Group 12
Jigyasa Bardalaye 240019637
Noor Karaman 1900014984
Teanna Puthucheary 240063221

Group 12
Jigyasa Bardalaye 240019637
Noor Karaman 1900014984
Teanna Puthucheary 240063221

Abstract

In analysing the sales data for Coles, a major Australian supermarket chain, we aim to derive insights that can support effective decision-making. We began by cleaning the dataset, removing entries with missing values to avoid the potential biases and inaccuracies associated with imputation (Alam et al., 2023). This process reduced the dataset from 682 to 611 observations. As the data covers only the first two quarters of 2023, we have proceeded cautiously, acknowledging that this limited timeframe may only partially capture seasonal trends or support firm conclusions.

To ensure clarity and accuracy in our visualisations, we prioritised graphical integrity by applying the theme_minimal of the ggplot2 package, maximising the data-ink ratio, and avoiding elements that might mislead or distract the audience. Following Tufte's (1983) principle of avoiding "chartjunk," we focused on delivering visuals that convey data without unnecessary embellishment, which is in line with best practices for objective data presentation (Ruder, 1983).

Group 12
Jigyasa Bardalaye 240019637
Noor Karaman 1900014984
Teanna Puthucheary 240063221

1: Boxplot Addressing Quality Issues of the Dataset

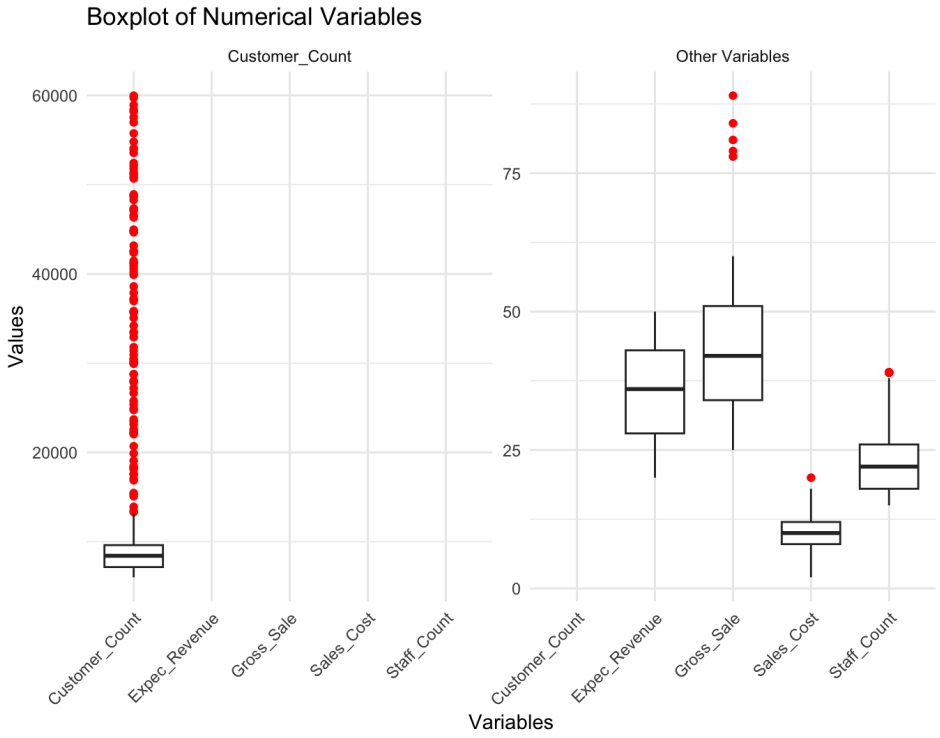


Figure 1

Suitability of the Visual Form

Figure 1 is a univariate analysis tool that visualises the distribution of numerical variables by showing its quartiles, median, and potential outliers. The schematic boxplot designed by Tukey (1977) aims to shed light and critique the overall quality of the dataset. Outliers, represented by a red dot for contrast, represent extreme values and can distort statistical measures such as the mean, which could lead to inaccurate analysis or misleading insights. Businesses must analyse the quality of their data before drawing strategic decisions from them. Using facet_wrap, to group variables according to their y-axis scales, allows for greater data visibility.

Insights

The variable Customer_Count has 153 outliers, comprising 25% of the total observations. This high number of outliers, along with the extreme values seen in variables like Sales_Cost and Staff_Count, is not unexpected; Dawson (2011) found that samples drawn from a normally distributed population typically contain some outliers. The extreme values for Customer_Count suggest a skewed distribution, as shown in Figure 1, where the data is left-skewed with a pronounced right-hand tail. While many stores have similar customer traffic, a few experience exceptionally high footfall; such skewness could distort the interpretation of average store performance if not addressed.

In Figure 1, two separate y-axes accommodate the wide range of values across variables. This variation in the scale of y-values highlights the need for normalisation in further analysis to prevent certain variables from disproportionately influencing statistical models. The presence of outliers may also reflect differences in in-store performance due to factors such as store size, location, and seasonal demand variations. For instance, larger stores are expected to have higher values in Customer_Count and Gross_Sale, which is essential for understanding store demand and capacity and ensuring that Coles can capture maximum revenue rather than losing sales to competitors.

Implications for Business Analytics

If the Customer_Count outliers are not addressed, they can impact the results of predictive models and skew averages, not only limiting the accuracy of insights drawn for carrying out exploratory data analysis but also the applicability for forecasting as the abnormal amount of Customers found in the first two quarters of 2023.

Potential for Additional Data Analysis

Incorporating time-based data to examine how variables like Customer_Count, Gross_Sales, and Sales_Cost fluctuate across seasons or during promotional events could provide valuable insights for demand forecasting and inventory planning. Segmenting the data into categories, such as urban versus rural stores, may further enhance analysis by revealing trends unique to each setting, which could support more tailored product allocation strategies. Additionally, applying log transformation to the data, due to the lack of normality, could improve the robustness of statistical models by reducing the influence of extreme values, leading to more accurate and reliable forecasts in future analyses (West, 2022)

Figure 1 indicates that the data is highly skewed and contains numerous outliers, which limits the reliability of insights gained from carrying out exploratory data analysis (EDA) using this dataset. Any conclusions drawn from this EDA should be interpreted cautiously. Coles should take additional steps to transform the data through either log transforming or subcategory stores according to size, which should unmask underlying trends.

Group 12
Jigyasa Bardalaye 240019637
Noor Karaman 1900014984
Teanna Puthucheary 240063221

2. Scatterplot revealing Novel Insights and Trends

To highlight novel insights, we decided to look at gross sales as our indicator of store performance given that the revenue and cost figures are expectations instead of observed figures. To better visualise the impact of store factors on gross sales, we take the natural logarithm of gross sales so that we can include a regression line and reduce the impact of outliers. Producing a heatmap (Figure 4) allowed us to see a distinct and significant relationship between store size and gross sales as presupposed from Kumar and Karande (2000), hence why we have chosen to further explore this relationship in Figure 2.

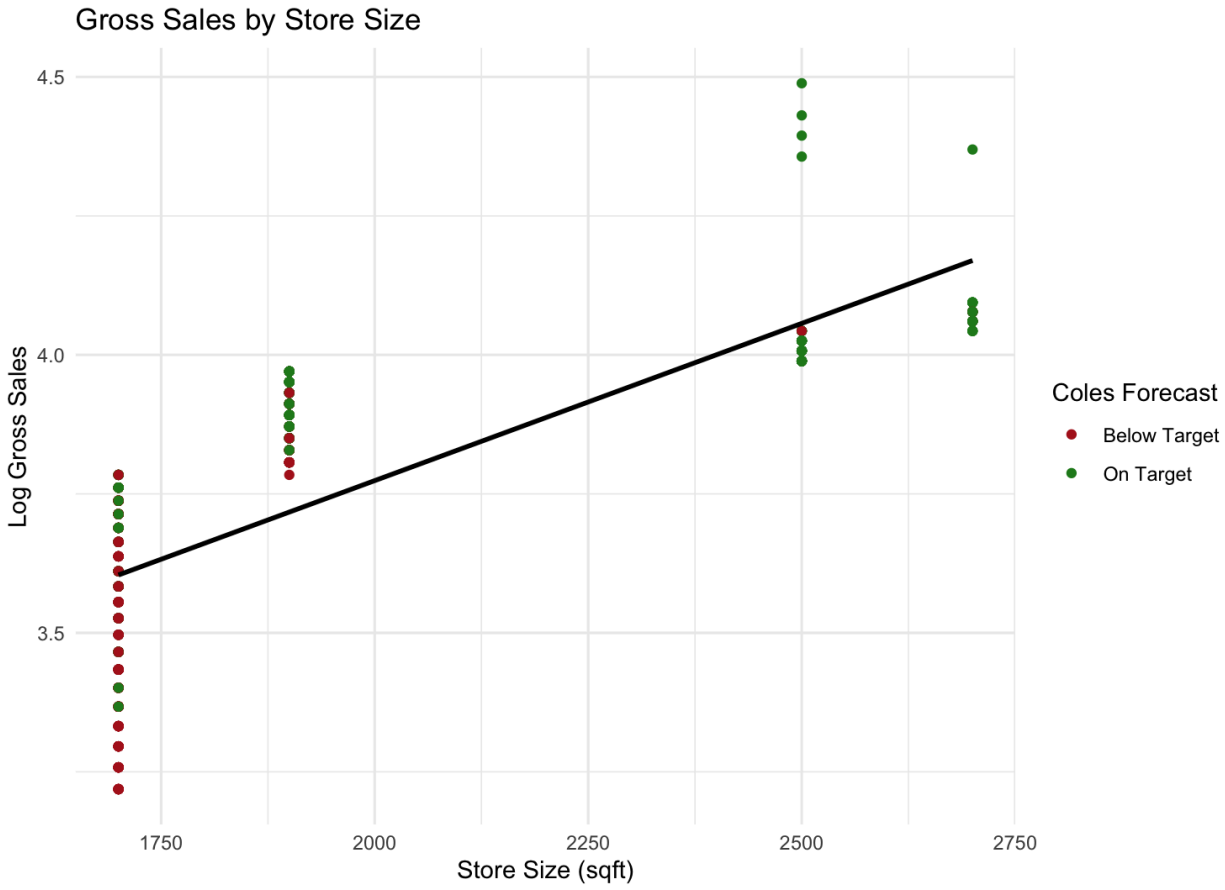


Figure 2

Suitability of the Visual Form

The plot introduces an innovative approach combining store size and sales performance with forecast status. This layered insight allows business analysts to gauge whether larger stores (which presumably incur higher costs) are indeed generating sufficient sales and meeting targets. A scatterplot is well-suited for visualising relationships between two continuous variables (store size and gross sales). Plotting each store individually maintains the granularity necessary to detect outliers or unique cases. The regression line adds value by providing a general trend, allowing viewers to interpret the correlation between store size and gross sales.

The colour-coded Coles Forecast categories bring in an additional layer of insight without cluttering the visualisation, making it easier to assess forecast status relative to store size and sales. The log transformation of Gross Sales addresses potential skewness, ensuring outliers or high sales volumes in large stores do not distort the analysis. This visualisation is crucial for business strategy. It helps stakeholders understand the potential impact of increasing store size on sales and examine whether stores that are below target also tend to have smaller sizes.

Insights

Figure 2 shows a positive correlation between store size and gross sales, as indicated by the upward slope of the regression line. This suggests that larger stores tend to generate higher sales, aligning with retail patterns where bigger stores can stock more products and serve more customers (Kumar & Karande, 2000). Log-transforming Gross Sales clarifies this trend, especially if the raw sales data was skewed, by highlighting proportional sales increases across store sizes and aiding comparison between smaller and larger stores.

Stores marked as “On Target” (in green) cluster around larger sizes, suggesting that bigger stores are more likely to meet sales targets, possibly due to economies of scale that offset higher operational costs. However, stores "Below Target" (in red) appear across various sizes, including larger ones, indicating that size alone doesn’t ensure forecast success. These underperforming larger stores may face operational inefficiencies or local factors—such as demographics, competition, or market conditions—that hinder performance despite their size advantage.

While the regression line shows a positive trend, the relationship between store size and sales may begin to plateau for very large stores. If this trend were confirmed, it would imply diminishing returns on increasing store size beyond a certain threshold. For example, the increase in sales per additional square foot might reduce at larger store sizes, meaning that at some point, adding more space may not proportionally increase revenue. This is highlighted in the literature where diminishing returns are mentioned, that after a certain store size, any further increase in size will not lead to a proportional increase in sales per square foot (Kumar & Karande, 2000). They further argue that other store factors and ‘socioeconomic characteristics of the trade area’ are crucial to further understanding the contributing factors to sales across stores.

Implications for Business Analytics:

Figure 2 offers insights into store planning and expansion strategies. While larger stores generally achieve higher sales, size alone doesn't guarantee meeting targets. Investigating underperforming large stores could help refine strategies, such as optimising layout, targeting marketing, or identifying an optimal store size that balances revenue with operational costs. Figure 2 highlights both the advantages and limits of expanding store size, providing actionable guidance for store performance and planning.

If further analysis confirms a sales plateau beyond a certain size, strategic decisions could shift from expansion toward improving operational efficiency and enhancing customer experience. Business leaders might also consider complementary approaches, like online integration, to drive sales without increasing physical store space.

Business analysts could identify the "break-even" size—where additional space no longer significantly contributes to sales—and calculate the optimal sales-to-cost ratio for different store sizes. This could inform decisions on expanding, scaling down, or relocating stores. Smaller, consistently underperforming stores could be candidates for closure, while high-performing ones might warrant expansion.

Larger stores that are "Below Target" present opportunities for targeted interventions. Analytics teams could investigate these stores further, focusing on areas such as local marketing, operational efficiency, and competitive positioning to address challenges and improve performance.

Group 12
Jigyasa Bardalaye 240019637
Noor Karaman 1900014984
Teanna Puthucheary 240063221

3. Bar Graph Evaluating the Business Implications for Cole Supermarket Sales

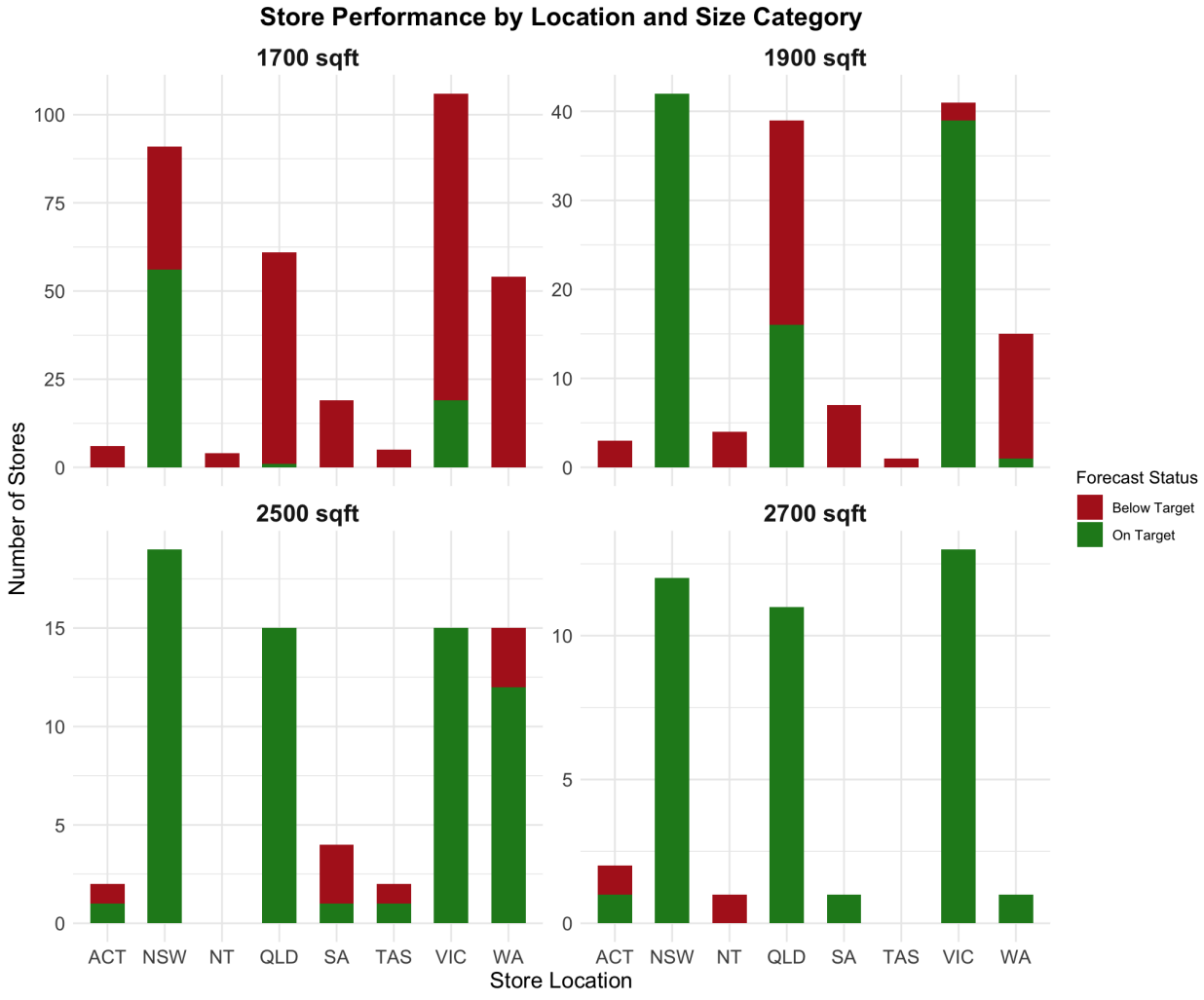


Figure 3

Suitability of the Visual Form

This analysis uses a stacked bar chart segmented by store size (in square feet) and grouped by store location (state) within each panel. Each bar reflects the number of stores classified as either ‘On Target’ (green) or ‘Below Target’ (red) based on revenue performance. The x-axis represents store locations, while the y-axis indicates the number of stores in each performance category. This design aligns with Tukey’s (1977) concept of Exploratory Data Analysis by visually revealing “location-specific trends and outliers.”

By dividing the chart into four panels by store size, we can clearly observe patterns that may otherwise be obscured if locations were analysed as a single group. The visualisation suggests that certain states, such as NSW and VIC, have notably more stores, highlighting the

importance of analysing urban versus rural locations separately. Evaluating these diverse locations together may overlook nuanced factors that affect performance.

The stacked bar chart format allows us to compare two categorical variables — Forecast Status (On Target vs. Below Target) and Store Location — across different store sizes, which resonates with Cleveland and McGill's (1984) research on graphical perception. The faceted layout enhances readability, aiding in cross-category comparison, while the colour-coding provides quick visual cues on performance status, as Few (2006) recommended for effective dashboard design. This structure supports informed business decisions related to store optimisation by revealing how location and store size correlate with performance outcomes.

Insights

Figure 3 reveals that smaller stores, particularly in NSW and VIC, have more 'Below Target' stores, suggesting that factors such as limited inventory capacity or staffing constraints may impact their revenue performance. In contrast, larger stores tend to have more stores performing 'On Target,' likely benefiting from increased customer capacity and broader product offerings. Cities like NSW and VIC consistently have a higher proportion of stores meeting revenue targets, especially in the 2500 and 2700 sqft categories, potentially due to higher demand or customer traffic in these regions.

Conversely, QLD and WA have more stores performing 'Below Target' across various sizes, suggesting possible challenges, such as lower customer demand or regional economic factors that might impact sales. This indicates a need for tailored, location-specific strategies to improve store performance.

Implications for Business Analytics

The observed correlation between store size and performance suggests opportunities for resource reallocation. Smaller stores, which frequently fall 'Below Target,' may benefit from strategies like increased staffing, localised marketing efforts, or promotional activities to attract more customers. For larger stores consistently meeting revenue targets, additional investments—such as expanding inventory or adding services—could further enhance revenue potential.

For stores in QLD and WA, which typically perform below expectations, customised strategies are essential to improve performance. Increased local marketing efforts can boost brand visibility and attract more customers, addressing the lower sales figures. Enhancing the customer experience through initiatives to drive repeat visits can foster customer loyalty and improve overall store performance. Additionally, adjusting product offerings to better align with regional preferences will ensure that inventory meets local demand, further increasing the likelihood of improved sales. These targeted strategies, tailored to the unique needs of each region, can help Coles overcome challenges in these underperforming areas.

Furthermore, the consistent underperformance of smaller stores in these regions may warrant an evaluation of whether the current store size meets market demand. Resizing or restructuring these stores might improve performance or establish distinct revenue targets for smaller stores to reflect their unique challenges.

Potential for Additional Data Analysis

This analysis opens several avenues for further exploration. One key area to investigate is customer demographics and foot traffic to determine whether high-traffic stores consistently perform better, potentially guiding marketing campaigns for underperforming locations. Additionally, examining operational costs and profit margins across different store sizes would offer valuable insights, as the current analysis did not account for these factors in the expected revenue calculations. Another area for investigation is the correlation between higher staff-to-customer ratios and improved performance, which could inform staffing adjustments to optimise store operations. Finally, reviewing seasonal or quarterly data could help tailor strategies to specific time-based trends, ensuring that Coles adapts to changing market conditions and maximises sales opportunities throughout the year.

By delving into these areas, Coles Supermarket can better understand the drivers behind store performance and develop targeted strategies for improvement across locations and store sizes.

Jigyasa Bardalaye 240019637
Noor Karaman 1900014984
Teanna Puthucheary 240063221

Conclusion

This analysis of Coles Supermarkets' store performance reveals a positive correlation between store size and gross sales, with larger stores generally achieving higher sales and more likely to meet revenue targets. However, size alone does not guarantee success; certain large stores still underperform, suggesting that factors like local demographics, competitive landscape, and operational efficiency play a role in performance. Location-specific trends also emerged, with stores in urban areas like NSW and VIC often performing better, while those in QLD and WA face more challenges. Therefore, urban and rural stores should be analysed separately.

The implications for Coles Supermarket are clear: optimising store size can boost sales, but a one-size-fits-all approach is inadequate. To enhance sales across store sizes and locations, Coles should implement location-based strategies, improve operations in underperforming stores (especially smaller ones), and adjust marketing strategies to regional demand.

However, this analysis has limitations, including a dataset covering only the first two quarters of 2023, which may not capture seasonal trends. Outliers and data cleaning reduced the sample size, affecting the generalisability of the results to the population. Additionally, the lack of detailed cost data limits the ability to assess profit margins and provide recommendations. Future analyses should use longer timeframes, incorporate a detailed breakdown of operational costs, and address data skewness for more robust insights.

Jigyasa Bardalaye 240019637
Noor Karaman 1900014984
Teanna Puthucheary 240063221

References

Alam, S., Ayub, M.S., Arora, S. and Khan, M.A., 2023. An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, *9*, p.100341.

Cleveland, W.S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Journal of the American Statistical Association, 79(387), 531-554.

Dawson, R., 2011. How significant is a boxplot outlier?. Journal of Statistics Education, 19(2).

Few, S. (2006). Information Dashboard Design: The Effective Visual Communication of Data. Sebastopol, CA: O'Reilly Media.

Kumar, V. and Karande, K. (2000) *The effect of retail store environment on retailer performance*, *Journal of Business Research*. Available at: https://econpapers.repec.org/article/eeejbrese/v_3a49_3ay_3a2000_3ai_3a2_3ap_3a167-181.htm (Accessed: 11 November 2024).

Ruder, M.A., 1983. The visual display of quantitative information [online]

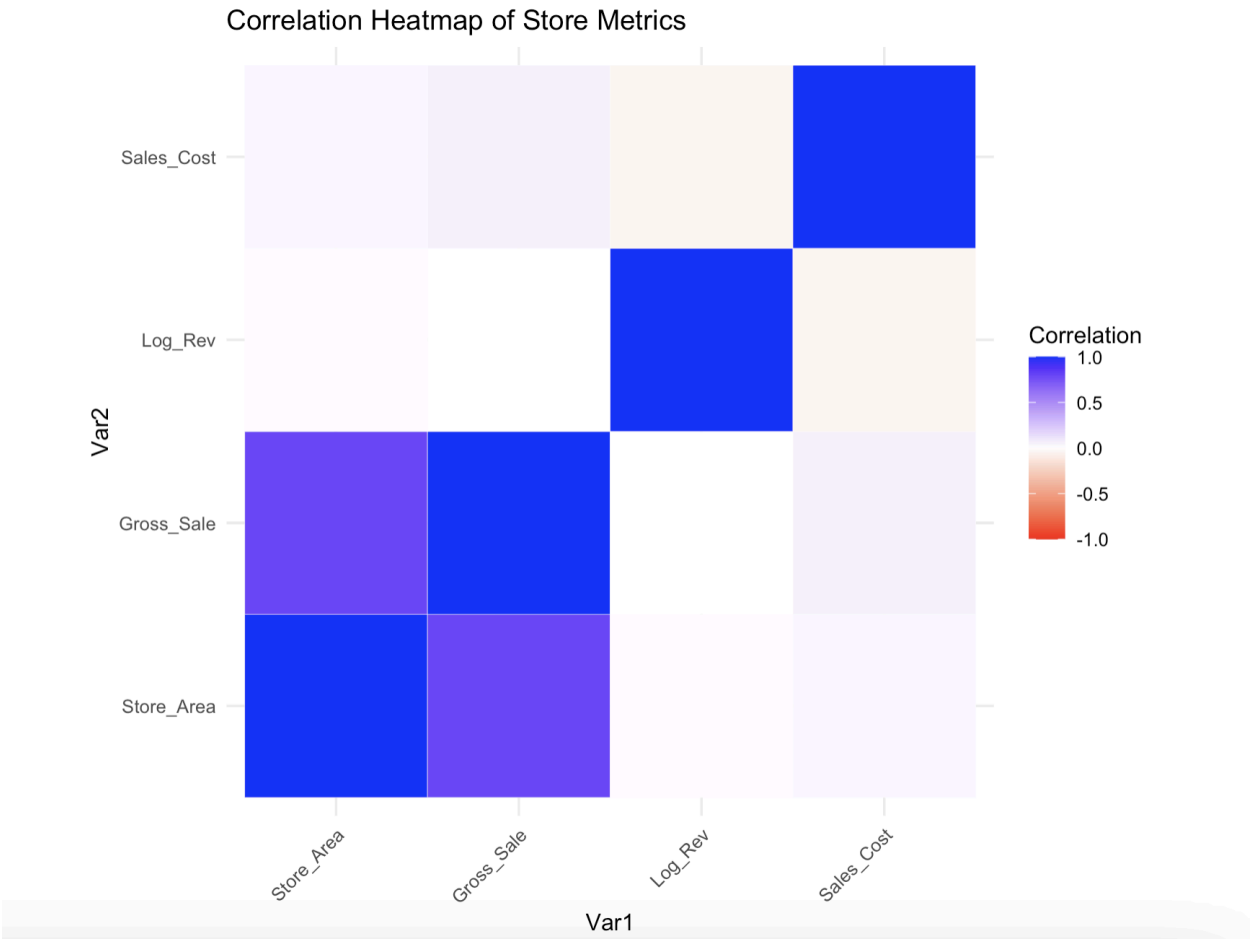
Tufte, E.R., 1985. The visual display of quantitative information. The Journal for Healthcare Quality (JHQ), 7(3), p.15.

Tukey, J.W., 1977. Exploratory data analysis. Reading/Addison-Wesley.

West, R.M., 2022. Best practice in statistics: The use of log transformation. Annals of Clinical Biochemistry, 59(3), pp.162-165.

Group 12
Jigyasa Bardalaye 240019637
Noor Karaman 1900014984
Teanna Puthucheary 240063221

Appendix



This shows quite a significant positive relationship between gross sales and store area, leading to the further exploration of store area vs gross sales in Plot 2.