~a Kituva-James Report~

Superstore Sales and Profit Analysis Report

Superstore Sales and Profit Analysis Report

1. Introduction This report presents a comprehensive analysis of the Superstore sales and profit data. The primary objective is to identify key trends, performance drivers, and areas for improvement across various business dimensions, including sales over time, product categories, customer segments, and geographical regions. The insights derived from this analysis aim to support strategic decision-making for optimizing operations and enhancing profitability.

2. Data Overview and Preprocessing The analysis was performed on the "Orders" sheet from the superstore single.xls dataset. The initial data inspection revealed:

- **Dataset Size:** The dataset contains 9994 rows and 21 columns.
- **Data Types:** Columns like Order Date and Ship Date are handled as datetime objects upon loading, which is crucial for time-series analysis.
- **Missing Values:** No significant missing values were found that required extensive imputation, ensuring data integrity.
- **Duplicate Values:** 1 duplicate rows were found and 5193 duplicate 'Order ID' entries were noted, indicating potential multiple items per order, which is handled appropriately by grouping.
- **Outliers:** Potential outliers were identified in numerical columns like Sales and Profit using the Interquartile Range (IQR) method. These were noted but not capped in this analysis to preserve actual transaction values, as they represent genuine high/low sales or profit figures.
- **Consistency:** Minor inconsistencies in whitespace and casing in object columns were identified, but these generally do not impede the overall analysis.

3. Key Performance Indicators (KPIs) Several key performance indicators were calculated to provide a high-level overview of the Superstore's financial health:

- **Overall Profit Margin:** 12.48% This indicates that for every dollar of sales, the Superstore retains 12.48 cents as profit.
- **Average Order Value:** $229.86 On average, each order placed has a value of $229.86.
- **Top 10 Customers by Sales and Profit** Understanding top customers helps in identifying valuable customer segments for targeted marketing or loyalty programs.

    - **Top 10 Customers by Sales:** (Refer to console output for detailed table) The highest-spending customer is Sean Miller with sales of $25043.05. This highlights the importance of retaining high-value customers.
    - **Top 10 Customers by Profit:** (Refer to console output for detailed table) The most profitable customer is Sean Miller with a profit of $8275.00. It's important to note if high-sales customers are also high-profit customers.
- **Top 10 Products by Sales and Profit** Identifying top-performing products is crucial for inventory management and sales strategies.
    - **Top 10 Products by Sales:** (Refer to console output for detailed table) Products like Canon imageCLASS 2200 Advanced Copier and Apple Smart Phone, Cordless are the highest revenue generators.
    - **Top 10 Products by Profit:** (Refer to console output for detailed table) Canon imageCLASS 2200 Advanced Copier and HP Designjet T210 Plotter contribute most significantly to the bottom line. Discrepancies between top sales and top profit products suggest varying profit margins.
- **Sales and Profit per Customer Segment** Analyzing performance by customer segment (Consumer, Corporate, Home Office) helps tailor strategies.
    - **Sales and Profit by Customer Segment:** (Refer to console output for detailed table) The Consumer segment generates the highest sales, followed by Corporate. Similarly, Consumer is the most profitable segment.
- **Average Sales and Profit per Quantity** These metrics provide insights into the efficiency and profitability at a per-unit level.
    - **Average Sales per Unit Quantity:** $106.39
    - **Average Profit per Unit Quantity:** $14.17 These figures help understand the average revenue and profit generated from each item sold.

4. Sales and Profit Trends Over Time The time-series analysis reveals the Superstore's performance evolution:

- **Monthly Sales Trend** (Refer to Superstore_Analysis_Plots.pdf - Page 1): Sales generally show an upward trend over the years, with noticeable seasonality. There are clear peaks, often towards the end of the year (e.g., November-December), suggesting strong holiday season performance. Some dips are also visible, which could correspond to off-peak seasons.
- **Monthly Profit Trend** (Refer to Superstore_Analysis_Plots.pdf - Page 2): Profit trends largely mirror sales trends, indicating that higher sales usually translate to higher profits. However, there are instances where profit dips significantly even with

moderate sales, suggesting potential issues with discounts or cost management during those periods.

5. Product Category and Sub-Category Analysis Understanding performance at the product level is critical for portfolio management.

- **Sales by Product Category** (Refer to Superstore_Analysis_Plots.pdf - Page 3): Technology and Furniture are the top two categories by total sales, followed closely by Office Supplies.
- **Profit by Product Category** (Refer to Superstore_Analysis_Plots.pdf - Page 4): Technology and Office Supplies are the most profitable categories. Furniture, despite high sales, has a significantly lower profit contribution, indicating potential margin issues.
- **Top 15 Sub-Categories by Sales** (Refer to Superstore_Analysis_Plots.pdf - Page 5): Phones, Chairs, and Storage are consistently high in terms of sales volume.
- **Top 15 Sub-Categories by Profit** (Refer to Superstore_Analysis_Plots.pdf - Page 6): Copiers and Phones are highly profitable.
- **Sub-Categories with Negative Profit** (Refer to Superstore_Analysis_Plots.pdf - Page 7): Tables, Bookcases, and Supplies consistently show negative profit. This is a critical area for immediate investigation, as these products are losing money.

6. Impact of Discount on Profit The analysis of discount groups reveals a clear inverse relationship:

- **Total Profit by Discount Group** (Refer to Superstore_Analysis_Plots.pdf - Page 8): Orders with "No Discount (0%)" contribute the vast majority of the total profit. As the discount percentage increases, the total profit generally decreases, turning significantly negative for higher discount ranges (e.g., 60-70%, 70-80%). This strongly suggests that aggressive discounting is detrimental to profitability.

7. Regional Performance Geographical analysis helps identify strong and weak markets.

- **Total Sales by Region** (Refer to Superstore_Analysis_Plots.pdf - Page 9): The West and East regions lead in total sales, indicating their larger market size or stronger sales operations.
- **Total Profit by Region** (Refer to Superstore_Analysis_Plots.pdf - Page 10): Similarly, West and East are the most profitable regions. Central and South regions have lower sales and profit, suggesting opportunities for growth or efficiency improvements.

8. Numerical Analysis: Correlation and Regression

8.1. Correlation Analysis A correlation matrix was generated for the key numerical features: Sales, Quantity, Discount, and Profit. This matrix helps to understand the linear relationships between these variables.

- **Correlation Matrix Heatmap** (Refer to Superstore_Analysis_Plots.pdf - Page 11): The heatmap visually represents the strength and direction of these correlations. Key observations include:
    - **Sales and Profit:** There is a strong positive correlation between Sales and Profit (correlation coefficient of approximately 0.48). This is expected, as higher sales generally lead to higher profits.
    - **Discount and Profit:** There is a strong negative correlation between Discount and Profit (correlation coefficient of approximately -0.22). This confirms that as discounts increase, profit tends to decrease, reinforcing the findings from the "Impact of Discount on Profit" section.
    - **Quantity and Profit:** There is a weak positive correlation between Quantity and Profit (correlation coefficient of approximately 0.07). While more items sold can contribute to profit, its direct linear relationship is not as strong as Sales.

8.2. Multiple Linear Regression A multiple linear regression model was built to predict 'Profit' (dependent variable) based on 'Sales', 'Quantity', and 'Discount' (independent variables). This model helps quantify the impact of each independent variable on profit, assuming a linear relationship.

- **Regression Results Table** (Refer to Superstore_Analysis_Plots.pdf - Page 12): The key results from the OLS (Ordinary Least Squares) regression model are summarized in the table.
    - **R-squared:** The model has an R-squared value of **0.273**, meaning approximately 27.3% of the variance in Profit can be explained by the independent variables (Sales, Quantity, and Discount). The Adjusted R-squared is also 0.273, indicating that the model's explanatory power is robust and not overly inflated by the number of predictors.
    - **F-statistic and Prob (F-statistic):** The F-statistic is 1249.00 with a p-value of 0.000. This highly significant p-value (less than 0.05) indicates that the overall regression model is statistically significant and that at least one of the independent variables is useful in predicting Profit.
    - **Coefficients and P-values:**

- **Intercept (const):** The intercept is **34.9721** (p-value 0.000). This represents the estimated profit when Sales, Quantity, and Discount are all zero.
- **Sales:** The coefficient for Sales is **0.1800** (p-value 0.000). This suggests that for every one-unit increase in Sales, Profit is estimated to increase by $0.18, holding Quantity and Discount constant. This relationship is highly statistically significant.
- **Quantity:** The coefficient for Quantity is **-2.9622** (p-value 0.001). This indicates that for every one-unit increase in Quantity sold, Profit is estimated to decrease by $2.96, holding Sales and Discount constant. This negative relationship, while statistically significant, might seem counter-intuitive at first. It could imply that selling more units, especially at lower prices or with higher associated costs (which are not directly modeled here), can sometimes lead to lower per-unit profit or that higher quantities are often associated with higher discounts.
- **Discount:** The coefficient for Discount is **-233.4570** (p-value 0.000). This is a strong negative coefficient, indicating that for every one-unit increase in Discount (e.g., from 0.0 to 1.0, or 0% to 100%), Profit is estimated to decrease by $233.46, holding Sales and Quantity constant. This reinforces the significant negative impact of discounts on profitability.

8.3. Random Forest Regression with Hyperparameter Tuning (5-Fold Cross-Validation) To explore potential non-linear relationships and improve predictive power, a Random Forest Regressor model was applied and its hyperparameters were tuned using GridSearchCV with **5-fold cross-validation**. This process helps find a more robust and better-performing model by systematically testing different parameter combinations and evaluating them across multiple data splits.

- **Best Parameters found by GridSearchCV:** {'max_features': 0.8, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
- **Best R-squared score from cross-validation:** 0.79
- **Model Evaluation (Tuned Random Forest Regression on Test Set):**
  - **Mean Squared Error (MSE):** 35948.67
  - **R-squared (R2):** 0.26

After tuning with 5-fold cross-validation, the Random Forest model shows a notable improvement in its cross-validation R-squared score (0.79), indicating that the chosen

parameters lead to a better fit on the training data during cross-validation. However, the R-squared on the *test set* is **0.26**. This suggests a potential issue, possibly overfitting during the tuning process (where the model performs well on cross-validation folds but poorly on unseen test data) or that the linear relationships captured by the simple linear regression model are still more robust for generalization in this specific dataset. While the model can explain 79% of the variance in profit during cross-validation, its ability to generalize to completely new data is less strong, similar to the linear regression model.

- **Feature Importances (Tuned Random Forest):**
    - **Sales:** 0.806418
    - **Discount:** 0.151278
    - **Quantity:** 0.042304

The feature importances from the tuned Random Forest model continue to highlight Sales as the most influential factor (approximately 80.6% importance), followed by Discount (about 15.1%), and then Quantity (about 4.2%). This consistent ranking across models reinforces the critical role of these variables in determining profit.

9. Conclusion and Recommendations The Superstore analysis provides valuable insights into its operational and financial performance.

Key Takeaways:

- Overall, the Superstore is profitable, but there are clear areas of strength and weakness.
- Technology and Office Supplies are strong performers in both sales and profit.
- Furniture sales are high, but profitability is a concern.
- Specific sub-categories (Tables, Bookcases, Supplies) are consistently unprofitable.
- High discounts severely erode profit margins.
- West and East are the dominant regions.
- Both linear and non-linear models confirm that Sales and Discount have a statistically significant relationship with Profit. While the tuned Random Forest model shows strong performance during cross-validation, its generalization to the test set is comparable to the linear model, suggesting that further investigation into model complexity or additional features might be beneficial to consistently capture non-linearities for better predictive power on unseen data. Sales remains the most influential factor, followed closely by Discount.

Recommendations:

- **Optimize Product Portfolio:**
    - **Investigate Unprofitable Sub-Categories:** Conduct a deep dive into Tables, Bookcases, and Supplies to understand the root causes of negative profit (e.g., high acquisition costs, shipping damages, pricing errors). Consider discontinuing or re-evaluating these products if profitability cannot be improved.
    - **Improve Furniture Profitability:** Explore strategies to enhance Furniture margins, such as negotiating better supplier prices, optimizing shipping costs, or slightly adjusting pricing.
- **Strategic Discounting:**
    - **Re-evaluate Discount Policy:** Implement stricter controls on high discounts. Discounts should be used strategically to clear old inventory or drive specific customer behavior, not as a general sales tactic.
    - **Analyze Discount Effectiveness:** For any future discounting, track the actual incremental sales and profit generated to ensure a positive ROI. The regression analysis highlights the significant negative impact of discounts on profit, urging careful consideration.
- **Regional Focus:**
    - **Leverage Strong Regions:** Continue to invest in West and East regions, replicating successful strategies in other areas.
    - **Develop Growth Strategies for Weaker Regions:** Analyze Central and South regions to understand their unique challenges and opportunities. This might involve localized marketing, different product assortments, or sales team training.
- **Customer Retention:**
    - **Nurture Top Customers:** Implement loyalty programs or personalized offers for the top-spending and most profitable customers to ensure their continued business.
- **Continuous Monitoring:**
    - Regularly review these KPIs and trends to adapt strategies and ensure sustained growth and profitability.

This report, combined with the detailed plots in Superstore_Analysis_Plots.pdf, provides a robust foundation for strategic decision-making within the Superstore.