PCA

QUESTION:

PERFORM PRINCIPAL COMPONENT ANALYSIS AND PERFORM CLUSTERING USING FIRST 3 PRINCIPAL COMPONENT SCORES (BOTH HIERARCHICAL AND K MEAN CLUSTERING(SCREE PLOT OR ELBOW CURVE) AND OBTAIN OPTIMUM NUMBER OF CLUSTERS AND CHECK WHETHER WE HAVE OBTAINED SAME NUMBER OF CLUSTERS WITH THE ORIGINAL DATA (CLASS COLUMN WE HAVE IGNORED AT THE BEGINNING WHO SHOWS IT HAS 3 CLUSTERS)

SUMMARY OF THE DATA:

	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols	Flavanoids	Nonflavanoids	Proanthocyanins	Color	Hue	Dilution	Proline
Min. :	11.03	0.740	1.360	10.60	70.00	0.980	0.340	0.1300	0.410	1.280	0.4800	1.270	278.0
1st Qu.:1.	12.36	1.603	2.210	17.20	88.00	1.742	1.205	0.2700	1.250	3.220	0.7825	1.938	500.5
Median :	13.05	1.865	2.360	19.50	98.00	2.355	2.135	0.3400	1.555	4.690	0.9650	2.780	673.5
Mean :	13.00	2.336	2.367	19.49	99.74	2.295	2.029	0.3619	1.591	5.058	0.9574	2.612	746.9
3rd Qu.:	13.68	3.083	2.558	21.50	107.00	2.800	2.875	0.4375	1.950	6.200	1.1200	3.170	985.0
Max. :	14.83	5.800	3.230	30.00	162.00	3.880	5.080	0.6600	3.580	13.000	1.7100	4.000	1680.0

The Initial Clustering was

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

PERFORM HIREARCHICAL CLUSTERING:

HIERARCHICAL CLUSTERING AFTER NORMALIZATION ON ORIGINAL DATA

I performed Hierarchical Clustering in my Original Data (after normalization) with all possible methods, Here I come up with the result below, Each methods and weight of all possible 3 clusters.

```
    method  C1 C2 C3
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
```

C1 C2 C3 are my 3 clusters with size of the corresponding cluster groups. From the above Table we can see that The methods “single”, “average”, “mcquitty”, “ward.D”, “ward. D2” seems to be good enough for clustering purpose. I may Consider my Ward.D method for my clustering purpose.

HIERARCHICAL CLUSTERING AFTER PERFORMING PCA TO THE DATA

After Performing PCA and considering only 3 components of the Principal components, I come up with the all clusters i.e.

```
   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 this Table I can say that Ward.D and Ward.D2 is still performing fantastic for my clustering model. Here again I am considering my Ward.D linkage method for my final model.

ACCURACY OF MY MODEL (WITH PCA V/S WITHOUT PCA)

Here all the table convey the cluster allocation After PCA (on row) v/s Before PCA (on column)

In case of Word.D we come up with 0.955 accuracy

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

Records Miss classified  67  70  74  79  84  96 122 131
```

In case of Word.D2 we come up with 0.949 accuracy.

```
   1  2  3
 1 59  6  0
 2  0 64  1
 3  0  1 47
 Records Miss classified 67  70  74  79  84  96 122 131 135
```

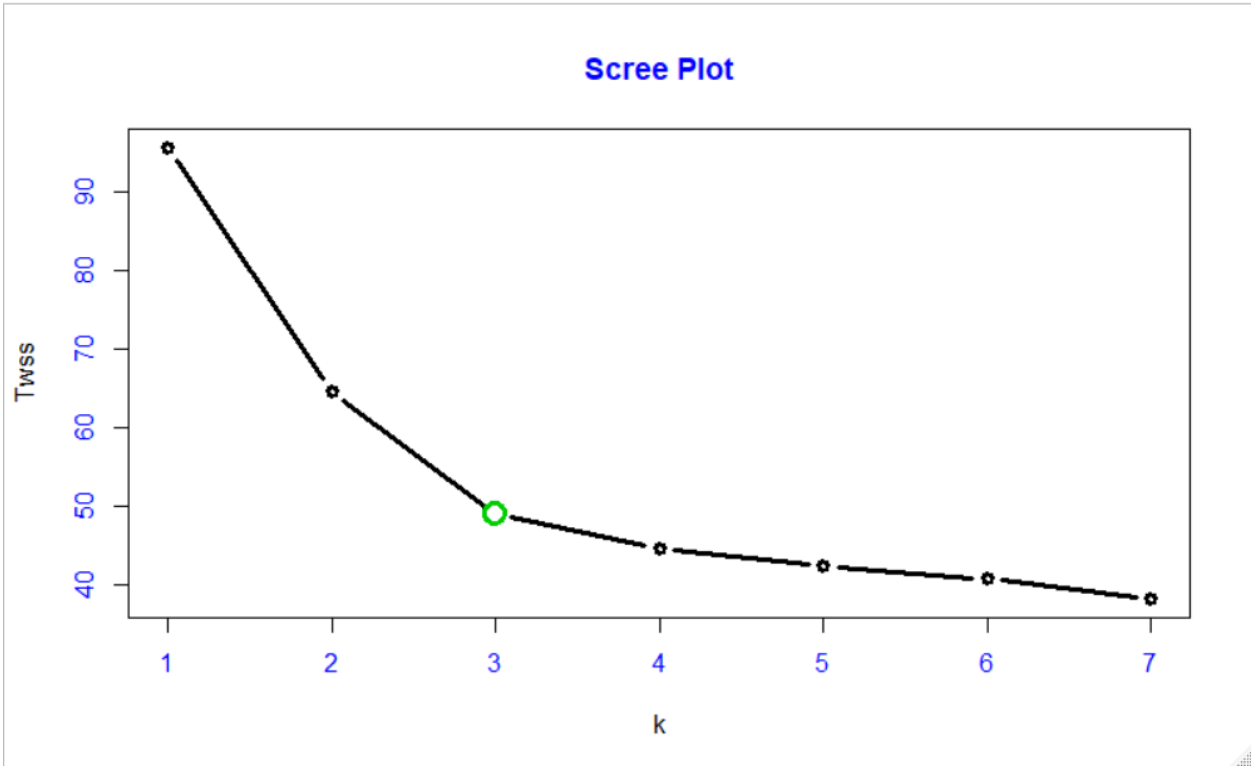
NB:

I may be making a little bit confusion here that why I am doing the classification thing here in a unsupervised learning model. Let me clear that, here we are calculating whether After performing PCA we are getting the same Clusters groups as before PCA or not. That’s why I may call it as a classification problem, and assuming that cluster groups are classes of the groups which we consider before the PCA, and doing clustering in PCA model and getting result from it is clusters, which are again classes. But here the class number is not relevant for our classification, our moto is to just see whether these are in same clusters or not after PCA.

K-MEANS CLUSTERING

K-MEANS CLUSTERING ON THE ORIGINAL DATA AFTER NORMALIZATION

Performing the Scree plot to determine the Best K value, based on the elbow point.



From the Scree plot, I can say that, Its better to go for my k i.e. optimum number of cluster as 3.

Performing the K-Means clustering with k = 3 we come up with the cluster sizes as

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

It seems like The clusters are well distributed over the three groups.

COMPARE WITH HIERARCHICAL V/S K-MEANS

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

Here we can see that the maximum member of our group-2 K-means cluster are same as in group-1 Hierarchical cluster.

length(union(which(km_o$cluster==2),which(df_norm$Type==1)))-
length(intersect(which(km_o$cluster==2),which(df_norm$Type==1)))

Except 6 observations from Kmeans group 2 all same as the cluster group 1 in Hierarchical clustering We may consider Kmeans cluster of group 2 as our group 1 cluster in hierarchical clustering.

length(union(which(km_o$cluster==1),which(df_norm$Type==3)))-
length(intersect(which(km_o$cluster==1),which(df_norm$Type==3)))

Except 3 observations from Kmeans group 1 all same as the cluster group 3 in Hierarchical clustering We may consider Kmeans cluster of group 1 as our group 3 cluster in hierarchical clustering.

length(union(which(km_o$cluster==3),which(df_norm$Type==2)))-
length(intersect(which(km_o$cluster==3),which(df_norm$Type==2)))

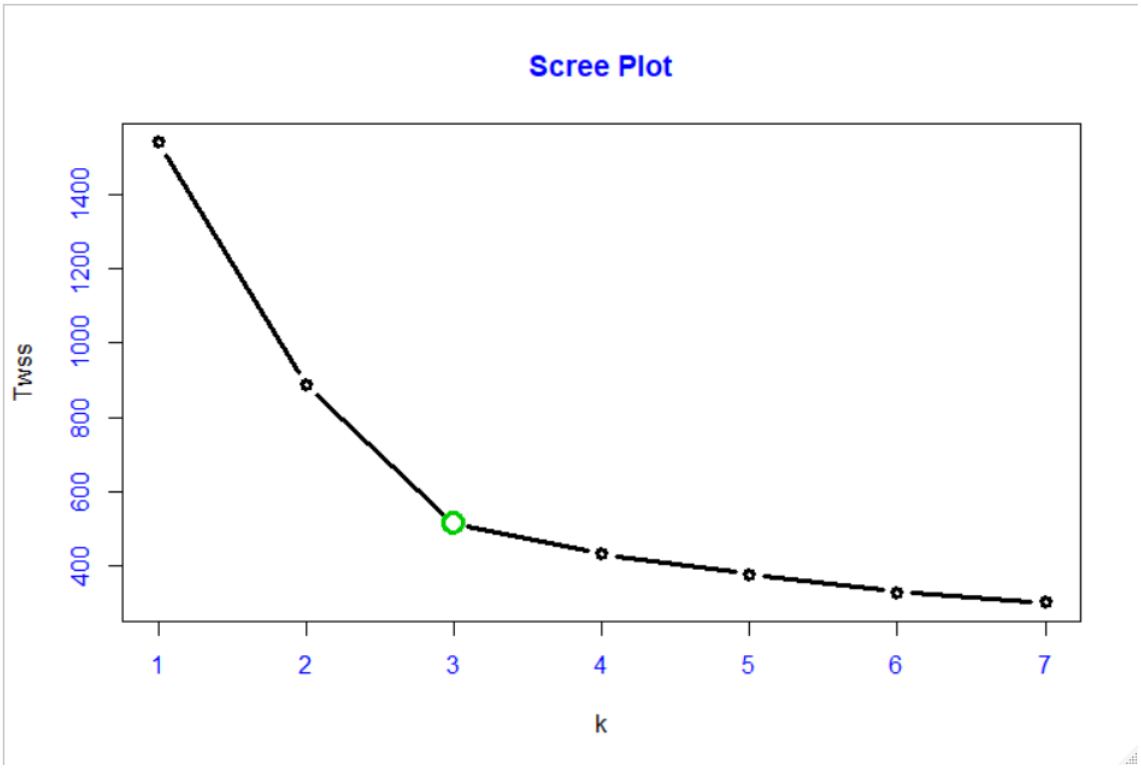
Except 9 observations from Kmeans group 3 all same as the cluster group 2 in Hierarchical clustering We may consider Kmeans cluster of group 3 as our group 2 cluster in hierarchical clustering.

After encoding this we can see that

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

Very less number of K-Means differ from Hierarchical cluster groups.

K-MEANS CLUSTERING ON THE PCA DATA



After performing The Scree plot in the PCA data, we come up with the same optimum cluster for k =3., And encoding for comparison with all the other clusters we come up with the table given below that.

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

Here we can see that we are losing our 0.039 i.e. 4% of our information, after considering the PCA.

COMPARISON WITH EACH AND EVERY METHOD OF CLUSTERING. CONCLUSION

Clustering Methods in Comparison	Proportions of getting Same kind cluster
KMeans V/S Hierarchical	0.9494382
PCA_KMeans V/S Hierarchical	0.96067
PCA_KMeans V/S PCA_Hierarchical	0.96067 (Same as above)
PCA_KMeans V/S KMeans	0.9775281
Hierarchical V/S PCA_Hierarchical	0.9494382

For Particular this data set we can say that, overall performance of clustering is same for considering 3 PCA and also for the Original data, as we are Losing less than 4% of our Information of the data set after clustering.

So we may consider to go for clustering where we need to be quick responsive to a project, But the project must not from a pharmaceutical company, as we cant trade our accuracy in such fields.

Sorry for not creating any clustering graph here, as Here our only focus is on PCA.