K Means Clustering Project

Guan-Yuan Wang

2020/9/4

K Means Clustering Project

Usually when dealing with an unsupervised learning problem, its difficult to get a good measure of how well the model performed. For this project, we will use data from the UCI archive based off of red and white wines (this is a very commonly used data set in ML).

We will then add a label to the a combined data set, we'll bring this label back later to see how well we can cluster the wine into groups.

Get the Data

Download the two data csv files from the UCI repository (or just use the downloaded csv files).

Use `read.csv` to open both data sets and set them as df1 and df2. Pay attention to what the separator (sep) is.

```
df1 <- read.csv("winequality-white.csv", sep = ";")
df2 <- read.csv("winequality-red.csv", sep = ";")
```

Now add a label column to both df1 and df2 indicating a label 'red' or 'white'.

```
df1$label <- "white"
df2$label <- "red"
```

Check the head of df1 and df2.

```
head(df1)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
## 4           7.2             0.23        0.32            8.5     0.058
## 5           7.2             0.23        0.32            8.5     0.058
## 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
```

```
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality label
## 1       6 white
## 2       6 white
## 3       6 white
## 4       6 white
## 5       6 white
## 6       6 white
```

```
head(df2)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality label
## 1       5   red
## 2       5   red
## 3       5   red
## 4       6   red
## 5       5   red
## 6       5   red
```

Combine df1 and df2 into a single data frame called wine.

```
wine <- rbind(df1, df2)
str(wine)
```

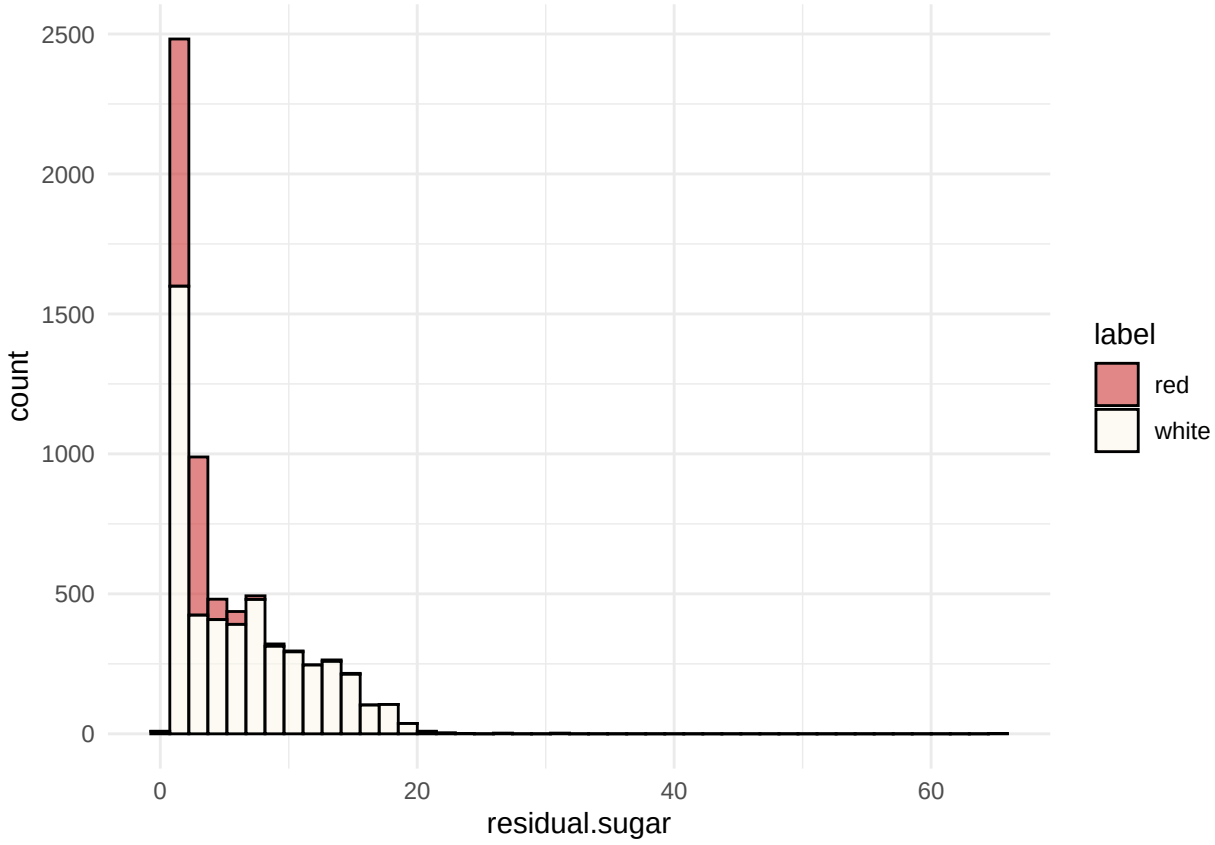
```
## 'data.frame':    6497 obs. of  13 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ label               : chr  "white" "white" "white" "white" ...
```

EDA

Let's explore the data a bit and practice our ggplot2 skills!

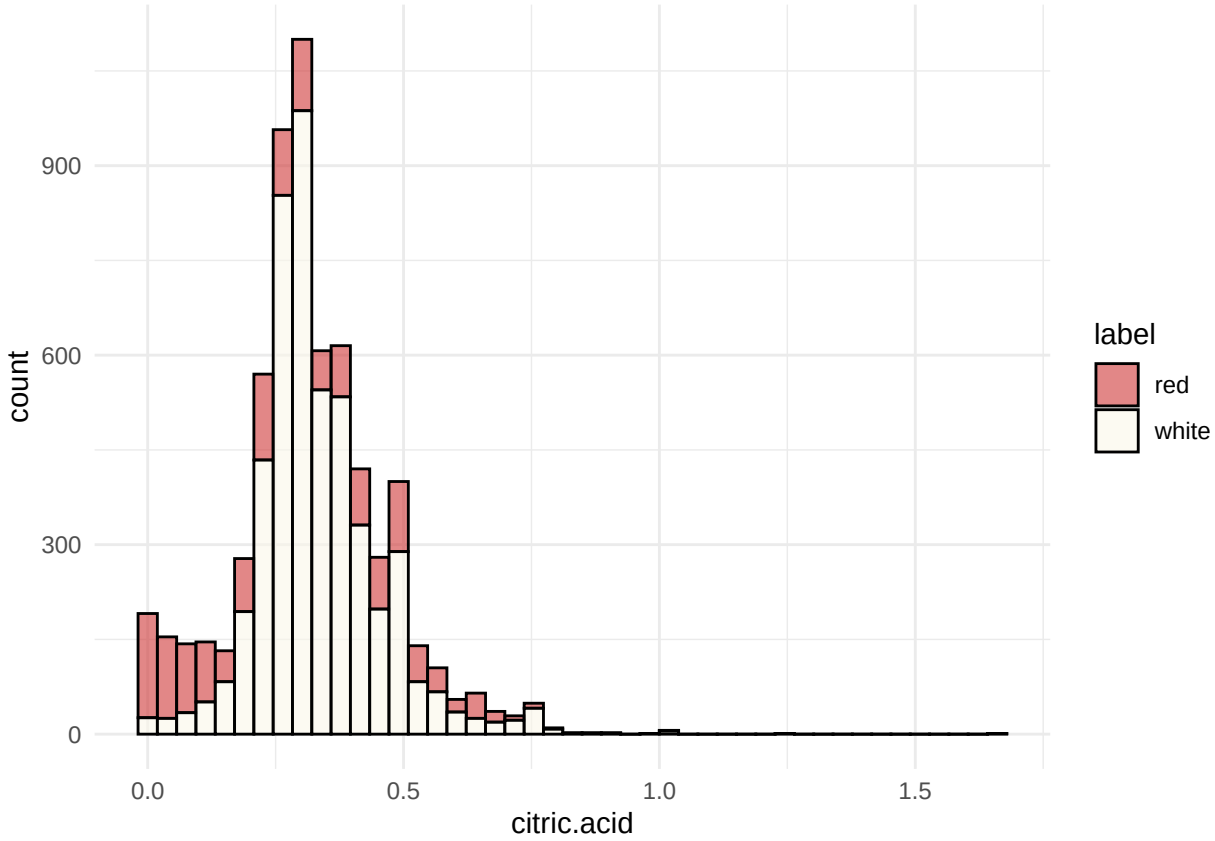
Create a Histogram of residual sugar from the wine data. Color by red and white wines.

```
ggplot(wine, aes(residual.sugar)) +
  geom_histogram(aes(fill = label), color = "black",
                 alpha = 0.6, bins = 45) +
  scale_fill_manual(values = c("#d03737", "#faf7ea"))
```



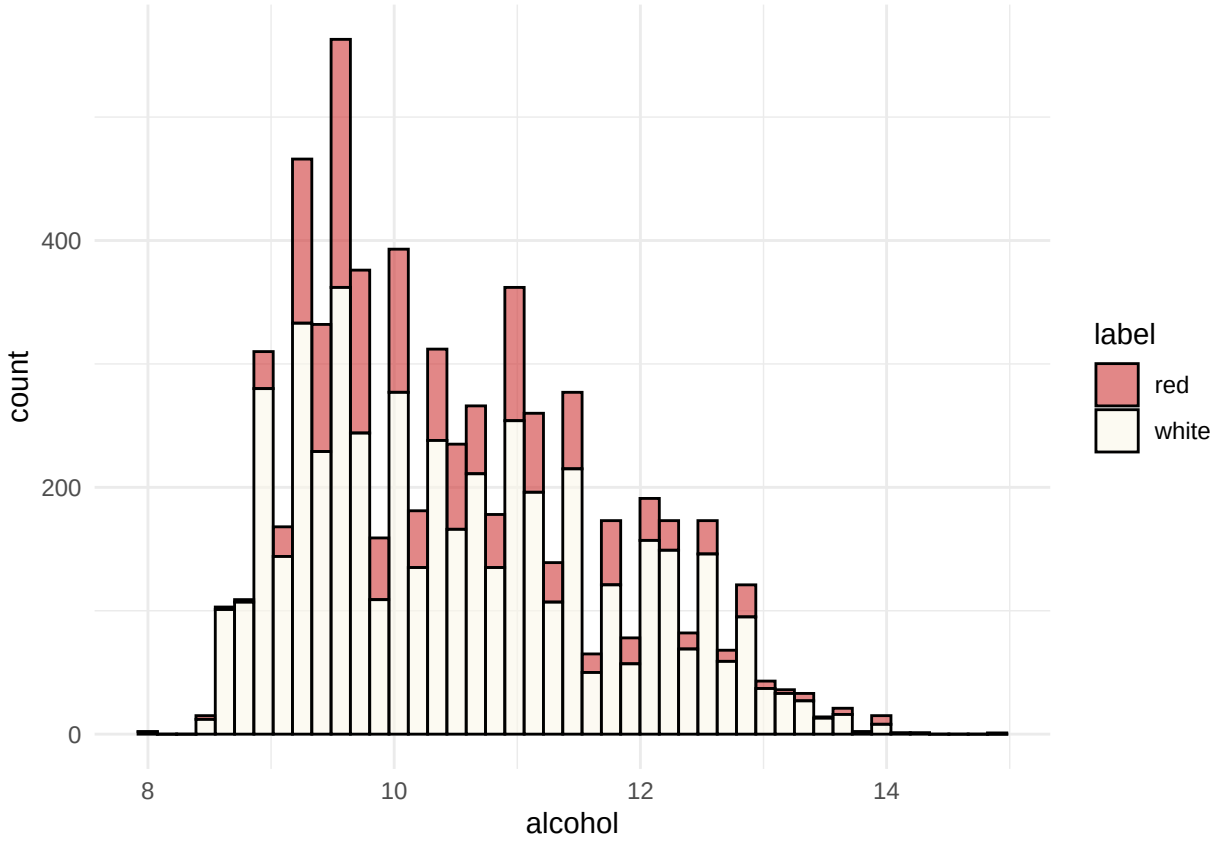
Create a Histogram of citric.acid from the wine data. Color by red and white wines.

```
ggplot(wine, aes(citric.acid)) +
  geom_histogram(aes(fill = label), color = "black",
                 alpha = 0.6, bins = 45) +
  scale_fill_manual(values = c("#d03737", "#faf7ea"))
```



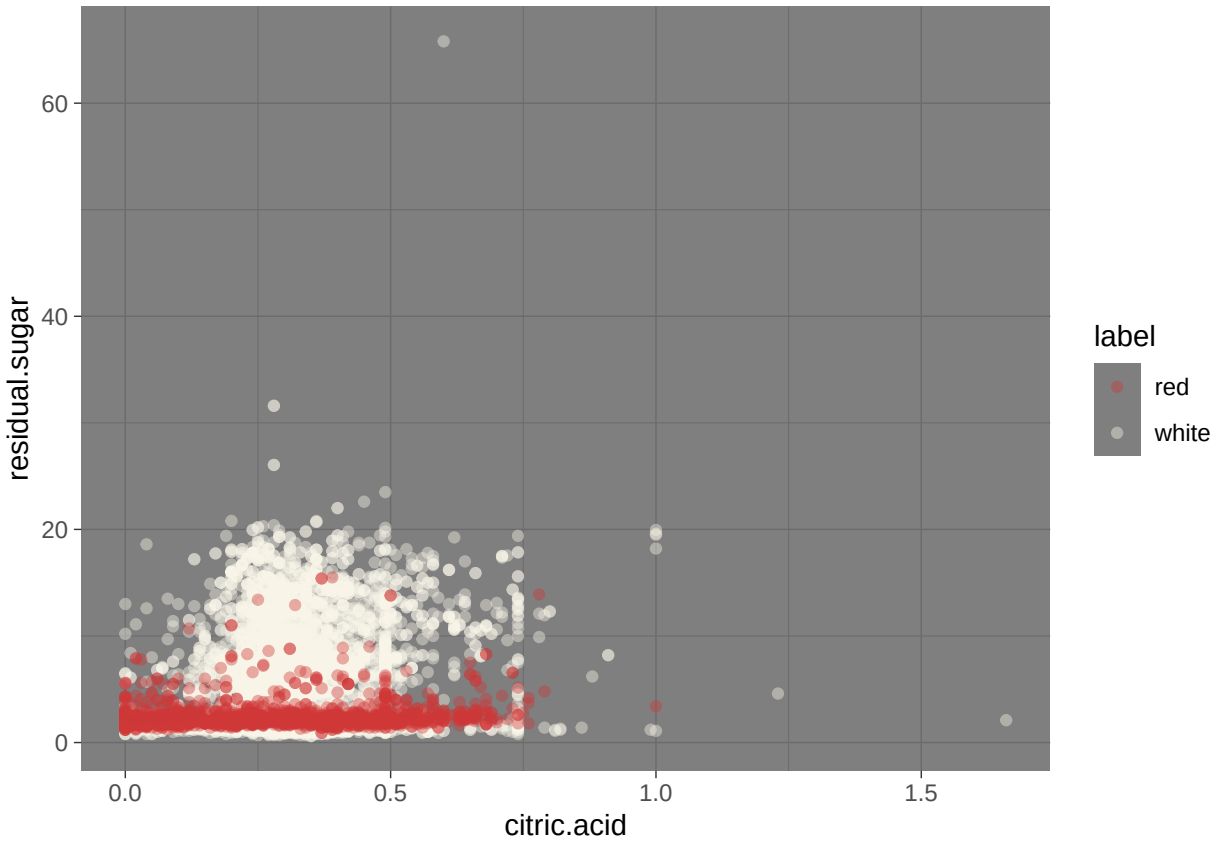
Create a Histogram of alcohol from the wine data. Color by red and white wines.

```
ggplot(wine, aes(alcohol)) +
  geom_histogram(aes(fill = label), color = "black",
                 alpha = 0.6, bins = 45) +
  scale_fill_manual(values = c("#d03737", "#faf7ea"))
```



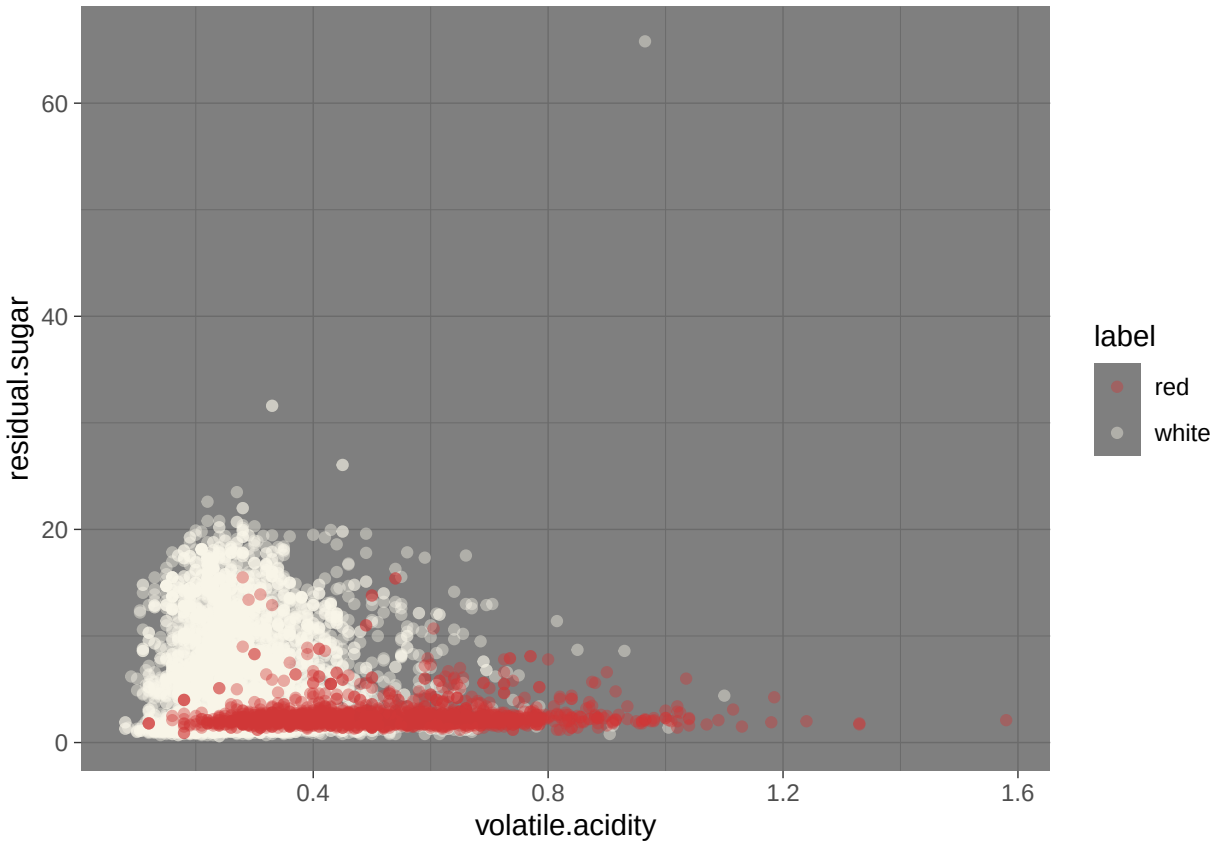
Create a scatterplot of residual.sugar versus citric.acid, color by red and white wine.

```
ggplot(wine, aes(citric.acid, residual.sugar)) +
  geom_point(aes(color = label), alpha = 0.4) +
  scale_color_manual(values = c("#d03737", "#faf7ea")) +
  theme_dark()
```



Create a scatterplot of volatile.acidity versus residual.sugar, color by red and white wine.

```
ggplot(wine, aes(volatile.acidity, residual.sugar)) +
  geom_point(aes(color = label), alpha = 0.4) +
  scale_color_manual(values = c("#d03737", "#faf7ea")) +
  theme_dark()
```



Feel free to explore the data as you see fit, we'll go ahead and move on!

Grab the wine data without the label and call it clus.data

```
clus.data <- wine[, -ncol(wine)]
```

Check the head of clus.data

```
head(clus.data)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
## 4           7.2             0.23        0.32            8.5     0.058
## 5           7.2             0.23        0.32            8.5     0.058
## 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
```

```
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

Building the Clusters

Call the kmeans function on clus.data and assign the results to wine.cluster.

```
wine.cluster <- kmeans(clus.data, 2)
```

Print out the wine.cluster Cluster Means and explore the information.

```
wine.cluster$centers
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides
## 1      6.904812        0.2871659   0.3397642       7.244809 0.04859257
## 2      7.623219        0.4086378   0.2908725       3.076425 0.06580983
##   free.sulfur.dioxide total.sulfur.dioxide   density       pH sulphates
## 1            39.75590            155.69246 0.9947903 3.190808 0.4999485
## 2            18.39868             63.26318 0.9945736 3.254882 0.5724145
##    alcohol  quality
## 1 10.25932 5.824343
## 2 10.79722 5.810541
```

Evaluating the Clusters

You usually won't have the luxury of labeled data with KMeans, but let's go ahead and see how we did!

Use the `table()` function to compare your cluster results to the real results. Which is easier to correctly group, red or white wines?

```
table(wine.cluster$cluster, wine$label)
```

```
##
##      red white
##   1   85  3604
##   2 1514  1294
```