

# Unsupervised Learning

Brady Lamson

2022-03-30

## Exercise 1

```
my.data <- data.frame(X1 = c(3, 5, 4, 7),
                      X2 = c(6, 4, 9, 9),
                      X3 = c(1, 7, 2, 1))
rownames(my.data) <- c("Obs1", "Obs2", "Obs3", "Obs4")

my.data_dist <- dist(my.data, method = "euclidean")
my.data_dist
```

## Obs1 Obs2 Obs3  
## Obs2 6.633250  
## Obs3 3.316625 7.141428  
## Obs4 5.000000 8.062258 3.162278

- a) The distance between Obs1 and Obs 2 is **6.633**.
  - b) Obs4 and Obs3 have a distance of **3.16** which appears to be the shortest.
  - c) Obs4 and Obs3 would be merged in the first step.
- 

## Exercise 2

```
arr_dist <- dist(USArrests, method = "euclidean")
#arr_dist
```

The distance between **Florida** and **Alabama** is 102.001618.

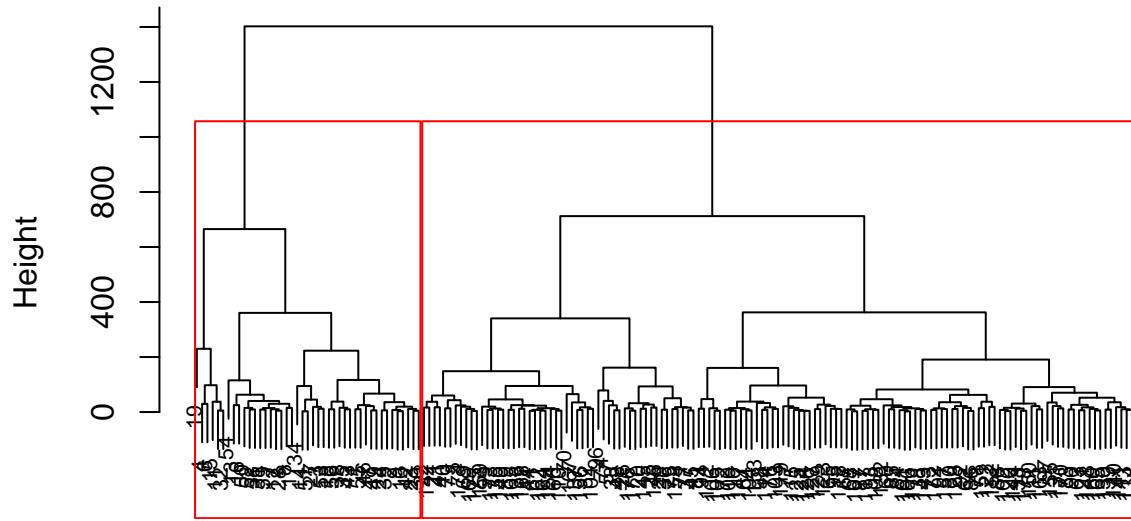
---

## Exercise 3

```
wine <- rattle::wine %>% select(-Type)

wine_dist <- dist(wine, method = "euclidean")
wine_hclust <- hclust(wine_dist)
plot(wine_hclust, cex = 0.7)
rect.hclust(wine_hclust, k = 2, border = "red")
```

## Cluster Dendrogram



```
wine_dist  
hclust (*, "complete")  
  
my_clusters <- cutree(wine_hclust, k = 2)  
my_clusters %>% table()  
  
## .  
##   1    2  
##  43  135
```

### Exercise 4

```
my.x1 <- c(5.2, 4.6, 5.9, 6.8, 10.5, 10.7, 8.6, 10.5, 14.1, 16.4, 14.3, 12.4)  
my.x2 <- c(3.6, 4.7, 2.2, 4.5, 7.2, 7.3, 7.1, 9.9, 6.3, 4.2, 6.2, 3.3)  
my.data <- data.frame(x1 = my.x1, x2 = my.x2)  
  
# So that everyone has the same randomly selected starting cluster centers:  
set.seed(27)  
  
# Carry out the k means cluster analysis with k = 3:  
my_kmclust <- kmeans(my.data, centers = 3)  
my_kmclust$cluster %>% table()  
  
## .  
## 1 2 3  
## 4 4 4
```

We have three clusters each with **4 observations** inside of them.

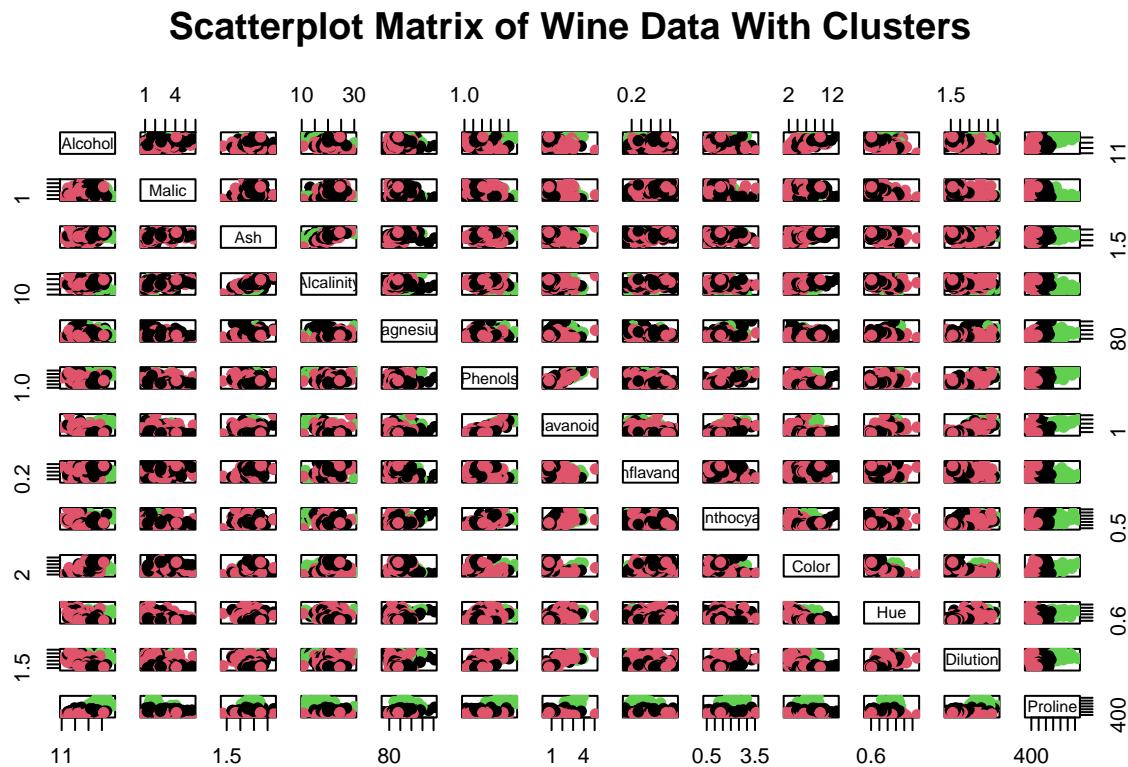
---

### Exercise 5

```
set.seed(20)

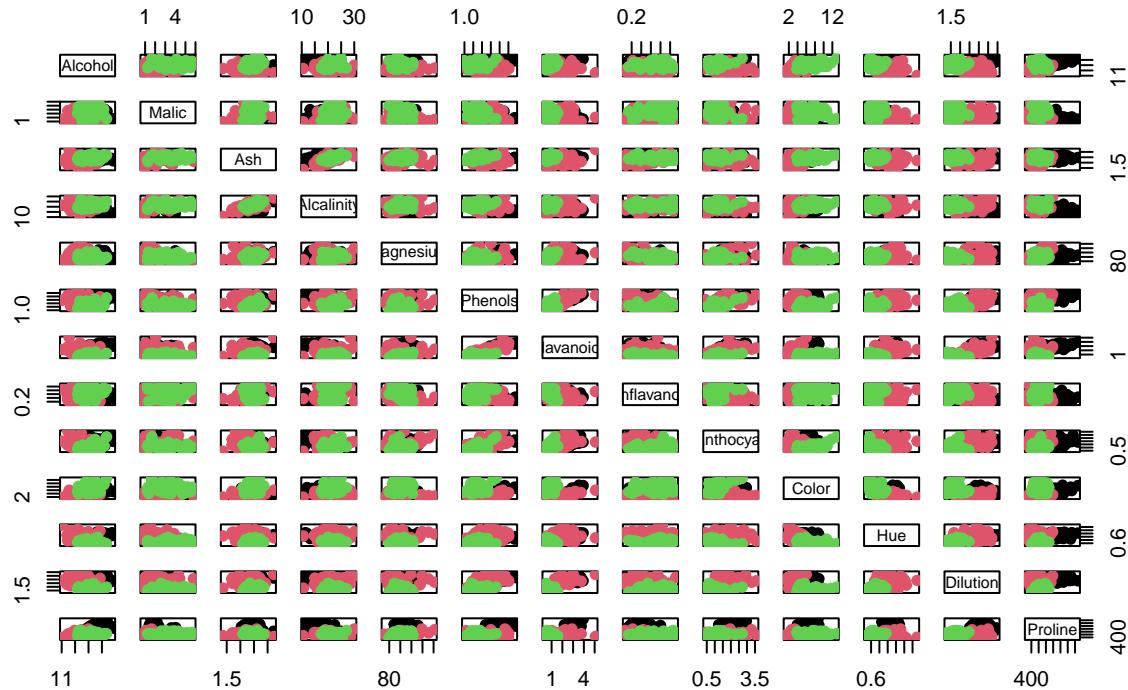
wine_kmclust <- kmeans(wine, centers = 3)
kmclusters <- wine_kmclust$cluster

wine %>%
  pairs(
    col = kmclusters,
    main = "Scatterplot Matrix of Wine Data With Clusters",
    pch = 19
  )
```



```
wine %>%
  pairs(
    col = rattle::wine$type,
    main = "Scatterplot Matrix of Wine Data With Types",
    pch = 19
  )
```

## Scatterplot Matrix of Wine Data With Types



- a) The clusters don't appear to correspond to type very well. With one exception, the bottom row does a fairly decent job of separating out the groups.

```

# Standardize each of the 13 variables:
wine2_std <- scale(wine, center = TRUE, scale = TRUE)

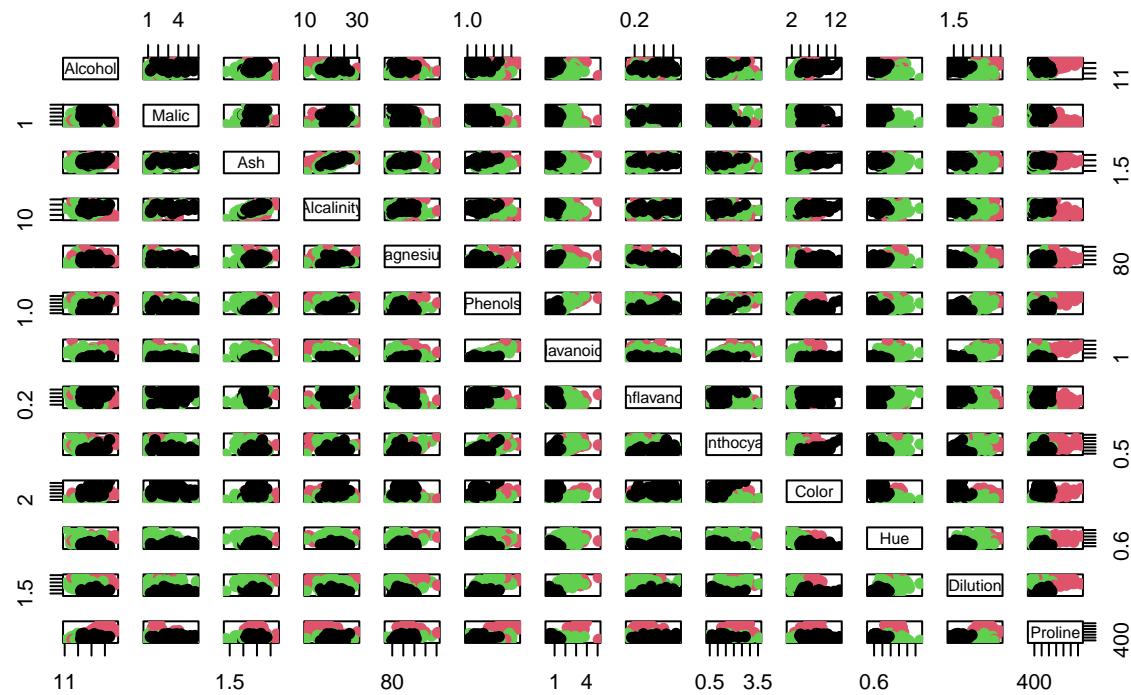
# So that everyone has the same randomly selected starting cluster centers:
set.seed(20)

# Carry out the k means cluster analysis with k = 3:
wine_kmclust_std <- kmeans(wine2_std, centers = 3)

my.clusters_std <- wine_kmclust_std$cluster
pairs(wine,
      col = my.clusters_std,
      main = "Scatterplot Matrix of Wine Data With Clusters",
      pch = 19)

```

## Scatterplot Matrix of Wine Data With Clusters



Yes, the clusters here do seem to be corresponding to wine types.