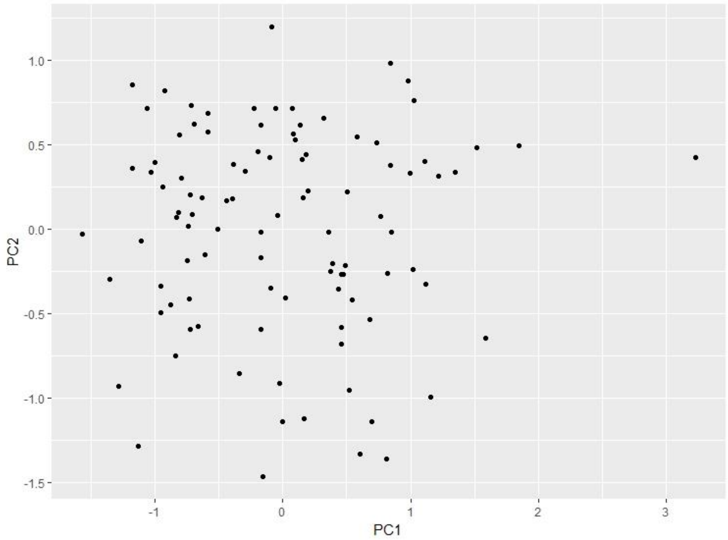
STA Home Work-3

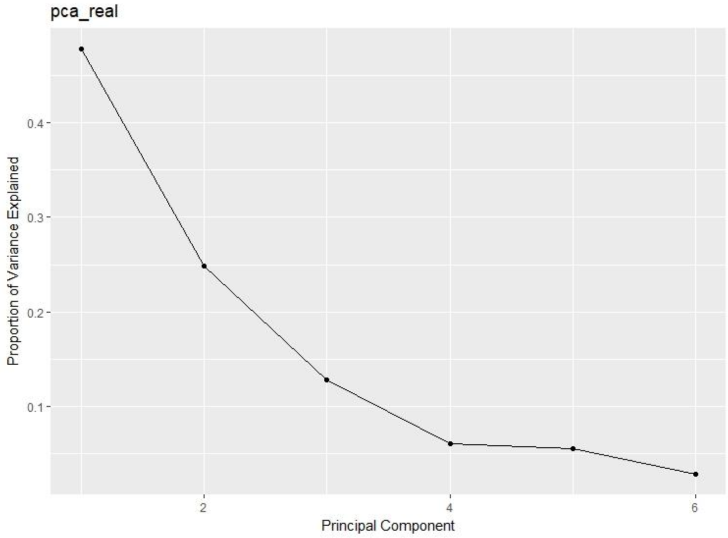
1) Perform the PCA on the real notes, fake notes and the original data which has both real and fake notes.

Genuine Notes:

After performing the PCA on the Genuine notes data, these are the graphs obtained.



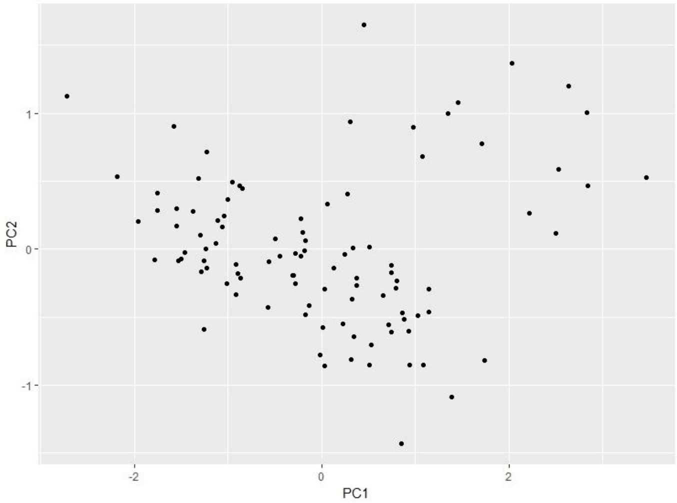
Inference: The above plot showing all principal components of PC1 and PC2.



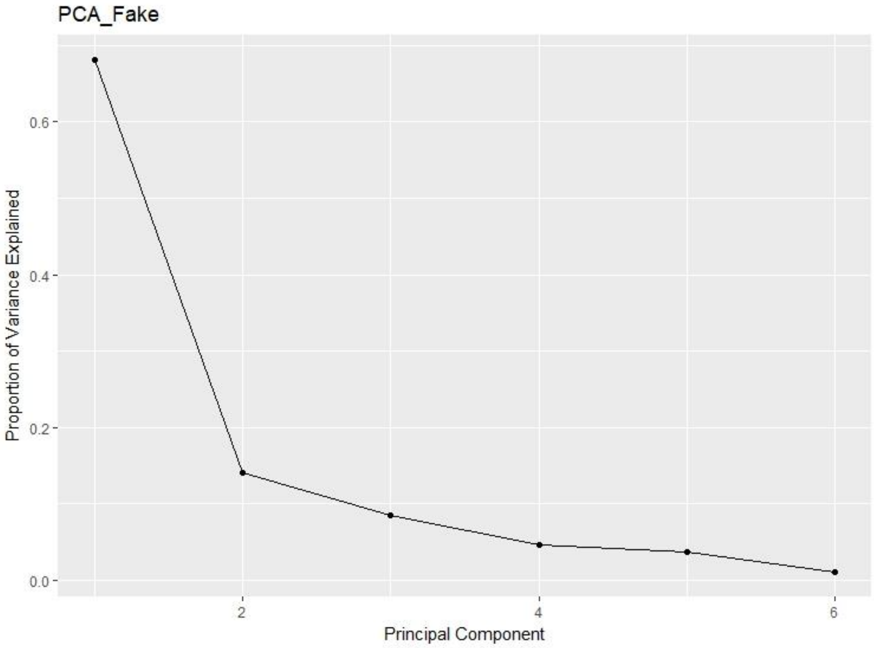
Inference: The above plot is obtained after performing PCA on the genuine dataset. The first principal component explains 57% of variance and second principal component explains 25% of variance.

Counterfeit Notes:

After performing PCA on the fake notes data, gives the following graphs.



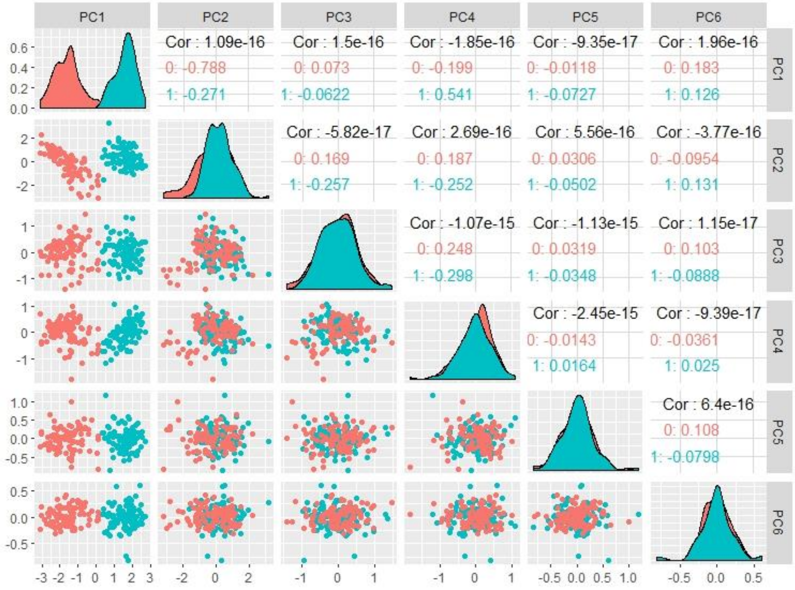
Inference: The above plot represents the principal components for counterfeit notes. As we can see that there is a difference between above genuine principal components plot and this plot. By this we can say that there is the difference between the data which implies real and fake notes.



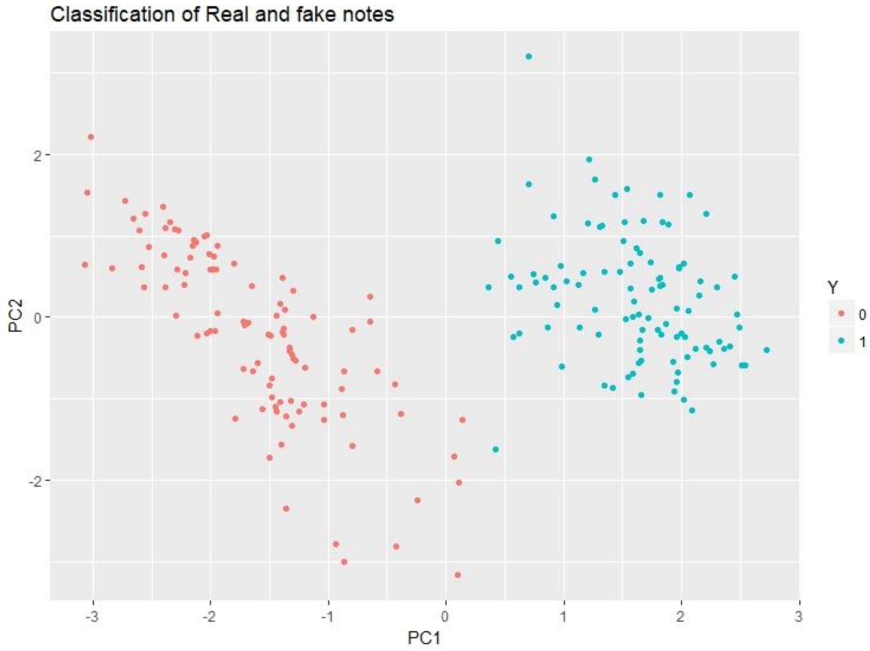
Inference: The above plot obtained for performing PCA on the counterfeit notes and we can see that variance has dropped from PC1 to PC2 i.e 67% to15% which is very different compared to the variances obtained for the genuine notes plot.

Combined data:

Performed PCA on the combined data and the below are the plots obtained.



Inference: For the combined data, we can see that PC1 clearly shows the difference the two classes which clearly shows two different types of notes.



Inference: There is a clear separation between the classes. Red represents genuine notes and green represents counterfeit notes. The principal components in the above figure is obtained from the features (height.left, height.right) and diagonal.

2)

Hierarchical Clustering:

a) Load the primate scapulae data using load function and before applying hierarchical clustering on the data, remove 10th and 11th columns (predictors) in order to remove non-numerical data and get only numerical data.

b) Apply hierarchical clustering on new data using Complete, Average, Single linkage.

c) As there are categorical columns in the given data, using those we got five classes, so five groupings have been considered.

```
> rate_com
[1] 0.4761905
> rate_sin
[1] 0.2
> rate_ave
[1] 0.4761905
```

According to the obtained misclassification rates, single linkage has very less misclassification error and performance for complete and average linkage is less. So for this dataset, hierarchical clustering using single linkage the better result.

We can't expect the result because the choice of linkage is always dependent on the dataset.

K-Means Clustering:

a) After checking the data, as there are many NA values, replaced NA values with median of the column data.

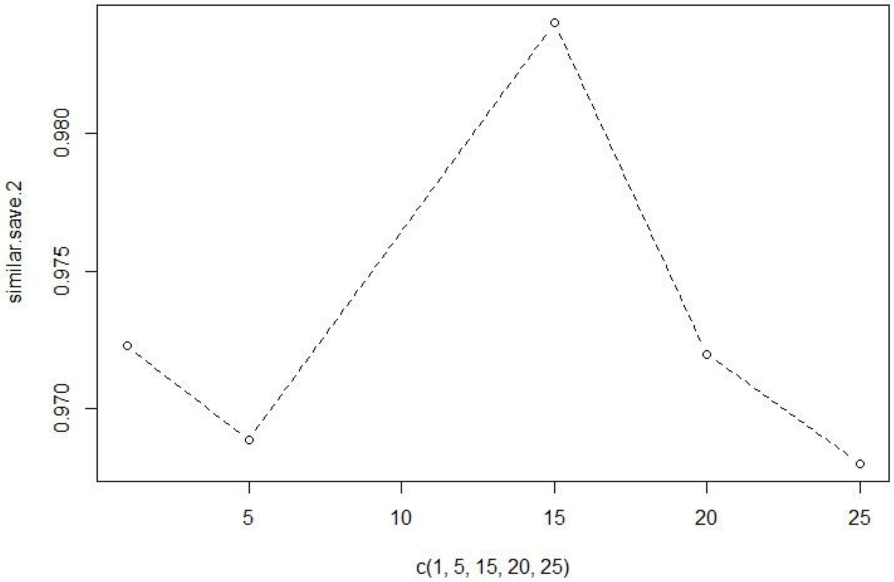
b) Considered groupings as five because there are five classes and misclassification rate obtained after performing k-means clustering is `0.6380952.`

c) Performance of K-means clustering is better than the performance of hierarchical clustering.

3)

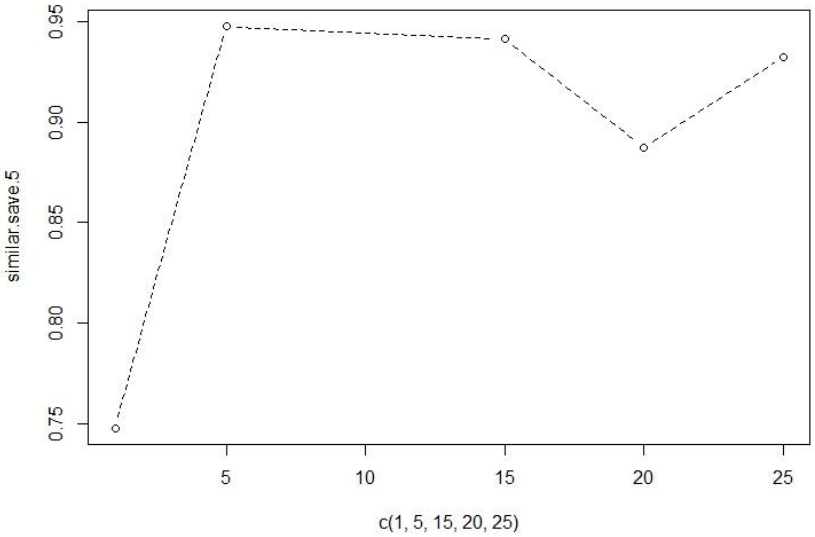
Fitting the dataset with k=2 and iterating over different values (1,5,15,20,25) of radius for fitting SOM's, we get the following results.

The similarity between K-means and self-organizing map are 0.9722798, 0.9688703, 0.9840219 0.9719952, 0.9680199



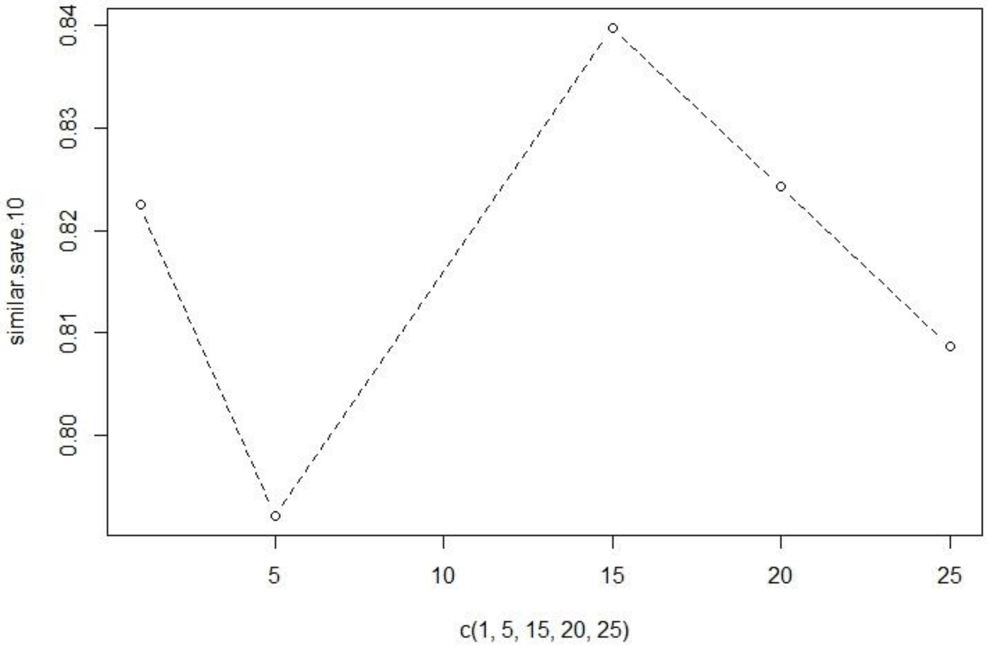
For k=5

The similarity between K-means and self-organizing map are 0.7477794, 0.9477588, 0.9413405 0.8874391, 0.9326391



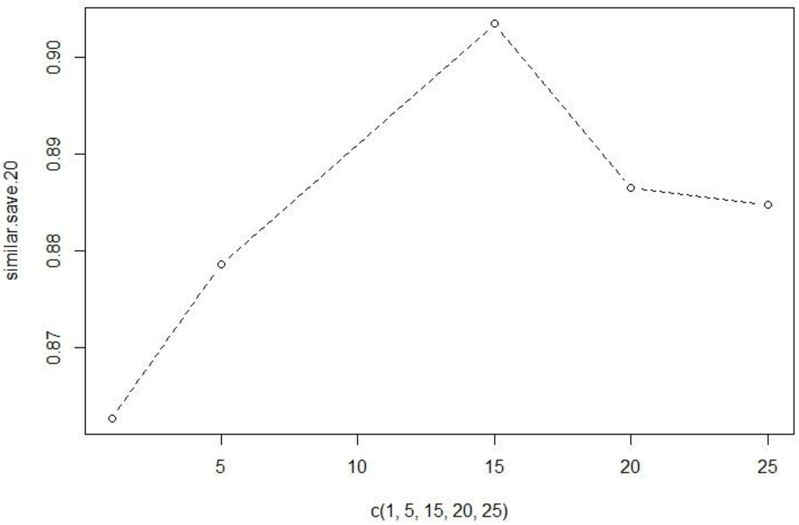
For k=10

The similarity between K-means and self-organizing map are 0.8225253,0.7921999,0.8397527, 0.8242807, 0.8086616



For k=20

The similarity between K-means and self-organizing map are 0.8626662, 0.8785215, 0.9034115, 0.8864460, 0.8847249.



By looking at all the above results, we can conclude that SOM solution becomes more similar to K-means solution when SOM radius is small and when k equals 20, the assumption is not confirmed but for other values of k, the assumption is confirmed.