BA-W3-Assignment-112073515

2024-03-09

```r
# Import Data
CCData <- read.csv("HCA Exercise.csv")

# Select the variable
cluster.initial <- data.frame(CCData$BALANCE_FREQUENCY,
                              CCData$ONEOFF_PURCHASES,
                              CCData$CASH_ADVANCE,
                              CCData$CREDIT_LIMIT)
```

Check assumption

```r
# Assumption 1: Representative sample
# Assumption 2: Check outliers
summary(cluster.initial)
```

```
##  CCData.BALANCE_FREQUENCY CCData.ONEOFF_PURCHASES CCData.CASH_ADVANCE
##  Min.   :0.0000           Min.   :    0.0         Min.   :    0
##  1st Qu.:0.8889           1st Qu.:    0.0         1st Qu.:    0
##  Median :1.0000           Median :   38.0         Median :    0
##  Mean   :0.8774           Mean   :  592.5         Mean   :  979
##  3rd Qu.:1.0000           3rd Qu.:  577.8         3rd Qu.: 1114
##  Max.   :1.0000           Max.   :40761.2         Max.   :47137
##  CCData.CREDIT_LIMIT
##  Min.   :   50
##  1st Qu.: 1600
##  Median : 3000
##  Mean   : 4494
##  3rd Qu.: 6500
##  Max.   :30000
```

```r
# Assumption 3: No multicollinearity issues
cor(cluster.initial)
```

```
##                          CCData.BALANCE_FREQUENCY CCData.ONEOFF_PURCHASES
## CCData.BALANCE_FREQUENCY               1.00000000              0.10425675
## CCData.ONEOFF_PURCHASES                0.10425675              1.00000000
## CCData.CASH_ADVANCE                    0.09931213             -0.03134124
## CCData.CREDIT_LIMIT                    0.09584272              0.31972368
##                          CCData.CASH_ADVANCE CCData.CREDIT_LIMIT
## CCData.BALANCE_FREQUENCY          0.09931213          0.09584272
## CCData.ONEOFF_PURCHASES          -0.03134124          0.31972368
## CCData.CASH_ADVANCE               1.00000000          0.30398503
## CCData.CREDIT_LIMIT               0.30398503          1.00000000
```

```r
# Assumption 4: All variables are measured on a metric and comparable scale
str(cluster.initial)
```

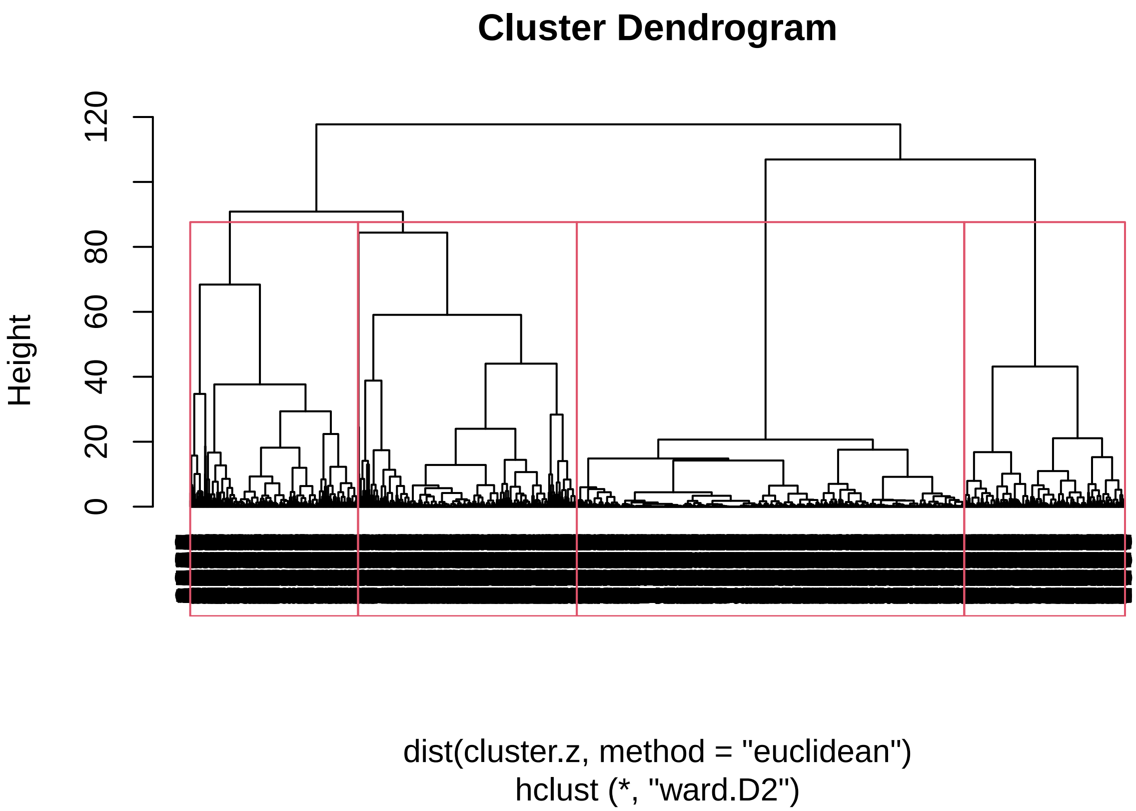
```
## 'data.frame':    8949 obs. of  4 variables:
##  $ CCData.BALANCE_FREQUENCY: num  0.818 0.909 1 0.636 1 ...
##  $ CCData.ONEOFF_PURCHASES : num  0 0 773 1499 16 ...
##  $ CCData.CASH_ADVANCE     : num  0 6443 0 206 0 ...
##  $ CCData.CREDIT_LIMIT     : num  1000 7000 7500 7500 1200 1800 13500 2300 7000 11000 ...
```

Standardized variables and create new data frame with standardized variables

```r
z_BALANCE_FREQUENCY <- scale(cluster.initial$CCData.BALANCE_FREQUENCY)
z_ONEOFF_PURCHASES <- scale(cluster.initial$CCData.ONEOFF_PURCHASES)
z_ADVANCE <- scale(cluster.initial$CCData.CASH_ADVANCE)
z_CREDIT_LIMIT <- scale(cluster.initial$CCData.CREDIT_LIMIT)

cluster.z <- data.frame(z_BALANCE_FREQUENCY,
                        z_ONEOFF_PURCHASES,
                        z_ADVANCE,
                        z_CREDIT_LIMIT)
```

Conducting hierarchical cluster analysis

```r
# Calculate distance matrix. Then conduct hierarchical cluster analysis
hcc <- hclust(dist(cluster.z, method = "euclidean"), method = "ward.D2")

# Plot dendogram
plot(hcc,hang= -1)

# Setting the number of groups and draw dendogram with red borders around the nGroup clusters
nGroup <- 4
rect.hclust(hcc, k=nGroup)
```




```r
# Cut tree into nGroup clusters
groups <- cutree(hcc, k=nGroup)

# Add groups to my dataset
hcc.groups <- cbind(CCData, groups)

#export new dataset
write.csv(hcc.groups, "hcc.groups4.csv")
```

Label	Groups(count)	BF(AVG)	OP(AVG)	CA(AVG)	CL(AVG)
1	3709	0.984153372	228.5862739	277.052222	2126.211917
2	1607	0.947169346	348.5761481	4218.316812	6358.058494
3	2094	0.952712311	1722.634097	261.6481505	8337.744204
4	1539	0.44450991	186.5694022	264.0653468	3026.687308
-					
Amount	8949	0.877350132	592.5035725	978.959616	4494.44945

Groups(count): Display the number of people in each group.
BF(AVG): Average per group of the variable BALANCE_FREQUENCY.
OP(AVG): Average per group of the variable ONEOFF_PURCHASES.
CA(AVG): Average per group of the variable CASH_ADVANCE.
CL(AVG): Average per group of the variable CREDIT_LIMIT.

Interpretation

I first took four variables that I thought could be included in the case, and then made sure that the sample representativeness, covariance, outliers, and data attributes were all integers or numeric values. Then I standardized the data of each variable and classified them into four clusters by cluster analysis method after experimentation, and finally produced this pivot table.

Task3: Identify and examine the characteristics of one specific cluster that emerges from the analysis, shedding light on its unique traits and behaviors.

Cluster 1: The largest number of people with the most frequent balance updates and the lowest credit card limits indicate that this group may make low-value purchases more frequently.

Cluster 2: With the highest average funding requirement, this group may be the most in need of lending services.

Cluster 3: This group has the highest spending and credit card limit, so it can be judged that the spending power of this group is the highest among all groups.

Cluster 4: This group has the lowest number of credit card purchases and is not included in the credit card marketing strategy for the time being.

Task4: Assuming the role of a data analyst within Horizon Financial, propose actionable strategies based on the insights gained from the hierarchical cluster analysis to enhance customer services and drive profitability within the credit card division.

As mentioned above, because Cluster 4 was determined to be a less frequent credit card user, a strategy was developed for Cluster 1 through Cluster 3 groups.

Cluster 1: Focusing on the promotion of credit cards that offer more rewards the more times they buy, and cooperating with low-priced channels.

Cluster 2: Offering low interest rate options on loans while relaxing the restrictions on the loan period.

Cluster 3: For this group, we offer services that are only available to those with high spending power, focusing on high rebates on spending, exclusive discounts at special stores, and priority financial services.