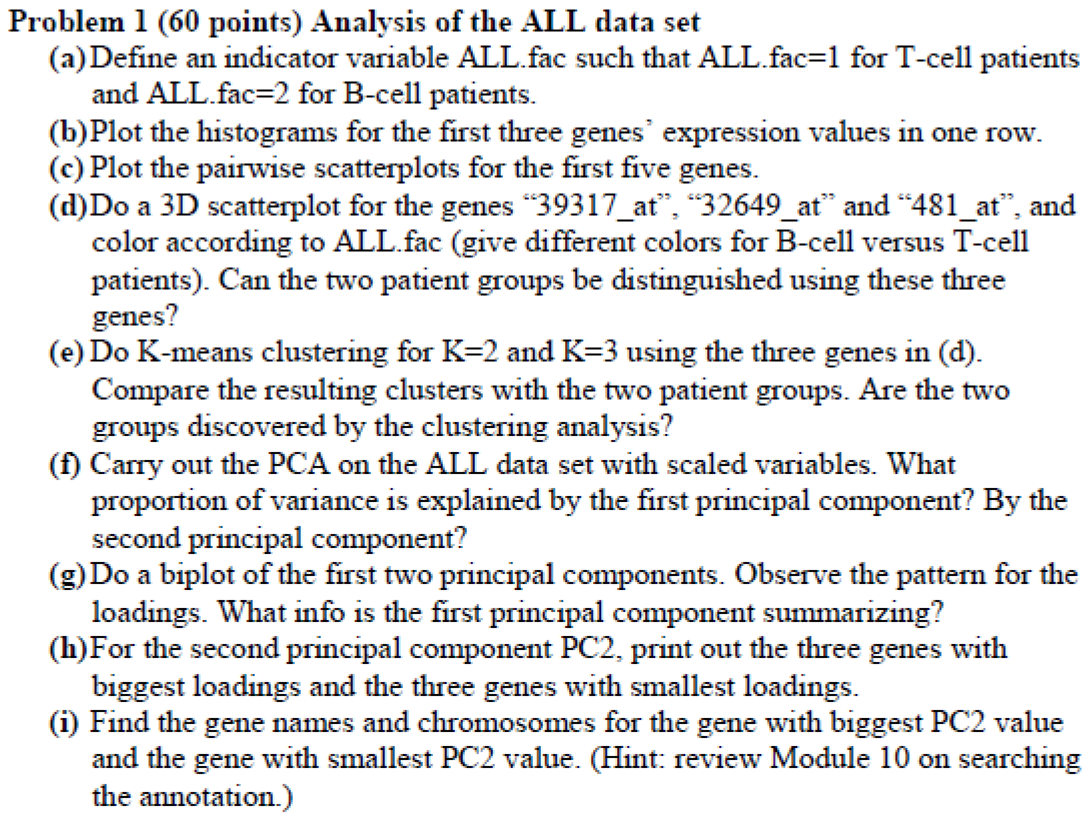
Solutions to homework module 12



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.)

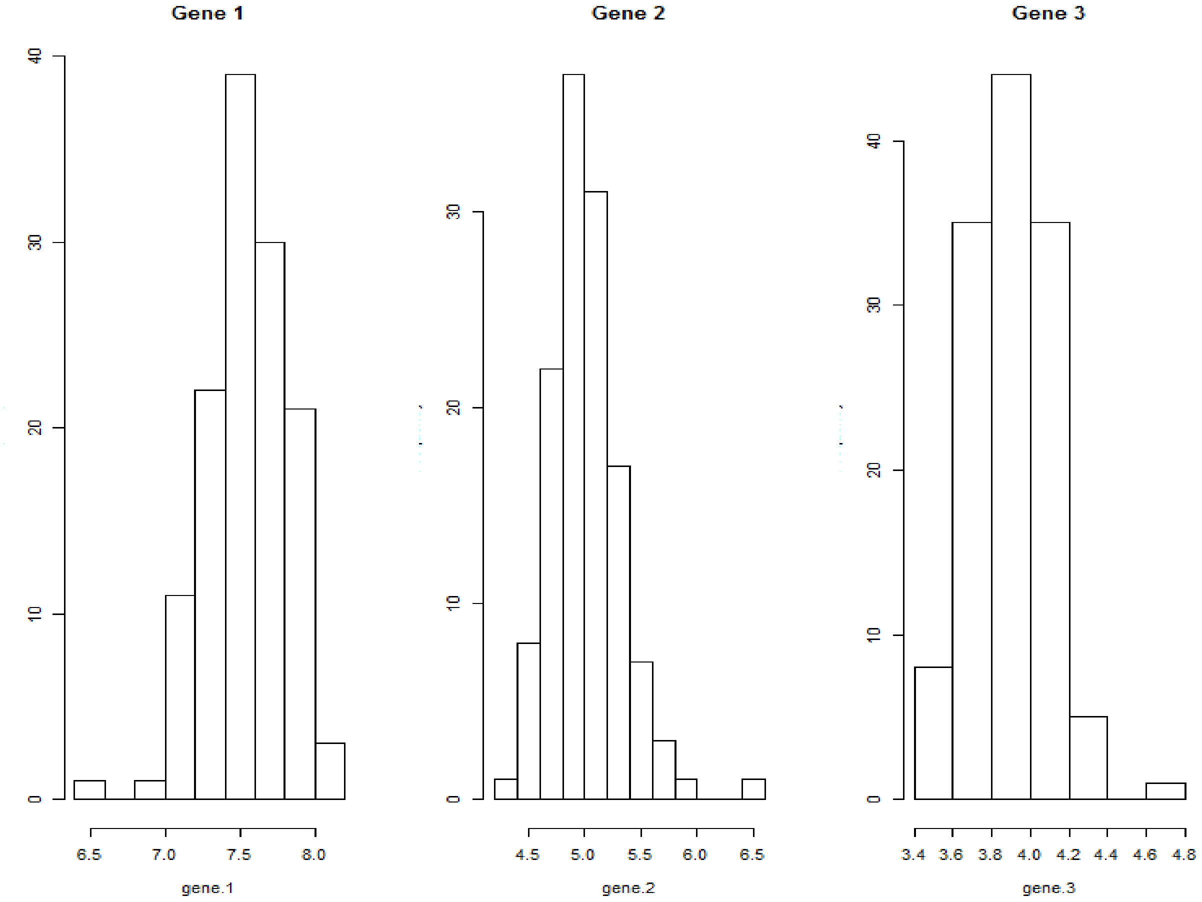
Solutions:

1a)

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

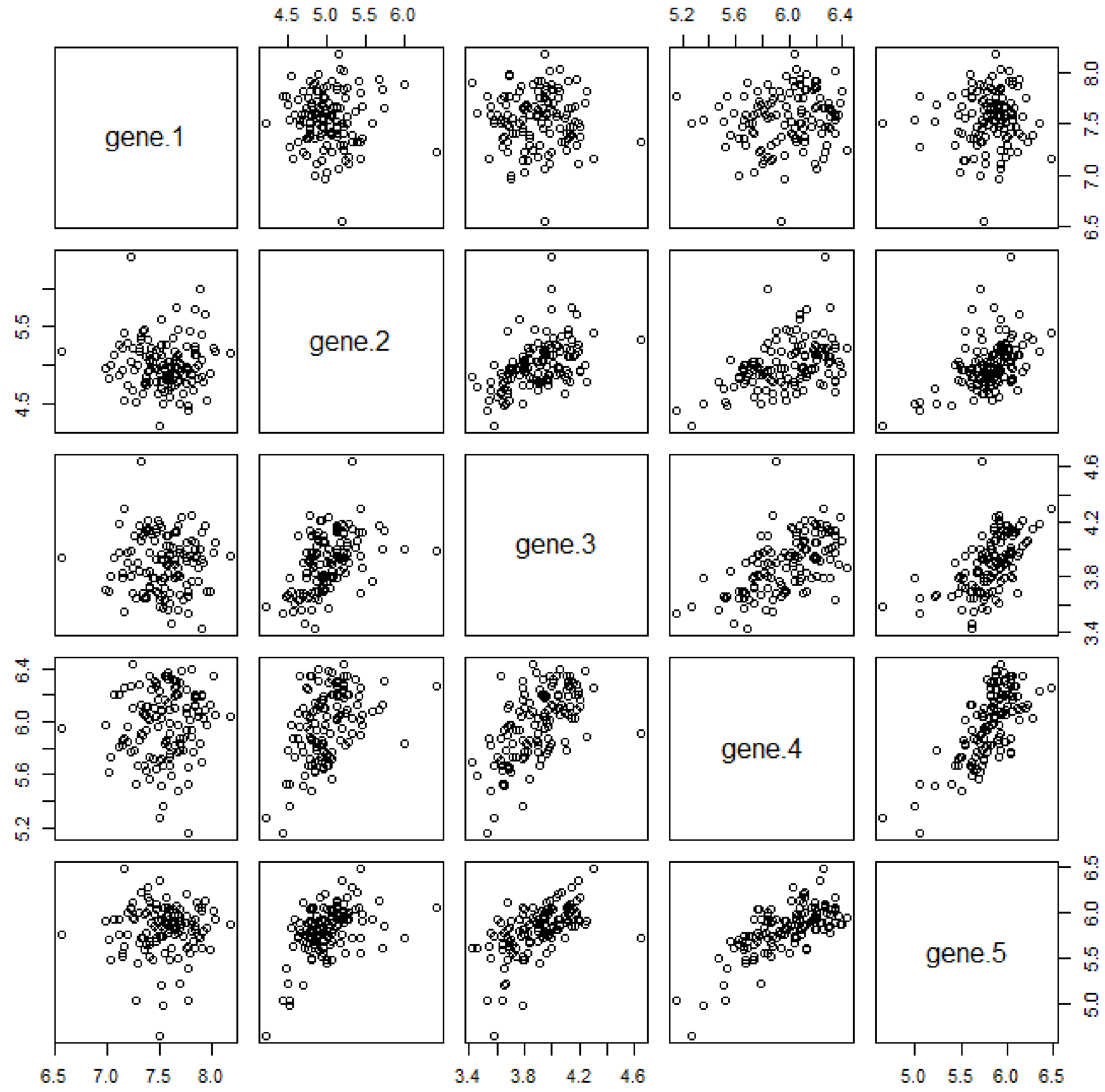
Solutions to homework module 12

1b)



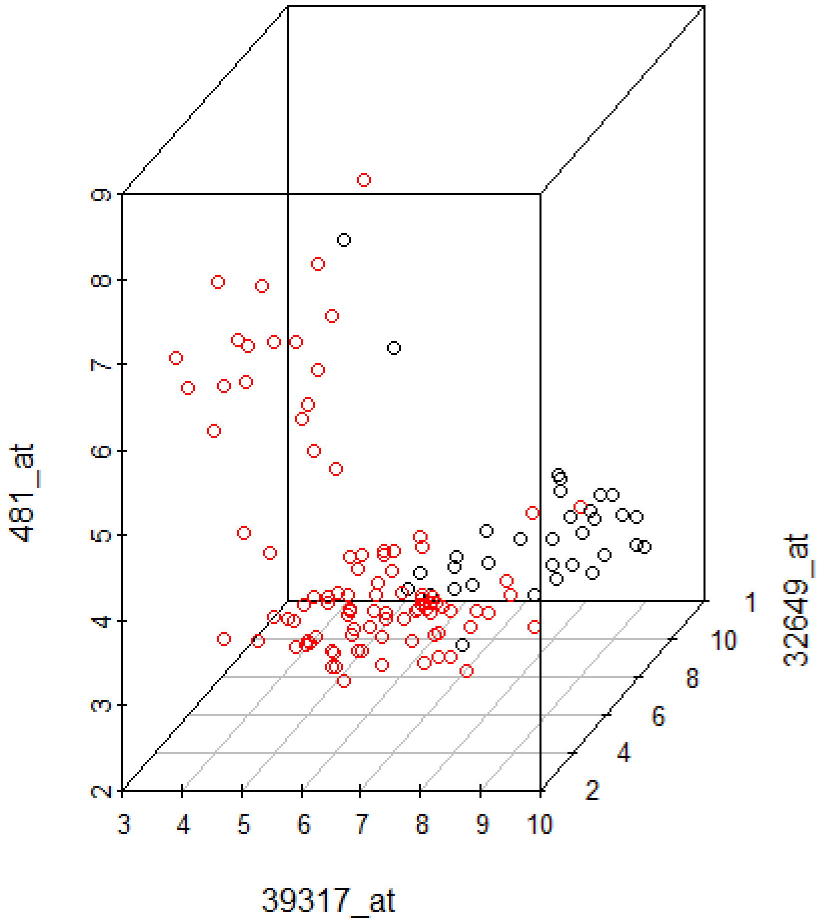
Solutions to homework module 12

1c)



Solutions to homework module 12

1d)



Yes, the two patient groups can be vaguely distinguished using these three genes

1e)

```
> table(ALL.fac,cluster1$cluster)

ALL.fac  1  2
      4 31  2
      3 21 74
> table(ALL.fac,cluster2$cluster)

ALL.fac  1  2  3
      4 28  2  3
      3  5 20 70
```

Solutions to homework module 12

- ALL has 95 B-cell, 33 T-cell
- I don't think the clusters have been exactly divided but they are approx. divided based on B/T-cell. When k=2, 2 T-cell and 74 B-cell have been clustered into cluster-1 and in cluster-2, there are 31 T-cell and 21 from B-cell
- When k=3, 2 T-cell and 20 B-cell have been clustered into cluster-1, In cluster-2, there are 28 T-cell and 5 from B-cell , in cluster-3, there are 3 T-cell and 70 from B-cell.

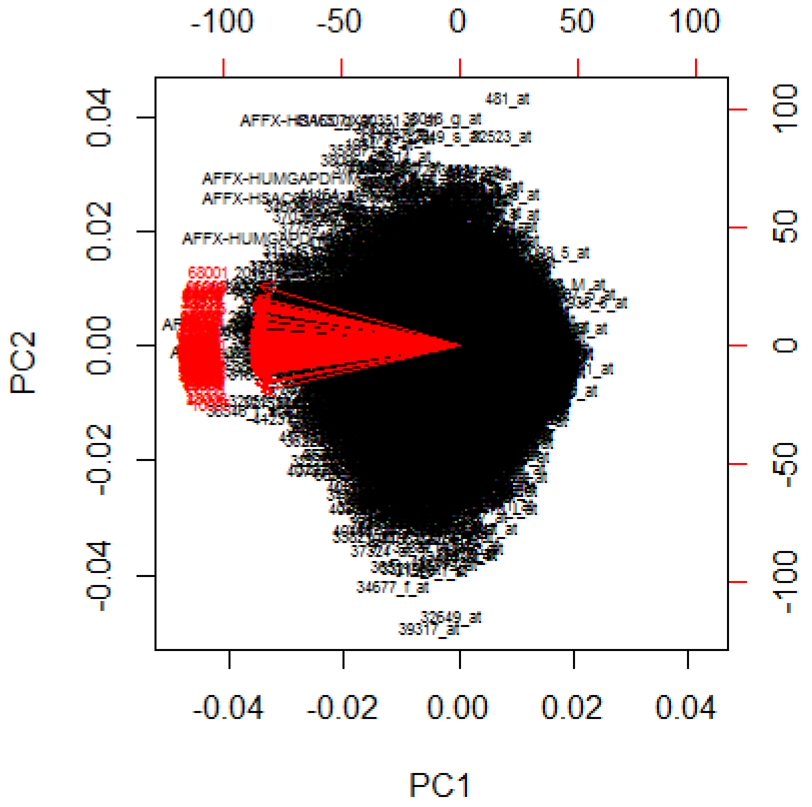
1f)

```
> P.ALL <- prcomp(data, scale=TRUE)
> summary(P.ALL)
Importance of components:
                           PC1     PC2     PC3     PC4     PC5     PC6     PC
7    PC8     PC9    PC10    PC11
Standard deviation     10.9450 1.10132 0.93237 0.75341 0.62938 0.57412 0.5319
7 0.5065 0.45455 0.44240 0.41633
Proportion of Variance  0.9359 0.00948 0.00679 0.00443 0.00309 0.00258 0.0022
1 0.0020 0.00161 0.00153 0.00135
Cumulative Proportion   0.9359 0.94536 0.95215 0.95658 0.95968 0.96225 0.9644
6 0.9665 0.96808 0.96961 0.97096
```

- Percentage of variance explained by first principle component is 93.6%
- Percentage of variance explained by second principle component = 0.95%

1g)

Solutions to homework module 12



The first component summarizing that PC1 is the average of the patient due to large avg expression values

1h) the genes with the biggest data loading are

```
> dimnames(data)[[1]][[gene.order[1]]]
[1] "481_at"
> dimnames(data)[[1]][[gene.order[2]]]
[1] "38018_g_at"
> dimnames(data)[[1]][[gene.order[3]]]
[1] "41165_g_at"
```

The gemnes with the smallest data loading are

```
> dimnames(data)[[1]][[gene.order[12623]]]
[1] "34677_f_at"
> dimnames(data)[[1]][[gene.order[12624]]]
[1] "32649_at"
> dimnames(data)[[1]][[gene.order[12625]]]
[1] "39317_at"
```

Solutions to homework module 12

1i)

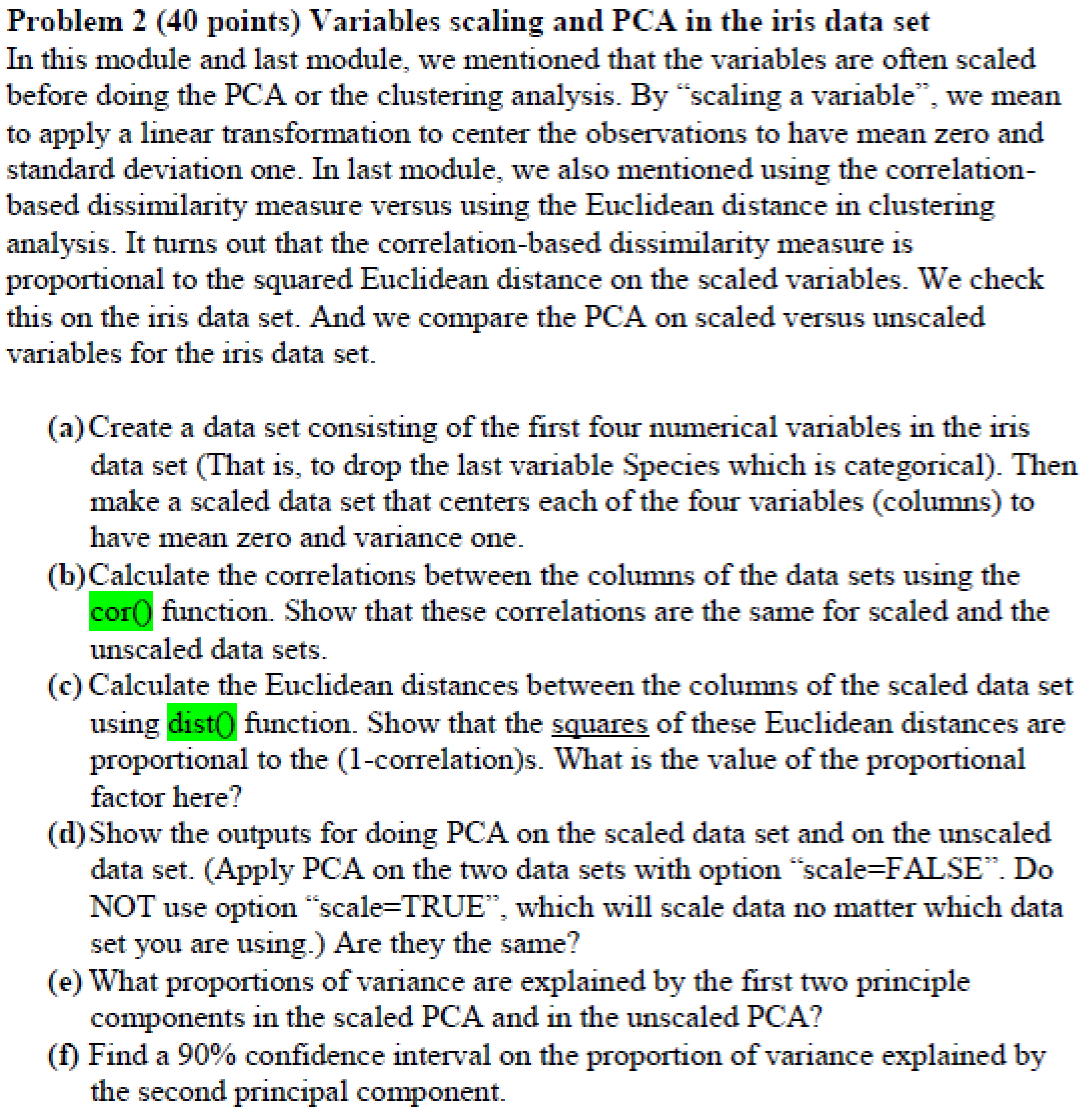
Gene name and chromosomes for gene with biggest PC2 value :

```
> genename
[1] "SNF related kinase"
> chromosomes
       3
```

Gene name and the chromosomes for gene with smallest PC2 value:

```
> genenamelow
[1] "cytidine monophospho-N-acetylneuraminic acid hydroxylase, pseudogene"
> chromosomeslow
        6
```

Solutions to homework module 12



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.

Solutions:

2a)

```
> iris.data <- iris[1:4]
> mean <- mean(iris.data[,1])
> sd <- sd(iris.data[,1])
> Sepal.Length <- NULL
> for (i in 1:150){Sepal.Length[i] <- (iris.data[i,1]-mean)/sd}
> mean <- mean(iris.data[,2])
> sd <- sd(iris.data[,2])
> Sepal.Width <- NULL
> for (i in 1:150){Sepal.Width[i] <- (iris.data[i,2]-mean)/sd}
> mean <- mean(iris.data[,3])
> sd <- sd(iris.data[,3])
> Petal.Length <- NULL
> for (i in 1:150){Petal.Length[i] <- (iris.data[i,3]-mean)/sd}
> mean <- mean(iris.data[,4])
```

Solutions to homework module 12

```
> sd <- sd(iris.data[,4])
> Petal.Width <- NULL
> for (i in 1:150){Petal.Width[i] <- (iris.data[i,4]-mean)/sd}
> scaled.data <- cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)
> xxx=data.frame(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)
```

2b)

```
> cor.scaled <- cor(scaled.data)
> cor.unscaled <- cor(iris.data)
> cor.scaled
             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.unscaled
             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
> all.equal(cor.scaled, cor.unscaled)
[1] TRUE
```

True indicates correlation

2c)

```
> euclidian.sqdist
             Sepal.Length Sepal.Width Petal.Length
Sepal.Width     333.03580
Petal.Length     38.21737   425.67515
Petal.Width      54.25354   407.10553     11.06610
> cor.data<-as.dist(1-cor(scaled.data))
> cor.data
             Sepal.Length Sepal.Width Petal.Length
Sepal.Width    1.11756978
Petal.Length   0.12824622  1.42844010
Petal.Width    0.18205887  1.36612593   0.03713457
> prop.factor<-euclidian.sqdist/cor.data
> prop.factor # propotional factor
             Sepal.Length Sepal.Width Petal.Length
Sepal.Width           298
Petal.Length          298         298
Petal.Width           298         298          298
```

2d)

```
> pca.unscaled <- prcomp(iris.data, scale=FALSE)
> pca.unscaled
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
```

Solutions to homework module 12

```
Petal.Length  0.85667061  0.17337266 -0.07623608 -0.4798390
Petal.Width   0.35828920  0.07548102 -0.54583143  0.7536574
> pca.scaled <- prcomp(scaled.data, scale=FALSE)
> pca.scaled
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
```

They are not same

2e)

```
> summary(pca.unscaled)
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(pca.scaled)
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
```

Percentage of var explained by first 2 components in scaled PCA = 95.81%

Percentage of var explained by first 2 components in unscaled PCA = 97.77%

2f)

```
> quantile(sdevs[,2], c(0.05,0.95))
       5%       95%
0.8420434 1.0434446
```

90% CI on the proportion of variance explained by second principle component is (0.8404, 1.0434)

Solutions to homework module 12