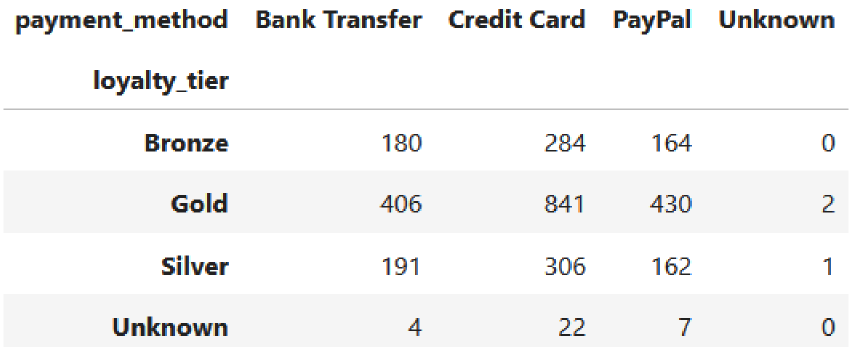


Delivery performance by price band:

Figure 4



Preferred payment method by loyalty tier:

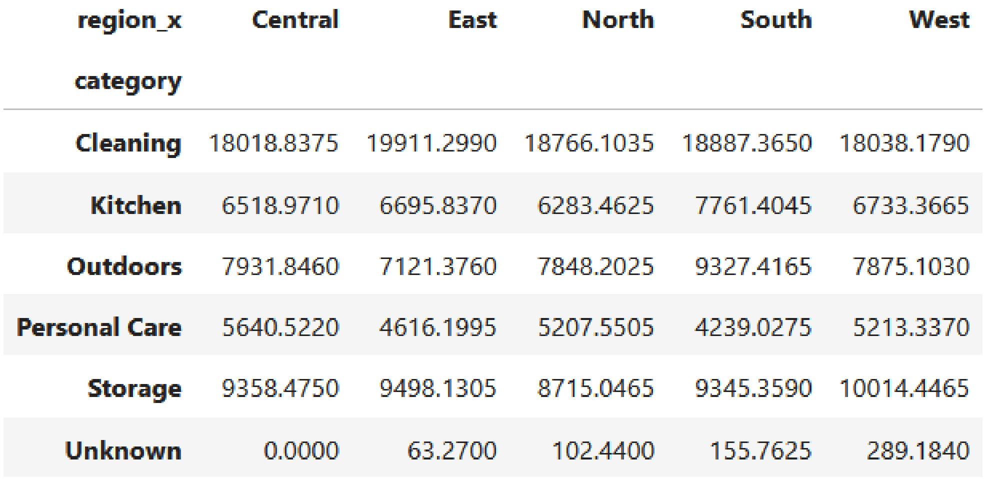


| payment_method | Bank Transfer | Credit Card | PayPal | Unknown |
|---|---|---|---|---|
| loyalty_tier | | | | |
| **Bronze** | 180 | 284 | 164 | 0 |
| **Gold** | 406 | 841 | 430 | 2 |
| **Silver** | 191 | 306 | 162 | 1 |
| **Unknown** | 4 | 22 | 7 | 0 |

5. Business questions:

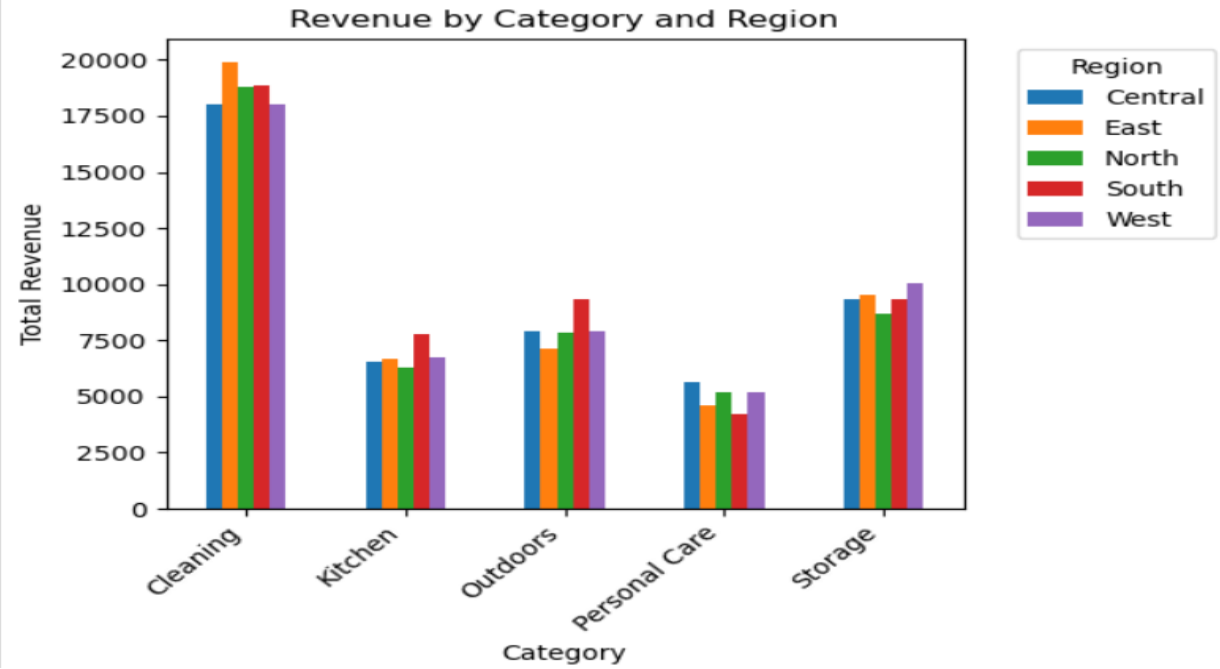
1. Which product categories drive the most revenue, and in which regions? Cleaning is the highest category for revenue, in all regions. Storage is always the second highest. Outdoors is always the third highest. Kitchen is always fourth,

followed by Personal Care.



| region_x | Central | East | North | South | West |
|---|---|---|---|---|---|
| category | | | | | |
| Cleaning | 18018.8375 | 19911.2990 | 18766.1035 | 18887.3650 | 18038.1790 |
| Kitchen | 6518.9710 | 6695.8370 | 6283.4625 | 7761.4045 | 6733.3665 |
| Outdoors | 7931.8460 | 7121.3760 | 7848.2025 | 9327.4165 | 7875.1030 |
| Personal Care | 5640.5220 | 4616.1995 | 5207.5505 | 4239.0275 | 5213.3370 |
| Storage | 9358.4750 | 9498.1305 | 8715.0465 | 9345.3590 | 10014.4465 |
| Unknown | 0.0000 | 63.2700 | 102.4400 | 155.7625 | 289.1840 |

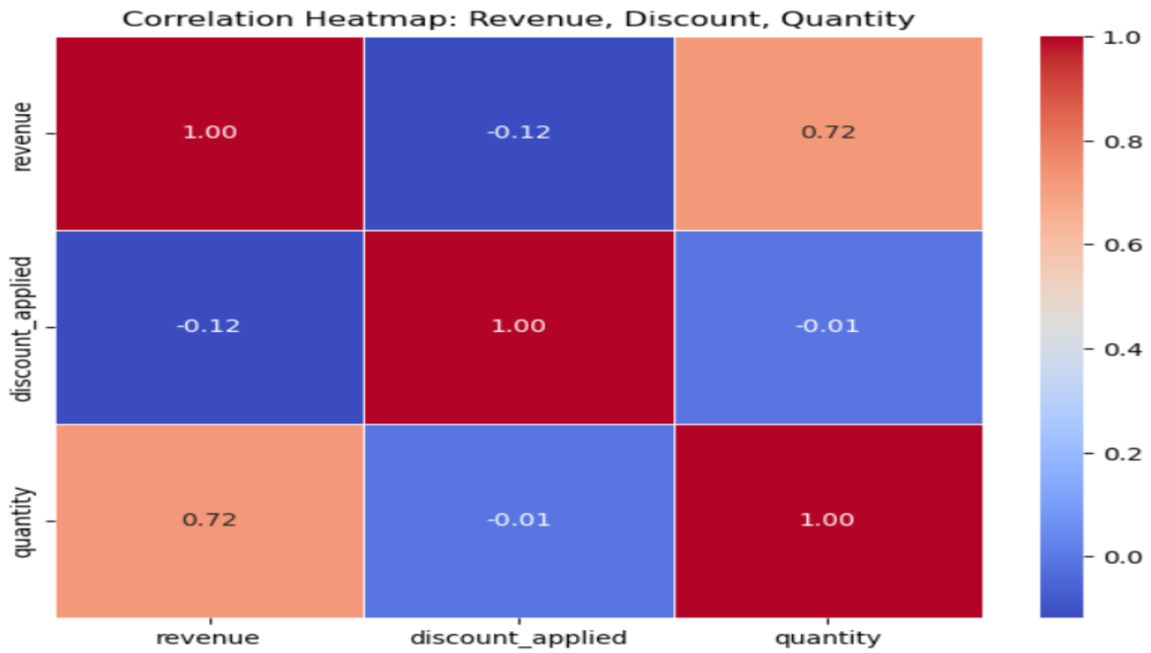
Figure 5



2. Do discounts lead to more items sold?

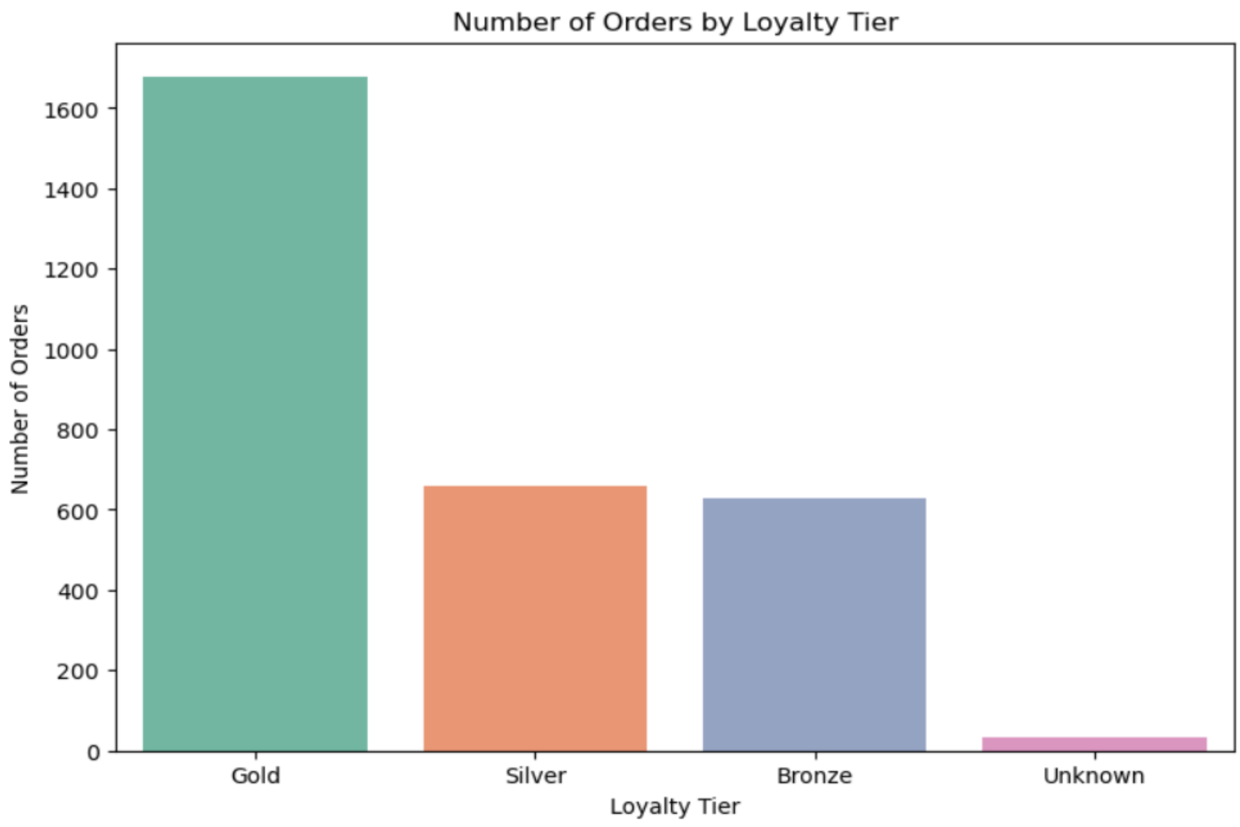
Discounts do not definitively lead to more items sold. There is no correlation under the Pearson coefficient on the heatmap.

Figure 6



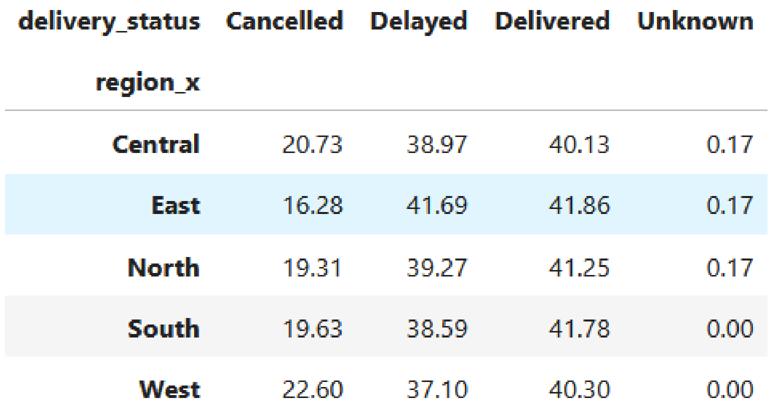
3. Which loyalty tier generates the most value? Gold generates the most value, placing twice as many orders as other tiers, and the difference is bigger for revenue.

Figure 7



4. Are certain regions struggling with delivery delays?

East Region has the most delays, at 41.69%, West has the fewest at 37.1%.



| delivery_status | Cancelled | Delayed | Delivered | Unknown |
|---|---|---|---|---|
| region_x | | | | |
| Central | 20.73 | 38.97 | 40.13 | 0.17 |
| East | 16.28 | 41.69 | 41.86 | 0.17 |
| North | 19.31 | 39.27 | 41.25 | 0.17 |
| South | 19.63 | 38.59 | 41.78 | 0.00 |
| West | 22.60 | 37.10 | 40.30 | 0.00 |

5. Do customer signup patterns influence purchasing activity?

Customer signup patterns do influence purchasing activity. The most revenue was from Gold customers who signed up in the second half of 2024. Also, although revenue is more often from Female customers, revenue per customer is slightly higher for Males.

Figure 8

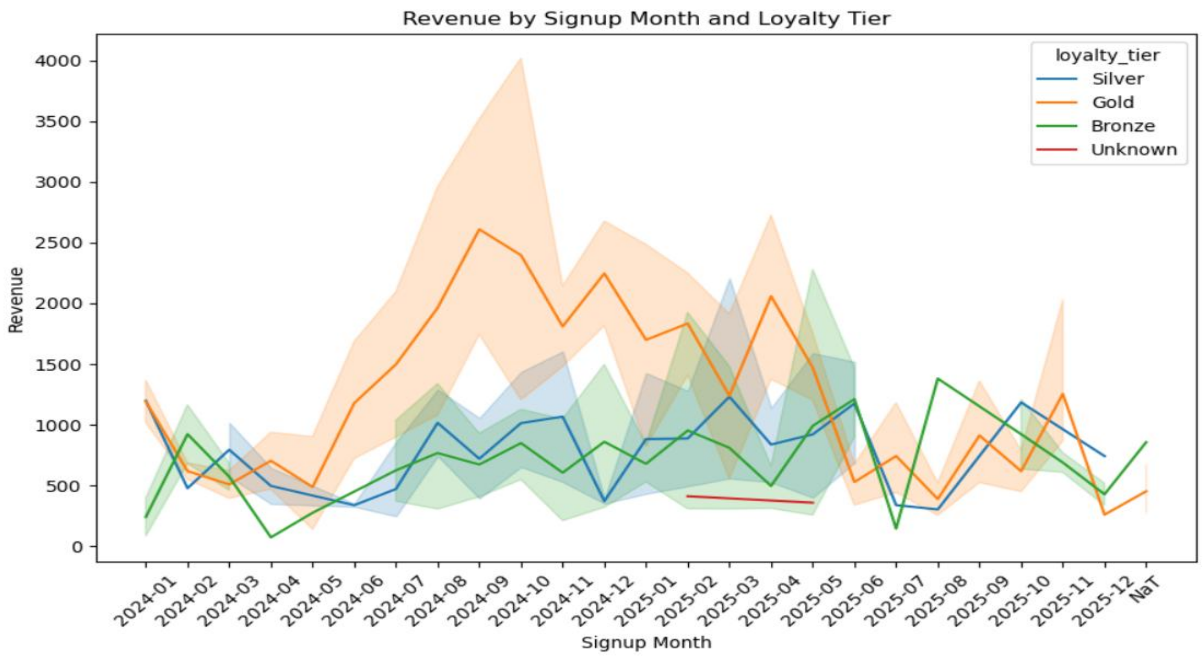
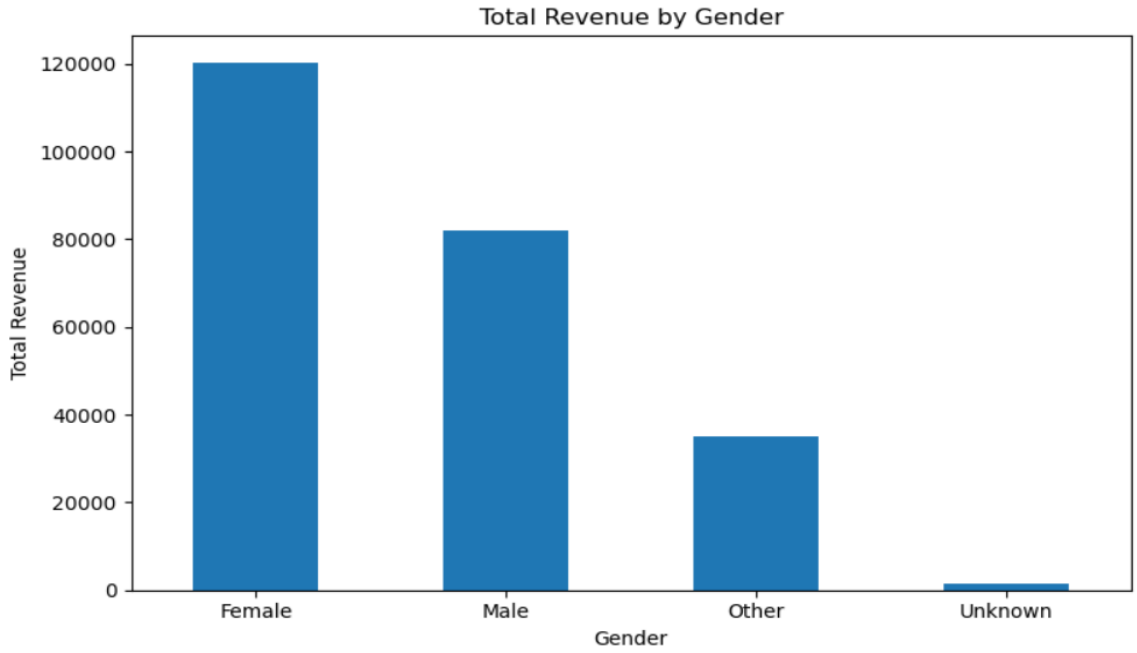
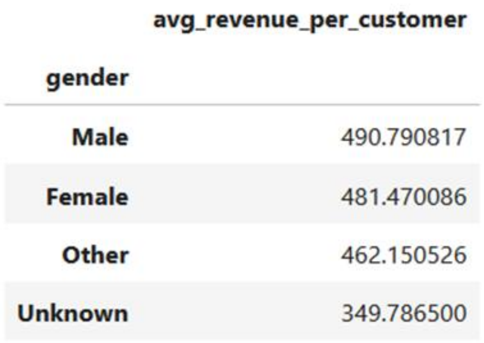


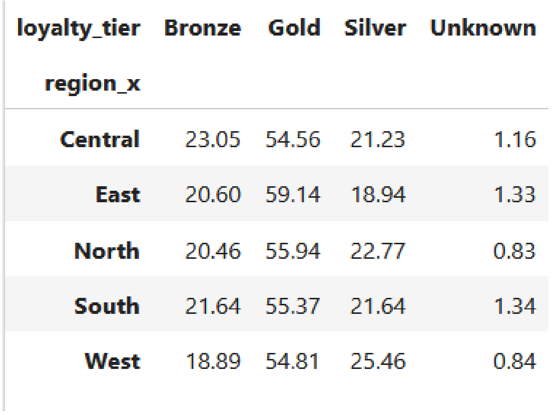
Figure 9





| | avg_revenue_per_customer |
|---|---|
| **gender** | |
| **Male** | 490.790817 |
| **Female** | 481.470086 |
| **Other** | 462.150526 |
| **Unknown** | 349.786500 |

You are most likely to be a Gold customer in East Region.



| loyalty_tier | Bronze | Gold | Silver | Unknown |
|---|---|---|---|---|
| **region_x** | | | | |
| **Central** | 23.05 | 54.56 | 21.23 | 1.16 |
| **East** | 20.60 | 59.14 | 18.94 | 1.33 |
| **North** | 20.46 | 55.94 | 22.77 | 0.83 |
| **South** | 21.64 | 55.37 | 21.64 | 1.34 |
| **West** | 18.89 | 54.81 | 25.46 | 0.84 |

6. Recommendations

- Focus marketing on Eastern region, as they are more likely to sign up as Gold mermbers.
- Try and address delivery delays, particularly in East Region where they are worst.
- Look at the marketing strategy that was used in late 2024 and replicate as a large amount of our revenue is from gold members that signed up in this period.

## 7. Data Issues

There were many inconsistent data issues, far too many as detailed in the table earlier. There needs to be fixed choices to prevent these (in Excel for example). This will help increase the speed and ease of data processing.

There were a large number of missing order dates, 60%. It's not possible to track order patterns and demand if you don't know when the orders happened.

References:

Matplotlib (2025). Stacked bar chart https://matplotlib.org/stable/gallery/lines_bars_and_markers/bar_stacked.html

Chatgpt (2026)

Uptrail (2025). Week 2 Recording.