ANA 515 Assignment 4

Siliang Gong

Dec 11, 2021

Introduction

In this assignment, I will work on the Mall customers dataset. It contains information about people visiting the mall, including gender, age, annual income of each customer. The purpose of the data analysis is to segmentate customers based on the age, gender, interests. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base.

Data Import and Description

The data set is available from https://www.kaggle.com/shwetabh123/mall-customers.

First, I read the data into R using the following code:

```
# read the data into a dataframe

customer_data <- read.csv("Mall_Customers.csv")

# A glimpse of data

head(customer_data, 5)
```

```
##   CustomerID Gender 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
```

The customer segmentation data has 200 rows and 5. There are 5 variables in the data set, with variable names as below:

```
# Variable names

colnames(customer_data)
```

```
## [1] "CustomerID"             "Gender"
## [3] "Age"                    "Annual.Income..k.."
## [5] "Spending.Score..1.100."
```

The summary statistics of `Age`, `Gender`, `Annual.Income..k..` and `Spending.Score..1.100.` are:

```
# Summary statistics

summary(customer_data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.75   36.00   38.85   49.00   70.00
```

```r
summary(customer_data$Gender)
```

```
## Female   Male
##    112     88
```

```r
summary(customer_data$Annual.Income..k..)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   41.50   61.50   60.56   78.00  137.00
```

```r
summary(customer_data$Spending.Score..1.100.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   34.75   50.00   50.20   73.00   99.00
```

Data Preparation

The data set is clean and there are no missing observations or errors. It is possible that the variable types for `Age`, `Annual.Income..k..` and `Spending.Score..1.100.` have been converted to numeric (interger) from character. The values for these three variables should be positive. If there are any negative observations, then it is possible that recording errors occur during data collection.

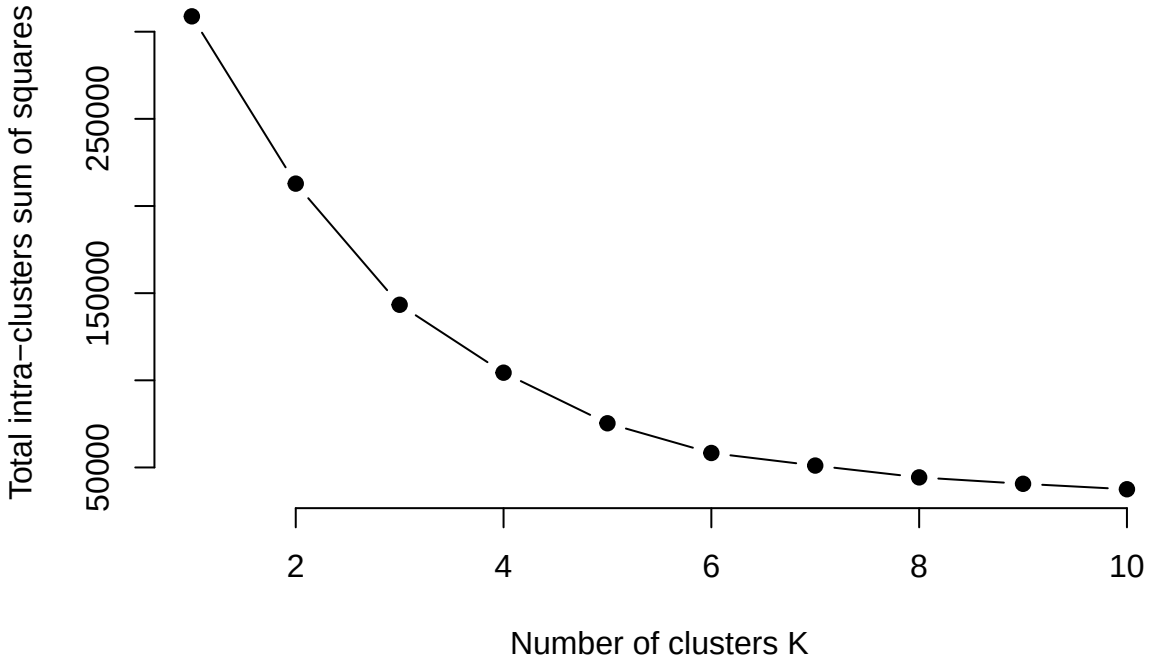
Data Analysis: Modeling and Outputs

K-means algorithm could be applied to segmentate customers into clusters. Suppose there are k clusters, the algorithm starts by selecting k observations randomly from the sample to be the initial center for the clusters. Then the remaining observations are assigned to the closest center, where the Euclidean Distance is used as the metric. When the assignment is complete, the new mean of each cluster is recalculated based on observations falling into the cluster. After the new centers are identified, each observation is reassigned to the closest center. This process will be repeated until the cluster assignments stop altering.

One important step in k-means algorithm is to identify the number of clusters. This could be done as follows: first let the number of clusters varies from 1 to 10; then calculate the total intra-cluster sum of square (iss); then proceed to plot iss based on the number of k clusters. This plot denotes the appropriate number of clusters required in our model. In the plot, the location of a bend or a knee is the indication of the optimum number of clusters. The folowing R code could used to implement this procedure.

```r
# Find the number of clusters for k-means algorithm

library(purrr)
set.seed(123)
# function to calculate total intra-cluster sum of square
iss <- function(k) {
  kmeans(customer_data[,3:5],k,iter.max=100,nstart=100,algorithm="Lloyd" )$tot.withinss
}
k.values <- 1:10
iss_values <- map_dbl(k.values, iss)
plot(k.values, iss_values,
    type="b", pch = 19, frame = FALSE,
    xlab="Number of clusters K",
    ylab="Total intra-clusters sum of squares")
```



From the above plot, one can see that 5 is the appropriate number of clusters since it apppears at the bend in the elbow plot.

Now apply k-means algorithm with 5 clusters to the data set in R:

```
k5<- kmeans(customer_data[,3:5], 5, iter.max=100,nstart=50,algorithm="Lloyd")
k5
```

```
## K-means clustering with 5 clusters of sizes 79, 36, 39, 23, 23
##
## Cluster means:
##        Age Annual.Income..k.. Spending.Score..1.100.
## 1 43.08861           55.29114               49.56962
## 2 40.66667           87.75000               17.58333
## 3 32.69231           86.53846               82.12821
## 4 25.52174           26.30435               78.56522
## 5 45.21739           26.30435               20.91304
##
## Clustering vector:
##   [1] 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5
##  [36] 4 5 4 5 4 5 4 5 4 5 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [71] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [106] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 2 3 1 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3
## [141] 2 3 1 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
## [176] 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3
##
## Within cluster sum of squares by cluster:
## [1] 30138.051 17669.500 13972.359  4622.261  8948.609
```

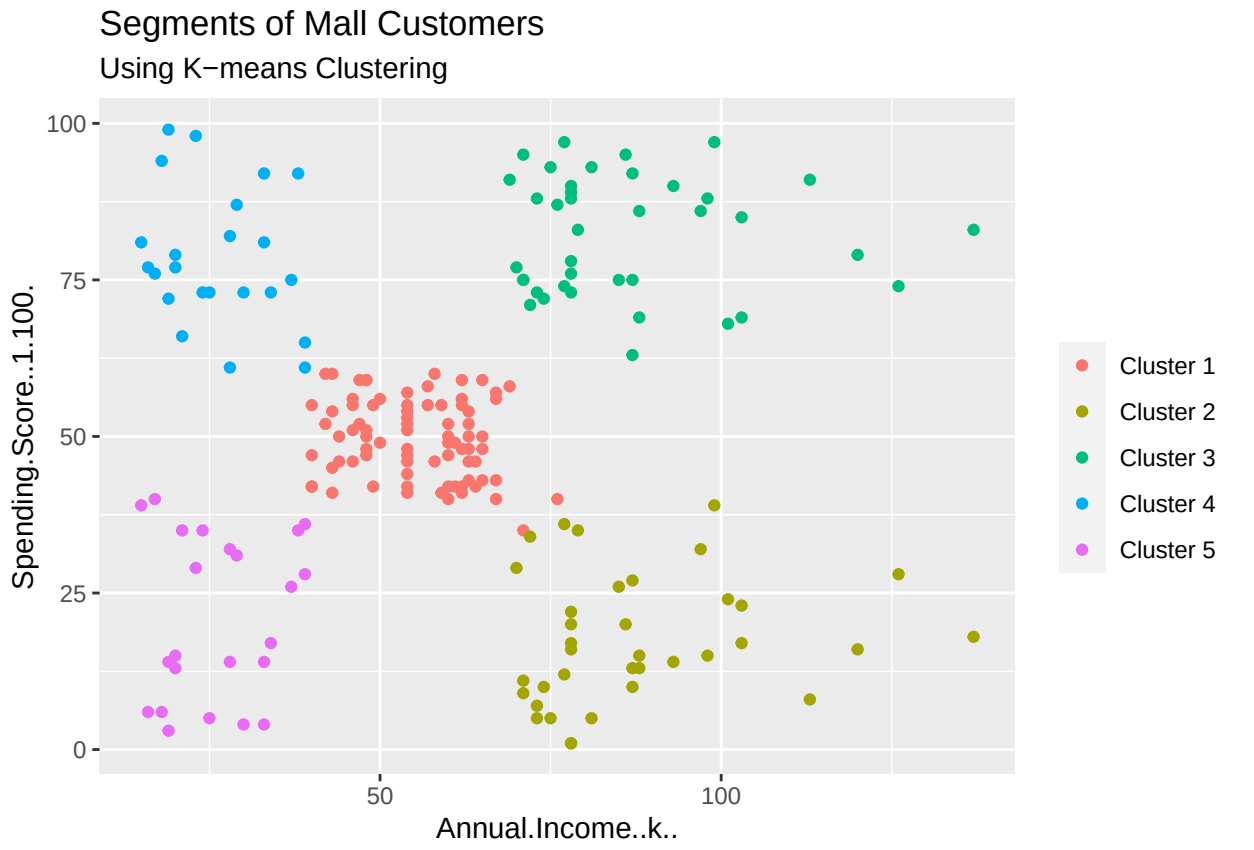
```
##  (between_SS / total_SS =  75.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

The R outputs show information including the cluster assignment for each observation, the cluster means, the total sum of squares, the intra-cluster sum of squares, etc.

Data Visualization

Now we visualize the clustering results in the following plots.

```
set.seed(1)
ggplot(customer_data, aes(x =Annual.Income..k.., y = Spending.Score..1.100.)) +
  geom_point(stat = "identity", aes(color = as.factor(k5$cluster))) +
  scale_color_discrete(name=" ",
                       breaks=c("1", "2", "3", "4", "5"),
                       labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5")) +
  ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")
```



From the above plot, one can see that the 5 clusters are well separated:

- Cluster 1: this cluster represents customers with median annual income and spending scores.
- Cluster 2: this cluster comprises of customers with high annual income and low spending scores.
- Cluster 3: this cluster represents customers with high annual income and high spending scores.

- Cluster 4: this cluster comprises of customers with low annual income and high spending scores.
- Cluster 5: this cluster comprises of customers with low annual income and low spending scores.