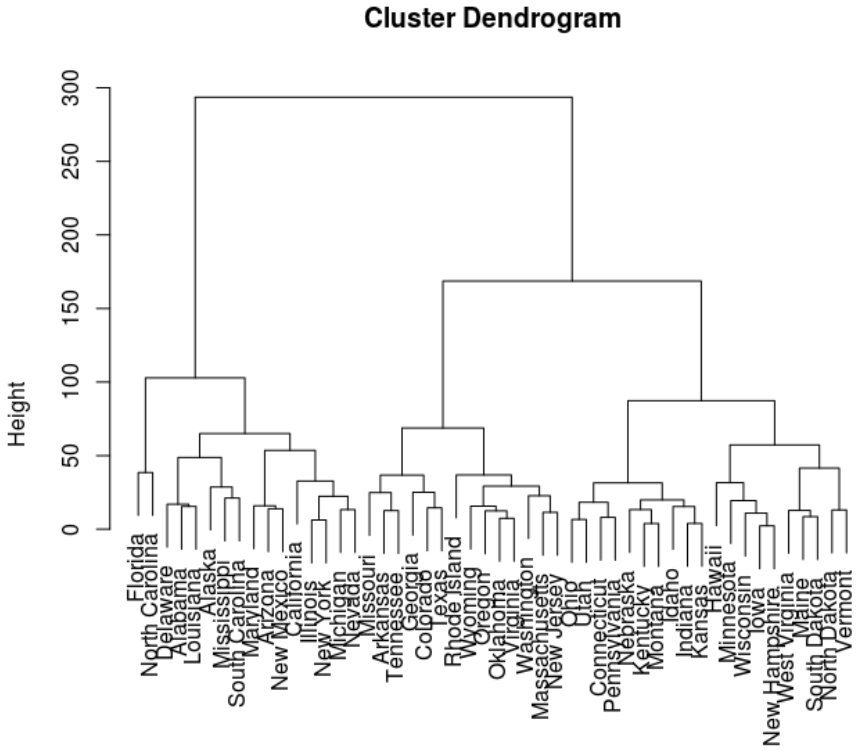
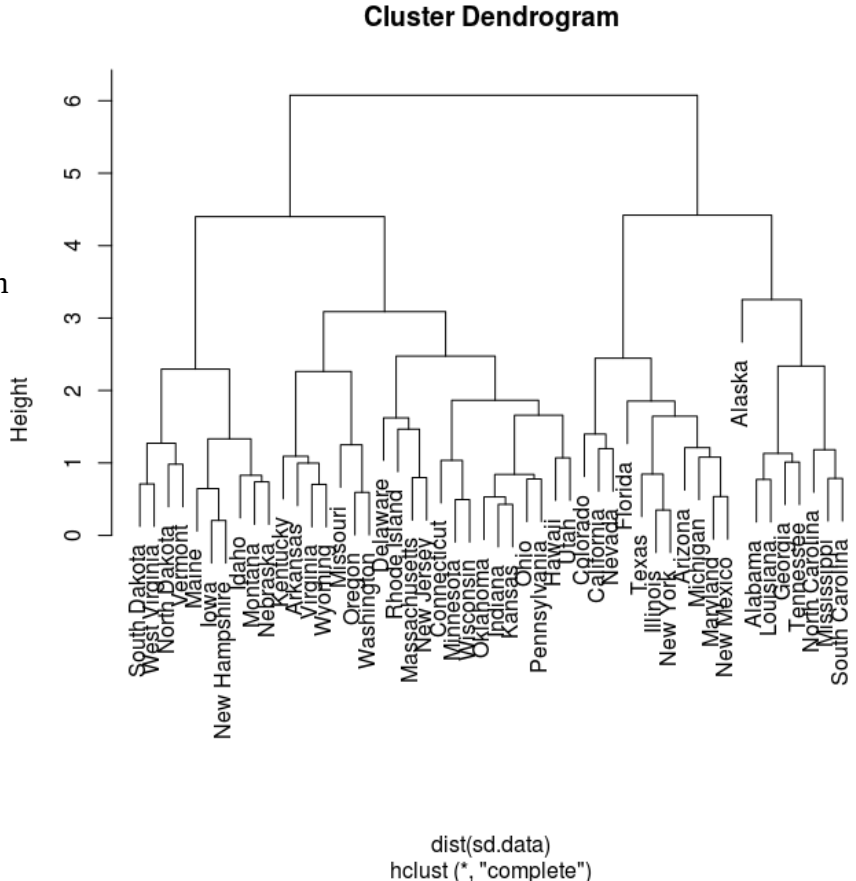
Assignment -2

Q.1)
At first, using hierarchical clustering, euclidean distance is calculated. The resultant clustered states are as below:



There are more than 3 clusters in the dendrogram. Now, clustering is done to make only 3 clusters. Resultant dendrogram is:

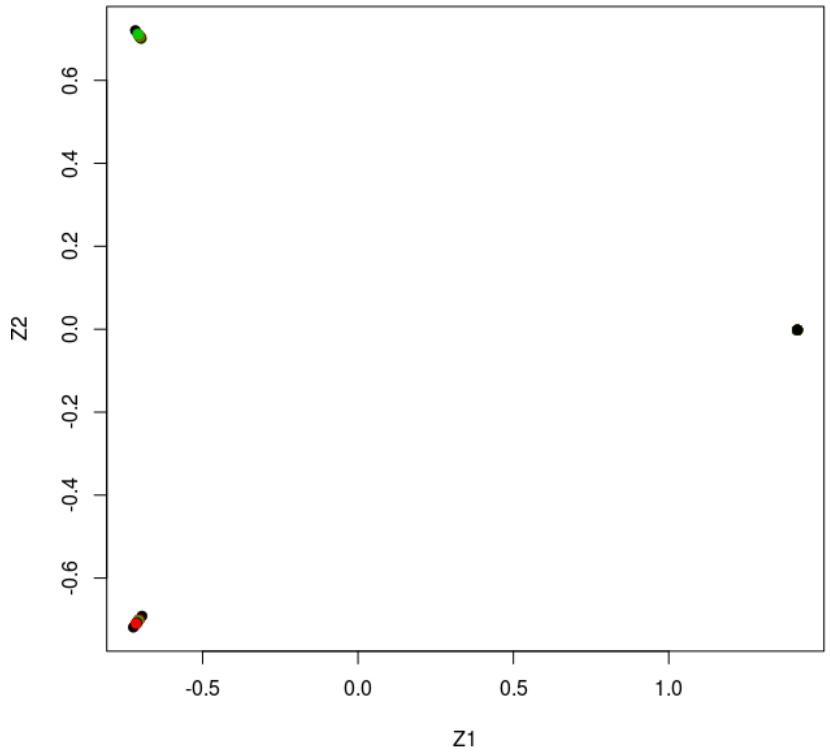


Now clustering is done with scaling data by subtracting mean and diving by SD from each entry of the data.
The table for scaled and unscaled clustering is shown below:

```
      1  2  3
1     6  9  1
2     2  2 10
3     0  0 20
```

Shows that most of the states which are in cluster 3 according to unscaled data remains in cluster 3 even after scaling. Also for 1st cluster, good proportion is data chooses cluster 1 again.

Q.2)
After several attempt for randomly created matrix and trying different values to be subtracted from some entries, the graph finally shows that 3 clusters are seperated.




```
true.labels  1  2  3
           1  0  0 20
           2  0 20  0
           3 20  0  0
```

K-means clustering for k=3 shows that 3 classes are properly separated in the 20 rows per cluster.

```
true.labels  1  2
           1  0 20
           2 20  0
           3  0 20
```

K-means clustering for k=2, one class is merged totally in either class.

```
true.labels  1  2  3  4
           1 20  0  0  0
           2  0 12  8  0
           3  0  0  0 20
```

K-means clustering for k=4 divides one of the cluster into 2 clusters. Remaining 2 clusters are unaffected.

```
true.labels  1  2  3
           1  0 20  0
           2  0  0 20
           3 20  0  0
```

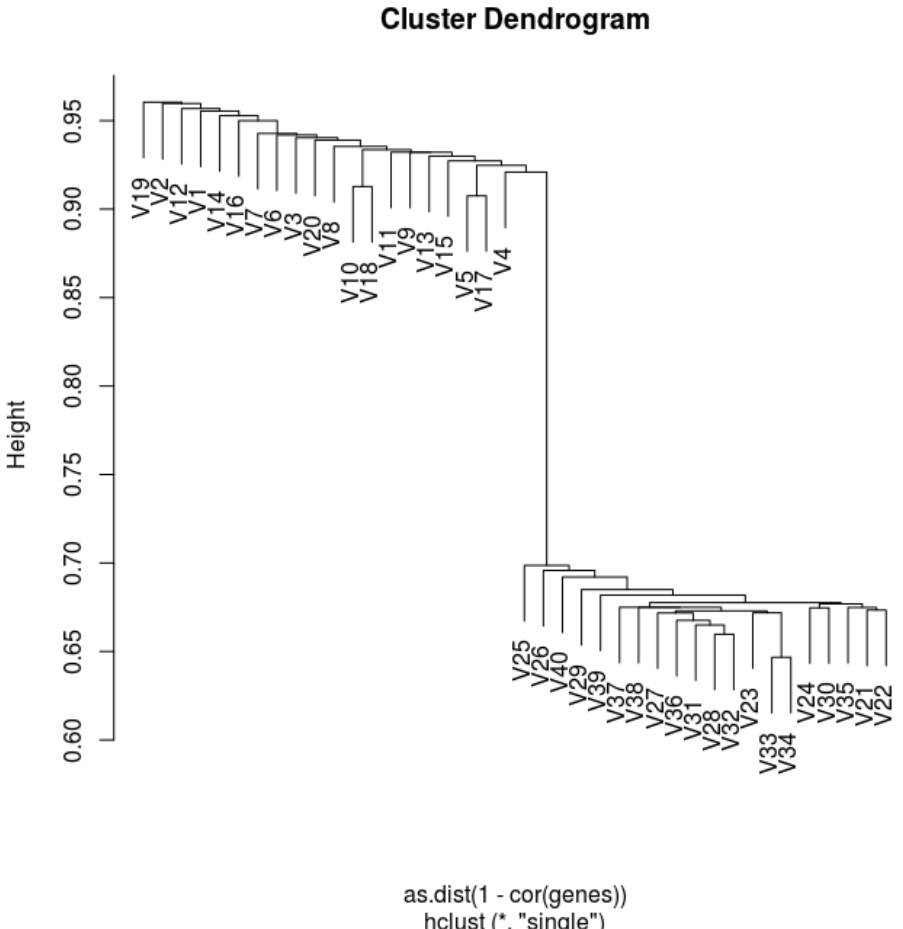
Now, clustering is done for 2 principle components only i.e. 60x2 matrix. Data is perfectly clustered again. But not same as the previous observations.

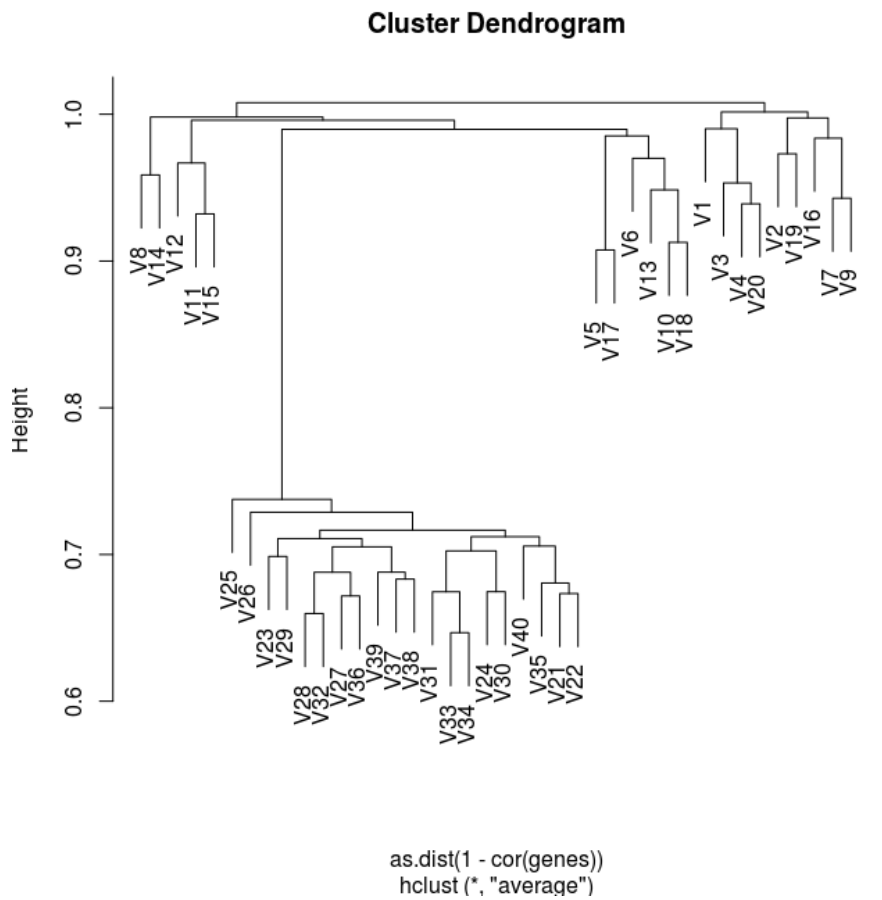
```
true.labels 1 2 3
           1 6 5 9
           2 3 9 8
           3 6 6 8
```

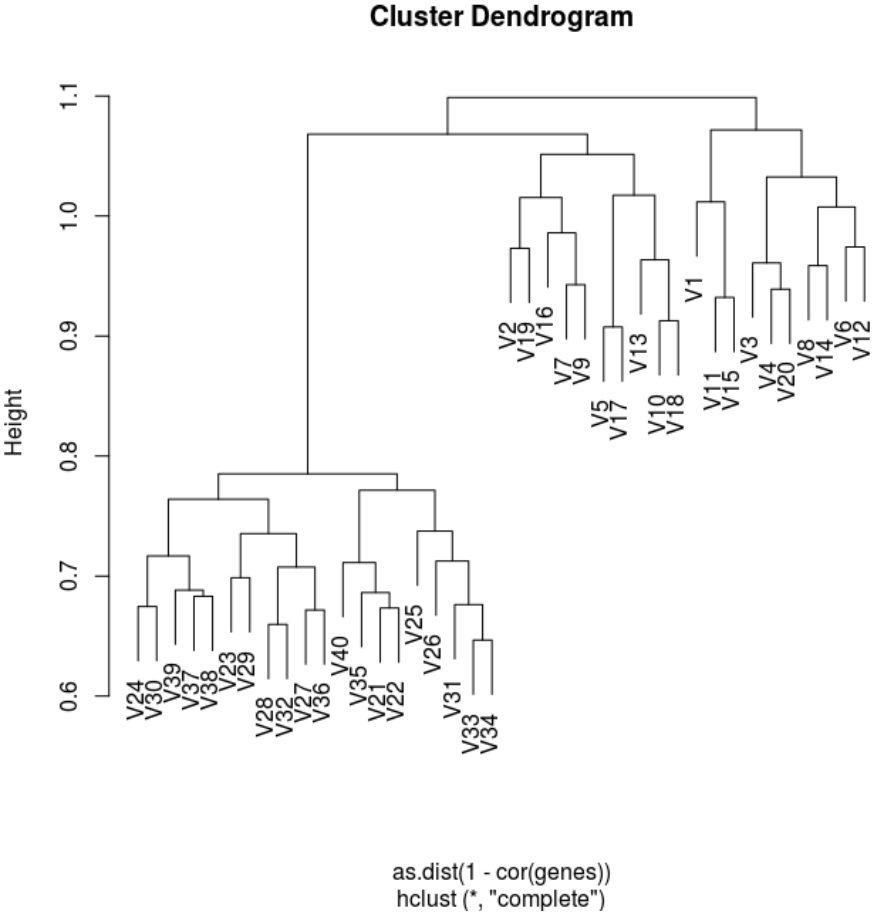
After scaling data, clustering is not perfectly done.

Q.3)

The figures below show that, clusters depend on type of linkage used.There are different clusters based on linkage methods.




Cluster Dendrogram
Height
1.0
0.9
0.8
0.7
0.6
V8
V14
V12
V11
V15
V5
V17
V6
V13
V10
V18
V1
V3
V4
V20
V2
V19
V16
V7
V9
V25
V26
V23
V29
V28
V32
V27
V36
V39
V37
V38
V31
V33
V34
V24
V30
V40
V35
V21
V22
as.dist(1 - cor(genes))
hclust (*, "average")


Cluster Dendrogram
Height
1.1
1.0
0.9
0.8
0.7
0.6
V2
V19
V16
V7
V9
V5
V17
V13
V10
V18
V1
V11
V15
V3
V4
V20
V8
V14
V6
V12
V24
V30
V39
V37
V38
V23
V29
V28
V32
V27
V36
V40
V35
V21
V22
V25
V26
V31
V33
V34
as.dist(1 - cor(genes))
hclust (*, "complete")

To get different genes, PCA is applied the data

```
865  68 911 428 624  11 524 803 980 822 529 765 801 771 570
```

These are the genes which differs the most in two clusters.