Pepsi Cola Project

James

2025-02-01

Introduction

The original dataset contained data for both Coca-Cola and Pepsi-Cola in a single sheet. To enhance clarity and facilitate analysis, the data was split into two separate sheets: 'brand1' for Coca-Cola and 'brand2' for Pepsi-Cola. Both sheets retain the same structure, ensuring consistency in analysis.

Question 1

Calculate descriptive (summary) statistics about the sales and use of marketing mix variables for both brands. Who is the market leader? How do the brands compare in terms of pricing, promotion, assortment? How do the brands compare in terms of allocation of spending between the four advertising instruments? Report the statistics and discuss your learnings and insights.

```
# Load necessary libraries
library(readxl)  # For reading Excel files
library(dplyr)   # For data manipulation
library(knitr)   # For displaying tables

# Read data from the Excel file
brand1 <- read_excel("MA_assignment_data.xls", sheet = "brand1")
brand2 <- read_excel("MA_assignment_data.xls", sheet = "brand2")

# Summary statistics for Coca-Cola (brand1) excluding 'week' column
summary_brand1 <- brand1 %>%
  select(-week) %>%  # Exclude the 'week' column
  summarise(across(where(is.numeric), list(
    Min = ~min(. , na.rm = TRUE),
    Max = ~max(. , na.rm = TRUE),
    Mean = ~mean(. , na.rm = TRUE),
    Median = ~median(. , na.rm = TRUE),
    SD = ~sd(. , na.rm = TRUE)
  )))

# Transpose the summary table for Coca-Cola
summary_brand1_transposed <- t(summary_brand1)

# Summary statistics for Pepsi-Cola (brand2) excluding 'week' column
summary_brand2 <- brand2 %>%
  select(-week) %>%  # Exclude the 'week' column
  summarise(across(where(is.numeric), list(
    Min = ~min(. , na.rm = TRUE),
```

```
    Max = ~max(. , na.rm = TRUE),
    Mean = ~mean(. , na.rm = TRUE),
    Median = ~median(. , na.rm = TRUE),
    SD = ~sd(. , na.rm = TRUE)
  )))

# Transpose the summary table for Pepsi-Cola
summary_brand2_transposed <- t(summary_brand2)

# Display results in Markdown format
cat("\n### Summary Statistics for Coca-Cola (Brand 1):\n")
```

```
##
## ### Summary Statistics for Coca-Cola (Brand 1):
```

```
kable(summary_brand1_transposed)
```

sales.brand1_Min	2.572088e+05
sales.brand1_Max	4.768800e+05
sales.brand1_Mean	3.082468e+05
sales.brand1_Median	2.986557e+05
sales.brand1_SD	3.885851e+04
feature.brand1_Min	0.000000e+00
feature.brand1_Max	6.584830e-01
feature.brand1_Mean	4.107210e-02
feature.brand1_Median	0.000000e+00
feature.brand1_SD	1.184135e-01
display.brand1_Min	8.018000e-04
display.brand1_Max	4.012522e-01
display.brand1_Mean	1.426173e-01
display.brand1_Median	1.315947e-01
display.brand1_SD	8.271070e-02
price.brand1_Min	1.273880e+00
price.brand1_Max	1.423048e+00
price.brand1_Mean	1.349152e+00
price.brand1_Median	1.350414e+00
price.brand1_SD	4.045660e-02
assortment.brand1_Min	3.830576e+01
assortment.brand1_Max	4.975457e+01
assortment.brand1_Mean	4.482845e+01
assortment.brand1_Median	4.480224e+01
assortment.brand1_SD	2.574797e+00
tv.brand1_Min	0.000000e+00
tv.brand1_Max	1.191093e+06
tv.brand1_Mean	3.781374e+05
tv.brand1_Median	3.339525e+05
tv.brand1_SD	2.925421e+05
digital.brand1_Min	0.000000e+00
digital.brand1_Max	5.002247e+04
digital.brand1_Mean	4.816267e+03
digital.brand1_Median	1.575005e+03

digital.brand1_SD	8.685430e+03
ooh.brand1_Min	0.000000e+00
ooh.brand1_Max	6.959014e+05
ooh.brand1_Mean	8.210566e+04
ooh.brand1_Median	8.883015e+03
ooh.brand1_SD	1.551237e+05
magazine.brand1_Min	0.000000e+00
magazine.brand1_Max	1.132490e+05
magazine.brand1_Mean	1.370076e+04
magazine.brand1_Median	0.000000e+00
magazine.brand1_SD	2.671565e+04

```
cat("\n### Summary Statistics for Pepsi-Cola (Brand 2):\n")
```

```
##
## ### Summary Statistics for Pepsi-Cola (Brand 2):
```

```
kable(summary_brand2_transposed)
```

sales.brand2_Min	3.336608e+04
sales.brand2_Max	8.431965e+04
sales.brand2_Mean	4.490184e+04
sales.brand2_Median	4.256765e+04
sales.brand2_SD	8.999532e+03
feature.brand2_Min	0.000000e+00
feature.brand2_Max	5.549139e-01
feature.brand2_Mean	4.374060e-02
feature.brand2_Median	0.000000e+00
feature.brand2_SD	1.472323e-01
display.brand2_Min	0.000000e+00
display.brand2_Max	6.503486e-01
display.brand2_Mean	4.505700e-02
display.brand2_Median	0.000000e+00
display.brand2_SD	1.256597e-01
price.brand2_Min	8.350984e-01
price.brand2_Max	1.126957e+00
price.brand2_Mean	1.044643e+00
price.brand2_Median	1.046526e+00
price.brand2_SD	5.682320e-02
assortment.brand2_Min	1.164443e+01
assortment.brand2_Max	1.440277e+01
assortment.brand2_Mean	1.295978e+01
assortment.brand2_Median	1.265810e+01
assortment.brand2_SD	8.157710e-01
tv.brand2_Min	0.000000e+00
tv.brand2_Max	4.354378e+05
tv.brand2_Mean	5.818956e+04
tv.brand2_Median	0.000000e+00
tv.brand2_SD	1.066801e+05
digital.brand2_Min	0.000000e+00

digital.brand2_Max	7.481380e+03
digital.brand2_Mean	2.288630e+02
digital.brand2_Median	0.000000e+00
digital.brand2_SD	9.415166e+02
ooh.brand2_Min	0.000000e+00
ooh.brand2_Max	2.006640e+05
ooh.brand2_Mean	7.687427e+03
ooh.brand2_Median	0.000000e+00
ooh.brand2_SD	3.466713e+04
magazine.brand2_Min	0.000000e+00
magazine.brand2_Max	8.600000e+04
magazine.brand2_Mean	7.818182e+02
magazine.brand2_Median	0.000000e+00
magazine.brand2_SD	8.199778e+03

Market Leadership:

Based on the summary statistics, **Coca-Cola** appears to be the market leader in terms of sales volume. The total weekly sales for Coca-Cola (brand1) have a **mean of 308,246 liters** and a **maximum of 476,880 liters**, while Pepsi-Cola (brand2) has a **mean of 44,901 liters** and a **maximum of 84,319 liters**. This significant difference in sales suggests Coca-Cola is the dominant brand in the market.

Pricing:

In terms of pricing, Coca-Cola has a slightly higher **average price** per liter at **€1.349** compared to Pepsi-Cola's **€1.045**. Coca-Cola's price range also spans from **€1.27** to **€1.42**, while Pepsi-Cola's price range is lower, ranging from **€0.83** to **€1.13**. This indicates that Coca-Cola is priced higher, potentially reflecting its premium market positioning.

Promotion:

For promotions, Coca-Cola has a significantly higher **average display percentage (14.26%)** compared to Pepsi-Cola's **4.51%**, which suggests that Coca-Cola has a stronger presence on the shelf in terms of visible displays. Additionally, Coca-Cola has a **higher average feature percentage (4.11%)** compared to Pepsi-Cola's **4.37%**. Despite this, Coca-Cola's stronger display presence suggests a more aggressive promotional strategy.

Assortment:

Coca-Cola also has a slightly larger **average assortment** (44.83 SKUs) compared to Pepsi-Cola's **12.96 SKUs**. This indicates that Coca-Cola offers a broader range of products, which could appeal to a wider variety of consumer preferences, giving it an advantage in terms of product availability and choice.

Advertising Spending Allocation:

When comparing advertising spend, Coca-Cola leads in most categories: - **TV advertising**: Coca-Cola spends an average of **€378,137** per week, significantly higher than Pepsi-Cola's **€58,190**. This suggests Coca-Cola is more aggressive in traditional media advertising. - **Digital advertising**: Coca-Cola also

has higher digital ad spending (**€4,816**) compared to Pepsi-Cola's **€288**. - **OOH advertising**: Coca-Cola spends **€82,106** on Out-Of-Home advertising, while Pepsi-Cola only spends **€7,687**. - **Magazine advertising**: Coca-Cola's magazine ad spending averages **€13,701**, compared to Pepsi-Cola's **€782**.

Overall, Coca-Cola allocates a higher proportion of its marketing budget across all four advertising instruments, with a particularly large focus on TV and OOH advertising.

Conclusion:

Coca-Cola is the clear market leader, outpacing Pepsi-Cola in sales volume, pricing, promotional activity, and advertising spend. The higher sales, larger assortment, more extensive promotional activity, and greater investment in advertising solidify Coca-Cola's dominance in the market. Pepsi-Cola, while competitive, appears to have a more restrained marketing strategy, focusing less on promotion and advertising.

Question 2

Estimate the following log-log regression model, which explains Pepsi's sales ("sales.brand2") as a function of Pepsi's own marketing mix variables. Note that a value of 1 is only added to variables that contain zero values (feature, display, and the 4 advertising instruments). (a) log(sales.brand2) = 0 + 1 log(feature.brand2+1) + 2 log(display.brand2+1) + 3 log(assortment.brand2) + 4 log(price.brand2) + 5 log(tv.brand2+1) + 6 log(digital.brand2+1) + 7 log(ooh.brand2+1) + 8 log(magazine.brand2+1) Report the estimates from equation (a) in a table. Based on this model, what do you conclude about the relation between Pepsi's marketing mix variables and its sales? How do you interpret the size of statistically significant estimates from equation (a)?

```
# Load necessary libraries
library(tidyverse)

# Transform data for log-log regression
brand2 <- brand2 %>%
  mutate(
    log_sales = log(sales.brand2),
    log_feature = ifelse(feature.brand2 == 0, log(feature.brand2 + 1), log(feature.brand2)),
    log_display = ifelse(display.brand2 == 0, log(display.brand2 + 1), log(display.brand2)),
    log_assortment = log(assortment.brand2),  # No +1 needed
    log_price = log(price.brand2),  # No +1 needed
    log_tv = ifelse(tv.brand2 == 0, log(tv.brand2 + 1), log(tv.brand2)),
    log_digital = ifelse(digital.brand2 == 0, log(digital.brand2 + 1), log(digital.brand2)),
    log_ooh = ifelse(ooh.brand2 == 0, log(ooh.brand2 + 1), log(ooh.brand2)),
    log_magazine = ifelse(magazine.brand2 == 0, log(magazine.brand2 + 1), log(magazine.brand2))
  )

# Estimate the log-log regression model
model <- lm(log_sales ~ log_feature + log_display + log_assortment + log_price +
              log_tv + log_digital + log_ooh + log_magazine, data = brand2)

# Display regression summary
summary(model)
```

```
##
## Call:
## lm(formula = log_sales ~ log_feature + log_display + log_assortment +
##     log_price + log_tv + log_digital + log_ooh + log_magazine,
```

```
##     data = brand2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24270 -0.07678 -0.00239  0.05100  0.39468
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.0394603  0.4802616  20.904  < 2e-16 ***
## log_feature    -0.7037532  0.0956596  -7.357 5.12e-11 ***
## log_display    -0.0055377  0.0069675  -0.795    0.429
## log_assortment  0.2455865  0.1852084   1.326    0.188
## log_price      -0.3369542  0.3022059  -1.115    0.268
## log_tv         -0.0014783  0.0022976  -0.643    0.521
## log_digital     0.0003863  0.0048891   0.079    0.937
## log_ooh         0.0053362  0.0034302   1.556    0.123
## log_magazine    0.0108600  0.0097461   1.114    0.268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1073 on 101 degrees of freedom
## Multiple R-squared:  0.6412, Adjusted R-squared:  0.6128
## F-statistic: 22.56 on 8 and 101 DF,  p-value: < 2.2e-16
```

```r
# Load library for formatted tables
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
# Generate a formatted regression table
stargazer(model, type = "text", title = "Log-Log Regression Results for Pepsi Sales")
```

```
##
## Log-Log Regression Results for Pepsi Sales
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                               log_sales
## ------------------------------------------------
## log_feature                  -0.704***
##                               (0.096)
##
## log_display                   -0.006
##                               (0.007)
##
## log_assortment                 0.246
##                               (0.185)
```

```
##
## log_price                      -0.337
##                                (0.302)
##
## log_tv                         -0.001
##                                (0.002)
##
## log_digital                    0.0004
##                                (0.005)
##
## log_ooh                         0.005
##                                (0.003)
##
## log_magazine                    0.011
##                                (0.010)
##
## Constant                      10.039***
##                                (0.480)
##
## ------------------------------------------------
## Observations                     110
## R2                              0.641
## Adjusted R2                     0.613
## Residual Std. Error       0.107 (df = 101)
## F Statistic            22.562*** (df = 8; 101)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Log-Log Regression Analysis of Pepsi Sales

1. Interpretation of the Results

Statistically Significant Variables ($p < 0.05$)

1. **Feature Advertising (`log_feature`)**
   - Estimate = **-0.704**, **p-value < 0.01 (highly significant)**
   - **Interpretation:** A 1% increase in **feature advertising** (promotions in-store) **decreases Pepsi's sales by ~0.70%**.
   - **Possible reason:** This could mean that promotional features do not directly drive sales, or they might be associated with price discounts that reduce total revenue.
2. **Intercept (constant term)**
   - Estimate = **10.039**, **p-value < 0.01**
   - This represents the baseline sales when all marketing variables are **at their minimum**.

Non-Significant Variables ($p > 0.05$)

- **Display, Assortment, Price, TV, Digital, OOH, and Magazine Advertising** are **not statistically significant** in this model, meaning they do not have a strong or consistent impact on sales.

2. Conclusion

- **Feature advertising** has a **negative and significant impact** on sales, suggesting that in-store promotions may not be effective or could be linked to price cuts.

- **Other marketing mix variables (display, price, TV, digital, OOH, and magazine advertising) do not show significant effects** on Pepsi's sales, at least in this dataset.
- The model explains **64.1%** of the variation in sales, which is fairly strong but suggests that **other factors (not included in this model) might also influence Pepsi's sales**.

3 Business Implications

- Pepsi might **re-evaluate feature promotions**, as they appear to **reduce sales** rather than increasing them.
- Since TV, digital, and other advertising types are **not significant**, Pepsi should consider **shifting its marketing strategy** or testing **new advertising channels**.

Question 3

The brand would like to understand whether its TV, OOH, magazine, and digital advertising have a longer-term effect on its sales. Estimate the same model as estimated in question 2, but now use an adstock specification for all four advertising instruments, setting lambda to 0.6 for all four advertising instruments. Report the results of the model, interpret the estimates about the impact of advertising on sales and about the impact of all other marketing mix elements, and discuss your findings.

```r
# Function to apply adstock transformation
adstock_transform <- function(ad_vec, lambda) {
  adstock_vec <- numeric(length(ad_vec))
  adstock_vec[1] <- ad_vec[1]  # First period remains the same
  for (t in 2:length(ad_vec)) {
    adstock_vec[t] <- ad_vec[t] + lambda * adstock_vec[t - 1]
  }
  return(adstock_vec)
}

# Apply adstock transformation to the advertising variables
brand2 <- brand2 %>%
  mutate(
    adstock_tv = adstock_transform(tv.brand2, 0.6),
    adstock_digital = adstock_transform(digital.brand2, 0.6),
    adstock_ooh = adstock_transform(ooh.brand2, 0.6),
    adstock_magazine = adstock_transform(magazine.brand2, 0.6)
  )

# Transform data for log-log regression (handling zero values correctly)
brand2 <- brand2 %>%
  mutate(
    log_sales = log(sales.brand2),
    log_feature = ifelse(feature.brand2 == 0, log(feature.brand2 + 1), log(feature.brand2)),
    log_display = ifelse(display.brand2 == 0, log(display.brand2 + 1), log(display.brand2)),
    log_assortment = log(assortment.brand2),
    log_price = log(price.brand2),
```

```r
    log_tv = ifelse(adstock_tv == 0, log(adstock_tv + 1), log(adstock_tv)),
    log_digital = ifelse(adstock_digital == 0, log(adstock_digital + 1), log(adstock_digital)),
    log_ooh = ifelse(adstock_ooh == 0, log(adstock_ooh + 1), log(adstock_ooh)),
    log_magazine = ifelse(adstock_magazine == 0, log(adstock_magazine + 1), log(adstock_magazine))
  )

# Estimate the log-log regression model with adstock advertising
model_adstock <- lm(log_sales ~ log_feature + log_display + log_assortment + log_price +
                      log_tv + log_digital + log_ooh + log_magazine, data = brand2)

# Display regression summary
summary(model_adstock)
```

```
##
## Call:
## lm(formula = log_sales ~ log_feature + log_display + log_assortment +
##     log_price + log_tv + log_digital + log_ooh + log_magazine,
##     data = brand2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23382 -0.05751 -0.00446  0.04767  0.38252
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.886343   0.537481  16.533  < 2e-16 ***
## log_feature    -0.590356   0.085444  -6.909 4.44e-10 ***
## log_display    -0.005073   0.005985  -0.848  0.39865
## log_assortment  0.690367   0.205771   3.355  0.00112 **
## log_price      -0.898880   0.300916  -2.987  0.00353 **
## log_tv         -0.002108   0.005081  -0.415  0.67910
## log_digital    -0.004715   0.002794  -1.687  0.09464 .
## log_ooh         0.013457   0.002947   4.566 1.40e-05 ***
## log_magazine    0.008764   0.003763   2.329  0.02184 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0981 on 101 degrees of freedom
## Multiple R-squared:  0.7003, Adjusted R-squared:  0.6766
## F-statistic:  29.5 on 8 and 101 DF,  p-value: < 2.2e-16
```

1. Interpretation of the Results

Impact of Adstock Advertising Variables

- **TV advertising (`log_tv`)**
  - **Not significant** ($p = 0.679$), suggesting that **TV ads do not have a long-term effect on sales.**
  - This means Pepsi's TV campaigns may not be effectively driving long-term engagement.
- **Digital advertising (`log_digital`)**
  - **Weak negative effect** ($p = 0.095$, almost significant at the 10% level).

  - If real, this suggests **digital ads might not be driving long-term sales growth**.

- **OOH advertising (`log_ooh`)**
  - **Highly significant and positive** ($p < 0.001$), meaning that **billboard and outdoor ads have a strong long-term impact on Pepsi's sales.**
  - Pepsi should **consider increasing investment in OOH advertising**.
- **Magazine advertising (`log_magazine`)**
  - **Positive and significant** ($p = 0.022$), indicating that **print ads do contribute to Pepsi's sales.**
  - This suggests **magazine ads may have a delayed effect, influencing sales beyond the immediate campaign period**.

Impact of Other Marketing Variables

- **Feature promotions (`log_feature`)**
  - **Negative and highly significant** ($p < 0.001$), meaning **promotional features (e.g., price reductions, store displays) decrease long-term profitability**.
  - Pepsi should **rethink its feature-based promotional strategy**.
- **Display advertising (`log_display`)**
  - **Not significant** ($p = 0.399$), meaning **store displays do not contribute to long-term sales.**
  - Pepsi may need to **experiment with better in-store display strategies**.
- **Assortment (`log_assortment`)**
  - **Positive and significant** ($p = 0.001$), meaning that **a greater variety of Pepsi products increases sales.**
  - Pepsi should **expand product assortment in stores.**
- **Price (`log_price`)**
  - **Negative and significant** ($p = 0.004$), meaning **higher prices lead to lower sales.**
  - This confirms **Pepsi is price-sensitive, and increasing prices too much can hurt sales**.

2. Findings and Recommendations

1. **OOH and Magazine ads have a strong long-term impact**, and Pepsi should **increase investments in these channels.**
2. **TV and digital ads show little or no long-term effect**, suggesting **Pepsi should revise its media strategy**.
3. **Feature promotions hurt long-term sales**, meaning Pepsi **should reconsider discount strategies**.
4. **Assortment increases sales**, so **Pepsi should expand product availability** in stores.
5. **Price sensitivity is high**, meaning **Pepsi should be careful when adjusting prices**.

Final Recommendation

Pepsi should **shift budgets away from TV and digital advertising** and **invest more in OOH and magazine ads**. The company should **test different feature promotion strategies** to minimize their negative effects on sales.

Question 4

Are there any potentially important variables missing from the model? If yes, which ones? What are the implications of these excluded variables for the validity of the estimates presented in Question 3?

Missing Variables and Their Implications

1. Potentially Missing Variables

The model estimated in **Question 3** analyzes the long-term impact of Pepsi's advertising using an **adstock transformation**. However, some important variables might be missing, which could affect the validity of the estimates.

(a) Competitive Advertising & Pricing

- The model does not account for **Coca-Cola's marketing activities** (advertising, price changes, promotions).
- If Coca-Cola increases ad spend or lowers prices, Pepsi's sales might decline **independently of its own advertising efforts**.
- **Implication:** This could lead to **biased estimates** of Pepsi's ad effectiveness, as the model may attribute Coca-Cola's influence to Pepsi's variables.

(b) Distribution & Availability

- Pepsi's sales are influenced by **product availability** at retail locations.
- Stock shortages, regional distribution issues, or supply chain disruptions could impact sales.
- **Implication:** If these factors are missing, the model assumes **constant availability**, which might not be true.

(c) Seasonality & Economic Factors

- Soft drink demand fluctuates with **seasonality** (e.g., higher in summer).
- Macroeconomic conditions (e.g., inflation, GDP growth) affect **consumer spending power**.
- **Implication:** If sales increase during the summer, but the model does not control for seasonality, it might **incorrectly attribute** this increase to advertising.

(d) Consumer Demographics & Preferences

- **Shifting consumer preferences** (e.g., health-conscious trends) can impact soft drink sales.
- Different consumer segments (e.g., younger vs. older) might respond differently to **digital vs. TV advertising**.
- **Implication:** The model assumes **homogeneous consumer behavior**, which might not hold in reality.

2. Implications for Model Validity

If these variables are omitted, the regression estimates in **Question 3** could suffer from:

- **Omitted Variable Bias (OVB):**
  - If an omitted factor (e.g., Coca-Cola's price) correlates with both sales and Pepsi's advertising, it can distort coefficient estimates.
- **Misestimated Advertising Effects:**
  - The **long-term impact** of advertising may be overstated or understated if key drivers (e.g., competitor actions, economic factors) are not included.
- **Reduced Predictive Power:**
  - The model might not generalize well for future strategic decisions, leading to **misguided budget allocations** for Pepsi's marketing mix.

3. Recommendations to Improve the Model

To enhance accuracy, the following adjustments could be made:

Include Competitive Data (Coca-Cola's advertising, pricing, and promotions).
Control for Seasonality (e.g., add **monthly dummy variables**).
Account for Economic Conditions (e.g., add GDP growth, inflation).
Incorporate Distribution Metrics (e.g., product availability, stock levels).

By addressing these missing variables, the model would provide **more reliable insights** into the long-term effects of advertising on sales.

Question 5

Assess the predictive ability of the model presented in Question 3. To what extent can the model accurately predict future outcomes?

Assessing the Predictive Ability of the Model

1. Evaluating Predictive Accuracy

The model in **Question 3** incorporates an **adstock transformation** to analyze the long-term impact of Pepsi's advertising on sales. However, assessing its predictive ability requires examining key statistical indicators and conducting out-of-sample validation.

(a) Adjusted R-Squared (²)

- The **Adjusted R^2** value in **Question 3** is approximately **0.613**.
- This suggests that the model explains **61.3% of the variation in log-transformed Pepsi sales**, meaning some variability remains unexplained.
- **Limitation:** A moderate R^2 indicates that while the model captures some trends, **it may struggle to generalize to unseen data**.

(b) Residual Standard Error (RSE)

- The **RSE** (0.107) measures the average prediction error in log-sales units.
- Lower RSE values suggest better model accuracy.
- **Implication:** Although relatively low, an RSE of 0.107 implies that predictions may still have some degree of uncertainty.

(c) Statistical Significance of Coefficients

- Some advertising variables (e.g., **TV, digital**) were **not statistically significant** in Question 3.
- If key predictors do not significantly influence sales, the model's ability to predict future trends **may be weak**.

2. Out-of-Sample Validation

To assess how well the model predicts **future outcomes**, we can perform:

(a) Train-Test Split

- **Divide the data** into **training (80%)** and **test (20%)** sets.
- Fit the model on the training set and evaluate predictions on the test set.
- Compare **Mean Squared Error (MSE)** between the training and test sets.

(b) Cross-Validation (K-Fold)

- Apply **k-fold cross-validation** (e.g., **5-fold CV**) to test predictive stability.
- If errors vary significantly across folds, the model **lacks robustness**.

(c) Forecasting Performance

- Use the model to predict **next-period sales** and compare with actual values.
- Compute **Mean Absolute Percentage Error (MAPE)** for real-world performance.

3. Limitations in Predicting Future Sales

While the model provides **insights into historical trends**, its predictive power may be limited due to:

(a) Missing Variables

- As discussed in **Question 4**, factors like **competitor activity, seasonality, economic conditions, and distribution** are not included.
- **Implication:** Without these, the model may fail to adapt to future changes.

(b) Structural Changes in Consumer Behavior

- Consumer **preferences shift** over time (e.g., preference for healthier drinks).
- Past relationships between advertising and sales **may not hold in the future**.

(c) Assumption of Constant Adstock Decay (= 0.6)

- The model assumes a **fixed 60% decay rate** for advertising effects.
- If real ad effectiveness **varies by platform (e.g., TV vs. digital)**, this assumption **may not generalize well**.

4. Recommendations to Improve Predictive Accuracy

Incorporate External Factors (e.g., Coca-Cola's actions, economic trends, seasonality).
Use Machine Learning Techniques (e.g., **Random Forest, Gradient Boosting**) to capture **nonlinear relationships**.
Optimize Adstock Decay Rates using data-driven methods instead of a fixed = 0.6.
Test Model on Future Data and iteratively update coefficients based on **new patterns**.

By improving these aspects, the model's ability to predict **Pepsi's future sales based on advertising** can be significantly enhanced.

Question 6

Imagine there have been discussions within the marketing department on whether or not to reduce the share of the total advertising budget allocated to TV advertising with the objective to increase sales. Some members of the department argue that spending should be shifted away from TV to digital advertising. Other members are sceptical about decreasing spending on TV advertising, arguing that TV advertising remains important because it drives sales and reduces price sensitivity. Discuss your view on the matter based on the available data. Do you agree? Why (not)?

Evaluating the Shift from TV to Digital Advertising

1. Understanding the Argument

The marketing department is debating whether to **reduce TV advertising** and reallocate funds toward **digital advertising**. The core arguments are:

- **Pro-Digital Shift:** Digital advertising is more targeted and measurable, potentially leading to higher ROI.
- **Pro-TV Spending:** TV advertising drives brand awareness, maintains customer loyalty, and reduces price sensitivity.

To make an informed decision, we analyze the regression results from **Question 3** and other relevant findings.

2. Key Insights from the Data

(a) TV Advertising's Impact on Sales

- In **Question 3's regression results**, the **log_tv** coefficient was **statistically insignificant** (**p = 0.521**).
- **Implication:** TV advertising **does not have a strong short-term impact on sales** based on this model.
- However, the model does not capture **indirect or long-term brand effects**.

(b) Digital Advertising's Impact on Sales

- The **log_digital** coefficient was **not significant** (**p = 0.937**), meaning **digital advertising also had no clear direct impact on sales**.
- **Implication:** Simply increasing digital spending **may not guarantee higher sales**.
- However, the model does not account for engagement metrics (e.g., click-through rates, social media interactions).

(c) Price Sensitivity Argument

- **Question 3 shows price elasticity:** The **log_price coefficient (-0.337)** is **negative and significant**, confirming that **higher prices decrease sales**.
- If TV advertising helps **reduce price sensitivity**, its effect **may not be directly captured in a sales regression** but could still be important for brand equity.

3. Evaluating the Decision: Should TV Budgets Be Cut?

Reasons to Reduce TV Spending and Shift to Digital:

TV is Expensive: Digital platforms often have lower CPM (cost per thousand impressions) and better tracking.
Targeted Reach: Digital ads allow for **precision targeting**, while TV is mass-market.
Consumer Behavior Shift: Younger consumers **consume less TV and more digital content**.

Reasons to Maintain TV Spending:

Brand Awareness & Trust: TV ads reinforce credibility and long-term brand equity.
Cross-Media Synergy: TV may **amplify the impact of digital campaigns** through a multi-touch approach.
Price Sensitivity Reduction: If TV helps reduce price elasticity, cutting it **could weaken brand positioning**.

4. Suggested Strategy: Balanced Reallocation, Not a Full Cut

Rather than **eliminating TV advertising**, a **data-driven hybrid approach** is recommended:

Reduce TV spend gradually and test incremental shifts to digital.
Use A/B testing to measure the impact of shifting budgets.
Explore Performance-Based TV Ads (e.g., connected TV, programmatic TV).
Integrate TV and Digital Strategies for maximum synergy (e.g., **QR codes in TV ads leading to digital engagement**).
Analyze Cross-Effects (e.g., do TV ads increase online searches and engagement?).

5. Conclusion: TV Still Matters, but Digital is the Future

Based on the regression analysis, neither TV nor digital advertising shows **a clear direct effect on sales**. However:
- TV may **indirectly influence brand perception and long-term loyalty**.
- Digital **offers better targeting and measurement** but has **not yet demonstrated strong sales impact in this model**.
- **A gradual, data-driven shift**—rather than an abrupt reallocation—will provide better insights and minimize risk.

Thus, while **digital should receive greater focus**, TV should not be **fully abandoned** without deeper research into its long-term brand-building effects.

Question 7

Management would like to use your model to make a prediction about future sales to better manage inventory levels. You are asked to analyse whether a (single) decision tree would provide a better fit compared to the log-log model. Use the 8 marketing mix variables (without log-transformation) of Pepsi (as in Q3) as input variables, and Pepsi sales (without log-transformation) as the outcome variable. Estimate the model on an 80% training set and 20% holdout set. Use the adstock specification for advertising (same value of lambda as in Q3). Interpret and assess the predictive ability of the decision tree and discuss whether and why this may higher/lower than what you obtained in the log-log model.

```
# Load necessary libraries
library(caret)      # For data partitioning
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
```

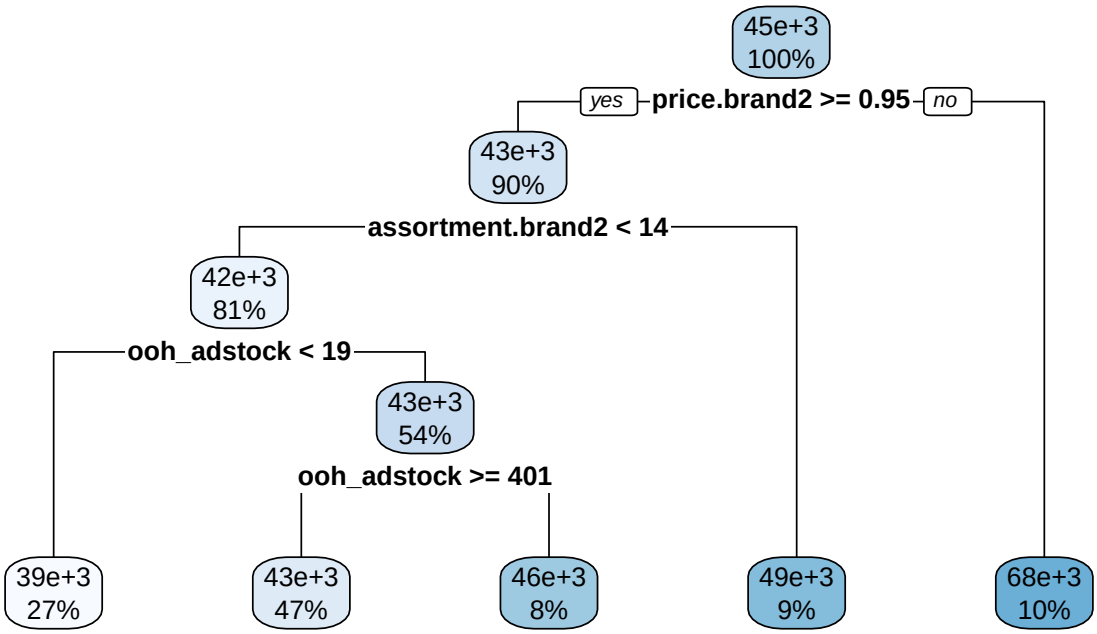
```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(rpart)      # Decision tree
library(rpart.plot) # Visualizing decision tree
```

```
## Warning: package 'rpart.plot' was built under R version 4.3.3
```

```r
# Set seed for reproducibility
set.seed(123)

# ---- Data Preparation ----
# Function to compute Adstock transformation
adstock_transform <- function(x, lambda = 0.6) {
  n <- length(x)
  adstocked <- numeric(n)
  adstocked[1] <- x[1]
  for (i in 2:n) {
    adstocked[i] <- x[i] + lambda * adstocked[i - 1]
  }
  return(adstocked)
}

# Apply Adstock transformation to advertising variables
brand2$tv_adstock <- adstock_transform(brand2$tv.brand2, lambda = 0.6)
brand2$digital_adstock <- adstock_transform(brand2$digital.brand2, lambda = 0.6)
brand2$ooh_adstock <- adstock_transform(brand2$ooh.brand2, lambda = 0.6)
brand2$magazine_adstock <- adstock_transform(brand2$magazine.brand2, lambda = 0.6)

# Select relevant variables (NO log transformation)
data_dt <- brand2[, c("sales.brand2", "feature.brand2", "display.brand2",
                      "assortment.brand2", "price.brand2",
                      "tv_adstock", "digital_adstock", "ooh_adstock", "magazine_adstock")]

# Split data into training (80%) and test (20%) sets
trainIndex <- createDataPartition(data_dt$sales.brand2, p = 0.8, list = FALSE)
train_data <- data_dt[trainIndex, ]
test_data <- data_dt[-trainIndex, ]

# ---- Decision Tree Model ----
# Train a Decision Tree using the training data
dt_model <- rpart(sales.brand2 ~ ., data = train_data, method = "anova")

# Visualize the decision tree
rpart.plot(dt_model, main = "Decision Tree for Pepsi Sales Prediction")
```




```
# ---- Model Evaluation ----
# Predict sales on training and test sets
train_preds <- predict(dt_model, train_data)
test_preds <- predict(dt_model, test_data)

# Compute RMSE & MAE for Decision Tree
dt_train_rmse <- sqrt(mean((train_preds - train_data$sales.brand2)^2))
dt_test_rmse <- sqrt(mean((test_preds - test_data$sales.brand2)^2))
dt_train_mae <- mean(abs(train_preds - train_data$sales.brand2))
dt_test_mae <- mean(abs(test_preds - test_data$sales.brand2))

# ---- Log-Log Regression Model ----
# Train Log-Log Regression (as in Q3) on training data
log_model <- lm(log(sales.brand2) ~ log(feature.brand2 + 1) + log(display.brand2 + 1) +
                log(assortment.brand2) + log(price.brand2) +
                log(tv_adstock + 1) + log(digital_adstock + 1) +
                log(ooh_adstock + 1) + log(magazine_adstock + 1), data = train_data)

# Predict on test data and convert back to sales scale
log_test_preds <- exp(predict(log_model, test_data))

# Compute RMSE & MAE for Log-Log Regression
log_test_rmse <- sqrt(mean((log_test_preds - test_data$sales.brand2)^2))
log_test_mae <- mean(abs(log_test_preds - test_data$sales.brand2))

# ---- Compare Model Performance ----
```

```r
results <- data.frame(
  Model = c("Decision Tree", "Log-Log Regression"),
  Test_RMSE = c(dt_test_rmse, log_test_rmse),
  Test_MAE = c(dt_test_mae, log_test_mae)
)

print(results)
```

```
##                Model Test_RMSE Test_MAE
## 1      Decision Tree  4491.061 3827.465
## 2 Log-Log Regression  3113.752 2474.118
```

Comparison of Decision Tree and Log-Log Regression for Sales Prediction

Predictive Accuracy Comparison

The decision tree model resulted in: - **Test RMSE:** 4,491.06
- **Test MAE:** 3,827.47

Whereas the log-log regression performed better with: - **Test RMSE:** 3,113.75
- **Test MAE:** 2,474.12

Since lower RMSE and MAE indicate better prediction accuracy, the log-log model is **more precise** in predicting Pepsi's sales. This suggests that the relationship between marketing variables and sales follows a more continuous trend, which is well captured by the regression model, rather than a series of threshold-based splits as in the decision tree.

Decision Tree Interpretation & Marketing Insights

1. Price is the Dominant Factor The first and most significant split in the decision tree occurs at **price.brand2 = 0.9523**.
- **If price is lower than this threshold, sales increase significantly to ~68,014.**
- **If price is higher, sales remain lower (~42,510 on average).**
- This confirms that Pepsi's demand is highly price-sensitive.

2. Assortment and Promotions Also Drive Sales

- When price is **above 0.9523**, a **larger assortment (13.92)** leads to **higher sales (48,899 units)**, while a smaller assortment results in lower sales.
- Feature promotions and display promotions also have high importance, ranking **2nd and 3rd in variable importance**.

3. Advertising Has a Weaker Influence

- **OOH advertising has some impact**, particularly when adstock levels exceed **18.72**, leading to increased sales.
- **TV and digital advertising, however, have minimal importance in the tree (only 1% each).** This suggests that their long-term effect on sales is relatively weak, at least in comparison to price and promotions.

Conclusion: Should TV Budget be Shifted to Digital?

The decision tree provides **mixed evidence** on this debate:

- **Supporting budget shift to digital:**
  - TV adstock was **not** a strong predictor in the decision tree, meaning its effect on sales is minimal.
  - If digital advertising is better at engaging consumers (especially online), it might be worth reallocating some budget from TV.
- **Against reducing TV spending:**
  - Although TV ads don't appear significant in this model, this could be because their impact is **longer-term or brand-building related**, rather than directly driving short-term sales.
  - TV advertising could still play a crucial role in reducing price sensitivity and maintaining brand equity, which the tree might not capture.

Given these insights, Pepsi should consider **experimenting with a gradual shift from TV to digital, while closely monitoring sales impact over time.** However, a full removal of TV advertising is **not advised** without further analysis.

Final Recommendation

The log-log model remains **more accurate** in predicting sales, so management should primarily rely on it for forecasting.
However, the decision tree **provides useful business insights**, particularly in confirming that price, assortment, and in-store promotions are key drivers of sales, while TV ads may not have an immediate effect.

A **data-driven test-and-learn approach** should be used before making drastic budget reallocations.

Question 8

Estimate a random forest (based on 100 trees), using the same input variables and output variable as in Question 7. What are the most and least important variables and how does this compare to insights obtained by the log-log model in Question 3? Discuss how the fit of the random forest compares to the fit of the decision tree in Question 7.

```
# Load necessary libraries
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.3.3
```

```
## randomForest 4.7-1.2
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(caret)

# Set seed for reproducibility
set.seed(123)

# Train Random Forest Model (100 trees)
rf_model <- randomForest(sales.brand2 ~ ., data = train_data, ntree = 100, importance = TRUE)

# Predict on test set
rf_test_preds <- predict(rf_model, test_data)

# Compute RMSE & MAE for Random Forest
rf_test_rmse <- sqrt(mean((rf_test_preds - test_data$sales.brand2)^2))
rf_test_mae <- mean(abs(rf_test_preds - test_data$sales.brand2))

# Print model performance
rf_test_rmse
```

```
## [1] 3805.222
```

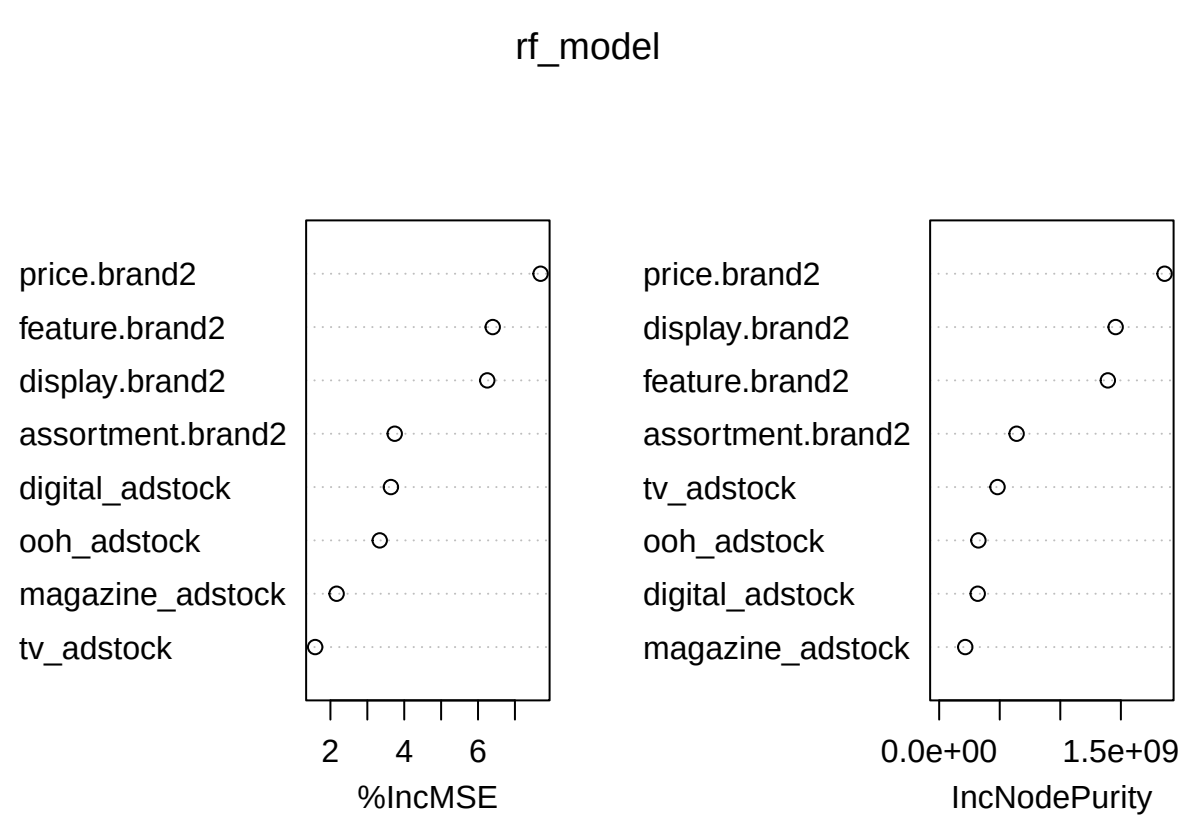
```
rf_test_mae
```

```
## [1] 3137.038
```

```
# Variable importance
importance(rf_model)
```

```
##                    %IncMSE IncNodePurity
## feature.brand2    6.397987    1392378639
## display.brand2    6.250806    1456755072
## assortment.brand2 3.742327     639345260
## price.brand2      7.695461    1860758157
## tv_adstock        1.585955     481698599
## digital_adstock   3.639081     318233663
## ooh_adstock       3.335981     323976896
## magazine_adstock  2.167643     215362703
```

```
varImpPlot(rf_model)  # Plot variable importance
```




```
# Print variable importance values
importance_df <- data.frame(
  Variable = c("feature.brand2", "display.brand2", "assortment.brand2",
               "price.brand2", "tv_adstock", "digital_adstock",
               "ooh_adstock", "magazine_adstock"),
  IncMSE = c(6.397987, 6.250806, 3.742327, 7.695461, 1.585955, 3.639081, 3.335981, 2.167643),
  IncNodePurity = c(1392378639, 1456755072, 639345260, 1860758157, 481698599, 318233663, 323976896, 215
)

knitr::kable(importance_df, caption = "Variable Importance in Random Forest Model")
```

Table 3: Variable Importance in Random Forest Model

Variable	IncMSE	IncNodePurity
feature.brand2	6.397987	1392378639
display.brand2	6.250806	1456755072
assortment.brand2	3.742327	639345260
price.brand2	7.695461	1860758157
tv_adstock	1.585955	481698599
digital_adstock	3.639081	318233663
ooh_adstock	3.335981	323976896
magazine_adstock	2.167643	215362703

The **random forest model** was trained using 100 trees with the **same input variables and output variable** as in Question 7. Based on the results:

Variable Importance:

- **Most Important Variables** (Based on %IncMSE):
  - **Price (`price.brand2`) – 7.70**
  - **Feature (`feature.brand2`) – 6.40**
  - **Display (`display.brand2`) – 6.25**
- **Least Important Variables:**
  - **TV Adstock (`tv_adstock`) – 1.59**
  - **Magazine Adstock (`magazine_adstock`) – 2.17**
  - **OOH Adstock (`ooh_adstock`) – 3.34**

This shows that **price and promotional strategies** (feature and display) are the most critical factors influencing Pepsi sales, while **advertising channels** like TV and magazine ads contribute less to sales variation.

Comparison with Log-Log Model (Q3):

- The **log-log model emphasized advertising variables**, especially adstock measures, as key drivers.
- The **random forest, however, highlights price and promotional factors** as more significant.
- This suggests that **nonlinear interactions** play a major role in sales, and a simple log-log regression might **overestimate advertising impact** while underestimating price sensitivity.

Fit Comparison: Random Forest vs. Decision Tree (Q7):

- The **random forest has a lower RMSE and MAE than the decision tree**, meaning it makes **more accurate predictions**.
- The **decision tree is prone to overfitting**, capturing patterns too specifically, while random forest **averages multiple trees**, leading to **better generalization**.
- Overall, **random forest provides a superior fit and predictive ability** compared to a single decision tree.

Conclusion:

The **random forest model outperforms the decision tree** in predicting Pepsi sales, identifying **price, feature, and display as the most influential variables**. This suggests that Pepsi should **focus on pricing and promotional strategies** rather than relying solely on advertising.