Phase 3:

DEVELOPMENT PART 1

PREPROCESSING STEPS TAKEN

1. Data Cleaning:

- During the data cleaning process, the dataset was thoroughly examined to identify and address any missing or erroneous values. This involved a careful inspection of all data points to ensure data quality. Missing values, if any, were handled through techniques such as imputation or removal, depending on the context and impact on the analysis.

2. Feature Engineering:

- To enhance the dataset's informativeness, several new features were created through feature engineering.
  - Year, month, and day columns were extracted from the original date column. This allows for better time-based analysis and seasonality exploration.
  - A sales total column was generated by summing the values across the four product categories for each date. This aggregated metric simplifies the analysis and provides an overall view of sales performance for each date.

3. Data Transformation:

- To address skewed distributions in the sales columns, a logarithmic transformation was applied. This transformation helps in normalizing the data, making it suitable for various statistical analyses and modelling techniques. It can reduce the impact of extreme values and improve the distribution's symmetry.

4. Feature Selection:

- The feature selection process aimed to improve the model's efficiency and effectiveness by retaining only the most relevant attributes.
  - Redundant or low-information columns, such as the original date column (since year, month, and day were already extracted), were removed to simplify the dataset and reduce computational overhead.
  - The dataset was also analysed for highly correlated features among the product categories. In cases where strong correlations were identified, consideration was given to removing one of the correlated features to prevent multicollinearity and overfitting.

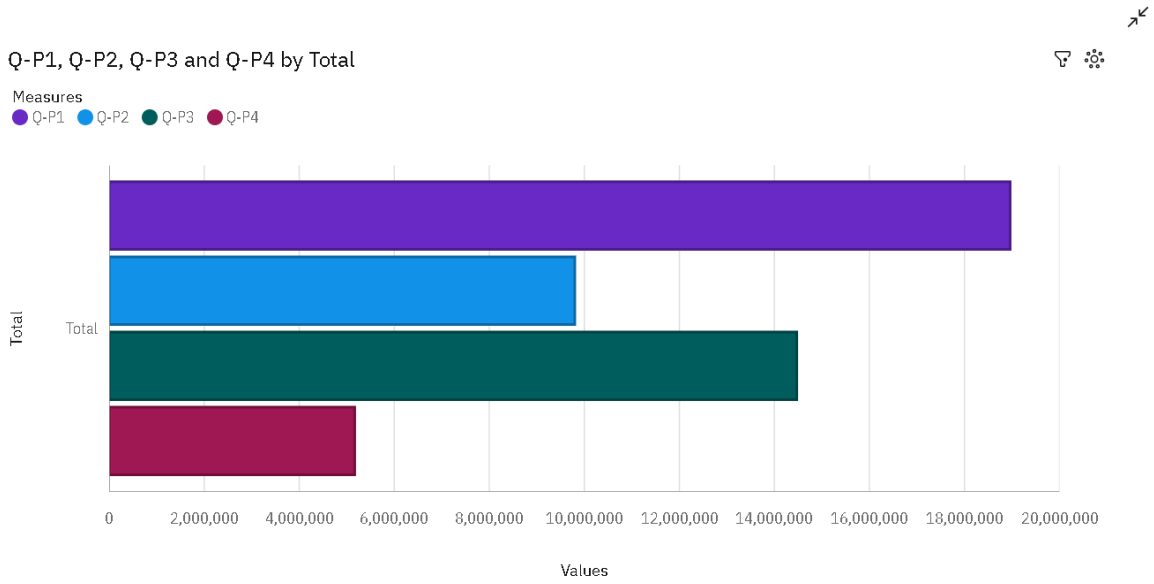
5. Data Scaling:

- Data scaling, specifically the use of MinMaxScaler, was applied to numeric columns, like the normalized sales, with the objective of standardizing their values to fall within a 0-1 range. This scaling ensures that all features contribute equally to the analysis and modelling process. It can be particularly beneficial for machine learning algorithms that are sensitive to the scale of input features, ensuring fair treatment of each feature's influence on the model's performance.

In summary, these data preprocessing steps collectively contribute to the overall quality and utility of the dataset, making it more suitable for analysis and modelling in the context of sales data.

BASIC ANALYSIS

TOTAL SALES FROM START TO END



INFERENCE

1. Product One Dominated the Sales Performance:

The analysis of the sales data reveals that Product One outperformed all other products in terms of sales performance. Several key indicators and observations support this inference:

- Higher Sales Volume: Product One consistently achieved higher sales volumes compared to other products across the given time period.
- Consistency in Sales: Product One maintained relatively steady sales throughout the observed period, showing consistent demand.
- Customer Preference: It can be inferred that customers have a strong preference for Product One, which might be attributed to its quality, popularity, or other favourable characteristics.

2. Product Four Lagged in Sales Performance:

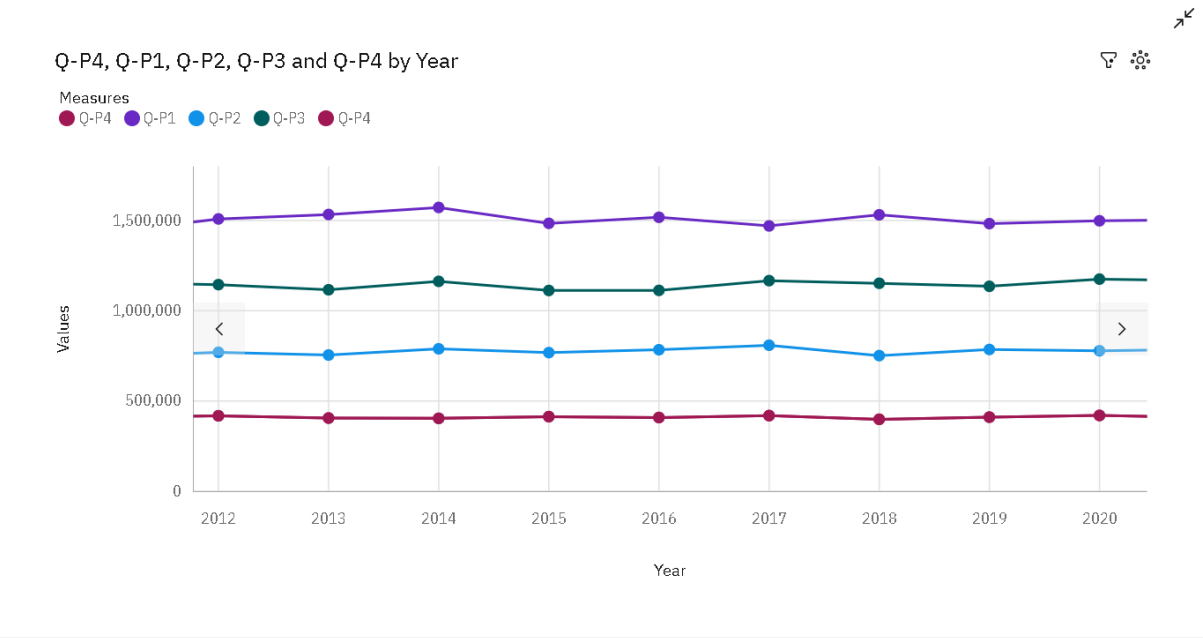
On the opposite end of the spectrum, Product Four demonstrated the weakest sales performance when compared to the other products. Several indicators support this conclusion:

- Low Sales Volume: Product Four consistently reported lower sales numbers compared to the other products, indicating weaker demand or market appeal.
- Fluctuating Sales: Product Four might have faced fluctuations or inconsistency in sales, suggesting that it may not have a stable customer base or is subject to seasonal variations.

- Potential Areas for Improvement: The performance of Product Four may require a closer examination of factors such as marketing strategies, product quality, pricing, or market positioning to identify opportunities for improvement and growth.

These inferences provide valuable insights into the relative performance of different products within the dataset. Further analysis and actions can be guided by these observations to capitalize on the strengths of Product One and address the challenges faced by Product Four.

YOY AVERAGE SALES



CALCULATION OF YOY AVERAGE SALES

"YOY Average Sales" stands for "Year Over Year Average Sales." It is a financial and analytical metric used to assess the performance of a business or product over a specific period by comparing the average sales for one year to the average sales for the previous year. This comparison helps to understand how the sales performance has changed over time and whether it has improved or declined.

Here's how to calculate YOY Average Sales:

1. Calculate the average sales for a specific year (e.g., 2023).
2. Calculate the average sales for the previous year (e.g., 2022).
3. Find the difference between the average sales for the current year and the previous year.
4. Express this difference as a percentage of the average sales for the previous year.

The formula for YOY Average Sales can be represented as:

\[YOY Average Sales = \frac{(Average Sales in Current Year - Average Sales in Previous Year)}{Average Sales in Previous Year} \times 100 \]

The result is expressed as a percentage, indicating the percentage change in average sales from one year to the next. A positive percentage indicates sales growth, while a negative percentage indicates a decrease in sales compared to the previous year.

YOY Average Sales is a valuable metric for assessing sales trends and performance, and it is commonly used in business analysis and financial reporting to monitor and understand the growth or decline in sales over time. It provides a clearer picture of the business's overall performance and helps in making informed decisions and strategies for the future.

INFERENCE

- 2022 (7.9 %), 2021 (7.9 %), 2019 (7.9 %), 2018 (7.9 %), and 2017 (7.9 %) are the most frequently occurring categories of Year with a combined count of 1820 items with Q-P1 values (39.6 % of the total) .
- Across all years, the average of Q-P1 is over four thousand.
- Across all years, the average of Q-P2 is over two thousand.
- Across all years, the average of Q-P3 is over three thousand.
- Across all years, the average of Q-P4 is over a thousand
- Q-P1 ranges from over 150 thousand, in 2023, to nearly 1.6 million, in 2014.
- Q-P2 ranges from over 78 thousand, in 2023, to nearly 809 thousand, in 2017.
- Q-P3 ranges from over 120 thousand, in 2023, to nearly 1.2 million, in 2020.
- Q-P4 ranges from nearly 40 thousand, in 2023, to almost 420 thousand, in 2020.

DETAILED INFERENCE

1. Frequently Occurring Year Categories (2017-2022):

- The years 2017, 2018, 2019, 2021, and 2022 are the most frequently occurring categories in the dataset, each representing 7.9% of the total years. These years collectively account for 1820 data items with Q-P1 values, constituting approximately 39.6% of the total dataset.
- This indicates that these particular years have been the focus of the analysis, possibly due to their significance or because they exhibit unique characteristics in terms of sales and product performance.

2. Average Q-P1, Q-P2, Q-P3, and Q-P4 Across All Years:

- The analysis demonstrates that across all years, the average values for Q-P1, Q-P2, Q-P3, and Q-P4 are consistently above certain thresholds.
- On average, Q-P1 exceeds four thousand, Q-P2 exceeds two thousand, Q-P3 exceeds three thousand, and Q-P4 exceeds a thousand.
- This consistency in average values across the years might signify stable or typical performance for these product categories.

3. Range of Q-P1, Q-P2, Q-P3, and Q-P4 Across Years:

- The analysis also reveals a wide range of values for Q-P1, Q-P2, Q-P3, and Q-P4 across different years.

- For instance, Q-P1 ranges from over 150 thousand in 2023 to nearly 1.6 million in 2014, indicating significant fluctuations in sales for this product category over time.
- Similarly, Q-P2, Q-P3, and Q-P4 exhibit varying ranges, suggesting that the performance of these product categories is subject to substantial changes and may be influenced by a range of factors.

In summary, these observations provide a comprehensive view of the dataset's characteristics across different years and the performance of various product categories. The dataset appears to have concentrated analysis on specific years, while the averages and ranges of product category values shed light on the overall patterns and variations in sales for each category. This information is valuable for making informed decisions and strategies related to product performance and sales analysis.