Nishtha Patel

Code

#Introduction

To begin the analysis, I:

Imported the dataset into R using read.csv().

Inspected the data structure with functions like str(), summary(), and glimpse().

Cleaned column names using janitor::clean_names().

Parsed date columns (Order Date, Ship Date) using lubridate::mdy().

Checked for missing values, duplicates, and outliers.

Created summary statistics and visualizations for key variables such as Sales, Profit, Discount, and Category.

Hide

```
library(tidyverse)
library(janitor)
library(lubridate)
library(ggplot2)
library(summarytools)

superstore <- read.csv("~/Aly- 6040 DataMining Adam/superstore.csv")

colnames(superstore)

# Explore the structure
str(superstore)

# Summary statistics
summary(superstore)
nrow(superstore)

# Check for missing values
colSums(is.na(superstore))
#Check for duplicate row
sum(duplicated(superstore))
superstore <- superstore %>% distinct()
```

What I did in the context of exploration?

In the data exploration phase, I first cleaned column names using janitor. I inspected the structure of the dataset using ‘str()’ and ‘glimpse()’. Then, I converted ‘Order.Date’ and ‘Ship.Date’ to Date format using lubridate. I checked for missing values, unique categorical values, and computed summary statistics. Finally, I created visualizations to understand sales and profit distributions, as well as category-wise and region-wise performance.

How many entries are in the dataset?

The dataset contains 9,994 entries.

Was there missing data? Duplications? How clean was the data?

There were no missing values. However, I found 17 duplicate rows, which I removed to ensure accuracy.

Hide

```
#Finding the outliers
superstore[superstore$Sales > quantile(superstore$Sales, 0.99), ]
superstore[superstore$Profit < quantile(superstore$Profit, 0.01), ]
```

Were there outliers or suspicious data?

Yes, there were clear outliers in the Sales and Profit columns. I noticed significant outliers in both the Sales and Profit columns. For example, Sales values exceed $22,000 and some Profits are negative and extremely low. These may be legitimate large orders or errors.

Sales had extreme values, ranging from $0.44 to $22,638, which is unusually high for single transactions.

Profit ranged from -6599.98 to 8399.98, showing heavy losses and profits, which is worth deeper business analysis.

A few orders had high discounts (up to 80%) but still had negative profits, suggesting potential issues in pricing or strategy.

Hide

```
# Unique values per categorical column
sapply(superstore[, sapply(superstore, is.character)], unique)

superstore$order_date <- mdy(superstore$Order.Date)
superstore$ship_date <- mdy(superstore$ship.Date)

boxplot(superstore$Sales, main = "Boxplot of Sales", col = "lightblue")
boxplot(superstore$Profit, main = "Boxplot of Profit", col = "lightgreen")
```

To better understand the structure of the dataset, I inspected all categorical columns to view their unique values. This helped verify the consistency and correctness of categorical data such as region, segment, and ship mode.

I converted the ‘Order Date’ and ‘Ship Date’ fields from character format to proper date objects. This was necessary for any time-based analysis, such as exploring trends over time or calculating delivery times.

1) Boxplot for Sales

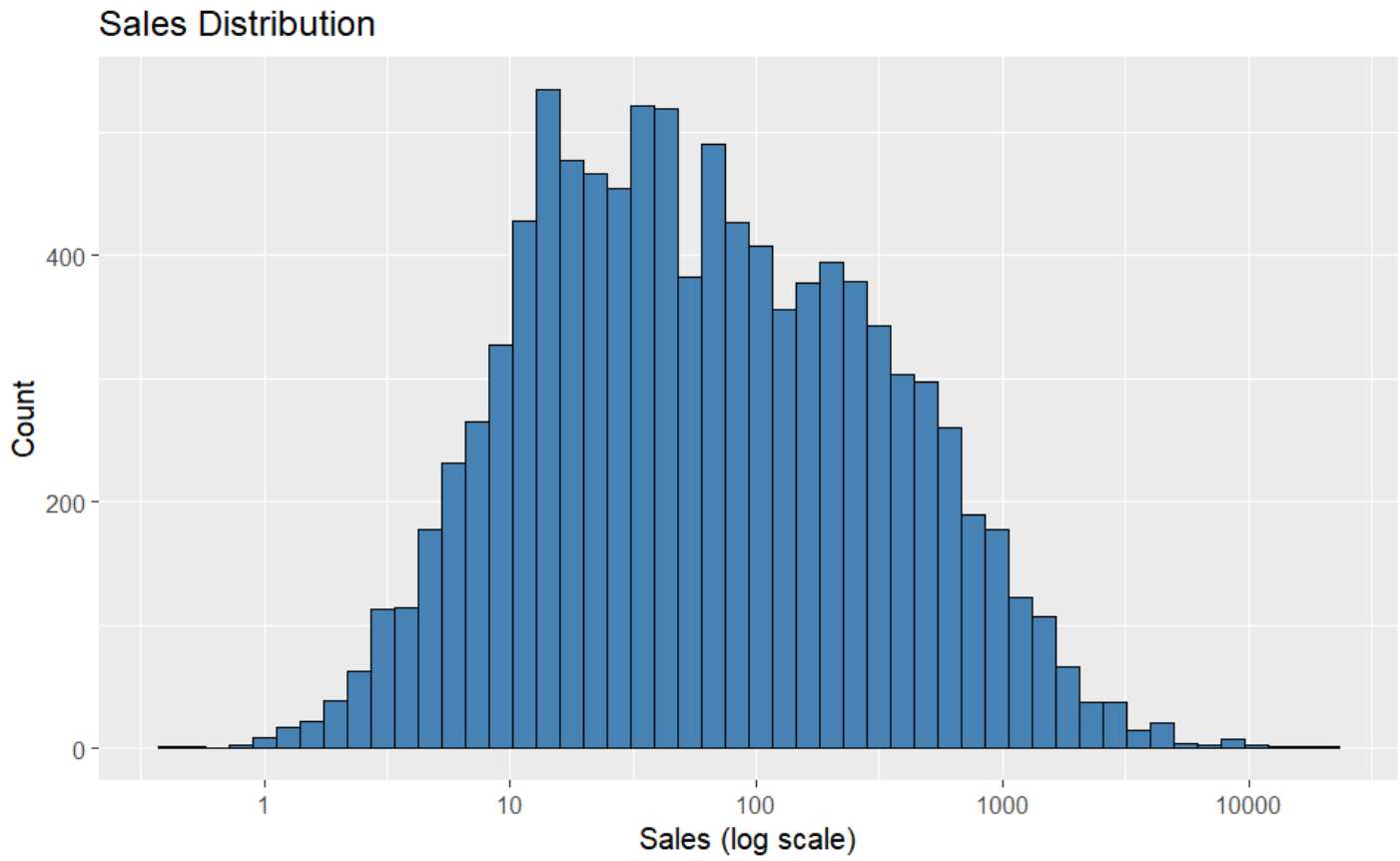
This boxplot shows the distribution of sales values in the dataset. The majority of the sales are concentrated near the lower end, while a few extremely high sales values appear as outliers (points outside the whiskers). This indicates that while most orders are small to moderate, some unusually large sales occur, which could significantly affect the average.

2) Boxplot for Profit

The profit boxplot reveals a similar pattern. Most profits are low or moderate, but there are both high positive and large negative outliers, meaning some products or orders result in heavy losses or large profits. The negative outliers are especially important, as they may indicate problem areas in the business, such as over-discounting or inefficient shipping.

Hide

```
# Visualization 1: Sales distribution
ggplot(superstore, aes(x = Sales)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "black") +
  scale_x_log10() +
  labs(title = "Sales Distribution", x = "Sales (log scale)", y = "Count")
```



This histogram illustrates the distribution of sales across all transactions in the Superstore dataset. Due to the large variation in sales amounts—from small orders to very large ones—the x-axis is transformed using a logarithmic scale. This helps to visualize the spread more clearly and reduce the skewness caused by extreme values.

Interpretation:

The histogram shows a right-skewed distribution (positively skewed), indicating that most sales are on the lower end (between $10 and $100).

A log transformation was applied to the sales values to compress the wide range and highlight the density of smaller sales.

The majority of sales fall between $10 and $1,000.

There are a few high-value outliers beyond $1,000, but they occur far less frequently.

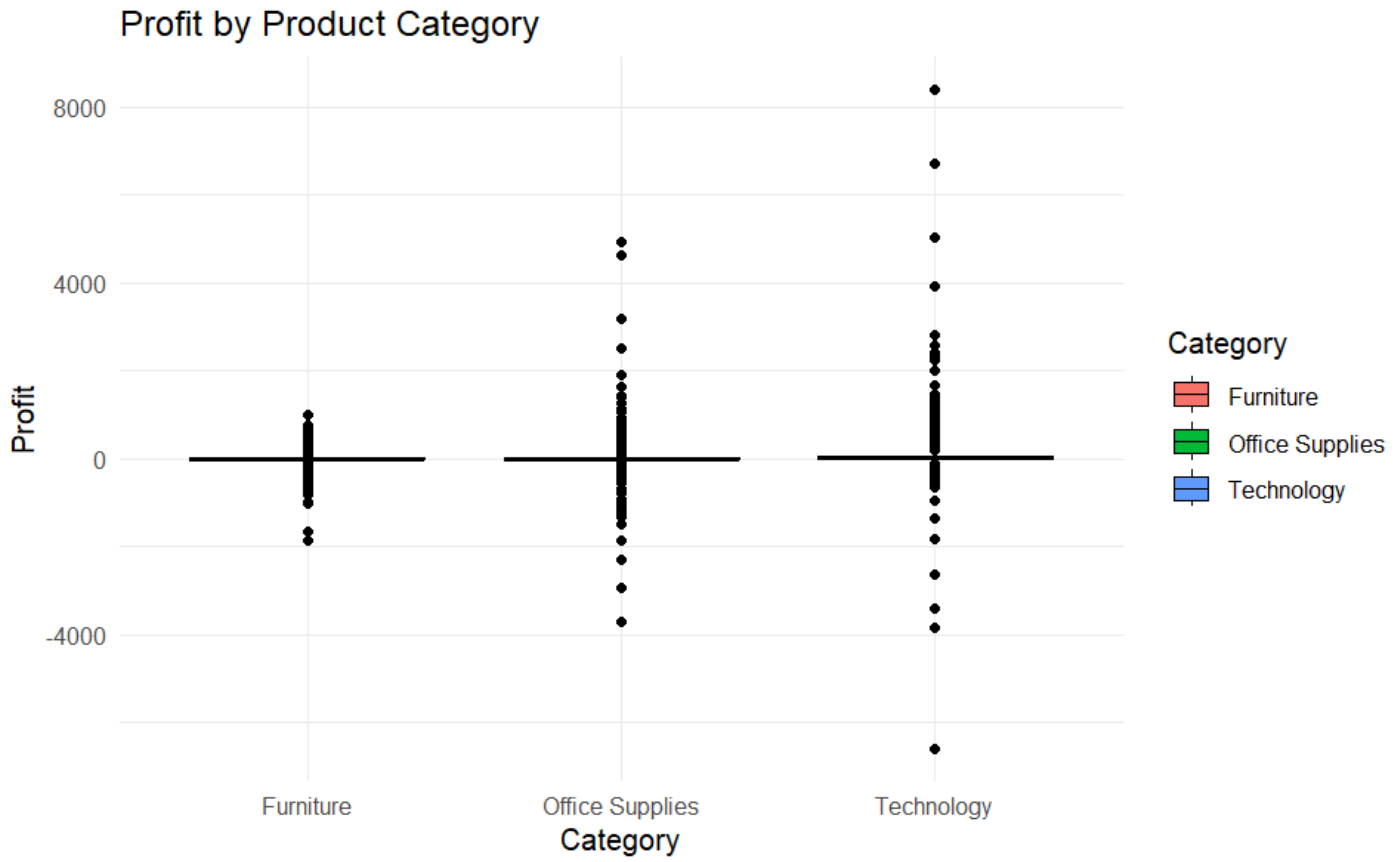
The shape of the histogram suggests that the business processes a large volume of small to mid-range sales, while high-ticket sales are rare.

Why it matters:

Understanding this distribution helps in pricing strategy, inventory planning, and identifying target customer segments. The long tail of high-value transactions may represent bulk or premium purchases, which could be explored further for strategic marketing.

Hide

```
ggplot(superstore, aes(x = Category, y = Profit, fill = Category)) +
  geom_boxplot(color = "black") +  # add border color for contrast
  labs(title = "Profit by Product Category", x = "Category", y = "Profit") +
  theme_minimal() +  # use a cleaner theme
  scale_fill_manual(values = c("Furniture" = "#F8766D",
                               "Office Supplies" = "#00BA38",
                               "Technology" = "#619CFF"))  # custom colors
```



This boxplot visualizes the distribution of profits for each product category—Furniture, Office Supplies, and Technology—within the Superstore dataset. The chart helps compare the variability, central tendency, and outliers in profit margins across these categories.

Interpretation:

Technology generally shows the highest variability in profit. It has many high-profit outliers, indicating that this category includes some very profitable transactions.

Furniture and Office Supplies have more compact distributions of profit, though they also include both positive and negative profit values, reflecting some losses.

All categories show outliers, both on the positive and negative side. This indicates that while most transactions yield moderate profits, some result in significant gains or losses.

The median profit appears relatively similar across categories but is slightly higher for Technology.

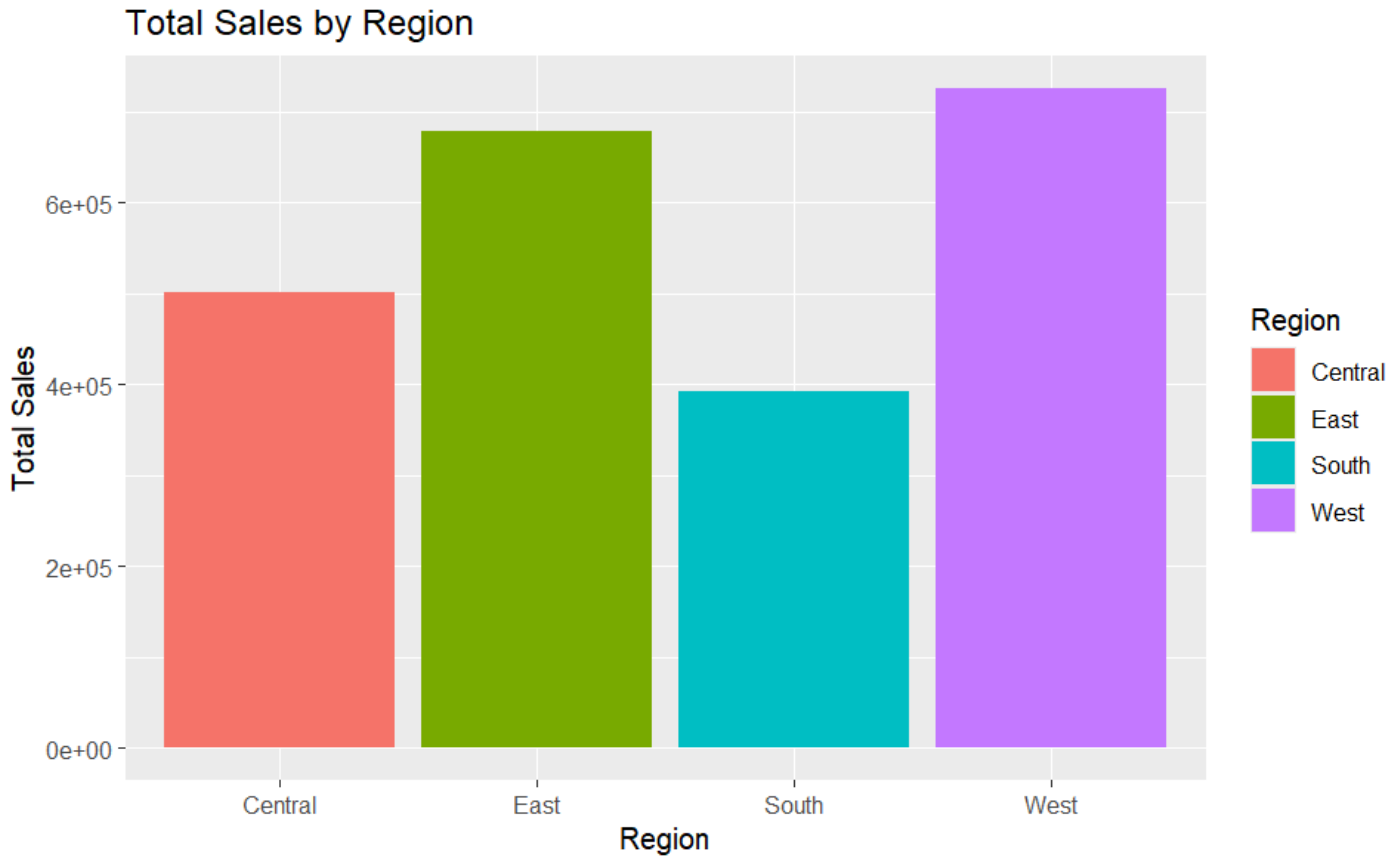
Some transactions in each category resulted in losses, as shown by the presence of points below zero.

Why it matters:

Understanding which product categories are more profitable or riskier helps in decision-making related to marketing, inventory management, and pricing strategies. For example, the wide profit spread in Technology may suggest high reward but also potential risk, making it crucial to monitor sales performance closely.

Hide

```
# Visualization 3: Sales by Region
ggplot(superstore, aes(x = Region, y = Sales, fill = Region)) +
  geom_bar(stat = "summary", fun = sum) +
  labs(title = "Total Sales by Region", x = "Region", y = "Total Sales")
```



The bar chart titled “Total Sales by Region” visualizes the total sales performance across four regions—Central, East, South, and West—based on the data from a business dataset. Each bar represents the cumulative sales for one region, with distinct colors enhancing regional comparison.

Interpretation:

The West region has the highest total sales, slightly surpassing the East.

The East region also performs strongly, with total sales close to the West.

The Central region shows moderate sales, falling behind both East and West.

The South region has the lowest total sales, indicating potential underperformance or smaller market size.

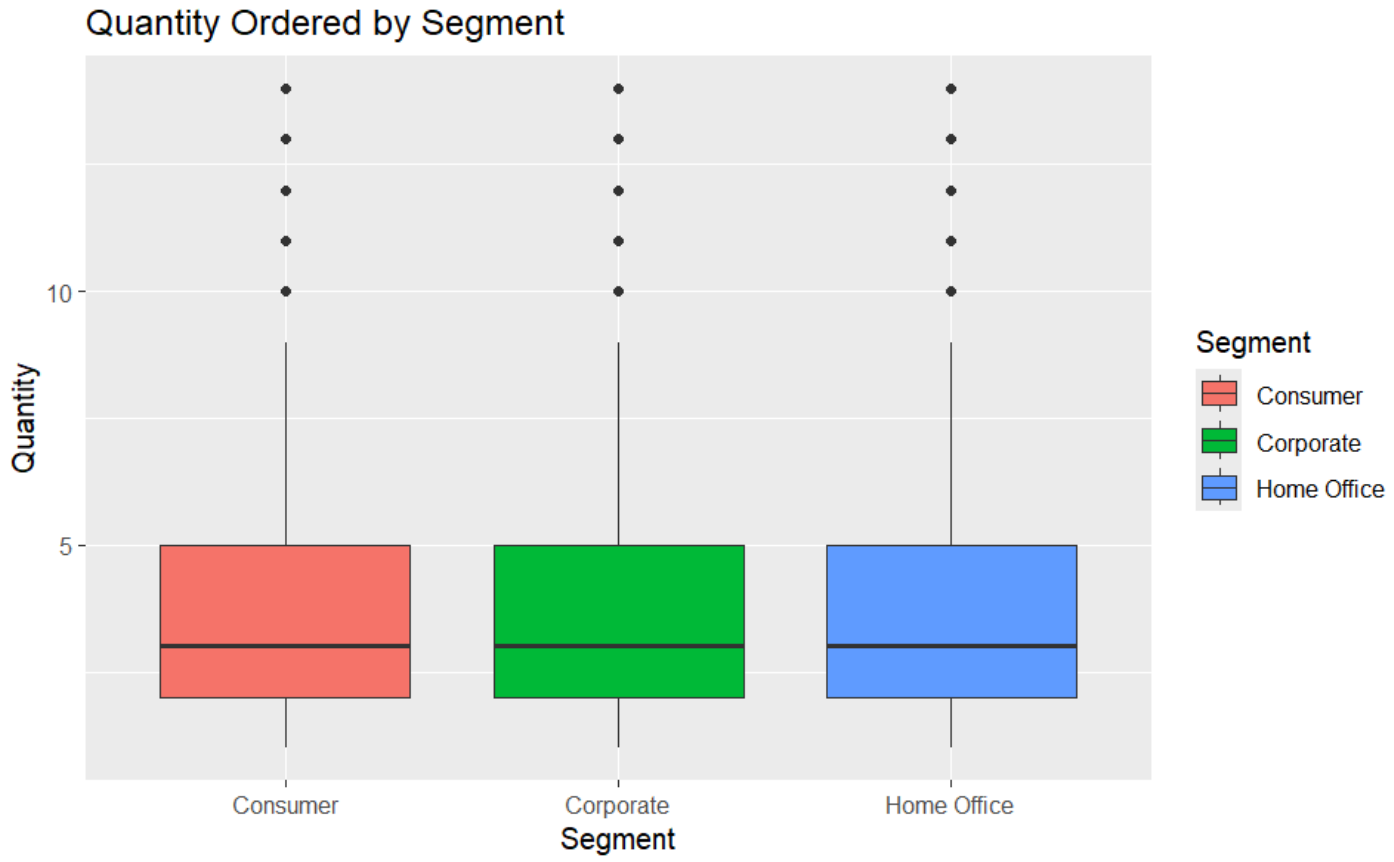
This suggests that business strategies may be more effective in the West and East, while South may require further analysis or improvement initiatives.

Why it matters

Understanding total sales by region is crucial for strategic decision-making, as it highlights high-performing areas like the West and East, which can guide investment and expansion. It also identifies underperforming regions like the South, where marketing and resource allocation may need adjustment. This data-driven approach supports decisions on budget distribution, staffing, and promotions. By visualizing these disparities, businesses can improve performance and profitability more effectively.

Hide

```
# Visualization 4: Quantity by Segment
ggplot(superstore, aes(x = Segment, y = Quantity, fill = Segment)) +
  geom_boxplot() +
  labs(title = "Quantity Ordered by Segment", x = "Segment", y = "Quantity")
```



This boxplot visualization shows the quantity of orders across three customer segments: Consumer, Corporate, and Home Office. Each box represents the distribution of order quantities for its respective segment, with the median, interquartile range (IQR), and outliers clearly displayed.

Interpretation

The median quantity ordered is similar across all segments, hovering around 5 units.

The interquartile range (IQR), which represents the middle 50% of data, is consistent across segments, indicating comparable variability in order quantities.

Outliers are present in all three segments, suggesting occasional instances of unusually high quantities ordered.

There are no significant differences in the overall distribution of order quantities between the segments.

Why It Matters

Understanding order quantities by segment helps businesses identify purchasing patterns and tailor their strategies. For example:The (example:The) similarity across segments suggests uniform demand behavior, which could simplify inventory management and marketing efforts.Identifying outliers may help investigate specific cases (e.g., bulk orders) that could indicate opportunities for targeted promotions or operational adjustments.

What did you find? What intrigued you about the data? Why does that matter?

I found that the majority of sales are small amounts, but a few very large transactions skew the average. High discounts don't always mean high profit. The data hints at interesting trends in sales strategies and product profitability. This matters because it can guide decisions on pricing and inventory.

What would your proposed next steps be? How do you plan to approach the cleansing of the data?

Data Cleansing Plan:

Remove duplicated rows.

Standardize categorical values (e.g., "Standard Class" vs "standard class").

Flag and analyze outliers using IQR and Z-score methods.

Normalize date formats and extract year/month for time series trends.

Remove unnecessary columns like Row.ID for analysis.

Advanced Steps:

Conduct time series analysis on sales trends. Analyze customer segmentation based on buying behavior. Use predictive modeling to forecast profit or identify at-risk transactions. This exercise taught me how to explore a dataset systematically: loading, cleaning, identifying issues, and summarizing key insights. Each step — especially visualization — helped in understanding data quality and patterns. These skills are fundamental to analytics workflows.

#Final Conclusion:

In this analysis of the Superstore dataset, I explored and cleaned the data to ensure its accuracy and consistency. Key steps included inspecting the data structure, handling missing values, removing duplicates, and identifying outliers in the sales and profit columns. Through visualizations, I revealed insights into sales distribution, profit variability by category, regional sales performance, and order quantities across customer segments.

The findings show that while most sales are relatively low, a few high-value transactions significantly impact the dataset's overall performance. Profitability varies greatly by category, with the Technology category showing higher variability in profit. Regional performance highlights the West and East regions as the strongest performers, while the South may require attention. Additionally, customer segment analysis suggests relatively consistent purchasing behavior across Consumer, Corporate, and Home Office segments.

Moving forward, I propose a data cleansing strategy to standardize values, flag and analyze outliers, and perform time series analysis to uncover trends. Advanced steps could include customer segmentation analysis and predictive modeling to forecast future performance.

This analysis provided valuable insights into the dataset’s structure, patterns, and anomalies. By refining these findings, businesses can make more informed decisions about sales strategies, inventory management, and customer targeting.