MACQUARIE University BUSINESS SCHOOL

SYDNEY · AUSTRALIA

Techniques of Business Analytics

Group Assignment

Semester 2 (2024)

Unit Convenor : Dr Narges Safari

Aakansh Kulyana(48157309)

Bhavesh Hemant Vasnani(48218294)

Bibek Adhikari(48489093)

Word Count:5300

Table of Contents

Report of Key Findings

This report aims to provide meaningful insights, through visualisations and modelling, performance and recommendations particularly in the areas of sales, customer demographics, and product inventory from the provided data of Lumina Tech Lighting (Australia). It is addressed to the management team of the company to help them make right decisions to maximise the companies value. It also discusses data cleaning processes undertaken to extract meaningful visualisation and statistical analyses to identify valuable insights.

Exploratory Data Analysis

Approach to Data Cleaning: Targeted, Question-Driven Process

Given the substantial volume of data, we adopted a reverse approach to data cleaning, focusing on efficiency and relevance. Instead of initiating a broad, preliminary data cleaning phase which is the usual way, we first identified the key questions and insights we aimed to address. We were careful on drafting the questions, as they would have to present valuable and useful insight about the company performance. After drafting the questions, we were able to target specific data subsets and relevant attributes important to answering their questions. This selective cleaning process enables us to prioritize accuracy and completeness in the most impactful areas, reducing unnecessary processing of unrelated data. This approach streamlined the overall data preparation, enhancing both the speed and precision of our analysis by ensuring that only relevant data underwent cleaning and transformation.

Figuring out the data

After loading the files(2012 and 2013) in the Jupyter lab, functions such as shape and describe were used to understand the overview of the data. By doing this, we understood general sense about the data such as column headers. Because the data was huge, we had to refer to meta data to further enhance our understanding.

Targeted columns for data cleaning

As mentioned earlier, given the volume of data, we will clean the columns that are pertinent to answer the questions that we have set to answer. Therefore, the important columns for data cleaning are value_sales, value_cost, value_quantitiy, customer_code, item_source_class, technology_group_code, customer_disctrict_code, currency, business_area_code, environment_group_code, abc_class_volume, invoice_date, and order_date.

Missing values

Checking for missing values is essential in data analytics because missing or incomplete data can impact the accuracy and reliability of analytical results. Therefore, to maintain data quality, improve model performance, and prepare for downstream analysis, it is important to ensure there are no missing values. In our analysis, we found that the column of item_source_class had no values. This task was performed on data from both 2012 and 2013.

Utilising the unique Function for Effective Data Cleaning

Unique function was used to identify distinct values and ensuring consistency in data cleaning tasks. It is important to detect variations in data entries, like spelling variations, abbreviations, or capitalization differences, as it can lead to inaccuracies in the analysis. Inconsistent data can skew results and reduce the accuracy of models and analyses. Whitespace were observed in the columns such as technology_group_code, business_area_code, and environment_group_code, and spelling errors were present in the currency column. Identified issues were fixed. This task was performed on data from both 2012 and 2013.

Dealing with Empty rows

Dealing with empty rows is essential in data cleaning because they can significantly affect the quality, accuracy, and performance of data analysis or modelling. It is important to handle empty rows as doing so will prevent inaccurate analysis, improve model performance and ensures consistent data structure. This task was performed on data from both 2012 and 2013.

Checking for duplicate rows

Checking for duplicate rows is essential in data cleaning because duplicates can distort analytical results, leading to inaccurate insights and potentially costly errors. Therefore, it is important to deal with it; if left unaddressed, duplicates can skew metrics which can lead to biased or misleading findings. Duplicate rows were checked based on invoice number, as it is possible for other things to repeat.

Conversion of date

Date conversion is a critical step in data analysis because it ensures that date and time values are in a consistent, usable format, enabling accurate analysis and meaningful insights. In this case, raw data was in accounting format, making it difficult for analysis in python. Converting dates into a standard format allows us to measure trend over time and other useful visualisations. For the analysis, invoice_date and order_date were converted to YY/MM/DD format.

Outliers Treatment

Outlier treatment is one of the most important parts of EDA because outliers can significantly impact the quality, reliability, and interpretability of insights drawn from the data.

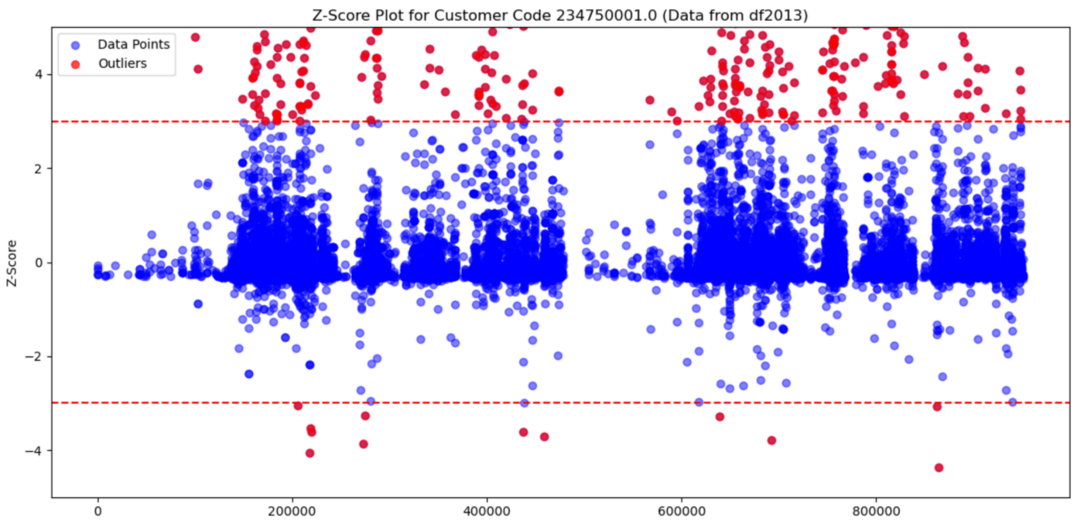
To deal with outliers, we went through each customer order and identified outliers as on a per-customer basis, as it is a detailed and personalised analysis of each customer's transaction history to detect any unusual or abnormal values in their orders. This approach allows for a granular examination of each customers purchasing behaviour, making it possible to detect specific anomalies that could indicate error, fraud, or unique purchasing patterns.

Z-score method was used for outlier detection and removal because it provides a straightforward, statistically sound way to identify values that significantly deviate from the mean of a dataset.

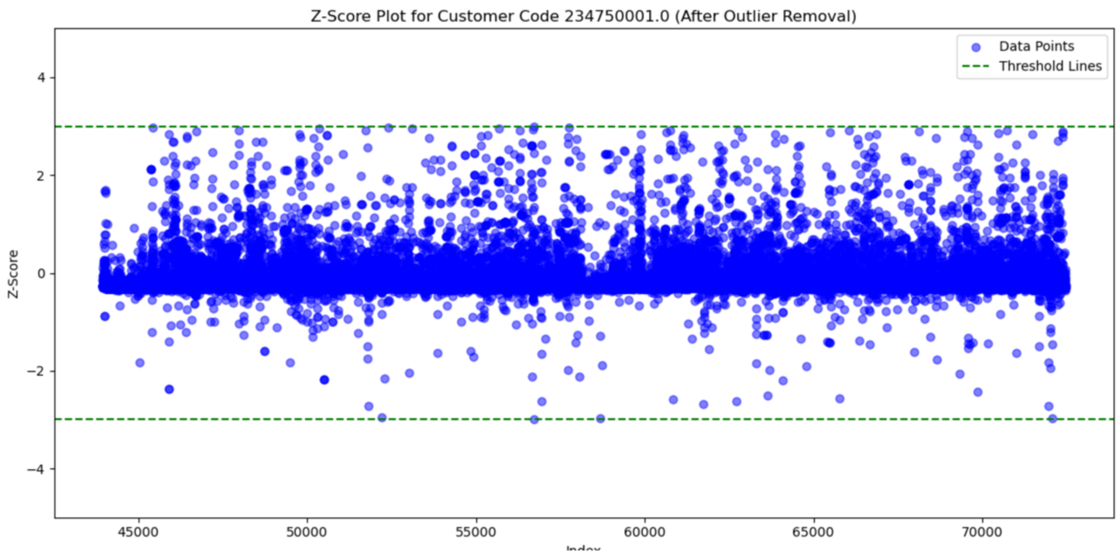
Minor adjustments had to be made to ensure the code ran smoothly. We converted customer_code values in the filter to float to ensure matching, as the customer codes in the data file appeared in floating-point format. This avoids mismatches when filtering by customer code. Also, we added a check for the value_sales column to handle cases where it might be missing or contain all null values for certain customer codes. If value_sales data were absent for any selected customer, the code would display a message in the plot instead of attempting calculations in missing data. These changes ensured the code was compatible with the dataset structure while maintaining the same logic for Z-score calculations and plotting.

Outliers were removed as the data had massive volume.

Following is an example of outlier removal for customer code 234750001.0



With outliers 1



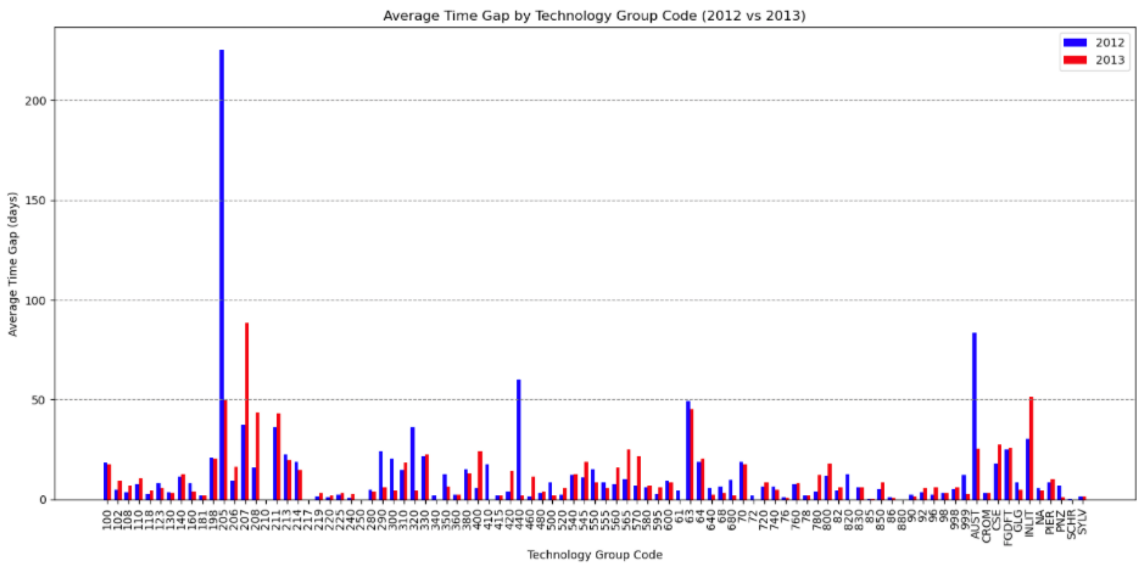
Without Outliers 1

Transformation of data

Transforming data is crucial in data analysis because it prepares the data for more accurate and meaningful interpretation, improving the effectiveness of analytical model and visualisations. Skewness for all the numerical data columns were checked and transformed when necessary. When the distribution's skewness is greater than 1, it is not useful for effective analytical models; therefore, such data columns were transformed.

Visualisations

1.Average time Gap by Technology Group Code



Average Time Gap: The average time gap is the difference between the invoice date and the order date for each group. This metric represents the typical duration between when an order is placed and when it is invoiced, averaged across all instances in each technology category. It is important to management as they need to increase efficiency to increase the profitability.

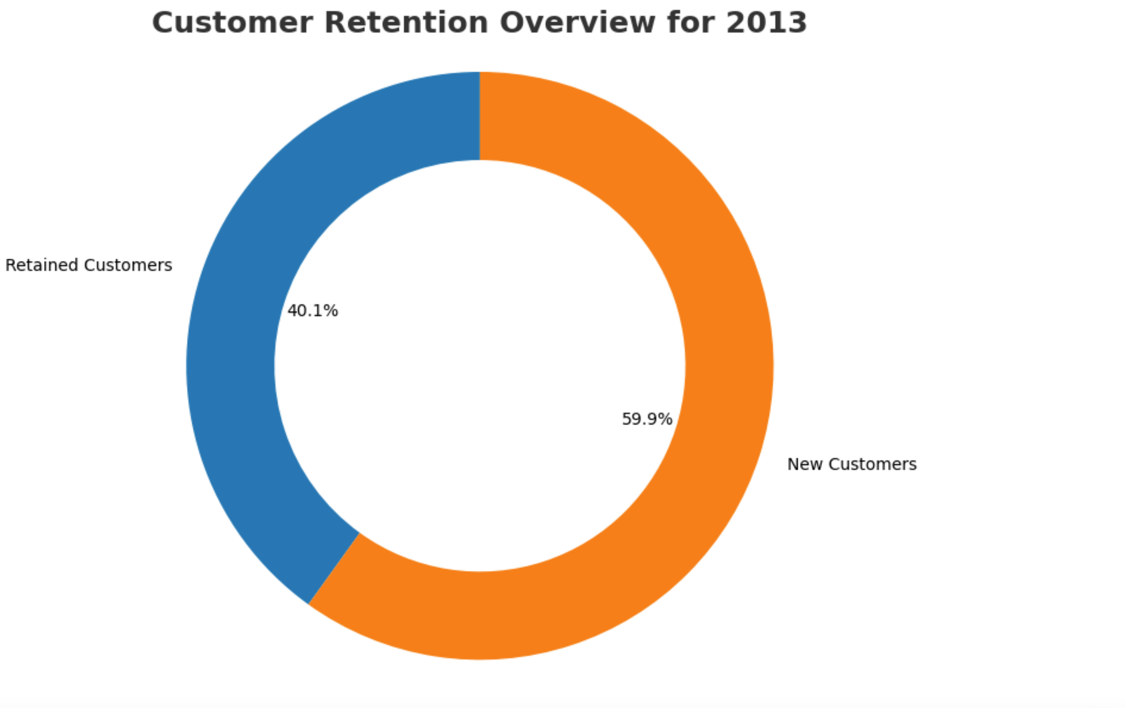
Technology Group Code: It is a code that identifies the specific technology category associated with each item. It serves to classify and group products or services based on their technology type, making it ready for segmented analysis by category.

Clustered Bar Graph is a useful tool for comparing average time gap for each technology group code for two years(2012 and 2013) side-by-side for easy comparison.

Insights:

- Bar Graph suggests noticeable variation in the average time gap across different technology group codes.
- Technology group code 205 saw a tall spike in 2012, suggesting possible wide difference between order date and invoice date. However, the reason can range from inefficiency to unavailability of the product from suppliers.
- Based on the comparison between the average time gaps for 2012 and 2013, we can see decrease in time gaps across most categories, suggesting improved efficiency. However, more ratios need to be looked at to draw a meaningful conclusion.

2.Customer Retention Rate of 2013



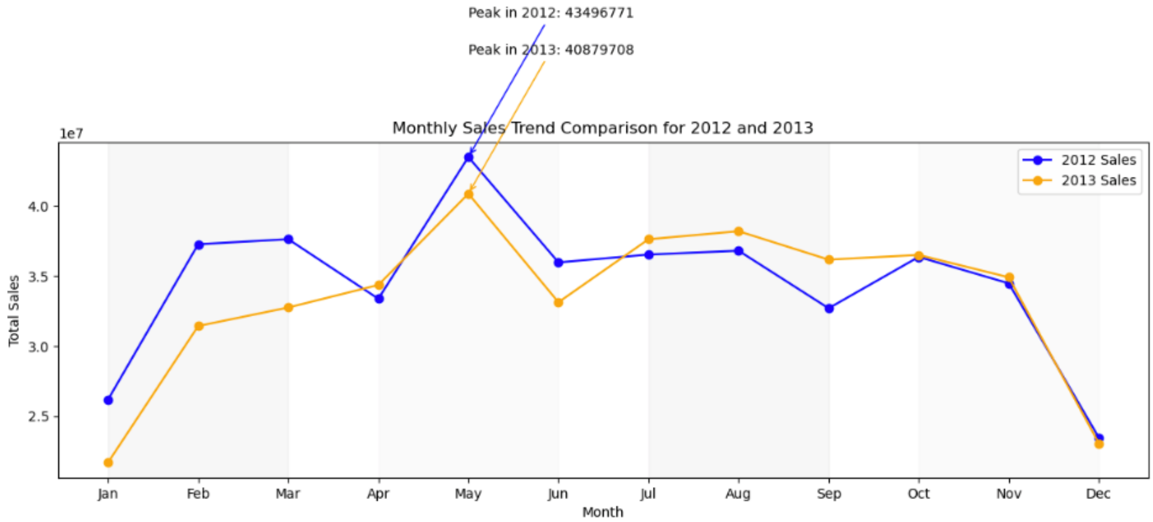
Customer Retention Rate is a metric that measures the percentage of customers a business retains over a specific period.

It is important because high retention means customers are satisfied and likely to return, which is important for long-term success. Furthermore, it is generally cheaper to retain existing customers than to acquire new ones; therefore, high retention can reduce marketing and acquisition cost. High Retention also indicates revenue growth, as future efforts will add to the present base of the customers.

Customer Retention Rate for 2013 for Lumina Tech Lighting is 40.1%, suggesting 40.1% of the customers in 2012 ordered from Lumina Tech Lighting in 2013 as well.

To find out if the retention rate is good or bad, it is important to compare the retention rate with industry benchmark. Data shows customer retention rate for B2B companies in similar sectors like manufacturing, electronics, or building materials vary but are often around 67%, meaning Lumina's retention rate would likely fall below average for this industry(Zippia, 2023).

3.Monthly Sales Trend Comparison of 2012 and 2013



Line graph illustrating monthly sales trends for 2012 and 2013 represents total sales per month for each year, allowing for a clear comparison on monthly performance of the firm.

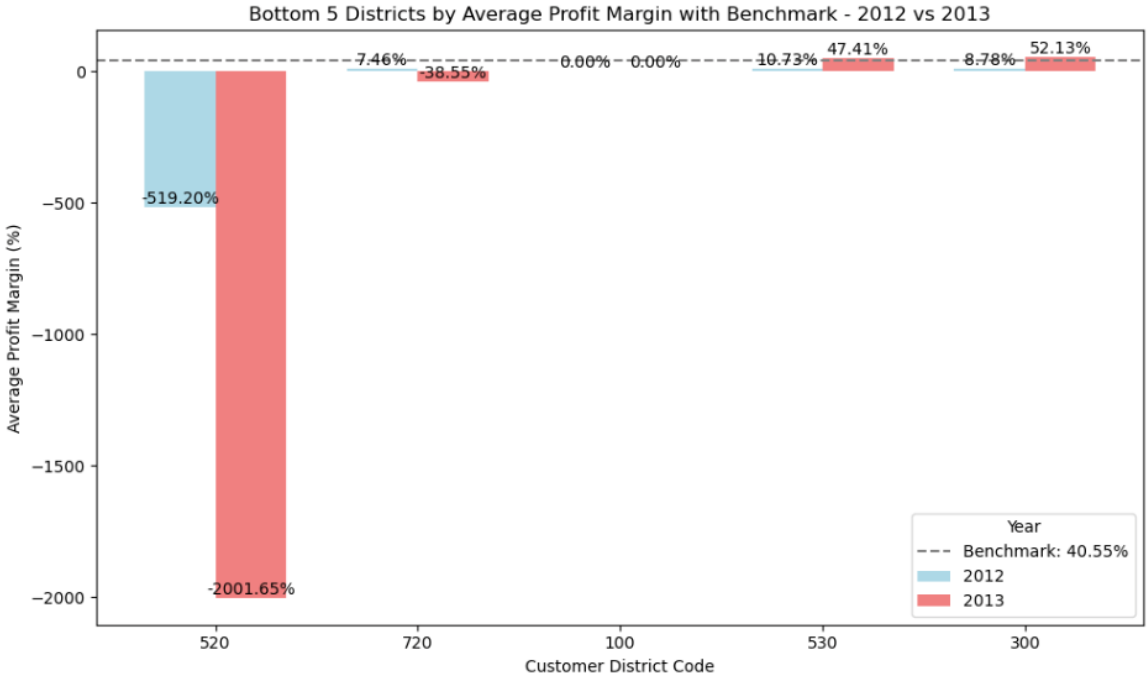
This graph is valuable to management as it helps identify seasonal patterns, allowing for improved planning and resource allocation. Be looking which months trend to perform better or worse, management can better prepare for demand fluctuations, manage inventory, and adjust staffing levels are necessary. Additionally, year-over-year comparisons enables management to evaluate the effectiveness of strategies and campaigns, revealing any growth or decline in sales performance.

The graph provides several key insights for the years 2012 and 2013. Both years reached peak sales in June, with 2012 reaching approximately 43,496,771 and 2013 peak slightly lower at around 40,879,708. This suggest that June is consistently a high-demand month, indicating a seasonal trend. Both years also show significant drop in sales in December, which might be due to seasonal factors or budget cycles. Notably, 2013 lags slightly behind 2012 in most months, suggesting a decline in overall performance of the firm.

Taking pointers from previous visualisations about unimproved efficiency, low retention and adding the decline of sales, these indicators point to a need for the company to reassess its strategies in customer relationship management, operational improvements, and sales initiatives to reverser these trends and strengthen its profitability.

4.Bottom 5 districts by Average profit margin

Tracking performance of district relative to the average profit margin of the firm is crucial for identifying areas that need improvement, optimizing resource allocation, and maintaining competitiveness. By pinpointing underperforming districts, management can make targeted interventions, while high-performing areas can receive further support to drive further growth. This monitoring allows for more strategic resource distribution, goal setting, and long-term planning, ensuring that each district is aligned with company’s average profit margin benchmark.



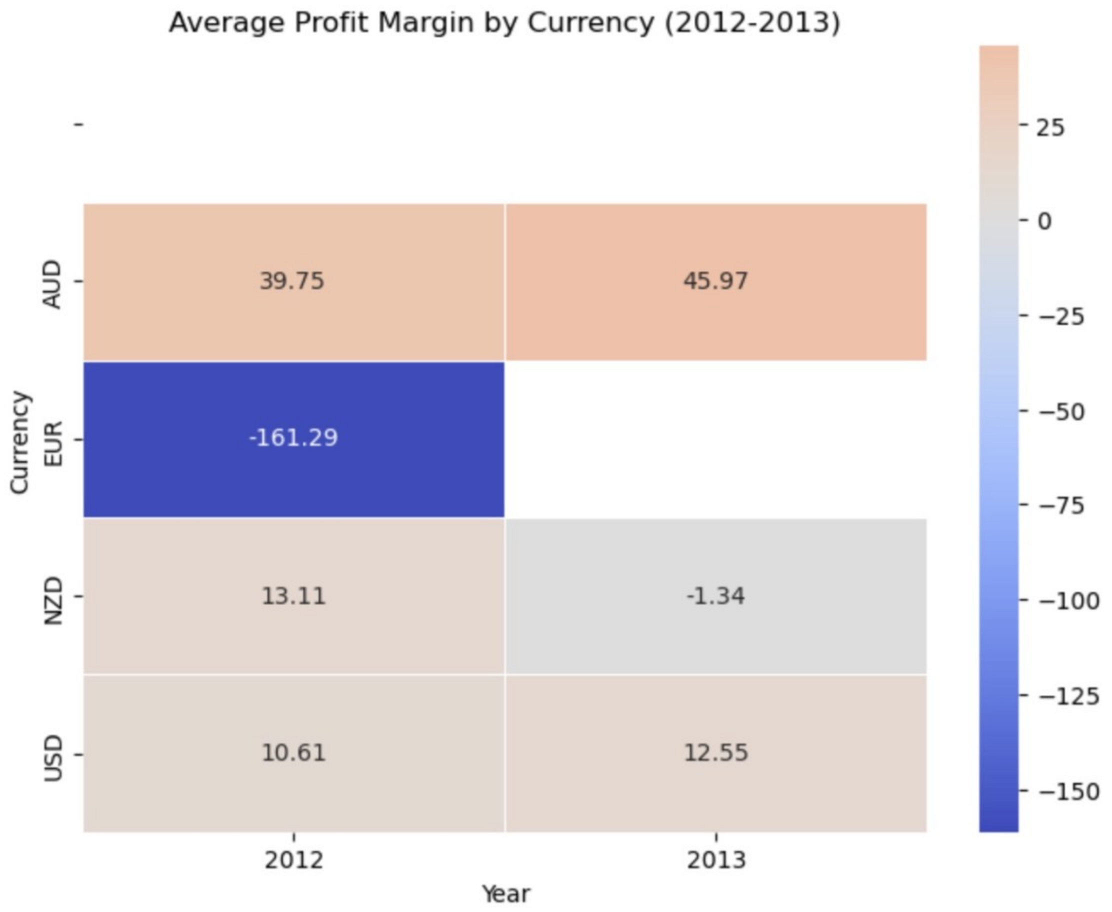
The chart displays the bottom 5 districts by average profit margin for the years 2012 and 2013, comparing each district's performance to a benchmark profit margin of 40.55%. The districts shown are 520 (Inlite - NZ), 720 (Intercompany Sales), 530 (South Island - NZ), and 310 (Tasmania).

The chart reveals that District 520 (Inlite - NZ) faces critical financial challenges, with an alarming drop in profit margin in 2013. District 720 (Intercompany Sales) and District 530 (South Island - NZ) also struggle with negative profit margins, though to a lesser extent. In contrast, District 310 (Tasmania) has shown a significant positive shift, surpassing the benchmark in 2013. These insights highlight areas that need intervention to improve profitability, particularly in New Zealand, while Tasmania's performance may serve as a model for other districts.

5.Average Profit Margin by Currency (2012-2013)

This heat map illustrates the average profit margin by currency for transactions conducted in AUD (Australian Dollar), EUR (Euro), NZD (New Zealand Dollar), and USD (United States

Dollar) over the years 2012 and 2013.



This chart is important to management because it reveals profitability across various currencies for a company operating primarily from its branches in Australia and New Zealand. By showing average profit margins for transactions in AUD, NZD, EUR, and USD, management gains insights into the financial performance of its international transactions. Even though the company doesn't have physical branches in Europe or the United States, it is still involved in transactions in EUR and USD, possibly through exports, international suppliers, or online sales. Understanding the profitability of these transactions is crucial for evaluating the viability and financial impact of maintaining international customer and supplier relationships.

In 2012, the effect of different currencies on profit margins was quite varied. The Australian Dollar (AUD) and New Zealand Dollar (NZD) both contributed positively, with AUD at 39.75% and NZD at 13.11%. In contrast, the Euro (EUR)had a strong negative impact, with a profit margin of -161.29%, indicating significant losses. The US Dollar (USD) showed a smaller positive margin at 10.61%. Moving to 2013, AUD improved further to 45.97%, suggesting stronger profitability, and USD saw a slight increase to 12.55%. However, NZD declined to -1.34%, marking a shift to a negative margin. Data for EUR in 2013 is not available, so its performance that year remains unclear. Overall, AUD and USD saw gains, while NZD's impact worsened, and EUR had a large loss in 2012 with no additional data for comparison, suggesting closure of EUR operation.

Test Sub Sample Differences

Question 1: Has the Profit Margin Changed over the year?

Profit margin is a financial metric that shows the percentage of revenue that remains as profit after all expenses have been deducted.

It is important for managers to understand the changes in profit margin over the year. A positive shift in profit margin indicates that there has been increase in operational efficiency whereas negative shift in profit margin indicates reduced operational efficiency. Furthermore, positive change in profit margin means strategies and campaigns ran by managers are delivering results as expectations and vice versa.

To find out if the average profit margin has changed from one year to another(increased) in this case, a two-sample independent t-test is an appropriate statistical test to use.

To use two-sample t-test, following assumptions must be met:

- Two samples must be independent of each other.
- There should be sufficiently large number of observations in each year, the two-sample t-test can be applied even if the data isn’t perfectly normal.

Null Hypothesis: The mean profit margin for the two years in the same(no significant change).

Alternative Hypothesis: The mean profit margin for the two years is different (there has been a significant change).

Outcome: With the t-statistics of -2.28 and p-value of 0.022, we can reject the null hypothesis and conclude that there is significant difference between profit margin between 2012 and 2013.

Conclusion: While the sales has decreased in 2013, increase in profit margin could indicate operational efficiency and successful pricing strategy. This also indicates that the company has been focused on cost-cutting measures, optimized operations, or improved pricing for higher profitability per unit sold.

Question 2: Has there been a change in customer Discount?

Although discount might have small impact on profit margin, it can be used for many other things such as understanding customer behaviour, evaluating promotional effectiveness and forecasting and budgeting.

Monitoring discount trends helps managers understand if customers are becoming reliant on discounts to make purchases. If discount rates increase over time, it may indicate that customers are less willing to pay full price, which can affect pricing strategies and customer perceptions of value.

Similarly, to find out the difference between customer discount in two years 2012 and 2013, we will perform independent t-test as above.

Null Hypothesis: The mean average discount for the two years in the same(no significant change).

Alternative Hypothesis: The average discount for the two years is different (there has been a significant change).

Outcome: With a t-statistics of 13.42 and p-value of near to zero, we can reject the null-hypothesis and conclude that the average discount has changed over the year.

Conclusion: This change in discount rates could have multiple strategic implications. An increase in discounts might suggest that customers are increasingly reliant on discounts, potentially lowering their willingness to pay full price. This trend could influence the company’s pricing strategies and customer value perception. Additionally, it highlights the need to monitor discount dependency and assess whether customers are making purchasing decisions based primarily on discounts. This insight could inform future budgeting and forecasting, ensuring that promotional strategies align with customer behaviour patterns and the company's profitability goals.

Inference

1.Analysis of cost per item based on business area

Importance of the insight: Understanding the cost per item across different business areas is essential for companies aiming to optimize their operations and profitability. By pinpointing how much it costs to produce or acquire each unit within distinct sections of the business, companies can identify areas where expenses are higher than expected and explore underlying causes. This level of insight enables more precise adjustments to lower costs, secure more favorable terms, or enhance operational efficiency. In turn, a clear grasp of item-specific costs by area helps shape effective pricing strategies, supports profitability goals, and highlights which parts of the business contribute most effectively to financial success.

Method used: To analyse cost per item based on business area, we use Multiple linear regression. It is a statistical method used to predict the value of dependent variable based on two or more independent variables. In this case, we use multiple linear regression to find the correlation between independent variables and a dependent variable. For this model, independent variables are region and warehouse and the dependent variable is lead time.

Steps taken during the regression: Following steps were taken before the regression:

- One-Hot coding of business_area_code, to include categorical variables like business_area_code, in a regression, we need to convert them into numerical values.
- After one-hot coding, the new columns were converted to integers to ensure they’re in numeric format which is required for regression.

Result of Multiple linear regression:

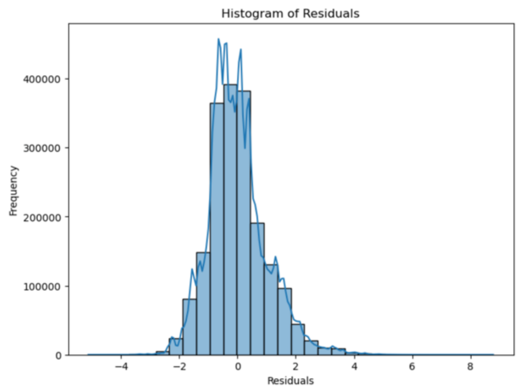
The regression output reveals that numerous business areas significantly impact cost_per_item_log, as indicated by the low p-values (mostly under 0.05). The coefficients show how each business area influences the cost, with positive values pointing to higher costs and negative values indicating lower costs relative to the base category. The R-squared value of 0.483 suggests that nearly 48.3% of the variation in cost_per_item_log can be explained by differences across business areas. When coefficients are statistically significant, it suggests that there’s a meaningful relationship between independent variables and the dependent variable.

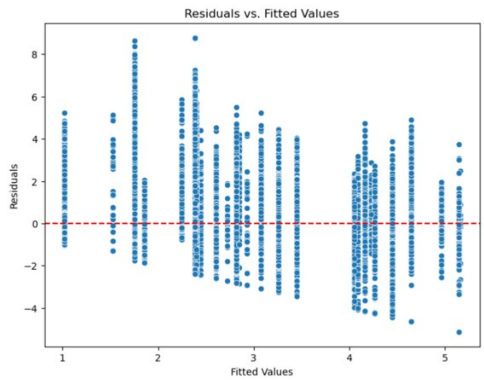
```
                            OLS Regression Results
==============================================================================
Dep. Variable:      cost_per_item_log   R-squared:                       0.483
Model:                            OLS   Adj. R-squared:                  0.483
Method:                 Least Squares   F-statistic:                 9.042e+04
Date:                Wed, 06 Nov 2024   Prob (F-statistic):               0.00
Time:                        02:12:42   Log-Likelihood:            -2.6994e+06
No. Observations:             1905355   AIC:                         5.399e+06
Df Residuals:                 1905327   BIC:                         5.399e+06
Df Model:                          27
Covariance Type:                  HC3
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            4.5793      0.046    100.376      0.000       4.490       4.669
business_920     0.1058      0.073      1.458      0.145      -0.036       0.248
business_930     0.5764      0.198      2.906      0.004       0.188       0.965
business_940    -0.4164      0.057     -7.285      0.000      -0.528      -0.304
business_945     0.3824      0.091      4.206      0.000       0.204       0.561
business_950    -0.3493      0.062     -5.589      0.000      -0.472      -0.227
business_960    -0.3166      0.082     -3.880      0.000      -0.476      -0.157
business_970    -2.7255      0.058    -47.209      0.000      -2.839      -2.612
business_980    -2.1392      0.047    -45.321      0.000      -2.232      -2.047
business_985    -1.8551      0.051    -36.632      0.000      -1.954      -1.756
business_999    -1.5025      0.074    -20.216      0.000      -1.648      -1.357
business_COM    -2.1953      0.046    -48.040      0.000      -2.285      -2.106
business_DLT    -1.9754      0.046    -43.149      0.000      -2.065      -1.886
business_EXL     0.5676      0.051     11.157      0.000       0.468       0.667
business_FLD    -1.3179      0.046    -28.746      0.000      -1.408      -1.228
business_HLB    -0.3158      0.046     -6.837      0.000      -0.406      -0.225
business_IAE    -1.6534      0.145    -11.395      0.000      -1.938      -1.369
business_IAI    -3.0564      0.109    -27.961      0.000      -3.271      -2.842
business_LCP    -2.3353      0.048    -48.838      0.000      -2.429      -2.242
business_LMP    -3.5597      0.046    -78.013      0.000      -3.649      -3.470
business_OTH    -2.8235      0.046    -61.751      0.000      -2.913      -2.734
business_PEN    -1.7310      0.046    -37.277      0.000      -1.822      -1.640
business_RWY    -0.1307      0.046     -2.823      0.005      -0.221      -0.040
business_SAE    -0.5302      0.046    -11.591      0.000      -0.620      -0.441
business_SUR    -1.7651      0.046    -38.665      0.000      -1.855      -1.676
business_TAL    -2.1788      0.046    -47.221      0.000      -2.269      -2.088
business_TRO    -0.4936      0.046    -10.768      0.000      -0.583      -0.404
business_URB    -1.1321      0.046    -24.451      0.000      -1.223      -1.041
==============================================================================
Omnibus:                   196955.498   Durbin-Watson:                   0.964
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           329353.124
Skew:                           0.737   Prob(JB):                         0.00
Kurtosis:                       4.406   Cond. No.                         400.
==============================================================================
```

Correlation between dependent and independent variables:

The heatmap below indicates that most business area codes have minimal correlation with cost_per_item_log, with values near zero. However, business_area_code_LMP stands out with a moderate negative correlation (-0.6), suggesting it’s linked to lower costs. A few areas, such as business_area_code_FLD and business_area_code_SUR, show slight positive correlations, indicating a small increase in cost per item. Overall, the impact of business area codes on cost per item is generally weak, with only a few codes showing notable correlations.

Robustness of the model:





	Feature	VIF
0	business_920	1.0
1	business_930	1.0
2	business_940	1.0
3	business_945	1.0
4	business_950	1.0
5	business_960	1.0
6	business_970	1.0
7	business_980	1.0
8	business_985	1.0
9	business_999	1.0
10	business_COM	1.0
11	business_DLT	1.0
12	business_EXL	1.0
13	business_FLD	1.0
14	business_HLB	1.0
15	business_IAE	1.0
16	business_IAI	1.0
17	business_LCP	1.0
18	business_LMP	1.0
19	business_OTH	1.0
20	business_PEN	1.0
21	business_RWY	1.0
22	business_SAE	1.0
23	business_SUR	1.0
24	business_TAL	1.0
25	business_TRO	1.0
26	business_URB	1.0

The provided model fails the homoscedasticity test.

Conclusion: Correlation from the model can still be informative because it does not consider the variance of residuals. However, they should be interpreted with caution, as heterosdasticity indicate that these relationships might vary across different levels of the independent variables, potentially undermining the stability of these correlations.

2.Effect of Inventory Classification and Order type on Quantity ordered

Importance: For a lighting company, understanding how inventory classification and order type affect the quantity ordered is essential for efficient management and meeting customer needs. By identifying which types of inventories are commonly ordered in large amounts, management can focus on keeping high-demand items in stock, reducing unnecessary storage expenses and ensuring products are readily available. Recognizing patterns in different order types, such as bulk or emergency requests, enables the company to prepare appropriately, enhancing delivery speed and boosting customer satisfaction. This insight also supports better resource allocation, as warehouse space and staff efforts can be concentrated on high-demand items, ultimately resulting in cost savings, faster delivery times, and greater operational efficiency—all of which help the company improve profitability and stay competitive.

Method Used: Multiple linear regression. The dependent variable is the value_quatity and the independent variables are item_type and abc_class_code.

Step taken before the regression: Categorical variables (item_type and abc_class_code) were converted into numeric columns through a process called one-hot encoding. This creates separate columns for each category in the variable, allowing them to be used in the regression model. For example, each unique value in the item_type gets its own column, with a 1 indicating presence and 0 indicating absence.

Result of the test:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     value_quantity_log   R-squared:                       0.193
Model:                            OLS   Adj. R-squared:                  0.193
Method:                 Least Squares   F-statistic:                 1.202e+04
Date:                Sun, 03 Nov 2024   Prob (F-statistic):               0.00
Time:                        14:53:26   Log-Likelihood:           -2.9796e+06
No. Observations:             1953239   AIC:                         5.959e+06
Df Residuals:                 1953199   BIC:                         5.960e+06
Df Model:                          39
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3839      0.033     41.837      0.000       1.319       1.449
abc_B         -0.5464      0.018    -29.951      0.000      -0.582      -0.511
abc_C         -0.6915      0.026    -27.041      0.000      -0.742      -0.641
abc_D         -0.4653      0.017    -26.626      0.000      -0.500      -0.431
abc_E         -0.6179      0.014    -44.958      0.000      -0.645      -0.591
abc_G         -1.4663      0.119    -12.354      0.000      -1.699      -1.234
abc_H         -0.7614      0.028    -27.413      0.000      -0.816      -0.707
abc_I         -0.8713      0.020    -43.228      0.000      -0.911      -0.832
abc_J         -0.6103      0.014    -45.142      0.000      -0.637      -0.584
abc_U         -0.1511      0.014    -10.995      0.000      -0.178      -0.124
order_AES      1.4659      0.033     44.307      0.000       1.401       1.531
order_CDG     -0.8548      0.032    -26.856      0.000      -0.917      -0.792
order_COA      1.2614      0.082     15.384      0.000       1.101       1.422
order_COP     -0.8047      0.044    -18.377      0.000      -0.890      -0.719
order_CPR     -0.7990      0.055    -14.552      0.000      -0.907      -0.691
order_CRD     -0.8057      0.031    -26.299      0.000      -0.866      -0.746
order_CRP     -0.8242      0.072    -11.436      0.000      -0.965      -0.683
order_CRR     -0.8384      0.031    -27.345      0.000      -0.899      -0.778
order_CSH      0.9823      0.039     25.162      0.000       0.906       1.059
order_EDI      1.2968      0.030     42.696      0.000       1.237       1.356
order_EDS      1.5819      0.061     25.852      0.000       1.462       1.702
order_EXP      2.6179      0.034     77.880      0.000       2.552       2.684
order_MIN      1.3682      0.054     25.441      0.000       1.263       1.474
order_NOH      1.9895      0.031     63.274      0.000       1.928       2.051
order_NOR      1.4378      0.030     47.587      0.000       1.379       1.497
order_NOS      1.3149      0.038     34.837      0.000       1.241       1.389
order_OBS      1.5748      0.185      8.496      0.000       1.212       1.938
order_PGS      1.2040      0.323      3.733      0.000       0.572       1.836
order_PME      0.4876      0.078      6.225      0.000       0.334       0.641
order_PMO      1.0394      0.032     32.528      0.000       0.977       1.102
order_PPD      1.0352      0.157      6.586      0.000       0.727       1.343
order_PPO      1.1353      0.032     35.739      0.000       1.073       1.198
order_PRD      1.3496      0.031     43.151      0.000       1.288       1.411
order_PRO      1.1069      0.031     35.721      0.000       1.046       1.168
order_PUP      1.1454      0.031     36.667      0.000       1.084       1.207
order_SPC      0.7386      0.034     21.713      0.000       0.672       0.805
order_SPL      0.7543      0.101      7.488      0.000       0.557       0.952
order_WDC      0.8433      0.353      2.388      0.017       0.151       1.535
order_ZCG     -0.7717      0.040    -19.356      0.000      -0.850      -0.694
order_ZCR     -0.7747      0.043    -18.147      0.000      -0.858      -0.691
==============================================================================
Omnibus:                   248545.797   Durbin-Watson:                   1.169
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           409071.125
Skew:                           0.882   Prob(JB):                         0.00
Kurtosis:                       4.385   Cond. No.                         688.
==============================================================================
```

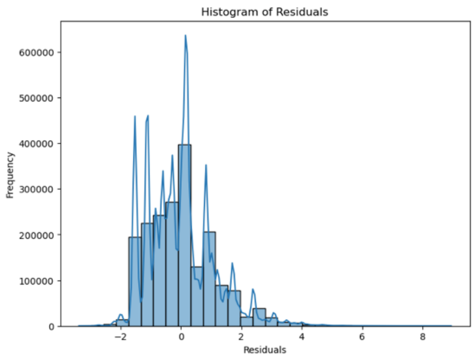
Interpretation:

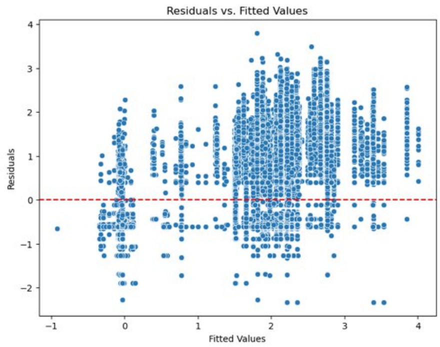
Model Fit: The R-squared value is 0.193, which means the model explains about 19.3% of the variation in the quantity sold. This is relatively low, indicating that other factors outside the model influence quantity sold.

Significance of Variables: Most variables have very low p-values(close to 0), indicating they are statistically significant in predicting quantity sold. Significant variables have a meaningful effect on the correlation of dependent and independent variables, even if the overall model fit is modest.

Coefficient Interpretation: Positive coefficients (e.g., order_EXP at 2.6179) suggest that these variables have a meaningful effect on the outcome, suggesting correlation.

Robustness Test:

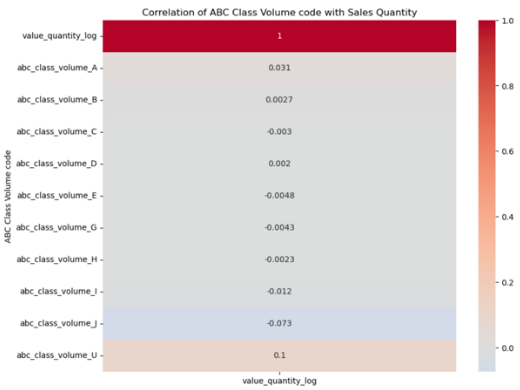


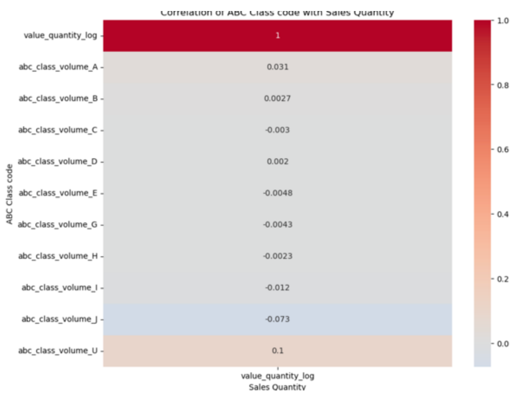


	Feature	VIF
0	abc_B	2.089266
1	abc_C	1.347453
2	abc_D	2.402837
3	abc_E	25.793697
4	abc_G	1.011795
5	abc_H	1.278643
6	abc_I	1.741020
7	abc_J	204.309838
8	abc_U	26.862526
9	order_AES	1.802032
10	order_CDG	2.624690
11	order_COA	1.028105
12	order_COP	1.163080
13	order_CPR	1.079352
14	order_CRD	7.238617
15	order_CRP	1.037325
16	order_CRR	6.882821
17	order_CSH	1.261441
18	order_CSO	1.000268
19	order_EDI	17.678553
20	order_EDS	1.058194
21	order_EXP	1.752957
22	order_MIN	1.083078
23	order_NOH	3.174683
24	order_NOR	210.614640
25	order_NOS	1.319836
26	order_OBS	1.004918
27	order_PGS	1.001596
28	order_PM0	1.000399
29	order_PME	1.032268
30	order_PMO	2.537615
31	order_PPD	1.006913
32	order_PPO	2.700908
33	order_PRD	3.501587
34	order_PRO	4.420489
35	order_PSA	1.079749
36	order_PUP	3.572534
37	order_SPC	1.735989
38	order_SPL	1.017900
39	order_WDC	1.001330
40	order_ZC2	1.000133
41	order_ZCG	1.242688
42	order_ZCR	1.179590
43	order_ZD3	1.000133
44	order_ZOP	1.000665

- The residuals are found to be nearly normal with skewness of 0.88.
- The model suggests potential heteroscedasticity.
- The model has potential multi-collinearity issues, independent variables such as abc_E, abc_J and order_NOR show signs of correlation due to high VIFs.

Correlation:





The correlation heatmaps indicate that there is little linear relationship between value_quantity_log (the dependent variable) and the abc_class_volume_code independent variables. Most correlation values are close to zero, showing minimal connection between them. The only variable with a somewhat notable correlation is abc_class_volume_J, with a weak negative correlation of about -0.073, suggesting a slight inverse relationship with value_quantity_log. Overall, these weak correlations suggest that changes in these independent variables do not strongly explain or predict value_quantity_log on their own, hinting that additional variables or a non-linear analysis may be more effective for understanding value_quantity_log.

Conclusion: When the assumptions of the model as violated, correlations may still show associations, but they could be misleading, as the observed relationships might no hold consistency across different level of data.

Prediction Model

Model to predict Sales based on technology_group code, item_type_code, abc_class code, abc_class_volume, business_chain _code.

Step taken before the regression: One-hot encoding was performed on categorical variables.

Result of the test:

Dep. Variable:	value_sales_log	R-squared:	0.227
Model:	OLS	Adj. R-squared:	0.227
Method:	Least Squares	F-statistic:	1.222e+04
Date:	Sun, 03 Nov 2024	Prob (F-statistic):	0.00
Time:	16:52:31	Log-Likelihood:	-3.6853e+06
No. Observations:	1953239	AIC:	7.371e+06
Df Residuals:	1953191	BIC:	7.371e+06
Df Model:	47		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.8949	0.036	161.642	0.000	5.823	5.966
business_920	0.2706	0.065	4.144	0.000	0.143	0.399
business_945	0.5486	0.134	4.087	0.000	0.286	0.812
business_970	-1.6738	0.056	-30.074	0.000	-1.783	-1.565
business_980	-0.9957	0.042	-23.495	0.000	-1.079	-0.913
business_985	-0.9386	0.072	-13.074	0.000	-1.079	-0.798
business_999	-0.5074	0.059	-8.626	0.000	-0.623	-0.392
business_COM	-0.9579	0.036	-26.267	0.000	-1.029	-0.886
business_DLT	-0.5506	0.037	-15.069	0.000	-0.622	-0.479
business_EXL	0.4506	0.051	8.898	0.000	0.351	0.550
business_FLD	-0.6547	0.037	-17.904	0.000	-0.726	-0.583
business_HLB	-0.5939	0.038	-15.550	0.000	-0.669	-0.519
business_IAE	-0.7524	0.251	-2.993	0.003	-1.245	-0.260
business_IAI	-1.7107	0.215	-7.959	0.000	-2.132	-1.289
business_LCP	-0.9577	0.042	-23.054	0.000	-1.039	-0.876
business_LMP	-1.5757	0.036	-43.511	0.000	-1.647	-1.505
business_OTH	-1.3314	0.036	-36.660	0.000	-1.403	-1.260
business_PEN	-0.8685	0.040	-21.640	0.000	-0.947	-0.790
business_RWY	0.1777	0.038	4.667	0.000	0.103	0.252
business_SAE	-0.4018	0.038	-10.712	0.000	-0.475	-0.328
business_SUR	-0.7768	0.036	-21.419	0.000	-0.848	-0.706
business_TAL	-1.3634	0.037	-36.662	0.000	-1.436	-1.291
business_TRO	-0.2276	0.037	-6.109	0.000	-0.301	-0.155
business_URB	-0.3116	0.037	-8.309	0.000	-0.385	-0.238
item_2	-0.9110	0.019	-47.379	0.000	-0.949	-0.873
item_3	-1.8539	0.024	-75.970	0.000	-1.902	-1.806
item_4	-1.4891	0.012	-123.364	0.000	-1.513	-1.465
item_5	-0.9506	0.006	-164.672	0.000	-0.962	-0.939
item_6	-0.7874	0.006	-129.316	0.000	-0.799	-0.775
item_7	-1.0257	0.005	-188.196	0.000	-1.036	-1.015
item_8	-1.2871	0.019	-67.147	0.000	-1.325	-1.250
item_9	-0.2796	0.010	-28.538	0.000	-0.299	-0.260
abc_B	-0.2144	0.005	-46.487	0.000	-0.223	-0.205
abc_C	-0.4020	0.005	-75.638	0.000	-0.412	-0.392
abc_D	-0.4774	0.005	-102.450	0.000	-0.487	-0.468
abc_E	-0.8343	0.010	-79.757	0.000	-0.855	-0.814
abc_F	0.1204	0.020	6.013	0.000	0.081	0.160
abc_G	-0.0942	0.006	-16.170	0.000	-0.106	-0.083
abc_I	-0.1860	0.008	-22.028	0.000	-0.203	-0.169
abc_J	-0.3564	0.004	-88.942	0.000	-0.364	-0.349
abc_U	-0.6331	0.006	-105.544	0.000	-0.645	-0.621
env_D	3.0564	0.072	42.492	0.000	2.915	3.197
env_I	1.4913	0.189	7.911	0.000	1.122	1.861
env_M	1.1657	0.142	8.214	0.000	0.888	1.444
env_P	0.9964	0.004	255.438	0.000	0.989	1.004
env_R	-0.1347	0.006	-22.334	0.000	-0.147	-0.123
env_S	1.2635	0.004	352.419	0.000	1.256	1.270
env_Z	0.5665	0.007	82.728	0.000	0.553	0.580

Omnibus:	269506.141	Durbin-Watson:	1.484

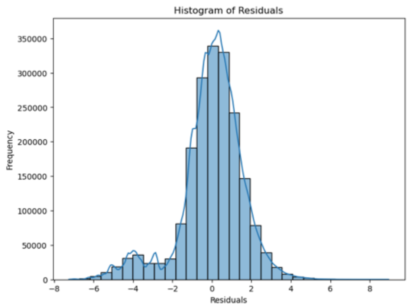
Interpretation:

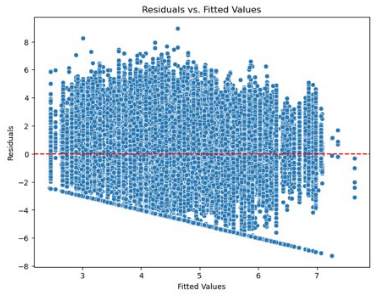
Model Fit: The R-squared is 0.227, which means the model explains about 22.7% of the variation in the sales. This is relatively low, indicating that other factors not included in the model likely influence sales.

Significant variables: Most variables have low p-values(close to 0), meaning they are statistically significant in predicting sales. This implies that these features have a meaningful effect on sales.

Effect of variables: Variables with positive coefficients (e.g., env_M with 3.0564, env_N with 2.7791) increase sales. This means these categories are associated with higher sales. Likewise, negative coefficients (e.g., business_980 with -.09957, item_4 with -1.4891) decrease sales. These categories are linked to lower sales.

Robustness Test:





- With the skewness of -0.865, residuals distribution is close to normal and is acceptable.
- The residuals vs. fitted values plot shows a funnel-like pattern, where the spread of residuals decreases as fitted values increase, indicating heteroscedasticity rather than homoscedasticity. Ideally, residuals should display a consistent spread across all fitted values, but here, the variance is wider for smaller fitted values and narrows for larger ones. This suggests that the variance of errors is not constant, which violates the homoscedasticity assumption.

Model for Sales:

value_sales_log=5.8949+(0.2706×business_920)+(0.5486×business_945)−(1.6738×business_970)−(0.9957×business_980)−(0.9386×business_985)−(0.5074×business_999)−(0.9579×business_COM)−(0.5506×business_DLT)+(0.4506×business_EXL)−(0.6547×covariance_FLD)−(0.5939×business_HLB)−(0.7524×business_IAE)−(1.7107×business_TAI)−(0.9577×business_LCP)−(0.1347×abc_C)−(0.8334×abc_E)+(3.0564×env_D)+(1.4913×env_I)+(0.9964×env_P)

Highlights from the model:

The regression model shows the following relationships for value_sales_log based on business area, environment group, and ABC classification codes:

1. Business Area Codes:

- Positive impact on sales: business_920 (Flood), business_945 (Architectural - Exterior), business_EXL(Healthcare Lighting).
- Negative impact on sales: business_970 (Lamps), business_980 (Trade/Retail - Interior), business_985(Trade/Retail - Exterior), business_COM (Components), business_DLT (Downlight), business_FLD (

Flood), business_HLB (Highbay/Lowbay), business_IAE (Inlite Architectural Exterior), business_TAI (Track & Linear Systems), business_LCP (Lighting Control).

2. Environment Group Codes:

- Positive impact on sales: env_D (Diginet Brand), env_I (Inlite Brand), env_P (Pierlite Brand).
- These codes increase value_sales_log, indicating that certain brands or product lines are linked to higher sales.

3. ABC Classification Codes:

- Negative impact on sales: abc_C (Low Sellers), abc_E (End of Life).
- Items in these classifications tend to have lower sales, as expected for low-selling or end-of-life products.

Conclusion: Model for Sales is derived with R-squared value of 0.193 is derived with statistically significant variables; however, the model is not completely reliable as it violates some key assumptions of multiple linear regression.

Higher Likelihood of Losing Customers

Logistic Regression is a statistical method used for predicting the probability of a binary outcome, churned or not churned in this case. It is particularly useful for modelling situations where the response variable is categorical, especially with two possible outcomes. The model predicts the probability of an event (churn in this case) occurring ,usually with a threshold(e.g., 0.5) to classify the outcome into one of two categories.

Steps taken before performing Logistic Regression:

- We defined a function to check if a customer from 2012 appears in 2013. If they don't, they are labelled as churned (1); otherwise, they are not churned (0). Separate column for created for the same purpose.
- We used one-hot encoding to transform categorical variables (such as business_area_code and others) into numeric format. This step allows us to include categorical variables in the regression.

Building the Initial Logistic Regression Model:

After encoding, we created a logistic regression model to predict the probability of churn. We inspected the p-values of each variable in the model summary to determine statistical significance. Variables with high p-values (>0.05) were deemed statistically insignificant, which means they did not contribute meaningfully to predicting churn.

Based on p-values, we removed the insignificant variables form the model, this process is called feature selection. This helps simplify the model and improve its interpretability by focusing in variables with a significant effect on churn.

Result of the test:

Logit Regression Results

Dep. Variable:	is_churned	No. Observations:	1953239
Model:	Logit	Df Residuals:	1953205
Method:	MLE	Df Model:	33
Date:	Sun, 03 Nov 2024	Pseudo R-squ.:	0.02969
Time:	21:03:15	Log-Likelihood:	-9.4880e+05
converged:	True	LL-Null:	-9.7783e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4135	0.006	-246.464	0.000	-1.425	-1.402
value_sales_log	0.0192	0.001	17.373	0.000	0.017	0.021
business_940	0.7906	0.062	12.747	0.000	0.669	0.912
business_950	0.3826	0.120	3.190	0.001	0.148	0.618
business_960	0.9890	0.167	5.920	0.000	0.662	1.316
business_980	1.6622	0.028	58.842	0.000	1.607	1.718
business_985	1.0132	0.080	12.721	0.000	0.857	1.169
business_999	0.3158	0.066	4.755	0.000	0.186	0.446
business_OTH	-0.2390	0.007	-34.139	0.000	-0.253	-0.225
business_PEN	0.3314	0.025	13.317	0.000	0.283	0.380
business_RWY	0.9546	0.016	60.106	0.000	0.923	0.986
business_TAL	0.2670	0.013	20.444	0.000	0.241	0.293
business_TRO	-0.1642	0.017	-9.914	0.000	-0.197	-0.132
business_URB	-0.2931	0.017	-16.779	0.000	-0.327	-0.259
env_D	0.7441	0.098	7.563	0.000	0.551	0.937
env_I	0.8189	0.120	6.846	0.000	0.584	1.053
env_P	0.0808	0.005	14.986	0.000	0.070	0.091
env_R	0.1562	0.006	24.488	0.000	0.144	0.169
env_S	0.2909	0.005	58.300	0.000	0.281	0.301
district_210	-0.6872	0.012	-57.070	0.000	-0.711	-0.664
district_300	-0.0950	0.005	-17.805	0.000	-0.106	-0.085
district_310	-0.3544	0.015	-23.736	0.000	-0.384	-0.325
district_400	-0.3786	0.006	-63.987	0.000	-0.390	-0.367
district_410	-0.4216	0.011	-38.711	0.000	-0.443	-0.400
district_500	0.2098	0.007	31.652	0.000	0.197	0.223
district_510	-0.5813	0.019	-30.963	0.000	-0.618	-0.544
district_520	-0.3610	0.038	-9.528	0.000	-0.435	-0.287
district_530	-1.8050	0.030	-59.317	0.000	-1.865	-1.745
district_535	-1.2478	0.026	-47.945	0.000	-1.299	-1.197
district_540	-1.2478	0.017	-74.715	0.000	-1.281	-1.215
district_545	2.1203	0.322	6.575	0.000	1.488	2.752
district_600	0.3388	0.007	51.106	0.000	0.326	0.352
district_710	0.0184	0.015	1.255	0.209	-0.010	0.047
district_720	-0.9028	0.012	-75.738	0.000	-0.926	-0.879

Interpretation:

- Pseudo R-squared: 0.0296, meaning the model accounts for about 2.96% of the variation in churn. This is relatively low, suggesting that while certain variables significantly predict churn, there are likely other influential factors outside this model.

Key Variables and Interpretation

- Intercept (const): Represents the base level of churn probability when all other variables are set to zero. It has a negative coefficient of -1.4135.
- value_sales_log: With a coefficient of 0.0192, this variable shows a slight positive association with churn, meaning higher sales values marginally increase the chance of churn.

Significant Variables

Most variables display very low p-values (P>|z| close to 0), indicating they are statistically significant predictors of churn. Key variables with noteworthy coefficients include:

1. Business Area Codes (e.g., business_940, business_980, business_OTH):

- Positive Coefficients (e.g., business_980 with 1.6622): Business areas with positive coefficients suggest a higher likelihood of churn. For example, customers in business_980 are more prone to churn.

- Negative Coefficients (e.g., business_OTH with -0.2390): Business areas with negative coefficients indicate a reduced likelihood of churn, suggesting customers in these areas are more likely to stay.

2. Environment Group Codes (e.g., env_D, env_S):

- Positive Coefficients (e.g., env_S with 0.2909): This implies that certain environmental groups are associated with a higher risk of churn.
- Negative Coefficients (e.g., env_D with -0.7441): This suggests that customers in this environmental group are less likely to churn.

3. Customer District Codes (e.g., district_210, district_720):

- Positive Coefficients (e.g., district_545 with 2.1203): Customers in this district are significantly more likely to churn.
- Negative Coefficients (e.g., district_720 with -0.9028): Customers in this district have a lower probability of leaving.

Robustness Test:

Interpretation of model’s performance metrics:

Accuracy: The model has an accuracy of 80.01%, indicating that it correctly predicts whether a customer churned or not 80.01% of the time. However, accuracy can be misleading in imbalanced datasets (e.g., when most customers don’t churn), as it can reflect the model's tendency to predict the majority class.

Precision: The precision is 0.5644 (or 56.44%), showing that when the model predicts a customer will churn, it’s accurate 56.44% of the time. This suggests some ability to identify churned customers, though there are also a notable number of false positives.

Recall: The recall is 0.0130 (or 1.3%), meaning the model only identifies 1.3% of actual churned customers. This low recall suggests the model misses most of the true churn cases, making it inadequate for capturing most actual churns.

F1 Score: The F1 Score is 0.0255, which is very low. The F1 Score balances precision and recall, so this low result highlights the model’s weak performance in effectively identifying churned customers.

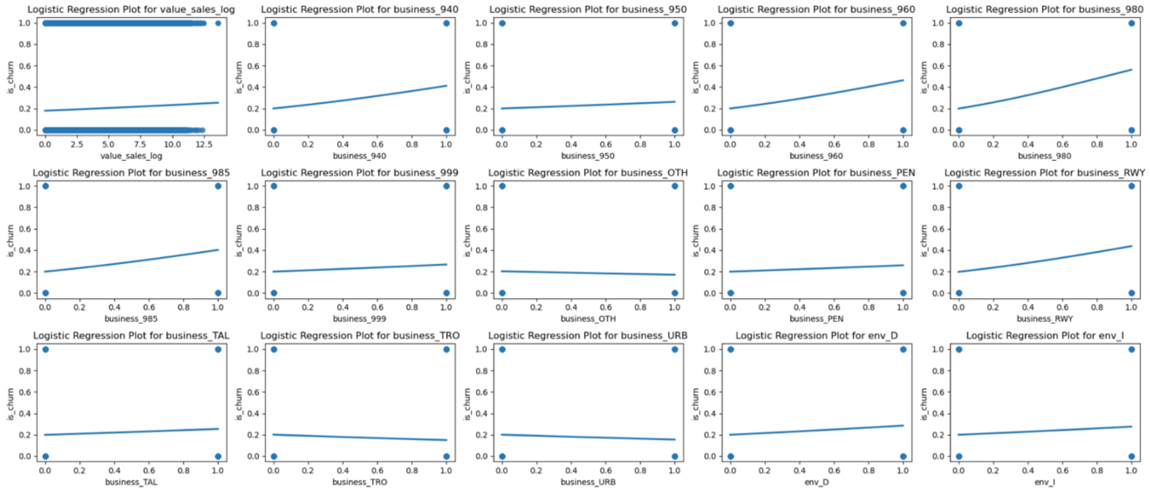
ROC AUC Score: The ROC AUC score is 0.6120, indicating the model's ability to differentiate between churned and non-churned customers. An AUC of 1.0 would be perfect, while 0.5 suggests random guessing. A score of 0.6120 implies weak discriminatory ability.

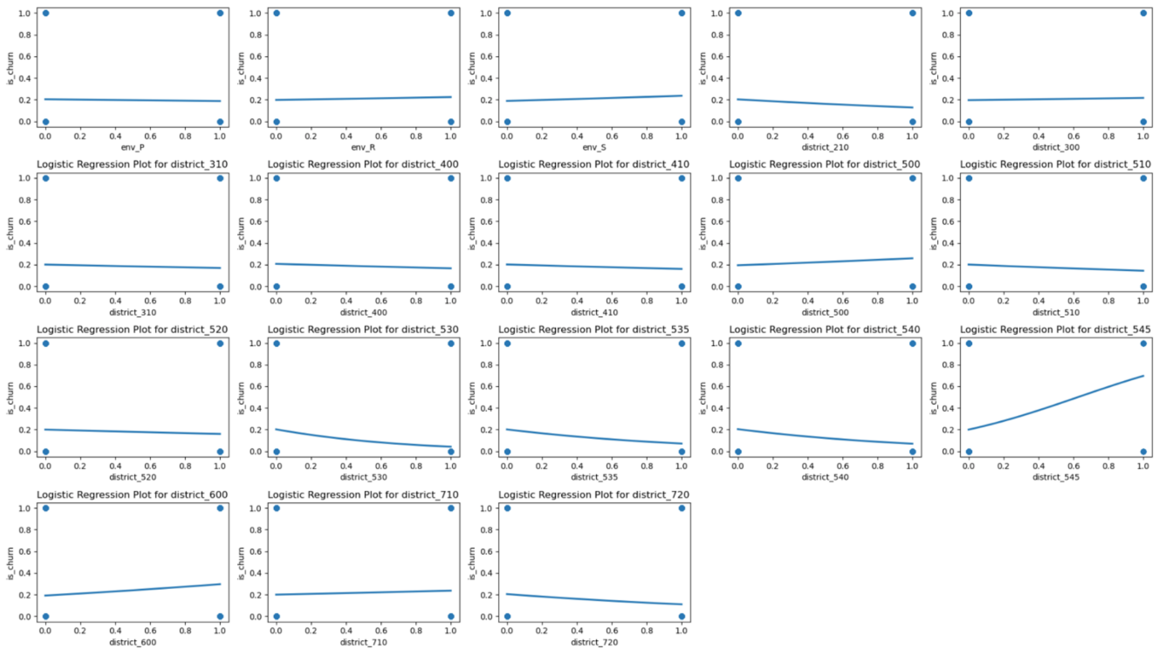
Confusion Matrix:

- True Negatives (467,306): Correctly predicted non-churned customers.
- False Positives (1,183): Non-churned customers incorrectly predicted as churned.
- False Negatives (115,950): Actual churned customers that the model failed to identify.
- True Positives (1,533): Correctly predicted churned customers.

Linearity for the Logit regression:

The following graph shows the relationships between each feature and the log-odds are linear.





Checking for Multicollinearity:

	Feature	VIF
0	value_sales_log	7.771623
1	business_920	1.009169
2	business_930	1.000398
3	business_940	1.055085
4	business_945	1.002355
5	business_950	1.022259
6	business_960	1.001643
7	business_970	1.007175
8	business_980	1.034565
9	business_985	1.003859
10	business_999	1.006200
11	business_COM	2.198762
12	business_DLT	1.586824
13	business_EXL	1.023747
14	business_FLD	1.535534
15	business_HLB	1.160874
16	business_IAE	2.283831
17	business_IAI	3.876391
18	business_LCP	1.114859
19	business_LMP	4.400710
20	business_OTH	2.175349
21	business_PEN	1.043317
22	business_RWY	1.216088
23	business_SAE	1.290758
24	business_SUR	3.606535
25	business_TAL	1.146782
26	business_TRO	1.353246
27	business_URB	1.162329
28	env_D	1.089597
29	env_I	5.169290
30	env_M	1.065308
31	env_P	2.251321
32	env_R	1.571737
33	env_S	2.177968
34	env_Z	3.349258
35	district_210	1.170527
36	district_300	2.043991
37	district_310	1.085986
38	district_400	1.827141
39	district_410	1.178302
40	district_500	1.429973
41	district_510	1.065252
42	district_520	1.173482
43	district_530	1.558975
44	district_535	1.451876
45	district_540	2.293322
46	district_545	1.001047
47	district_600	1.415133
48	district_710	1.089536
49	district_720	1.403948

A VIF more than 10 indicates potential multicollinearity issues. Here, all the features have VIF score of less than 10.

Conclusion: The model has identified the features that result in higher likelihood of losing customers. The model has high accuracy, but this is primarily because most customers don't churn, and the model is biased towards predicting non-churn. Its low recall and F1 Score indicate that it struggles to capture actual churned customers effectively, making it unreliable for churn prediction.

References

Zippia. (2023, January 1). *Customer retention statistics: Trends and insights for 2023*. Zippia. Retrieved from https://www.zippia.com/advice/customer-retention-statistics/

Appendix

The names and descriptions of the variables that used for the regression models:

Variable	Description
Business Area Codes	
business_920	920 - Flood
business_945	945 - Architectural - Exterior
business_970	970 - Lamps
business_980	980 - Trade/Retail - Interior
business_985	985 - Trade/Retail - Exterior
business_999	999 - Other
business_COM	COM - Components
business_DLT	DLT - Downlight
business_EXL	EXL - Healthcare Lighting
business_FLD	FLD - Flood
business_HLB	HLB - Highbay/Lowbay
business_IAE	IAE - Inlite Architectural Exterior
business_TAI	TAI - Track & Linear Systems
business_LCP	LCP - Lighting Control
business_RWY	RWY - Roadway
business_SUR	SUR - Surface
business_TRO	TRO - Troffer
business_OTH	OTH - Other
business_URB	URB - Urban Amenity
Environment Group Codes	
env_D	D - Diginet Brand
env_I	I - Inlite Brand
env_P	P - Pierlite Brand
env_S	S - Sylvania Lighting Brand
env_R	R - Retail Brand

ABC Classification Codes

abc_C	C - Low Sellers
abc_E	E - End of Life

Customer District Codes

district_210	210 - Act/Riverina
district_300	300 - Melbourne
district_310	310 - Tasmania
district_400	400 - Brisbane
district_410	410 - Townsville
district_500	500 - Adelaide
district_510	510 - Darwin
district_520	520 - Inlite - Nz
district_530	530 - South Island - Nz
district_535	535 - Central Region - Nz
district_540	540 - Northern Region - Nz
district_545	545 - Head Office Nz
district_600	600 - Perth
district_710	710 - Head Office Sales
district_720	720 - Intercompany Sales