ANA 515 Assignment 4 Data Analytics Project

Saif Hossain

2022-10-12

1 Business Problem

A mall has collected information about its customers and wants to know which gender and at what income level they have the greatest spending habit which they classified as spending score. The score ranges from 1 - 100. That way a greater customer service can be provided to those VIP customers.

2 Dataset retreival

The data set was collected from Kaggle as a csv file, since this is a relatively small data set it was stored in my personal machine and below is a an example of what the data set looks like:

```
# 3

# Importing the data set and saving it as a variable from my personal machine.

mall_data <- read.csv("C:/Users/saiii/OneDrive/Desktop/McDaniel College/ANA 515/Week 8/Mall_Customers.c

head(mall_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
```

```
#4

# Getting columns of my dataframe

colnames(mall_data)
```

```
## [1] "CustomerID"             "Genre"                  "Age"
## [4] "Annual.Income..k.."     "Spending.Score..1.100."
```

```
# Characteristics of the data. And Inline code below.

dimensions <- dim(mall_data)
```

```
#Summary of the dataset

summary(mall_data)
```

```
##    CustomerID        Genre                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
```

This dataframe has 200 rows and 5 columns. The names of the columns and a brief description of each are in the table below:

```
#5

#Cleaning the column name
mall_data <- rename(mall_data, Sex = 'Genre')
mall_data <- rename(mall_data, Spending_Score_1_To_100 = 'Spending.Score..1.100.')

#Dropping columns such as customer id because it has no value to analysis and sex for discrimination pu

mall_data_2 <- select(mall_data,
                      Age,
                      Annual.Income..k..,
                      Spending_Score_1_To_100)

head(mall_data_2)
```

```
##   Age Annual.Income..k.. Spending_Score_1_To_100
## 1  19                 15                      39
## 2  21                 15                      81
## 3  20                 16                       6
## 4  23                 16                      77
## 5  31                 17                      40
## 6  22                 17                      76
```

```
#Making sure there is no missing values in my data set.
sum(is.na(mall_data_2))
```

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

```
#6

# Bringing the wss plot under function

wssplot <- function(data, nc=15, seed=1234){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")

}
```

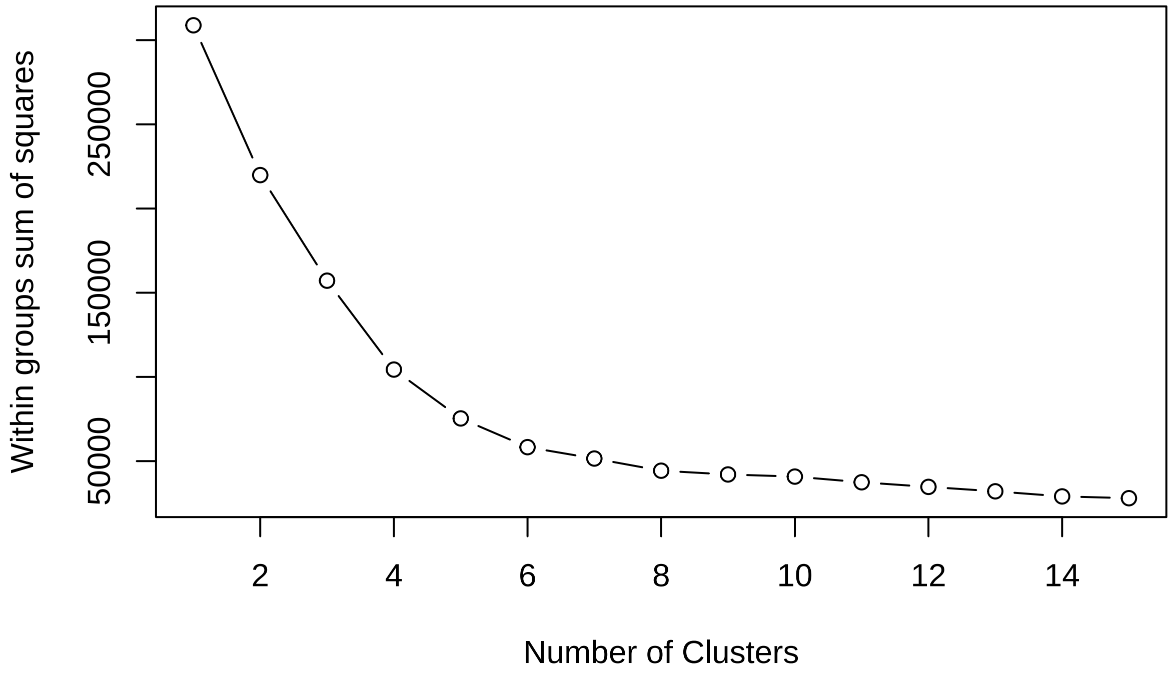
Modeling the data

K-Means cluster was chosen for modeling to find groups who has the highest spending score and group those variables of data.

```
# 6 & 7

# wss plot to choose the maximum number of clusters, using the elbow method 4 is a good number of clust

wssplot(mall_data_2)
```



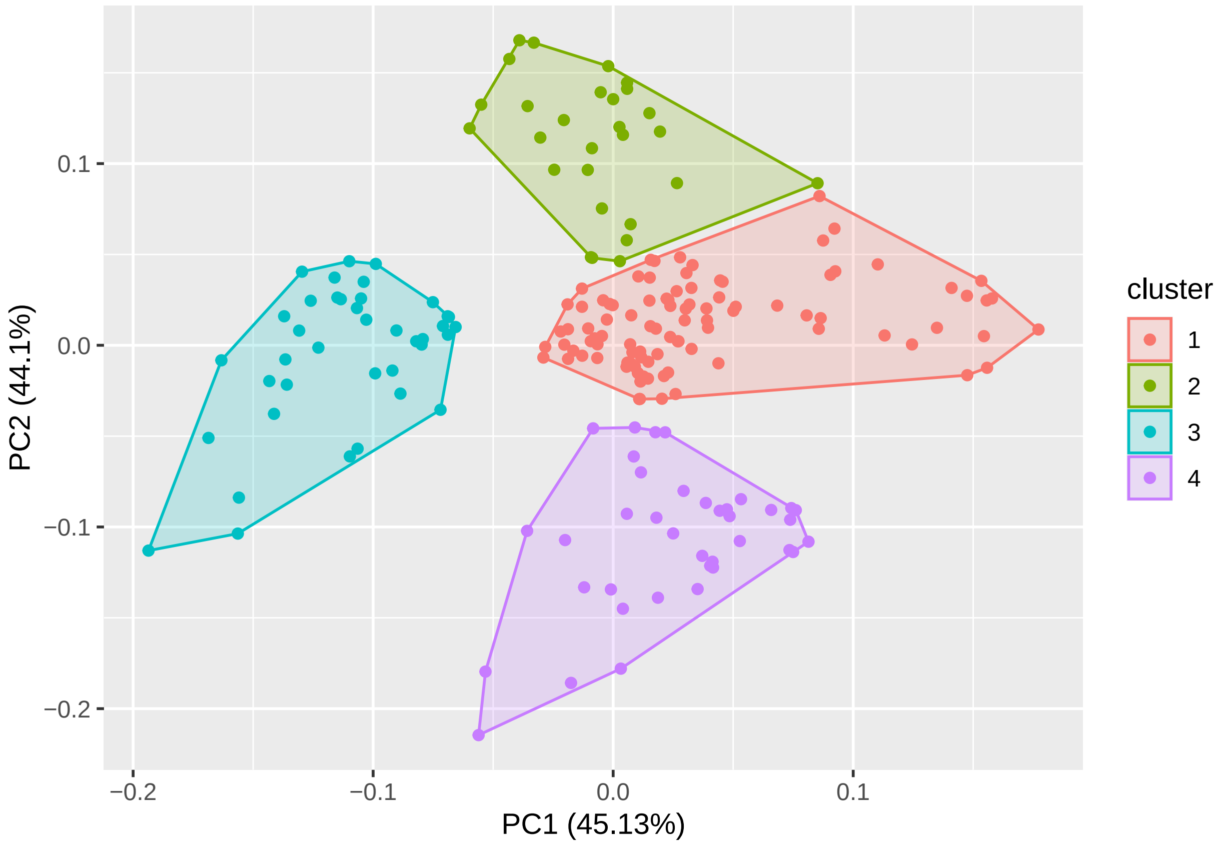
```
# K-Means cluster


KM <- kmeans(mall_data_2, 4)
```

```
# 8 & 9

#Clustering the plot with values of 4 because of elbow shape.


autoplot(KM, mall_data_2,frame = TRUE)
```



```
#Cluster centers describes the age and annual income for maximum spending score.

KM$centers
```

```
##        Age Annual.Income..k.. Spending_Score_1_To_100
## 1 44.89474           48.70526                42.63158
## 2 24.82143           28.71429                74.25000
## 3 32.69231           86.53846                82.12821
## 4 40.39474           87.00000                18.63158
```

Summary

Average Spending for cluster 1 is 42.63 out of 100 when average age is 44.59 and income is 48.71 K

Average Spending for cluster 2 is 74.25 out of 100 when average age is 24.83 and income is 28.71 K

Average Spending for cluster 3 is 82.12 out of 100 when average age is 32.69 and income is 86.53 K

Average Spending for cluster 4 is 18.63 out of 100 when average age is 40.39 and income is 87.00 K

Result

The mall should target people of age 32, with an income of 86.54 K for maximum profit.