MACQUARIE
University

LuminaTech Lighting: Data-Driven Insights into Sales, Profitability, and Customer Retention

- **Group Number** = 46
- **Member_01** = "Pradyot Jain", "48479985"
- **Member_02** = "Rakshitha Kundapura Raghavendra", "48355631"
- **Member_03** = "Harshit Arora", "47971614"

- **Word Count** = 6129

Contents

1.0 Data Description for LuminaTech Lighting Project.

This dataset contains sales, customer, and product-related information for LuminaTech Lighting. The data spans multiple years and provides key insights into sales performance, customer behavior, and product inventory management.

The datasets for the years 2012 and 2013 contain the following number of rows and columns:

- 2012 Dataset: 1,037,205 rows and 41 columns
- 2013 Dataset: 951,177 rows and 41 columns

2.0 Load the Datasets and Show the Basic Information.

We first load both datasets in the jupyter notebook and then we showed the basic information like number of unique values, column names in the data, etc.

3.0 Section 01: Clean the Data

3.01 Dropping Unnecessary Columns:

Justification for Dropping Unnecessary Columns:

We removed columns irrelevant to our analysis objectives, focusing on features that contribute to customer behavior insights, sales prediction, or churn analysis.

- **Internal/Admin Codes**: Columns like light_source, contact_method_code, and commission_group_code are used internally and hold no predictive value for customer or sales analysis.
- **System Timestamps**: dss_update_time records update times, not relevant to understanding sales patterns.
- **Line-Item Details**: line_number specifies invoice line items, which are unnecessary for high-level insights.
- **Inventory Fields**: Columns like item_source and abc_class_volume focus on inventory classifications, unrelated to sales or churn.

Dropping accounting_date: This column duplicates invoice_date, creating redundancy. Keeping only invoice_date preserves time-based information for analysis.

Dropping market_segment: With identical values across rows, this column doesn't contribute to customer segmentation or market insights, simplifying our dataset for meaningful analysis.

3.02 Missing Values:

There are no missing Values in the Datasets.

3.03 Handling Currency Column and converting the other currency to AUD:

1. **Dropping Rows with Missing Currency Values**
    - **Reasoning**: Currency information is essential for accurate financial analysis. Without knowing the currency, interpreting transaction values becomes unreliable.
    - **Justification**: Removing rows with missing currency values ensures the dataset includes only complete financial data, preventing inaccuracies in aggregated calculations. This step avoids misleading results, such as treating values as AUD by default, which could distort monetary analysis.
2. **Replacing 'AUS' with 'AUD'**
    - **Reasoning**: The abbreviation 'AUS' is likely a variant for the Australian Dollar and should be standardized to 'AUD'.
    - **Justification**: Consistent naming avoids duplicate categories in currency analysis, aligns with ISO 4217 standards, and simplifies aggregation. This replacement enhances clarity and reduces complexity.
3. **Converting All Values to AUD for Uniformity**
    - **Reasoning**: Converting multiple currencies to AUD allows for consistent analysis.
    - **Justification**: A single currency standard enables accurate calculations and comparisons across records, making aggregated analysis more reliable and easier to interpret.

These steps ensure a clean, standardized dataset, supporting reliable analysis.

3.04 Duplicate Rows:

We are dropping the duplicate rows, since the datasets are very big, dropping them will not affect our analysis.

3.05 Intercompany Transactions and Head Office Sales:

Justification for Excluding Intercompany Transactions:

1. **Identified Intercompany Transaction Codes**
    - **710 - Head Office Sales**: Transactions by the head office for internal transfers, not external sales.
    - **720 - Intercompany Sales**: Internal transactions within the corporate group, not generating revenue from outside customers.

- **520 - Inlite - NZ** and **545 - Head Office NZ**: Regional branches in New Zealand, indicating internal transfers rather than external sales.

2. **Reason for Exclusion**: Intercompany transactions are internal and do not reflect genuine revenue. Including them would inflate revenue and costs, misrepresenting the company's performance. Excluding these transactions allows us to focus on revenue generated from external customers, providing an accurate view of actual sales and operational costs.
3. **Implementation**: We flagged and filtered out rows with customer_district_code values 710, 720, 520, and 545, ensuring our analysis reflects only external operations and true revenue.

3.06 Convert DataType:

Justification for Converting Data Types:

Converting the data types ensures that each column is optimized for its specific purpose:

1. **Date Columns**: Converting columns like invoice_date, and order_date to datetime format allows for more accurate time-based calculations and comparisons, such as analyzing trends over time or calculating the time gap between order and invoice.
2. **Categorical Columns**: All columns with object datatype are converted to the category data type to reduce memory usage and improve processing efficiency. This also ensures that the dataset reflects the categorical nature of these variables, which are identifiers and should not be treated as numeric or continuous data.

By converting these data types appropriately, we enhance the dataset's integrity and ensure that our analysis will be both efficient and accurate.

3.07 Error Checking in Time gap between Order Date and Invoice Date:

Justification for Time Gap Check Between order_date and invoice_date:

1. Objective: The time gap between order_date and invoice_date was analyzed to identify potential data entry errors or unusual processing delays. This check ensures consistency in the data by confirming that invoices are generated within a reasonable timeframe after orders are placed.

2. Methodology:

- Time Gap Calculation: For each transaction, the difference in days (time_gap_days) between invoice_date and order_date was computed.
- Unusual Gap Flagging: Rows with negative time gaps (indicating that the invoice_date was recorded before the order_date), or excessively large gaps (more than 30 days) were flagged as unusual_gap = True.

3. Results:

- **2012 Dataset**: No rows were flagged with unusual time gaps in external_sales_data_2012, indicating reasonable and consistent time gaps between order and invoice dates in 2012.
- **2013 Dataset:** Similarly, no rows were flagged with unusual time gaps in external_sales_data_2013, suggesting the 2013 transactions also meet expected standards for time gap consistency.

4. Conclusion: Since there are no unusual time gaps in either dataset, we conclude that the order_date and invoice_date fields are consistent and do not require further cleaning for date-related anomalies. This consistency enhances the dataset's reliability for further analysis, ensuring that the order and invoicing processes align with expected business timelines.

3.08 Handling Zero Values in Sales, Cost and Quantity and Negative Sales:

Justification for Handling and Flagging Different Transaction Scenarios:

In this analysis, multiple cases were identified where value_sales, value_cost, and value_quantity contain zero or negative values, each representing distinct transaction types. Properly flagging and handling these cases is essential to ensure accurate revenue calculations and meaningful insights into business operations. Below is a summary of each case and the handling strategy:

Case 1: value_sales is negative, value_cost and value_quantity are zero:

- **Interpretation**: Likely represents refunds or returns without associated cost or restocking.
- **Handling Strategy**: Flagged as is_refund and excluded from revenue calculations to prevent negative sales from affecting total revenue.

Case 2: value_sales is positive, value_cost and value_quantity are zero:

- **Interpretation**: Represents service sales or non-inventory transactions (e.g., consulting fees).
- **Handling Strategy**: Flagged as is_service_sale and included in revenue calculations, as these represent legitimate sales.

Case 3: All three values are negative:

- **Interpretation**: Likely an inventory adjustment or reversal, where revenue, cost, and quantity are adjusted.
- **Handling Strategy**: Flagged as is_inventory_adjustment and excluded from revenue calculations.

Case 4: value_sales and value_quantity are positive, value_cost is zero:

- **Interpretation**: May represent a zero-cost sale or data entry error.
- **Handling Strategy**: Flagged as is_zero_cost_sale, included in revenue calculations, and reviewed separately.

Case 5: value_sales and value_cost are positive, value_quantity is zero:

- **Interpretation**: Could represent a non-quantity-based sale, such as licensing fees.
- **Handling Strategy**: Flagged as is_non_quantity_sale and included in revenue calculations.

Case 6: value_sales is zero, value_cost and value_quantity are negative:

- **Interpretation**: Likely an inventory write-down or stock adjustment.
- **Handling Strategy**: Flagged as is_stock_adjustment and excluded from revenue calculations.

Case 7: value_sales is zero, value_cost and value_quantity are positive:

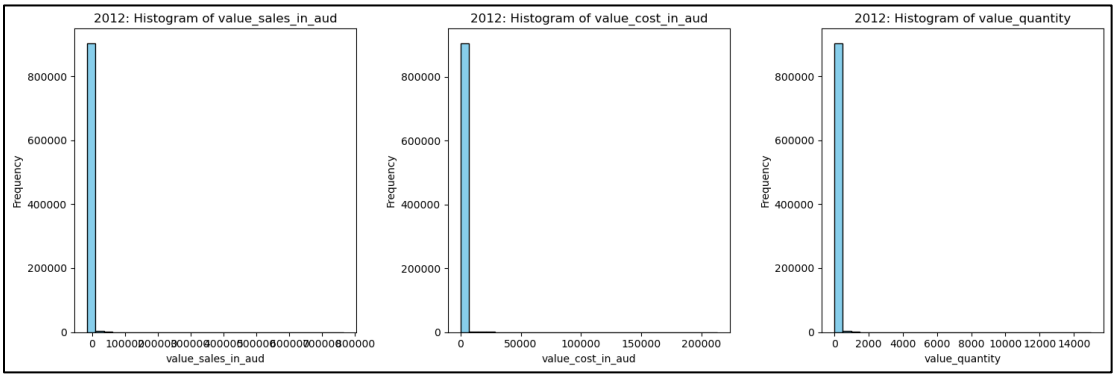
- **Interpretation**: Represents promotional items or free samples.
- **Handling Strategy**: Flagged as is_promotion and excluded from revenue calculations.

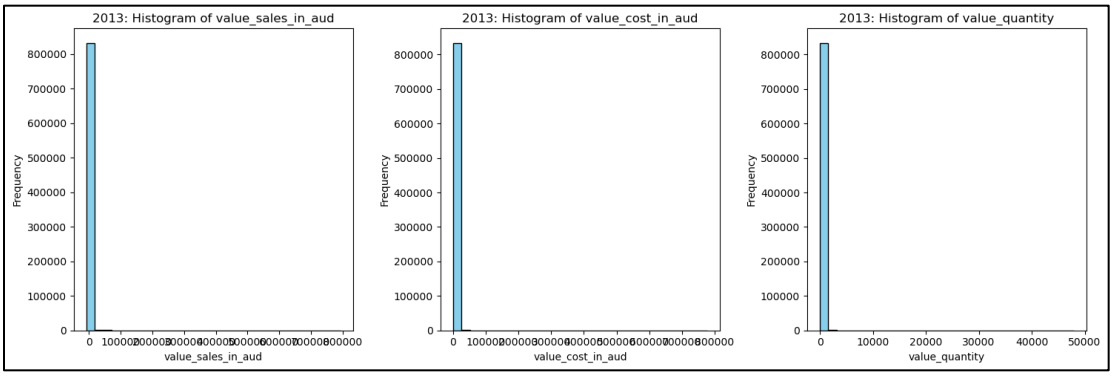
Summary: Each case was flagged within the 2012 and 2013 datasets, allowing separation for detailed analysis. Excluding refunds, inventory adjustments, stock adjustments, and promotions ensures that only legitimate revenue-generating transactions contribute to a true representation of net sales.

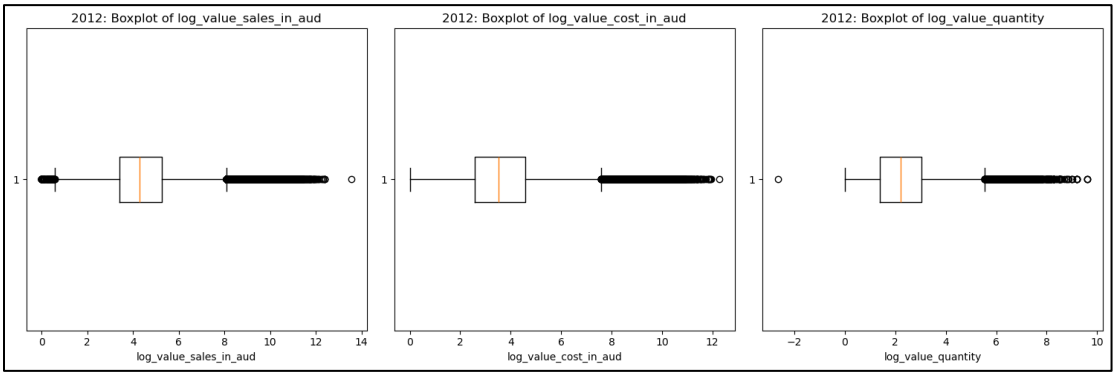
3.09 Unit Price and Unit Cost:

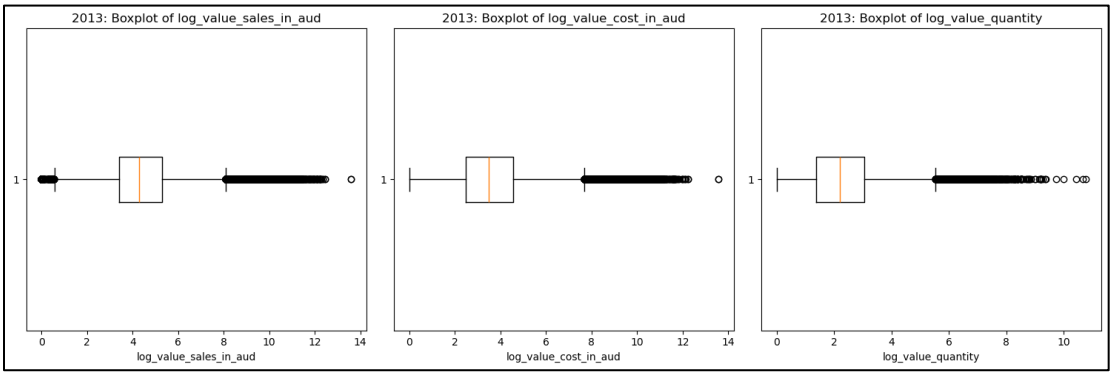
We created Unit Price and Unit Cost column for both the datasets.

3.10 Outliers and Skewness:









Justification for Detecting and Handling Outliers:

1. Identifying Skewness and Outliers: In our initial data exploration, we observed high skewness in the value_sales_in_aud, value_cost_in_aud, and value_quantity columns in both the 2012 and 2013 datasets. High skewness often results in long-tailed distributions, introducing extreme values far from the mean. These may represent legitimate high-value transactions or potential data entry errors, so careful handling of these outliers is essential.

2. Why Handle Skewness Before Detecting Outliers?

Given the skewness, standard outlier detection methods like Z-score would likely flag excessive outliers, as these methods assume a normally distributed dataset. To address this, we applied a log transformation to each variable, which compresses large values and makes the distribution more symmetric, allowing for more reliable outlier detection.

3. Choosing IQR for Outlier Detection and Removal:

After transformation, we chose the Interquartile Range (IQR) method for outlier detection and removal based on these reasons:

- **IQR Method:** This method is robust for non-normally distributed data and identifies outliers as values falling outside 1.5 times the IQR below the first quartile or above the third quartile. It is well-suited for transformed data and skewed distributions.
- **Removal:** Since outliers represented less than 5% of each column, we removed these rows, minimizing any impact on analysis and preventing extreme values from distorting the dataset.

4. Benefits of Outlier Removal:

- **Data Integrity:** Removing a small percentage of extreme values helps create a dataset that better represents typical values and prevents skewed results.
- **Minimal Impact on Analysis**: With fewer than 5% of data points removed, the overall structure and patterns are preserved.
- **Improved Reliability:** Removing outliers enhances the accuracy of statistical methods, especially those sensitive to extremes.

5. Using Robust Metrics in Analysis:

For summary statistics and further analysis, we recommend median-based or quantile-based metrics. These approaches are resistant to outliers and provide stable, representative results.

Summary:

1. **Address Skewness:** Log transformation prepares data for more accurate outlier detection.
2. **Apply IQR and Remove Outliers:** Less than 5% of data points removed for accurate representation of typical values.
3. **Use Robust Metrics:** Median-based or quantile-based metrics improve analysis stability.

This approach preserves dataset integrity, minimizes extreme value influence, and provides a solid foundation for reliable analysis.

3.11 Dropping Extra columns and other new columns created that are not needed:

Justification for Creating cleaned_sales_data_2012 and cleaned_sales_data_2013:

To enhance data clarity and streamline analysis, we created cleaned sales data 2012 and cleaned sales data 2013 by removing columns no longer necessary for our analysis. Below is the rationale for excluding each:

Columns Removed:

1. **value_sales, value_cost, value_quantity**: Original values were replaced by log-transformed versions that better handle skewness.
2. **value_sales_in_aud, value_cost_in_aud**: Retained only log-transformed versions to reduce redundancy.
3. **is_service_sale, is_zero_cost_sale, is_non_quantity_sale**: These flags filtered specific transactions; after ensuring relevant rows are included or excluded, they are no longer needed.

Benefits of the Cleaned Datasets:

- **Reduced Complexity**: Focuses only on necessary columns for analysis.
- **Enhanced Readability**: Eases data interpretation by removing redundant columns.
- **Improved Efficiency**: Reduces dataset size, optimizing storage and processing.

The cleaned datasets provide a streamlined and analysis-ready view, focusing on essential variables.

4.0 Section 2: Exploratory Insights

4.01 Insight 1: Top-Selling Products:

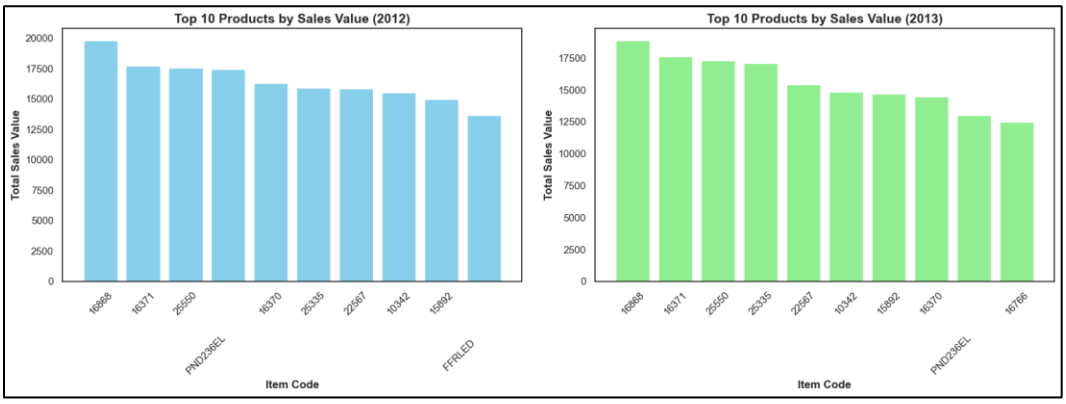
Objective: This analysis aims to identify the top-selling products by sales value and quantity for 2012 and 2013. These insights help management understand product performance in terms of revenue and volume, informing inventory, marketing, and sales strategies.

Method Used:

1. **Data Aggregation**: Total sales value and quantity sold were calculated for each product by grouping the data by item_code for each year.
2. **Top 10 Products Selection**:
    - **Sales Value**: The top 10 products contributing the most to revenue in 2012 and 2013 were identified.
    - **Quantity Sold**: The top 10 products with the highest sales volume for each year were selected.
3. **Visualization**: Bar charts were created to visualize the top products based on sales value and quantity sold for each year.

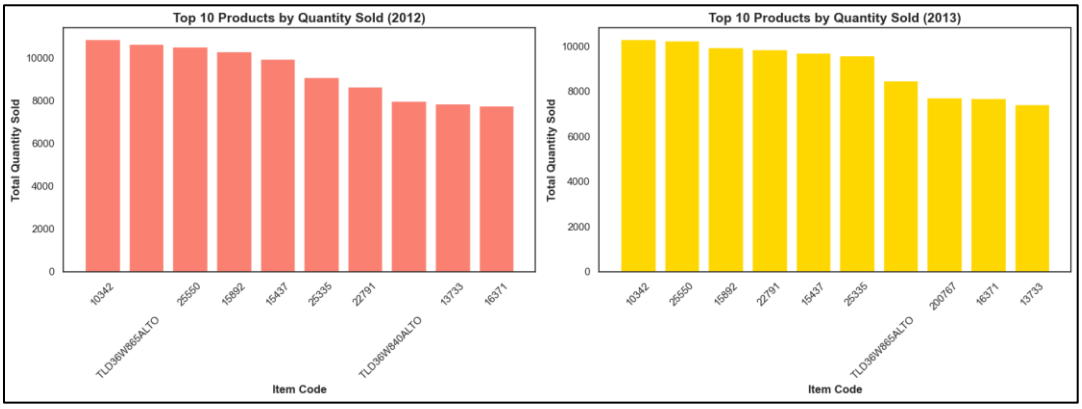
Insights

1. **Top Products by Sales Value**:



- **2012**: Product 16868 led in sales value, followed by 25550, PND236EL, and 16371. These products consistently rank high, indicating they are significant revenue contributors.
- **2013**: Product 16868 again topped the list in sales value, maintaining its previous year's position. Products like 25550, 16371, and PND236EL also remained prominent, suggesting stable demand across the two years.

2. **Top Products by Quantity Sold**:



- **2012**: Product 10342 had the highest quantity sold, significantly outpacing other items. This may indicate high demand, possibly due to a lower price point or broad application. Other products with high quantities sold include TLD36W865ALTO, 15892, and 25550.
- **2013**: Product 10342 continued to rank first in quantity sold, demonstrating its ongoing popularity. High-quantity products included 15892, 25550, and TLD36W865ALTO, showing consistency in demand.

3. **Consistency Across Years**: Products such as 16868, 10342, 16371, and 25550 consistently rank high in both sales value and quantity sold for 2012 and 2013.

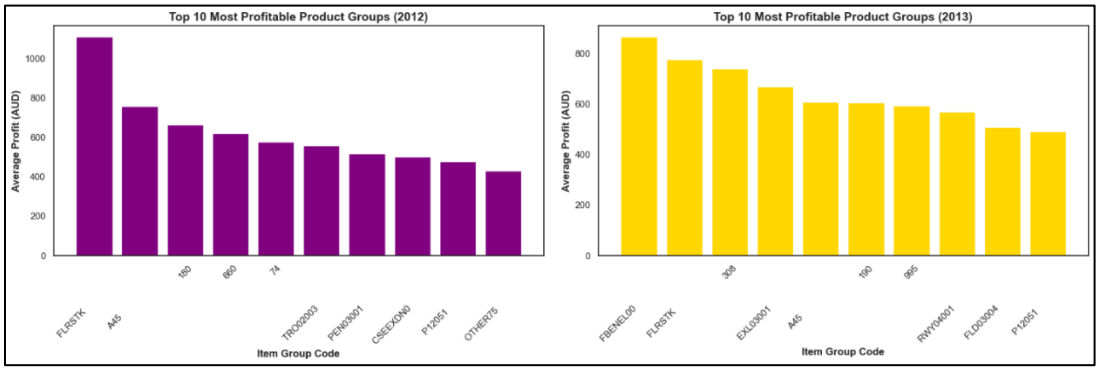
This suggests that these items are core products in the inventory with sustained customer demand.

Value in Management:

1. **Inventory Management**: High-demand items like 10342 require careful stock management to prevent shortages. Products with high sales values, such as 16868, may need prioritized availability due to their substantial revenue impact.
2. **Sales and Marketing Strategy**: Revenue-driving products, like 16868 and 16371, could benefit from aggressive promotion to maintain or increase their contributions. For high-quantity, lower-value items like 10342, management could consider bulk promotions to boost profitability.
3. **Product Line Optimization**: Products excelling in both sales value and quantity (such as 10342 and 25550) are likely essential to the business. Management might view these products as strategic assets, ensuring their consistent availability, while considering lower-performing products for potential discontinuation.

By examining top products by sales value and quantity, management gains actionable insights into product demand and revenue generation, aiding in strategic decisions across inventory, promotions, and product line optimization.

4.02 Insight 2: Product Profitability by Item Group Code: -



Objective: This analysis aims to identify the top 10 most profitable product groups in 2012 and 2013. By understanding which product groups yield the highest average profit, management can make informed decisions about resource allocation, marketing focus, and inventory prioritization.

Method Used:

1. **Profit Calculation**: We calculated the Profit from unit price, unit cost and quantity.
2. **Aggregation by Item Group**: The calculated profits were aggregated by item_group_code to find the average profit for each product group.

3. **Top 10 Selection**: From the aggregated data, the top 10 product groups with the highest average profit were selected for each year.

4. Visualization: The top 10 product groups were visualized in a bar chart for each year, allowing for a clear comparison of the most profitable product groups across 2012 and 2013.

Insights:

- **Top Profitable Product Groups**: In both years, FLRSTK emerges as the most profitable product group, indicating a consistently high demand or high margin for items within this group.
- **Shifts in Profitability**: Some product groups, such as A45, appear in the top 10 for both years, indicating stable profitability. However, there are changes in other product groups, such as 308 and A45, which suggests fluctuating profitability across years.
- **Average Profit Range**: The average profit values vary significantly across product groups, with the highest group (FLRSTK) achieving over 5,000 AUD in both years, while other groups achieve lower yet still substantial profits.

Value for Management:

- **Inventory and Resource Allocation**: By focusing on the most profitable product groups, management can prioritize these items for inventory replenishment, ensuring stock availability to meet demand.
- **Targeted Marketing**: High-profit groups can be given special attention in marketing campaigns to maximize revenue.
- **Product Development and Strategy**: Understanding which product groups consistently deliver high profits allows management to invest in expanding or improving these categories. For example, if FLRSTK consistently yields high returns, the business might explore similar or complementary products within this group.
- **Pricing and Cost Control**: Identifying product groups with fluctuating profitability can guide cost control and pricing strategy adjustments to improve margins.

This analysis provides a data-driven foundation for strategic decision-making, enabling management to focus on the most lucrative areas of the product portfolio.

4.03 Insight 3: Sales Trends Over Time: -

Objective: The objective of this analysis is to compare the monthly sales trends for 2012 and 2013. By analyzing sales patterns over time, we aim to identify seasonal peaks, growth patterns, or potential drops in sales across the two years.

Method Used:

1. **Data Aggregation**: Monthly sales values were aggregated by summing the value_sales for each month in both 2012 and 2013.
2. **Visualization**: A line plot was created to display monthly sales trends for both years, with months on the x-axis and total sales value on the y-axis.
3. **Comparison**: Each year's sales trend is represented by a separate line, highlighting seasonal fluctuations and differences between the two years.

Insights:

- **Seasonal Peaks**: Both years show a sales peak around May and June, indicating a potential seasonal trend with high demand mid-year.
- **Comparison Across Years**: Sales in 2012 are consistently higher than in 2013, especially in the first half of the year. This difference could be due to changes in market conditions, demand, or business strategies.
- **Late-Year Drop**: Both years exhibit a significant drop in December sales, suggesting a typical end-of-year slowdown.

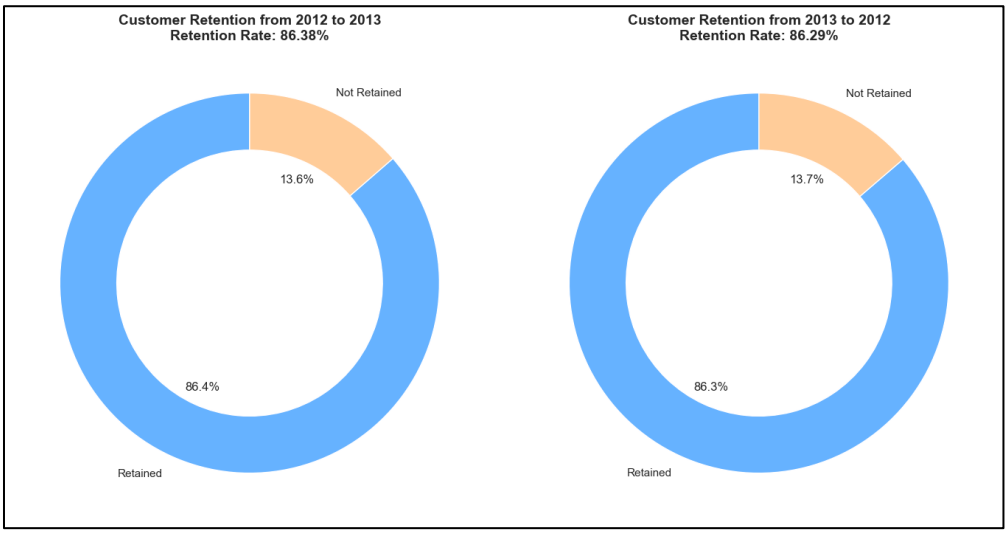
Value for Management:

- **Inventory Planning**: Knowing that sales peak mid-year, management can ensure sufficient inventory during high-demand months to prevent stockouts and meet customer needs.

- **Staffing and Resources**: Seasonal trends enable management to allocate additional resources and staffing during peak periods for smooth operations.
- **Sales Strategy**: The decline in 2013 compared to 2012 may prompt an analysis of contributing factors. Management can use this insight to adjust marketing and sales strategies to boost future sales.
- **Budgeting and Forecasting**: Understanding monthly sales trends supports more accurate budgeting and forecasting, helping management make informed financial and operational decisions.

This analysis provides actionable insights for inventory management, resource allocation, and sales strategies, aligning business practices with customer demand patterns to enhance overall performance.

4.04 Insight 4: Customer Retention Rate: -



Objective: Determine the customer retention rate between 2012 and 2013, which indicates how many customers from 2012 made purchases again in 2013, and vice versa. Understanding customer retention is crucial for evaluating customer loyalty and long-term revenue potential.

Method Used:

1. **Identify Unique Customers**: First, we identified unique customers in 2012 and 2013.
2. **Calculate Retention**: We then calculated the retention rate by finding the intersection of customers in both years, i.e., customers who made purchases in both 2012 and 2013.
3. **Visualization**: The results are visualized using a bar chart, showing the number of retained and not retained customers for each year-to-year retention calculation.

Insights:

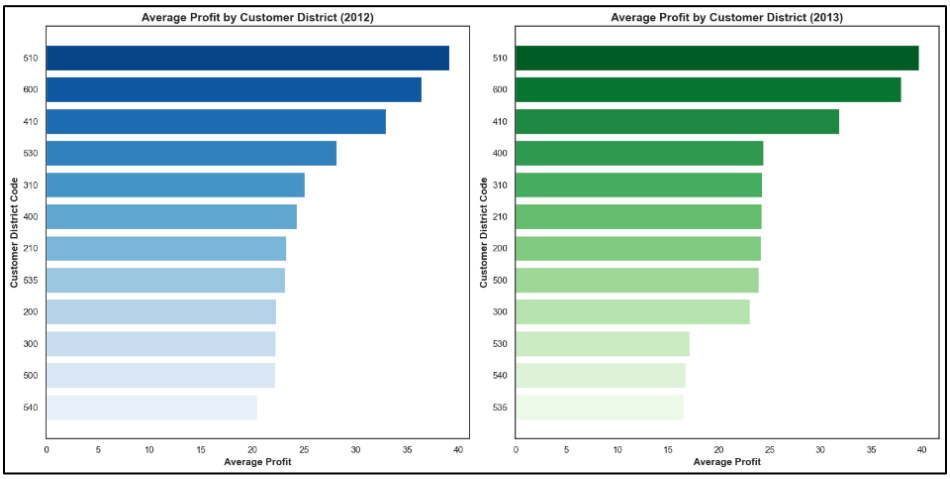
- **Retention from 2012 to 2013**: About 86.38% of customers from 2012 returned to make purchases in 2013, indicating strong customer loyalty.
- **Retention from 2013 to 2012**: Similarly, 86.29% of customers from 2013 were also present in 2012, reinforcing that a large majority of customers are consistently returning year over year.
- **Customer Churn**: The relatively small percentage of customers who did not return highlights the low churn rate, which is beneficial for long-term revenue stability.

Value to Management:

1. **Customer Loyalty Analysis**: High retention rates indicate strong customer loyalty, suggesting that the company's products and services are satisfactory to its customer base.
2. **Revenue Stability**: With a high retention rate, management can expect stable revenue from a core group of loyal customers, making financial forecasting more reliable.
3. **Customer Engagement Strategies**: The small percentage of customers who did not return could represent an opportunity. Management can develop strategies, such as targeted marketing or loyalty programs, to re-engage these customers and further reduce churn.

This retention analysis supports strategic decision-making for customer relationship management and highlights the importance of maintaining high retention rates for long-term business success.

4.05 Insight 5: Profitability by Customer District: -



Objective: The objective of this analysis is to identify customer districts with the highest average profit for 2012 and 2013. By examining district-level performance, we aim to uncover regional profitability insights and determine which locations contribute the most to the company's bottom line.

Method Used:

1. **Profit Calculation**: For each transaction, profit was calculated as the difference between unit_price_in_aud and unit_cost_in_aud, multiplied by the quantity sold.
2. **Aggregation by District**: Profits were aggregated by customer_district_code to compute the average profit per district for both 2012 and 2013.
3. **Top 10 Districts**: The top 10 customer districts with the highest average profit were selected for each year.
4. **Visualization**: A bar chart was created for each year to display the top 10 customer districts by average profit, enabling a direct comparison between 2012 and 2013.

Insights:

- **Consistent High-Profit Districts**: Districts like Perth (600) and Darwin (510) consistently rank high in average profit across both years, highlighting strong performance.
- **Regional Differences**: Major Australian cities, including Sydney (200) and Brisbane (400), are among the top 10 but show varying profit levels between 2012 and 2013.
- **Profit Variations in New Zealand**: New Zealand districts, such as South Island - NZ (530) and Central Region - NZ (535), rank in the top 10 in 2012 but have different dynamics in 2013, indicating shifting performance.

Value for Management:

- **Resource Allocation**: High-profit districts like Perth and Darwin could benefit from focused resource allocation to enhance revenue generation.
- **Targeted Marketing Campaigns**: Regional profitability insights support tailored marketing, promoting products in high-demand areas like Sydney and Brisbane.
- **Evaluating New Zealand Performance**: The variable profitability in New Zealand can inform decisions to focus on profitable areas like South Island - NZ while exploring growth opportunities elsewhere.
- **Strategic Decision-Making**: Identifying high-profit districts aids in optimizing distribution, regional pricing, and customer engagement strategies.

This analysis provides actionable insights for management to guide regional focus, allocate resources, and implement strategies to maximize profitability in Australia and New Zealand.

5.0 Section 3: Test Sub Sample Differences

5.01 Question 1: Is There a Significant Difference in Average Profit Between High-Performing and Low-Performing Customer Districts?

Hypothesis:

- **Null Hypothesis (H0)**: There is no significant difference in average profit between high-performing and low-performing customer districts.
- **Alternative Hypothesis (H1)**: There is a significant difference in average profit between high-performing and low-performing customer districts.

Explanation of Test: This test evaluates whether high-performing districts (top 5 by average profit) generate significantly different profits from low-performing districts (bottom 5 by average profit). By analyzing profit differences, management can pinpoint regions contributing most to profitability and develop strategies for lower-performing areas.

Method:

1. **Data Segmentation**: Customer districts were divided into "high-performing" (top 5) and "low-performing" (bottom 5) categories based on average profit.
2. **Two-Sample T-Test**: An independent two-sample t-test was conducted on profit values for high- and low-performing districts for each year (2012, 2013) and the combined dataset.

Test Results:

- **2012**
  - High-Performing Group Size: 150,921 and Low-Performing Group Size: 538,722
  - T-Statistic: 68.24, P-Value: 0.0
- **2013**
  - High-Performing Group Size: 285,102 and Low-Performing Group Size: 318,197
  - T-Statistic: 54.49, P-Value: 0.0
- **Combined (2012 & 2013)**
  - High-Performing Group Size: 584,399 and Low-Performing Group Size: 658,564
  - T-Statistic: 75.27, P-Value: 0.0

With p-values of 0.0 across all tests, we reject the null hypothesis, confirming a statistically significant profit difference between high- and low-performing districts.

Insights for Management:

- **Resource Allocation**: Prioritize marketing, resources, and sales in high-performing districts to maximize returns.
- **Targeted Improvements**: Address challenges in low-performing districts to boost profitability.
- **Strategic Planning**: Profit distribution insights enable data-driven decisions for expanding, investing, or restructuring operations in specific regions.

This analysis provides actionable insights for optimizing district-specific strategies and driving overall profitability.

5.02 Question 2: Is There a Difference in Sales Volume Between High-Priced and Low-Priced Products?

Hypothesis:

- **Null Hypothesis (H0)**: There is no significant difference in the average sales volume between high-priced and low-priced products.
- **Alternative Hypothesis (H1)**: There is a significant difference in the average sales volume between high-priced and low-priced products.

Explanation of Test: This test evaluates whether a product's price (high vs. low) impacts its average sales volume. By examining the relationship between price and volume, we can gain insights into whether higher-priced products experience reduced sales volume or if volume remains unaffected by price. This understanding can help in setting optimal pricing strategies.

Method:

1. **Data Segmentation**: Products were divided into "high-priced" and "low-priced" categories based on the median unit price within each year (2012 and 2013).
2. **Two-Sample T-Test**: An independent two-sample t-test was conducted on log_value_quantity (sales volume) for high- and low-priced products in each year and the combined dataset.

Test Results:

- **2012**
    - High-Priced Avg Sales Volume: 1.67 and Low-Priced Avg Sales Volume: 2.69
    - T-Statistic: -526.50, P-Value: 0.00000
- **2013**
    - High-Priced Avg Sales Volume: 1.67 and Low-Priced Avg Sales Volume: 2.69
    - T-Statistic: -511.82, P-Value: 0.00000
- **Combined (2012 & 2013)**
    - High-Priced Avg Sales Volume: 1.67 and Low-Priced Avg Sales Volume: 2.70
    - T-Statistic: -737.38, P-Value: 0.00000

With p-values well below 0.05, we reject the null hypothesis, concluding a significant difference in sales volume between high- and low-priced products.

Value for Management:

- **Pricing Strategy**: This relationship indicates that price increases may reduce sales volume, informing competitive pricing strategies.
- **Product Positioning**: Prioritize lower-priced, higher-volume products in inventory planning.
- **Revenue Optimization**: Consider price adjustments or promotions on high-priced products to boost volume without heavily impacting profitability.

This analysis offers insights to balance pricing, volume, and inventory, supporting revenue growth objectives.

6.0 Section 4: Inference

6.01 Question 1: What are the primary factors that drive sales revenue?

Objective: The objective of this analysis is to identify the main factors driving sales revenue by examining the relationship between unit_price_in_aud, log_value_cost_in_aud, and log_value_quantity on log_value_sales_in_aud. Understanding these drivers can help management optimize pricing, cost control, and inventory forecasting.

Method Used: A multiple linear regression analysis was conducted for each year (2012 and 2013) and the combined dataset (2012 & 2013). Using log-transformed values for sales revenue, cost, and quantity controls for skewness, enhancing the model's reliability in estimating each factor's contribution to overall sales revenue.

Steps:

1. **Dependent Variable**: log_value_sales_in_aud (log-transformed sales revenue).
2. **Independent Variables**:
    - unit_price_in_aud: selling price per unit.
    - log_value_cost_in_aud: log-transformed cost per sale.
    - log_value_quantity: log-transformed quantity sold.
3. **Model Fit**: An Ordinary Least Squares (OLS) regression was applied to the data.

Results:

2012 Results:

- **R-squared**: 0.945, explaining 94.5% of the sales revenue variability.
- **Significant Variables**: All predictors (unit price, cost, and quantity) are highly significant ($p < 0.05$).
- **Coefficients**: unit_price_in_aud: 0.001, log_value_cost_in_aud: 0.8842 and log_value_quantity: 0.0487

2013 Results:

- **R-squared**: 0.938, explaining 93.8% of revenue variance.
- **Significant Variables**: All predictors remain significant.
- **Coefficients**: unit_price_in_aud: 0.001, log_value_cost_in_aud: 0.8654 and log_value_quantity: 0.0516

Combined Results (2012 & 2013):

- **R-squared**: 0.941, explaining 94.1% of the variability.
- **Significant Variables**: All predictors remain highly significant.
- **Coefficients**: unit_price_in_aud: 0.001, log_value_cost_in_aud: 0.8748 and log_value_quantity: 0.0503

Robustness Evidence:

1. **Residual Analysis**: The residuals' mean, skewness, and kurtosis were analyzed. Despite minor skewness and kurtosis, high R-squared values indicate the model's effectiveness.
2. **Variance Inflation Factor (VIF)**: VIF values confirmed low multicollinearity, ensuring model reliability.

Value for Management:

1. **Pricing Strategy**: The unit price's modest positive coefficient suggests that price increases can raise revenue without severely impacting sales volume.
2. **Cost Management**: The strong association between cost and revenue highlights the importance of cost control to maximize profitability.
3. **Inventory and Sales Forecasting**: The positive link between quantity and revenue suggests that higher volumes drive revenue, supporting demand forecasting and inventory prioritization.

These findings offer actionable insights for management in pricing, cost control, and inventory strategy to optimize revenue.

6.02 Question 2: What factors influence the variability in unit price across different transactions?

Objective: This analysis aims to understand the factors influencing unit price across transactions by examining relationships between log_value_quantity, log_value_sales_in_aud, and profit on unit_price_in_aud. Insights gained can help optimize pricing strategies and maximize revenue.

Method: A multiple regression model was used to assess how quantity sold, sales revenue, and gross margin relate to unit price. The regression was performed for each year (2012 and 2013) and for the combined dataset (2012 & 2013).

Steps:

1. **Dependent Variable**: unit_price_in_aud.
2. **Independent Variables**: log_value_quantity, log_value_sales_in_aud, profit.
3. **Model Fit**: Ordinary Least Squares (OLS) regression assessed predictor impacts on unit price.

Results:

2012:

- **R-squared**: 0.774 (77.4% variability explained).
- **Significant Variables**: All predictors are significant ($p < 0.05$).
- **Coefficients**:
    - log_value_quantity: -10.9777 (suggesting bulk discount effects).
    - log_value_sales_in_aud: -0.7579 (slight negative relationship).
    - profit: 1.4096 (indicating premium pricing).

2013:

- **R-squared**: 0.762 (76.2% variability explained).
- **Coefficients**:
    - log_value_quantity: -11.5157 (bulk discount effect).
    - log_value_sales_in_aud: -0.2228 (weaker negative effect than 2012).
    - profit: 1.4263 (positive relationship with unit price).

Combined (2012 & 2013):

- **R-squared**: 0.768 (76.8% variability explained).
- **Coefficients**:
    - log_value_quantity: -11.2405.
    - log_value_sales_in_aud: -0.4945.
    - profit: 1.4176.

Robustness Evidence:

1. **Residual Analysis**: Residuals analyzed for normality and homoscedasticity.
2. **Variance Inflation Factor (VIF)**: Low VIF values confirm no multicollinearity.

Value for Management:

1. **Pricing Strategy**: Negative log_value_quantity coefficients suggest volume discounts could be optimized.

2. **Revenue Optimization**: Positive profit coefficients highlight potential for premium pricing.
3. **Sales Volume Influence**: Higher volumes associate with lower prices, supporting bulk sales strategies.

These insights inform pricing and revenue decisions aligned with demand patterns and profitability goals.

7.0 Section 5: Prediction Model

Objective: This analysis aims to build a predictive model to accurately forecast sales prices, using data from 2012 and 2013 to predict 2014 sales values. After evaluating various models, the Random Forest Regressor showed superior predictive capabilities and was selected as the final model.

Steps Taken in Developing the Prediction Model

1. **Feature Selection and Preparation**: Relevant features were chosen based on their potential impact on sales prices, including **Unit Price in AUD, Value Price Adjustment, Customer District Code, Fiscal Month, Item Group Code, Log-transformed Quantity and Cost values, Warehouse Code, Item Type and Class Code and Salesperson and Order Type Codes**

These features encompass product attributes, customer location, and transaction details essential for accurate forecasting.

2. **Data Cleaning and Encoding**:
    - **Encoding Categorical Variables**: Label Encoding was applied to make categorical data suitable for the model.
    - **Handling Missing or Infinite Values**: Replaced or removed incomplete values to ensure a clean, reliable dataset.
3. **Model Selection and Training**: Multiple models were evaluated, including:
    - Random Forest Regressor
    - Decision Tree Regressor
    - Linear Regression
    - K-Nearest Neighbors (KNN) Regressor

Each model was assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). The Random Forest Regressor achieved the highest accuracy and lowest error rates, making it the optimal choice.

4. **Sampling and Full Training on 2012-2013 Data**: A 50% sample was used initially to reduce processing time. Once the Random Forest Regressor was confirmed as the best model, it was trained on the entire 50% sample of 2012-2013 data to maximize learning.
5. **Prediction for 2014 Sales**:
    - **Data Preparation**: Applied the same encoding and feature engineering as 2012-2013 data.
    - **Generate Predictions**: Used the trained model to forecast 2014 sales.
    - **Interpretation**: Analyzed results to interpret trends and assess business strategies.

Conclusion: The Random Forest Regressor provides an accurate sales forecasting model, enabling management to make data-driven decisions for 2014 sales strategies. Utilizing a 50% sample effectively balanced dataset management and model performance.

8.0 Section 6: Higher Likelihood of Losing Customers

Objective: The objective of this churn analysis is to identify factors associated with a higher likelihood of customer churn. By understanding these factors, management can take proactive steps to retain at-risk customers and develop targeted marketing strategies to reduce churn.

Churn Definition: For this analysis, churn is defined using a **Time-Based Churn** criterion. A customer is classified as "churned" if they have not purchased within a 60-day threshold. This timeframe reflects inactivity beyond typical customer behavior, assuming that most customers generally make purchases within two months.

Steps in Churn Analysis:

1. **Feature Selection**: We selected various customer-related features to capture different aspects of customer behavior and characteristics:
    - **Purchase Frequency**: Number of purchases made by each customer.
    - **Average Order Value**: Average sales value per order.
    - **Average Quantity Purchased**: Average quantity purchased per order.
    - **Days Since Last Purchase**: Days since each customer's last recorded purchase.
    - **Customer District Code**: Geographic information representing customer location.
    - **Order Type Code**: Type of orders placed by the customer.
    - **Salesperson Code**: Salesperson associated with the customer’s transactions.
    - **Warehouse Code**: Warehouse from which purchases were fulfilled.
    - **Fiscal Month**: Month of the fiscal year in which purchases were made.
    - **Item Group Code**: Product group of items purchased.

2. **Churn Label Creation**: Using the Time-Based Churn definition, we created a binary churn label based on the days_since_last_purchase feature. Customers with no purchase in the last 60 days were labeled as 1 (churned), while customers with more recent purchases were labeled as 0 (active).
3. **Model Training and Evaluation**: We used a Random Forest Classifier to assess the relationship between selected features and customer churn. The model was trained on a balanced set of active and churned customers to ensure effective differentiation between the two groups.
4. **Feature Importance Analysis**: After training the model, we examined feature importance to understand which factors most contribute to predicting churn, helping to identify customer characteristics associated with churn likelihood.

Results and Interpretation:

- **Confusion Matrix and Classification Report**: The model showed perfect classification for active customers but failed to identify any churned customers due to a lack of churned examples. The churn target distribution only included active customers (0), indicating that none of the customers met the 60-day churn threshold. The classification report confirms this, suggesting the chosen 60-day churn threshold may not capture meaningful churn behavior.
- **Feature Importances**: All features showed zero importance in predicting churn, consistent with the lack of churned customers. This indicates insufficient variability in the target variable and suggests that the 60-day threshold may be too short.

Conclusion and Recommendations:

1. **Refine Churn Definition**: The analysis suggests revisiting the churn definition. A longer threshold, such as 180 days or one year, may better capture infrequent purchase behavior.
2. **Consider Alternative Churn Criteria**: Given the 60-day period yielded no churned customers, alternative churn criteria, such as a Monetary-Based Churn definition (e.g., declining spending), could capture more meaningful attrition patterns.
3. **Management Value**: Defining churn accurately is crucial for understanding at-risk customer behavior. Properly identifying these customers allows management to intervene strategically, potentially reducing churn rates. By identifying high-risk factors, the company can develop targeted retention strategies, such as personalized outreach, promotions, or adjusting service models to better meet customer needs.

In summary, this analysis provides a foundation for churn analysis but suggests that refining the time-based threshold could improve model effectiveness in identifying at-risk customers. Future steps should consider testing various churn definitions for a more accurate assessment.

9.0 Conclusion

This comprehensive analysis of LuminaTech Lighting's data provides valuable insights across key areas of sales, profitability, customer retention, and predictive modeling. By examining the data from 2012 and 2013, we implemented rigorous cleaning, feature selection, and data preparation processes to ensure the reliability and accuracy of our results.

In **sales analysis**, we identified top-performing products and regions, allowing management to optimize inventory, focus marketing strategies, and enhance product offerings. The **profitability insights** by customer district and item group highlighted the most profitable segments, guiding strategic resource allocation and pricing decisions.

Our **predictive modeling efforts** for 2014 sales successfully leveraged the Random Forest Regressor, demonstrating strong predictive accuracy and enabling management to make data-driven decisions for future sales forecasts. Additionally, our churn analysis emphasized the need to refine the definition of churn to more accurately capture at-risk customers, offering actionable recommendations to improve customer retention strategies.

This project equips LuminaTech Lighting with data-driven tools for strategic decision-making, ultimately enhancing operational efficiency, profitability, and customer loyalty. Moving forward, refining models and expanding definitions for churn analysis will further improve the accuracy and impact of insights for sustained growth and competitive advantage.