# K-Means_Clustering

Poro Burman

8/26/20

## Introduction

The purpose of this script is to find a pattern amongst the customers visiting a Mall.
The dataset used in this script consists of customers data collected by a Mall.
Each row consists of a customer. Each column consists of a customer's variable,
including his/her spending score.

## 1. Data pre-processing

Import the dataset

```
dataset = read.csv('Mall_Customers.csv')
```

Select the variables to perform clustering on

```
X <- dataset[4:5]
```

print top 5 dataset values

```
head(dataset)
```

```
##   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
```

## 2. Select number of clusters

```
set.seed(6)

wcss <- vector()
```

```r
for(i in 1:10) wcss[i] <- sum(kmeans(X, i)$withinss)

plot(1:10,
     wcss,
     type = 'b',
     main = paste('Clusters of clients'),
     xlab = "Number of clusters",
     ylab = "WCSS")
```


from the elbow plot, I'm selecting 5 clusters for further downstream analysis.

## 3. Training K-Means model on the dataset

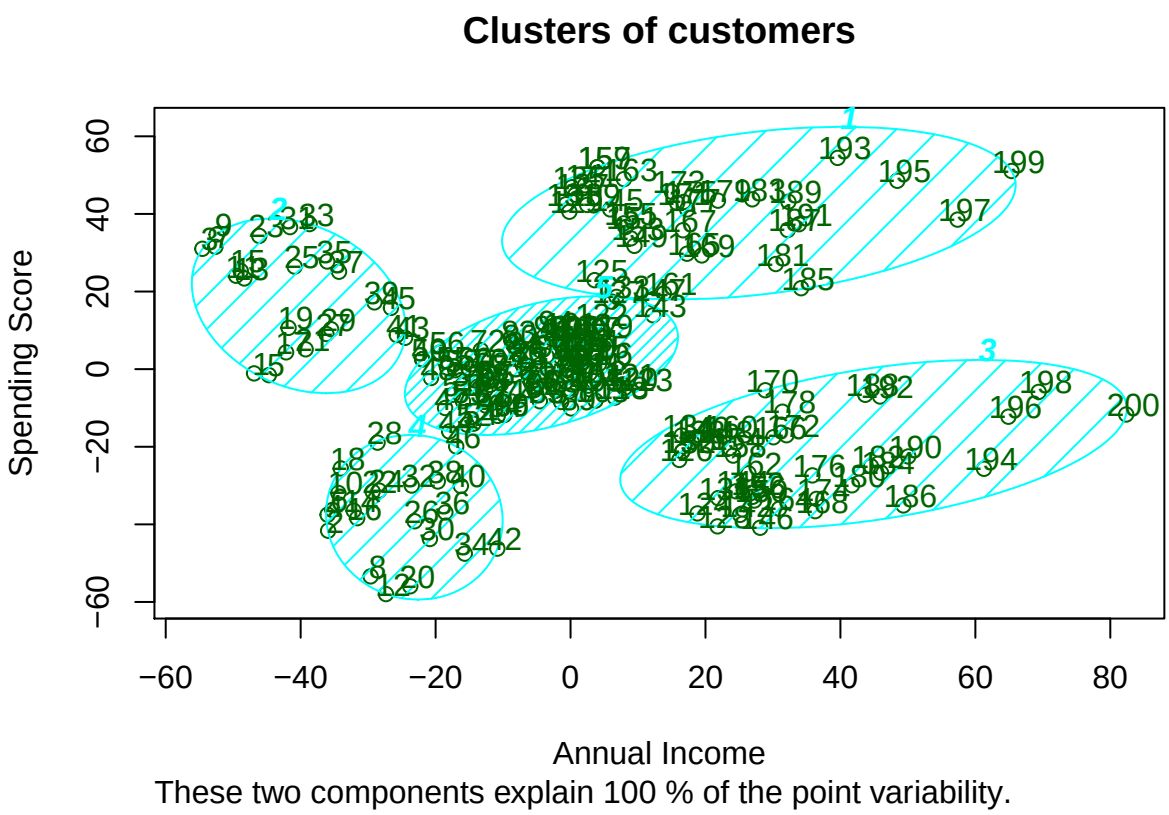
```r
set.seed(29)

kmeans <- kmeans(X, 5, iter.max = 300, nstart = 10)
```

## 4. Visualize the clusters

```r
library(cluster)

clusplot(X,
         kmeans$cluster,
         lines = 0,
         shade = TRUE,
         labels = 2,
         plotchar = FALSE,
         span = TRUE,
         main = paste('Clusters of customers'),
         xlab = "Annual Income",
         ylab = "Spending Score")
```




As you can see, different clusters show the customers who have different annual incomes and their related spending scores. This can be used to target certain specific customers based upon these clusters.