superstore.k-clust

2025-07-03

# Introduction

This project applies **K-Means clustering** to segment customers based on their behavior using the Superstore dataset.  
By grouping similar customers, businesses can target them more effectively, tailor marketing strategies, and increase profitability.

We’ll use R and the following tools: - tidyverse for data wrangling - ggplot2 and factoextra for visualizations - kmeans() for clustering

# Loading and Cleaning the Data

library(janitor)  
  
# Load data  
df <- read\_csv("Sample - Superstore.csv")  
  
# Clean column names  
df <- clean\_names(df)  
  
# Remove duplicates  
df <- df[!duplicated(df), ]  
  
# Convert date columns  
df$order\_date <- as.Date(df$order\_date, format = "%m/%d/%Y")  
df$ship\_date <- as.Date(df$ship\_date, format = "%m/%d/%Y")  
  
# Check structure  
str(df)

#We grouped data by customer to create features like total sales, total profit, average discount, and total number of orders. This allows us to cluster customers based on their overall behavior.

#Scaling the data ```# Remove IDs before scaling scaled\_data <- scale(customer\_data[, -c(1, 2)])

# Check scaled values

head(scaled\_data)

We scale the data to ensure all variables are on the same scale, which is essential for K-Means to work properly.  
  
```set.seed(123)  
kmeans\_result <- kmeans(scaled\_data, centers = 4)  
  
# Add cluster labels  
customer\_data$cluster <- as.factor(kmeans\_result$cluster)

We use kmeans() to segment the customers into 4 clusters. The results are added back to the customer dataset.

#Visualing the data

library(factoextra)  
  
fviz\_cluster(kmeans\_result, data = scaled\_data,  
 geom = "point",  
 palette = "jco",  
 ggtheme = theme\_minimal())

This plot shows how customers are grouped in 2D space. Each color represents a different cluster.

#Cluster Profiles (Summary Table)

customer\_data %>%  
 group\_by(cluster) %>%  
 summarise(  
 avg\_sales = mean(total\_sales),  
 avg\_profit = mean(total\_profit),  
 avg\_discount = mean(avg\_discount),  
 avg\_orders = mean(total\_orders)  
 )

This table helps us understand the behavior of each customer segment.

#Business Insights ## 🧠 Cluster Insights

* **Cluster 1 – High-Value Customers**  
  High sales and profit, low discount usage, frequent orders. Ideal for loyalty campaigns.
* **Cluster 2 – Discount-Driven Loss-Makers**  
  High discount usage, low or negative profit. Review pricing or offer strategy.
* **Cluster 3 – Casual Buyers**  
  Low spend, low order count. Target with reactivation campaigns.
* **Cluster 4 – Reliable Regulars**  
  Medium spend and profit, frequent orders. Potential for upselling.

# Conclusion

Using K-Means Clustering in R, we segmented customers based on behavioral data from the Superstore dataset.  
This approach enables targeted marketing and strategic customer engagement.

Next steps could include: - Creating dashboards in Tableau - Applying more advanced clustering like DBSCAN - Analyzing churn risk by segment