Project Coversheet

May 30, 2025

Full Name	Jude Spellacy
Email	judespellacy@gmail.com
Contact Number	07891187181
Project Title	week 2

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.

Project 2: Data Cleaning, Analysis, and Business Insights

1 Data Cleaning (SQL Queries)

Importing the data to postgresql sorted out the date format which was inconsistent, by setting the data type directly to SQL's date format. The dates were then verified to ensure no month day values were flipped.

```sql
SELECT * FROM sales_data
```

Listing 1: SQL query to view in order to then update rows

After viewing the data it was immediately clear there was some missing values in the data.

- Two null emails.
- Two null phone numbers.
- Two null Discount percentages.

```sql
UPDATE sales_data SET email = 'no_email@email.com' WHERE email IS NULL;
```

Listing 2: Updating null emails to filler value.

```sql
UPDATE sales_data SET phone = '0000000000' WHERE phone IS NULL;
```

Listing 3: Updating null phone numbers to filler value.

For both missing values in the discount percentage, it was clear from a glance that filling in the mean value would lead to anomalous data. The discount percentage seemed to correlate heavily with revenue.

```sql
SELECT CORR(revenue, discount) AS correlation FROM sales_data;
```

Listing 4: Pearson correlation between revenue and discounts which outputs 0.703.

Using linear regression between revenue and discount percentage to obtain an estimated relation of $\hat{y} = 0.00591X + 3.91739$. Then applying that to the two points of 500 revenue, it was estimated to be a discount percentage of 7.

```sql
UPDATE sales_data SET discount = '7' WHERE discount IS NULL and revenue = 500;
```

Listing 5: Updating null discount percentage to filler value.

- There was a duplicated row with only a different order ID.
- There was another row repeated, but with a different date. This was left in as a repeat customer.

```sql
DELETE FROM sales_data WHERE order_id = 104;
```

Listing 6: Deleting duplicate row of John Doe.

Using repeat customer information to fill in a missing email, there was a phone number of 9898989898 from the same duplication, but this was assumed fake.

```sql
UPDATE sales_data set email = 'alice@email.com' WHERE customer_name = 'Alice Smith';
UPDATE sales_data set phone = '0000000000' WHERE customer_name = 'Alice Smith';
```

Listing 7: Replacing missing email and replacing fake phone number with filler value.

2 Visualisation and Data Summary

2.1 Data Summary

Summarising Revenue data the following was obtained:

Category	Electronics	Furniture	Clothing
Mean	2100	2150	566.67
Median	2100	2150	500
Mode	3000/1200	2500/1800	500
Range	1800	700	200
Standard Deviation	1272.8	495.0	115.5
Total	4200	4300	1700

The mean discount percentage per category is given in:

Electronics (%)	Furniture (%)	Clothing (%)
15	20	6.33

Then looking at the sales distribution for the year by month:

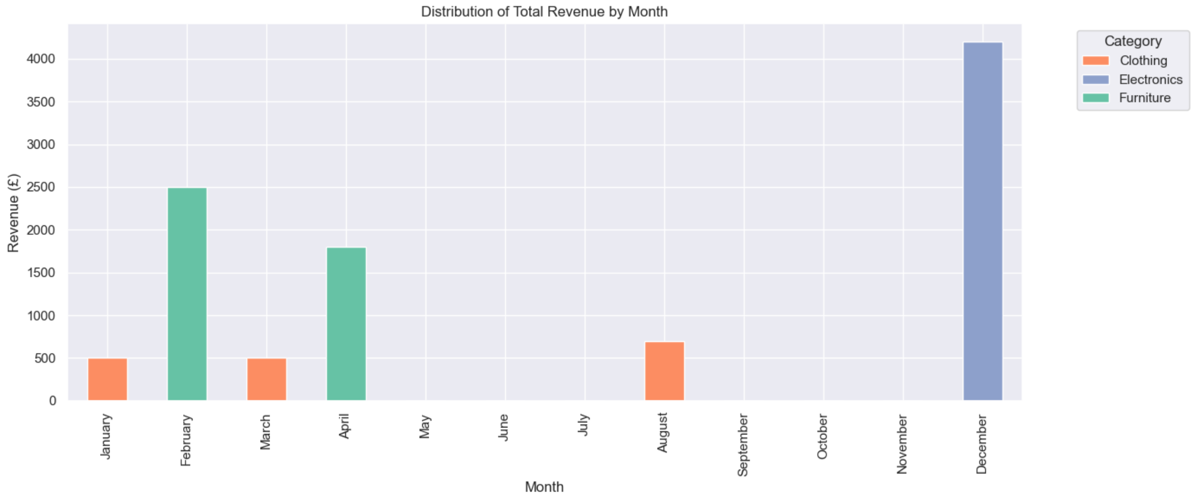

Figure 1: Revenue by month.

2.2 Visualisations

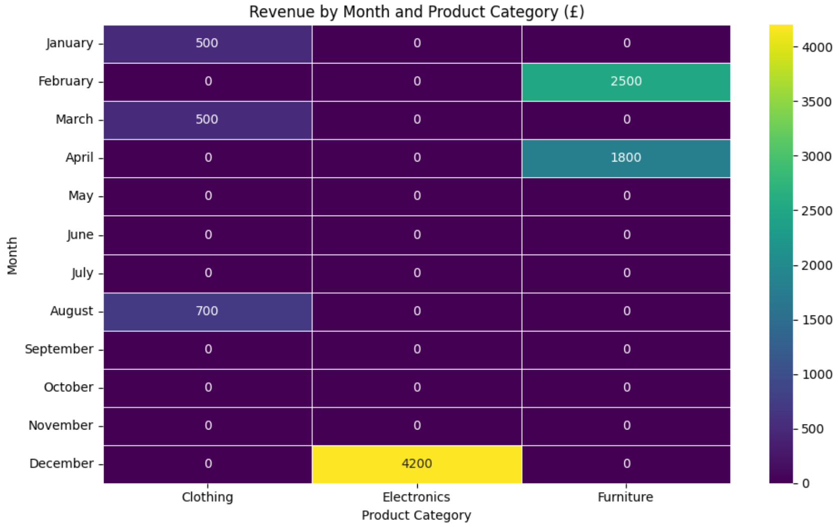

Figure 2: Revenue by month and product category (colouring skewed by inconsistent sampling).

Examining the relationship between revenue and month gives shows a concentration in revenue towards the beginning of the year. Sampling is very sparse and skewed so this could be the main reason for the distribution. Then examining the revenue against discount corroborates the previous assumption of a linear relationship. Removing the data point for furniture at a 25% discount may be argued to more accurately represent the distribution, but the sample size is to small to dismiss it as an outlier.

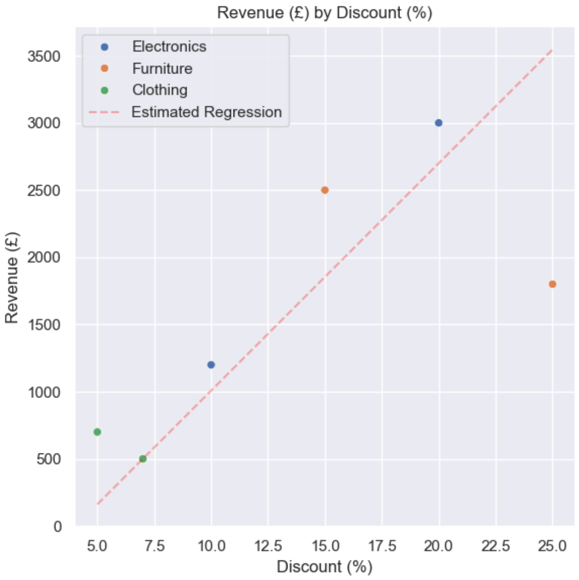

Figure 3: Revenue against discount percentage.

Looking at revenue from individual transactions as a histogram, shows an almost uniform distribution over the sample.

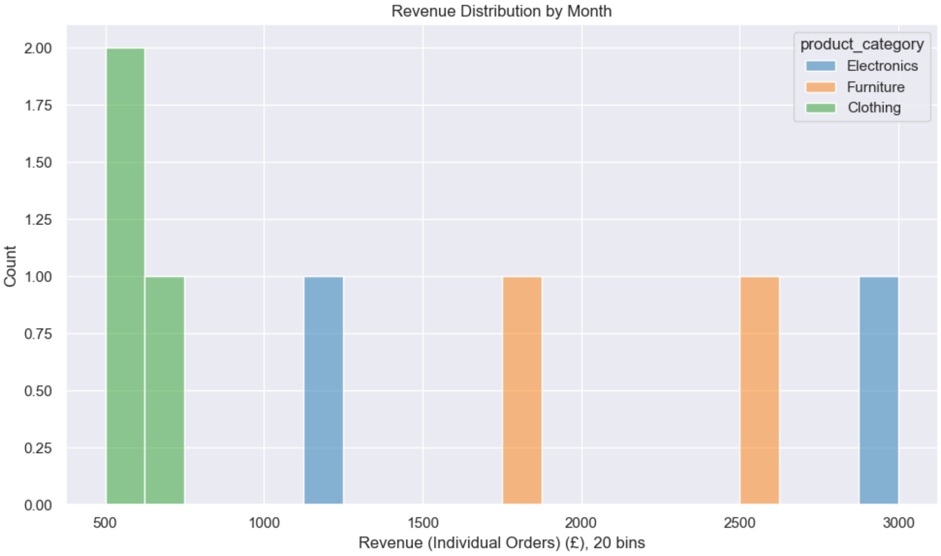

Figure 4: Order size distribution.

3 Key Findings and Report

Monthly Revenue

- Sample mainly focuses on Winter and early Spring, revenue distribution absent through most of the rest of the year.
- Large revenue generated in December in Electronics possibly due to seasonal holidays.
- Furniture sales within the sample peak in February with £2500 and show a secondary peak at £1800.
- Clothing sales in the sampled months are relatively even, they do not generate large revenue comparatively.

Discount Effects on Revenue

- Over all categories there is a clear positive linear relationship between discount percentage and revenue in the data analysed.
- Discounts in Furniture over the limited sample did not result in an increase in revenue.
- Discounts in Electronics over the sample generated a significant increase in revenue.
- Clothing does not have enough data within the category to conclusively say whether revenue has a relationship with discount.

Order Size Distribution

- Even distribution in the quantity of individual order sizes over the range of revenues, over all categories.
- In the limited sample of clothing potentially a relationship between revenue and quantity of transactions, with cheaper clothes orders resulting in more orders within the data.
- Over the range of revenue values from transactions there is also an even distribution, no large price gaps.

Summary Insights

- Transaction sizes are even across the range and over all categories likely to be relatively insensitive in quantity to size.
- Over the whole set of categories within the data, there is evidence of a potential positive linear relationship between discount percentage and revenue.
- Electronics potentially the only/most significantly effected category to discount percentage on revenue.
- Electronics and Furniture generate the most revenue over all categories.
- Likely seasonal relationship for revenue in Electronics over Winter.

Recommended Actions

- Larger or broader discounts on Electronics over the winter in the lead-up to December could generate a larger revenue. Pushing marketing in Electronics over the same date range could compound this.
- Focusing on discounts in Furniture and Clothing is potentially not worth it.
- Increasing product range on low revenue clothing could potentially increase the quantity of sales, more information would be needed for Summer and Autumn, but it could generate a more even revenue stream.