Problem 1: (40 points)

Clustering analysis on the "CCND3 Cyclin D3" gene expression values of the Golub et al. (1999) data.

(a) Conduct hierarchical clustering using single linkage and Ward linkage. Plot the cluster dendrogram for both fit. Get two clusters from each of the methods. Use function table() to compare the clusters with the two patient groups ALL/AML. Which linkage function seems to work better here?

(b) Use *k*-means cluster analysis to get two clusters. Use table() to compare the two clusters with the two patient groups ALL/AML.

(c) Which clustering approach (hierarchical versus k-means) produce the best matches to the two diagnose groups ALL/AML?

(d) Find the two cluster means from the k-means cluster analysis. Perform a bootstrap on the cluster means. Do the confidence intervals for the cluster means overlap? Which of these two cluster means is estimated more accurately?

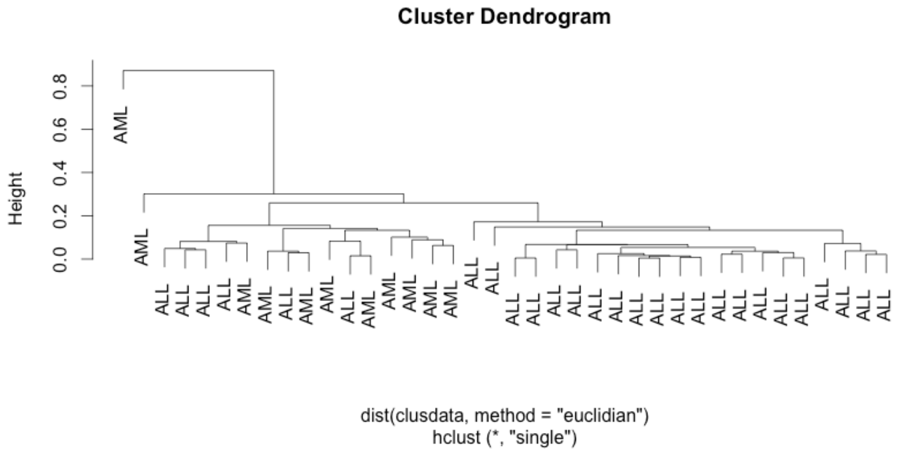
(e) Produce a plot of K versus SSE, for K=1, …, 30. How many clusters does this plot suggest?

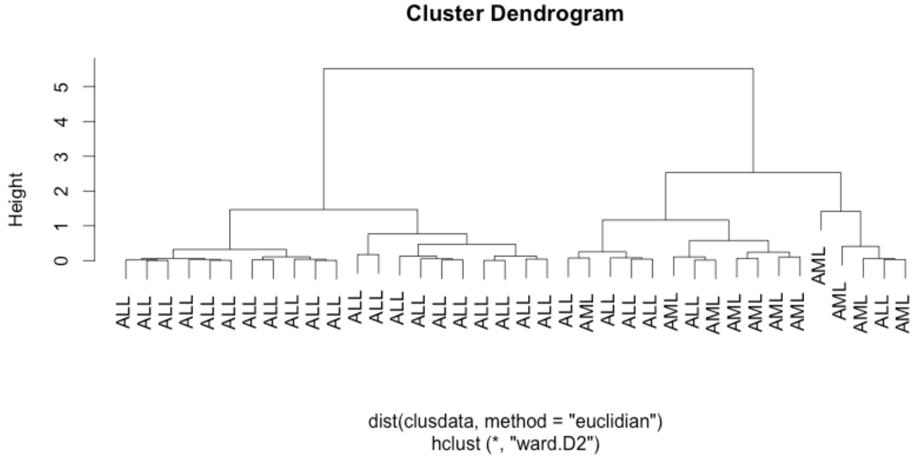
A)

Rscript:

```
# Answer 1
data(golub, package="multtest")
grep("CCND3 Cyclin D3",golub.gnames[,2])
clusdata <- data.frame(golub[1042,])
gol.fac <- factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
# Answer 1a
hcALL.sing<-hclust(dist(clusdata,method="euclidian"),method="single")
hcALL.ward<-hclust(dist(clusdata,method="euclidian"),method="ward.D2")
par(mfrow=c(1,2))
plot(hcALL.sing, labels=gol.fac)
plot(hcALL.ward, labels=gol.fac)
cALL.2sing<- cutree(hcALL.sing, k=2)
cALL.2ward<- cutree(hcALL.ward, k=2)
table(gol.fac,cALL.2sing)
table(gol.fac,cALL.2ward)
```

Answer:





```
> table(gol.fac,cALL.2sing)
       cALL.2sing
gol.fac  1  2
    ALL 27  0
    AML 10  1
> table(gol.fac,cALL.2ward)
       cALL.2ward
gol.fac  1  2
    ALL 21  6
    AML  0 11
```

Using the table function we can see and Ward linkage method is better than single linkage method as in single linkage all the clusters are in one group i.e. there are 10 wrong groups where as in Ward linkage all the AML is correctly placed and only 6 ALL are in incorrect clusters.

B)
Rscript:

```
clusters.km <- kmeans(clusdata, centers=2) #Do K-means with K=3 clusters
table(gol.fac, clusters.km$cluster)
```

Answer:

```
gol.fac  1  2
    ALL  5 22
    AML 10  1
```

C)
Answer:
Comparing hierarchical clustering versus k-means clustering, k-means seems to be better than hierarchical clustering as it gives just 6 incorrect clusters. But for the given data kmeans clustering and ward linkage seems to be working in a same way.

D)

```
cl.2mean <- kmeans(clusdata, centers=2, nstart = 10)
cl.2mean$centers
initial <-cl.2mean$centers
n <- dim(clusdata)[1]; nboot <-1000
boot.cl <- matrix(NA,nrow=nboot,ncol=2)
for (i in 1:nboot){
  dat.star <- clusdata[sample(1:n,replace=TRUE),]
  cl <- kmeans(dat.star, centers=initial)
  boot.cl[i,] <- c(cl$centers[,1])
}
apply(boot.cl,2,mean)
quantile(boot.cl[,1],c(0.025,0.975))
quantile(boot.cl[,2],c(0.025,0.975))
```

Answer:

```
Cluster means:
  golub.1042...
1     0.738366
2     2.045689


> apply(boot.cl,2,mean)
[1] 0.6887524 2.0319940
> quantile(boot.cl[,1],c(0.025,0.975))
     2.5%     97.5%
0.1850939 1.0711327
```

```
   2.5%    97.5% 
1.847179 2.197224
```

Two cluster means from the k-means cluster analysis is 0.738366 and 2.045689.
Two cluster means after bootstrapping is 0.6887524 and 2.0319940

Confidence interval for cluster mean overlap is:
bootstrap 95% CI for first coordinate of the cluster mean

```
2.5%     97.5% 
0.1850939 1.0711327
```

bootstrap 95% CI for second coordinate of the cluster mean

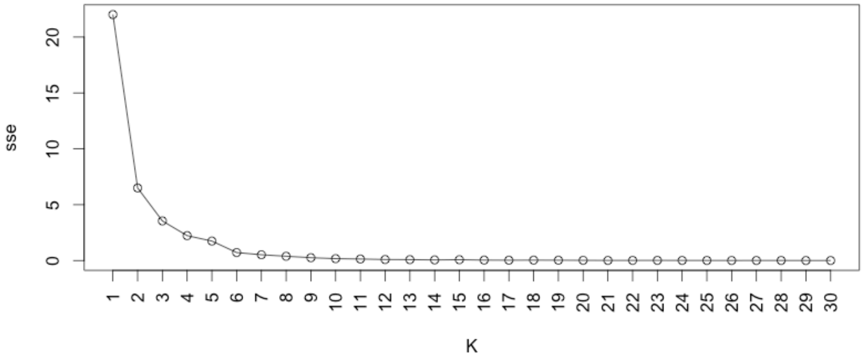
```
   2.5%    97.5% 
1.847179 2.197224
```

The confidence intervals for the cluster means doesn't overlap. Second cluster seems to be more accurate as 2.032 and 2.046.

E)
Rscript:

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

Answer:



As there is a drop from 1 to 2 and 2 to 3 there can be 2 or 3 clusters.

Problem 2 (30 points):
Cluster analysis on part of Golub data.
(a) Select the oncogenes and antigens from the Golub data. (Hint: Use grep()).

(b) On the selected data, do clustering analysis for the genes (not for the patients). Using K-means and K-medoids with K=2 to cluster the genes. Use table() to compare the resulting two clusters with the two gene groups oncogenes and antigens for each of the two clustering analysis.

(c) Use appropriate tests (from previous modules) to test the marginal independence in the two by two tables in (b). Which clustering method provides clusters related to the two gene groups?

(d) Plot the cluster dendrograms for this part of golub data with single linkage and complete linkage, using Euclidean distance.

A)
Rscript:

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

Answer:

```
> oncogene
 [1]  501  502  503  587  758  766  775  805  817  819  938 1067 1090 1111
[15] 1211 1268 1542 1596 1615 1735 1747 1750 1788 1818 1820 1837 1839 2004
[29] 2291 2302 2488 2517 2661 2681 2692 2703 2714 2715 2892 2981 2990 2993
> antigen
 [1]  166  313  388  497  504  514  527  540  548  614  646  664  685  763
[15]  808  826  832  833  834  872  885  890  892  893  926  936  947 1008
[29] 1010 1075 1087 1208 1258 1279 1287 1412 1422 1467 1531 1616 1645 1719
[43] 1748 1752 1756 1760 1781 1789 1798 1806 1808 1827 1852 1863 1882 1893
[57] 1908 1911 1964 2007 2170 2171 2231 2371 2546 2581 2613 2653 2672 2749
[71] 2761 2855 2989 3026 3047
```

B)
Rscript:

```
library(cluster)
combined<-c(oncogene,antigen)
golub.data<-data.frame(golub[combined,])
data.fac<-factor(c(rep("oncogene",length(oncogene)),rep("antigen",length(antigen))))
clusters <- kmeans(golub.data, centers=2)
table(data.fac,clusters$cluster)
cluster.kmed <- pam(golub.data, k=2)
table(data.fac,cluster.kmed$cluster)
```

Answer:

```
data.fac    1  2
  antigen  41 34
  oncogene 22 20

data.fac    1  2
  antigen  49 26
  oncogene 29 13
```

C)

Null hypothesis: Two by two tables are marginally independent

Rscript:

```
chisq.test(table1)
chisq.test(table2)
```

Answer:

```
Pearson's Chi-squared test with Yates' continuity correction

data:  table1
X-squared = 0.002, df = 1, p-value = 0.9644

Pearson's Chi-squared test with Yates' continuity correction

data:  table2
X-squared = 0.0418, df = 1, p-value = 0.838
```

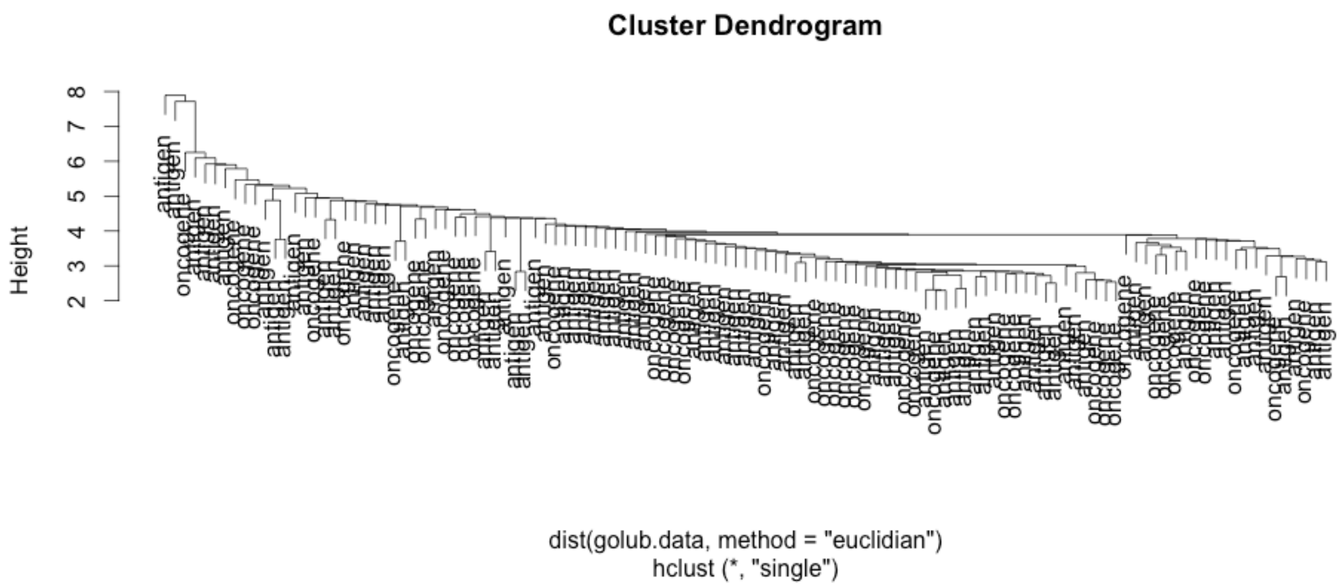
Here p values are greater than 0.05 so we fail to reject the null hypothesis of marginal independence hence we conclude that both the clustering method doesn’t provides clusters related to the two gene groups

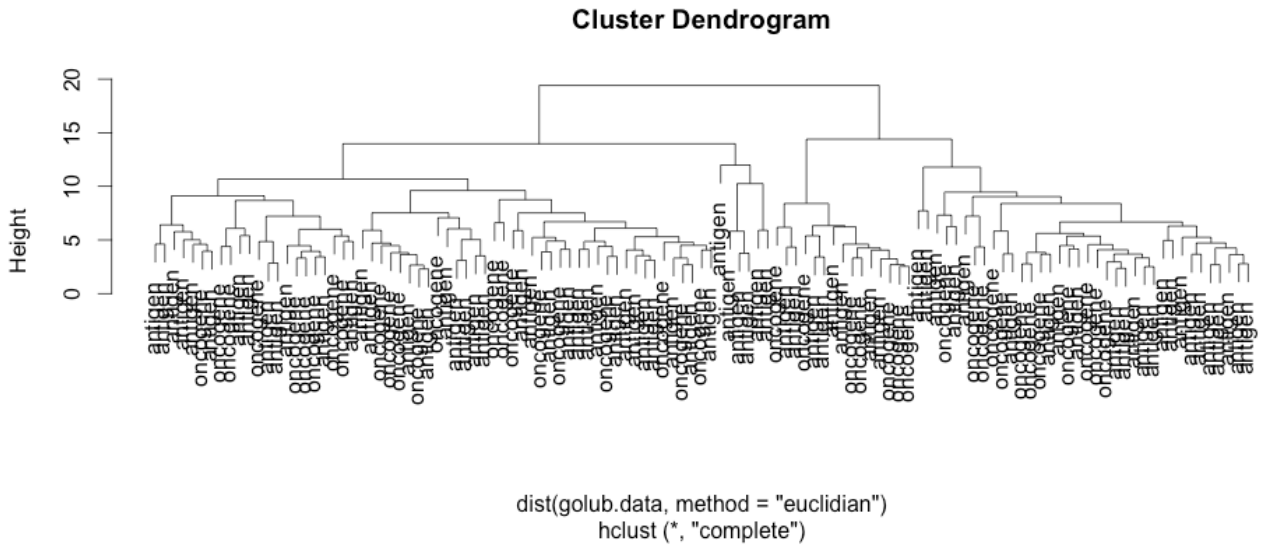
D)

Rscript:

```
hc.sing<-hclust(dist(golub.data,method="euclidian"),method="single")
plot(hc.sing,labels=data.fac)
hc.comp<-hclust(dist(golub.data,method="euclidian"),method="complete")
plot(hc.comp,labels=data.fac)
```

Answer:


Cluster Dendrogram
Height
2
3
4
5
6
7
8
antigen
oncogene
dist(golub.data, method = "euclidian")
hclust (*, "single")


Cluster Dendrogram
Height
0
5
10
15
20
antigen
oncogene
dist(golub.data, method = "euclidian")
hclust (*, "complete")

Problem 3 (30 points):

Clustering analysis on NCI60 cancer cell line microarray data (Ross et al. 2000)

We use the data set in package ISLR from r-project (Not Bioconductor). You can use the following commands to load the data set.

```
install.packages('ISLR')
library(ISLR)
ncidata<-NCI60$data
ncilabs<-NCI60$labs
```

The ncidata (64 by 6830 matrix) contains 6830 gene expression measurements on 64 cancer cell lines. The cancer cell lines labels are contained in ncilabs. We do clustering analysis on the 64 cell lines (the rows).

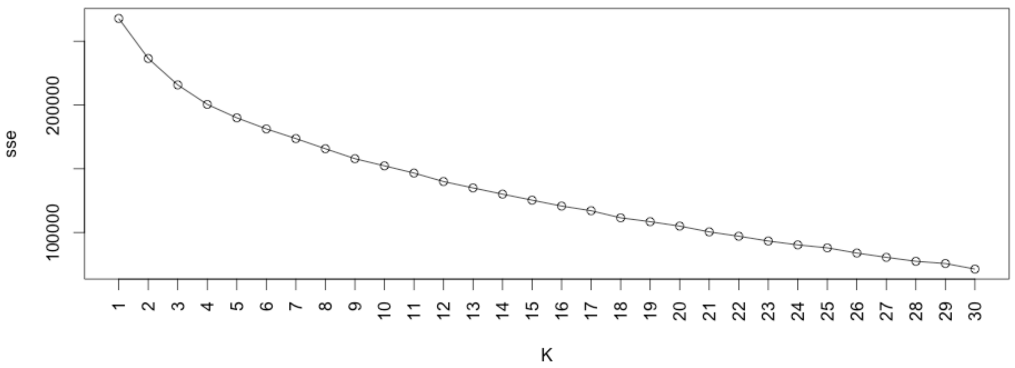
(a) Using k-means clustering, produce a plot of K versus SSE, for K=1,…, 30. How many clusters appears to be there?

(b) Do K-medoids clustering (K=7) with 1-correlation as the dissimilarity measure on the data. Compare the clusters with the cell lines. Which type of cancer is well identified in a cluster? Which type of cancer is not grouped into a cluster? According to the clustering results, which types of cancer are most similar to ovarian cancer?

Rscript:

```
install.packages('ISLR')
library(ISLR)
ncidata<-NCI60$data
ncilabs<-NCI60$labs
K<-(1:30); sse<-rep(NA,length(K))
for (k in K) {
  sse[k]<-kmeans(ncidata, centers=k,nstart = 10)$tot.withinss
}
plot(K, sse, type='o', xaxt='n')
axis(1, at = K, las=2)
```

Answer:



Sharp decline seems to stop at around 7-8 suggesting there could be 7-8 clusters.

B)

Rscript:

```
clusters.7medoid <- pam(as.dist(1-cor(t(ncidata))),k=7)
table(factor(ncilabs), clusters.7medoid$cluster)
table(NCI60$labs)
```

Answer:

```
           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
```

```
    BREAST         CNS       COLON K562A-repro K562B-repro
         7           5           7           1           1
  LEUKEMIA MCF7A-repro MCF7D-repro    MELANOMA       NSCLC
         6           1           1           8           9
   OVARIAN    PROSTATE       RENAL     UNKNOWN
         6           2           9           1
```

- Which type of cancer is well identified in a cluster?
  According to the clusters of NCI60$labs colon cancer and leukemia seems to be well identified in cluster 4 and cluster 7 respectively.

- Which type of cancer is not grouped into a cluster?
  All the cancer is grouped into a cluster apart from 1 which is unknown.

- which types of cancer are most similar to ovarian cancer?
  NSCLC and prostate cancer are most similar to ovarian cancer.