Kyle Beitz
STAT 4300

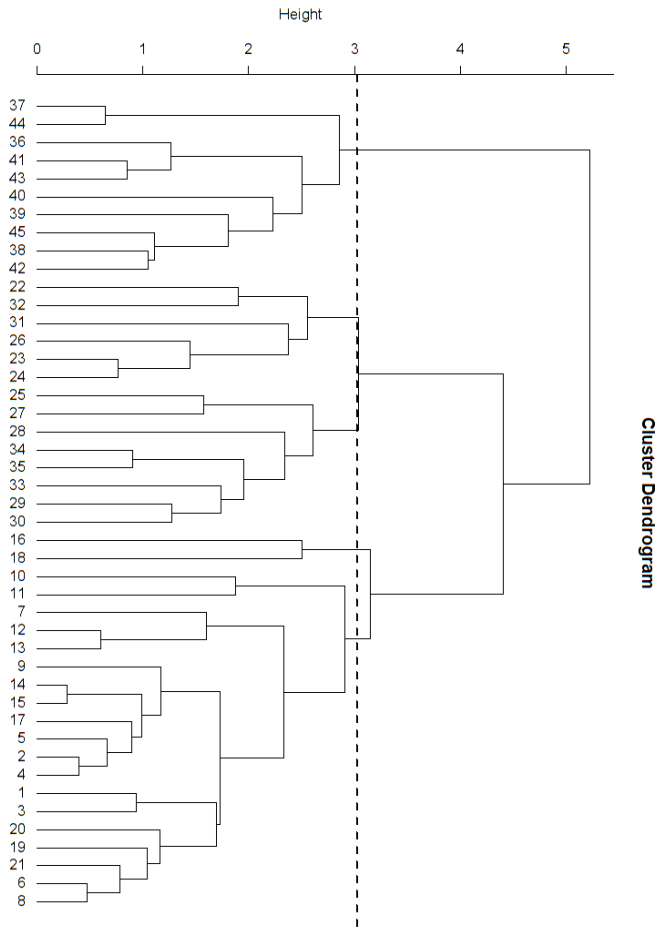
Homework 4

Confusion matrix:

```
                G.class (cv)
RF    Kiln 1  Kiln 2  Kiln 4  Kiln 5
  Kiln 1: 21       0       0       0
  Kiln 2:  0      12       0       0
  Kiln 3:  0       2       0       0
  Kiln 4:  0       0       1       4
  Kiln 5:  0       0       4       1
          correct.rate   error.rate
             0.8222222    0.1777778
```

```
                G.class (cv)
SVM   Kiln 1  Kiln 2  Kiln 4  Kiln 5
  Kiln 1: 21       0       0       0
  Kiln 2:  0      12       0       0
  Kiln 3:  0       2       0       0
  Kiln 4:  0       0       0       5
  Kiln 5:  0       0       5       0
          correct.rate   error.rate
             0.8444444    0.1555556
```

- H.4.1:
  - (a):
    - Please see the confusion matrices above.
  - (b):
    - From the confusion matrices above we can see that the Support Vector Machine (SVM) algorithm performed best versus the Random Forest (RF) algorithm in predicting the kilns. The SVM model had a correct rate of 84% where the RF model had a correct rate of 82%. The SVM model also had a better error rate at 15% where the RF model had an error rate of 17%. From the confusion matrices we can also determine that Kiln 3 cannot be well classified in both models and both confusions matrices predicted Kiln 3 to be the same as Kiln 2.



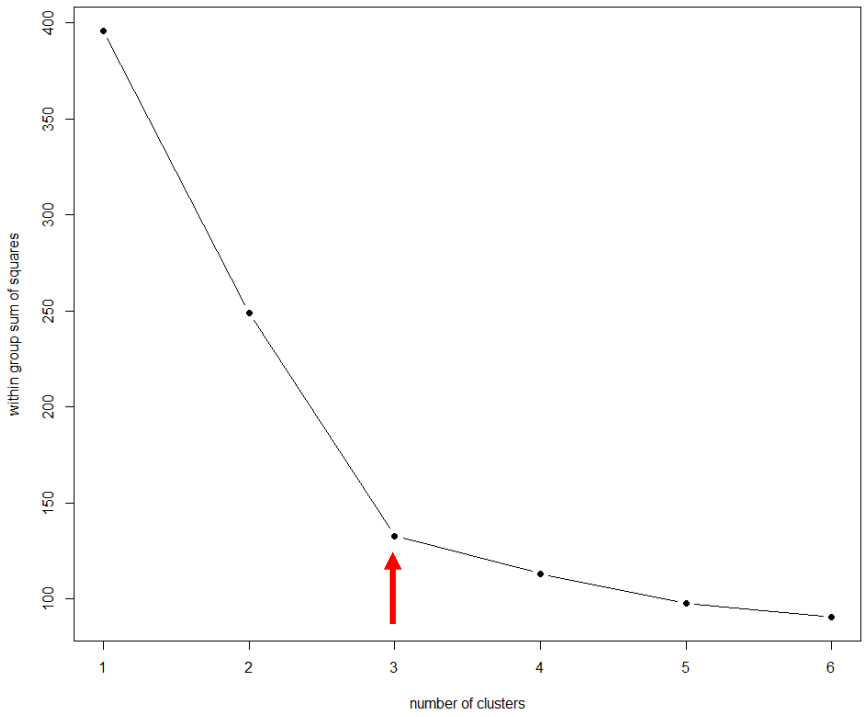
kc$cluster

[1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3

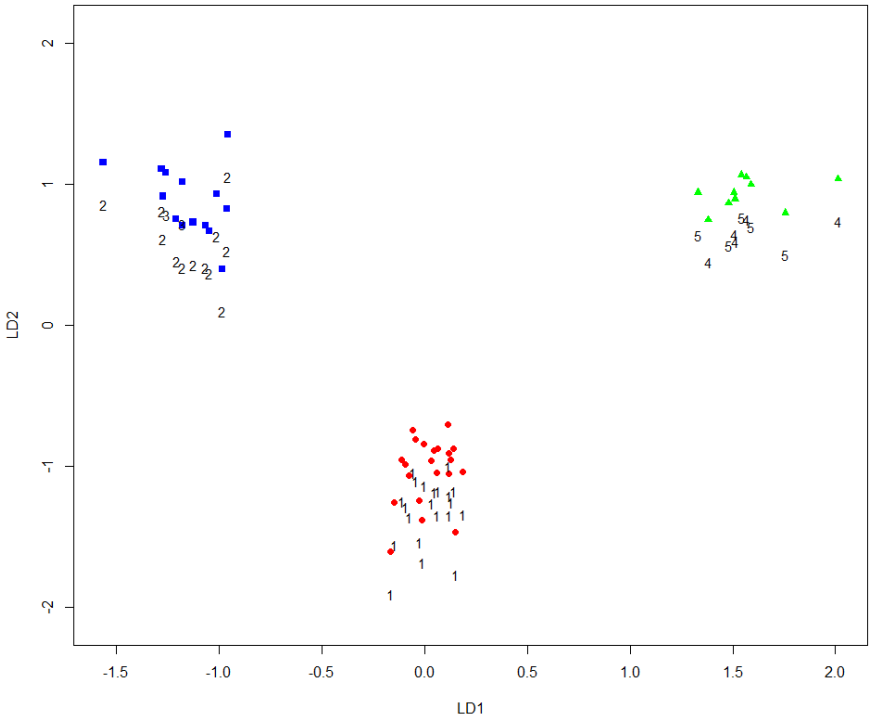
Kc$centers

	Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO
1	-1.2109353	0.1877267	1.314160	-0.6587807	-0.09635432	1.1584702	-1.0778543	1.00706411	-0.1953565
2	0.4477072	0.6951290	-0.370851	0.9366328	0.57687920	-0.1139204	0.3389812	0.01349852	0.2118580
3	0.7551242	-1.7225881	-1.061037	-1.0446359	-1.07655028	-1.3826254	0.7971355	-1.43823665	-0.1714027

F: NA 25.35615 **41.55753** 34.11977 30.44068 26.30431



Linear Discriminant Score



```
     kiln
cluster  1  2  3  4  5
     1  21  0  0  0  0
     2   0 12  2  0  0
     3   0  0  0  5  5

G.class (randomForest w/ G = as.factor(groups)
G     1  2  3
  1  14  0  0
  2   0 21  0
  3   0  0 10
correct.rate   error.rate
           1            0
```

- H.4.2:
  - (a):
    - As we can see from the dendrogram above are pruning line separates the clusters around a height of 3 leaving us with 5 clusters. We can tell that we have 5 clusters because the pruning line cuts off 5 lines from the other side of the dendrogram. The correlation between original distances and cophenetic distances is 0.8752
  - (b):
    - From the data above we can see that the Pseudo-F criterion values from 1 to 6 group k-means are 25.35615 41.55753, 34.11977, 30.44068 and 26.30431. We can see that the highest Pseudo-F generated at cluster 3. After viewing the sum of squares graph (scree diagram), we can see that we have an elbow at 3 which indicates we should have 3 clusters. Both the Pseudo-F and scree diagram correspond to a cluster of 3. From the linear discriminant score graph we can see that kilns 1 are all in the same cluster and come from region 1, kilns 2 and 3 are in the same cluster and come from region 2 and kilns 4 and 5 are in the same cluster and come from the same region.
  - (c):
    - When performing supervised learning using RF (random forest) with the 3 cluster groups as the grouping factor we can see from the confusion matrix and error rates above that we perfectly classified the kilns with a correct rate of 1, as opposed to the RF confusion matrix and error rates in H.4.1 where we had a correct rate of 0.82 and an error rate of 0.18.