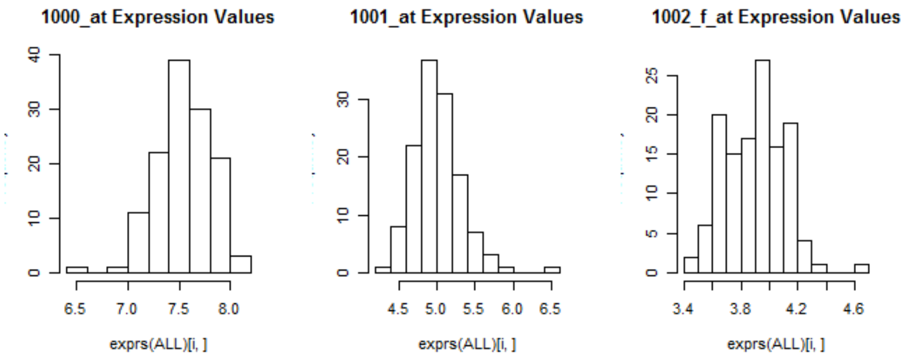
Charles Valentine
Homework 12
11/23/2016

Note:
All code is in script hw12.R --- I have made use of cat and print statements to display information easily!

Problem 1

a) See code.
b) Histograms for the first three genes' expression values.

OUTPUT

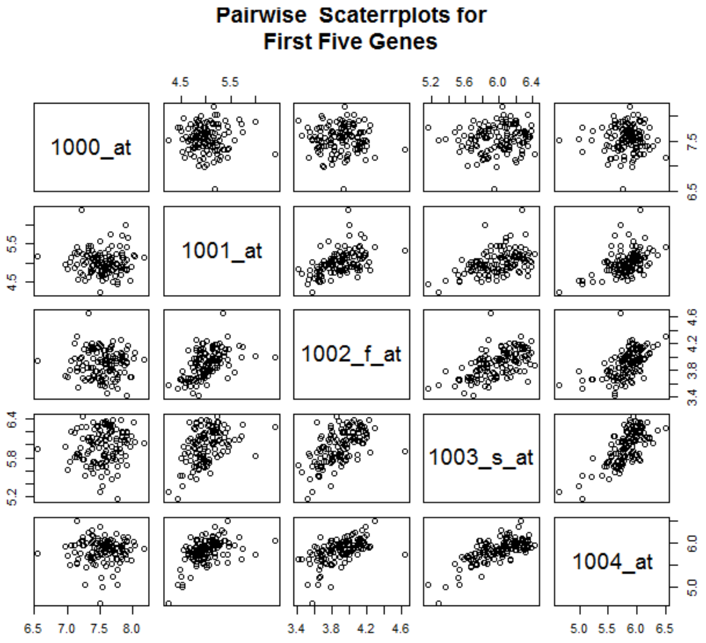

1000_at Expression Values

1001_at Expression Values

1002_f_at Expression Values

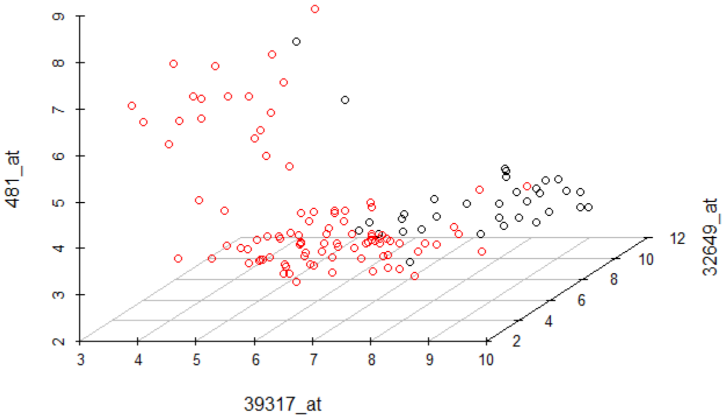
c) Pairwise scatterplots for the first five genes' expression values.

OUTPUT


Pairwise Scaterrplots for
First Five Genes

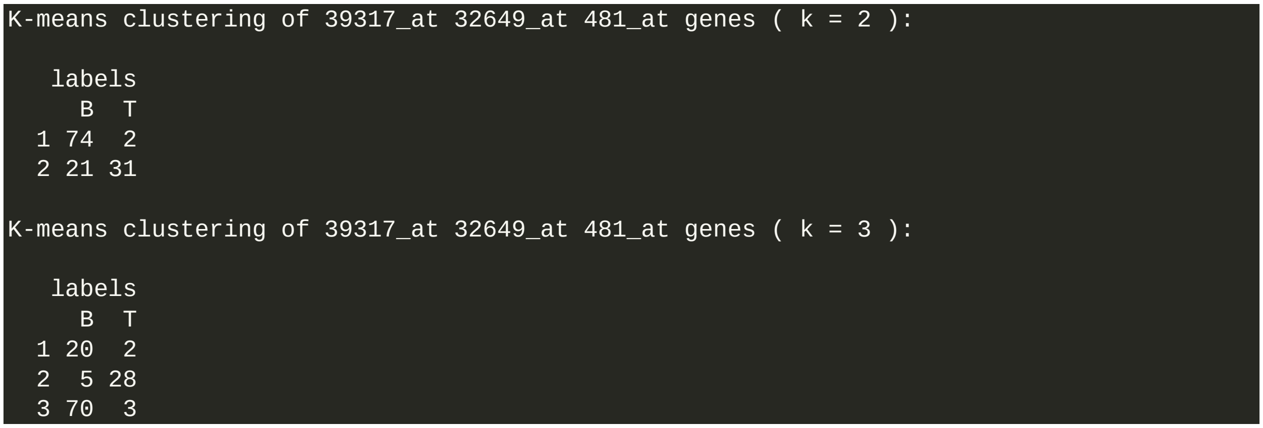
d) 3D-scatterplot for the genes *3917_at*, *32649_at*, and *481_at* colored to represent patients with labels B-cell or T-cell. T-cell patients are labeled black and B-cell patients are labeled red. The two groups can be distinguished as they appear to cluster together. There is no linear boundary in which the groups can be separated and a clustering algorithm will have to be used to attempt to classify the two patient groups in an unsupervised manner.

OUTPUT



e) The two groups (B-cell and T-cell patients) are discovered roughly with a *k* = 2 and K-means clustering. Only 2 out of 33 B-cell patients were misclassified as T-cell patients and only 21 out of 95 T-cell patients were misclassified as B-cell patients.
The clustering shows improvement when *k* = 3 as the amount of misclassified B-cell patients lessens to 5 out of 95. This is coupled with a slight increase in misclassification of T-cell patients (5 out of 33).

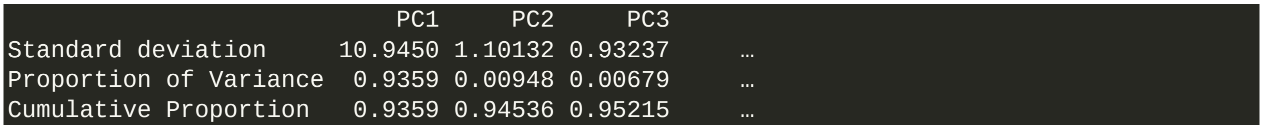
OUTPUT



```
K-means clustering of 39317_at 32649_at 481_at genes ( k = 2 ):

   labels
     B  T
  1 74  2
  2 21 31

K-means clustering of 39317_at 32649_at 481_at genes ( k = 3 ):

   labels
     B  T
  1 20  2
  2  5 28
  3 70  3
```

f) PCA was carried out on the ALL dataset with scaled variables. The proportion of the variance that can be attributed to the first principal component is 93.59%. The proportion of the variance that can be attributed to the second principal component is 0.95%.

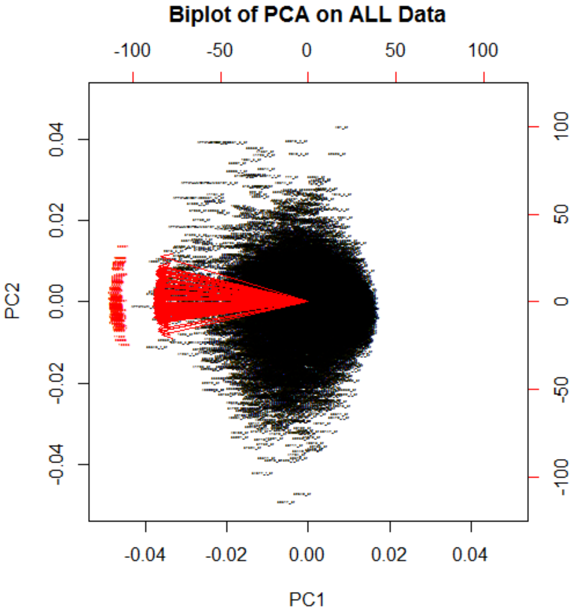
OUTPUT



```
                           PC1     PC2     PC3
Standard deviation     10.9450 1.10132 0.93237     …
Proportion of Variance  0.9359 0.00948 0.00679     …
Cumulative Proportion   0.9359 0.94536 0.95215     …
```

g) A biplot of the first two principal components of the ALL data is shown below. The loadings (red) are roughly all pointing in the same direction and are all of equivalent length. This indicates to us that PC1 explains nearly all of the variance of the gene data. It is likely that PC2 represents the distinction between T-cell vs. B-cell patients.

OUTPUT



h) For the PCA of the entire ALL dataset the biggest and smallest PC2 values and their corresponding genes were found.

OUTPUT

```
The three genes with the biggest PC2 values are:

39317_at
32649_at
34677_f_at

The three genes with the smallest PC2 values are:

41165_g_at
38018_g_at
481_at
```

i) The names and the chromosomes were then found for the genes in Problem 1h using the *annotation* routine in R.

OUTPUT

```
Gene with biggest PC2 value
 ID: 39317_at
 Chromosome: 6
 Name: cytidine monophospho-N-acetylneuraminic acid hydroxylase, pseudogene

Gene with smallest PC2 value
 ID: 481_at
 Chromosome: 3
 Name: SNF related kinase
```

Problem 2

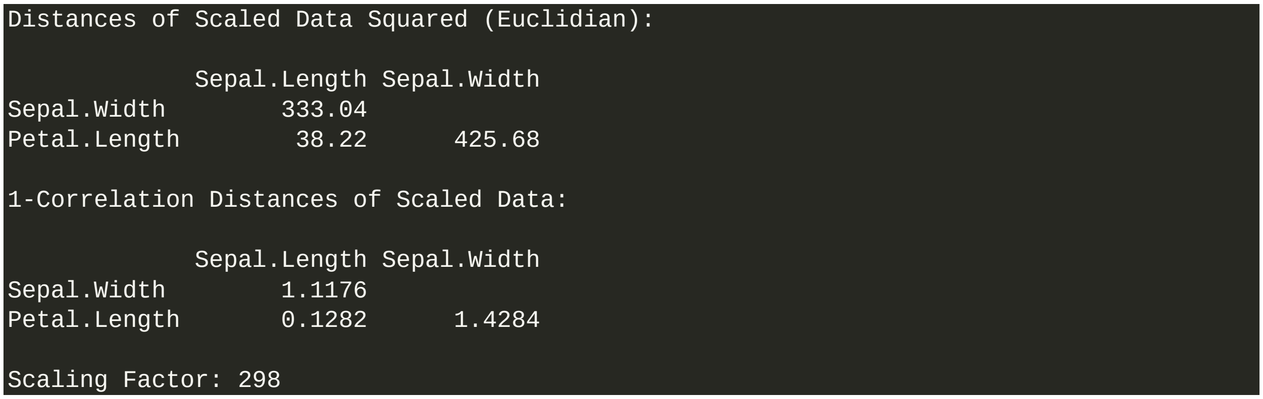
a) See code for implementation.
b) The correlations between the columns of the Iris data were compared for scaled and unscaled data. The two tables shown below indicate that the pairwise correlations are unaffected by data scaling.

OUTPUT

```
Pairwise correlation of unscaled data:
             Sepal.Length Sepal.Width Petal.Length
Sepal.Length       1.0000     -0.1176       0.8718
Sepal.Width       -0.1176      1.0000      -0.4284
Petal.Length       0.8718     -0.4284       1.0000


Pairwise correlation of scaled data:
             Sepal.Length Sepal.Width Petal.Length
Sepal.Length       1.0000     -0.1176       0.8718
Sepal.Width       -0.1176      1.0000      -0.4284
Petal.Length       0.8718     -0.4284       1.0000
```

c) The Euclidean distances between the columns of the scaled data were determined using the *dist* routine in R. Distances between the columns of the scaled data were also calculated using the $1 - \text{correlation}$ metric. The squared Euclidean distances were then compared to the $1 - \text{correlation}$ distances and a scaling factor of 298 was computed.

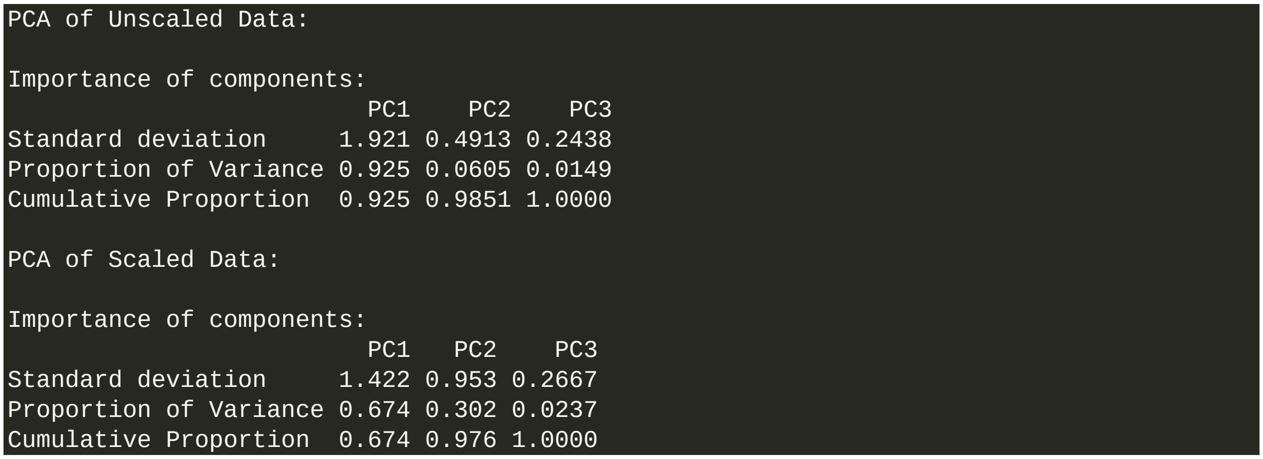
OUTPUT



```
Distances of Scaled Data Squared (Euclidian):

             Sepal.Length Sepal.Width
Sepal.Width        333.04
Petal.Length        38.22      425.68

1-Correlation Distances of Scaled Data:

             Sepal.Length Sepal.Width
Sepal.Width        1.1176
Petal.Length       0.1282      1.4284

Scaling Factor: 298
```

d) The outputs for the scaled and unscale PCS on the Iris data are not the same.

OUTPUT

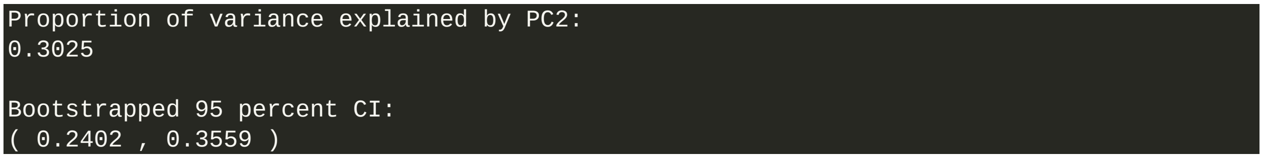


```
PCA of Unscaled Data:

Importance of components:
                         PC1    PC2    PC3
Standard deviation     1.921 0.4913 0.2438
Proportion of Variance 0.925 0.0605 0.0149
Cumulative Proportion  0.925 0.9851 1.0000

PCA of Scaled Data:

Importance of components:
                         PC1   PC2    PC3
Standard deviation     1.422 0.953 0.2667
Proportion of Variance 0.674 0.302 0.0237
Cumulative Proportion  0.674 0.976 1.0000
```

e) In the unscaled data as presented in Problem 2d the first principal component (PC1) explain 92.5% of the variance and the second principal component (PC2) explain 6.05% of the variance. This is in contrast to the PCA on the scale data in which PC1 explain 67.4% of the variance and PC2 explain 30.2% of the variance.

f) The 90% confidence intervals on the proportion of the variance explained by PC2 are presented below. They were determined using a bootstrap method.

OUTPUT



```
Proportion of variance explained by PC2:
0.3025

Bootstrapped 95 percent CI:
( 0.2402 , 0.3559 )
```