Customer Segmentation

Dataset: Mall Customers dataset (Age, Income, Spending Score).
Tools: R for analysis, Tableau for visualization.
Algorithm: K-means clustering (Unsupervised Machine Learning)
Optimal k Selection: Chose k=6 based on the Elbow Method

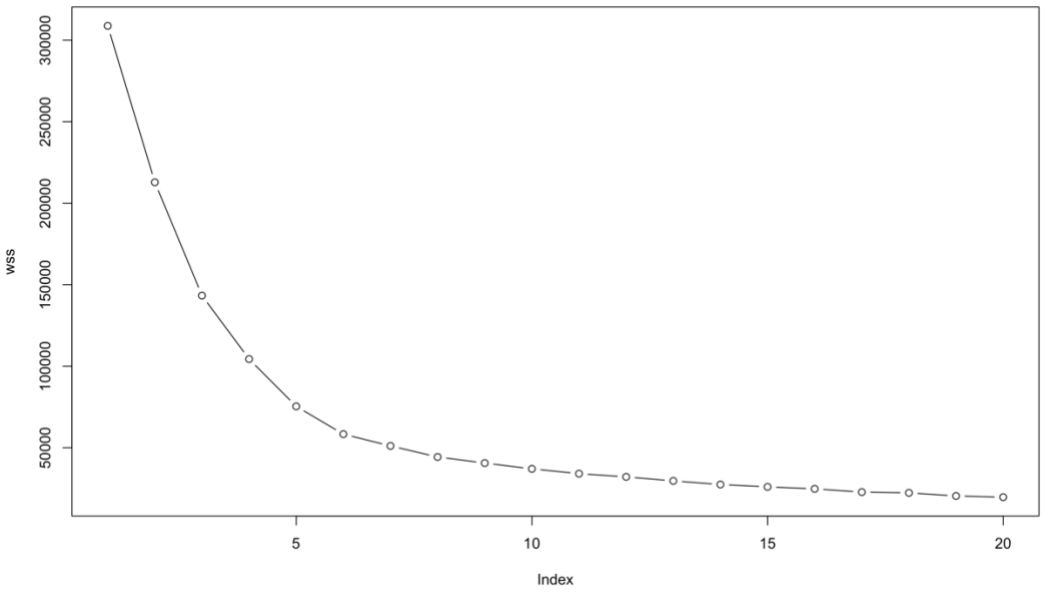
Visualization:
3D Scatter Plots: Generated in R using plotly to visualize clusters across Age, Income, and Spending Score.
Tableau Dashboard: Interactive visualizations of clusters, including histograms, scatter plots, and summary tables.
https://public.tableau.com/shared/HZYRQGT2P?:display_count=n&:origin=viz_share_link

R Code:

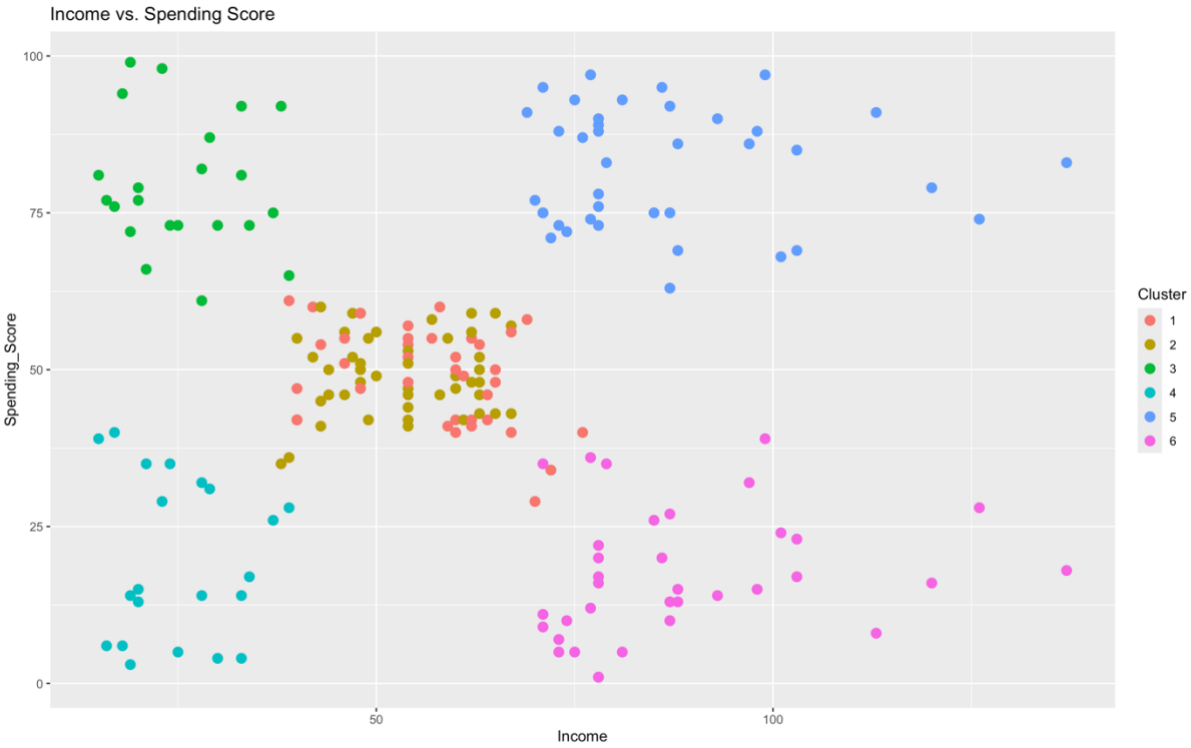
```
# install.packages(c("tidyverse", "ggplot2", "plotly"))

# Load libraries
library(tidyverse)
library(ggplot2)
library(plotly)

# Load data
customer_data <- read.csv("Mall_Customers.csv")

# Preprocess data
df <- customer_data[, c(3,4,5)]
colnames(df) <- c("Age", "Income", "Spending_Score")


# Elbow method for optimal clusters
set.seed(123)
KM<- kmeans(df,4,nstart = 50)

wss<- nrow(df-1)*sum(apply(df,2,var))
for (i in 1:20) {
  wss[i] <- kmeans(df, centers = i, nstart = 50)$tot.withinss}

#create a plot
plot(wss,type = "b")


# Apply K-means (k=6)
set.seed(123)
```

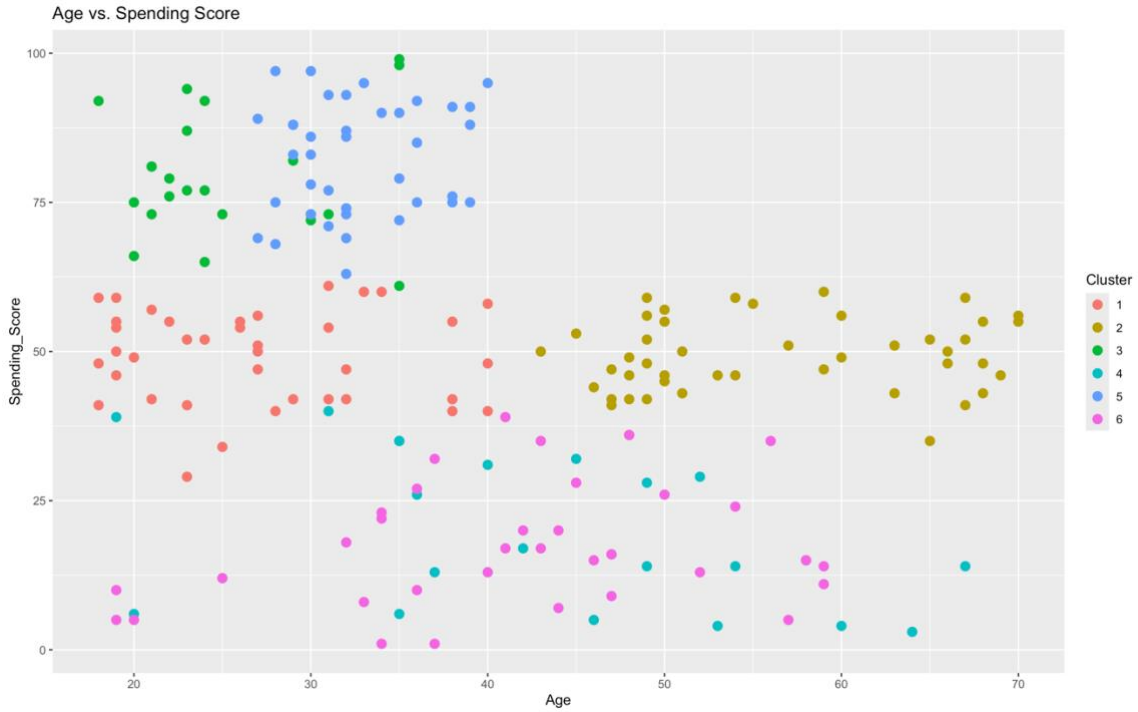
```
KM_Model <- kmeans(df, centers = 6, nstart = 50)
df$Cluster <- as.factor(KM_Model$cluster)

# 3D visualization
plot_ly(df,
        x = ~Age,
        y = ~Income,
        z = ~Spending_Score,
        color = ~Cluster,
        colors = c("blue", "green", "red", "purple", "orange"),
        type = "scatter3d",
        mode = "markers",
        marker = list(size = 5)) %>%
  layout(title = "Customer Segmentation (3D Clusters)")

# Pairwise analysis
ggplot(df, aes(x = Income, y = Spending_Score, color = Cluster)) +
  geom_point(size = 3) +
  labs(title = "Income vs. Spending Score")

ggplot(df, aes(x = Age, y = Spending_Score, color = Cluster)) +
  geom_point(size = 3) +
  labs(title = "Age vs. Spending Score")

# Cluster summary
cluster_summary <- df %>%
  group_by(Cluster) %>%
  summarise(
    Avg_Age = mean(Age),
    Avg_Income = mean(Income),
    Avg_Spending = mean(Spending_Score),
    Count = n())

View(cluster_summary)

# Save results
write.csv(df, "segmented_customers.csv", row.names = FALSE)
```

Visualization in R




```
```


Income vs. Spending Score
Spending_Score
100
75
50
25
0
50
100
Income
Cluster
1
2
3
4
5
6

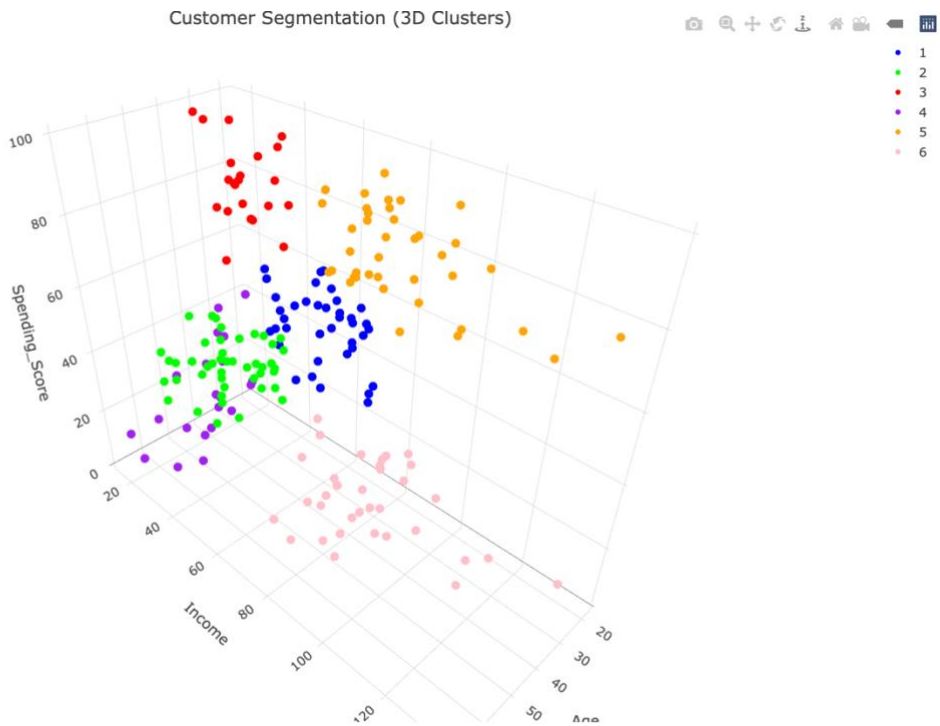


cluster_summary

Filter

	Cluster	Avg_Age	Avg_Income	Avg_Spending	Count
1	1	27.00000	56.65789	49.13158	38
2	2	56.15556	53.37778	49.08889	45
3	3	25.27273	25.72727	79.36364	22
4	4	44.14286	25.14286	19.52381	21
5	5	32.69231	86.53846	82.12821	39
6	6	41.68571	88.22857	17.28571	35

Showing 1 to 6 of 6 entries, 5 total columns



Key Results

Identified 6 Clusters:
Cluster 1: Moderate income, moderate spending
Cluster 3: Young, low income, high spending
Cluster 6: High income, very low spending

Actionable Insights:
High-Income, Low-Spenders : Ideal for retention campaigns (e.g., loyalty programs).
Young, High-Spenders : Target with trendy or budget-friendly promotions.