Load dataset / libraries

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.2

## Warning: package 'tibble' was built under R version 4.1.2

## Warning: package 'tidyr' was built under R version 4.1.2

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'purrr' was built under R version 4.1.2

## Warning: package 'dplyr' was built under R version 4.1.2

## Warning: package 'stringr' was built under R version 4.1.2

## Warning: package 'forcats' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.3

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.1.2
```

```
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.1.3

## Loading required package: grid
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Warning: package 'modeltools' was built under R version 4.1.1
```

```
## Loading required package: stats4
```

```
Pharma <- read.csv('C:/Users/lmszr/Documents/School/Fundamentals of Machine Learning/Pharmaceuticals.csv
set.seed(123)
```

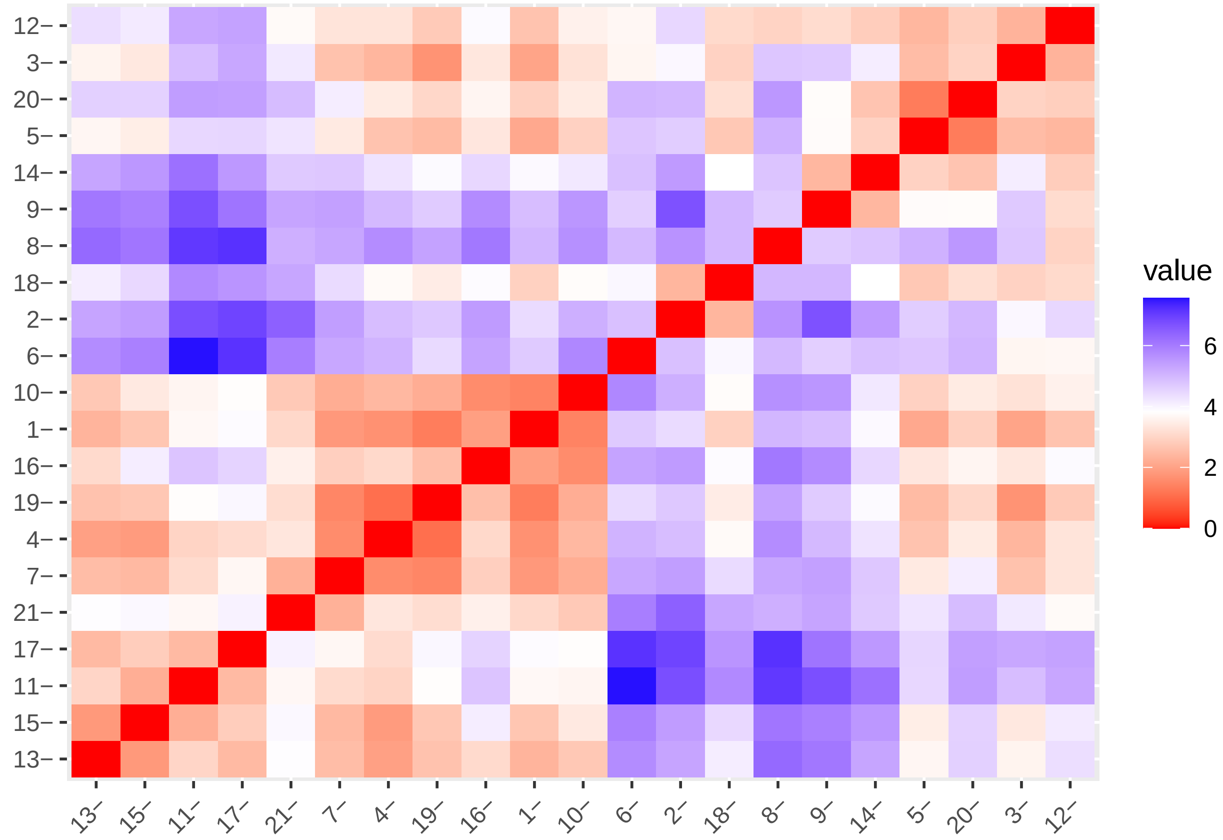
a. Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.

```
Pharma_s <- Pharma[, c(3:11)]
summary(Pharma_s)
```

```
##    Market_Cap          Beta           PE_Ratio          ROE      
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9  
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9  
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6  
##  Mean   : 57.65   Mean   :0.5257   Mean   :25.46   Mean   :25.8  
##  3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90   3rd Qu.:31.0  
##  Max.   :199.47   Max.   :1.1100   Max.   :82.50   Max.   :62.9  
##       ROA        Asset_Turnover    Leverage        Rev_Growth   
##  Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17  
##  1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38  
##  Median :11.20   Median :0.6    Median :0.3400   Median : 9.37  
##  Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37  
##  3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87  
##  Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21  
##  Net_Profit_Margin
##  Min.   : 2.6     
##  1st Qu.:11.2     
##  Median :16.1     
##  Mean   :15.7     
##  3rd Qu.:21.1     
##  Max.   :25.5     
```

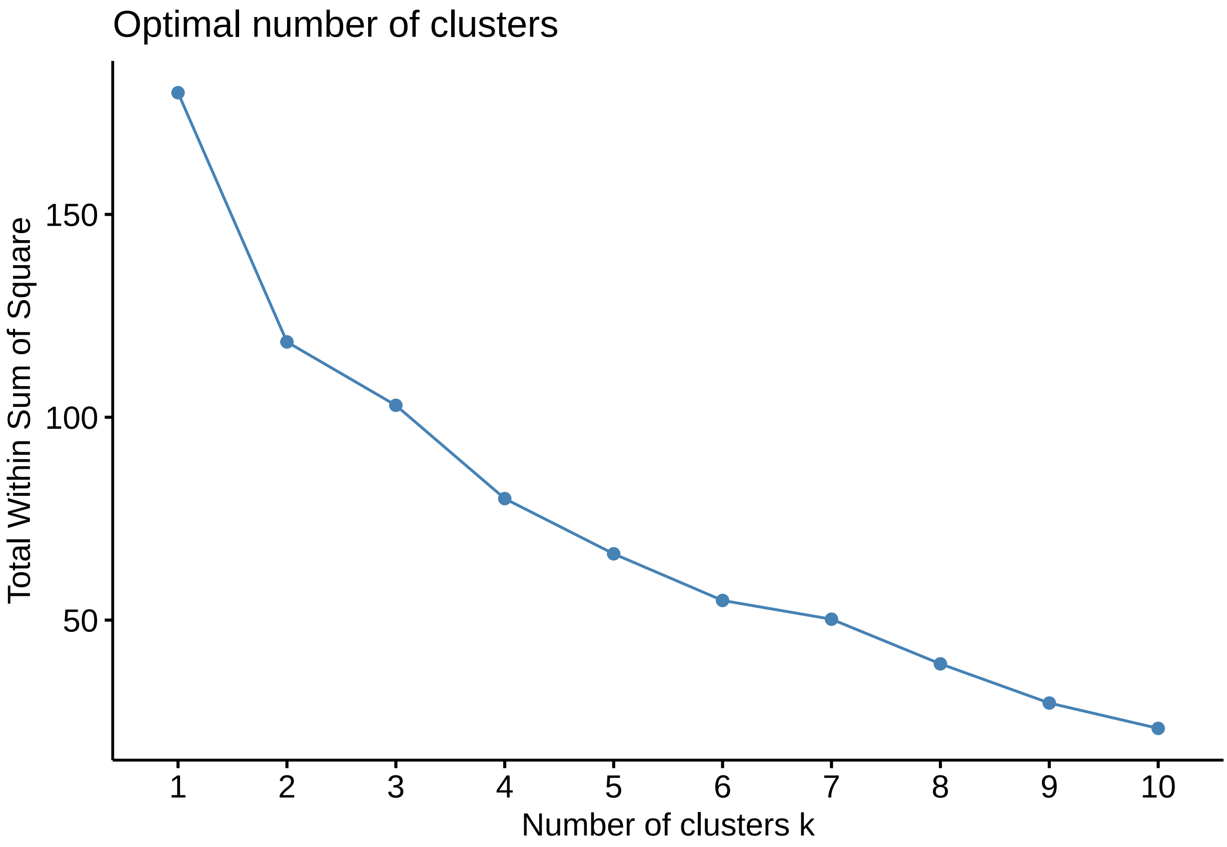
Scaling the data frame

```
Pharma_s <-scale(Pharma_s)
distance <- get_dist(Pharma_s)
fviz_dist(distance)
```

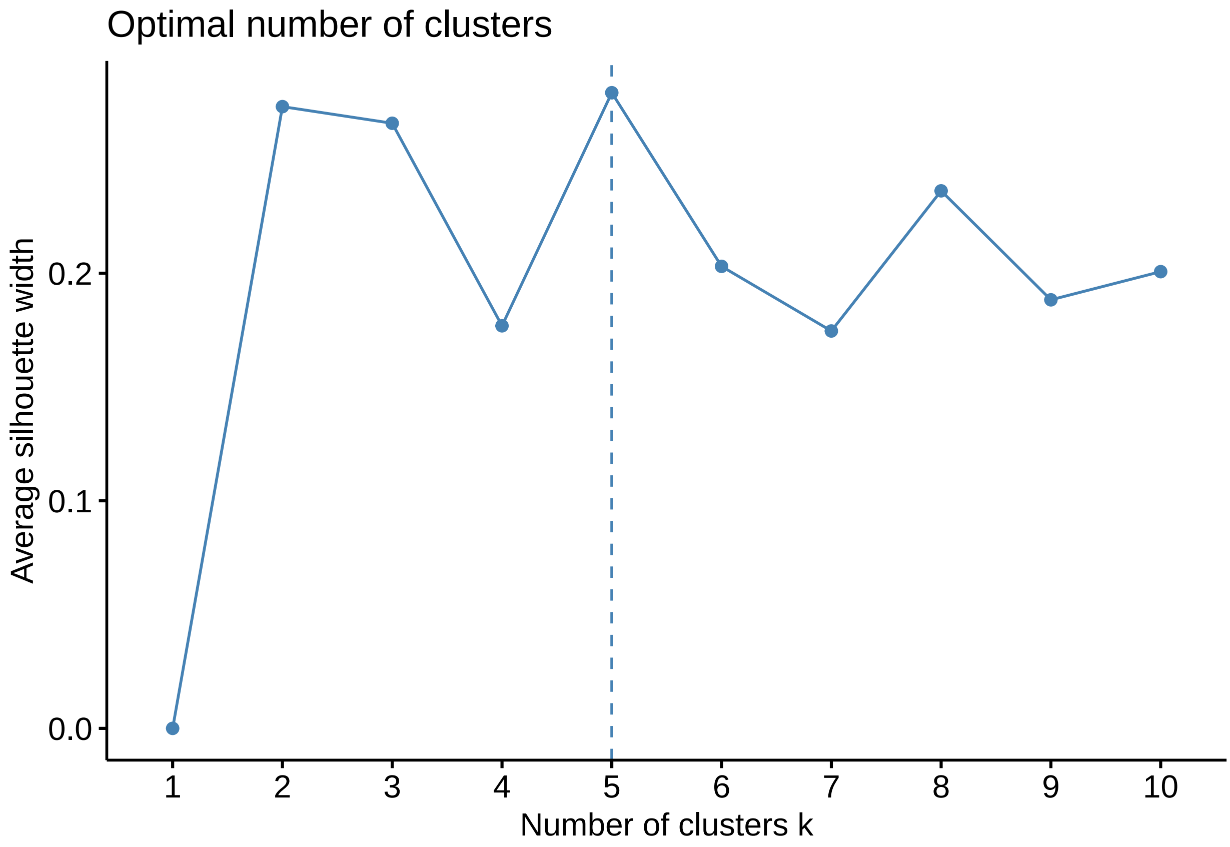


Determine best K

```
fviz_nbclust(Pharma_s, kmeans, method = "wss")
```



```
fviz_nbclust(Pharma_s, kmeans, method = "silhouette")
```



Silouhette says K =5 is optimal, but looking at WSS k=4 could be better, so I tried both.

Cluster the data

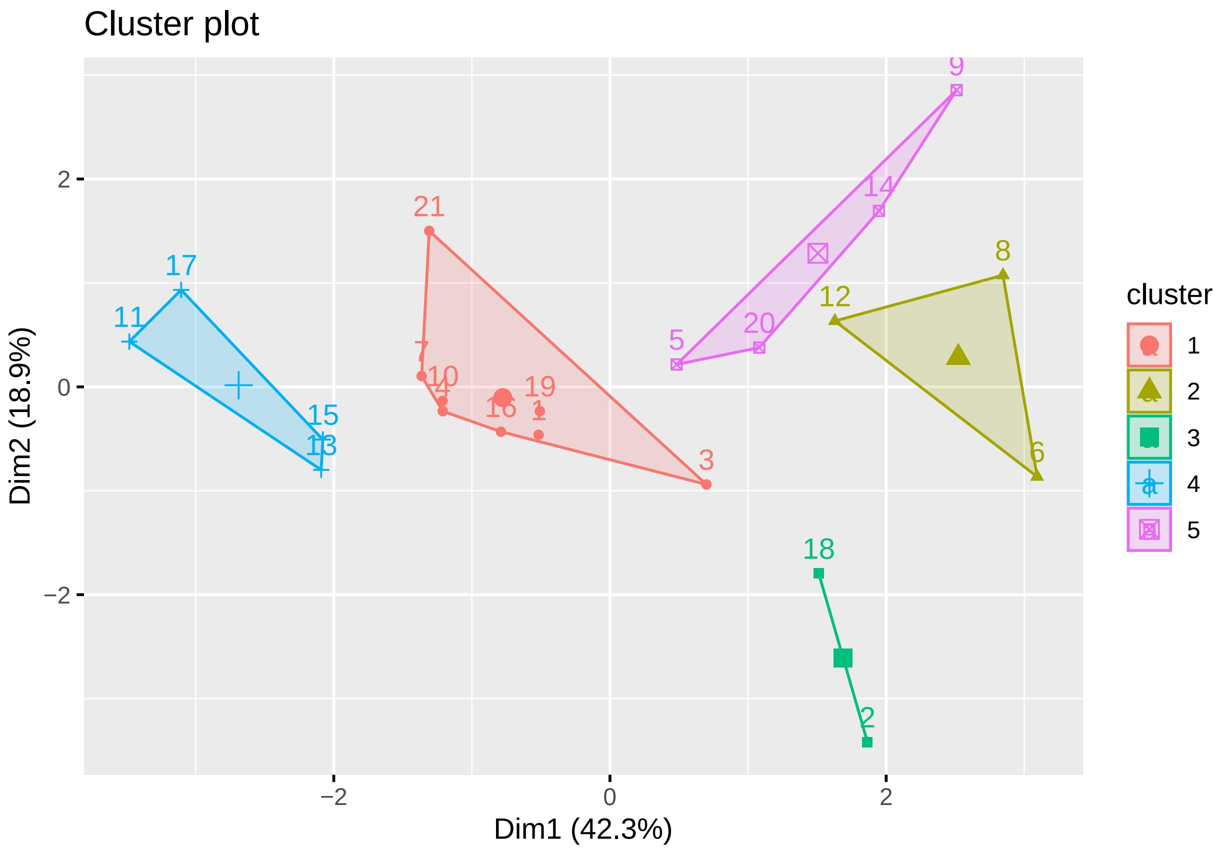
```
k5 <-kmeans(Pharma_s, centers = 5, nstart = 30)
k5$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 5 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
```

```
k5$size
```

```
## [1] 8 3 2 4 4
```

```
fviz_cluster(k5, data = Pharma_s)
```



I don't think the suggested 5 clusters is that useful as cluster 5 only consists of 2 data points. The analyst is looking for an overview of the pharmaceutical market and it is not as meaningful in that sense, so I tried this with K = 4:

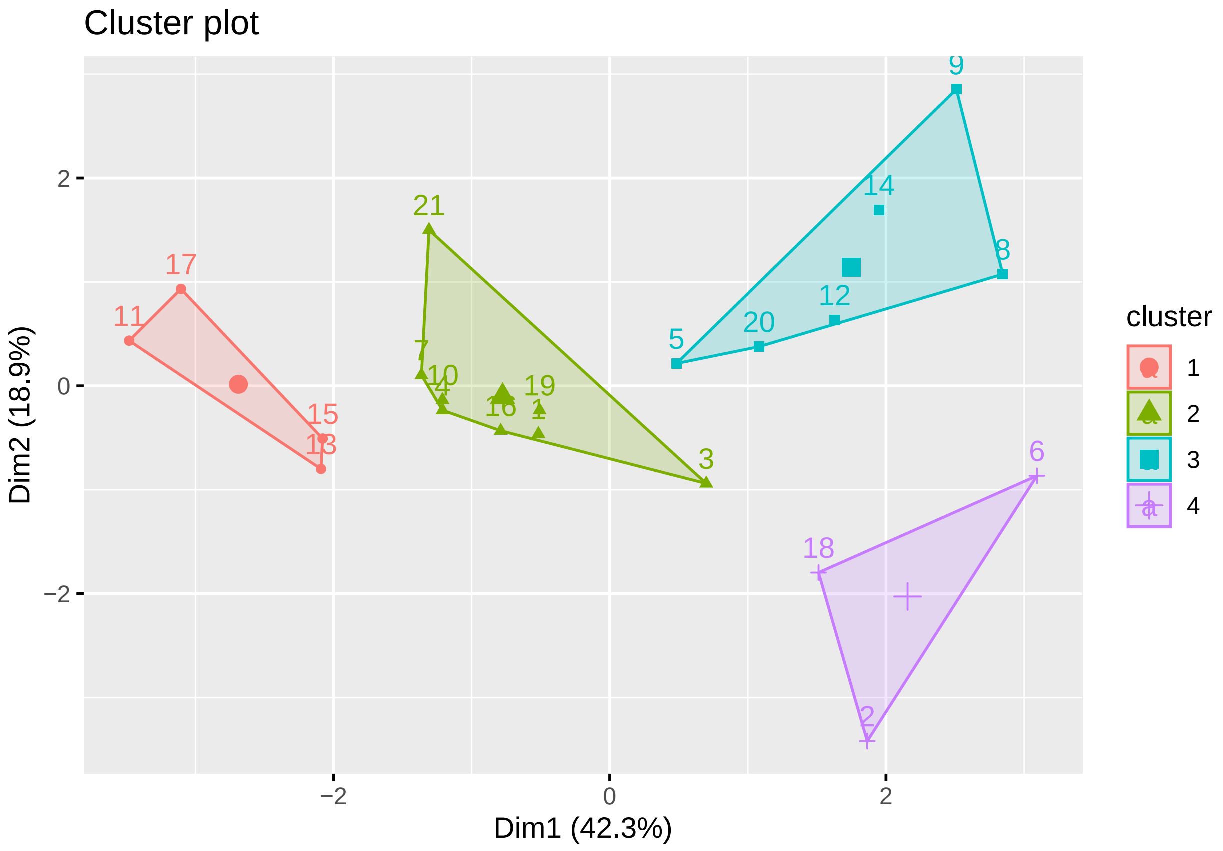
```
k4 <-kmeans(Pharma_s, centers = 4, nstart = 30)
k4$centers
```

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

```
k4$size
```

```
## [1] 4 8 6 3
```

```
fviz_cluster(k4, data = Pharma_s)
```



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

Group 1: 2, 6, 18

Market_Cap: Lower than average Beta: Higher than average PE_Ratio: Higher than average (highest of all groups) ROE: Lower than average (lowest of all groups) ROA: Lower than average (lowest of all groups) Asset_turnover: Around average Leverage: Lower than average Rev_Growth: Lower than average Net_Profit_Margin: Lower than average (lowest of all groups)

Group 2: 1, 3, 4, 7, 10, 16, 19, 21

Market_Cap: Slightly lower than average Beta: Lower than average (lowest of all groups) PE_Ratio: Lower than average ROE: Higher than average ROA: Higher than average Asset_turnover: Slightly higher than average Leverage: Lower than average Rev_Growth: Lower than average (lowest of all groups) Net_Profit_Margin: Higher than average

Group 3: 17, 13, 15, 11

Market_Cap: Higher than average (highest of all groups) Beta: Lower than average PE_Ratio: Lower than average ROE: Higher than average (highest of all groups) ROA: Higher than average (highest of all groups) Asset_turnover: Higher than average (highest of all groups) Leverage: Lower than average Rev_Growth: Higher than average Net_Profit_Margin: Higher than average

Group 4: 5, 8, 9, 12, 14, 20

Market_Cap: Lower than average (lowest of all groups) Beta: Higher than average (highest of all groups) PE_Ratio: Lower than average (lowest of all groups) ROE: Lower than average ROA: Lower than average

Asset_turnover: Lower than average (lowest of all groups) Leverage: Higher than average (highest of all groups) Rev_Growth: Higher than average (highest of all groups) Net_Profit_Margin: Lower than average

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

```
table(Pharma[c(2, 6, 18), c(12:14)])
```

```
## , , Exchange = NYSE
## 
##                         Location
## Median_Recommendation CANADA GERMANY US
##          Hold              0       1  1
##          Moderate Buy      1       0  0
```

Group 1: They are all on NYSE, 2/3 are Hold, All in different countries

```
table(Pharma[c(1, 3, 4, 7, 10, 16, 19, 21), c(12:14)])
```

```
## , , Exchange = NYSE
## 
##                         Location
## Median_Recommendation SWITZERLAND UK US
##         Hold                    1  0  3
##         Moderate Buy            0  0  1
##         Moderate Sell           0  1  1
##         Strong Buy              0  1  0
```

Group 2: All on NYSE, 4/8 are Hold, 2/8 are Moderate Sell, then 1/8 each Moderate Buy and Strong Buy, 5/8 in the US, 2/8 in UK, and 1 in Switzerland.

```
table(Pharma[c(17, 13, 15, 11), c(12:14)])
```

```
## , , Exchange = NYSE
## 
##                         Location
## Median_Recommendation UK US
##          Hold          1  1
##          Moderate Buy  0  2
```

Group 3:

```
table(Pharma[c(5, 8, 9, 12, 14, 20), c(12:14)])
```

```
## , , Exchange = AMEX
## 
##                         Location
## Median_Recommendation FRANCE IRELAND US
##         Hold               0       0  1
##         Moderate Buy       0       0  0
##         Moderate Sell      0       0  0
```

```
##
## , , Exchange = NASDAQ
##
##                      Location
## Median_Recommendation FRANCE IRELAND US
##         Hold               0       0  0
##         Moderate Buy       0       0  1
##         Moderate Sell      0       0  0
##
## , , Exchange = NYSE
##
##                      Location
## Median_Recommendation FRANCE IRELAND US
##         Hold               0       0  0
##         Moderate Buy       1       0  1
##         Moderate Sell      0       1  1
```

Group 4: All on different Exchanges, 4/6 are in the US, 3/6 are Moderate Buy, 2/6 are Moderate Sell, and 1/6 is Hold.

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

Group 1: Medium-Low Market Cap and Low Rev Growth Group 2: Medium Market Cap and Lowest Rev Growth Group 3: Highest Market Cap and High Rev Growth Group 4: Lowest Market Cap and Highest Rev Growth