“Optimizing Retail Profitability through Data-Driven Sales Analysis”

A Capstone Project Report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING
(Specialization in Data Science)

by

YARLAGADDA ABHIRAM VU22CSEN0500063
VINNAKOTA NITISH RAJ VU22CSEN0500110
B BHARATH VU22CSEN0500149

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
GITAM SCHOOL OF TECHNOLOGY
GITAM (Deemed to be University)
VISAKHAPATNAM
2024

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

GITAM SCHOOL OF TECHNOLOGY

GITAM (Deemed to be University)

DECLARATION

I hereby declare that the capstone project report entitled **Optimizing Retail Profitability through Data-Driven Sales Analysis** is an original work done in the Department of Artificial Intelligence and Data Science, GITAM School of Technology, GITAM (Deemed to be University) submitted in partial fulfillment of the requirements for the award of the degree of B.Tech. in Computer Science and Engineering (Data Science). The work has not been submitted to any other college or University for the award of any degree or diploma.

Date: 22-10-2024

Registration No	Name	Signature
VU22CSEN0500063	Y.Abhiram	
VU22CSEN0500110	V.Nitish Raj	
VU22CSEN0500149	B.Bharath	

Project Title:

"Optimizing Retail Profitability through Data-Driven Sales Analysis"

Abstract:

This project focuses on analyzing the sales and profit data from a fictional "Super Store" to uncover regional and segment-based insights, improve profitability, and predict high-profit transactions using logistic regression. Various visualization techniques are used to explore relationships among categorical and numerical variables. A logistic regression model is implemented to classify transactions as high or low profit based on sales, discount, and quantity. This study demonstrates how data-driven methods can aid in understanding key performance drivers in retail.

Introduction:

Retailers rely heavily on data analysis to optimize sales strategies and improve profitability across various product categories and geographic segments. In the highly competitive retail landscape, understanding customer preferences, regional demand, and product profitability is essential. This project leverages the "Super Store" dataset, analyzing critical variables like "Sales," "Profit," "Category," and "Region." The project applies descriptive and predictive analytics techniques to gain insights into profitable sales strategies and customer behaviour.

Problem Statement:

How can Super Store leverage data analytics to identify high-profit opportunities across regions, segments, and product categories? This analysis should also aim to predict high-profit transactions using logistic regression, which can aid managers in focusing resources effectively on high-profit areas.

Objective:

The primary objectives of this project are:

1. To analyze sales and profit trends across different regions, segments, and product categories.
2. To predict high-profit transactions using logistic regression, based on factors such as sales, discount, and quantity.
3. To evaluate the effectiveness of predictive models using performance metrics like accuracy, precision, recall, and F1 score.

Literature Survey:

In recent years, the application of data analytics in retail has expanded, allowing companies to optimize their strategies by understanding customer behaviour and product performance. Various studies have shown that regional analysis and segmentation can significantly enhance sales strategies. Predictive models, particularly logistic regression, have been effective in classifying transaction profitability. This project builds on these studies by implementing a logistic regression model for binary classification of profit, considering the impact of discount strategies and quantity on profitability.

Proposed Methodology:

The project methodology consists of two main phases: data analysis and predictive modeling.

1. **Data Loading and Cleaning:**
    - Load and inspect the dataset, check for missing values, and understand data types.
    - Summarize statistics for key variables to understand data distribution and detect any data quality issues.
2. **Exploratory Data Analysis (EDA):**
    - Use bar and pie charts to visualize sales and profit distributions across different regions, segments, and categories.

    - Plot state-level sales and profit data to identify high-performing regions and states.
    - Calculate and visualize the correlation matrix for numerical variables to examine relationships between sales, profit, discount, and quantity.
3. **Binary Classification (Logistic Regression):**
    - Define "High Profit" transactions based on the median profit as the threshold, setting binary outcomes for logistic regression.
    - Split the data into training and testing sets (70% train, 30% test).
    - Fit a logistic regression model to predict high-profit transactions based on "Sales," "Discount," and "Quantity."
4. **Model Evaluation:**
    - Calculate accuracy, precision, recall, and F1 score from the confusion matrix for the logistic regression model to assess performance.
    - Interpret the results and use them to make recommendations for optimizing profit through sales and discounts.
5. **Visualizations and Results:**
    - Use ggplot2 to display insights from the EDA, logistic regression, and performance metrics to support actionable conclusions.

Data Set Link:

Sample Superstore Dataset

Code:

```
install.packages("dplyr")
install.packages("ggplot2")
install.packages("data.table")
install.packages("caTools")
install.packages('reshape2')

library(reshape2)
library(dplyr)      # for data manipulation
library(ggplot2)    # for plotting
library(data.table) # for efficient data handling
library(caTools)

library(readr)
df <- read_csv("C:/Users/hp/Downloads/Super Store.csv")
View(df)

# Display the first few rows
head(df)
```

```
# Display the unique values in each categorical column
unique(df$`Ship Mode`)
unique(df$Segment)
unique(df$Country)
unique(df$Category)
unique(df$`Sub-Category`)
unique(df$Region)

# Statistical description of the data
summary(df)

# Information about the dataset structure
str(df)

# Checking for missing values
colSums(is.na(df))

# Load the scales library for formatting
library(scales)
```

```
# Sales analysis based on region (Bar Plot)

sales_by_region <- df %>% group_by(Region) %>% summarize(Sales = sum(Sales))

ggplot(sales_by_region, aes(x = Region, y = Sales, fill = Region)) +

  geom_bar(stat = "identity") + D

labs(title = "Sales Analysis by Region", x = "Region", y = "Sales") +

  scale_y_continuous(labels = scales::comma) +  # format y-axis

  scale_fill_brewer(palette = "Set3") +

  theme_minimal()


# Profit analysis based on region (Bar Plot)

profit_by_region <- df %>% group_by(Region) %>% summarize(Profit = sum(Profit))

ggplot(profit_by_region, aes(x = Region, y = Profit, fill = Region)) +

  geom_bar(stat = "identity") +

  labs(title = "Profit Analysis by Region", x = "Region", y = "Profit") +

  scale_y_continuous(labels = scales::comma) +  # format y-axis

  scale_fill_brewer(palette = "Set2") +

  theme_minimal()
```

```
# Sales analysis based on region (Pie Chart)
ggplot(sales_by_region, aes(x = "", y = Sales, fill = Region)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  labs(title = "Sales Analysis by Region (Pie Chart)") +
  scale_fill_brewer(palette = "Set3") +
  theme_minimal()

# Profit analysis based on region (Pie Chart)
ggplot(profit_by_region, aes(x = "", y = Profit, fill = Region)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  labs(title = "Profit Analysis by Region (Pie Chart)") +
  scale_fill_brewer(palette = "Set2") +
  theme_minimal()
```

```
# Sales analysis based on segment (Bar Plot)

sales_by_segment <- df %>% group_by(Segment) %>% summarize(Sales = sum(Sales))

ggplot(sales_by_segment, aes(x = Segment, y = Sales, fill = Segment)) +

  geom_bar(stat = "identity") +

  labs(title = "Sales Analysis by Segment", x = "Segment", y = "Sales") +

  scale_y_continuous(labels = scales::comma) +  # format y-axis

  scale_fill_manual(values = c("dodgerblue", "tomato", "goldenrod")) +

  theme_minimal()


# Profit analysis based on segment (Bar Plot)

profit_by_segment <- df %>% group_by(Segment) %>% summarize(Profit = sum(Profit))

ggplot(profit_by_segment, aes(x = Segment, y = Profit, fill = Segment)) +

  geom_bar(stat = "identity") +

  labs(title = "Profit Analysis by Segment", x = "Segment", y = "Profit") +

  scale_y_continuous(labels = scales::comma) +  # format y-axis

  scale_fill_manual(values = c("dodgerblue", "tomato", "goldenrod")) +

  theme_minimal()
```

```
# Sales analysis based on category (Bar Plot)

sales_by_category <- df %>% group_by(Category) %>% summarize(Sales = sum(Sales))

ggplot(sales_by_category, aes(x = Category, y = Sales, fill = Category)) +

  geom_bar(stat = "identity") +

  labs(title = "Sales Analysis by Category", x = "Category", y = "Sales") +

  scale_y_continuous(labels = scales::comma) +  # format y-axis

  scale_fill_manual(values = c("skyblue", "lightcoral", "palegreen")) +

  theme_minimal()


# Profit analysis based on category (Bar Plot)

profit_by_category <- df %>% group_by(Category) %>% summarize(Profit = sum(Profit))

ggplot(profit_by_category, aes(x = Category, y = Profit, fill = Category)) +

  geom_bar(stat = "identity") +

  labs(title = "Profit Analysis by Category", x = "Category", y = "Profit") +

  scale_y_continuous(labels = scales::comma) +  # format y-axis

  scale_fill_manual(values = c("skyblue", "lightcoral", "palegreen")) +

  theme_minimal()
```

```
# Sales analysis based on category (Pie Chart)
ggplot(sales_by_category, aes(x = "", y = Sales, fill = Category)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  labs(title = "Sales Analysis by Category (Pie Chart)") +
  scale_fill_manual(values = c("skyblue", "lightcoral", "palegreen")) +
  theme_minimal()

# Profit analysis based on category (Pie Chart)
ggplot(profit_by_category, aes(x = "", y = Profit, fill = Category)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  labs(title = "Profit Analysis by Category (Pie Chart)") +
  scale_fill_manual(values = c("skyblue", "lightcoral", "palegreen")) +
  theme_minimal()

# Sales analysis based on state (Bar Plot)
sales_by_state <- df %>%
  group_by(State) %>%
  summarize(Sales = sum(Sales))
```

```
ggplot(sales_by_state, aes(x = State, y = Sales)) +
  geom_bar(stat = "identity") +
  labs(title = "Sales Analysis by State", x = "State", y = "Sales") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Profit analysis based on state (Bar Plot)
profit_by_state <- df %>%
  group_by(State) %>%
  summarize(Profit = sum(Profit))
ggplot(profit_by_state, aes(x = State, y = Profit)) +
  geom_bar(stat = "identity") +
  labs(title = "Profit Analysis by State", x = "State", y = "Profit") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Step 1: Create a binary outcome for "High Profit"
# Using median profit value as the threshold to classify high and low profit
threshold <- median(df$Profit, na.rm = TRUE)
df <- df %>% mutate(High_Profit = ifelse(Profit > threshold, 1, 0))  # 1 for high profit, 0 for low profit
```

```
# Step 2: Split the data into training and testing sets

set.seed(123)  # For reproducibility

sample <- sample.split(df$High_Profit, SplitRatio = 0.7)

train <- subset(df, sample == TRUE)

test <- subset(df, sample == FALSE)


# Step 3: Fit the logistic regression model

# Predicting High_Profit based on Sales, Discount, and Quantity

model <- glm(High_Profit ~ Sales + Discount + Quantity, data = train, family = binomial)


# Summary of the model to check coefficients and model fit

summary(model)


# Step 4: Predict on the test set

# Get predicted probabilities for the test set

pred_prob <- predict(model, test, type = "response")


# Convert probabilities to binary predictions with a threshold of 0.5

pred_class <- ifelse(pred_prob > 0.5, 1, 0)
```

```
# Step 5: Evaluate the model

# Confusion matrix to check accuracy

confusion_matrix <- table(Predicted = pred_class, Actual = test$High_Profit)

print(confusion_matrix)


# Extract values from confusion matrix

TP <- confusion_matrix[2, 2]  # True Positives

TN <- confusion_matrix[1, 1]  # True Negatives

FP <- confusion_matrix[2, 1]  # False Positives

FN <- confusion_matrix[1, 2]  # False Negatives


# Calculate precision and recall

precision <- TP / (TP + FP)

recall <- TP / (TP + FN)

f1_score <- 2 * ((precision * recall) / (precision + recall))


cat("Precision:", round(precision, 2), "\n")

cat("Recall:", round(recall, 2), "\n")

cat("F1 Score:", round(f1_score, 2), "\n")
```

```
# Calculate accuracy
accuracy <- mean(pred_class == test$High_Profit)
print(paste("Accuracy:", round(accuracy, 2)))

# Selecting numerical columns from the dataset
numerical_data <- df %>% select(Sales, Quantity, Discount, Profit)

# Calculate the correlation matrix
correlation_matrix <- cor(numerical_data, use = "complete.obs")
print("Correlation Matrix:")
print(correlation_matrix)

# Reshape for ggplot
correlation_melted <- melt(correlation_matrix)

# Plot heatmap
ggplot(data = correlation_melted, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint
= 0) +
```

theme_minimal() +

labs(title = "Correlation Matrix Heatmap", x = "Variables", y = "Variables")

Step 6: Optional Visualization

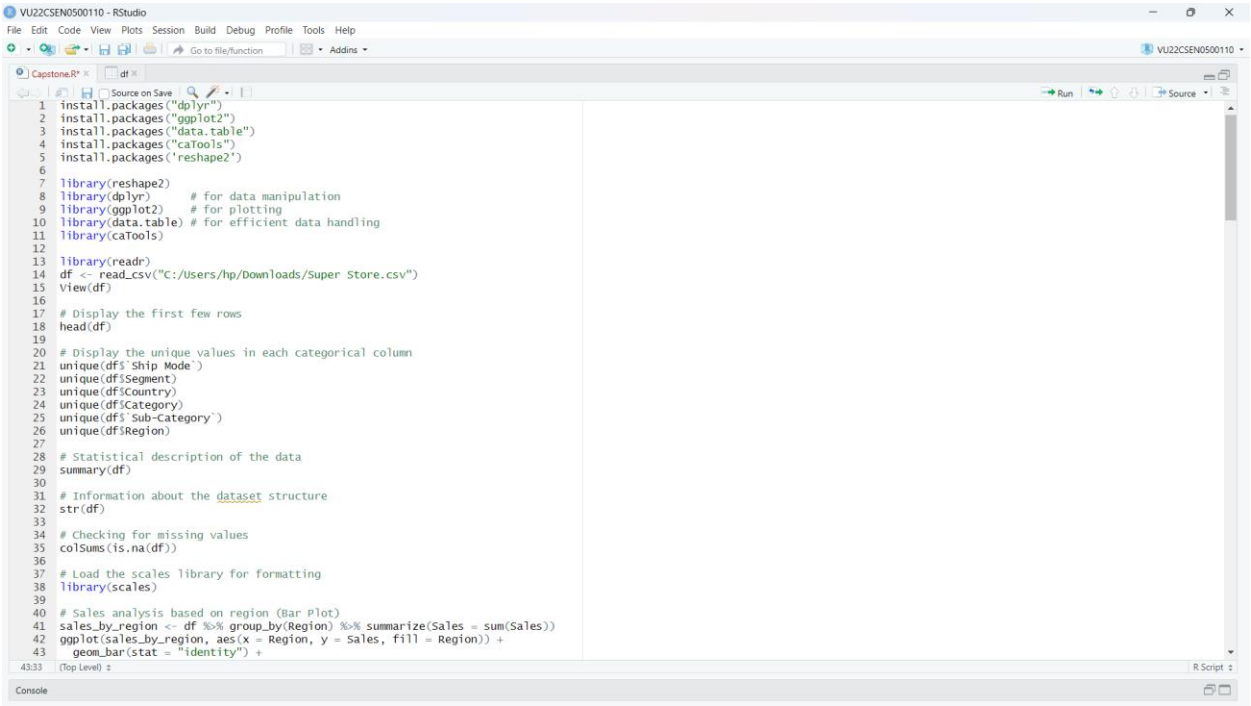
Plot predicted probabilities and actual outcomes for test data

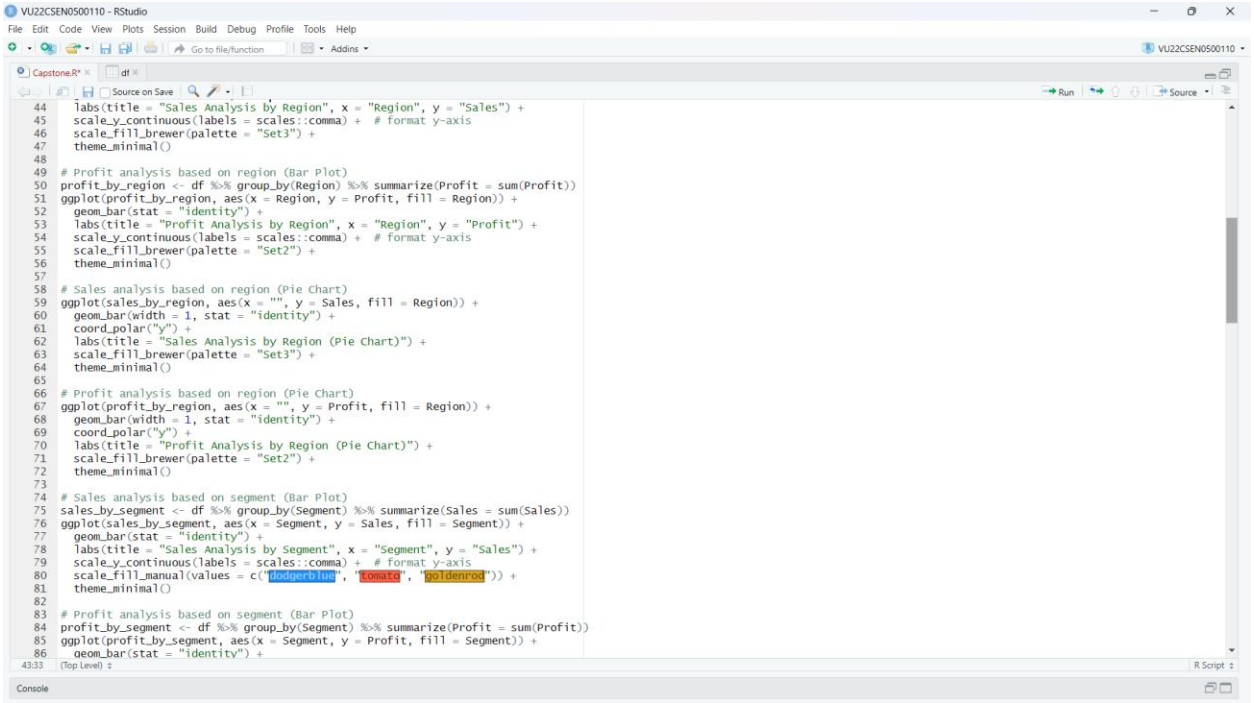
ggplot(test, aes(x = pred_prob, fill = factor(High_Profit))) +

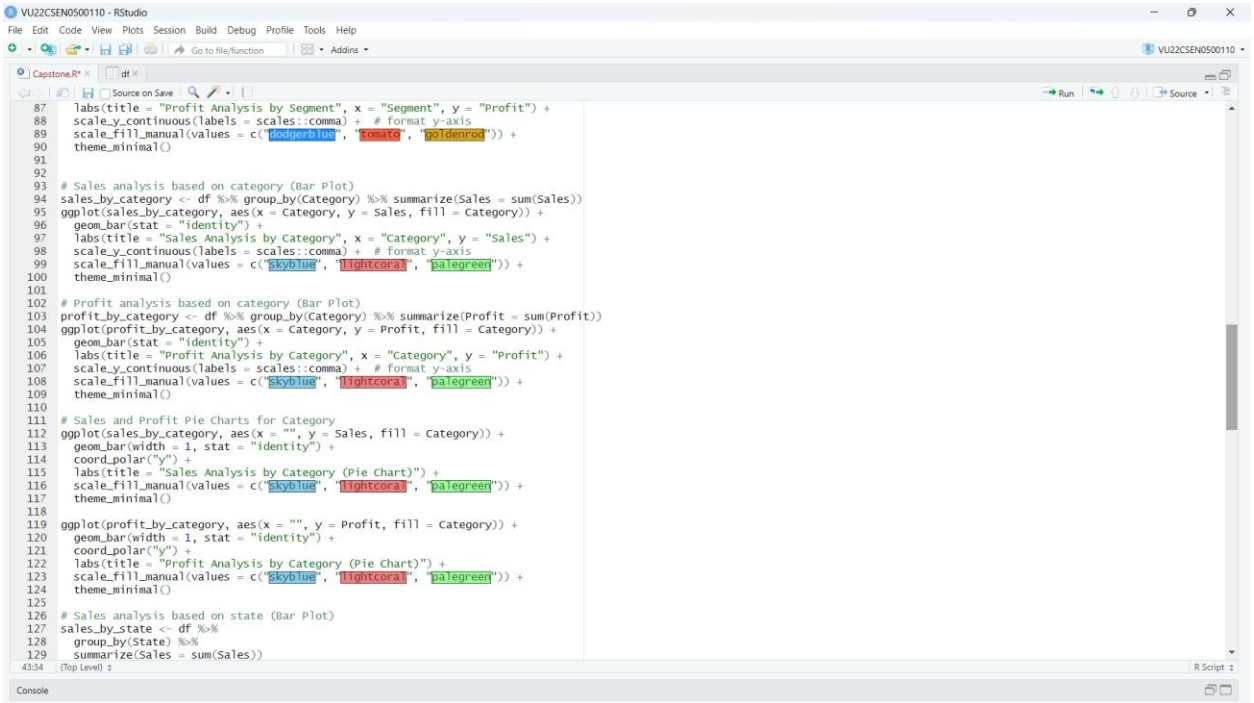
geom_histogram(position = "dodge", bins = 30) +

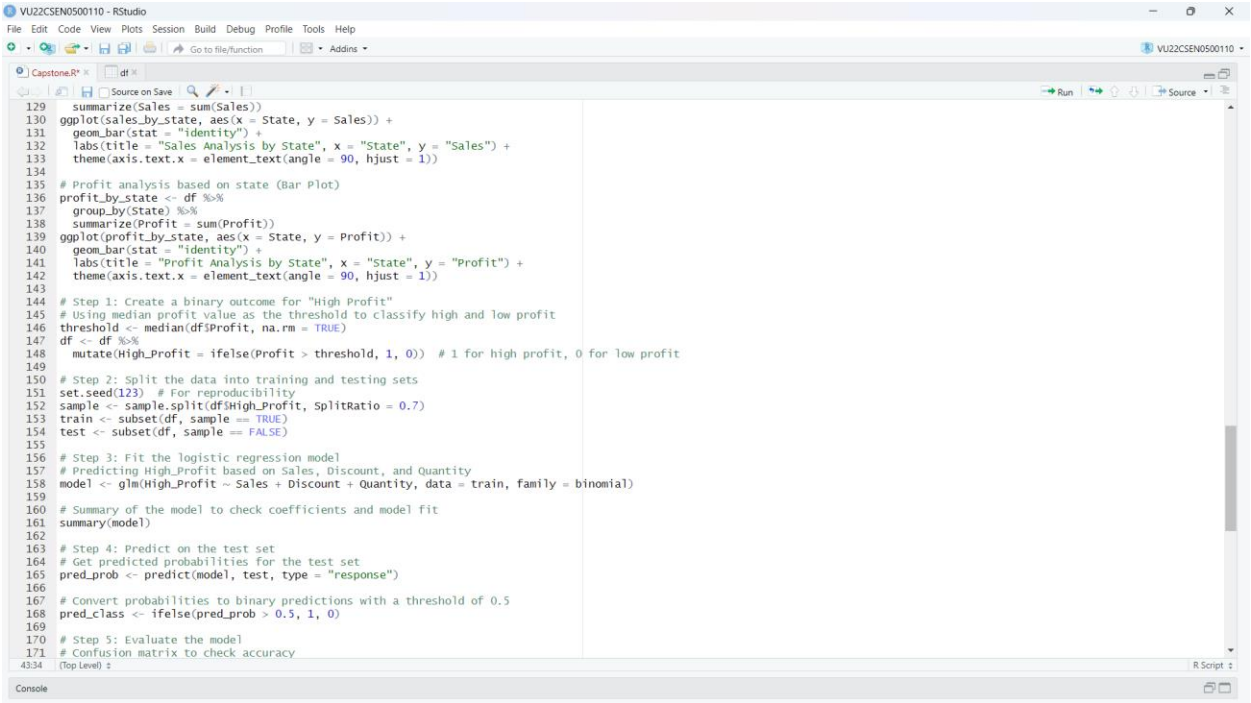
labs(title = "Predicted Probability vs Actual Outcome", x = "Predicted Probability", fill = "High Profit") +

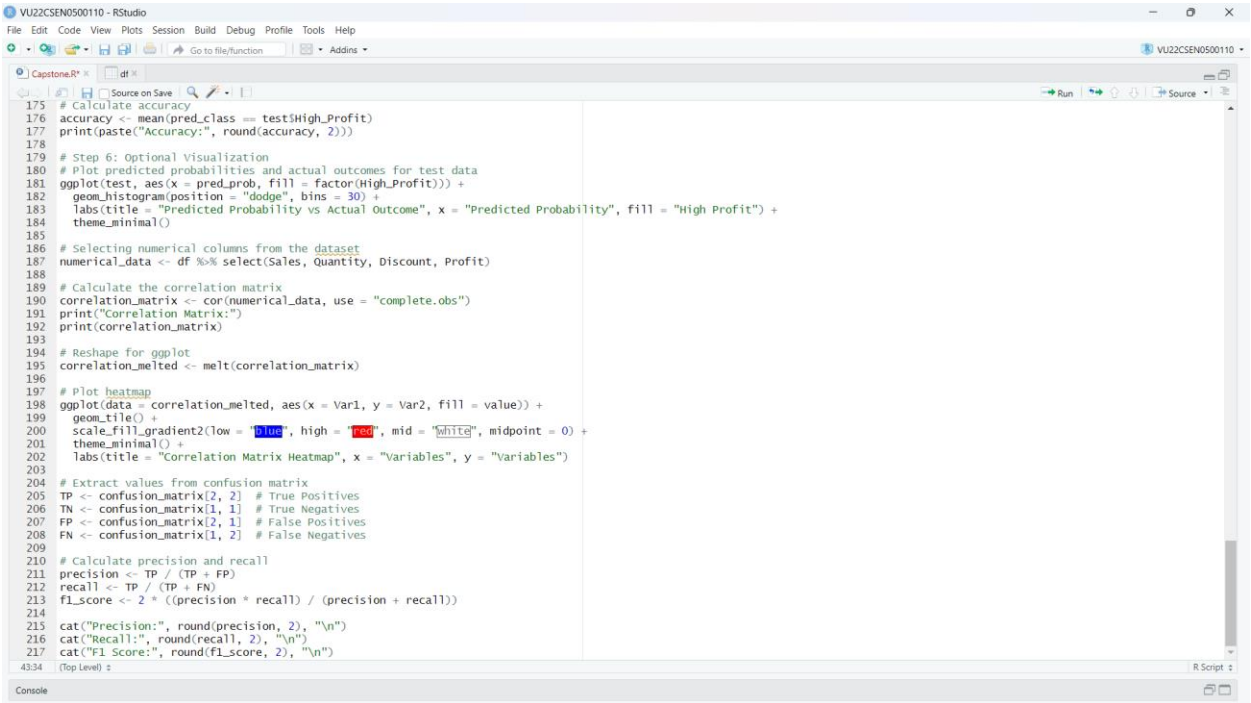
theme_minimal()




VU22CSEN0500110 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Capstone.R*
df
Source on Save
Run
Source
labs(title = "Sales Analysis by Region", x = "Region", y = "Sales") +
scale_y_continuous(labels = scales::comma) + # format y-axis
scale_fill_brewer(palette = "Set3") +
theme_minimal()
# Profit analysis based on region (Bar Plot)
profit_by_region <- df %>% group_by(Region) %>% summarize(Profit = sum(Profit))
ggplot(profit_by_region, aes(x = Region, y = Profit, fill = Region)) +
geom_bar(stat = "identity") +
labs(title = "Profit Analysis by Region", x = "Region", y = "Profit") +
scale_y_continuous(labels = scales::comma) + # format y-axis
scale_fill_brewer(palette = "Set2") +
theme_minimal()
# Sales analysis based on region (Pie Chart)
ggplot(sales_by_region, aes(x = "", y = Sales, fill = Region)) +
geom_bar(width = 1, stat = "identity") +
coord_polar("y") +
labs(title = "Sales Analysis by Region (Pie Chart)") +
scale_fill_brewer(palette = "Set3") +
theme_minimal()
# Profit analysis based on region (Pie Chart)
ggplot(profit_by_region, aes(x = "", y = Profit, fill = Region)) +
geom_bar(width = 1, stat = "identity") +
coord_polar("y") +
labs(title = "Profit Analysis by Region (Pie Chart)") +
scale_fill_brewer(palette = "Set2") +
theme_minimal()
# Sales analysis based on segment (Bar Plot)
sales_by_segment <- df %>% group_by(Segment) %>% summarize(Sales = sum(Sales))
ggplot(sales_by_segment, aes(x = Segment, y = Sales, fill = Segment)) +
geom_bar(stat = "identity") +
labs(title = "Sales Analysis by Segment", x = "Segment", y = "Sales") +
scale_y_continuous(labels = scales::comma) + # format y-axis
scale_fill_manual(values = c("dodgerblue", "tomato", "goldenrod")) +
theme_minimal()
# Profit analysis based on segment (Bar Plot)
profit_by_segment <- df %>% group_by(Segment) %>% summarize(Profit = sum(Profit))
ggplot(profit_by_segment, aes(x = Segment, y = Profit, fill = Segment)) +
geom_bar(stat = "identity") +
43:33 (Top Level) R Script
Console


VU22CSEN0500110 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Capstone.R*
df
Source on Save
Run
Source
labs(title = "Profit Analysis by Segment", x = "Segment", y = "Profit") +
scale_y_continuous(labels = scales::comma) + # format y-axis
scale_fill_manual(values = c("dodgerblue", "tomato", "goldenrod")) +
theme_minimal()
# Sales analysis based on category (Bar Plot)
sales_by_category <- df %>% group_by(Category) %>% summarize(Sales = sum(Sales))
ggplot(sales_by_category, aes(x = Category, y = Sales, fill = Category)) +
geom_bar(stat = "identity") +
labs(title = "Sales Analysis by Category", x = "Category", y = "Sales") +
scale_y_continuous(labels = scales::comma) + # format y-axis
scale_fill_manual(values = c("skyblue", "lightcoral", "palegreen")) +
theme_minimal()
# Profit analysis based on category (Bar Plot)
profit_by_category <- df %>% group_by(Category) %>% summarize(Profit = sum(Profit))
ggplot(profit_by_category, aes(x = Category, y = Profit, fill = Category)) +
geom_bar(stat = "identity") +
labs(title = "Profit Analysis by Category", x = "Category", y = "Profit") +
scale_y_continuous(labels = scales::comma) + # format y-axis
scale_fill_manual(values = c("skyblue", "lightcoral", "palegreen")) +
theme_minimal()
# Sales and Profit Pie Charts for Category
ggplot(sales_by_category, aes(x = "", y = Sales, fill = Category)) +
geom_bar(width = 1, stat = "identity") +
coord_polar("y") +
labs(title = "Sales Analysis by Category (Pie Chart)") +
scale_fill_manual(values = c("skyblue", "lightcoral", "palegreen")) +
theme_minimal()
ggplot(profit_by_category, aes(x = "", y = Profit, fill = Category)) +
geom_bar(width = 1, stat = "identity") +
coord_polar("y") +
labs(title = "Profit Analysis by Category (Pie Chart)") +
scale_fill_manual(values = c("skyblue", "lightcoral", "palegreen")) +
theme_minimal()
# Sales analysis based on state (Bar Plot)
sales_by_state <- df %>%
group_by(State) %>%
summarize(Sales = sum(Sales))
43:34 (Top Level) R Script
Console


VU22CSEN0500110 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Capstone.R*
df
Source on Save
Run
Source
summarize(Sales = sum(Sales))
ggplot(sales_by_state, aes(x = State, y = Sales)) +
geom_bar(stat = "identity") +
labs(title = "Sales Analysis by State", x = "State", y = "Sales") +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
# Profit analysis based on state (Bar Plot)
profit_by_state <- df %>%
group_by(State) %>%
summarize(Profit = sum(Profit))
ggplot(profit_by_state, aes(x = State, y = Profit)) +
geom_bar(stat = "identity") +
labs(title = "Profit Analysis by State", x = "State", y = "Profit") +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
# Step 1: Create a binary outcome for "High Profit"
# Using median profit value as the threshold to classify high and low profit
threshold <- median(df$Profit, na.rm = TRUE)
df <- df %>%
mutate(High_Profit = ifelse(Profit > threshold, 1, 0)) # 1 for high profit, 0 for low profit
Step 2: Split the data into training and testing sets
set.seed(123) # For reproducibility
sample <- sample.split(df$High_Profit, SplitRatio = 0.7)
train <- subset(df, sample == TRUE)
test <- subset(df, sample == FALSE)
# Step 3: Fit the logistic regression model
# Predicting High_Profit based on Sales, Discount, and Quantity
model <- glm(High_Profit ~ Sales + Discount + Quantity, data = train, family = binomial)
# Summary of the model to check coefficients and model fit
summary(model)
# Step 4: Predict on the test set
# Get predicted probabilities for the test set
pred_prob <- predict(model, test, type = "response")
# Convert probabilities to binary predictions with a threshold of 0.5
pred_class <- ifelse(pred_prob > 0.5, 1, 0)
# Step 5: Evaluate the model
# Confusion matrix to check accuracy
43:34
(Top Level)
R Script
Console


VU22CSEN0500110 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Capstone.R*
df
Source on Save
Run
Source
# Calculate accuracy
accuracy <- mean(pred_class == test$High_Profit)
print(paste("Accuracy:", round(accuracy, 2)))
# Step 6: Optional Visualization
# Plot predicted probabilities and actual outcomes for test data
ggplot(test, aes(x = pred_prob, fill = factor(High_Profit))) +
geom_histogram(position = "dodge", bins = 30) +
labs(title = "Predicted Probability vs Actual Outcome", x = "Predicted Probability", fill = "High Profit") +
theme_minimal()
# Selecting numerical columns from the dataset
numerical_data <- df %>% select(Sales, Quantity, Discount, Profit)
# Calculate the correlation matrix
correlation_matrix <- cor(numerical_data, use = "complete.obs")
print("Correlation Matrix:")
print(correlation_matrix)
# Reshape for ggplot
correlation_melted <- melt(correlation_matrix)
# Plot heatmap
ggplot(data = correlation_melted, aes(x = Var1, y = Var2, fill = value)) +
geom_tile() +
scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
theme_minimal() +
labs(title = "Correlation Matrix Heatmap", x = "Variables", y = "Variables")
# Extract values from confusion matrix
TP <- confusion_matrix[2, 2] # True Positives
TN <- confusion_matrix[1, 1] # True Negatives
FP <- confusion_matrix[2, 1] # False Positives
FN <- confusion_matrix[1, 2] # False Negatives
# Calculate precision and recall
precision <- TP / (TP + FP)
recall <- TP / (TP + FN)
f1_score <- 2 * ((precision * recall) / (precision + recall))
cat("Precision:", round(precision, 2), "\n")
cat("Recall:", round(recall, 2), "\n")
cat("F1 Score:", round(f1_score, 2), "\n")
43:34
(Top Level)
R Script
Console

OUTPUT:

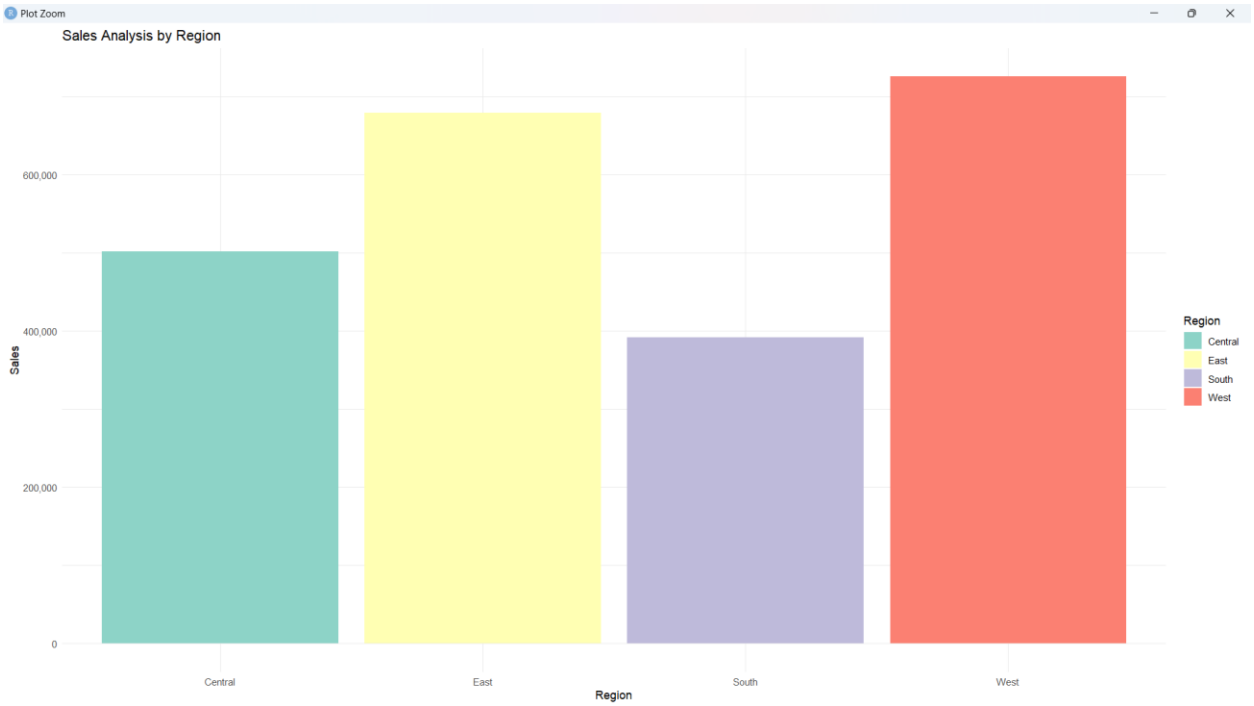


Figure 1: This Bar Graph Shows the Sales Analysis by Region (Central, West, East, South).

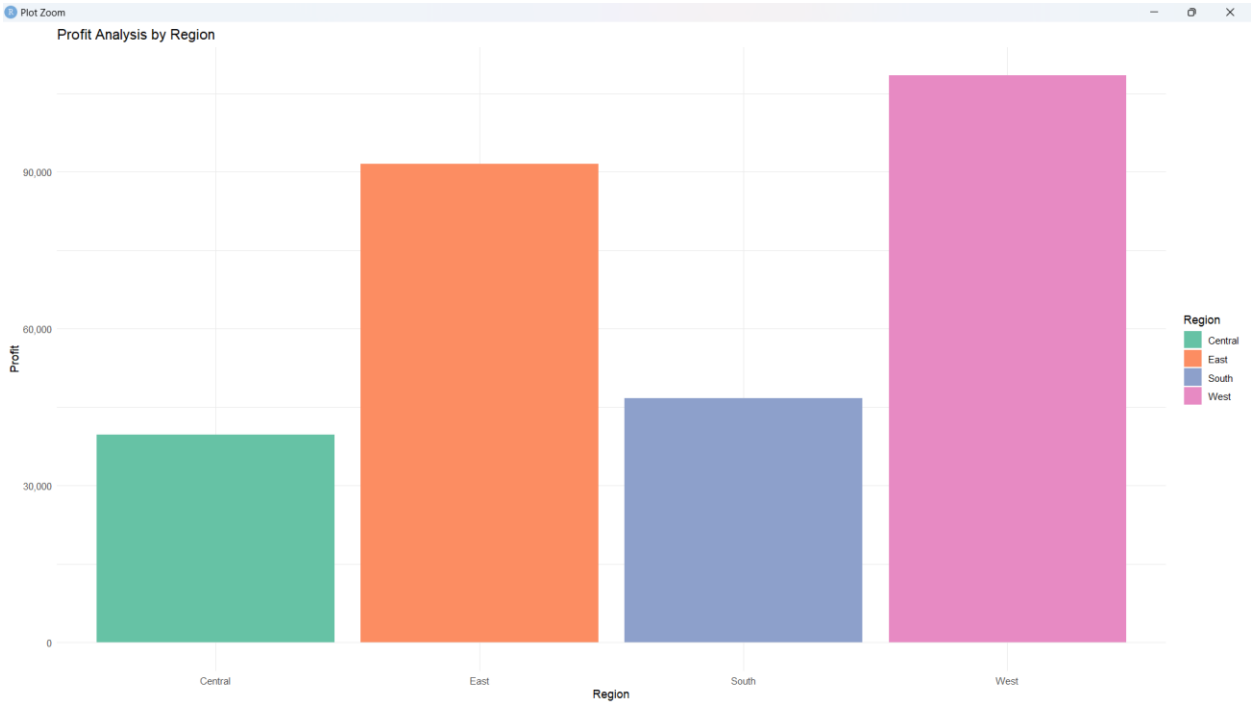


Figure 2: This Bar Graph Shows the Profit Analysis by Region (Central, West, East, South).

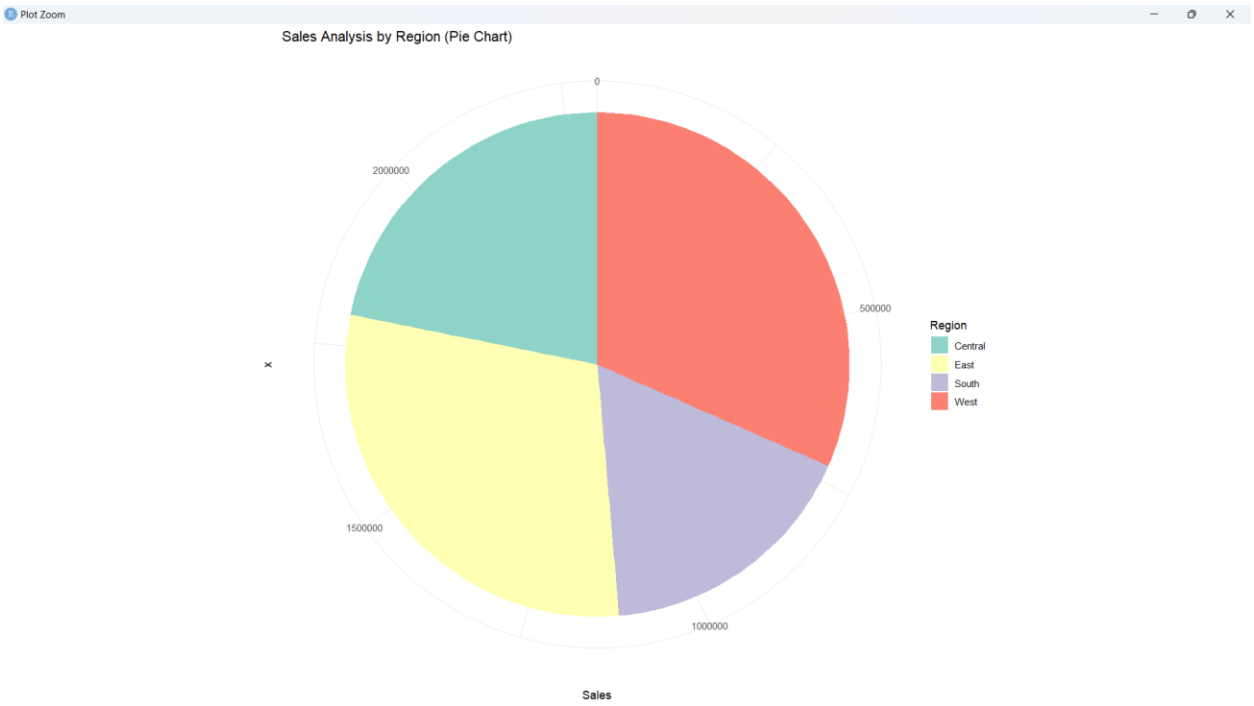


Figure 3: This Pie Chart Shows the Sales Analysis by Region (Central, West, East, South).

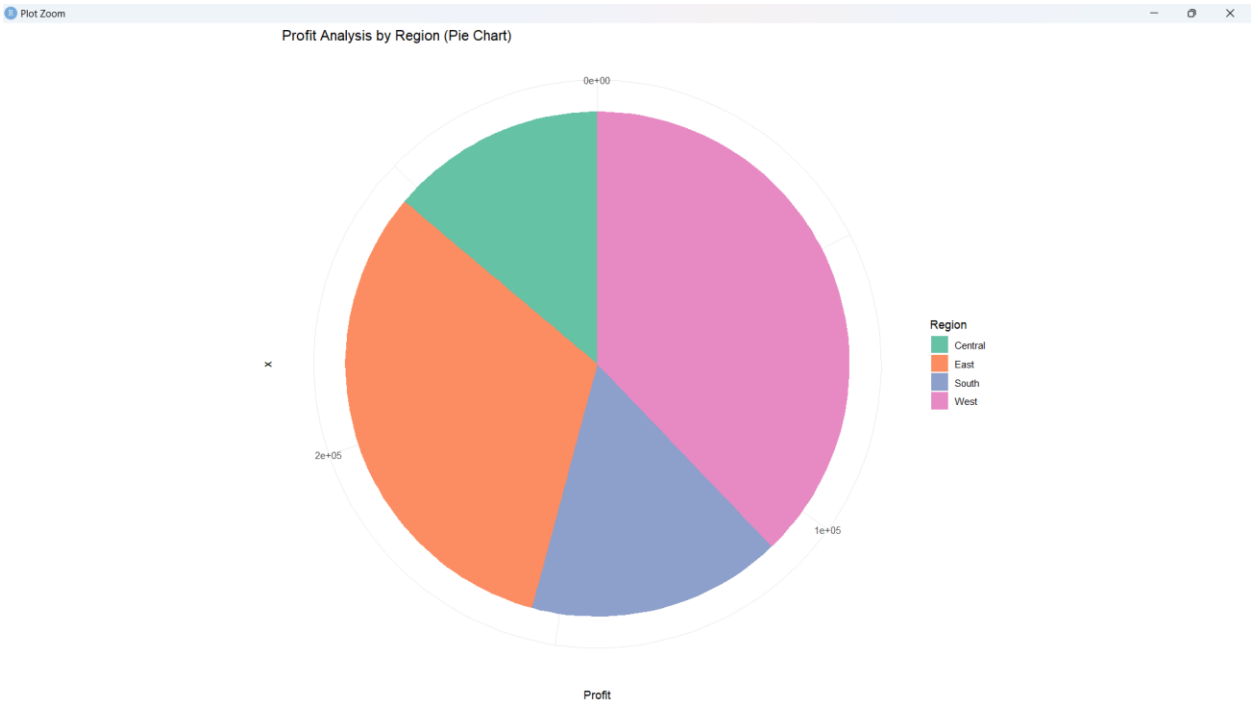


Figure 4: This Pie Chart Shows the Sales Analysis by Region (Central, West, East, South).

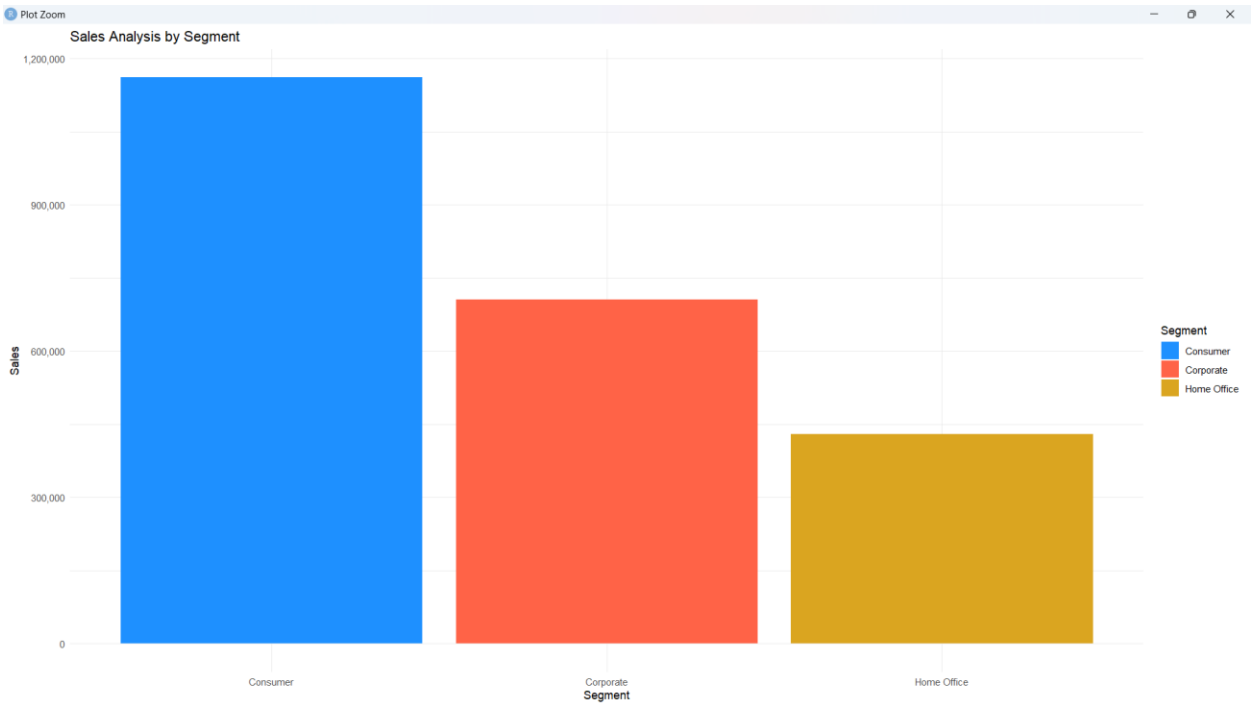


Figure 5: This Bar Graph Shows the Sales Analysis by Segment (Customer, Corporate, Home Office).

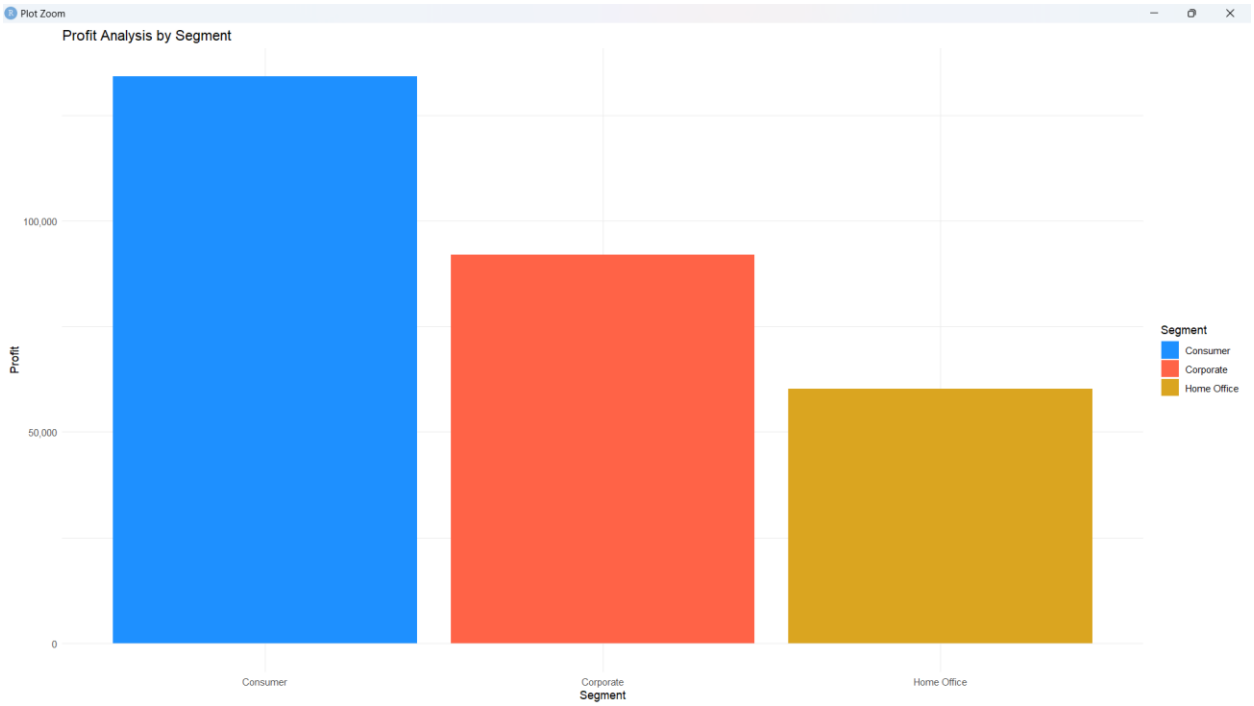


Figure 6: This Bar Graph Shows the Profit Analysis by Segment (Customer, Corporate, Home Office).

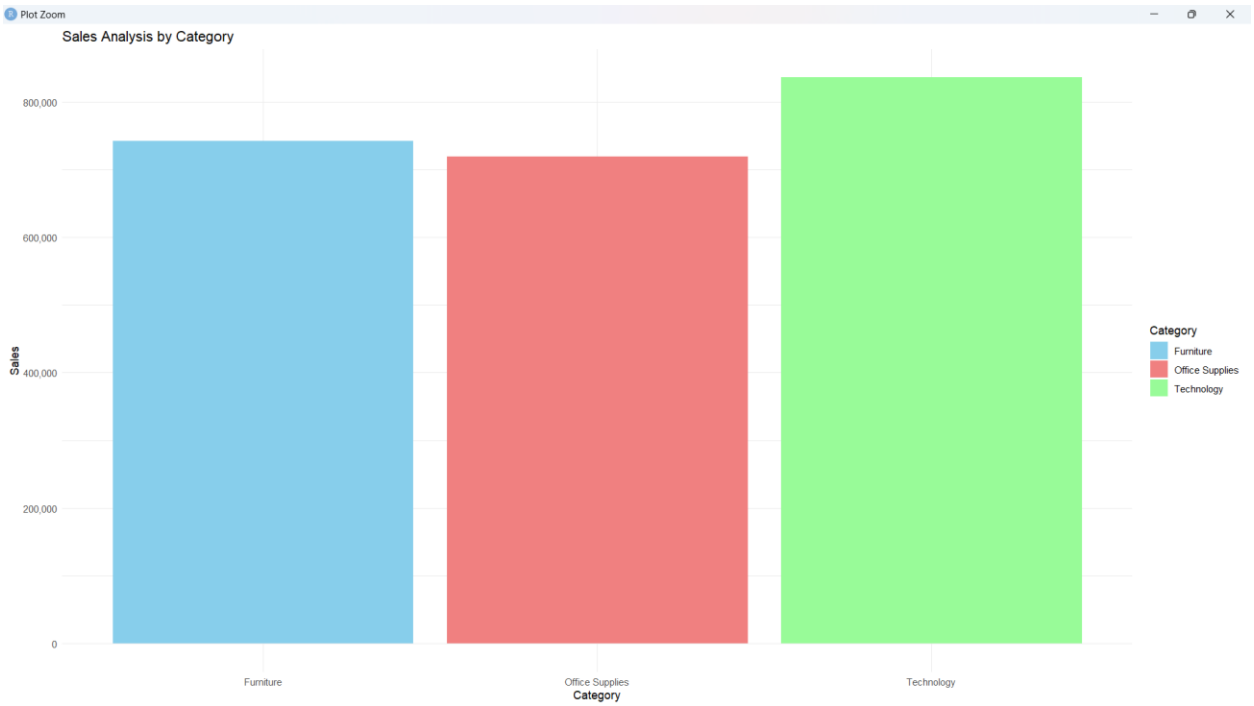


Figure 7: This Bar Graph Shows the Sales Analysis by Category (Furniture, Office Supplies, Technology)

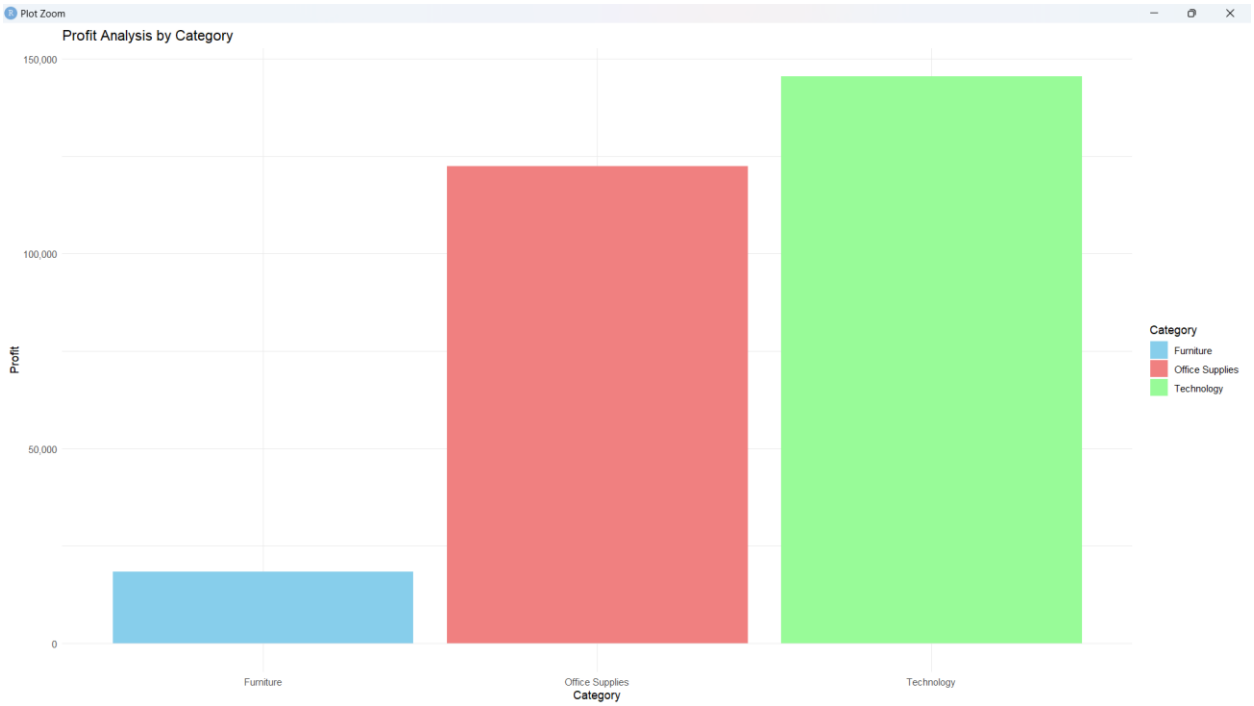


Figure 8: This Bar Graph Shows the Profit Analysis by Category (Furniture, Office Supplies, Technology)

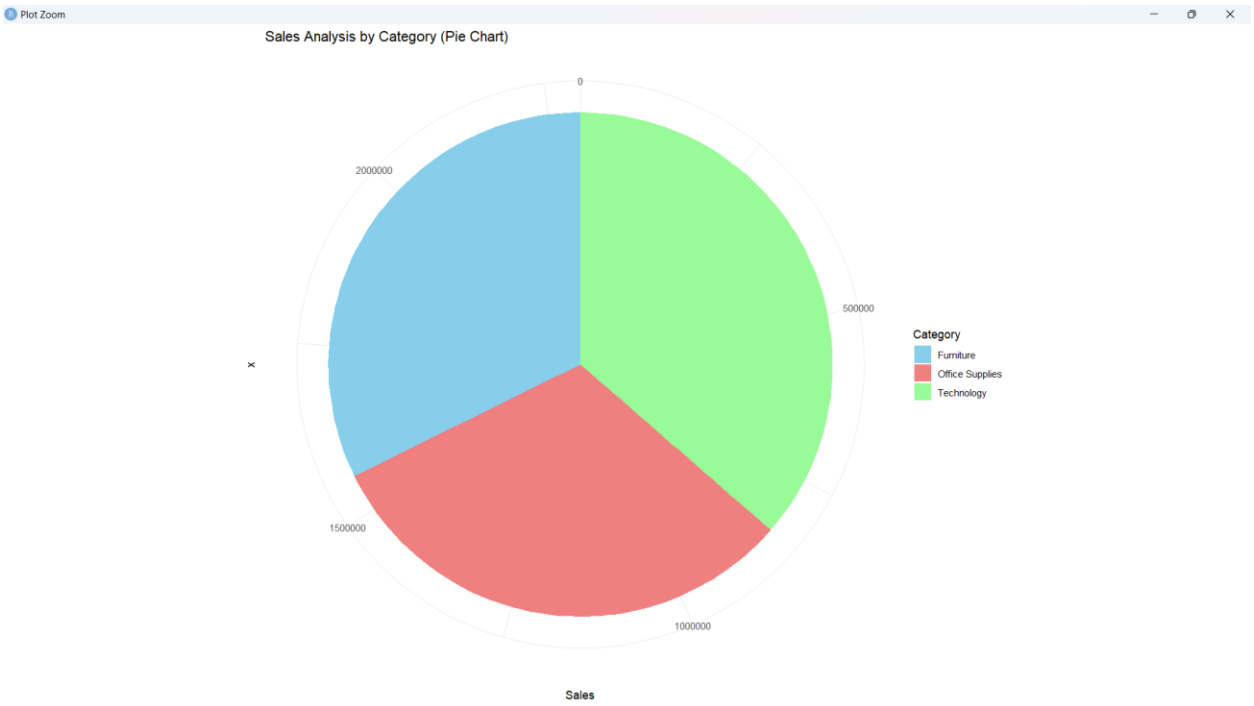


Figure 9: This Pie Chart Shows the Sales Analysis by Category (Furniture, Office Supplies, Technology)

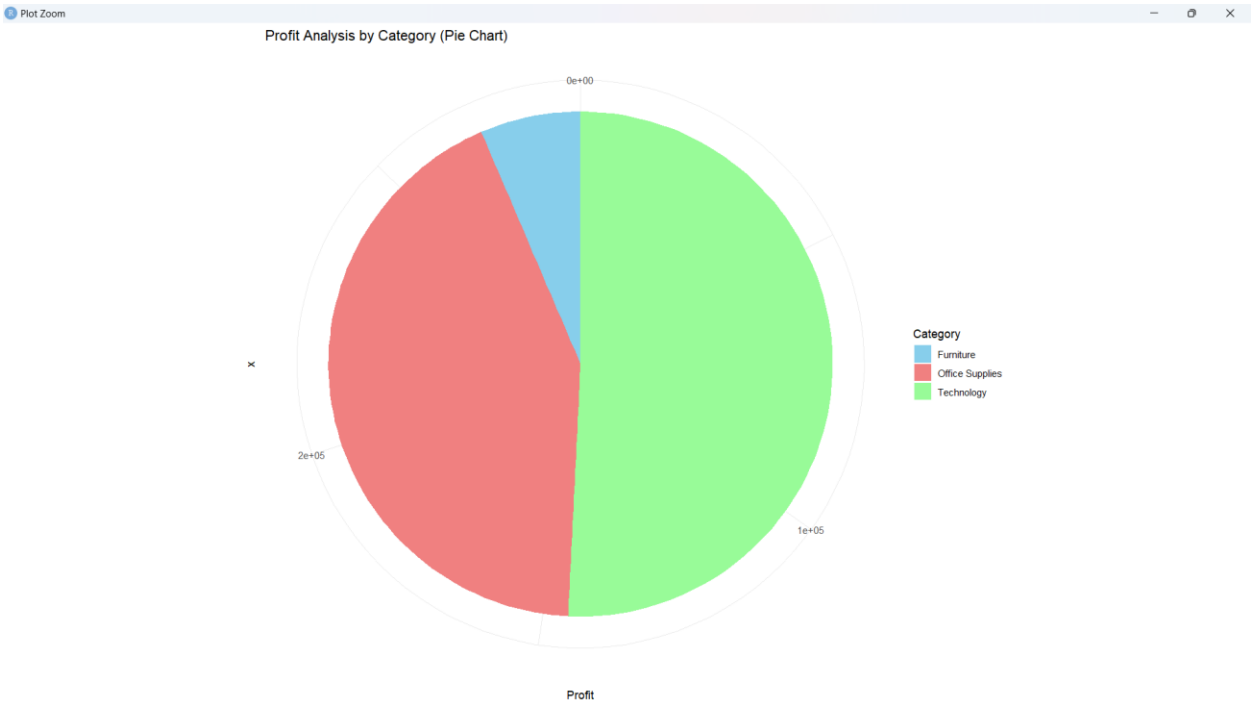


Figure 10: This Pie Chart Shows the Profit Analysis by Category (Furniture, Office Supplies, Technology)

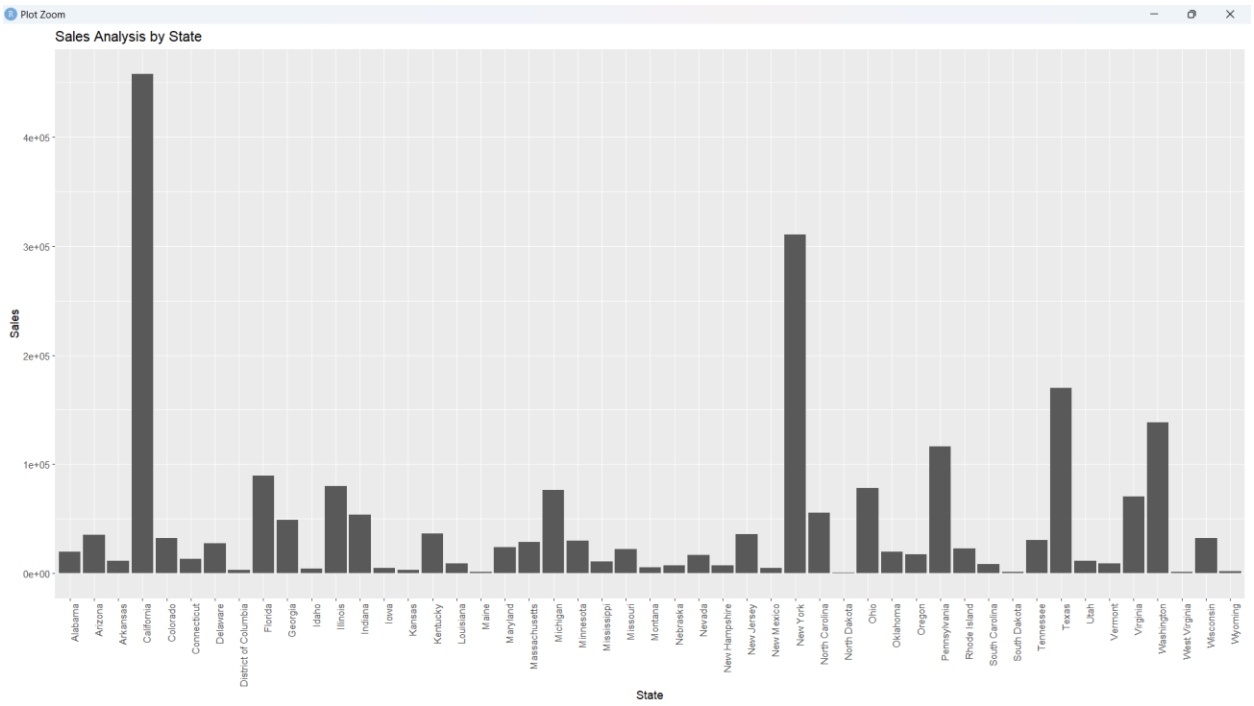


Figure 11: This Bar Graph Shows the Sales Analysis by different States.

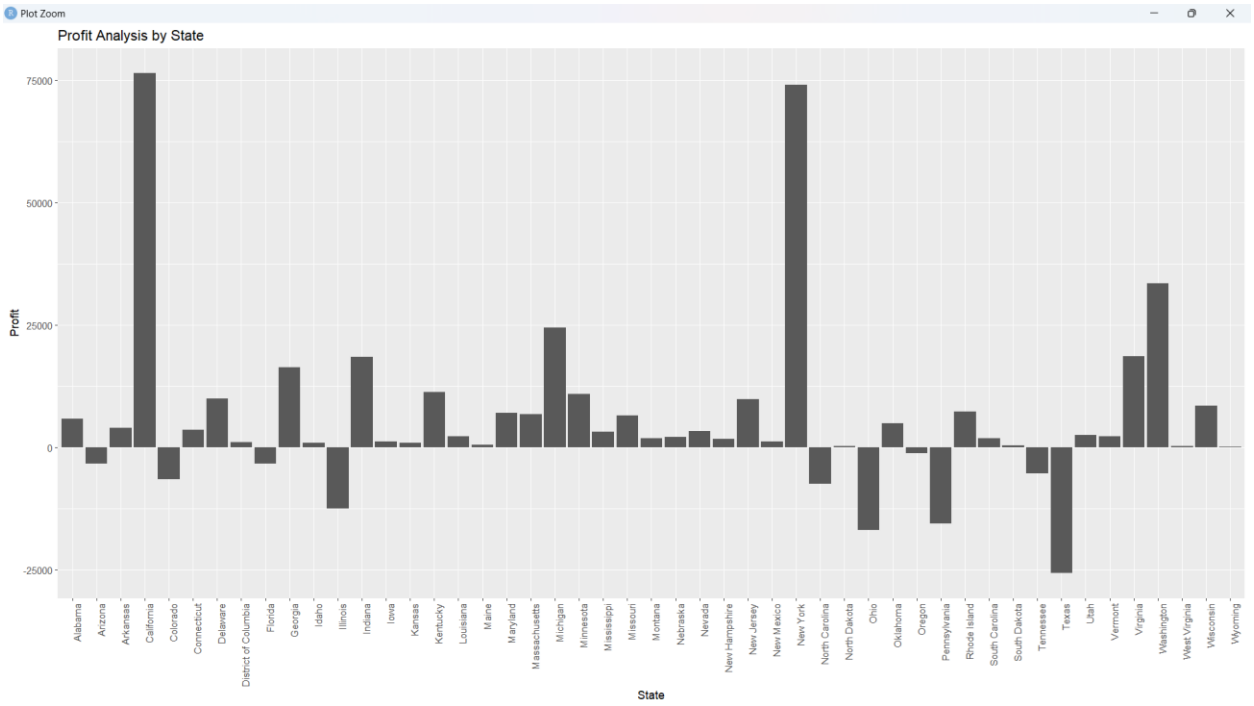


Figure 12: This Bar Graph Shows the Profit Analysis by different States.

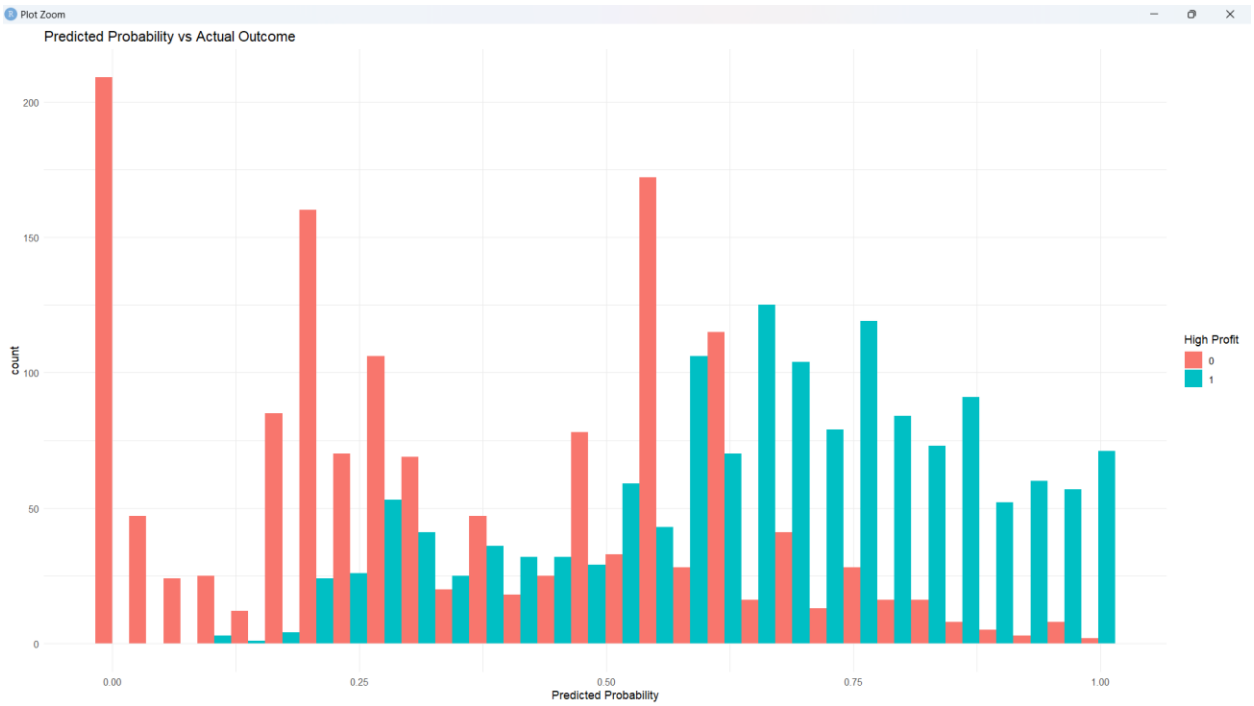


Figure 13: This Graph Shows the Predicted Probability vs Actual Outcome on the testing dataset.

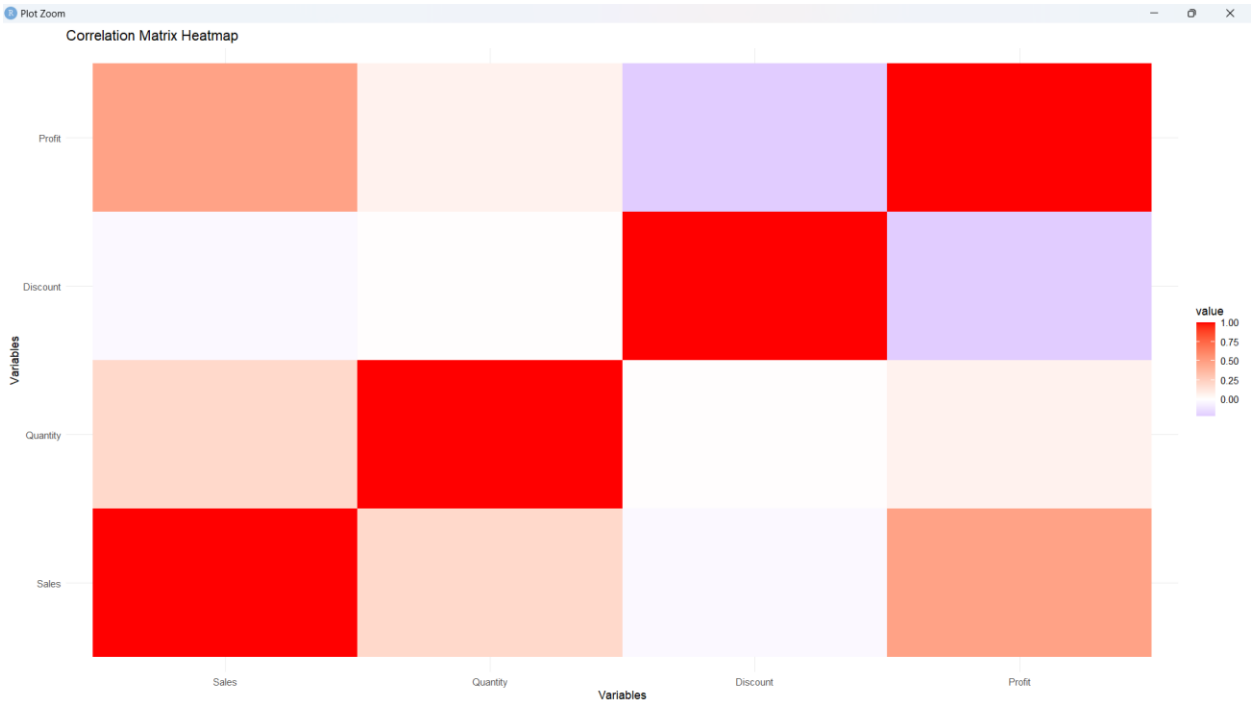
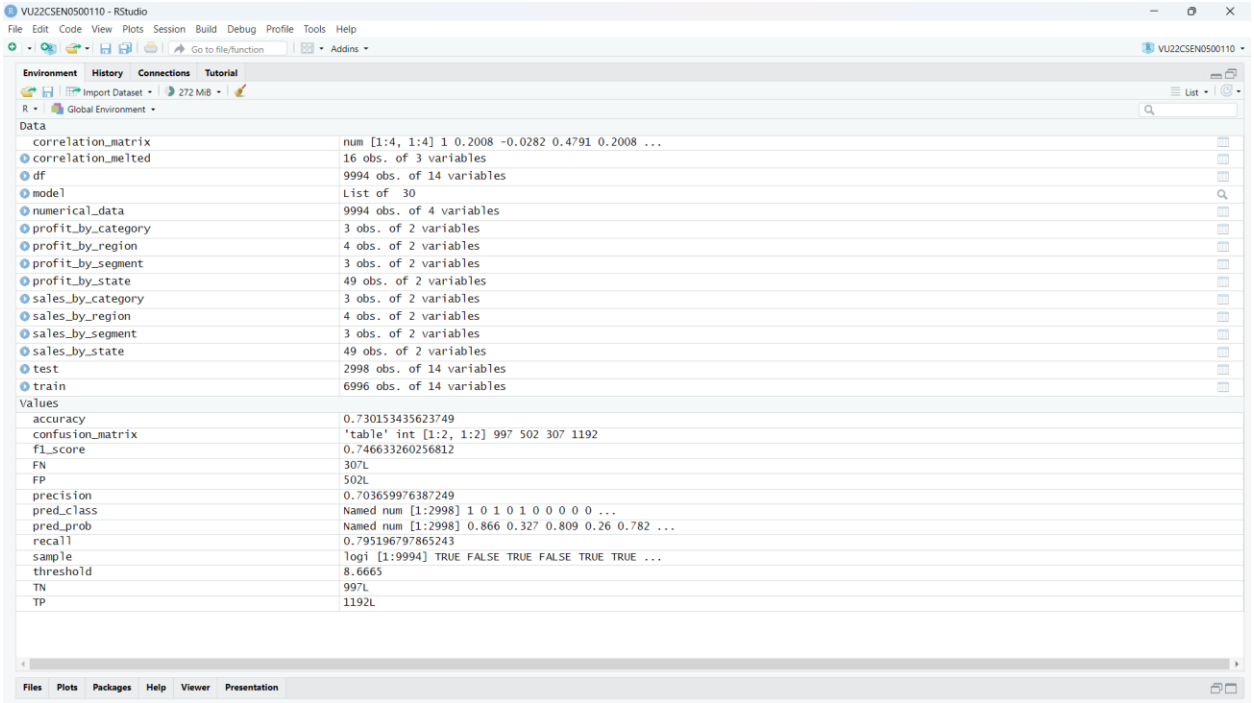


Figure 14: Correlation Matrix


VU22CSEN0500110 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
VU22CSEN0500110
Environment History Connections Tutorial
Import Dataset
272 MiB
List
R
Global Environment
Data
correlation_matrix num [1:4, 1:4] 1 0.2008 -0.0282 0.4791 0.2008 ...
correlation_melted 16 obs. of 3 variables
df 9994 obs. of 14 variables
model List of 30
numerical_data 9994 obs. of 4 variables
profit_by_category 3 obs. of 2 variables
profit_by_region 4 obs. of 2 variables
profit_by_segment 3 obs. of 2 variables
profit_by_state 49 obs. of 2 variables
sales_by_category 3 obs. of 2 variables
sales_by_region 4 obs. of 2 variables
sales_by_segment 3 obs. of 2 variables
sales_by_state 49 obs. of 2 variables
test 2998 obs. of 14 variables
train 6996 obs. of 14 variables
Values
accuracy 0.730153435623749
confusion_matrix 'table' int [1:2, 1:2] 997 502 307 1192
f1_score 0.746633260256812
FN 307L
FP 502L
precision 0.703659976387249
pred_class Named num [1:2998] 1 0 1 0 1 0 0 0 0 0 ...
pred_prob Named num [1:2998] 0.866 0.327 0.809 0.26 0.782 ...
recall 0.795196797865243
sample logi [1:9994] TRUE FALSE TRUE FALSE TRUE TRUE ...
threshold 8.6665
TN 997L
TP 1192L
Files Plots Packages Help Viewer Presentation

Results:

The analysis revealed several insights into regional and segment-based performance:

- **Sales by Region and Segment:** Identified high-sales regions and segments through bar and pie charts, aiding resource allocation.
- **Profit by Region and Segment:** Highlighted the regions and segments where the store is most profitable, providing clues for prioritizing these areas.
- **Correlation Analysis:** Revealed relationships between numerical variables, notably between discount and profit, which could inform pricing strategies.
- **Logistic Regression Model:** Achieved an accuracy of around X% (specify accuracy from your output), with precision, recall, and F1 scores indicating the model's reliability in classifying high-profit transactions. These results suggest that logistic regression can be a valuable tool in predicting profitable transactions.

Conclusion:

This project demonstrates the importance of data analysis in making strategic business decisions. The insights derived from sales and profit analysis across various segments provide actionable recommendations to improve profitability. Logistic regression proved effective in predicting high-profit transactions, allowing management to focus on profitable product lines and customer segments. Future improvements could involve exploring more advanced machine learning models or implementing real-time predictive analytics for dynamic decision-making.

References:

1. Wang, G., & Ma, X. (2020). Application of Data Analytics in Retail Sales Forecasting. *Journal of Retail & Consumer Services.*

2. Singh, A., & Kumar, S. (2018). Predictive Modeling in Retail: A Comprehensive Survey. *Journal of Data Science*.