Home Equity Loan Customer Profiling

K-Means Clustering

Julio Garcia Rengifo

02 May 2023

Problem

The aim of this project is to, through a K-means clustering analysis, create portraits of a large group of home equity loan customers in one bank by divide the customers into different segments based on their financial and career data. Such customer profiling is expected to provide senior management of the bank with a better understanding of different distinct features of different customer segments.

Data

```
# import data
library(readr)
hmeq_profile <- read_csv("hmeq_profile.csv")
head(hmeq_profile)
```

```
## # A tibble: 6 x 12
##    LOAN MORTDUE  VALUE REASON JOB     YOJ DEROG DELINQ CLAGE  NINQ  CLNO DEBTINC
##   <dbl>   <dbl>  <dbl> <chr>  <chr> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1  1700   30548  40320 HomeI~ Other     9     0      0 101.      1     8    37.1
## 2  1800   28502  43034 HomeI~ Other    11     0      0  88.8     0     8    36.9
## 3  2300  102370 120953 HomeI~ Offi~     2     0      0  91.0     0    13    31.6
## 4  2400   34863  47471 HomeI~ Mgr      12     0      0  70.5     1    21    38.3
## 5  2400   98449 117195 HomeI~ Offi~     4     0      0  93.8     0    13    29.7
## 6  2900  103949 112505 HomeI~ Offi~     1     0      0  96.1     0    13    30.1
```

```
ncol(hmeq_profile)
```

```
## [1] 12
```

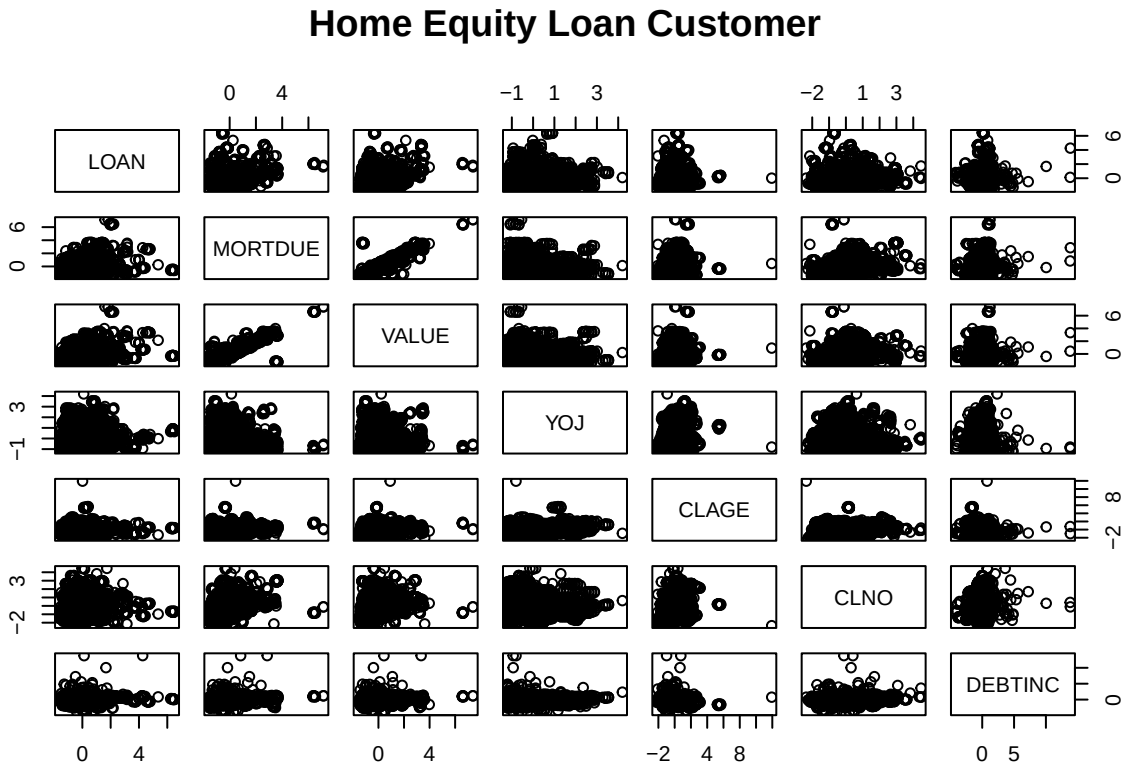
```
nrow(hmeq_profile)
```

```
## [1] 3364
```

```
# wrangle data
# standardize selected attributes in this dataset
library(dplyr)
hmeq_profile.std<-hmeq_profile %>% mutate_at(vars(-REASON, -JOB, -DEROG, -DELINQ, -NINQ), scale)
head(hmeq_profile.std)
```

```
## # A tibble: 6 x 12
##   LOAN[,1] MORTDUE[,1] VALUE[,1] REASON  JOB    YOJ[,1] DEROG DELINQ CLAGE[,1]
##      <dbl>       <dbl>     <dbl> <chr>   <chr>    <dbl> <dbl>  <dbl>     <dbl>
## 1    -1.60      -1.01    -1.23   HomeImp Other  -0.0145     0      0    -0.961
## 2    -1.60      -1.06    -1.18   HomeImp Other   0.249      0      0    -1.11
## 3    -1.55       0.579    0.246  HomeImp Office -0.936      0      0    -1.09
## 4    -1.54      -0.918   -1.10   HomeImp Mgr     0.380      0      0    -1.34
## 5    -1.54       0.492    0.177  HomeImp Office -0.673      0      0    -1.05
## 6    -1.49       0.614    0.0914 HomeImp Office -1.07       0      0    -1.03
## # i 3 more variables: NINQ <dbl>, CLNO <dbl[,1]>, DEBTINC <dbl[,1]>
```

```
# create the scatterplot matrix showing the relationships within different pairs of two
# attributes.
pairs(select(hmeq_profile.std, LOAN, MORTDUE, VALUE, YOJ, CLAGE, CLNO, DEBTINC),
      main = "Home Equity Loan Customer")
```




```
# remove those records with DEBTINC>10 or CLAGE>10
hmeq_profile.std.filtered<-hmeq_profile.std %>% filter(DEBTINC<10 & CLAGE<10)
#check whether the records are removed.
max(hmeq_profile.std.filtered$DEBTINC)
```

```
## [1] 7.228546
```

```
max(hmeq_profile.std.filtered$CLAGE)
```

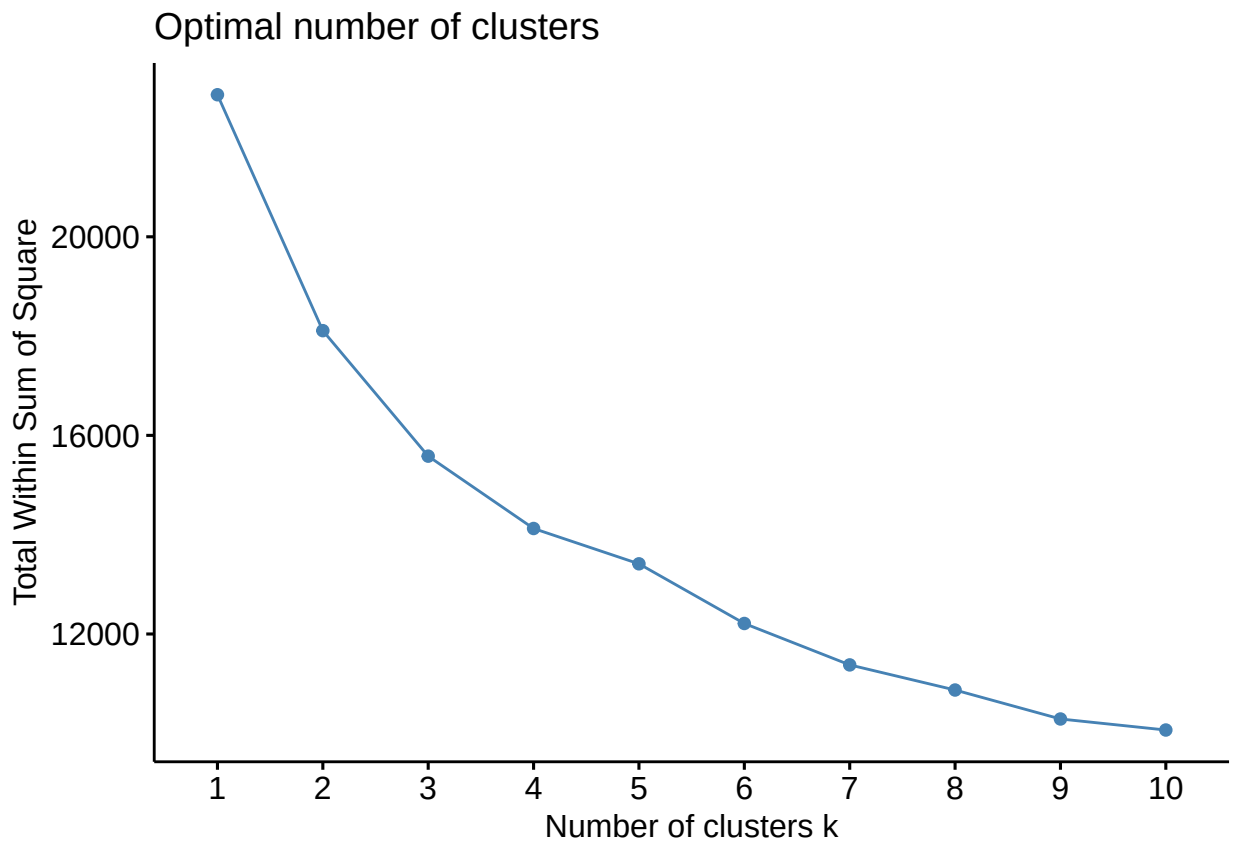
```
## [1] 5.663375
```

We must standardize the selected attributes to ensure they are all of the same value scale

Because Clage and Debtinc both have irregular outliers which fall above 10, much more than any of the values in the other datasets AND much more than any of the values within its respective dataset, if the records identified are not removed, then the clustering process will be impacted in a way that will provide misleading results.)

Analysis

```
# select optimal number of clusters
library(factoextra)
set.seed(2020)
fviz_nbclust(select(hmeq_profile.std.filtered, LOAN, MORTDUE, VALUE, YOJ, CLAGE, CLNO, DEBTINC),
             kmeans, method = "wss")
```




```
hmeq_profile.std.filtered.selected<-hmeq_profile.std.filtered%>%
                                   select(LOAN, MORTDUE, VALUE, YOJ, CLAGE, CLNO, DEBTINC)
set.seed(2020)
hmeq_profile_kmeans <- kmeans(hmeq_profile.std.filtered.selected, centers =6)
hmeq_profile_kmeans$centers
```

```
##          LOAN     MORTDUE       VALUE         YOJ      CLAGE       CLNO
## 1 -0.21155189 -0.49245741 -0.28451748 -0.31416947  1.3040939 -0.4682619
## 2 -0.08643112 -0.46879474 -0.37297061  1.69755725  0.3281573 -0.1942727
## 3  1.16524934  1.68494719  1.87524562  0.01550506  0.4505304  0.2352339
## 4 -0.21048096 -0.62024959 -0.65966130 -0.37274068 -0.7526329 -0.7801163
## 5 -0.26147610 -0.06470187 -0.16529532 -0.52708133 -0.6641182 -0.1746072
## 6 -0.09863087  0.13784575 -0.05614349 -0.10704846  0.3157937  1.5162376
##       DEBTINC
## 1 -0.37126676
## 2 -0.05777951
## 3  0.29347096
## 4 -0.85400799
## 5  0.46599574
## 6  0.11453731
```

```
hmeq_profile_kmeans$size
```

```
## [1] 380 513 456 574 922 515
```

interesting features of some clusters produced from the analysis

In cluster 1 around 380 clients "Clage" values are significantly higher these customers may be older, while all other attributes relatively comparable. Cluster 3 show 456 clients have relatively high loan values, mortgage payments, and property values this can be useful for marketing home equity loans because its seems that loan values are materially lower than property values for these clients, it also looks like the age attributes for cluster 3 shows that the age for this group is about half of cluster 1 (potentially). cluster 4 has the lowest debt to income ratios

```
# read all cluster assignments into a vector called ClusterID
ClusterID<-hmeq_profile_kmeans$cluster
# merge data records in hmeq_profile.std.filtered with their ClusterID
hmeq_profile.std.filtered.K<-cbind(hmeq_profile.std.filtered, ClusterID)
head(hmeq_profile.std.filtered.K)
```

```
##        LOAN    MORTDUE       VALUE  REASON    JOB         YOJ DEROG DELINQ
## 1 -1.604941 -1.0134437 -1.22754515 HomeImp  Other -0.01451784     0      0
## 2 -1.595745 -1.0588142 -1.17795467 HomeImp  Other  0.24875989     0      0
## 3 -1.549770  0.5792252  0.24578927 HomeImp Office -0.93598991     0      0
## 4 -1.540575 -0.9177576 -1.09688135 HomeImp    Mgr  0.38039876     0      0
## 5 -1.540575  0.4922762  0.17712271 HomeImp Office -0.67271218     0      0
## 6 -1.494600  0.6142399  0.09142655 HomeImp Office -1.06762878     0      0
##        CLAGE NINQ       CLNO    DEBTINC ClusterID
## 1 -0.9608358    1 -1.5035319  0.3745450         4
## 2 -1.1142741    0 -1.5035319  0.3457804         4
## 3 -1.0873740    0 -0.9708991 -0.3203139         5
## 4 -1.3350680    1 -0.1186866  0.5191718         5
## 5 -1.0533125    0 -0.9708991 -0.5601048         5
## 6 -1.0256385    0 -0.9708991 -0.5136590         5
```

```
# the "janitor" package has some nice functions to creating table summary that
# play nicely with the %>% pipe in dplyr
library(janitor)
```

```
hmeq_profile.std.filtered.K%>%
  tabyl(ClusterID, REASON, JOB)
```

```
## $Mgr
##  ClusterID DebtCon HomeImp
##          1      18      16
##          2      60      19
##          3      57      17
##          4      96      17
##          5      68      24
##          6      52       6
##
## $Office
##  ClusterID DebtCon HomeImp
##          1      27       7
##          2      33      63
##          3      22      22
##          4      76      12
##          5     183      37
##          6      71      24
##
## $Other
##  ClusterID DebtCon HomeImp
##          1     113      60
##          2     169      89
##          3      74       8
##          4     166      86
##          5     306      70
##          6      99      45
##
## $ProfExe
##  ClusterID DebtCon HomeImp
##          1      72      49
##          2      58      22
##          3     146      58
##          4      70      49
##          5     112      85
##          6     130      47
##
## $Sales
##  ClusterID DebtCon HomeImp
##          1       9       0
##          2       0       0
##          3       8       0
##          4       0       0
##          5      15       7
##          6      13       1
##
## $Self
##  ClusterID DebtCon HomeImp
##          1       0       9
##          2       0       0
##          3      17      27
```

```
##          4        1        1
##          5       15        0
##          6       11       16
```

interesting features of some clusters in terms of REASON and JOB

The Profexe job title is much more evenly distributed across home improvement as the reason for using a home equity loan. Within that same category cluster 3, 5, and 6 have a much higher occurrence of debt consolidation being the reason for a home equity. The "other Job category seems to have highest number of individuals who are using the home equity loan for debt consolidation, but the category"other" can have so many different inputs and much more generic which may not give much utility to the debt consolidation attribution in cluster five for this job title. the sales category seem to be the least insightful across all clusters.