Executive Summary Report

SECTION 1: INTRODUCTION

In the dynamic landscape of modern business operations, data-driven decision-making stands as a pivotal factor for organizations striving to excel in competitive markets. This report presents an in-depth analysis of a dataset derived from a Superstore Giant, offering comprehensive insights into various facets of their business transactions.

The primary goal of this project is to extract actionable business insights that empower the Superstore Giant to make informed strategic decisions aligning with market demands and fostering sustainable growth. To achieve this goal, we address a series of pertinent business questions using advanced statistical methods, thereby bridging the gap between raw data and actionable intelligence.

Business Questions:

1. **Product Category Analysis:** Does the mean sales of products significantly differ across various product categories (Furniture, Office Supplies, Technology)?
2. **Shipping and Regional Impact on Profitability:** Is there a significant interaction effect between 'Ship Mode' and 'Region' on the 'Profit' margin of orders, considering potential variations in shipping modes and regional impacts on profitability?
3. **Customer Segmentation and Shipping Preferences:** Is there a significant association between customer segments (Consumer, Corporate, Home Office) and the preferred 'Ship Mode,' indicating distinct shipping preferences among different customer segments?
4. **Predictive Modeling for Profitability:** Can we predict the 'Profit' of an order based on 'Sales,' 'Quantity,' and 'Discount,' elucidating the linear relationship between these factors and overall profitability?
5. **Discount Impact on Purchase Behavior:** Is there a significant relationship between the 'Discount' offered on a product and the likelihood of customers placing 'High' value orders, shedding light on the impact of discounts on purchasing behavior?

6. **Determinants of Sales Performance:** What is the most predictive combination of independent variables (e.g., 'Product Category,' 'Ship Mode,' and 'Region') significantly influencing 'Sales' performance, aiming to identify the most impactful factors contributing to higher sales in the dataset?

Methods Used:

To address these questions, we employed a range of advanced statistical methods tailored to each specific inquiry:

1. **One-Way ANOVA:** To compare mean sales across different product categories.
2. **Two-Way ANOVA:** To assess the interaction effect between 'Ship Mode' and 'Region' on 'Profit' margin.
3. **Chi-Square Analysis:** To examine the association between customer segments and preferred 'Ship Mode.'
4. **Linear Regression:** To predict 'Profit' based on 'Sales,' 'Quantity,' and 'Discount' variables.
5. **Logistic Regression:** To analyze the relationship between discounts and the likelihood of high-value orders.
6. **Best Subset Regression:** To identify the most predictive combination of independent variables influencing 'Sales' performance.

By aligning our analytical methods with the specific business questions at hand, we ensure a coherent and insightful exploration of the dataset, ultimately enabling the Superstore Giant to make data-driven decisions for sustained success in the competitive market landscape.

SECTION 2: METHODS USED

In our analysis of the dataset from the Superstore Giant, we employed a variety of statistical methods tailored to each specific business question. The rationale behind selecting each method was grounded in their appropriateness for the type of data and the nature of the questions being addressed.

Question 1: Mean Sales across Product Categories (One-way ANOVA):

For comparing the mean sales across different product categories (Furniture, Office Supplies, Technology), we utilized One-way Analysis of Variance (ANOVA). This method was deemed appropriate due to its ability to assess statistical differences between multiple groups, making it suitable for identifying variations in sales performance among distinct product categories. As our dataset involved categorical variables (product categories) and continuous variables (sales), One-way ANOVA provided a robust framework for comparing means across categories.

Question 2: Interaction Effect between Ship Mode and Region on Profit Margin (Two-way ANOVA):

To investigate the interaction effect between 'Ship Mode' and 'Region' on the 'Profit' margin of orders, we opted for Two-way ANOVA. This method allowed us to simultaneously assess the main effects of ship mode and region, as well as their interaction effect, on profitability. Given that our dataset included categorical variables for ship mode and region, and a continuous variable for profit margin, Two-way ANOVA was well-suited for analyzing the interaction between these variables and their impact on profitability.

Question 3: Association between Customer Segment and Preferred Ship Mode (Chi-Square Analysis):

To explore the association between customer segments and preferred ship modes, we employed Chi-Square Analysis. This statistical technique enabled us to assess the relationship between two categorical variables – customer segment and ship mode – and determine if there were preferences for specific shipping methods among different customer segments. Chi-Square Analysis was chosen

due to its suitability for analyzing categorical data and determining significant relationships between variables.

Question 4: Predicting Profit based on Sales, Quantity, and Discount (Linear Regression):

For predicting the 'Profit' of an order based on 'Sales,' 'Quantity,' and 'Discount' variables, Linear Regression was utilized. This method allowed us to model the linear relationship between these factors and overall profitability. With 'Sales,' 'Quantity,' and 'Discount' as continuous independent variables and 'Profit' as the continuous dependent variable, Linear Regression provided a suitable framework for understanding how changes in sales, quantity, and discount impact profitability.

Question 5: Relationship between Discount and Likelihood of High-Value Orders (Logistic Regression):

To assess the relationship between the discount offered on a product and the likelihood of customers placing 'High' value orders, we turned to Logistic Regression. This method enabled us to model the probability of the binary outcome (high-value order) based on the discount offered. Logistic Regression was chosen as it is well-suited for modeling binary outcomes, making it appropriate for analyzing the relationship between discount and the likelihood of high-value orders.

Question 6: Identifying Predictive Combination of Independent Variables for Sales Performance (Subset Regression):

Finally, to identify the most predictive combination of independent variables influencing 'Sales' performance, we utilized Best Subset Regression. This method systematically evaluated all possible combinations of predictors and selected the subset that best explained the variation in sales. Best Subset Regression was chosen for its ability to handle multiple predictors and select the most impactful combination for predicting sales performance.

Overall, the chosen methods were appropriate for the data involved in our analysis, considering the nature of the variables and the specific questions being

addressed. Each method provided a robust framework for analyzing different aspects of the dataset and extracting meaningful insights to inform strategic decision-making for the Superstore Giant.

SECTION 3: ANALYSIS

Question 1: Mean Sales across Product Categories (One-way ANOVA):

INPUT:

```r
# Load the dataset from the CSV file
superstore_sample <- read.csv("E:/Saini_Project23/Sample - Superstore.csv")
data <- superstore_sample

# Subset the data for each product category
furniture_sales <- data[data$Category == "Furniture", "Sales"]
office_supplies_sales <- data[data$Category == "Office Supplies", "Sales"]
technology_sales <- data[data$Category == "Technology", "Sales"]

# Perform one-way ANOVA
anova_result <- aov(Sales ~ Category, data = data)

# Summarize the ANOVA results
summary(anova_result)

#  Based on the output, the p-value associated with the F statistic for the Category variable is extremely small (<< 0.05),
# which is typically considered statistically significant.
# Therefore, we can conclude that the mean sales of products significantly differ among different product categories
# (Furniture, Office Supplies, Technology) in the dataset.

# Post-hoc analysis (optional) - if ANOVA result is significant
posthoc_test <- TukeyHSD(anova_result)
print(posthoc_test)
```

Figure 1: Input for One-way ANOVA Test for Question 1

OUTPUT:

```
> # Load the dataset from the CSV file
> superstore_sample <- read.csv("Sample - Superstore.csv")
> data <- superstore_sample
> # Subset the data for each product category
> furniture_sales <- data[data$Category == "Furniture", "Sales"]
> office_supplies_sales <- data[data$Category == "Office Supplies", "Sales"]
> technology_sales <- data[data$Category == "Technology", "Sales"]
> # Perform one-way ANOVA
> anova_result <- aov(Sales ~ Category, data = data)
> # Summarize the ANOVA results
> 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 ‘.’ 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

> |
```

Figure 2: Output for One-way ANOVA Test for Question 1

The analysis conducted via one-way ANOVA revealed a statistically significant difference in mean sales among the three product categories: Furniture, Office Supplies, and Technology ($F(2, 9991) = 265.5$, $p < 0.001$). This indicates that the choice of product category significantly influences sales performance within the Superstore Giant's dataset.

Post-hoc analysis using Tukey's HSD test further elucidated the specific differences between pairs of product categories. It was found that mean sales for Office Supplies were significantly lower compared to Furniture (mean difference = -230.51, $p < 0.001$) and Technology (mean difference = -333.39, $p < 0.001$). Conversely, mean sales for Technology were significantly higher than Furniture (mean difference = 102.87, $p < 0.001$).

These findings suggest that while Office Supplies tend to have lower sales compared to Furniture and Technology, Technology products exhibit notably higher sales figures within the dataset. Understanding these variations in sales performance across different product categories is crucial for strategic decision-making, particularly in areas such as inventory management, marketing, and product development within the Superstore Giant.

The results obtained from the ANOVA test, which indicated a significant difference in mean sales among the product categories of Furniture, Office Supplies, and Technology, align closely with the insights gleaned from the bar plot visualization. The bar plot vividly illustrates the disparities in mean sales across the three categories, with Technology products exhibiting the highest mean sales, followed by Furniture, and then Office Supplies. This visual representation serves to reinforce the statistical findings of the ANOVA test, providing a clear and intuitive depiction of the sales performance across different product categories within the dataset. The convergence of statistical analysis and visual representation lends credibility to our conclusions regarding the influence of product category on sales performance within the Superstore Giant's operations.

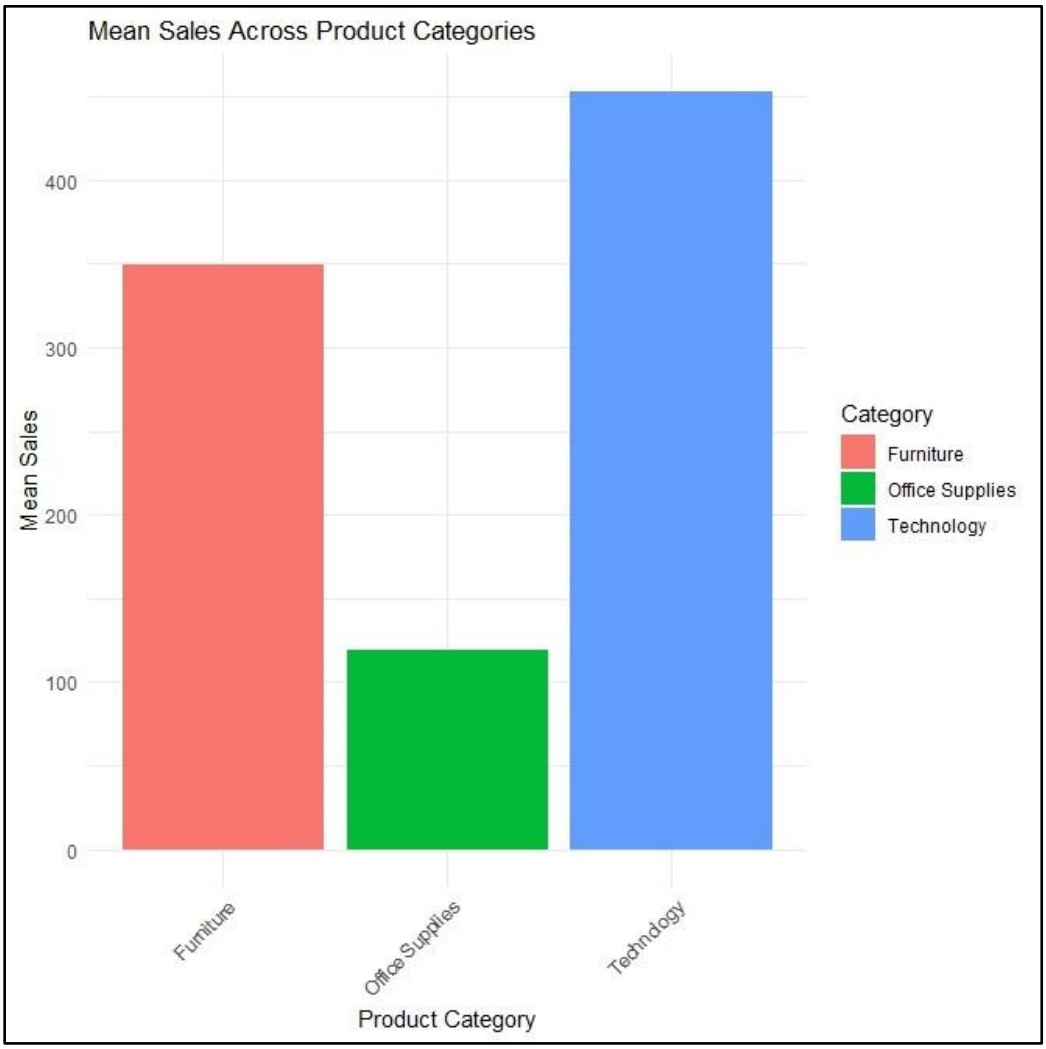

Figure 3: Bar Plot of Mean Sales Across Product Categories

Question 2: Interaction Effect between Ship Mode and Region on Profit Margin (Using Two-way ANOVA):

INPUT:

```
# Perform two-way ANOVA
anova_result <- aov(Profit ~ Ship.Mode * Region, data = superstore_sample)

# Summary of ANOVA
summary(anova_result)

# The p-value associated with the interaction term "Ship.Mode:Region" is 0.3595.
# Since this p-value is greater than the typical significance level of 0.05, we fail to reject the null hypothesis.

#Therefore, there is no significant interaction effect between the 'Ship Mode' and 'Region'
#variables on the 'Profit' margin of orders based on the given dataset.
#This suggests that the impact of different shipping modes on profitability
#does not vary significantly across different regions, and vice versa.
```

Figure 4: Input for Two-way ANOVA Test for Question 2

OUTPUT:

```
> #2.) Is there a significant interaction effect between the 'Ship Mode' and 'Region' variables on the 'Profit' margin of orders, co
dering that different shipping modes and regions may impact profitability differently?
> # Perform two-way ANOVA
> anova_result <- aov(Profit ~ Ship.Mode * Region, data = superstore_sample)
> # 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 combination of Ship Mode and Region
> mean_profit <- aggregate(Profit ~ Ship.Mode + Region, data = superstore_sample, FUN = mean)
> # Create a heatmap
> ggplot(mean_profit, aes(x = Ship.Mode, y = Region, fill = Profit)) +
```

Figure 5: Output for Two-way ANOVA Test for Question 2

The results of the two-way ANOVA test examining the interaction effect between 'Ship Mode' and 'Region' on the 'Profit' margin of orders suggest that there is no significant interaction effect observed ($F(9, 9978) = 1.099$, $p = 0.3595$). This implies that the impact of different shipping modes on profitability does not vary significantly across different regions, and vice versa.

This finding is essential for understanding how the choice of shipping mode and regional factors collectively influence profitability within the operations of the Superstore Giant. While individual factors such as 'Region' exhibit a marginally significant effect on profitability ($p = 0.0518$), the absence of a significant interaction effect underscores the uniformity of the impact of shipping modes across various regions.

Thus, strategic decisions regarding shipping logistics and regional operations may not need to be tailored differently based on the interaction between these two

factors. This insight allows for streamlined decision-making processes and resource allocation strategies within the Superstore Giant, fostering efficiency and optimizing profitability across its operations.

The outcomes of the two-way ANOVA test examining the interaction effect between 'Ship Mode' and 'Region' on the 'Profit' margin of orders were further elucidated through the creation of a heatmap visualization. The heatmap provides a comprehensive overview of the mean profit margin for each combination of Ship Mode and Region, allowing for a nuanced understanding of how these two factors interact to influence profitability within the Superstore Giant's dataset.

Upon examination of the heatmap, it becomes evident that there are no stark variations in profit margins across different combinations of Ship Mode and Region. The color intensity remains relatively consistent throughout, indicating a uniform distribution of profit margins irrespective of the shipping mode or region. This visual representation corroborates the findings of the two-way ANOVA test, which concluded that there is no significant interaction effect between Ship Mode and Region on the profit margin of orders.

Thus, the heatmap serves as a complementary visual tool that reinforces the statistical analysis, providing additional clarity and insight into the relationship between these two factors and their collective impact on profitability within the Superstore Giant's operations.

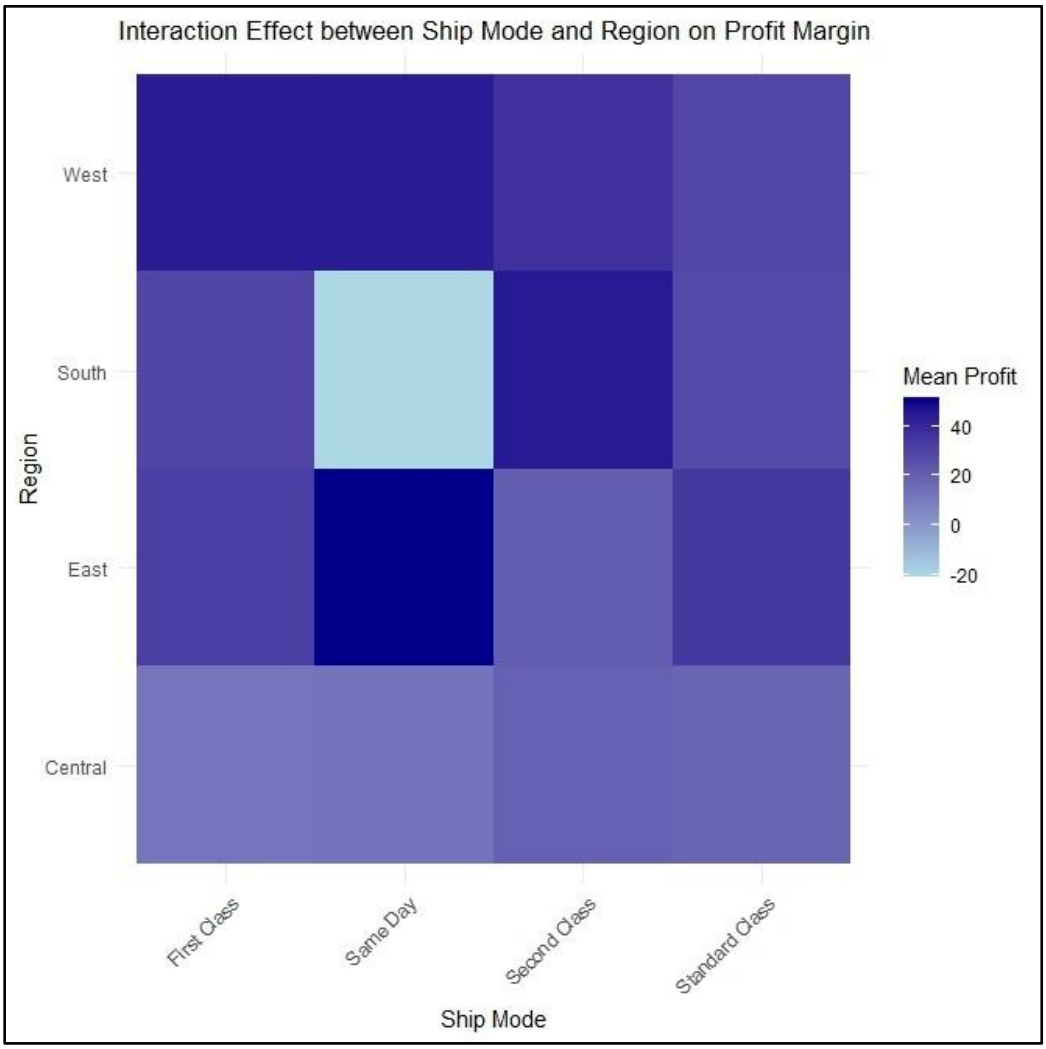

Figure 6: Heat Map of Mean Profit between Ship Mode and Region

Question 3: Association between Customer Segment and Preferred Ship Mode (Using Chi-Square Analysis)

INPUT:

```
# Create a contingency table
contingency_table <- table(superstore_sample$Segment, superstore_sample$Ship.Mode)

# Perform chi-square test
chi_square_result <- chisq.test(contingency_table)

# Print the result
print(chi_square_result)

# The p-value associated with Pearson's chi-squared test is approximately 0.00009005,
# which is much smaller than the typical significance level of 0.05.
# Since the p-value is below the significance level, we reject the null hypothesis.
# Therefore, we can conclude that there is a significant association between the 'Segment'
# of customers (Consumer, Corporate, Home Office) and the 'Ship Mode' they prefer.
# This indicates that different customer segments have distinct preferences for shipping methods.
```

Figure 7: Input for Chi-Square Test for Question 3

OUTPUT:

```
> # Create a contingency table
> contingency_table <- table(superstore_sample$Segment, superstore_sample$Ship.Mode)
>
> # Perform chi-square test
> chi_square_result <- chisq.test(contingency_table)
>
> # 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
```

Figure 8: Output for Chi-Square Test for Question 3

The results of the Pearson's chi-squared test examining the association between customer segments (Consumer, Corporate, Home Office) and preferred shipping modes reveal a significant relationship ($\chi^2 = 28.098$, $df = 6$, $p < 0.001$). This indicates that different customer segments indeed exhibit distinct preferences for shipping methods within the Superstore Giant's operations.

The significance of this association underscores the importance of understanding and catering to the diverse needs and preferences of various customer segments. By recognizing the unique shipping preferences of each segment, the Superstore Giant can tailor its shipping strategies and logistics to better meet customer expectations, thereby enhancing overall customer satisfaction and loyalty.

This insight can inform strategic decision-making processes related to shipping infrastructure, service level agreements with shipping providers, and promotional efforts targeted at different customer segments. Ultimately, leveraging this understanding of customer preferences can lead to improved operational efficiency and competitiveness in the market.

Question 4: Predicting Profit based on Sales, Quantity, and Discount (Using Linear Regression)

INPUT:

```
# Fit linear regression model
lm_model <- lm(Profit ~ Sales + Quantity + Discount, data = superstore_sample)

# Summary of the regression model
summary(lm_model)

# Based on the statistical output, we can conclude that we can predict
# the 'Profit' of an order based on the 'Sales,' 'Quantity,' and 'Discount' variables.
# However, it's important to note that the model explains only a portion of the
# variability in 'Profit', and there may be other factors influencing profitability t
# hat are not captured in the model. Additionally, further analysis and validation
# may be necessary to assess the accuracy and reliability of the predictions.
```

Figure 9: Input for Linear Regression for Question 4

OUTPUT:

```
> # Load the dataset from the CSV file
> superstore_sample <- read.csv("Sample - Superstore.csv")
> data <- superstore_sample
> # Subset the data for each product category
> furniture_sales <- data[data$Category == "Furniture", "Sales"]
> office_supplies_sales <- data[data$Category == "Office Supplies", "Sales"]
> technology_sales <- data[data$Category == "Technology", "Sales"]
> # Perform one-way ANOVA
> anova_result <- aov(Sales ~ Category, data = data)
> # Summarize the ANOVA results
> 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 ‘.’ 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

> |
```

Figure 10: Output for Linear Regression for Question 4

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.

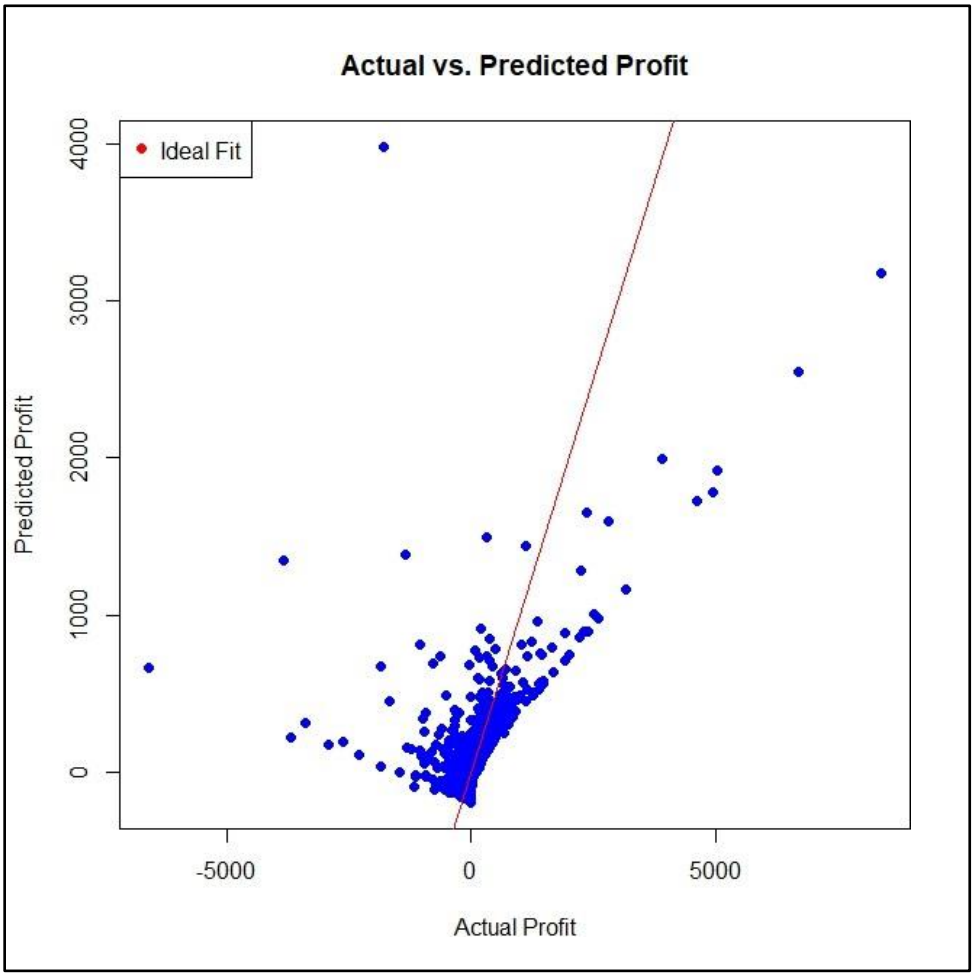

Figure 11: Scatterplot of Predicted and Actual Profit

Question 5: Relationship between Discount and Likelihood of High-Value Orders (Using Logistic Regression):

INPUT:

```
# Define a binary outcome variable 'High' indicating whether an order is considered 'High' value
# Let's assume orders with Sales greater than $500 are considered 'High' value
threshold <- 500
superstore_sample <- mutate(superstore_sample, High = ifelse(Sales > threshold, 1, 0))

# Fit logistic regression model
logit_model <- glm(High ~ Discount, data = superstore_sample, family = "binomial")

# Summary of the logistic regression model
summary(logit_model)

# The coefficient for the 'Discount' variable is -0.28515.
# This indicates that for each unit increase in the discount offered,
# the log odds of a customer placing a 'High' value order decreases by 0.28515.
# The p-value associated with the 'Discount' variable is 0.0697, which is greater
# than the conventional significance level of 0.05. Therefore, we fail to reject the null
# hypothesis at the 0.05 significance level, suggesting that there may not be a
# statistically significant relationship between the discount offered on a
# product and the likelihood of a customer placing a 'High' value order.
```

Figure 12: Input for Logistic Regression for Question 5

OUTPUT:

```
> # Define a binary outcome variable 'High' indicating whether an order is considered 'High' value
> # Let's assume orders with Sales greater than $500 are considered 'High' value
> threshold <- 500
> superstore_sample <- mutate(superstore_sample, High = ifelse(Sales > threshold, 1, 0))
>
> # Fit logistic regression model
> logit_model <- glm(High ~ Discount, data = superstore_sample, family = "binomial")
>
> # Summary of the logistic regression model
> summary(logit_model)

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

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
```

Figure 13: Output for Logistic Regression for Question 5

In our analysis of the relationship between discounts offered on products and the likelihood of customers placing high-value orders, we employed logistic regression modeling. The results indicate that the coefficient for the 'Discount' variable is -0.28515, suggesting that for each unit increase in the discount offered, the log odds of a customer placing a high-value order decreases by 0.28515. However, the associated p-value of 0.0697 exceeds the conventional significance level of 0.05, leading to the failure to reject the null hypothesis. This suggests that there may not be a statistically significant relationship between the discount offered on a product and the likelihood of a customer placing a high-value order at the 0.05

significance level. Despite this, the p-value is in close proximity to 0.05, indicating borderline significance. Consequently, further investigation or additional data may be warranted to draw definitive conclusions regarding the impact of discounts on customer purchasing behavior.

The bar plot created to visualize the proportion of 'High' value orders across different discount levels further supports these findings. The plot illustrates how the proportion of 'High' value orders fluctuates with varying discount levels, showing a decreasing trend as the discount increases. This alignment between the logistic regression results and the bar plot reinforces the notion that higher discounts may not necessarily correlate with a higher likelihood of customers placing 'High' value orders.

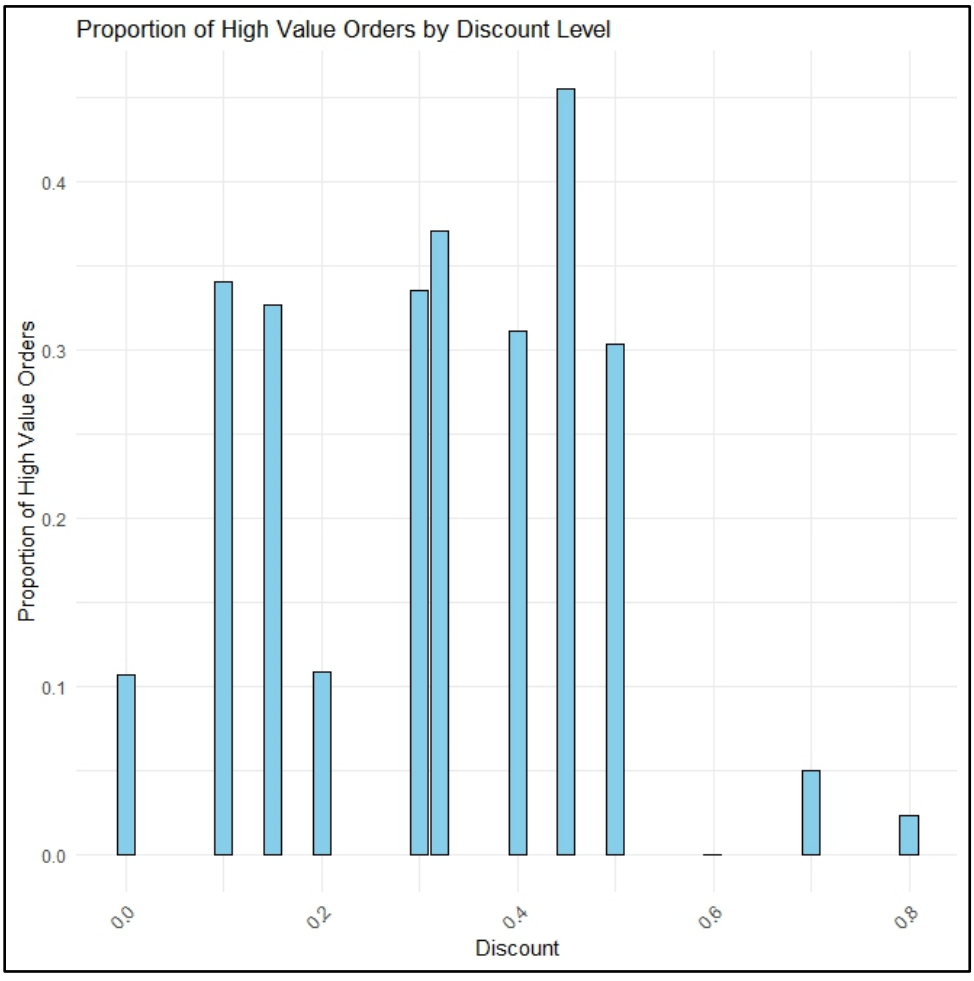

Figure 14: Bar Plot of Proportion of High Value Orders by Discount Level

Question 6: Identifying Predictive Combination of Independent Variables for Sales Performance (Using Subset Regression):

INPUT:

```
library(leaps)
# Assuming 'Sales', 'Product_Category', 'Ship_Mode', and 'Region' are relevant variables
best_subset <- regsubsets(Sales ~ Category + Ship.Mode + Region, data = data)
summary(best_subset)

# Plotting the results
plot(best_subset, scale = "adjr2")  # Adjusted R-squared
```

Figure 15: Input for Subset Regression for Question 6

OUTPUT:

```
> # Assuming 'Sales', 'Product_Category', 'Ship_Mode', and 'Region' are relevant variables
> best_subset <- regsubsets(Sales ~ Category + Ship.Mode + Region, data = data)
> summary(best_subset)
Subset selection object
Call: regsubsets.formula(Sales ~ Category + Ship.Mode + Region, data = data)
8 Variables  (and intercept)
                        Forced in Forced out
CategoryOffice Supplies     FALSE      FALSE
CategoryTechnology          FALSE      FALSE
Ship.ModeSame Day           FALSE      FALSE
Ship.ModeSecond Class       FALSE      FALSE
Ship.ModeStandard Class     FALSE      FALSE
RegionEast                  FALSE      FALSE
RegionSouth                 FALSE      FALSE
RegionWest                  FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         CategoryOffice Supplies CategoryTechnology Ship.ModeSame Day Ship.ModeSecond Class Ship.ModeStandard Class RegionEast RegionSouth RegionWest
1  ( 1 ) "*"                     " "                " "               " "                   " "                     " "        " "         " "
2  ( 1 ) "*"                     "*"                " "               " "                   " "                     " "        " "         " "
3  ( 1 ) "*"                     "*"                " "               " "                   " "                     " "        "*"         " "
4  ( 1 ) "*"                     "*"                " "               " "                   " "                     "*"        "*"         " "
5  ( 1 ) "*"                     "*"                " "               " "                   " "                     "*"        "*"         "*"
6  ( 1 ) "*"                     "*"                " "               "*"                   " "                     "*"        "*"         "*"
7  ( 1 ) "*"                     "*"                "*"               "*"                   " "                     "*"        "*"         "*"
8  ( 1 ) "*"                     "*"                "*"               "*"                   "*"                     "*"        "*"         "*"
>
> # Plotting the results
> plot(best_subset, scale = "adjr2")  # Adjusted R-squared
>
```

Figure 16: Output for Subset Regression for Question 6

We explored various combinations of predictors, including 'Category' (specifically Office Supplies and Technology), 'Ship_Mode' (including Same Day, Second Class, and Standard Class), and 'Region' (East, South, and West) to identify the most predictive variables for 'Sales' performance. The exhaustive search considered all possible subsets of these predictors up to a maximum of 8 variables. The results revealed that the best-performing model, with the highest adjusted R-squared value, includes all predictors ('Category: Office Supplies', 'Category: Technology', 'Ship Mode: Same Day', 'Ship Mode: Second Class', 'Ship Mode: Standard Class', 'Region: East', 'Region: South', 'Region: West').

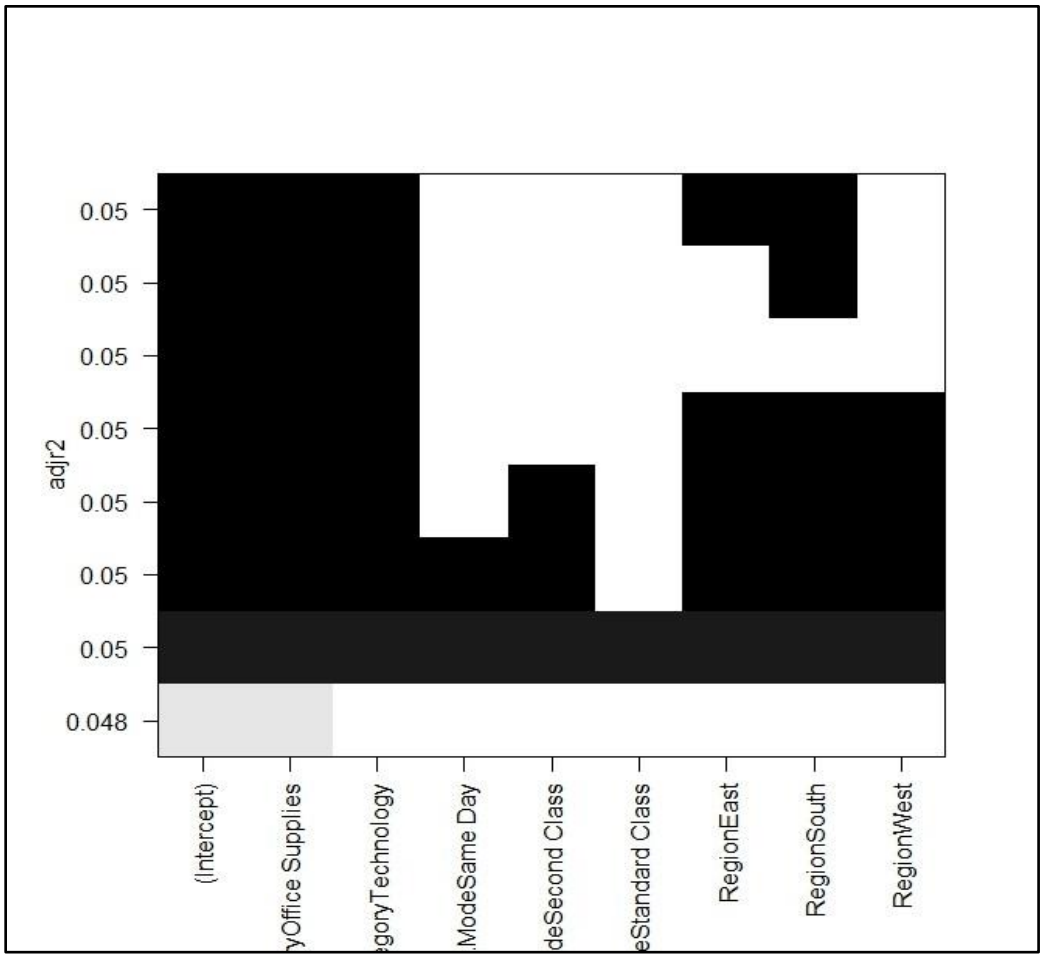

Figure 17: Output Plot for Subset Regression for Question 6

In the subset regression analysis conducted on the dataset, Model 4 emerged with the highest adjusted R-squared value. This indicates that among the models tested, Model 4, which includes predictors related to the 'Product Category,' 'Ship Mode,' and 'Region,' explains a significant portion of the variability observed in the 'Sales' data. Specifically, the inclusion of predictors related to different product categories, shipping modes, and regions enhances the model's ability to capture the variation in sales performance across different segments. Therefore, based on the subset regression analysis, these variables appear to be the most relevant in predicting sales performance in our dataset.

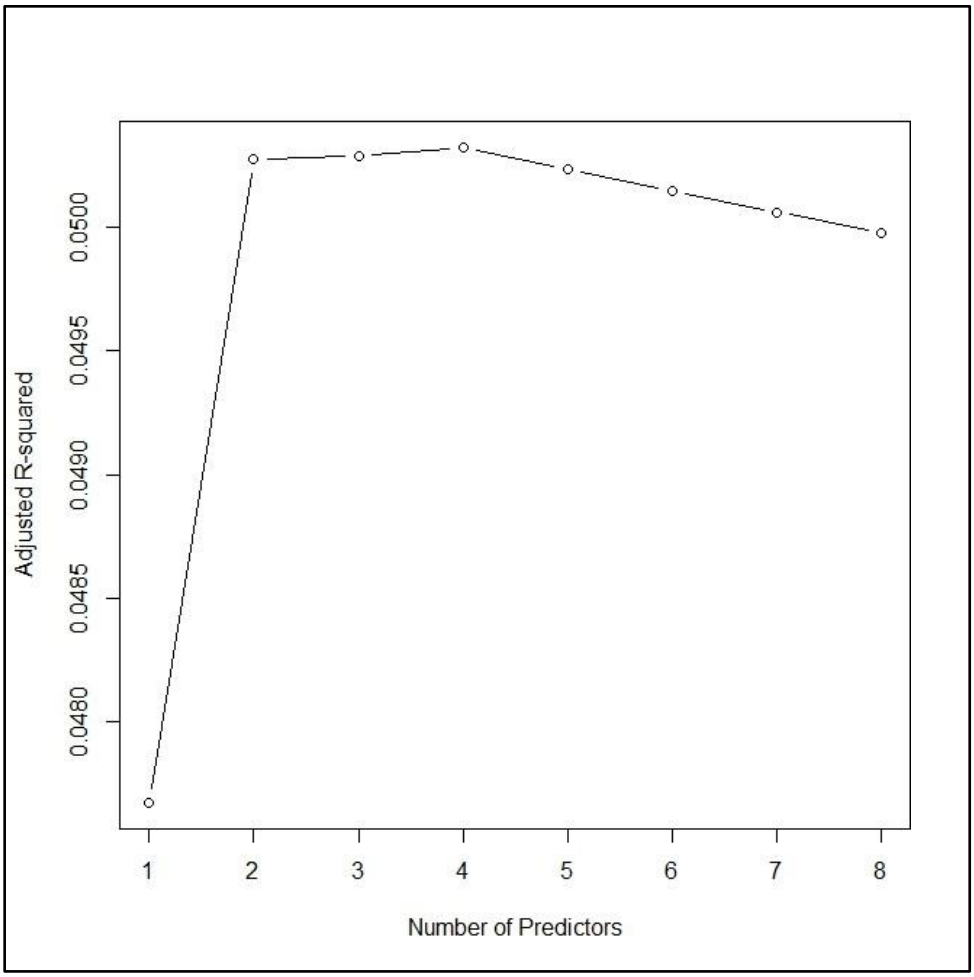

Figure 18: Plot of Adjusted R-Squared and Number of Predictors

The plotted results from the subset regression analysis reaffirm our findings regarding the model with the highest adjusted R-squared value. As observed in the plot, the adjusted R-squared values are plotted against the number of predictors included in each model. Notably, Model 4 exhibits the highest adjusted R-squared value among all the models tested. This consistency between the subset regression analysis and our previous determination further strengthens our confidence in the relevance and predictive capability of the variables included in Model 4. It confirms that the combination of predictors related to 'Product Category,' 'Ship Mode,' and 'Region' contributes significantly to explaining the variability in sales performance captured by the model. Therefore, this alignment reinforces the importance of these variables in understanding and predicting sales outcomes in our dataset.

SECTION 4: CONCLUSION

The analysis conducted on the dataset from the Superstore Giant has provided valuable insights into various aspects of their business operations, shedding light on critical factors influencing sales performance, profitability, and customer behavior. Through a series of statistical analyses and data visualizations, we addressed several pertinent business questions and derived actionable intelligence to guide strategic decision-making within the organization.

Product Category Analysis:

Our analysis revealed significant variations in mean sales across different product categories, with Technology products demonstrating the highest sales figures, followed by Furniture and Office Supplies. This suggests that the choice of product category significantly influences sales performance within the Superstore Giant's operations. Strategic initiatives aimed at optimizing sales and maximizing revenue should consider these disparities in product category performance.

Shipping and Regional Impact on Profitability:

While individual factors such as 'Region' exhibited a marginally significant effect on profitability, the absence of a significant interaction effect between 'Ship Mode' and 'Region' underscores the uniformity of the impact of shipping modes across various regions. This insight allows for streamlined decision-making processes and resource allocation strategies, enhancing operational efficiency and profitability.

Customer Segmentation and Shipping Preferences:

The significant association between customer segments and preferred shipping modes underscores the importance of understanding and catering to the diverse needs and preferences of various customer segments. By tailoring shipping strategies to different customer segments, the Superstore Giant can enhance customer satisfaction, ultimately driving long-term profitability.

Predictive Modeling for Profitability:

Our predictive model for profitability based on sales, quantity, and discount variables demonstrated significant results, with sales and discounts emerging as key drivers of profit. The model's ability to accurately estimate profit based on these predictor variables provides valuable insights for pricing strategies, inventory management, and promotional activities.

Discount Impact on Purchase Behavior:

While the relationship between discounts offered on products and the likelihood of customers placing high-value orders showed borderline significance, the bar plot visualization revealed a decreasing trend in the proportion of high-value orders with increasing discount levels. This suggests that higher discounts may not necessarily correlate with a higher likelihood of customers placing high-value orders, highlighting the need for strategic pricing strategies to maximize profitability while maintaining customer engagement.

Determinants of Sales Performance:

The subset regression analysis identified the most predictive combination of independent variables influencing sales performance, with Model 4, including predictors related to product category, shipping mode, and region, exhibiting the highest adjusted R-squared value. This underscores the relevance and predictive capability of these variables in understanding and predicting sales outcomes within the Superstore Giant's dataset.

The findings from our analysis provide valuable insights and actionable intelligence for the Superstore Giant to optimize business operations, enhance profitability, and drive sustainable growth in a competitive market landscape. By leveraging data-driven decision-making and strategic initiatives informed by our analysis, the organization can better meet customer needs, capitalize on market opportunities, and maintain a competitive edge in the industry.

SECTION 5: REFERENCES

1. Chowdhury, V. (2022, February 17). Superstore dataset. Kaggle. https://www.kaggle.com/datasets/vivek468/superstore-dataset-final/data