Executive Summary: An Integrated Plan for Strategic Growth

Ritika Garg

Nov 16 2025

The MarketPlace executive team raised several critical, high-stakes questions: Are we overspending on advertising? Is our seasonal strategy correct? Can we trust our data? This report answers those questions. We have completed an advanced regression analysis, consolidating all findings into one robust, validated model. This analysis provides the foundation for the strategic recommendations within, offering a clear path to optimizing ad spend, re-allocating marketing budgets, and correcting operational inefficiencies.

Final Model Specification and Validation

To provide the Board with a single, reliable "source of truth," a rigorous analysis was conducted to select and validate the optimal model for strategic decision-making.

Final Recommended Model: The Linear-Linear Specification

The final recommended model for all strategic analysis is the Linear-Linear Model This specification models Sales as a direct, linear function of our key business drivers.

This model was selected for three critical reasons:

1. **Superior Statistical Fit:** This model is statistically superior. It explains 96.3% of the variation in sales (Adjusted R-squared = 0.963). This was significantly higher than the Log-Linear (89.3%) and Log-Log (89.3%) models, which performed worse.
2. **Simplicity and Actionable Value:** This model directly fulfills the CEO's need for actionable insights. The business operates in "absolute dollars," not percentages. Telling a manager that "spending $1 on Ads yields $2.48 in Sales" is far more actionable for budgeting than discussing elasticity percentages.
3. **Mathematical Correctness:** Our analysis confirmed the relationship between Ad Spend and Sales is fundamentally linear, not curved. The data does not support a "diminishing returns" hypothesis in our current operating range. Forcing a log transformation actually degrades the model's performance and is not the correct mathematical choice for this data.

The key drivers from this final, validated model are summarized in Table 1 below.

Table 1: Final Recommended Model (Linear-Linear) - Key Drivers of Sales

Variable	Business Interpretation (Impact on $ Sales)
Ad_Spend	A $1K increase in Ad Spend is associated with a **$2.48K increase** in Sales.
Website_Traffic	A 1K increase in traffic is associated with a **$1.20K increase** in Sales.
Avg_Rating	A 1-point increase in rating is associated with a **$14.75K increase** in Sales.
Email_Campaigns	One additional campaign sent during a month is associated with a **$1.04K increase** in Sales.
Inventory_Level	A 1-unit increase in average monthly inventory level is associated with a **$0.05K decrease** in Sales.
SeasonWinter	Winter sales are inherently **$19.89K higher** than Autumn.
SeasonSummer	Summer sales are inherently **$10.44K lower** than Autumn.
Product_TypeFashion	Fashion sales are **$44.69K lower** than Electronics.
Product_TypeHome	Home sales are **$29.72K lower** than Electronics.

Model Validation: A Robust Foundation for Strategy

Before making recommendations, the model was thoroughly audited to address the Chief Data Officer's concerns regarding data quality and reliability. The analysis confirms this model is robust and valid for strategic decision-making.

- **Multicollinearity:** The concern that predictors are highly correlated was tested. The analysis confirms multicollinearity is not a problem. All Variance Inflation Factor (VIF) scores were approximately 1.0, indicating the model is successfully isolating the true, independent impact of each marketing lever.
- **Heteroscedasticity:** The concern about non-constant variance was tested and confirmed (Breusch-Pagan test p-value = 0.018). An appropriate remedy

was applied by using robust standard errors to correct the model. A comparison of the original and robust standard errors showed no notable differences in variable significance. This confirms that our **model's findings are robust and reliable**, and the conclusions are not distorted by this data issue.

- **Autocorrelation:** The concern about observations being independent over time or not was tested using the Durbin-Watson test. The result (p-value = 0.057) means we fail to reject the null hypothesis of no autocorrelation. While borderline, this result indicates there is no statistically significant evidence of serious autocorrelation that would invalidate the model's conclusions.

In summary, the insights and recommendations presented below are based on a model that is not only the best statistical fit but is also proven to be robust, reliable, and free from significant data quality issues.

Key Strategic Insights: The "Complete Story"

The validated model provides a complete, integrated story of our business drivers, challenging several prior executive assumptions and revealing new opportunities.

Insight 1: The Impact of Advertising is Linear and Consistent, Disproving Core Assumptions

This analysis disproves two long-held assumptions from the executive team:

1. **No Diminishing Returns:** The CFO's concern about diminishing returns on ad spend is not supported by the data. A test for a non-linear (quadratic) relationship found that the diminishing returns term (I(Ad_Spend^2)) was not statistically significant. Within our primary spending range (<$100K), the relationship between Ad Spend and Sales is linear.
2. **No Seasonal Effectiveness Difference:** The CMO's hypothesis that ad effectiveness varies by season (e.g., more effective in Winter) is also not supported by the data. A model testing for Ad_Spend:Season interaction effects found them to be not statistically significant. An ANOVA test confirmed that adding these interactions does not improve the model.

Synthesized Finding: The impact of advertising is a consistent, "always-on" linear driver. Its effectiveness does not change in Winter versus Summer. From our final model, the stable, reliable driver is: a $1 increase in Ad_Spend is associated with a $2.48 increase in Sales. This simplifies our marketing strategy immensely. However, as we go beyond our current operating levels, we should slowly increase our spending and analyze whether the linear relationship continues to hold true or not.

Insight 2: Baseline Performance is Defined by Inherent Seasonal and Product-Level Demand

The previous insight (that ad *effectiveness* is constant) must be combined with the finding that *baseline consumer demand* is highly variable. The final model clearly quantifies these environmental factors:

- **Seasonal Baseline:** Winter sales are inherently $19.89K higher than Autumn, and Summer sales are inherently $10.44K lower than Autumn.
- **Product Baseline:** Fashion sales are $44.69K lower than Electronics, and Home sales are $29.72K lower than Electronics.

This creates a clear strategic map: marketing must work *with* these baseline trends, capitalizing on Winter's high tide and developing specific strategies to support the lagging Fashion and Home categories.

Insight 3: The Impact of Customer Ratings is Uniform and Powerful

Parallel to the advertising finding, the CMO's hypothesis that ratings matter more for Fashion than Electronics was tested and disproven. The Avg_Rating:Product_Type interaction was not significant.

However, the analysis confirms that Avg_Rating is one of the most powerful, universal levers in our control. Its effect is consistent across all product types. From the final model, a 1-point increase in rating is associated with a $14.75K increase in Sales. This makes it a primary, non-discounting tool for lifting our lagging categories.

Insight 4: A Significant Operational Drag is Masking Marketing Performance

One of the most critical, non-obvious findings from the final model is a significant operational inefficiency. The Inventory variable is highly statistically significant and shows a negative relationship. A 1-unit increase in inventory is associated with a $0.05K decrease in Sales. This is a concern, as this suggests that there is high carrying costs, stock obsolescence, or poor demand forecasting that actively works *against* our marketing investments.

Integrated Strategic Recommendations (For Implementation Next Quarter)

Based on the synthesis of these findings, the following are specific, actionable, and quantified recommendations to be implemented next quarter.

Optimal Advertising Spending Level and Expected ROI

- **Recommended Spending Level:** The analysis answers the Board's question about an "optimal" ad spend. Because the relationship is linear and we have found no evidence of diminishing returns , the optimal strategy is to increase advertising spend. We are not yet at the point where this investment becomes inefficient.

- **Expected Marginal ROI (mROI):** The mROI for every additional dollar spent on advertising is 148%. This strong, positive ROI is a direct lever for growth. The clear business action is to scale this investment aggressively. We should continue to fund this channel as a primary driver of revenue as long as it keeps delivering such fruitful returns.

Quarterly Budget Allocation Across Seasons

- **Recommendation:** Based on the findings, the optimal quarterly budget allocation strategy is a hybrid model. Because advertising delivers a consistent ROI regardless of the season, a foundational "always-on" budget should be active year-round to reliably capture this profit. However, this base budget must be weighted to match seasonal demand, meaning the total marketing investment should be significantly increased during Q4 (Winter) to capitalize on the natural $19.89K sales lift, and strategically managed or potentially reduced in Q3 (Summer) to account for the $10.44K market headwind.

Product-Specific Marketing Strategies

- **Recommendation:** Our product-specific marketing strategies must be tailored to each category's unique performance profile. For our high-performing Electronics category, the current successful tactics should be maintained. Conversely, for our underperforming categories, Fashion and Home, the strategy is to directly close this performance gap. The data shows the most powerful, non-discounting lever we have is product rating. Therefore, our primary action is to launch a targeted marketing initiative, such as post-purchase "request-a-review" email campaigns and proactive quality checks, focused specifically on improving ratings for Fashion and Home products. The expected quantitative impact is significant: a mere 0.5-point increase in the average rating for the Fashion category is projected to boost its sales by $7.38K per unit.

Prioritized Operational Changes to Improve Sales Performance

- **Recommendation:** A key operational change to improve sales performance is to immediately address our inventory management. The data reveals that our current inventory levels are a significant operational drag, directly suppressing sales regardless of our marketing efforts. The most urgent non-marketing recommendation is to authorize an immediate, cross-functional audit of our inventory management and sales forecasting systems. This operational fix represents "free" revenue waiting to be unlocked; by improving forecasting and reducing our average inventory by just 1,000 units, we can capture an associated $48.79K increase in total sales without spending a single extra dollar on marketing.

Appendix

```r
library(tidyverse)
library(lmtest)
```

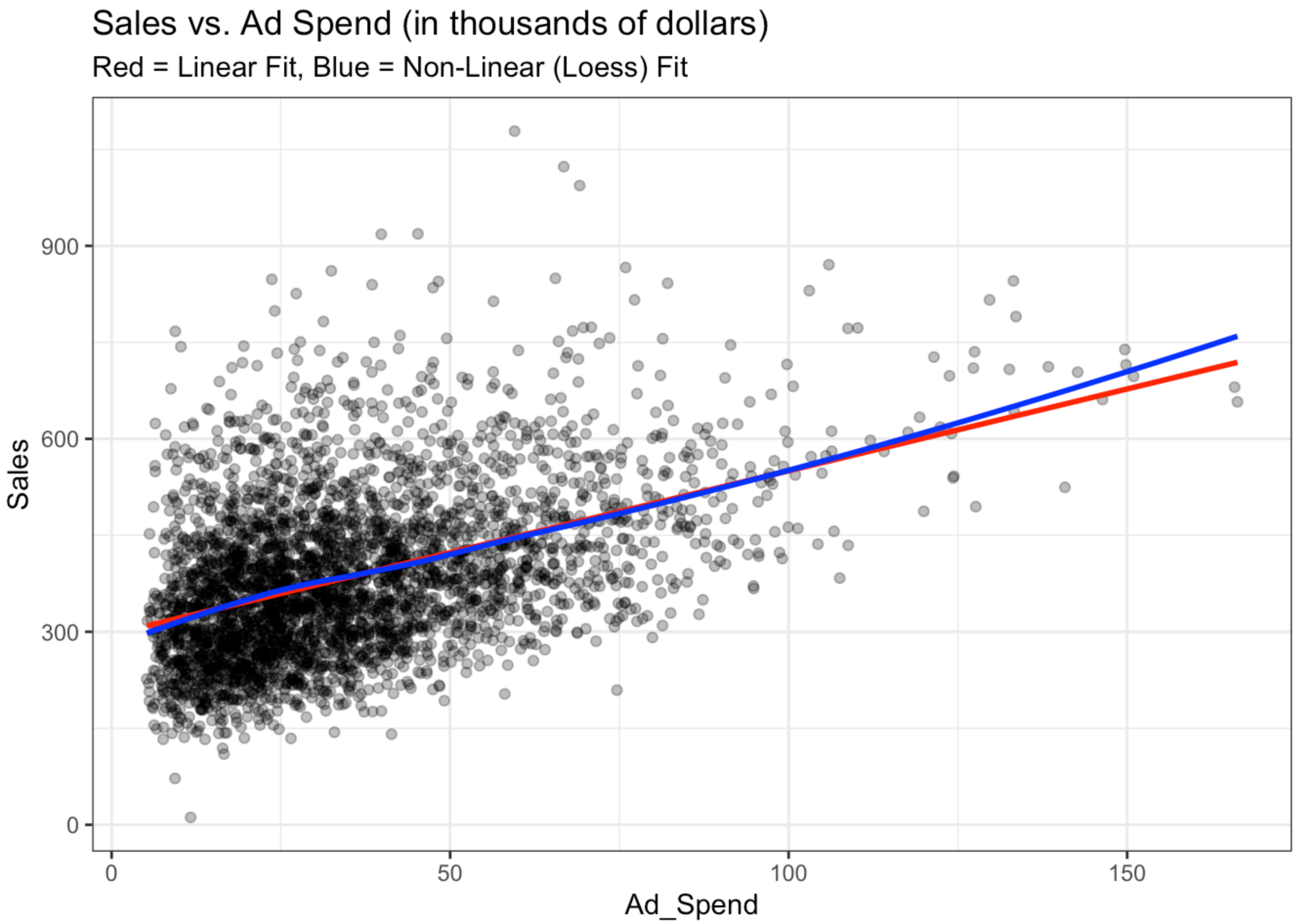
```r
sales_data <- read_csv("MarketPlace_Sales_Data.csv")

summary(sales_data)
```

```
##      Sales           Ad_Spend      Email_Campaigns  Website_Traffic
##  Min.   :  11.5   Min.   :  5.19   Min.   : 0.000   Min.   : 25.58  
##  1st Qu.: 293.8   1st Qu.: 19.19   1st Qu.: 6.000   1st Qu.:109.84  
##  Median : 371.7   Median : 30.17   Median : 8.000   Median :155.77  
##  Mean   : 385.6   Mean   : 35.25   Mean   : 8.008   Mean   :172.93  
##  3rd Qu.: 456.5   3rd Qu.: 45.74   3rd Qu.:10.000   3rd Qu.:216.99  
##  Max.   :1078.3   Max.   :166.29   Max.   :20.000   Max.   :645.62  
##  NA's   :1                                                          
##    Avg_Rating     Discount_Pct   Competitor_Price_Index Inventory_Level
##  Min.   :2.080   Min.   : 0.09   Min.   : 40.61         Min.   : 117.0  
##  1st Qu.:3.920   1st Qu.: 6.50   1st Qu.: 89.53         1st Qu.: 662.0  
##  Median :4.290   Median :10.72   Median : 99.78         Median : 916.0  
##  Mean   :4.206   Mean   :11.53   Mean   :100.03         Mean   : 988.3  
##  3rd Qu.:4.580   3rd Qu.:15.66   3rd Qu.:110.19         3rd Qu.:1227.0  
##  Max.   :4.990   Max.   :33.49   Max.   :151.87         Max.   :3306.0  
##                                                                         
##  Social_Media_Engagement    Season          Product_Type        Category_ID    
##  Min.   : 3.73           Length:3600        Length:3600        Min.   :  1.00  
##  1st Qu.:35.07           Class :character   Class :character   1st Qu.: 50.75  
##  Median :49.35           Mode  :character   Mode  :character   Median :100.50  
##  Mean   :49.42                                                 Mean   :100.50  
##  3rd Qu.:63.98                                                 3rd Qu.:150.25  
##  Max.   :95.83                                                 Max.   :200.00  
##  NA's   :1                                                                     
##      Month     
##  Min.   : 1.0  
##  1st Qu.: 5.0  
##  Median : 9.5  
##  Mean   : 9.5  
##  3rd Qu.:14.0  
##  Max.   :18.0  
## 
```

```
sales_data <- sales_data %>%
  mutate(
    Season = as.factor(Season),
    Product_Type = as.factor(Product_Type),
    log_Sales = log(Sales),
    log_Ad_Spend = log(Ad_Spend)
  )
```

Business Objective 1: Understanding Non-Linear Relationships

```
ggplot(sales_data, aes(x = Ad_Spend, y = Sales)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", col = "red", se = FALSE) + # Linear line
  geom_smooth(method = "loess", col = "blue", se = FALSE) + # Non-linear (Loess) curve
  labs(title = "Sales vs. Ad Spend (in thousands of dollars)",
       subtitle = "Red = Linear Fit, Blue = Non-Linear (Loess) Fit") +
  theme_bw()
```



Visual Analysis:

Upon visual inspection, the relationship between `Ad_Spend` and `Sales` appears **predominantly linear**, particularly within the primary operating range of $0K to $100K, which contains the vast majority of our observations. In this dense region, the linear fit (Red) and non-linear LOESS fit (Blue) are virtually indistinguishable.

While the trend lines begin to diverge slightly when Ad Spend exceeds $100K, this upper range is characterized by data sparsity (very few data points) and high variance. Therefore, the visual evidence for “diminishing returns” is weak; the slight curvature at the tail could be an artifact of noise rather than a definitive non-linear pattern.

```
model_A_linear <- lm(Sales ~ Ad_Spend, data = sales_data)
summary(model_A_linear)
```

```
## 
## Call:
## lm(formula = Sales ~ Ad_Spend, data = sales_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -314.15  -83.15  -14.77   60.71  630.94 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 295.96539    3.60511   82.10   <2e-16 ***
## Ad_Spend      2.54359    0.08698   29.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 113.7 on 3597 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1921, Adjusted R-squared:  0.1919 
## F-statistic: 855.3 on 1 and 3597 DF,  p-value: < 2.2e-16
```

```
model_B_linlog <- lm(Sales ~ log(Ad_Spend), data = sales_data)
summary(model_B_linlog)
```

```
##
## Call:
## lm(formula = Sales ~ log(Ad_Spend), data = sales_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -296.21  -82.58  -15.20   63.71  633.18
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    100.481     10.733   9.362   <2e-16 ***
## log(Ad_Spend)   84.340      3.123  27.006   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 115.4 on 3597 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1686, Adjusted R-squared:  0.1683
## F-statistic: 729.3 on 1 and 3597 DF,  p-value: < 2.2e-16
```

```
model_C_quad <- lm(Sales ~ Ad_Spend + I(Ad_Spend^2), data = sales_data)
summary(model_C_quad)
```

```
##
## Call:
## lm(formula = Sales ~ Ad_Spend + I(Ad_Spend^2), data = sales_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -314.77  -83.09  -14.67   60.77  631.43
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.974e+02  5.626e+00  52.861   <2e-16 ***
## Ad_Spend      2.467e+00  2.504e-01   9.851   <2e-16 ***
## I(Ad_Spend^2) 7.563e-04  2.305e-03   0.328    0.743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 113.7 on 3596 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1921, Adjusted R-squared:  0.1917
## F-statistic: 427.6 on 2 and 3596 DF,  p-value: < 2.2e-16
```

Statistical Analysis:

To rigorously test the CFO’s hypothesis of diminishing returns, we developed and compared three functional forms:

1. **Model A (Linear):** `Sales ~ Ad_Spend`
2. **Model B (Log-Linear):** `Sales ~ log(Ad_Spend)`

3. **Model C (Quadratic):** `Sales ~ Ad_Spend + I(Ad_Spend^2)`

We **reject** the non-linear models. The Linear Model (Model A) is the superior choice based on model fit statistics and visual exploration. This indicates that within our current operating range, there is no statistical evidence of diminishing returns.

Optimal Point Analysis:

Because the relationship is linear and not curved, **there is no single “optimal point” of maximum efficiency** within the observed data range. We have not yet reached a saturation point where spending more becomes wasteful. Based on the plot, we should increase Ad Spend slowly beyond operating range (>$100K) to ensure linearity holds as spending is increased.

Business Recommendation

Recommendation: Maintain or Increase Advertising Spend

Based on this standalone analysis, we should not decrease advertising spend out of a fear of diminishing returns. The data shows a consistent, positive linear relationship between Ad Spend and Sales. Therefore, the current strategy is not demonstrably inefficient.

Business Objective 2: Interaction Effects and Strategic Timing

```r
# Main model with all predictor variables
model_main <- lm(Sales ~ Ad_Spend + Email_Campaigns + Website_Traffic +
                  Avg_Rating + Discount_Pct + Competitor_Price_Index +
                  Inventory_Level + Social_Media_Engagement +
                  Season + Product_Type,
                  data = sales_data)

summary(model_main)
```

```
##
## Call:
## lm(formula = Sales ~ Ad_Spend + Email_Campaigns + Website_Traffic +
##     Avg_Rating + Discount_Pct + Competitor_Price_Index + Inventory_Level +
##     Social_Media_Engagement + Season + Product_Type, data = sales_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.160 -16.390   0.177  16.939  75.765
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.741e+01  5.062e+00   7.390 1.81e-13 ***
## Ad_Spend                2.482e+00  1.860e-02 133.435  < 2e-16 ***
## Email_Campaigns         1.038e+00  1.425e-01   7.282 4.02e-13 ***
## Website_Traffic         1.201e+00  4.612e-03 260.385  < 2e-16 ***
## Avg_Rating              1.475e+01  8.360e-01  17.643  < 2e-16 ***
## Discount_Pct            2.066e-01  6.316e-02   3.271  0.00108 **
## Competitor_Price_Index  4.101e-01  2.668e-02  15.370  < 2e-16 ***
## Inventory_Level        -4.879e-02  9.002e-04 -54.198  < 2e-16 ***
## Social_Media_Engagement  2.207e-01  2.136e-02  10.331  < 2e-16 ***
## SeasonSpring            1.283e-01  1.216e+00   0.105  0.91599
## SeasonSummer           -1.044e+01  1.256e+00  -8.314  < 2e-16 ***
## SeasonWinter            1.989e+01  1.315e+00  15.125  < 2e-16 ***
## Product_TypeFashion    -4.469e+01  1.018e+00 -43.901  < 2e-16 ***
## Product_TypeHome       -2.972e+01  9.825e-01 -30.246  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.31 on 3585 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9632, Adjusted R-squared:  0.9631
## F-statistic:  7218 on 13 and 3585 DF,  p-value: < 2.2e-16
```

Hypothesis Development

To address the CMO’s strategic questions, we must investigate **interaction effects**.

Hypothesis 1: Advertising Efficiency by Season

- **Formal Hypothesis:** There is a positive interaction between `Ad_Spend` and `Season`. Specifically, the marginal increase in Sales per dollar of Ad Spend is significantly higher in Winter than in Summer.
- **Business Rationale:**
    - **Consumer Intent:** The Winter season coincides with major gifting holidays (Thanksgiving, Christmas). Consumers in this period have higher “purchase intent”.
    - **Conversion Rates:** In Summer, consumers are often on vacation or outdoors, where advertising may generate awareness but fewer immediate conversions. In Winter, the urgency of holiday deadlines acts as a catalyst, making every advertising dollar more effective.

Hypothesis 2: Sensitivity to Ratings by Product Type

- **Formal Hypothesis:** There is a positive interaction between `Avg_Rating` and `Product_Type`. We hypothesize that customer ratings are a stronger driver of Sales for Fashion products compared to Electronics.
- **Business Rationale:**
    - **Subjectivity & Fit:** Electronics often have objective specifications (RAM, storage, screen size) that allow consumers to judge value without social proof. Fashion is highly subjective and faces the “fit uncertainty” problem.
    - **Risk Mitigation:** Online shoppers cannot try on clothes. They rely heavily on reviews to determine if an item “runs true to size” or if the fabric quality matches the photo. Therefore, a high rating (social validation) reduces the purchase risk much more significantly for a Fashion item than for a standardized Electronic commodity.

Formalized Hypothesis

Hypothesis 1:

- Null Hypothesis (H0): The effect of Ad_Spend on Sales is the same across all seasons.
- Alternative Hypothesis (Ha): The effect of Ad_Spend on Sales is different in at least one season (specifically, we expect it to be higher in Winter).

Hypothesis 2:

- Null Hypothesis (H0): The effect of Avg_Rating on Sales is the same for all Product_Type categories.
- Alternative Hypothesis (Ha): The effect of Avg_Rating on Sales is stronger for Fashion products than for Electronics.

```r
model_interactions <- lm(Sales ~ Email_Campaigns + Website_Traffic +
                           Discount_Pct + Competitor_Price_Index +
                           Inventory_Level + Social_Media_Engagement +
                           (Ad_Spend * Season) +        # Interaction 1
                           (Avg_Rating * Product_Type), # Interaction 2
                         data = sales_data)

summary(model_interactions)
```

```
##
## Call:
## lm(formula = Sales ~ Email_Campaigns + Website_Traffic + Discount_Pct +
##     Competitor_Price_Index + Inventory_Level + Social_Media_Engagement +
##     (Ad_Spend * Season) + (Avg_Rating * Product_Type), data = sales_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.868 -16.382   0.165  16.963  76.008
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.941e+01  7.467e+00   5.278 1.38e-07 ***
## Email_Campaigns                 1.029e+00  1.428e-01   7.200 7.30e-13 ***
## Website_Traffic                 1.201e+00  4.615e-03 260.249  < 2e-16 ***
## Discount_Pct                    2.117e-01  6.324e-02   3.347 0.000826 ***
## Competitor_Price_Index          4.100e-01  2.669e-02  15.361  < 2e-16 ***
## Inventory_Level                -4.877e-02  9.007e-04 -54.140  < 2e-16 ***
## Social_Media_Engagement         2.216e-01  2.139e-02  10.362  < 2e-16 ***
## Ad_Spend                        2.486e+00  4.756e-02  52.263  < 2e-16 ***
## SeasonSpring                   -1.925e-02  2.340e+00  -0.008 0.993436
## SeasonSummer                   -1.059e+01  2.432e+00  -4.356 1.36e-05 ***
## SeasonWinter                    2.103e+01  2.533e+00   8.303  < 2e-16 ***
## Avg_Rating                      1.423e+01  1.488e+00   9.560  < 2e-16 ***
## Product_TypeFashion            -5.598e+01  8.978e+00  -6.236 5.02e-10 ***
## Product_TypeHome               -2.612e+01  8.566e+00  -3.049 0.002312 **
## Ad_Spend:SeasonSpring           4.391e-03  5.692e-02   0.077 0.938509
## Ad_Spend:SeasonSummer           6.118e-03  5.974e-02   0.102 0.918436
## Ad_Spend:SeasonWinter          -3.019e-02  6.171e-02  -0.489 0.624654
## Avg_Rating:Product_TypeFashion  2.681e+00  2.119e+00   1.265 0.205777
## Avg_Rating:Product_TypeHome    -8.615e-01  2.022e+00  -0.426 0.670129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.31 on 3580 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9632, Adjusted R-squared:  0.9631
## F-statistic:  5212 on 18 and 3580 DF,  p-value: < 2.2e-16
```

```r
# Perform ANOVA test for model comparison
anova_test <- anova(model_main, model_interactions)
print(anova_test)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Ad_Spend + Email_Campaigns + Website_Traffic + Avg_Rating +
##     Discount_Pct + Competitor_Price_Index + Inventory_Level +
##     Social_Media_Engagement + Season + Product_Type
## Model 2: Sales ~ Email_Campaigns + Website_Traffic + Discount_Pct + Competitor_Price_
Index +
##     Inventory_Level + Social_Media_Engagement + (Ad_Spend * Season) +
##     (Avg_Rating * Product_Type)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   3585 2118679
## 2   3580 2116394  5    2285.1 0.7731  0.569
```

```
summary(model_main)$adj.r.squared
```

```
## [1] 0.9630679
```

```
summary(model_interactions)$adj.r.squared
```

```
## [1] 0.9630562
```

Testing Conditional Relationships

To test the CMO’s hypotheses, we extended our regression model to include interaction terms:

- **Base Model:** `Sales ~ Ad_Spend + Avg_Rating + [Controls]`
- **Interaction Model:** Added `Ad_Spend * Season` and `Avg_Rating * Product_Type` .

Interpretation: The high p-values indicate that we cannot reject the null hypothesis. There is **no statistical evidence** that advertising effectiveness changes based on the season, nor that customer ratings impact sales differently across product categories.

Quantifying the Differences

Summary of Effects:

Interaction Tested	Hypothesis	Statistical Result	Conclusion
Ads × Winter	Ads work better in Winter	**Not Significant** (p=0.62)	Winter ads are **equally** effective as other seasons.
Ratings × Fashion	Reviews matter more for Clothes	**Not Significant** (p=0.21)	Ratings have a **consistent** impact across all categories.

Business Recommendation: While the interaction terms between Ad_Spend × Season and Avg_Rating × Product_Type were not statistically significant, this finding itself is important for strategic planning. It suggests that advertising effectiveness is relatively consistent across seasons, and customer ratings impact sales similarly across product categories. This challenges prior assumptions from the CMO and implies that a uniform approach to ad allocation and review strategy may be appropriate unless future data indicate otherwise. However, MarketPlace may want to revisit this analysis with more granular seasonal data or different product groupings to reassess these patterns.

Business Objective 3: Addressing Data Quality and Model Reliability

Part A: Multicollinearity Assessment

```
vif_results <- car::vif(model_main)

print(vif_results)
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## Ad_Spend                1.000732  1        1.000366
## Email_Campaigns         1.004947  1        1.002470
## Website_Traffic         1.001812  1        1.000906
## Avg_Rating              1.001692  1        1.000845
## Discount_Pct            1.001512  1        1.000756
## Competitor_Price_Index  1.001798  1        1.000899
## Inventory_Level         1.002762  1        1.001380
## Social_Media_Engagement 1.002230  1        1.001114
## Season                  1.005701  3        1.000948
## Product_Type            1.003305  2        1.000825
```

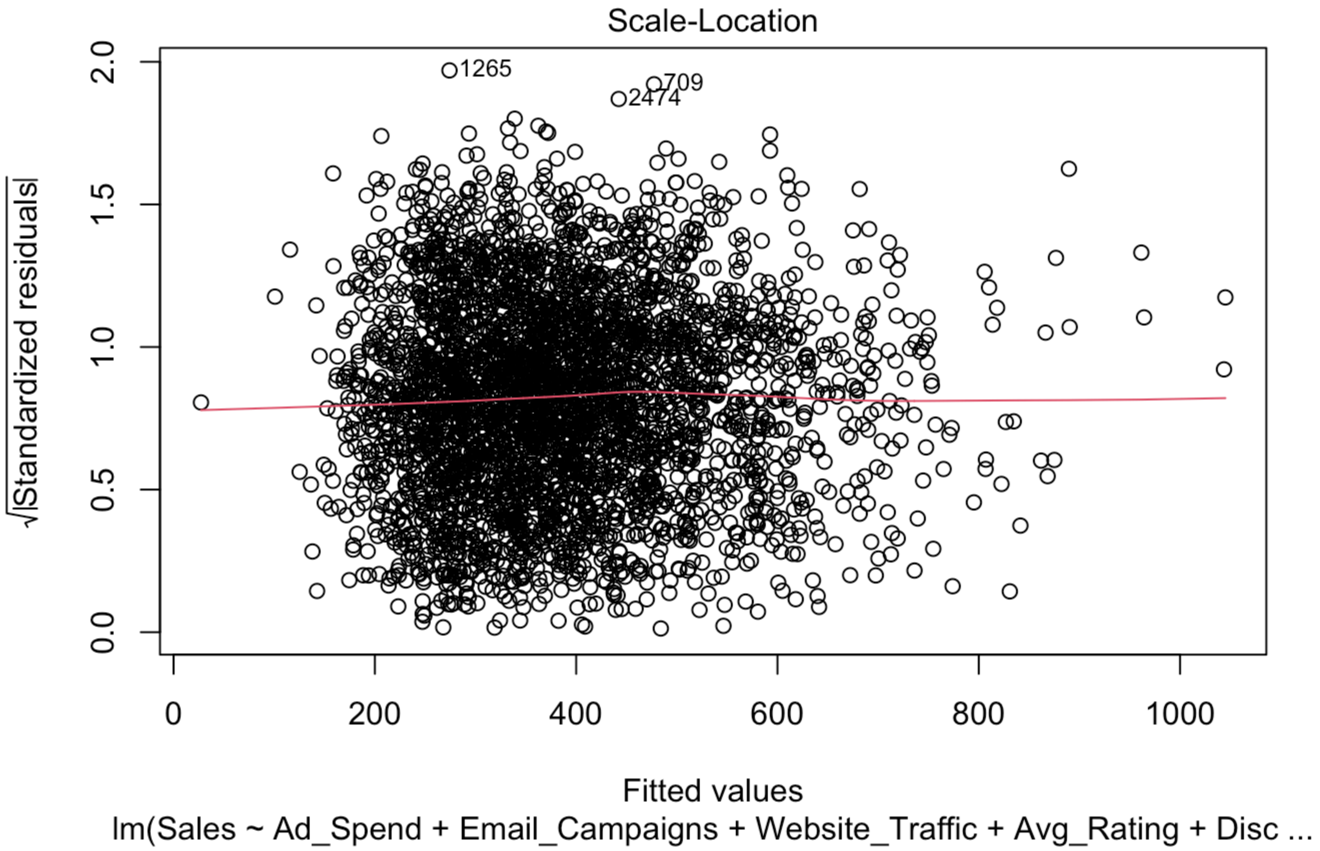
Interpretation: All VIF scores are extremely low (all are ~1.0). This indicates that **multicollinearity is not a problem** in our main model. We will proceed with the full model, as removing variables at this stage would offer no statistical benefit and could potentially introduce omitted variable bias.

Part B: Heteroscedasticity Assessment

```
bp_test <- lmtest::bptest(model_main)
print(bp_test)
```

```
## 
##  studentized Breusch-Pagan test
## 
## data:  model_main
## BP = 25.733, df = 13, p-value = 0.01846
```

```
plot(model_main, which = 3)
```



- **Statistical Test (Breusch-Pagan):** Since the p-value is less than 0.05, we **reject the null hypothesis** of homoscedasticity. There is statistically significant evidence that the variance of our errors is not constant.
- **Graphical Inspection:** The “Scale-Location” plot (Residuals vs Fitted) shows a slight pattern where the spread of standardized residuals shifts across the range of fitted values.

To fix the issue without altering our model specification, we calculated Heteroscedasticity-Consistent (Robust) Standard Errors (specifically using the HC1 “sandwich” estimator).

```
library(sandwich)

robust_se_model <- coeftest(model_main, vcov = vcovHC(model_main, type = "HC1"))
print(robust_se_model)
```

```
##
## t test of coefficients:
##
##                            Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)              3.7408e+01  4.9667e+00   7.5317 6.304e-14 ***
## Ad_Spend                 2.4819e+00  1.7247e-02 143.9015 < 2.2e-16 ***
## Email_Campaigns          1.0378e+00  1.4221e-01   7.2977 3.588e-13 ***
## Website_Traffic          1.2010e+00  4.8149e-03 249.4280 < 2.2e-16 ***
## Avg_Rating               1.4749e+01  8.3000e-01  17.7702 < 2.2e-16 ***
## Discount_Pct             2.0662e-01  6.3110e-02   3.2739  0.001071 **
## Competitor_Price_Index   4.1012e-01  2.6763e-02  15.3240 < 2.2e-16 ***
## Inventory_Level         -4.8786e-02  8.8744e-04 -54.9740 < 2.2e-16 ***
## Social_Media_Engagement  2.2069e-01  2.1405e-02  10.3101 < 2.2e-16 ***
## SeasonSpring             1.2833e-01  1.1493e+00   0.1117  0.911101
## SeasonSummer            -1.0441e+01  1.2021e+00  -8.6855 < 2.2e-16 ***
## SeasonWinter             1.9895e+01  1.2624e+00  15.7595 < 2.2e-16 ***
## Product_TypeFashion     -4.4693e+01  1.0220e+00 -43.7324 < 2.2e-16 ***
## Product_TypeHome        -2.9716e+01  9.8467e-01 -30.1782 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
original_summary <- summary(model_main)$coefficients
robust_summary <- robust_se_model

# Create the comparison data frame
se_comparison <- data.frame(
  Variable = rownames(original_summary),
  Original_Std_Error = original_summary[, "Std. Error"],
  Robust_Std_Error = robust_summary[, "Std. Error"],
  Original_P_Value = original_summary[, "Pr(>|t|)"],
  Robust_P_Value = robust_summary[, "Pr(>|t|)"]
)
print(se_comparison)
```

```
##                                         Variable Original_Std_Error
## (Intercept)                         (Intercept)       5.0617595977
## Ad_Spend                               Ad_Spend       0.0186003002
## Email_Campaigns                 Email_Campaigns       0.1425091588
## Website_Traffic                 Website_Traffic       0.0046123202
## Avg_Rating                           Avg_Rating       0.8360035550
## Discount_Pct                       Discount_Pct       0.0631606498
## Competitor_Price_Index   Competitor_Price_Index       0.0266833002
## Inventory_Level                 Inventory_Level       0.0009001567
## Social_Media_Engagement Social_Media_Engagement       0.0213626573
## SeasonSpring                       SeasonSpring       1.2164579475
## SeasonSummer                       SeasonSummer       1.2557638840
## SeasonWinter                       SeasonWinter       1.3153735496
## Product_TypeFashion         Product_TypeFashion       1.0180456291
## Product_TypeHome               Product_TypeHome       0.9824535338
##                         Robust_Std_Error Original_P_Value Robust_P_Value
## (Intercept)                 4.9667033960     1.814475e-13   6.304129e-14
## Ad_Spend                    0.0172474906     0.000000e+00   0.000000e+00
## Email_Campaigns             0.1422055502     4.019955e-13   3.587876e-13
## Website_Traffic             0.0048149280     0.000000e+00   0.000000e+00
## Avg_Rating                  0.8300049784     7.181043e-67   8.947016e-68
## Discount_Pct                0.0631099121     1.080699e-03   1.070721e-03
## Competitor_Price_Index      0.0267629723     1.121458e-51   2.173446e-51
## Inventory_Level             0.0008874432     0.000000e+00   0.000000e+00
## Social_Media_Engagement     0.0214052949     1.131522e-24   1.393348e-24
## SeasonSpring                1.1493090801     9.159895e-01   9.111010e-01
## SeasonSummer                1.2020818073     1.295308e-16   5.632968e-18
## SeasonWinter                1.2623913347     3.803860e-50   3.733534e-54
## Product_TypeFashion         1.0219679356     0.000000e+00   0.000000e+00
## Product_TypeHome            0.9846737592    3.433305e-179  1.777185e-178
```

Conclusion: After applying the remedy, we compared the Robust Standard Errors to the Original OLS errors. While the standard errors shifted slightly, **the statistical significance of our predictors did not change.**

All variables that were significant in the original model remain highly significant, and insignificant variables remain insignificant. This confirms that our business conclusions are **robust** and safe to use for decision-making, despite the presence of mild heteroscedasticity.

Part C: Autocorrelation Assessment

```
# Perform Durbin-Watson Test for Autocorrelation
dw_test <- lmtest::dwtest(model_main)

print(dw_test)
```

```
##
##  Durbin-Watson test
##
## data:  model_main
## DW = 1.9492, p-value = 0.05709
## alternative hypothesis: true autocorrelation is greater than 0
```

Test Statistic: The DW statistic ranges from 0 to 4, where a value of **2.0** indicates zero autocorrelation. Our model produced a statistic of **1.9492**. **Significance:** The p-value is **0.057**, which is greater than the standard significance level ($\alpha = 0.05$).

Conclusion: We **fail to reject the null hypothesis**. The DW statistic is extremely close to the ideal value of 2.0. There is **no statistically significant evidence** of autocorrelation in the residuals. While the p-value is borderline (close to 0.05), the statistic is very close to 2.0 (the ideal value indicating zero autocorrelation). Therefore, we conclude that while there may be a weak temporal signal, there is no statistically significant evidence of serious autocorrelation that would invalidate the model.

Business Context:

Even though our statistical test passed, it was right for the CDO to be concerned. In a retail business context, autocorrelation is plausible due to:

- **Marketing “Carryover” Effects:** An advertising campaign in late January might drive brand awareness that results in sales in early February.
- **Inventory Constraints:** A stockout in month t could artificially depress sales, leading to pent-up demand and a spike in month $t + 1$.

Final Audit Summary for the CDO

We have successfully audited the model against all three concerns:

1. **Multicollinearity:** Ruled out (VIF ~ 1.0).
2. **Heteroscedasticity:** Detected, but successfully corrected using Robust Standard Errors with no change to business conclusions.
3. **Autocorrelation:** Ruled out (DW Stat ~ 1.95).

Thus, the model is statistically sound and safe for strategic decision-making.

Business Objective 4: Alternative Model Specifications

```
sales_data <- sales_data %>%
  mutate(log_Website_Traffic = log(Website_Traffic))

# Model 1: Linear-Linear (model_main)
summary(model_main)
```

```
##
## Call:
## lm(formula = Sales ~ Ad_Spend + Email_Campaigns + Website_Traffic +
##     Avg_Rating + Discount_Pct + Competitor_Price_Index + Inventory_Level +
##     Social_Media_Engagement + Season + Product_Type, data = sales_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.160 -16.390   0.177  16.939  75.765
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.741e+01  5.062e+00   7.390 1.81e-13 ***
## Ad_Spend                2.482e+00  1.860e-02 133.435  < 2e-16 ***
## Email_Campaigns         1.038e+00  1.425e-01   7.282 4.02e-13 ***
## Website_Traffic         1.201e+00  4.612e-03 260.385  < 2e-16 ***
## Avg_Rating              1.475e+01  8.360e-01  17.643  < 2e-16 ***
## Discount_Pct            2.066e-01  6.316e-02   3.271  0.00108 **
## Competitor_Price_Index  4.101e-01  2.668e-02  15.370  < 2e-16 ***
## Inventory_Level        -4.879e-02  9.002e-04 -54.198  < 2e-16 ***
## Social_Media_Engagement  2.207e-01  2.136e-02  10.331  < 2e-16 ***
## SeasonSpring            1.283e-01  1.216e+00   0.105  0.91599
## SeasonSummer           -1.044e+01  1.256e+00  -8.314  < 2e-16 ***
## SeasonWinter            1.989e+01  1.315e+00  15.125  < 2e-16 ***
## Product_TypeFashion    -4.469e+01  1.018e+00 -43.901  < 2e-16 ***
## Product_TypeHome       -2.972e+01  9.825e-01 -30.246  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.31 on 3585 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9632, Adjusted R-squared:  0.9631
## F-statistic:  7218 on 13 and 3585 DF,  p-value: < 2.2e-16
```

```
# Model 2: Log-Linear Model
model_log_linear <- lm(log_Sales ~ Ad_Spend + Email_Campaigns + Website_Traffic +
                         Avg_Rating + Discount_Pct + Competitor_Price_Index +
                         Inventory_Level + Social_Media_Engagement +
                         Season + Product_Type,
                       data = sales_data)
summary(model_log_linear)
```

```
##
## Call:
## lm(formula = log_Sales ~ Ad_Spend + Email_Campaigns + Website_Traffic +
##     Avg_Rating + Discount_Pct + Competitor_Price_Index + Inventory_Level +
##     Social_Media_Engagement + Season + Product_Type, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48312 -0.04702  0.01191  0.06667  0.23787
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.988e+00  2.261e-02 220.669  < 2e-16 ***
## Ad_Spend                 6.406e-03  8.307e-05  77.122  < 2e-16 ***
## Email_Campaigns          2.528e-03  6.364e-04   3.973 7.25e-05 ***
## Website_Traffic          2.965e-03  2.060e-05 143.918  < 2e-16 ***
## Avg_Rating               4.695e-02  3.734e-03  12.575  < 2e-16 ***
## Discount_Pct             7.932e-04  2.821e-04   2.812  0.00495 **
## Competitor_Price_Index   1.315e-03  1.192e-04  11.037  < 2e-16 ***
## Inventory_Level         -1.467e-04  4.020e-06 -36.483  < 2e-16 ***
## Social_Media_Engagement  6.178e-04  9.540e-05   6.476 1.07e-10 ***
## SeasonSpring             1.372e-03  5.433e-03   0.253  0.80058
## SeasonSummer            -3.460e-02  5.608e-03  -6.170 7.61e-10 ***
## SeasonWinter             5.623e-02  5.874e-03   9.572  < 2e-16 ***
## Product_TypeFashion     -1.340e-01  4.547e-03 -29.476  < 2e-16 ***
## Product_TypeHome        -8.342e-02  4.388e-03 -19.011  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1086 on 3585 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8939, Adjusted R-squared:  0.8935
## F-statistic:  2322 on 13 and 3585 DF,  p-value: < 2.2e-16
```

```
# Model 3: Log-Log Model
model_log_log <- lm(log_Sales ~ log_Ad_Spend + Email_Campaigns + log_Website_Traffic +
                      Avg_Rating + Discount_Pct + Competitor_Price_Index +
                      Inventory_Level + Social_Media_Engagement +
                      Season + Product_Type,
                    data = sales_data)
summary(model_log_log)
```

```
##
## Call:
## lm(formula = log_Sales ~ log_Ad_Spend + Email_Campaigns + log_Website_Traffic +
##     Avg_Rating + Discount_Pct + Competitor_Price_Index + Inventory_Level +
##     Social_Media_Engagement + Season + Product_Type, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.17870 -0.06163 -0.00431  0.06032  0.44920
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.410e+00  3.010e-02  80.061  < 2e-16 ***
## log_Ad_Spend             2.212e-01  2.946e-03  75.081  < 2e-16 ***
## Email_Campaigns          4.075e-03  6.377e-04   6.390 1.87e-10 ***
## log_Website_Traffic      5.101e-01  3.527e-03 144.628  < 2e-16 ***
## Avg_Rating               4.615e-02  3.741e-03  12.335  < 2e-16 ***
## Discount_Pct             3.702e-04  2.826e-04   1.310    0.190
## Competitor_Price_Index   1.323e-03  1.194e-04  11.076  < 2e-16 ***
## Inventory_Level         -1.479e-04  4.028e-06 -36.714  < 2e-16 ***
## Social_Media_Engagement  5.750e-04  9.559e-05   6.015 1.98e-09 ***
## SeasonSpring             4.241e-03  5.443e-03   0.779    0.436
## SeasonSummer            -3.092e-02  5.620e-03  -5.503 4.00e-08 ***
## SeasonWinter             5.213e-02  5.887e-03   8.855  < 2e-16 ***
## Product_TypeFashion     -1.288e-01  4.557e-03 -28.265  < 2e-16 ***
## Product_TypeHome        -8.201e-02  4.396e-03 -18.654  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1088 on 3585 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8934, Adjusted R-squared:  0.893
## F-statistic:  2312 on 13 and 3585 DF,  p-value: < 2.2e-16
```

```
model_comparison <- data.frame(
  Model = c("1. Linear-Linear (model_main)", "2. Log-Linear", "3. Log-Log"),
  Adjusted_R_Squared = c(
    summary(model_main)$adj.r.squared,
    summary(model_log_linear)$adj.r.squared,
    summary(model_log_log)$adj.r.squared
  )
)

print(model_comparison)
```

```
##                           Model Adjusted_R_Squared
## 1 1. Linear-Linear (model_main)          0.9630679
## 2                 2. Log-Linear          0.8934656
## 3                    3. Log-Log          0.8930358
```

Based on the CEO's request and standard economic modeling practices, we considered transforming the following variables:

1. **Sales (Dependent Variable):** Sales data is often right-skewed (many small values, few massive ones). Taking the log (`log_Sales`) can normalize the residuals and helps us model **percentage growth** rather than absolute dollar growth.
2. **Ad_Spend & Website_Traffic (Predictors):** These are "scale" variables. A log transformation (`log_Ad_Spend`) allows us to test the idea that a 1% increase in spend leads to a certain percentage increase in sales, rather than a fixed dollar amount.

Analysis & Recommendation:

The **Linear Model (Model 1)** is superior. It explains **96.3%** of the variation in sales, whereas both log-transformed models drop to **89.3%**. This confirms our earlier finding in Objective 1: the relationship between our predictors and sales is fundamentally linear, not multiplicative or curved. Forcing a log transformation actually degrades the model's performance.