K-MEANS CLUSTERING

Ahmed Basha

29 Januar 2018

```
# Importing the dataset
dataset = read.csv("bank-full.csv",
                   header = TRUE,
                   sep = ";",
                   stringsAsFactors = TRUE)

#save y attribute
y_var = dataset$y

#filter data with two attributes (month, duration)
dataset = dataset[11:12]
```

```
# Encoding categorical data - (month)
dataset$month = factor(dataset$month,
                   levels = c('jan', 'feb', 'mar', 'apr', 'may', 'jun',
                              'jul', 'aug', 'sep', 'oct', 'nov', 'dec'),
                   labels = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12))

# Convert month attribute datatype to numeric
dataset$month = as.numeric(levels(dataset$month))[dataset$month]
```

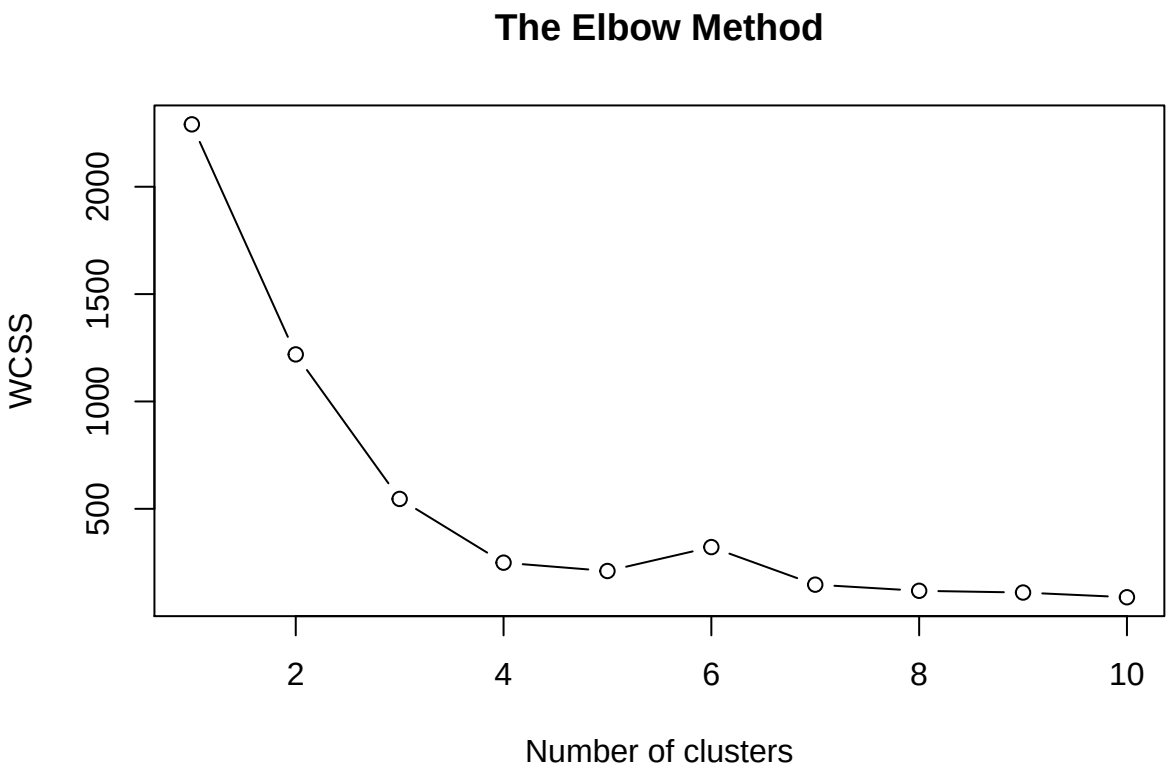
```
#MinMax Normalization
dataset$month <- scales::rescale(dataset$month, to=c(0,1))
dataset$duration <- scales::rescale(dataset$duration, to=c(0,1))

#see the structure of data
str(dataset)
```

```
## 'data.frame':    45211 obs. of  2 variables:
##  $ month   : num  0.364 0.364 0.364 0.364 0.364 ...
##  $ duration: num  0.0531 0.0307 0.0155 0.0187 0.0403 ...
```

```
# Fitting K-Means to the dataset
set.seed(29)
kmeans_MinMax = kmeans(x = dataset, centers = 2)
```

```
#Cluster component
km_min_max <- kmeans_MinMax$cluster
```

```
#Using the elbow method to find the optimal number of clusters
set.seed(6)
wcss = vector()
for (i in 1:10) wcss[i] = sum(kmeans(dataset, i)$withinss)
plot(1:10,
     wcss,
     type = 'b',
     main = paste('The Elbow Method'),
     xlab = 'Number of clusters',
     ylab = 'WCSS')
```




```
#confusion matrix of clustering compared to y
table(y_var,km_min_max)
```

```
##       km_min_max
## y_var     1     2
##    no  23689 16233
##    yes  2879  2410
```

```
#table of clustering results
table(km_min_max)
```

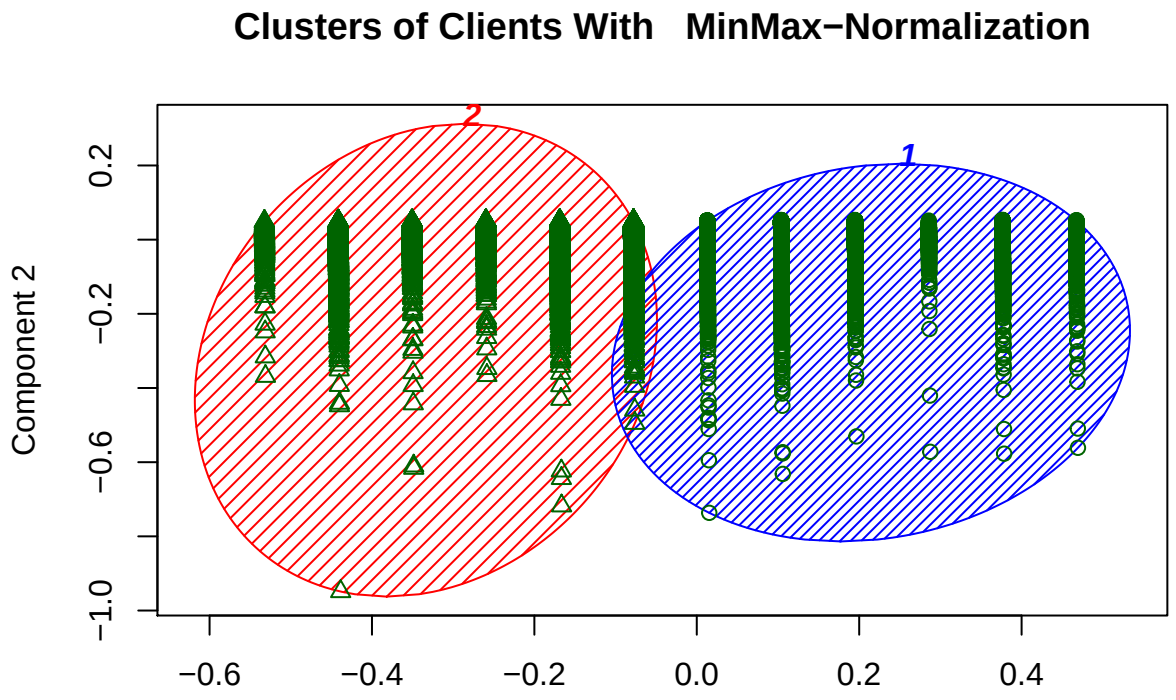
```
## km_min_max
##     1     2
## 26568 18643
```

```
# Visualising the clusters
#install.packages('cluster')
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 3.4.3
```

```
clusplot(dataset,
         km_min_max,
         lines = 0,
         shade = TRUE,
         color = TRUE,
         labels = 4,
         plotchar = TRUE,
         span = TRUE,
```

```
        main = paste('Clusters of Clients With   MinMax-Normalization'))
```



These two components explain 100 % of the point variability.

==========

```
#Z-score Normalization
dataset$month <- scale(dataset$month)
dataset$duration <- scale(dataset$duration)
```

```
#see the structure of data
str(dataset)
```

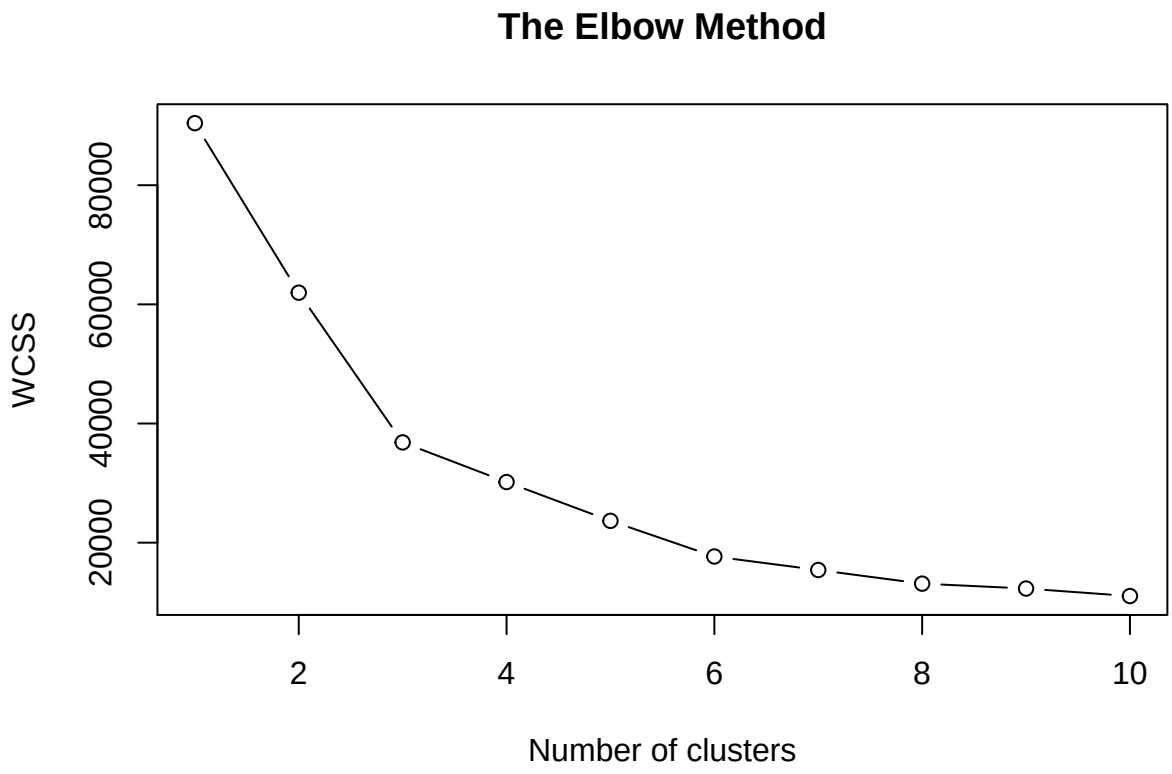
```
## 'data.frame':    45211 obs. of  2 variables:
##  $ month   : num [1:45211, 1] -0.475 -0.475 -0.475 -0.475 -0.475 ...
##   ..- attr(*, "scaled:center")= num 0.468
##   ..- attr(*, "scaled:scale")= num 0.219
##  $ duration: num [1:45211, 1] 0.011 -0.416 -0.707 -0.645 -0.234 ...
##   ..- attr(*, "scaled:center")= num 0.0525
##   ..- attr(*, "scaled:scale")= num 0.0524
```

```
# Fitting K-Means to the dataset
set.seed(29)
kmeans_Z_Score = kmeans(x = dataset, centers = 2)
```

```
km_z_score <- kmeans_Z_Score$cluster
```

```
#Using the elbow method to find the optimal number of clusters
set.seed(6)
wcss = vector()
```

```
for (i in 1:10) wcss[i] = sum(kmeans(dataset, i)$withinss)
plot(1:10,
     wcss,
     type = 'b',
     main = paste('The Elbow Method'),
     xlab = 'Number of clusters',
     ylab = 'WCSS')
```




```
#confusion matrix of clustering compared to y
table(y_var,km_z_score)
```

```
##      km_z_score
## y_var     1     2
##   no  23689 16233
##   yes  2879  2410
```

```
#table of clustering results
table(km_z_score)
```

```
## km_z_score
##     1     2
## 26568 18643
```

```
# Visualising the clusters
#install.packages('cluster')
library(cluster)
clusplot(dataset,
         km_z_score,
```

```
        lines = 0,
        shade = TRUE,
        color = TRUE,
        labels = 4,
        plotchar = TRUE,
        span = TRUE,
        main = paste('Clusters of Clients With Z-Score-Normalization'))
```