Clustering Stock Returns

Suzana Iacob

01/12/2019

Problem statement

We look at a portfolio of stocks aiming to find a diversification strategy. The data contains monthly returns from some stocks among the S&P500 from March 2006 through February 2016.

```
data = read.csv("returns.csv")
returns = data[,3:122]
```

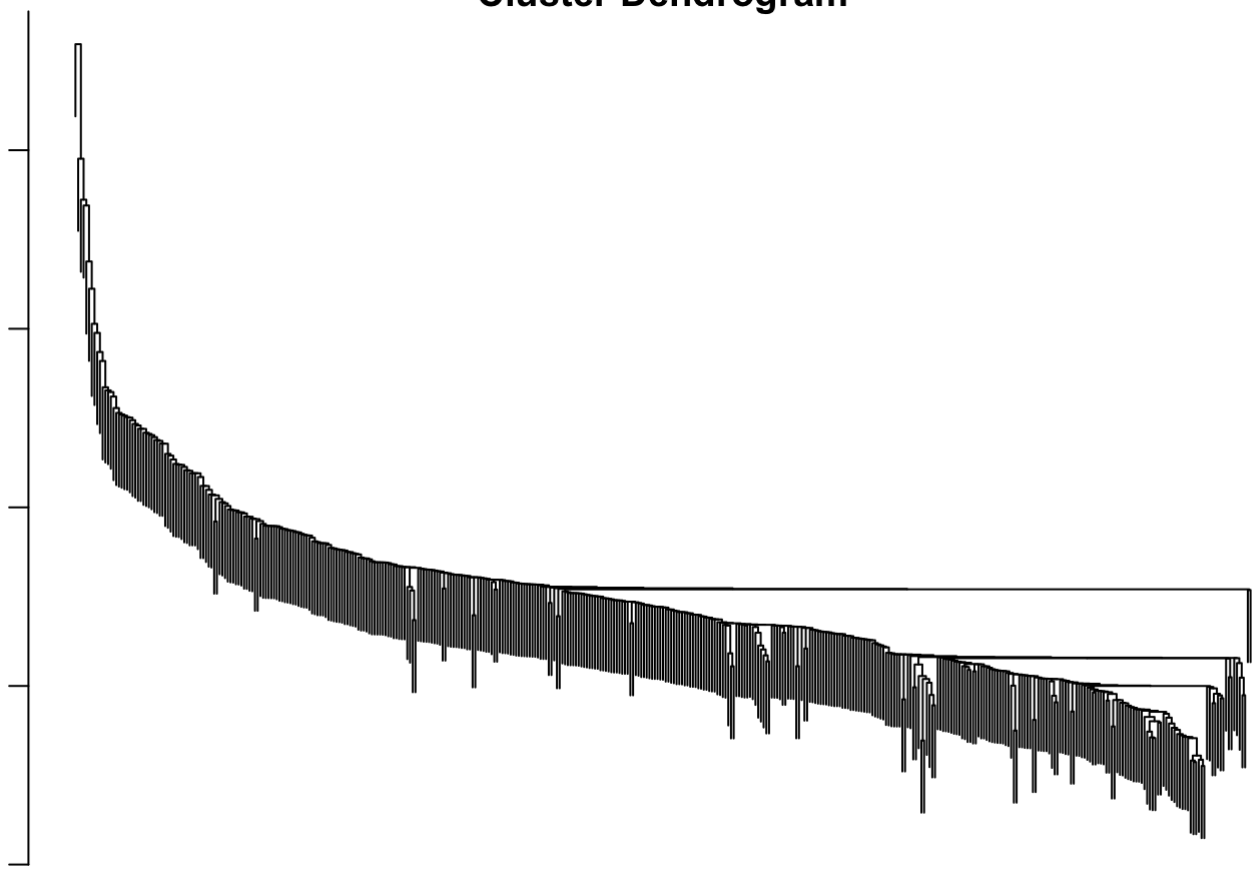
Hierarchical Clustering

We cluster the data using hierarchical clustering, using the Euclidean distance and the four cluster link- age metrics (single linkage, complete linkage, average linkage, and Ward D2 linkage). We plot the resulting dendograms

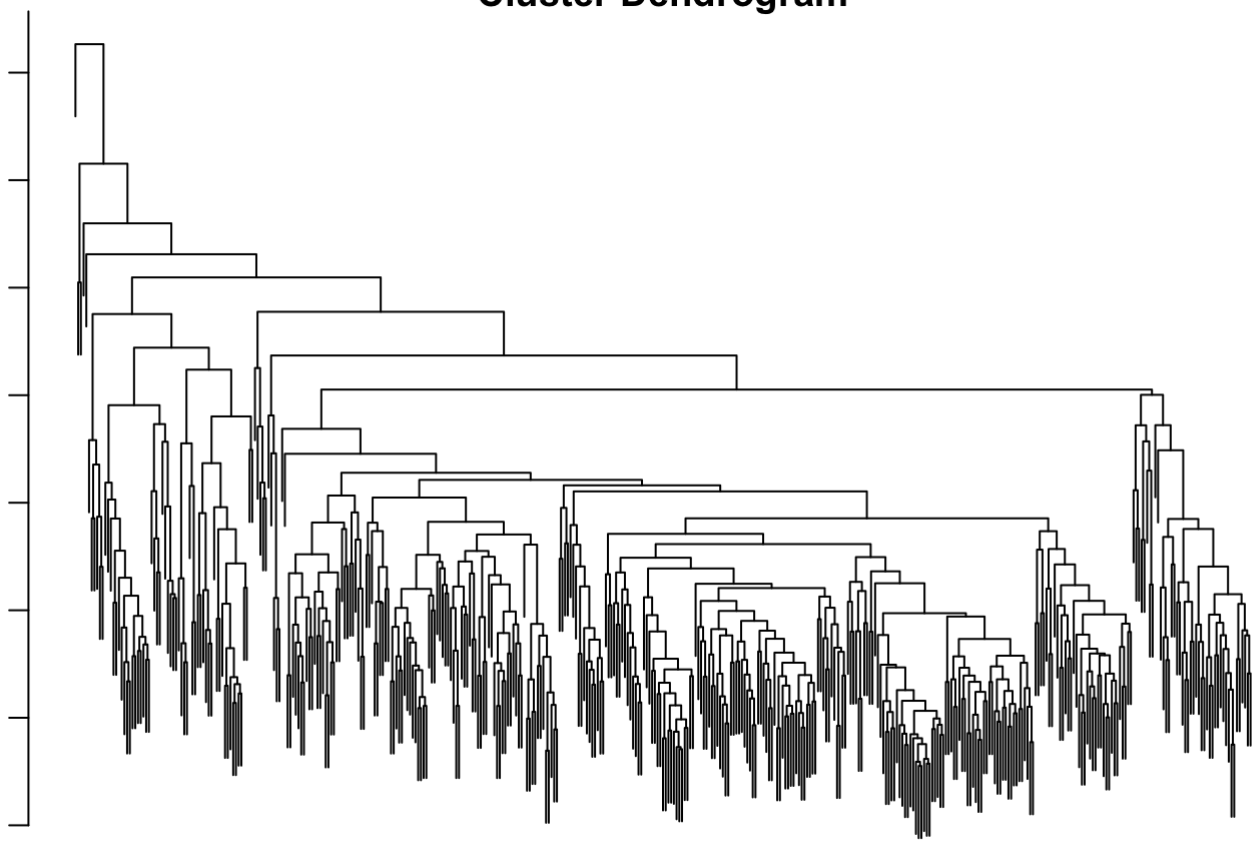
```
d <- dist(returns)
class(d)
```

```
## [1] "dist"
```

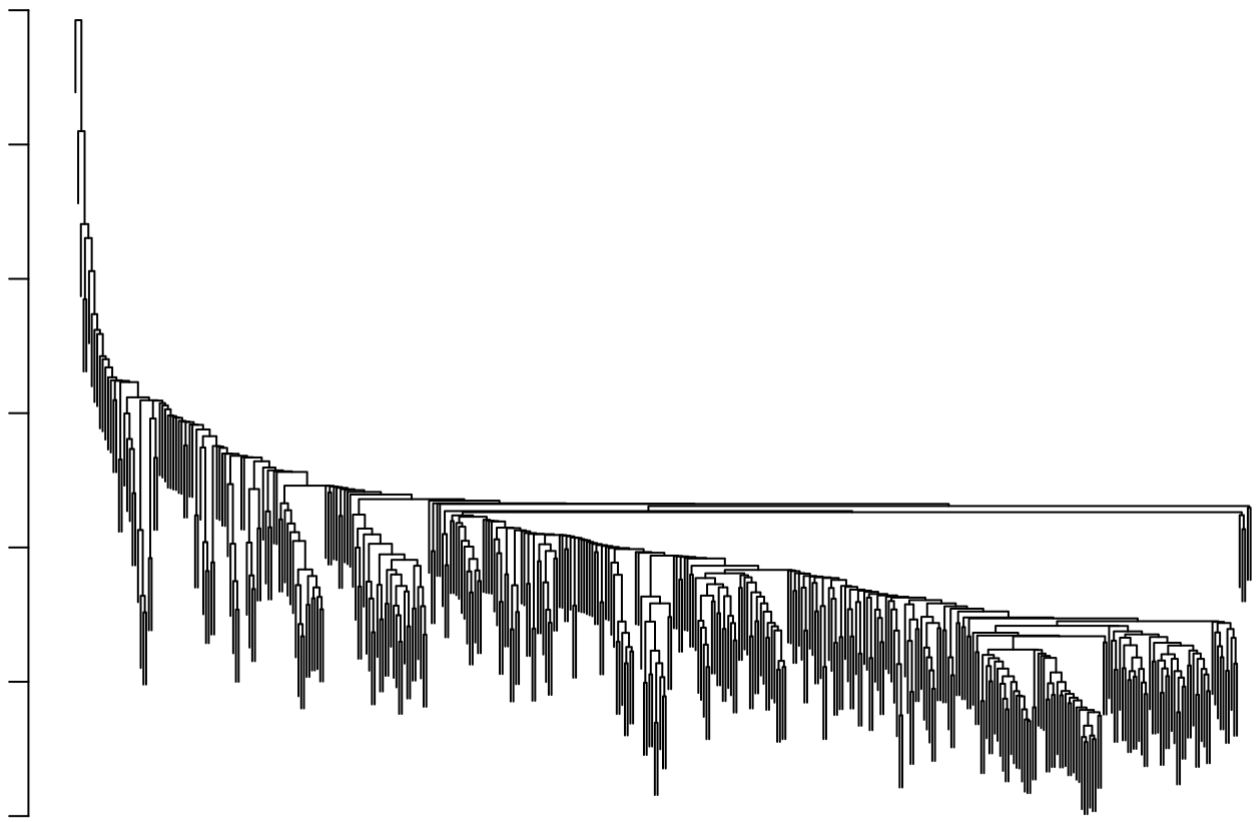
```
hclust.mod1 <- hclust(d, method="single")
par(mar = c(1,1,1,1))
plot(hclust.mod1, labels=F, ylab="Dissimilarity", xlab = "", sub = "")
```



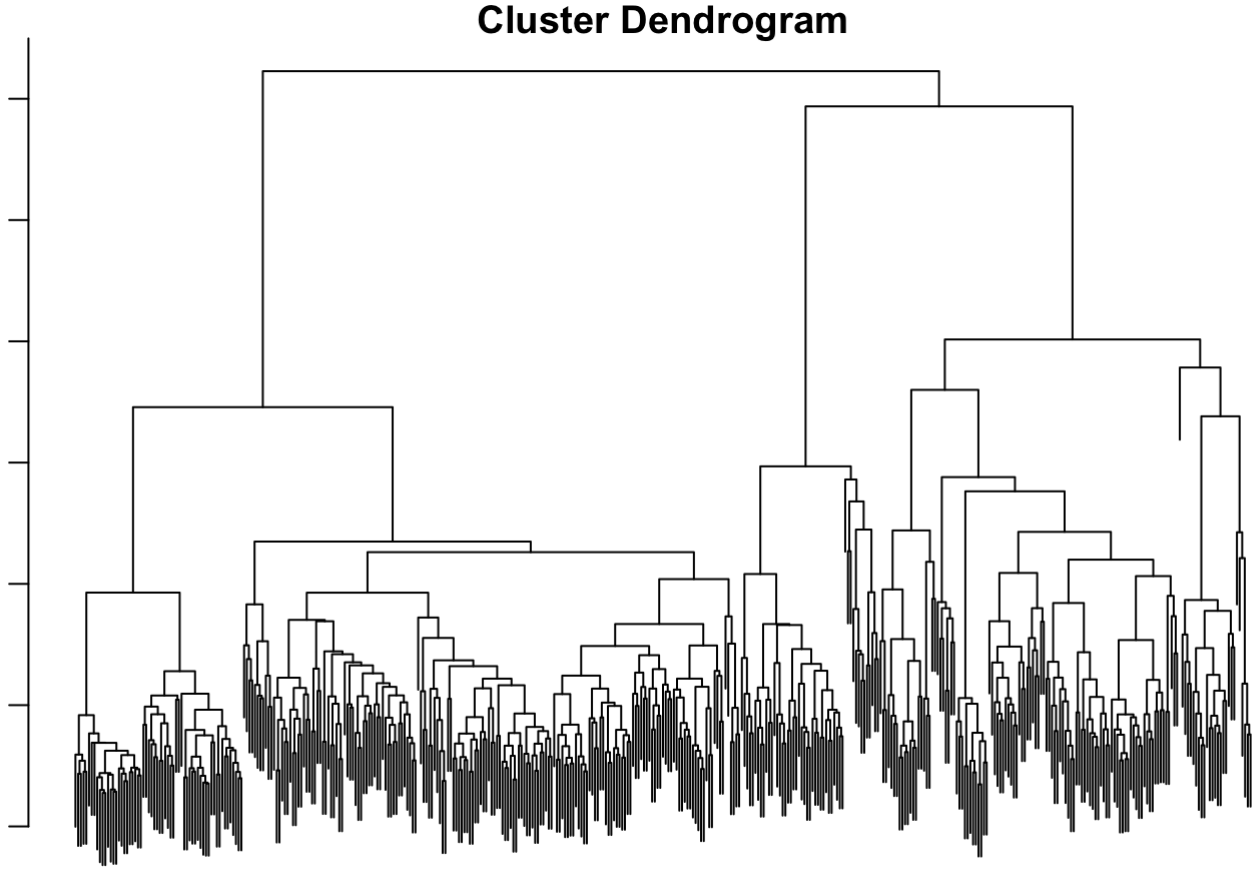

```
hclust.mod2 <- hclust(d, method="complete")
par(mar = c(1,1,1,1))
plot(hclust.mod2, labels=F, ylab="Dissimilarity", xlab = "", sub = "")
```




```
hclust.mod3 <- hclust(d, method="average")
par(mar = c(1,1,1,1))
plot(hclust.mod3, labels=F, ylab="Dissimilarity", xlab = "", sub = "")
```




```
hclust.mod4 <- hclust(d, method="ward.D2")
par(mar = c(1,1,1,1))
plot(hclust.mod4, labels=F, ylab="Dissimilarity", xlab = "", sub = "")
```

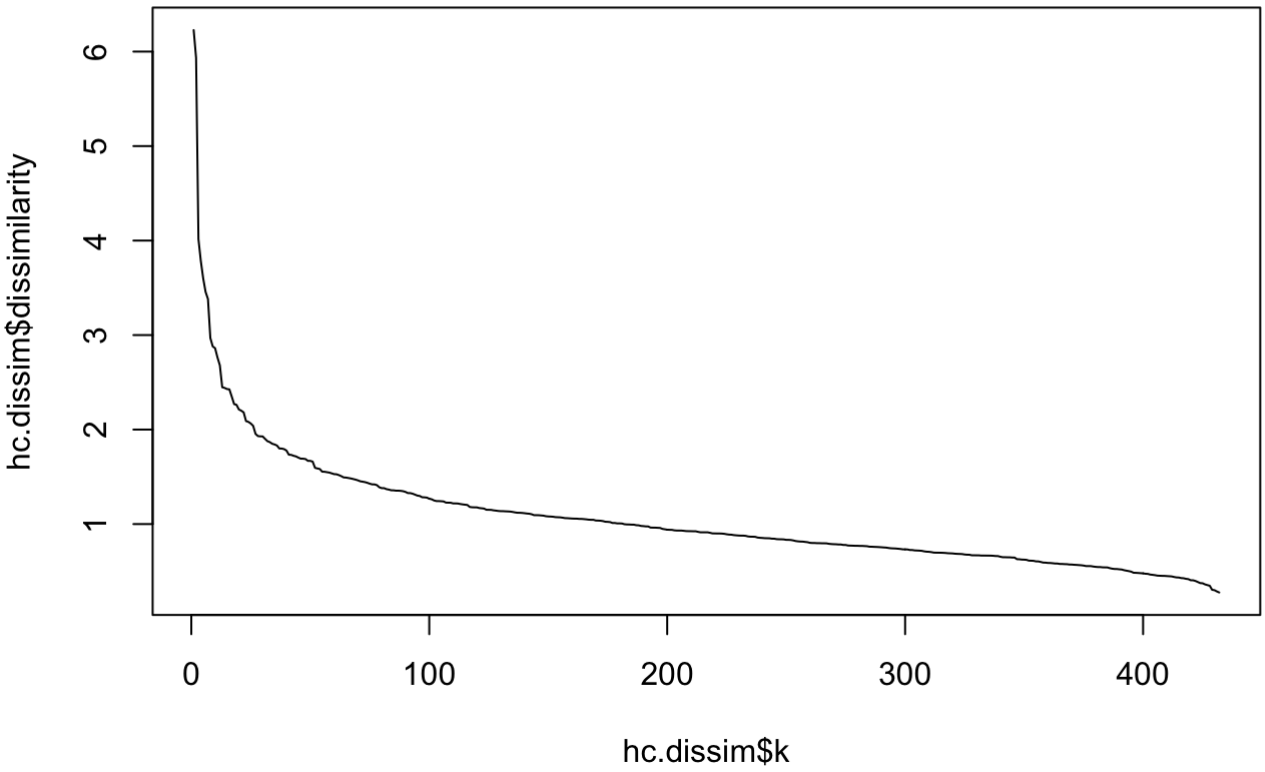


Scree Plot for deciding the number of clusters

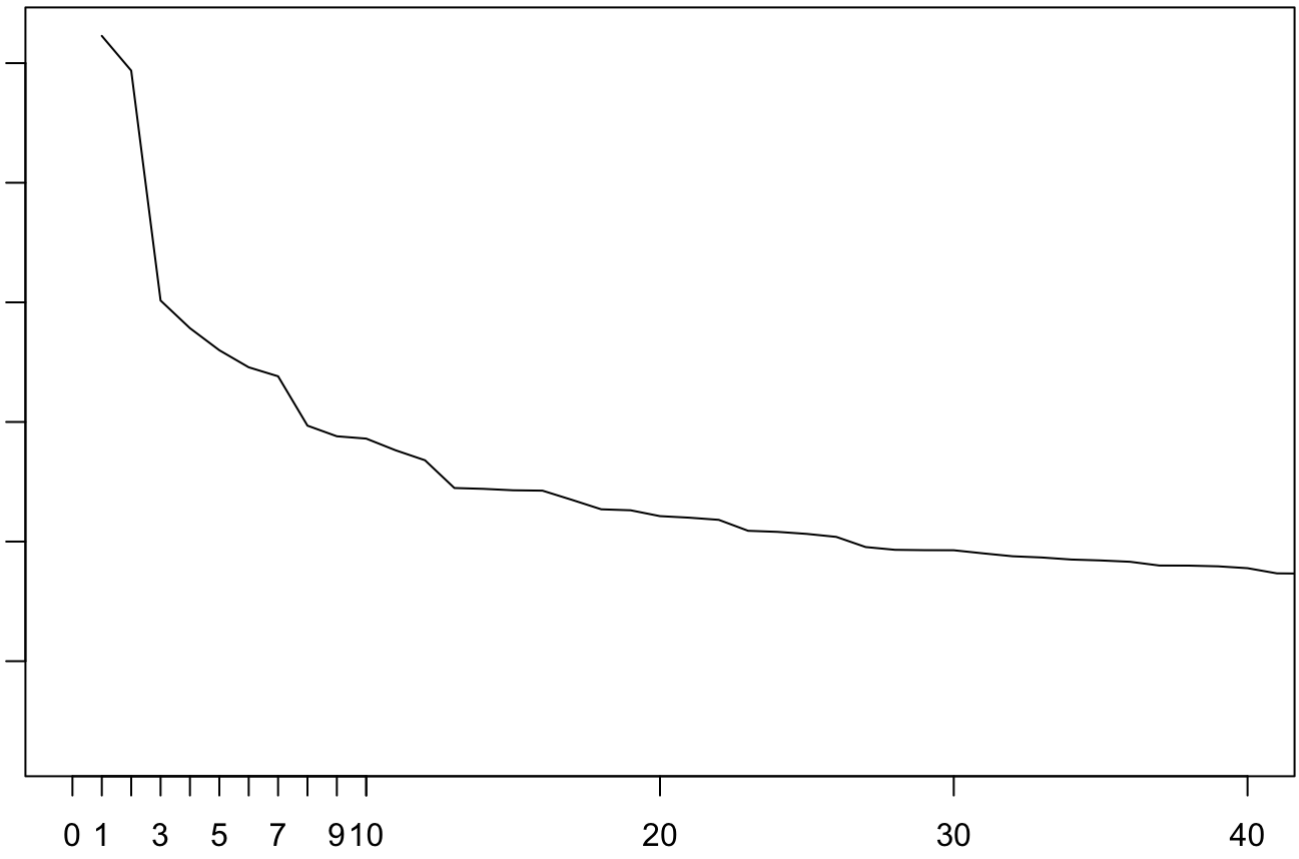
```
hc.dissim <- data.frame(k = seq_along(hclust.mod4$height),
                        dissimilarity = rev(hclust.mod4$height))
head(hc.dissim)
```

```
##   k dissimilarity
## 1 1      6.227624
## 2 2      5.937484
## 3 3      4.015477
## 4 4      3.784050
## 5 5      3.599681
## 6 6      3.457118
```

```
plot(hc.dissim$k, hc.dissim$dissimilarity, type="l")
```




```
par(mar = c(4,1,1,1))
plot(hc.dissim$k, hc.dissim$dissimilarity, type="l", xlim=c(0,40))
axis(side = 1, at = 1:10)
```



We want to choose a number of clusters where the line “pivots”, which can be argued happens from 3 to 15 approximately. We also want a small number of clusters so that it is interpretable. We choose 8.

Analysing the clusters

We look at the number of companies in each cluster from each industry sector

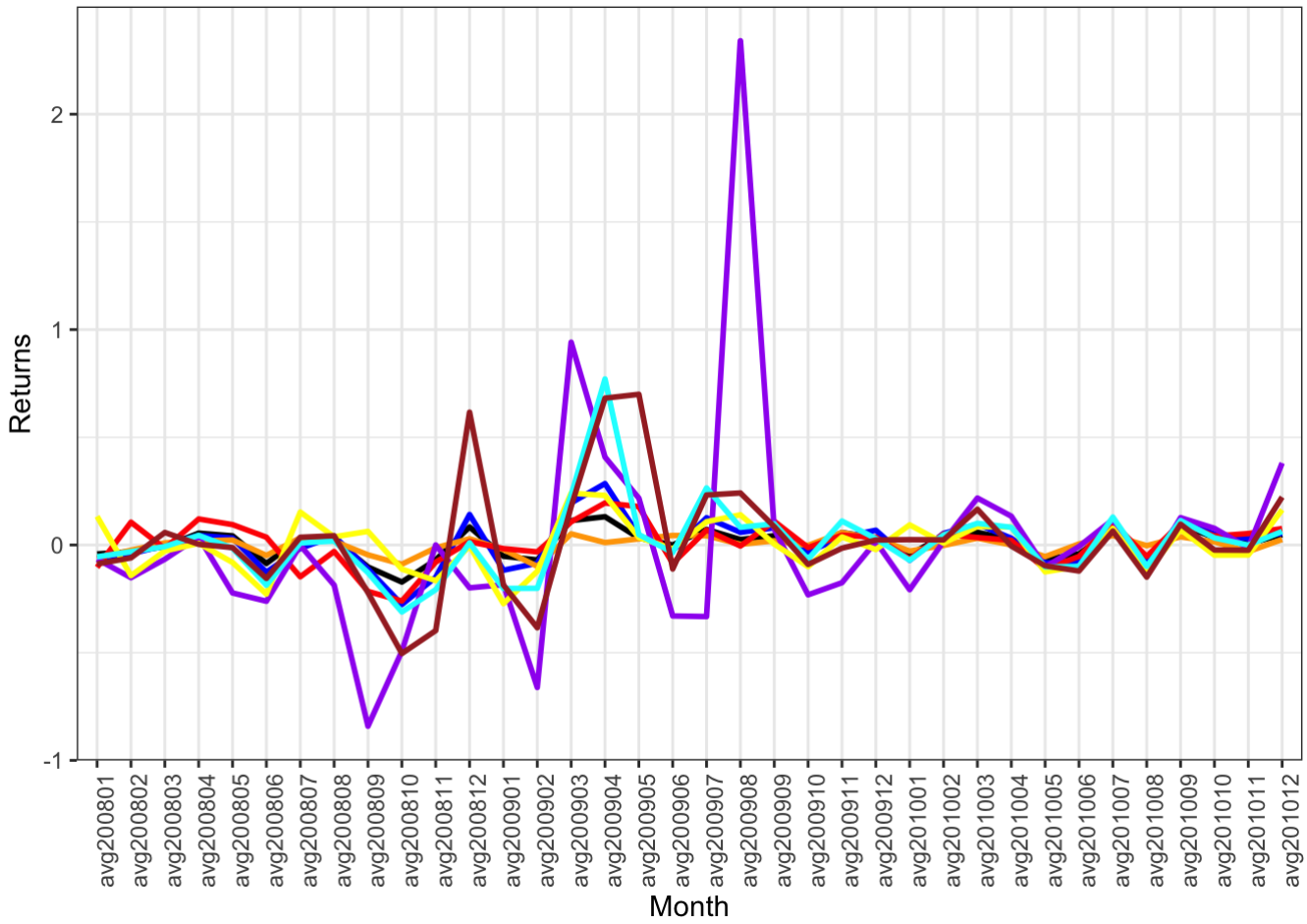
```
h.clusters <- cutree(hclust.mod4, 8)
table(h.clusters, data$Industry)
```

```
##
## h.clusters Consumer Discretionary Consumer Staples Energy Financials
##          1                      29               12      3         18
##          2                      22                3      2         32
##          3                       1               16      0          2
##          4                       3                0     33          0
##          5                       0                0      0          1
##          6                       4                1      0         16
##          7                      10                0      0          4
##          8                       0                0      0          5
##
## h.clusters Health Care Industrials Information Technology Materials
##          1           27          37                     38        16
##          2            0          12                     17         1
##          3           15           1                      0         0
##          4            0           4                      1         7
##          5            0           0                      0         0
##          6            0           0                      0         0
##          7            1           1                      0         4
##          8            1           0                      0         0
##
## h.clusters Telecommunications Services Utilities
##          1                           2         1
##          2                           0         0
##          3                           2        25
##          4                           1         2
##          5                           0         0
##          6                           0         0
##          7                           0         0
##          8                           0         0
```

We see here how many companies we have per cluster per industry. Cluster 1 has a little of everything, cluster 2 has more Consumer Discretionary and Financials. Cluster 7 has almost exclusively Consumer Discretionary, while cluster 8 Financials.

We plot the returns.

```
ggplot(data=transposed, aes(x=month, group=1)) +
  geom_line(aes(y=X1),lwd=1, color = "black") +
  geom_line(aes(y=X2),lwd=1, color = "blue") +
  geom_line(aes(y=X3),lwd=1, color = "orange") +
  geom_line(aes(y=X4),lwd=1, color = "red") +
  geom_line(aes(y=X5),lwd=1, color = "purple") +
  geom_line(aes(y=X6),lwd=1, color = "yellow") +
  geom_line(aes(y=X7),lwd=1, color = "cyan") +
  geom_line(aes(y=X8),lwd=1, color = "brown") +
  theme_bw() +
  xlab("Month") +
  ylab("Returns") + theme(axis.text.x=element_text(angle=90))
```



We see some clusters are much more volatile than others such as cluster 5 (the purple line) which only has one stock. This is a great argument for portfolio deiversification. Overall the stocks move in the same way based on market movements.

Average returns by cluster in Oct 2008 - all clusters underperform (probably related to the financial crisis)

```
avg_returns$avg200810
```

```
## [1] -0.17205266 -0.27638585 -0.08925413 -0.26023813 -0.49329651 -0.11201718
## [7] -0.31099412 -0.50388025
```

Average returns by cluster March 2009 - all clusters give positive returns.

```
avg_returns$avg200903
```

```
## [1] 0.11329471 0.19541286 0.05148726 0.10755846 0.94117177 0.24063092
## [7] 0.23100897 0.17827922
```

K-means algorithm

```
set.seed(177)
km <- kmeans(returns, centers = 8, iter.max=100)
km.centroids <- km$centers
km.clusters <- km$cluster
table(km.clusters)
```

```
## km.clusters
##   1   2   3   4   5   6   7   8
##  30  34  55  84   8   2 142  78
```

The clusters are somewhat similar, we have one cluster with 142 observations and one with 2 (verus 183 and 1 in hierarchical clustering). But we see that the observations are more dispersed within the clusters, if in hierarchincal we had some very large and some very small clusters, here they are more homogenuous.

```
km$tot.withinss
```

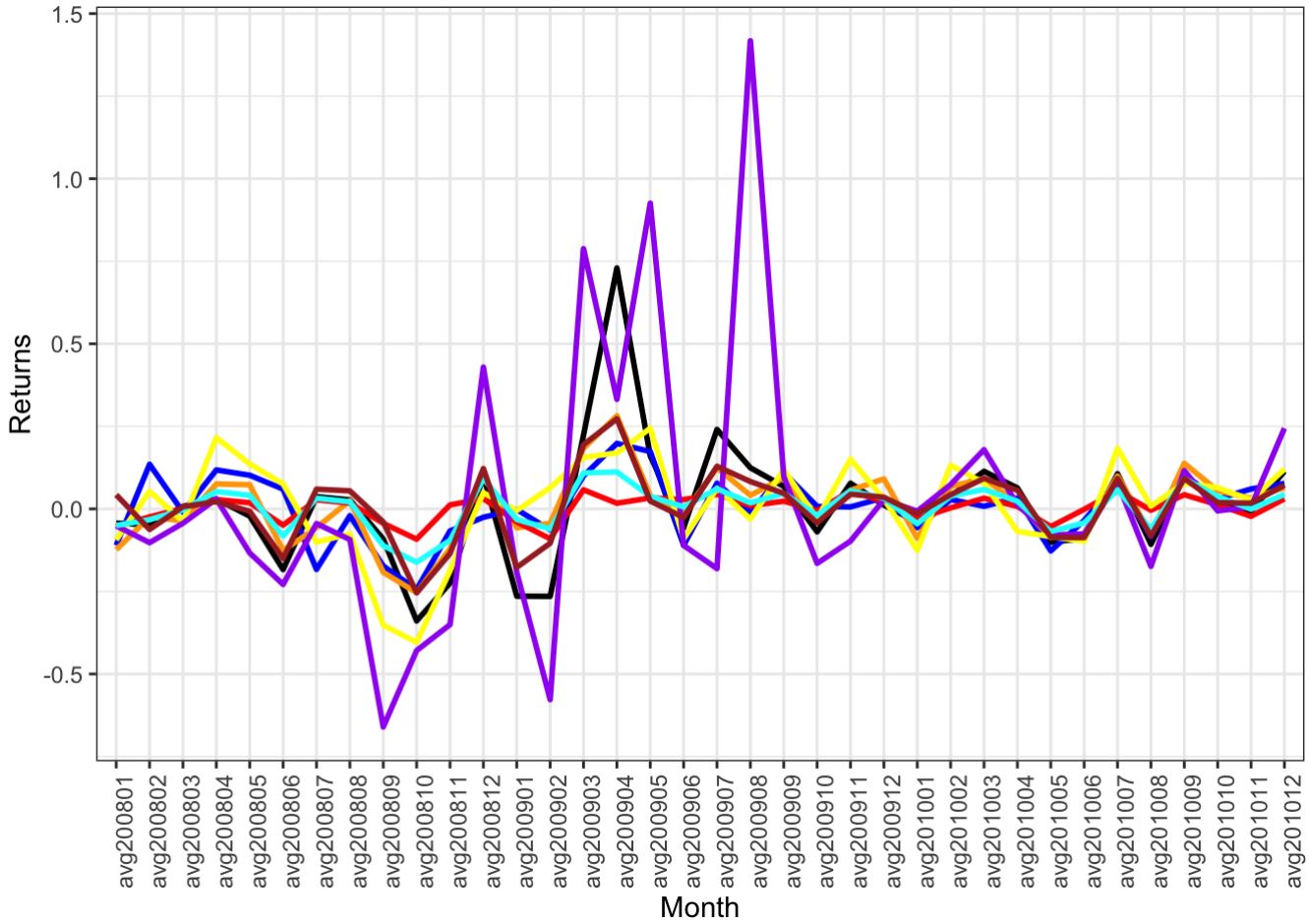
```
## [1] 258.5645
```

```
table(km.clusters, data$Industry)
```

```
##
## km.clusters Consumer Discretionary Consumer Staples Energy Financials
##           1                      9                0      0          14
##           2                      0                0     32           0
##           3                      9                0      2           4
##           4                      4               21      2           6
##           5                      0                0      1           0
##           6                      0                0      0           2
##           7                     21                7      1          15
##           8                     26                4      0          37
##
## km.clusters Health Care Industrials Information Technology Materials
##           1           1           2                     0         4
##           2           0           2                     0         0
##           3           2          11                    25         2
##           4          13           3                     1         3
##           5           0           1                     0         5
##           6           0           0                     0         0
##           7          27          31                    29        10
##           8           1           5                     1         4
##
## km.clusters Telecommunications Services Utilities
##           1                            0         0
##           2                            0         0
##           3                            0         0
##           4                            4        27
##           5                            1         0
##           6                            0         0
##           7                            0         1
##           8                            0         0
```

When visualising per industry, we find cluster 3 with Information Technology and Industrials, clusters 7 and 8 with a little of everything and cluster 2 with energy. This has both similarities and differences to hierarchical clustering. Here the clusters are more mixed.

```
avg_returns2 = aggregate(reduced, by=list(km.clusters), mean) %>% select(-Group.1)
transposed2 = data.frame(t(avg_returns2))
transposed2$month = row.names(transposed2)
ggplot(data=transposed2, aes(x=month, group=1)) +
  geom_line(aes(y=X1),lwd=1, color = "black") +
  geom_line(aes(y=X2),lwd=1, color = "blue") +
  geom_line(aes(y=X3),lwd=1, color = "orange") +
  geom_line(aes(y=X4),lwd=1, color = "red") +
  geom_line(aes(y=X5),lwd=1, color = "yellow") +
  geom_line(aes(y=X6),lwd=1, color = "purple") +
  geom_line(aes(y=X7),lwd=1, color = "cyan") +
  geom_line(aes(y=X8),lwd=1, color = "brown") +
  theme_bw() +
  xlab("Month") +
  ylab("Returns") + theme(axis.text.x=element_text(angle=90))
```



The plots are very similar to hierarchincal clustering.

Average returns by cluster in Oct 2008 - all clusters underperform (probably related to the financial crisis)

```
avg_returns2$avg200810
```

```
## [1] -0.33926920 -0.24440021 -0.25347300 -0.09169947 -0.40377965 -0.42903754
## [7] -0.16150913 -0.25416202
```

Average returns by cluster March 2009 - all clusters give positive returns.

```
avg_returns2$avg200903
```

```
## [1] 0.22342970 0.09988884 0.18333571 0.05773507 0.15577384 0.78797719
## [7] 0.10884172 0.19457246
```

Discussion

No matter how we cluster the stocks, a mixed portfolio will be stable over time, while a more specialized selection may give higher returns but will also incur high risks. As a portfolio manager, it depends on the risk apetitie of the client to decide the best strategy for investment. The stock that stands our in cluster 5 (hierarchcal) is a financial company and it significantly outperforms the market in 2009, but has been stable since. Portfolio managers should not rely solely on this analysis and they require company-specific and industry specific knowledge.