# 1. Load Required Libraries

if (!require("tidyverse")) install.packages("tidyverse")

if (!require("factoextra")) install.packages("factoextra")

library(tidyverse)

library(ggplot2)

library(scales)

library(factoextra)

# 2. Load and Clean the Data

superstore <- read\_csv("Sample - Superstore.csv")

superstore <- superstore %>%

rename\_all(~make.names(.)) %>%

mutate(

Category = as.factor(Category),

Sub.Category = as.factor(Sub.Category),

Region = as.factor(Region)

)

# 3. Summary Statistics

summary\_stats <- superstore %>%

summarise(

Total\_Sales = sum(Sales),

Total\_Profit = sum(Profit),

Avg\_Profit\_Margin = mean(Profit / Sales, na.rm = TRUE)

)

print(summary\_stats)

# 4. EDA Visualizations

# 4.1 Sales by Region

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

geom\_bar(stat = "summary", fun = sum) +

scale\_y\_continuous(labels = dollar) +

labs(title = "Total Sales by Region", x = "Region", y = "Sales ($)")

# 4.2 Profit by Category

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

geom\_bar(stat = "summary", fun = sum) +

scale\_y\_continuous(labels = dollar) +

labs(title = "Total Profit by Category", x = "Category", y = "Profit ($)")

# 4.3 Profit by Sub-Category

ggplot(superstore, aes(x = reorder(Sub.Category, Profit), y = Profit, fill = Category)) +

geom\_bar(stat = "summary", fun = sum) +

coord\_flip() +

scale\_y\_continuous(labels = dollar) +

labs(title = "Profit by Sub-Category", x = "Sub-Category", y = "Profit ($)")

# 4.4 Sales vs Profit Scatter Plot

ggplot(superstore, aes(x = Sales, y = Profit, color = Category)) +

geom\_point(alpha = 0.6) +

scale\_x\_continuous(labels = dollar) +

scale\_y\_continuous(labels = dollar) +

labs(title = "Sales vs Profit", x = "Sales ($)", y = "Profit ($)")

# 5. Outlier Removal + Boxplots (Profit & Sales)

# 5.1 Function to remove outliers by Category

remove\_outliers\_by\_group <- function(df, group\_col, value\_col) {

df %>%

group\_by(!!sym(group\_col)) %>%

filter({

q1 <- quantile(.data[[value\_col]], 0.25, na.rm = TRUE)

q3 <- quantile(.data[[value\_col]], 0.75, na.rm = TRUE)

iqr <- q3 - q1

lower <- q1 - 1.5 \* iqr

upper <- q3 + 1.5 \* iqr

.data[[value\_col]] >= lower & .data[[value\_col]] <= upper

}) %>%

ungroup()

}

# 5.2 Apply to Profit and Sales

filtered\_profit <- remove\_outliers\_by\_group(superstore, "Category", "Profit")

filtered\_sales <- remove\_outliers\_by\_group(superstore, "Category", "Sales")

# 5.3 Boxplot of Profit (outliers removed)

ggplot(filtered\_profit, aes(x = Category, y = Profit, fill = Category)) +

geom\_boxplot(alpha = 0.7) +

scale\_y\_continuous(labels = dollar) +

labs(title = "Boxplot of Profit by Category (Outliers Removed)", x = "Category", y = "Profit ($)") +

theme\_minimal()

# 5.4 Boxplot of Sales (outliers removed)

ggplot(filtered\_sales, aes(x = Category, y = Sales, fill = Category)) +

geom\_boxplot(alpha = 0.7) +

scale\_y\_continuous(labels = dollar) +

labs(title = "Boxplot of Sales by Category (Outliers Removed)", x = "Category", y = "Sales ($)") +

theme\_minimal()

# 6. Linear Regression

lm\_model <- lm(Profit ~ Sales + Discount, data = superstore)

summary(lm\_model)

# 6.1 Profit vs Discount

ggplot(superstore, aes(x = Discount, y = Profit)) +

geom\_point(alpha = 0.5, color = "steelblue") +

geom\_smooth(method = "lm", se = TRUE, color = "red") +

labs(title = "Linear Regression: Profit vs Discount", x = "Discount", y = "Profit ($)")

# 6.2 Profit vs Sales

ggplot(superstore, aes(x = Sales, y = Profit)) +

geom\_point(alpha = 0.5, color = "darkgreen") +

geom\_smooth(method = "lm", se = TRUE, color = "black") +

labs(title = "Linear Regression: Profit vs Sales", x = "Sales ($)", y = "Profit ($)")

# 7. K-Means Clustering

# 7.1 Select and Scale Data

cluster\_data <- superstore %>%

select(Sales, Profit, Discount) %>%

drop\_na() %>%

filter(Sales < quantile(Sales, 0.99), Profit > quantile(Profit, 0.01))

cluster\_scaled <- scale(cluster\_data)

# 7.2 Apply K-means

set.seed(123)

kmeans\_result <- kmeans(cluster\_scaled, centers = 3, nstart = 25)

cluster\_data$Cluster <- as.factor(kmeans\_result$cluster)

# 7.3 Visualize Clusters

fviz\_cluster(kmeans\_result, data = cluster\_scaled,

geom = "point",

ellipse.type = "convex",

palette = "jco",

ggtheme = theme\_minimal(),

main = "K-means Clustering: Sales, Profit, Discount")