Project 1

Introduction to Data Analysis

1) Cleaned Dataset (Excel)

- Formatted the data as a table.
- **Handling Missing/Inaccurate Values** -
  Used formula: Quantity * Price [e.g. =**F**2*__G__2] to fix any numerical calculation inaccuracies in Total_Amount.
- **Dropping Redundant Columns** – None; all columns can be used to draw insights.
- **Duplicates** - Removed duplicates using inbuilt action, excluding Transaction_ID and Date column. Transaction_ID is a primary key so by default should be excluded from duplicates, and Date is related to it so by proxy also removed. Two rows were removed: 1006 and 1011.
- **Data Type Formatting:**
  - **Quantitative:**
    - **Numerical (Whole Number) –** Transaction_ID & Quantity.
    - **Currency (2dp + GBP)** - Price & Total_Amount.
  - **Qualitative**:
    - **Text** – Product.
    - **Categorical/List (Text)** - Category, Payment_Method, & Region.
  - **Ordinal**:
    - **Date** – Date.
- **Expanding the Dataset** – Sales (Total_Amount) already exists.
  No cost column so a profit column can't be derived from available data.
- **Data Modelling** – None; it's a small dataset so it is not required.
- **Note:** Online version of Excel was used, so there may be some inconsistencies in data formatting.

2) Exploratory Data Analysis (Python & Excel)

Due to the small sample size, it's difficult to make any claims with strong supporting evidence due to potential bias. However, we can make several inferences based on what we already have.

Payment_Method	Region
Debit Card	East
Debit Card	East
Debit Card	East
Debit Card	East
Debit Card	East
Credit Card	North
Credit Card	North
Credit Card	North
Credit Card	North
Credit Card	North
Cash	South
Cash	South
Cash	South
Cash	South
Cash	South
PayPal	West
PayPal	West
PayPal	West

Region has a direct relationship with the payment method.

This becomes visually apparent when sorting either Payment_Method or Region on a spreadsheet.
This suggests either one of the following:

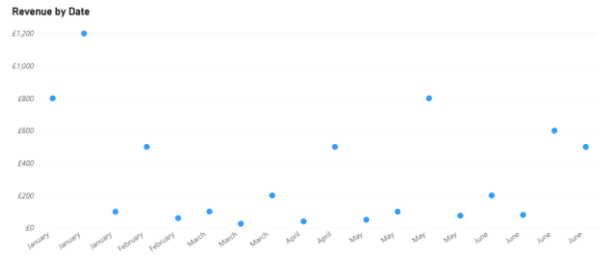
A) Customers strongly prefer using a certain type of payment method for a particular region.

B) A particular region only; allows for a particular type of payment method to be used.

Due to a lack of data, it's difficult to determine whether scenario **A**, or **B** is true. But if a customer ever uses a payment method not associated with a particular region, we can prove **B** to be false (e.g. Cash in West).

Correlations:

	Transaction_ID	Quantity	Price	Total_Amount
Transaction_ID	1.000000	0.080667	-0.074129	-0.200142
Quantity	0.080667	1.000000	-0.447867	-0.268290
Price	-0.074129	-0.447867	1.000000	0.916645
Total_Amount	-0.200142	-0.268290	0.916645	1.000000



Asides from Pricing and Total Amount, there are no correlations between the quantitative data.

When inputted into a correlation matrix, no value exceeds the (-)0.5 range, suggesting no strong correlations within the data provided.
Similarly, there are no visual indicators suggesting that the Date has any correlation with other numerical columns when inputted into a scatter graph (the example uses Total_Amount).

	Transaction_ID	Date	Quantity	Price	Total_Amount
count	18.00	18	18.00	18.00	18.00
mean	1010.72	2024-04-01 16:00:00	1.56	283.89	329.44
min	1001.00	2024-01-05 00:00:00	1.00	20.00	25.00
25%	1005.50	2024-02-17 00:00:00	1.00	31.25	76.25
50%	1011.00	2024-04-05 12:00:00	1.00	150.00	150.00
75%	1015.75	2024-05-09 12:00:00	2.00	500.00	500.00
max	1020.00	2024-06-10 00:00:00	4.00	800.00	1200.00
std	6.15	NaN	0.92	287.67	343.58

The average spending of a customer is around £329, with a current **maximum** of £1200, and **minimum** of £25. A **standard deviation** of £344 indicates a high variability in total spending.

The upper quartile and median are consistent for Price and Total Amount, meaning **customers typically do not purchase more than 1 high-cost product**. Whereas it differs for the lower quartile, suggesting that **customers typically buy cheaper items in bulk** – being more than double indicates at least; 2 items are purchased together.

An **average** customer purchases 2 items when rounded, with a **maximum** of 4 items.

Transactions with more than 1 product sold:

	Category	Product	Quantity
6	Books	Book	3
7	Books	Book	2
9	Books	Book	4
10	Electronics	Smartphone	2
11	Clothing	Shoes	2
13	Clothing	T-Shirt	3

Based off quantity, **Books account of the most bulk purchases**, with Electronics being the least likely (only Smartphones).

The MEAN of [Price] by [Category]:

Category	
Books	20.00
Clothing	37.50
Electronics	445.45

Name: Price, dtype: float64

The MEAN of [Price] by [Category] & [Product]:

Category	Product	
Books	Book	20.0
Clothing	Shoes	50.0
	T-Shirt	25.0
Electronics	Headphones	100.0
	Laptop	800.0
	Smartphone	600.0
	Smartwatch	200.0
	Tablet	500.0

Name: Price, dtype: float64

The SUM of [Quantity] by [Category] & [Product]:

Category	Product	
Books	Book	9
Clothing	Shoes	3
	T-Shirt	4
Electronics	Headphones	2
	Laptop	2
	Smartphone	3
	Smartwatch	2
	Tablet	3

Name: Quantity, dtype: int64

Electronics generates the most sales being around £400 more than the next category on average, despite the (mostly) single quantity purchases per transaction. With **Laptops being the highest earner** – despite also having the least purchases.

Count of Transaction_ID	Region				
Category	East	North	South	West	Grand Total
Books	0	3	0	0	3
Clothing	0	0	3	1	4
Electronics	5	2	2	2	11
Grand Total	5	5	5	3	18

Sum of Total_Amount	Region				
Category	East	North	South	West	Grand Total
Books	£ -	£ 180.00	£ -	£ -	£180.00
Clothing	£ -	£ -	£ 225.00	£ 25.00	£250.00
Electronics	£1,900.00	£1,600.00	£1,800.00	£200.00	£5,500.00
Grand Total	£1,900.00	£1,780.00	£2,025.00	£225.00	£5,930.00

On the other hand, Electronics are extremely popular in the East; but doesn't make any sales in any other category. This suggests that either products in other categories are not sold in the East, or there is no demand for anything else.

Customers only purchase Books in the North. This suggests that either Books are exclusively sold in the North or aren't in demand in any other region.

The West is the least profitable region, by around £1600 and should be investigated.

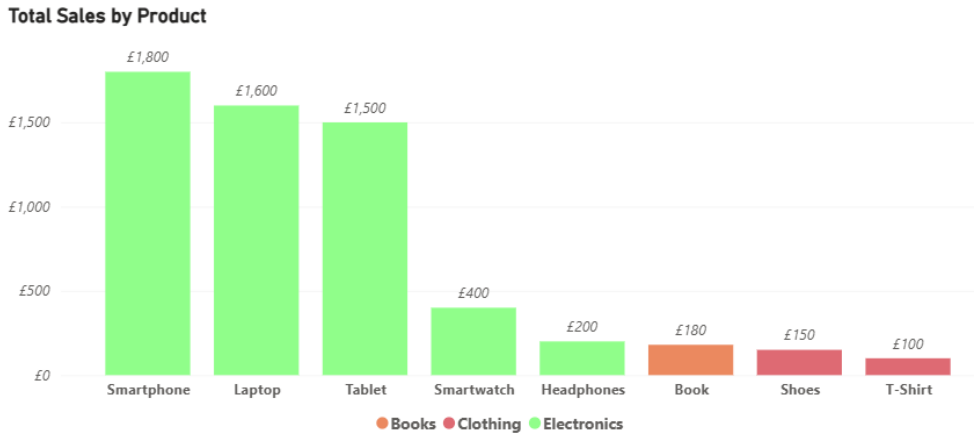
Total revenue by MONTH:

Date	
2024-01	2100
2024-02	560
2024-03	325
2024-04	540
2024-05	1025
2024-06	1380

January generated the most sales.

We can see that the Total_Amount shows a downward trend from January to March (the lowest month), before showing an upward increase to June.
The most concerning months range from February to April, which are around four times less than January.

3) Data Visualizations (Excel & Power BI)



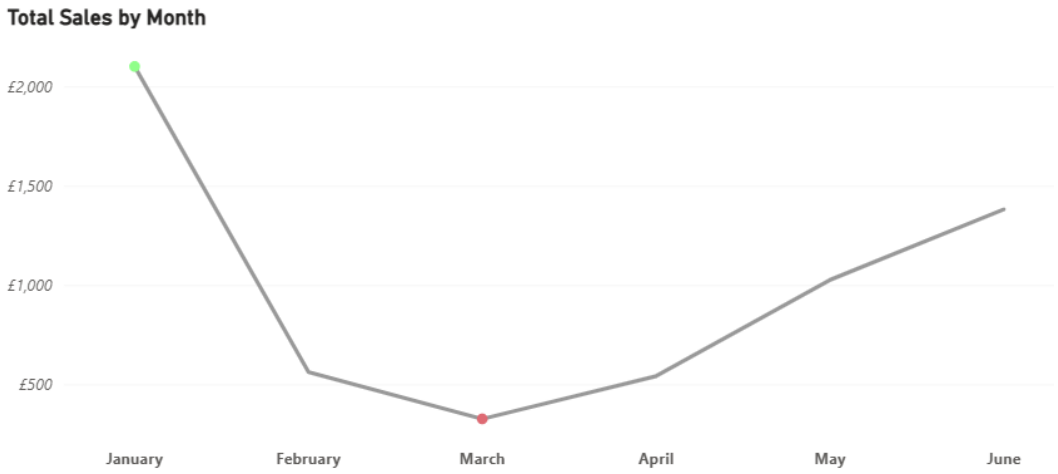
The figure above depicts the total sales of each product line (for comparison) over the 6-month duration and are labelled by category. It's very clear that electronics has the best performance in total sales, and it isn't even close – although it's also worth mentioning that it has the most diverse product line.

We've previously established that customers are more willing to purchase a higher quantity of items if the price is low, but despite being greatly more expensive on average; electronics has exceeded the performance in every other product line, with the highest performer (Smartphone) generating almost 8x more sales than the total of the next best performing category (Clothing).

Sum of Total_Amount	Region				
Category	East	North	South	West	Grand Total
Books	£ -	£ 180.00	£ -	£ -	£180.00
Clothing	£ -	£ -	£ 225.00	£ 25.00	£250.00
Electronics	£1,900.00	£1,600.00	£1,800.00	£200.00	£5,500.00
Grand Total	**£1,900.00**	**£1,780.00**	**£2,025.00**	**£225.00**	**£5,930.00**

A conditional formatted pivot table was derived to determine the distribution of sales by product category and region. We can see that the east, north, and south contribute roughly the same number of sales, with electronics being the main driving force. However, in the west, which is by far the worst performing region; coincidentally also generated the least sales in electronics. Evidently, we can see a pattern that the performance of electronic sales is directly tied to the performance of a region.

One thing to note is that certain categories made no sales in specific regions, with the east only having made sales in electronics, which is worth further investigation.



Moving onto date, a line graph was plotted to determine whether there is any relationship between total sales and their month. Ideally, I would have liked to use a legend for the three respective categories, however clothing and books made no sales on certain months, making it impossible to plot three lines on the graph.

From the graph we can determine the best and worst performing month, being January and March respectively. The line shows a parabola facing down; indicating poor sales performance during the transition between winter to spring but showing signs of improved performance nearing summer.

4) Final Data Insights Report

From our findings, we can draw several conclusions:

- **Electronics** is the best performing category in sales by a wide margin.
- **Smartphones** are the product with the best sales performance followed shortly by Laptops and Tablets.
- **January** is the best performing month in sales, with the months leading into spring (**March**) having the worst performance.
- The **West** has the worst sales performance of all four regions by a wide margin.

Electronics has demonstrated itself to be the biggest influencer in terms of sales, having sold in every month and region. Due to the substantially higher average price, electronics can generate far more in terms of sales than every other category despite the lower quantity. On average; for every single unit of electronics sold, more than 11 units of clothing (22 for books) is needed to match the sales performance.

It's evident that, electronics should be the main focal point of attention to improve sales performance, whether it's marketing, availability/stock, or expanding the product line. Although the same can be applied to other categories, the amount of resources required to match the cash cow is unproportional to that of electronics, it raises the question whether it is worth the effort.

Clothing and books combined accounts for approximately 7% of total sales which isn't much, however there are two suggestions to consider based on the circumstances; if there are no plans to discontinue their sales.

The first is to increase the prices of products in both categories. It's clear that customers are willing to purchase items in higher quantities if the price is low, so naturally a price increase will scale much better than electronics. The average total sales per transaction is roughly £329 meaning that patrons are relatively wealthy, so it's worth considering experimentally bumping up the prices to determine a good middle ground where the pricing can be increased to where the quantity of sales does not decrease.

A more stable approach however; is to sell clothing and books in each region if they are not already. The prime example of this is the east, which only made sales in electronics,

or books which only made sales in the north. If the reason for 0 sales is no demand, then this proposal can be ignored, otherwise there is room for increased sales for both categories.

Marketing efforts should be focused on smartphones, laptops, and tablets as they have the best performance in sales, with over triple the performance of the next best product – these products also happen to have a low quantity of transactions. An increase in purchases for any of these 3 products will substantially increase sales performance.

January has the best performance of the 6-months with the most likely reason being post-holiday sales, so more marketing will have a significant impact. The months following leading into spring (February to April) sees a sharp decline in sales with March being the worst, whereas it begins to increase as summer approaches. Discounts and spring seasonal promotions to incentivize purchases should be held between February to end of April, as this period is undoubtedly the worst for sales. Loyalty programs can also aid during this period; with bonuses that can encourage spending during low sales periods due to fear of missing out.

Regarding regions, it's worth considering increasing the marketing of electronics in the west, as it’s the only region where the sales are poor. Additionally, payment method is currently being associated with region. If it’s a customer preference, then this is not an issue. However, if it’s intentionally specific to a region, unless there is a justifiable reason for this, then this rule should be abolished as it’s limiting accessibility to payment methods; which in turn can reduce sales.

Customer IDs are present in the dataset which can potentially be traced; therefore, data masking is required – the most simply solution is to use a hash function for Customer_ID. Consent from the customer should also be acquired; if there are any plans to sell this data due to ethical concerns in the GDPR, stating clearly: what the data is used for, who it will be shared with, and how long it is retained in the system before being removed.