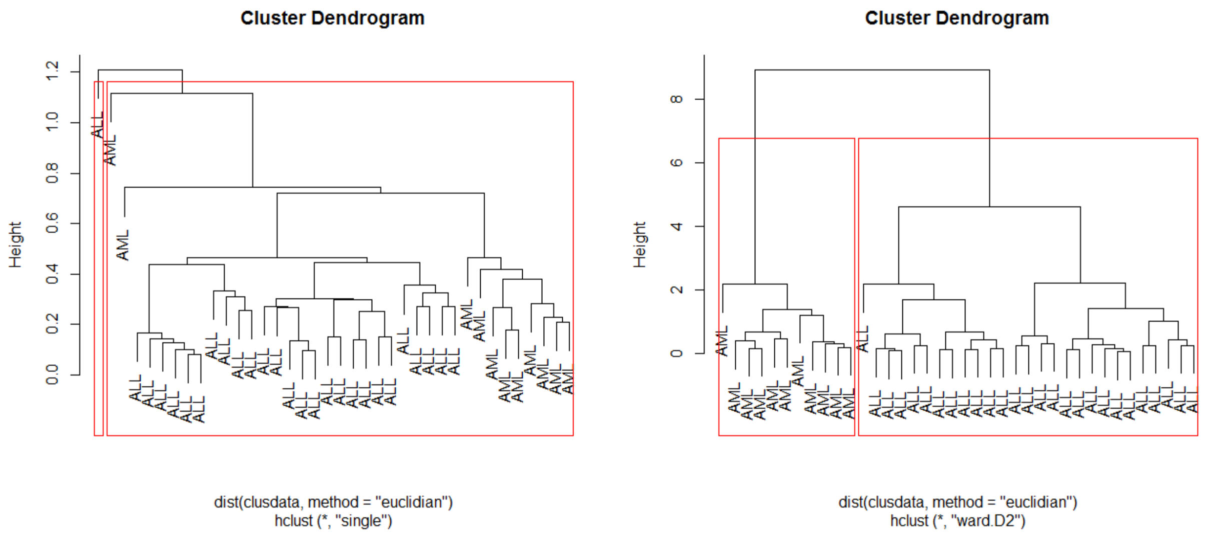
Charles Valentine
Homework 11
11/20/2016

Note:
All code is in script hw11.R --- I have made use of cat and print statements to display information easily!

Problem 1

a) Hierarchical clustering of the CCND3 Cyclin D3 and Zyxin gene expression data of the Golub 1999 set were compared using single and Ward.D3 linkage. Two clusters were compared and the output visualized. The Ward.D3 visually created two clustered that separated AML and ALL patients. The *table* function confirms this as all AML and ALL labels are assigned correctly.

OUTPUT



```
         single.k2
gol.fac   1  2
    ALL 26  1
    AML 11  0

         ward.k2
gol.fac   1  2
    ALL 27  0
    AML  0 11
```

b) K-means clustering with k=2 was also used and the *table* function was run to compare the accuracy of clustering assignment. Similar to Ward.D2 the clusters were accurately created.

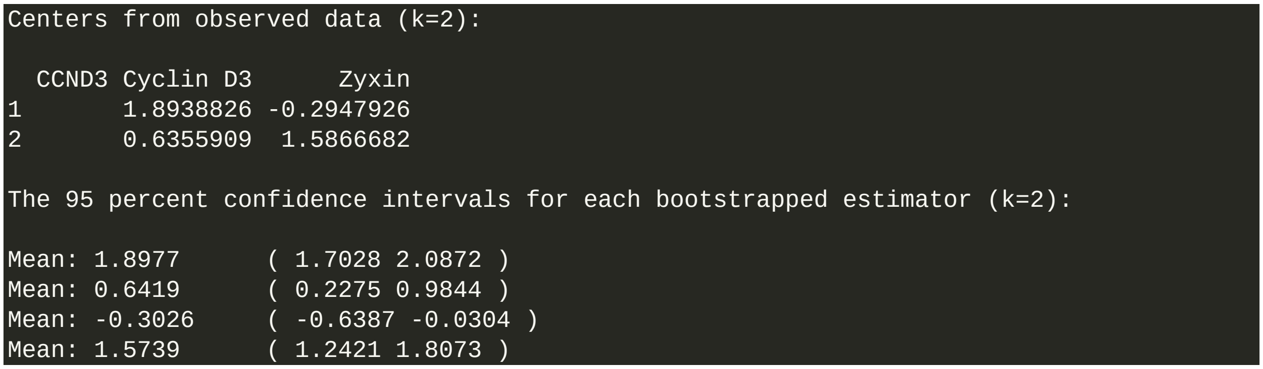
OUTPUT

```
gol.fac   1  2
    ALL 27  0
    AML  0 11
```

c) The Ward.D2 and K-means clustering routines both accurately defined the clusters. Single linkage routine failed to do so.

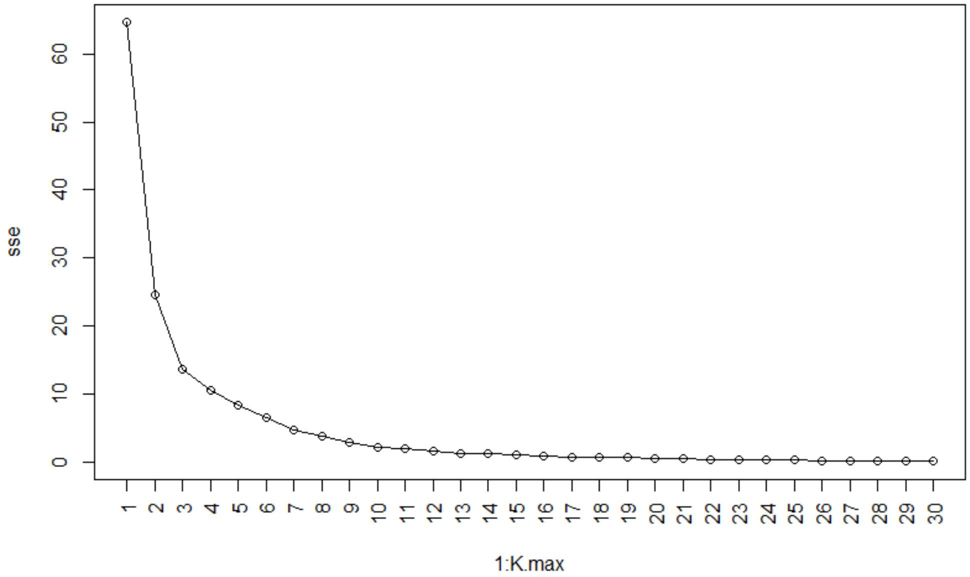
d) The centers of the clusters created in Problem 1b were found and bootstrapped. The bootstrapped point estimators for the means of the clusters were very accurate as evidence by relatively small 95% confidence intervals. No confidence intervals overlap which indicates that the cluster assignment is robust.

OUTPUT



```
Centers from observed data (k=2):

  CCND3 Cyclin D3      Zyxin
1       1.8938826 -0.2947926
2       0.6355909  1.5866682

The 95 percent confidence intervals for each bootstrapped estimator (k=2):

Mean: 1.8977      ( 1.7028 2.0872 )
Mean: 0.6419      ( 0.2275 0.9844 )
Mean: -0.3026     ( -0.6387 -0.0304 )
Mean: 1.5739      ( 1.2421 1.8073 )
```

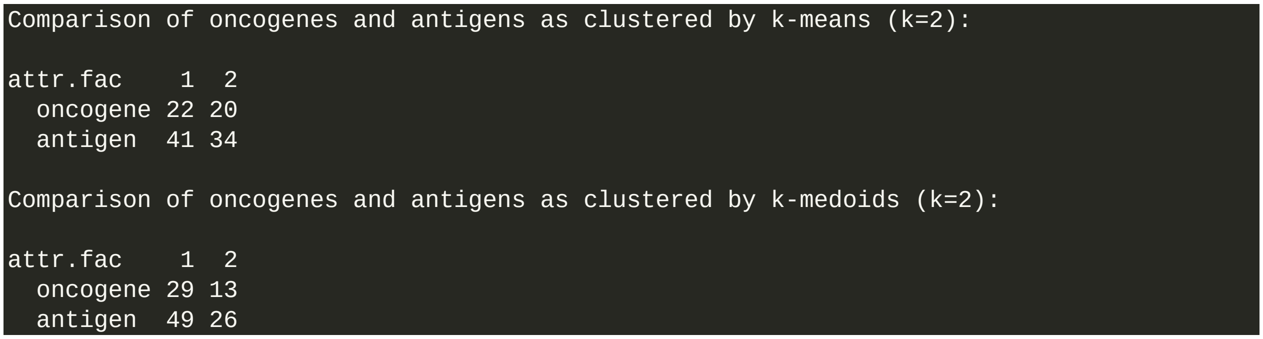
e) A plot of *K* versus *SSE* was created for K=1, …, 30. The plot suggests that there may be an optimal cluster assignment for K = {2, 3} as SSE tends to taper off at K = 3. If we examine the Ward.D2 linkage from Problem 1a we can visually see that there may be a robust explanation for three clusters which would sub-define ALL patients into two groups.



Problem 2

a) See code for implementation.
b) K-means and K-mediods with K=2 were run on the Golub data for all oncogenes and antigen expression values. Both routines struggled to assign the gene expression values to the appropriate groups (oncogene *vs.* antigen). However, the K-medoid routine does appear to work a little better.

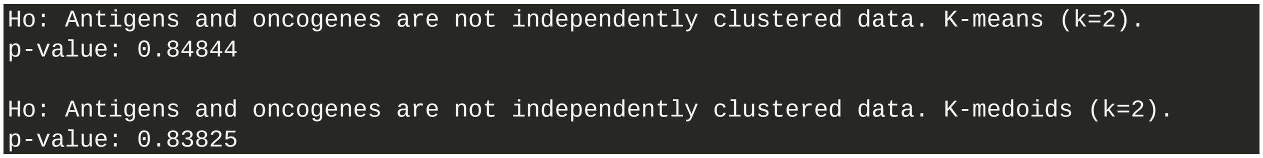
OUTPUT



```
Comparison of oncogenes and antigens as clustered by k-means (k=2):

attr.fac    1  2
  oncogene 22 20
  antigen  41 34

Comparison of oncogenes and antigens as clustered by k-medoids (k=2):

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

c) The appropriate test for these contingency tables is the Fisher's exact test. This test was implemented and the following p-values were calculated. We confirm our expectation that both routines performed poorly on assigning gene expression values to the proper cluster label. The p-value for the K-medoids routine is slightly smaller indicating better performance. We must reject the null hypothesis and claim there are no structural differences between the expression values between oncogenes and antigens.

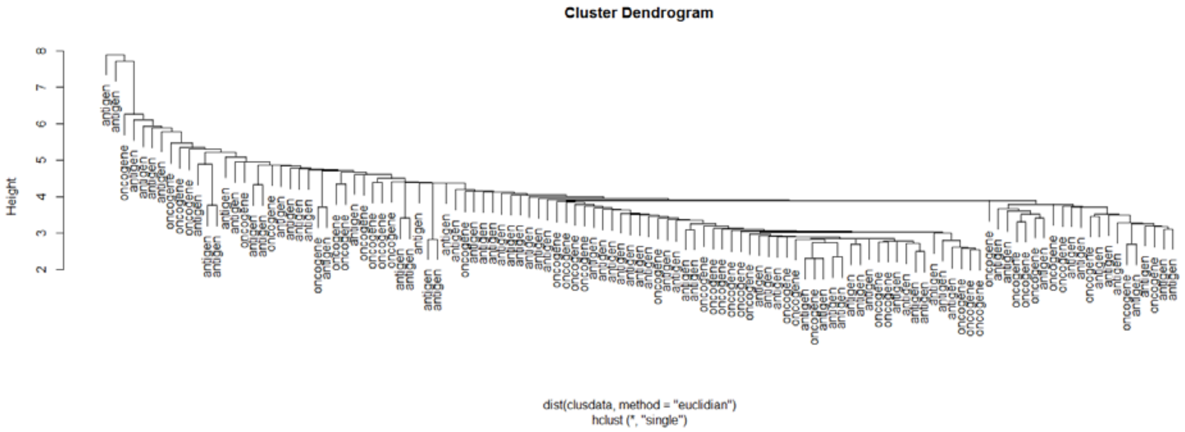
OUTPUT

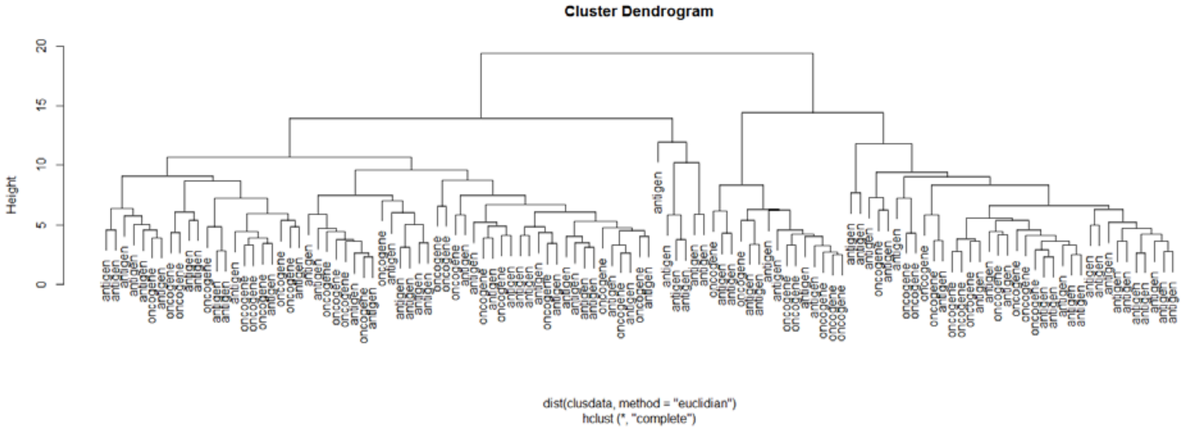


```
Ho: Antigens and oncogenes are not independently clustered data. K-means (k=2).
p-value: 0.84844

Ho: Antigens and oncogenes are not independently clustered data. K-medoids (k=2).
p-value: 0.83825
```

d) The cluster dendrograms for the gene expression values selected for in Problem 2a are shown below. The figure above uses the metric of Euclidean distance and the method of single linkage. The figure below uses the metric of Euclidean distance and the method of complete linkage. It is easy to see the routines behave differently in clustering the data yet there is no visual clusters that satisfy the labelling scheme of the gene expression data.

OUTPUT

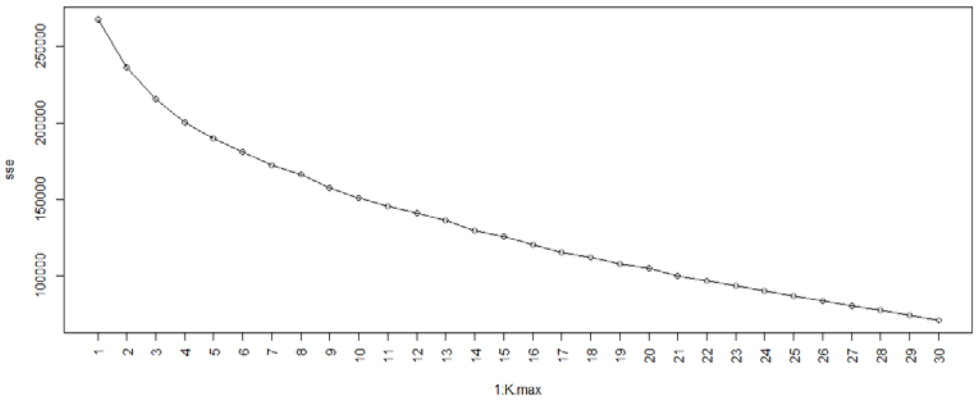




Problem 3

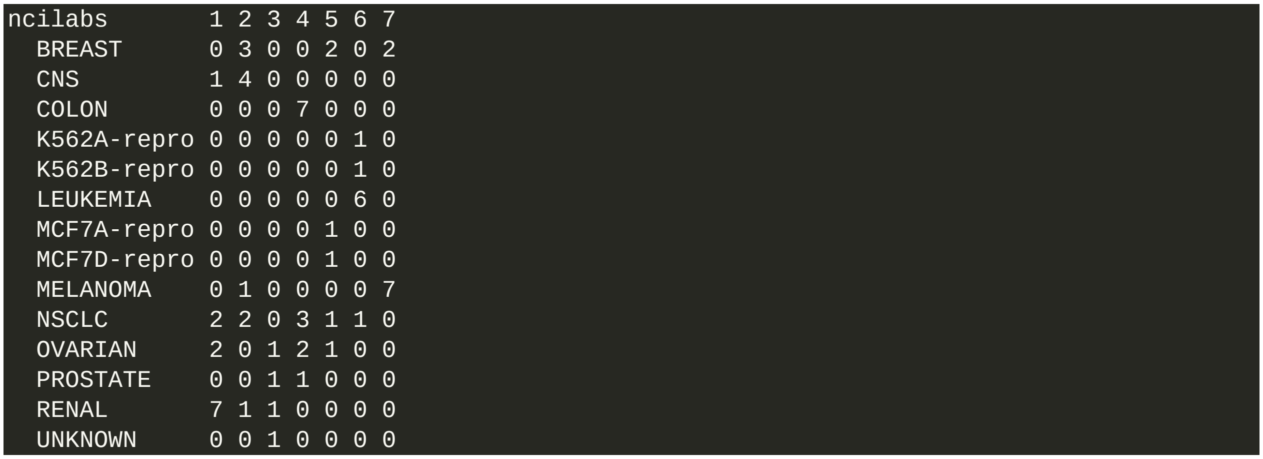
a) A plot of K versus SSE was created for K=1, …, 30 on the ISLR NCIdata. The plot suggests that there may be no optimal cluster number as there is a steady tapering of the SSE for all observed values of K. A slight dampening of this drop off is evident at K=6 and may indicate K=7 for Problem 3b as being a decent choice. Incidentally, we may find different meaningful behavior for any low value of K.

OUTPUT



b) K-medoids clustering (K=7) was performed on the ISLR NCIdata with 1-correlation as the dissimilarity measure. The clusters, as compared to the cell lines, appear to have binned some cell lines to sole clusters (success) and others to multiple clusters (bridges between clustered cell lines). The colon cell line, for example, is successively defined in the fourth cluster. The NSCLC cell line, however, is less specifically in a cluster as it assigned across clusters 1, 2, 4, 5, and 6. Other cell lines that are well defined in a cluster are CNS, K562A-repro, K562B-repro, leukemia, MCF7A-repro, MCF7D-repro, melanoma, renal, and unknown. The other types of cancer not well defined are breast and prostate. The cancer most similar to ovarian (clusters near) is potentially NSCLC as it shares a similar density of samples in the same clusters.

OUTPUT



```
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
```