Data Quality: Lecture 6

Philipp Drebes

03.04.2023

Task 01

A data set *learning* is given. The dataset *learning* shows the relationship between learning effort in self-study [hours per week] and success on the final exam [index 0 to 10] in a master's program.

Run a "Cluster Based" outlier detection applying k-means clustering after Barai & Dey (2017). Calculate solutions with two different numbers of clusters. x = 3 and x = 4 in kmeans(datas, centers = x, nstart = 10)

How can the solutions be interpreted?

Load & Explore

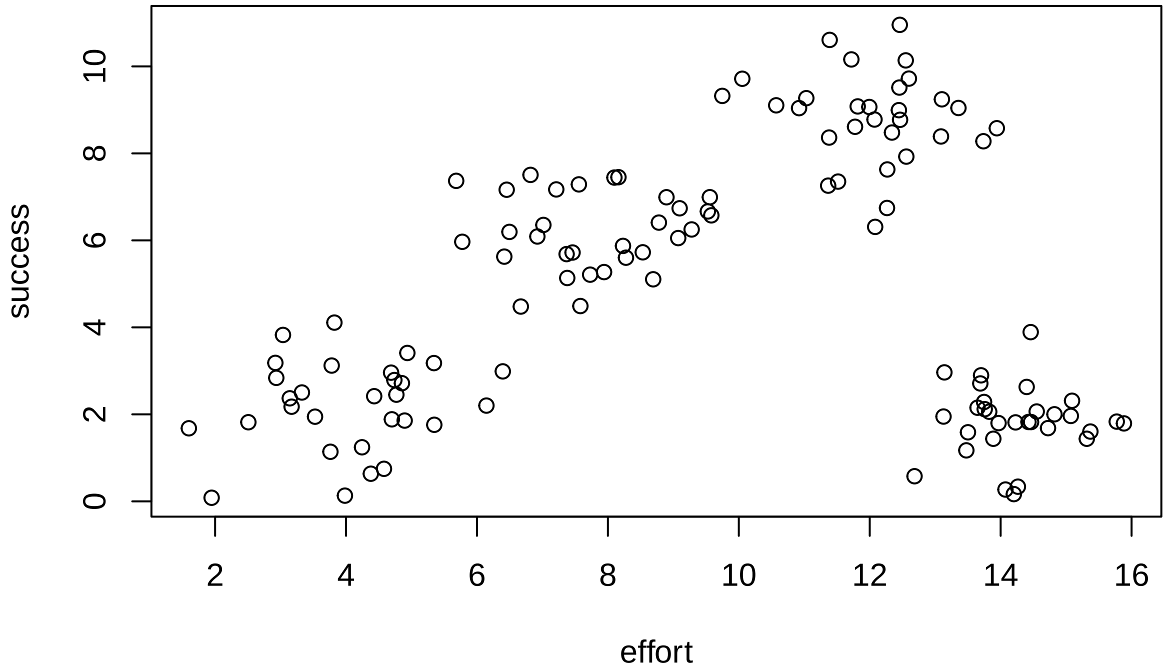
```
library(readxl)
library(MVA)
library(cluster)

data <- read_excel("data/learning.xlsx")
any(is.na(data))
```

```
## [1] FALSE
```

```
summary(data)
```

```
##      effort          success
##  Min.   : 1.599   Min.   : 0.08273
##  1st Qu.: 6.332   1st Qu.: 1.99214
##  Median : 9.664   Median : 4.48437
##  Mean   : 9.571   Mean   : 4.80295
##  3rd Qu.:13.194   3rd Qu.: 7.35565
##  Max.   :15.884   Max.   :10.95476
```

```
plot(data)
```

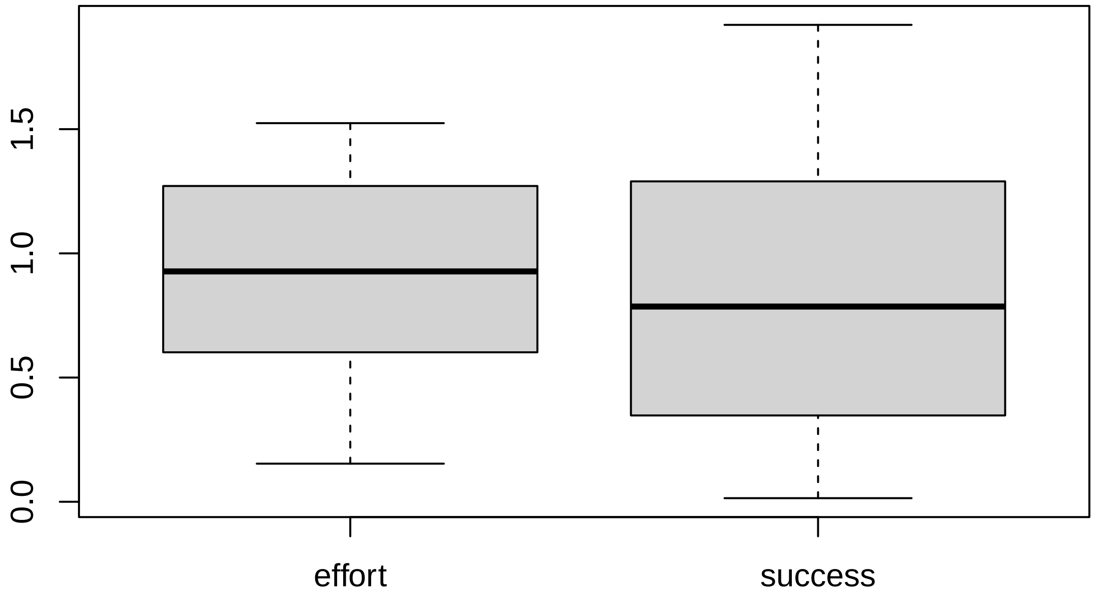


Preparation

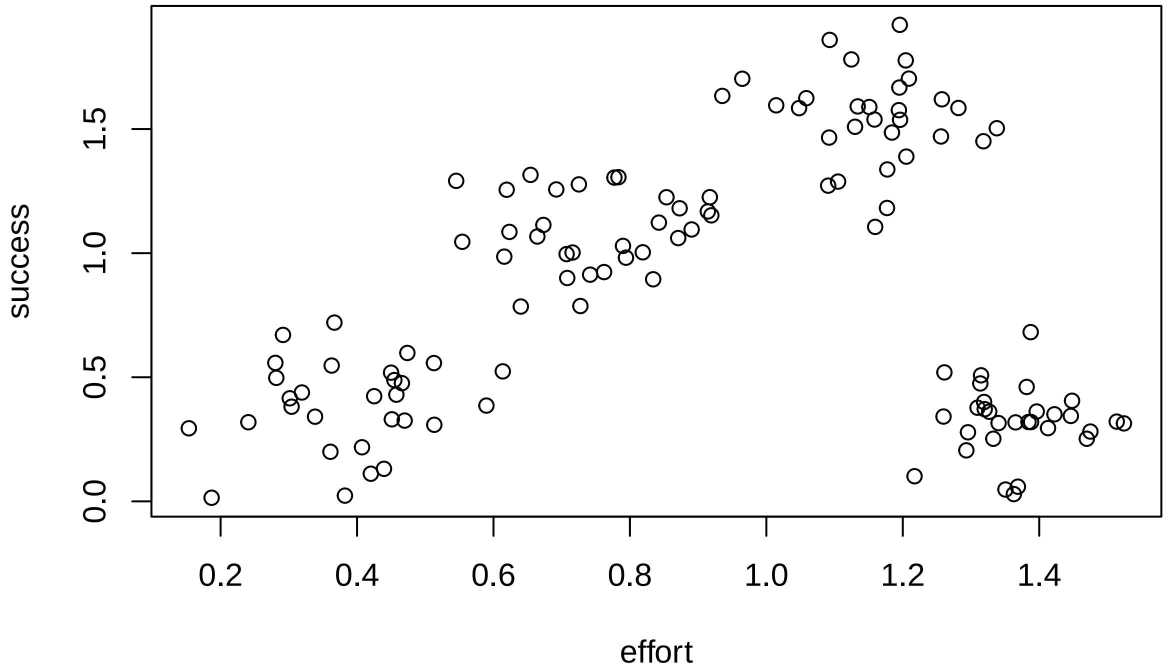
```
data.scaled <- scale(data, center = FALSE, scale = TRUE)
dists <- dist(data.scaled)
summary(data.scaled)
```

```
##      effort          success
##  Min.   :0.1534   Min.   :0.0145
##  1st Qu.:0.6076   1st Qu.:0.3491
##  Median :0.9274   Median :0.7859
##  Mean   :0.9184   Mean   :0.8418
##  3rd Qu.:1.2662   3rd Qu.:1.2891
##  Max.   :1.5242   Max.   :1.9199
```

```
boxplot(data.scaled)
```



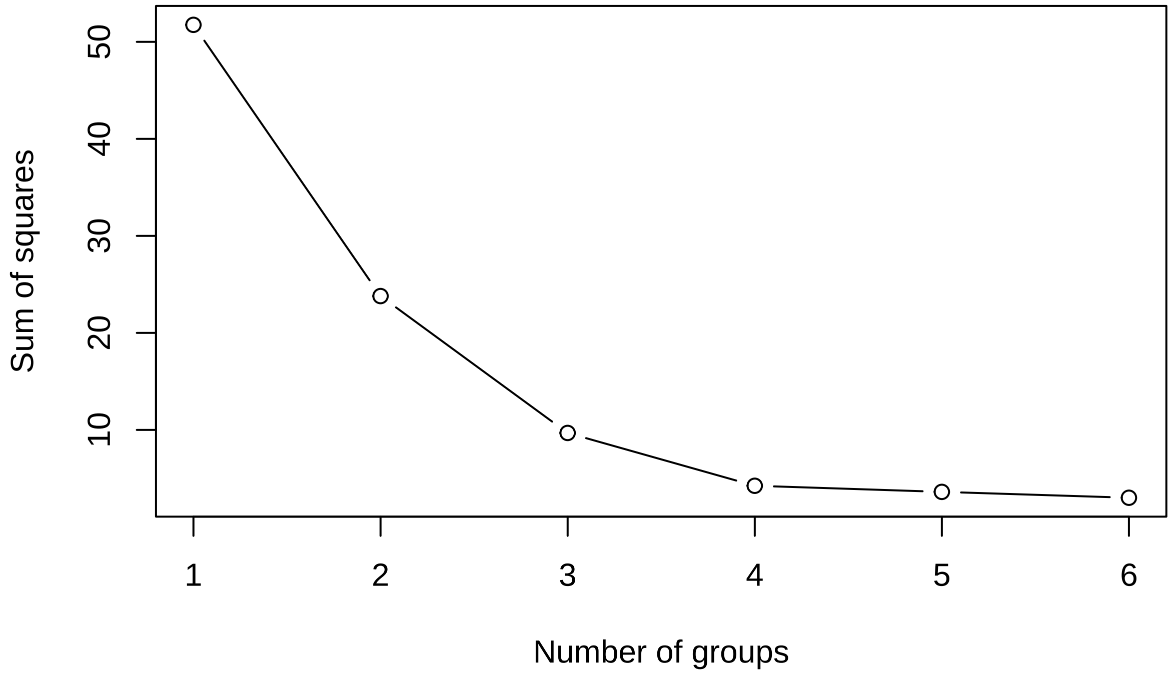
```
plot(data.scaled)
```



Clustering

Scree plot

```
reps <- rep(0, 6)
for (i in 1:6) reps[i] <- sum(kmeans(data.scaled, centers = i, nstart = 20)$withinss)
plot(1:6, reps, type = "b", xlab = "Number of groups", ylab = "Sum of squares")
```



4 Clusters

```
km.4 <- kmeans(data.scaled, centers = 4, nstart = 10)
km.4.groups <- km.4$cluster
km.4.groups
```

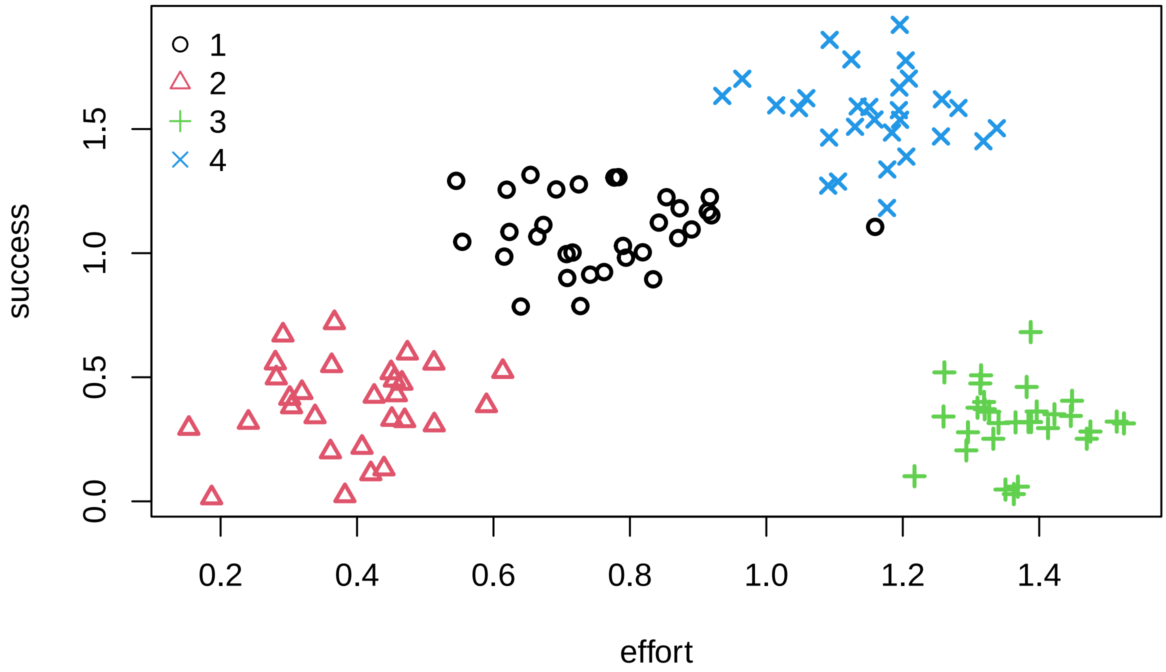
```
##   [1] 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4
##  [75] 4 4 4 4 1 4 4 4 4 4 4 4 4 4 4 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [112] 3 3 3 3 3 3 3 3 3
```

```
cluster.size.4 <- cbind(sum(km.4.groups == 1), sum(km.4.groups == 2),
                        sum(km.4.groups == 3), sum(km.4.groups == 4))
cluster.size.4
```

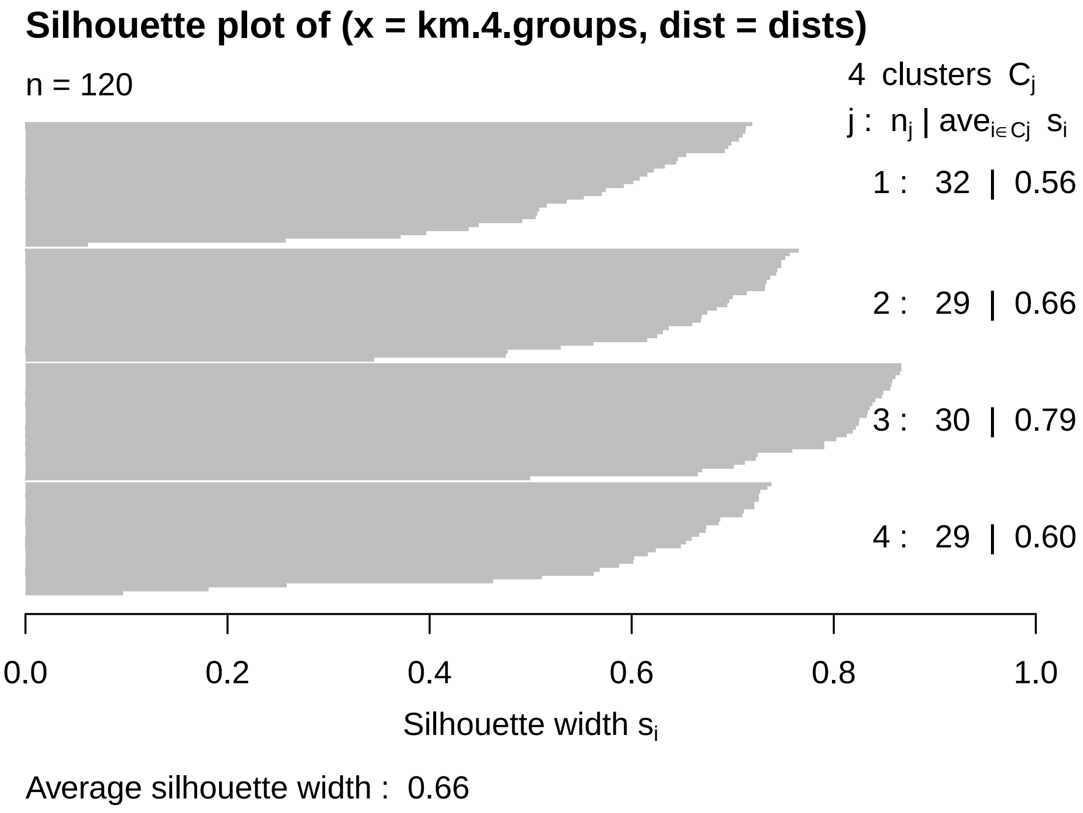
```
##      [,1] [,2] [,3] [,4]
## [1,]   32   29   30   29
```

Plot / Silhouette plot

```
plot(data.scaled, pch = km.4.groups, col=km.4.groups, lwd=2)
legend("topleft", legend = 1:4, pch = 1:4, col=1:4, bty="n")
```



```
plot(silhouette(km.4.groups, dists))
```



3 Clusters

```
km.3 <- kmeans(data.scaled, centers = 3, nstart = 10)
km.3.groups <- km.3$cluster
km.3.groups
```

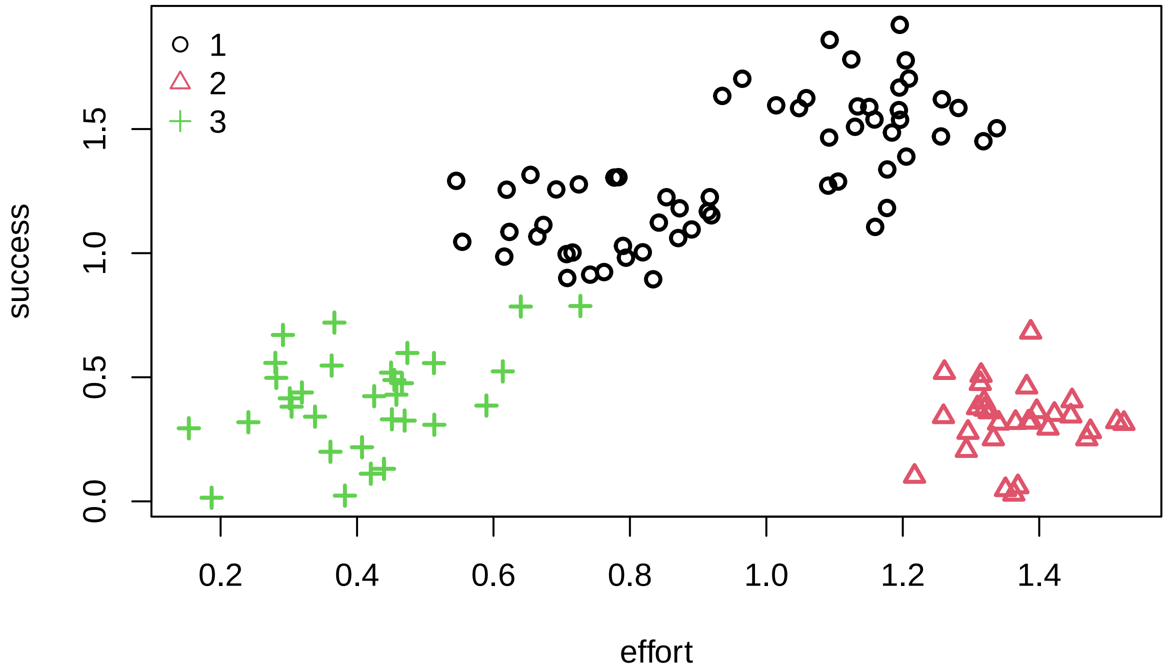
```
##   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [112] 2 2 2 2 2 2 2 2 2
```

```
cluster.size.3 <- cbind(sum(km.3.groups == 1), sum(km.3.groups == 2),
                        sum(km.3.groups == 3))
cluster.size.3
```

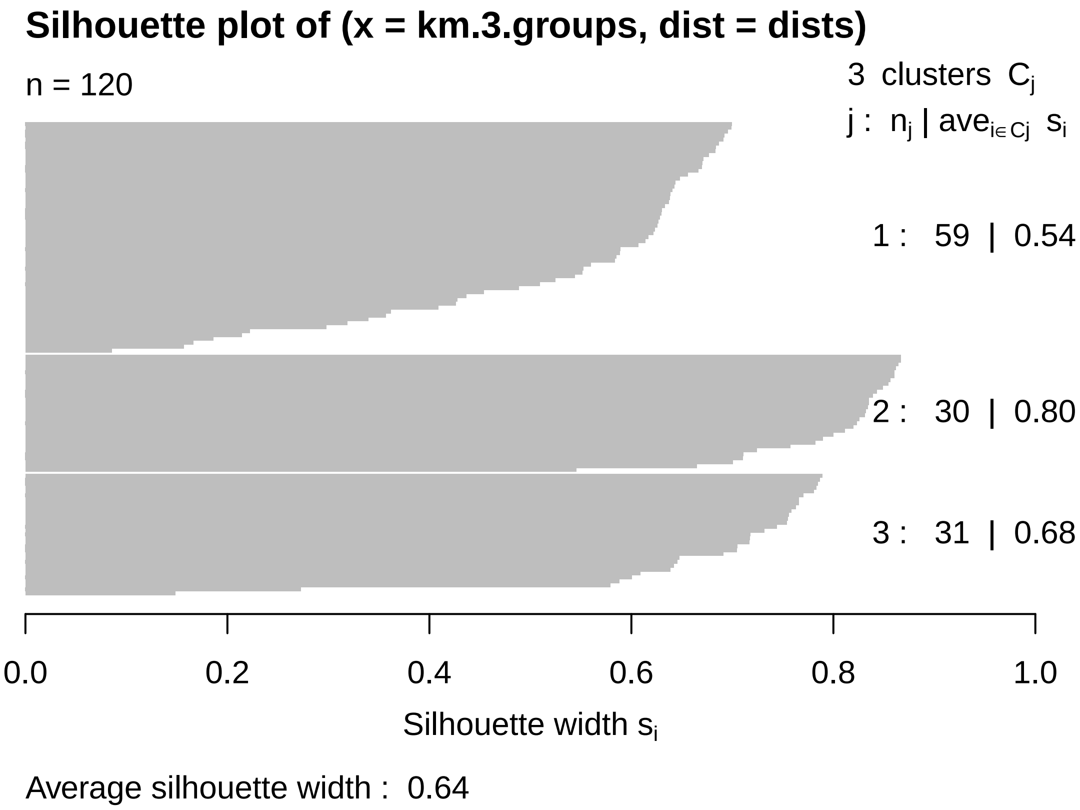
```
##      [,1] [,2] [,3]
## [1,]   59   30   31
```

Plot / Silhouette plot

```
plot(data.scaled, pch = km.3.groups, col=km.3.groups, lwd=2)
legend("topleft", legend = 1:3, pch = 1:3, col=1:3, bty="n")
```



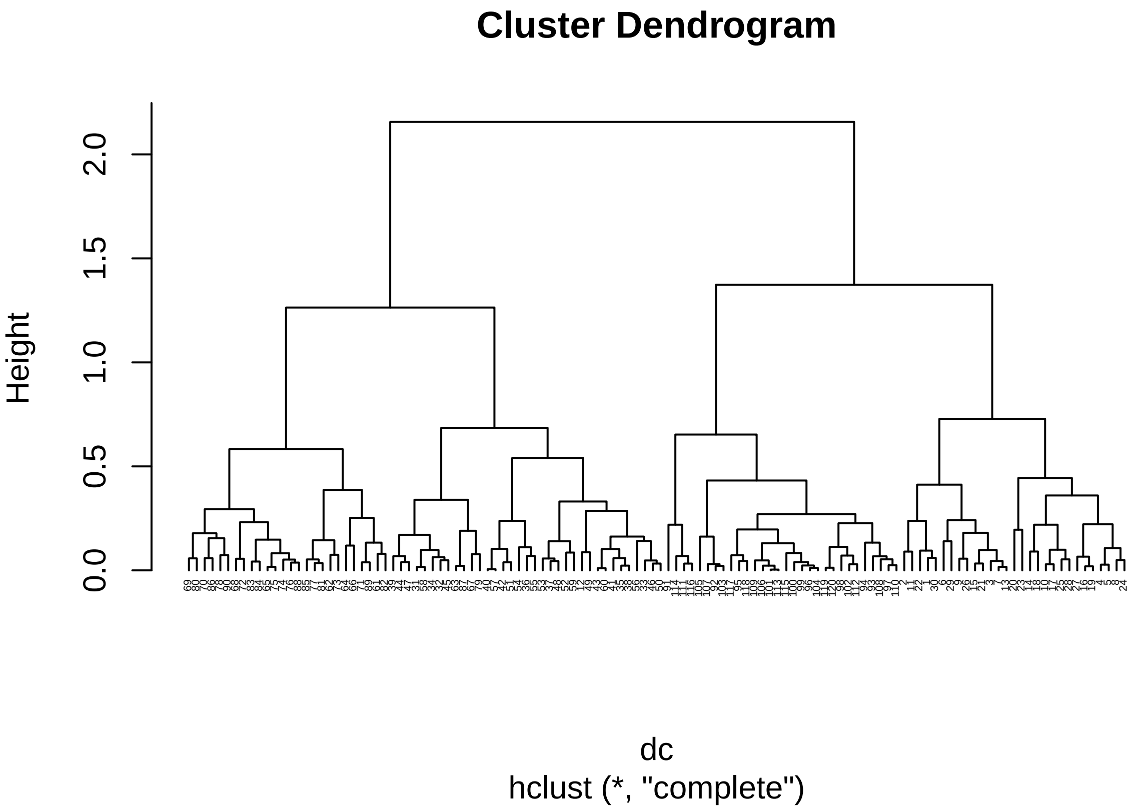
```
plot(silhouette(km.3.groups, dists))
```



Hierarchical cluster analysis

```
dc <- dist(data.scaled, method = "euclidean")
dc
```

```
cc <- hclust(dc, method = "complete")
plot(cc,cex = 0.3, hang = -1)
```



Interpretation

According to the numbers you could argue for 3 and 4 clusters. However, 4 clusters seems to be more appropriate considering the domain of the data set.

1. Those who learn little and have a bad grade
2. Those who learn an average amount and have an average grade
3. Those who learn a lot and have good performance
4. Those who learn a lot and still have bad performance (maybe not an appropriate learning technique)