PCA

Example-wine dataset

Summary of data ➔

	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols
Min.	11.03	0.74	1.36	10.6	70	0.98
1st Qu.	12.36	1.603	2.21	17.2	88	1.742
Median	13.05	1.865	2.36	19.5	98	2.355
Mean	13.00	2.336	2.367	19.49	99.74	2.295
3rd Qu.	13.68	3.083	2.558	21.5	107	2.8
Max.	14.83	5.80	3.230	30.00	162	3.88

	Flavanoids	Nonflavanoids	Proanthocyanins	Color	Hue	Dilution	Proline
Min.	0.34	0.13	0.41	1.28	0.48	1.27	278
1st Qu.	1.205	0.27	1.25	3.22	0.7825	1.938	500.5
Median	2.135	0.34	1.555	4.69	0.965	2.78	673.5
Mean	2.029	0.3619	1.591	5.058	0.9574	2.612	746.9
3rd Qu.	2.875	0.4375	1.95	6.2	1.12	3.17	985
Max.	5.08	0.66	3.58	13	1.71	4	1680

Clustering before normalization of data ➔

```
1  2  3
59 71 48
```

Hierarchical Clustering after normalization of data ➔

```
method  v2 v3 v4
1   single  59 71 48
2 complete  76 54 48
3  average  59 71 48
4 mcquitty  59 71 48
5   ward.D  59 71 48
6  ward.D2  59 71 48
7 centroid 129  1 48
8   median 129  1 48
```

From the above data we can say that single, average, mcquitty, ward.d, ward.D2 seems good enough for clustering.

Hierarchical clustering after performing PCA ➔

```
method   V2 V3 V4
1   single 174  3  1
2 complete 106 22 50
3  average 125  1 52
4 mcquitty 174  3  1
5   ward.D  65 65 48
6  ward.D2  65 66 47
7 centroid 176  1  1
8   median 174  3  1
```

From the above information we can infer that ward.D and ward.D2 are performing good for my clustering model.

Accuracy of model with PCA and without PCA ➔

Cluster allocation after PCA (on row) v/s before PCA (on column)

For ward.D2

```
  1  2  3
  1 59  6  0
  2  0 64  2
  3  0  1 46
```

Accuracy ➔0.949

Mis-classified ➔67,70,74,79,84,96,122,131,135

For ward.D

```
     1  2  3
  1 59  6  0
  2  0 64  1
  3  0  1 47
```

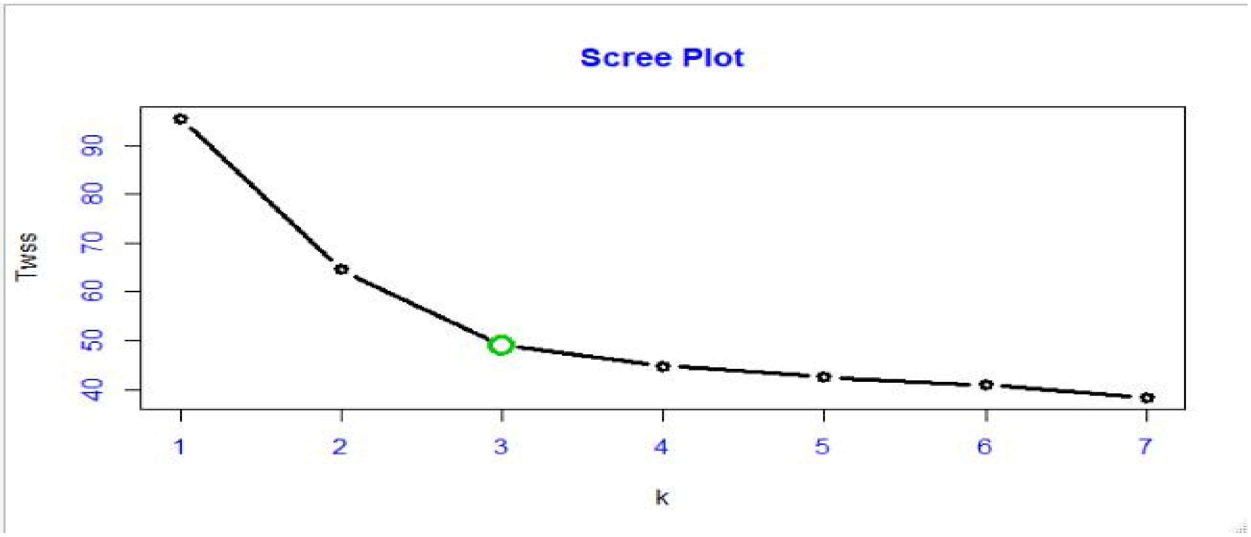
Accuracy ➔0.955

Mis-classified ➔67,70,74,79,84,96,122,131

Now we are doing classification here in a unsupervised learning to calculate whether after performing PCA we are getting the same groups of cluster as before PCA or not.

But here class number is not relevant for our classification, we are going to see just whether these are same cluster or not after PCA.

K-means clustering after normalization of data



After performing k-means clustering with k=3, we are getting below cluster size as,

```
1  2  3
51 65 62
```

Clusters are distributed over the three groups.

Comparison of Hierarchical and k-means

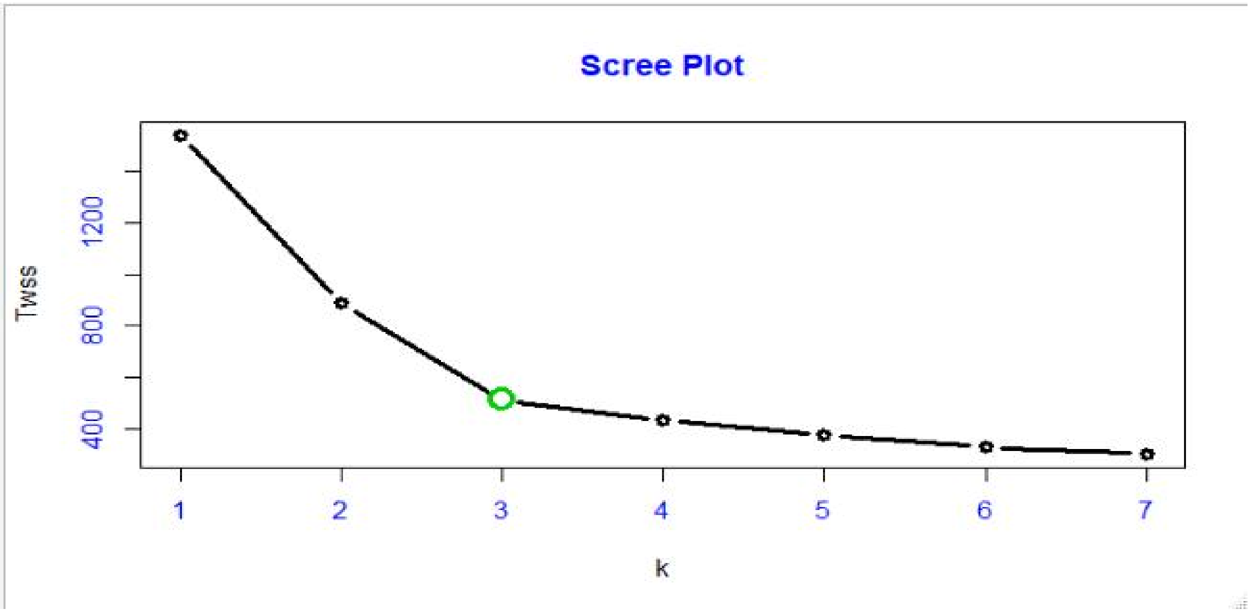
```
                 HierarchicalGroup
KmeansClusterGroup  1  2  3
                 1  0  3 48
                 2 59  6  0
                 3  0 62  0
```

From the above comparison, maximum number of group-2 k-means clusters are same in group-1 Hierarchical cluster.

After excepting some observations form k-means and encoding we can get following groups.

```
            Hierarchical_groups
Kmeans_Groups  1  2  3
            1 59  6  0
            2  0 62  0
            3  0  3 48
```

K-means clustering on PCA data ➔



From the above scree plot optimum cluster for k=3 and comparison with all the other clusters.

```
         KmeansOriginal
KmeansPCA  1  2  3
        1 62  1  0
        2  3 61  0
        3  0  0 51
```

From the above information we are losing our 4% of our information, after considering the PCA.

Comparison with each and every method of clustering

Clustering Method in comparison	Proportion of getting same cluster
K-means v/s Hierarchical	0.9494382
PCA_Kmeans v/s Hierarchical	0.96067
PCA_kmeans v/s PCA_Hierarchical	0.96067
PCA_kmeans v/s kmeans	0.9775281
Hierarchical v/s PCA-Hierarchical	0.9494382