Assignment4.R

Jason

2021-03-21

```
#Assignment 4 Fundamentals of Machine Learning
#Data comes From Pharmaceuticals.csv

library(utils)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(class)
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

```
library(FNN)
```

```
##
## Attaching package: 'FNN'

## The following objects are masked from 'package:class':
##
##     knn, knn.cv
```

```
library(e1071)
library(reshape2)

WD<-setwd("C:/Users/Jason/Documents/MSBA/Fundamentals for Machine Learning/Assignment4")
Drugs<-read.csv("Pharmaceuticals.csv", header = TRUE)
```

```
DrugNum <- Drugs[ , c(3, 4, 5, 6, 7, 8, 9, 10, 11)]

NormDN <- scale(DrugNum) #z-score normalization

#using 4 clusters based on the 4 median recommendations in the data
Kclus <- kmeans(NormDN, centers = 4, nstart = 10)

DrugNum$Cluster <-Kclus$cluster #adding cluster back to dataset

DrugNum$MedRec <- Drugs[, c(12)] #adding variables 10-12 back to dataset
DrugNum$Location <-Drugs[ , c(13)]
DrugNum$Exchange <- Drugs[ , c(14)]

Df <- DrugNum[order(DrugNum$Cluster),] #ordering Data by cluster to spot trends
Df
```

```
##    Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 5       47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## 8        0.41 0.85     26.0 24.1  4.3            0.6     3.51       6.38
## 9        0.78 1.08      3.6 15.1  5.1            0.3     1.07      34.21
## 12       2.60 0.65     19.9 21.4  6.8            0.6     1.45      13.99
## 14       1.20 0.75     28.6 11.2  5.4            0.3     0.93      30.37
## 20       3.26 0.24     18.4 10.2  6.8            0.5     0.20      29.18
## 2        7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## 6       16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
## 18      56.24 0.40     56.5 13.5  5.7            0.6     0.35      15.00
## 1       68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## 3        6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## 4       67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## 7       51.33 0.50     13.9 34.8 15.1            0.9     0.57       2.70
## 10      73.84 0.18     27.9 31.0 13.5            0.6     0.53       6.21
## 16      96.65 0.19     21.6 17.9 11.2            0.5     0.06      -2.69
## 19      34.10 0.51     18.9 22.6 13.3            0.8     0.00       8.56
## 21      48.19 0.63     13.1 54.9 13.4            0.6     1.12       0.36
## 11     122.11 0.35     18.0 62.9 20.3            1.0     0.34      21.87
## 13     173.93 0.46     28.4 28.6 16.3            0.9     0.10       9.37
## 15     132.56 0.46     18.9 40.6 15.0            1.1     0.28      17.35
## 17     199.47 0.65     23.6 45.6 19.2            0.8     0.16      25.54
##    Net_Profit_Margin Cluster        MedRec    Location Exchange
## 5                 12.9       1  Moderate Buy      FRANCE     NYSE
## 8                  7.5       1  Moderate Buy          US   NASDAQ
## 9                 13.3       1 Moderate Sell     IRELAND     NYSE
## 12                11.0       1          Hold          US     AMEX
## 14                21.3       1  Moderate Buy          US     NYSE
## 20                15.1       1 Moderate Sell          US     NYSE
## 2                  5.5       2  Moderate Buy      CANADA     NYSE
## 6                  2.6       2          Hold     GERMANY     NYSE
## 18                 7.3       2          Hold          US     NYSE
## 1                 16.1       3  Moderate Buy          US     NYSE
## 3                 11.2       3    Strong Buy          UK     NYSE
## 4                 18.0       3 Moderate Sell          UK     NYSE
## 7                 20.6       3 Moderate Sell          US     NYSE
## 10                23.4       3          Hold          US     NYSE
## 16              22.4       3          Hold SWITZERLAND     NYSE
## 19              17.6       3          Hold          US     NYSE
## 21              25.5       3          Hold          US     NYSE
## 11              21.1       4          Hold          UK     NYSE
## 13              17.9       4  Moderate Buy          US     NYSE
## 15              14.1       4          Hold          US     NYSE
## 17              25.2       4  Moderate Buy          US     NYSE
```

```
#The clusters are largely influenced by Market_Cap, ROE, and ROA, and Asset Turnover

GD <- Df %>% group_by(Cluster, MedRec) %>% count(MedRec,  name = "count") #grouping by cluster and Medi
GD
```

```
## # A tibble: 11 x 3
## # Groups:   Cluster, MedRec [11]
##    Cluster MedRec        count
##      <int> <chr>         <int>
##  1       1 Hold              1
##  2       1 Moderate Buy      3
##  3       1 Moderate Sell     2
##  4       2 Hold              2
##  5       2 Moderate Buy      1
##  6       3 Hold              4
##  7       3 Moderate Buy      1
##  8       3 Moderate Sell     2
##  9       3 Strong Buy        1
## 10       4 Hold              2
## 11       4 Moderate Buy      2
```

```
#no apparent correlation

ED <- Df %>% group_by(Cluster, Exchange) %>% count(Exchange,  name = "count") #grouping by cluster and
ED
```

```
## # A tibble: 6 x 3
## # Groups:   Cluster, Exchange [6]
##   Cluster Exchange count
##     <int> <chr>    <int>
## 1       1 AMEX         1
## 2       1 NASDAQ       1
## 3       1 NYSE         4
## 4       2 NYSE         3
## 5       3 NYSE         8
## 6       4 NYSE         4
```

```
#no apparent correlation

CD <- Df %>% group_by(Cluster, Location) %>% count(Location,  name = "count") #grouping by cluster and
CD
```

```
## # A tibble: 11 x 3
## # Groups:   Cluster, Location [11]
##    Cluster Location    count
##      <int> <chr>       <int>
##  1       1 FRANCE          1
##  2       1 IRELAND         1
##  3       1 US              4
##  4       2 CANADA          1
##  5       2 GERMANY         1
##  6       2 US              1
##  7       3 SWITZERLAND     1
##  8       3 UK              2
##  9       3 US              5
## 10       4 UK              1
## 11       4 US              3
```

```
#no apparent correlation

#there is no pattern between the clusters and variables 10-12

Df2<-Df[, c(1:10)]
clusNames <- Df2 %>% group_by(Cluster) %>% summarize(across(everything(), list(mean)))
clusNames
```

```
## # A tibble: 4 x 10
##   Cluster Market_Cap_1 Beta_1 PE_Ratio_1 ROE_1 ROA_1 Asset_Turnover_1 Leverage_1
## *   <int>        <dbl>  <dbl>      <dbl> <dbl> <dbl>            <dbl>      <dbl>
## 1       1         9.24  0.648       19.4  17.3  5.98            0.483      1.25
## 2       2        26.9   0.64        55.6  10.1  4.2             0.7        0.317
## 3       3        55.8   0.414       20.3  28.7 12.7             0.738      0.371
## 4       4       157.    0.48        22.2  44.4 17.7             0.95       0.22
## # ... with 2 more variables: Rev_Growth_1 <dbl>, Net_Profit_Margin_1 <dbl>
```

```
#Cluster 1 = "Large_Cap, Large return on Investments" ROE and ROA is highest
#Cluster 2 = "Medium_Cap, Medium Return on Investments" ROE and ROA is 2nd highest
#Cluster 3 = "Small-Medium Cap, Small Return on Investments"ROE and ROA is worst
#Cluster 4 = "Micro Cap, Medium Return on Investments" ROE and ROA is 3rd

#Cluster 3 is worst performing cluster
```