Project Coversheet

Full Name	Mohammed Fahmidur Rahman
Email	20frahman@gmail.com
Contact Number	07429427096
Date of Submission	14/10/2025
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

Sales & Customer Behaviour Insights

This project involves analysing a retail sales dataset provided by a mid-sized e-commerce company. As part of the Data and Insights team, my task was to audit the dataset, clean it, and extract actionable insights that can help the company understand sales performance, customer purchasing behaviour, and regional delivery efficiency. The dataset includes key details such as order dates, product categories, order values, discount applied, customer regions, loyalty tiers, and delivery status.

Through this analysis, the report aims to uncover trends in revenue generation, identify which product categories and regions drive the most sales, assess the impact of discounts on customer purchases, and highlight potential operational issues such as delivery delays. The insights will support data-driven decision-making for marketing, sales, and logistics teams.

Data Cleaning Summary:

Before analysing the data, a thorough cleansing of the data was required to ensure accuracy and consistency.

Data Type Conversions

- The order_date column was originally stored as a string, preventing time-based analysis. This was converted to datetime format using pd.to_datetime()
- quantity, unit_price, and discount were converted to numeric types to allow calculations and summary statistics.

Text Standardisation

Several categorical columns contained inconsistent capitalisation and spacing. For example, delivery_status had entries such as “Delivered” and “delivered”. These were standardised using .str.strip().str.title(). Loyalty tiers such as “gold”, “Gold”, and “GOLD” were unified to “Gold”.

Duplicate Removal and Missing Values

- One duplicate order based on order_id was identified and removed.

Missing values were handled as follows:

- Missing loyalty_tier or region were filled with “Unknown” to retain records while marking them as incomplete.
- Missing quantity or unit_price were treated as zero where appropriate, and missing discounts were set to zero.

Before

```
order_id             1
customer_id          2
product_id           5
quantity             3
unit_price           1
order_date           3
delivery_status      3
payment_method       3
region               0
discount_applied   517
dtype: int64
```

After

```
order_id           0
customer_id        0
product_id         0
quantity           0
unit_price         0
order_date         0
delivery_status    0
payment_method     0
region             0
discount_applied   0
dtype: int64
```

Following these cleaning steps, and applying this logic to all 3 databases, each dataset now contains no missing values, as demonstrated in the image above.

Feature Engineering Summary:

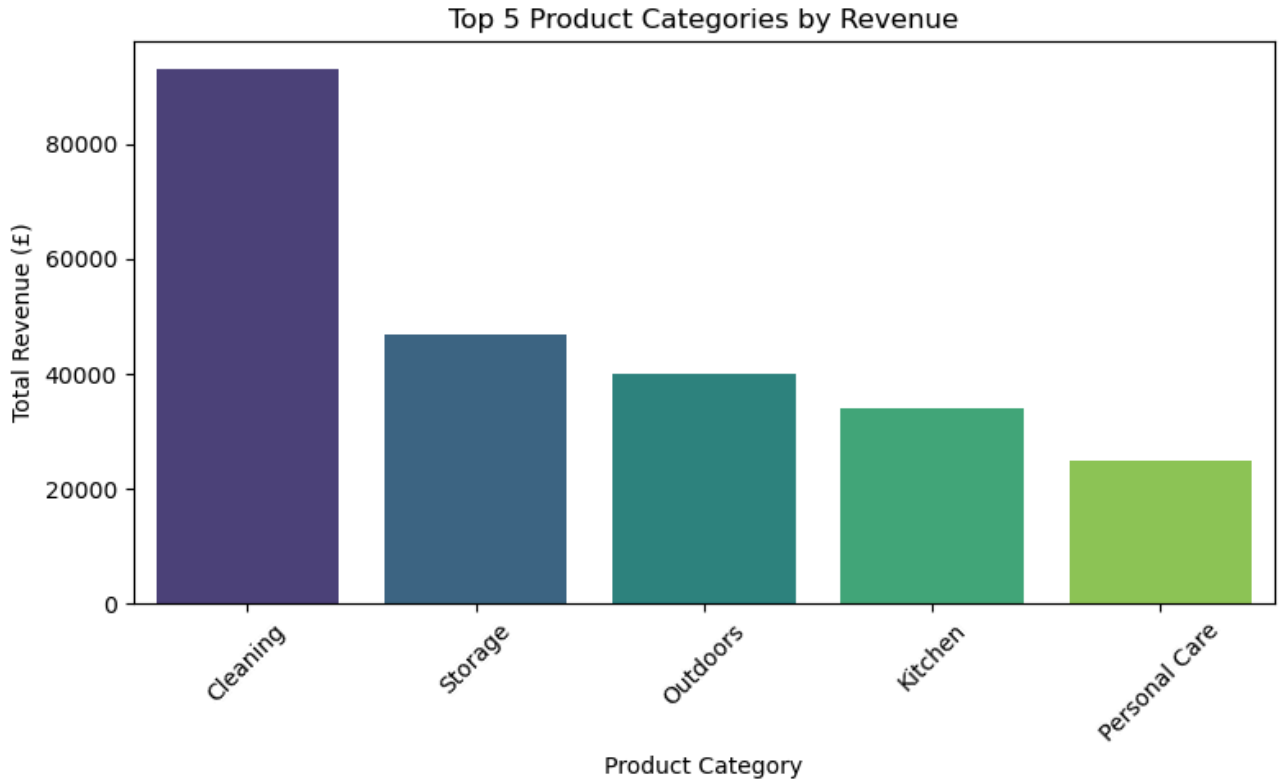
Several new features were created to support analysis and visualisation:

- **Revenue**: Calculated as quantity * unit_price * (1 - discount).
- **Order Week**: Extracted from order_date to allow weekly trend analysis.
- **Price Band**: Categorised order revenue into bands such as Low, Medium, High.
- **Email Domain**: Extracted from customer email addresses to study domain patterns.
- **Is Late**: Boolean indicating whether delivery was delayed.

	revenue	order_week	price_band	days_to_order	email_domain	is_late
0	117.7500	23	High	246	mills-logan.com	False
1	94.6000	23	Medium	140	morgan.com	True
2	25.2280	23	Medium	74	walters-smith.com	False
3	26.2080	23	High	327	gmail.com	False
4	38.0960	23	High	107	hotmail.com	True
5	102.3030	23	High	230	yahoo.com	True
6	146.4425	23	High	259	moore.com	False
7	37.6800	23	Low	325	whitehead-hernandez.biz	False
8	72.2160	23	Medium	64	herring.com	False
9	26.3160	23	High	293	russell.com	True
10	19.9600	23	Low	246	mcknight.info	False

Key Findings and Trends:

1. **Highest Revenue by Category:** Throughout the year, the Cleaning category consistently generated the highest revenue, significantly outperforming all other categories. Categories such as Storage and Outdoors contributed far less, highlighting a clear dominance of Cleaning products in overall sales. This suggests that demand for cleaning items drives the bulk of revenue, while other categories play a much smaller role.



2. **Revenue by Region:** In Week 23, the highest revenue was recorded in the South region at £49,560, closely followed by East (£47,799) and Central (£47,143). The North region showed a slightly lower contribution (£46,759), while the entry labelled “Nrth” with only £19 is clearly a data inconsistency and should be merged with “North” to ensure accurate analysis. Overall, revenue is relatively balanced across the main regions, with South slightly outperforming others during this period.

	region_x	order_week	revenue
0	Central	23	47143.2815
1	East	23	47799.4580
2	North	23	46758.9775
3	Nrth	23	19.5120
4	South	23	49560.5725

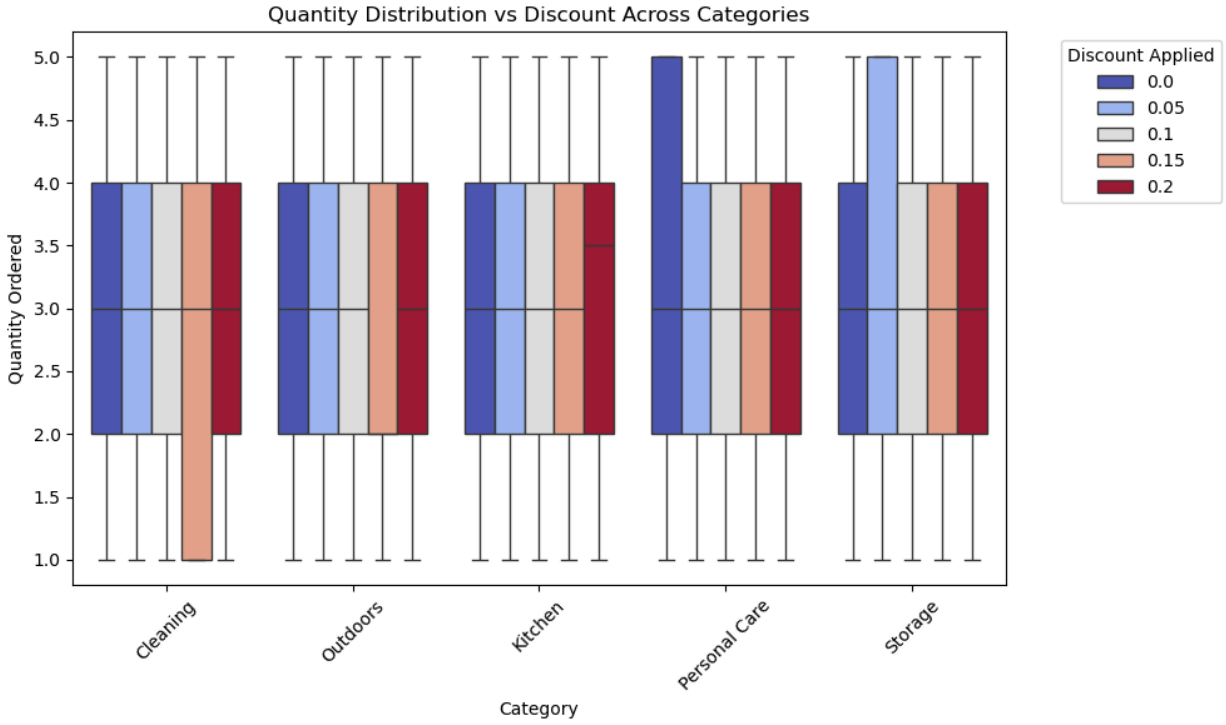
Business Question Answers:

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

Cleaning drives most of the revenue in the South Region.

2. Do discounts lead to more items sold?

Discounts show some influence on quantity sold, but not uniformly. We can see in the graph that when a discount was applied on Personal Care items, it had a higher maximum value for the quantity ordered.



3. Which loyalty tier generates the most value?

GOLD customers dominate transactions, mainly using Credit Card and Paypal, while BRONZE and SILVER tiers show much lower activity. Overall, Credit Cards are the most popular method across all tiers.

4. Are certain regions struggling with delivery delays?

Delayed orders are generally highest in the East and North regions, with Medium and High price bands seeing the largest proportion of delays (around 40–44%). The South region shows relatively lower delay rates for Low and Medium bands but spikes for High-priced orders. Overall, higher-priced orders tend to experience slightly more delays across most regions.

5. Do customer signup patterns influence purchasing activity?

Customer order frequency appears higher for long-standing loyalty tiers (Gold/Silver), suggesting that retention and loyalty schemes impact purchasing behaviour.

Recommendations:

- Focus on cleaning and high-revenue categories: Since cleaning clearly dominates revenue, consider promotional campaigns or inventory optimization for this category.
- Address delivery delays in key regions: Prioritize improving logistics in the East and North, especially for Medium and High price bands, to reduce the 40–44% delay rates.

Data Issues or Risk:

One key data quality issue identified is the inconsistent naming of regions and loyalty tiers, such as “Nrth” vs “North” and “SLLVER” vs “SILVER,” which can lead to inaccurate aggregations and misleading insights. This could be fixed at the source by implementing stricter data entry validation and standardized dropdowns for regions and loyalty tiers, as well as adding automated checks to flag or correct inconsistencies before the data is stored.