Module 11 Homework

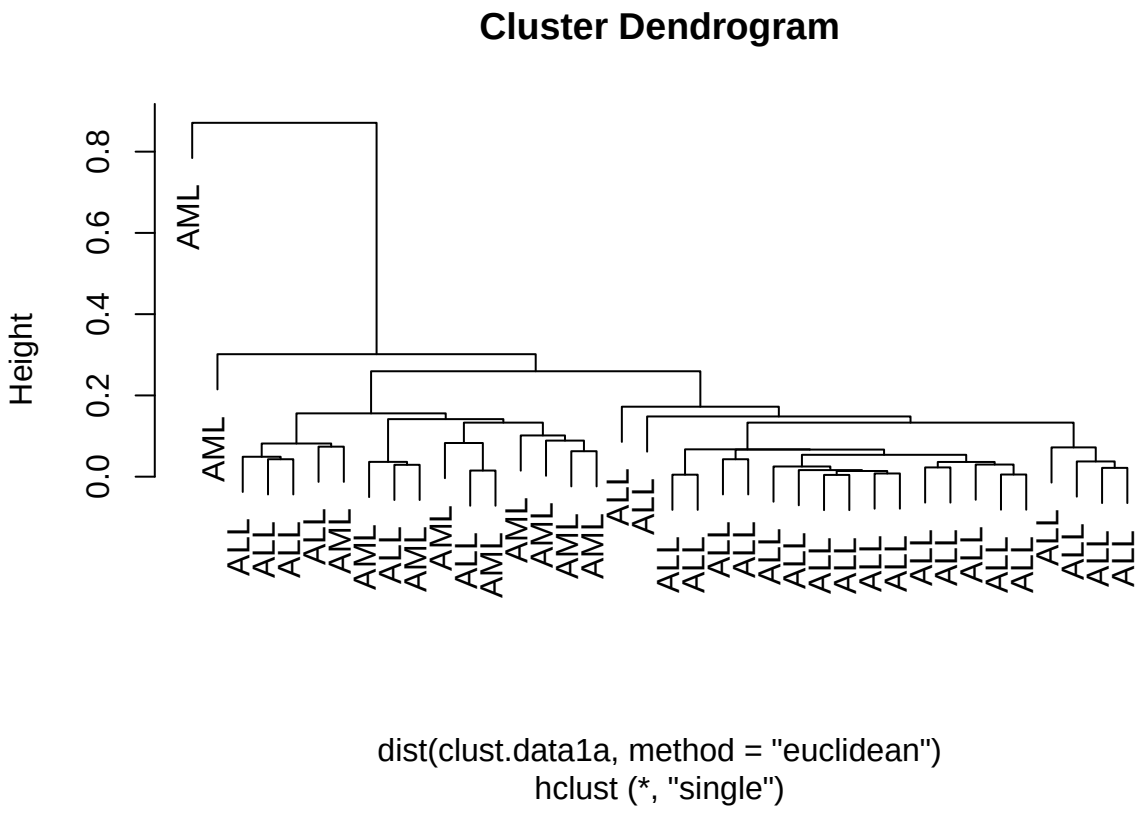
Yanhe Wen

Question 1

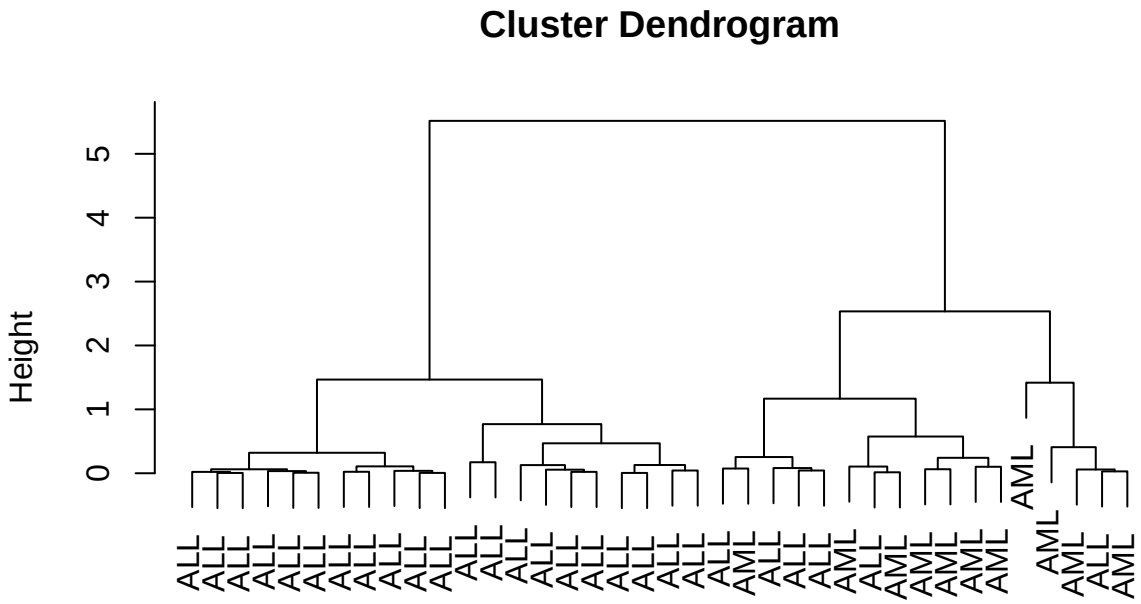
```
library(multtest)
data("golub")
```

(a)

```
gol.fac <- factor(golub.cl, levels=0:1, labels = c("ALL","AML"))
ccnd3 <- golub[grep("CCND3 Cyclin D3", golub.gnames[,2]),]
clust.data1a <- data.frame(ccnd3)
hc.single <- hclust(dist(clust.data1a, method = "euclidean"), method = "single")
hc.ward <- hclust(dist(clust.data1a, method = "euclidean"), method = "ward.D2")
plot(hc.single, labels=gol.fac)
```



```
plot(hc.ward, labels=gol.fac)
```




```
table(cutree(hc.single, k = 2), gol.fac)
```

```
##    gol.fac
##     ALL AML
##   1  27  10
##   2   0   1
```

```
table(cutree(hc.ward, k = 2), gol.fac)
```

```
##    gol.fac
##     ALL AML
##   1  21   0
##   2   6  11
```

```
cat("From the plots, we can tell that ward linkage is better.")
```

```
## From the plots, we can tell that ward linkage is better.
```

(b)

```
cl.2means1b <- kmeans(clust.data1a, centers = 2, nstart = 10)
table(cl.2means1b$cluster, labels=gol.fac)
```

```
##    labels
##     ALL AML
##   1  22   1
##   2   5  10
```

(c)

From the table we can tell the outcome are the same, both of them have been clustered into the right catagories. Therefore, kmeans and hierarchical are the same.

(d)

```
initial <- cl.2means1b$centers
n <- dim(clust.data1a)[1]
nboot <- 2000
boot.cl <- matrix(NA, nrow = nboot, ncol = 4) # column 4 to store CI for each mean, here is 2 means
for(i in 1:nboot){
  dat.star <- clust.data1a[sample(1:n,replace=TRUE),]
  cl <- kmeans(dat.star, initial, nstart = 10)
  boot.cl[i,] <- c(cl$centers[1,], cl$centers[2,])
}
apply(boot.cl,2,mean)
```

```
## [1] 2.0271770 0.6890578 2.0271770 0.6890578
```

```
quantile(boot.cl[,1],c(0.025,0.975))
```

```
##     2.5%    97.5%
## 1.822141 2.197109
```

```
quantile(boot.cl[,2],c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.1666047 1.0524180
```

```
quantile(boot.cl[,3],c(0.025,0.975))
```

```
##     2.5%    97.5%
## 1.822141 2.197109
```

```
quantile(boot.cl[,4],c(0.025,0.975))
```

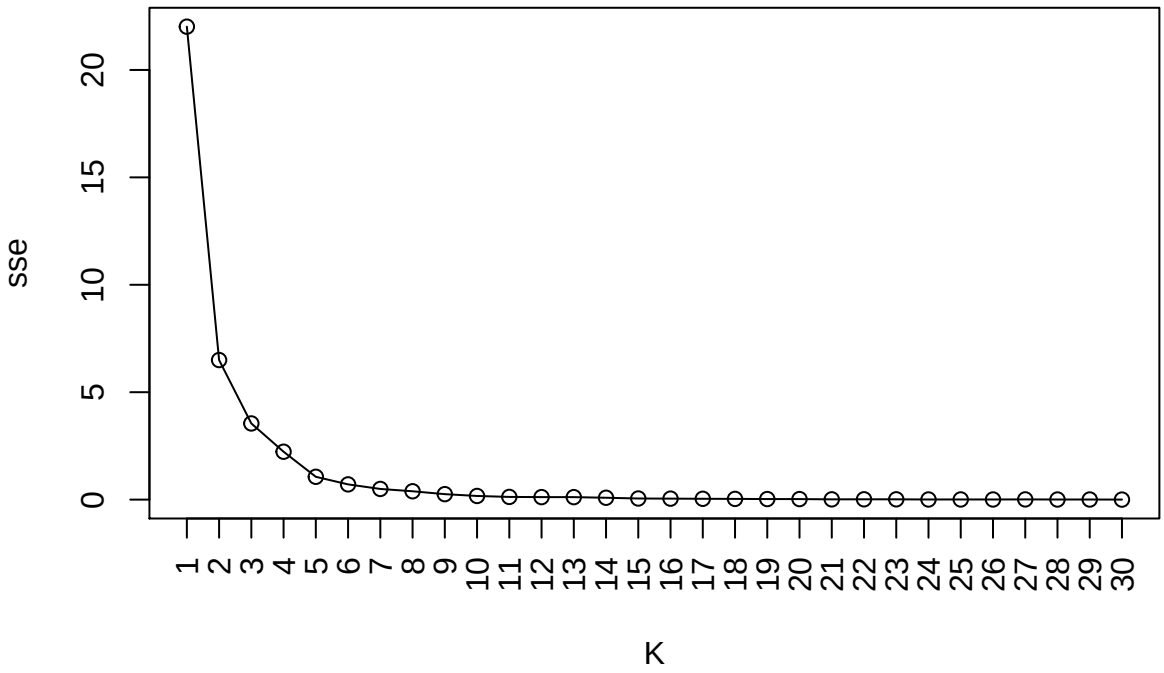
```
##      2.5%     97.5%
## 0.1666047 1.0524180
```

```
cat("There is no overlap.", quantile(boot.cl[,1],c(0.025,0.975)), "is more accurate.")
```

```
## There is no overlap. 1.822141 2.197109 is more accurate.
```

(e)

```
K <- c(1:30)
sse <- rep(NA,length(K))
for(k in K){
  sse[k] <- kmeans(clust.data1a, centers = k, nstart = 10)$tot.withinss
}
plot(K, sse, type = 'o', xaxt = 'n')
axis(1, at = K, las = 2)
```




```
cat("From the plot, it suggests 4 clusters.")
```

```
## From the plot, it suggests 4 clusters.
```

Question 2

(a)

```
data2 <- golub[c(grep("oncogene",golub.gnames[,2]),grep("antigen",golub.gnames[,2])),]
```

(b)

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 3.2.5
```

```
cl.2meansData2 <- kmeans(data2, centers = 2, nstart = 10)
cl.2mediodsData2 <- pam(dist(data2, method = "euclidean"), k = 2)
table(cl.2meansData2$cluster)
```

```
##
##  1  2
## 54 63
```

```
table(cl.2mediodsData2$clustering)
```

```
##
##  1  2
## 78 39
```

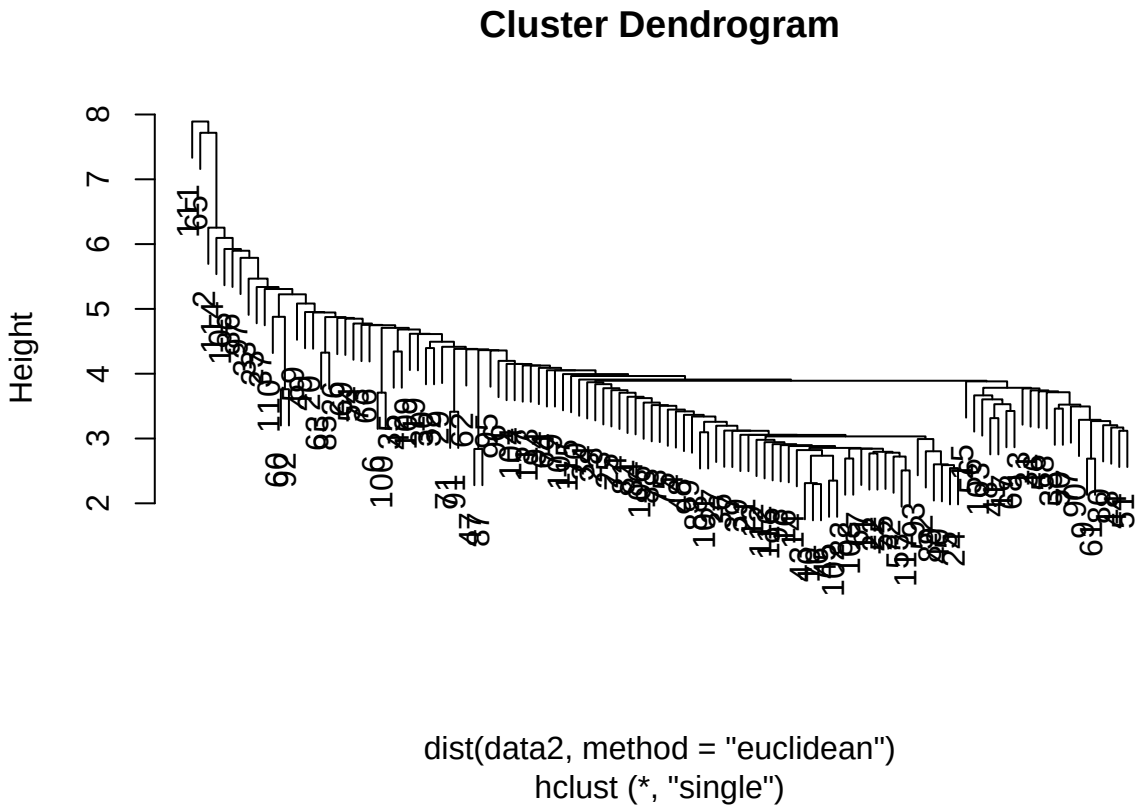
(c)

```
oncogenes <- golub[grep("oncogene",golub.gnames[,2]),]
antigens <- golub[grep("antigen",golub.gnames[,2]),]
# t.test(oncogenes, antigens)
```

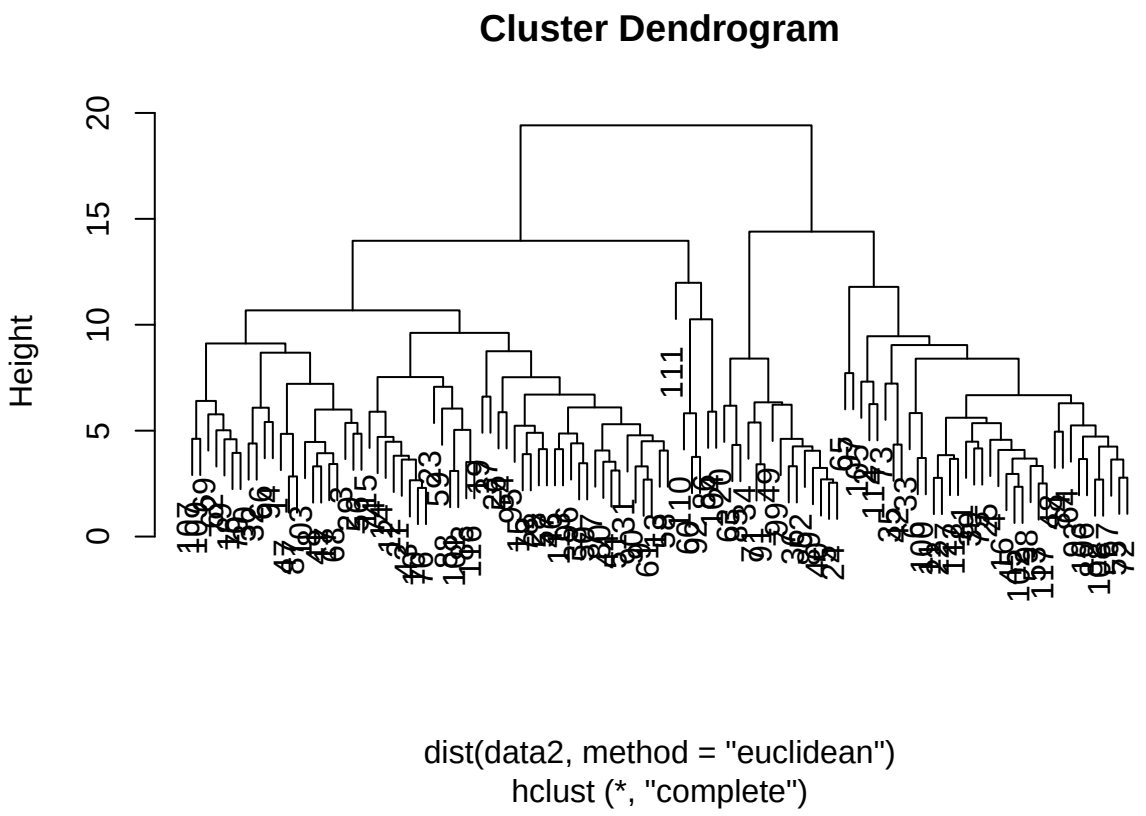
The 2 medoids method provides the more accurate clusters from the actual data.

(d)

```
hc.single2 <- hclust(dist(data2, method = "euclidean"), method = "single")
hc.ward2 <- hclust(dist(data2, method = "euclidean"), method = "complete")
plot(hc.single2)
```



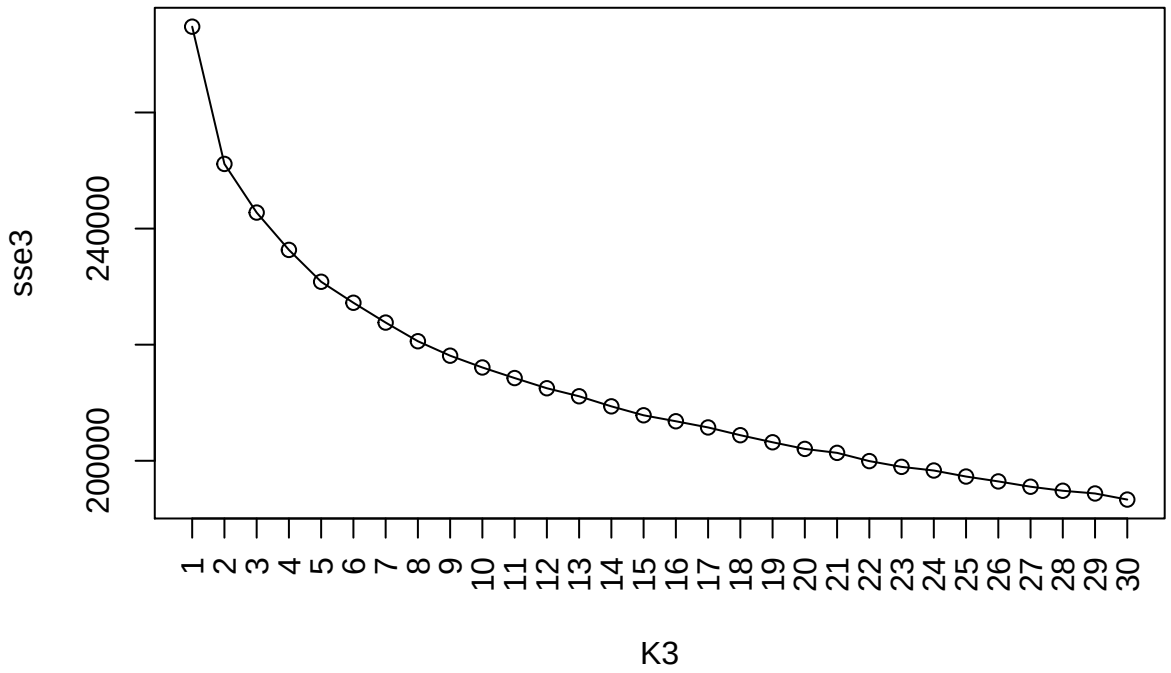

```
plot(hc.ward2)
```



Question 3

(a)

```
library(ISLR)
ncidata <- NCI60$data
ncilabs <- NCI60$labs
K3 <- c(1:30)
sse3 <- rep(NA,length(K3))
for(k in K3){
  sse3[k] <- kmeans(t(ncidata), centers = k, nstart = 10)$tot.withinss
}
plot(K3, sse3, type = 'o', xaxt = 'n')
axis(1, at = K3, las = 2)
```




```
cat("It needs 1 cluster.")
```

```
## It needs 1 cluster.
```

(b)

```
cl.7medoids <- pam(as.dist(1-cor(t(ncidata))), k = 7)
table(ncilabs,cl.7medoids$clustering)
```

```
##
## ncilabs       1 2 3 4 5 6 7
##   BREAST      0 3 0 0 2 0 2
##   CNS         1 4 0 0 0 0 0
##   COLON       0 0 0 7 0 0 0
##   K562A-repro 0 0 0 0 0 1 0
##   K562B-repro 0 0 0 0 0 1 0
##   LEUKEMIA    0 0 0 0 0 6 0
##   MCF7A-repro 0 0 0 0 1 0 0
##   MCF7D-repro 0 0 0 0 1 0 0
##   MELANOMA    0 1 0 0 0 0 7
##   NSCLC       2 2 0 3 1 1 0
##   OVARIAN     2 0 1 2 1 0 0
##   PROSTATE    0 0 1 1 0 0 0
##   RENAL       7 1 1 0 0 0 0
##   UNKNOWN     0 0 1 0 0 0 0
```

From the table, we can tell that COLON and LEUKEMIA are well identified. NSCLC is the most similar one to ovarian.