You have been hired to determine the impact of advertising on gross sales revenue for “Four Musketeers” candy bars. Four Musketeers has the same price and basically the same ingredients as competing candy bars, so you theorize that advertising is the only thing affecting sales. You decide to build a model of sales as a function of advertising.
Data: Import Week12Candy.xlsx

1. Drop the first four observations.

```
Week12Candy <- Week12Candy[-c(1:4), ]
```

2. Summarize the data. What is contained in the dataset? Any “red flags?”

```
# Calculate summary statistics using base R
mean_OBS <- mean(Week12Candy$OBS, na.rm = TRUE)
sd_OBS <- sd(Week12Candy$OBS, na.rm = TRUE)
min_OBS <- min(Week12Candy$OBS, na.rm = TRUE)
max_OBS <- max(Week12Candy$OBS, na.rm = TRUE)
n_OBS <- sum(!is.na(Week12Candy$OBS))

mean_SALES <- mean(Week12Candy$SALES, na.rm = TRUE)
sd_SALES <- sd(Week12Candy$SALES, na.rm = TRUE)
min_SALES <- min(Week12Candy$SALES, na.rm = TRUE)
max_SALES <- max(Week12Candy$SALES, na.rm = TRUE)
n_SALES <- sum(!is.na(Week12Candy$SALES))

mean_AD <- mean(Week12Candy$AD, na.rm = TRUE)
sd_AD <- sd(Week12Candy$AD, na.rm = TRUE)
min_AD <- min(Week12Candy$AD, na.rm = TRUE)
max_AD <- max(Week12Candy$AD, na.rm = TRUE)
n_AD <- sum(!is.na(Week12Candy$AD))

# Combine the results into a data frame for better display
summary_data <- data.frame(
  Statistic = c("Mean", "SD", "Min", "Max", "N"),
  OBS = c(mean_OBS, sd_OBS, min_OBS, max_OBS, n_OBS),
  SALES = c(mean_SALES, sd_SALES, min_SALES, max_SALES, n_SALES),
  AD = c(mean_AD, sd_AD, min_AD, max_AD, n_AD)
)

# Use gt to display the summary table in a nice format
library(gt)
summary_table <- summary_data %>%
  gt() %>%
  tab_header(
      title = "Summary Statistics for OBS, SALES, and AD"
  ) %>%
  fmt_number(
      columns = c(OBS, SALES, AD),
      decimals = 2
  ) %>%
  cols_label(
      Statistic = "Statistic",
      OBS = "OBS",
      SALES = "SALES",
      AD = "AD"
  )
```

```
# Display the table
summary_table
```

Summary Statistics for OBS, SALES, and AD

Statistic	OBS	SALES	AD
Mean	17.00	644.80	63.92
SD	7.36	191.20	11.97
Min	5.00	360.00	40.00
Max	29.00	900.00	80.00
N	25.00	25.00	25.00

No red flags – first 4 rows correctly dropped. Mean, sd, min and max seem correct and within expectations.

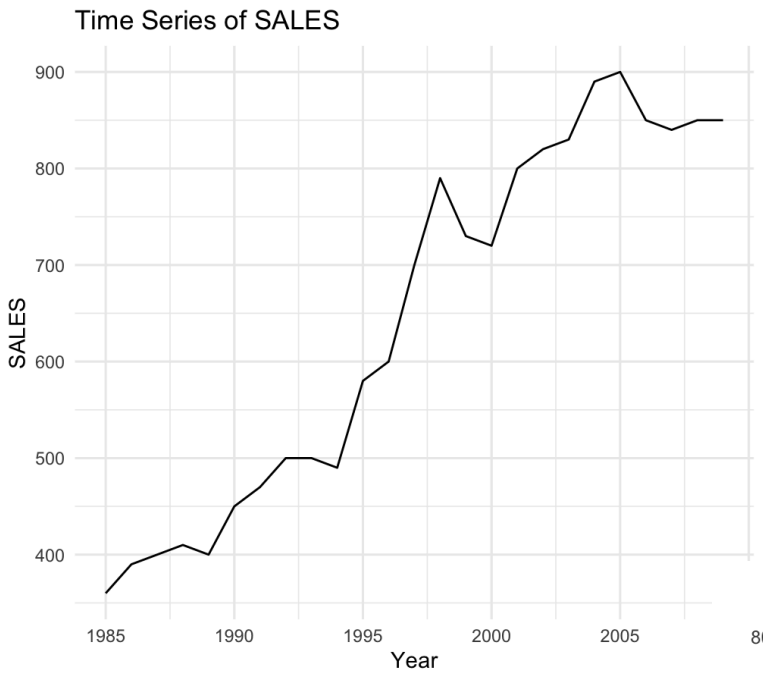
3. Are SALES and AD significantly correlated? Explain what the correlation coefficient tells you in this case. Given your finding, what additional information will a regression model provide?

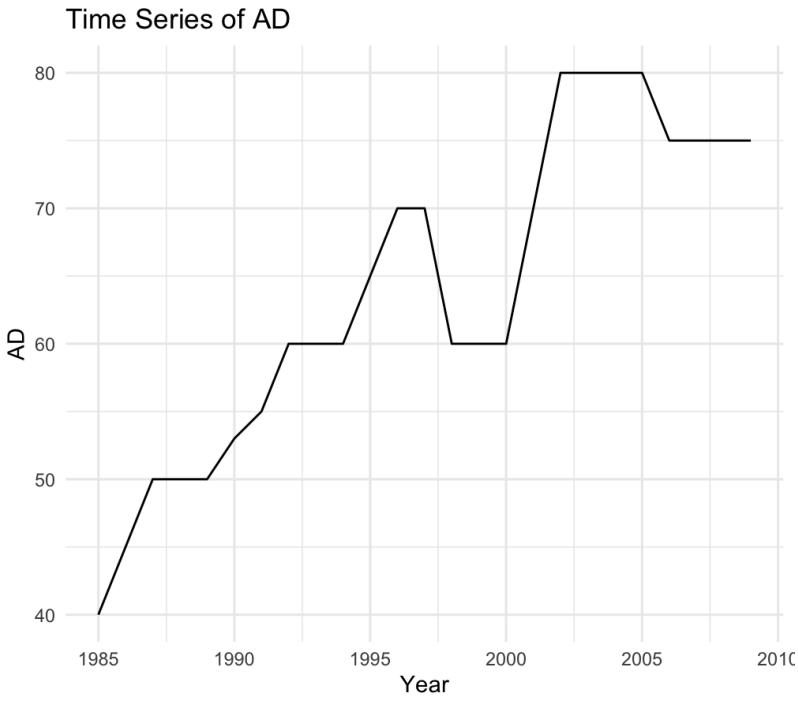
```
# Perform the Pearson correlation test between SALES and AD
correlation_test <- cor.test(Week12Candy$SALES, Week12Candy$AD, method
= "pearson")

# Display the result
correlation_test
Pearson's product-moment correlation

data:  Week12Candy$SALES and Week12Candy$AD
t = 10.192, df = 23, p-value = 5.336e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7933359 0.9575948
sample estimates:
      cor
0.9048288
```

Sales and AD is significantly correlated. Correlation is 0.9. Correlation coefficients tell us that Sales and AD are positively (linear relationship) and strongly correlated which means that they move together. An increase in sales correlates with an increase in AD and vice versa. Regression model may tell us how much an increase in AD correlates with an increase in Sales which we cannot observe with correlation.

4. Create a line graph to show how SALES and AD grew over time.

```
# Load ggplot2 if you haven't already
library(ggplot2)

# Create the time series plot for SALES
ggplot(Week12Candy, aes(x = YEAR, y = SALES)) +
  geom_line() +
  labs(title = "Time Series of SALES", x = "Year", y = "SALES") +
  theme_minimal()


# Load ggplot2 if not already loaded
library(ggplot2)

# Create the time series plot for AD
ggplot(Week12Candy, aes(x = YEAR, y = AD)) +
  geom_line() +
  labs(title = "Time Series of AD", x = "Year", y = "AD") +
  theme_minimal()
```





5. Estimate a simple regression using AD to predict SALES. How well does this regression appear to do? Why can't we stop here?

```
# Fit the linear regression model
model <- lm(SALES ~ AD, data = Week12Candy)

# Summary of the model
summary(model)
# Extract the coefficients and statistics
coefs <- summary(model)$coefficients

# Create a nice-looking table with gt
library(gt)

model_table <- data.frame(
  Variable = rownames(coefs),
  Estimate = coefs[, "Estimate"],
  Std_Error = coefs[, "Std. Error"],
  t_value = coefs[, "t value"],
  p_value = coefs[, "Pr(>|t|)"]
)

model_table %>%
  gt() %>%
  tab_header(
      title = "Simple Regression: SALES ~ AD"
  ) %>%
  fmt_number(
      columns = vars(Estimate, Std_Error, t_value, p_value),
      decimals = 3
  ) %>%
  cols_label(
      Variable = "Variable",
      Estimate = "Estimate",
      Std_Error = "Standard Error",
      t_value = "t-Statistic",
      p_value = "p-value"
  )
```

Residual standard error: 83.16 on 23 degrees of freedom
Multiple R-squared: 0.8187, Adjusted R-squared: 0.8108
F-statistic: 103.9 on 1 and 23 DF, p-value: 5.336e-10

Simple Regression: SALES ~ AD

Variable	Estimate	Standard Error	t-Statistic	p-value
(Intercept)	−278.917	92.147	−3.027	0.006
AD	14.451	1.418	10.192	0.000

Every $1 increase in advertising is associated with a $14.45 increase in sales, on average. The intercept is not meaningful since we do not expect negative sales at 0 ADs. Adjusted R^2 is 81% which is extremely good. 81% of the variability in the sales is explained by ADs. The model is statistically significant overall with a high 103.9 F-stat and less than 0.01 p-value

We cant stop here since this is a time-series data. We have to account for time and validate independence, trend, seasonality.

6. Estimate a *linear trend* model to predict sales. Hint: Linear trend will be a model that uses sales as a dependent variable and an ID variable/observation number as an independent variable to denote the time period. What does this model show? Then run a plot to show the trend over time.

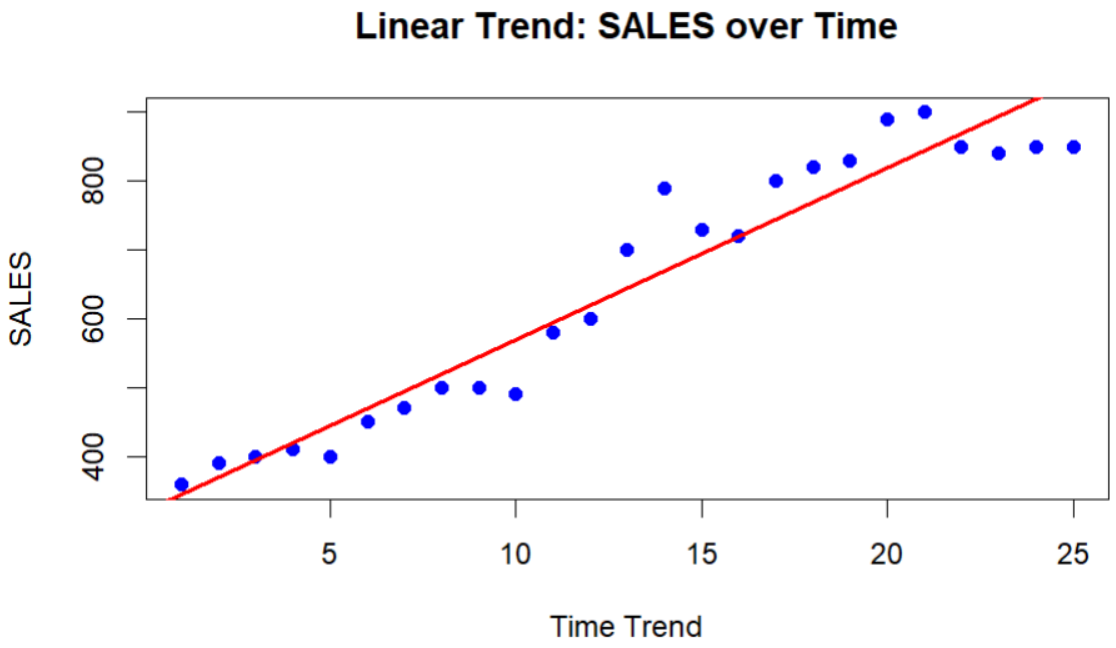
```
# Create the TIMETREND variable as an integer sequence (if not already
present)
Week12Candy$TIMETREND <- 1:nrow(Week12Candy)
# Fit the linear trend model: SALES = beta0 + beta1 * TIMETREND
trend_model <- lm(SALES ~ TIMETREND, data = Week12Candy)

# Display the summary of the model
summary(trend_model)
# Plot the actual SALES values and the fitted trend line
plot(Week12Candy$TIMETREND, Week12Candy$SALES, main = "Linear Trend
Model: SALES vs TIMETREND",
      xlab = "Time Period (TIMETREND)", ylab = "SALES", pch = 19, col
= "blue")
abline(trend_model, col = "red", lwd = 2)  # Add the fitted trend line
```

```
Call:
lm(formula = SALES ~ TIMETREND, data = Week12Candy)

Residuals:
    Min      1Q  Median      3Q     Max
-94.985 -24.708  -9.662  35.169 120.185

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  319.600     21.737   14.70 3.47e-13 ***
TIMETREND     25.015      1.462   17.11 1.41e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.72 on 23 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.924
F-statistic: 292.7 on 1 and 23 DF,  p-value: 1.411e-14
```

The regression output shows a strong linear relationship between time and candy bar sales. The coefficient for the time trend (TIMETREND) is 25.015, which means that, on average, sales increase by approximately 25 units each period. The intercept value is 319.600, suggesting the estimated starting level of sales when the time trend begins. Both of these values are statistically significant, with p-values well below 0.001, indicating that the trend is not due to random chance.

The model explains a large portion of the variability in sales, as shown by the R-squared value of 0.9271 and an adjusted R-squared of 0.924. These values suggest that over 92% of the changes in sales can be attributed to the passage of time. The F-statistic is also very high and statistically significant, reinforcing that the model as a whole fits the data well.

However, while the model clearly identifies a strong upward trend in sales, it does not include advertising or any other explanatory variables. This limits its usefulness for understanding causal relationships—specifically, whether advertising is driving the increase in sales. Thus, while informative about overall trends, this model alone cannot address the core question.



The graph shows a linear trend of candy bar sales over time, with each blue dot representing actual sales for a given time period, and the red line representing the fitted linear regression trend. The upward slope of the red line confirms a consistent increase in sales over time, which aligns with the regression result where the time trend variable (TIMETREND) had a strong positive coefficient.

However, towards the later periods, many points lie below the red line, suggesting that actual sales may have flattened or grown more slowly than the predicted trend. This deviation hints that the linear model may not fully capture recent sales behavior, possibly due to saturation, external effects, or the influence of other factors like advertising. While the trend is strong overall, this visual pattern raises questions about whether a simple linear time trend is sufficient for long-term forecasting.

Shows the secular trend in sales over time but does not answer the question of whether ads play a role.

7. Estimate a distributed lag model. Test for serial correlation. How would you interpret the results, and is serial correlation a major concern? Explain.

```
df$diff1 <- c(rep(NA, 1), diff(df$AD, lag = 1))
df$diff2 <- c(rep(NA, 2), diff(df$AD, lag = 2))
df$diff3 <- c(rep(NA, 3), diff(df$AD, lag = 3))
df$diff4 <- c(rep(NA, 4), diff(df$AD, lag = 4))
# Fit the distributed lag model: SALES = beta0 + beta1 * AD + beta2 *
diff1 + beta3 * diff2 + beta4 * diff3 + beta5 * diff4
distributed_lag_model <- lm(SALES ~ AD + diff1 + diff2 + diff3 +
diff4, data = Week12Candy)

# Display a summary of the model
summary(distributed_lag_model)
# Load the gt package if not already loaded
library(gt)

# Extract coefficients and statistics from the model
coefs <- summary(distributed_lag_model)$coefficients

# Create a dataframe for the table
model_table <- data.frame(
  Variable = rownames(coefs),
  Estimate = coefs[, "Estimate"],
  Std_Error = coefs[, "Std. Error"],
  t_value = coefs[, "t value"],
  p_value = coefs[, "Pr(>|t|)"]
)

# Create a gt table
model_table %>%
  gt() %>%
  tab_header(
      title = "Distributed Lag Model: SALES ~ AD + diff1 + diff2 +
diff3 + diff4"
  ) %>%
  fmt_number(
      columns = vars(Estimate, Std_Error, t_value, p_value),
      decimals = 3
  ) %>%
  cols_label(
      Variable = "Variable",
      Estimate = "Estimate",
      Std_Error = "Standard Error",
      t_value = "t-Statistic",
      p_value = "p-value"
  )

# Install lmtest if not already installed
install.packages("lmtest")

# Load the lmtest package
library(lmtest)

# Perform the Breusch-Godfrey test for serial correlation
bg_test <- bgtest(distributed_lag_model)

# Print the results of the test
```

```
Bg_test
```

Distributed Lag Model: SALES ~ AD + diff1 + diff2 + diff3 + diff4				
Variable	Estimate	Standard Error	t-Statistic	p-value
(Intercept)	−376.850	105.700	−3.565	0.003
AD	16.438	1.565	10.501	0.000
diff1	−1.582	6.086	−0.260	0.798
diff2	−4.382	6.235	−0.703	0.493
diff3	−0.339	5.979	−0.057	0.955
diff4	−4.057	3.696	−1.098	0.290

Breusch-Godfrey test for serial correlation of order up to 1
data: distributed_lag_model
LM test = 9.9641, df = 1, p-value = 0.001596

This model assumes that current sales might be influenced not just by current ad spending, but by past changes in advertising (i.e., momentum or delayed impact). essentially checking for short-term lag effects in ad spending on sales. The model indicates that current advertising (AD) has a significant, positive impact on SALES.

The lagged differences (diff1-diff4) are not statistically significant, suggesting past changes in AD don't additionally affect SALES. Overall, only the current level of advertising appears to drive sales in this specification. I don't have to worry about past ad expenditure. Ad spending now or current is what matters. Immediate use actionable.

The null hypothesis for BG test for serial correlation: No serial correlation in residuals. Since $p < 0.05$, you reject the null → there is serial correlation in residuals. Serial correlation in regression residuals means:

The error terms are not independent. It violates OLS assumptions, making, Standard errors biased → leading to misleading t-stats and p-values. Confidence intervals and hypothesis tests unreliable. May indicate important dynamics are missing from the model (e.g., time trend, omitted lags, or autocorrelated shocks).

8. Estimate a dynamic model.

```
# Install nlme package if it's not already installed
install.packages("nlme")

# Load the nlme package
library(nlme)

# Create lagged variables as you did before
Week12Candy$diff1 <- c(rep(NA, 1), diff(Week12Candy$AD, lag = 1))
Week12Candy$diff2 <- c(rep(NA, 2), diff(Week12Candy$AD, lag = 2))
Week12Candy$diff3 <- c(rep(NA, 3), diff(Week12Candy$AD, lag = 3))
Week12Candy$diff4 <- c(rep(NA, 4), diff(Week12Candy$AD, lag = 4))

# Remove rows with NA values from any of the variables involved in the
model
Week12Candy_clean <- na.omit(Week12Candy[, c("SALES", "AD", "diff1",
"diff2", "diff3", "diff4")])

# Fit the Prais-Winsten regression with AR(1) autocorrelation
structure
prais_model <- gls(SALES ~ AD + diff1 + diff2 + diff3 + diff4,
                   data = Week12Candy_clean,
                   correlation = corAR1())  # AR(1) correlation
structure

# Display the results
summary(prais_model)
```

	Value <chr>	Std.Error <chr>	t-value <chr>	p-value <chr>
(Intercept)	15.033479	430865.7	0.0000349	1.0000
AD	9.821099	5.0	1.9835907	0.0659
diff1	-2.047635	2.7	-0.7447439	0.4679
diff2	-3.244258	2.8	-1.1782776	0.2570
diff3	-1.533678	2.7	-0.5668436	0.5792
diff4	-0.007189	2.6	-0.0028066	0.9978

The (Intercept) is not significant, implying no strong baseline once the other variables are considered.

AD is borderline significant ($p \approx 0.066$), suggesting a modest effect of current advertising on sales.

The diff terms (diff1-diff4) aren't significant, indicating that past changes in AD add little predictive power.

9. Estimate a dynamic model predicting SALES using AD and L1_SALES (which is a lag for sales).

```
Week12Candy$L1_SALES <- c(NA, head(Week12Candy$SALES, -1))
Week12Candy_dyn <- na.omit(Week12Candy[, c("SALES", "L1_SALES", "AD",
"diff1", "diff2", "diff3", "diff4")])
dyn_model_ols <- lm(SALES ~ L1_SALES + AD + diff1 + diff2 + diff3 +
diff4,
                    data = Week12Candy_dyn)

summary(dyn_model_ols)
library(gt)
library(broom)

tidy(dyn_model_ols) %>%
  gt() %>%
  tab_header(
      title = "Dynamic Model Estimation Results"
  )
```

Dynamic Model Estimation Results				
term	estimate	std.error	statistic	p.value
(Intercept)	-67.7507939	102.6427541	-0.6600641	0.5199252084
L1_SALES	0.7438577	0.1752287	4.2450689	0.0008159113
AD	3.9199659	3.1373362	1.2494568	0.2319841130
diff1	-0.4708767	4.1735346	-0.1128244	0.9117715836
diff2	-1.5433676	4.3194363	-0.3573076	0.7261903226
diff3	1.7134485	4.1205067	0.4158344	0.6838349551
diff4	-0.9352358	2.6346071	-0.3549811	0.7278945242

L1_SALES is highly significant, suggesting that previous sales strongly predict current sales.

AD and the advertising terms aren't significant, indicating little additional explanatory power once you include lagged sales.

The intercept is not significant, meaning there's no strong baseline level of sales after controlling for L1_SALES and advertising.

10. What did the line graph in step 4 tell us about potential nonstationarity (i.e., mean is not constant) of the SALES variable? Perform a Dickey-Fuller test to see formally if SALES exhibits nonstationarity.

```
install.packages("tseries")  # only if not yet installed

library(tseries)

adf.test(Week12Candy$SALES)
```

```
        [1] "tseries"

                Augmented Dickey-Fuller Test

        data:  Week12Candy$SALES
        Dickey-Fuller = -0.90838, Lag order = 2, p-value = 0.9341
        alternative hypothesis: stationary
```

The ADF test's high p-value (~0.9341) means we fail to reject the null hypothesis of a unit root.

Hence, the SALES series appears nonstationary, consistent with the null of nonstationarity.

è This is not significant so we CANNOT reject H_o of nonstationarity.

11. What would you conclude about which model makes the most sense, and why? How well do any of the models do in allowing you to understand the role of advertising in candy bar sales?

The dynamic model is best because it takes into account that past sales heavily influence current sales. While the simpler models show that advertising has an immediate positive effect, they miss the ongoing momentum in sales that the dynamic model captures.

The dynamic model makes the most sense because it includes lagged sales (L1_SALES), which strongly predicts current sales. This reflects real-world patterns where sales depend on past performance. While simpler models show that advertising has a strong short-term effect, they miss how sales carry over from one period to the next. Overall, advertising does have an effect, but it's not enough on its own momentum from past sales matters more.

The dynamic model explicitly includes the past value of SALES, which is how nonstationarity often manifests (through persistence or trends). This structure soaks up much of the nonstationary behaviour, reducing its harmful effects on the rest of the model.

The simple regression ignores nonstationarity completely. So, the strong relationship it finds between AD and SALES might be spurious.

The distributed lag model also fails to address the underlying trend or persistence in SALES, and still suffers from serial correlation, making the nonstationarity even more damaging.