Group 4 Project

Laura Larregui, Angela Garcia, John Christman

11/18/2019

```
readURL <- function(inputURL)  #Begin function named readURL that takes a URL
{
  csvFile <- read.csv(url(inputURL), sep = ';')  #assign the results of the
URL call as a csv file to a dataframe named csvFile. Added sep = ';' to
seperate the data into columns
  return(csvFile)  # return the dataframe
}

redWine <- readURL("https://archive.ics.uci.edu/ml/machine-learning-
databases/wine-quality/winequality-red.csv")
whiteWine <- readURL("https://archive.ics.uci.edu/ml/machine-learning-
databases/wine-quality/winequality-white.csv")

str(redWine)

## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58
0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069
0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36
3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57
0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...

str(whiteWine)

## 'data.frame':    4898 obs. of  12 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 ...
```

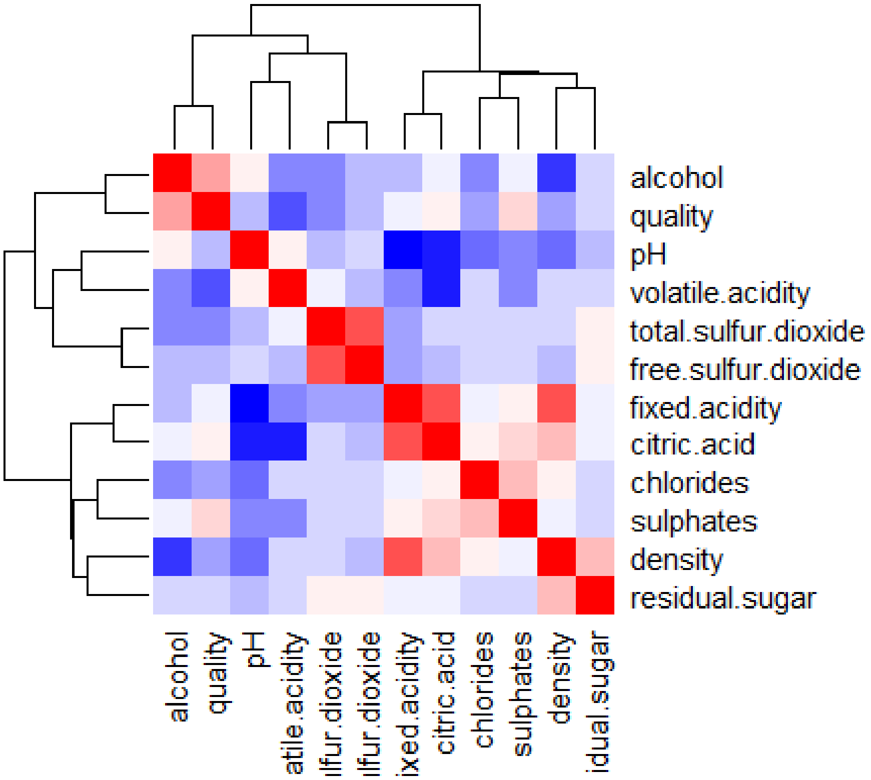
```
summary(redWine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

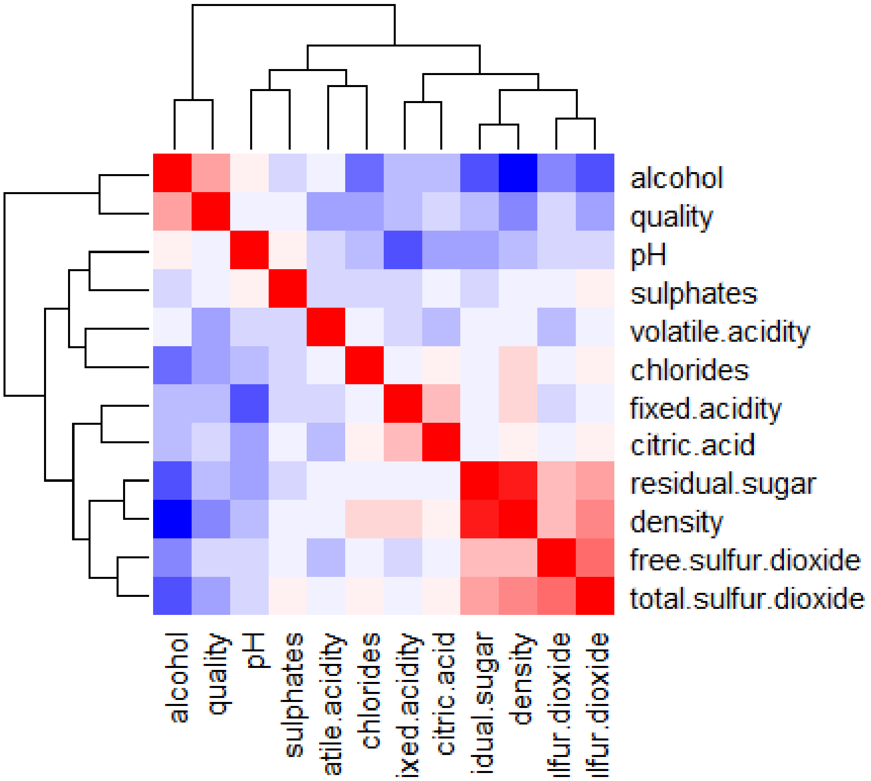
```
summary(whiteWine)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800  
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0       
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0       
##  Median :0.04300   Median : 34.00      Median :134.0       
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4       
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0       
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0       
##     density            pH          sulphates         alcohol     
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00  
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50  
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40  
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51  
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40  
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20  
##     quality     
##  Min.   :3.000  
##  1st Qu.:5.000  
##  Median :6.000  
##  Mean   :5.878  
##  3rd Qu.:6.000  
##  Max.   :9.000  
```

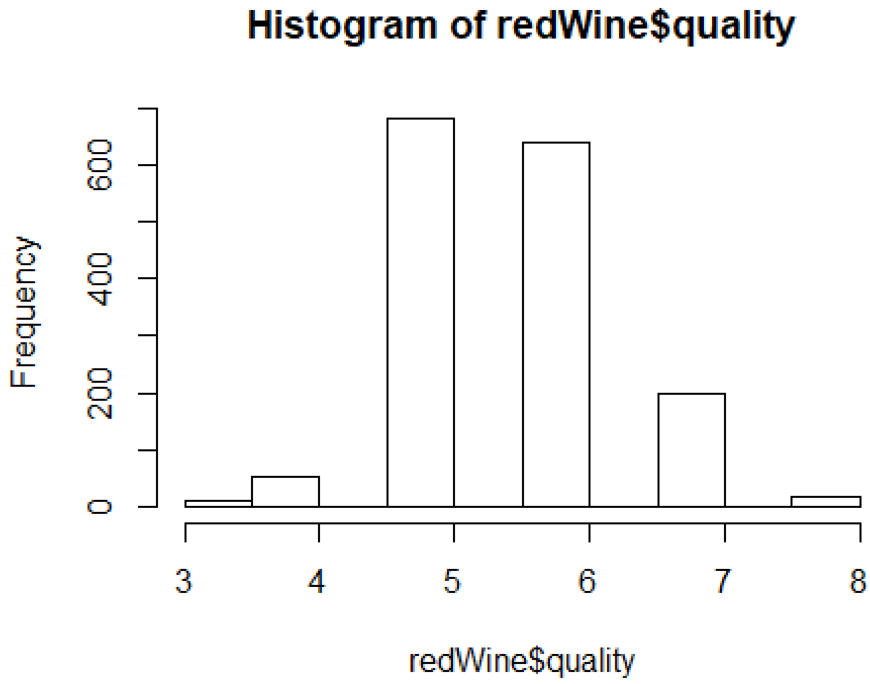
```
#THe datasets only have one column of data.  The column names are separated
by periods the data by semi-colons
#1. Create columns
#2.  separate the data into the columns
#3.  Verify no NAs
redWine <- na.omit(redWine)
whiteWine<-na.omit (whiteWine)

#1.  Create visulaizations for the data
#heat maps, histograms and scatter plots?

#Heatmaps
red_cor <- cor(redWine)
white_cor <- cor(whiteWine)
col<- colorRampPalette(c("blue", "white", "red"))(20)
heatmap(x = red_cor, col = col, symm = TRUE)
```



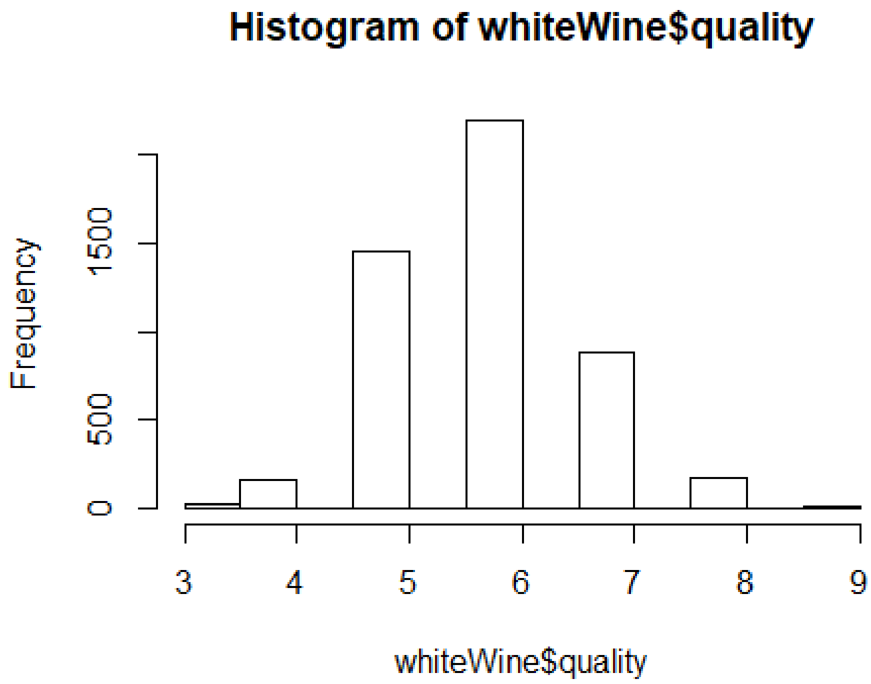
```r
heatmap(x = white_cor, col = col, symm = TRUE)
```



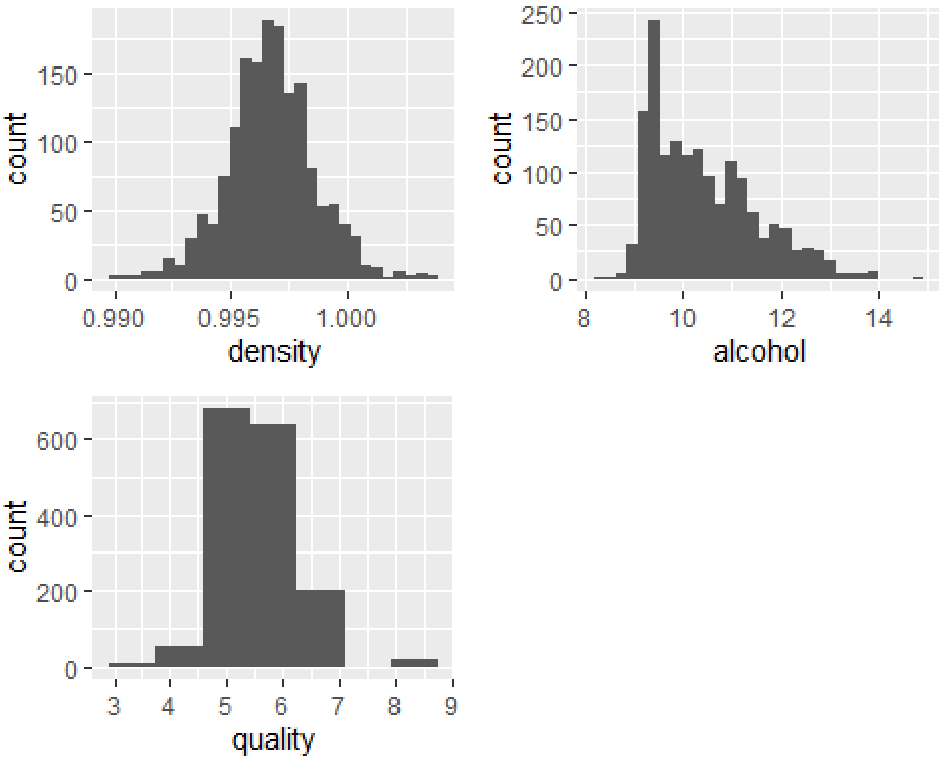
```
#Histograms
hist(redWine$quality)
```



```
hist(whiteWine$quality)

library(grid)
```

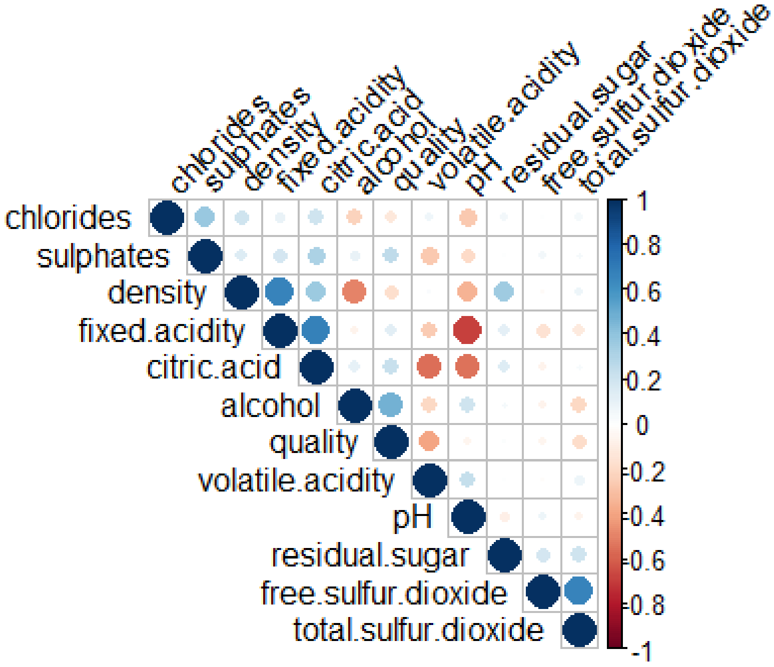


```r
library(gridExtra)
library (ggplot2)
h1 <- ggplot(aes(density), data = redWine) + geom_histogram(bins = 30)

h2 <- ggplot(aes(alcohol), data = redWine) + geom_histogram(bins = 30)

h3 <- ggplot(aes(quality), data = redWine) + geom_histogram(bins = 7)


grid.arrange(h1,h2,h3,ncol=2)
```

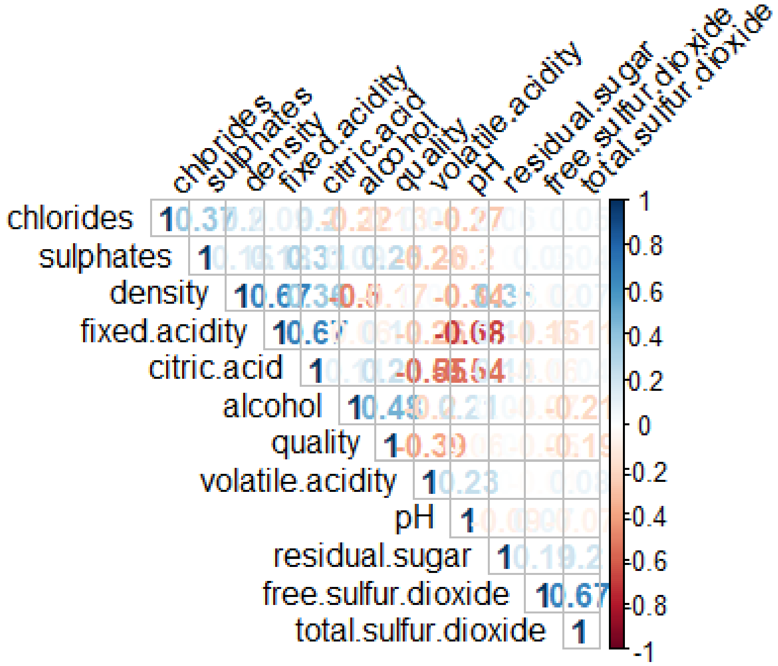


```
#1. Create the correlation matrix

#Red Wine Correlation Matrix

#install.packages("corrplot")
library(corrplot)

## corrplot 0.84 loaded

red_cor <- cor(redWine)
round(red_cor, 2)

##                      fixed.acidity volatile.acidity citric.acid
## fixed.acidity                 1.00            -0.26        0.67
## volatile.acidity             -0.26             1.00       -0.55
## citric.acid                   0.67            -0.55        1.00
## residual.sugar                0.11             0.00        0.14
## chlorides                     0.09             0.06        0.20
## free.sulfur.dioxide          -0.15            -0.01       -0.06
## total.sulfur.dioxide         -0.11             0.08        0.04
## density                       0.67             0.02        0.36
## pH                           -0.68             0.23       -0.54
## sulphates                     0.18            -0.26        0.31
## alcohol                      -0.06            -0.20        0.11
## quality                       0.12            -0.39        0.23
##                      residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                  0.11      0.09               -0.15
## volatile.acidity               0.00      0.06               -0.01
## citric.acid                    0.14      0.20               -0.06
## residual.sugar                 1.00      0.06                0.19
## chlorides                      0.06      1.00                0.01
## free.sulfur.dioxide            0.19      0.01                1.00
## total.sulfur.dioxide           0.20      0.05                0.67
## density                        0.36      0.20               -0.02
## pH                            -0.09     -0.27                0.07
## sulphates                      0.01      0.37                0.05
## alcohol                        0.04     -0.22               -0.07
## quality                        0.01     -0.13               -0.05
##                      total.sulfur.dioxide density    pH sulphates alcohol
## fixed.acidity                       -0.11    0.67 -0.68      0.18   -0.06
## volatile.acidity                     0.08    0.02  0.23     -0.26   -0.20
## citric.acid                          0.04    0.36 -0.54      0.31    0.11
## residual.sugar                       0.20    0.36 -0.09      0.01    0.04
## chlorides                            0.05    0.20 -0.27      0.37   -0.22
## free.sulfur.dioxide                  0.67   -0.02  0.07      0.05   -0.07
## total.sulfur.dioxide                 1.00    0.07 -0.07      0.04   -0.21
## density                              0.07    1.00 -0.34      0.15   -0.50
## pH                                  -0.07   -0.34  1.00     -0.20    0.21
## sulphates                            0.04    0.15 -0.20      1.00    0.09
## alcohol                             -0.21   -0.50  0.21      0.09    1.00
## quality                             -0.19   -0.17 -0.06      0.25    0.48
##                      quality
## fixed.acidity           0.12
## volatile.acidity       -0.39
## citric.acid             0.23
## residual.sugar          0.01
## chlorides              -0.13
## free.sulfur.dioxide    -0.05
## total.sulfur.dioxide   -0.19
## density                -0.17
## pH                     -0.06
## sulphates               0.25
## alcohol                 0.48
## quality                 1.00
```

```
corrplot(red_cor, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



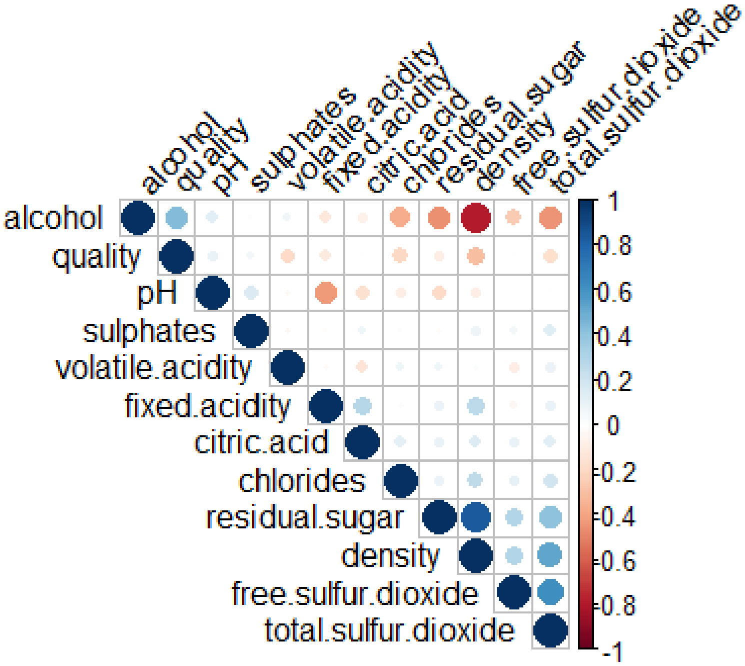
```
#Positive correlations are displayed in blue and negative correlations in red
color. Color intensity and the size of the circle are proportional to the
correlation coefficients.

#Correlation matrix with numbers
corrplot(red_cor, method = 'number', type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



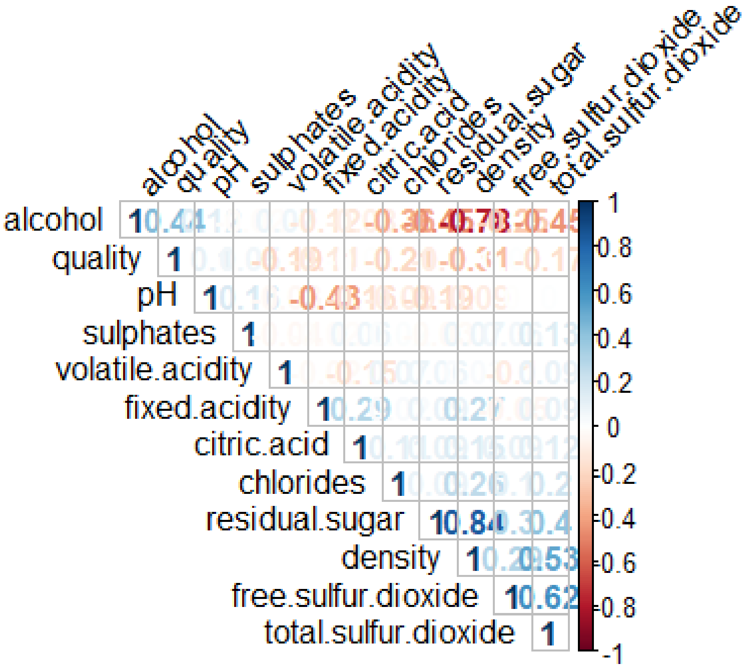
```
#White Wine Correlation Matrix

#install.packages("corrplot")
library(corrplot)
white_cor <- cor(whiteWine)
round(white_cor, 2)
```

```
##                      fixed.acidity volatile.acidity citric.acid
## fixed.acidity                 1.00            -0.02        0.29
## volatile.acidity             -0.02             1.00       -0.15
## citric.acid                   0.29            -0.15        1.00
## residual.sugar                0.09             0.06        0.09
## chlorides                     0.02             0.07        0.11
## free.sulfur.dioxide          -0.05            -0.10        0.09
## total.sulfur.dioxide          0.09             0.09        0.12
## density                       0.27             0.03        0.15
## pH                           -0.43            -0.03       -0.16
## sulphates                    -0.02            -0.04        0.06
## alcohol                      -0.12             0.07       -0.08
## quality                      -0.11            -0.19       -0.01
##                      residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                  0.09      0.02               -0.05
## volatile.acidity               0.06      0.07               -0.10
## citric.acid                    0.09      0.11                0.09
## residual.sugar                 1.00      0.09                0.30
## chlorides                      0.09      1.00                0.10
## free.sulfur.dioxide            0.30      0.10                1.00
## total.sulfur.dioxide           0.40      0.20                0.62
## density                        0.84      0.26                0.29
## pH                            -0.19     -0.09                0.00
## sulphates                     -0.03      0.02                0.06
## alcohol                       -0.45     -0.36               -0.25
## quality                       -0.10     -0.21                0.01
##                      total.sulfur.dioxide density    pH sulphates alcohol
## fixed.acidity                        0.09    0.27 -0.43     -0.02   -0.12
## volatile.acidity                     0.09    0.03 -0.03     -0.04    0.07
## citric.acid                          0.12    0.15 -0.16      0.06   -0.08
## residual.sugar                       0.40    0.84 -0.19     -0.03   -0.45
## chlorides                            0.20    0.26 -0.09      0.02   -0.36
## free.sulfur.dioxide                  0.62    0.29  0.00      0.06   -0.25
## total.sulfur.dioxide                 1.00    0.53  0.00      0.13   -0.45
## density                              0.53    1.00 -0.09      0.07   -0.78
## pH                                   0.00   -0.09  1.00      0.16    0.12
## sulphates                            0.13    0.07  0.16      1.00   -0.02
## alcohol                             -0.45   -0.78  0.12     -0.02    1.00
## quality                             -0.17   -0.31  0.10      0.05    0.44
##                      quality
## fixed.acidity          -0.11
## volatile.acidity       -0.19
## citric.acid            -0.01
## residual.sugar         -0.10
## chlorides              -0.21
## free.sulfur.dioxide     0.01
## total.sulfur.dioxide   -0.17
## density                -0.31
## pH                      0.10
## sulphates               0.05
## alcohol                 0.44
## quality                 1.00
```

```
corrplot(white_cor, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



```
#Correlation matrix with numbers
corrplot(white_cor, method = 'number',type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



```
##Reference: http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software

#Machine learning techniques to see if we can train the system to pick a good wine
```