STAT 4601
Nolan Hodge
100925862

Assignment #2

K-means Clustering & Principal Component Analysis

Question 1

To ensure my test set and my training set used equivalent classes of “Cultivar” I used stratified random sampling with the strata classified by the “Cultivar” types (1, 2, and 3)

```
for (i in 1:3) {
  print(paste("Percentage of Cultivar type",as.integer(i),"in Training Set",(length(which(d.wine.trainset$Cultivar==i))/nrow(d.wine.trainset))*100), quote=F)
}
for (i in 1:3) {
  print(paste("Percentage of Cultivar type",as.integer(i),"in Test Set:",(length(which(d.wine.testset$Cultivar==i))/nrow(d.wine.testset))*100), quote=F)
}
```

Out:

[1] Percentage of Cultivar type 1 in Training Set 33.0508474576271

[1] Percentage of Cultivar type 2 in Training Set 39.8305084745763

[1] Percentage of Cultivar type 3 in Training Set 27.1186440677966

[1] Percentage of Cultivar type 1 in Test Set: 33.3333333333333

[1] Percentage of Cultivar type 2 in Test Set: 40

[1] Percentage of Cultivar type 3 in Test Set: 26.6666666666667

As you can see the representation of cultivar is fairly constant across both sets.

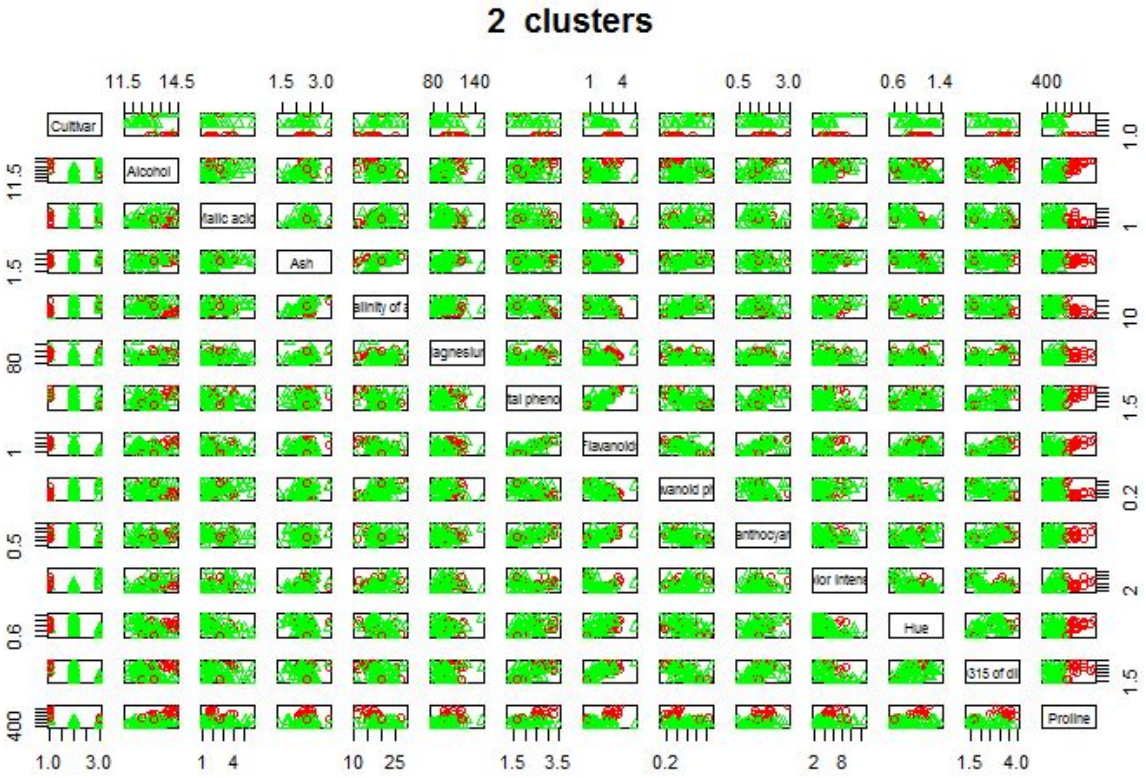
Question 2

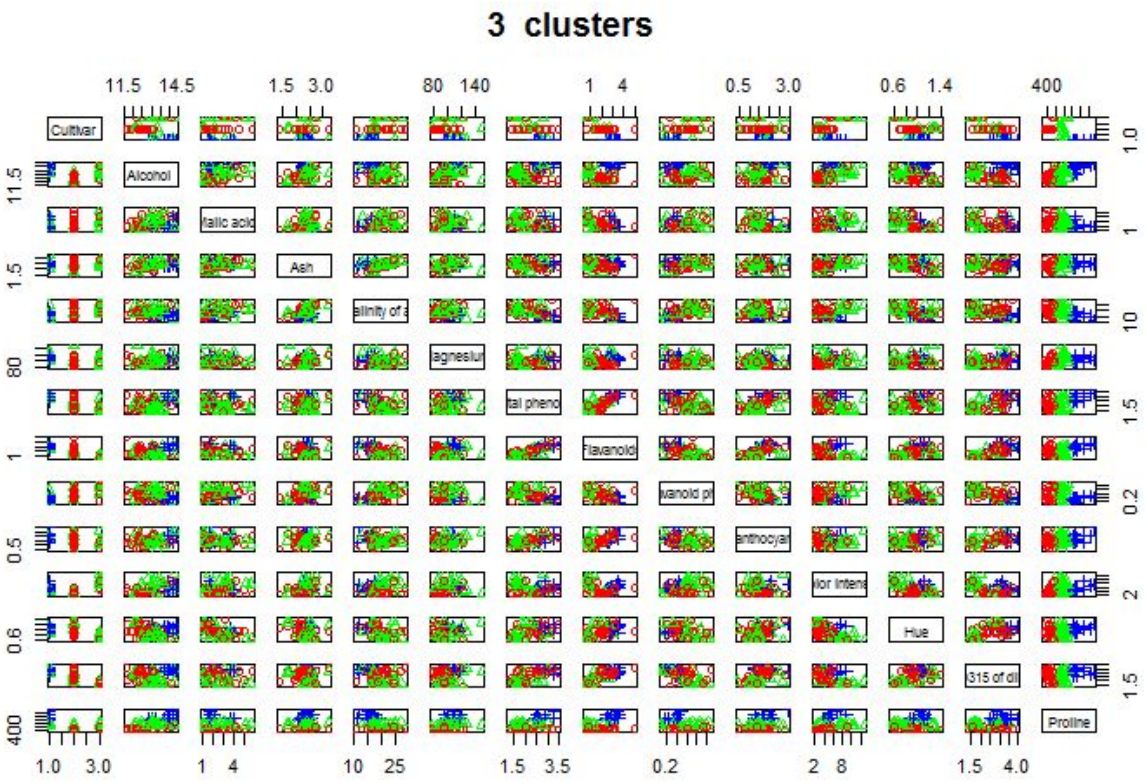
Here is output from a k-means clustering algorithm run 13 times with SSE and Davies-Bouldin averages applied.

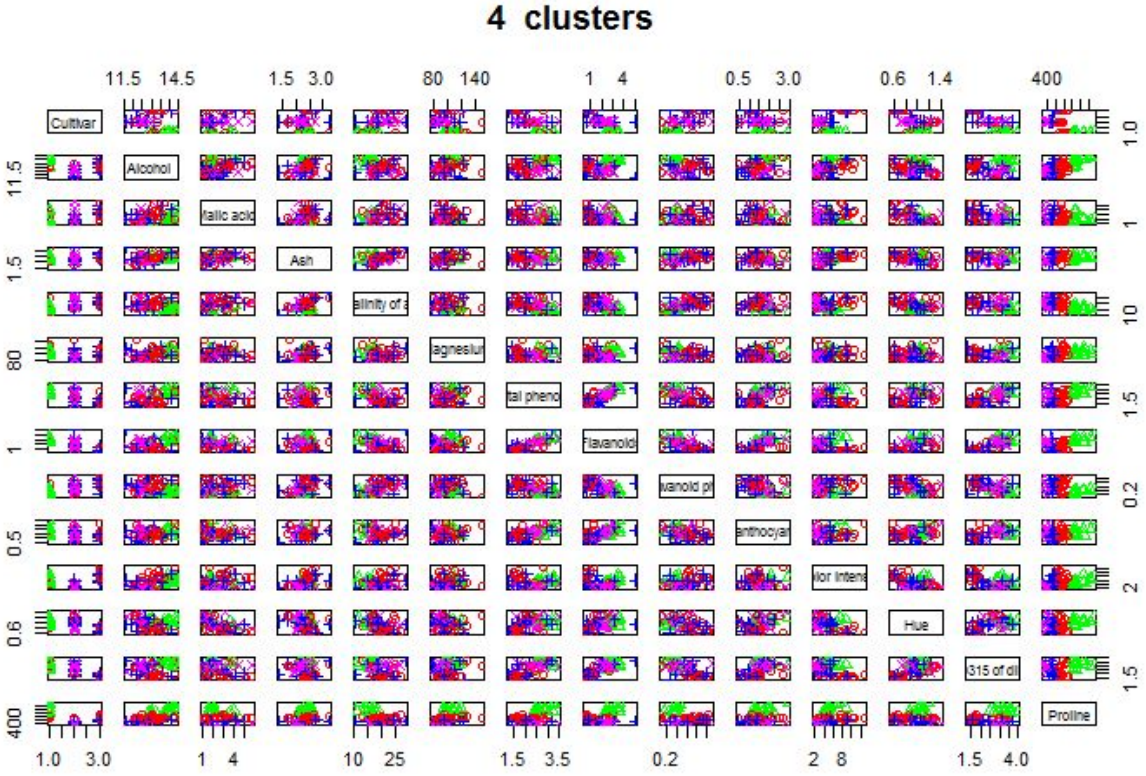
Each runthrough was done 10 times with 10 different seeds applied.

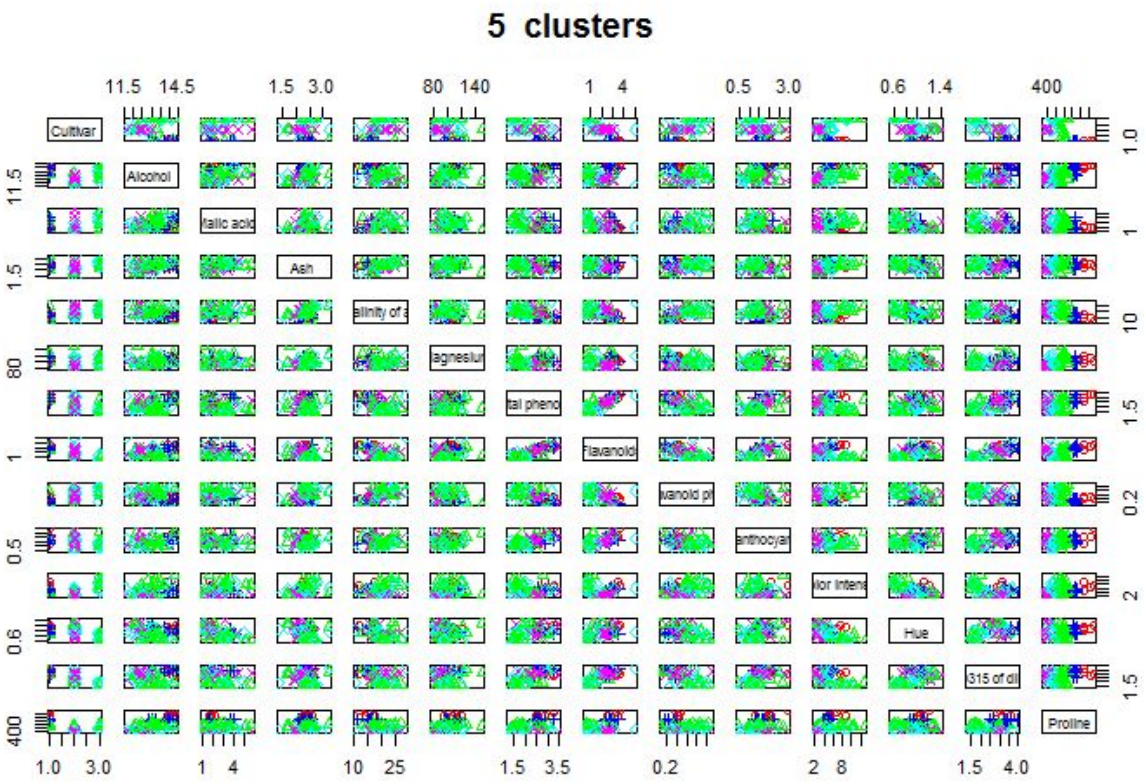
Test Set:

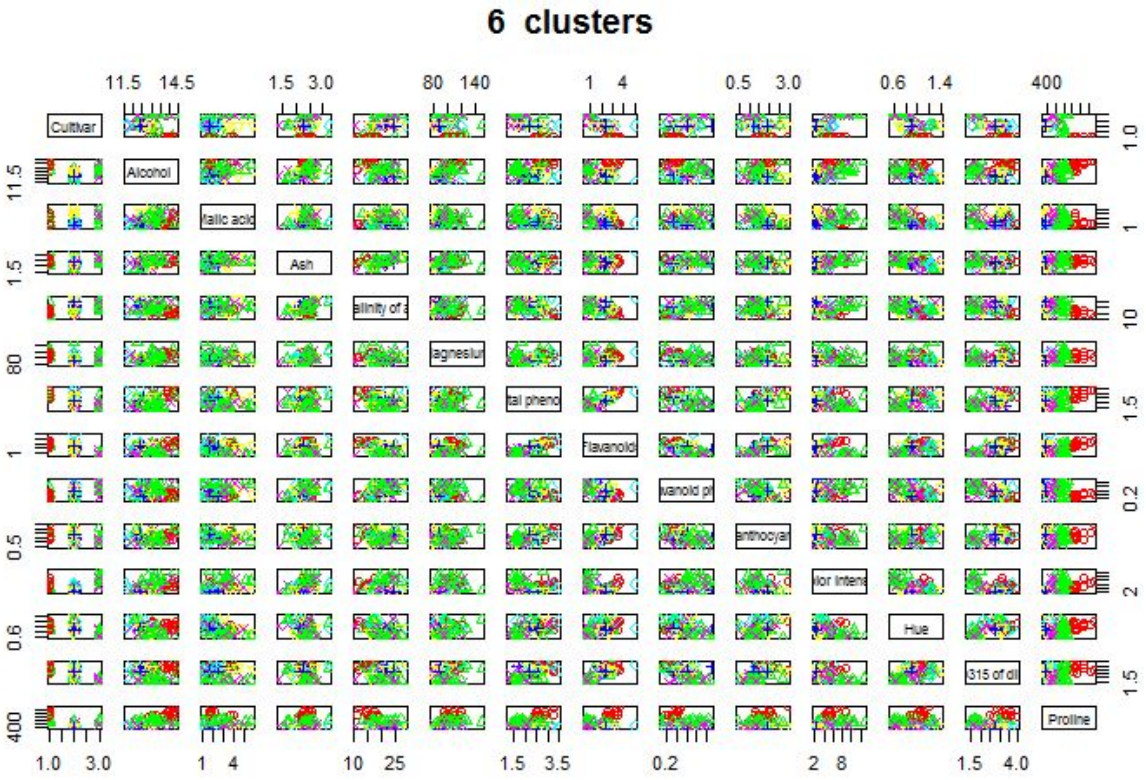
Seeds: 1 - 8

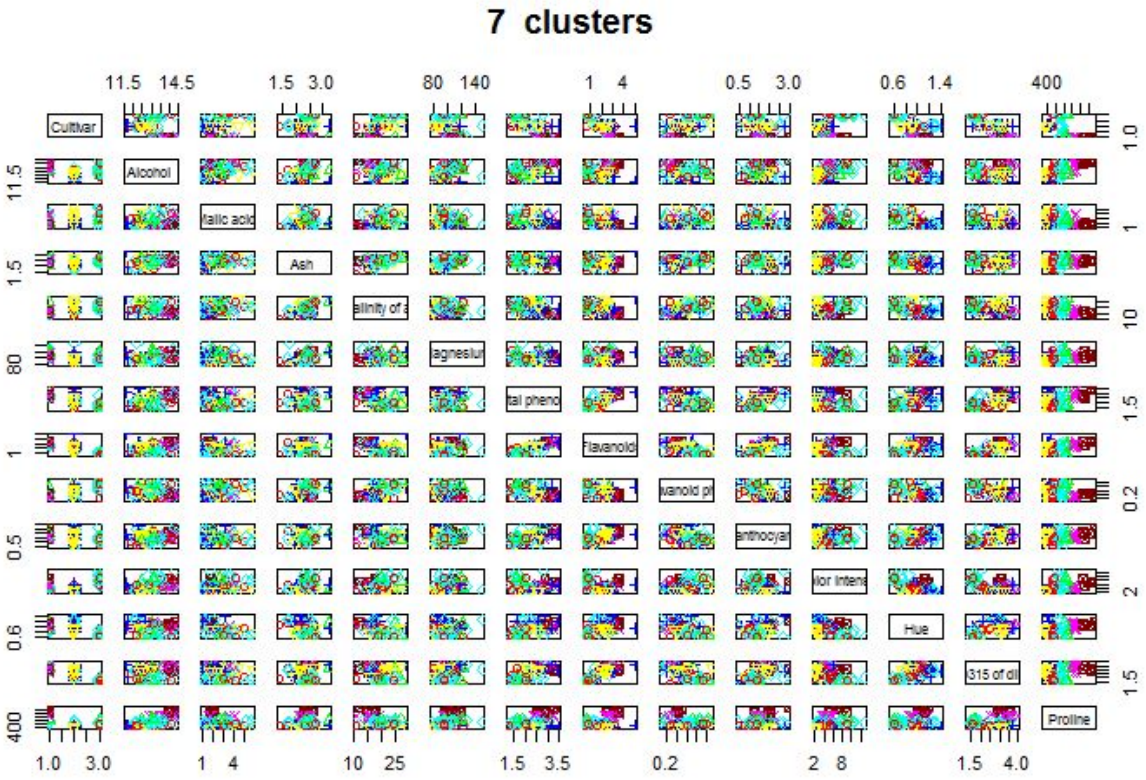


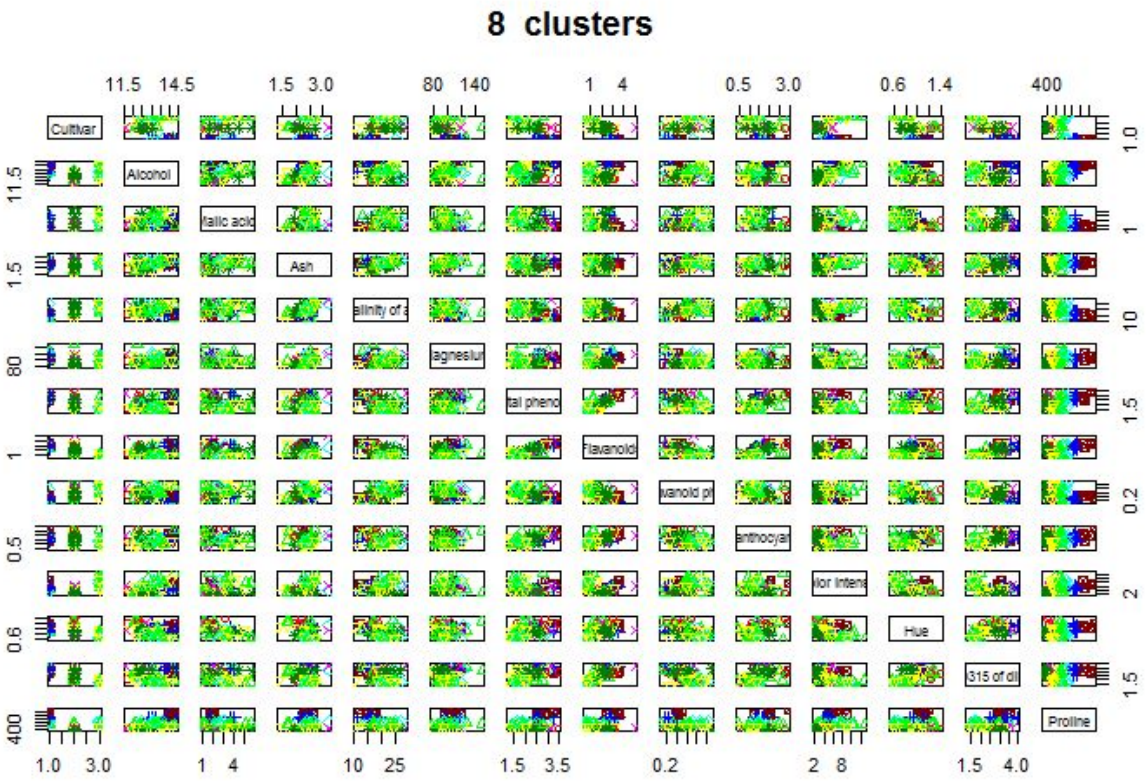

3 clusters
Cultivar
Alcohol
Malic acid
Ash
Alkalinity of
Magnesium
Total phenols
Flavanoids
Nonflavanoid
Proanthocyan
Color intens
Hue
315 of dil
Proline


4 clusters
Cultivar
Alcohol
Malic acid
Ash
Alkalinity of a
Magnesium
Total phenols
Flavanoids
Nonflavanoid ph
Proanthocyani
Color intens
Hue
OD315 of di
Proline


5 clusters
Cultivar
Alcohol
Malic acid
Ash
Flavanoid
Hue
Proline

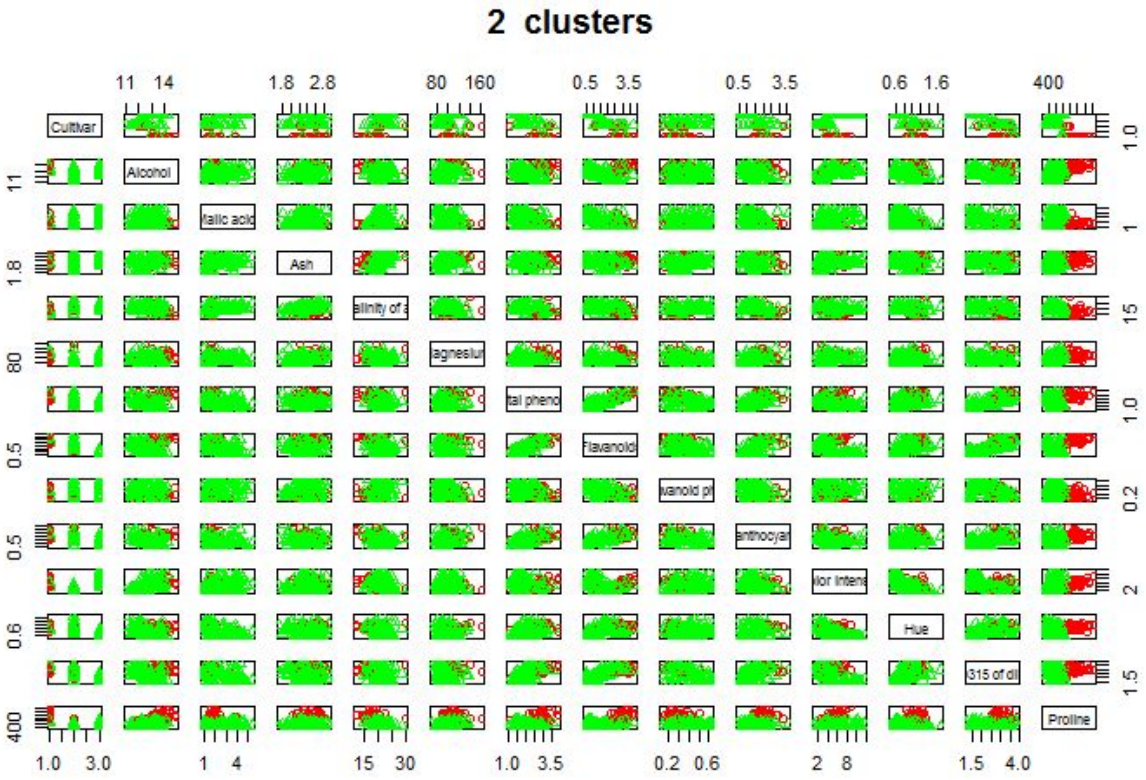

6 clusters
Cultivar
Alcohol
Malic acid
Ash
Alcalinity of a
Magnesium
Total phenol
Flavanoids
Nonflavanoid p
Proanthocyan
Color intens
Hue
OD315 of dil
Proline

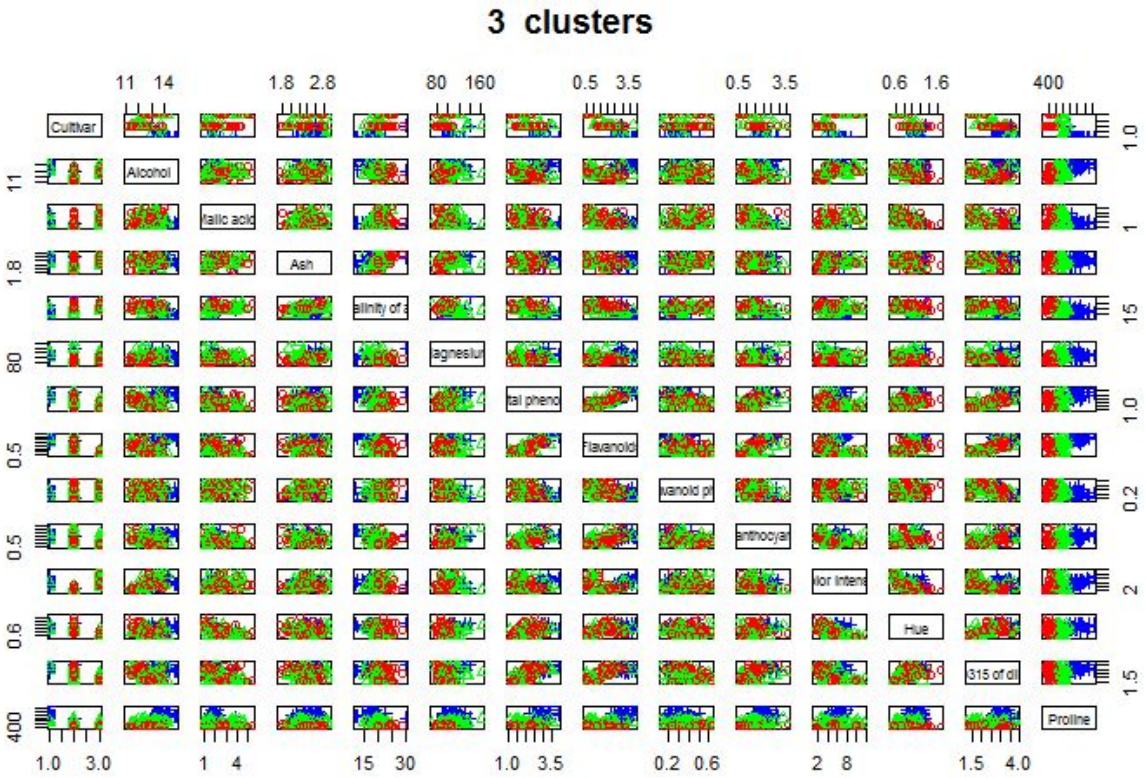

7 clusters
Cultivar
Alcohol
Malic acid
Ash
Flavanoids
Hue
Proline

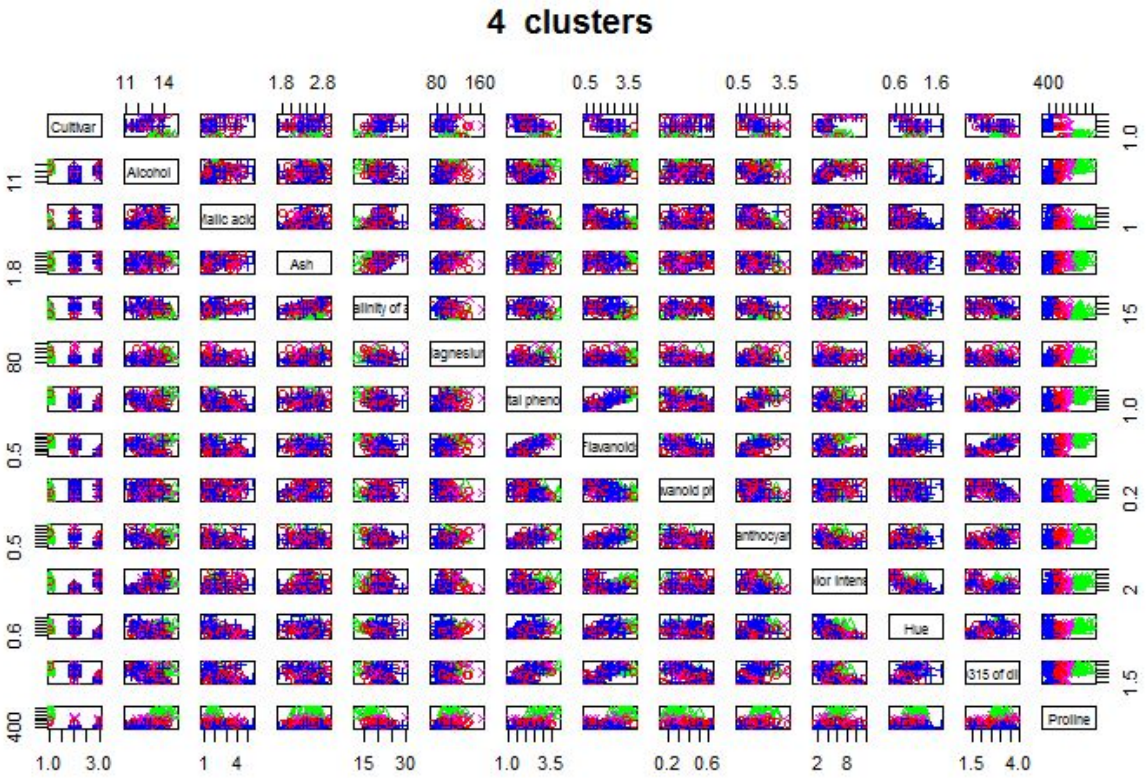

8 clusters
Cultivar
Alcohol
Malic acid
Ash
Alkalinity of a
Magnesium
Total pheno
Flavanoids
Nonflavanoid p
Proanthocyan
Color intens
Hue
OD315 of di
Proline

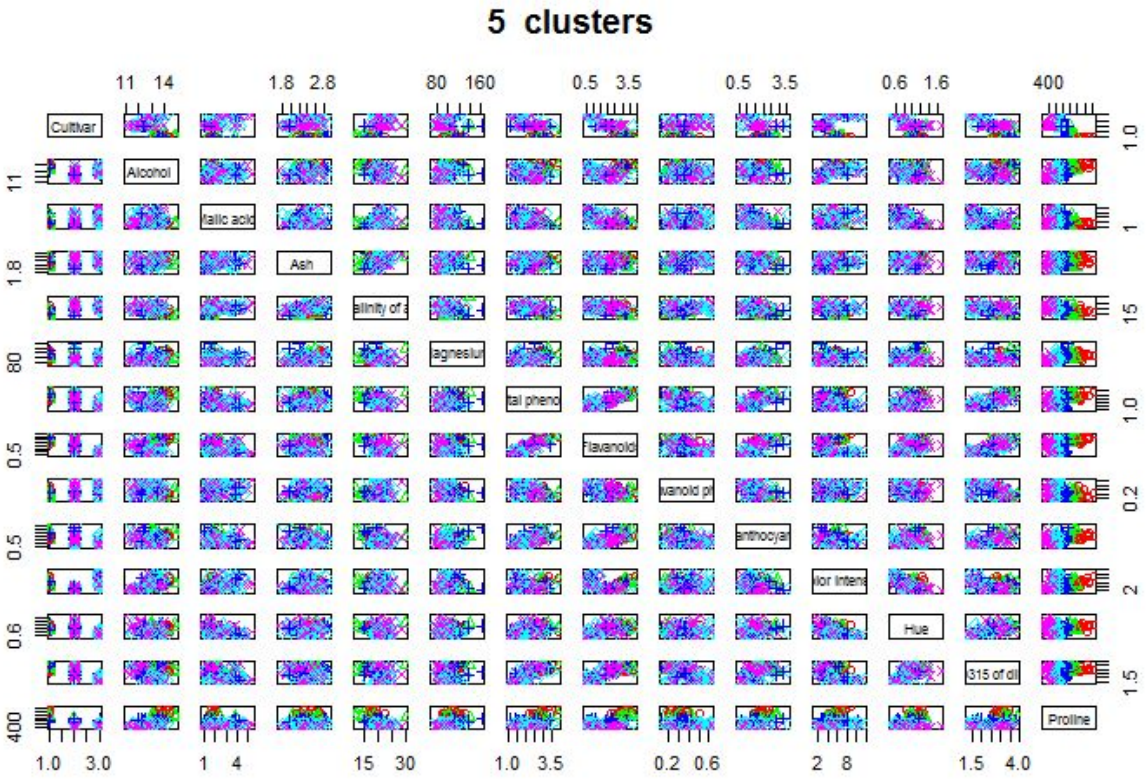
Training Set:

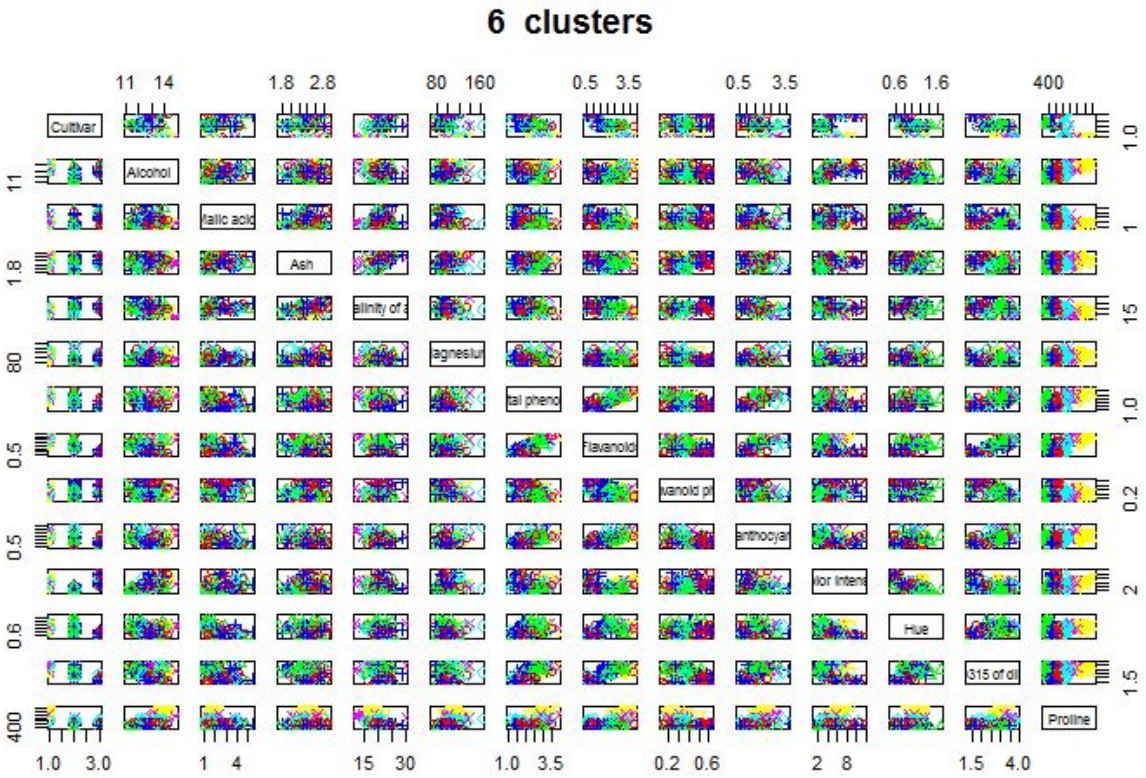
Seeds: 1 - 8

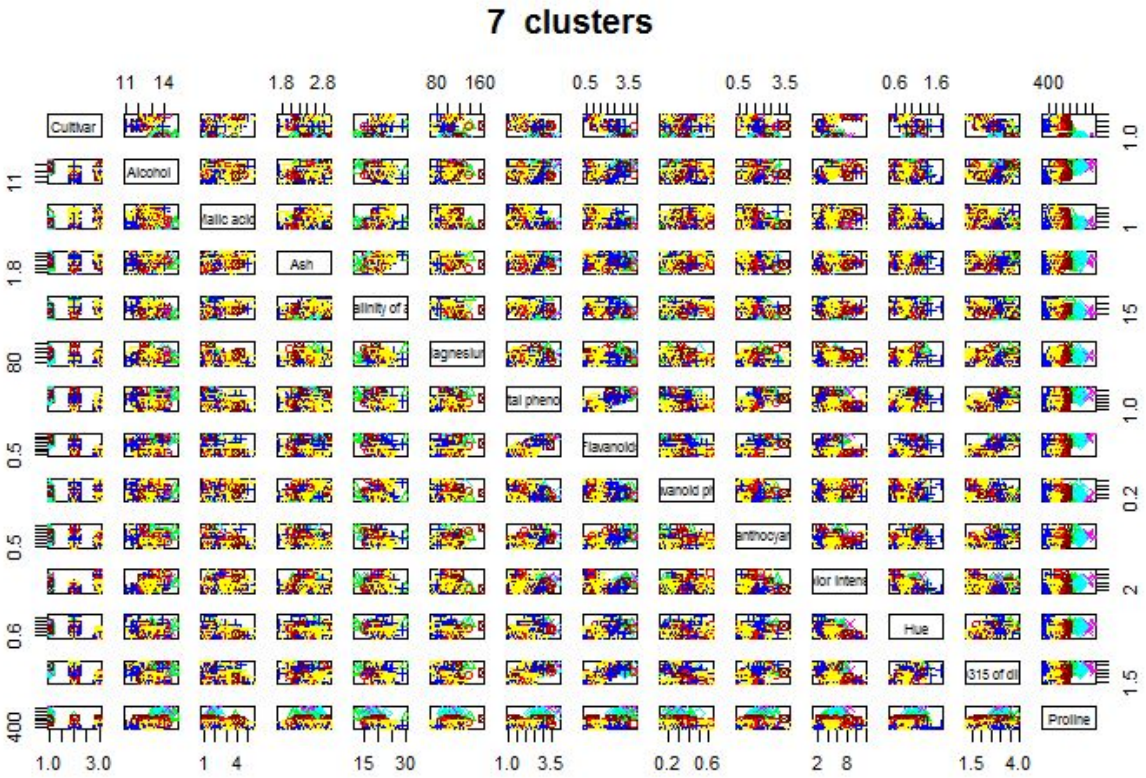

2 clusters
Cultivar
Alcohol
Malic acid
Ash
Flavanoid
Hue
Proline


3 clusters
Cultivar
Alcohol
Malic acid
Ash
Alcalinity of a
Magnesium
Total phenol
Flavanoids
Nonflavanoid ph
Proanthocyani
Color intens
Hue
OD315 of dil
Proline

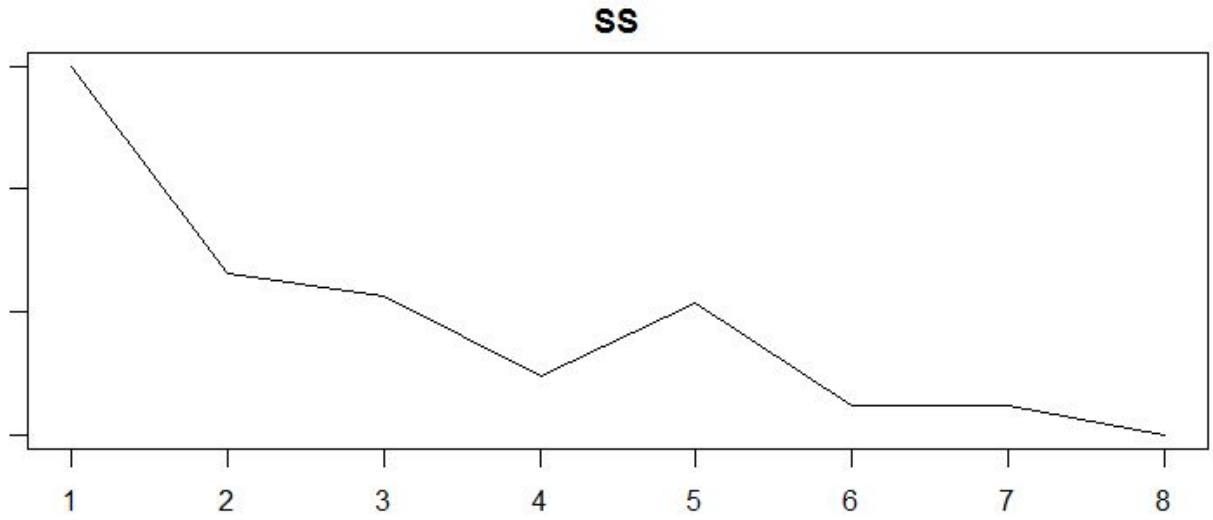

4 clusters
Cultivar
Alcohol
Malic acid
Ash
Flavanoid
Hue
Proline


5 clusters
Cultivar
Alcohol
Malic acid
Ash
Alkalinity of
Magnesium
Total phenol
Flavanoids
Nonflavanoid
Proanthocyan
Color intens
Hue
315 of dil
Proline


6 clusters
Cultivar
Alcohol
Malic acid
Ash
Flavanoid
Hue
Proline


7 clusters
Cultivar
Alcohol
Malic acid
Ash
Alcalinity of a
Magnesium
Total phenol
Flavanoids
Nonflavanoid ph
Proanthocyani
Color intens
Hue
OD315 of dil
Proline

Here is the standard error plot, the elbow shows 4 clusters for the test data.



Here is the standard error plot for the training data.