K Means Clustering

Suhail Shaikh

12/19/2019

Q.1) Use the K-means method to cluster the prospects dataset. Set the number of clusters to four. How many points are in each cluster? What are cluster means and variances?

```
library(xlsx)
Data=read.xlsx("prospect.xls",sheetName="Sheet1")
str(Data)
```

```
## 'data.frame':    4701 obs. of  9 variables:
##  $ ID       : Factor w/ 4701 levels "000595865","001038701",..: 3268 631
1086 686 3889 867 3492 817 109 403 ...
##  $ AGE      : num  37 46 45 38 34 69 46 28 37 46 ...
##  $ INCOME   : num  57 71 65 50 44 60 42 63 59 57 ...
##  $ SEX      : Factor w/ 3 levels "","F","M": 2 3 3 2 3 2 2 2 3 3 ...
##  $ MARRIED  : num  0 1 1 0 0 0 1 0 1 1 ...
##  $ OWNHOME  : num  0 0 1 0 0 0 0 1 0 1 ...
##  $ LOC      : Factor w/ 8 levels "A","B","C","D",..: 2 2 6 1 6 8 2 5 2 5
...
##  $ CLIMATE  : Factor w/ 3 levels "10","20","30": 2 2 2 1 2 3 2 2 2 2 ...
##  $ FICO..700: num  0 0 1 0 0 0 1 1 1 1 ...
```

```
#changing columnname of a single column
colnames(Data)[ncol(Data)] <- "FICO"

#Converting variables to factor
Data$MARRIED <- as.factor(Data$MARRIED)
Data$OWNHOME <- as.factor(Data$OWNHOME)
Data$FICO <- as.factor(Data$FICO)
str(Data)
```

```
## 'data.frame':    4701 obs. of  9 variables:
##  $ ID     : Factor w/ 4701 levels "000595865","001038701",..: 3268 631
1086 686 3889 867 3492 817 109 403 ...
##  $ AGE    : num  37 46 45 38 34 69 46 28 37 46 ...
##  $ INCOME : num  57 71 65 50 44 60 42 63 59 57 ...
##  $ SEX    : Factor w/ 3 levels "","F","M": 2 3 3 2 3 2 2 2 3 3 ...
##  $ MARRIED: Factor w/ 2 levels "0","1": 1 2 2 1 1 1 2 1 2 2 ...
##  $ OWNHOME: Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 2 1 2 ...
##  $ LOC    : Factor w/ 8 levels "A","B","C","D",..: 2 2 6 1 6 8 2 5 2 5 ...
##  $ CLIMATE: Factor w/ 3 levels "10","20","30": 2 2 2 1 2 3 2 2 2 2 ...
##  $ FICO   : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 2 2 2 ...
```

```r
DataNew <- Data
DataNew$ID <- NULL
DataNew$LOC <-NULL
str(DataNew)
```

```
## 'data.frame':    4701 obs. of  7 variables:
##  $ AGE    : num  37 46 45 38 34 69 46 28 37 46 ...
##  $ INCOME : num  57 71 65 50 44 60 42 63 59 57 ...
##  $ SEX    : Factor w/ 3 levels "","F","M": 2 3 3 2 3 2 2 2 3 3 ...
##  $ MARRIED: Factor w/ 2 levels "0","1": 1 2 2 1 1 1 2 1 2 2 ...
##  $ OWNHOME: Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 2 1 2 ...
##  $ CLIMATE: Factor w/ 3 levels "10","20","30": 2 2 2 1 2 3 2 2 2 2 ...
##  $ FICO   : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 2 2 2 ...
```

```r
summary(DataNew)
```

```
##       AGE             INCOME       SEX      MARRIED     OWNHOME
##  Min.   :18.00   Min.   : 15.00    : 106   0   :1937   0   :3089
##  1st Qu.:38.00   1st Qu.: 35.00   F:2161   1   :2658   1   :1506
##  Median :44.00   Median : 50.00   M:2434   NA's: 106   NA's: 106
##  Mean   :44.23   Mean   : 47.69
##  3rd Qu.:50.00   3rd Qu.: 61.00
##  Max.   :75.00   Max.   :116.00
##  NA's   :106     NA's   :106
##  CLIMATE     FICO
##  10: 871   0   :2695
##  20:2932   1   :1900
##  30: 898   NA's: 106
##
##
##
##
```

```r
#Treating na
nrow(DataNew)
```

```
## [1] 4701
```

```r
DataNew1=na.omit(DataNew)
nrow(DataNew1)
```

```
## [1] 4595
```

```r
#As there are only 106 rows i.e. 2% rows with na values remove them

#install.packages("clustMixType")
library(clustMixType)

km4 = kproto(DataNew1, k = 4, lambda = NULL, iter.max=100, nstart=1, na.rm =
TRUE, verbose = TRUE)
```

```
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
```

```
km4
```

```
## Numeric predictors: 2
## Categorical predictors: 5
## Lambda: 373.7569
##
## Number of Clusters: 4
## Cluster sizes: 1339 798 1235 1223
## Within cluster error: 876072.2 523360.5 829174.3 821250.6
##
## Cluster prototypes:
##        AGE   INCOME SEX MARRIED OWNHOME CLIMATE FICO
## 1 49.04630 57.19866   M       1       0      20    1
## 2 39.66541 63.53008   M       0       1      20    1
## 3 36.77004 42.87045   F       0       0      20    0
## 4 49.48160 31.82420   F       1       0      20    0
```

```
#number of pts in each cluster
table(km4$cluster)
```

```
##
##    1    2    3    4
## 1339  798 1235 1223
```

```
#Cluster means(prototype)
km4$centers
```

```
##        AGE   INCOME SEX MARRIED OWNHOME CLIMATE FICO
## 1 49.04630 57.19866   M       1       0      20    1
## 2 39.66541 63.53008   M       0       1      20    1
## 3 36.77004 42.87045   F       0       0      20    0
## 4 49.48160 31.82420   F       1       0      20    0
```

Q.2) What is the best value of k for this data set?

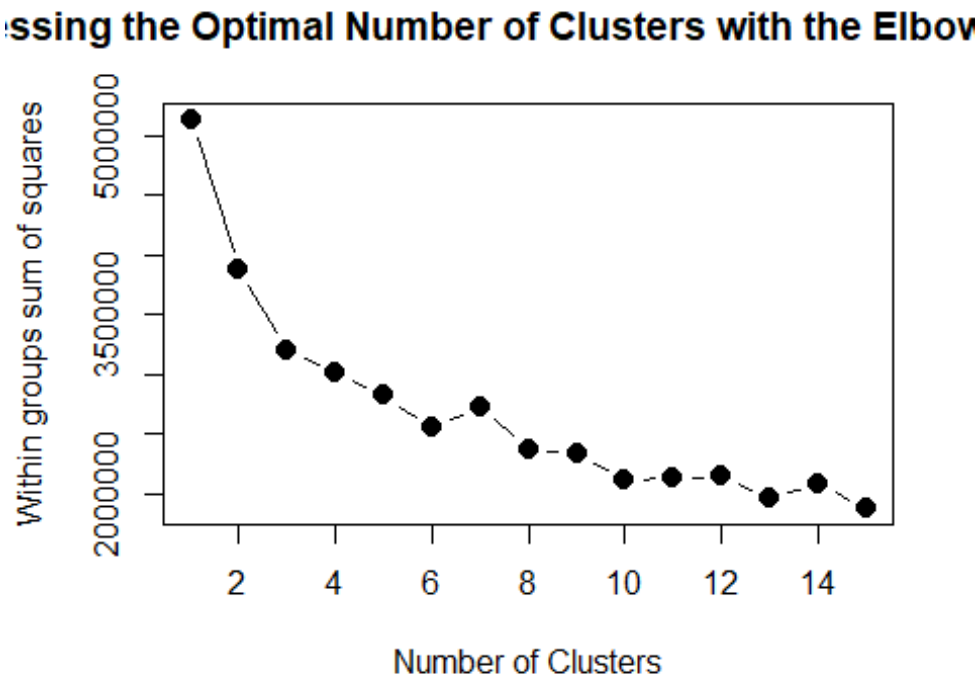
```
#Check for the optimal number of clusters given the data
wss <- (nrow(DataNew1)-1)*sum(apply(DataNew1[,c(1,2)],2,var))
for (i in 1:15) wss[i] <- sum(kproto(DataNew1, k=i)$withinss)
#i is number of clusters   #sum adds the distance within all the clusters
```

```
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
```

```
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
```

```
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
##
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569
```

```
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum
of squares",
     main="Assessing the Optimal Number of Clusters with the Elbow Method",
pch=20, cex=2)
```




```
#hence the Optimal Number of Clusters with the Elbow Method is 8


#Perform K-Means with the optimal number of clusters identified from the
Elbow method
km8 = kproto(DataNew1, k = 3, lambda = NULL, iter.max=100, nstart=1, na.rm =
TRUE, verbose = TRUE)

## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 373.7569

km8

## Numeric predictors: 2
## Categorical predictors: 5
## Lambda: 373.7569
##
## Number of Clusters: 3
```

```
## Cluster sizes: 1427 1720 1448
## Within cluster error: 973828.6 1261117 971486.3
##
## Cluster prototypes:
##        AGE   INCOME SEX MARRIED OWNHOME CLIMATE FICO
## 1 46.04765 58.86615   M       1       1      20    1
## 2 46.30465 32.28779   F       1       0      20    0
## 3 39.98550 54.98273   M       0       0      20    0

# install.packages("factoextra")
# library(factoextra)
# fviz_cluster(km8, geom = "point", data = DataNew1) + ggtitle("k=8")
```

Q.3) What is the Silhouette measure of the clusters obtained by best k in part (c)

```
# function to compute average silhouette for k clusters
library(cluster)
str(DataNew1)

## 'data.frame':    4595 obs. of  7 variables:
##  $ AGE    : num  37 46 45 38 34 69 46 28 37 46 ...
##  $ INCOME : num  57 71 65 50 44 60 42 63 59 57 ...
##  $ SEX    : Factor w/ 3 levels "","F","M": 2 3 3 2 3 2 2 2 3 3 ...
##  $ MARRIED: Factor w/ 2 levels "0","1": 1 2 2 1 1 1 2 1 2 2 ...
##  $ OWNHOME: Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 2 1 2 ...
##  $ CLIMATE: Factor w/ 3 levels "10","20","30": 2 2 2 1 2 3 2 2 2 2 ...
##  $ FICO   : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int  45 46 53 149 171 220 291 292
307 373 ...
##   ..- attr(*, "names")= chr  "45" "46" "53" "149" ...

mins <- apply(DataNew1[,c(1,2)], 2, min)
maxs <- apply(DataNew1[,c(1,2)], 2, max)
#str(scaled_data_12)
scaled_data_12 <- as.data.frame(scale(DataNew1[,c(1,2)], center = mins, scale
= maxs- mins))   #scale() function is used for normalization
scaled_data <- data.frame(scaled_data_12,DataNew1[,c(3,4,5,6,7)])
str(scaled_data)

## 'data.frame':    4595 obs. of  7 variables:
##  $ AGE    : num  0.333 0.491 0.474 0.351 0.281 ...
##  $ INCOME : num  0.416 0.554 0.495 0.347 0.287 ...
##  $ SEX    : Factor w/ 3 levels "","F","M": 2 3 3 2 3 2 2 2 3 3 ...
##  $ MARRIED: Factor w/ 2 levels "0","1": 1 2 2 1 1 1 2 1 2 2 ...
##  $ OWNHOME: Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 2 1 2 ...
##  $ CLIMATE: Factor w/ 3 levels "10","20","30": 2 2 2 1 2 3 2 2 2 2 ...
##  $ FICO   : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 2 2 2 ...

#Here i have scaled the data and in question c) i have used unscaled data
```

```
km8_1 = kproto(scaled_data, k = 3)#, lambda = NULL, iter.max=100, nstart=1,
na.rm = TRUE, verbose = TRUE)
```

```
## # NAs in variables:
##     AGE  INCOME     SEX MARRIED OWNHOME CLIMATE    FICO
##       0       0       0       0       0       0       0
## 0 observation(s) with NAs.
##
## Estimated lambda: 0.05904227
```

```
km8_1
```

```
## Numeric predictors: 2
## Categorical predictors: 5
## Lambda: 0.05904227
##
## Number of Clusters: 3
## Cluster sizes: 1681 1358 1556
## Within cluster error: 197.3989 155.1145 179.6044
##
## Cluster prototypes:
##         AGE    INCOME SEX MARRIED OWNHOME CLIMATE FICO
## 1 0.5803250 0.2962110   M       1       0      20    0
## 2 0.5002971 0.3569387   F       1       1      20    1
## 3 0.2955396 0.3243847   M       0       0      20    0
```

```
ss <- silhouette(km8_1$cluster, dist(scaled_data))
```

```
## Warning in dist(scaled_data): NAs introduced by coercion
```

```
mean(ss[ ,3])
```

```
## [1] 0.01813101
```