Homework 12

Andrew Boschee

No Collaborators. OUtside Resources: rpkgs.datanovia.com, rdocumentation.com, R Graphics Cookbook

1. Question 10.7.8 pg 416: On the USArrests data, calculate PVE in two ways:

a. Using the sdev output of the prcomp() function, as was done in section 10.2.3
b. By applying Equation 10.8 directly. That is, use the prcomp() function to compute the principal component loadings. Then, use those loadings in Equation 10.8 to obtain the PVE.

Table 1: Mean Values of Variables

Murder	7.788
Assault	170.760
UrbanPop	65.540
Rape	21.232

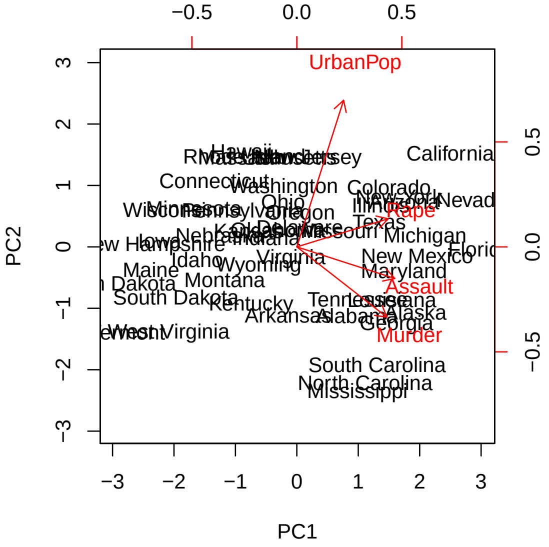
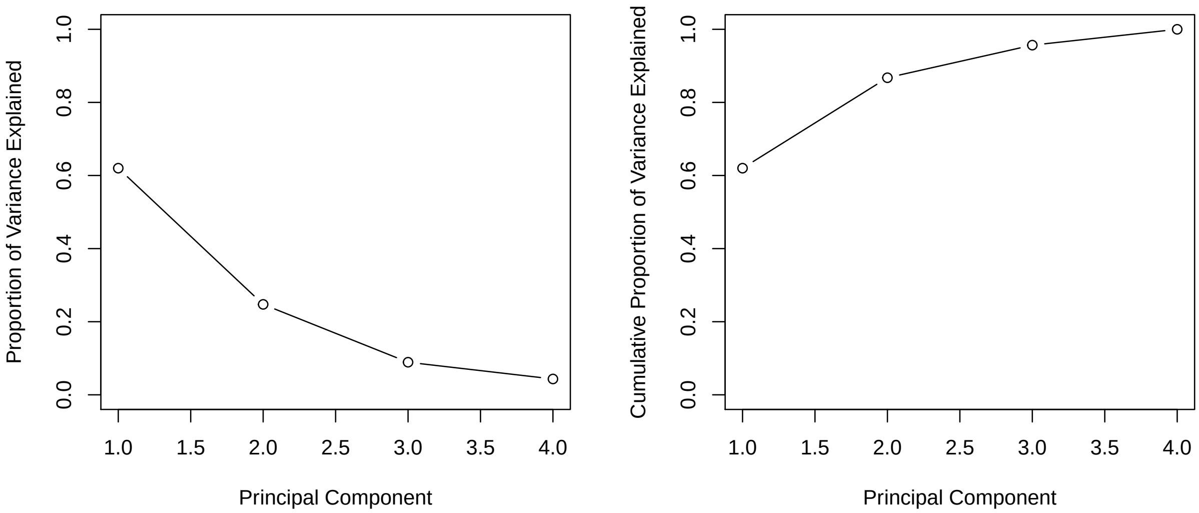


Table 2: Standard Deviation of Each Variable

	Standard Deviation
1	1.5748783
2	0.9948694
3	0.5971291
4	0.4164494

Table 3: Proportion of Variance by Component

	Proportion of Variance
1	0.6200604
2	0.2474413
3	0.0891408
4	0.0433575

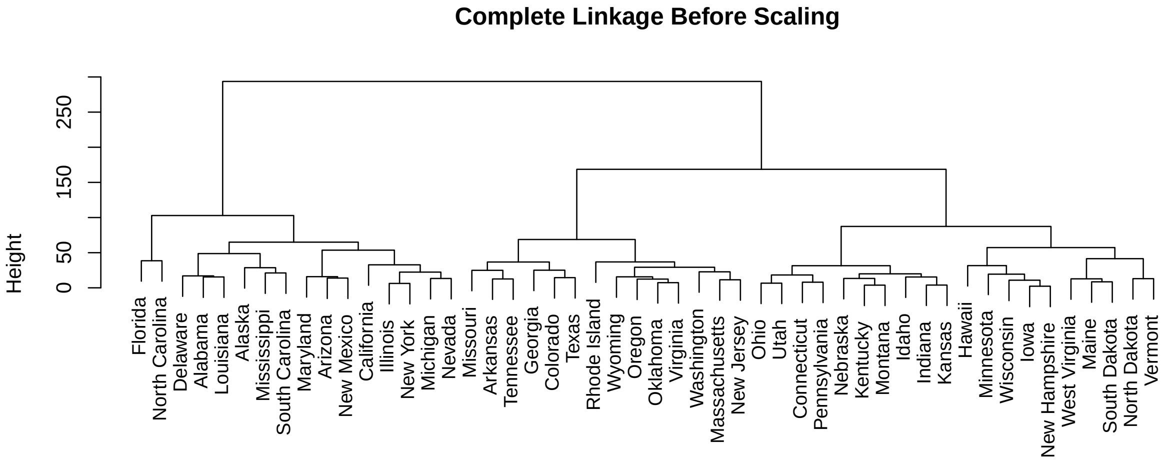


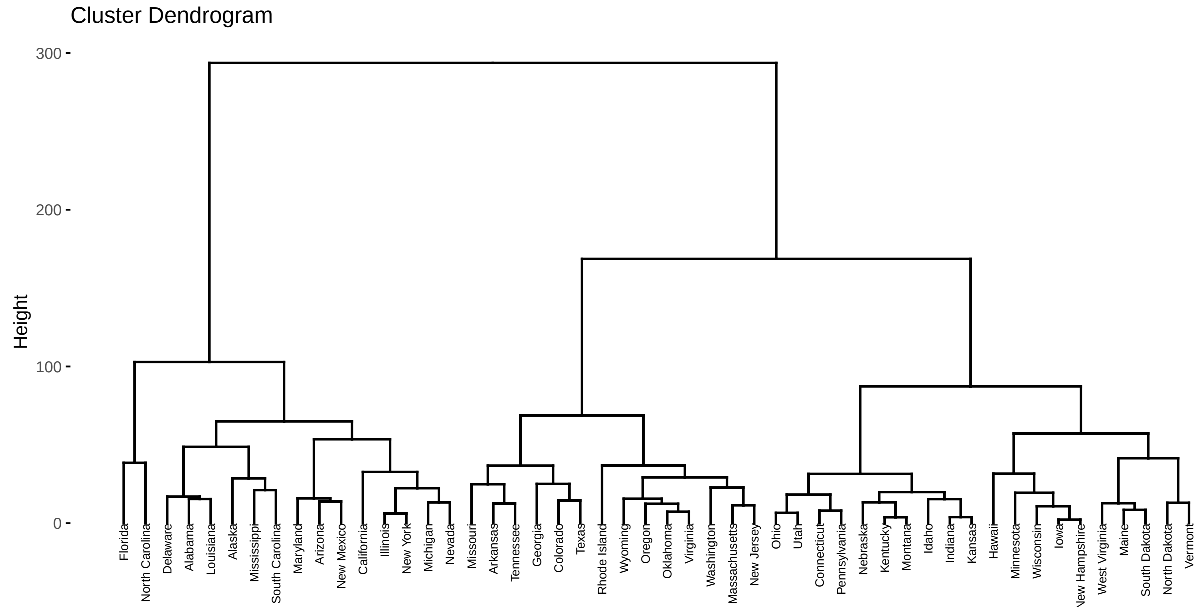
For part A the sdev function was used to find the standard deviation for each principal component. The proportion of variance was then calculated for each component with values given in table 3

2. Question 10.7.9 pg 416: Consider the USArrests data. We will now perform hierarchical clustering on the states.

a. Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

Data was loaded and the dist function was applied with euclidean method. Output was stored in distMat variable and hclust function applied model with complete method and stored in variable for plotting. Used plot() function and fviz_dend() function to see all states

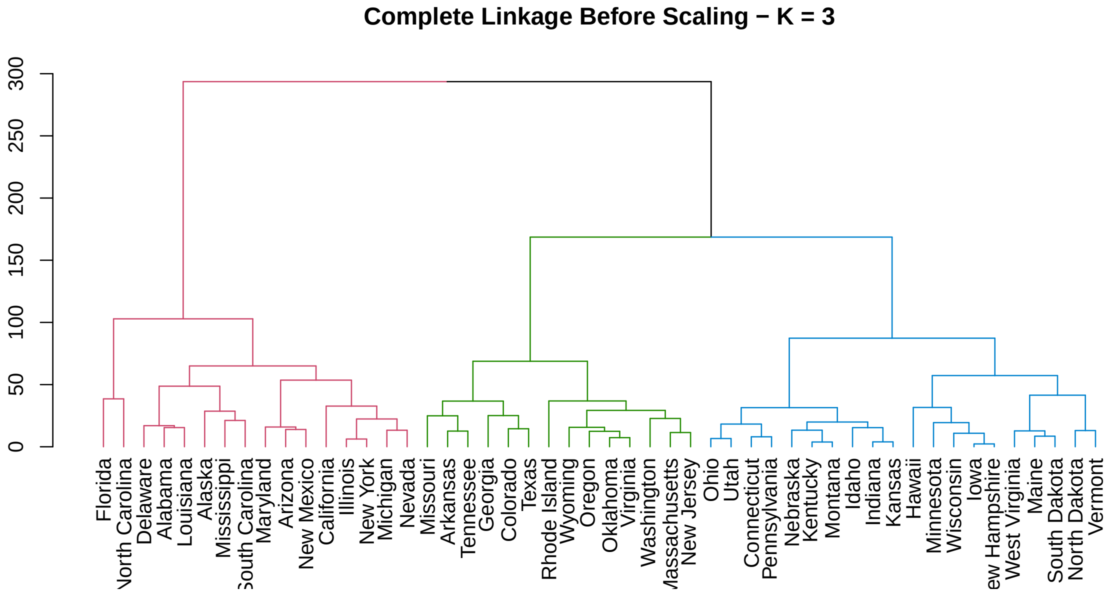




b. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

Using the cutree() function with second argument set to 3, for three clusters. Plots are now adjusted with color_branches() function to color code clusters. Prior to the graph, output from cutree shows every state with the corresponding cluster that it falls under.

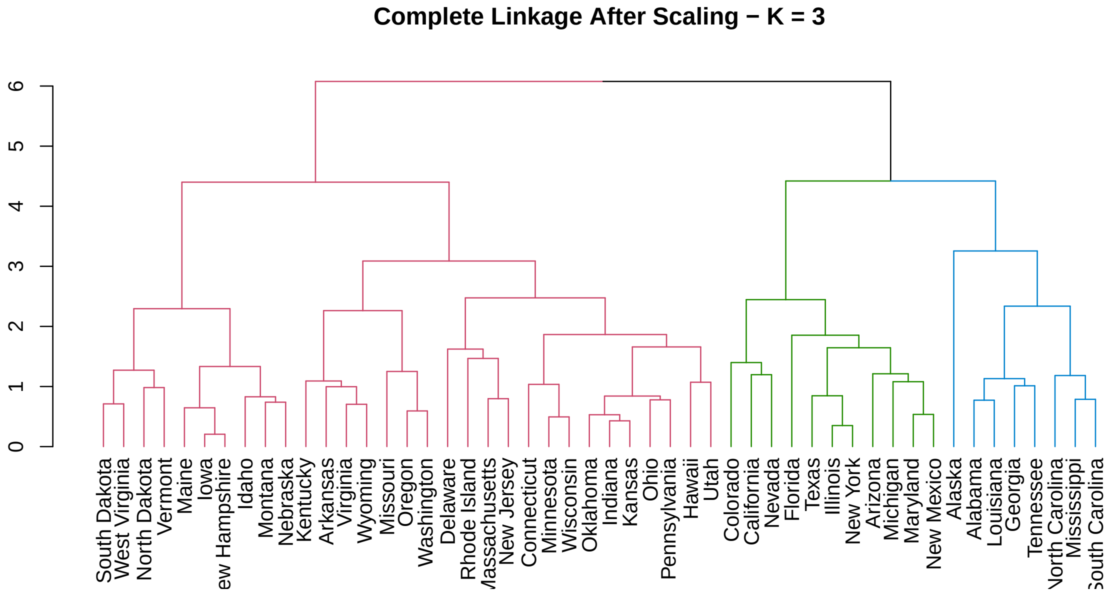
```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              3              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              3              3              2
```

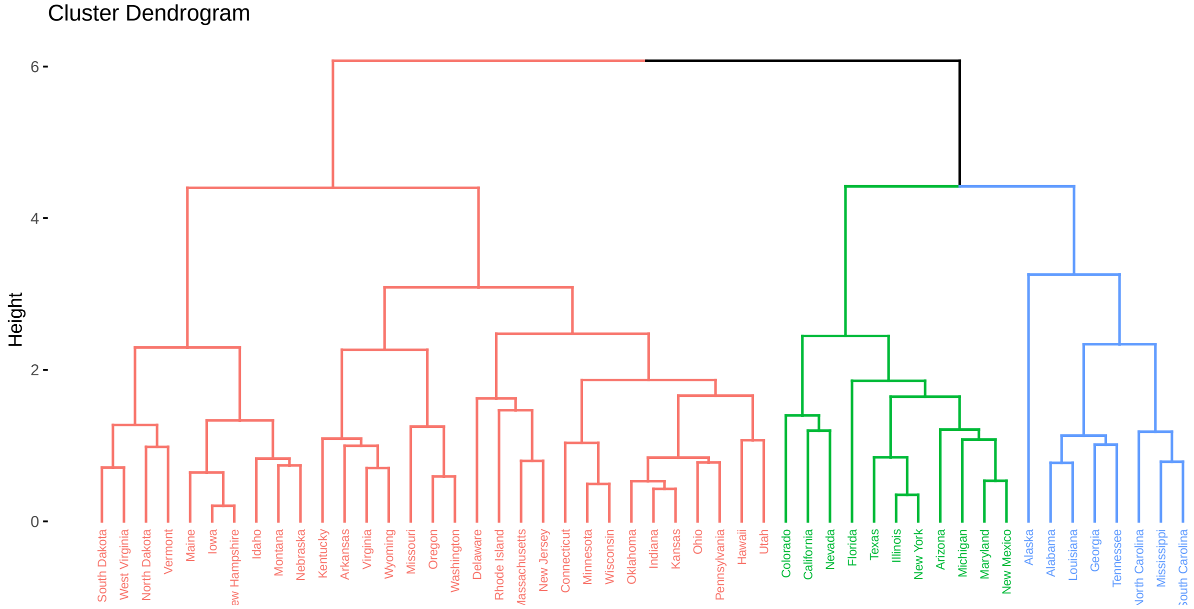


Complete Linkage Before Scaling − K = 3

c. Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

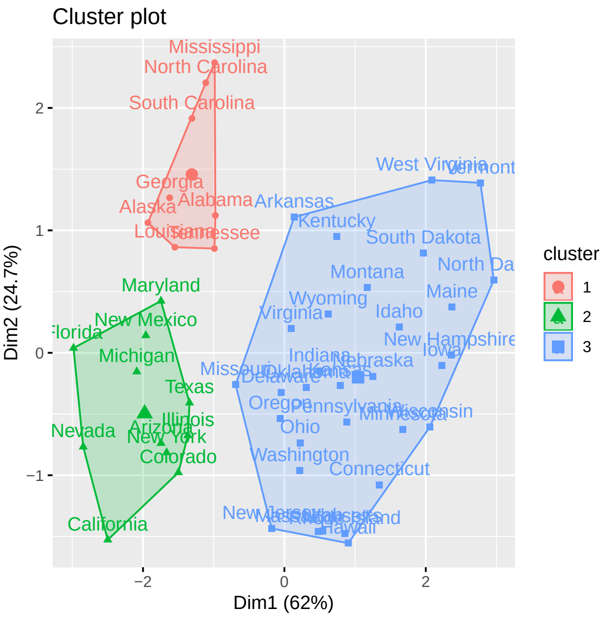
Steps from prior question repeated after modifying the dataset applying the scale() function. Can see with the scaling, that there are now many more states in the first cluster

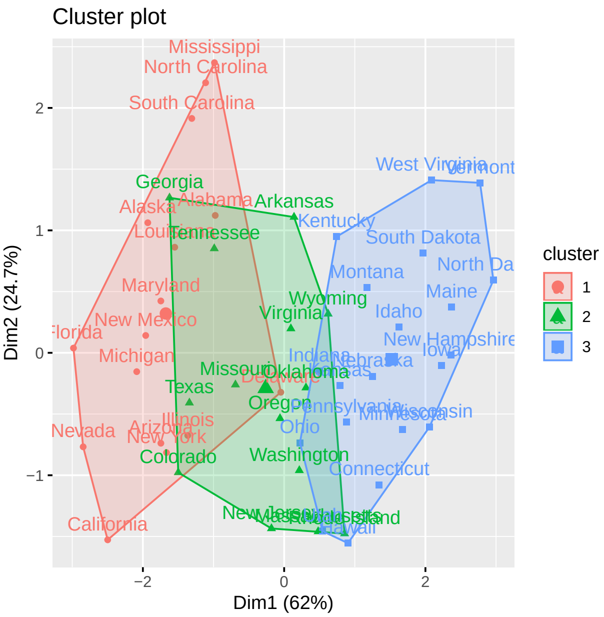




d. What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

From the plot below, we can see that there are much more defined clusters after scaling is put into place. If using this plot as visual aid, the scaled version would be very helpful, while the non-scaled would cause much confusion and is hard to read.


Cluster plot


Cluster plot

3. Question 10.7.11 pg 417: On the book website, www.StatLearning.com, there is a gene expression data set (ch10Ex11.csv) that consists of 40 tissue samples with measurements on 1000 genes. the first 20 samples are from healthy patients, while the second 20 are from a diseased group.

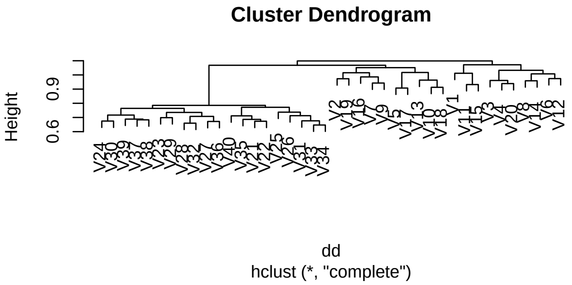
a. Load in the data using read.csv(). You will need to select header = F.

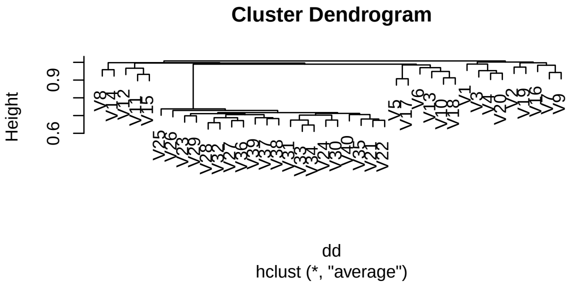
Used read.csv function as specifed and stored as geneData variable with first five columns and rows shown.

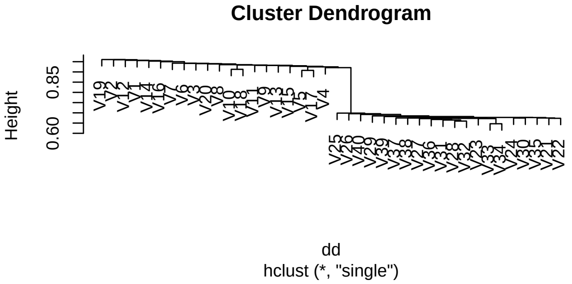
```
##           V1         V2         V3         V4         V5
## 1 -0.9619334  0.4418028 -0.9750051  1.4175040  0.8188148
## 2 -0.2925257 -1.1392670  0.1958370 -1.2811210 -0.2514393
## 3  0.2587882 -0.9728448  0.5884858 -0.8002581 -1.8203980
## 4 -1.1521320 -2.2131680 -0.8615249  0.6309253  0.9517719
## 5  0.1957828  0.5933059  0.2829921  0.2471472  1.9786680
```

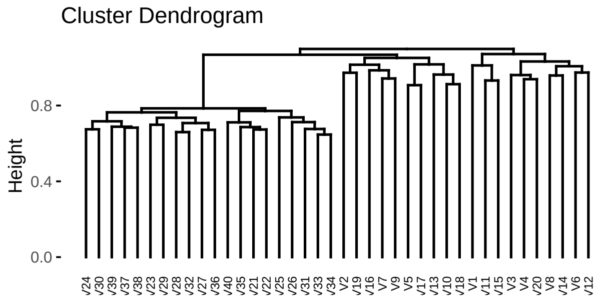
b. Apply hierarchical clustering to the samples using correlation based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used?

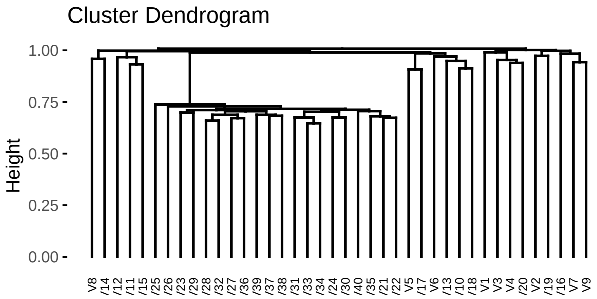
The complete and single methods both do a good job separating into two groups. The average method ends up giving an output with three groups.

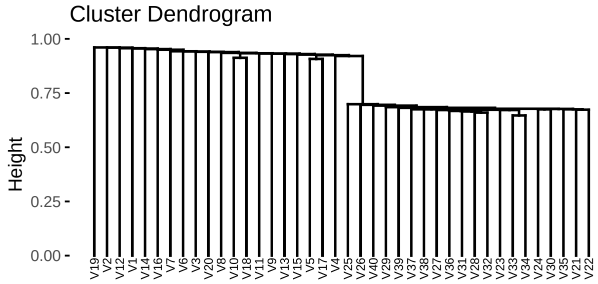

Cluster Dendrogram
Height
0.6
0.9
V24
V30
V39
V37
V38
V23
V29
V28
V32
V27
V36
V40
V35
V21
V22
V25
V26
V31
V33
V34
V2
V19
V16
V7
V9
V5
V17
V13
V10
V18
V1
V11
V15
V3
V4
V20
V8
V14
V6
V12
dd
hclust (*, "complete")


Cluster Dendrogram
Height
0.6
0.9
V8
V14
V12
V11
V15
V25
V26
V23
V29
V28
V32
V27
V36
V39
V37
V38
V31
V33
V34
V24
V30
V40
V35
V21
V22
V5
V17
V6
V13
V10
V18
V1
V3
V4
V20
V2
V19
V16
V7
V9
dd
hclust (*, "average")


Cluster Dendrogram
Height
0.60
0.85
V19
V2
V12
V1
V14
V16
V7
V6
V3
V20
V8
V10
V18
V11
V9
V13
V15
V5
V17
V4
V25
V26
V40
V29
V39
V37
V38
V27
V36
V31
V28
V32
V23
V33
V34
V24
V30
V35
V21
V22
dd
hclust (*, "single")


Cluster Dendrogram
Height
0.0
0.4
0.8
/24
/30
/39
/37
/38
/23
/29
/28
/32
/27
/36
/40
/35
/21
/22
/25
/26
/31
/33
/34
V2
/19
/16
V7
V9
V5
/17
/13
/10
/18
V1
/11
/15
V3
V4
/20
V8
/14
V6
/12


Cluster Dendrogram
Height
0.00
0.25
0.50
0.75
1.00
V8
/14
/12
/11
/15
/25
/26
/23
/29
/28
/32
/27
/36
/39
/37
/38
/31
/33
/34
/24
/30
/40
/35
/21
/22
V5
/17
V6
/13
/10
/18
V1
V3
V4
/20
V2
/19
/16
V7
V9


Cluster Dendrogram
Height
0.00
0.25
0.50
0.75
1.00
V19
V2
V12
V1
V14
V16
V7
V6
V3
V20
V8
V10
V18
V11
V9
V13
V15
V5
V17
V4
V25
V26
V40
V29
V39
V37
V38
V27
V36
V31
V28
V32
V23
V33
V34
V24
V30
V35
V21
V22

c. Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer this question, and apply it here.

Performed PCA to find where the most variance occurs and order descending. To show results the top ten are stored in vector and output below.

Table 4: Top 10 Differing Genes

865
68
911
428
624
11
524
803
980
822