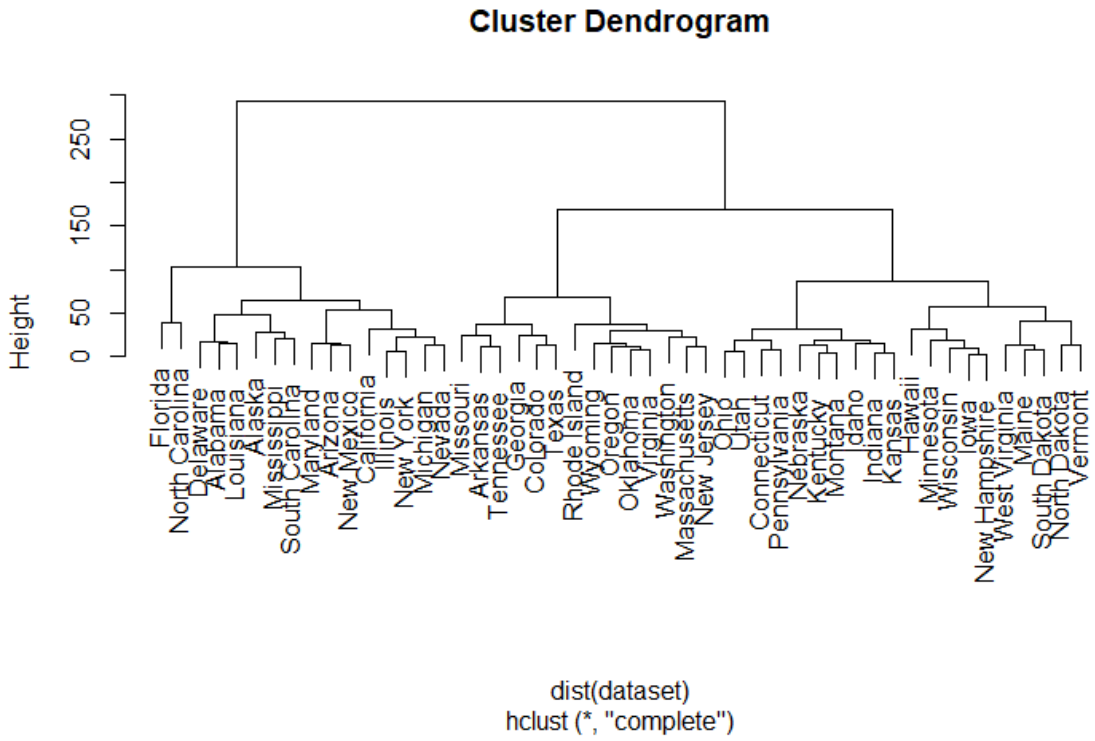
Sai Kiran Putta

Question – 9

a)

We obtain the following hierarchical cluster structure by using complete linkage and Euclidean distance.



b) Cutting the above hierarchical cluster structure to obtain 3 clusters overall using the cutree() method results in the following cluster distribution.

States in Cluster – 1

```
> cluster_df[which(cluster_df$clusters == 1),]$cities
 [1] Alabama        Alaska         Arizona        California     Delaware       Florida        Illinois
 [8] Louisiana      Maryland       Michigan       Mississippi    Nevada         New Mexico     New York
[15] North Carolina South Carolina
```

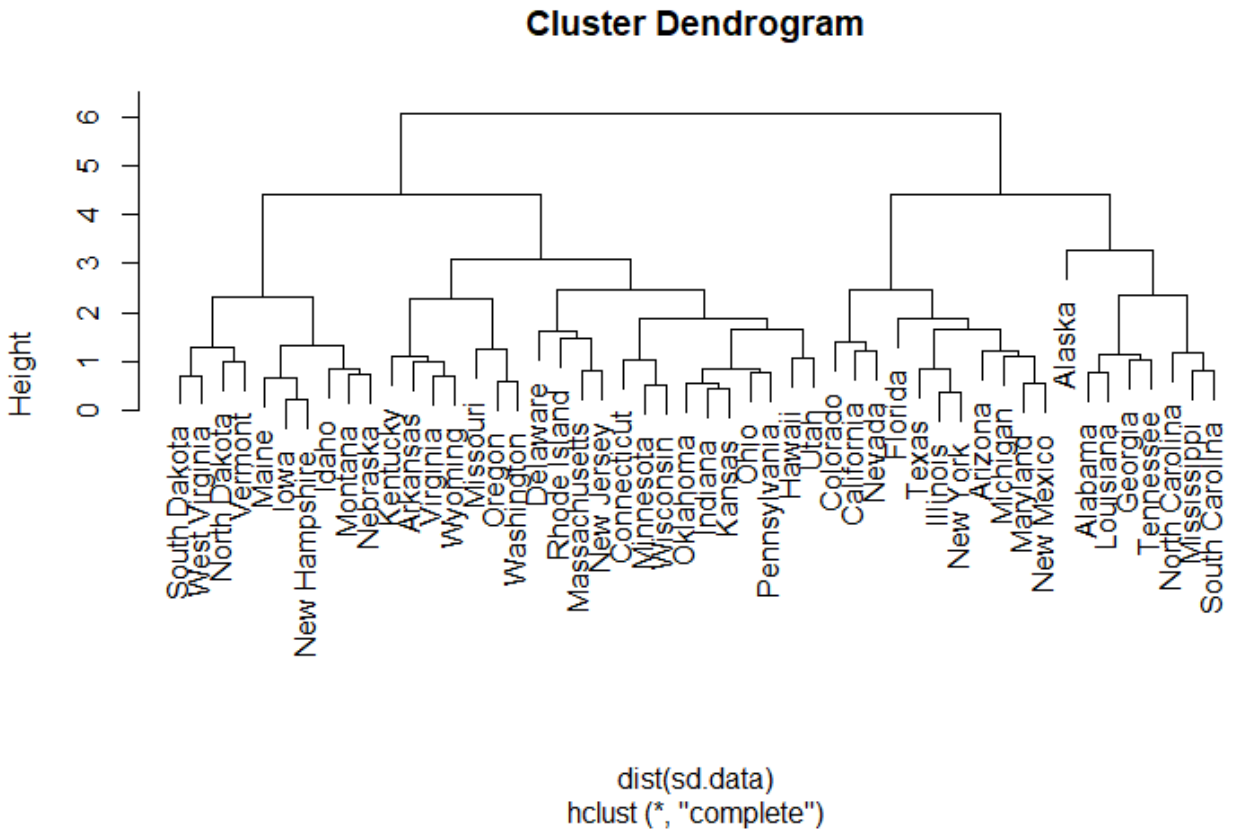
States in Cluster – 2

```
 cluster_df[which(cluster_df$clusters == 2),]$cities
[1] Arkansas       Colorado       Georgia        Massachusetts Missouri       New Jersey     Oklahoma
[8] Oregon         Rhode Island   Tennessee      Texas         Virginia       Washington     Wyoming
```

States in Cluster – 3

```
 cluster_df[which(cluster_df$clusters == 3),]$cities
[1] Connecticut    Hawaii         Idaho          Indiana        Iowa           Kansas         Kentucky
[8] Maine          Minnesota      Montana        Nebraska       New Hampshire North Dakota   Ohio
15] Pennsylvania   South Dakota   Utah           Vermont        West Virginia  Wisconsin
```

c) We obtain the following hierarchical cluster structure by using complete linkage method and Euclidean distance after the features are normalized.



d) Cutting the above hierarchical cluster structure to obtain 3 clusters overall using the cutree() method results in the following cluster distribution.
From the first look at the above plot we can see that we do not clearly have a 3 cluster cut. We have a good 2 or 4 cluster cut.

Cluster – 1

```
> new_cluster_df[which(new_cluster_df$newclusters == 1),]$cities
[1] Alabama        Alaska         Georgia        Louisiana      Mississippi    North Carolina South Carolina
[8] Tennessee
```

Cluster – 2

```
> new_cluster_df[which(new_cluster_df$newclusters == 2),]$cities
 [1] Arizona    California Colorado   Florida    Illinois   Maryland   Michigan   Nevada     New Mexico
[10] New York   Texas
```

Cluster – 3

```
> new_cluster_df[which(new_cluster_df$newclusters == 3),]$cities
 [1] Arkansas      Connecticut   Delaware      Hawaii        Idaho         Indiana       Iowa
 [8] Kansas        Kentucky      Maine         Massachusetts Minnesota     Missouri      Montana
[15] Nebraska      New Hampshire New Jersey    North Dakota  Ohio          Oklahoma      Oregon
[22] Pennsylvania  Rhode Island  South Dakota  Utah          Vermont       Virginia      Washington
[29] West Virginia Wisconsin     Wyoming
```

Observation: We can clearly see that the number of states in cluster – 3 significantly higher than other clusters.

```
> table(clusters, newclusters)
        newclusters
clusters  1  2  3
       1  6  9  1
       2  2  2 10
       3  0  0 20
```

In the above picture, clusters signify the clusters when hierarchical clustering is used and newclusters signify the cluster when hierarchical clustering is used after scaling the features.

We can clearly see a big variation in the cluster assignment.

Irrespective of the type of distance metric used, it is important to scale the features first since if one of the feature range is much greater than others, then it's impact on the overall distance is big. Hence dominating the cluster assignment overall.

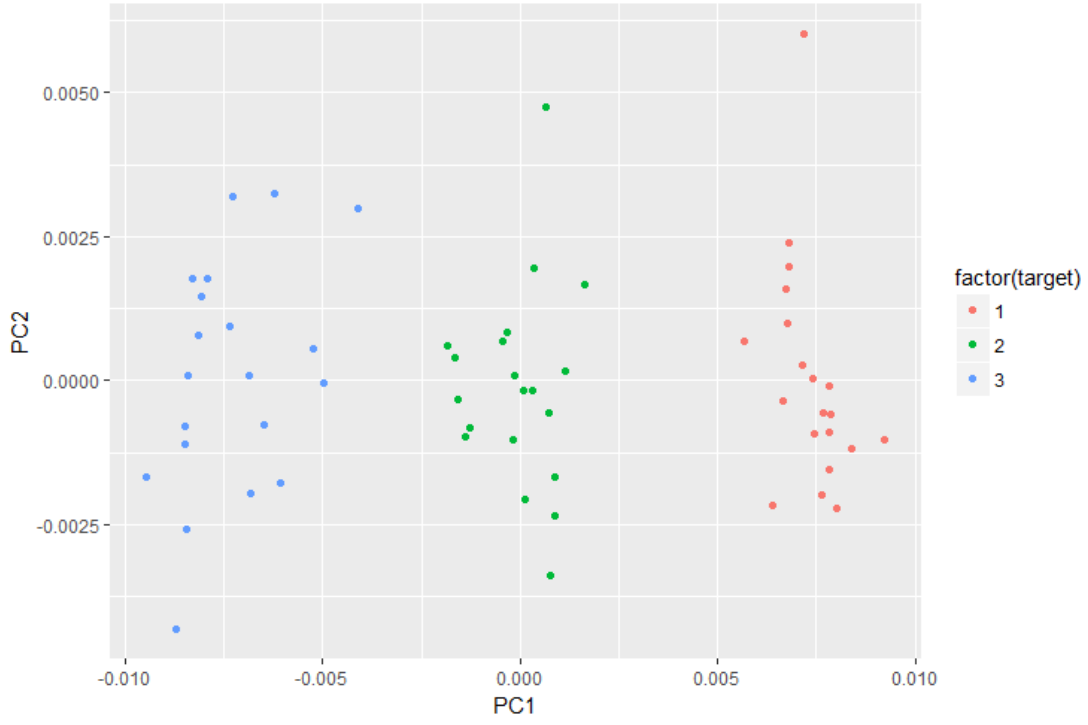
Question – 10

a)

Creating a random dataset using the normalized distribution function – rnorm with different mean so that there is a good separation between the data points.

b)

After doing a PCA and plotting the first two principal components by target value we get the following figure.



c)
Performing the K-Means clustering on the dataset with k – 3 and comparing with the original classes gives the following.

```
> table(data$target, kmeans_obj$cluster)

    1  2  3
1  20  0  0
2   0  0 20
3   0 20  0
```

Inference : It can be clearly seen that there is a separation for class – 1 observations and class – 2 and class – 3 are interchanged.

Following has to be noted that, ***cluster name/number assignment in un-supervised learning context is not possible – That is in the above table, cluster – 2 can be called as cluster – 3 and vice versa.***
If this is the case, then we can see that cluster – 2 has all class – 2 observations and cluster – 3 has all class – 3 observations.

For further analysis and inference, the above point will be considered from now on.

d)
Performing the K-Means clustering on the dataset with k – 2 and comparing with the original classes gives the following.

```
> table(data$target, kmeans_obj$cluster)

    1  2
1   0 20
2  20  0
3  20  0
```

Inference: One cluster has one class' observations and the other cluster has the observations from the remaining two classes.

e)
Performing the K-Means clustering on the dataset with k – 4 and comparing with the original classes gives the following.

```
> table(data$target, kmeans_obj$cluster)

    1  2  3  4
1   0 20  0  0
2   0  0 20  0
3  10  0  0 10
```

Inference: Two of the each clusters has unique class' observations and the other two clusters has half of the remaining class in them.

f)
Performing the K-Means clustering on the principal component data with k – 3 and comparing with the original classes gives the following.

```
> table(data$target, kmeans_obj$cluster)

    1  2  3
1   0  0 20
2  20  0  0
3   0 20  0
```

Inference: Each cluster has observations from unique class.

g)
Performing the K-Means clustering on the scaled data with k – 3 and comparing with the original classes gives the following.

```
> table(data$target, kmeans_obj$cluster)

    1  2  3
1   0 20  0
2  20  0  0
3   0  0 20
```

Inference: Each cluster has observations from unique class.
When these results are compared with the clustering results that we obtained in b) we do not see any difference.

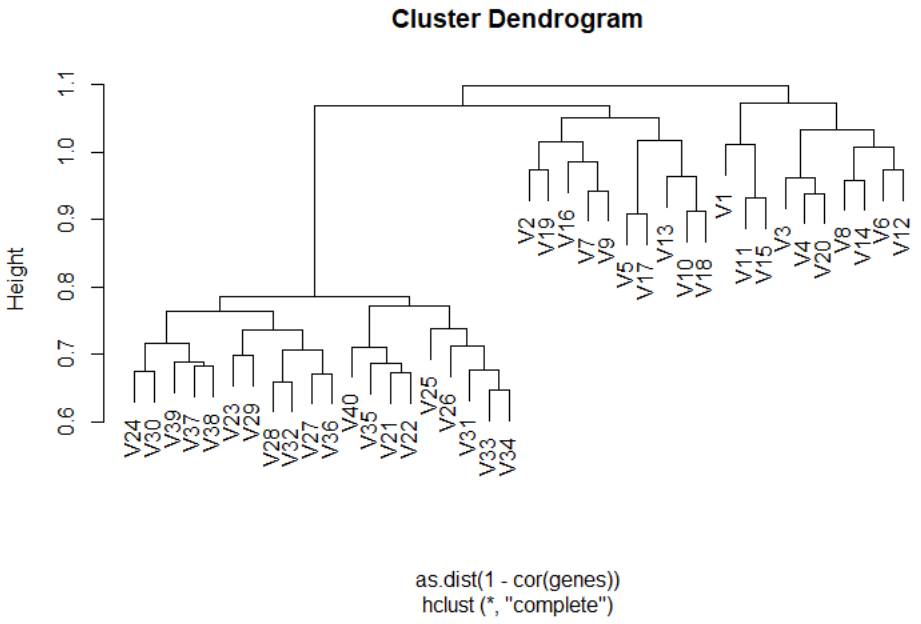
Question – 11

a)

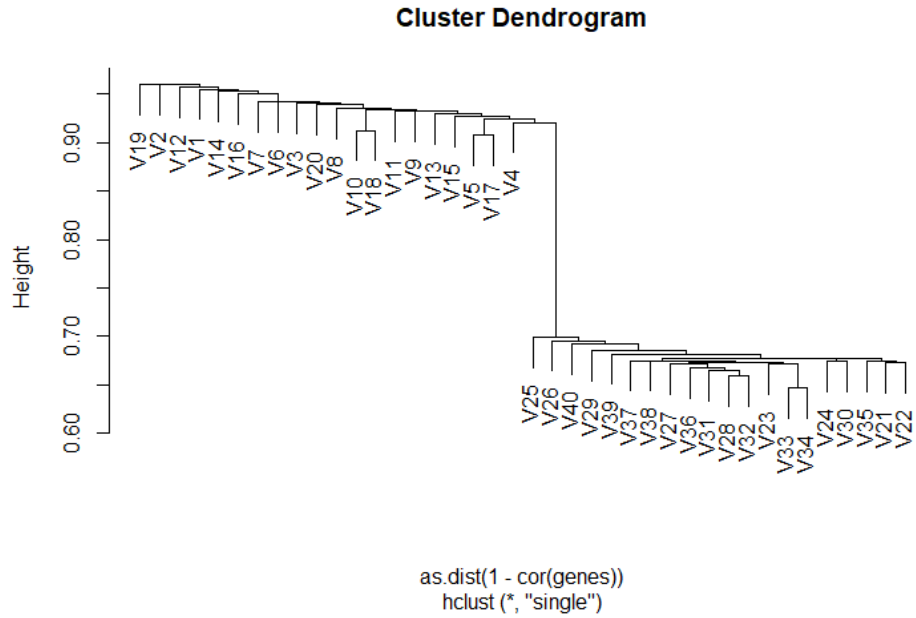
Loading the data using specified method and arguments.

b)

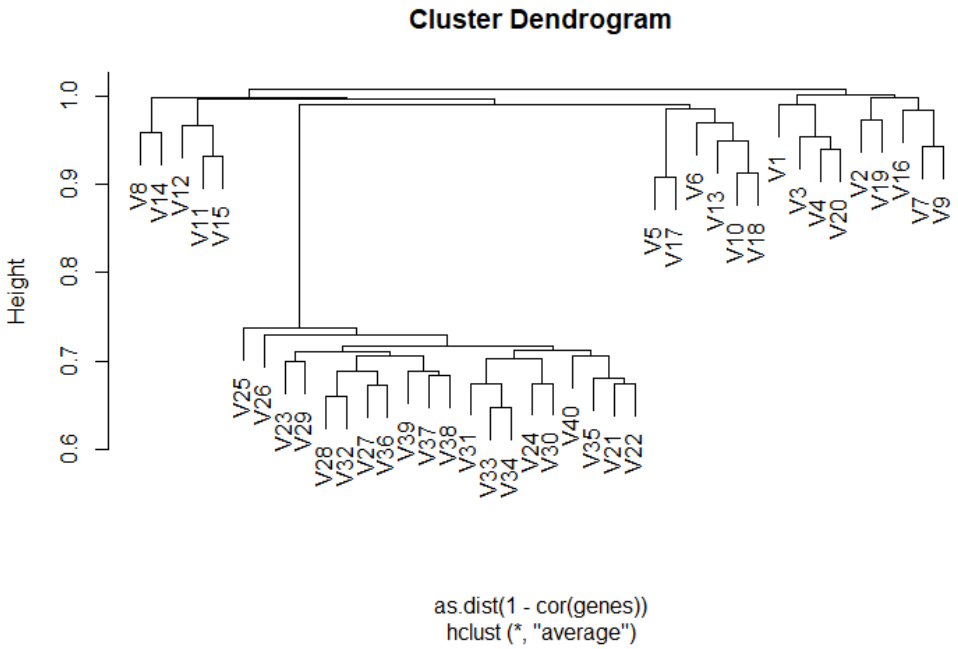
Creating a hierarchical clustering using correlation based distance and complete linkage gives the following plot.



Creating a hierarchical clustering using correlation based distance and complete linkage gives the following plot.



Creating a hierarchical clustering using correlation based distance and average linkage gives the following plot.



Overall Inference: We get different structures when different linkages are used. We obtain two clusters when we use complete linkage, single linkage and three clusters when we use average linkage.

c)
We can use PCA to see which genes are different. Examining the absolute values of the total loading for each gene.
We get the following as the 10 most different genes.

```
> index[1:10]
 [1] 865  68 911 428 624  11 524 803 980 822
```