Assignment_4

Kristen Durkin

Git Hub Link: https://github.com/kdurkin5/64060-002-kdurkin5/tree/be54afd8723d5c0de8f8e9efa6e6599b7a45d1ad/Assignment4

Load Data

```
df <- read.csv("Pharmaceuticals.csv", stringsAsFactors = FALSE)
head(df)
```

```
##   Symbol                Name Market_Cap Beta PE_Ratio  ROE  ROA
Asset_Turnover
## 1    ABT Abbott Laboratories      68.44 0.32     24.7 26.4 11.8
0.7
## 2    AGN      Allergan, Inc.       7.58 0.41     82.5 12.9  5.5
0.9
## 3    AHM        Amersham plc       6.30 0.46     20.7 14.9  7.8
0.9
## 4    AZN     AstraZeneca PLC      67.63 0.52     21.5 27.4 15.4
0.9
## 5    AVE             Aventis      47.16 0.32     20.1 21.8  7.5
0.6
## 6    BAY            Bayer AG      16.90 1.11     27.9  3.9  1.4
0.6
##   Leverage Rev_Growth Net_Profit_Margin Median_Recommendation Location
Exchange
## 1     0.42       7.54              16.1          Moderate Buy       US
NYSE
## 2     0.60       9.16               5.5          Moderate Buy   CANADA
NYSE
## 3     0.27       7.05              11.2            Strong Buy       UK
NYSE
## 4     0.00      15.00              18.0         Moderate Sell       UK
NYSE
## 5     0.34      26.81              12.9          Moderate Buy   FRANCE
NYSE
## 6     0.00      -3.17               2.6                  Hold  GERMANY
NYSE
```

```
#select numeric variables (1-9)
data_num <- df[,1:9]
# take all the rows but only the first 9 columns from the data set.
summary (data_num)
```

```
##     Symbol              Name             Market_Cap          Beta
##  Length:21          Length:21          Min.   :  0.41   Min.   :0.1800
```

```
##  Class :character   Class :character   1st Qu.:  6.30   1st Qu.:0.3500
##  Mode  :character   Mode  :character   Median : 48.19   Median :0.4600
##                                        Mean   : 57.65   Mean   :0.5257
##                                        3rd Qu.: 73.84   3rd Qu.:0.6500
##                                        Max.   :199.47   Max.   :1.1100
##     PE_Ratio          ROE            ROA        Asset_Turnover    Leverage
##  Min.   : 3.60   Min.   : 3.9   Min.   : 1.40   Min.   :0.3    Min.
:0.0000
##  1st Qu.:18.90   1st Qu.:14.9   1st Qu.: 5.70   1st Qu.:0.6    1st
Qu.:0.1600
##  Median :21.50   Median :22.6   Median :11.20   Median :0.6    Median
:0.3400
##  Mean   :25.46   Mean   :25.8   Mean   :10.51   Mean   :0.7    Mean
:0.5857
##  3rd Qu.:27.90   3rd Qu.:31.0   3rd Qu.:15.00   3rd Qu.:0.9    3rd
Qu.:0.6000
##  Max.   :82.50   Max.   :62.9   Max.   :20.30   Max.   :1.1    Max.
:3.5100

#check to make sure the data is looking correct

# Keep only numeric columns
data_numeric_only <- df[, sapply(df, is.numeric)]

# Scale the numeric data
data_scaled <- scale(data_numeric_only)
#This standardized the numeric data with a mean of 0 and standard deviation
of 1

# Show a few rows of the scaled data
head(data_scaled)

##      Market_Cap        Beta    PE_Ratio         ROE        ROA
Asset_Turnover
## [1,]  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121
0.0000000
## [2,] -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871
0.9225312
## [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700
0.9225312
## [4,]  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259
0.9225312
## [5,] -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -
0.4612656
## [6,] -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -
0.4612656
##        Leverage Rev_Growth Net_Profit_Margin
## [1,] -0.2120979 -0.5277675        0.06168225
## [2,]  0.0182843 -0.3811391       -1.55366706
## [3,] -0.4040831 -0.5721181       -0.68503583
```

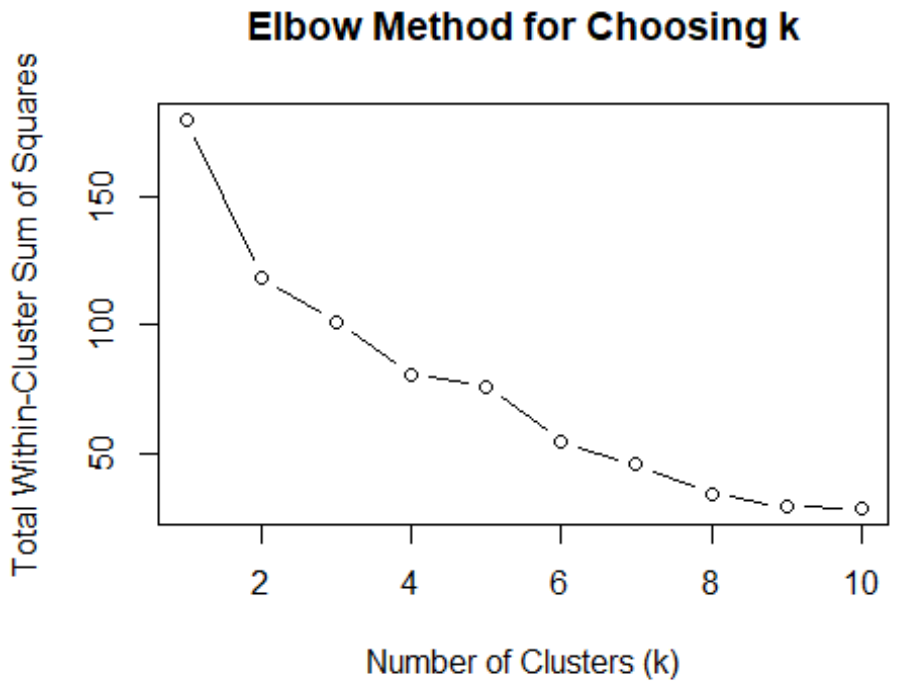
```
## [4,] -0.7496565  0.1474473        0.35122600
## [5,] -0.3144900  1.2163867       -0.42597037
## [6,] -0.7496565 -1.4971443       -1.99560225
```

#choosing number of clusters with Elbow Method #This method helps determine the best number of clusters (k) for K-Means. #It calculates how much total variation exists within clusters (WSS) for different k values. # The “elbow” of the plot should show where adding more clusters doesn’t improve much.

```
wss <- numeric(10)
#Within-Cluster Sum of Squares will create an empty numeric vector to store
results

# Loop through cluster numbers 1 to 10.
# For each k, run K-Means and record the total within-cluster sum of squares.
for (k in 1:10) {
  km <- kmeans(data_scaled, centers = k)
  wss[k] <- km$tot.withinss
  # This runs the K-Means algorithm using the scaled data and will test how
close the clustering is
}

# Plot the elbow chart
plot(1:10, wss, type = "b", main = "Elbow Method for Choosing k",
     xlab = "Number of Clusters (k)",
     ylab = "Total Within-Cluster Sum of Squares")
```



```
#This draws the elbow chart
#The chart shows a big drop from 1-4 but after the improvements become
smaller indicating diminishing returns
```

#Run K-Means w/ Chosen k

#Since the elbow plot suggests k = 4, we'll run the K-Means algorithm.

```
set.seed(8)
k <- 4
km_result <- kmeans(data_scaled, centers = k, nstart = 25)

#set.seed(8) Ensures reproducibility
# 'nstart = 25' runs the algorithm 25 times with different random starting
points to make sure the result is stable and reliable.

# Display the main results from the K-Means model.
km_result
```

```
## K-means clustering with 4 clusters of sizes 4, 3, 8, 6
##
## Cluster means:
##    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431   1.153164e+00
## 2 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838   1.480297e-16
## 3 -0.03142211 -0.4360989 -0.3172485  0.1950459  0.4083915   1.729746e-01
## 4 -0.82617719  0.4775991 -0.3696184 -0.5631589 -0.8514589  -9.994088e-01
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.4680782  0.4671788         0.5912425
## 2 -0.3443544 -0.5769454        -1.6095439
## 3 -0.2744931 -0.7041516         0.5569544
## 4  0.8502201  0.9158889        -0.3319956
##
## Clustering vector:
##  [1] 3 2 3 3 4 2 3 4 4 3 1 4 1 4 1 3 1 2 3 4 3
##
## Within cluster sum of squares by cluster:
## [1]  9.284424 14.938904 21.879320 32.143356
##  (between_SS / total_SS =  56.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

#K-means clustering with 4 clusters of company sizes (1) 4,(2) 3, (3) 8, (4) 6

#cluster 1 is showing large, efficient and profitable companies. They look to have avg returns but a healthy profit margin

#cluster 2 is showing small-to-mid size companies with an high p/e ratio but looks to show low profitability

#cluster 3 is more moderate with average sized companies. They rare trending with a little below average market risk (lower beta)

#cluster 4 are smaller but leveraged companies. They have higher growth opporunity but lower profit margins.

#SS number = 56.5% meaning the variation is explained by the cluster differences and they are not random.

#Add Cluster Labels to the data

```
df$Cluster <- km_result$cluster
head(df)
```

```
##   Symbol                Name Market_Cap Beta PE_Ratio  ROE  ROA
Asset_Turnover
## 1    ABT Abbott Laboratories      68.44 0.32     24.7 26.4 11.8
0.7
## 2    AGN      Allergan, Inc.       7.58 0.41     82.5 12.9  5.5
0.9
## 3    AHM        Amersham plc       6.30 0.46     20.7 14.9  7.8
0.9
## 4    AZN     AstraZeneca PLC      67.63 0.52     21.5 27.4 15.4
0.9
## 5    AVE             Aventis      47.16 0.32     20.1 21.8  7.5
0.6
## 6    BAY            Bayer AG      16.90 1.11     27.9  3.9  1.4
0.6
##   Leverage Rev_Growth Net_Profit_Margin Median_Recommendation Location
Exchange
## 1     0.42       7.54              16.1          Moderate Buy       US
NYSE
## 2     0.60       9.16               5.5          Moderate Buy   CANADA
NYSE
## 3     0.27       7.05              11.2            Strong Buy       UK
NYSE
## 4     0.00      15.00              18.0         Moderate Sell       UK
NYSE
## 5     0.34      26.81              12.9          Moderate Buy   FRANCE
NYSE
## 6     0.00      -3.17               2.6                  Hold  GERMANY
NYSE
##   Cluster
## 1       3
## 2       2
## 3       3
## 4       3
```

```
## 5       4
## 6       2
```

#We need to know which firm belongs to which cluster so we can compare groups.

Cluster Diagnostics

```
km_result$centers
```

```
##    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431   1.153164e+00
## 2 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838   1.480297e-16
## 3 -0.03142211 -0.4360989 -0.3172485  0.1950459  0.4083915   1.729746e-01
## 4 -0.82617719  0.4775991 -0.3696184 -0.5631589 -0.8514589  -9.994088e-01
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.4680782  0.4671788         0.5912425
## 2 -0.3443544 -0.5769454        -1.6095439
## 3 -0.2744931 -0.7041516         0.5569544
## 4  0.8502201  0.9158889        -0.3319956
```

#Sizes tell us if any cluster is too tiny/huge; centers show the “average” profile of each cluster in standardized units (after scale()).

#cluster means - original values

```
cluster_means_raw <- aggregate(
  x = df[, sapply(df, is.numeric)],
  by = list(Cluster = df$Cluster),
  FUN = mean, na.rm = TRUE
)
cluster_means_raw
```

```
##   Cluster Market_Cap      Beta PE_Ratio     ROE       ROA Asset_Turnover
## 1       1  157.01750 0.4800000 22.22500 44.4250 17.700000      0.9500000
## 2       2   26.90667 0.6400000 55.63333 10.1000  4.200000      0.7000000
## 3       3   55.81000 0.4137500 20.28750 28.7375 12.687500      0.7375000
## 4       4    9.23500 0.6483333 19.43333 17.3000  5.983333      0.4833333
##    Leverage Rev_Growth Net_Profit_Margin Cluster
## 1 0.2200000  18.532500         19.575000       1
## 2 0.3166667   6.996667          5.133333       2
## 3 0.3712500   5.591250         19.350000       3
## 4 1.2500000  23.490000         13.516667       4
```

#Centers above are in z-scores; business interpretation is easier in the original units (ie, Market_Cap billions, margins %, etc.).

Cluster Means - Scaled Values

```
scaled_df <- as.data.frame(data_scaled)
scaled_df$Cluster <- df$Cluster
```

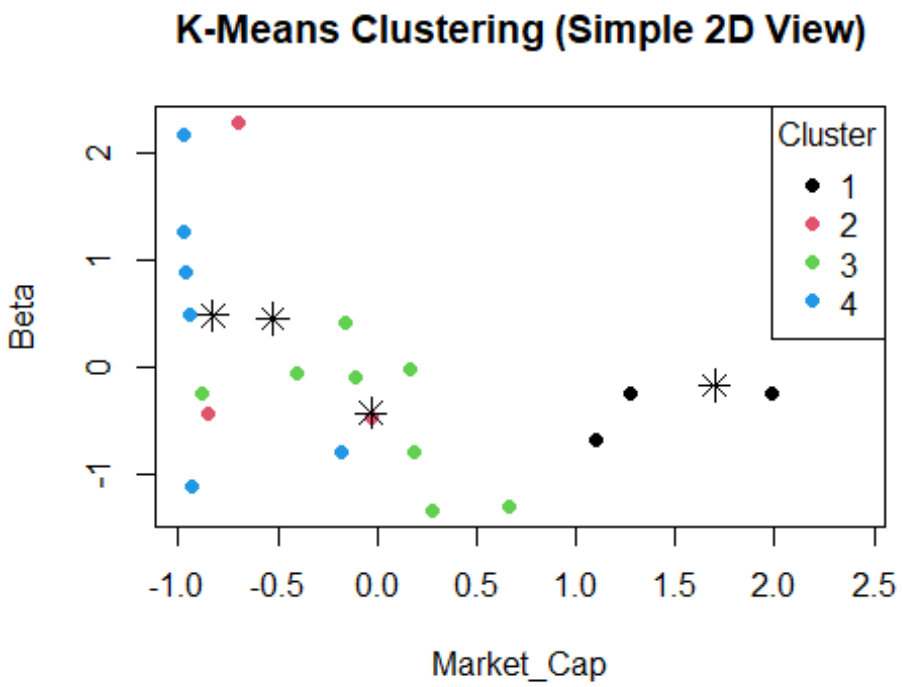
```r
cluster_means_scaled <- aggregate(
  x = scaled_df[, !names(scaled_df) %in% "Cluster"],
  by = list(Cluster = scaled_df$Cluster),
  FUN = mean
)
cluster_means_scaled
```

```
##   Cluster  Market_Cap       Beta   PE_Ratio        ROE        ROA
## 1       1  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431
## 2       2 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838
## 3       3 -0.03142211 -0.4360989 -0.3172485  0.1950459  0.4083915
## 4       4 -0.82617719  0.4775991 -0.3696184 -0.5631589 -0.8514589
##   Asset_Turnover   Leverage Rev_Growth Net_Profit_Margin
## 1   1.153164e+00 -0.4680782  0.4671788         0.5912425
## 2   1.480297e-16 -0.3443544 -0.5769454        -1.6095439
## 3   1.729746e-01 -0.2744931 -0.7041516         0.5569544
## 4  -9.994088e-01  0.8502201  0.9158889        -0.3319956
```

#Z scores show us how far above/below the dataset avg of each cluster sits.

#Cluster Plot

```r
plot(
  data_scaled[, 1], data_scaled[, 2],
  col = df$Cluster,
  pch = 16,
  main = "K-Means Clustering (Simple 2D View)",
  xlab = colnames(data_scaled)[1],
  ylab = colnames(data_scaled)[2]
)
points(km_result$centers[, 1:2], pch = 8, cex = 1.5)
legend("topright", legend = sort(unique(df$Cluster)),
       col = sort(unique(df$Cluster)), pch = 16, title = "Cluster")
```



#plot() creates a simple scatterplot.

#data_scaled[, 1] - this is the first numeric variable in the scaled dataset (on the x-axis).

#data_scaled[, 2] - the second numeric variable (on the y-axis).

#col = df$Cluster - colors each point according to which cluster it belongs to (so points in the same cluster share a color).

#pch = 16 - sets the plotting character (a filled circle).

#main - adds a title to the plot.

#xlab / ylab - automatically labels the axes using the dataset's first two column names.

Quick visual check that clusters are reasonably separated

#Interpreation & Conclusion
The cat() function prints text in the export

```
cat("

** The K-Means analysis grouped the pharmaceutical firms into four distinct
clusters based on their financial performance metrics. Each cluster
represents a unique combination of company size, profitability, leverage, and
growth pattern:
```

```
** Cluster 1 – High-Performer Large Caps:These companies show strong Market Cap, ROE, and ROA, with high efficiency (Asset Turnover) and healthy profit margins. They operate with lower leverage, indicating stability and consistent profitability.

** Cluster 2 – High-Valuation, Low-Profit Firms: Smaller companies with very high P/E ratios but weak profitability (low ROE, ROA) and negative profit margins. These firms may be valued highly by investors but Are still developing stable earnings.

** Cluster 3 – Stable Mid-Range Performers: Moderate in size, with average profitability and slightly below-average market risk.These companies represent balanced, dependable performers with consistent financials.

** Cluster 4 – Leveraged Growth Firms: Smaller firms with high leverage and strong revenue growth potential but thinner profit margins. They may be pursuing aggressive expansion financed by debt, prioritizing growth over immediate profitability.

** Overall: The four-cluster model explains about 56.5% of the total variation, showing that the segmentation captures meaningful financial differences. This classification  helps identify leaders, stable performers, and high-risk/high-reward firms across the pharmaceutical industry."
)
```