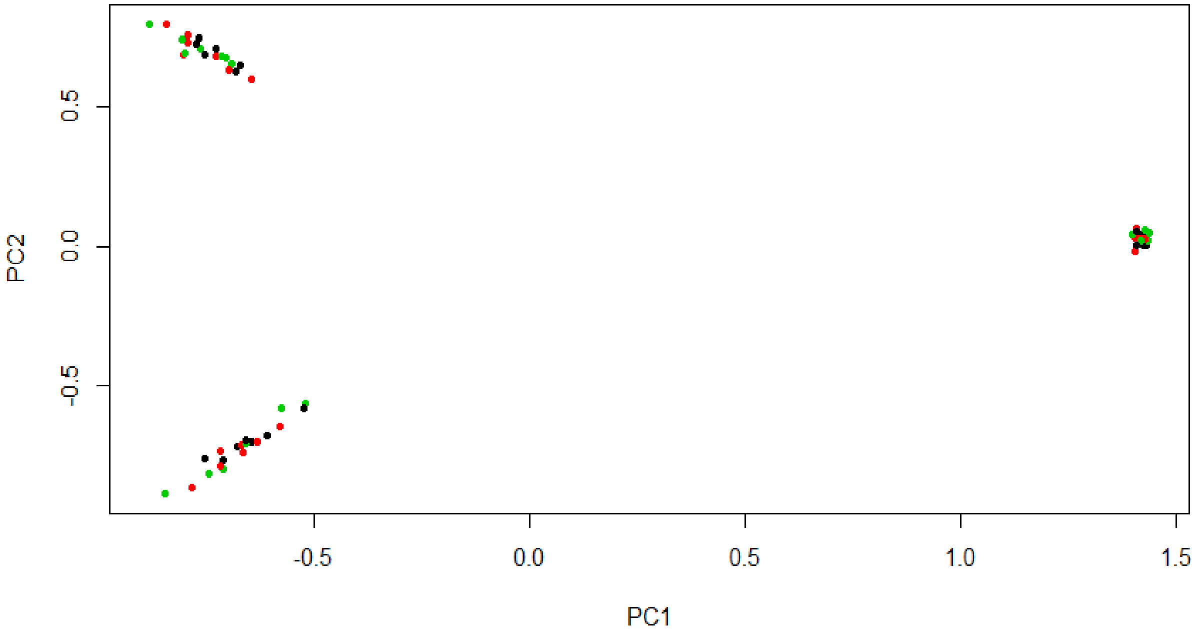
2.)

a.) Generates a simulated data set with 20 observations in each of three classes, and 50 variables. Created a random dataset using the normalized distribution function

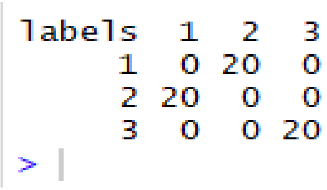
b.)

After doing PCA and plotting the first two principal components.



c.) Performing k-means clustering on the dataset with k=3 and comparing with original classes.

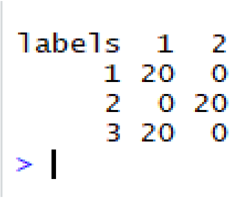
Clustered into 3 classes



```
labels  1  2  3
     1  0 20  0
     2 20  0  0
     3  0  0 20
> 
```

d.) Performing k-means clustering on the dataset with k=2 and comparing with original classes.

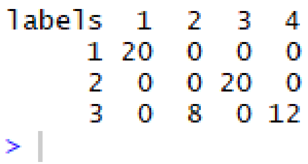
original 3 classes are now clustered into 2 only



```
labels  1  2
     1 20  0
     2  0 20
     3 20  0
>
```

e.) Performing k-means clustering on the dataset with k=4 and comparing with original classes.

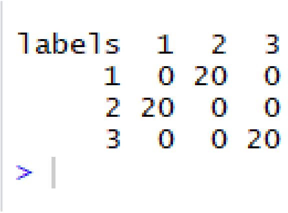
original 3 classes are now classified into 4 clusters



```
labels  1  2  3  4
     1 20  0  0  0
     2  0  0 20  0
     3  0  8  0 12
>
```

f.) Performing k-means clustering on the PCA data with k=3 and comparing with original classes.

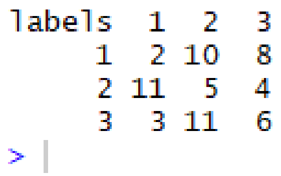
Now observations are perfectly clustered once again



```
labels  1  2  3
     1  0 20  0
     2 20  0  0
     3  0  0 20
>
```

g.) Performing k-means clustering on the Scaled data with k=3 and comparing with original classes.

Observations are not perfectly clustered and the results are worse than unscaled clustering. Scaling affects the distance between the observations.



```
labels  1  2  3
     1  2 10  8
     2 11  5  4
     3  3 11  6
>
```