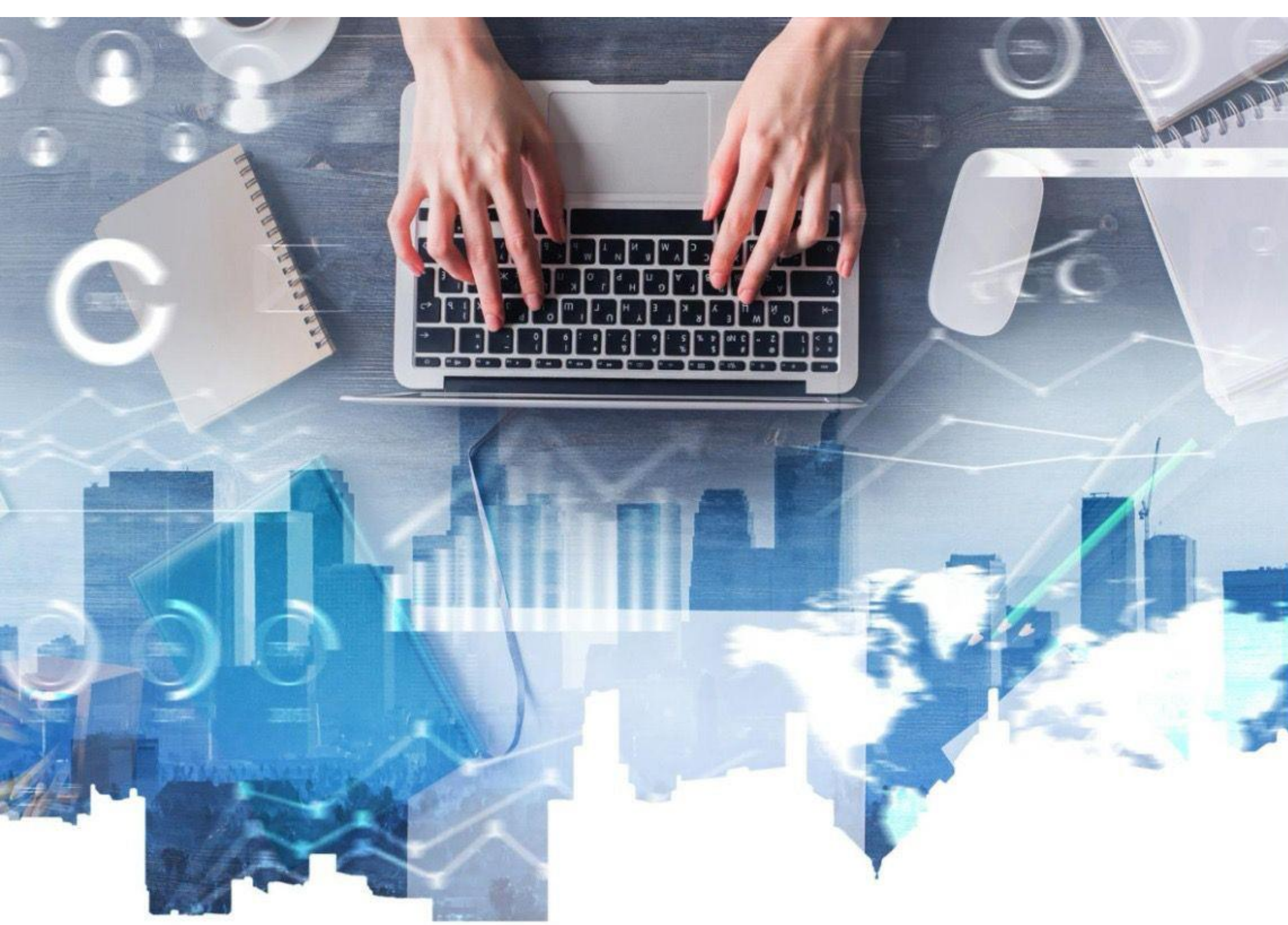


ANALYZING SALES AND CUSTOMER INSIGHTS

A DATA-DRIVEN APPROACH TO ENHANCING SALES PERFORMANCE AND CUSTOMER RETENTION AT LUMINATECH LIGHTING.

Table of Contents

1. Project Overview *3*

1.1 Problem Statement **3**

1.2 Dataset description **3**

2. Clean the dataset **3**

2.1 Data Consolidation **4**

2.2 Column and Data Type Adjustments **4**

2.3. Handling Null & Meaningless Values **4**

2.4. Standardizing Categorical Data **4**

2.5. Negative Values Investigation and Removal **4**

2.6. Handling Specific Scenarios: **4**

2.7. Outlier Detection and Removal **5**

2.8. Skewness and Transformation **6**

2.9. Final Dataset Cleanup **6**

3. Exploratory Insights: **7**

3.1 Time gap between order date and invoice date **7**

3.2 Customer Retention Rates **8**

3.3 The impact of currency on profit margins **9**

3.4 Sales Trend by Currency **10**

3.5 Sales Distribution by Customer District Code **11**

Summary Insights **13**

Overall Recommendations **13**

4. Test Sub Sample Differences **14**

Question 1: Is There a Significant Difference in Average Transaction Value Between High Retention and Low Retention Periods? **14**

Question 2: Are Profit Margins Different Between High and Low Retention Periods? **14**

5. Interference **15**

Question 1: What Business Factors Most Significantly Influence Sales Value? **15**

Question 2: What business factors most significantly influence profit? **18**

6. Sales Prediction **19**

Selecting Features for Sales Prediction **19**

Linear Regression for Sales Prediction **20**

7. Customer Churn Analysis **23**

Logistic Regression for Customer Churn Analysis **23**

SMOTE **24**

ROC AUC **24**

Recursive Feature Elimination **25**

8. Conclusion **27**

1. Project Overview

1.1 Problem Statement

LuminaTech Lighting aims to strengthen its market position and improve its overall business performance. However, the company cannot identify the different factors that will determine sales, customer loyalty, and efficiency in maintaining inventories. The company has a huge amount of data regarding its sales, customer demographics, and inventory but lacks the insight to make better decisions.

In a competitive lighting market, understanding customer behavior, product demand, and sales patterns is essential to retaining customers, attracting new ones, and managing inventory effectively. This project will clean, analyze, and visualize LuminaTech's data to reveal trends and identify factors affecting sales and retention. The goal is to provide actionable recommendations that help the company make informed decisions and support growth.

1.2 Dataset description

The dataset contains detailed information related to LuminaTech Lighting's sales transactions, customer demographics, and inventory characteristics. Key variables include:

- **Dates and Periods** (accounting dates, fiscal and calendar year/months, providing temporal context for each transaction),
- **Identifiers and Codes** (unique identifiers for companies, customers, items, and various business areas, as well as codes for business chains, sales channels, and product categories),
- **Product Details** (information on product categories, item types, environmental and technology groups, and inventory classifications, e.g., ABC classes),
- **Sales and Financials** (transaction values, costs, quantities, currency, and price adjustments, which offer insights into financial performance and sales trends),
- **Order and Invoice Information** (details on order types, order and invoice dates, and customer order numbers for transaction tracking).

2. Clean the dataset

2.1 Data Consolidation

Two datasets (data_2012 and data_2013) were combined into a single dataset with pd.concat, resulting in 1,988,382 rows and 41 columns. This unification enabled a holistic approach to data cleaning and analysis across both years.

2.2 Column and Data Type Adjustments

For date-time display, we set the display option to show all columns (pd.set_option("display.max_columns", None)). For date columns, we converted accounting_date, invoice_date, and order_date columns to a datetime format to allow time-based operations. We also changed several columns (fiscal_year, fiscal_month, calendar_year, calendar_month, calendar_day, value_sales, value_cost, and value_quantity) to numeric types to enable arithmetic and statistical operations.

2.3. Handling Null & Meaningless Values:

The item_source_class column was removed as it contained only NaN values, offering no useful information for analysis. We identified **543** duplicate rows, which were removed to avoid redundancy and skew in subsequent analysis. We also dropped column dss_update_time since it a datetime column but contains only 1 unique value. Order-type_code is the a code that categorizes the type of order, we removed order type that marked **Do not use** (ZC2, 5TN, ZD3...) in the order type description as it doesn't have any contribution to analysis of dataset.

2.4. Standardizing Categorical Data:

Most unique values in categorical columns have space, we need to strip them for further join. In this case, we will strip all the column with Object type, since this may happen to all of them including those unique id columns. We also replaced mislabeled currency values ('AUS' with 'AUD') to ensure consistency. Entries with blank currency values were removed, as they constituted only two rows.

2.5. Negative Values Investigation and Removal:

Value_sales, value_cost, and value_quantity had negative values, suggesting returns, refunds, or adjustments. Joining with Metadata: Merged data with a metadata file containing descriptions for order_type_code to understand the nature of negative values.

2.6. Handling Specific Scenarios:

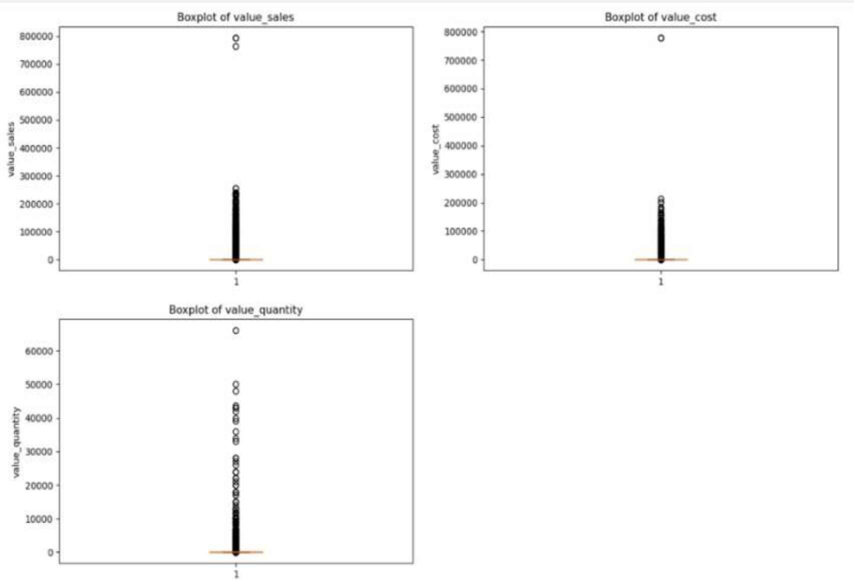
Transactions with order type codes such as NOR - Normal Order and NOH - Normal Order Head Office Sales that have negative values fall into two distinct scenarios. **First**, they may represent internal transactions between entities within the same corporate group or a multinational company (indicate by the value in business_chain_l1_code = 'INTERCO'). **Second**, they could be error entries. In both cases, these transactions should either be removed or kept separate in a different dataset, as they do not reflect genuine external sales. The pricing in these transactions (e.g., transfer pricing) may differ from market prices, and they follow different business rules compared to regular customer sales. Including them could distort actual sales trends and financial metrics. In this case, I will keep these internal transactions in a variable called internal_transaction for further analysis if needed.

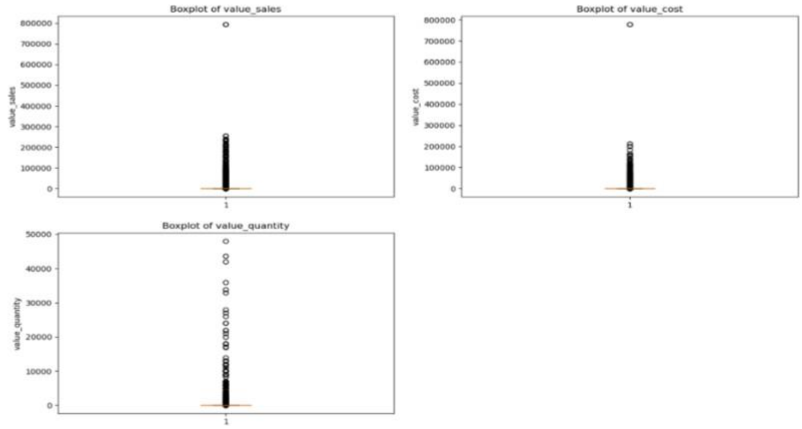
As for credit transaction, they are credited transaction and will be negative because the credit nature (returns, refunds, credits). These transactions are dependent on the original sales transactions, which means they happen after the original sales. Including them in sales analysis or use them in prediction would skew the pattern of regular sales. That is why we will separate them from the dataset and keep them in another variable called credit_transactions for further analysis if needed. **Additionall**y, we will also remove transactions with value_quantity = 0 or value_quantity that has decimal value (0.1, 0.2, etc,..) or value_sales = 0 or value_cost = 0, since this make no sense and could skew the dataset. Furthermore, regarding value_quantity, there are entries with decimal values. As this column should contain only whole numbers, the entries with decimal values are removed.

2.7. Outlier Detection and Removal

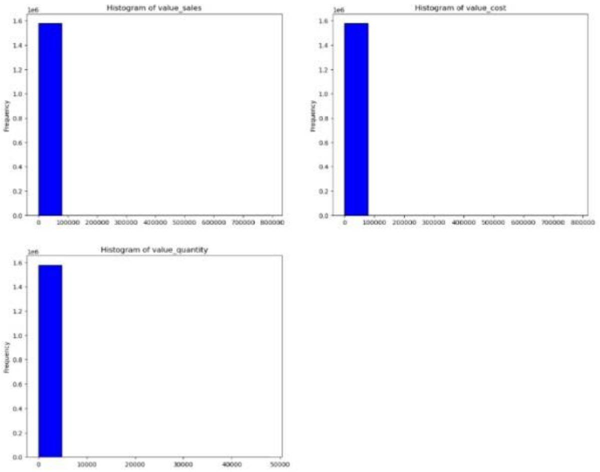
We still have plenty of outliers after addressing the negative values. We cannot identify the extreme outliers by calculating the interquartile range then the lower, upper limits for the whole columns (value_sales, value_cost, value_quantity). Because each item likely has its own price range, and an outlier for one item might be normal for another. For example: a luxury lighting product will be an outlier if we compared it to the standard lighting accessory. That is why in this case, we will use the item_code (A unique identifier for the item being sold) to calculate the interquartile range and upper, lower limit for each unique item.

To identify the outlier, we group the value_sales, value_cost and value_quantity by item_code and currency. Then in each item_group, we calculate the maximum, minimum, median, 25th quartile and 75th quartile. From there, we can calculate the interquartile IQR = Q3 – Q1, and its upper limit U = Q3 + 1.5*IQR, and lower limit L = Q1 – 1.5*IQR. Any data point lies outside this range will be considered as extreme outliers.



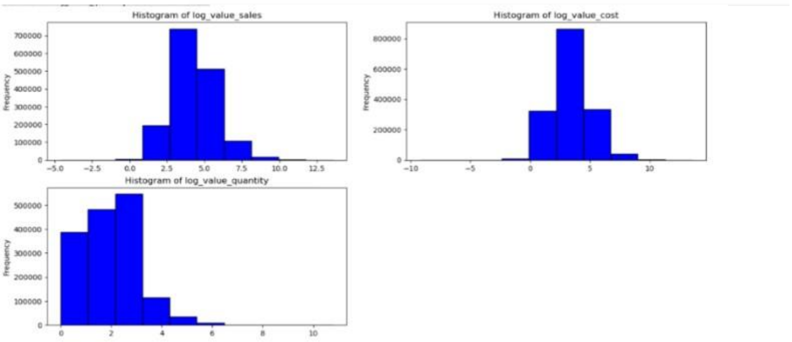


After applying methodology, we have below diagram. We can still see outliers in the boxplots, this is because different items naturally have different price ranges. When plotted together, expensive items might appear as outliers compared to cheaper items. High-volume items will show as outliers compared to low-volume items. We will keep these outliers because it preserves the natural variation between different products, maintains legitimate business patterns and keeps valid high-value or high-volume transactions.



2.8. Skewness and Transformation

We found high skewness in value_sales (95.42), value_cost (189.01), and value_quantity (193.54). We can see that the skewness of value_sales, value_cost and value_quantity is so high, reflecting that the distribution of these values are not normal in diagram. Applying logarithmic transformation to reduce skewness for these fields, making the distribution more symmetrical with its new skewness log_value_sales (0.48), log_value_cost (0.31), and log_value_quantity (0.39). Post-transformation, skewness values fell between 0 and 0.5, indicating a nearly symmetrical distribution, which enhances the reliability of subsequent analyses.

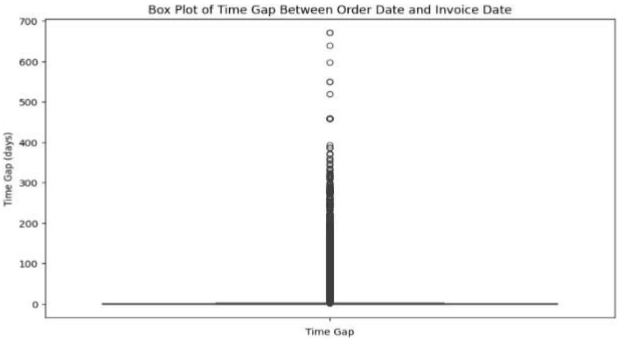


2.9. Final Dataset Cleanup

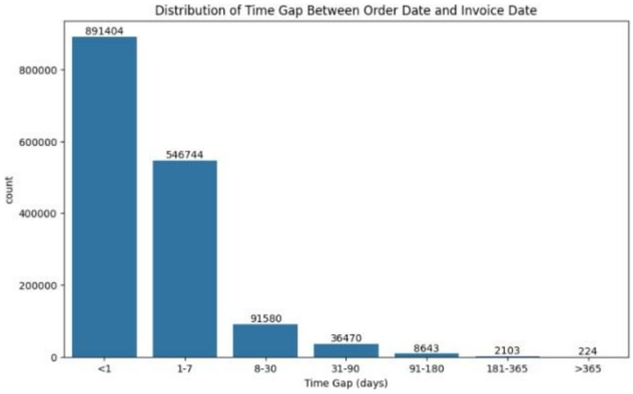
We also removed temporary columns used for outlier detection (e.g., lower_limit_sales, upper_limit_sales). After all steps, the cleaned dataset contained 1,748,010 rows and 39 columns, free from duplicates, erroneous entries, and extreme outliers.

3. Exploratory Insights:

3.1 Time gap between order date and invoice date

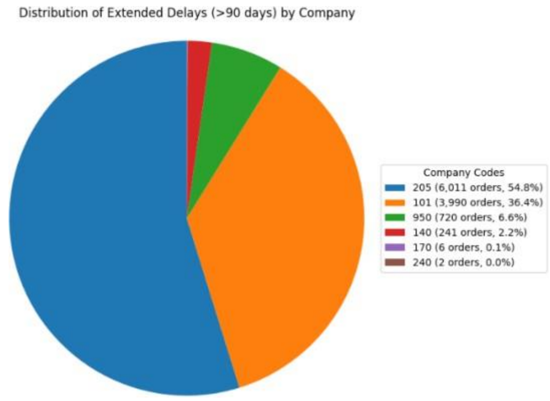


The interquartile range appears near the bottom, indicating that most orders are processed quickly. However, there are numerous outliers extending all the way up to about 900 days. Since the large number of outliers makes it difficult to understand the patterns in the data, we need to create time gap segments to make it easier to understand and find the reason behind this. The time gap segmentation will be created as follow: <1 day, 1-7 days, 8-30 days, 31-90 days, 91-180 days, 181-365 days, > 365 days



The chart shows that most orders are invoiced quickly, with 891,414 orders processed in less than 1 day and 546,752 orders within 1-7 days. The number of orders drops sharply in the subsequent time windows, with 91,600 orders taking 8-30 days and 36,545 orders processed in 31-90 days. However, a small number of orders (10,986 in total) are delayed by more than 90 days. Although this represents a small proportion of the overall orders, these significant delays may negatively impact the customer experience. So we will investigate closer on these extreme delays.

Investigate on the extreme delays (>90 days) Looking at the pie chart, we can see that company 205 leads with 54.7% of all extended delays, accounting for 6,011 orders, followed by Company 101 with 36.3% (3,990 orders). Company 950 takes third place with 736 extreme delay orders. Other companies have minimal contributions to these delays.



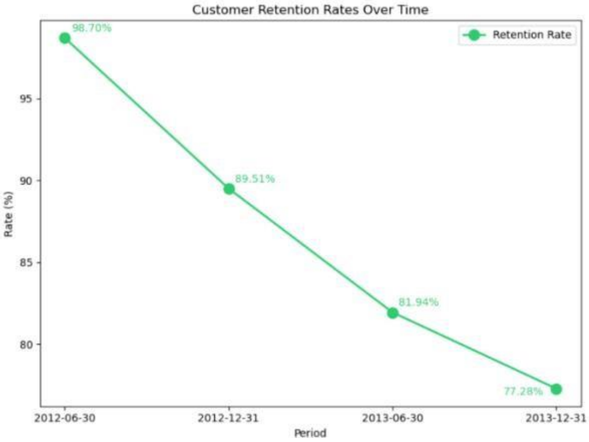
Recommendation: Since the extreme delay orders mostly come from companies 205, 101, and 950, the focus should be on identifying and addressing the specific steps causing these delays in their processes. After the issues in these companies are resolved, the improved processes can be applied to other companies to improve overall efficiency

3.2 Customer Retention Rates

The customer retention rates formula is:

$$CRR = \frac{\text{Customers at the End of the Period} - \text{New customer acquired during the period}}{\text{Customers at the Start of the Period}} \times 100$$

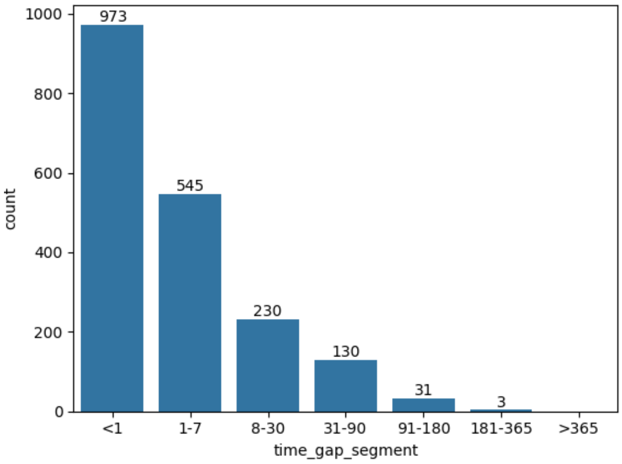
For this analysis, I will use a 6-month period. To calculate the number of customers at the beginning of the period (2012-01-01), I will get all unique customers who placed orders before this date. Next, I will calculate the total number of unique customers who made orders from 2012-01-02 to 2012-06-30. The customer retention rate will be calculated by dividing the number of remaining customers at the end of the period by the total number of unique customers at the start. This process will be repeated for subsequent 6-month intervals, continuing until the final period on 2013-12-31.



It can be seen that retention rates have been steadily dropping over time. They started strong at 98.7% on 2012-06-30 but fell to 77.28% by the end of 2013, which is a total drop of 21.42%. This decline suggests there might be some issues causing customers to leave, so we will inspect the reasons behind it.

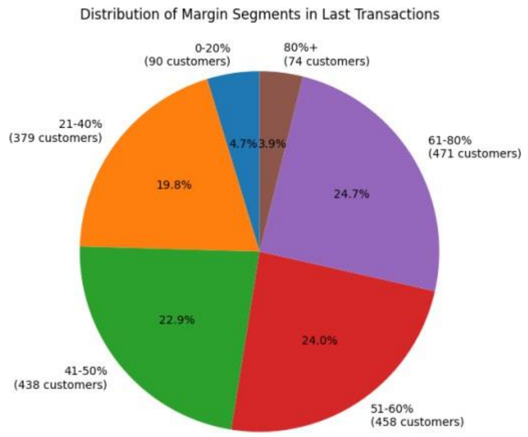
To have clearer view about the reason why they are not coming back, we need to extract the last transaction that they made.

Overall, most of the last transactions for churned customers were processed within a day (973). Another 545 took 1-7 days. Moderate delays happened for 230 transactions in the 8-30 day range and 130 in the 31-90 day range, while only 34 transactions had extreme delays (over 90 days). This suggests that the time gap between the order date and invoice date probably is not the reason why customers are churning.



Move on to value_sales and value_cost, it can be seen that the value_sales is doubled the value_cost, we need to check the profit margin of these transactions. Profit margin can be calculated by Gross Profit Margin = [(Value Sales – Value Cost) / Value Sales] x 100. After that, we need to create margin segmentation to make it easier to find the pattern in the data. The margin segmentation will be created as follow: 0-20% (Low

Margin), 21-40% (Moderate Margin), 41-50% (Moderate Margin), 51-60% (High Margin), 61-80% (High Margin), +80% (Very High Margin).



Overall, almost 53% of churned customers had pretty high margins, over 50% on their last purchase. To be specific: 24.0% (458 customers) had margins between 51-60%, 24.7% (471 customers) had margins in the 61-80% range, and 3.9% (74 customers) were above 80%. This pattern suggests that high margins may play a significant role in customer churn. However, it’s important to note that around 47% of churned customers had margins below 50%, indicating that factors beyond pricing could be driving churn.

Recommendations: Given the concentration of churned customers in higher-margin segments (above 50%), this could point to potential pricing issues, such as customers finding cheaper alternatives or showing sensitivity to price. Balancing margin optimization with retention strategies could help address these concerns and reduce churn.

3.3 The impact of currency on profit margins

Methodology: First, we need to group by currency to get the total value sales and total value cost. Then, we calculate the margin and profit margin by this formula Gross Profit Margin = [(Value Sales – Value Cost) / Value Sales] x 100. Additionally, to know if the currency has a statistical significant to profit margin, we use t-test in this case. Since the sample size for each currency group in this case is large (> 30), we don't need to follow the normality assumption. So we don't need to do the transformation to normalise the profit margin.

Analysis and Findings:

Currency pairs	**AUD & USD**	**AUD & NZD**	**AUD & EUR**	**USD & NZD**	**USD & EUR**	**NZD & EUR**
T-statistic	13.37	1.59	7.69	-11.45	6.45	7.54
p-value	0.00	0.11	0.00	0.00	0.00	0.00

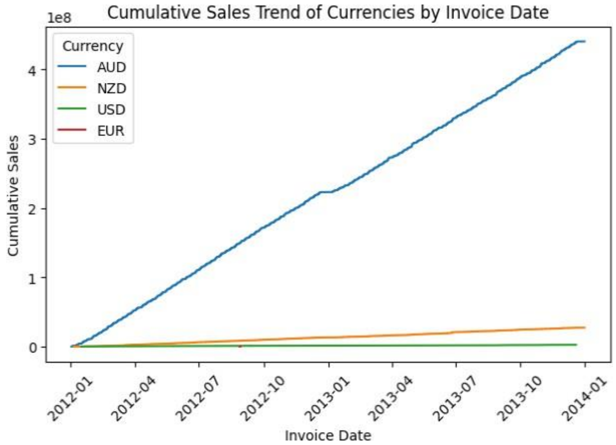
The t-test results revealed significant regional patterns across currency groups. The analysis highlights no significant difference was found between **AUD and NZD**’s mean profit margins, indicating that profit margins are similar across Australia and New Zealand. This similarity suggests a uniform approach to pricing and cost structure within the Oceania region. For **USD and EUR**, Both currencies exhibited statistically significant differences in mean profit margins when compared to AUD and NZD, as well as to each other. This differentiation indicates distinct pricing or cost dynamics in the North American and European markets, possibly influenced by varying competitive pressures, demand factors, and regional economic conditions. The findings delineate three distinct profit margin patterns across Oceania (AUD/NZD), North America (USD), and Europe (EUR). These patterns underscore the importance of tailoring pricing strategies to regional market conditions, with Oceania presenting consistent margins and North America and Europe each displaying unique margin characteristics.

Recommendations: To maximize profitability across regions, the company should consider adjusting its pricing strategies based on these regional insights. For **Oceania (AUD/NZD)**, the alignment in profit margins between Australia and New Zealand suggests that a harmonized pricing strategy could be implemented, promoting consistency across the Oceania region while potentially reducing administrative complexities. For **North America and Europe (USD/EUR)**, given the significant margin differences observed, pricing strategies for these regions should account for local market conditions, such as cost structures, demand elasticity, and competitive pricing. Targeted pricing adjustments in these regions could help optimize profitability while aligning with market-specific dynamics.

3.4 Sales Trend by Currency

Methodology To perform the analysis, we separated the dataset by currency to calculate cumulative sales over time. The invoice dates were converted to datetime format and sorted chronologically, allowing for a smooth calculation of cumulative sales in each currency group. Finally, a line graph was generated to visualize the cumulative sales trends for each currency over the specified period.

Results The analysis of cumulative sales reveals notable differences in performance across the four currency markets. The **AUD Market** recorded the highest sales, with a total of AUD 440.39 million over the period. The cumulative sales curve for AUD demonstrates a steady, upward trajectory, indicating consistent growth and high sales volume. For **NZD Market** ranks second in sales, totaling NZD 27.37 million. However, the growth trend is relatively flat, suggesting slower and more gradual sales increases compared to the AUD market. In **USD Market**, total sales reached USD 2.60 million. Similar to the NZD market, the cumulative sales line remains flat, indicating minimal sales growth over the two-year period. **The EUR Market** reported the lowest sales volume, totaling only EUR 0.17 million. All transactions in the EUR market were conducted in August 2012, leading to a flat cumulative sales trend over the entire period.



The comparison of cumulative sales trends shows that the AUD market is the company's primary revenue driver, contributing significantly more than other regions. Meanwhile, the flat lines for NZD, USD, and EUR indicate comparatively low sales volumes and growth.

Insights In terms of sales concentration, the AUD market dominates the company’s sales, indicating a heavy reliance on this region. The steady upward trend suggests a stable market with consistent sales, which could support additional growth initiatives. As for growth potential, The NZD and USD markets exhibit only slight growth, while the EUR market has minimal recorded activity. These trends highlight an opportunity to reassess and possibly expand efforts in these regions to stimulate additional sales. However, regarding market risks, we learned that heavy reliance on a single region for the majority of sales presents a risk of revenue volatility due to market-specific economic, political, or regulatory changes in Australia. Diversifying across regions can help balance these risks by establishing revenue sources in additional markets.

Recommendations

- Sustain and Enhance AUD Market Investments: Given the AUD market’s significant contribution to overall sales, continued investment in this region is recommended to maintain and potentially boost growth. Marketing, customer engagement, and tailored product offerings in Australia could help sustain this upward trend.
- Strategic Diversification: To reduce revenue concentration risk, management should consider strategies to increase sales in the NZD, USD, and EUR regions. Targeted campaigns, region-specific product promotions, or partnerships in these areas could help foster growth. Expanding in the NZD market, which ranks second, may offer a relatively higher return on investment given its existing sales foundation.
- Re-evaluation of the EUR Market Strategy: Given the EUR market’s limited activity, the company may need to assess whether there is potential to expand within this region or if resources should be reallocated to higher-performing regions. If market analysis suggests a viable opportunity, targeted initiatives could help establish a stronger foothold in the EUR region.

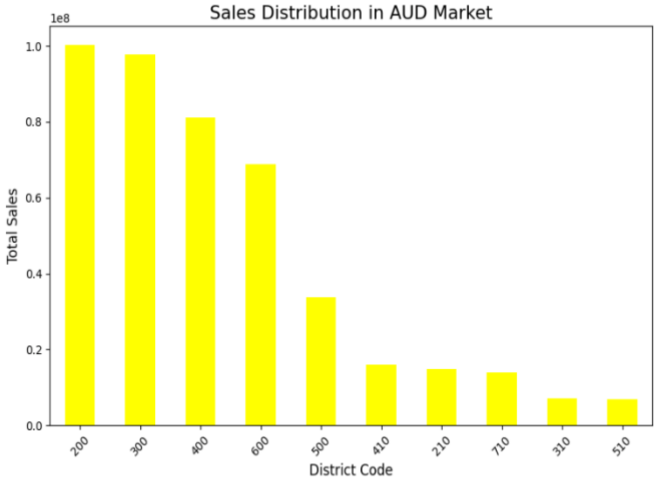
3.5 Sales Distribution by Customer District Code

Methodology To assess sales distribution by customer district, the dataset was filtered by currency, and total sales were computed for each district. Districts were ranked by sales volume, and visualizations were created to highlight differences across the AUD, NZD, EUR, and USD markets. Bar charts were generated for the top-performing districts in each currency market to illustrate the varying contributions from each district.

Results

3.5.1 Sales Distribution in AUD Market

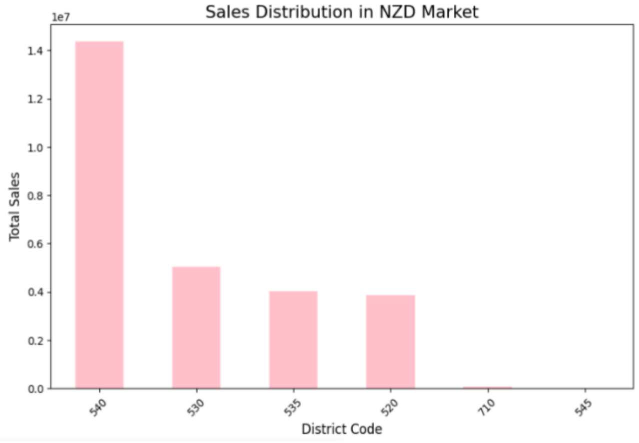
The AUD market is the most substantial, with total sales reaching AUD 440.39 million.



- Top Districts by Sales: District 200 leads with AUD 100.19 million, followed by District 300 (AUD 97.78 million) and District 400 (AUD 81.03 million). Together, these top three districts account for 63% of total AUD market sales, indicating high regional concentration.
- Customer Base: Districts 200, 300, and 400 also rank highest by unique customer count, with District 200 alone serving 963 customers. This suggests these regions have a large and potentially loyal customer base.
- Underperforming Districts: Districts 540, 720, and 510 had the lowest sales, indicating underperformance in these areas.

3.5.2 Sales Distribution in NZD Market

Total sales in the NZD market amounted to NZD 27.37 million.



- Top Districts by Sales: District 540 stands out with NZD 14.37 million, accounting for 53% of total NZD sales. Districts 520 and 530 follow with NZD 5.03 million and NZD 4.03 million, respectively.

3.5.3 Sales Distribution in EUR Market

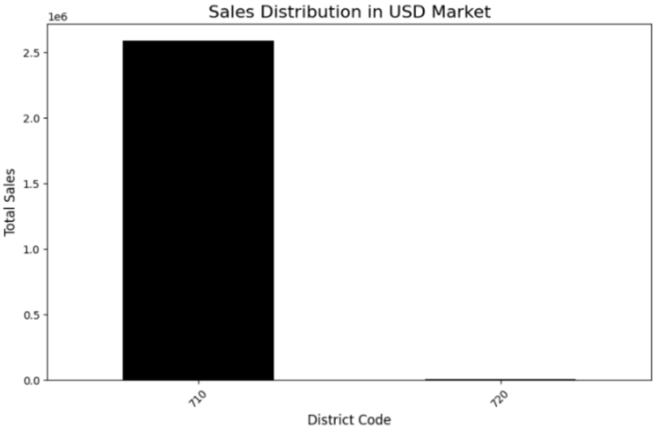
Sales in the EUR market are relatively low, totaling only EUR 0.17 million. As given the lack of variation, no bar chart was generated for EUR sales distribution. And we can see from the EUR market that:

- Limited Activity: All transactions are confined to a single district, indicating minimal market presence.
- Profitability Concerns: The EUR market incurred a loss of EUR 0.12 million when subtracting cost from sales, raising questions about sustainability.

3.5.4 Sales Distribution in USD Market

Total sales in the USD market were USD 2.60 million.

- Top District: District 710 generated nearly all USD sales, totaling USD 2.59 million, while District 720 contributed only USD 5,619.
- Customer Base: Given the reliance on District 710, customer diversification is limited, indicating potential market concentration risks.



Summary Insights

The sales distribution analysis by customer district code provides key insights into regional performance and market concentration. The AUD and NZD markets show potential for further growth with concentrated efforts in top-performing districts. However, the EUR and USD markets exhibit limited activity, suggesting opportunities for strategic reassessment or reallocation of resources.

Overall Recommendations

- In the AUD market, Continue to support districts 200, 300, and 400 by allocating more resources, such as marketing, staffing, and product availability. Since they are already performing well, further investments could yield even higher returns. This market has a number of customer districts with potential. Although sales in other customer districts lag behind the top three, continue to support and encourage the growth of other customer districts.
- In the NZD market, customer districts 540, 520, and 530 contribute significantly to the regional market's total sales. Continue to support their growth by allocating additional resources, such as sales staff, marketing, or inventory management tools. Customer districts 710 and 545 underperform compared to other regional customer districts. Continue to support districts 710 and 545 by increasing marketing and brand exposure.
- In the EUR market, there is only one customer district operating at a low level. Customer district 710 was not profitable in the EUR market, and as a result, pricing and sales strategy may not be suitable and require re-evaluation. If perforamnce in the EUR market improves, consider expanding the number of customer districts to diversiy the revenue stream. However, if losses continue, it may be more beneficial to cease operations in the EUR market, and reallocate resources to other regional markets.
- In the USD market, allocate additional resources to customer district 710 given its profitability and dominance in the region. Consider increasing the marketing budget, expanding sales staff, or enhancing inventory management to sustain and grow its

performance. However, relying heavily on one district may expose the business to regional risks. Diversifying strategies to increase sales in other districts such as 720 could help balance the revenue stream and mitigate risk. GIven the success of customer district 710, customer district 720 could be a district with high growth potential.

4. Test Sub Sample Differences

4.1. Question 1: Is There a Significant Difference in Average Transaction Value Between High Retention and Low Retention Periods?

Methodology:

- Period Definition: The high retention period is defined from January 1, 2012, to June 30, 2012, while the low retention period is from July 1, 2013, to December 31, 2013.
- Statistical Test: A two-sample t-test was conducted to compare the mean transaction values in both periods, assuming unequal variances.

Results:

- T-statistic: -3.88 | P-value: 0.0001 (significant at α = 0.05)
- Average Transaction Values:
  - High Retention Period: $281.57
  - Low Retention Period: $302.85

Interpretation: The p-value indicates a statistically significant difference in average transaction values between the high and low retention periods. The higher average transaction value during the low retention period suggests that while high-spending customers are retained, the company might be losing lower-spending, possibly more price-sensitive customers.

Recommendations:

- **Diversified Pricing**: To appeal to both high- and low-spending segments, consider balancing pricing strategies or offering a range of products that cater to different spending capacities.
- **Retention Strategies**: Implement retention initiatives specifically targeting lower-spending customers to foster a broader customer base.
- **Customer Segmentation Focus**: Recognize and nurture the core high-spending segment while also working to retain a diverse customer demographic.

4.2. Question 2: Are Profit Margins Different Between High and Low Retention Periods?

Methodology:

- **Profit Margin Calculation**: For each transaction, profit margin was calculated as (value_sales−value_cost)/value_sales×100.
- **Statistical Test**: A two-sample t-test was performed to assess any differences in mean profit margins between the two periods.

Results:

- T-statistic: -0.42 | P-value: 0.67 (not significant at α = 0.05)
- **Average Profit Margins**:
    - High Retention Period: 50.57%
    - Low Retention Period: 51.13%

Interpretation: The p-value suggests no statistically significant difference in profit margins between high and low retention periods. This finding indicates that the observed retention decline is likely not due to changes in pricing or cost structure. Thus, the company should explore other factors affecting customer loyalty, such as customer service quality, market competition, or product relevance.

Recommendations:

- **Customer Experience Focus**: Improve customer engagement, gather feedback, and enhance overall customer satisfaction to address retention without altering the pricing model.
- **Non-Pricing Strategies**: Consider strategies such as personalized marketing, loyalty programs, or value-added services to build customer loyalty, as the current profit margins are stable across periods.

5. Inference

5.1. Question 1: What Business Factors Most Significantly Influence Sales Value?

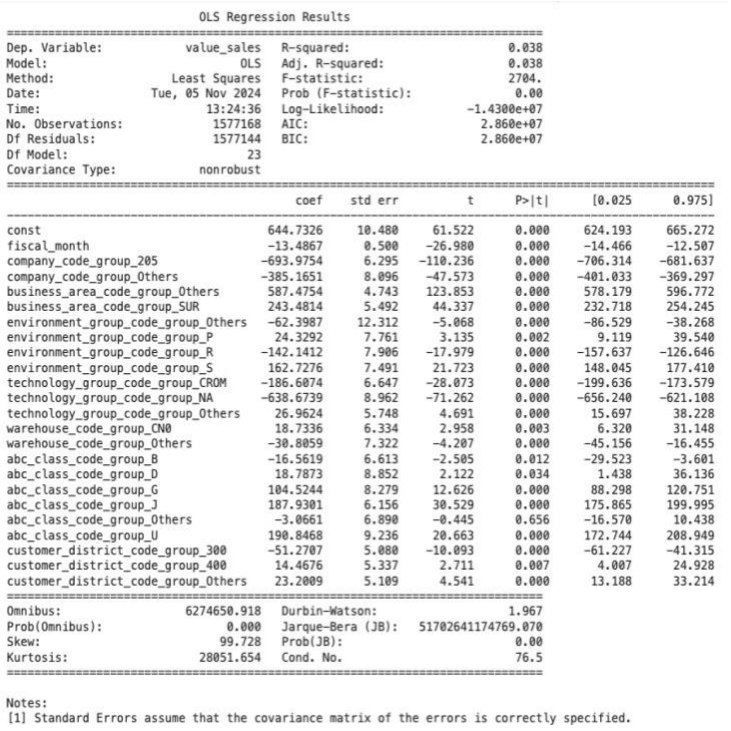
Methodology

To understand the impact of various business factors on sales value, we conducted a multi-step analysis involving data preprocessing, feature engineering, and multiple regression analysis. Given the numerous categories within factors such as company_code, business_area_code, environment_group_code, technology_group_code, warehouse_code, abc_class_code, and customer_district_code, we grouped minority categories (each representing less than 10% of the total) into an "Others" category. This reduction in categorical levels helped prevent overfitting and manage computational complexity in the one-hot encoding step, which we used to prepare the data for regression analysis.

After encoding, we assessed multicollinearity by calculating the Variance Inflation Factor (VIF) for each independent variable. Variables with high VIF values, particularly warehouse_code_group_Unk (unknown), were removed to maintain model stability and ensure more reliable estimates.

Results

Model Fit Statistics:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            value_sales   R-squared:                       0.038
Model:                            OLS   Adj. R-squared:                  0.038
Method:                 Least Squares   F-statistic:                     2704.
Date:                Tue, 05 Nov 2024   Prob (F-statistic):               0.00
Time:                        13:24:36   Log-Likelihood:            -1.4300e+07
No. Observations:             1577168   AIC:                         2.860e+07
Df Residuals:                 1577144   BIC:                         2.860e+07
Df Model:                          23
Covariance Type:            nonrobust
=============================================================================================================
                                          coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
const                                 644.7326     10.480     61.522      0.000     624.193     665.272
fiscal_month                          -13.4867      0.500    -26.980      0.000     -14.466     -12.507
company_code_group_205               -693.9754      6.295   -110.236      0.000    -706.314    -681.637
company_code_group_Others            -385.1651      8.096    -47.573      0.000    -401.033    -369.297
business_area_code_group_Others       587.4754      4.743    123.853      0.000     578.179     596.772
business_area_code_group_SUR          243.4814      5.492     44.337      0.000     232.718     254.245
environment_group_code_group_Others   -62.3987     12.312     -5.068      0.000     -86.529     -38.268
environment_group_code_group_P         24.3292      7.761      3.135      0.002       9.119      39.540
environment_group_code_group_R       -142.1412      7.906    -17.979      0.000    -157.637    -126.646
environment_group_code_group_S        162.7276      7.491     21.723      0.000     148.045     177.410
technology_group_code_group_CROM     -186.6074      6.647    -28.073      0.000    -199.636    -173.579
technology_group_code_group_NA       -638.6739      8.962    -71.262      0.000    -656.240    -621.108
technology_group_code_group_Others     26.9624      5.748      4.691      0.000      15.697      38.228
warehouse_code_group_CN0               18.7336      6.334      2.958      0.003       6.320      31.148
warehouse_code_group_Others           -30.8059      7.322     -4.207      0.000     -45.156     -16.455
abc_class_code_group_B                -16.5619      6.613     -2.505      0.012     -29.523      -3.601
abc_class_code_group_D                 18.7873      8.852      2.122      0.034       1.438      36.136
abc_class_code_group_G                104.5244      8.279     12.626      0.000      88.298     120.751
abc_class_code_group_J                187.9301      6.156     30.529      0.000     175.865     199.995
abc_class_code_group_Others            -3.0661      6.890     -0.445      0.656     -16.570      10.438
abc_class_code_group_U                190.8468      9.236     20.663      0.000     172.744     208.949
customer_district_code_group_300      -51.2707      5.080    -10.093      0.000     -61.227     -41.315
customer_district_code_group_400       14.4676      5.337      2.711      0.007       4.007      24.928
customer_district_code_group_Others    23.2009      5.109      4.541      0.000      13.188      33.214
==============================================================================
Omnibus:                  6274650.918   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.000   Jarque-Bera (JB):   51702641174769.070
Skew:                          99.728   Prob(JB):                         0.00
Kurtosis:                   28051.654   Cond. No.                         76.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

R-squared and Adjusted R-squared: Both R-squared and Adjusted R-squared are 0.038, indicating that the model explains only 3.8% of the variance in the dependent variable (value_sales). This suggests that the model does not capture much of the variability in the sales values, meaning there may be other unaccounted variables influencing value_sales.

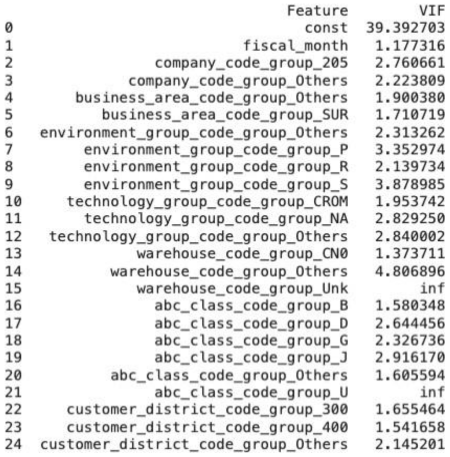
F-statistic and Prob (F-statistic): The F-statistic is 2704, with a p-value of 0.00, which indicates that the overall model is statistically significant. This implies that at least one of the independent variables is significantly related to value_sales.

Correlation with dependent variable (value_sales)

Based on the OLS regression analysis of sales value, most independent variables demonstrate statistical significance (p < 0.05), with abc_class_code_group_Others being the only exception. The analysis reveals compelling correlations, with company_code_group_205 (coef: -693.98, t-stat: -110.24), business_area_code_group_Others (coef: 587.48, t-stat: 123.85), and technology_group_code_group_NA (coef: -638.67, t-stat: -71.26) showing the strongest relationships with value_sales. This indicates that business_area_code, technology_group_code, and company_code are the primary business factors influencing sales variations.

Company_code impacts on value_sales: both have negative correlation to value_sales. Business_area impacts on profit: both have positive correlation to profit margins. Environment_group_code impacts on profit: group P and S have positive correlation, while group R and Others have negative correlation. Technology_group_code impacts on profit: only group Other has positive correlation, while group CROM and NA have negative correlation. abc_class_code impacts on profit: all group have positive correlation. Warehouse_code impacts on profit: group CN0 has a weak positive correlation, while group

Other has negative correlation. customer_district_code: group 300 has negative correlation, while group Other and group 400 have weak positive correlation.



```
                              Feature        VIF
0                               const  39.392703
1                        fiscal_month   1.177316
2              company_code_group_205   2.760661
3           company_code_group_Others   2.223809
4     business_area_code_group_Others   1.900380
5        business_area_code_group_SUR   1.710719
6   environment_group_code_group_Others 2.313262
7     environment_group_code_group_P   3.352974
8     environment_group_code_group_R   2.139734
9     environment_group_code_group_S   3.878985
10  technology_group_code_group_CROM   1.953742
11    technology_group_code_group_NA   2.829250
12 technology_group_code_group_Others  2.840002
13         warehouse_code_group_CN0   1.373711
14      warehouse_code_group_Others   4.806896
15         warehouse_code_group_Unk        inf
16           abc_class_code_group_B   1.580348
17           abc_class_code_group_D   2.644456
18           abc_class_code_group_G   2.326736
19           abc_class_code_group_J   2.916170
20      abc_class_code_group_Others   1.605594
21           abc_class_code_group_U        inf
22  customer_district_code_group_300   1.655464
23  customer_district_code_group_400   1.541658
24 customer_district_code_group_Others 2.145201
```

Robustness Checks

We used the VIF calculation to ensure minimal multicollinearity in the final model. The F-statistic (2704, p-value < 0.05) indicates that the model is statistically significant, although low R-squared values suggest that additional factors may be influencing sales.

Insights and Recommendations

The analysis of company codes reveals a consistent negative correlation with sales performance, suggesting the need for a comprehensive review of sales strategies and operational practices. In contrast, business areas show positive correlations, indicating successful market performance and potential opportunities for expansion. These findings suggest that while company-specific challenges exist, certain business areas have developed effective strategies that could be replicated across the organization.

Environmental group analysis shows mixed results, with groups P and S demonstrating positive correlations while groups R and Others show negative correlations. Similarly, technology group analysis reveals that only the 'Others' group maintains a positive correlation, while groups CROM and NA show negative correlations. This suggests a need for careful market trend analysis and potential realignment of product offerings with customer needs.

Warehouse operations show interesting insights, with CN0 showing a weak positive correlation while other warehouses demonstrate negative correlations. This variance in performance indicates an opportunity to study and replicate successful practices from CN0 across the warehouse network. ABC classification analysis reveals positive correlations across all groups, suggesting effective product categorization and potential for optimizing inventory management strategies.

Regional analysis through customer district codes shows varied performance, with group 300 showing negative correlation while groups Others and 400 demonstrate weak positive correlations. This geographic variation in performance suggests the need for customized regional strategies and potential market-specific approaches to improve sales performance across all regions.

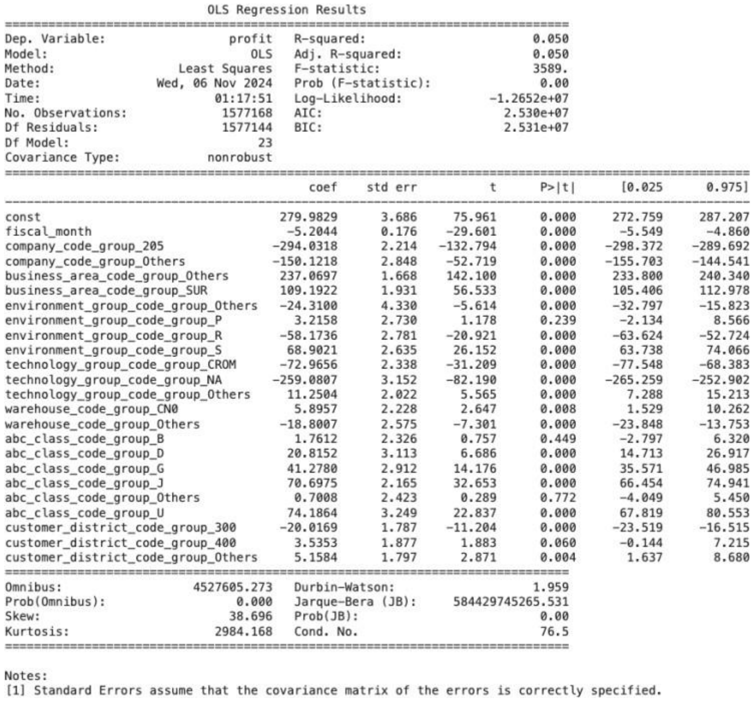
While these variables are statistically significant, the R-squared (0.038) indicates the model explains only 3.8% of the variance in value_sales. This suggests there might be other important factors not included in the model

5.2. Question 2: What business factors most significantly influence profit?

Methodology

Using Ordinary Least Squares (OLS) regression, we modeled profit as the dependent variable, with the chosen business factors as independent variables. To simplify the analysis and reduce overfitting risks, categorical values were grouped by major categories with proportions of at least 10% and mapped into "Others" groups for minor categories. After one-hot encoding these grouped categories, the Variance Inflation Factor (VIF) was used to assess multicollinearity. A final multiple regression analysis determined the significance and direction of influence for each variable on profit margins.

Model Fit and Statistical Results



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 profit   R-squared:                       0.050
Model:                            OLS   Adj. R-squared:                  0.050
Method:                 Least Squares   F-statistic:                     3589.
Date:                Wed, 06 Nov 2024   Prob (F-statistic):               0.00
Time:                        01:17:51   Log-Likelihood:            -1.2652e+07
No. Observations:             1577168   AIC:                         2.530e+07
Df Residuals:                 1577144   BIC:                         2.531e+07
Df Model:                          23
Covariance Type:            nonrobust
=======================================================================================================
                                          coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
const                                 279.9829      3.686     75.961      0.000     272.759     287.207
fiscal_month                           -5.2044      0.176    -29.601      0.000      -5.549      -4.860
company_code_group_205               -294.0318      2.214   -132.794      0.000    -298.372    -289.692
company_code_group_Others            -150.1218      2.848    -52.719      0.000    -155.703    -144.541
business_area_code_group_Others       237.0697      1.668    142.100      0.000     233.800     240.340
business_area_code_group_SUR          109.1922      1.931     56.533      0.000     105.406     112.978
environment_group_code_group_Others   -24.3100      4.330     -5.614      0.000     -32.797     -15.823
environment_group_code_group_P          3.2158      2.730      1.178      0.239      -2.134       8.566
environment_group_code_group_R        -58.1736      2.781    -20.921      0.000     -63.624     -52.724
environment_group_code_group_S         68.9021      2.635     26.152      0.000      63.738      74.066
technology_group_code_group_CROM      -72.9656      2.338    -31.209      0.000     -77.548     -68.383
technology_group_code_group_NA       -259.0807      3.152    -82.190      0.000    -265.259    -252.902
technology_group_code_group_Others     11.2504      2.022      5.565      0.000       7.288      15.213
warehouse_code_group_CN0                5.8957      2.228      2.647      0.008       1.529      10.262
warehouse_code_group_Others           -18.8007      2.575     -7.301      0.000     -23.848     -13.753
abc_class_code_group_B                  1.7612      2.326      0.757      0.449      -2.797       6.320
abc_class_code_group_D                 20.8152      3.113      6.686      0.000      14.713      26.917
abc_class_code_group_G                 41.2780      2.912     14.176      0.000      35.571      46.985
abc_class_code_group_J                 70.6975      2.165     32.653      0.000      66.454      74.941
abc_class_code_group_Others             0.7008      2.423      0.289      0.772      -4.049       5.450
abc_class_code_group_U                 74.1864      3.249     22.837      0.000      67.819      80.553
customer_district_code_group_300      -20.0169      1.787    -11.204      0.000     -23.519     -16.515
customer_district_code_group_400        3.5353      1.877      1.883      0.060      -0.144       7.215
customer_district_code_group_Others     5.1584      1.797      2.871      0.004       1.637       8.680
==============================================================================
Omnibus:                  4527605.273   Durbin-Watson:                   1.959
Prob(Omnibus):                  0.000   Jarque-Bera (JB):     584429745265.531
Skew:                          38.696   Prob(JB):                         0.00
Kurtosis:                    2984.168   Cond. No.                         76.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

R-squared and Adjusted R-squared: Both R-squared and Adjusted R-squared are 0.050, indicating that the model explains only 5.0% of the variance in the dependent variable (profit). This suggests that the model does not capture much of the variability in the profit, meaning there may be other unaccounted variables influencing profit.

F-statistic and Prob (F-statistic): The F-statistic is 3589, with a p-value of 0.00, which indicates that the overall model is statistically significant. This implies that at least one of the independent variables is significantly related to value_sales.

Correlation with dependent variable (profit)

Based on the OLS regression analysis of profit margins, most independent variables demonstrate statistical significance ($p < 0.05$), with notable exceptions being abc_class_code_group_Others, abc_class_code_group_B, environment_group_code_group_P, and customer_district_code_group_400. The analysis reveals that company_code_group_205 (coef: -294.03, t-stat: -132.79), technology_group_code_group_NA (coef: -259.08, t-stat: -82.19), and business_area_code_group_Others (coef: 237.01, t-stat: 142.10) exhibit the strongest correlations with profit margins, suggesting these are the most influential factors on profitability.

Company_code impacts on profit: both have negative correlation to profit margins. Business_area impacts on profit: both have positive correlation to profit margins. Environment_group_code impacts on profit: only group S have positive correlation, while group R and Others have negative correlation. Technology_group_code impacts on profit: only group Other has positive correlation, while group CROM and NA have negative correlation. Warehouse_code impacts on profit: group CN0 has a weak positive correlation, while group Other has negative correlation. abc_class_code impacts on profit: all group have positive correlation.

Insights and Recommendations

Company codes show a consistent negative impact on profits, indicating a need for comprehensive review of pricing strategies and operational efficiency. Conversely, business areas demonstrate positive correlations, suggesting potential opportunities for targeted investment and expansion. Technology groups show mixed results, with group 'Others' showing positive correlation while 'CROM' and 'NA' groups show negative correlations, highlighting the need for technological assessment and modernization in underperforming segments.

ABC classification analysis reveals consistently positive correlations, particularly for Class J (made-to-order) products, suggesting opportunities for inventory optimization and strategic pricing. Warehouse operations analysis shows superior performance in CN0 compared to others, indicating potential best practices that could be replicated across the network. The geographical analysis (customer district code) shows mixed performance across regions, with group 300 showing negative correlation while group Others shows weak positive correlation, suggesting the need for region-specific strategies.

While these variables are statistically significant, the R-squared (0.050) indicates the model explains only 5.0% of the variance in profit margin. This suggests there might be other important factors not included in the model

6. Sales Prediction

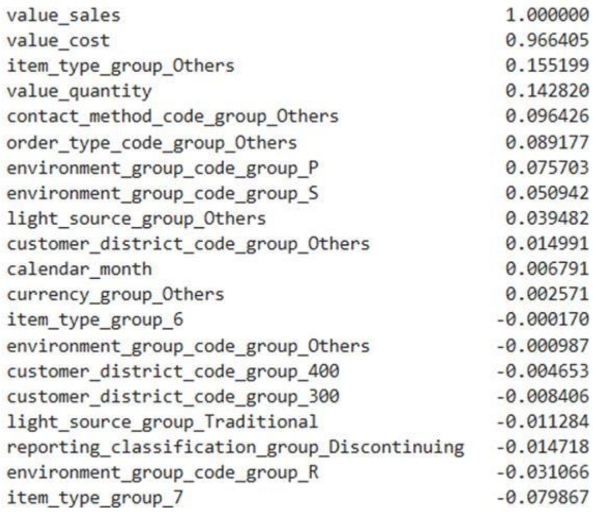
6.1. Selecting Features for Sales Prediction

For Sales Prediction, we will be selecting features that can potentially influence value_sales.

Product-related	**Transaction-related**	**Sales-related**	**Customer-related**
item_type environment_group_code reporting_classification light_source	order_type_code currency value_quantity value_cost	contact_method_code salesperson_code	customer_district_code

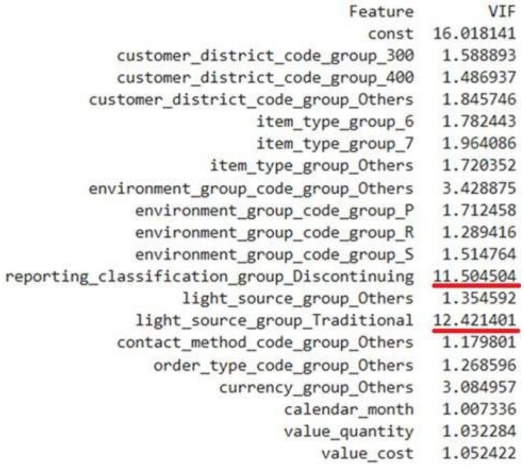
The features selected ensure that at least one (1) of data column characteristics has been included that can potentially influence value_sales (target variable).

Produced correlation matrix and checked the features that have high correlation to value_sales.

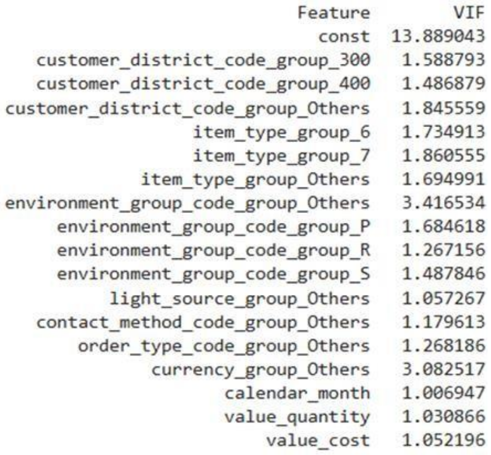


```
value_sales                                    1.000000
value_cost                                     0.966405
item_type_group_Others                         0.155199
value_quantity                                 0.142820
contact_method_code_group_Others               0.096426
order_type_code_group_Others                   0.089177
environment_group_code_group_P                 0.075703
environment_group_code_group_S                 0.050942
light_source_group_Others                      0.039482
customer_district_code_group_Others            0.014991
calendar_month                                 0.006791
currency_group_Others                          0.002571
item_type_group_6                             -0.000170
environment_group_code_group_Others           -0.000987
customer_district_code_group_400              -0.004653
customer_district_code_group_300              -0.008406
light_source_group_Traditional                -0.011284
reporting_classification_group_Discontinuing  -0.014718
environment_group_code_group_R                -0.031066
item_type_group_7                             -0.079867
```

Calculated VIF to check for multicollinearity to ensure the model's predictability performance not reduced. Reporting_classification_group_Discontinuing & light_source_group_Traditional had VIF above 5 so these will be dropped.



```
                                     Feature        VIF
                                       const  16.018141
            customer_district_code_group_300   1.588893
            customer_district_code_group_400   1.486937
         customer_district_code_group_Others   1.845746
                           item_type_group_6   1.782443
                           item_type_group_7   1.964086
                      item_type_group_Others   1.720352
         environment_group_code_group_Others   3.428875
              environment_group_code_group_P   1.712458
              environment_group_code_group_R   1.289416
              environment_group_code_group_S   1.514764
reporting_classification_group_Discontinuing  11.504504
                   light_source_group_Others   1.354592
              light_source_group_Traditional  12.421401
            contact_method_code_group_Others   1.179801
                order_type_code_group_Others   1.268596
                       currency_group_Others   3.084957
                              calendar_month   1.007336
                              value_quantity   1.032284
                                  value_cost   1.052422
```



```
                            Feature        VIF
                              const  13.889043
   customer_district_code_group_300   1.588793
   customer_district_code_group_400   1.486879
customer_district_code_group_Others   1.845559
                  item_type_group_6   1.734913
                  item_type_group_7   1.860555
             item_type_group_Others   1.694991
environment_group_code_group_Others   3.416534
     environment_group_code_group_P   1.684618
     environment_group_code_group_R   1.267156
     environment_group_code_group_S   1.487846
          light_source_group_Others   1.057267
   contact_method_code_group_Others   1.179613
       order_type_code_group_Others   1.268186
              currency_group_Others   3.082517
                     calendar_month   1.006947
                     value_quantity   1.030866
                         value_cost   1.052196
```

6.2. Linear Regression for Sales Prediction

Linear Regression provides an easy-to-understand relationship between the features and the target variable. Coefficients are easy to interpret that show how the features affect the target variables. If there is a strong correlation between a feature and target variable, this model can provide accurate predictions that can also serve as our baseline for comparison.

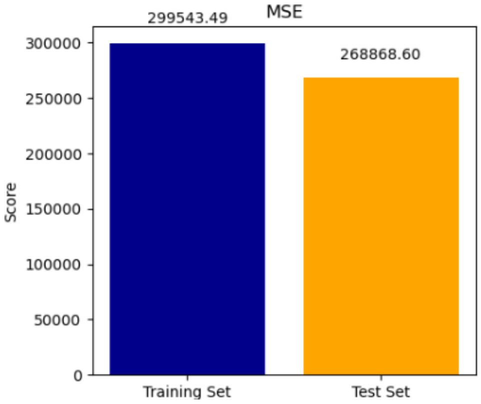
Mean Squared Error (MSE) measures the squared difference between the predicted sales value and actual sales value. As the range of our value_sales is between 0.01 to almost 800,000, results show that the model is off by 268,868.60. To further interpret this value, we can look at Root Mean Squared Error for a clearer understanding.

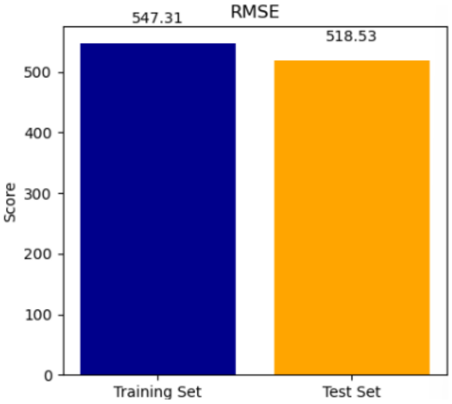
Root Mean Squared Error is simply the square root of MSE that represents the average prediction error of the model. A score of 518.53 suggests that the average deviation between the predicted sales value and actual sales values is about 518.53. As this value is relatively low, this indicates that the model has performed well by keeping the prediction errors minor.

R-Squared indicates how the model was able to explain the variance in value_sales. This means that the model was able to explain 93.3% of the variability in value_sales, successfully capturing the relationship between the features and value_sales.

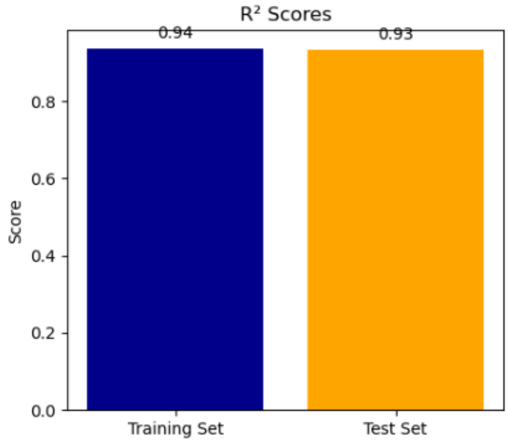
Mean Absolute Error is the average absolute difference between predicted sales value and actual sales value that is less sensitive to outliers as opposed to MSE. This suggests that the model's predictions deviate from the actual values by around 102.93 units.

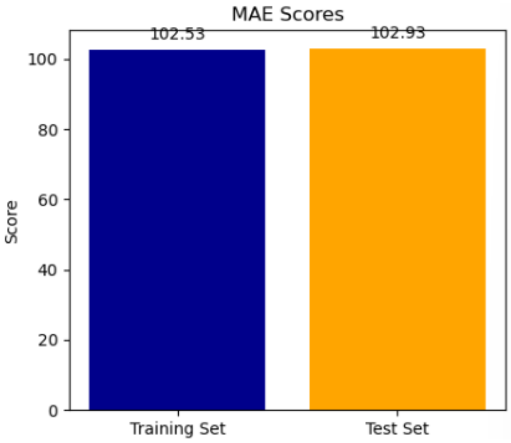
As the model had already produced good results, this raises a question that the model could potentially be overfitting. To verify, we examine the scores on the training set to ensure consistency of results.





From the results shown in MSE and RMSE bar charts, the test set scores are slightly lower than the training scores, suggesting that the model generalizes well to unseen data. The slight difference between the two sets means that the model is neither overfitting nor underfitting. The model performed well with minimal errors considering the wide range of values in value_sales.





The R2 scores and MAE scores show almost similar results for both testing and training set. This suggests that the model is performing well on generalizing on unseen data. The model is both accurate and interpretable, producing reliable predictions with minimal error relative to the scale of value_sales.

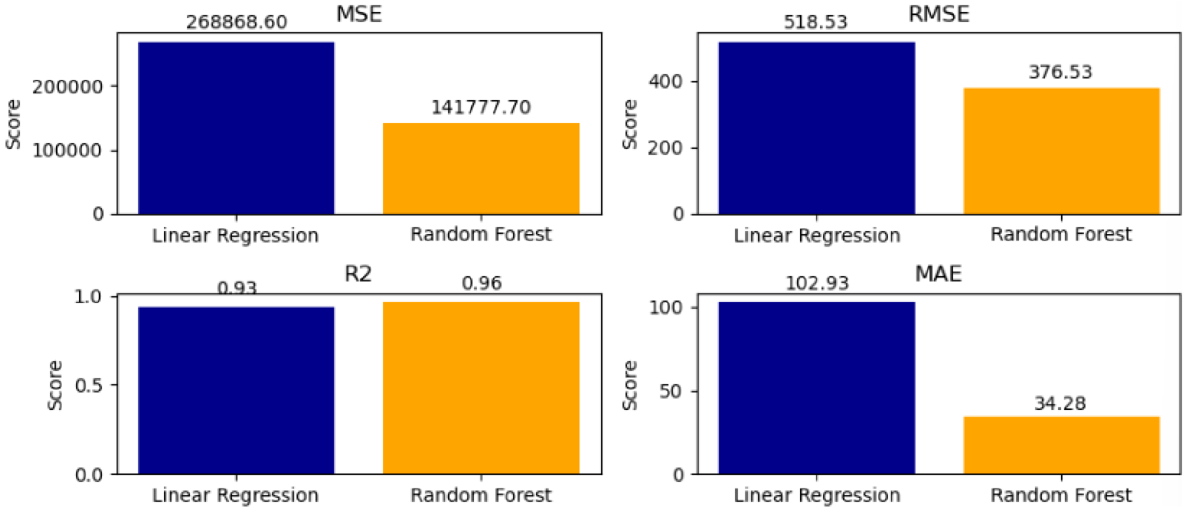
The linear regression model produced good scores in evaluation metrics indicating strong performance, reliability and accuracy due to low average prediction error, captured most of the variance of values_sales, demonstrated a strong fit and minimal absolute error on average. Consistent results across testing and training set, confirming the model is not overfitting and generalizes well to unseen data.

Random Forest for Sales Prediction

Random forest is a flexible and powerful model that can prevent overfitting, capture complex relationships and provide insights into feature importance which makes it suitable for value_sales prediction.

Overall Insight:

The Random Forest evaluation scores suggest that this model has also performed well similar to the Linear Regression Model.



The Mean Squared Error of 141,777.70 is lower compared to the previous model's MSE, suggesting a significant reduction in the prediction error.

The average prediction error is around 376.53 units, also a significant improvement compared to the previous mode's RMSE, suggesting that this model has more accurate predictions.

Random Forest was able to explain 96% of the variance in value_sales, more variability captured, another improvement compared to 93% from the previous model.

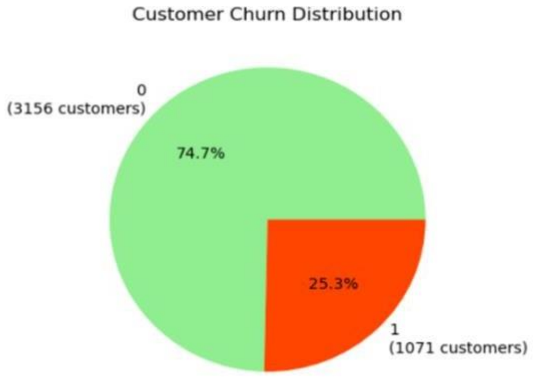
Lastly, a much lower score of MAE with 34.28. A significant improvement as well compared to previous model's MAE of 102.93, suggesting that this model's predictions are only off by around 34 units.

Recommendation:

As the Random Forest outperformed the linear regression model across all metrics, it is recommended for the management to use Random Forest model for sales prediction for 2014. With its lower MSE, RMSE and MAE scores and a higher R squared, this means that the model has provided a more accurate prediction of sales due to its ability to reduce overfitting and capturing complex relationships, making it more reliable.

7. Customer Churn Analysis

In continuation of retention analysis, this section focuses on the customer churn. Churned customers were identified by setting a threshold of 180 days of no activity with the business. Other factors/features that contribute to customer churn will be identified in this section.



The chart above shows that out of 4227 customers, 3156 or 74.7% customers are still doing business with the company while 1071 or 25.3% customers are no longer engaged with the company.

Selected features that can potentially impact our customer churn by having at least 1 characteristics of data columns:

Product-related	Transaction-related	Sales-related	Customer-related
item_type environment_group_code reporting_classification light_source	order_type_code currency time_gap_segment value_sales value_quantity value_cost time_gap	contact_method_code salesperson_code	customer_district_code

7.1. Logistic Regression for Customer Churn Analysis

Logistic Regression is easy to interpret that helps in understanding the relationship between features and customer churn. Logistic regression can also help in identifying which features are most impactful in predicting customer churn. Above all, this model can serve as a baseline model for comparison in case improvement is required.

Logistic Regression model is prepared by (1) Labeling major and minor categorical values, (2) Selecting the features, (3) Performing one-hot encoding for categorical features, (4) Combining numerical and categorical features in one data frame, (5) Calculating VIF, (6) Removing VIF with > 5 scores, (7) Checking new VIF scores, and (8) Splitting the data into training and testing sets.

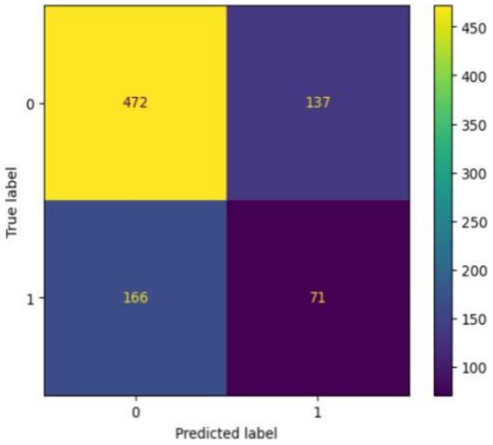
As the dataset is imbalanced, where non-customer churners are more than 50% compared to customer churners, we have used a parameter class_weight ='balanced' to ensure that the model is not learning/analyzing the non-churners only but also the customers who have churned. Solver saga has also been used as this can handle large datasets, provides faster results and is good for regularization that prevents overfitting.

With these parameters, the model showed an accuracy score of 57.45%. This suggests that the model was only able to correctly predict a little bit of more than half of customers who will churn. While the F1 score suggests that 35.48% of the predictions are effective in balancing precision and recall. This means that the model still struggles to accurately identify customer churners.

7.2. SMOTE

Synthetic Minority Oversampling Technique or SMOTE is commonly used to improve the model's performance in situations where the dataset is imbalanced by enhancing the minority class (number of churned customers). After applying SMOTE, the accuracy has improved to 64.85% indicating that the model is now making more correct predictions. However, a drop in F1 score has also been observed to 31.91%. This suggests that while the model is now better at predicting correctly, it struggles in balancing between the precision and recall, possibly predicting more false positives or failing to identify true positives. Let's see the numbers in the confusion matrix to gain more understanding.

Confusion Matrix



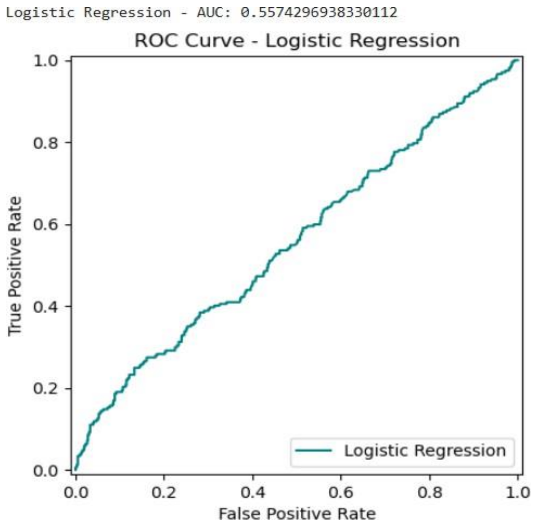
Here's what the confusion matrix has shown:

- True Negative (472) - Non-churn customers CORRECTLY predicted as non-churn
- False Positive (137) - Non-churn customers INCORRECTLY predicted as CHURN
- False Negative (166) - Churn customers INCORRECTLY predicted as NON-CHURN
- True Negative (71) - Churn customers CORRECTLY predicted as churn

This suggests a high rate of false positives where many non-churners are incorrectly labeled as churned but the identification of actual churn is relatively good but there is still a room for improvement.

7.3. ROC AUC

The ROC or Receiver Operating Characteristic Curve shows the true positive rate against the false positive rate. This will only be considered good if the curve moves towards the top-left corner of graph, suggesting a high recall with low positive rate. AUC or Area Under the Curve summarizes the ROC Curve. A score of 1 is a perfect model and a score 0.5 suggests that the model has no predictive power. With a score of 55% suggests that the model is only slightly better at guessing prediction.


```
Logistic Regression - AUC: 0.5574296938330112
```

As improvement is still required despite feature engineering, parameters tuning and SMOTE, we can also use Random Forest. Random Forest is also used for imbalanced dataset where we can also use class weight and smote to further address imbalance and be more effective in capturing the minority class patterns.

Random Forest for Customer Churn Analysis

At the initial scores of Random Forest, it seems that the model is correctly predicting 63%, an improvement compared to logistic regression. However, F1 score is still low with 32.40% indicating that the model is still struggling finding the balance between the churned customers and non-churner customers despite the SMOTE application.

Another improvement technique can be implemented such as Grid Search Cross-Validation, which is also used for random forest to fine=tune hyperparameters and find the best combination of settings to improve the model's performance.

GridSearch CV results show that the best parameters to use in random forest are:

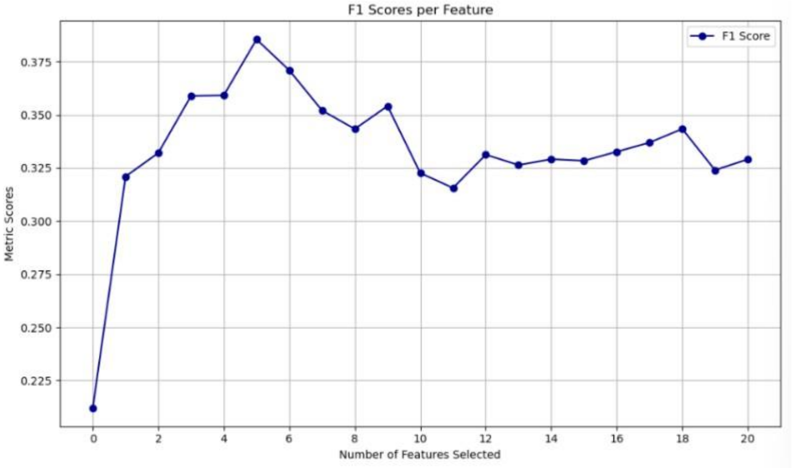
min_estimators: 300	max_depth: 20.	min_samples_split: 2
min_samples_leaf: 1	max_features: sqrt.	class_weight: balanced

After applying the best parameters, there was a slight decrease on the accuracy score with 62.88%, indicating that the model is now less correct in classifying a slightly lower proportion of churned and non-churned customers overall. A slight improvement in F1 score with 32.90%, suggesting that it now a little bit more effective at handling actual churners and minimizing false positives.

As the dataset is imbalanced, focusing on getting the highest F1 score is a must to ensure a great balance between precision and recall as achieving high accuracy score can be misleading to the model's effectiveness at identifying the churned customers.

7.4. Recursive Feature Elimination

Recursive Feature Elimination (RFE) identifies which features and the number of features are needed to optimize the F1 score of our model. The line chart shows that the number of features to select with the highest F1 score is 6 with 38.54%. While this is the optimal F1 score, this suggests that the model is slightly effective at predicting the churned customers and still struggle with false positives or false negatives.

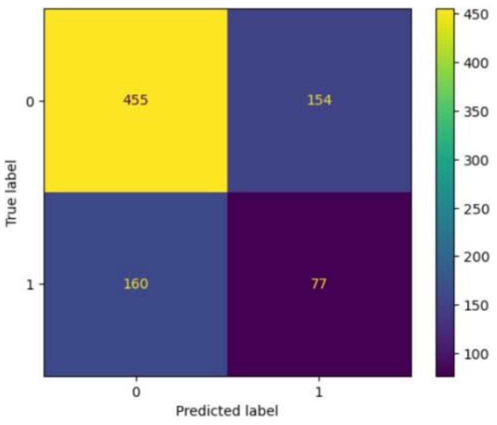


Confusion Matrix

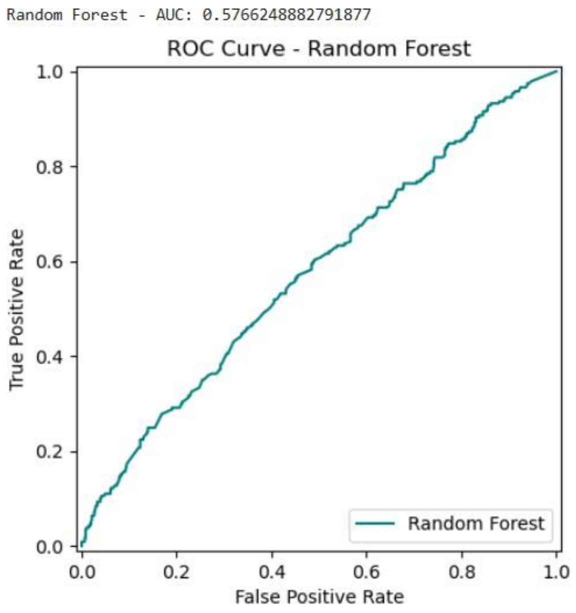
The new confusion matrix shows:

- True Negative (455) - Non-churn customers CORRECTLY predicted as non-churn

- False Positive (154) - Non-churn customers INCORRECTLY predicted as CHURN
- False Negative (160) - Churn customers INCORRECTLY predicted as NON-CHURN
- True Negative (7) - Churn customers CORRECTLY predicted as churn



The model's accuracy is correctly predicting 62.88% of cases overall. The F1 score of 38.54% suggests a moderate balance between precision and recall. The model is correct 33% of the time at predicting churned customers. The recall score suggests that the model captures 32.48% of actual churners but fails to classify a great portion.



```
Random Forest - AUC: 0.5766248882791877
```

Random Forest's AUC score of 57.66% suggests that this model is still not better at distinguishing churn and non-churners compared to logistic regression model. The curve is still not close to the top-left corner, which means no good balance between true positives and false positives.

Although the scores have not significantly improved, it is clear that random forest has performed better than logistic regression. Random Forest has provided a better balance between correct identification of churners while minimizing false positives.

Overall Insight:

At the start of this analysis, we had 23 features selected that could potentially influence customer churn. By performing VIF and RFE, we have narrowed down the features into 7 features namely:

Product-related	Transaction-related
item_type_group_7 environment_group_code_group_P environment_group_code_group_S light_source_group_LED	order_type_code_group_Others value_quantity time_gap

Majority of the features affecting customer-churn are product-related and transaction-related concerns. This means that customer churn is heavily influenced by the characteristics of products being offered to customers.

Recommendation:

By focusing on product-related and transaction-related features, the company may be able to address customer churning and improve customer retention rates through enhanced product quality, having more variety of products, ensuring availability and streamlining the process of customer ordering and product delivery. These strategies can help improve customer satisfaction and loyalty. Moreover, building stronger relationships with at-risk customers can be done, gathering customer feedback and continuous monitoring of the features or churn metrics can also provide valuable insights should there be any refinement needed for business strategies.

8. Conclusion

This report analyzes LuminaTech Lighting's sales, customer retention, and profitability based on various business factors. Through data cleaning, exploration, testing, inferencing, predictive modeling, and churn analysis, we've identified trends to inform the strategic decisions.

Key areas for your team to consider:

1. **Sales Growth and Market Risks**: AUD is the company's main revenue source and shows steady growth. However, reliance on a single market increases risk. Expanding in NZD, USD, and EUR markets could reduce this dependency.
2. **Pricing Adjustments by Region**: Currency impacts profitability differently across regions. Tailoring pricing strategies for Oceania, North America, and Europe may help boost margins.
3. **Customer Retention and Churn**: Retention rates are dropping, with high-margin customers among those churning. Focusing on engagement, pricing strategies, and product variety could help retain these valuable customers.
4. **Factors Impacting Sales and Profit**: Business area, company code, and technology group affect sales and profit in different ways. Invest in high-performing areas and address challenges in low-performing segments.
5. **Sales Prediction Model**: The Random Forest model effectively predicts sales, helping LuminaTech Lighting plan future revenue and resource needs.
6. **Churn Insights**: Product features and transaction factors like order type and delivery impact churn. Addressing product quality, variety, and delivery speed could improve retention.

Recommendations:

- Expand efforts beyond the AUD market to diversify revenue.
- Align pricing with regional demands for competitive advantage.
- Focus retention strategies on high-margin customers.
- Improve performance in underperforming business areas.
- Use the Random Forest model for accurate sales forecasting.
- Enhance product and delivery processes to lower churn.