Assignment\_4

Deepak

2023-10-29

knitr::opts\_chunk$set(echo = TRUE)

#install.packages("httr")  
#install.packages("readr")  
#install.packages("factoextra")  
#install.packages("flexclust")  
library(httr)  
library(readr)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ purrr 1.0.2  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(factoextra)

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

library(ISLR)  
library(flexclust)

## Loading required package: grid  
## Loading required package: lattice  
## Loading required package: modeltools  
## Loading required package: stats4

library(caret)

##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift  
##   
## The following object is masked from 'package:httr':  
##   
## progress

#Importing Data set

#importing Data set and converting   
getwd()

## [1] "C:/Users/durga/OneDrive/Documents"

pharma<-read.csv("C:/Users/durga/Downloads/Pharmaceuticals.csv")  
#summarize the Data  
#str(pharma)  
head(pharma,10)

## Symbol Name Market\_Cap Beta PE\_Ratio ROE ROA  
## 1 ABT Abbott Laboratories 68.44 0.32 24.7 26.4 11.8  
## 2 AGN Allergan, Inc. 7.58 0.41 82.5 12.9 5.5  
## 3 AHM Amersham plc 6.30 0.46 20.7 14.9 7.8  
## 4 AZN AstraZeneca PLC 67.63 0.52 21.5 27.4 15.4  
## 5 AVE Aventis 47.16 0.32 20.1 21.8 7.5  
## 6 BAY Bayer AG 16.90 1.11 27.9 3.9 1.4  
## 7 BMY Bristol-Myers Squibb Company 51.33 0.50 13.9 34.8 15.1  
## 8 CHTT Chattem, Inc 0.41 0.85 26.0 24.1 4.3  
## 9 ELN Elan Corporation, plc 0.78 1.08 3.6 15.1 5.1  
## 10 LLY Eli Lilly and Company 73.84 0.18 27.9 31.0 13.5  
## Asset\_Turnover Leverage Rev\_Growth Net\_Profit\_Margin Median\_Recommendation  
## 1 0.7 0.42 7.54 16.1 Moderate Buy  
## 2 0.9 0.60 9.16 5.5 Moderate Buy  
## 3 0.9 0.27 7.05 11.2 Strong Buy  
## 4 0.9 0.00 15.00 18.0 Moderate Sell  
## 5 0.6 0.34 26.81 12.9 Moderate Buy  
## 6 0.6 0.00 -3.17 2.6 Hold  
## 7 0.9 0.57 2.70 20.6 Moderate Sell  
## 8 0.6 3.51 6.38 7.5 Moderate Buy  
## 9 0.3 1.07 34.21 13.3 Moderate Sell  
## 10 0.6 0.53 6.21 23.4 Hold  
## Location Exchange  
## 1 US NYSE  
## 2 CANADA NYSE  
## 3 UK NYSE  
## 4 UK NYSE  
## 5 FRANCE NYSE  
## 6 GERMANY NYSE  
## 7 US NYSE  
## 8 US NASDAQ  
## 9 IRELAND NYSE  
## 10 US NYSE

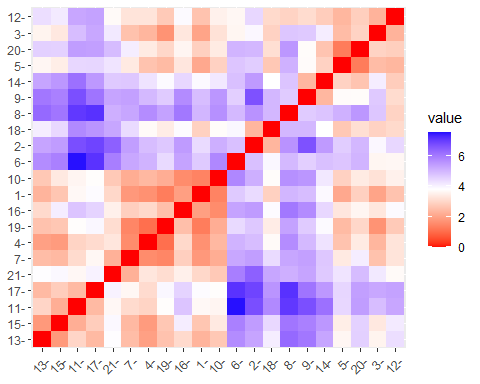
set.seed(23)  
#Data frame Z Score scaling  
pharma\_scaled <- scale(pharma[,3:11])  
summary(pharma\_scaled)

## Market\_Cap Beta PE\_Ratio ROE   
## Min. :-0.9768 Min. :-1.3466 Min. :-1.3404 Min. :-1.4515   
## 1st Qu.:-0.8763 1st Qu.:-0.6844 1st Qu.:-0.4023 1st Qu.:-0.7223   
## Median :-0.1614 Median :-0.2560 Median :-0.2429 Median :-0.2118   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.2762 3rd Qu.: 0.4841 3rd Qu.: 0.1495 3rd Qu.: 0.3450   
## Max. : 2.4200 Max. : 2.2758 Max. : 3.4971 Max. : 2.4597   
## ROA Asset\_Turnover Leverage Rev\_Growth   
## Min. :-1.7128 Min. :-1.8451 Min. :-0.74966 Min. :-1.4971   
## 1st Qu.:-0.9047 1st Qu.:-0.4613 1st Qu.:-0.54487 1st Qu.:-0.6328   
## Median : 0.1289 Median :-0.4613 Median :-0.31449 Median :-0.3621   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.8430 3rd Qu.: 0.9225 3rd Qu.: 0.01828 3rd Qu.: 0.7693   
## Max. : 1.8389 Max. : 1.8451 Max. : 3.74280 Max. : 1.8862   
## Net\_Profit\_Margin   
## Min. :-1.99560   
## 1st Qu.:-0.68504   
## Median : 0.06168   
## Mean : 0.00000   
## 3rd Qu.: 0.82364   
## Max. : 1.49416

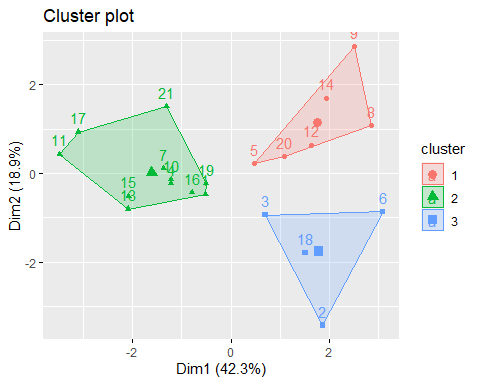
# Data Frame Range Scaling   
pharma\_range <- scale(pharma[,3:11])

1. 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.

set.seed(23)  
dst\_rows <- get\_dist(pharma\_scaled)  
fviz\_dist(dst\_rows) #To visualize distance between matrix rows



cluster1 <- kmeans(pharma\_scaled, centers = 3, nstart = 15) # HEre taking K=3 & nstart=15  
fviz\_cluster(cluster1, data = pharma\_scaled)



print(cluster1)

## K-means clustering with 3 clusters of sizes 6, 11, 4  
##   
## Cluster means:  
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.8261772 0.4775991 -0.3696184 -0.5631589 -0.8514589 -0.9994088  
## 2 0.6733825 -0.3586419 -0.2763512 0.6565978 0.8344159 0.4612656  
## 3 -0.6125361 0.2698666 1.3143935 -0.9609057 -1.0174553 0.2306328  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 0.8502201 0.9158889 -0.3319956  
## 2 -0.3331068 -0.2902163 0.6823310  
## 3 -0.3592866 -0.5757385 -1.3784169  
##   
## Clustering vector:  
## [1] 2 3 3 2 1 3 2 1 1 2 2 1 2 1 2 2 2 3 2 1 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 32.14336 43.30886 20.54199  
## (between\_SS / total\_SS = 46.7 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

The k-means algorithm was used to divide the 21 enterprises into three groups with no variable weights. We chose k=3 since that is the optimal k indicated by the silhouette approach.

fviz\_nbclust(pharma\_scaled, kmeans, method = "wss") # WSS method (ELBOW METHOOD)



fviz\_nbclust(pharma\_scaled, kmeans, method = "silhouette") #SILHOUETTE Method (To find best K value)



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

I did not use the WSS approach since the graph did not show a distinct elbow and was extremely unclear. The graph does not indicate the elbow/knee position, and it flattens out more than once at k = 4 and 6, respectively, and I chose the silhouette approach since it is apparent to display the ideal cluster K = 5.

#let's look at the mean value from actual data by clusters  
aggregate(pharma[3:11], by=list(cluster=cluster1$cluster), mean)

## cluster Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## 1 1 9.23500 0.6483333 19.43333 17.3 5.983333 0.4833333 1.2500000  
## 2 2 97.11364 0.4336364 20.95455 35.7 14.954545 0.8000000 0.3254545  
## 3 3 21.75500 0.5950000 46.90000 11.3 5.100000 0.7500000 0.3050000  
## Rev\_Growth Net\_Profit\_Margin  
## 1 23.49000 13.51667  
## 2 10.16455 20.17273  
## 3 7.01000 6.65000

actual\_data <- cbind(pharma, cluster = cluster1$cluster)  
tibble(actual\_data)

## # A tibble: 21 × 15  
## Symbol Name Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 ABT Abbott … 68.4 0.32 24.7 26.4 11.8 0.7 0.42  
## 2 AGN Allerga… 7.58 0.41 82.5 12.9 5.5 0.9 0.6   
## 3 AHM Amersha… 6.3 0.46 20.7 14.9 7.8 0.9 0.27  
## 4 AZN AstraZe… 67.6 0.52 21.5 27.4 15.4 0.9 0   
## 5 AVE Aventis 47.2 0.32 20.1 21.8 7.5 0.6 0.34  
## 6 BAY Bayer AG 16.9 1.11 27.9 3.9 1.4 0.6 0   
## 7 BMY Bristol… 51.3 0.5 13.9 34.8 15.1 0.9 0.57  
## 8 CHTT Chattem… 0.41 0.85 26 24.1 4.3 0.6 3.51  
## 9 ELN Elan Co… 0.78 1.08 3.6 15.1 5.1 0.3 1.07  
## 10 LLY Eli Lil… 73.8 0.18 27.9 31 13.5 0.6 0.53  
## # ℹ 11 more rows  
## # ℹ 6 more variables: Rev\_Growth <dbl>, Net\_Profit\_Margin <dbl>,  
## # Median\_Recommendation <chr>, Location <chr>, Exchange <chr>, cluster <int>

by(actual\_data, factor(actual\_data$cluster), summary)#intensive statistical cluster analysis

## factor(actual\_data$cluster): 1  
## Symbol Name Market\_Cap Beta   
## Length:6 Length:6 Min. : 0.410 Min. :0.2400   
## Class :character Class :character 1st Qu.: 0.885 1st Qu.:0.4025   
## Mode :character Mode :character Median : 1.900 Median :0.7000   
## Mean : 9.235 Mean :0.6483   
## 3rd Qu.: 3.095 3rd Qu.:0.8250   
## Max. :47.160 Max. :1.0800   
## PE\_Ratio ROE ROA Asset\_Turnover   
## Min. : 3.60 Min. :10.20 Min. :4.300 Min. :0.3000   
## 1st Qu.:18.77 1st Qu.:12.18 1st Qu.:5.175 1st Qu.:0.3500   
## Median :20.00 Median :18.25 Median :6.100 Median :0.5500   
## Mean :19.43 Mean :17.30 Mean :5.983 Mean :0.4833   
## 3rd Qu.:24.52 3rd Qu.:21.70 3rd Qu.:6.800 3rd Qu.:0.6000   
## Max. :28.60 Max. :24.10 Max. :7.500 Max. :0.6000   
## Leverage Rev\_Growth Net\_Profit\_Margin Median\_Recommendation  
## Min. :0.2000 Min. : 6.38 Min. : 7.50 Length:6   
## 1st Qu.:0.4875 1st Qu.:17.20 1st Qu.:11.47 Class :character   
## Median :1.0000 Median :28.00 Median :13.10 Mode :character   
## Mean :1.2500 Mean :23.49 Mean :13.52   
## 3rd Qu.:1.3550 3rd Qu.:30.07 3rd Qu.:14.65   
## Max. :3.5100 Max. :34.21 Max. :21.30   
## Location Exchange cluster   
## Length:6 Length:6 Min. :1   
## Class :character Class :character 1st Qu.:1   
## Mode :character Mode :character Median :1   
## Mean :1   
## 3rd Qu.:1   
## Max. :1   
## ------------------------------------------------------------   
## factor(actual\_data$cluster): 2  
## Symbol Name Market\_Cap Beta   
## Length:11 Length:11 Min. : 34.10 Min. :0.1800   
## Class :character Class :character 1st Qu.: 59.48 1st Qu.:0.3350   
## Mode :character Mode :character Median : 73.84 Median :0.4600   
## Mean : 97.11 Mean :0.4336   
## 3rd Qu.:127.33 3rd Qu.:0.5150   
## Max. :199.47 Max. :0.6500   
## PE\_Ratio ROE ROA Asset\_Turnover Leverage   
## Min. :13.10 Min. :17.9 Min. :11.20 Min. :0.50 Min. :0.0000   
## 1st Qu.:18.45 1st Qu.:26.9 1st Qu.:13.35 1st Qu.:0.65 1st Qu.:0.0800   
## Median :21.50 Median :31.0 Median :15.00 Median :0.80 Median :0.2800   
## Mean :20.95 Mean :35.7 Mean :14.95 Mean :0.80 Mean :0.3255   
## 3rd Qu.:24.15 3rd Qu.:43.1 3rd Qu.:15.85 3rd Qu.:0.90 3rd Qu.:0.4750   
## Max. :28.40 Max. :62.9 Max. :20.30 Max. :1.10 Max. :1.1200   
## Rev\_Growth Net\_Profit\_Margin Median\_Recommendation Location   
## Min. :-2.690 Min. :14.10 Length:11 Length:11   
## 1st Qu.: 4.455 1st Qu.:17.75 Class :character Class :character   
## Median : 8.560 Median :20.60 Mode :character Mode :character   
## Mean :10.165 Mean :20.17   
## 3rd Qu.:16.175 3rd Qu.:22.90   
## Max. :25.540 Max. :25.50   
## Exchange cluster   
## Length:11 Min. :2   
## Class :character 1st Qu.:2   
## Mode :character Median :2   
## Mean :2   
## 3rd Qu.:2   
## Max. :2   
## ------------------------------------------------------------   
## factor(actual\_data$cluster): 3  
## Symbol Name Market\_Cap Beta   
## Length:4 Length:4 Min. : 6.30 Min. :0.4000   
## Class :character Class :character 1st Qu.: 7.26 1st Qu.:0.4075   
## Mode :character Mode :character Median :12.24 Median :0.4350   
## Mean :21.75 Mean :0.5950   
## 3rd Qu.:26.73 3rd Qu.:0.6225   
## Max. :56.24 Max. :1.1100   
## PE\_Ratio ROE ROA Asset\_Turnover Leverage   
## Min. :20.7 Min. : 3.90 Min. :1.400 Min. :0.60 Min. :0.0000   
## 1st Qu.:26.1 1st Qu.:10.65 1st Qu.:4.475 1st Qu.:0.60 1st Qu.:0.2025   
## Median :42.2 Median :13.20 Median :5.600 Median :0.75 Median :0.3100   
## Mean :46.9 Mean :11.30 Mean :5.100 Mean :0.75 Mean :0.3050   
## 3rd Qu.:63.0 3rd Qu.:13.85 3rd Qu.:6.225 3rd Qu.:0.90 3rd Qu.:0.4125   
## Max. :82.5 Max. :14.90 Max. :7.800 Max. :0.90 Max. :0.6000   
## Rev\_Growth Net\_Profit\_Margin Median\_Recommendation Location   
## Min. :-3.170 Min. : 2.600 Length:4 Length:4   
## 1st Qu.: 4.495 1st Qu.: 4.775 Class :character Class :character   
## Median : 8.105 Median : 6.400 Mode :character Mode :character   
## Mean : 7.010 Mean : 6.650   
## 3rd Qu.:10.620 3rd Qu.: 8.275   
## Max. :15.000 Max. :11.200   
## Exchange cluster   
## Length:4 Min. :3   
## Class :character 1st Qu.:3   
## Mode :character Median :3   
## Mean :3   
## 3rd Qu.:3   
## Max. :3

Recommendations, Location and Exchange of cluster

#Cluster median recommendation  
T\_Recom <- table(actual\_data$cluster, actual\_data$Median\_Recommendation)   
names(dimnames(T\_Recom)) <- c("Cluster", "Recommendation")  
TR <- addmargins(T\_Recom)  
TR

## Recommendation  
## Cluster Hold Moderate Buy Moderate Sell Strong Buy Sum  
## 1 1 3 2 0 6  
## 2 6 3 2 0 11  
## 3 2 1 0 1 4  
## Sum 9 7 4 1 21

The data do not show a clear link between clusterMedian Recommendation. There are 21 recommendations in total, with 1 strong buy, 7 moderate buys, 9 holds, and 4 moderate sells.

#Cluster-based location breakdown  
T\_Location <- table(actual\_data$cluster, actual\_data$Location)  
names(dimnames(T\_Location)) <- c("Cluster", "Location")  
Tlocation <- addmargins(T\_Location)  
Tlocation

## Location  
## Cluster CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US Sum  
## 1 0 1 0 1 0 0 4 6  
## 2 0 0 0 0 1 2 8 11  
## 3 1 0 1 0 0 1 1 4  
## Sum 1 1 1 1 1 3 13 21

We cannot deduce any association between cluster Location from the findings. A total of 21 firms are divided into 13 in the United States, three in the United Kingdom, and one each in Canada, France, Germany, Ireland, and Switzerland.

#Exchange breakdown by cluster  
T\_Exchange <- table(actual\_data$cluster, actual\_data$Exchange)  
names(dimnames(T\_Exchange)) <- c("Cluster", "Exchange")  
Texchange <- addmargins(T\_Exchange)  
Texchange

## Exchange  
## Cluster AMEX NASDAQ NYSE Sum  
## 1 1 1 4 6  
## 2 0 0 11 11  
## 3 0 0 4 4  
## Sum 1 1 19 21

The results show that there is no link between clusterExchange. There are 21 corporations in all, divided into 1 Amex, 1 Nasdaq, and 19 NYSE.

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

fviz\_nbclust(pharma\_range, FUN = kmeans, method = "silhouette")



fviz\_nbclust(pharma\_range, kmeans, method = "wss")

 We also perform tests to determine the best k using range normalization. The ideal k is 2 from the silhouette and 6 from the elbow (not clear). We’ll stick with z-score normalization data because the k from range normalization isn’t as good.

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

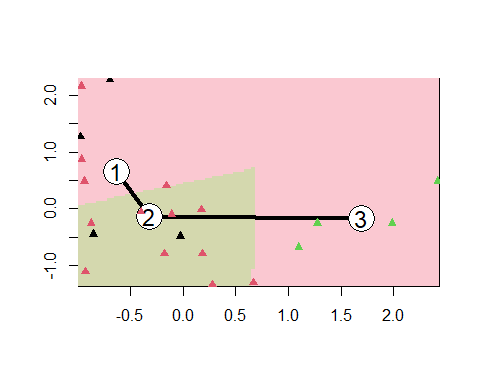
set.seed(11)  
cluster2 = kcca(pharma\_scaled, k=3, kccaFamily("kmeans"))  
cluster2

## kcca object of family 'kmeans'   
##   
## call:  
## kcca(x = pharma\_scaled, k = 3, family = kccaFamily("kmeans"))  
##   
## cluster sizes:  
##   
## 1 2 3   
## 4 13 4

clusters(cluster2)

## [1] 2 1 2 2 2 1 2 1 2 2 3 2 3 2 3 2 3 1 2 2 2

#Apply the predict() function  
clusters\_index <- predict(cluster2)  
image(cluster2)  
points(pharma\_scaled, col=clusters\_index, pch=17, cex=1.0)

 To run kmeans cluster on k =3, we use the kcca algorithm instead of kmeans from basic R.

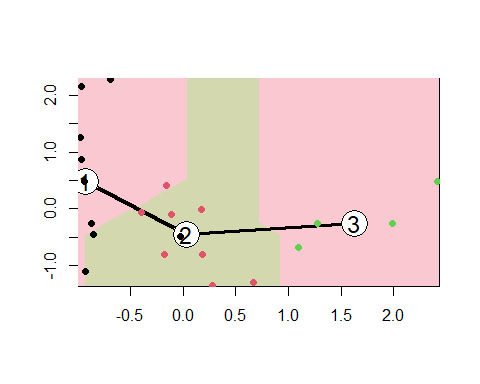
set.seed(11)  
cluster2 = kcca(pharma\_scaled, k=3, kccaFamily("kmedians"))  
cluster2

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = pharma\_scaled, k = 3, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2 3   
## 9 8 4

clusters(cluster2)

## [1] 2 1 1 2 2 1 2 1 1 2 3 1 3 1 3 2 3 1 2 1 2

#Apply the predict() function  
clusters\_index <- predict(cluster2)  
image(cluster2)  
points(pharma\_scaled, col=clusters\_index, pch=16, cex=1.0)



1. Now that the clustering is complete, there are some insights we can pull from the output. Particularly, by using both the WSS and Silhouette methods, we could accurately determine that 5 clusters were needed as they both returned 5 as the optimum point.
2. we can make some general inferences about the clusters:

Cluster 1 had high ROE, ROA, Asset\_Turnover, and Net\_Profit\_Margin, but low Market\_Cap and Rev\_Growth. Cluster 2 had very high Beta and Leverage, but very low Market\_Cap, ROE, ROA, Net\_Profit\_Margins\_, and Revenue Growth, which is likely why is is the furthest away from cluster 4. cluster 3 is the oddest of the bunch with only 2 members. This cluster has a VERY high PE\_Ratio as well as a positive Asset\_Turnover, but is low in every other category. While having low leverage,Beta, and PE\_Ratio: cluster 4 held high Market\_cap, ROE, ROA, Asset\_Turnover, Revenue Growth, and Net\_Profit\_Margin which together set it apart from its closest neighbor, cluster 1. Cluster 5 has the a very high Rev\_Growth and positive Beta and Leverage, while maintaining low numbers in the other categories.

1. Looking at the last three columns that were not used in the clustering, there seems to be no consistent patterns within the clusters. Between most points, you will find that while they both may have the same exchange, the location or recommendation would be different, or visa versa. Though generally speaking, almost all were in the NYSE exchange anyways.
2. Cluster 1: Medium Market\_cap,ROE,ROA,Asset\_Turnover,Leverage, Net\_Profit\_Margin, and Rev\_Growth: “Medium”

Cluster 2: very high Beta and Leverage, very low ROA and Net\_Profit\_Margin: “High beta, low assets”

Cluster 3: Extreme PE\_Ratio and low Net\_Profit\_Margin:“High Price Earnings ratio, but low new profits”

Cluster 4: highest Market\_Cap, ROE, ROA, Asset\_Turnover, and Net\_Profit\_Margin: “great asset management with small negatives”

Cluster 5: small positive Beta with highest Rev\_Growth and slightly negative Net\_profit\_margin: “bad asset management with good growth”