A Data-Driven Framework for Optimizing MarketPlace's $20 Million Marketing Budget

Ritika Garg | 20 October 2025

Table of Contents

Executive Summary

This report presents a comprehensive statistical analysis of MarketPlace's sales data to identify key performance drivers, quantify the impact of marketing investments, and provide a data-driven framework for the allocation of the $20 million annual marketing budget. The findings are intended to equip the executive team with actionable insights for strategic planning, resource allocation, and sustained growth.

The analysis has produced a highly robust multiple regression model that successfully explains 96.3% of the variation in monthly sales. This level of predictive accuracy provides a reliable and powerful tool for strategic decision-making, moving beyond intuition to a quantitative understanding of the factors that shape MarketPlace's revenue.

The model confirms that sales are most powerfully and positively influenced by Website Traffic and Digital Advertising Spend. However, the analysis reveals a more complex ecosystem where success is equally dependent on other factors. Key findings include:

- **Customer Satisfaction is a Major Financial Lever:** A one-point increase in average product ratings is associated with a monthly sales increase of approximately $14,750.
- **Market Positioning Matters:** Competitive pricing and promotional discounts are proven, significant drivers of sales performance.

This analysis provides definitive statistical evidence to address the CFO's concerns regarding marketing effectiveness. The marketing portfolio, comprising Digital Ad Spend, Email Campaigns, and Social Media Engagement, is not a cost center but a significant and indispensable driver of revenue.

- **Digital Advertising:** Demonstrates a strong Return on Ad Spend (ROAS) of approximately 2.48:1. For every $1 invested in digital advertising, the model predicts a return of $2.48 in sales revenue.
- **Joint Impact:** As a group, these three marketing activities are proven to be jointly significant, explaining over 83% of the sales variation that other operational and market factors fail to account for.

Strategic Recommendations:

1. **Endorse and Optimize Marketing Investment:** Fully endorse the proposed marketing budget. The data provides a clear justification for investment, particularly in digital advertising. The allocation strategy should leverage the model to optimize spending across channels based on their demonstrated return.
2. **Prioritize and Invest in the Customer Experience:** The substantial financial impact of average rating warrants strategic investment in initiatives aimed at improving product

quality, enhancing customer service, and actively managing online reputation. This is a highly efficient path to revenue growth.

3. **Diagnose Underperforming Product Categories:** The Fashion category systematically underperforms relative to Electronics and Home goods, even after controlling for marketing support. A dedicated strategic review is required to diagnose the root causes, such as product assortment, pricing, or market fit, before allocating further marketing resources to this segment.

This report will now detail the methodology, findings, and business implications for each stage of the analysis presented in 9 parts.

PART 1: Establishing a Sales Driver Model

The construction of the model began with the creation of dummy variables for categorical data. The choice of a "reference category" for each is a strategic decision that frames the interpretation of the results around MarketPlace's core business realities.

- **Season:** Winter was selected as the reference category. As this period typically includes the holiday shopping season, it represents MarketPlace's peak sales performance. By setting Winter as the baseline, the model's coefficients for Spring, Summer, and Fall directly measure the expected sales decrease relative to this critical peak period. This provides an immediate and intuitive understanding of seasonal performance gaps.
- **Product Type:** Electronics was chosen as the reference category. Electronics often serve as high-volume, high-revenue anchor products that drive significant traffic and sales for e-commerce platforms. Using this category as the baseline allows for a direct comparison of how Home and Fashion products perform against the company's primary revenue engine, offering a more actionable insight.

Model 1: The Full Predictive Framework

Using these strategically defined variables, *Model 1* was constructed. This comprehensive model includes all 13 available predictors, spanning marketing, operational, product, and categorical factors. The goal of this initial model is to capture the maximum amount of information available to explain the variation in monthly sales.

PART 2: Validating the Strategic Analytics Framework

The test is structured around two competing hypotheses:

- **Null Hypothesis:** This hypothesis states that none of the 13 predictor variables in the *Model 1* have a significant linear relationship with monthly sales. In business terms, this would mean that the collective data-gathering and analytics efforts have failed, and the chosen metrics are meaningless for predicting sales performance.

- **Alternative Hypothesis:** This hypothesis posits that at least one of the predictor variables has a significant linear relationship with monthly sales. In business terms, this confirms that the data-driven model is valid and that the factors being tracked are meaningful drivers of business outcomes.

We overwhelmingly reject the null hypothesis. There is extremely strong statistical evidence that the overall regression model is valid and highly significant.

This result is a critical validation of the analytics team's entire approach. It provides the executive team with the statistical confidence needed to rely on this model for high-stakes decisions, including the $20 million marketing budget allocation. The finding confirms that the factors being collected and tracked are indeed systematically related to revenue performance. This gives the team a definitive "green light" to proceed with a deeper analysis of the model, confident that its insights are grounded in a statistically reliable framework.

PART 3: Identifying Key Performance Levers

To directly address the CFO's concerns about marketing expenditures, a specific hypothesis test was conducted on the ad spend coefficient.

- **Null Hypothesis:** After controlling for all other variables, advertising spend has no statistically significant effect on monthly sales.
- **Alternative Hypothesis:** After controlling for all other variables, advertising spend has a statistically significant effect on monthly sales.

The t-statistic for Ad Spend is 133.44, with a p-value of essentially zero. We reject the null hypothesis. There is overwhelming statistical evidence that advertising spend is a significant and positive driver of sales. This finding provides a clear, data-driven justification for the marketing budget, confirming that these expenditures generate a measurable return.

Additionally, we found that all 13 predictors in the model are statistically significant at the 5% level. In other words, there are no non-significant predictors in this model. This finding is important as it indicates that the analytics team has collected a highly relevant set of metrics.

PART 4: Quantifying the Impact of Key Business Drivers

Statistical significance confirms which variables matter, but interpreting their coefficients is what translates this knowledge into actionable business strategy. This section quantifies the real-world impact of four key predictors based on *Model 1*, providing clear financial and strategic implications for the executive team.

4.1 Impact of Advertising Spend

- **Key Finding:** For every $1,000 increase in monthly digital advertising spend, sales are predicted to increase by $2,482, holding all other variables constant.

- **Business Implication:** This coefficient translates to a Return on Ad Spend (ROAS) of approximately 2.48:1. For every dollar spent on advertising, MarketPlace generates about $2.48 in sales revenue. This is a strong, positive return that provides a clear financial justification for the marketing budget. It directly addresses the CFO's concerns by demonstrating that advertising is not merely a cost but a highly effective revenue-generating investment. While this strong linear relationship supports increasing the ad spend budget, it is advisable to do so incrementally while monitoring ROAS to identify the point of diminishing returns.

4.2 Value of Customer Satisfaction

- **Key Finding:** For every one-point increase in the average customer product rating (on its 1-5 scale), monthly sales are predicted to increase by $14,749, holding all other variables constant.
- **Business Implication:** This result quantifies the immense financial value of customer satisfaction and product quality. It demonstrates that a seemingly small improvement, such as raising a product's average rating from 3.5 to 4.5 stars, can have a substantial impact on the bottom line. This provides a data-driven business case for investing in initiatives aimed at improving product quality, enhancing customer service, and actively managing online reviews.

4.3 Understanding the Summer Sales Slump

- **Key Finding:** Compared to Winter (the reference season), Summer sales are predicted to be lower by $30,335 per month, controlling for all other factors.
- **Business Implication:** This quantifies a significant and costly seasonal slump. The model predicts that even for the same product with identical marketing support and pricing, sales will be substantially lower in the summer months. This insight is critical for seasonal strategy. It strongly suggests that inventory planning must be adjusted to avoid overstocking and costly markdowns during the summer. Furthermore, it indicates that marketing campaigns should be specifically tailored to boost demand during this slower period, perhaps through summer-specific promotions or by strategically shifting a portion of the ad budget to counter the downturn.

4.4 Challenge of the Fashion Category

- **Key Finding:** Compared to Electronics (the reference category), Fashion products are predicted to have lower sales by $44,693 per month, holding all other variables constant.
- **Business Implication:** This highlights a major structural challenge. The model shows that even if a Fashion product and an Electronics product had the exact same ad spend, website traffic, discount, and customer rating, the Fashion product is still predicted to sell nearly $45k less. A strategic review is recommended to diagnose this core

underperformance rather than simply attempting to fix it by adjusting the marketing levers already being measured.

PART 5: Selecting Optimal Model for Decision-Making

A key task in analytics is to balance a model's predictive power with its simplicity (parsimony). To this end, a second, simpler model (*Model 2*) was constructed, including only a subset of the original predictors: Ad Spend, Website Traffic, Rating, Season and Product Type. This simpler model was then compared against the comprehensive *Model 1* to determine which is better suited for guiding business decisions at MarketPlace.

Model 2: A Simplified Framework

The Standard Error of the regression (SE) measures the typical size of a prediction error. A lower SE indicates a more accurate model. *Model 1* has a substantially lower Standard Error. In business terms, *Model 1*'s sales forecasts are, on average, off by about $24,310. *Model 2*'s forecasts are off by about $33,950. The nearly 40% increase in average prediction error for *Model 2* represents a significant loss of accuracy and reliability for business planning.

Adjusted R-squared measures the proportion of variance in sales that is explained by the model, penalizing the score for adding variables that do not significantly improve the fit. A higher Adjusted R-squared is better. *Model 1* has a higher Adjusted R-squared. It explains 96.3% of the variation in sales, while the simpler *Model 2* only explains 92.8%. The significant drop in Adjusted R-squared for *Model 2* confirms that the variables that were removed were not redundant; they were adding real explanatory power.

Table: Comparison of error rate for Model 1 and Model 2

	Standard Error	**Adjusted R-squared**
Model 1 (Full Model)	$24,310	96.3%
Model 2 (Simplified Model)	$33,950	92.8%

Based on the error rate comparisons, *Model 1* is strongly recommended as the primary tool for the team to use for all business analysis and strategic decision-making. Here are some key results to justify this recommendation:

1. **Superior Statistical Performance:** *Model 1* is unequivocally superior on both key performance metrics. Its lower Standard Error translates to significantly more accurate and reliable sales forecasts, which is critical for budgeting and planning. Its higher Adjusted R-squared confirms that it provides a more complete and accurate picture of the business environment.
2. **Strategic Decision-Making Needs:** The team's primary goal is to provide actionable insights to guide the $20 million budget allocation. *Model 1* is the only model that serves

this purpose comprehensively. It allows the team to quantify the individual and collective impact of all key marketing, operational, and market factors. The strategic value of this completeness far outweighs the minor parsimony of *Model 2*.

PART 6: Forecasting Performance and Managing Risk

This part calculates the forecast for a new product with the following characteristics: Ad_Spend = $25,000, Email_Campaigns = 10, Website_Traffic = 150,000 visitors, Avg_Rating = 4.5, Discount_Pct = 15%, Competitor_Price_Index = 105, Inventory_Level = 1,000 units, Social_Media_Engagement = 70, Season = Fall, and Product_Type = Home.

Based on these characteristics, the key findings based on *Model 1* are:

- The predicted monthly sales for this new product is $339,460.
- We are 95% confident that the actual sales for this specific product will fall between $291,730 and $387,190. The wide range quantifies the inherent uncertainty of any single new product. The supply chain team can use the lower bound ($291.7k) as a conservative estimate to avoid overstocking, while the marketing team can use the upper bound ($387.2k) to set optimistic sales targets and ensure sufficient inventory is available to meet a high-demand scenario.
- The 95% confidence interval for the average sales of all products with these characteristics is from $336,900 to $342,025. This predicts the long-term average performance of this entire product profile. The narrowness of the interval (only about $5.1k wide) gives the executive team a highly reliable and precise estimate. This is the number to use in long-term financial models.

The prediction interval is significantly wider than the confidence interval because it must account for two sources of uncertainty: (1) the uncertainty in estimating the model's regression line and (2) the inherent, random variability of any single new product around that line. The confidence interval only accounts for the first source of uncertainty. In short, it is much easier to predict an average than a single outcome.

- **When to use the Prediction Interval:** For **tactical, individual-level decisions**. A practical example is using the range [$291.7k, $387.2k] to set the initial inventory order for this one specific product launch.
- **When to use the Confidence Interval:** For **strategic, long-term average decisions**. A practical example is using the range [$336.9k, $342.0k] when the CFO builds the annual budget and needs to forecast the expected long-term revenue from this entire product category strategy.

PART 7: A Data-Driven Case for Marketing Investment

To provide a definitive answer to the CFO's question about the collective value of marketing investments, a partial F-test was conducted. This test specifically evaluates whether the three main marketing variables (Ad Spend, Email Campaigns, Social Media Engagement), as a group, significantly improve the model's ability to explain sales, even after all other factors are taken into account. To measure, we

Model 3: Reduced Model for Partial F-test

Out of the 13 predictor variables, *Model 3* includes all predictors *except* Ad Spend, Email Campaigns, and Social Media Engagement. The hypothesis is stated as follows:

- **Null Hypothesis:** The three marketing variables, as a group, add no significant explanatory power to the model once all other factors are included.
- **Alternative Hypothesis:** At least one of the marketing variables adds significant explanatory power to the model.

The null hypothesis reflects the concern that marketing's impact is redundant or illusory. The alternative hypothesis posits that the marketing portfolio provides significant, unique insight into sales performance.

With a p-value of essentially zero, which is far below the significance level of 0.05, we strongly reject the null hypothesis. The partial F-test provides extremely strong statistical evidence that the group of marketing variables jointly adds significant explanatory power to the sales model. This means these marketing activities, taken together, are crucial and non-redundant drivers of sales performance. They provide unique information that cannot be explained by other factors like website traffic or pricing alone. The recommendation is to maintain, and potentially increase, investment in these areas, as they are demonstrably linked to revenue. Reducing this budget would, according to this robust statistical evidence, almost certainly lead to a significant decrease in sales.

PART 8: Isolating the Unique Value of Marketing Initiatives

While the partial F-test confirmed that the marketing portfolio is significant as a group, the coefficient of partial determination goes a step further. It quantifies precisely how much of the "unexplained" sales variance is captured by these marketing variables after all other factors have already been considered. This metric isolates the unique and incremental contribution of the marketing function.

After controlling for all other variables in the model, the three main marketing controllable variables (Ad Spend, Email Campaigns, and Social Media Engagement) collectively explain 83.34% of the sales variance that those other factors failed to account for.

This extremely high value is a powerful testament to the unique value of the marketing function. It demonstrates that marketing activities provide substantial and irreplaceable insight into sales performance. They are not redundant with other metrics. This finding solidifies the argument that these marketing variables are critically important and must remain in any model used for analysis, forecasting, and budget allocation. They represent the key levers that the company can proactively pull to influence revenue, and their impact is both significant and unique.

PART 9: Strategic Insights from Categorical Variables

A deep dive into the model's categorical variables provides specific, actionable recommendations for inventory, marketing, and product strategy. By analyzing the coefficients for Season and Product Type, we can develop more nuanced and effective business plans.

Part A: Seasonal Analysis

Based on the coefficients from *Model 1*, the four seasons can be ranked from highest to lowest expected sales, holding all other factors constant. The values represent the expected change in monthly sales compared to the Winter peak season followed by spring, fall and finally summer.

Table: Ranking of Seasons by Expected Sales Impact (Relative to Winter)

Season	**Sales Adjustment vs. Winter**
Winter	$0 (Reference season)
Spring	-$19,770
Fall	-$19,890
Summer	-$30,340

The model clearly shows a significant performance gap between the peak Winter season and the Summer trough.

- **Key Finding:** Sales in Summer are expected to be $30,340 lower per month than in Winter, all else being equal.
- **Business Implication:** This confirmed seasonal slump is critical for operational planning. It necessitates a proactive strategy to reduce inventory levels leading into the summer to avoid costly overstocking. From a marketing perspective, it suggests the need for a dedicated "Summer Strategy" involving targeted promotions or a potential shift in ad budget to mitigate this predictable downturn.

While the point estimates for Spring (-$19.77k) and Fall (-$19.89k) are slightly different, a formal statistical test was conducted to see if this difference is meaningful.

- **Key Finding:** We cannot conclude there is a statistically significant difference in sales between Spring and Fall.

- **Business Implication:** For practical business purposes, MarketPlace can treat Spring and Fall as having a similar impact on sales. This simplifies planning, allowing for similar inventory levels, staffing, and marketing approaches during these two "shoulder" seasons, freeing up resources to focus on managing the more extreme Winter peak and Summer trough.

Part B: Product Type Analysis

The model's coefficients also allow for a clear ranking of product categories by their baseline sales performance, relative to the high-performing Electronics category. Electronics is a top followed by Home and then Fashion.

Table: Ranking of Product Types by Expected Sales Impact (Relative to Electronics)

Product Type	Sales Adjustment vs. Electronics
Electronics	$0 (Reference season)
Home	-$29,720
Fashion	-$44,690

A comparison between the Home and Fashion categories reveals a substantial and statistically significant difference in performance.

- **Key Finding:** Home products are expected to generate $14,980 more in monthly sales than Fashion products, even after controlling for all other factors like marketing spend and seasonality.
- **Business Implication:** This finding warrants careful consideration. While Home goods generate more revenue for the same level of input, this does not automatically mean they are more profitable or have higher growth potential. The Fashion category might have higher margins or appeal to a different, valuable customer segment. The recommendation is not to simply shift investment away from Fashion. Instead, MarketPlace should launch a strategic investigation to understand the root cause of this underperformance. Is it an issue with product assortment, pricing strategy, on-site conversion rates, or overall market demand? Answering these questions is essential before making any major changes to budget allocation between the categories.

Appendix

Data Setup

```
dataset <- read.csv("MarketPlace_Sales_Data.csv")
dataset <- dataset[, 1:11]
head(dataset, 10)
```

```
##     Sales Ad_Spend Email_Campaigns Website_Traffic Avg_Rating Discount_Pct
## 1  431.59     9.84              14          257.97       3.85         9.38
## 2  502.24     9.55               7          255.41       3.40        10.01
## 3  408.97    18.55               8          236.68       2.63         2.39
## 4  300.93    56.57               7           56.35       4.10        13.89
## 5  294.82    14.00              13          122.79       3.99         4.21
## 6  360.82    27.15               5          152.42       4.76        22.17
## 7  347.69    10.30               9          261.69       4.16         4.70
## 8  276.15     6.79               9          167.70       3.71        22.55
## 9  393.03    60.54               9          130.73       3.91         6.56
## 10 208.80    14.11               4           99.58       4.41         8.82
##    Competitor_Price_Index Inventory_Level Social_Media_Engagement Season
## 1                   85.62             697                   27.26 Winter
## 2                  105.69             625                   67.88 Winter
## 3                  106.04             874                   40.99 Spring
## 4                  103.77             872                   38.75 Spring
## 5                  112.59            1203                   58.97 Spring
## 6                  114.66             545                   41.46 Summer
## 7                   94.00            1729                   16.90 Summer
## 8                   83.50             677                   48.91 Summer
## 9                  115.98             616                   48.47   Fall
## 10                  90.63            1233                   65.41   Fall
##    Product_Type
## 1          Home
## 2          Home
## 3          Home
## 4          Home
## 5          Home
## 6          Home
## 7          Home
## 8          Home
## 9          Home
## 10         Home
```

PART 1

Question 1.1: Dummy variables for Season

```
dataset$Season_Spring <- ifelse(dataset$Season == "Spring", 1, 0)
dataset$Season_Summer <- ifelse(dataset$Season == "Summer", 1, 0)
dataset$Season_Fall <- ifelse(dataset$Season == "Fall", 1, 0)
```

Interpretation: Choosing 'Winter' as the reference category is a strategic decision. Winter often represents the peak sales season for an e-commerce company, likely driven by holiday shopping. By setting this peak season as the reference, the model's coefficients for Season_Spring, Season_Summer, and Season_Fall will directly measure sales performance relative to the company's most critical period, making the seasonal impact immediately clear.

Question 1.2: Dummy variables for Category

```
dataset$Product_Type_Home <- ifelse(dataset$Product_Type == "Home", 1, 0)
dataset$Product_Type_Fashion <- ifelse(dataset$Product_Type == "Fashion", 1, 0)
```

Interpretation: Selecting 'Electronics' as the reference category is a strategic decision. Electronics are often high-ticket, high-volume products that can serve as a primary driver of revenue and website traffic. By using Electronics as the baseline, the coefficients for Product_Type_Home and Product_Type_Fashion will directly measure how their sales performance compares against this core, high-revenue category. This allows MarketPlace to quickly assess if other categories are underperforming or overperforming relative to this key business driver.

Question 1.3: Build Model 1 using ALL predictors

```
Model1 <- lm(Sales ~ Ad_Spend + Email_Campaigns + Website_Traffic +
                     Avg_Rating + Discount_Pct + Competitor_Price_Index +
                     Inventory_Level + Social_Media_Engagement +
                     Season_Spring + Season_Summer + Season_Fall +
                     Product_Type_Home + Product_Type_Fashion,
                     data = dataset)
summary(Model1)
```

```
##
## Call:
## lm(formula = Sales ~ Ad_Spend + Email_Campaigns + Website_Traffic +
##     Avg_Rating + Discount_Pct + Competitor_Price_Index + Inventory_Level +
##     Social_Media_Engagement + Season_Spring + Season_Summer +
##     Season_Fall + Product_Type_Home + Product_Type_Fashion, data = dataset)
##
## 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)             5.730e+01  5.019e+00  11.418  < 2e-16 ***
## 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 ***
## Season_Spring          -1.977e+01  1.111e+00 -17.794  < 2e-16 ***
## Season_Summer          -3.034e+01  1.155e+00 -26.258  < 2e-16 ***
## Season_Fall            -1.989e+01  1.315e+00 -15.125  < 2e-16 ***
## Product_Type_Home      -2.972e+01  9.825e-01 -30.246  < 2e-16 ***
## Product_Type_Fashion   -4.469e+01  1.018e+00 -43.901  < 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
```

```
coef_table <- summary(Model1)$coefficients
kable(coef_table,
      digits = 3,
      caption = "Model 1: Full Regression Model of Monthly Sales",
      col.names = c("Estimate", "Std. Error", "t-value", "p-value"))
```

Model 1: Full Regression Model of Monthly Sales

	Estimate	Std. Error	t-value	p-value
(Intercept)	57.302	5.019	11.418	0.000
Ad_Spend	2.482	0.019	133.435	0.000
Email_Campaigns	1.038	0.143	7.282	0.000
Website_Traffic	1.201	0.005	260.385	0.000
Avg_Rating	14.749	0.836	17.643	0.000
Discount_Pct	0.207	0.063	3.271	0.001
Competitor_Price_Index	0.410	0.027	15.370	0.000
Inventory_Level	-0.049	0.001	-54.198	0.000

	Estimate	Std. Error	t-value	p-value
Social_Media_Engagement	0.221	0.021	10.331	0.000
Season_Spring	-19.766	1.111	-17.794	0.000
Season_Summer	-30.335	1.155	-26.258	0.000
Season_Fall	-19.895	1.315	-15.125	0.000
Product_Type_Home	-29.716	0.982	-30.246	0.000
Product_Type_Fashion	-44.693	1.018	-43.901	0.000

PART 2

Question 2.1: Global F-Test Hypotheses

- **Null Hypothesis (H_0): $\beta_1 = \beta_2 = \cdots = \beta_{13} = 0$**
  - The model has no explanatory power. None of the 13 predictor variables (advertising spend, website traffic, season, etc.) have a linear relationship with sales.
- **Alternative Hypothesis (H_1): At least one $\beta_j \neq 0$ (for $j = 1, \ldots, 13$)**
  - At least one of the predictor variables in the model has a significant linear relationship with sales.

- **Business Interpretation:**
  - In business terms, the null hypothesis (H_0) suggests that the entire set of metrics Tracy's team has collected is useless for predicting monthly sales.
  - The alternative hypothesis (H_1) posits that the team's data-driven model is valid and that at least one of the tracked factors is a meaningful driver of sales performance.

Question 2.2: Conduct Global F-test for Model 1

```
dat_used <- model.frame(Model1)

# Components for the Global F-test
n  <- nrow(dat_used)             # Total number of observations used in the model
k  <- 13                         # Number of predictors in the model
y  <- dat_used$Sales             # The Sales values corresponding to the NA-free rows

# Sum of Squares
sst <- sum( (y - mean(y))^2 )         # Total Sum of Squares (based on used data)
sse <- sum( residuals(Model1)^2 ) # Error/Residual Sum of Squares
ssr <- sst - sse                      # Regression Sum of Squares (SST - SSE)

# Degrees of Freedom
df_reg <- k                             # k
df_res <- n - k - 1                     # n - k - 1
df_tot <- n - 1                         # n - 1

# Mean Squares and F-statistic
msr <- ssr / df_reg
mse <- sse / df_res
Fval <- msr / mse
pval <- pf(Fval, df_reg, df_res, lower.tail = FALSE)

kable(
  data.frame(
    Source  = c("Regression","Residuals","Total"),
    Df      = c(df_reg, df_res, df_tot),
    `Sum Sq`= c(ssr, sse, sst),
    `Mean Sq`= c(msr, mse, NA),
    `F value`= c(Fval, NA, NA),
    `Pr(>F)`= c(pval, NA, NA)
  ),
  digits = 4,
  caption = "ANOVA Table with Single Overall F (Multiple Regression Model)"
)
```

ANOVA Table with Single Overall F (Multiple Regression Model)

Source	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
Regression	13	55456265	4265866.5750	7218.238	0
Residuals	3585	2118679	590.9845	NA	NA
Total	3598	57574945	NA	NA	NA

```
qf(0.95, df1 = 13, df2 = 3585)
```

```
## [1] 1.722883
```

Question 2.3: Global F-Test Conclusion

- **Statistical Conclusion:** The calculated F-statistic of 7218.24 is vastly greater than the critical F-value of 1.72. Correspondingly, the p-value is essentially zero, which is far less than the significance level of $\alpha = 0.05$. Therefore, we **reject the null hypothesis (H_0)**.

- **Interpretation for Tracy's Team:** The rejection of the null hypothesis is a crucial first step. It provides strong statistical evidence that the overall regression model is valid and that the chosen set of predictors, when taken together, has a highly significant ability to explain the variation in monthly sales.
- **Business Implications:** This finding is a major validation of the analytics team's efforts. It confirms for the executive team that the factors being tracked, from marketing spend to operational metrics, are indeed related to revenue performance. This result provides Tracy with a green light to proceed with the analysis, confident that the model is a reliable tool for identifying key sales drivers and informing strategic decisions, such as the $20 million marketing budget allocation.

PART 3

Question 3.1: Comprehensive coefficient table for Model 1

```
summary_m1 <- summary(Model1)

coef_table <- summary_m1$coefficients

p_values <- coef_table[, 4]
significance_indicator <- ifelse(p_values < 0.05, "Yes", "No")

coef_report_df <- data.frame(
Variable = rownames(coef_table),
Estimate = coef_table[, 1],
Std_Error = coef_table[, 2],
t_statistic = coef_table[, 3],
P_value = p_values,
Significant_at_0_05 = significance_indicator,
row.names = NULL # Remove default R row names
)

kable(
coef_report_df,
digits = 4,
caption = "Comprehensive Coefficient Table for Model 1",
align = 'lrrrrr'
)
```

Comprehensive Coefficient Table for Model 1

Variable	Estimate	Std_Error	t_statistic	P_value	Significant_at_0_05
(Intercept)	57.3022	5.0186	11.4179	0.0000	Yes
Ad_Spend	2.4819	0.0186	133.4355	0.0000	Yes
Email_Campaigns	1.0378	0.1425	7.2821	0.0000	Yes
Website_Traffic	1.2010	0.0046	260.3848	0.0000	Yes
Avg_Rating	14.7494	0.8360	17.6427	0.0000	Yes
Discount_Pct	0.2066	0.0632	3.2713	0.0011	Yes
Competitor_Price_Index	0.4101	0.0267	15.3698	0.0000	Yes
Inventory_Level	-0.0488	0.0009	-54.1975	0.0000	Yes
Social_Media_Engagement	0.2207	0.0214	10.3307	0.0000	Yes
Season_Spring	-19.7663	1.1108	-17.7944	0.0000	Yes

Variable	Estimate	Std_Error	t_statistic	P_value	Significant_at_0_05
Season_Summer	-30.3354	1.1553	-26.2581	0.0000	Yes
Season_Fall	-19.8947	1.3154	-15.1247	0.0000	Yes
Product_Type_Home	-29.7157	0.9825	-30.2464	0.0000	Yes
Product_Type_Fashion	-44.6931	1.0180	-43.9009	0.0000	Yes

Question 3.2: Effect of Advertising Spend

- **Null Hypothesis ($H_0 : \beta_1 = 0$):**
  - After controlling for all other variables in the model, advertising spend has no statistically significant effect on monthly sales.
- **Alternative Hypothesis ($H_1 : \beta_1 \neq 0$):**
  - After controlling for all other variables, advertising spend has a statistically significant effect on monthly sales.

```
ad_spend_row <- coef_table["Ad_Spend", ]

t_statistic_ad_spend <- ad_spend_row["t value"]
p_value_ad_spend <- ad_spend_row["Pr(>|t|)"]

print(paste("Ad_Spend t-statistic:", round(t_statistic_ad_spend, 4)))
```

```
## [1] "Ad_Spend t-statistic: 133.4355"
```

```
print(paste("Ad_Spend p-value:", format(p_value_ad_spend, scientific = TRUE)))
```

```
## [1] "Ad_Spend p-value: 0e+00"
```

- **Test Statistics:**
  - **t-statistic:** 133.4355
  - **p-value:** < 2.2e-16 (or essentially 0)

- **Statistical Conclusion:** Since the p-value (approx. 0) is far less than the significance level of $\alpha = 0.05$, we reject the null hypothesis (H_0). There is overwhelming statistical evidence that advertising spend has a significant linear relationship with sales, even after accounting for all other factors.

- **Business Meaning (Advertising Effectiveness):** This result provides a clear, data-driven justification for the marketing budget. It demonstrates that advertising expenditures are not just a cost center but a significant and positive driver of revenue. This finding directly addresses the CFO's concerns, confirming that marketing expenditures are generating statistically significant returns.

Question 3.3: Summary of Findings

- **Statistically Significant Predictors:**
  - Based on the coefficient table, **all** 13 predictors in the model are statistically significant. Their p-values are all well below 0.05.

- **Non-Significant Predictors:**
  - There are **no** non-significant predictors in this model at the 5% significance level.

- **Business Interpretation:**
  - The finding that every single variable is significant is powerful. It indicates that the analytics team has collected a highly relevant set of metrics, and every factor plays a role in explaining sales.
  - While all factors matter, the **most impactful drivers** (those with the highest absolute t-statistics) appear to be **Website_Traffic** (t=260.4), **Ad_Spend** (t=133.4), **Inventory_Level** (t=-54.2), and the **Product_Type** dummies (t=-43.9

for Fashion, t=-30.2 for Home).
    - This tells Tracy's team that website traffic and ad spend are exceptionally strong positive drivers, while inventory levels and product category differences are also critical factors in determining sales performance.

PART 4

Question 4.1: Interpretation of Ad_Spend Coefficient

```
coef_ad_spend <- coef_table["Ad_Spend", "Estimate"]
cat("Coefficient Value:", coef_ad_spend)
```

```
## Coefficient Value: 2.481939
```

- **Interpretation:** For every $1,000 increase in monthly digital advertising spend (a 1-unit increase in `Ad_Spend`), sales are predicted to **increase** by $2,481.90, holding all other variables constant.
- **Business Implication:** This coefficient suggests a **Return on Ad Spend (ROAS) of approximately 2.48:1**. For every $1 spent on advertising, MarketPlace generates about $2.48 in sales revenue, which is a strong, positive return. This evidence provides a clear justification for the marketing budget and supports the case for increasing ad spend, as it appears to be a highly effective revenue driver.
    - **Recommendation:** The team should recommend increasing the ad spend budget, but with a caveat: this linear relationship may not hold indefinitely. The best approach is to increase spending incrementally while closely monitoring the ROAS to find the optimal investment level.

Question 4.2: Interpretation of Avg_Rating Coefficient

```
coef_avg_rating <- coef_table["Avg_Rating", "Estimate"]
cat("Coefficient Value:", coef_avg_rating)
```

```
## Coefficient Value: 14.74939
```

- **Interpretation:** For every **one-point increase** in the average customer product rating (on its 1-5 scale), monthly sales are predicted to **increase** by $14,749.40, holding all other variables constant.
- **Business Implication (Importance of Customer Satisfaction):** This result quantifies the immense value of customer satisfaction and product quality. It shows that a seemingly small improvement in customer ratings—which might be the difference between an average 3.5-star product and a 4.5-star product—can have a substantial impact on revenue.
    - This provides a strong, data-driven business case for investing in initiatives aimed at improving product quality, enhancing customer service, and actively managing online reviews and reputation. It proves to the executive team that customer satisfaction is not just a "soft" metric but a powerful and direct financial lever.

Question 4.3: Interpretation of Season_Summer Coefficient

```
coef_season_summer <- coef_table["Season_Summer", "Estimate"]
cat("Coefficient Value:", coef_season_summer)
```

```
## Coefficient Value: -30.3354
```

- **Interpretation:** Compared to **Winter** (the reference season), Summer sales are predicted to be **lower** by $30,335.40 per month, controlling for all other factors.

- **Business Implication (Seasonal Strategy):** This quantifies a significant and costly seasonal slump. The model predicts that, even for the exact same product with the same marketing support, sales will be substantially lower in the summer.
    - This insight is critical for seasonal strategy. It strongly suggests that MarketPlace should adjust inventory planning to avoid overstocking in the summer.
    - It also indicates that marketing campaigns should be tailored to boost demand during this slower period, perhaps through summer-specific promotions, different product features, or by shifting a portion of the ad budget to this timeframe to counter the slump.

Question 4.4: Interpretation of Product_Type_Fashion Coefficient

```
coef_product_fashion <- coef_table["Product_Type_Fashion", "Estimate"]
cat("Coefficient Value:", coef_product_fashion)
```

```
## Coefficient Value: -44.69313
```

- **Interpretation:** Compared to **Electronics** (the reference category), Fashion products are predicted to have **lower** sales by $44,693.10 per month, holding everything else constant.

- **Business Implication (Product Strategy):** This highlights a major structural challenge. The model shows that even if a Fashion product and an Electronics product had the **exact same** ad spend, website traffic, discount percentage, and customer rating, the Fashion product is **still** predicted to sell $44.69k less.
    - This finding pinpoints that the underperformance is not just a simple marketing or pricing issue (as those are accounted for). It points to a more fundamental problem, such as a mismatch in product-market fit, lower baseline consumer demand, or a less effective sales conversion funnel for that category. It is recommended to have a "strategic review" to focus on this core underperformance, not just on tweaking the marketing levers we are already measuring.

PART 5

Question 5.1: Build Model 2 with Reduced Predictors

```
Model2 <- lm(Sales ~ Season_Spring + Season_Summer + Season_Fall  + Product_Type_Home +
                            Product_Type_Fashion +
                            Ad_Spend + Website_Traffic + Avg_Rating, data = dataset)
summary(Model2)
```

```
## 
## Call:
## lm(formula = Sales ~ Season_Spring + Season_Summer + Season_Fall + 
##     Product_Type_Home + Product_Type_Fashion + Ad_Spend + Website_Traffic + 
##     Avg_Rating, data = dataset)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -136.760  -22.003    1.224   23.940   98.302 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           73.17274    5.30226   13.80   <2e-16 ***
## Season_Spring        -20.06259    1.55018  -12.94   <2e-16 ***
## Season_Summer        -31.69613    1.61159  -19.67   <2e-16 ***
## Season_Fall          -20.40885    1.83441  -11.13   <2e-16 ***
## Product_Type_Home    -28.28912    1.37072  -20.64   <2e-16 ***
## Product_Type_Fashion -43.43353    1.42047  -30.58   <2e-16 ***
## Ad_Spend               2.48348    0.02597   95.62   <2e-16 ***
## Website_Traffic        1.20136    0.00644  186.56   <2e-16 ***
## Avg_Rating            14.29123    1.16700   12.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 33.95 on 3590 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9281, Adjusted R-squared:  0.928 
## F-statistic:  5796 on 8 and 3590 DF,  p-value: < 2.2e-16
```

```
coef_table <- summary(Model2)$coefficients
kable(coef_table,
      digits = 3,
      caption = "Model 2: Regression Model of Monthly Sales with Reduced Predictors",
      col.names = c("Estimate", "Std. Error", "t-value", "p-value"))
```

Model 2: Regression Model of Monthly Sales with Reduced Predictors

	Estimate	Std. Error	t-value	p-value
(Intercept)	73.173	5.302	13.800	0
Season_Spring	-20.063	1.550	-12.942	0
Season_Summer	-31.696	1.612	-19.668	0
Season_Fall	-20.409	1.834	-11.126	0
Product_Type_Home	-28.289	1.371	-20.638	0
Product_Type_Fashion	-43.434	1.420	-30.577	0
Ad_Spend	2.483	0.026	95.620	0
Website_Traffic	1.201	0.006	186.560	0
Avg_Rating	14.291	1.167	12.246	0

Question 5.2: Compare Standard Error (SE)

```
se_m1 <- summary(Model1)$sigma
print(se_m1)
```

```
## [1] 24.31017
```

```
se_m2 <- summary(Model2)$sigma
print(se_m2)
```

```
## [1] 33.94863
```

- **Which model has lower SE?**
  - **Model 1** has a substantially lower Standard Error (24.31) compared to Model 2 (33.95).
- **Interpretation (What SE tells us):**
  - The Standard Error of the regression (or Residual Standard Error) represents the **typical size of a prediction error**. It's the average distance that the observed sales values fall from the regression line.
  - In business terms, Model 1's predictions are, on average, off by about **$24,310**. Model 2's predictions are, on average, off by about **$33,949**.
  - The 40% increase in average prediction error for Model 2 indicates a **significant loss of accuracy**. A lower SE is better, as it means the model's forecasts are more precise and reliable for business planning.

Question 5.3: Compare Adjusted R-squared

```
adj_r_sq_m1 <- summary(Model1)$adj.r.squared
print(adj_r_sq_m1)
```

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

```
adj_r_sq_m2 <- summary(Model2)$adj.r.squared
print(adj_r_sq_m2)
```

```
## [1] 0.9279769
```

- **Which model has higher Adjusted R^2?**
  - **Model 1** has a higher Adjusted R-squared.
- **Interpretation (What Adjusted R^2 tells us):**
  - **Adjusted R-squared** measures the proportion of the variance in Sales that is explained by the model's predictors, but it crucially **penalizes the score for adding predictors that don't improve the model fit** significantly.
  - Model 1 explains 96.31% of the variation in sales, while Model 2 (the simpler model) only explains 92.80%.
  - The fact that the Adjusted R^2 dropped significantly for Model 2 confirms that the variables we removed were **not** useless. They were adding real, significant explanatory power to the model, which is why the more complex Model 1 is a better fit for the data.

Question 5.4: Model Recommendation and Justification

- **Recommendation: Model 1** is the strongly recommended model for Tracy's team to use for business decisions.
- **Justification:**
  - **Statistical Performance:** Model 1 is superior on both key metrics. Its **lower Standard Error** ($24.31k vs. $33.95k) translates to significantly more accurate and reliable sales forecasts. Its **higher Adjusted R-squared** (96.31%

  - vs. 92.80%) confirms that the variables dropped in Model 2 were, in fact, statistically important and added real explanatory power. The performance drop in Model 2 is too significant to ignore for the sake of simplicity.
  - **Business Practicality & Completeness:** The variables excluded from Model 2 are all actionable business levers. Removing them from the model would make it "simpler" but also "useless" for many of its intended purposes. For example, without `Discount_Pct` in the model, the team cannot analyze the effectiveness of promotional pricing strategies.
  - **Strategic Decision-Making Needs:** The team's goal is not only to *predict* sales but also to *provide actionable insights for the $20 million budget allocation*. Model 1 is the only one that serves this dual purpose. It allows Tracy to quantify the ROAS for `Ad_Spend`, `Email_Campaigns`, and `Social_Media_Engagement` individually. The completeness of Model 1 is essential for this strategic analysis and far outweighs the minor parsimony of Model 2.

PART 6

Question 6.1: Calculate Point Prediction

```
new_product_data <- data.frame(
  Ad_Spend = 25,
  Email_Campaigns = 10,
  Website_Traffic = 150,
  Avg_Rating = 4.5,
  Discount_Pct = 15,
  Competitor_Price_Index = 105,
  Inventory_Level = 1000,
  Social_Media_Engagement = 70,
  # Set dummy variables according to Season = Fall and Product_Type = Home
  Season_Spring = 0,         # Not Spring
  Season_Summer = 0,     # Not Summer
  Season_Fall = 1,       # Is Fall
  Product_Type_Home = 1,       # Is Home
  Product_Type_Fashion = 0        # Not Fashion
)

point_prediction <- predict(Model1, newdata = new_product_data)
print(paste("Predicted Sales: $", round(point_prediction, 2), "thousand"))
```

```
## [1] "Predicted Sales: $ 339.46 thousand"
```

Finding: The predicted monthly sales for this new product is $339,460.

Question 6.2: 95% Prediction Interval

```
# 95% Prediction Interval for a single new product's sales
prediction_interval <- predict(Model1, newdata = new_product_data, interval = "prediction", level =
0.95)
print(prediction_interval)
```

```
##        fit      lwr      upr
## 1 339.4607 291.7286 387.1928
```

- **Interpretation Statement:** We are 95% confident that the **actual sales for this single new product** in its first month will fall between **$291,730** and **$387,190**.

- **Business Planning Explanation:**

  - This interval is crucial for **tactical planning and risk assessment** for this *specific product launch*.

    - The wide range ($291.7k to $387.2k) quantifies the uncertainty and risk for a *single* month.
    - For business planning, the team should use the **lower bound ($291.7k)** as a conservative estimate or “worst-case scenario” for their initial revenue and cash flow projections. The **upper bound ($387.2k)** could be used to set optimistic sales targets or to ensure they have enough initial inventory to meet a high-demand scenario.

Question 6.3: 95% Confidence Interval for Mean Sales

```r
# 95% Confidence Interval for the average sales of products with these characteristics
confidence_interval <- predict(Model1, newdata = new_product_data, interval = "confidence", level = 0.95)
print(confidence_interval)
```

```
##        fit      lwr      upr
## 1 339.4607 336.8961 342.0253
```

- **Interpretation Statement:** The 95% confidence interval for the **average sales** of *all* products with these characteristics is from **$336,900** to **$342,025**.

- **Strategic Planning Explanation:**

    - This interval is used for **high-level strategic planning and forecasting**. It doesn’t predict one single month; it predicts the *long-term average* performance of this entire product profile.
    - The interval is very narrow (only about $5.1k wide), which gives Tracy a highly reliable and precise estimate.
    - For strategic planning, she can use this range to confidently forecast the revenue for an entire portfolio. For example, if MarketPlace plans to launch 20 similar products, she can reliably budget for an average sales figure between $336.9k and $342.0k *per product*. This is the number to use in long-term financial models to assess the overall profitability of this product strategy.

Question 6.4: Prediction vs. Confidence Interval

- **Why is the prediction interval wider than the confidence interval?**
    - The **Confidence Interval** (CI) is narrow because it only accounts for *one* source of uncertainty: our uncertainty about the **true average** sales (i.e., how well we’ve estimated the regression line).
    - The **Prediction Interval** (PI) is much wider because it must account for *two* sources of uncertainty:
        1. The same uncertainty about the regression line (from the CI).
        2. The inherent, random, and unpredictable variability of any **single new observation** around that line (known as the irreducible error).
    - In short, predicting an *average* is much safer than predicting a *single specific outcome*, so the PI’s range must be wider to account for that extra individual-level risk.
- **When should Tracy use the Prediction Interval?**
    - **For: Tactical, individual-level decisions.**
    - **Practical Business Example:** Tracy should use the PI ([$291.7k, $387.2k]) for planning the **specifics of this single product launch**. For example, she should use the lower bound ($291.7k) to set a conservative revenue budget for this *one product* or use the upper bound ($387.2k) to decide on the initial inventory order to avoid a stockout if it’s a hit.
- **When should Tracy use the Confidence Interval?**
    - **For: Strategic, long-term average decisions.**
    - **Practical Business Example:** Tracy should use the CI ([$336.9k, $342.0k]) for **high-level financial planning**. When the CFO asks, “What is the expected long-term profitability of this *type* of product strategy?” this is the number to use. It gives a very precise estimate of the *average revenue* she can expect from all future products that share these characteristics, which is perfect for building an annual budget or a long-term business case.

PART 7

Question 7.1: State the hypotheses

Full Model (Model 1) Equation:

$$Sales = \beta_0 + \beta_1 \text{Ad\_Spend} + \beta_2 \text{Email\_Campaigns} + \beta_3 \text{Website\_Traffic} + \beta_4 \text{Avg\_Rating} + \ldots + \beta_{13} \text{Product\_Type\_Fashion} + \epsilon$$

Reduced Model (Model 3 - excludes the 3 marketing variables) Equation:

$$Sales = \beta_0 + \beta_3 \text{Website\_Traffic} + \beta_4 \text{Avg\_Rating} + \beta_5 \text{Discount\_Pct} + \ldots + \beta_{13} \text{Product\_Type\_Fashion} + \epsilon$$

Statistical Hypotheses:

- **Null Hypothesis (H_0):** $\beta_{\text{Ad\_Spend}} = \beta_{\text{Email\_Campaigns}} = \beta_{\text{Social\_Media\_Engagement}} = 0$ (i.e., $\beta_1 = \beta_2 = \beta_8 = 0$)
- **Alternative Hypothesis (H_1):** At least one of $\beta_{\text{Ad\_Spend}}$, $\beta_{\text{Email\_Campaigns}}$, or $\beta_{\text{Social\_Media\_Engagement}}$ is not equal to 0.

Business Interpretation:

- H_0**:** After accounting for website traffic, ratings, pricing, inventory, season, and product type, the combined effect of Ad Spend, Email Campaigns, and Social Media Engagement does *not* significantly help explain sales. Essentially, these specific marketing activities, as a group, add no extra explanatory value.

- H_1: At least one of these three marketing activities significantly adds to the explanation of sales, even when controlling for all other factors in the model. The marketing portfolio provides significant additional insight.

Question 7.2: Build Model 3 (Reduced Model)

```
Model3_Reduced <- lm(Sales ~ Website_Traffic + Avg_Rating + Discount_Pct +
                       Competitor_Price_Index + Inventory_Level + Season_Spring +
                       Season_Summer + Season_Fall + Product_Type_Home +
                       Product_Type_Fashion,
                       data = dataset)

summary_m3 <- summary(Model3_Reduced)
r_sq_m3 <- summary_m3$r.squared
sse_m3 <- sum(residuals(Model3_Reduced)^2)

cat("R-squared for Model 3 (Reduced):", round(r_sq_m3, 4), "\n")
```

```
## R-squared for Model 3 (Reduced): 0.7791
```

```
cat("SSE for Model 3 (SSE_R):", round(sse_m3, 2), "\n")
```

```
## SSE for Model 3 (SSE_R): 12716986
```

Question 7.3: Calculate the partial F-statistic

```
partial_f_test_result <- anova(Model3_Reduced, Model1)
print(partial_f_test_result)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Website_Traffic + Avg_Rating + Discount_Pct + Competitor_Price_Index +
##     Inventory_Level + Season_Spring + Season_Summer + Season_Fall +
##     Product_Type_Home + Product_Type_Fashion
## Model 2: Sales ~ Ad_Spend + Email_Campaigns + Website_Traffic + Avg_Rating +
##     Discount_Pct + Competitor_Price_Index + Inventory_Level +
##     Social_Media_Engagement + Season_Spring + Season_Summer +
##     Season_Fall + Product_Type_Home + Product_Type_Fashion
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1   3588 12716986
## 2   3585  2118679  3  10598306 5977.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
partial_F_stat <- partial_f_test_result$F[2]
partial_p_value <- partial_f_test_result$`Pr(>F)`[2]

cat("\nPartial F-statistic:", round(partial_F_stat, 4), "\n")
```

```
##
## Partial F-statistic: 5977.769
```

```
cat("p-value:", format(partial_p_value, scientific = TRUE, digits=4), "\n")
```

```
## p-value: 0e+00
```

Question 7.4: Conclusion and Business Recommendation

Statistical Conclusion: Since the p-value (essentially 0) is much smaller than the significance level $\alpha = 0.05$, we **reject the null hypothesis (H_0)**.

Business Recommendation to the CFO: The partial F-test provides extremely strong statistical evidence (F = 5977.8, p < 0.001) that the group of marketing variables (Ad Spend, Email Campaigns, Social Media Engagement) *jointly* adds significant explanatory power to the sales model, even after considering all other factors. This means these marketing activities, taken together, are crucial drivers of sales performance. The recommendation is to **maintain or potentially increase investment** in these areas, as they are demonstrably linked to revenue. Reducing this budget would likely lead to a significant decrease in sales, according to the model.

PART 8

Question 8.1: Define Full and Reduced Models & Report SSE

Full Model (Model 1): Includes all 13 predictors.

Reduced Model (Model 3): Excludes the 3 marketing variables (Ad_Spend, Email_Campaigns, Social_Media_Engagement). This is the same reduced model used in the Part 7 partial F-test.

```
sse_full_model1 <- sum(residuals(Model1)^2)
cat("SSE for the Full Model (SSE_F - Model 1):", round(sse_full_model1, 2))
```

```
## SSE for the Full Model (SSE_F - Model 1): 2118679
```

```
sse_reduced_model3 <- sum(residuals(Model3_Reduced)^2)
cat("SSE for the Reduced Model (SSE_R - Model 3):", round(sse_reduced_model3, 2))
```

```
## SSE for the Reduced Model (SSE_R - Model 3): 12716986
```

Question 8.2: Calculate Partial R^2

```
partial_r_squared_marketing <- (sse_reduced_model3 - sse_full_model1) / sse_reduced_model3

cat("Partial R^2 =",
    paste0("(", round(sse_reduced_model3, 2), " - ", round(sse_full_model1, 2), ") / ",
           round(sse_reduced_model3, 2)))
```

```
## Partial R^2 = (12716985.85 - 2118679.48) / 12716985.85
```

```
cat("Coefficient of Partial Determination for Marketing Variables:", round(partial_r_squared_marketin
g, 4))
```

```
## Coefficient of Partial Determination for Marketing Variables: 0.8334
```

Question 8.3: Interpret the result

Interpretation: After controlling for all other variables in the model (including website traffic, average rating, pricing, inventory, season, and product type), the three main marketing controllable variables (Ad Spend, Email Campaigns, and Social Media Engagement) collectively explain **83.34%** of the sales variance that those other factors failed to account for. This extremely high value demonstrates that our marketing activities provide substantial and unique insight into sales performance. They are not redundant with other metrics.

Business Decision-Making Implication: This is a very high value and indicates that these three marketing activities provide substantial and unique explanatory power *beyond* what the other operational and market factors capture. Removing them would significantly weaken the model's ability to explain sales variations. Therefore, these marketing variables are critically important, provide unique insights, and should **absolutely remain** in the model used for analysis and decision-making regarding budget allocation and sales forecasting. They represent key levers the company can pull to influence revenue.

PART 9 (A: Seasonal Analysis)

Question 9.1: Rank all four seasons

```
coefs_model1 <- summary(Model1)$coefficients

season_coeffs <- c(
  Winter = 0, # Reference category explicitly set to 0
  Spring = coefs_model1["Season_Spring", "Estimate"],
  Summer = coefs_model1["Season_Summer", "Estimate"],
  Fall = coefs_model1["Season_Fall", "Estimate"]
)

season_ranking <- sort(season_coeffs, decreasing = TRUE)

season_table <- data.frame(
  Season = names(season_ranking),
  Expected_Sales_Adjustment_vs_Winter_Thousand_USD = unname(season_ranking),
  row.names = NULL
)

kable(season_table,
      digits = 2,
      caption = "Ranking of Seasons by Expected Sales Impact (Relative to Winter)",
      col.names = c("Season", "Sales Adjustment vs. Winter ($k)"),
      align = 'lr')
```

Ranking of Seasons by Expected Sales Impact (Relative to Winter)

Season	Sales Adjustment vs. Winter ($k)
Winter	0.00
Spring	-19.77
Fall	-19.89
Summer	-30.34

Interpretation: Based on the coefficients of Model 1, holding all other factors constant:

1. **Winter** has the highest expected sales (as it’s the reference category with a 0 adjustment).
2. **Spring** and **Fall** have similar, lower expected sales compared to Winter (approx. $19.8k lower).
3. **Summer** has the lowest expected sales (approx. $30.3k lower than Winter).

Question 9.2: Calculate Winter vs. Summer difference

```
winter_summer_diff <- season_coeffs["Winter"] - season_coeffs["Summer"]
abs(winter_summer_diff)
```

```
##  Winter
## 30.3354
```

```
summer_p_value <- coefs_model1["Season_Summer", "Pr(>|t|)"]
summer_p_value
```

```
## [1] 3.82288e-139
```

Business Implication: The model confirms a significant seasonal slump in Summer compared to the peak Winter season. On average, Summer sales are expected to be about **$30.34k lower** than Winter sales, even when controlling for marketing, product type, and other factors. This statistically significant difference is crucial for **inventory planning** (avoiding Summer overstock) and **marketing strategy** (potentially shifting budget or using targeted promotions to mitigate the Summer downturn).

Question 9.3: Test Spring vs. Fall

```
# H0: Season_Spring - Season_Fall = 0
spring_fall_test <- linearHypothesis(Model1, "Season_Spring - Season_Fall = 0")
spring_fall_p_value <- spring_fall_test$`Pr(>F)`[2]
spring_fall_p_value
```

```
## [1] 0.9159895
```

Business Implication: Although the point estimates differ slightly (-19.77k for Spring vs. -19.89k for Fall), the difference is **not statistically significant** (p = 0.916). For practical business purposes, MarketPlace can treat Spring and Fall as having a similar impact on sales compared to Winter. This could simplify seasonal planning, allowing for similar inventory levels and marketing approaches during these two “shoulder” seasons.

PART 9 (B: Product Type Analysis)

Question 9.4: Rank all three product types

```
product_coeffs <- c(
  Electronics = 0, # Reference category explicitly set to 0
  Home = coefs_model1["Product_Type_Home", "Estimate"],
  Fashion = coefs_model1["Product_Type_Fashion", "Estimate"]
)

product_ranking <- sort(product_coeffs, decreasing = TRUE)

product_table <- data.frame(
  Product_Type = names(product_ranking),
  Sales_Adjustment_vs_Electronics_Thousand_USD = unname(product_ranking),
  row.names = NULL
)

kable(product_table,
      digits = 2,
      caption = "Ranking of Product Types by Expected Sales Impact (Relative to Electronics)",
      col.names = c("Product Type", "Sales Adjustment vs. Electronics ($k)"),
      align = 'lr')
```

Ranking of Product Types by Expected Sales Impact (Relative to Electronics)

Product Type	Sales Adjustment vs. Electronics ($k)
Electronics	0.00
Home	-29.72
Fashion	-44.69

Interpretation: Based on Model 1 coefficients, holding all other factors constant:

1. **Electronics** has the highest expected sales (reference category).
2. **Home** products have significantly lower expected sales than Electronics (approx. $29.72k less).
3. **Fashion** products have the lowest expected sales, significantly lower than both Electronics (approx. $44.69k less) and Home goods.

Question 9.5: Home vs. Fashion comparison

```
home_fashion_diff <- product_coeffs["Home"] - product_coeffs["Fashion"]
home_fashion_diff
```

```
##     Home 
## 14.97747
```

```
# H0: Product_Type_Home - Product_Type_Fashion = 0
home_fashion_test <- linearHypothesis(Model1, "Product_Type_Home - Product_Type_Fashion = 0")
home_fashion_p_value <- home_fashion_test$`Pr(>F)`[2]
home_fashion_p_value
```

```
## [1] 1.435708e-50
```

Business Implication: Home products are expected to generate significantly more sales (approx. **$14.98k more per month**) than Fashion products, even after controlling for marketing spend, seasonality, ratings, etc. This difference is statistically significant (p < 0.001).

Should MarketPlace invest more in one category over the other? Not necessarily based *solely* on this coefficient. While Home goods generate more revenue *on average* for the *same* level of other inputs, this doesn’t automatically mean they are more *profitable* or have higher *growth potential*. The Fashion category might have higher margins, faster inventory turnover, or appeal to a different, valuable customer segment.

Recommendation: MarketPlace should investigate *why* Fashion underperforms relative to Home goods in terms of sales volume generated per unit of input (like ad spend or traffic). Is it product assortment, pricing strategy within the category, conversion rates on the website for fashion items, or overall market demand? Simply shifting investment *away* from Fashion might be premature without understanding the root cause and the category’s overall contribution to profitability and customer acquisition. A deeper dive into the Fashion category’s specific performance metrics is warranted.