# Sales: Uncovering Sales Trends - An Exploratory Analysis

# Executive Summary

This report presents an in-depth analysis of a synthetic sales dataset designed for educational and testing purposes. The dataset simulates sales transactions across various products, regions, sales representatives, and customer segments, spanning the years 2022 to 2024. The primary focus of this analysis is on transactions occurring between 2023 and 2024 to identify relevant patterns and insights during this period.

The analysis was conducted using Microsoft Excel, employing tools such as Pivot Tables, Charts, and standard Excel functions to organise, summarise, and visualise the data. The report highlights the process of cleaning and preparing the dataset, conducting exploratory analysis, and generating business-oriented insights based on key performance indicators.

Key objectives of the analysis include investigating variations in performance among sales representatives, exploring regional differences in revenue contributions, examining product category performance, assessing the influence of discounts on sales outcomes, and comparing the profitability of online versus retail sales channels.

**Findings:**

- **Sales Representatives:** The analysis revealed that sales representatives exhibited varying levels of performance, with some consistently delivering higher revenue. Notably, Eve and David emerged as top performers, showing significant revenue growth in 2024.
- **Regional Performance:** The North and East regions were the top contributors to overall revenue, with the East region slightly surpassing the North in 2024. The South and West regions also experienced significant revenue growth from 2023 to 2024.
- **Product Categories:** Food and Clothing were identified as the most profitable product categories, with Food demonstrating the largest relative increase in profit from 2023 to 2024. Electronics was the least profitable category.
- **Discounts:** The analysis indicated that strategic discounting positively impacted sales profitability. In 2024, a more aggressive average discount rate of 12.46% aligned with a significant 58.85% rise in average profit.
- **Sales Channels:** Online sales yielded slightly higher average profit margins than retail sales, highlighting potential efficiencies in the online model. However, retail remained a substantial contributor to total sales.

**Constraints:** Several challenges were encountered during the analysis, including the lack of unique customer identifiers, variance in profit due to randomisation, ambiguity in product identifiers, and the granularity of time data. These challenges impacted the depth and accuracy of the insights derived.

**Recommendations:** To enhance the quality, depth, and accuracy of future analyses, the report proposes several recommendations, including introducing unique customer and transaction identifiers, assigning globally unique product identifiers, standardising variable relationships, including additional customer and product metadata, improving time granularity, and ensuring consistent data generation rules.

**Conclusion:** Despite the challenges posed by the synthetic nature of the data and its limitations, the analysis provided valuable insights into sales performance, customer behaviour, and regional trends. By addressing these challenges and implementing the recommendations outlined, future analyses can achieve greater accuracy, depth, and relevance, ultimately contributing to improved business outcomes.

## Project Introduction

This report presents an analysis of a synthetic sales dataset designed for educational and testing purposes. The dataset simulates sales transactions across various products, regions, sales representatives, and customer segments, spanning the years 2022 to 2024. For the purpose of this analysis, focus has been placed on transactions occurring between 2023 and 2024 to identify relevant patterns and insights during this period.

Although the dataset does not represent real business operations, it has been carefully structured to reflect realistic sales scenarios. As such, it offers a valuable opportunity to explore sales performance, customer behaviour, and regional trends, while applying core data analysis techniques in a controlled environment.

The analysis was conducted using Microsoft Excel, employing tools such as Pivot Tables, Charts, and standard Excel functions to organise, summarise, and visualise the data. This report highlights the process of cleaning and preparing the dataset, conducting exploratory analysis, and generating business-oriented insights based on key performance indicators.

## Objectives

The primary aim of this analysis is to conduct an exploratory investigation into various sales-related factors to uncover patterns, trends, and potential relationships within the dataset. Rather than testing specific hypotheses, the focus is on gaining a deeper understanding of the data and identifying insights that may inform future business strategies. In particular, this exploration includes:

- Investigating variations in performance among sales representatives.
- Exploring regional differences in revenue contributions.
- Examining how different product categories perform.
- Looking into how discounts may influence overall sales outcomes.
- Observing customer spending behaviour between new and returning buyers.
- Comparing the profitability of online versus retail sales channels.
- Identifying recurring trends in monthly sales over the analysis period.

# Dataset Description

This dataset comprises sales transaction data generated for analysis and practice purposes. Each record represents an individual sales transaction, characterised by the following attributes in Table 1 below.

*Table 1 Sales Dataset Column Descriptions*

| Feature Name | Description |
|---|---|
| Product_ID | Unique identifier for each product sold. Randomly generated between 1000 and 1100. |
| Sale_Date | The date when the sale occurred. Dates range from 2022 to 2024. |
| Sales_Rep | The sales representative responsible for the transaction (Alice, Bob, Charlie, David, Eve). |
| Region | The geographical region where the sale took place (North, South, East, West). |
| Sales_Amount | Total revenue from the transaction (between £200 to £260,000), including discounts. Returns/refunds not considered in this dataset so Sales_Amount = Revenue. |
| Quantity_Sold | Number of units sold in the transaction (Randomly generated between 1 to 50 units). |
| Product_Category | Category of the product sold (Electronics, Furniture, Clothing, Food). |
| Unit_Cost | Cost per unit of the product sold (£50 to £5000). |
| Unit_Price | Selling price per unit of the product (£150 to £5500). Always higher than Unit_Cost. |
| Customer_Type | Indicates if the customer is "New" or "Returning." |
| Discount | Discount applied to the unit price (0% to 30%). |
| Payment_Method | Method customer used to pay (Bank Transfer, Credit Card, Cash). |
| Sales_Channel | Channel through which the product was sold (Online or Retail). |

# Dataset Constraints

Several limitations should be considered when interpreting the findings presented in this report. These constraints, stemming primarily from the structure and synthetic nature of the dataset, are outlined below.

## Lack of Unique Customer Identifiers

The dataset does not include a Customer_ID column, which limits the ability to distinguish between individual new and returning customers. While new customers are assumed to be unique, the aggregation of transactions by date, rather than by a specific Transaction_ID or Customer_ID, prevents accurate tracking of customer behaviour or purchase history over time. As a result, this also means that another interesting piece of analysis, being Average Sale Value across Sales Reps is not possible to derive.

This limitation also affects the analysis of returning customer data, as it is not possible to determine how many units a specific customer ordered on a given day. Instead, the data only reflects the total quantity of a product sold on that date. The absence of timestamps or transaction identifiers compounds this issue by restricting the granularity of the analysis to the daily level. For example, a customer recorded as new on 01/01/2022 and making a purchase again on 02/01/2022 would be considered returning, despite the lack of evidence that these transactions relate to the same individual.

### Variance in Profit Due to Randomisation

Profit figures in the dataset were derived using randomly generated values for Unit_Cost, Unit_Price, and Discount, leading to significant variability. This volatility made it difficult to conduct reliable profit-based analysis. As a result, revenue being comparatively stable across the dataset was used as the primary performance metric throughout the report.

Additionally, due to the randomised nature of the dataset, key relationships between variables such as Sales_Amount, Quantity_Sold, Unit_Price, Unit_Cost and Discount required manual validation to ensure logical coherence. This further limits the extent to which findings can be generalised or used for predictive insights.

### Ambiguity in Product Identifiers

The Product_ID field contains randomly assigned values between 1000 and 1100, which are reused across four different Product_Category types. This structure does not offer a unique identifier for each individual product, thus limiting the ability to perform detailed, product-specific analysis. A more effective approach would include a globally unique product identifier that links directly to both category and product-level details.

## Methodology

This section outlines the data preparation and processing steps undertaken to facilitate subsequent analysis.

### Data Preparation

The dataset was initially imported from a CSV file. Several columns were then added to the dataset to enhance its structure and analytical potential: *Index*, *New_Product_ID*, *Year*, *Month*, *Profit_Margin*, and *Region_and_Sales_Rep*.

**Index**
An Index column was added to assign a unique identifier to each row, as the original dataset lacked any unique key. This enabled easier tracking and potential referencing of individual transactions during analysis and QA.

**New_Product_ID**
The original Product_ID values were randomly generated between 1000–1100 and reused across multiple product categories, which limited the ability to distinguish individual items. To resolve this, a New_Product_ID column was created using an XLOOKUP combined with the TEXTJOIN function. This added a layer of category-specific identification (e.g., clt-1045 for Clothing and fnt-1045 for Furniture) to ensure each product-category combination was uniquely identifiable.

**Year and Month**
The Sale_Date column was broken down into Year and Month fields using date formulas. This allowed for more granular time-series analysis and the ability to track trends on a monthly or yearly basis.

**Region_and_Sales_Rep**
A combined column named Region_and_Sales_Rep was created using the CONCAT function to merge the Region and Sales_Rep columns. This was useful in understanding sales performance patterns by specific representatives within each region and streamlined visualisations and filtering.

**Profit_Margin**
A new column for Profit_Margin was calculated based on the following formula below:

*Equation 1 Profit Calculation*

$$Unit\ Price\ \times (1 - Discount)) - Unit\ Cost\ \times Quantity\ Sold$$

This calculated the actual profit made on each transaction. A detailed QA process, discussed below, was implemented to verify the accuracy and logic of this calculation. Conditional formatting and manual verification were used to identify anomalies such as negative profit margins, which were investigated and confirmed to be valid given the randomised nature of the dataset.

These data preparation steps ensured the dataset was not only clean and structured, but also analytically robust and tailored for the specific business questions explored in this report.

## Profit Calculation QA

To ensure the accuracy and reliability of the profit calculation formula used in this analysis, several quality assurance (QA) steps were implemented:

1. **Manual Spot Checks**

A sample of rows was manually checked using known values to verify that the formula logic worked as intended. These checks confirmed that the calculated profit per transaction accurately reflected Equation 1.

**2. Test Case Validation**

Specific test cases were constructed using fixed values (e.g., zero discounts, known margins) to validate the formula against expected profit outcomes. All test cases passed successfully, confirming the consistency of the calculation across different scenarios.

**3. Automated Cross-Check Column**

A secondary QA column was introduced using an IF formula to cross-verify calculated profit against a separately derived result, accounting for rounding differences. This allowed quick identification of any discrepancies in logic or data input.

**4. Conditional Formatting and Anomaly Review**

Conditional formatting was applied to flag any transactions where the calculated profit was negative. This revealed 12 instances where profit margins were below zero. Shown in Table 2 below. Each of these rows was individually reviewed and validated. Upon inspection, the negative margins were found to be accurate and logical, resulting from a combination of high discount percentages and the randomised nature of the dataset values. This highlights the impact of synthetic data generation on profitability patterns and further emphasises the importance of validating anomalies before drawing conclusions.

*Table 2 Profit QA and Conditional Formatting*

| Index | Unit_Cost | Unit_Price | Profit_Margin | Discount |
|---|---|---|---|---|
| 462 | £ 3,291.86 | £ 3,324.13 | -£ 17.48 | 1.00% |
| 765 | £ 470.52 | £ 493.99 | -£ 41.80 | 5.00% |
| 920 | £ 3,764.74 | £ 3,793.97 | -£ 313.55 | 1.00% |
| 945 | £ 4,043.30 | £ 4,120.57 | -£ 205.66 | 2.00% |
| 1128 | £ 3,686.81 | £ 3,710.92 | -£ 181.99 | 1.00% |
| 1205 | £ 4,093.61 | £ 4,117.88 | -£ 84.54 | 1.00% |
| 1429 | £ 4,398.16 | £ 4,439.12 | -£ 30.88 | 1.00% |
| 1489 | £ 3,290.89 | £ 3,317.75 | -£ 6.32 | 1.00% |
| 1675 | £ 4,680.35 | £ 4,758.11 | -£ 156.62 | 2.00% |
| 1922 | £ 4,781.42 | £ 4,864.34 | -£ 617.77 | 2.00% |
| 1956 | £ 2,407.80 | £ 2,502.76 | -£ 159.66 | 4.00% |
| 2424 | £ 1,687.35 | £ 1,719.27 | -£ 46.84 | 2.00% |

These QA measures ensured that the profit metric used throughout the analysis was both technically sound and accurately reflected within the context of the dataset.

## Cell Formatting

To ensure accuracy in analysis and consistency across calculations, appropriate data types and cell formats were applied to each column within Excel. Proper formatting not only enhances readability but also ensures that Excel functions and formulas operate as expected, particularly when performing numerical operations or filtering and sorting data.

Each column in the dataset was reviewed and assigned a suitable cell format based on the nature of the values it contained. The table below outlines the formatting decisions made:

*Table 3 Cell Format for Columns in Dataset*

| Column Name | Cell Format |
|---|---|
| Index | Number |
| Product_ID | General |
| new_product_id | General |
| Sale_Date | Short Date |
| Year | General |
| month | General |
| Sales_Rep | Text |
| Region | Text |
| Sales_Amount | Accounting |
| Quantity_Sold | Number |
| Product_Category | Text |
| Unit_Cost | Accounting |
| Unit_Price | Accounting |
| Profit_Margin | Accounting |
| Customer_Type | Text |
| Discount | Percentage |
| Payment_Method | Text |
| Sales_Channel | Text |
| Region_and_Sales_Rep | Text |

Special consideration was given to financial figures such as Sales_Amount, Unit_Cost, Unit_Price, and Profit_Margin, which were formatted using the Accounting format to improve clarity when comparing values. The Discount field was expressed as a Percentage to align with its role in profit calculations. Meanwhile, all categorical fields like Sales_Rep, Region, and Product_Category were assigned the **Text** format to avoid automatic misinterpretation by Excel (e.g., converting codes into dates or removing leading zeroes).

By applying structured formatting across the dataset, it ensured that analytical operations such as conditional formatting, filtering, pivoting, and formula applications were both effective and error-free.

## Extracting Data into Tables

To support the analysis in this report, a combination of PivotTables and custom data tables was used to extract and organise insights from the dataset. Each approach served a specific purpose in the overall methodology.

Initially, PivotTables were employed as a rapid exploratory tool. These enabled quick aggregation and visualisation of key metrics such as total revenue (Sales_Amount) by Sales Representative and Month, providing early insights into performance trends in a time-series format. The simplicity and flexibility of PivotTables also inspired additional areas of inquiry, such as revenue breakdowns by Region or Sales_Channel, allowing patterns to emerge and guiding deeper analysis.

However, as the project progressed, most of the final analysis was conducted using formula-based data tables. These tables were created on separate worksheet tabs and relied heavily on functions such as SUMIFS, which enabled more customised and dynamic aggregation of the data. This approach offered several advantages:

- Greater control over grouping and filtering, such as isolating trends by Year, Month, Product Category, or other dimensions.
- Support for advanced calculations, including metrics like annual moving averages and profit margins.
- Compatibility with statistical methods, allowing for the integration of basic hypothesis testing and the calculation of confidence intervals, where appropriate.

By building these formula-driven tables, the analysis could adapt more easily to evolving questions, test different assumptions, and present more tailored insights than PivotTables alone could offer. The structured design of these tables also made it easier to automate summaries and create linked visualisations across multiple dashboards within the workbook.

In summary, while PivotTables played an important role in the initial data exploration and trend identification, custom data tables provided the analytical depth and flexibility required for more advanced and business-focused insights presented throughout this report.

## Calculations

The following calculated fields and derived metrics were used in the analysis:

- **Annual Moving Average:** A 12-point moving average was calculated using the AVERAGE function in Excel. This metric was used to smooth out short-term fluctuations in sales data and identify longer-term trends.
- **Average Profit per Unit:** Calculated by dividing the total profit by the Quantity_Sold. This metric provides the average profit generated per unit sold. The data tables allowed for varying granularity in the calculation, providing monthly and yearly average profit figures.
- **Average Revenue:** Calculated by dividing the total revenue (Sales_Amount) by the Quantity_Sold. This metric represents the average revenue generated per unit sold. Similar to Average Profit, the data tables enabled calculations with monthly and yearly granularity.
- **Effective Units Sold:** This derived metric was calculated by dividing Total Profit by Marginal Profit per Unit, which accounts for the impact of discounts. It helped approximate how many units would have to be sold at a given profit margin to reach the

total profit reported. This calculation added another layer of interpretation regarding efficiency and pricing strategies within the dataset.

- **Average Discount Rate:** To accurately assess discount trends, weighted average discount rates were calculated for both 2023 and 2024. The formula used multiplies each transaction's discount by the number of units sold, summing these values only for the specified year. This total is then divided by the total quantity sold that year. This approach gives a clearer picture of how discounts were distributed across all sales volumes, rather than relying on a simple average. The result shows the average discount experienced per unit sold in each respective year, helping to compare the real impact of discount strategies over time.

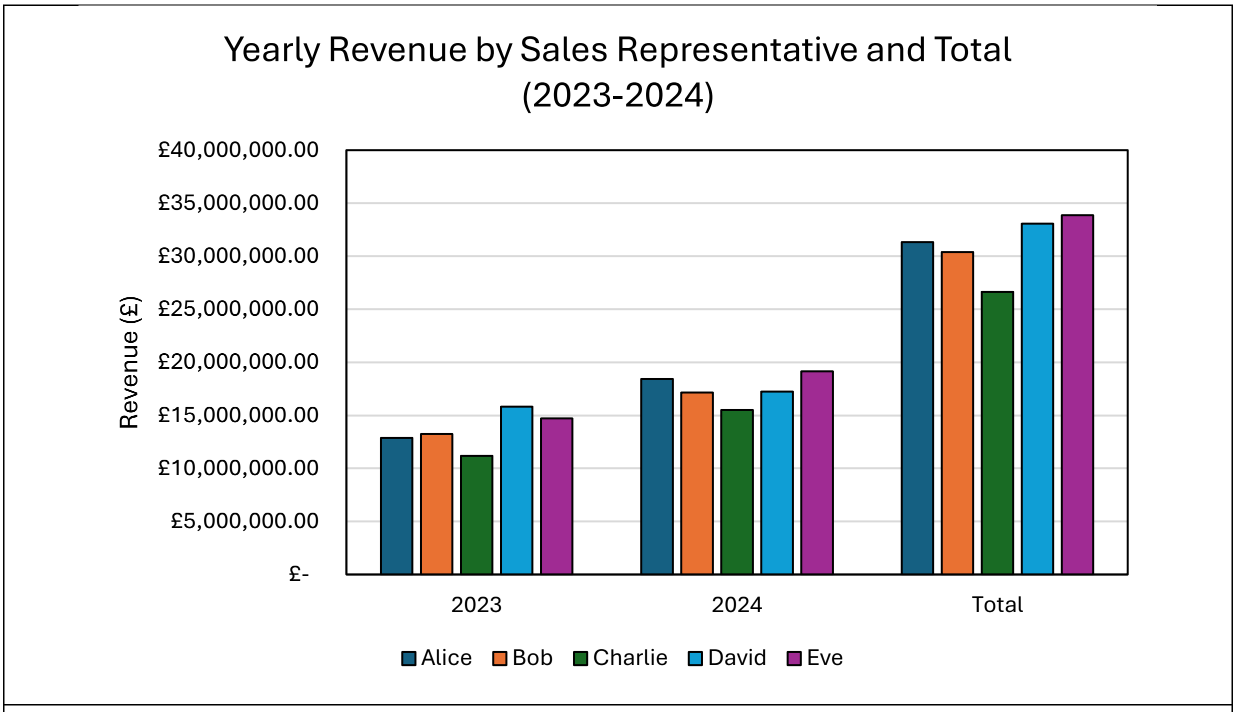
# Analysis & Findings

## Performance Across Different Sales Representatives

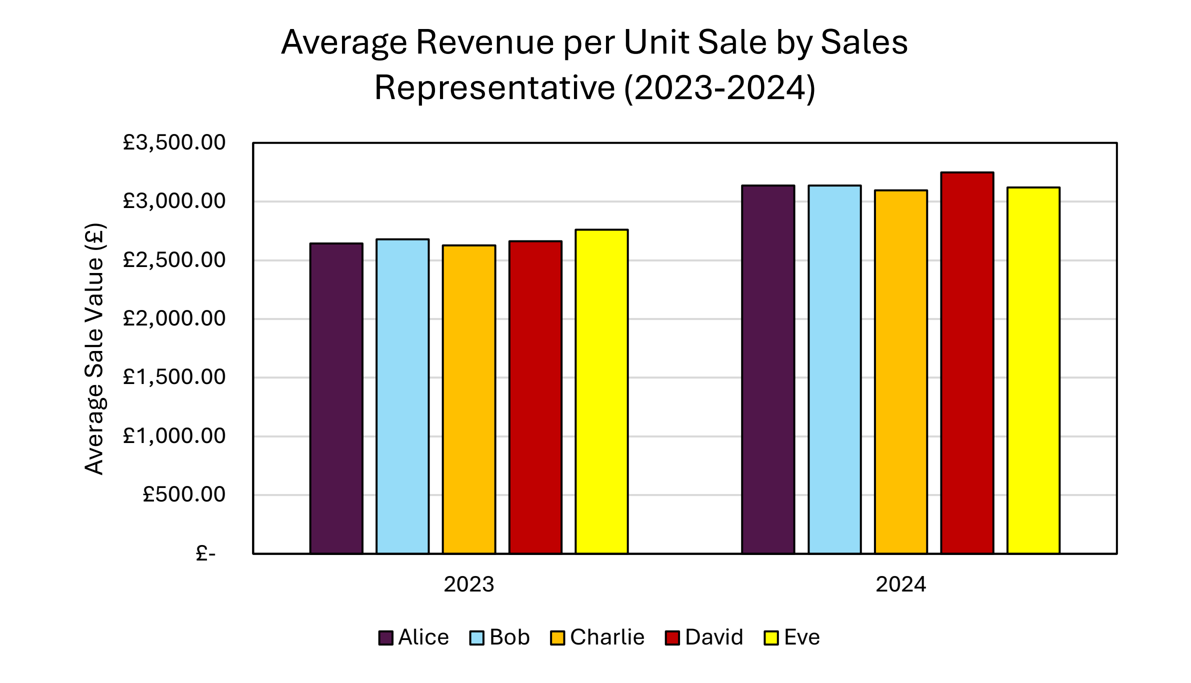
To evaluate the contribution and performance of individual sales representatives, several key metrics were examined, including total sales, average sale value, and monthly performance trends. This analysis aimed to identify top performers and understand how sales efforts vary across the team.

**Total Revenue by Sales Representative**

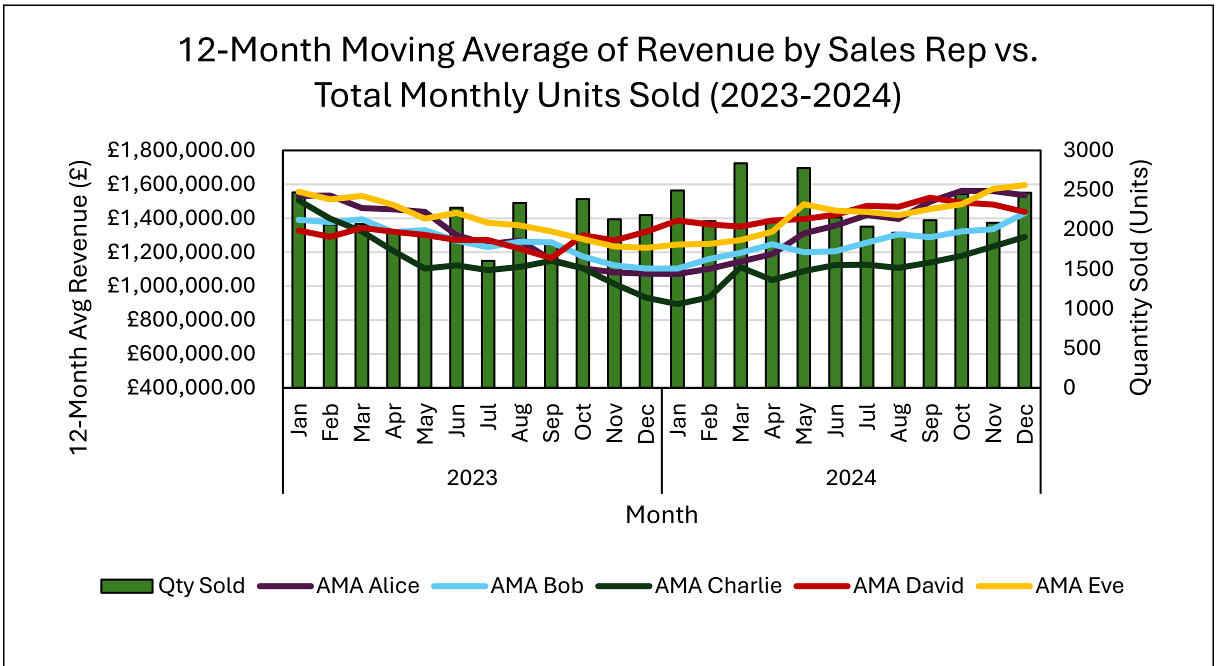
The total Sales_Amount generated by each sales representative was calculated to highlight their overall contribution to revenue. This metric provided a straightforward comparison of performance at a macro level. The results revealed which sales reps consistently delivered higher revenue across the entire dataset timeframe as well as allowing for a comparison between the years 2023 and 2024.



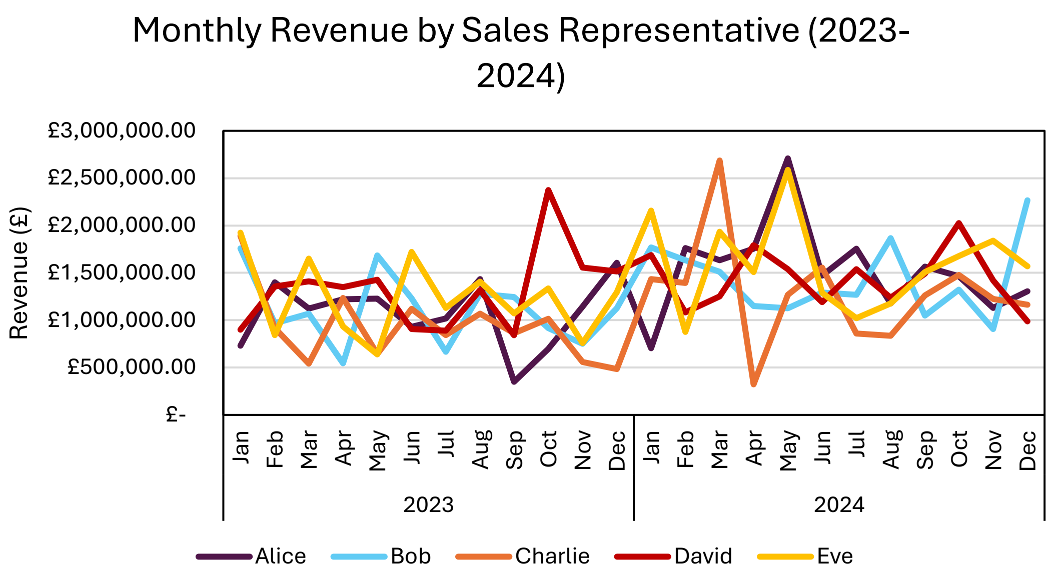

This chart illustrates the yearly revenue for each sales representative in 2023 and 2024, and a total for the sum of these two. For all sales reps, their total revenue in 2024 increased in comparison to 2023, notably with David having the highest revenue in 2023 of around £15.8m, and in 2024 Eve having the highest revenue of around £19.1m. Overall, between these two years, Eve has generated the most revenue whereas Charlie has generated the least by a considerable amount.




The chart highlights the change in average revenue per unit sale for each sales representative across 2023 and 2024. While the average revenue per unit was relatively consistent among all reps in 2023, ranging between £2,600 and £2,800, a notable increase is observed in 2024. David stands out as the top performer in 2024, with the highest average sale value exceeding £3,200, followed closely by Alice and Bob. This upward trend suggests potential improvements in product pricing, sales strategy, or overall sales efficiency among the reps.



This chart compares the 12-month moving average revenue for each sales representative against the total monthly quantity of units sold. It reveals seasonal peaks in unit sales, particularly in the first and last quarters of each year. Among the sales reps, Eve and David consistently show higher average monthly revenues, with David exhibiting a strong upward trend throughout 2024. Conversely, Charlie maintained the lowest moving average, indicating underperformance relative to peers. The overall increase in average revenue towards the end of 2024 suggests a positive trajectory in sales performance for most representatives.



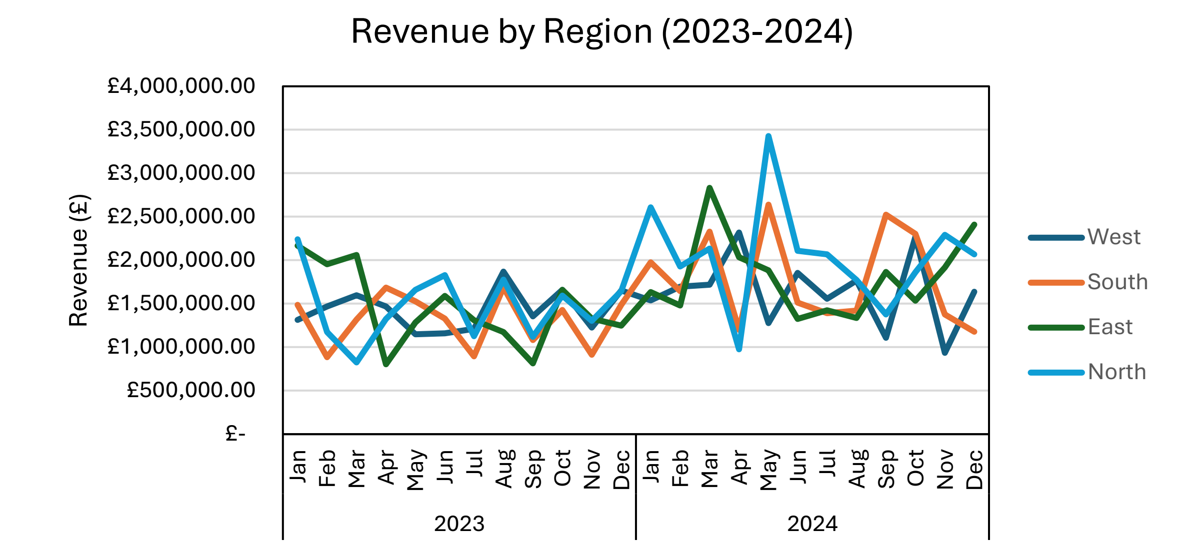

This chart illustrates monthly revenue fluctuations for each sales representative over a two-year period. Revenue patterns show significant volatility month-to-month for all reps, with clear spikes and dips suggesting seasonality or irregular purchasing behaviour. In 2023, revenue for most reps remained below £2 million, with Bob and Charlie showing more stable but lower figures throughout the year.

Moving into 2024, all sales reps demonstrated increased revenue peaks, particularly in Q2. Charlie had standout performances in March 2024, whilst Eve and Alice had standout performances in May 2024, exceeding £2.5 million in revenue. Bob showed a stronger performance in the early part of the year, particularly in January, and Bob had a strong finish to the year, with revenue peaking in December 2024.

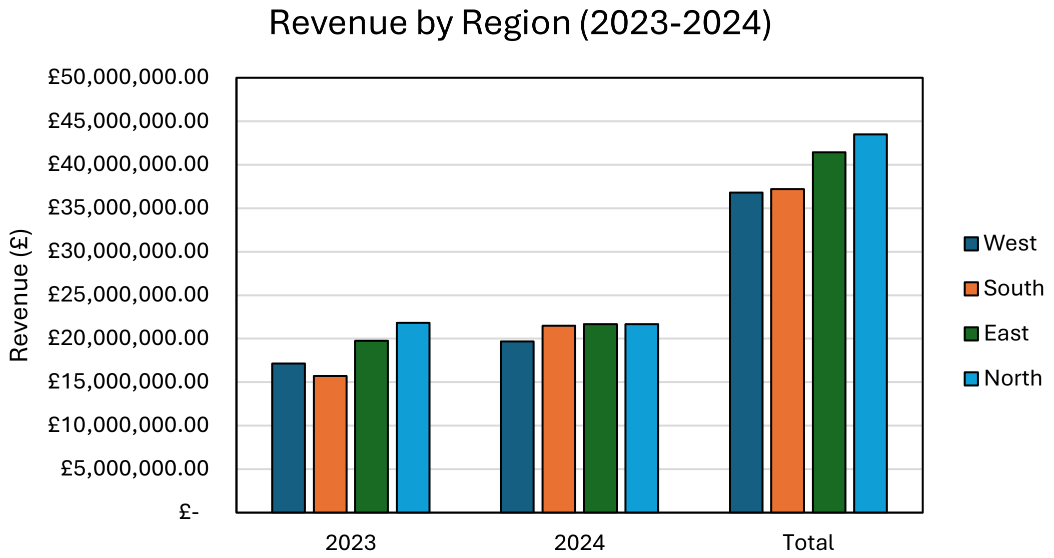
Overall, the chart highlights that while performance is inconsistent on a monthly basis, there are clear periods of high performance for each rep, offering insight into when individual reps may be most effective or when external factors (like promotions or seasonal trends) may have influenced sales.

## Revenue Contributions by Region

To better understand the geographic performance of sales efforts, an analysis was conducted to assess how each region contributed to overall revenue between 2023 and 2024. By breaking down total sales by region, the aim was to uncover regional strengths, disparities in performance, and potential areas for growth. This regional perspective provides critical context for evaluating sales strategy effectiveness and guiding future resource allocation. The charts below show this.

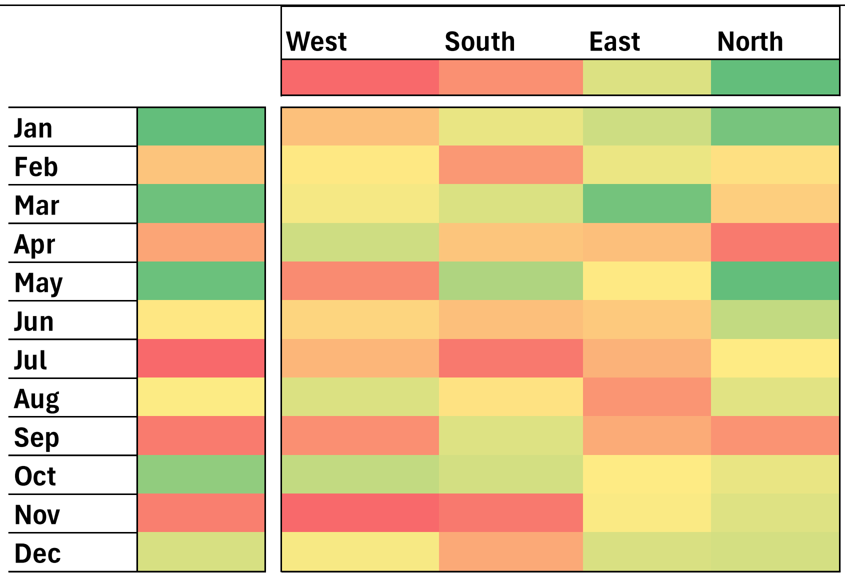



This chart shows a monthly plot of revenue performance across the four regions between 2023 and 2024 reveals distinct patterns, with the North region experiencing a notable surge in revenue during the spring of 2024, contrasting with more stable performance in the West. The South region showed the most fluctuation, while the East region had a strong start to 2024. Overall, each region displayed some degree of seasonality in its revenue generation.




This chart shows the total revenue by region. The North and East regions were the top contributors overall between 2023 and 2024, although the East slightly surpassed the North in 2024. Both the West and South regions experienced significant revenue growth from 2023 to 2024. In contrast, the South region generated the lowest revenue across 2023, however the West generated the least revenue across both years, despite the increase in 2024.

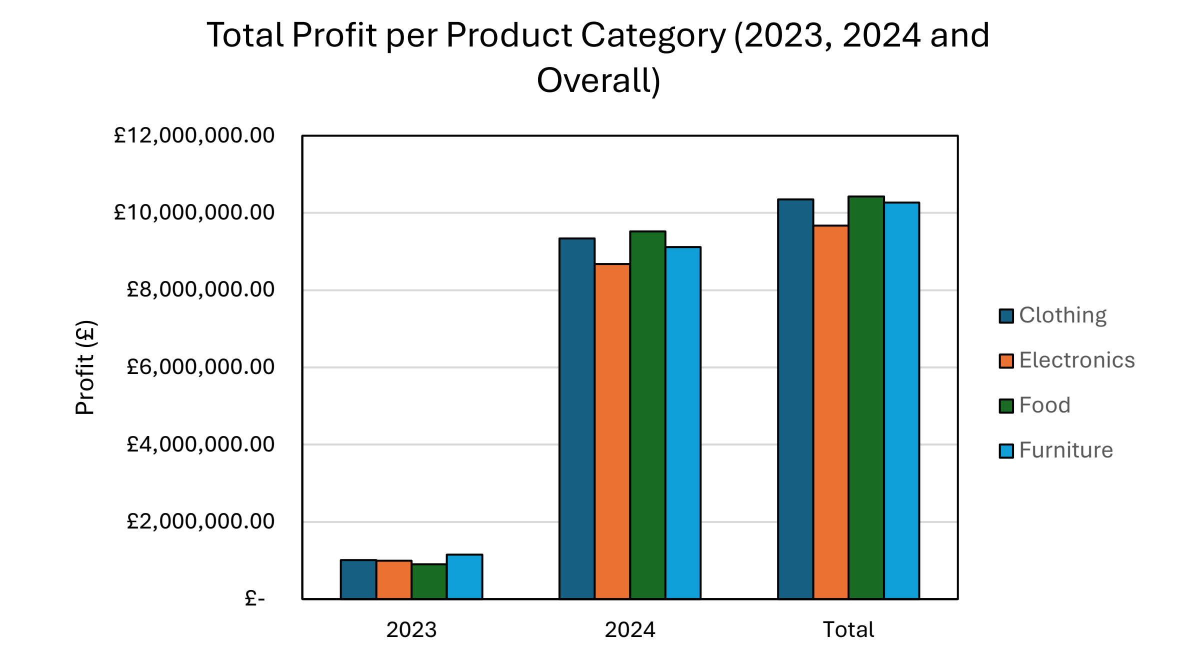
**Heatmap of Aggregated Monthly Revenue by Region (2023–2024): Summed Revenue for Each Month Across Two Years**



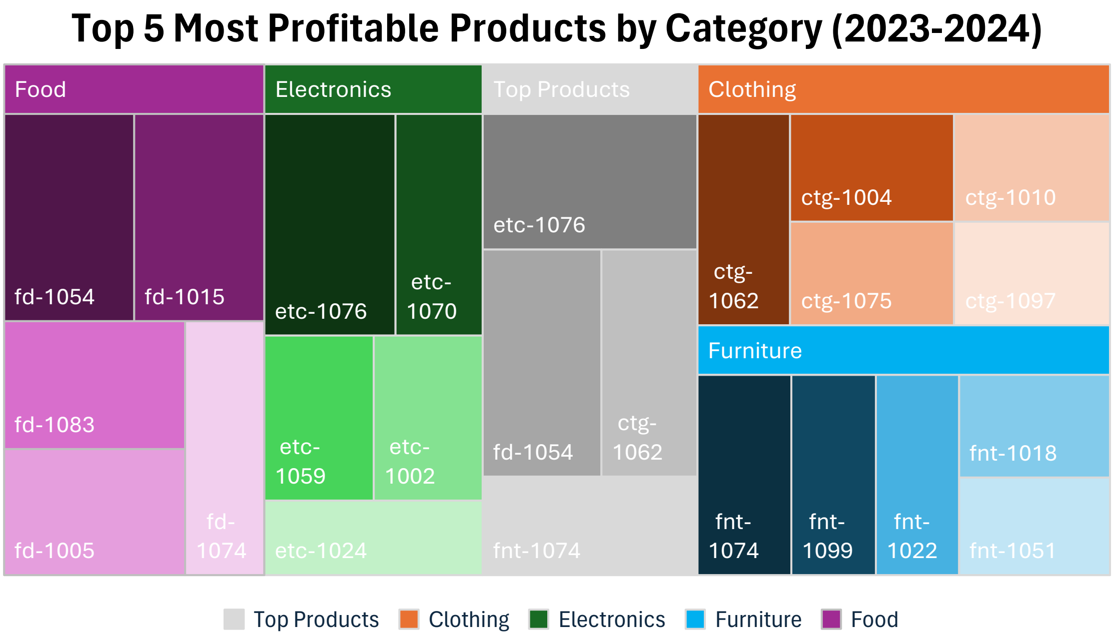

This heatmap illustrates combined monthly revenue by region for the years 2023 and 2024, where revenues for corresponding months across both years have been aggregated. The North region consistently contributed the highest overall revenue, although it experienced noticeable declines in April and September. The East region displayed strong seasonal peaks, particularly in Q1 and Q4, with March standing out as its most revenue generating month. In contrast, the South and West regions underperformed in terms of revenue generation. The South showed significant volatility, with a revenue peak in May and a relatively stable period from August to October. The West exhibited the greatest inconsistency throughout both years, with November marking its weakest performance. Overall, January, March, and May emerged as the top-performing months across all regions, while July generally recorded the lowest revenue contributions.

## Performance across Product_Category

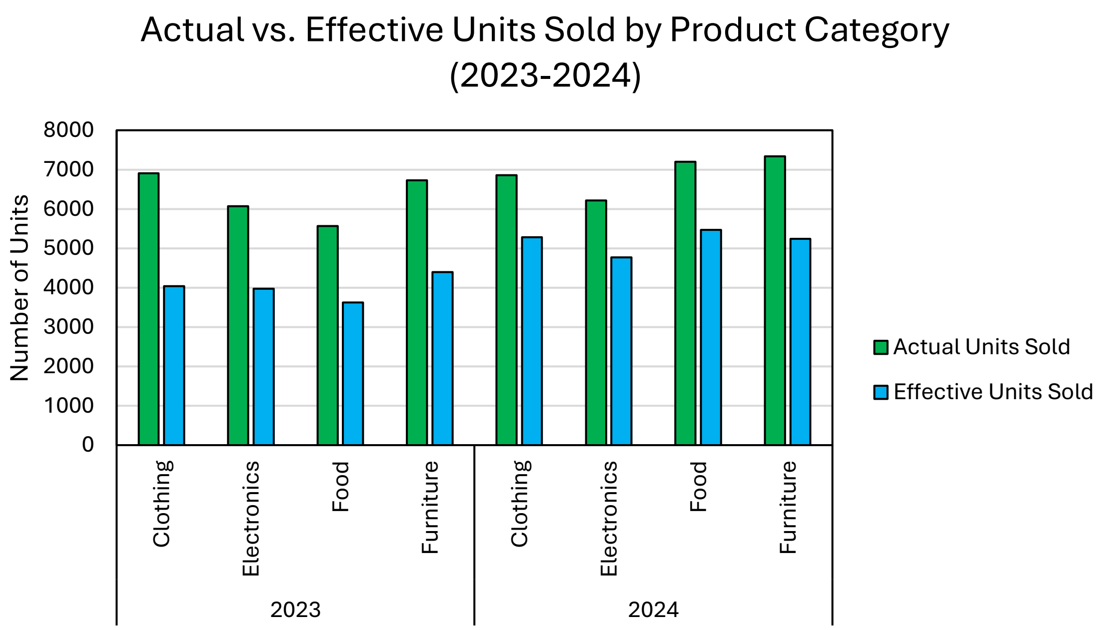
To gain insight into the strengths and weaknesses of the product portfolio, this section will analyse performance across different product categories. An understanding of the key drivers of success at the category level is crucial for informing decisions related to inventory management, marketing strategies, and future product development. The analysis presented herein will highlight key trends and identify areas of opportunity within our product offerings.




The total profit per product category between 2023 and 2024 reveals that Food and Clothing were the most profitable categories overall. Notably, all categories saw significant profit growth from 2023 to 2024, with Food demonstrating the largest relative increase, surpassing Clothing in profit in 2024 and in the overall total. Electronics is the overall least profitable category.




The tree map chart reveals that within the top 5 most profitable products by category between 2023 and 2024, etc-1076 in Electronics and fd-1054 in Food stand out as highly profitable, as seen appearing in the 'Top Products' overview. While Clothing's profitability is led by ctg-1062, the Furniture category shows a more distributed profitability among its leading products. The varying sizes of the rectangles highlight the differing levels of profitability across these top-performing products and categories. Food is the most profitable product category.




This chart shows the comparison of actual versus effective units sold in 2023 and 2024. This allows us to see the units sold that are lost/ or in other words not realised in profit due to discounts. This reveals that in 2023, actual sales are consistently higher across all product categories and years than the effective units. The Clothing category experiences the most significant difference, suggesting a higher rate of discounts on average. Conversely, Food shows the most efficient conversion from actual to effective sales, revealing less discounts applied on these products. All categories saw an reduction in the difference between actual and effective units sold in 2024.

## Discounts

To assess how discounts influenced sales performance, an analysis was carried out examining the relationship between discount levels and average profit. By comparing the average profits with and without discounts, insights were drawn into whether offering price reductions led to increased customer spending or if it negatively impacted profitability.

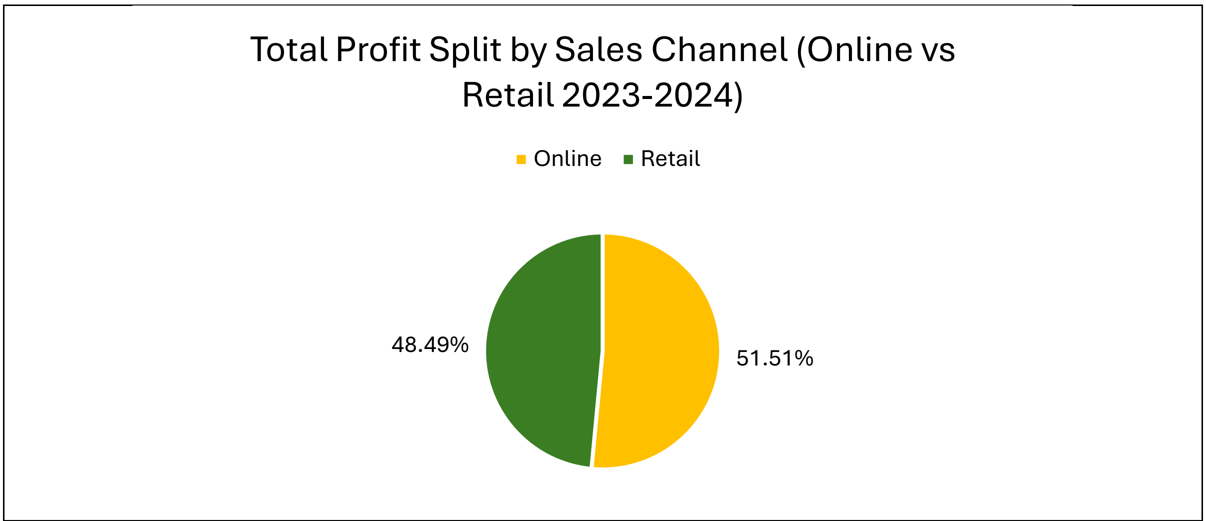
*Table 4 Comparison of Average Profit With and Without Discounts Across 2023–2024, Alongside Corresponding Discount Rate*

| Year | Average Profit (No Discount) | Average Profit (Discount) | % Change in Profit | Average Discount Rate |
|---|---|---|---|---|
| **2023** | £ 138.08 | £ 164.25 | (+) 18.95% | 4.86% |
| **2024** | £ 845.12 | £ 1,342.51 | (+) 58.85% | 12.46% |

While the average discount rate in 2023 was relatively modest at 4.86%, it corresponded to an 18.95% increase in average profit compared to non-discounted transactions. This suggests that discounting was effective in driving higher-value sales, potentially through increased volume or upselling. Similarly, in 2024, a more aggressive average discount rate of 12.46% aligned with a significant 58.85% rise in average profit, reinforcing the positive relationship between strategic discounting and sales profitability.

## Sales Channels (Online vs Retail)

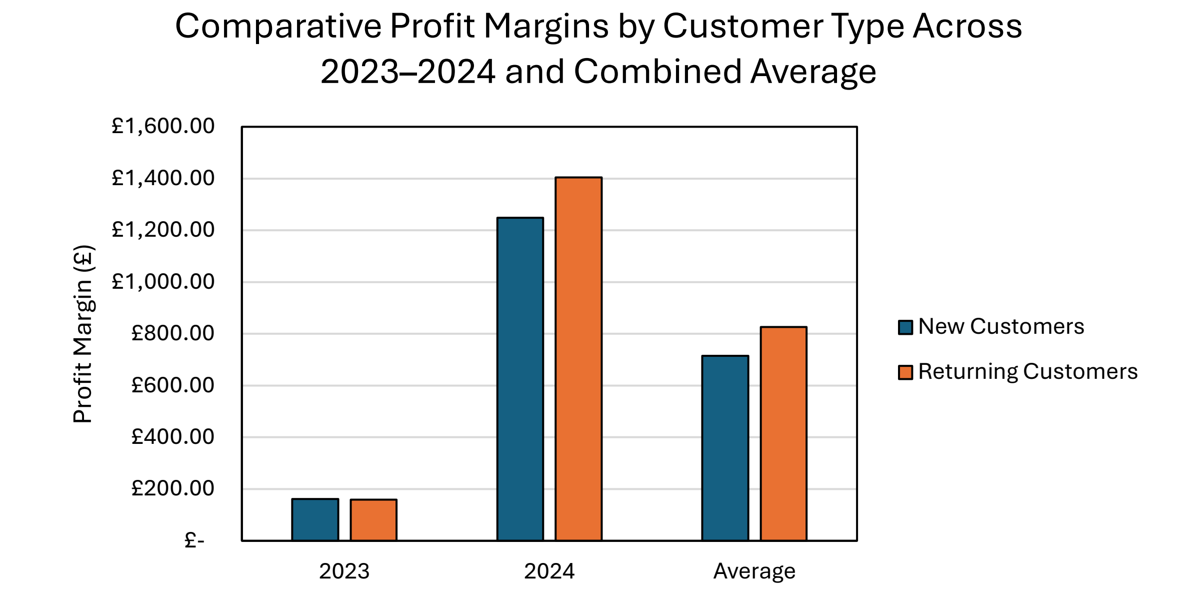
To evaluate the effectiveness of different sales strategies, this section explores the performance of Online and Retail sales channels in terms of profitability. By comparing profit margins across both channels, we can identify which model yields better returns and uncover potential areas for improvement.




This chart illustrates the overall profit split between Online and Retail sales channels across all product categories. The analysis indicates that Online sales yielded slightly higher average profit margins than Retail, highlighting potential efficiencies in the online model. Contributing factors may include reduced overhead costs, broader customer reach, and the strategic use of discounts. However, given the relatively narrow gap between the two, there may still be untapped potential for increased profitability within the Online channel. Despite this, Retail remains a substantial contributor to total sales, though Online sales appear more profitable on a per-transaction basis.

### New Customers vs Returning Customers

Understanding the profitability of different customer segments is essential for evaluating sales strategies and long-term business sustainability. This section explores the profit margins generated by new versus returning customers over a two-year period (2023–2024).




The chart compares profit margins between new and returning customers across 2023, 2024, and their combined average. While both customer types had similar profit margins in 2023, there was a notable increase in 2024, with returning customers generating significantly higher marginal profits. On average, returning customers consistently outperformed new customers, indicating their stronger contribution to overall profitability.

# Recommendations/Future Work

### Future Data Collection

To enhance the quality, depth, and accuracy of future analyses, the following recommendations are proposed for improving the structure and design of the dataset:

**1. Introduce Unique Customer and Transaction Identifiers**

Including a Customer_ID column would allow for the tracking of individual customers across multiple purchases, enabling richer analysis of customer retention, frequency of purchases, and lifetime value. Similarly, incorporating a unique Transaction_ID, alongside a timestamp, would enable transaction-level granularity, providing better visibility into purchase behaviour within a single day and supporting more detailed time-based trend analysis.

**2. Assign Globally Unique Product Identifiers**

A consistent and unique Product_ID for each distinct product, regardless of category, would facilitate more accurate and powerful product-level analysis. This would eliminate ambiguity and allow analysts to identify top-performing products, analyse product lifecycle trends, and assess category-level contribution with greater confidence.

**3. Standardise Variable Relationships**

To reduce variance and maintain logical consistency across key financial variables, such as Sales_Amount, Unit_Cost, Unit_Price, and Discount, it is important to apply business logic during data generation or collection. Ensuring these fields align with realistic pricing and discounting models would provide more meaningful insights and improve the reliability of profit-based analysis.

**4. Include Additional Customer and Product Metadata**

Enriching the dataset with attributes such as customer demographics (e.g. customer segment) or product details (e.g., brand, size, warranty) would open opportunities for deeper segmentation and targeted analysis. These variables are particularly valuable when assessing customer behaviour, market performance, or category trends.

**5. Improve Time Granularity**

Incorporating timestamps alongside dates would allow for more refined time-series analysis, including hourly or intraday patterns. This is particularly relevant for understanding peak sales periods, transaction volume trends, and campaign effectiveness during specific times of the day.

**6. Ensure Consistent Data Generation Rules**

When working with synthetic or randomised datasets, it is essential to define clear rules that govern how variables are generated and how they relate to one another. This ensures data integrity and reduces the need for manual verification of variable logic during analysis.

**7. Integrate External Data Sources**

Incorporating external data sources, such as market trends, competitor performance, and economic indicators, can provide a broader context for analysis. This can help identify external factors influencing sales performance and support more comprehensive strategic planning.

**8. Develop Predictive Models**

Utilising machine learning techniques to develop predictive models can enhance the ability to forecast future sales trends, customer behaviour, and product performance. These models can be trained on historical data and continuously updated with new information to improve accuracy.

**9. Develop an Interactive Dashboard**

Design a comprehensive and dynamic dashboard to consolidate key sales metrics, trends, and performance indicators in one view. Utilise pivot charts, slicers, and filters to allow users to interact with the data, such as drilling down by region, product category, sales channel, or time period. A well-structured dashboard enhances data accessibility and supports quick, data-driven decision-making by stakeholders.

# Conclusion

This analysis of the synthetic sales dataset has provided valuable insights into various aspects of sales performance, customer behaviour, and regional trends. Despite the challenges posed by the synthetic nature of the data and the limitations in its structure, the findings offer a comprehensive understanding of the dataset's dynamics.

Key observations include the importance of unique identifiers for customers and products, the impact of randomisation on profit analysis, and the need for more granular time data. Additionally, the analysis highlighted the significance of consistent data generation rules and the potential benefits of integrating external data sources and predictive models.

By addressing these challenges and implementing the recommendations outlined in the

Recommendations/Future Work section, future analyses can achieve greater accuracy, depth, and relevance. This will enable more informed decision-making and strategic planning, ultimately contributing to improved business outcomes.

Overall, this project has demonstrated the value of thorough data preparation, rigorous analysis, and thoughtful interpretation of results. The insights gained from this analysis can serve as a foundation for future research and practical applications in the field of sales analytics.