Group Assignment

BUSA8000

Techniques in Business Analytics

Group member

1. Sawitree Nilsanga – 47940352
2. Warinthon Thongkaew – 48113093
3. Tanyarat Boonprasit – 48155489

Submission Date of November 6, 2024, 23:55

Acknowledgement Statement by students

I acknowledge that I have only used GAITs (e.g., ChatGPT) in drafting and proofreading this assignment, which is permitted in the assignment instructions.

(6,953 words)

Introduction

The ability to analyse and interpret vast amounts of data is crucial for making informed decisions. As data analysts for LuminaTech Lighting, this project aims to provide actionable insights into sales performance, customer demographics, and inventory management. Using a comprehensive dataset encompassing variables such as accounting dates, customer district codes, product details, and financial transactions, our team applied a range of data analytics techniques to uncover trends, forecast sales, and identify factors influencing customer behaviour. This report presents a structured analysis covering data cleaning, exploratory insights, hypothesis testing, multiple regression, predictive modelling, and customer churn analysis. Each section is designed to offer valuable information for management, facilitating strategic planning, resource optimisation, and targeted interventions to drive business growth and improve operational efficiency. Through this project, we aim to support LuminaTech Lighting in leveraging data to enhance decision-making and maintain a competitive edge in the market.

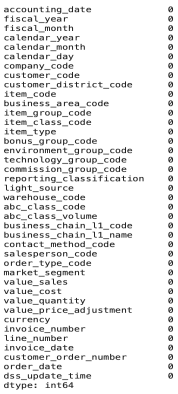
Section 1: Clean the dataset

As a Data Analyst at LuminaTech Lighting, Dataset must thoroughly clean and prepare the dataset—including accounting dates, customer codes, and product details—to ensure accurate analysis. This process lays the groundwork for reliable visualisations and statistical analyses. The steps to clean data are below.

1.1 Handling Missing Values:

To ensure data completeness and accuracy, first identify columns with missing values and choose to either drop or impute values based on their impact on the analysis. This approach keeps the dataset robust and ready for meaningful insights. In this case, the column 'item_source_class' has 1,988,382 missing values, which is the entire dataset length, indicating that it is unused. As a result, we decided to remove this column.

```
accounting_date                0
fiscal_year                    0
fiscal_month                   0
calendar_year                  0
calendar_month                 0
calendar_day                   0
company_code                   0
customer_code                  0
customer_district_code         0
item_code                      0
business_area_code             0
item_group_code                0
item_class_code                0
item_type                      0
bonus_group_code               0
environment_group_code         0
technology_group_code          0
commission_group_code          0
reporting_classification       0
light_source                   0
warehouse_code                 0
abc_class_code                 0
abc_class_volume               0
business_chain_l1_code         0
business_chain_l1_name         0
contact_method_code            0
salesperson_code               0
order_type_code                0
market_segment                 0
value_sales                    0
value_cost                     0
value_quantity                 0
value_price_adjustment         0
currency                       0
item_source_class        1988382
invoice_number                 0
line_number                    0
invoice_date                   0
customer_order_number          0
order_date                     0
dss_update_time                0
dtype: int64
```



```
accounting_date                0
fiscal_year                    0
fiscal_month                   0
calendar_year                  0
calendar_month                 0
calendar_day                   0
company_code                   0
customer_code                  0
customer_district_code         0
item_code                      0
business_area_code             0
item_group_code                0
item_class_code                0
item_type                      0
bonus_group_code               0
environment_group_code         0
technology_group_code          0
commission_group_code          0
reporting_classification       0
light_source                   0
warehouse_code                 0
abc_class_code                 0
abc_class_volume               0
business_chain_l1_code         0
business_chain_l1_name         0
contact_method_code            0
salesperson_code               0
order_type_code                0
market_segment                 0
value_sales                    0
value_cost                     0
value_quantity                 0
value_price_adjustment         0
currency                       0
invoice_number                 0
line_number                    0
invoice_date                   0
customer_order_number          0
order_date                     0
dss_update_time                0
dtype: int64
```

1.2 Correcting Inconsistent Data:

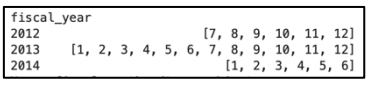
Addressed inconsistencies by standardising formats, particularly for dates and categorical fields, to ensure data consistency and accuracy.

- **accounting_date:** This column was converted to a standard DateTime format, allowing for consistent comparisons across fiscal_year and calendar_year. then created separate columns for accounting_year, accounting_month, and accounting_day and confirmed alignment with calendar dates.

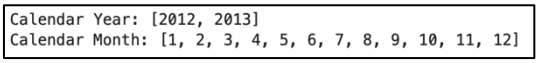
	accounting_date
0	2012-05-09
1	2012-02-16
2	2012-05-09
3	2012-05-18
4	2012-01-09

accounting_year	accounting_month	accounting_day
2012	5	9
2012	2	16
2012	5	9
2012	5	18
2012	1	9

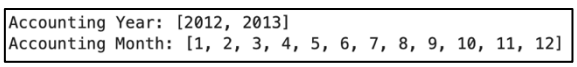
- **accounting, fiscal, and calendar:** Checked the unique values in each column to confirm consistency between the calendar and accounting dates, spanning 2012 and 2013 with a full set of months and days. However, the fiscal data follows a unique pattern aligned with the Australian financial calendar: fiscal year 2012 covers July to December 2012 and January to June 2013, while fiscal year 2013 includes July to December 2013 and January to June 2014. This structure aligns with the Australian fiscal year, so no further adjustments were necessary to maintain consistency across these date-related fields.



```
fiscal_year
2012                  [7, 8, 9, 10, 11, 12]
2013    [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
2014                     [1, 2, 3, 4, 5, 6]
```



```
Calendar Year: [2012, 2013]
Calendar Month: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
```

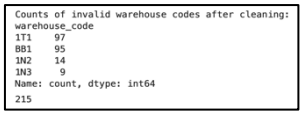


```
Accounting Year: [2012, 2013]
Accounting Month: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
```

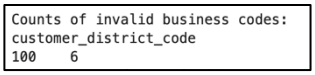
- **order_type_code:** the unique values in the order_type_code column to ensure they align with the expected categories in our

documentation. After inspection, 243 rows with "PME" as the order_type_code, suggesting an inconsistency or possible error. So, we removed these 243 rows from the dataset.

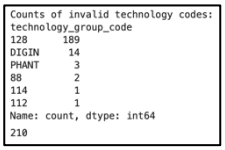
- **abc_class_volume:** The unique feature of columns is consistency and a uniform format.
- **abc_class_code:** The unique feature of columns is consistency and a uniform format.
- **warehouse_code:** The unique value in this column shows that some codes, like 'V0', 'N0', and 'Q0', appeared as invalid entries due to trailing spaces. So, we mapped these entries to the correct format by removing spaces. Then, rechecked the dataset and found that four codes—1T1, BB1, 1N2, and 1N3—were still inconsistent with the reference list, totalling 215 rows. Hence, we decided to remove them.



- **environment_group_code:** The unique value in this column shows that the column contained entries with extra spaces, which affected consistency. So, we created a mapping to replace the values with their correct, space-free format. After applying the replacement, we confirmed that all codes were correctly formatted and matched the reference list.
- **business_area_code:** The unique value in this column shows extra spaces, so we addressed inconsistencies by removing extra spaces found in some codes. This standardisation ensures that all codes are consistent and align with our reference list. After applying the replacement, we confirmed that all codes were correctly formatted and matched the reference list.
- **customer_district_code:** Examined this column and found six rows with an invalid code, 100, which did not match our reference list. To ensure data consistency, we removed these rows.



- **technology_group_code:** Reviewed this column for consistency against our reference list. We found several valid codes with extra spaces. After these adjustments, there are some codes that are invalid, totalling 210 rows. So, we decided to remove it.



- **currency:** There are some issues the code “AUS” was used instead of "AUD”, and a few entries were blank. Standardised "AUS" to "AUD" and removed rows with blank entries. Afterwards, we converted all currency values to AUD using average yearly exchange rates.
- **invoice_number:** Focused on ensuring the integrity of this column, which should contain unique, non-zero values. It found that 117 rows were zero, which likely represented missing, erroneous, or placeholder data, and we removed these rows for consistency.

Additionally, we checked for any blank spaces and removed leading and trailing whitespace from all string columns.

- **Item_code:** Found consulting services listed in the item_code column during data exploration. We decided to remove these entries to focus on actionable insights related to product sales, as consulting services typically follow unique revenue models, cost structures, and customer behaviours that differ from those of physical products.

1.3 Removing Duplicates

Identified and removed duplicate records from the dataset to ensure data accuracy. We expected that each row should be unique, representing a specific order item from a customer. We found 8,204 duplicate rows, which could skew analysis results. So we decided to remove it.

Duplicate rows based on all columns:

	accounting_date	fiscal_year	fiscal_month	calendar_year	cale
342	2012-05-31	2012	11	2012	
546	2012-03-09	2012	9	2012	
852	2012-02-07	2012	8	2012	
940	2012-01-05	2012	7	2012	
1257	2012-05-08	2012	11	2012	
...	...	...	...	...	
1985196	2013-09-19	2014	3	2013	
1985511	2013-08-16	2014	2	2013	
1986320	2013-10-18	2014	4	2013	
1987850	2013-10-18	2014	4	2013	
1988195	2013-11-01	2014	5	2013	

8204 rows × 43 columns

1.4 Data type conversion

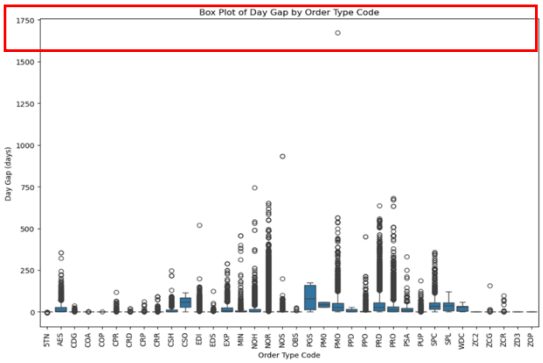
Identified and removed duplicate records from the dataset to ensure data accuracy. We expected each row to be unique, representing a specific order item from a customer. We found 8,204 duplicate rows, which could skew analysis results. So we decided to remove it.

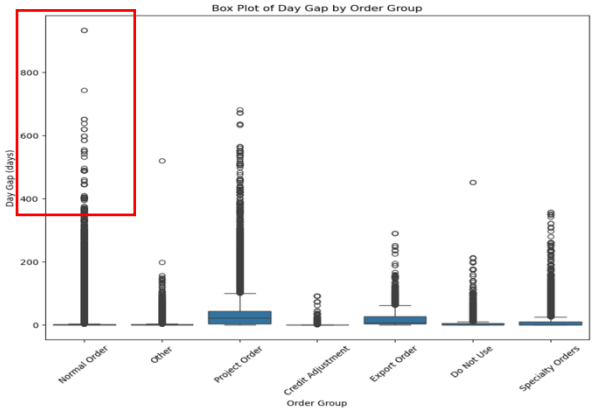
1.5 Handling Outliers

Outliers can distort analysis, especially in statistical modelling. We will identify them using visualisation techniques like boxplots and histograms, then decide whether to remove, transform, or keep these outliers based on the analysis context and objectives. The following steps outline this process.

- **outlier management in core metric:** Focus on key value columns ‘value_sales’, ‘value_cost’, and ‘value_quantity’ as these impact financial performance and operational decisions. Since ‘value_price_adjustment’ is mostly binary (0 or 1), indicating it's a categorical variable, it doesn't require outlier handling and can be used as-is for analysis and modelling. The boxplots show extreme outliers, so we set thresholds of 600,000 for ‘value_sales’ and ‘value_cost’ and 80,000 for ‘value_quantity’. Rows exceeding these limits, representing data entry errors, were identified and reviewed. Specifically, For Item Code GR99, reversed and extreme sales and cost entries were removed as errors. Item Code 200767’s high-value transactions were kept due to alignment with typical sales. Duplicate and negative entries for Item Code NS010015 were removed to ensure accuracy, while Item Code TJ01-0002-00’s high-quantity export transaction was retained as expected for exports.

- **Zero and Placeholder Values:** Managed zero values and outliers across key financial columns. Transactions with zero values in ‘value_sales’, ‘value_cost’, and ‘value_quantity’ were removed as they likely represented cancelled or placeholder entries with no financial impact.
- **Data Anomalies:** Addressed cases where ‘value_sales’ was less than ‘value_cost’ for product categories ‘C’, ‘G’, ‘I’, and ‘F’, removing entries without price adjustments to avoid unintentional losses. Additionally, negative quantities in normal orders (NOR, NOH, and NOS) were filtered out, as these orders should reflect straightforward sales. These steps improve data integrity, supporting accurate financial analysis.
- **Handling Day Gap Outliers:** Calculated the time difference between ‘accounting_date’ and ‘order_date’, identifying a maximum of 1,672 days, which is unrealistic for typical order processing and suggests data entry errors, so we removed it. Then, we addressed negative gaps, where ‘order_date’ was recorded after ‘accounting_date’, also likely due to entry mistakes, and removed these entries. Moreover, we grouped orders by type and set a 365-day limit for 'Normal Orders' (NOR, NOH, NOS) based on typical processing times. Any rows with a day gap over this limit were removed to ensure the data reflects realistic processing times, making the analysis more accurate.





1.6 Log Transformation Review

In the normalization and scaling process, we applied a log transformation using the log(1+x) method to columns with highly skewed distributions, such as ‘value_sales’, ‘value_cost’, and ‘value_quantity’. This approach reduces skewness by compressing high-value data points, resulting in a more symmetric distribution that improves the reliability of statistical analyses and model performance. Additionally, the transformation handles wide ranges effectively, minimizing the disproportionate influence of extreme values. By using log(1+x), we preserved zero values, transforming them to zero, which ensures completeness by retaining

all data points, even those with meaningful zero values. This process enhanced data consistency and interpretability, making the data more suitable for modelling and analysis.

```
Row counts for standard log transformation (data > 0):
value_sales_log: 1849971
value_cost_log: 1865684
value_quantity_log: 1867000

Row counts for log(1 + x) transformation (data >= 0):
value_sales_log_plus1: 1868753
value_cost_log_plus1: 1915339
value_quantity_log_plus1: 1915314
```

1.7 Handling Categorical Variables

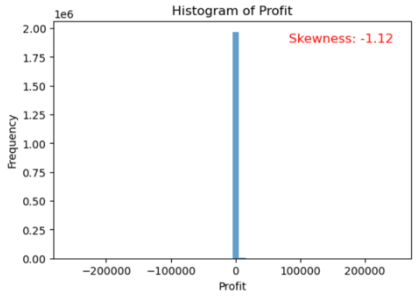
To prepare categorical data for modelling, we applied One-Hot Encoding for non-ordinal columns and Label Encoding for ordinal ones, such as 'abc_class_code', converting categorical values into numerical form for better model compatibility. We first grouped rare categories (those appearing in less than 5% of data points) into an 'Other' category in high-cardinality columns, which reduced dimensionality and minimised noise from underrepresented values. After grouping, we applied one-hot encoding using 'pd.get_dummies', with 'drop_first=True' to avoid multicollinearity by dropping the first category from each encoded variable. This approach streamlined the dataset while retaining key categorical insights, enhancing modelling efficiency and interpretability by consolidating infrequent categories and simplifying the structure of categorical data.

```
Number of unique groups after grouping rare categories with threshold 5%:
  business_area_code: 6 unique groups
  item_group_code: 2 unique groups
  item_class_code: 6 unique groups
  item_type: 5 unique groups
  bonus_group_code: 2 unique groups
  environment_group_code: 5 unique groups
  technology_group_code: 10 unique groups
  commission_group_code: 2 unique groups
  reporting_classification: 2 unique groups
  light_source: 3 unique groups
  warehouse_code: 5 unique groups
  abc_class_code: 8 unique groups
  abc_class_volume: 4 unique groups
  business_chain_l1_code: 7 unique groups
  contact_method_code: 2 unique groups
  order_type_code: 3 unique groups
```

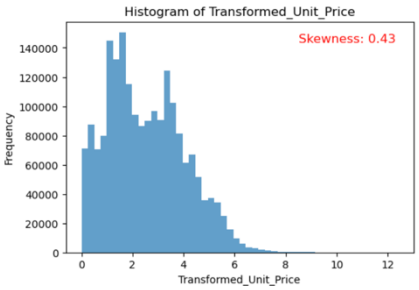
1.8 Feature Engineering

In the feature engineering stage, we created new variables, transformed data, and removed redundant features to enhance the dataset's analytical value and model performance. We generated Profit (difference between 'value_sales' and 'value_cost'), Unit Price (ratio of 'value_sales' to 'value_quantity', with zero substitution for undefined values), and Profit Margin (percentage of Profit relative to 'value_sales'), setting negative Profit Margin values to zero to focus on profitability. To address skewness, we applied a square root transformation to Profit and a log transformation to Unit Price, reducing extreme values and aligning distributions closer to normality. We also dropped redundant date-related columns, avoiding multicollinearity and simplifying the dataset. Histograms and skewness checks confirmed these transformations improved data suitability for modelling, ultimately enhancing interpretability and predictive power.

Before address skewness

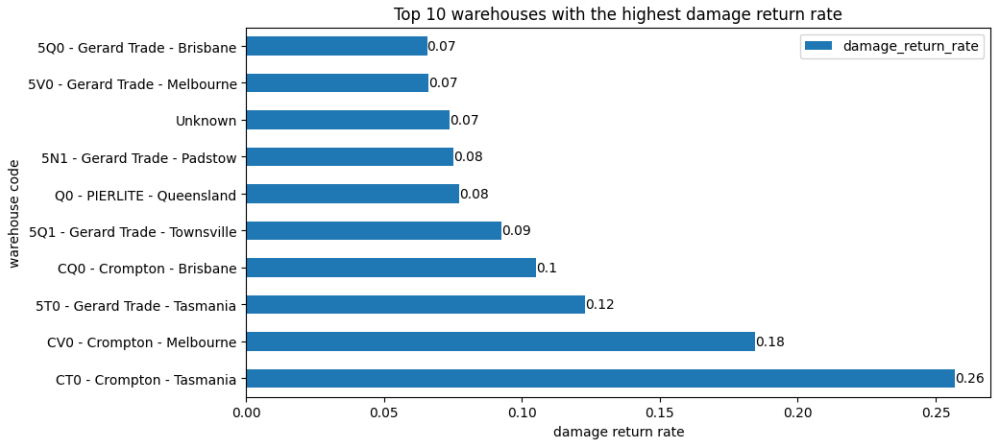


After address skewness



Section 2: Exploratory Insight

2.1 Warehouses with the highest damage and return rate



- **Explanation of the method used**

  To assess warehouse performance, we identified those with the highest damage or return rates. First, we filtered the dataset to include only transactions marked as returns or damaged (order type codes 'CDG' and 'CRR'). Next, we grouped these transactions by ‘warehouse_code’ to count returned or damaged items; we calculated the total orders per warehouse to establish a baseline. Then, we merged these counts to compute the damage return rate for each warehouse, which is defined as the ratio of returned/damaged items to the total number of orders. Lastly, we sorted the results to highlight the top 10 warehouses with the highest return rates and visualised this information using a bar chart.

- **Insight gained**

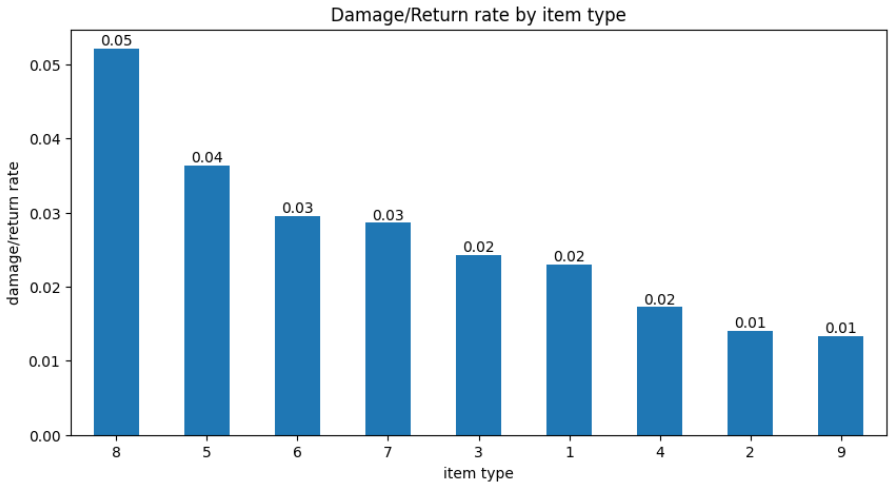
  The analysis revealed significant variability in damage return rates across warehouses. From the bar chart, the highest return rate was observed for warehouse code 'CT0 - Crompton - Tasmania', with a rate of 0.26, followed by 'CV0 - Crompton - Melbourne' at 0.18, and '5T0 - Gerard Trade - Tasmania' at 0.12. This insight indicates that specific warehouses have disproportionately high return rates, which could signal handling, packaging, or quality control issues.

- **Valuable for the management team**

  This analysis is valuable for management as it highlights warehouses with high damage return rates, indicating potential inefficiencies or quality control issues. The insight enables targeted investigations into warehouse processes to identify and address weaknesses. By focusing on the highlighted warehouses, management can better understand and mitigate factors contributing to higher return rates. Reducing these rates can improve customer satisfaction, lower return-related costs, and streamline operations. Additionally, this insight

supports resource allocation for process improvements in targeted locations, enhancing overall operational efficiency.

2.2 Damage/Return rate by item type



- **Explanation of the method used**

  To analyse the damage/return rate across different item types. First, we filtered the dataset for transactions marked as returns or damaged (identified by order type codes 'CDG' and 'CRR'). Then, we grouped these filtered transactions by item_type to calculate the count of returned or damaged items for each type. We also computed the total number of orders for each item type. By merging these datasets, we derived each item type's damage/return rate as the ratio of returned/damaged items to the total order count. Finally, we sorted and visualised the item types by damage/return rate using a bar chart.

- **Insight gained**

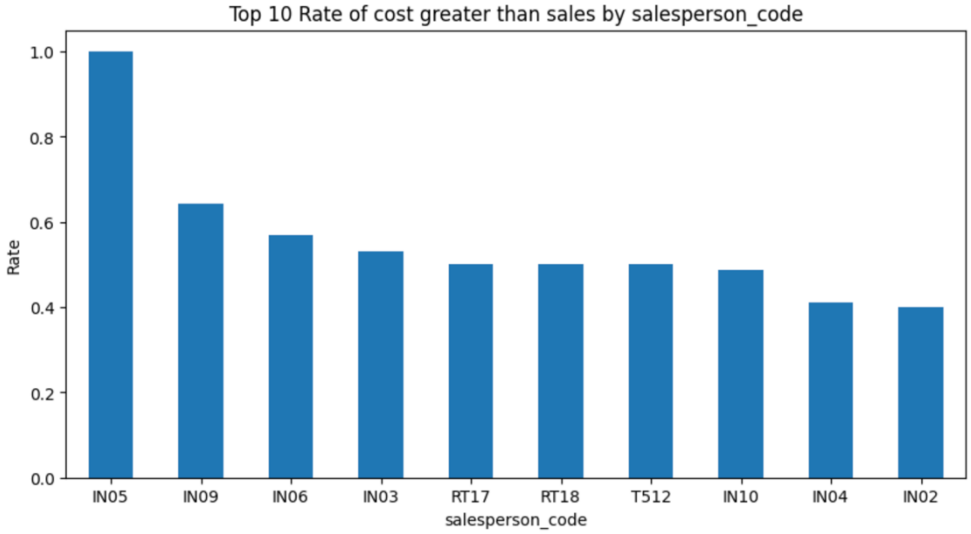
  The analysis reveals that item type ‘8’ has the highest damage/return rate at 5%, followed by item type ‘5’ at 4% and ‘6’ at 3%, respectively. This indicates that specific item types are more prone to returns or damages, which could be due to product fragility, packaging issues, or handling concerns. Meanwhile, item types 2, 4, and 9 have the lowest return rates, indicating they are more stable and align well with customer expectations.

- **Valuable for the management team**

  This insight is valuable for management in identifying and investigating item types with high return rates, suggesting a need for quality improvements. Management can collaborate with suppliers to enhance manufacturing processes, explore alternative materials, or adjust product design, especially for high-return items like item type 8. Inventory strategies may involve lowering stock levels for high-return items and reallocating resources toward reliable types, such as item types 2, 4, and 9, to optimise stock management. Management can

highlight these low-return, reliable items to build customer confidence, while bundling high-return items with low-return ones may mitigate return impacts. Additionally, enhancing product descriptions and providing proactive customer support can help align customer expectations, reducing returns. Overall, these insights enable management to streamline operations, improve customer satisfaction, and boost profitability by addressing return-related challenges through a data-driven approach.

2.3 Unprofitable Transactions by Salesperson



- **Explanation of the method used**

  This analysis aimed to identify salespeople with a high occurrence of transactions where 'value_cost' exceeds 'value_sales'. First, transactions where cost > sales were filtered, and counts were aggregated by 'salesperson_code'. Next, the rate of such transactions for each salesperson was calculated by dividing their unprofitable transaction count by their total transaction count. Finally, the top 10 salespeople by this rate were plotted to visualize a bar chart of those with the highest likelihood of engaging in unprofitable sales.

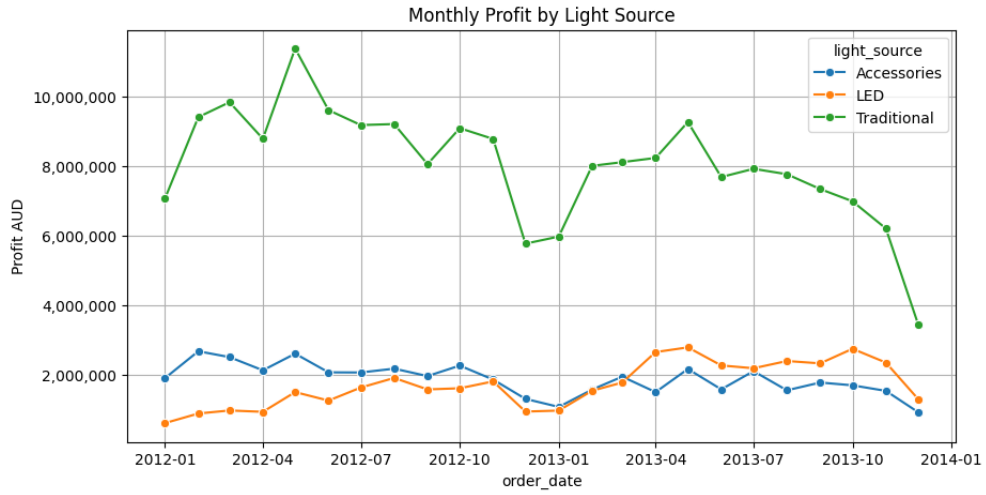
- **Insight gained**

  The chart reveals that salesperson IN05 has an exceptionally high rate of transactions where costs exceed sales, nearing 100%. This suggests a pattern where **IN05** is frequently involved in unprofitable sales, which may be due to factors like pricing inaccuracies, excessive discounts, or issues with specific product lines. Other salespeople, like IN09, IN06, and IN03, also show relatively high rates, although considerably lower than IN05. This trend suggests that certain salespeople are either regularly handling items with pricing or cost issues or might be offering discounts or adjustments that lead to unprofitable sales.

- **Valuable for the management team**

  This insight is valuable for management as it identifies specific salespeople and items associated with unprofitable sales. By understanding the reasons behind these cost > sales transactions—whether due to sales strategies, pricing policies, or product-specific challenges—management can take targeted actions. They might consider training on pricing policies, adjusting product costs, or revising discount strategies for certain items. Addressing these issues can help reduce financial losses and improve the profitability of sales operations. Additionally, management can provide focused support or resources to salespeople like IN05 to enhance decision-making and sales effectiveness. This approach will allow you to assess and mitigate factors contributing to unprofitable sales at both the salesperson and product levels.

2.4 Monthly Sales by light source



- **Explanation of the method used**

  This analysis investigates monthly profit trends across three light source categories: ‘Traditional’, ‘LED’, and ‘Accessories’. The dataset was grouped by month and light source, with monthly profits calculated for each category. The resulting data was visualised using a line chart, allowing for a clear comparison of profit patterns over time from January 2012 to January 2014. This approach highlights trends, seasonal fluctuations, and any shifts in profitability across the different light source types.

- **Insight gained**

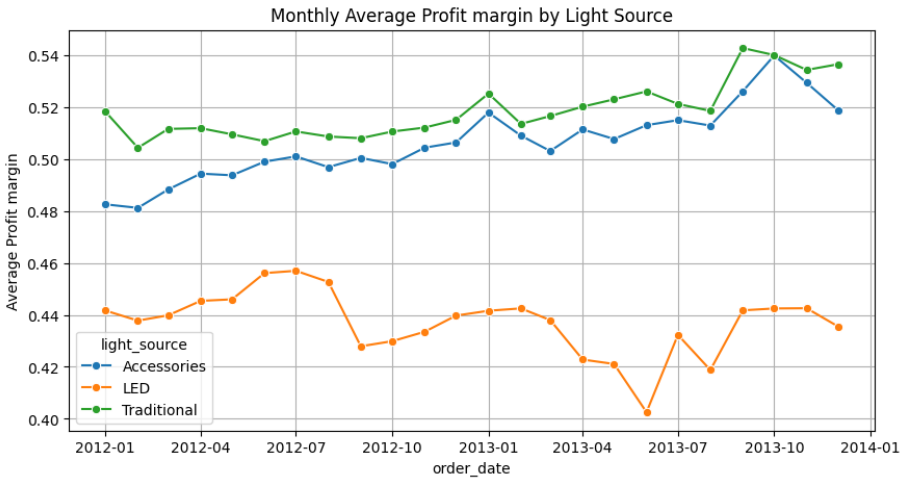
  The analysis reveals that ‘Traditional’ lighting consistently outperforms other categories in profit, suggesting it dominated the market during this period. However, profits for this category exhibit notable volatility, with peaks potentially linked to specific promotions or seasonal demand cycles. In contrast, ‘LED’ lights show a slow but steady profit growth,

indicating gradual consumer adoption, likely due to increasing awareness of energy efficiency. The ‘Accessories’ category maintains a lower and more stable profit level, suggesting either a smaller demand or a niche market for these products.

- **Valuable for the management team**

  These insights are valuable for management to inform inventory and marketing strategies. The dominance of ‘Traditional’ lighting suggests an opportunity to capitalise on this category during peak seasons, while the growth in LED sales underscores a shift toward sustainable options, warranting increased investment in LED products to capture a growing market. The stability of Accessories profits suggests that resources could be optimised by maintaining a steady stock level without aggressive marketing. Understanding these patterns will enable management to allocate resources effectively, plan seasonal promotions, and anticipate shifts toward energy-efficient lighting solutions, enhancing profitability and aligning with evolving consumer preferences.

2.5 Monthly Average Profit margin by light source



- **Explanation of the method used**

  This analysis aimed to examine monthly average profit margins across different light source categories such as ‘Traditional’, ‘LED and ‘Accessories’. We began by calculating profit and profit margin. Profit was determined as the difference between ‘value_sales’ and ‘value_cost’, while profit margin was calculated as Profit divided by ‘value_sales’. Additional filters were applied to ensure data accuracy; we removed records where ‘value_sales’, ‘value_cost’, or profit equalled zero and filtered out entries with negative or zero profit margins. Only orders from 2012 onward in Australian Dollars (AUD) were included. Grouping the dataset by month and light source, we calculated the average profit margin per category over time, visualising the results in a line chart to reveal trends across the period.

- **Insight gained**

  The ‘Traditional’ light source category consistently shows the highest average profit margin, exceeding 50% and trending slightly upward over time, indicating it’s a strong revenue driver. ‘Accessories’ maintain moderate and stable margins, while ‘LED’ profit margins exhibit volatility and a declining trend in mid-2013. This fluctuation in the LED category could signal increasing market competition or pricing adjustments impacting profitability.

- **Valuable for the management team**

  This analysis offers management valuable insights into profit margins across different light source categories, guiding strategic pricing, production, and marketing decisions. The stable, high margins in ‘Traditional’ lighting suggest a strong revenue base, making it ideal for continued investment and promotion. In contrast, the declining margins for ‘LED’ products indicate challenges that may be mitigated by negotiating supplier costs, optimising production, or adjusting pricing. The steady margins in ‘Accessories’ hint at a niche market, suggesting potential for bundling with other products to increase sales. Management can enhance operational efficiency and drive long-term profitability by focusing on high-margin categories, adjusting inventory to align with profitability, and targeting marketing efforts to capitalise on financially advantageous products.

Section 3: Test Sub-Sample Differences

3.1 Is there a significant difference in the log-transformed average value_sales between Top Sellers and Low Sellers categories?

- **Explanation of Tested**

  In this question, we aim to test whether there is a statistically significant difference in the log-transformed average sale values between two distinct product categories, top Sellers and Low Sellers. The sale values metric reflects the sales revenue generated by each product, and we apply a log transformation to this value to normalise the data and reduce the impact of extreme values or outliers. By focusing on Top Sellers (products classified as abc_class_code_A) and Low Sellers (abc_class_code_C), the goal is to determine if one category consistently achieves higher average sales. Understanding this difference can provide valuable insights into which product category is more financially successful, guiding management decisions on product promotion, stocking, and resource allocation.

- **How Being Tested**

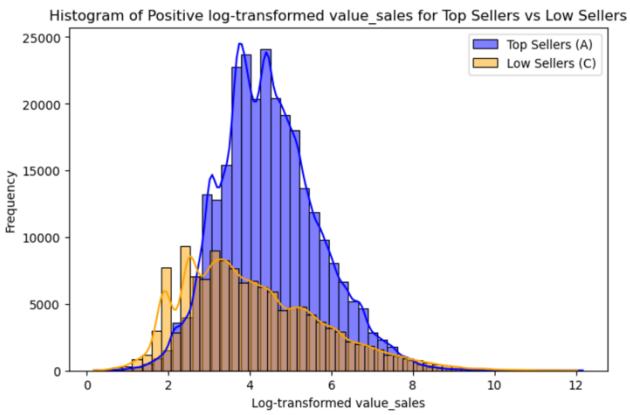
  To test for differences in the average log-transformed sale value between Top Sellers and Low Sellers, we first filter the data to include only these categories (abc_class_code_A for

Top Sellers and abc_class_code_C for Low Sellers). Then, we use Levene's test to determine if the variances in the log-transformed sales values are equal between the two groups. The null hypothesis in this analysis is that there is no difference in the mean log-transformed value_sales between top sellers and low sellers. If the variance is equal, we use a standard independent t-test. The test provides both a t-statistic, indicating the magnitude of the difference, and a p-value, informing us of the likelihood that this difference is due to chance. A p-value below 0.05 leads us to reject the null hypothesis, indicating a statistically significant mean difference.

- **Results Obtained**

The results from the t-test show a t-statistic of -5.99 and a p-value of approximately 2.13e-09. Given that the p-value is significantly below the 0.05 threshold, we conclude that there is a statistically significant difference in the mean log-transformed sale value between Top Sellers and Low Sellers *(reject the null hypothesis)*. This result suggests that, even after log transformation to adjust for scale and outliers, the average sales between these two categories differ meaningfully. The negative t-statistic indicates that, on average, Top Sellers have lower log-transformed value_sales than Low Sellers.

```
T-statistic: -5.987554554500018
P-value: 2.131954785068421e-09
The mean log-transformed value_sales are significantly different between Top Sellers and Low Sellers.
```



The histogram supports these findings, showing that Top Sellers have a more concentrated distribution of log-transformed sales values around the mid-range, while Low Sellers are more spread toward lower values. This visual evidence aligns with the t-test results, reinforcing the conclusion of a significant difference in average sales between the categories.

- **Valuable for Management**

The findings of this analysis provide critical insights that can directly influence management strategies and decision-making with a statistically significant difference in average value_sales between these categories. Top sellers show higher average sales; management can confidently invest more in advertising, optimise shelf space, and increase inventory for these products, knowing that they perform consistently well. Conversely, if Low

sellers have comparable sales, the t-test results can justify targeted promotions or bundling strategies to enhance their market appeal. This statistical confirmation removes the guesswork from strategic decisions, allowing management to use evidence-based insights when developing marketing and product development plans. The t-test further supports dynamic inventory and pricing strategies by highlighting differences in sales consistency, empowering management to experiment with tailored approaches for each category to optimise overall profitability. In summary, the t-test enhances management's ability to allocate resources effectively, refine product mix, and plan for sustainable growth based on reliable sales data.

3.2 Is there a statistically significant difference in the average Transformed_Profit between products with LED lights and those with Traditional lights?

- **Explanation of Tested**

  In this question, we aim to determine whether there is a statistically significant difference in the average Transformed_Profit between products with LED light sources and those with Traditional light sources. Transformed_Profit is used as a measure of profitability, where the transformation adjusts for data skewness or outliers, providing a more normalised view of profit distribution. By comparing these two product categories (LED vs. traditional), we can identify whether one light source consistently yields higher average profitability. This analysis can provide valuable insights into the financial performance of different product types, supporting data-driven decisions on product strategy and resource allocation.

- **How Being Tested**

  To test for differences in average Transformed_Profit between LED and Traditional products. First, we filter the data to include only rows where the light_source is either LED or Traditional, allowing us to focus on the relevant product categories. Next, we apply Levene's test to check for equality of variances. The outcome of this test informs our choice of statistical test. If variances are equal (p-value > 0.05), we use a standard independent t-test. Then, we conduct the t-test to compare the mean Transformed_Profit between the two groups. The null hypothesis is that there is no difference in the average Transformed_Profit between LED and traditional products. If the p-value from the test is below 0.05, we reject the null hypothesis, indicating a statistically significant difference in profitability.

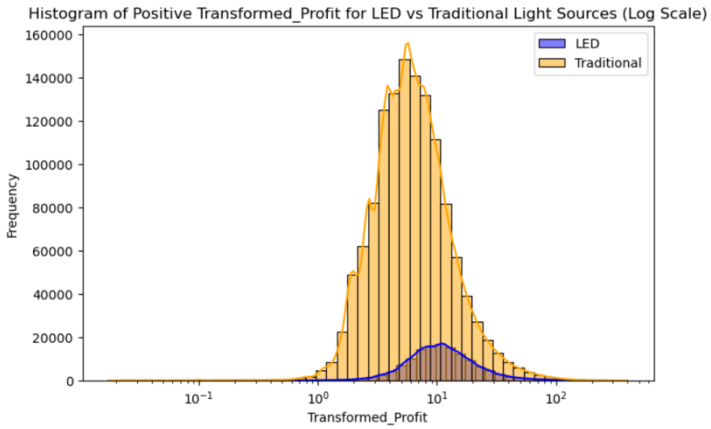
- **Results Obtained**

  The results from the t-test show a t-statistic of 90.08 and a p-value of 0.0, which is significantly below the 0.05 threshold. This indicates a statistically significant difference in the mean Transformed_Profit between products with LED and Traditional light sources,

suggesting that one category yields a higher average profit *(reject the null hypothesis)*. The positive t-statistic suggests that LED products have a higher average Transformed_Profit than Traditional products.

```
T-statistic: 90.08206518375088
P-value: 0.0
The mean Transformed_Profit is significantly different between LED and Traditional light sources.
```

The histogram further supports this finding by showing a clear difference in the distribution of Transformed_Profit between the two groups. The traditional light sources (in yellow) have a higher frequency at lower Transformed_Profit values, while the LED products (in blue) show a distribution that shifts towards higher values. This visual evidence aligns with the statistical test, reinforcing the conclusion that LED products are more profitable on average than traditional products.



- **Valuable for Management**

The findings of this analysis provide critical insights for management by confirming a statistically significant difference in profitability between traditional and LED light sources. This data-driven confirmation helps management make informed decisions to maximise the strengths of each category. For instance, with traditional products showing higher average profits, management can confidently prioritise this category in resource allocation, such as increasing advertising budgets, expanding inventory, or bundling Traditional products with complementary items to boost their appeal. The statistical significance also highlights areas where LED profitability might be improved; management could explore partnerships with sustainability initiatives to position LED products as a high-value, eco-friendly choice, potentially expanding their market. Additionally, understanding the profitability difference enables management to experiment with tailored pricing strategies, such as applying higher markups to traditional products due to their profitability and testing dynamic pricing for LED products based on demand trends. The t-test results offer a reliable basis for these decisions, reducing guesswork and ensuring that each strategy is backed by statistical evidence. This analysis enables management to optimise each category's strengths for sustainable growth, ensuring strategic decisions are based on solid evidence.

Section 4: Inference

4.1 What factors drive sales revenue across different customer segments?

We identify key drivers of sales revenue across customer segments to inform LuminaTech's strategies in marketing, inventory management, and pricing. By analyzing factors such as product quantity, cost, discounts, and region, LuminaTech can determine which variables most significantly impact revenue, allowing for targeted business strategies.

Explanation of Method Used

We applied multiple regression analysis on log-transformed sales revenue to examine factors like product quantity, cost, discounts, and region, reducing outlier influence and improving model interpretability. Higher costs positively impacted revenue, while discounts had a strong negative effect, suggesting a need to refine discount strategies. Regions like LMP and SUR emerged as key revenue contributors, highlighting areas for targeted marketing and inventory.

Present the Results

- Model Fit (R-squared & RMSE): The model explained 28.1% of revenue variance (R-squared = 0.281) with a moderate prediction error (RMSE = 415.45), suggesting that additional variables or more complex modelling could improve fit and accuracy.
- Interpretation of Coefficients:
  - value_quantity(-0.5968) - Slight revenue decrease associated with higher quantities, possibly due to bulk discounts.
  - value_cost(1.6039) - Positive correlation with revenue, as higher-cost items generate more revenue.
  - value_price_adjustment(-180.59) - Significant negative impact of discounts on revenue, suggesting insufficient volume boost to offset revenue loss.
  - day_gap(0.0342) - Weak positive relationship, suggesting that higher-value items or bulk orders take longer to process.
  - Regional Codes - Regions LMP (4.2870) and SUR (2.4689) have positive impacts, identifying them as strong revenue drivers, while DLT (-1.6112) and Other (-10.6006) show negative impacts, suggesting these areas may benefit from strategy adjustments or additional resources to stimulate growth.

Robustness Evidence

- Variance Inflation Factor (VIF): VIF values close to 1 for all features, indicating no multicollinearity and unique contribution of each variable.

- Cross-Validation: Variability in R-squared scores (0.01 to 0.86) suggests inconsistent model performance across subsets. Mean cross-validation R-squared (0.5601) indicates moderate explanatory power but also suggests further model refinement for stability.

Value for Management

This analysis provides strategic insights to help management optimize profitability. The negative impact of discounts highlights a need to review discount strategies, as they currently reduce revenue without sufficiently boosting volume. Strong revenue contributions in regions like LMP and SUR suggest that these areas could benefit from increased marketing and inventory focus to maximize returns. The weak association between quantity and revenue implies that high-volume sales may not yield higher profits, possibly due to bulk discounting, suggesting a review of bulk pricing practices. Finally, the variability in cross-validation results indicates that the model may need refinement to provide more stable, actionable insights across diverse customer segments. By focusing on targeted discounting, regional investment, and pricing adjustments , management can better align strategies with revenue optimization goals.

4.2 What Factors Influence Profit Margin Across Product Categories and Customer Segments?

This analysis identifies key drivers of profit margin across product types and customer segments, supporting LuminaTech's focus on high-margin products, effective discounting, and region-specific strategies.

Explanation of Method Used

Using multiple regression analysis, we examined how product quantity, cost, discounts, and region influence profit margin. This approach allows us to assess the unique impact of each factor, including distinctions between product types and regional contributions.

Present the Results

- Model Fit (R-squared & RMSE):
  - R-squared (0.2106): The model explains approximately 21.06% of the variance in profit margin, indicating that other unmodeled factors may also play a role.
  - RMSE (17.42): This suggests an average prediction error of 17.42 in profit margin, reflecting moderate prediction accuracy
- Coefficients:
  - Cost (0.0006): This small positive coefficient indicates a minimal positive impact on profit margin from higher costs, though the effect size is very small.

- Quantity (-0.0018)*:* The negative coefficient suggests a slight decrease in profit margin with higher sales quantities, possibly due to bulk discounts reducing profitability.
- Discounts (50.28)*:* A significant positive coefficient for discounts suggests that, when well-targeted, price adjustments (discounts) can increase profit margin, potentially by boosting high-margin sales.
- Product Types*:* Both have negative coefficients, with LED products reducing profit margins more than Traditional products, likely due to higher costs or competitive pricing in the LED market.
- Regions*:* LMP (11.00) and SUR (3.74) show strong positive impacts on profit margins, marking them as top profitable markets. DLT (5.32) and Other (2.01) also boost margins but to a lesser extent, suggesting moderate profitability.

Robustness Evidence

- VIF: VIF values are mostly low, indicating minimal multicollinearity. light_source_Traditional (5.03) and business_area_code_LMP (3.80) have slightly higher VIFs but remain below concerning thresholds, suggesting each predictor contributes uniquely to the model.
- Cross-Validation: R-squared scores range from 0.16 to 0.26, showing some variability but overall modest explanatory power. The mean R-squared score of 0.2056 aligns with the initial model, indicating that the selected features provide a consistent, though limited, explanation of profit margin variability.

Value for Management

This analysis provides strategic insights to help management optimize profitability. For instance, the positive impact of discounts on profit margins suggests that current discounting practices are effective and could be refined further to maximize returns, especially on high-margin products. High-profit regions, such as LMP and SUR, present opportunities for increased investment, where additional resources could drive better returns. The slight reduction in profit margin with higher sales quantities indicates a need to review bulk discount practices to prevent erosion of profitability. Additionally, LED products are less profitable than Traditional ones, likely due to higher costs or competitive pricing, suggesting adjustments in pricing or cost management may be beneficial. These insights enable to adopt targeted strategies in discounting, regional investments, and product-specific profitability, supporting improved margin optimization across segments.

Section 5: Prediction Model

To predicted sale price in 2014, we developed a prediction model to forecast daily sales for 2014, focusing on historical sales data from 2012 and 2013. By employing time series analysis and using the Holt-Winters Exponential Smoothing method, we aimed to capture key patterns and trends in the sales data, including seasonality and potential holiday effects. Our analysis and model forecast provide insights that are valuable for strategic planning, resource allocation, and demand forecasting.

- **Data Exploration and Preparation**

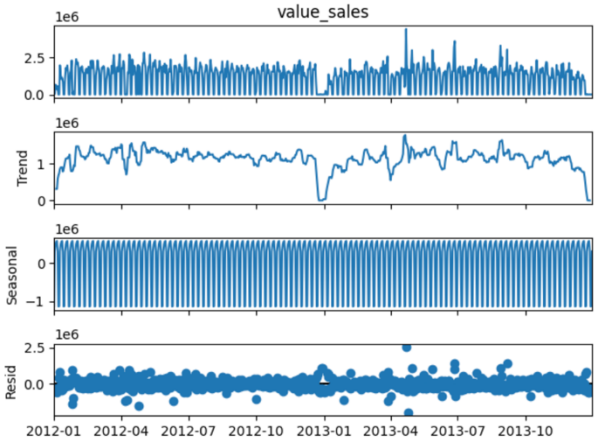
The dataset, covering 2012 to 2013, was initially cleaned by filtering for key columns (order_date and value_sales), converting 'order_date' to datetime, and ensuring all 'value_sales' were positive. We calculated daily sales by aggregating values on 'order_date', focusing on data within the specified two-year range. We filled gaps with zero sales to address missing dates, assuming this reflects typical B2B (Business to Business) operations, with no sales on weekends or holidays.

- **Time Series Decomposition**

To better understand the underlying components of our sales data, we performed an additive decomposition of the 'value_sales' time series. The decomposition chart is split into four components, each representing a different aspect of the sales data. This breakdown helps identify trends, seasonality, and irregular fluctuations, which are valuable for forecasting and understanding factors influencing sales trends.

Observed: The top plot shows actual 'value_sales' over two years, with regular peaks and troughs typical of B2B cycles, reflecting both trends and irregular fluctuations.

Trend: The trend line shows a gradual rise in early 2012, becoming more variable in 2013, with dips suggesting seasonal slowdowns or external demand factors.



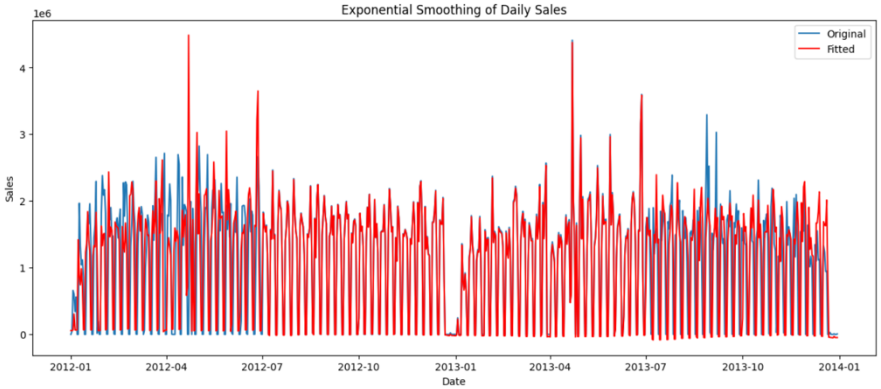
Seasonal: The seasonal component highlights consistent cycles, indicating periodic demand shifts likely tied to B2B patterns and seasonal influences, with strong, repetitive patterns each year.

Residual: The residual plot captures random fluctuations and outliers beyond trend and seasonality. While mostly random, larger deviations, especially in 2013, suggest some anomalies that the model does not fully explain.

- **Prediction Model**

For our forecasting model, we employed Holt-Winters Exponential Smoothing (Triple Exponential Smoothing). This is because the method is well-suited for data with trend and seasonality, as it considers three main components:

1. **Level:** Baseline value of the series.
2. **Trend:** Rate of change in sales over time.
3. **Seasonality:** Periodic patterns within the data.



We configured the model with a multiplicative trend and additive seasonality, which aligned well with the characteristics of our dataset.

This chart shows the results of applying Holt-Winters Exponential Smoothing to model and forecast daily sales data from 2012 to 2013. The model captures the general seasonality and trend in daily sales, following recurring patterns and aligning with typical B2B cycles, such as dips on weekends or holidays. However, while it tracks major peaks and troughs, it tends to underestimate extreme sales spikes, indicating limitations in accounting for sudden, irregular fluctuations. Overall, this makes the model suitable for predicting average sales and periodic trends, though further refinement or additional variables may be needed to improve accuracy for high-impact sales events.

- **Model Performance and Evaluation**

After fitting the model to our historical data, we evaluated its performance with the following metrics:

Root Mean Squared Error (RMSE)	644,673.90
Mean Absolute Error (MAE)	331,203.67
Mean Squared Error (MSE)	415,604,435,887.42
R-squared (R^2)	42.37%

While the model captured general trends and seasonality, the R^2 score indicates a significant amount of unexplained variance. This suggests that certain fluctuations may be driven by unexpected factors or external influences beyond typical B2B demand cycles, which the model does not fully capture.

- **Forecast for 2014**

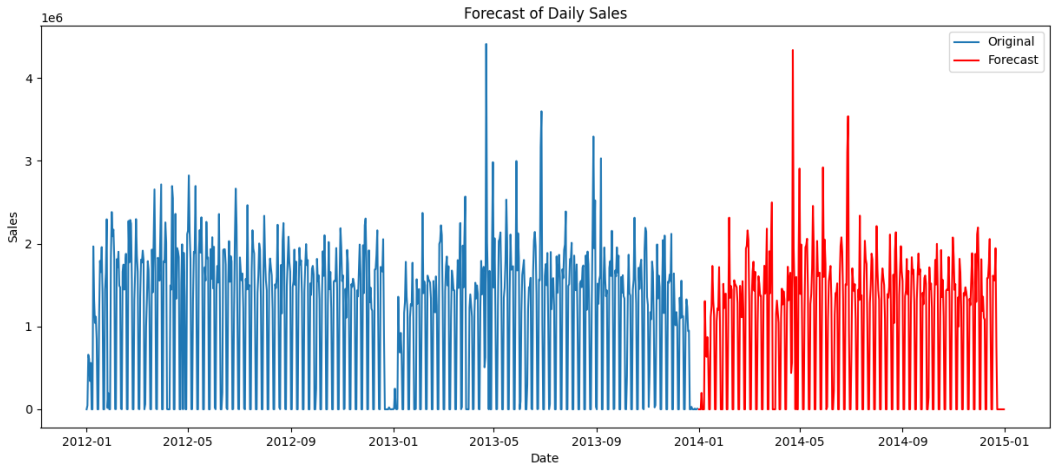
Using the fitted Holt-Winters model, we forecasted daily sales for 2014. The forecasted results were as follows:

- **Total Forecasted Sales for 2014**: 382,886,737.63 AUD
- **Comparison with Previous Years:**

  2012 Sales Total: 423,990,503.37 AUD

  2013 Sales Total: 397,950,946.62 AUD

The forecast suggests a potential decrease in sales for 2014, following the downward trend seen at the end of 2013. This insight could prompt management to explore strategies to counteract this decline, such as targeted marketing initiatives or product promotions.



The chart shows the forecast of daily sales for 2014, with the original sales data (in blue) from 2012 to 2013, followed by the forecasted values (in red) for 2014. The forecast model for 2014 sales closely follows the general daily pattern and seasonal cycles observed in 2012 and 2013, capturing typical B2B fluctuations such as regular dips on weekends or holidays. While the model effectively reflects recurring sales patterns, it underestimates extreme peaks seen in the original data, suggesting it may not fully capture high-impact, irregular events like promotions or special sales. Consequently, the forecasted values are more stable than the historical data, indicating a smoother, more predictable outlook but potentially less accuracy in anticipating sudden sales spikes.

Insights and Recommendations

The prediction model and time series analysis provide management with a data-driven outlook for 2014, supporting proactive strategies to optimize business performance. The strong seasonal patterns observed allow for precise demand planning, enabling strategic inventory and resource allocation during peak periods. However, the forecasted decline in 2014 sales suggests the need for targeted interventions, such as promotional campaigns or marketing efforts, particularly during slower periods, to sustain demand. While the model effectively captures general trends and seasonality, it falls short in predicting certain irregular sales spikes. To enhance accuracy, future improvements could incorporate external factors, like economic indicators or promotional events, that may influence demand. These insights equip management to address anticipated sales trends, ensuring preparedness for both high-demand cycles and potential downturns.

Section 6: Higher Likelihood of Losing Customers

To analyze the likelihood of losing customers, we created an RFM metric to identify churn-related features. RFM (Recency, Frequency, Monetary) analysis segments customers based on purchasing behaviour, helping businesses identify and target valuable customer segments.

- **Recency (R):** Measures how recently a customer made a purchase. More recent purchases indicate stronger engagement and lower churn risk. Calculated as the difference between a reference date and the latest ‘order_date’.
- **Frequency (F):** Measures purchase frequency, with higher frequencies indicating loyalty. Calculated by counting unique ‘customer_order_number’ values.
- **Monetary (M):** Measures total spending, with higher values indicating greater customer value. Calculated by summing ‘value_sales’ (gross profit).

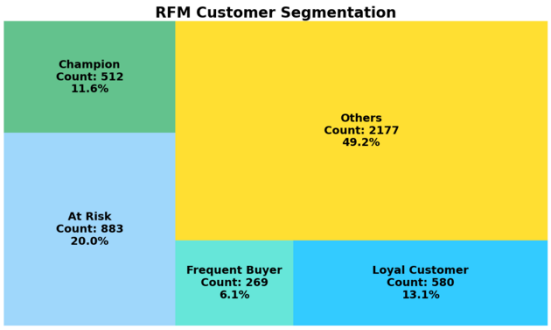
Using these metrics, we scored each component from 1 to 5 and combined them to create an **RFM score**, enabling segmentation as follows below.

Segment	Condition
Champion	RFM score of '555'
Loyal Customer	RFM score starting with '5'
Frequent Buyer	Middle RFM score is '5'
At Risk	RFM score starting with '1'
Others	Does not meet any of the above criteria

After scoring and segmenting customers, we analyzed the churn-prone characteristics of each segment to better understand the factors contributing to customer churn.

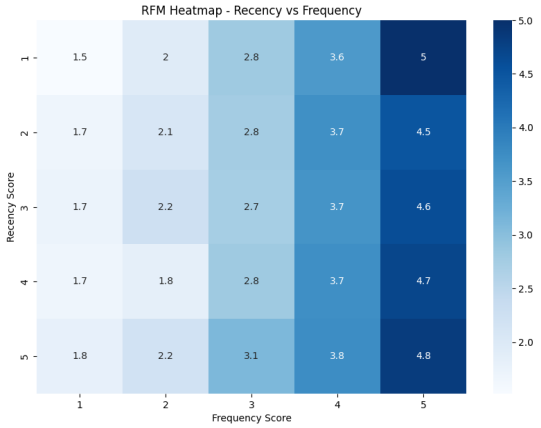
Treemap of Customer Segment Percentages

The treemap of customer segment counts shows that most customers fall into the "Others" category, which comprises nearly half of the customer base (49.2%). The "At Risk" segment also has a substantial portion at 20%, suggesting a notable number of customers may be close to churning. In contrast, high-value segments like "Champion" (11.6%) and "Frequent Buyer" (6.1%) are smaller, underscoring the importance of retaining these valuable customers. This distribution highlights the need to focus retention efforts on at-risk customers through loyalty programs or personalised offers to increase engagement and reduce churn.



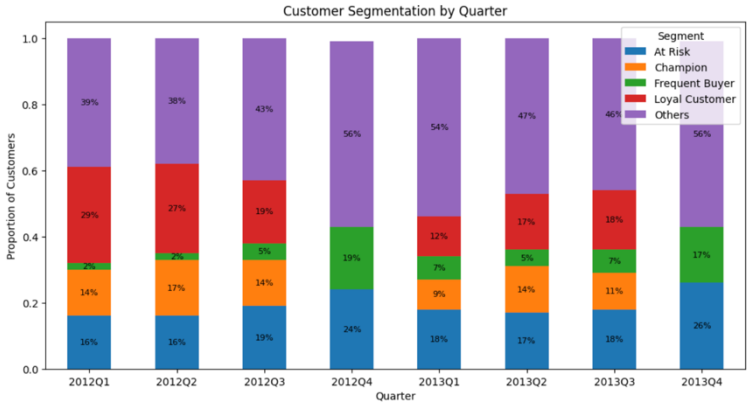
Heatmap of Recency and Frequency

The heatmap shows that customers with high Recency and Frequency scores contribute the most to revenue (darker cells), making them priority targets for retention. Customers with high frequency but low recency scores still have substantial monetary value, suggesting they could benefit from re-engagement. In contrast, low recency and frequency customers show low monetary value, indicating a higher churn risk. This analysis helps prioritise loyalty for engaged customers and re-engagement for frequent but less recent buyers.



Stacked bar chart of Customer Segmentation by Quarter

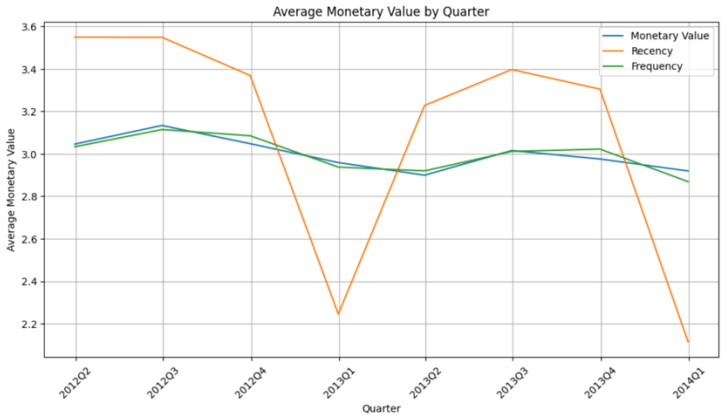
The chart of Customer Segmentation by Quarter reveals engagement shifts, with "Others" peaking in Q4 of 2012 and 2013, indicating reduced engagement toward year-end. The "At Risk" segment also grows in these quarters, suggesting seasonal disengagement and a higher likelihood of churn. High-value segments like "Champion" and



"Loyal Customer" are more prominent in early quarters but decline over time, indicating a need for consistent retention efforts.

Line chart of quarterly trends in average Monetary Value, Recency, and Frequency score.

The line chart of quarterly trends shows declining Monetary Value and Recency scores in Q1 2013 and Q4 2013, suggesting reduced spending and engagement during these periods. Meanwhile, Frequency remains relatively stable, though with a slight dip in Q1 2013, indicating consistent purchase patterns with occasional declines. These insights suggest that targeted re-engagement efforts, especially in Q1 and Q4, could help sustain customer engagement and spending.



In summary

Key features linked to customer churn include low Recency and Frequency scores, even when Monetary scores are high, as well as seasonal disengagement patterns. Customers with low recency and frequency scores, along with low spending, are likely to churn, particularly toward year-end when the "At Risk" segment grows. By focusing on these features, businesses can develop proactive strategies to retain at-risk customers and reduce churn.

Conclusion

This project provided a comprehensive analysis of LuminaTech Lighting's sales, customer behaviour, and operational performance through various data analytics techniques. By cleansing and preparing the dataset, we ensured data accuracy, forming a strong foundation for subsequent analyses. Exploratory insights, such as identifying high-return warehouses and analysing monthly sales trends, enabled targeted recommendations for improving efficiency and profitability. Hypothesis testing uncovered significant differences between product categories, guiding inventory and marketing strategies. Regression analysis identified key factors driving sales revenue, including the positive impact of product cost and regional contributions, and highlighted areas for refining discount strategies due to their negative impact on revenue. Additionally, the prediction model offered valuable forecasts for 2014, helping management plan for anticipated demand shifts. Customer churn analysis, based on RFM metrics, identified key features associated with customer retention, empowering management to develop proactive strategies to retain high-value customers and reduce churn risks. Overall, this report supports LuminaTech in making data-driven decisions to enhance business performance and customer satisfaction while strategically aligning with market trends and operational goals.