K-Means Clustering

Nabil Momin

2024-06-10

```r
library(corrgram)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(caTools)
library(Amelia)
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.2, built: 2024-04-10)
## ## Copyright (C) 2005-2024 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(rpart)
library(rpart.plot)
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(ISLR)
library(e1071)
library(cluster)


#### importing the csv file

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

head(df1)
```

```
##   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
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```r
head(df2)
```

```
##   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
```

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

head(df1)
```

```
##   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
```

```
head(df2)
```

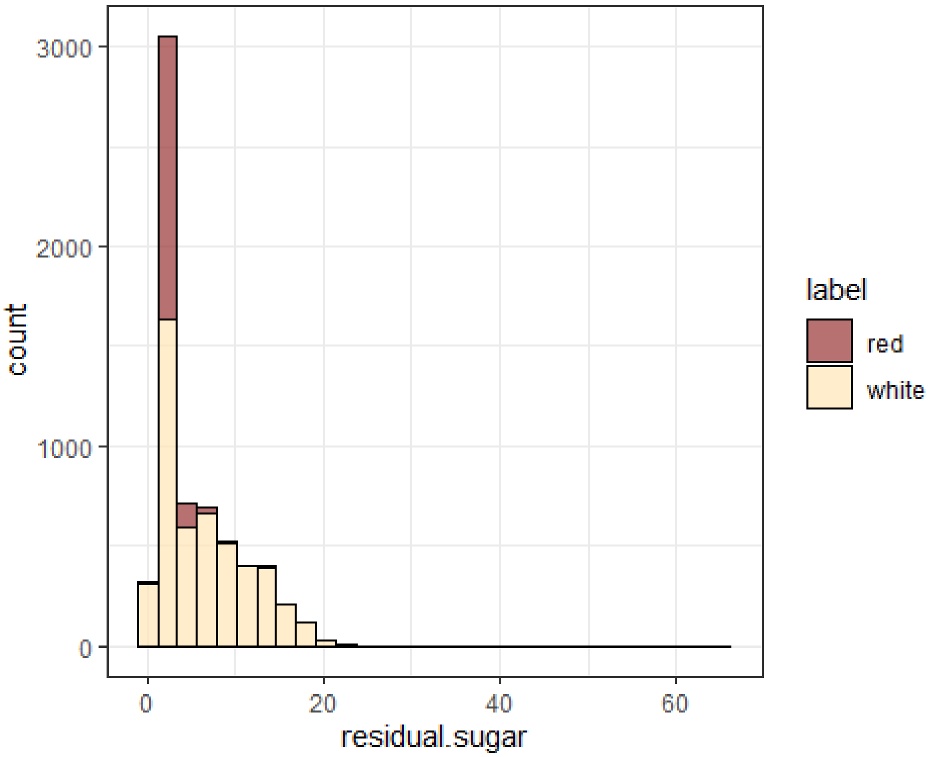
```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
## 4           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
```

```r
#### combining them together

wine <- rbind(df1,df2)

View(wine)

print(table(wine$label))
```

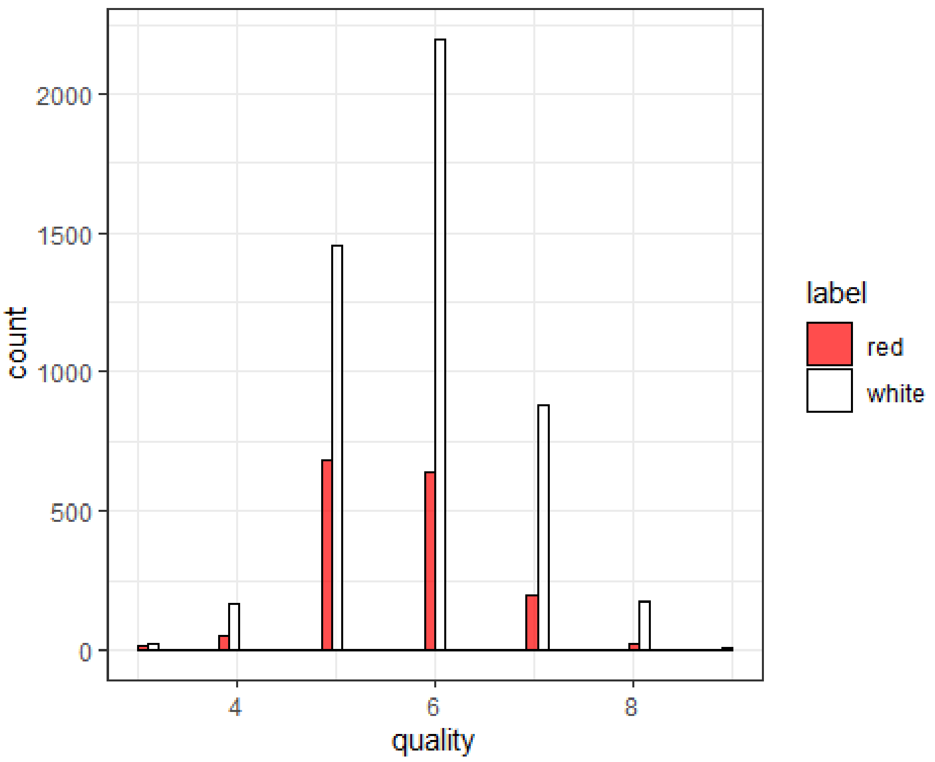
```
## 
##   red white 
##  1599  4898
```

```r
#### EDA time

ggplot(wine,aes(residual.sugar)) +
geom_histogram(aes(fill=label),color='black',alpha=0.7)  +
scale_fill_manual(values = c('#993333','#ffe5b4')) +theme_bw()
```

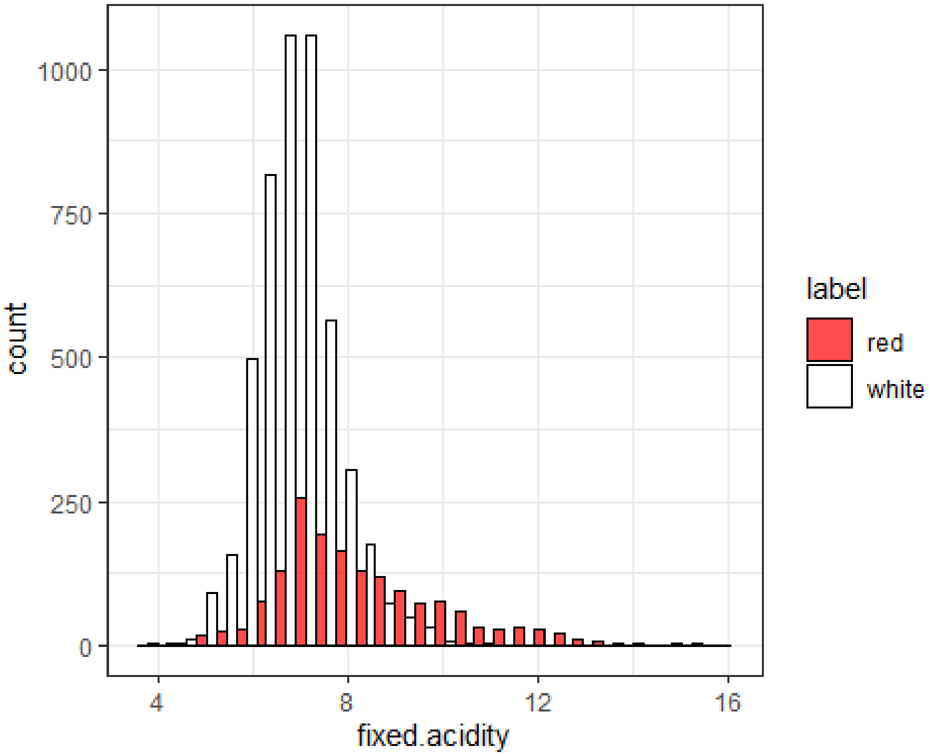
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



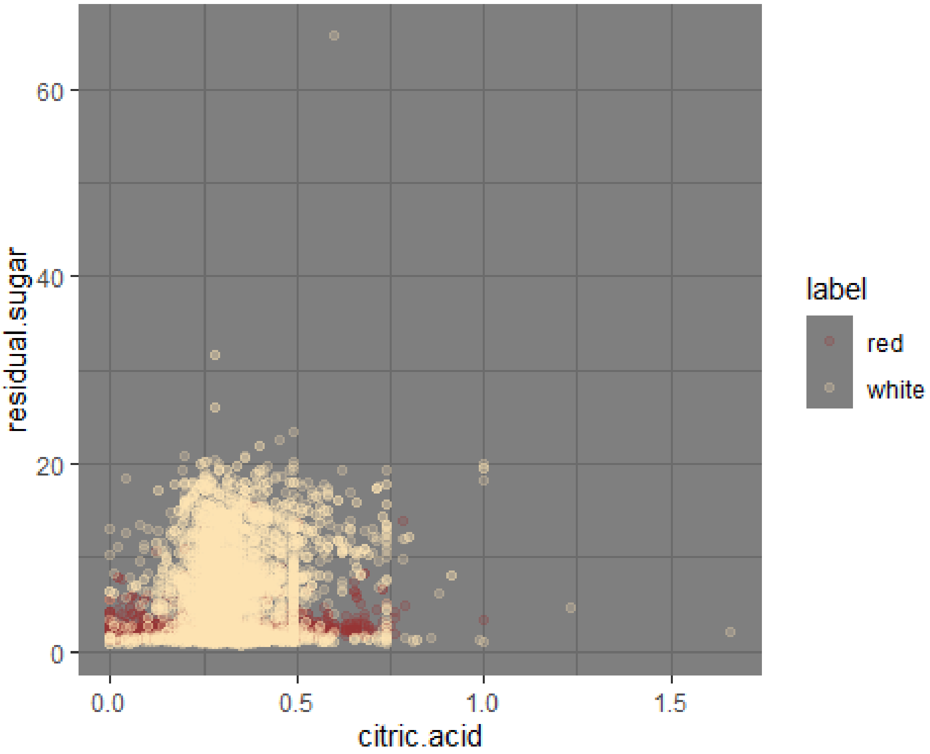
```
ggplot(wine,aes(quality)) +
geom_histogram(aes(fill=label),color='black',position='dodge',alpha=0.7)  +
scale_fill_manual(values = c('red','white')) +theme_bw()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



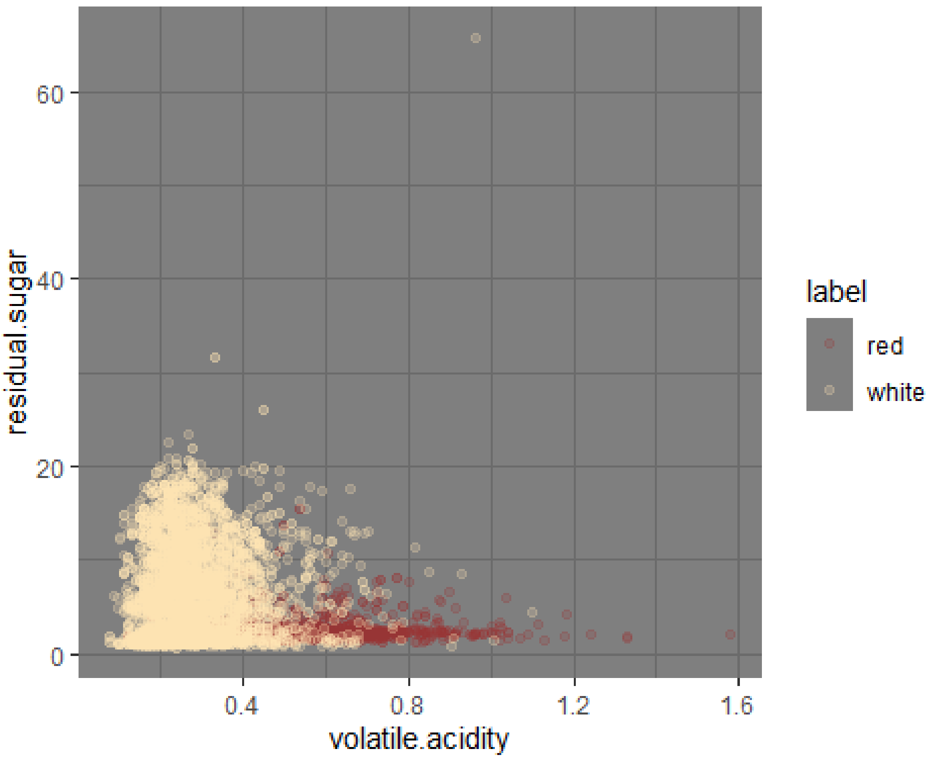
```
ggplot(wine,aes(fixed.acidity)) +
geom_histogram(aes(fill=label),color='black',position='dodge',alpha=0.7)  +
scale_fill_manual(values = c('red','white')) +theme_bw()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



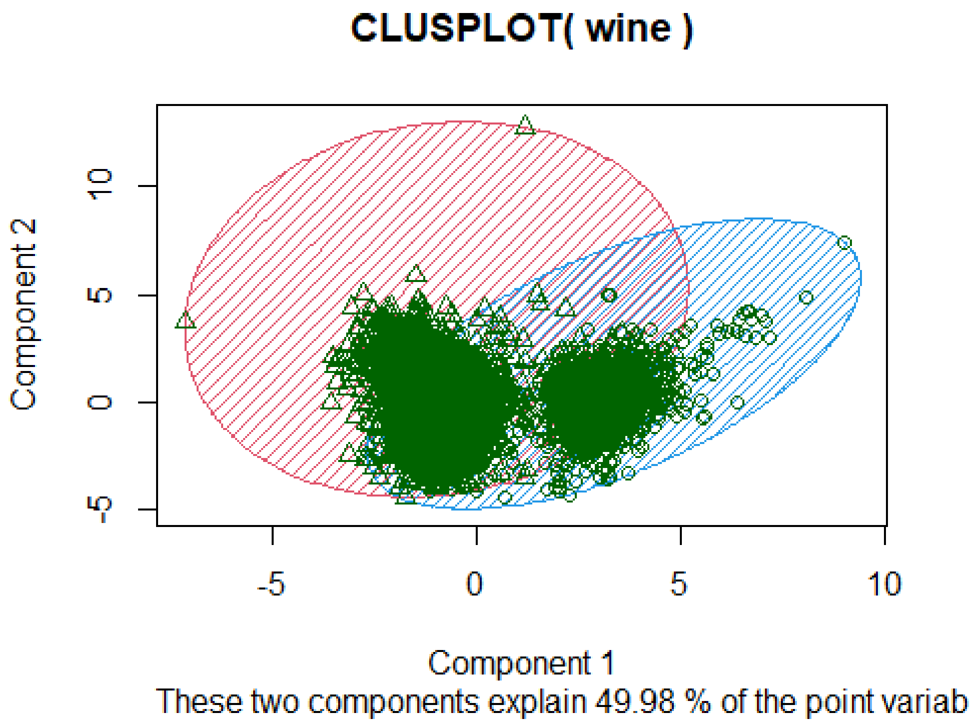
```
ggplot(wine,aes(citric.acid,residual.sugar)) + geom_point(aes(colour =
label),alpha=0.2) + scale_color_manual(values = c('#993333','#ffe5b4'))
+theme_dark()
```



```
ggplot(wine,aes(volatile.acidity,residual.sugar)) + geom_point(aes(colour =
label),alpha=0.2) + scale_color_manual(values = c('#993333','#ffe5b4'))
+theme_dark()
```



```
#### its going to be challenge to properly label them and separate them due
to the fact our closely packed each of them are
#### lets put the k means cluster in the act

model <- kmeans(wine[,1:12],2)

summary(model)

##              Length Class  Mode
## cluster      6497   -none- numeric
## centers        24   -none- numeric
## totss           1   -none- numeric
## withinss        2   -none- numeric
## tot.withinss    1   -none- numeric
## betweenss       1   -none- numeric
## size            2   -none- numeric
## iter            1   -none- numeric
## ifault          1   -none- numeric

table(wine$label,model$cluster)

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

clusplot(wine,model$cluster,color=T,labels = F,shade = T)
```



```
#### in this we had the privilege to know if the clustering is working by
using the label column but usually these are called
#### unsupervised clustering method meaning we cluster them without knowing
to which column to compare it to
```