K-MEANS CLUSTERING

23CSEG28

```
# Load necessary libraries
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.3.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.3

library(cluster)

## Warning: package 'cluster' was built under R version 4.3.3

library(factoextra)

## Warning: package 'factoextra' was built under R version 4.3.3

## Welcome! Want to learn more? See two factoextra-related books at https://g
oo.gl/ve3WBa

# Read the data
data <- read.csv("C:/Users/ADMIN/Downloads/Mall_Customers.csv")
summary(data)

##    CustomerID        Gender                Age         Annual.Income..k..
##  Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00
##  1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50
##  Median :100.50   Mode  :character   Median :36.00   Median : 61.50
##  Mean   :100.50                      Mean   :38.85   Mean   : 60.56
##  3rd Qu.:150.25                      3rd Qu.:49.00   3rd Qu.: 78.00
##  Max.   :200.00                      Max.   :70.00   Max.   :137.00
##  Spending.Score..1.100.
##  Min.   : 1.00
##  1st Qu.:34.75
##  Median :50.00
##  Mean   :50.20
##  3rd Qu.:73.00
##  Max.   :99.00
```

```
sum(is.na(data))
```

```
## [1] 0
```

```
# EDA
df <- select_if(data, is.numeric)


# Subsetting based on Age
young_adult <- df[df$Age <= 30, ]
middleage_adult <- df[df$Age > 30 & df$Age <= 55, ]
older_age <- df[df$Age > 55, ]

# Combine the subsetted data into one dataframe
subset1 <- bind_rows(
  mutate(young_adult, Age_Group = "Young Adult"),
  mutate(middleage_adult, Age_Group = "Middle Age Adult"),
  mutate(older_age, Age_Group = "Older Adult")
)
```

```
#univariate analysis
# Age Distribution of young adult
hist(young_adult$Age, main = "Age Distribution of young adult", xlab = "Age")
```

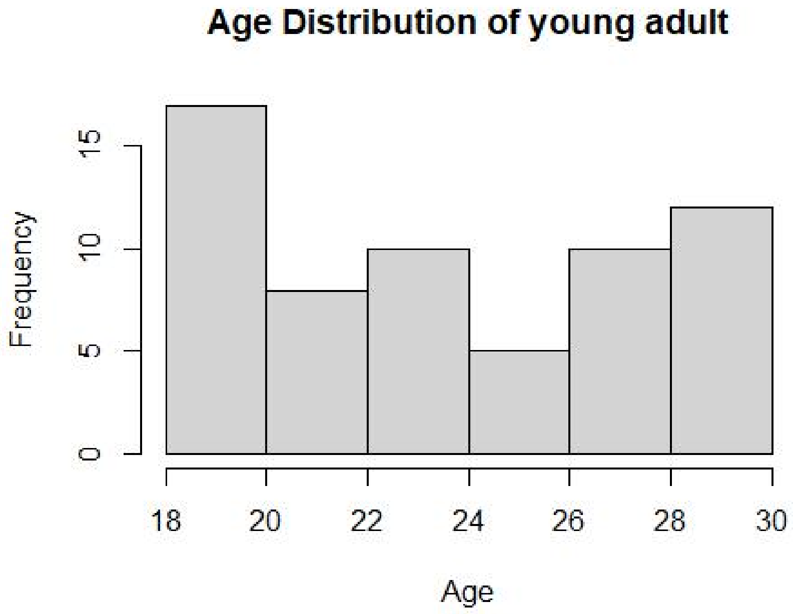

Fig 1.1

```
#Age Distribution of middle age adult

hist(middleage_adult$Age , main = "Age Distribution of middle age adult", xlab = "Age")
```

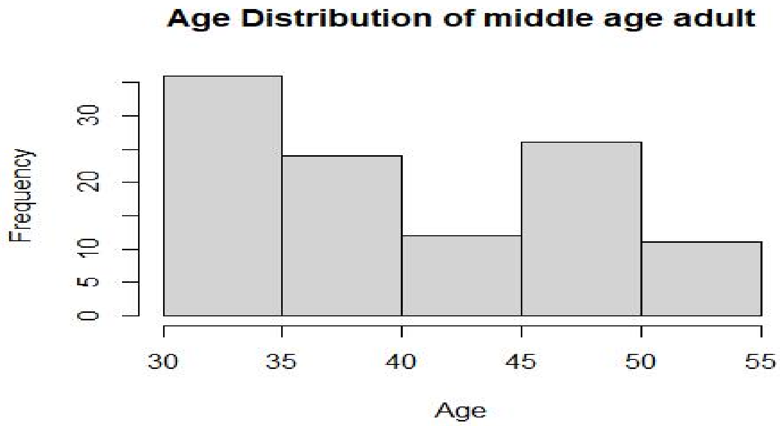

Fig 1.2

```
#Age Distribution of older age

hist(older_age$Age , main = "Age Distribution of older age", xlab = "Age")
```

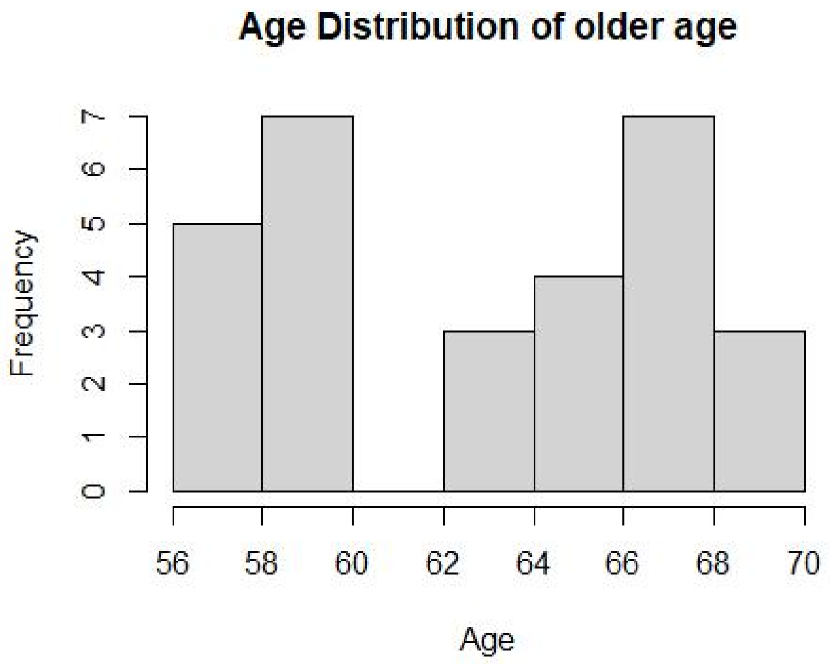

Fig 1.3

```
#subsetting based on spending score
low_spending <- subset(data, `Spending.Score..1.100.` <= 34.75)
average_spending <- subset(data, `Spending.Score..1.100.` > 34.75 & `Spending.Score..1.100.` <= 73)
high_spending <- subset(data, `Spending.Score..1.100.` > 73)

# Combine the low, average, and high spending subsets into one dataframe
subset2 <- bind_rows(
  mutate(low_spending, Spending_Group = "Low Spending"),
  mutate(average_spending, Spending_Group = "Average Spending"),
  mutate(high_spending, Spending_Group = "High Spending")
)
```

```r
#histogram for spending scores
par(mfrow=c(1,3))
hist(low_spending$`Spending.Score..1.100.`, main = "Low Spending Score", xlab
 = "Spending Score")
hist(average_spending$`Spending.Score..1.100.`, main = "Average Spending Scor
e", xlab = "Spending Score")
hist(high_spending$`Spending.Score..1.100.`, main = "High Spending Score", xl
ab = "Spending Score")
```

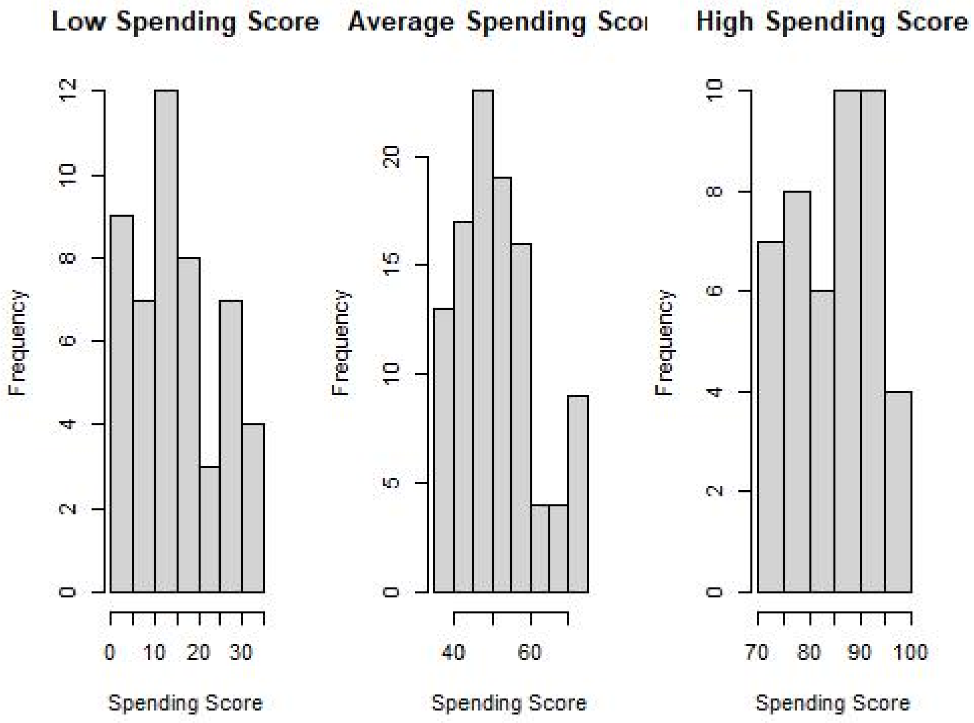

Fig 1.4

```r
# Box plot for spending score in each category
par(mfrow=c(1,3))
boxplot(low_spending$`Spending.Score..1.100.`, main = "Low Spending Score", ylab = "Spending Score")
boxplot(average_spending$`Spending.Score..1.100.`, main = "Average Spending Score", ylab = "Spending Score")
boxplot(high_spending$`Spending.Score..1.100.`, main = "High Spending Score", ylab = "Spending Score")
```

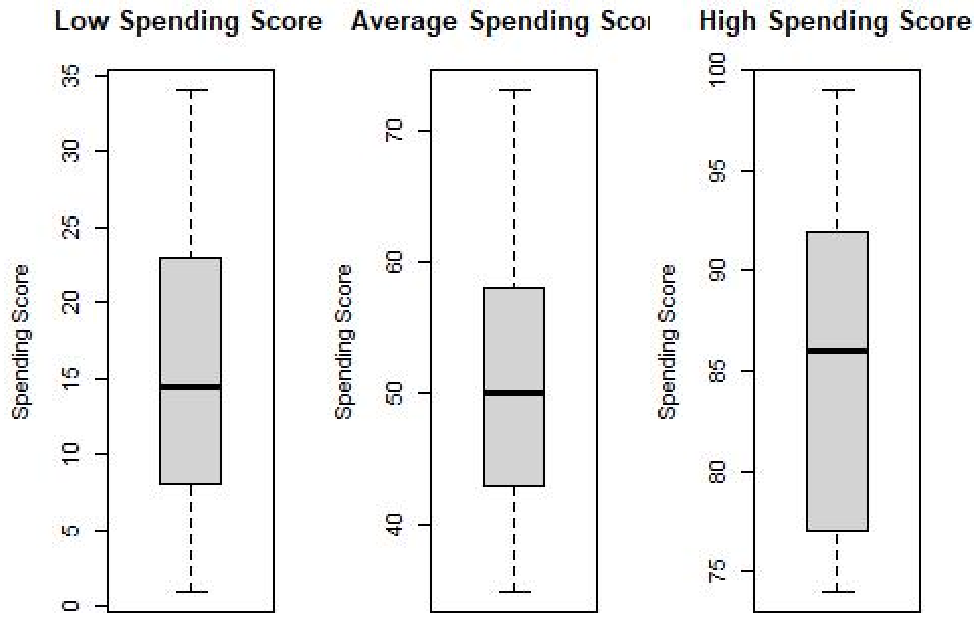

Fig 1.5

```
# Scatter plot of Annual Income vs Spending Score in each category
par(mfrow=c(1,3))
plot(low_spending$`Annual.Income..k..`, low_spending$`Spending.Score..1.100.`,
 main = "Low Spending Score",
     xlab = "Annual Income", ylab = "Spending Score", col = "blue")
plot(average_spending$`Annual.Income..k..`, average_spending$`Spending.Score..
1.100.`, main = "Average Spending Score",
     xlab = "Annual Income", ylab = "Spending Score", col = "green")
plot(high_spending$`Annual.Income..k..`, high_spending$`Spending.Score..1.100.
`, main = "High Spending Score",
     xlab = "Annual Income", ylab = "Spending Score", col = "red")
```

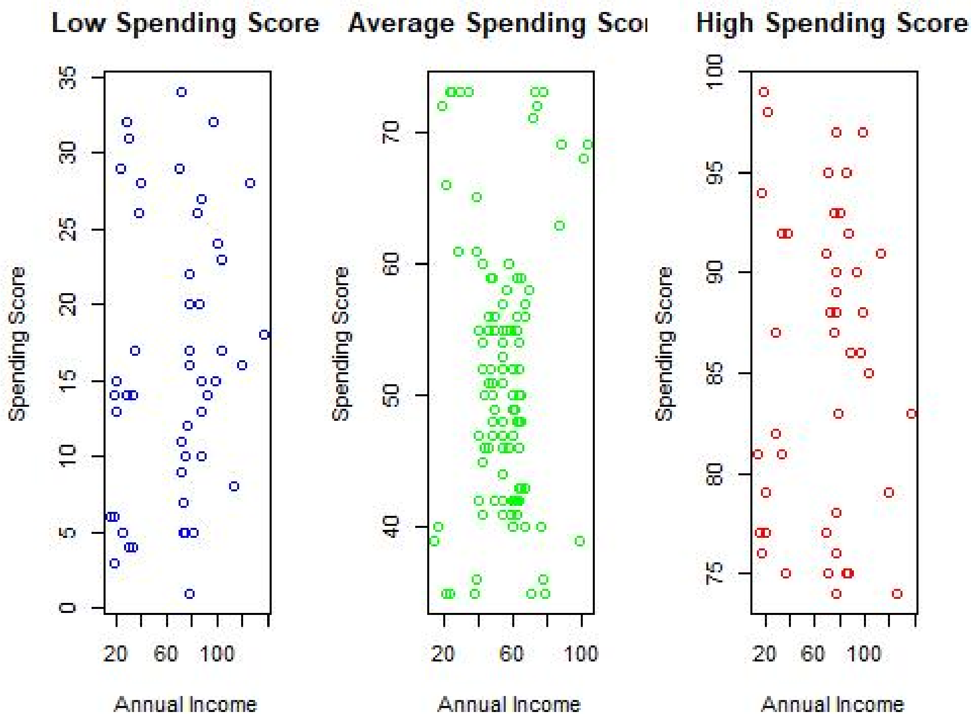

Fig 1.6

```
#subsetting based on anuual income
low_income <- subset(data, `Annual.Income..k..` <= 41.50)
average_income <- subset(data, `Annual.Income..k..` > 41.50 & `Annual.Income..
k..` <= 78.00)
high_income <- subset(data, `Annual.Income..k..` > 78.00)

# Combine the low, average, and high income subsets into one dataframe
subset3 <- bind_rows(
  mutate(low_income, Income_Group = "Low Income"),
  mutate(average_income, Income_Group = "Average Income"),
  mutate(high_income, Income_Group = "High Income")
)
```

```r
# Histogram for spending score distribution in each category
par(mfrow=c(1,3))
hist(low_income$`Spending.Score..1.100.`, main = "Low Income", xlab = "Spending Score")
hist(average_income$`Spending.Score..1.100.`, main = "Average Income", xlab = "Spending Score")
hist(high_income$`Spending.Score..1.100.`, main = "High Income", xlab = "Spending Score")
```

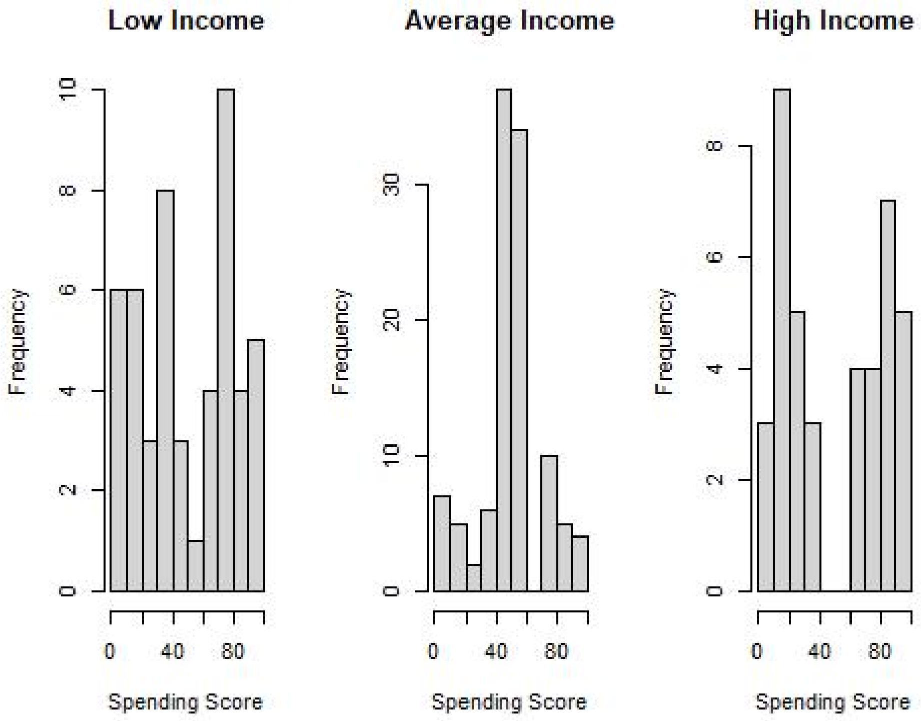

Fig 1.7

```r
# Box plot for spending score in each category
par(mfrow=c(1,3))
boxplot(low_income$`Spending.Score..1.100.`, main = "Low Income", ylab = "Spending Score")
boxplot(average_income$`Spending.Score..1.100.`, main = "Average Income", ylab = "Spending Score")
boxplot(high_income$`Spending.Score..1.100.`, main = "High Income", ylab = "Spending Score")
```

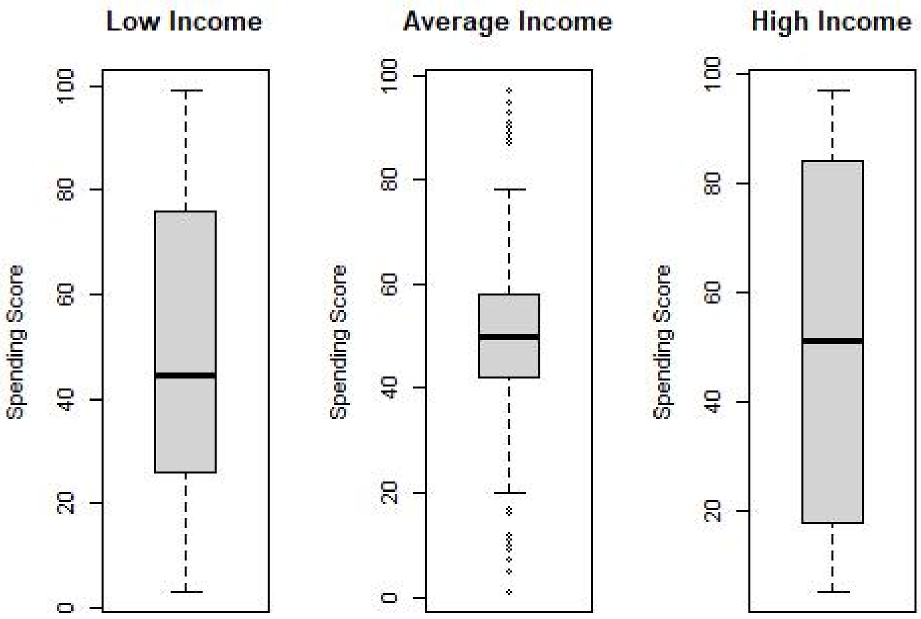

Fig 1.8

```
# Scatter plot of Annual Income vs Spending Score in each category
par(mfrow=c(1,3))
plot(low_income$`Annual.Income..k..`, low_income$`Spending.Score..1.100.`, ma
in = "Low Income",
     xlab = "Annual Income", ylab = "Spending Score", col = "blue")
plot(average_income$`Annual.Income..k..`, average_income$`Spending.Score..1.1
00.`, main = "Average Income",
     xlab = "Annual Income", ylab = "Spending Score", col = "green")
plot(high_income$`Annual.Income..k..`, high_income$`Spending.Score..1.100.`,
main = "High Income",
     xlab = "Annual Income", ylab = "Spending Score", col = "red")
```

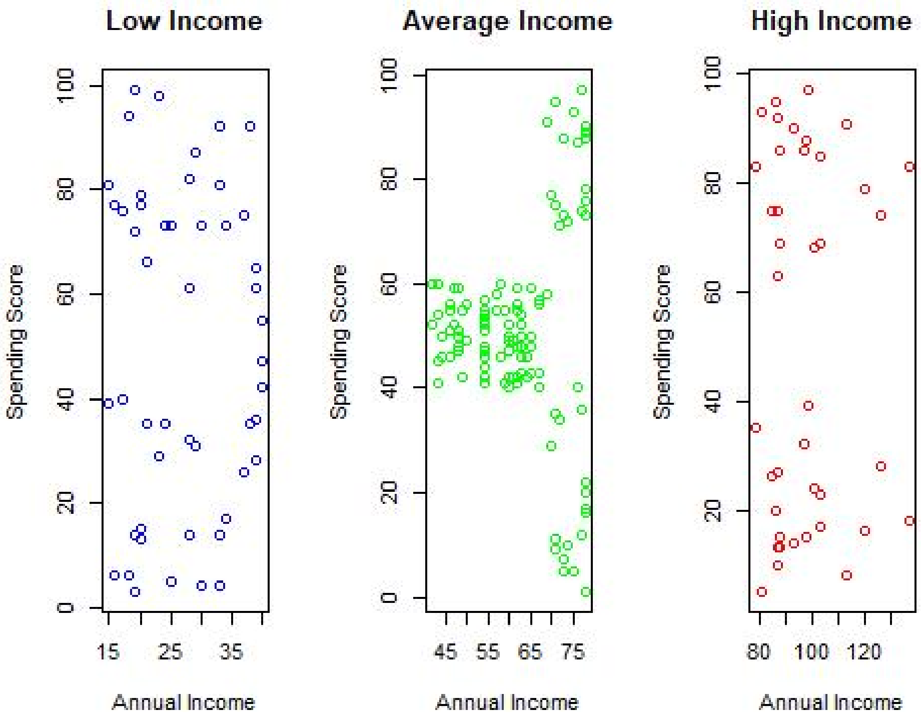

Fig 1.9

```
# Combine the subsetted data into one dataframe
subset1 <- bind_rows(
  mutate(young_adult, Age_Group = "Young Adult"),
  mutate(middleage_adult, Age_Group = "Middle Age Adult"),
  mutate(older_age, Age_Group = "Older Adult")
)

# Load necessary libraries
library(cluster)
library(factoextra)

# Function to calculate within-cluster sum of squares (WCSS)
calculate_wcss <- function(data, k_max = 10) {
```

```r
  wcss <- numeric(k_max)
  for (i in 1:k_max) {
    kmeans_result <- kmeans(data, centers = i, nstart = 25)
    wcss[i] <- kmeans_result$tot.withinss
  }
  return(wcss)
}

# Function to plot the elbow curve
plot_elbow_curve <- function(wcss, k_max = 10) {
  plot(1:k_max, wcss, type = "b", xlab = "Number of Clusters", ylab = "Within
-cluster Sum of Squares",
       main = "Elbow Curve for Optimal Number of Clusters")
}

# Perform K-means clustering on subset1
wcss1 <- calculate_wcss(subset1[, c("Age", "Annual.Income..k..", "Spending.Sc
ore..1.100.")])
plot_elbow_curve(wcss1)
```

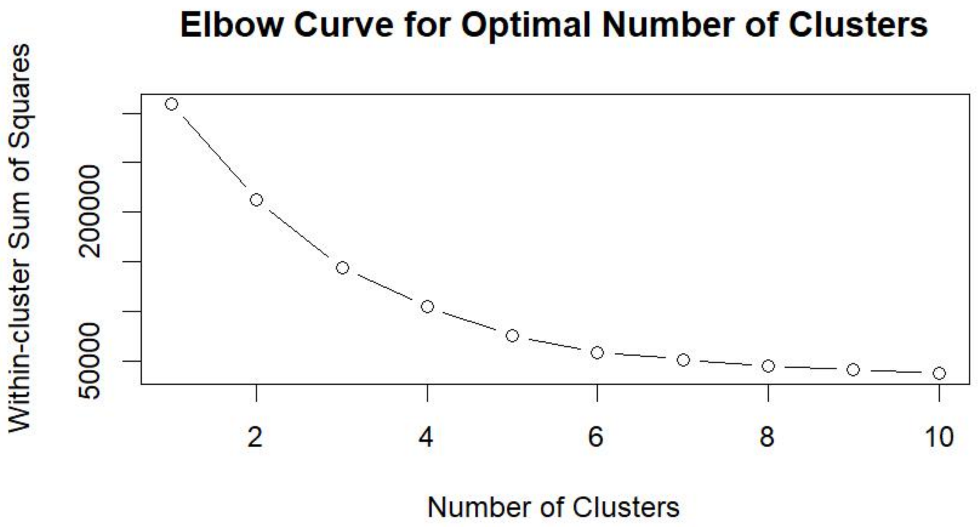

Fig 1.10

```r
kmeans_result1 <- kmeans(subset1[, c("Age", "Annual.Income..k..", "Spending.S
core..1.100.")], centers = 3, nstart = 25)

# Visualize the clusters for subset1
fviz_cluster(kmeans_result1, data = subset1[, c("Age", "Annual.Income..k..",
"Spending.Score..1.100.")], geom = "point")
```

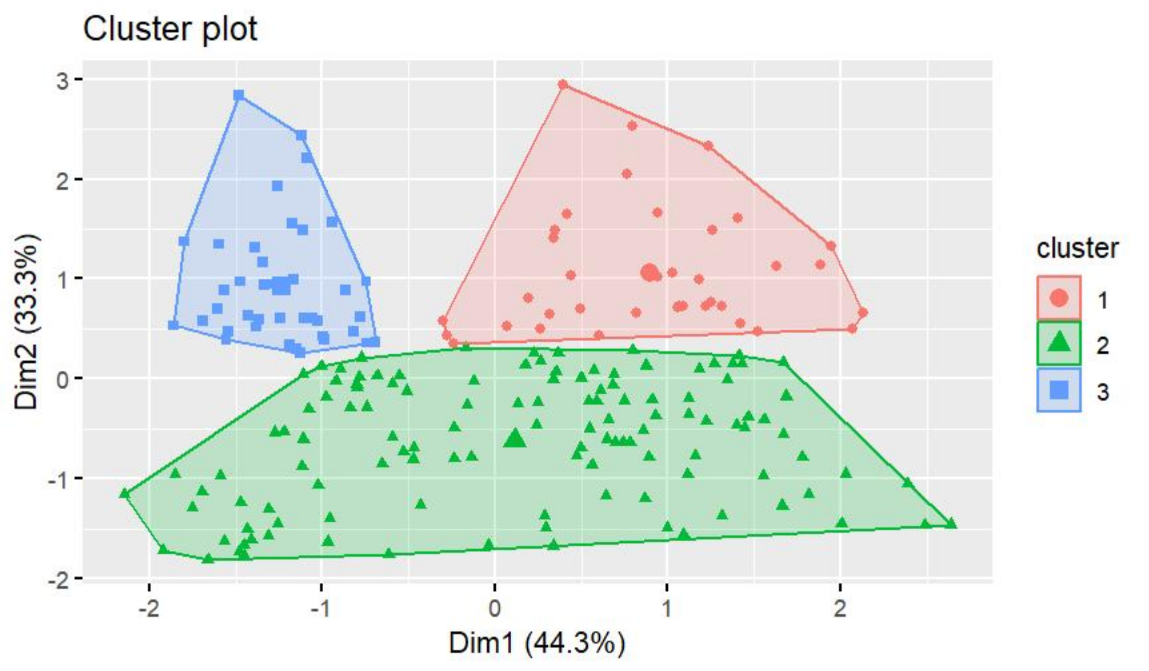

Fig 1.11

```r
# Calculate silhouette index for subset1
sil1 <- silhouette(kmeans_result1$cluster, dist(subset1[, c("Age", "Annual.In
come..k..", "Spending.Score..1.100.")]))

# Visualize silhouette index for subset1
fviz_silhouette(sil1)

##   cluster size ave.sil.width
## 1       1   39          0.60
## 2       2   38          0.4




                            0
## 3       3  123          0.58
```

```
# Perform K-means clustering on subset2
wcss2 <- calculate_wcss(subset2[, c("Age", "Annual.Income..k..", "Spending.Sc
ore..1.100.")])
kmeans_result2 <- kmeans(subset2[, c("Age", "Annual.Income..k..", "Spending.S
core..1.100.")], centers = 3, nstart = 25)

# Visualize the clusters for subset2
fviz_cluster(kmeans_result2, data = subset2[, c("Age", "Annual.Income..k..",
"Spending.Score..1.100.")], geom = "point")
```

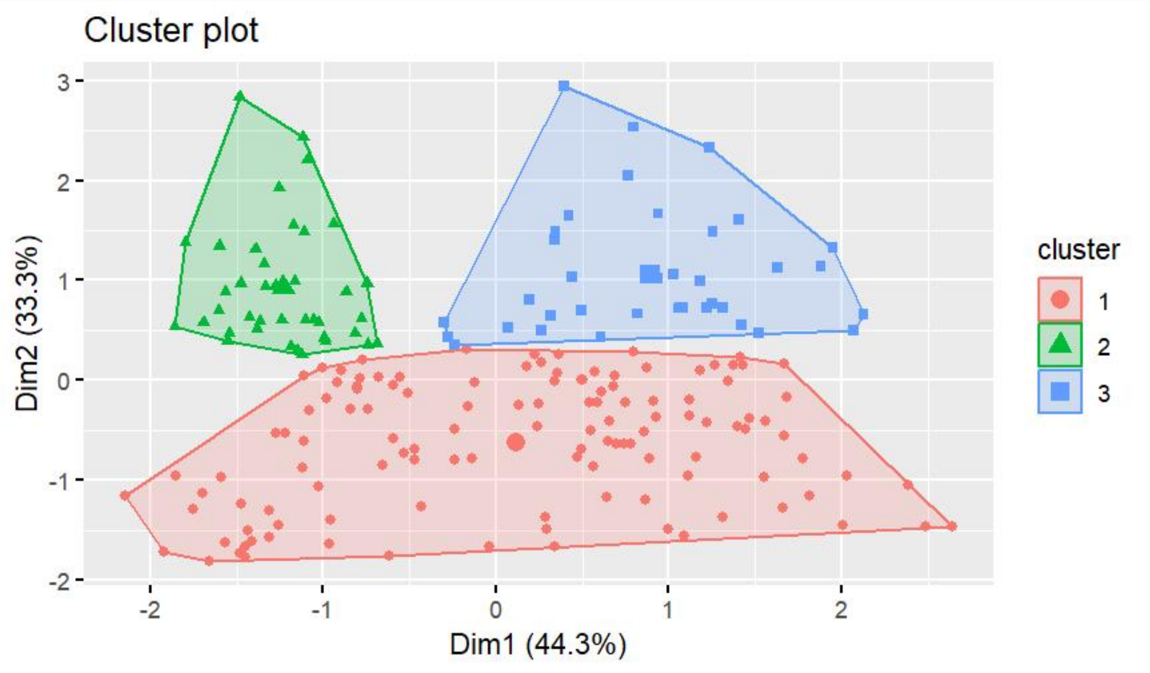

Fig 1.12

```
# Calculate silhouette index for subset2
sil2 <- silhouette(kmeans_result2$cluster, dist(subset2[, c("Age", "Annual.In
come..k..", "Spending.Score..1.100.")]))

# Visualize silhouette index for subset2
fviz_silhouette(sil2)

##   cluster size ave.sil.width
## 1       1   39          0.60
## 2       2  123          0.28
## 3       3   38          0.56
```

```r
# Perform K-means clustering on subset3
wcss3 <- calculate_wcss(subset3[, c("Age", "Annual.Income..k..", "Spending.Score..1.100.")])
kmeans_result3 <- kmeans(subset3[, c("Age", "Annual.Income..k..", "Spending.Score..1.100.")], centers = 3, nstart = 25)

# Visualize the clusters for subset3
fviz_cluster(kmeans_result3, data = subset3[, c("Age", "Annual.Income..k..",
"Spending.Score..1.100.")], geom = "point")
```

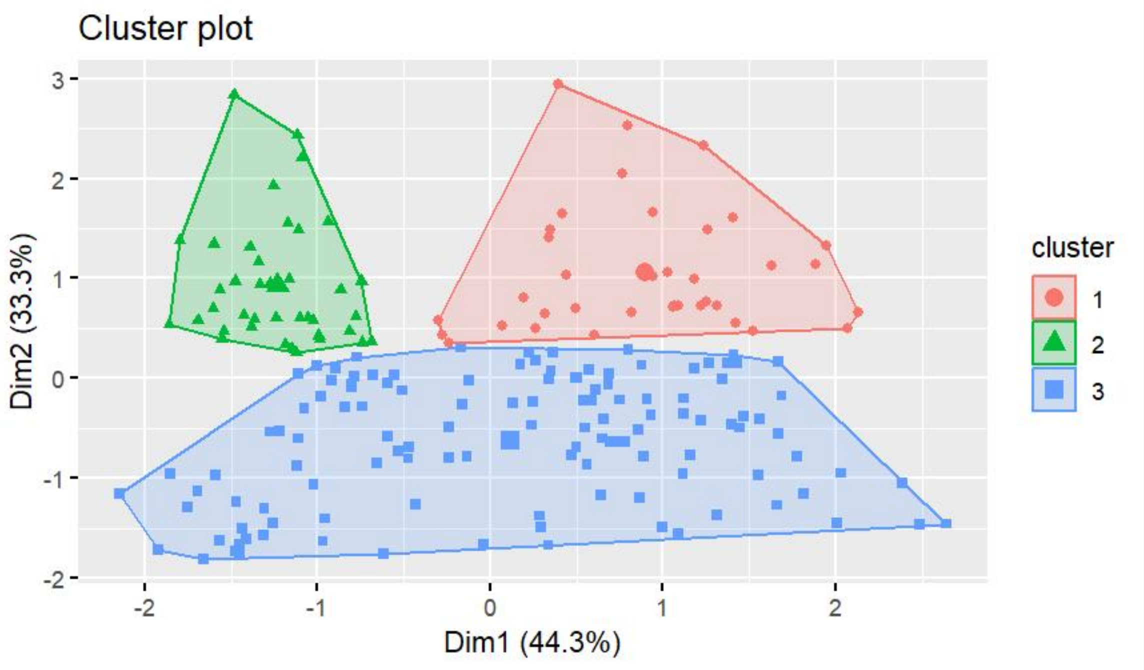

Fig 1.13

```r
# Calculate silhouette index for subset3
sil3 <- silhouette(kmeans_result3$cluster, dist(subset3[, c("Age", "Annual.Income..k..", "Spending.Score..1.100.")]))

# Visualize silhouette index for subset3
fviz_silhouette(sil3)
```

```
##   cluster size ave.sil.width
## 1       1   39          0.60
## 2       2   38          0.50
## 3       3  123          0.58
```