Assignment_4_Final

Nawwaf Albahar

2022-11-03

```
# load and preprocess data
Pharmaceuticals <- read_csv("/Users/nawwaf/Desktop/Kent/Kent Master_s/Machine
Learning/Pharmaceuticals.csv")
Pharmaceuticals
```

```
## # A tibble: 21 × 14
##    Symbol Name         Marke…¹  Beta PE_Ra…²   ROE   ROA Asset…³ Lever…⁴
Rev_G…⁵
##    <chr>  <chr>          <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl>
<dbl>
##  1 ABT    Abbott Labo…   68.4   0.32    24.7  26.4  11.8     0.7    0.42
7.54
##  2 AGN    Allergan, I…    7.58  0.41    82.5  12.9   5.5     0.9    0.6
9.16
##  3 AHM    Amersham plc    6.3   0.46    20.7  14.9   7.8     0.9    0.27
7.05
##  4 AZN    AstraZeneca…   67.6   0.52    21.5  27.4  15.4     0.9    0
15
##  5 AVE    Aventis        47.2   0.32    20.1  21.8   7.5     0.6    0.34
26.8
##  6 BAY    Bayer AG       16.9   1.11    27.9   3.9   1.4     0.6    0
-3.17
##  7 BMY    Bristol-Mye…   51.3   0.5     13.9  34.8  15.1     0.9    0.57
2.7
##  8 CHTT   Chattem, Inc    0.41  0.85    26    24.1   4.3     0.6    3.51
6.38
##  9 ELN    Elan Corpor…    0.78  1.08     3.6  15.1   5.1     0.3    1.07
34.2
## 10 LLY    Eli Lilly a…   73.8   0.18    27.9  31    13.5     0.6    0.53
6.21
## # … with 11 more rows, 4 more variables: Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>, and
## #   abbreviated variable names ¹Market_Cap, ²PE_Ratio, ³Asset_Turnover,
## #   ⁴Leverage, ⁵Rev_Growth
```

```
#Keep seprate to use latter for labeling to see if there is a pattern
Pharmaceuticals_Label <- Pharmaceuticals$Median_Recommendation
table(Pharmaceuticals_Label)
```

```
## Pharmaceuticals_Label
##          Hold  Moderate Buy Moderate Sell    Strong Buy
##             9             7             4             1
```

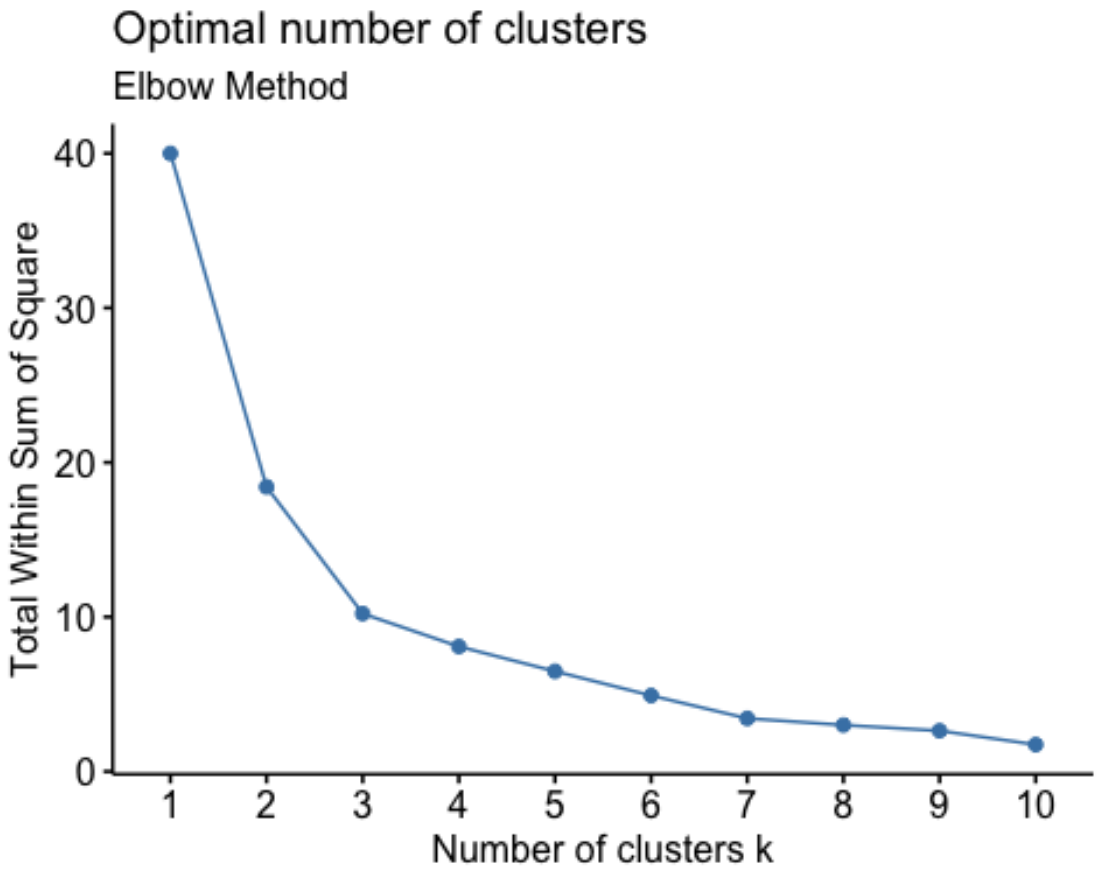
```
#Take the numircal columns only
Pharmaceuticals_Data <-- Pharmaceuticals[,3:11]
```

a. Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.

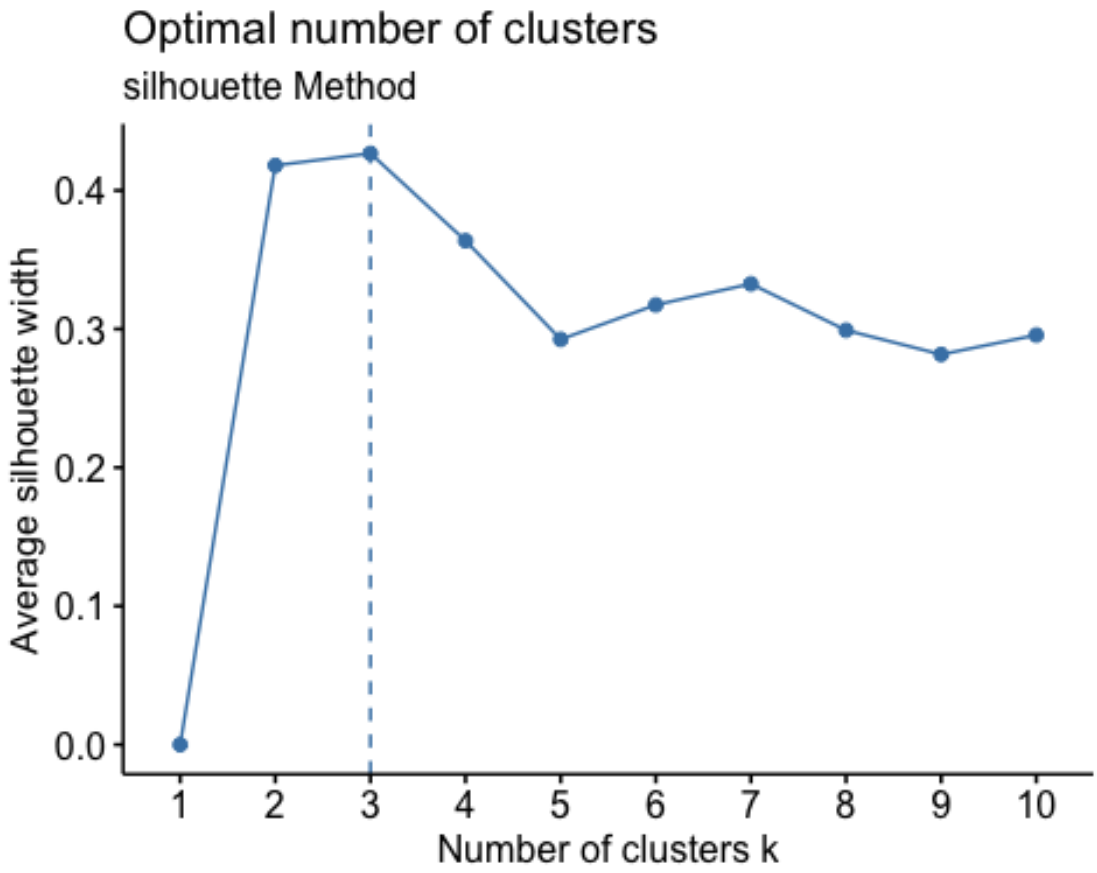
```
#assign the columns that we want to cluster by to the variable
Pharmaceuticals_Clustering_Columns
Pharmaceuticals_Clustering_Columns <- Pharmaceuticals_Data[,c(1,9)]
#Renove columns that we want to cluster by from the dataset
#Pharmaceuticals_Data <- Pharmaceuticals_Data[,-1]
#Pharmaceuticals_Data <- Pharmaceuticals_Data[,-8]

#normalize the dataset after removing the cluster we want to cluster by
Normalized_Pharmaceuticals <- sapply(Pharmaceuticals_Clustering_Columns,
scale)

#Calculate the distance
Normalized_Pharmaceuticals_Dist <- dist(Normalized_Pharmaceuticals)
Normalized_Pharmaceuticals_Dist
```

```
##            1         2         3         4         5         6         7
## 2  1.9203815
## 3  1.2968974 0.8689058
## 4  0.2898735 2.1630099 1.4727771
## 5  0.6079987 1.3144789 0.7438083 0.8520825
## 6  2.2373870 0.4696800 1.3229895 2.5013922 1.6523825
## 7  0.7453270 2.4191831 1.6255531 0.4840992 1.1755696 2.8052581
## 8  1.7507675 0.3284232 0.5727361 1.9687701 1.1461144 0.7979759 2.1772251
## 9  1.2308750 1.1943035 0.3335963 1.3469441 0.7937716 1.6536272 1.4076980
## 10 1.1162673 2.9528497 2.1874171 0.8297085 1.6636167 3.3153154 0.5741169
## 11 1.1913490 3.0775029 2.4862399 1.0427924 1.7880791 3.3423394 1.2101899
## 12 1.3661150 0.8424498 0.0701083 1.5392544 0.8136337 1.3031381 1.6827610
## 13 1.8208597 3.4100450 3.0371994 1.8139656 2.2934731 3.5519482 2.1321227
## 14 1.3944319 2.4102448 1.5416119 1.2401046 1.5012297 2.8622853 0.8620432
## 15 1.1357996 2.5031599 2.1993570 1.2573025 1.4686935 2.6394014 1.7036643
## 16 1.0739873 2.9904561 2.3000102 0.8335587 1.6760263 3.3100340 0.8205470
## 17 2.6310323 4.4423441 3.9264489 2.5030200 3.2044207 4.6440282 2.6232559
## 18 1.3571072 0.8744694 1.0389548 1.6421309 0.8673435 0.9816513 2.0285372
## 19 0.6289844 1.8986551 1.0845532 0.5753934 0.7501094 2.3046370 0.5435547
## 20 1.1226219 1.4648139 0.5965862 1.1839799 0.8207096 1.9190600 1.1727473
## 21 1.4735670 3.1256145 2.2934376 1.1901021 1.9202126 3.5303735 0.7486380
##            8         9        10        11        12        13        14
## 2
## 3
## 4
## 5
## 6
## 7
```

```
## 8
## 9  0.8838929
## 10 2.7278334 1.9807174
## 11 2.9339364 2.3873295 0.8951518
## 12 0.5346776 0.3518735 2.2469019 2.5549573
## 13 3.3584218 3.0366485 1.9025088 1.0098085 3.1069165
## 14 2.1030451 1.2191526 1.2801732 2.0634315 1.5698137 2.9926564
## 15 2.4691397 2.2519942 1.7356753 1.0815415 2.2673953 0.9130660 2.4956608
## 16 2.8022700 2.1446112 0.4179985 0.4774862 2.3650965 1.4863549 1.6373608
## 17 4.3374576 3.8449623 2.1612265 1.4604676 3.9960224 1.1947785 3.4350811
## 18 0.9531697 1.3159206 2.4718149 2.3845343 1.0750448 2.5772976 2.3310594
## 19 1.6430116 0.8675657 1.1140365 1.5937035 1.1404049 2.3864949 0.7956762
## 20 1.1591956 0.2775498 1.7465293 2.2246435 0.6249064 2.9434029 0.9454806
## 21 2.8616505 2.0275652 0.5422063 1.4285146 2.3426202 2.4382530 1.0259622
##           15        16        17        18        19        20
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16 1.4054633
## 17 2.0408137 1.8056595
## 18 1.6642970 2.4022069 3.6625716
## 19 1.7627497 1.2939483 3.0502993 1.6144580
## 20 2.2116297 1.9434866 3.6849601 1.4933864 0.6496824
## 21 2.2562765 0.9523509 2.5818444 2.7769239 1.2276662 1.7605743

#We need to decide the number of cluster
#This help in determining what K should be.
#It looks like 6 is the elbow as it curves upwaerd a little after it.
Nonetheless, It is not clear still
fviz_nbclust(Normalized_Pharmaceuticals, kmeans, method = "wss") +
labs(subtitle = "Elbow Method")
```




```
#if elbow method is not clear there is Another method for determining the K
value which is  silhouette method:
fviz_nbclust(Normalized_Pharmaceuticals, kmeans, method = "silhouette") +
labs(subtitle = "silhouette Method")
```



a. Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.

K-Means and euclidean distance is used for clustering as it is centroid-based clustering algorithm and Centroid-based algorithms are efficient and simple in clustering numircal data into sub-groups that share similar characteristics. More weight is given to the chosen columns that we are going to cluster by and that is by eliminating the rest of the columns. the number of cluster is decided by either the elbow method or silhouette as shown above 3 is the elbow area and the silhouette confrims that as well

```
#Now we run the Kmeans algorithm with K that we got from either of the
methods and 100 iterations to cluster our data
km <- kmeans(Normalized_Pharmaceuticals, centers = 3, nstart = 100)

#print the result of kmeans algortihm
km

## K-means clustering with 3 clusters of sizes 4, 9, 8
##
## Cluster means:
##   Market_Cap Net_Profit_Margin
```

```
## 1 -1.6955811        -0.5912425
## 2  0.7159913         0.9288621
## 3  0.0423004        -0.7493486
##
## Clustering vector:
##  [1] 3 2 2 3 2 2 3 2 2 3 1 2 1 3 1 3 1 2 3 2 3
##
## Within cluster sum of squares by cluster:
## [1] 2.687129 4.171884 3.357378
##  (between_SS / total_SS =  74.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

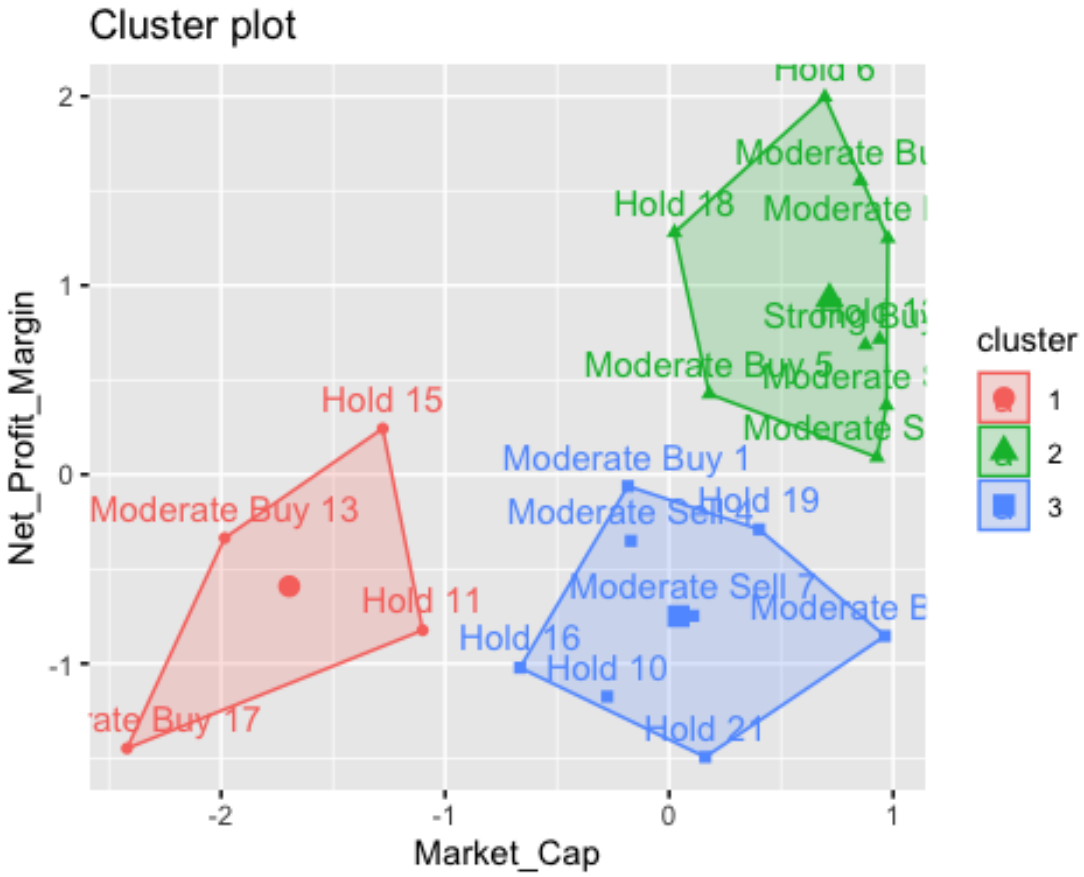
```
#Now assign the label back so that we see if there is a pattern based on it
rownames(Normalized_Pharmaceuticals) <-
paste(Pharmaceuticals_Label,1:dim(Pharmaceuticals)[1])
rownames(Normalized_Pharmaceuticals)
```

```
##  [1] "Moderate Buy 1"   "Moderate Buy 2"   "Strong Buy 3"     "Moderate
Sell 4"
##  [5] "Moderate Buy 5"   "Hold 6"           "Moderate Sell 7"  "Moderate
Buy 8"
##  [9] "Moderate Sell 9"  "Hold 10"          "Hold 11"          "Hold 12"
## [13] "Moderate Buy 13"  "Moderate Buy 14"  "Hold 15"          "Hold 16"
## [17] "Moderate Buy 17"  "Hold 18"          "Hold 19"          "Moderate
Sell 20"
## [21] "Hold 21"
```

b. Interpret the clusters with respect to the numerical variables used in forming the clusters.

The green cluster is those companies that have large market cap and high net profit. The blue one is those companies that also have medium to large market cap but they have low net profit. The red cluster is those compaines that have small market cap as well as low net profit.

```
#Vizualize the clusters
fviz_cluster(km, data = Normalized_Pharmaceuticals)
```



c. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

We don't see any clear pattern with the regards to variable 10 to 12, we used Median recommendation (across major brokerages) as label to illustrate that there is no clear pattern as can be seen below

d. Provide an appropriate name for each cluster using any or all of the variables in the dataset.

The green cluster is those companies that have large market cap and high net profit, suggested name is Stars
The blue one is those companies that also have large market cap but they have low net profit, suggested name is cash cow The red cluster is those compaines that have small market cap as well as low net profit, suggested name is dog

The suggested names for the clusters is inspired from the BCG matrix