Clustering Example 3: *k*-means Applied to Wine Data

Amy Wagaman, Nathan Carter

July 2020

Load necessary libraries.

```r
library(mosaic)
library(cluster)
```

Load wine dataset.

This requires you to have access to the `winedata.txt` file. It is available in the book's GitHub repository at the following URL.

https://github.com/ds4m/ds4m.github.io/tree/master/chapter-5-resources/winedata.txt

If you run this R code, place the data file in the same folder as the code file.

```r
winedata <- read.csv("winedata.txt")
```

See column names, data types, and some values in each column.

```r
glimpse(winedata)
```

```
## Observations: 178
## Variables: 14
## $ Cultivar             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ Alcohol              <dbl> 14.23, 13.20, 13.16, 14.37, 13.24, 14.20, 14.39, ...
## $ MalicAcid            <dbl> 1.71, 1.78, 2.36, 1.95, 2.59, 1.76, 1.87, 2.15, 1...
## $ Ash                  <dbl> 2.43, 2.14, 2.67, 2.50, 2.87, 2.45, 2.45, 2.61, 2...
## $ AlcalinityofAsh      <dbl> 15.6, 11.2, 18.6, 16.8, 21.0, 15.2, 14.6, 17.6, 1...
## $ Magnesium            <int> 127, 100, 101, 113, 118, 112, 96, 121, 97, 98, 10...
## $ TotalPhenols         <dbl> 2.80, 2.65, 2.80, 3.85, 2.80, 3.27, 2.50, 2.60, 2...
## $ Flavanoids           <dbl> 3.06, 2.76, 3.24, 3.49, 2.69, 3.39, 2.52, 2.51, 2...
## $ NonflavanoidPhenols <dbl> 0.28, 0.26, 0.30, 0.24, 0.39, 0.34, 0.30, 0.31, 0...
## $ Proanthocyanins      <dbl> 2.29, 1.28, 2.81, 2.18, 1.82, 1.97, 1.98, 1.25, 1...
## $ ColorIntensity       <dbl> 5.64, 4.38, 5.68, 7.80, 4.32, 6.75, 5.25, 5.05, 5...
## $ Hue                  <dbl> 1.04, 1.05, 1.03, 0.86, 1.04, 1.05, 1.02, 1.06, 1...
## $ ODDilutedWines       <dbl> 3.92, 3.40, 3.17, 3.45, 2.93, 2.85, 3.58, 3.58, 2...
## $ Proline              <int> 1065, 1050, 1185, 1480, 735, 1450, 1290, 1295, 10...
```

See summary statistics for all columns in wine data.

```r
favstats(~ Alcohol, data = winedata)
```

```
##    min      Q1 median      Q3   max     mean        sd   n missing
##  11.03 12.3625  13.05 13.6775 14.83 13.00062 0.8118265 178       0
```

```r
favstats(~ MalicAcid, data = winedata)
```

```
##   min     Q1 median     Q3 max     mean       sd   n missing
##  0.74 1.6025  1.865 3.0825 5.8 2.336348 1.117146 178       0
```

```r
favstats(~ Ash, data = winedata)
```

```
##   min   Q1 median     Q3  max     mean       sd   n missing
##  1.36 2.21   2.36 2.5575 3.23 2.366517 0.274344 178       0
```

```r
favstats(~ AlcalinityofAsh, data = winedata)
```

```
##   min   Q1 median   Q3 max     mean       sd   n missing
##  10.6 17.2   19.5 21.5  30 19.49494 3.339564 178       0
```

```r
favstats(~ Magnesium, data = winedata)
```

```
##  min Q1 median  Q3 max     mean       sd   n missing
##   70 88     98 107 162 99.74157 14.28248 178       0
```

```r
favstats(~ TotalPhenols, data = winedata)
```

```
##   min     Q1 median  Q3  max     mean       sd   n missing
##  0.98 1.7425  2.355 2.8 3.88 2.295112 0.625851 178       0
```

```r
favstats(~ Flavanoids, data = winedata)
```

```
##   min    Q1 median    Q3  max    mean        sd   n missing
##  0.34 1.205  2.135 2.875 5.08 2.02927 0.9988587 178       0
```

```r
favstats(~ NonflavanoidPhenols, data = winedata)
```

```
##   min   Q1 median     Q3  max      mean        sd   n missing
##  0.13 0.27   0.34 0.4375 0.66 0.3618539 0.1244533 178       0
```

```r
favstats(~ Proanthocyanins, data = winedata)
```

```
##   min   Q1 median   Q3  max     mean        sd   n missing
##  0.41 1.25  1.555 1.95 3.58 1.590899 0.5723589 178       0
```

```r
favstats(~ ColorIntensity, data = winedata)
```

```
##   min   Q1 median  Q3 max    mean       sd   n missing
##  1.28 3.22   4.69 6.2  13 5.05809 2.318286 178       0
```

```r
favstats(~ Hue, data = winedata)
```

```
##   min     Q1 median   Q3  max      mean        sd   n missing
##  0.48 0.7825  0.965 1.12 1.71 0.9574494 0.2285716 178       0
```

```r
favstats(~ ODDilutedWines, data = winedata)
```

```
##   min     Q1 median   Q3 max     mean        sd   n missing
##  1.27 1.9375   2.78 3.17   4 2.611685 0.7099904 178       0
```

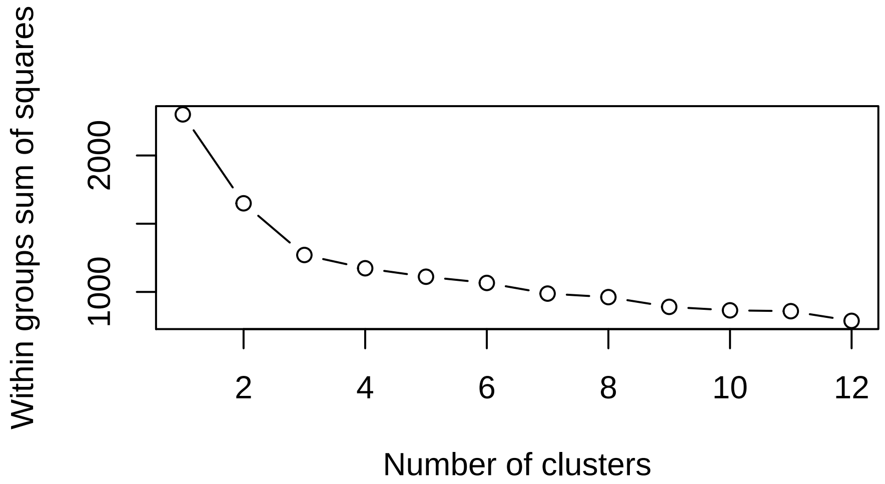
```r
favstats(~ Proline, data = winedata)
```

```
##  min    Q1 median  Q3  max     mean       sd   n missing
##  278 500.5  673.5 985 1680 746.8933 314.9075 178       0
```

Run k-means clustering for $k = 1$ to $k = 12$.

We plot the within-groups sum of squares for each run, so that we can assess which k value may be best. Note the choice of a random number seed, for reproducibility.

```r
set.seed(240)
wss <- rep(0, 12) #creates 12 copies of 0 to create an empty vector
for(i in 1:12) {
  wss[i] <- sum(kmeans(scale(winedata[, -c(1)]), centers = i)$withinss)
}
plot(1:12, wss, type = "b",
     xlab = "Number of clusters", ylab = "Within groups sum of squares")
```



Run k-means with the chosen value of $k = 3$.

And print the results, which include cluster centroids, the clustering partition as a vector, and the within-groups sum of squares.

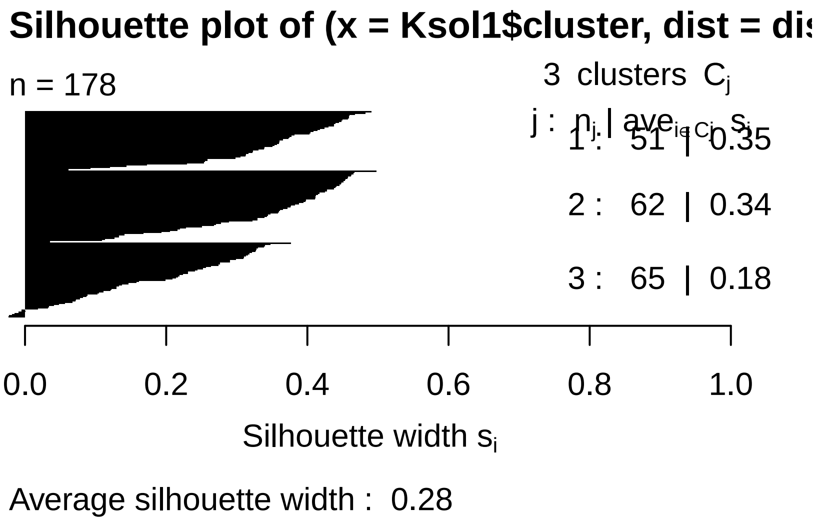
```r
set.seed(304)
Ksol1 <- kmeans(scale(winedata[, -c(1)]), centers = 3) #centers is the # of clusters
list(Ksol1) #so you can see what it gives you
```

```
## [[1]]
## K-means clustering with 3 clusters of sizes 51, 62, 65
##
## Cluster means:
##      Alcohol  MalicAcid        Ash AlcalinityofAsh   Magnesium TotalPhenols
## 1  0.1644436  0.8690954  0.1863726       0.5228924 -0.07526047  -0.97657548
## 2  0.8328826 -0.3029551  0.3636801      -0.6084749  0.57596208   0.88274724
## 3 -0.9234669 -0.3929331 -0.4931257       0.1701220 -0.49032869  -0.07576891
##    Flavanoids NonflavanoidPhenols Proanthocyanins ColorIntensity        Hue
## 1 -1.21182921          0.72402116     -0.77751312      0.9388902 -1.1615122
## 2  0.97506900         -0.56050853      0.57865427      0.1705823  0.4726504
## 3  0.02075402         -0.03343924      0.05810161     -0.8993770  0.4605046
##   ODDilutedWines    Proline
## 1     -1.2887761 -0.4059428
## 2      0.7770551  1.1220202
## 3      0.2700025 -0.7517257
##
## Clustering vector:
```

```
##   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 1 3 3 3 3 3 3 3 3 3 3 3 2
##  [75] 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [112] 3 3 3 3 3 3 3 1 3 3 2 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 326.3537 385.6983 558.6971
##  (between_SS / total_SS =  44.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Create corresponding silhouette plot.

```
kmeansSil <- silhouette(Ksol1$cluster, dist(scale(winedata[, -c(1)])))
silsum <- summary(kmeansSil)
plot(kmeansSil, col = "black")
```



Report average silhouette width more precisely.

```
summary(kmeansSil)$avg.width
```

```
## [1] 0.2848589
```

Compute average silhouette width for all possible values of k.

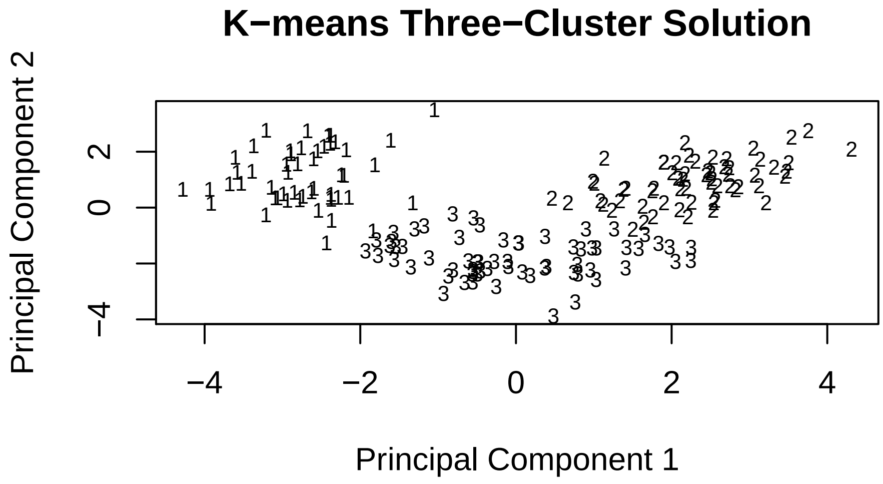
Thus k ranges from 2 to $n - 1$ (where n is the number of observations, here 178) and we report summary statistics for the collection of silhouette widths.

```
n <- 178
mydist <- dist(scale(winedata[, -c(1)]))
avgwidths <- rep(0, 176)

for (i in 2:(n-1)) {
  Ksol <- kmeans(scale(winedata[, -c(1)]), centers = i) #centers is the # of clusters
  kmeansSil <- silhouette(Ksol$cluster, mydist)
  avgwidths[i-1] <- summary(kmeansSil)$avg.width
}

summary(avgwidths)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.001421 0.074830 0.123755 0.110797 0.148152 0.284859
```

Plot the $k = 3$ solution in principal component space.

```
set.seed(1)
winePCAs <- princomp(winedata[, -c(1)], cor = TRUE)
plot(winePCAs$scores[, 1:2], type = "n",
     xlab = "Principal Component 1", ylab = "Principal Component 2",
     main="K-means Three-Cluster Solution")
text(winePCAs$scores[, 1:2], labels = Ksol1$cluster, cex = 0.7)
```



Compare found clusters to original wine cultivars.

```
tally(winedata$Cultivar ~ Ksol1$cluster)
```

```
##                   Ksol1$cluster
## winedata$Cultivar  1  2  3
##                 1  0 59  0
##                 2  3  3 65
##                 3 48  0  0
```

Perform k-means clustering with $k = 4$.

```
set.seed(304)
Ksol2 <- kmeans(scale(winedata[, -c(1)]), centers = 4) #centers is the # of clusters
list(Ksol2)
```

```
## [[1]]
## K-means clustering with 4 clusters of sizes 49, 56, 45, 28
##
```

```
## Cluster means:
##      Alcohol   MalicAcid        Ash AlcalinityofAsh   Magnesium TotalPhenols
## 1  0.1860184  0.90242582  0.2485092       0.5820616 -0.05049296   -0.9857762
## 2  0.9580555 -0.37748461  0.1969019      -0.8214121  0.39943022    0.9000233
## 3 -0.9051690 -0.53898599 -0.6498944       0.1592193 -0.71473842   -0.4537841
## 4 -0.7869073  0.04195151  0.2157781       0.3683284  0.43818899    0.6543578
##   Flavanoids NonflavanoidPhenols Proanthocyanins ColorIntensity        Hue
## 1 -1.2327174           0.7148253      -0.7474990      0.9857177 -1.1879477
## 2  0.9848901          -0.6204018       0.5575193      0.2423047  0.4799084
## 3 -0.2408779           0.3315072      -0.4329238     -0.9177666  0.5202140
## 4  0.5746004          -0.5429201       0.8888549     -0.7346332  0.2830335
##   ODDilutedWines    Proline
## 1    -1.29787850 -0.3789756
## 2     0.76926636  1.2184972
## 3     0.07869143 -0.7820425
## 4     0.60628629 -0.5169332
##
## Clustering vector:
##   [1] 2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 4 2 2 2 2 2 2 2 2 2 2 2 2
##  [38] 2 2 2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 4 3 4 2 3 3 4 3 4 3 4
##  [75] 4 3 3 3 4 4 3 3 3 1 4 3 3 3 3 3 3 3 3 4 4 4 4 3 4 4 3 3 4 3 3 3 3 3 3 4 4
## [112] 3 3 3 3 3 3 3 3 3 4 4 4 4 4 3 4 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 302.9915 268.5747 289.9515 307.0966
##  (between_SS / total_SS =  49.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Compare $k = 4$ solution to original wine cultivars.

```
tally(Ksol1$cluster ~ Ksol2$cluster)
```

```
##              Ksol2$cluster
## Ksol1$cluster  1  2  3  4
##             1 49  0  2  0
##             2  0 55  0  7
##             3  0  1 43 21
```