Problem 1 (60 points) Analysis of the ALL data set

(a) Define an indicator variable ALL.fac such that ALL.fac=1 for T-cell patients and ALL.fac=2 for B-cell patients.

(b) Plot the histograms for the first three genes' expression values in one row.

(c) Plot the pairwise scatterplots for the first five genes.

(d) Do a 3D scatterplot for the genes "39317_at", "32649_at" and "481_at", and color according to ALL.fac (give different colors for B-cell versus T-cell patients). Can the two patient groups be distinguished using these three genes?

(e) Do K-means clustering for K=2 and K=3 using the three genes in (d). Compare the resulting clusters with the two patient groups. Are the two groups discovered by the clustering analysis?

(f) Carry out the PCA on the ALL data set with scaled variables. What proportion of variance is explained by the first principal component? By the second principal component?

(g) Do a biplot of the first two principal components. Observe the pattern for the loadings. What info is the first principal component summarizing?

(h) For the second principal component PC2, print out the three genes with biggest loadings and the three genes with smallest loadings.

(i) Find the gene names and chromosomes for the gene with biggest PC2 value and the gene with smallest PC2 value. (Hint: review Module 10 on searching the annotation.)

A)

Rscript:

```
library("ALL"); data(ALL, package = "ALL");
ALL.fac <- factor(ALL$BT %in% c("B","B1","B2","B3","B4"), labels=c("1","2"))
ALL.fac
```

Answer:

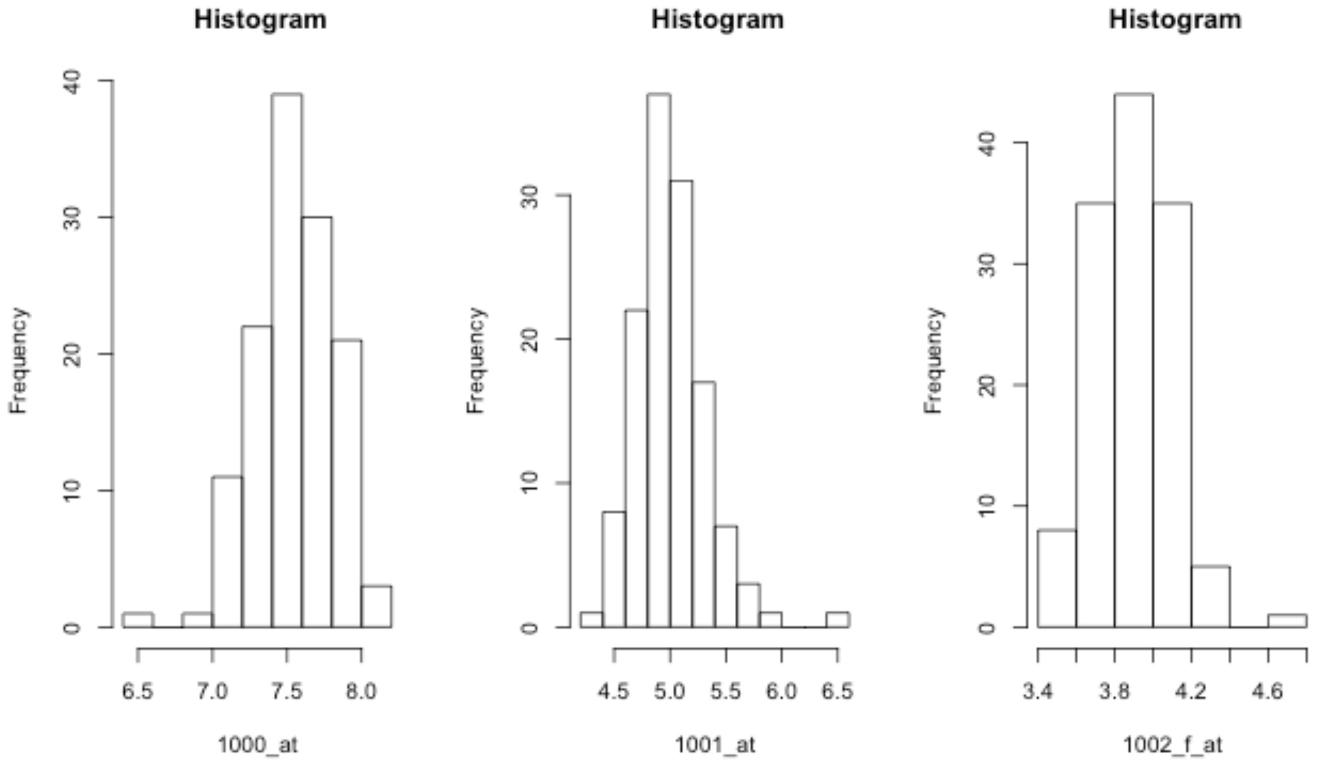
```
[1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [51] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1
[101] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Levels: 1 2
```

B)

Rscript:

```
ALLdata<-exprs(ALL)
gene<-ALLdata[c(1:3),]
varnames<- rownames(gene)[1:3]
par(mfrow=c(1,3))
for(i in 1:3){
  hist(gene[i,], xlab=varnames[i], main=" Histogram")
}
```

Answer:

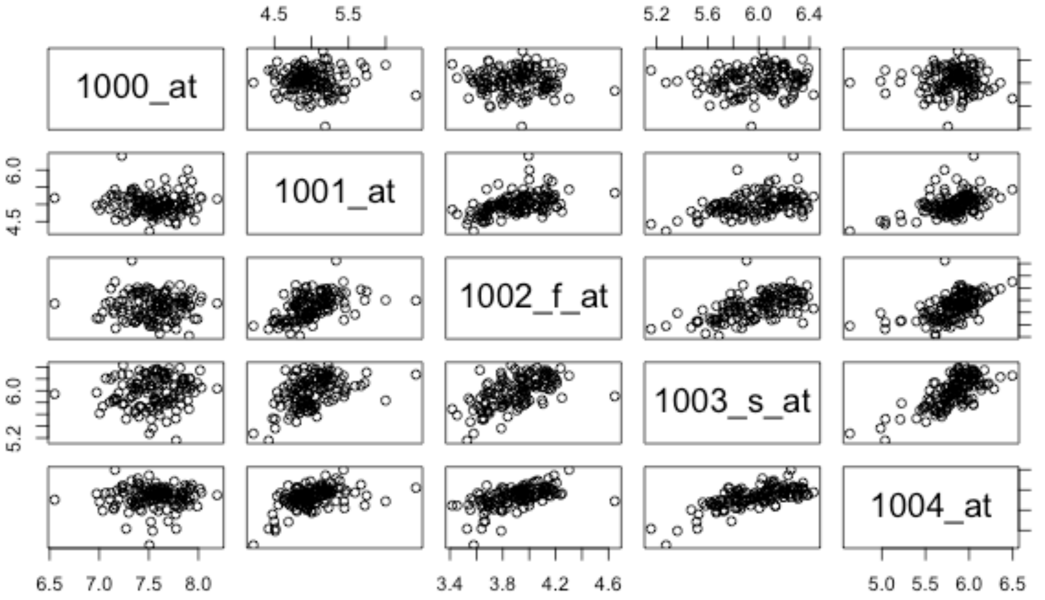


C)

Rscript:

```
ALLdata<-exprs(ALL)
genenew<-t(ALLdata[c(1:5),])
pairs(genenew)
```

Answer:

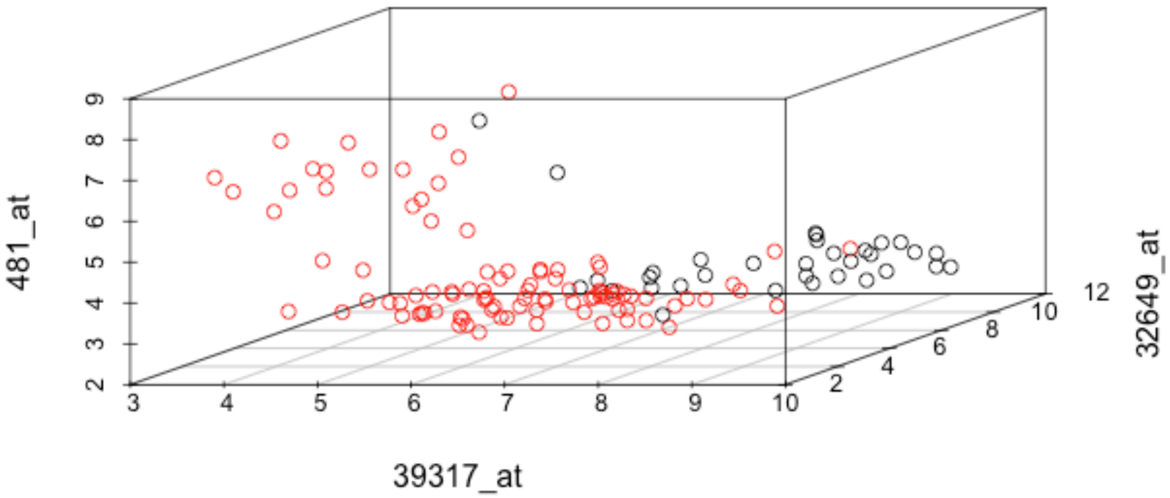


D)

Rscript:

```
require(scatterplot3d)
par(mfrow=c(1,1))
scatterplot3d(t(exprs(ALL[c("39317_at","32649_at","481_at"),])),color=ALL.fac)
```

Answer:



es, two patient groups be distinguished using these three genes

E)

Rscript:

```
ALL.fac <- factor(ALL.fac, levels=c(1,2), labels=c("T cell","B cell"))
clusdata1<-kmeans(t(exprs(ALL[c("39317_at","32649_at","481_at"),])),centers=2)
table(ALL.fac,clusdata1$cluster)
clusdata2<-kmeans(t(exprs(ALL[c("39317_at","32649_at","481_at"),])),centers=3)
table(ALL.fac,clusdata2$cluster)
```

Answer:

```
ALL.fac   1  2
  T cell  4 29
  B cell 85 10
```

```
ALL.fac   1  2  3
  T cell  2  3 28
  B cell 20 70  5
```

ALL data has 95 B cell and 33 Tcell patients. Clusters are not perfectly divided but they are approximately divided with 14 false positives.

When k =3 it gives me total of 33 t cell and 95 Bcell with three different clusters.

F)

Rscript:

```
pr.ALL <- prcomp(exprs(ALL), scale=TRUE)
summary(pr.ALL)
```

Answer:

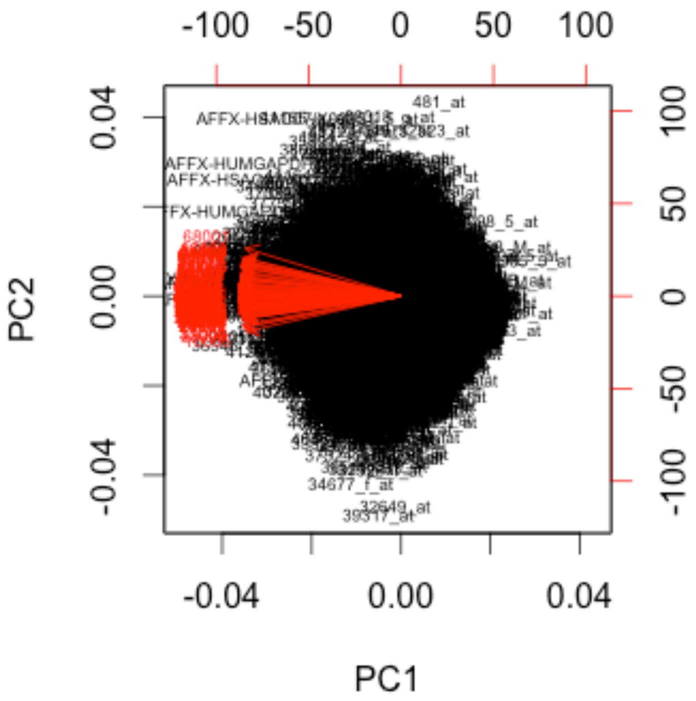
Proportion of variance explained by the first principal component is 94% (.9359)
and proportion of variance explained by the second principal component is .95%(0.00948)

G)

Rscript:

```
biplot(pr.ALL,cex=0.5)
```

Answer:



We see the red arrows have about the same horizontal lengths. This reflects the fact that PC1 is essentially the average of the patients

H)

Rscript:

```
o <- order(pr.ALL$x[,2],decreasing=T)
ALLdata[o[1:3],0]
dim(ALLdata)
ALLdata[o[12623:12625],0]
```

Answer:

The three genes with biggest loadings:
1.481_at
2.38018_g_at
3.41165_g_at

The three genes with Smallest loadings:
3. 34677_f_at
2. 32649_at
1. 39317_at

I)
Rscript:

```
source("http://bioconductor.org/biocLite.R")
annotation(ALL)
biocLite("hgu95av2.db")
library("hgu95av2.db")
chr<- as.list(hgu95av2CHR)
gene<-as.list(hgu95av2GENENAME)
chr[o[1]]
gene[o[1]]
chr[o[12625]]
gene[o[12625]]
```

Answer:
The gene names and chromosomes for the gene with biggest PC2 value is
"SNF related kinase" and Chr 3
The gene names and chromosomes for the gene with smallest PC2 value is
"cytidine monophospho-N-acetylneuraminic acid hydroxylase, pseudogene" and Chr 6.

Problem 2 (40 points) Variables scaling and PCA in the iris data set
In this module and last module, we mentioned that the variables are often scaled before doing the PCA or the clustering analysis. By "scaling a variable", we mean to apply a linear transformation to center the observations to have mean zero and standard deviation one. In last module, we also mentioned using the correlation-based dissimilarity measure versus using the Euclidean distance in clustering analysis. It turns out that the correlation-based dissimilarity measure is proportional to the squared Euclidean distance on the scaled variables. We check this on the iris data set. And we compare the PCA on scaled versus unscaled variables for the iris data set.

(a) Create a data set consisting of the first four numerical variables in the iris data set (That is, to drop the last variable Species which is categorical). Then make a scaled data set that centers each of the four variables (columns) to have mean zero and variance one.

(b) Calculate the correlations between the columns of the data sets using the cor() function. Show that these correlations are the same for scaled and the unscaled data sets.

(c) Calculate the Euclidean distances between the columns of the scaled data set using dist() function. Show that the squares of these Euclidean distances are proportional to the (1-correlation)s. What is the value of the proportional factor here?

(d) Show the outputs for doing PCA on the scaled data set and on the unscaled data set. (Apply PCA on the two data sets with option "scale=FALSE". Do NOT use option "scale=TRUE", which will scale data no matter which data set you are using.) Are they the same?

(e) What proportions of variance are explained by the first two principle components in the scaled PCA and in the unscaled PCA?

(f) Find a 90% confidence interval on the proportion of variance explained by the second principal component.

A)

Rscript:

```
data<-iris[,c(1:4)]
scaledData<-scale(data, center=T, scale=T)
```

B)

Rscript:

```
scaledCor<-cor(scaledData)
scaledCor
cor(data)
```

Answer:

```
> scaledCor
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
> cor(data)
```

```
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Correlation of scaled and unscaled data is same.

C)

Rscript:

```
scaleddata1<-rbind(scaledData[,1],scaledData[,2],scaledData[,3],scaledData[,4])
eucldist<-dist(scaleddata1,method="eucl")
eucldist^2
1-scaledCor
eucldist^2/sum(eucldist^2)
(1-scaledCor)/sum(1-scaledCor)
```

Answer:

```
eucldist^2
         1         2         3
2 333.03580
3  38.21737 425.67515
4  54.25354 407.10553  11.06610
> 1-scaledCor
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.0000000    1.117570   0.12824622  0.18205887
Sepal.Width     1.1175698    0.000000   1.42844010  1.36612593
Petal.Length    0.1282462    1.428440   0.00000000  0.03713457
Petal.Width     0.1820589    1.366126   0.03713457  0.00000000

> (1-scaledCor)/sum(1-scaledCor)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.00000000   0.1311832  0.015053874 0.021370542
Sepal.Width    0.13118323   0.0000000  0.167673998 0.160359399
Petal.Length   0.01505387   0.1676740  0.000000000 0.004358952
Petal.Width    0.02137054   0.1603594  0.004358952 0.000000000
> eucldist^2/sum(eucldist^2)
            1           2           3
2 0.262366470
3 0.030107748 0.335347996
4 0.042741084 0.320718799 0.008717904
```

The value of the proportional factor here seems to be 2.

D)

Rscript:

```
unscaledPCA<-prcomp(data,scale=F)
unscaledPCA
scaledPCA<-prcomp(scaledData,scale=F)
scaledPCA
```

```
Standard deviations:
[1] 2.0562689 0.4926162 0.2796596 0.1543862

Rotation:
                    PC1         PC2         PC3        PC4
Sepal.Length  0.36138659 -0.65658877  0.58202985  0.3154872
Sepal.Width  -0.08452251 -0.73016143 -0.59791083 -0.3197231
Petal.Length  0.85667061  0.17337266 -0.07623608 -0.4798390
Petal.Width   0.35828920  0.07548102 -0.54583143  0.7536574

> scaledPCA
Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:
                   PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

No PCA on scaled and unscaled data is not the same.

E)
Rscript:

```
summary(unscaledPCA)
summary(scaledPCA)
```

Answer:

```
> summary(unscaledPCA)
Importance of components:
                          PC1     PC2    PC3     PC4
Standard deviation     2.0563 0.49262 0.2797 0.15439
Proportion of Variance 0.9246 0.05307 0.0171 0.00521
Cumulative Proportion  0.9246 0.97769 0.9948 1.00000
> summary(scaledPCA)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

Proportions of variance are explained by the first two principle components in the unscaled PCA is 97.7%(.97769)

Proportions of variance are explained by the first two principle components in the scaled PCA is 95.81%(.9581)

F)
Rscript:

```
data2 <- scaledData;
```

```
nboot<-1000
sdevs <- array(dim=c(nboot,p))
pvar <- array(dim=c(nboot,p))
for (i in 1:nboot) {
  dat.star <- data2[sample(1:n,replace=TRUE),]
  sdevs[i,] <- prcomp(dat.star)$sdev
  pvar[i,]<- (sdevs[i,])^2/sum((sdevs[i,])^2)
}
quantile(pvar[,2], c(0.05,0.95))
```

Answer:
90% confidence interval on the proportion of variance explained by the second principal component is:

```
5%        95%
0.1873642 0.2655731
```