Introduction

- "In today's competitive business landscape, data-driven decision-making is crucial for success. This presentation includes an analyses of a dataset from a Superstore Giant to provide insights into their business transactions.
- Our goal is to extract actionable insights, empowering strategic decision-making aligned with market demands and fostering sustainable growth. Using advanced statistical methods, we bridge the gap between raw data and actionable intelligence."

Business Questions and Methods Used:

Product Category Analysis: Do sales significantly differ across Furniture, Office Supplies, and Technology?

Shipping and Regional Impact on Profitability: Does ship mode interact with region to impact profit margins?

Customer Segmentation and Shipping Preferences: Is there an association between customer segments and preferred ship mode?

Predictive Modeling for Profitability: Can we predict profit based on sales, quantity, and discount?

Discount Impact on Purchase Behavior: Does discount affect the likelihood of high-value orders?

Determinants of Sales Performance: What factors significantly influence sales performance?

Methodology and Justification

Q1 **One-Way ANOVA was selected to compare mean sales across product categories** due to its ability to detect statistical differences between multiple groups, allowing us to identify significant variations in sales performance among distinct product categories.

Q2 **Two-Way ANOVA was chosen to examine the interaction effect between 'Ship Mode' and 'Region' on 'Profit' margin.** This method enables us to simultaneously evaluate the impact of two categorical independent variables and their interaction effect on a continuous dependent variable, providing insights into how different shipping modes and regions affect profitability.

Q3 **Chi-Square Analysis was utilized to assess the association between customer segments and preferred ship modes due to its suitability for analyzing categorical data**. By determining whether there is a significant relationship between two categorical variables, this method helps us understand the shipping preferences of different customer segments.

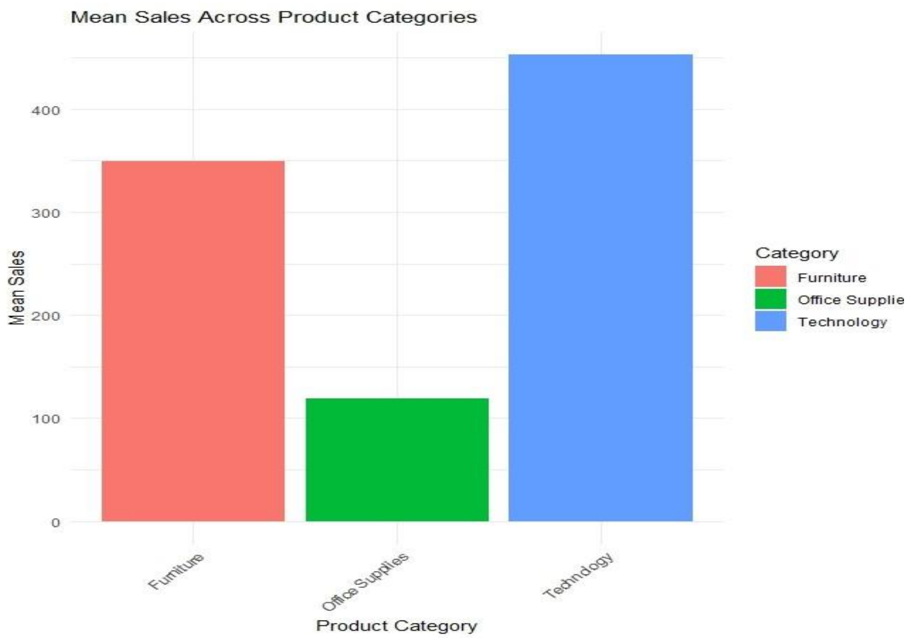
Q4 **Linear Regression was employed to predict 'Profit' based on 'Sales,' 'Quantity,' and 'Discount'** variables due to its capability to model the linear relationship between continuous independent variables and a continuous dependent variable. This method allows us to understand how changes in sales, quantity, and discounts impact overall profitability.

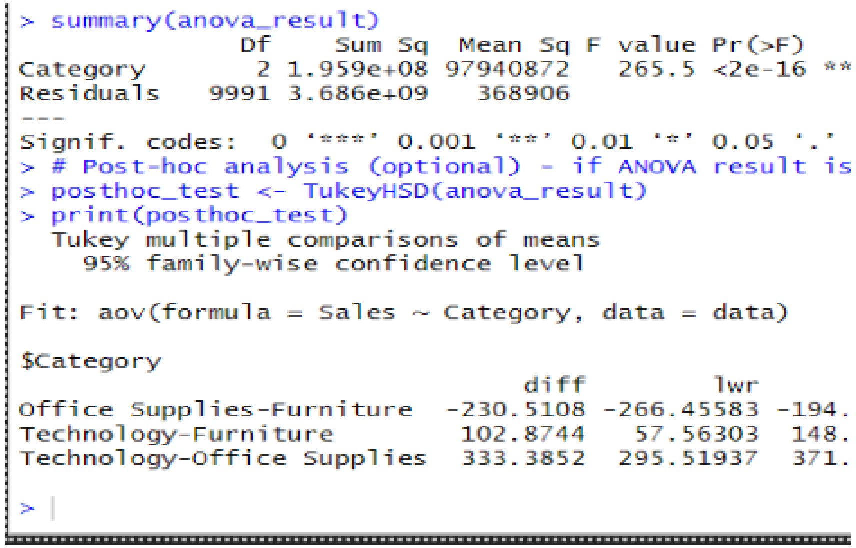
Q5 **Logistic Regression was chosen to assess the relationship between the 'Discount' offered on a product and the likelihood of a customer placing a 'High' value order.** By modeling binary outcomes, this method helps us understand the effect of discounts on purchasing behavior, providing valuable insights into customer preferences.

Q6 **Best Subset Regression was employed to identify the most predictive combination of independent variables influencing 'Sales' performance.** By systematically evaluating all combinations of predictors, this method helps us select the subset of variables that best explains the variation in sales, allowing us to identify the most impactful factors contributing to higher sales in the dataset.

Mean Sales across Product Categories (One-way ANOVA)

- Using one-way ANOVA, we found a significant difference in mean sales among Furniture, Office Supplies, and Technology categories (F(2, 9991) = 265.5, p < 0.001). Post-hoc Tukey's HSD test revealed Office Supplies had lower sales than Furniture (mean difference = -230.51, p < 0.001) and Technology (mean difference = -333.39, p < 0.001), while Technology products had higher sales than Furniture (mean difference = 102.87, p < 0.001). These results highlight the substantial influence of product category on sales.

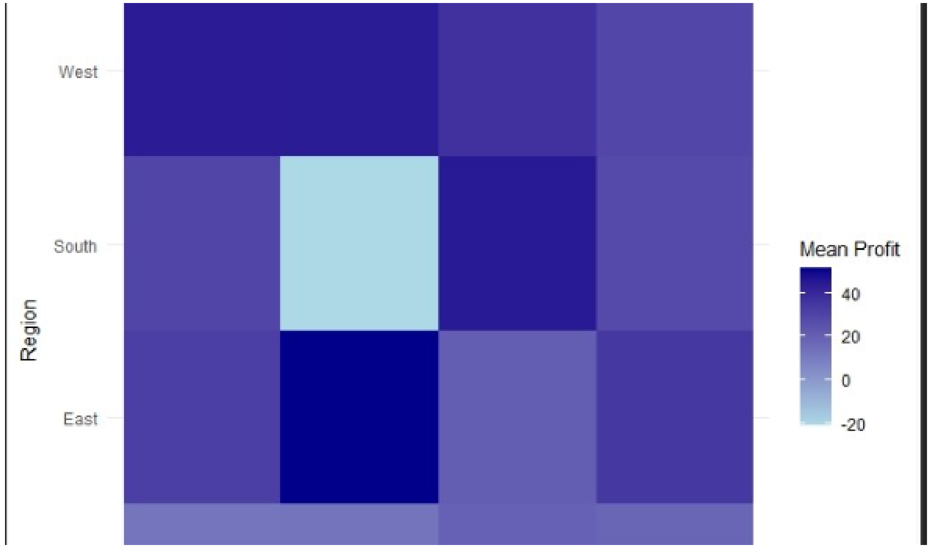


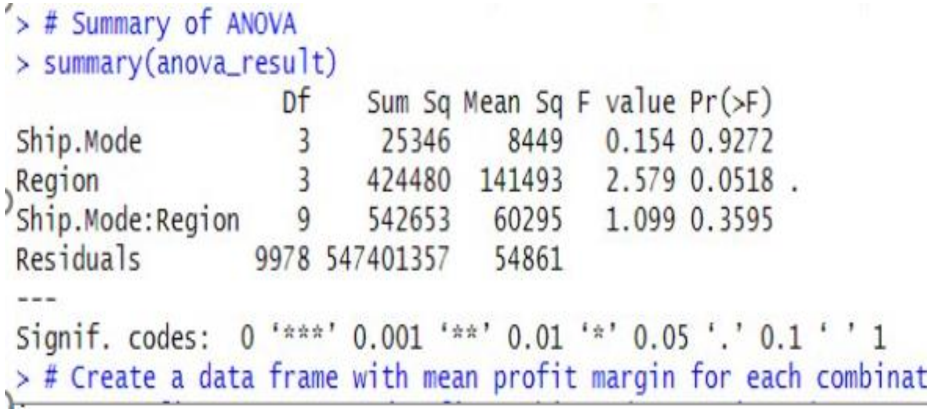


```
> summary(anova_result)
               Df    Sum Sq  Mean Sq F value Pr(>F)
Category        2 1.959e+08 97940872   265.5 <2e-16 **
Residuals    9991 3.686e+09   368906
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
> # Post-hoc analysis (optional) - if ANOVA result is
> posthoc_test <- TukeyHSD(anova_result)
> print(posthoc_test)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Sales ~ Category, data = data)

$Category
                                diff        lwr
Office Supplies-Furniture  -230.5108 -266.45583 -194.
Technology-Furniture        102.8744   57.56303  148.
Technology-Office Supplies  333.3852  295.51937  371.

> 
```

Interaction Effect between Ship Mode and Region on Profit Margin (Using Two-way ANOVA)

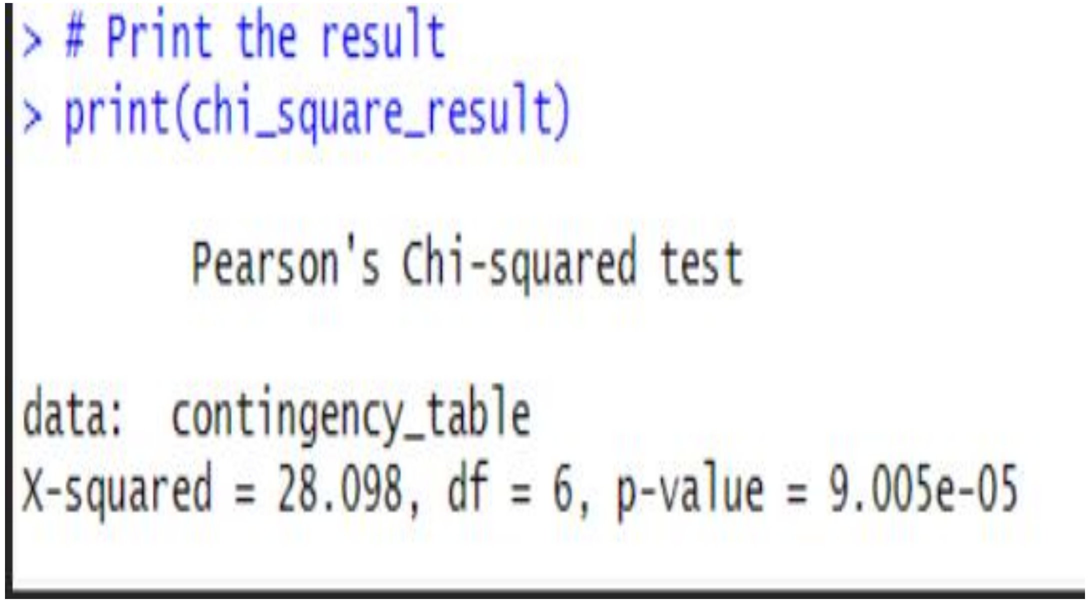
- The two-way ANOVA revealed no significant interaction between 'Ship Mode' and 'Region' on 'Profit' margin (F(9, 9978) = 1.099, p = 0.3595). This suggests consistent impact of shipping modes on profitability across regions. Although 'Region' marginally affected profitability (p = 0.0518), tailored strategies may not be necessary, streamlining decision-making. The heatmap visualization confirmed uniform profit margins across Ship Mode and Region combinations, supporting the ANOVA results. These insights offer valuable guidance for optimizing profitability within the Superstore Giant's operations.




```
> # Summary of ANOVA
> summary(anova_result)
                   Df    Sum Sq Mean Sq F value Pr(>F)
Ship.Mode           3     25346    8449   0.154 0.9272
Region              3    424480  141493   2.579 0.0518 .
Ship.Mode:Region    9    542653   60295   1.099 0.3595
Residuals        9978 547401357   54861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Create a data frame with mean profit margin for each combinat
```

Association between Customer Segment and Preferred Ship Mode (Using Chi-Square Analysis)

- The Pearson's chi-squared test found a significant association between customer segments (Consumer, Corporate, Home Office) and preferred shipping modes ($\chi^2 = 28.098$, df $= 6$, $p < 0.001$). This indicates distinct shipping preferences among different customer segments within the Superstore Giant's operations. Tailoring shipping strategies to align with each segment's preferences can enhance overall customer satisfaction and loyalty. This insight informs strategic decision-making related to shipping infrastructure, service agreements, and promotional efforts, ultimately improving operational efficiency and competitiveness in the market.

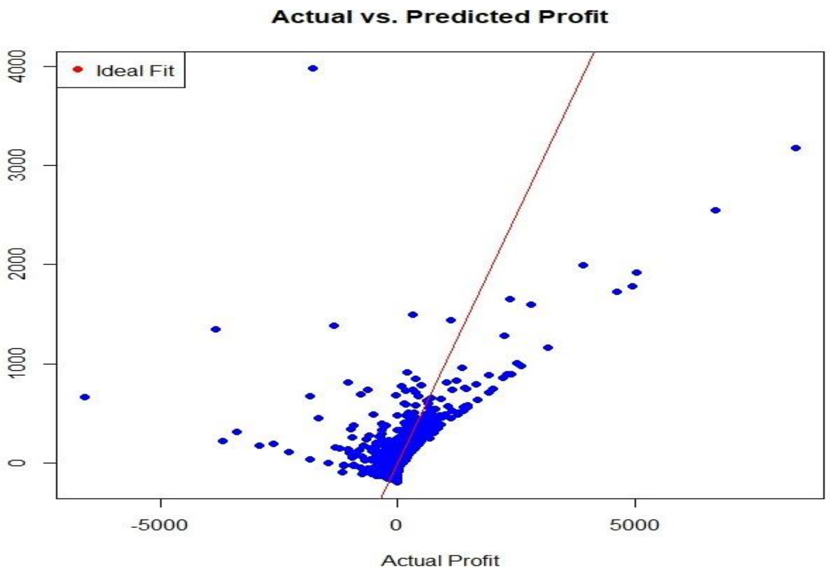


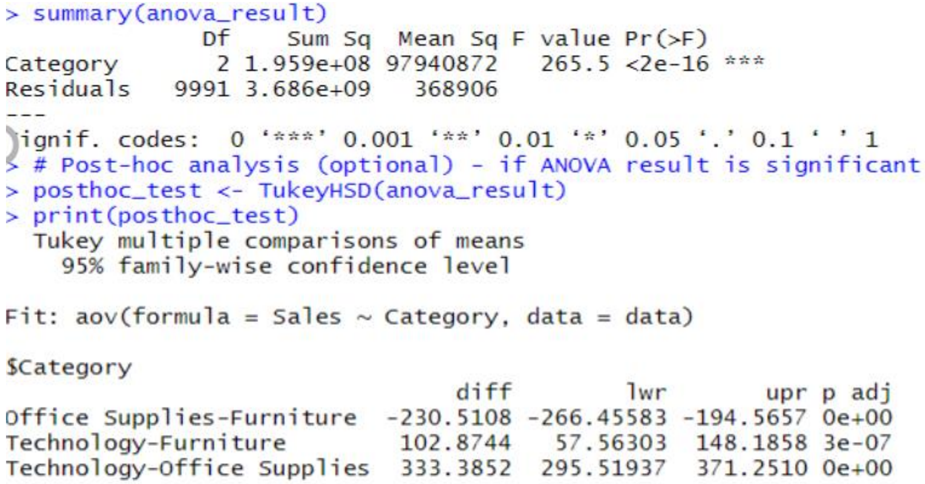
```
> # Print the result
> print(chi_square_result)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 28.098, df = 6, p-value = 9.005e-05
```

Predicting Profit based on Sales, Quantity, and Discount (Using Linear Regression)

- In the analysis of predicting profit based on sales, quantity, and discount variables using linear regression, the model yielded significant results. The coefficients of the variables indicate that sales and discounts have a notable impact on profit, as evidenced by their statistically significant p-values. Specifically, for every unit increase in sales, there is an increase in profit of 0.18 units, while for every unit increase in discount, profit decreases by 233.5 units. Additionally, the negative coefficient for quantity suggests that an increase in quantity is associated with a decrease in profit, although this relationship is less pronounced compared to sales and discount. The model's adjusted R-squared value of 0.2725 indicates that approximately 27.25% of the variability in profit can be explained by the combination of these three variables.

- The linear regression results are in harmony with the scatterplot created, showcasing a consistent depiction of the relationship between predicted profit, derived from the regression model, and actual profit. The scatterplot reveals a clear linear pattern, indicating that the model's predictions align closely with the observed profit values. This alignment reinforces the validity of the regression model's predictions and underscores its capability to accurately estimate profit based on the specified predictor variables—sales, quantity, and discount. The close correspondence between the regression results and the scatterplot further enhances confidence in the model's reliability and efficacy for profit prediction tasks.

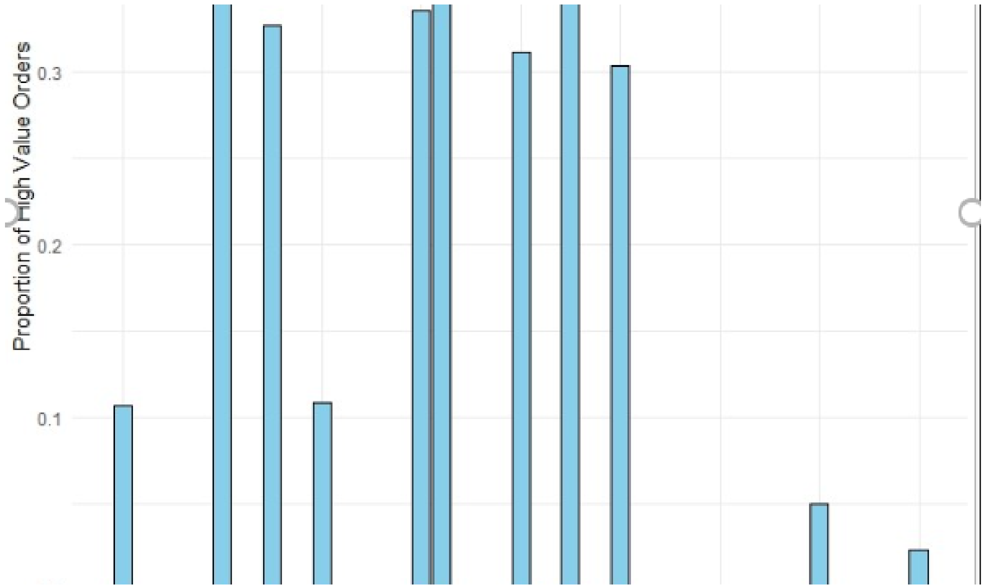


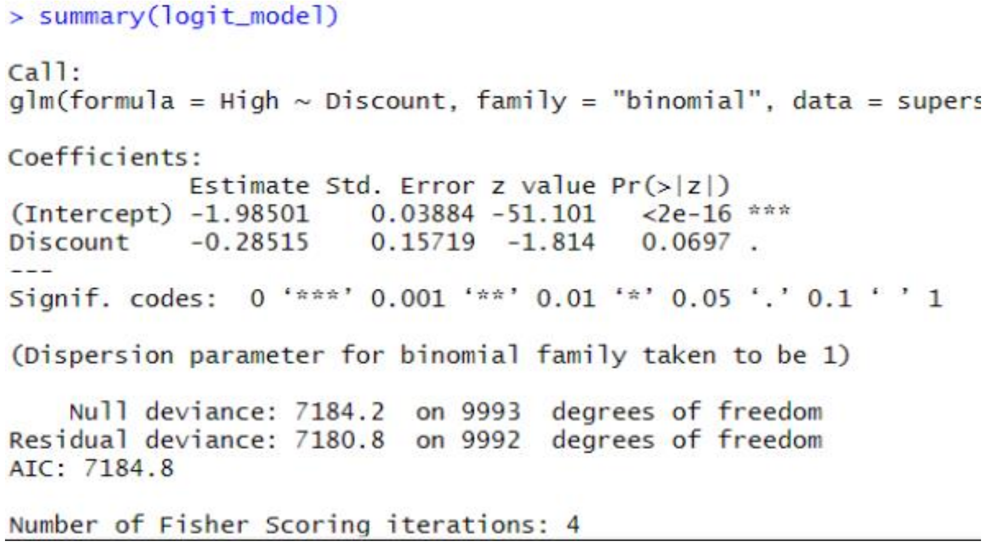


```
> summary(anova_result)
              Df    Sum Sq  Mean Sq F value Pr(>F)    
Category       2 1.959e+08 97940872   265.5 <2e-16 ***
Residuals   9991 3.686e+09   368906                   
---
ignif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Post-hoc analysis (optional) - if ANOVA result is significant
> posthoc_test <- TukeyHSD(anova_result)
> print(posthoc_test)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Sales ~ Category, data = data)

$Category
                                diff        lwr       upr p adj
Office Supplies-Furniture  -230.5108 -266.45583 -194.5657 0e+00
Technology-Furniture        102.8744   57.56303  148.1858 3e-07
Technology-Office Supplies  333.3852  295.51937  371.2510 0e+00
```

Relationship between Discount and Likelihood of High-Value Orders (Using Logistic Regression):

- Using logistic regression, we analyzed the relationship between product discounts and customers' likelihood of placing high-value orders. The coefficient for the 'Discount' variable (-0.28515) indicates that as discounts increase, the log odds of high-value orders decrease. While the associated p-value (0.0697) falls short of conventional significance levels, it's close, suggesting borderline significance. Further investigation or data collection may be needed.
- A bar plot visually represents how the proportion of 'High' value orders changes with varying discount levels, confirming the regression's findings. It shows a decreasing trend as discounts increase, implying that higher discounts may not always lead to more high-value orders.

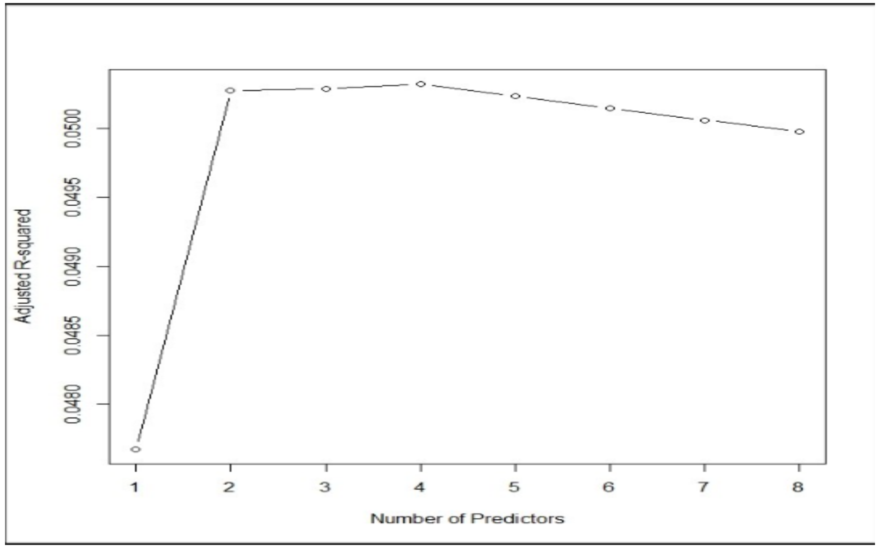


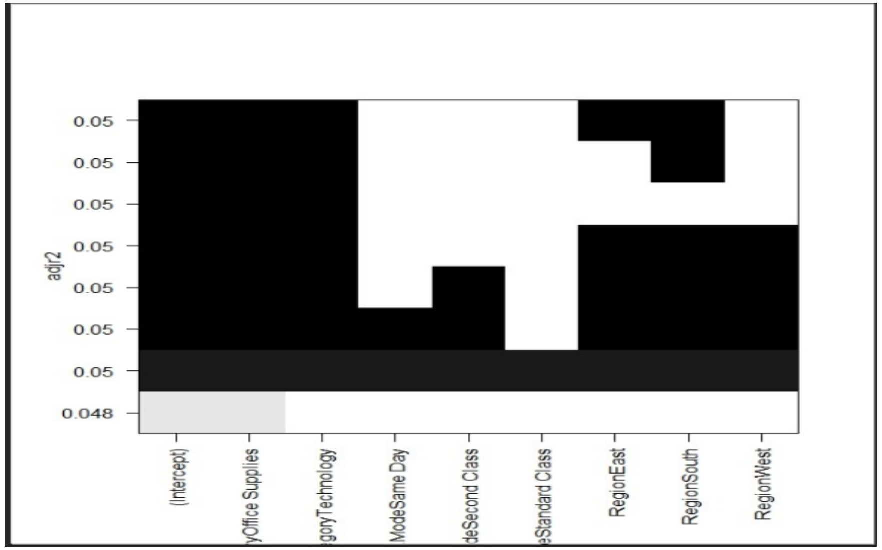


```
> summary(logit_model)

Call:
glm(formula = High ~ Discount, family = "binomial", data = supers

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.98501    0.03884 -51.101   <2e-16 ***
Discount    -0.28515    0.15719  -1.814   0.0697 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7184.2  on 9993  degrees of freedom
Residual deviance: 7180.8  on 9992  degrees of freedom
AIC: 7184.8

Number of Fisher Scoring iterations: 4
```

Identifying Predictive Combination of Independent Variables for Sales Performance (Using Subset Regression):

- Subset regression analysis identified the most predictive variables for 'Sales' performance, including 'Category' (Office Supplies, Technology), 'Ship_Mode' (Same Day, Second Class, Standard Class), and 'Region' (East, South, West). Model 4, incorporating all predictors, yielded the highest adjusted R-squared value, indicating its superior predictive power. This model, encompassing 'Product Category,' 'Ship Mode,' and 'Region,' effectively explains sales variability across segments. Plotted results reaffirmed Model 4's superiority, highlighting the importance of these variables in predicting sales outcomes.





Conclusion

Our analysis of the Superstore Giant's dataset offers actionable recommendations to enhance operational efficiency and profitability. By focusing on maximizing revenue from high-performing product categories, understanding regional trends, and tailoring shipping strategies to customer preferences, the organization can drive long-term profitability. Leveraging predictive models to inform pricing and promotional strategies while carefully evaluating discounting tactics will further optimize profitability. Ultimately, by implementing these recommendations and embracing data-driven decision-making, the Superstore Giant can maintain competitiveness and sustain growth in the market landscape.

References

- Chowdhury, V. (2022, February 17). Superstore dataset. Kaggle.

https://www.kaggle.com/datasets/vivek468/superstore-dataset-final/data

THANK YOU!!!