

# Sales Performance Regression Analysis Report

## 1. Objective of the Study

The objective of this analysis is to examine the relationship between sales performance and key business drivers, specifically:

- Advertising budget
- Salesperson experience
- Regional economic conditions

The study aims to determine:

1. Which factors significantly influence sales
2. Whether a simple linear regression is sufficient or a multiple regression model provides better explanatory power
3. How well sales can be predicted using these variables

## 2. Dataset Description

The dataset consists of sales performance data across multiple office locations, with the following variables:

- Advertising\_Budget: Amount spent on advertising
- Salesperson\_Experience: Years of experience of the salesperson
- Regional\_Economic\_Index: Indicator of regional economic strength
- Sales: Total sales (response variable)

Only numeric variables were used for statistical modeling to ensure validity of regression assumptions

## 3. Descriptive Statistics

Summary statistics indicate:

- Average advertising budget is approximately 9,792
- Mean salesperson experience is about 5 years
- Average regional economic index is around 51
- Mean sales value is approximately 494,219

The distributions of all numeric variables appear reasonably symmetric, with no extreme skewness or obvious anomalies based on histogram inspection.

## **4. Exploratory Data Analysis**

Histogram analysis shows:

- Advertising budgets vary moderately across locations
- Salesperson experience is centered around mid-level experience
- Regional economic conditions differ across offices
- Sales values are spread across a wide range, suggesting potential influence from multiple factors

These observations justify further regression modeling to quantify relationships.

## **5. Simple Linear Regression Model**

### **Model Specification**

A simple linear regression model was fitted with:

- Sales as the response variable
- Advertising\_Budget as the sole predictor

### **Key Results**

- Advertising budget has a strong positive and statistically significant effect on sales
- For every 1-unit increase in advertising budget, sales increase by approximately 49 units
- The model explains 99.05% of the variance in sales ( $R^2 = 0.9905$ )

### **Model Fit**

- Residual standard error is relatively small compared to the scale of sales
- Scatter plot with regression line confirms a strong linear relationship

This indicates that advertising spend alone is an extremely powerful predictor of sales in this dataset.

## **6. Multiple Linear Regression Model**

### **Model Specification**

A multiple regression model was fitted using:

- Advertising\_Budget
- Salesperson\_Experience
- Regional\_Economic\_Index

### **Key Findings**

- Advertising\_Budget remains highly significant

- Salesperson\_Experience and Regional\_Economic\_Index are not statistically significant
- $R^2$  increases only marginally to 0.9906
- Adjusted  $R^2$  does not improve, indicating no meaningful gain from added predictors

## **Interpretation**

Although the multiple regression model includes more variables, it does not provide a better explanation of sales compared to the simple model. This suggests that advertising budget already captures most of the explainable variation.

## **7. Model Comparison (ANOVA)**

An ANOVA comparison between the simple and multiple models shows:

- No statistically significant reduction in residual variance
- Additional predictors do not materially improve model performance

Therefore, the simpler model is preferred based on parsimony and interpretability.

## **8. Prediction Example**

Using the multiple regression model:

- Advertising budget = 40,000
- Salesperson experience = 5 years
- Regional economic index = 120

The predicted sales value is approximately 1,981,592, demonstrating the practical use of the model for forecasting.

## **9. Business Insights**

- Advertising investment is the dominant driver of sales
- Increasing salesperson experience or operating in stronger economic regions does not significantly boost sales unless advertising spend increases
- Organizations can confidently prioritize advertising allocation decisions using this model
- A simple regression model is sufficient for reliable sales forecasting in this context

## **10. Conclusion**

This analysis shows that sales performance is overwhelmingly driven by advertising budget, with minimal contribution from salesperson experience or regional economic conditions. The simple linear regression model provides an excellent fit, strong interpretability, and high predictive power. From both a statistical and business perspective, the simpler model is preferred for decision-making and forecasting applications.

## Appendix

```
df <- read.csv("Sales_Performance.csv", stringsAsFactors = FALSE)

colnames(df)

## [1] "Location"           "Advertising_Budget"
## [3] "Salesperson_Experience" "Regional_Economic_Index"
## [5] "Sales"

head(df)      # first 6 rows

##   Location Advertising_Budget Salesperson_Experience
## Regional_Economic_Index
## 1      Of-1             10993                      2
## 2      Of-2              9723                      4
## 3      Of-3             11295                      4
## 4      Of-4             13046                      3
## 5      Of-5              9532                      5
## 6      Of-6              9532                      6

##   Sales
## 1 543538
## 2 483763
## 3 575616
## 4 661315
## 5 479439
## 6 481476

# Select only numeric columns
numeric_df <- df[sapply(df, is.numeric)]


# Mean
colMeans(numeric_df)

##   Advertising_Budget Salesperson_Experience Regional_Economic_Index
##                   9792.32                      5.08                      50.67
##   Sales
##                   494219.02


# Median
apply(numeric_df, 2, median)
```

```

##      Advertising_Budget  Salesperson_Experience Regional_Economic_Index
##                         9746                               5                           51
##                         Sales
##                         496283

par(mfrow = c(2, 2)) # 4 plots on one screen

hist(df$Advertising_Budget,
     main = "Advertising Budget Distribution",
     xlab = "Advertising Budget",
     col = "lightblue", border = "white")

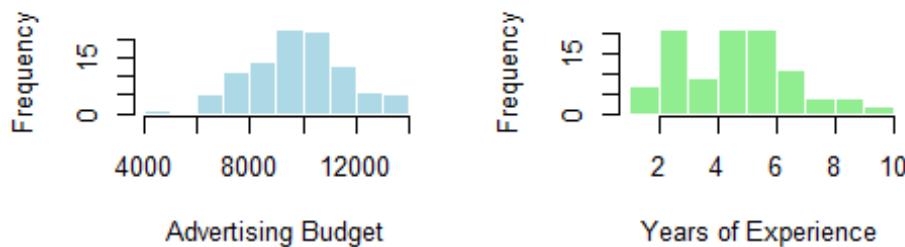
hist(df$Salesperson_Experience,
     main = "Salesperson Experience Distribution",
     xlab = "Years of Experience",
     col = "lightgreen", border = "white")

hist(df$Regional_Economic_Index,
     main = "Regional Economic Index Distribution",
     xlab = "Economic Index",
     col = "lightpink", border = "white")

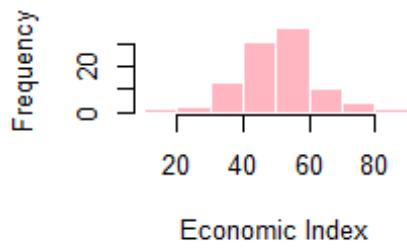
hist(df$Sales,
     main = "Sales Distribution",
     xlab = "Sales",
     col = "lightgoldenrod", border = "white")

```

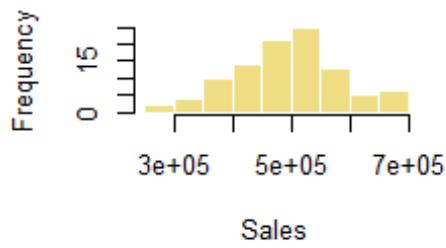
**Advertising Budget Distribution** **Salesperson Experience Distribution**



**Regional Economic Index Distribution**



**Sales Distribution**



```

summary(numeric_df)

## Advertising_Budget Salesperson_Experience Regional_Economic_Index
## Min.    : 4761      Min.    : 1.00      Min.    :18.00
## 1st Qu.: 8798      1st Qu.: 3.00      1st Qu.:43.00
## Median  : 9746      Median  : 5.00      Median  :51.00
## Mean    : 9792      Mean    : 5.08      Mean    :50.67
## 3rd Qu.:10812      3rd Qu.: 6.00      3rd Qu.:57.25
## Max.    :13705      Max.    :10.00      Max.    :89.00
## 
## Sales
## Min.    :263138
## 1st Qu.:441595
## Median :496283
## Mean   :494219
## 3rd Qu.:547615
## Max.   :679867

# 1. Fit simple Linear regression model
model_simple <- lm(Sales ~ Advertising_Budget, data = df)

# 2. View model summary (coefficients, R-squared, etc.)
summary(model_simple)

##
## Call:
## lm(formula = Sales ~ Advertising_Budget, data = df)
##
## Residuals:
##     Min      1Q      Median      3Q      Max
## -20334.2 -6872.8   -120.6   6469.5  20863.1
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.318e+04  4.836e+03   2.726   0.0076 **
## Advertising_Budget 4.912e+01  4.857e-01 101.145   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8777 on 98 degrees of freedom
## Multiple R-squared:  0.9905, Adjusted R-squared:  0.9904
## F-statistic: 1.023e+04 on 1 and 98 DF,  p-value: < 2.2e-16

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.5.2

ggplot(df, aes(x = Advertising_Budget, y = Sales)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Linear Regression: Sales ~ Advertising Budget",

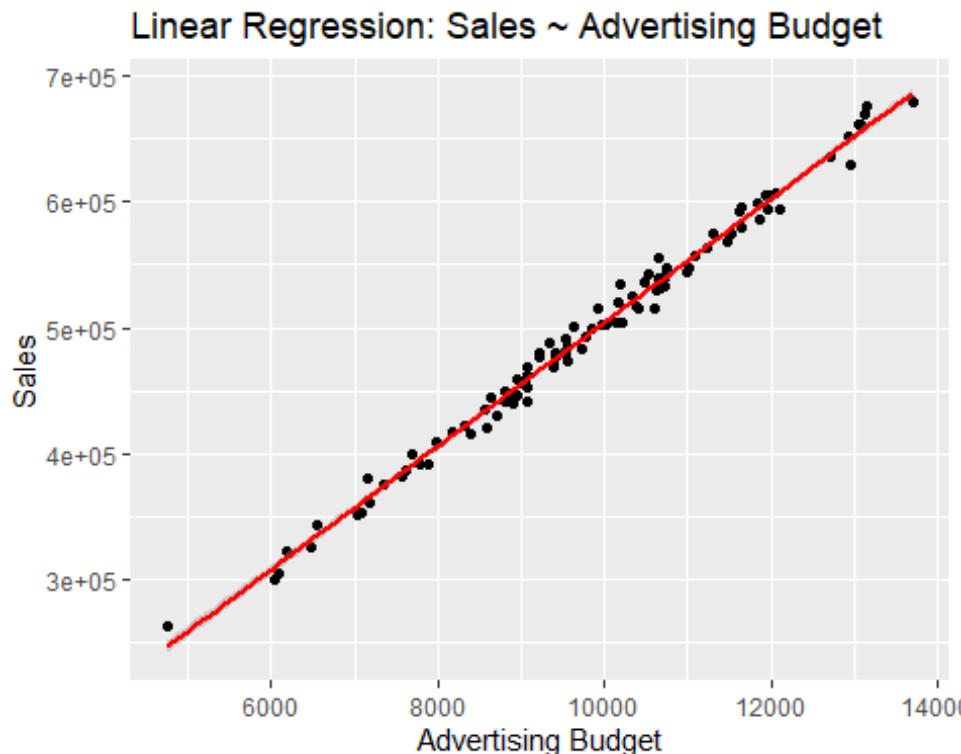
```

```

x = "Advertising_Budget",
y = "Sales")

## `geom_smooth()` using formula = 'y ~ x'

```



```

model_simple <- lm(Sales ~ Advertising_Budget, data = df)
summary(model_simple)

##
## Call:
## lm(formula = Sales ~ Advertising_Budget, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -20334.2  -6872.8   -120.6   6469.5  20863.1 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.318e+04  4.836e+03   2.726   0.0076 ** 
## Advertising_Budget 4.912e+01  4.857e-01 101.145  <2e-16 *** 
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 8777 on 98 degrees of freedom
## Multiple R-squared:  0.9905, Adjusted R-squared:  0.9904 
## F-statistic: 1.023e+04 on 1 and 98 DF,  p-value: < 2.2e-16

```

We fitted a simple linear regression model with Sales as the response and Advertising\_Budget as the predictor. The estimated intercept is 13,180 ( $t = 2.73$ ,  $p = 0.0076$ ), representing the predicted sales when advertising budget is zero. The slope for Advertising\_Budget is 49.12 ( $t = 101.15$ ,  $p < 0.001$ ), indicating that for each 1-unit increase in advertising budget, sales are expected to increase by about 49.12 units on average.

The model has an R-squared value of 0.9905, meaning that approximately 99.05% of the variation in sales is explained by advertising budget, which indicates an excellent model fit. The residual standard error is 8,777, suggesting that the model's predictions of sales deviate from the observed values by about 8,777 units on average. Overall, the model provides a very strong and statistically significant fit to the data.

multiple linear regression model

```
# 1. Fit multiple Linear regression model
model_multi <- lm(
  Sales ~ Advertising_Budget + Salesperson_Experience +
  Regional_Economic_Index,
  data = df
)

# 2. View model summary (coefficients, t-tests, R-squared, etc.)
summary(model_multi)

##
## Call:
## lm(formula = Sales ~ Advertising_Budget + Salesperson_Experience +
##     Regional_Economic_Index, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -20198.6  -6814.4    176.3   6203.9  21289.8 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 8696.1524  6592.1416   1.319   0.190    
## Advertising_Budget 49.1372    0.5023  97.827 <2e-16 ***
## Salesperson_Experience 413.4842   458.7332   0.901   0.370    
## Regional_Economic_Index 44.5160    83.2299   0.535   0.594    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 8818 on 96 degrees of freedom
## Multiple R-squared:  0.9906, Adjusted R-squared:  0.9903 
## F-statistic: 3379 on 3 and 96 DF,  p-value: < 2.2e-16

model_simple <- lm(Sales ~ Advertising_Budget, data = df)

summary(model_simple)
```

```

## 
## Call:
## lm(formula = Sales ~ Advertising_Budget, data = df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -20334.2  -6872.8   -120.6   6469.5  20863.1 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           1.318e+04  4.836e+03   2.726   0.0076 **  
## Advertising_Budget  4.912e+01  4.857e-01 101.145  <2e-16 ***  
## ---                
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 8777 on 98 degrees of freedom
## Multiple R-squared:  0.9905, Adjusted R-squared:  0.9904 
## F-statistic: 1.023e+04 on 1 and 98 DF,  p-value: < 2.2e-16 

summary(model_multi)

## 
## Call:
## lm(formula = Sales ~ Advertising_Budget + Salesperson_Experience +
##     Regional_Economic_Index, data = df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -20198.6  -6814.4    176.3   6203.9  21289.8 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           8696.1524  6592.1416   1.319   0.190    
## Advertising_Budget    49.1372    0.5023  97.827  <2e-16 ***  
## Salesperson_Experience 413.4842   458.7332   0.901   0.370    
## Regional_Economic_Index 44.5160    83.2299   0.535   0.594    
## ---                
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 8818 on 96 degrees of freedom
## Multiple R-squared:  0.9906, Adjusted R-squared:  0.9903 
## F-statistic: 3379 on 3 and 96 DF,  p-value: < 2.2e-16 

# Optional: ANOVA comparison
anova(model_simple, model_multi)

## Analysis of Variance Table
## 
## Model 1: Sales ~ Advertising_Budget
## Model 2: Sales ~ Advertising_Budget + Salesperson_Experience +
## Regional_Economic_Index

```

```
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    98 7549695465
## 2    96 7463888851  2  85806614 0.5518 0.5777
```

Although the multiple regression model has a very high R-squared (0.9906), only advertising budget is statistically significant, and the adjusted R-squared does not improve compared to the simple model. This suggests that advertising budget alone is sufficient to explain sales in this dataset, and the additional predictors do not materially improve the model.

```
new_data <- data.frame(
  Advertising_Budget      = 40000,
  Salesperson_Experience   = 5,
  Regional_Economic_Index = 120
)

predicted_sales <- predict(model_multi, newdata = new_data)
predicted_sales

##       1
## 1981592
```

Using the multiple linear regression model, for an advertising budget of 40,000, a salesperson with 5 years of experience, and a regional economic index of 120, the predicted sales are approximately 1,981,592 units (or dollars, depending on the scale of the data). This value was obtained using the predict() function in R on the fitted multiple regression model.