**Homework 07 Joseph Blubaugh** jblubau1@tamu.edu **STAT 636-720**

1. The Average method and Single method look similar while the complete method looks very different

Steps for Single Linkage

| | **A** | **B** | **C** | **D** |
|---|---|---|---|---|
| A | 0 | | | |
| B | 1 | 0 | | |
| C | 11 | 2 | 0 | |
| D | 5 | 3 | 4 | 0 |

| | **AB** | **C** | **D** |
|---|---|---|---|
| AB | 0 | | |
| C | 2 | 0 | |
| D | 3 | 4 | 0 |

| | **ABC** | **D** |
|---|---|---|
| ABC | 0 | |
| D | 3 | 0 |

Steps for Complete Linkage

| | **A** | **B** | **C** | **D** |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| A | 0 | | | |
| B | 1 | 0 | | |
| C | 11 | 2 | 0 | |
| D | 5 | 3 | 4 | 0 |

| | **AB** | **C** | **D** |
|---|---|---|---|
| AB | 0 | | |
| C | 11 | 0 | |
| D | 5 | 4 | 0 |

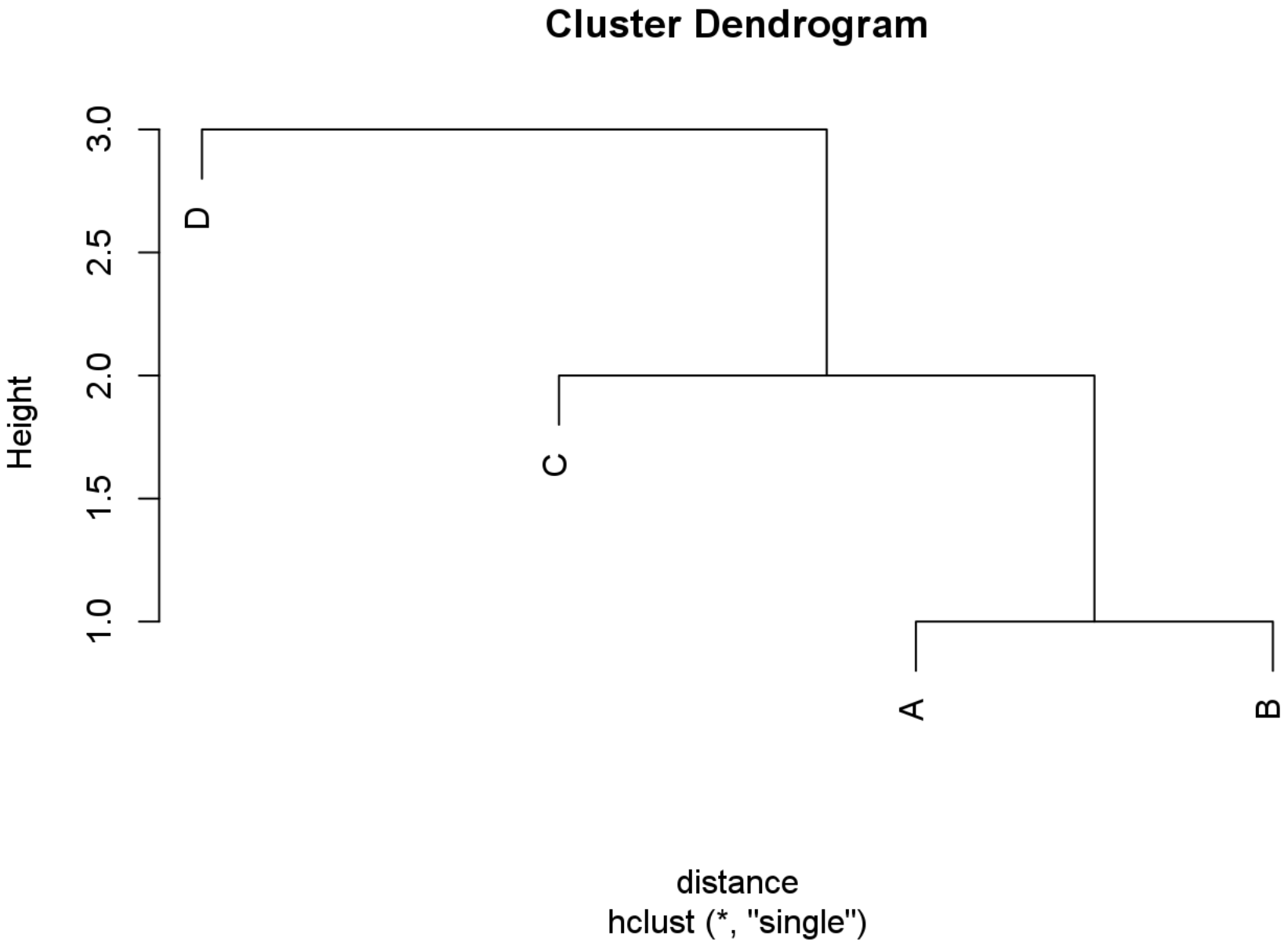
| | **AB** | **CD** |
|---|---|---|
| ABC | 0 | |
| D | 11 | 0 |

Steps for Average Linkage

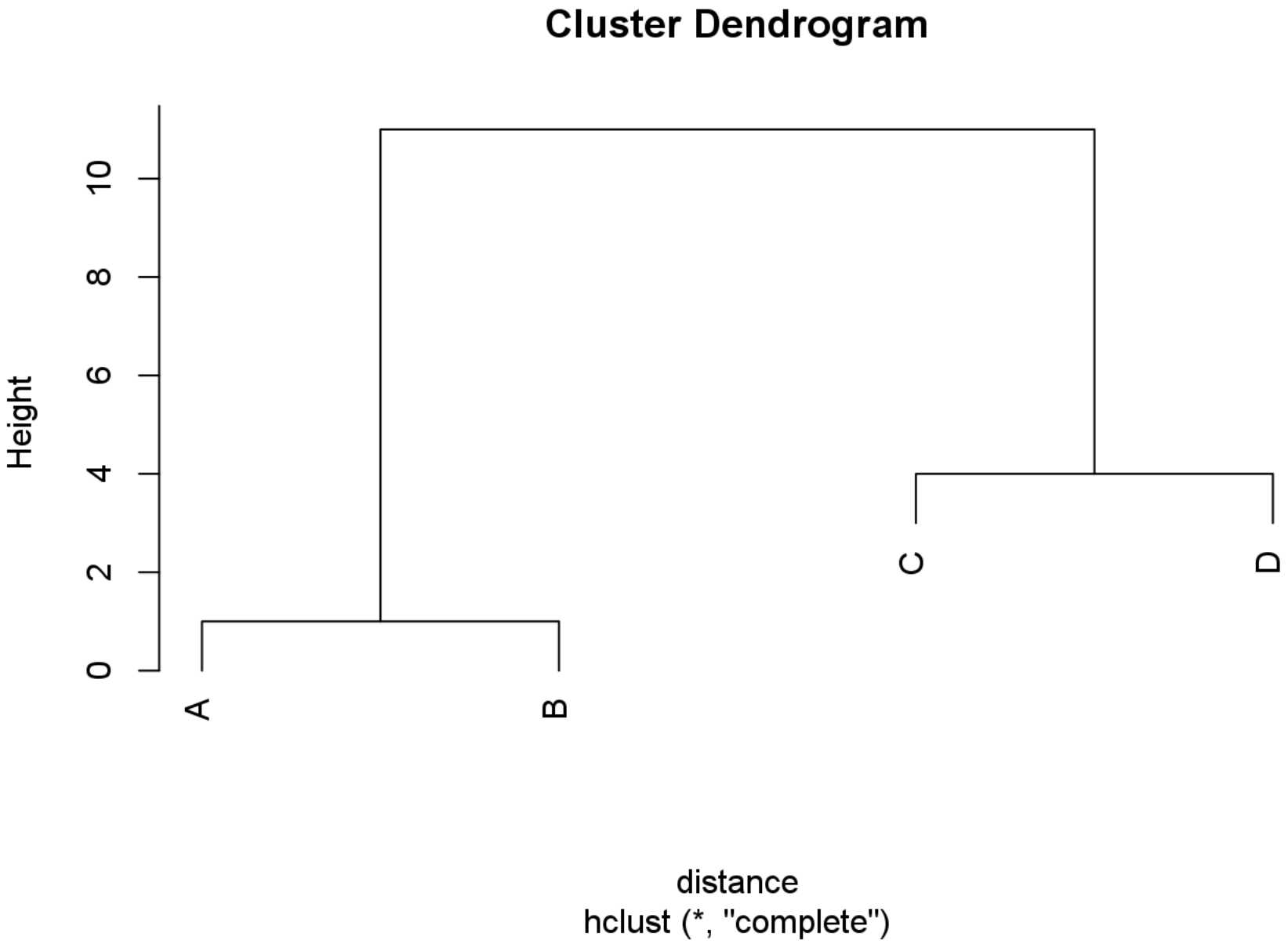
| | **A** | **B** | **C** | **D** |
|---|---|---|---|---|
| A | 0 | | | |
| B | 1 | 0 | | |
| C | 11 | 2 | 0 | |
| D | 5 | 3 | 4 | 0 |

| | AB | C | D |
|---|---|---|---|
| AB | 0 | | |
| C | 6.5 | 0 | |
| D | 4 | 4 | 0 |

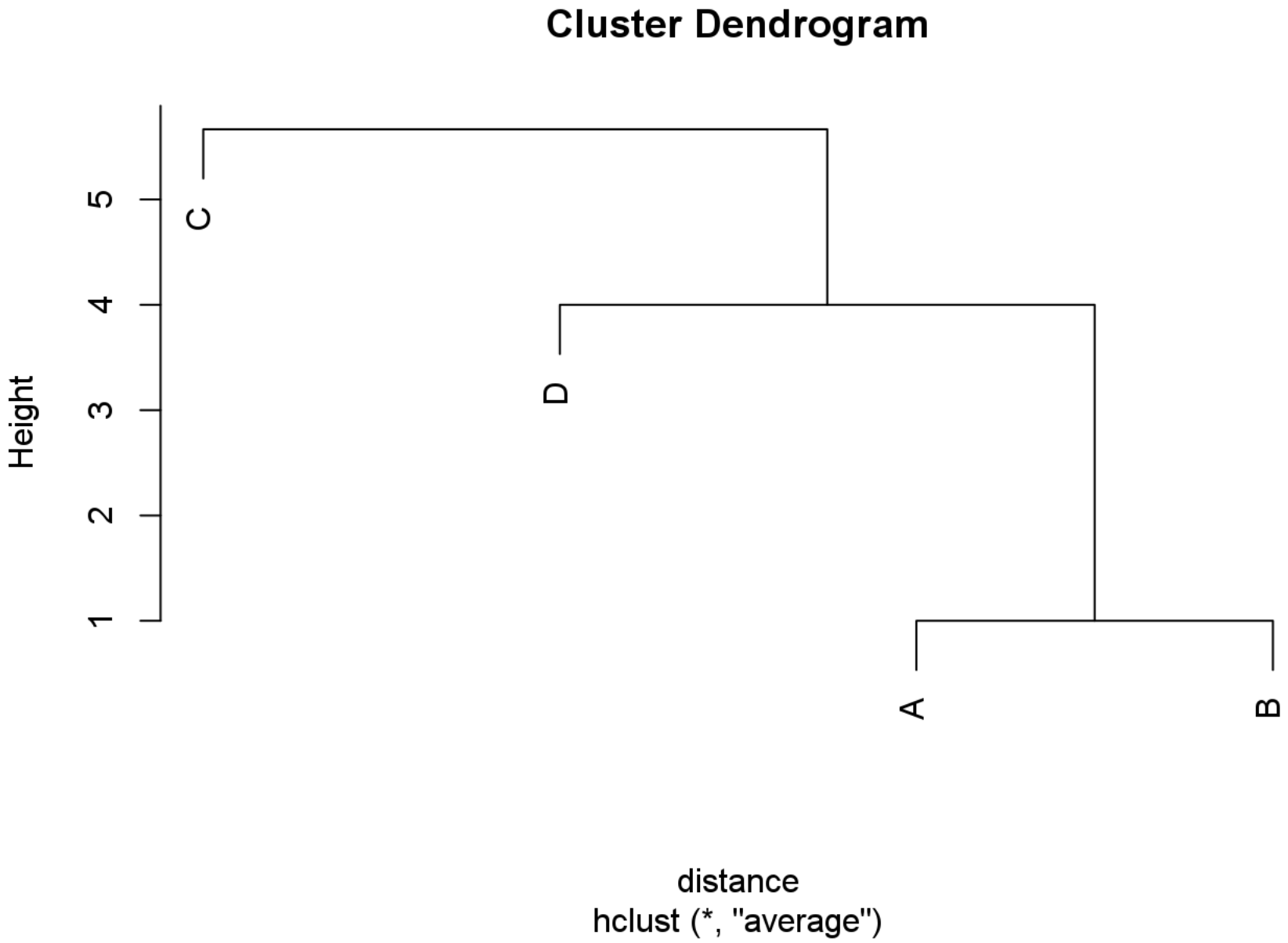
```
distance = data.frame(
  x1 = c("A", "A", "A", "B", "B", "C"),
  x2 = c("B", "C", "D", "C", "D", "D"),
  Distance = c(1, 11, 5, 2, 3, 4)
)

distance = with(distance, structure(Distance,
                           Size = 4,
                           Labels = c("A", "B", "C", "D"),
                           Diag = FALSE,
                           Upper = FALSE,
                           method = "user",
                           class = "dist"))

## Single
plot(hclust(distance, method = "single"))
```




```
## Complete
plot(hclust(distance, method = "complete"))
```




```
## Average
plot(hclust(distance, method = "average"))
```




2. Kmeans

a.

Data:

| Item | X1 | X2 |
|---|---|---|
| A | 5 | 4 |
| B | 1 | -2 |
| C | -1 | 1 |
| D | 3 | 1 |

Initial Clusters:

- AB (X1) = $(5+1)/2 = 3$
- AB (X2) = $(4-2)/2 = 1$
- CD (X1) = $(-1+3)/2 = 1$
- CD (X2) = $(1+1)/2 = 1$

Step 1:

- $d^2(A, AB) = (5-3)^2 + (4-1)^2 = 13$ **closest**
- $d^2(A, ACD) = (5+3)^2 + (4+2)^2 = 100$

- $d^2(B, AB) = (1-3)^2 + (-2-1)^2 = 13$ **no reassignment**
- $d^2(B, BCD) = (1+1)^2 + (-2+6)^2 = 20$

- $d^2(C, CAB) = (-1-7)^2 + (1-1)^2 = 64$
- $d^2(C, CD) = (-1-1)^2 + (1-1)^2 = 4$ **no reassignment**

- $d^2(D, CD) = (3-3)^2 + (1-1)^2 = 0$
- $d^2(D, DAB) = (3-5)^2 + (1-1)^2 = 4$ **no reassignment**

Complete iteration with no reassignments, clusters (AB) and (CD) are optimal

b.

| Item | X1 | X2 |
|---|---|---|
| A | 5 | 4 |
| B | 1 | -2 |
| C | -1 | 1 |
| D | 3 | 1 |

Initial Clusters:

- AC (X1) = $(5-1)/2 = 2$
- AC (X2) = $(4+1)/2 = 2.5$
- BD (X1) = $(1+3)/2 = 2$
- BD (X2) = $(-2+1)/2 = -.5$

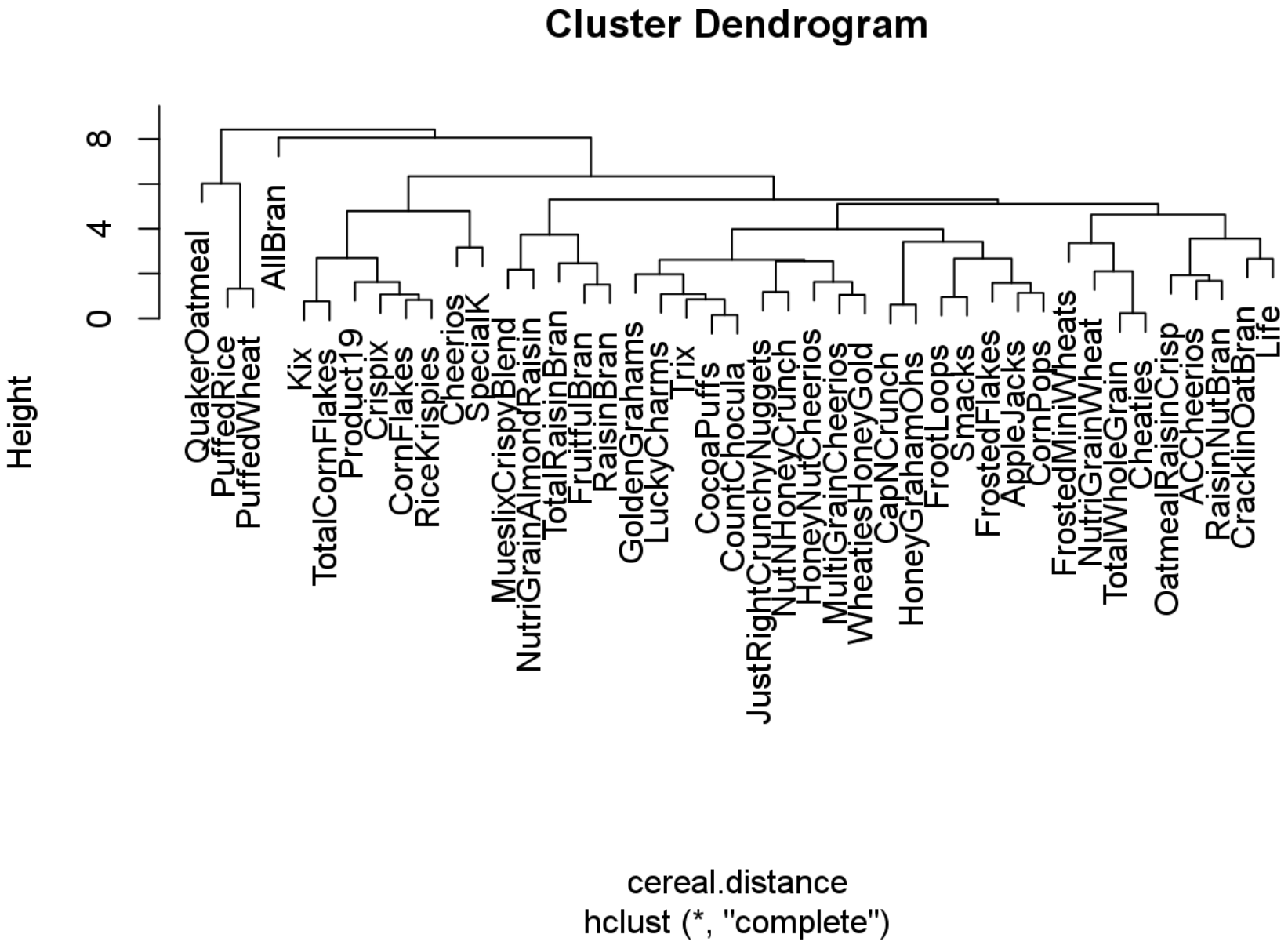
Step 1:

- $d^2(A, AC) = (5-2)^2 + (4-2.5)^2 = 11.25$ **No reassignment**
- $d^2(A, ABD) = (5+1)^2 + (5+0)^2 = 61$
- $d^2(B, BAC) = (1-3)^2 + (-2-7)^2 = 85$
- $d^2(B, BD) = (1-3)^2 + (-2+.5)^2 = 6.25$ **no reassignment**
- $d^2(C, AC) = (-1-2)^2 + (1-2.5)^2 = 11.25$ **no reassignment**
- $d^2(C, CBD) = (-1-5)^2 + (1-0)^2 = 37$
- $d^2(D, DAC) = (3-1)^2 + (1-0)^2 = 5$
- $d^2(D, BD) = (3-2)^2 + (1+.5)^2 = 3.25$ **no reassignment**

Complete iteration with no reassignment. It appears the cluster assignment can change depending on the starting groups.

3. The hierarchical clustering method creates many more distinct cluster than the 3 kmeans method using 3 clusters. The first 2 principal components make up around 60% of the variation. Two of the kmeans clusters appears close in proximity when plotting the first two principal components, but there is a 3rd cluster part of the kmeans which all appear to be outliers and may not actually be closely related to one another.

```
## Hierarchical Clustering
cereal = read.table("T11-9.DAT", quote="\"", comment.char="")
row.names(cereal) = cereal$V1
cereal = cereal[, -1]

## Scale the data since the variables have very different ranges
cereal.scale = scale(cereal[, 2:10])

## Create the distance matrix
cereal.distance = dist(cereal.scale)
plot(hclust(cereal.distance))
```

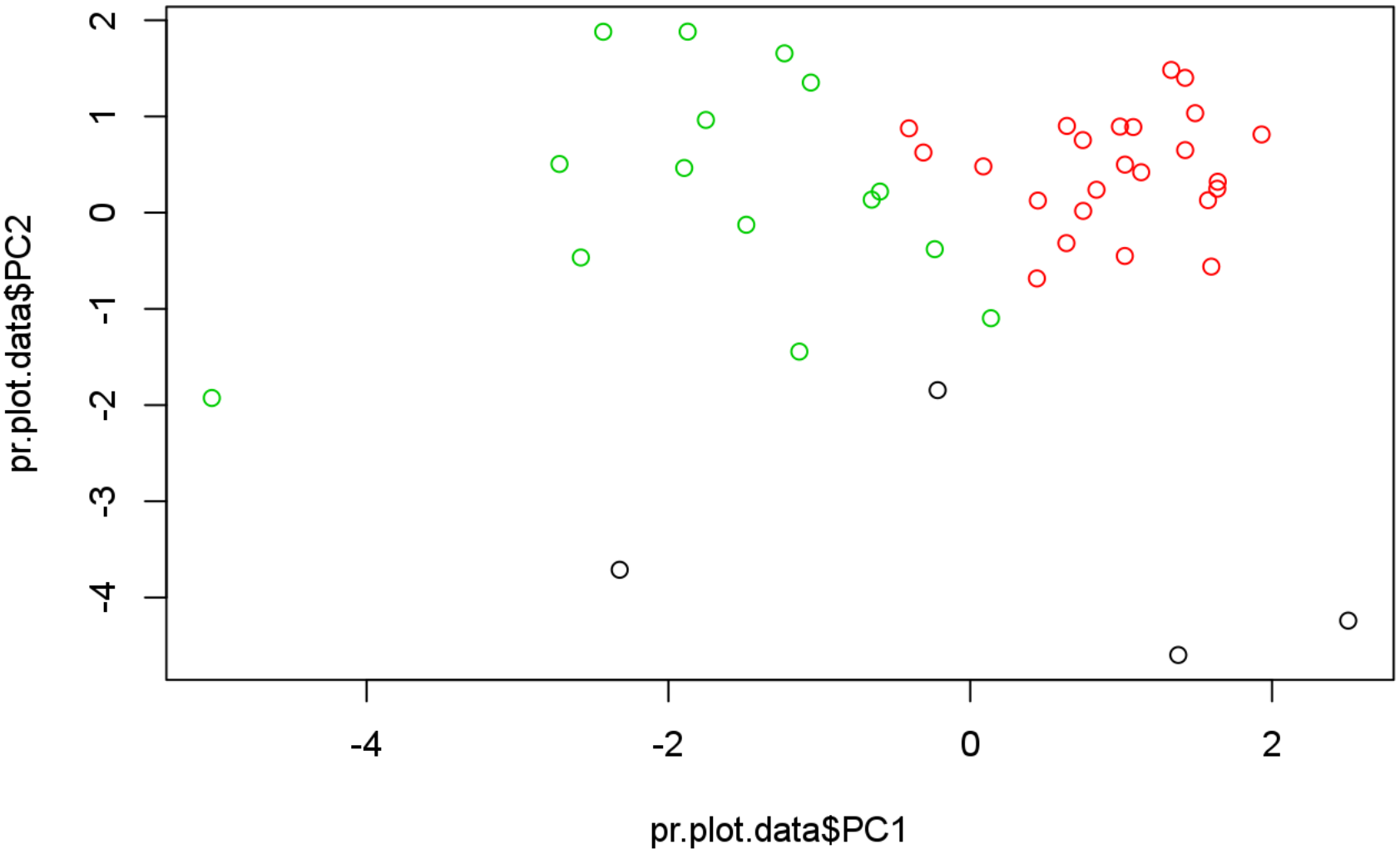



```
## Kmeans Clustering
mdl = kmeans(cereal.scale, centers = 3)

## Principal components
```

```
pr = prcomp(cereal.scale)
summary(pr)
```

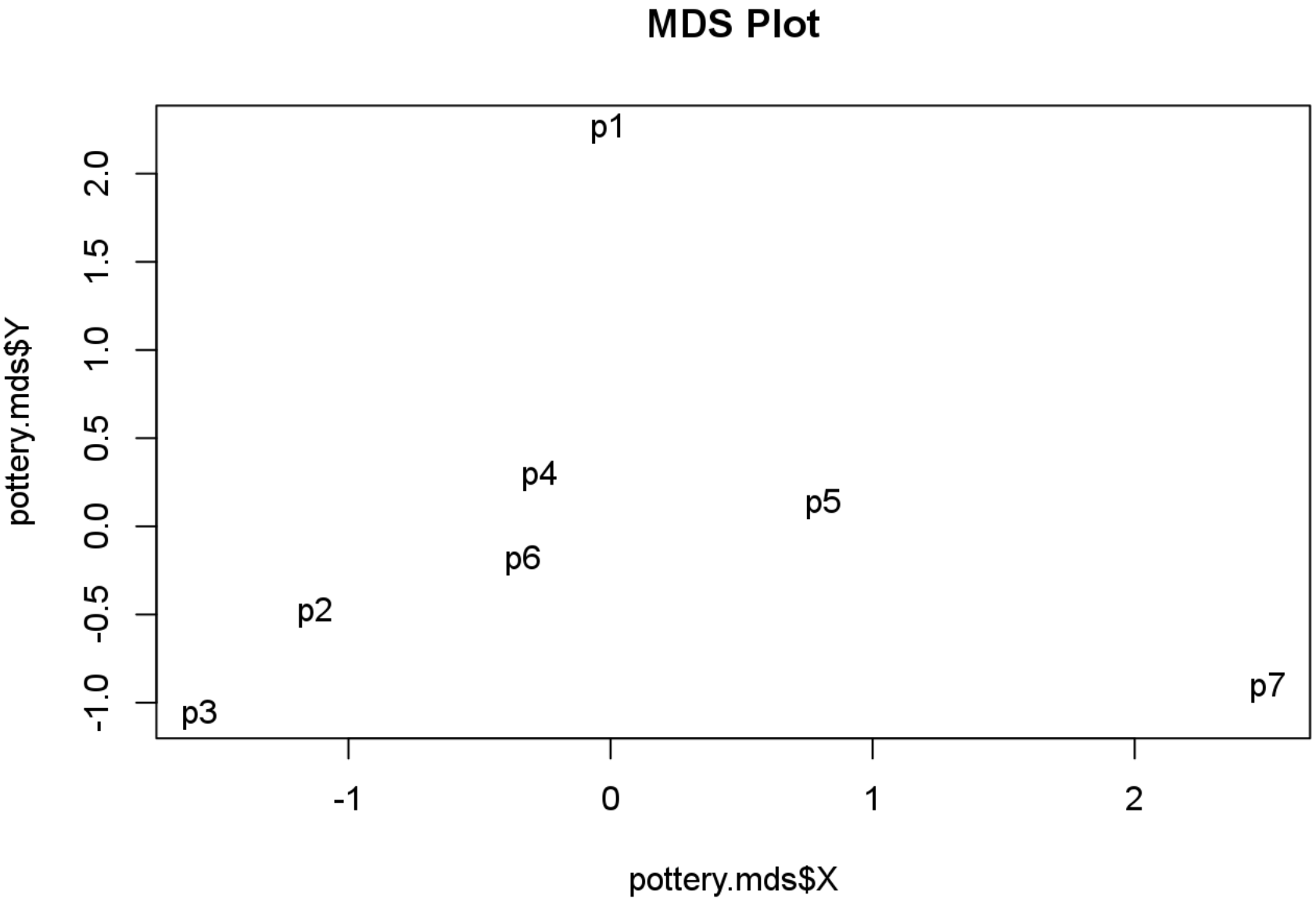
```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.5962 1.4575 1.3396 0.93439 0.85569 0.69217
Proportion of Variance 0.2831 0.2360 0.1994 0.09701 0.08136 0.05323
Cumulative Proportion  0.2831 0.5191 0.7185 0.81552 0.89687 0.95011
                           PC7     PC8     PC9
Standard deviation     0.59364 0.22878 0.21046
Proportion of Variance 0.03916 0.00582 0.00492
Cumulative Proportion  0.98926 0.99508 1.00000
```

```
## Apply principal components to the data to create new variables for plotting
pr.plot.data = data.frame(cereal.scale %*% pr$rotation)
pr.plot.data = pr.plot.data[, 1:2]
pr.plot.data$km.cluster = mdl$cluster

plot(x = pr.plot.data$PC1, y = pr.plot.data$PC2, col = factor(pr.plot.data$km.cluster))
```




4. The two plots looks similar in their dispersion of variables. They are not quite mirror images of each other, but the intial pattern is similar. It looks like P2 and P3 are heavily influenced by the first and 3rd principal compnoents and P1 is heavily influenced by the 2nd and 4th principal components.

```
pottery = read.table("T12-8.DAT", quote="\"", comment.char="")

pottery.mds = data.frame(cmdscale(dist(scale(pottery)), 2))
row.names(pottery.mds) = c("p1", "p2", "p3", "p4", "p5", "p6", "p7")
colnames(pottery.mds) = c("X", "Y")

plot(x = pottery.mds$X, y = pottery.mds$Y, type = "n", main = "MDS Plot")
text(x = pottery.mds$X, y = pottery.mds$Y, labels = row.names(pottery.mds))
```




```
biplot(prcomp(scale(pottery)), main = "BiPlot")
```