K-Means—Prospects-clustering.R

patriciamaya

2020-11-30

```r
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.0.2
```

```r
data <- read_excel("~/Downloads/prospect.xls")
head(data)
```

```
## # A tibble: 6 x 9
##   ID          AGE INCOME SEX   MARRIED OWNHOME LOC   CLIMATE `FICO>=700`
##   <chr>     <dbl>  <dbl> <chr>   <dbl>   <dbl> <chr> <chr>         <dbl>
## 1 700778094    37     57 F           0       0 B     20                0
## 2 138158771    46     71 M           1       0 B     20                0
## 3 229652047    45     65 M           1       1 F     20                1
## 4 150424460    38     50 F           0       0 A     10                0
## 5 828150627    34     44 M           0       0 F     20                0
## 6 185836923    69     60 F           0       0 H     30                0
```

```r
#exclude location and ID
data <- data[,-1]
data<- data[,-6]

summary(data)
```

```
##       AGE            INCOME           SEX               MARRIED      
##  Min.   :18.00   Min.   : 15.00   Length:4701        Min.   :0.0000  
##  1st Qu.:38.00   1st Qu.: 35.00   Class  :character   1st Qu.:0.0000  
##  Median :44.00   Median : 50.00   Mode   :character   Median :1.0000  
##  Mean   :44.23   Mean   : 47.69                      Mean   :0.5785  
##  3rd Qu.:50.00   3rd Qu.: 61.00                      3rd Qu.:1.0000  
##  Max.   :75.00   Max.   :116.00                      Max.   :1.0000  
##  NA's   :106     NA's   :106                         NA's   :106     
##     OWNHOME         CLIMATE            FICO>=700     
##  Min.   :0.0000   Length:4701        Min.   :0.0000  
##  1st Qu.:0.0000   Class  :character   1st Qu.:0.0000  
##  Median :0.0000   Mode   :character   Median :0.0000  
##  Mean   :0.3277                      Mean   :0.4135  
##  3rd Qu.:1.0000                      3rd Qu.:1.0000  
##  Max.   :1.0000                      Max.   :1.0000  
##  NA's   :106                         NA's   :106     
```

```r
Data <- data[complete.cases(data),]
Data_num <- Data[-c(4:5, 7)] #df with num variables only

#normalize the data
num_var <- unlist(lapply(Data_num, is.numeric))
```

```
num_var
```

```
##     AGE  INCOME     SEX CLIMATE
##    TRUE    TRUE   FALSE   FALSE
```

```
Data_norm <- Data_num[, num_var]
min <- apply(Data_norm, 2, min, na.rm = TRUE)
max <- apply(Data_norm, 2, max, na.rm = TRUE)
Data_scaled <- scale(Data_norm, center = min, scale = max - min)
summary(Data_scaled)
```

```
##       AGE             INCOME
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3509   1st Qu.:0.1980
##  Median :0.4561   Median :0.3465
##  Mean   :0.4602   Mean   :0.3237
##  3rd Qu.:0.5614   3rd Qu.:0.4554
##  Max.   :1.0000   Max.   :1.0000
```

```
#convert categorical variables to dummy variables
fac_var <- !num_var
fac_var
```

```
##     AGE  INCOME     SEX CLIMATE
##   FALSE   FALSE    TRUE    TRUE
```

```
fac_var <- as.logical(fac_var)
fac_var
```

```
## [1] FALSE FALSE  TRUE  TRUE
```

```
library(psych)
Data_fac <- as.data.frame(lapply(Data_num[,fac_var], dummy.code))

#combine normalized data and original dummy variables
Data_clean <- data.frame(Data_scaled, Data_fac, Data$`FICO>=700`, Data$OWNHOME, Data$MARRIED)
View(Data_clean)

set.seed(123)
kmModel <- kmeans(Data_clean, 4, nstart=100)
#kmModel

# The number of instnces in each cluster is:
kmModel$size
```

```
## [1]  933  800 1501 1361
```

```
# The cluster means, aka centroids, are:
kmModel$centers
```

```
##         AGE    INCOME SEX.M SEX.F CLIMATE.20 CLIMATE.30 CLIMATE.10
## 1 0.4634738 0.3259792     1     0          0  0.4876742  0.5123258
## 2 0.4508553 0.2702104     0     1          0  0.5300000  0.4700000
## 3 0.4588169 0.3736717     1     0          1  0.0000000  0.0000000
## 4 0.4650992 0.2984628     0     1          1  0.0000000  0.0000000
##   Data..FICO..700. Data.OWNHOME Data.MARRIED
## 1        0.3879957    0.2111468    0.5712755
## 2        0.4237500    0.3412500    0.5775000
## 3        0.4497002    0.3837442    0.5622918
## 4        0.3850110    0.3379868    0.6017634
```

```
# The variances within clusters are:
kmModel$withinss
```

```
## [1] 1124.819 1018.777 1179.812 1025.054
```

```
# The variance between clusters is:
kmModel$betweenss
```

```
## [1] 3939.794
```

```
#characteristics of each cluster
Cluster1 <- Data_clean[kmModel$cluster ==1, ]
summary(Cluster1)
```

```
##       AGE              INCOME           SEX.M       SEX.F     CLIMATE.20
##  Min.   :0.01754   Min.   :0.0000   Min.   :1   Min.   :0   Min.   :0
##  1st Qu.:0.35088   1st Qu.:0.2178   1st Qu.:1   1st Qu.:0   1st Qu.:0
##  Median :0.45614   Median :0.3663   Median :1   Median :0   Median :0
##  Mean   :0.46347   Mean   :0.3260   Mean   :1   Mean   :0   Mean   :0
##  3rd Qu.:0.57895   3rd Qu.:0.4356   3rd Qu.:1   3rd Qu.:0   3rd Qu.:0
##  Max.   :1.00000   Max.   :0.7129   Max.   :1   Max.   :0   Max.   :0
##    CLIMATE.30       CLIMATE.10     Data..FICO..700.  Data.OWNHOME
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000    Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000    1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :0.000    Median :0.0000
##  Mean   :0.4877   Mean   :0.5123   Mean   :0.388    Mean   :0.2111
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000    3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000    Max.   :1.0000
##   Data.MARRIED
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.5713
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
#Cluster 1 is entirely Male, and there are no occurrences of Climate 20.
#Occurrences are basically split between Climate 30 and 10.
#A little over half of occurrences are Married, but the majority do not own a home.

Cluster2 <- Data_clean[kmModel$cluster ==2, ]
summary(Cluster2)
```

```
##       AGE             INCOME           SEX.M       SEX.F     CLIMATE.20
##  Min.   :0.0000   Min.   :0.0000   Min.   :0   Min.   :1   Min.   :0
##  1st Qu.:0.3333   1st Qu.:0.1188   1st Qu.:0   1st Qu.:1   1st Qu.:0
##  Median :0.4561   Median :0.2723   Median :0   Median :1   Median :0
##  Mean   :0.4509   Mean   :0.2702   Mean   :0   Mean   :1   Mean   :0
##  3rd Qu.:0.5614   3rd Qu.:0.3960   3rd Qu.:0   3rd Qu.:1   3rd Qu.:0
##  Max.   :1.0000   Max.   :1.0000   Max.   :0   Max.   :1   Max.   :0
##    CLIMATE.30     CLIMATE.10   Data..FICO..700.  Data.OWNHOME
##  Min.   :0.00   Min.   :0.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.00   Median :0.00   Median :0.0000   Median :0.0000
##  Mean   :0.53   Mean   :0.47   Mean   :0.4238   Mean   :0.3412
##  3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00   Max.   :1.00   Max.   :1.0000   Max.   :1.0000
##   Data.MARRIED
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.5775
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
#Cluster 2 is entirely Female, and there are no occurrences of Climate 20.
#Occurrences are basically split between Climate 30 and 10.
#A little over half of occurrences are Married, but the majority do not own a home.
#However, more own a home than in Cluster 1. Income mean is lower than in Cluster 1,
#but Age mean is higher.

Cluster3 <- Data_clean[kmModel$cluster ==3, ]
summary(Cluster3)
```

```
##       AGE             INCOME             SEX.M        SEX.F     CLIMATE.20
##  Min.   :0.0000   Min.   :0.009901   Min.   :1   Min.   :0   Min.   :1
##  1st Qu.:0.3509   1st Qu.:0.267327   1st Qu.:1   1st Qu.:0   1st Qu.:1
##  Median :0.4561   Median :0.405941   Median :1   Median :0   Median :1
##  Mean   :0.4588   Mean   :0.373672   Mean   :1   Mean   :0   Mean   :1
##  3rd Qu.:0.5614   3rd Qu.:0.485148   3rd Qu.:1   3rd Qu.:0   3rd Qu.:1
##  Max.   :1.0000   Max.   :0.990099   Max.   :1   Max.   :0   Max.   :1
##    CLIMATE.30   CLIMATE.10 Data..FICO..700.  Data.OWNHOME     Data.MARRIED
##  Min.   :0    Min.   :0    Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0    1st Qu.:0    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0    Median :0    Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :0    Mean   :0    Mean   :0.4497   Mean   :0.3837   Mean   :0.5623
##  3rd Qu.:0    3rd Qu.:0    3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :0    Max.   :0    Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

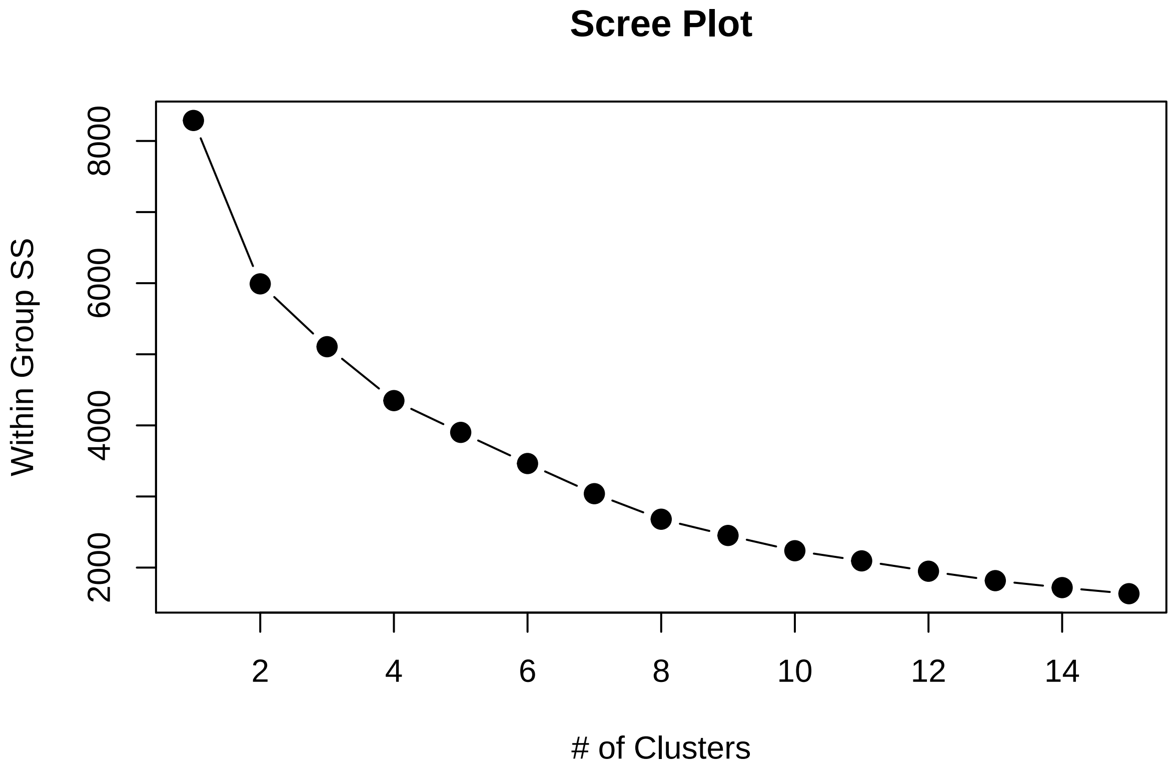
```
#Cluster 3 is entirely Male, and all occurrences are Climate 20.
#No occurrences are in Climate 30 or 10. Again, a little over half of occurrences are Married,
#but the majority do not own a home. More own a home than in Cluster 1 and 2 though.
#Income mean is the highest yet of Clusters 1-3.

Cluster4 <- Data_clean[kmModel$cluster ==4, ]
summary(Cluster4)
```

```
##       AGE             INCOME           SEX.M        SEX.F     CLIMATE.20
##  Min.   :0.0000   Min.   :0.0000   Min.   :0   Min.   :1   Min.   :1
##  1st Qu.:0.3509   1st Qu.:0.1782   1st Qu.:0   1st Qu.:1   1st Qu.:1
##  Median :0.4561   Median :0.2772   Median :0   Median :1   Median :1
##  Mean   :0.4651   Mean   :0.2985   Mean   :0   Mean   :1   Mean   :1
##  3rd Qu.:0.5789   3rd Qu.:0.4257   3rd Qu.:0   3rd Qu.:1   3rd Qu.:1
##  Max.   :1.0000   Max.   :0.7327   Max.   :0   Max.   :1   Max.   :1
##    CLIMATE.30   CLIMATE.10 Data..FICO..700.  Data.OWNHOME    Data.MARRIED
##  Min.   :0    Min.   :0    Min.   :0.000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0    1st Qu.:0    1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0    Median :0    Median :0.000   Median :0.000   Median :1.0000
##  Mean   :0    Mean   :0    Mean   :0.385   Mean   :0.338   Mean   :0.6018
##  3rd Qu.:0    3rd Qu.:0    3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :0    Max.   :0    Max.   :1.000    Max.   :1.000   Max.   :1.0000
```

```
#Cluster 4 is entirely Male, and all occurrences are in Climate 20.
#Like Cluster 3, no occurrences are in Climate 30 and 10.
#A little over half of occurrences are Married, and the mean here is the highest of all 4 Clusters.
#Still, the majority do not own a home. Age mean is the highest of all 4 Clusters,
#whereas Income mean is lower than in Cluster 2 and 1.

#BEST VALUE OF K
mydata <- Data_clean
wss <- (nrow(mydata)-1)*sum(apply(mydata, 2, var))
for (i in 1:15)
  wss[i] <- sum(kmeans(mydata, centers = i, nstart = 100)$withinss)
plot(1:15, wss, type ="b", main = "Scree Plot", xlab = "# of Clusters", ylab = "Within Group SS", pch =
```



```
#EVALUATE model -- Silhouette measure of the clusters obtained by best k
#rerun model with k=8 clusters
set.seed(123)
kmModel_bestk <- kmeans(Data_clean, 8, nstart = 100)

# The number of instnces in each cluster is:
kmModel_bestk$size
```

```
## [1] 424 376 478 819 455 826 675 542
```

```
# The cluster means, aka centroids, are:
kmModel_bestk$centers
```

```
##         AGE    INCOME SEX.M SEX.F CLIMATE.20 CLIMATE.30 CLIMATE.10
## 1 0.4466236 0.2572389     0     1          0          1          0
## 2 0.4556271 0.2848378     0     1          0          0          1
## 3 0.4613154 0.3184888     1     0          0          0          1
## 4 0.5222672 0.2969813     0     1          1          0          0
## 5 0.4657413 0.3338483     1     0          0          1          0
## 6 0.4481543 0.3246710     1     0          1          0          0
## 7 0.4718648 0.4336340     1     0          1          0          0
## 8 0.3787143 0.3007015     0     1          1          0          0
##   Data..FICO..700. Data.OWNHOME Data.MARRIED
## 1        0.4386792    0.3160377    0.5613208
## 2        0.4069149    0.3696809    0.5957447
## 3        0.3619247    0.2740586    0.5564854
## 4        0.4981685    0.3357753    1.0000000
## 5        0.4153846    0.1450549    0.5868132
## 6        0.0000000    0.2493947    0.4370460
## 7        1.0000000    0.5481481    0.7155556
## 8        0.2140221    0.3413284    0.0000000
```

```r
# The variances within clusters are:
kmModel_bestk$withinss
```

```
## [1] 324.9492 294.0895 353.8402 424.6883 299.9426 413.2596 328.5150 241.1409
```

```r
# The variance between clusters is:
kmModel_bestk$betweenss
```

```
## [1] 5607.831
```

```r
library(cluster)
ss <- silhouette(kmModel_bestk$cluster, dist(Data_clean))
mean(ss[ ,3])
```

```
## [1] 0.3706712
```