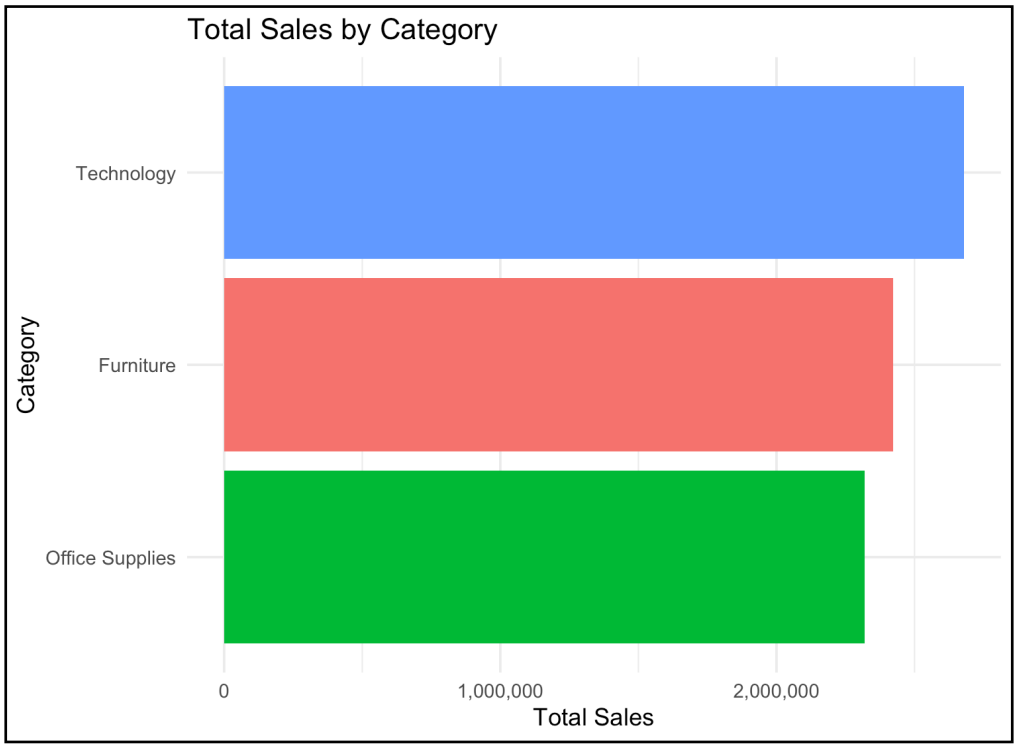
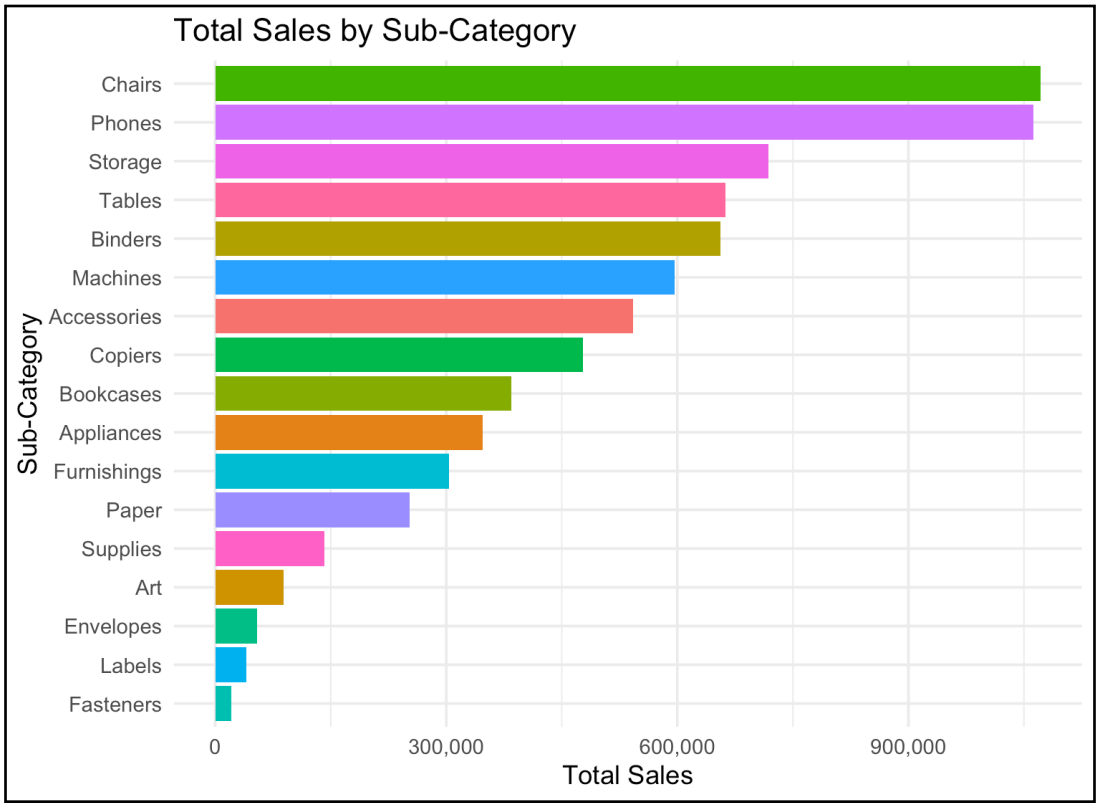
Prediction and Forecasting for Business Success

Authored by SHIH-HSUAN, YAN

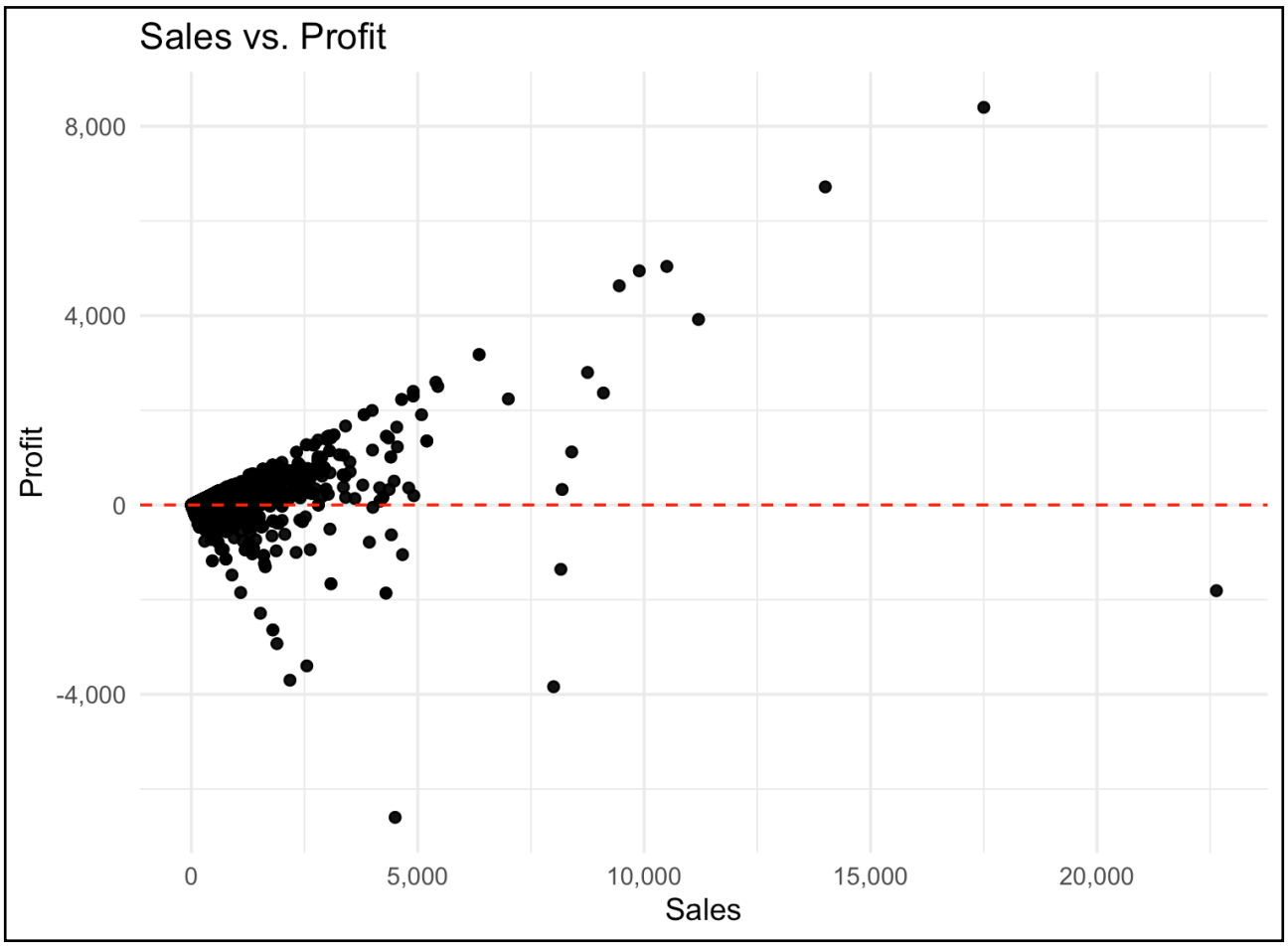
In this report, I will guide you through the dataset from a superstore which includes data from 2014 to 2024. There are a lot of columns in this data. However, given that I will bring you some interesting insights and also tell you about how the business can be successful, the main variables I will focus on are sales, profits, and discounts.Before I guide you to the main insight, let us take a look at how much sales come from different categories.



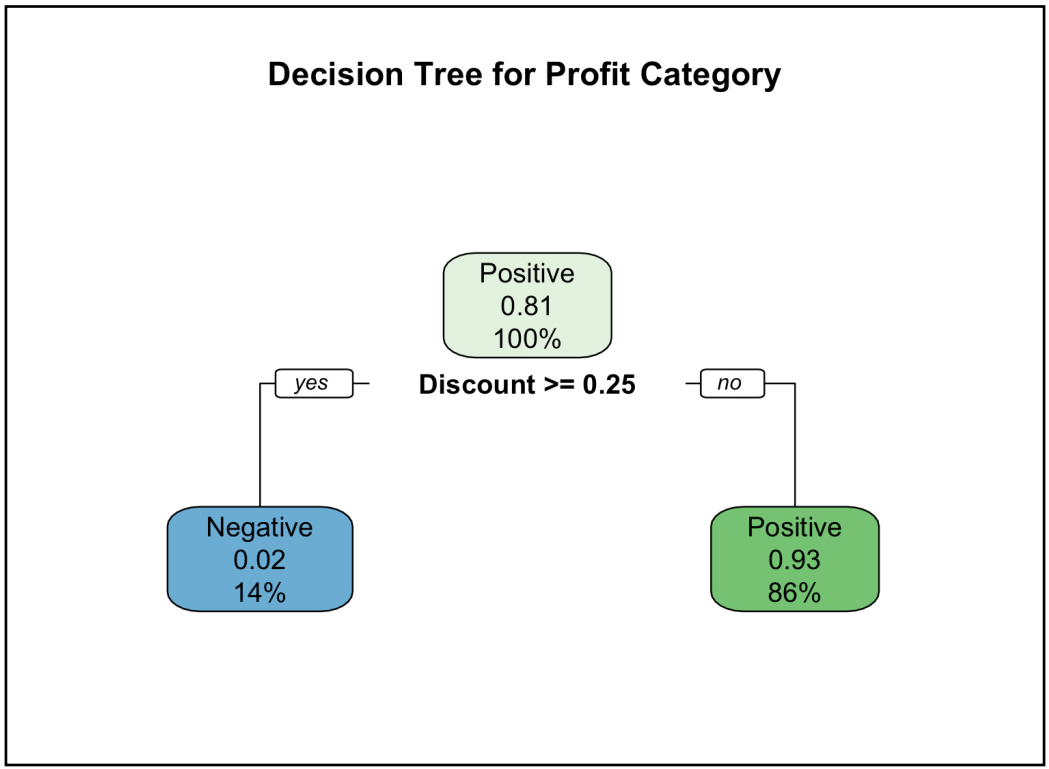
You can see that the category that makes the most sales is technology, which is pretty important in the 21st century. After technology, the sales are followed by furniture as the second category, and office supplies as the least when comparing to the previous two.These categories provide a comprehensive view of the product range and customer preferences, which is crucial for strategic planning.



Now, let us look a bit deeper from category to sub-category. You can see that although technology has the highest sales, the highest sales for sub-category is chairs, which is in the category of furniture. However, does higher sales indicate business success? This question is essential as it highlights the need to analyze profitability, not just revenue, to ensure sustainable growth.



Shown in this graph, just because a product has higher sales does not mean that it provides higher profits. Therefore, I will show you what factor affects the profit, which might bring a massive business insight. Understanding these factors can help the business make informed decisions about pricing and discount strategies.

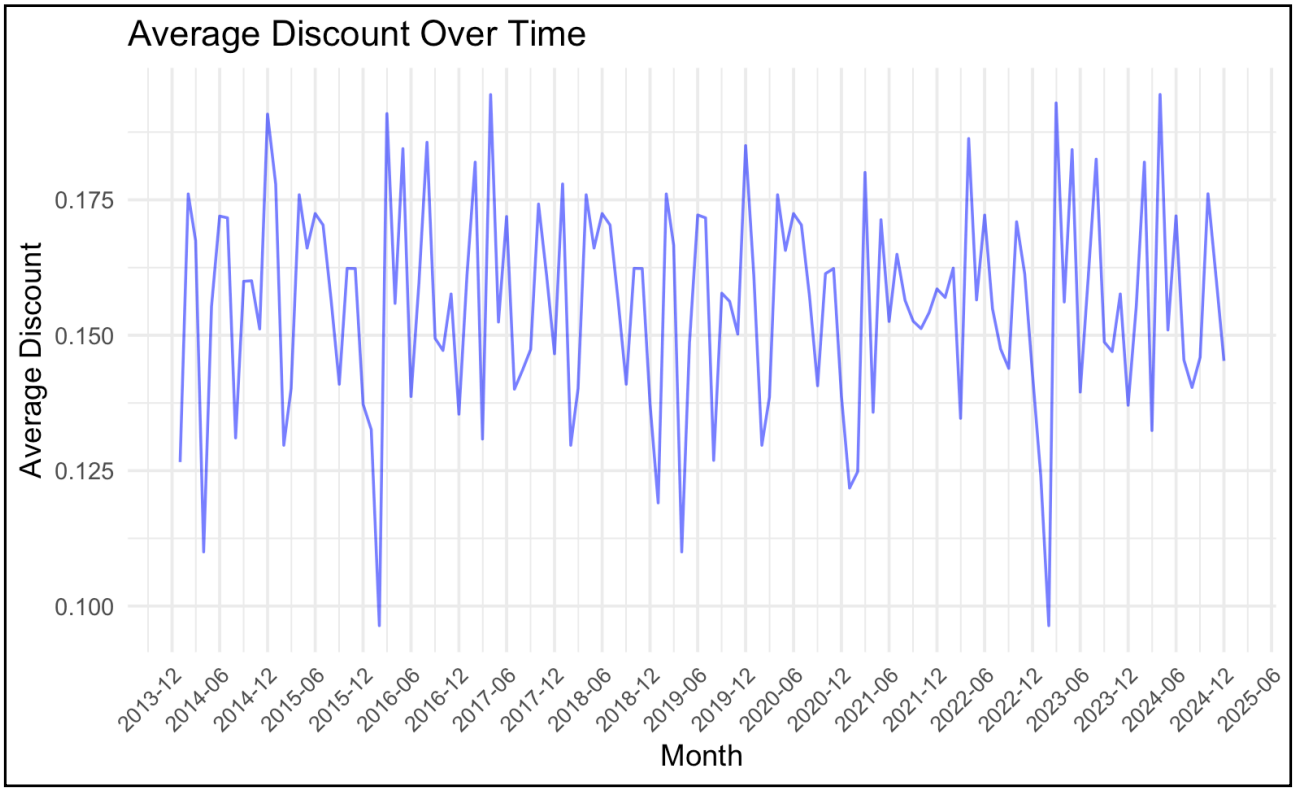


In this decision tree plot, you can see that even though I used multiple columns to predict whether the profit is positive or negative, the only factor that matters is if the discount is greater than 25 percent. This means that if a product's discount is greater than 25 percent, it has a very high chance of having a negative profit, which should be avoided. This insight can guide the business in setting optimal discount rates to maximize profit margins.

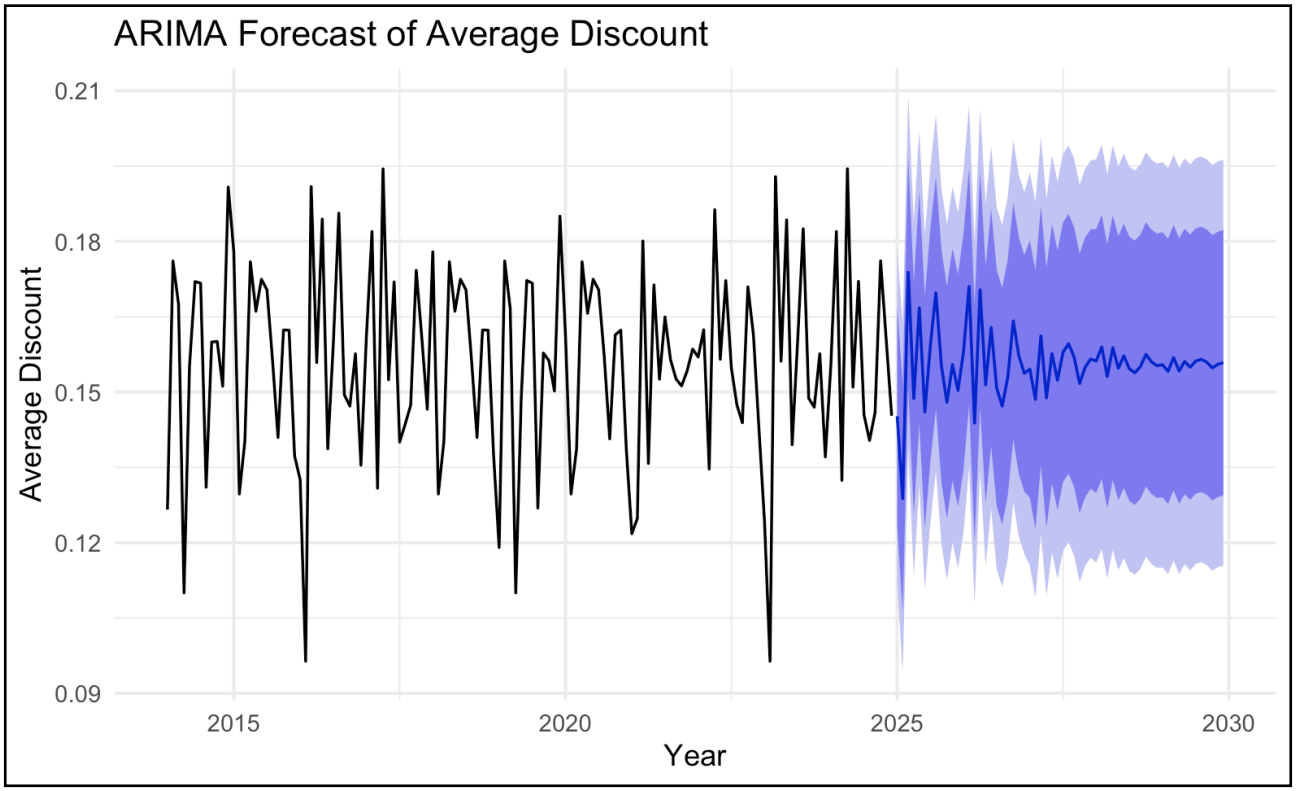
```
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative      869       19
  Positive      385     5223

               Accuracy : 0.9378
                 95% CI : (0.9317, 0.9436)
    No Information Rate : 0.807
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7755
```

```
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6930
            Specificity : 0.9964
         Pos Pred Value : 0.9786
         Neg Pred Value : 0.9313
             Prevalence : 0.1930
         Detection Rate : 0.1338
   Detection Prevalence : 0.1367
      Balanced Accuracy : 0.8447

       'Positive' Class : Negative
```

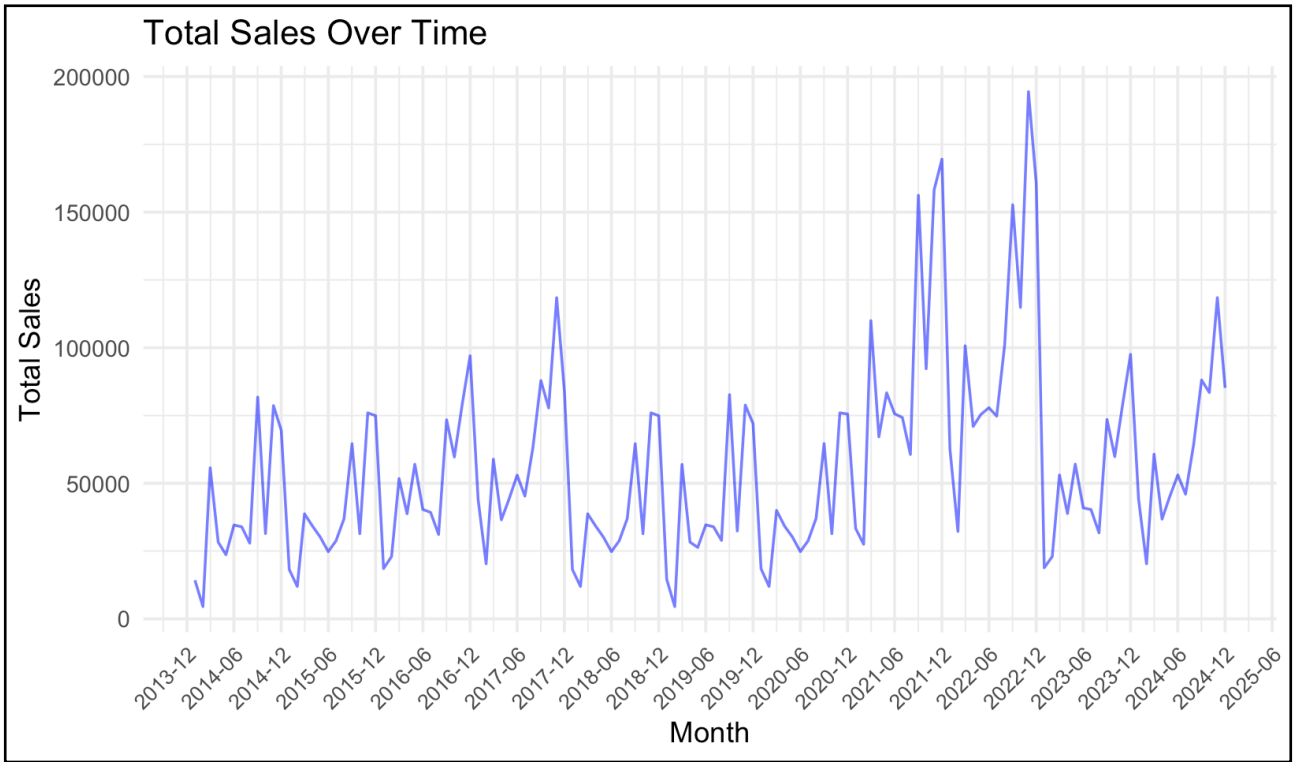
To prove that this analysis is reliable, the confusion matrix shows that the model has a 93.78% accuracy, which is pretty high. The model's sensitivity is 69.30%, meaning it correctly identifies 69.30% of the negative profit cases. The specificity is very high at 99.64%, indicating it almost always correctly identifies positive profit cases. The positive predictive value is 97.86%, meaning when the model predicts a negative profit, it is correct 97.86% of the time. The negative predictive value is 93.13%, meaning when the model predicts a positive profit, it is correct 93.13% of the time. The balanced accuracy, which is the average of sensitivity and specificity, is 84.47%. This shows the model is reliable in distinguishing between positive and negative profits. The high specificity and positive predictive value indicate that the model is particularly effective in identifying profitable products, which is crucial for strategic decision-making.



This graph shows the average discount over time by month from 2014 to 2024. It shows that the average discount varies from month to month, indicating some degree of fluctuation in discount rates offered by the superstore. However, this graph seems it is stationary which can apply for ARIMA to forecast the average discount for the next 5 years. The ARIMA model, which stands for AutoRegressive Integrated Moving Average, requires stationarity to make accurate forecasts. This characteristic makes ARIMA a powerful tool for predicting stable trends in time series data.

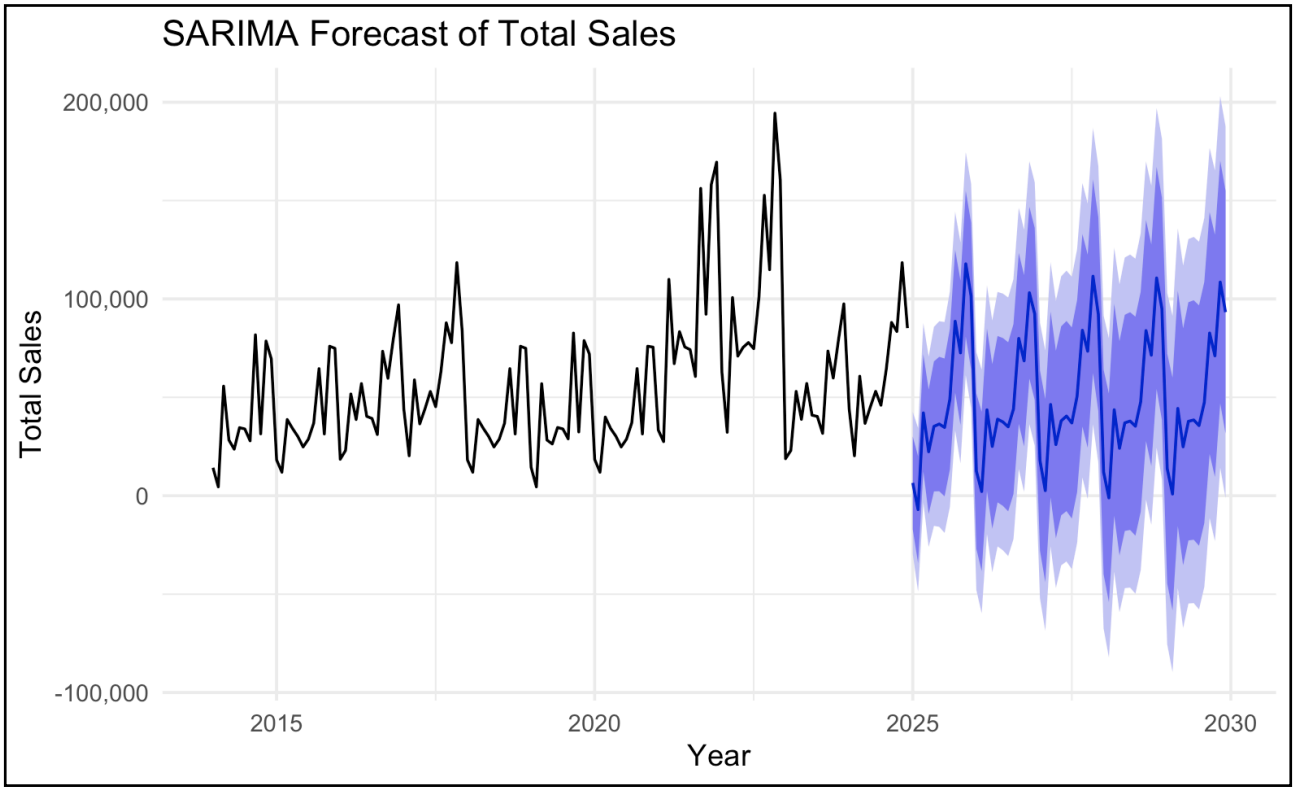


After forecasting the average discount, we observe that the forecast predicts a relatively stable trend in the average discount rates for the next 5 years, as shown by the blue shaded area representing the confidence intervals. This forecast can help the business plan its pricing and discount strategies more effectively, ensuring that discount rates are set at optimal levels to maintain profitability without overly reducing prices. By understanding the expected trends in discount rates, the business can better align its promotional activities and inventory management to the anticipated market conditions. This proactive approach ensures that the business remains competitive and responsive to market dynamics.



This graph shows the total sales over time by month from 2014 to 2024. We can see that total sales show a clear seasonal pattern, with peaks and troughs repeating each year. However, there is an anomaly from 2021 to 2022 where the pattern looks different, with increased volatility likely due to the impacts of COVID-19. Despite this period of greater volatility, the overall yearly pattern remains evident.

Given this clear seasonality, I chose to used the SARIMA (Seasonal AutoRegressive Integrated Moving Average) model to forecast total sales for the next 5 years. SARIMA is well-suited for time series data with seasonal patterns, making it ideal for this sales data that shows repeating annual trends. Analyzing these seasonal patterns helps in understanding consumer behavior and market trends over different periods.



After applying the SARIMA model, the forecast shows a continuation of the seasonal pattern, with expected peaks and troughs over the next 5 years. The blue shaded area represents the confidence intervals, indicating the range within which future sales are likely to fall. This forecast can help the business plan for seasonal fluctuations in sales, allowing for better inventory management, staffing, and promotional activities. By anticipating periods of high and low sales, the business can optimize its operations and marketing efforts to maximize revenue and efficiency. This foresight enables the business to allocate resources more effectively and enhance overall performance.

In conclusion, this analysis provides valuable insights into the sales, profit, and discount patterns of the superstore from 2014 to 2024. By examining the data, we identified key factors affecting profitability, such as the impact of high discounts on profit margins. The decision tree model highlighted the significance of maintaining discounts below 25 percent to avoid negative profits. Furthermore, the ARIMA and SARIMA models allowed us to forecast future trends in average discounts and total sales, respectively. These forecasts enable the business to plan strategically for the next 5 years, optimizing pricing strategies, inventory management, and promotional activities. Overall, this comprehensive analysis equips the business with the knowledge to make informed decisions, enhance profitability, and achieve sustainable growth.

Reference

Patel, M. (2024). Superstore Dataset 2014-2024. Kaggle. https://www.kaggle.com/datasets/mananapatel99/superstore-dataset-2014-2024

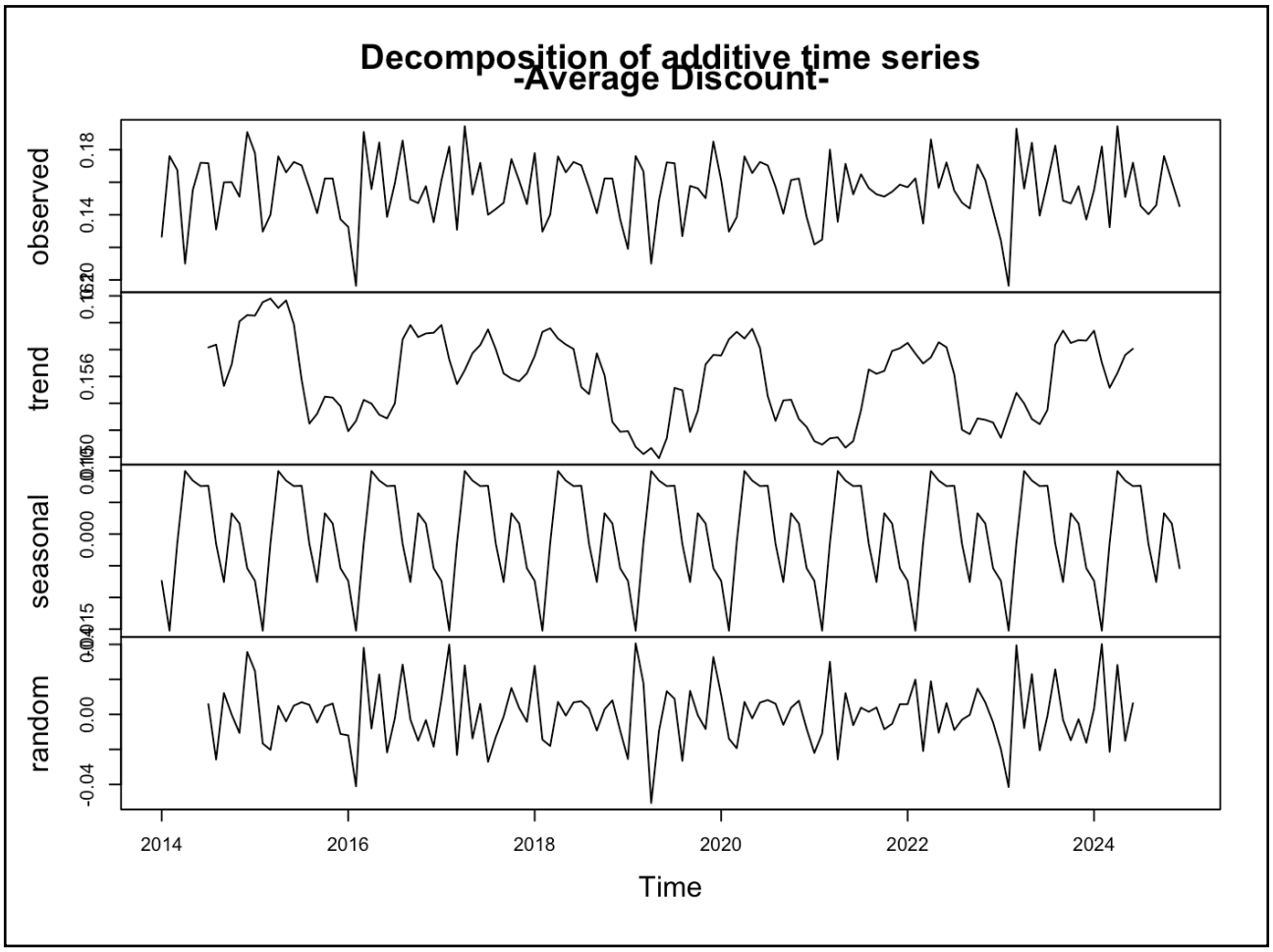
OpenAI. (2024). ChatGPT-4 (May 2024 version) [Large language model]. https://www.openai.com

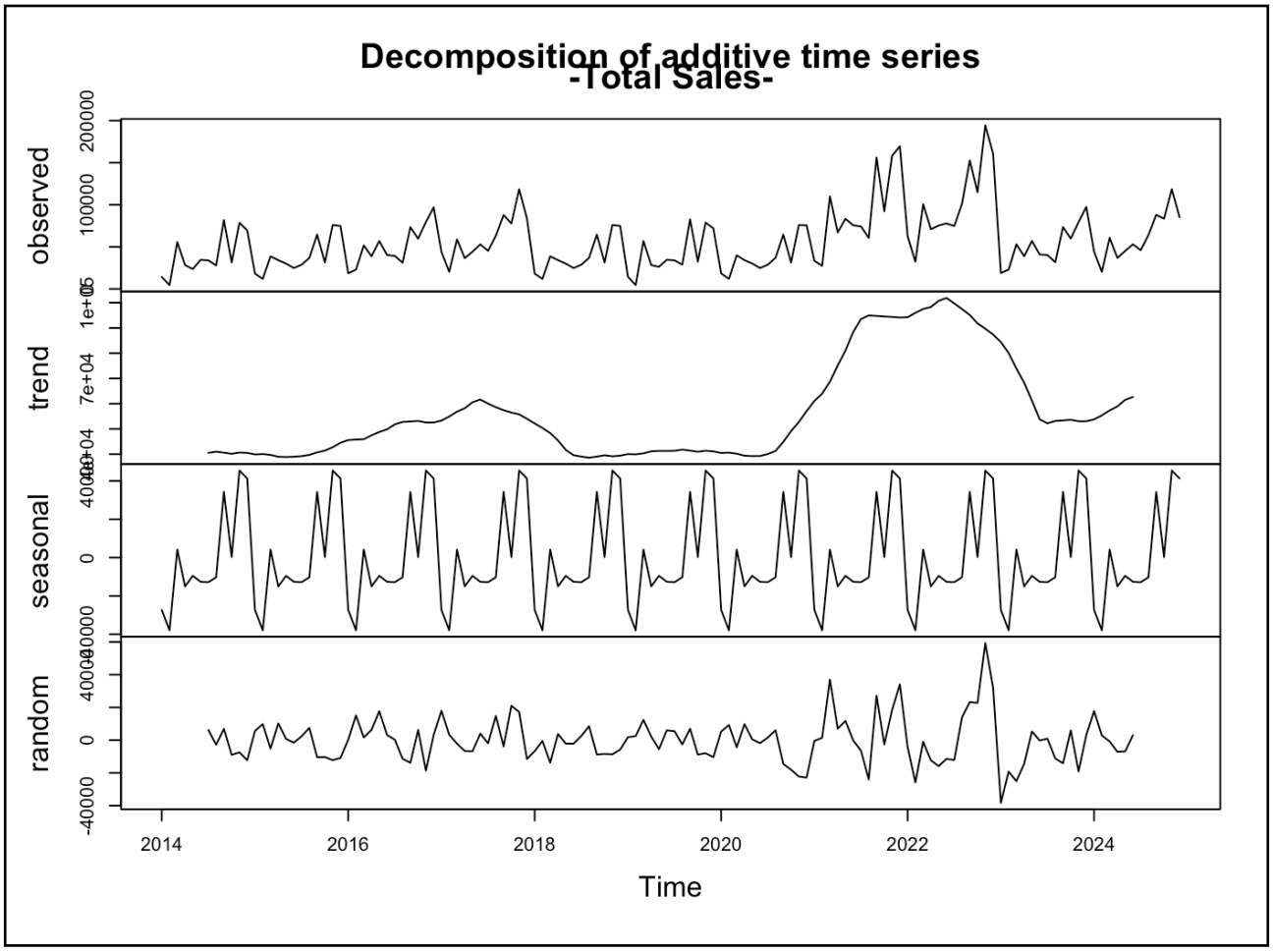
Appendix

```r
library(readr)
library(dplyr)
library(rpart)
library(rpart.plot)
library(caret)
library(corrplot)
library(ggplot2)
library(lubridate)
library(forecast)

# Read dataset
dataset <- read_csv("data/dataset.csv")

#########################################################
# Overview Dataset
#########################################################

# Remove NA values
dataset <- dataset %>% filter(!is.na(Category) & !is.na(`Sub-Category`))

# Bar plot of Sales by Category (transposed)
category_sales <- dataset %>%
  group_by(Category) %>%
  summarise(Total_Sales = sum(Sales, na.rm = TRUE))

ggplot(category_sales, aes(x = Total_Sales, y = reorder(Category, Total_Sales), fill = Category)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Total Sales by Category", x = "Total Sales", y = "Category") +
  theme(legend.position = "none") +
  scale_x_continuous(labels = scales::comma)

# Bar plot of Sales by Sub-Category (transposed)
subcategory_sales <- dataset %>%
  group_by(`Sub-Category`) %>%
  summarise(Total_Sales = sum(Sales, na.rm = TRUE))

```

```r
ggplot(subcategory_sales, aes(x = Total_Sales, y = reorder(`Sub-Category`, Total_Sales), fill = `Sub-Category`)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Total Sales by Sub-Category", x = "Total Sales", y = "Sub-Category") +
  theme(legend.position = "none") +
  scale_x_continuous(labels = scales::comma)

# Scatter plot of Sales vs. Profit with y = 0 line
ggplot(dataset, aes(x = Sales, y = Profit)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  theme_minimal() +
  labs(title = "Sales vs. Profit", x = "Sales", y = "Profit") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma)

#########################################################
# Decision Tree Prediction
#########################################################


# Create a binary target variable for Profit (Positive or Negative)
dataset <- dataset %>%
  mutate(Profit_Category = ifelse(Profit > 0, "Positive", "Negative"))

# Convert categorical variables to factors
dataset$Category <- as.factor(dataset$Category)
dataset$`Sub-Category` <- as.factor(dataset$`Sub-Category`)
dataset$`Ship Mode` <- as.factor(dataset$`Ship Mode`)
dataset$Segment <- as.factor(dataset$Segment)
dataset$Region <- as.factor(dataset$Region)
dataset$Profit_Category <- as.factor(dataset$Profit_Category)

# Select features and target for the decision tree
features <- dataset %>% select(Sales, Quantity, Discount, Category, `Sub-Category`, `Ship Mode`, Segment, Region)
target <- dataset$Profit_Category

```

```r
# Split the data into training and testing sets
set.seed(42)
train_index <- createDataPartition(target, p = 0.8, list = FALSE)
train_data <- dataset[train_index, ]
test_data <- dataset[-train_index, ]

# Train a decision tree model
decision_tree <- rpart(Profit_Category ~ Sales + Quantity + Discount + Category + `Sub-Category` + `Ship Mode` + Segment + Region,
                       data = train_data, method = "class")

# Predict on the test set
predictions <- predict(decision_tree, test_data, type = "class")

# Generate confusion matrix
confusion_matrix <- confusionMatrix(predictions, test_data$Profit_Category)


# Decomposing the time series for Average Discount
discount_decomp <- decompose(ts_discount)
plot(discount_decomp)
title(main = "-Average Discount-")

# Decomposing the time series for Total Sales
sales_decomp <- decompose(ts_sales)
plot(sales_decomp)
title(main = "-Total Sales-")

# Convert Order Date to Date format
dataset_discount$`Order Date` <- as.Date(dataset_discount$`Order Date`, format = "%Y/%m/%d")

# Aggregate discount by month
discount_over_time <- dataset_discount %>%
  mutate(Month = floor_date(`Order Date`, "month")) %>%
  group_by(Month) %>%
  summarise(Average_Discount = mean(Discount, na.rm = TRUE))

# Prepare the time series data for Discount
ts_discount <- ts(discount_over_time$Average_Discount, start = c(year(min(discount_over_time$Month)), month(min(discount_over_time$Month))), frequency = 12)
# Plot the original discount data
ggplot(discount_over_time, aes(x = Month, y = Average_Discount)) +
  geom_line(color = "blue", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Average Discount Over Time", x = "Month", y = "Average Discount") +
  scale_x_date(date_labels = "%Y-%m", date_breaks = "6 months") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8))

# Fit ARIMA model for Discount
arima_discount_model <- auto.arima(ts_discount)

# Forecast for the next 5 years (60 months) for Discount
forecast_horizon <- 5 * 12
arima_discount_forecast <- forecast(arima_discount_model, h = forecast_horizon)

# Plot the forecasted values for Discount
autoplot(arima_discount_forecast) +
  theme_minimal() +
  labs(title = "ARIMA Forecast of Average Discount", x = "Year", y = "Average Discount") +
  scale_y_continuous(labels = scales::comma)

############################################################
# SARIMA Forecasting for Sales
############################################################

# Remove NA values for Sales
dataset_sales <- dataset %>% filter(!is.na(Sales))

# Convert Order Date to Date format
dataset_sales$`Order Date` <- as.Date(dataset_sales$`Order Date`, format = "%Y/%m/%d")

# Aggregate sales by month
sales_over_time <- dataset_sales %>%
  mutate(Month = floor_date(`Order Date`, "month")) %>%
  group_by(Month) %>%
  summarise(Total_Sales = sum(Sales, na.rm = TRUE))

# Prepare the time series data for Sales
ts_sales <- ts(sales_over_time$Total_Sales, start = c(year(min(sales_over_time$Month)), month(min(sales_over_time$Month))), frequency = 12)

# Plot the original sales data
ggplot(sales_over_time, aes(x = Month, y = Total_Sales)) +
  geom_line(color = "blue", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Total Sales Over Time", x = "Month", y = "Total Sales") +
  scale_x_date(date_labels = "%Y-%m", date_breaks = "6 months") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8))

# Fit SARIMA model for Sales
sarima_sales_model <- auto.arima(ts_sales, seasonal = TRUE)

# Forecast for the next 5 years (60 months) for Sales
sarima_forecast_horizon <- 5 * 12
sarima_sales_forecast <- forecast(sarima_sales_model, h = sarima_forecast_horizon)

autoplot(sarima_sales_forecast) +
  theme_minimal() +
  labs(title = "SARIMA Forecast of Total Sales", x = "Year", y = "Total Sales") +
  scale_y_continuous(labels = scales::comma)
```


Decomposition of additive time series
-Average Discount-
observed
trend
seasonal
random
0.18
0.14
0.156
0.000
0.00
-0.04
2014
2016
2018
2020
2022
2024
Time


Decomposition of additive time series
-Total Sales-
observed
trend
seasonal
random
200000
100000
7e+04
0
-40000
2014
2016
2018
2020
2022
2024
Time