Chapter 20: KMeans

Exercise 2: Shopping

Yêu cầu: Thực hiện Kmeans để phân cụm dữ liệu theo yêu cầu sau:

- Cho dữ liệu shopping_data.csv
- Tạo data.frame với 2 cột: Annual Income (k$) và Spending Score (1-100)
- Trực quan hóa dữ liệu
- Áp dụng Elbow tìm k
- Áp dụng thuật toán K-Means để giải bài toán phân cụm theo K
- Vẽ hình, xem kết quả
- Nhận xét kết quả

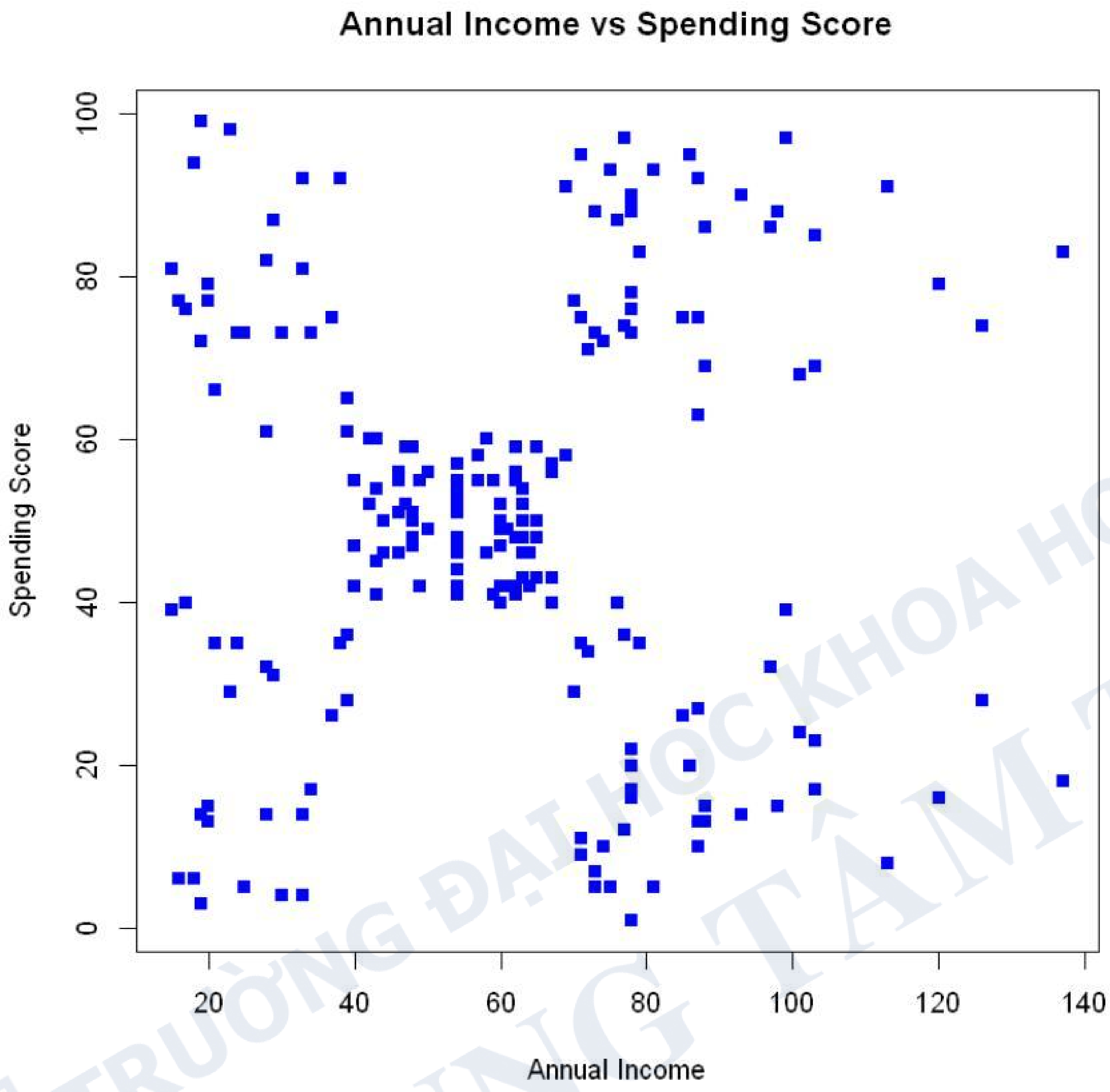
In [1]:

```
data <- read.csv("shopping_data.csv")
print(head(data))
print(is.data.frame(data))
print(ncol(data))
print(nrow(data))
```

```
  CustomerID  Genre Age Annual.Income..k.. Spending.Score..1.100.
1          1   Male  19                 15                     39
2          2   Male  21                 15                     81
3          3 Female  20                 16                      6
4          4 Female  23                 16                     77
5          5 Female  31                 17                     40
6          6 Female  22                 17                     76
[1] TRUE
[1] 5
[1] 200
```

In [2]:

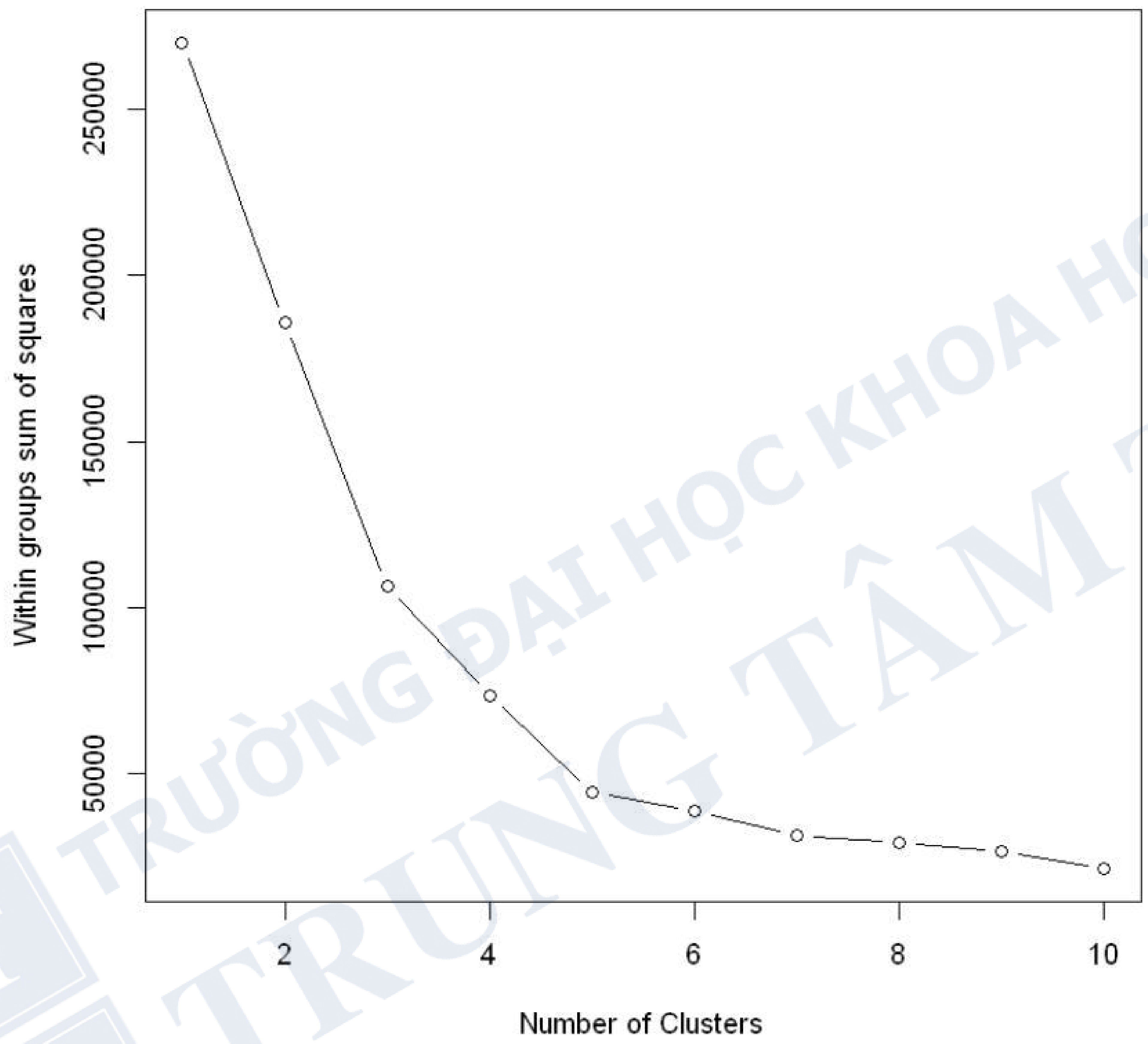
```
# Plot the chart
plot(x = data$Annual.Income..k..,y = data$Spending.Score..1.100.,
     xlab = "Annual Income",
     ylab = "Spending Score",
     main = "Annual Income vs Spending Score",
     pch = 15, col = "blue"
)
```



In [3]:

```
# finding k
# Determine number of clusters
mydata<-data
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 1:10) wss[i] <- sum(kmeans(mydata[, 4:5],
                                     centers=i)$withinss)
plot(1:10, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares")
```

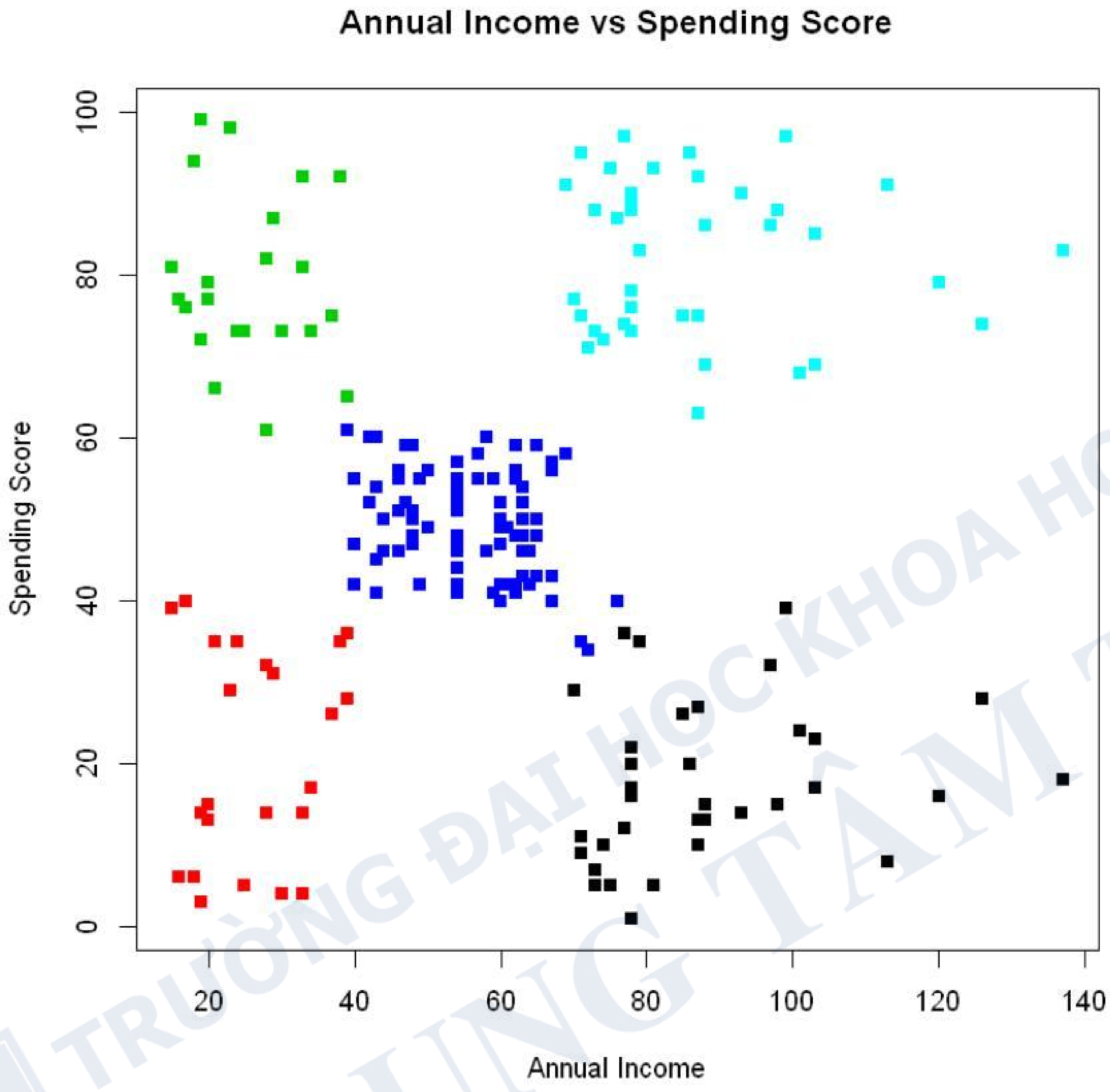
```
Warning message in FUN(newX[, i], ...):
"NAs introduced by coercion"
```



In [4]:
```
# clustering
set.seed(20)
dataCluster <- kmeans(mydata[, 4:5], 5, nstart = 20)
dataCluster
```

```
K-means clustering with 5 clusters of sizes 35, 23, 22, 81, 39

Cluster means:
  Annual.Income..k.. Spending.Score..1.100.
1           88.20000               17.11429
2           26.30435               20.91304
3           25.72727               79.36364
4           55.29630               49.51852
5           86.53846               82.12821

Clustering vector:
  [1] 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2
 [38] 3 2 3 2 3 2 4 2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
 [75] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[112] 4 4 4 4 4 4 4 4 4 4 4 4 5 1 5 4 5 1 5 1 5 4 5 1 5 1 5 1 5 1 5 4 5 1 5 1 5
[149] 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1
[186] 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5

Within cluster sum of squares by cluster:
[1] 12511.143  5098.696  3519.455  9875.111 13444.051
 (between_SS / total_SS =  83.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

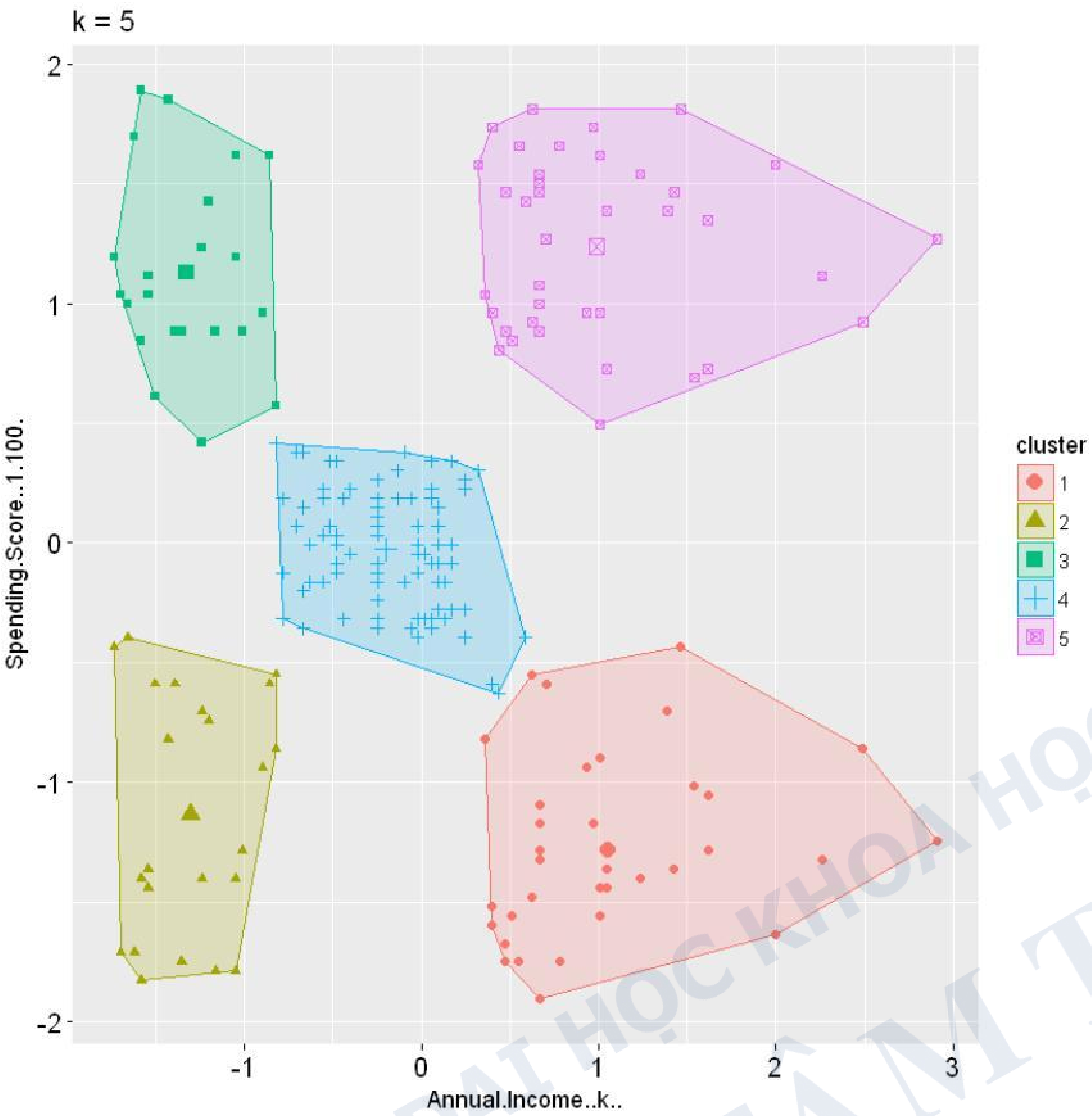
In [5]:
```
# Plot the chart
dataCluster$cluster <- as.factor(dataCluster$cluster)
plot(x = mydata$Annual.Income..k..,y = mydata$Spending.Score..1.100.,
     xlab = "Annual Income",
     ylab = "Spending Score",
     main = "Annual Income vs Spending Score", col = dataCluster$cluster,
     pch = 15
)
```



In [6]:
```
library(factoextra)
```

```
Loading required package: ggplot2
```

In [7]:
```
fviz_cluster(dataCluster, geom = "point", data = mydata[, 4:5]) +
ggtitle("k = 5")
```



In [ ]: