M2.851 Tipología y ciclo de vida de los datos - UOC

Práctica 2. Limpieza y validación de los datos

11/06/2018

Alberto Gómez (agomezma@uoc.edu)

Introducción

En este *notebook* se desarrolla la práctica 2 de limpieza y validación de datos de la asignatura **Tipología y ciclo de vida de los datos**.

Se ha elegido uno de los dataset propuestos, Red Wine Quality.

El desarrollo se ha hecho en un *notebook* de Jupyter con un *kernel* de R. (También se proporciona un fichero HTML con el desarrollo y el código en R exportado.)

De esta forma, se integran, en un único documento, las respuestas a las preguntas planteadas en el enunciado con el código en R con el que se realiza el trabajo y los gráficos y tablas que sirven para explicar los resultados.

1. Descripción del *dataset*

Inicialmente se eligió el *dataset* de Red Wine Quality que se puede descargar de la siguiente dirección:

https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/data (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/data)

El *dataset* original está en https://archive.ics.uci.edu/ml/datasets/wine+quality (https://archive.ics.uci.edu/ml/datasets/wine+quality) y realmente está compuesto de dos conjuntos de datos: uno de vinos blancos y otro de vinos tintos.

Para que el problema de integración tuviera más interés se ha decidido partir de los dos conjuntos de datos y unificarlos. Desde el punto de vista del análisis de los datos, realmente no tiene sentido unirlos en un único conjunto, porque las características quimicas de ambos tipos de vino son distintas, como veremos en el apartado de análisis.

Ambos conjuntos tienen las mismas 11 variables, que representan distintas características físico-químicas de 6497 vinos portugueses: 1 - fixed acidity 2 - volatile acidity 3 - citric acid 4 - residual sugar 5 - chlorides 6 - free sulfur dioxide 7 - total sulfur dioxide 8 - density 9 - pH 10 - sulphates 11 - alcohol

La variable de salida (12) es un valor entre 0 y 10 que representa la calidad (*quality*) del vino según un análisis sensorial: 12 - quality (score between 0 and 10)

En este *dataset* no hay identificadores para los vinos. Como indican los autores del conjunto de datos en su descripción, las clases no están equilibradas (hay muchos más vinos normales que muy bueno o muy malos) y los datos aparecen ordenados por su calidad.

1.1. Interés y objetivos del análisis

Este dataset es interesante para realizar pruebas de algoritmos de clasificación y predicción.

Sería muy útil poder predecir la calidad final del vino a partir de algunas de las características extraídas de un análisis químico. Además, también serviría para clasificar los vinos según sean sus características.

Con este conocimiento se podría, posteriormente, un productor de vinos podría actuar sobre la uva o sobre el proceso de vinificación para intentar obtener vinos de mejor calidad.

Los objetivos de este trabajo son:

- Conocer los datos de los data sets elegidos
- Integrarlos y limpiar los datos
- Hacer un análisis básico de los datos

1.2. Carga y resumen de los datos

Con el siguiente código se cargan los dos ficheros en formato CSV que se pueden descargar desde el repositorio https://archive.ics.uci.edu/ml/datasets/wine+quality (https://archive.ics.uci.edu/ml/datasets/wine+quality).

Además, se muestran, para cada dataset cargado:

- número de registros
- número de campos
- nombre y tipo de los campos
- primeros datos
- resumen de cada variable (máximo, mínimo, cuartiles, media y mediana)

En el resumen, si faltaran datos de alguna variable, aparecería el número de valores no disponibles.

In [1]:

```
# Los dos ficheros deben estar en el mismo directorio donde se ejecuta el código
 en R

# Carga del fichero de vinos blancos y resumen de datos
cat("VINOS BLANCOS \n")
cat("Fichero de datos: winequality-white.csv  \n")
whitewines = read.csv("winequality-white.csv", header = TRUE, sep=";")
cat("Número de registros, número de campos \n")
dim(whitewines)
cat("Tipos de los campos \n")
sapply(whitewines, function(x) class(x))
cat("Primeras filas \n")
head(whitewines)
cat("Resumen \n")
summary(whitewines)
cat("\n\n")

# Carga del fichero de vinos tintos y resumen de dato
cat("VINOS TINTOS \n")
cat("Fichero de datos: winequality-red.csv  \n")
redwines = read.csv("winequality-red.csv", header = TRUE, sep=";")
cat("Número de registros, número de campos \n")
dim(redwines)
cat("Tipos de los campos \n")
sapply(redwines, function(x) class(x))
cat("Primeras filas  \n")
head(redwines)
cat("Resumen \n")
summary(redwines)
cat("\n\n")
```

```
VINOS BLANCOS
Fichero de datos: winequality-white.csv
Número de registros, número de campos

    4898  12

Tipos de los campos
```

fixed.acidity
'numeric'
volatile.acidity
'numeric'
citric.acid
'numeric'
residual.sugar
'numeric'
chlorides
'numeric'
free.sulfur.dioxide
'numeric'
total.sulfur.dioxide
'numeric'
density
'numeric'
pH
'numeric'
sulphates
'numeric'
alcohol
'numeric'
quality
'integer'

Primeras filas

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	t
7.0	0.27	0.36	20.7	0.045	45	1
6.3	0.30	0.34	1.6	0.049	14	1
8.1	0.28	0.40	6.9	0.050	30	9
7.2	0.23	0.32	8.5	0.058	47	1
7.2	0.23	0.32	8.5	0.058	47	1
8.1	0.28	0.40	6.9	0.050	30	9

Resumen

```
 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    densi
ty
 Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :
0.9871
 1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:
0.9917
 Median :0.04300   Median : 34.00      Median :134.0        Median :
0.9937
 Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :
0.9940
 3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:
0.9961
 Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :
1.0390
       pH          sulphates         alcohol         quality
 Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
 1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
 Median :3.180   Median :0.4700   Median :10.40   Median :6.000
 Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
 3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
 Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000


VINOS TINTOS
Fichero de datos: winequality-red.csv
Número de registros, número de campos

    1599  12

Tipos de los campos
```

fixed.acidity
'numeric'
volatile.acidity
'numeric'
citric.acid
'numeric'
residual.sugar
'numeric'
chlorides
'numeric'
free.sulfur.dioxide
'numeric'
total.sulfur.dioxide
'numeric'
density
'numeric'
pH
'numeric'
sulphates
'numeric'
alcohol
'numeric'
quality
'integer'

Primeras filas

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	t
7.4	0.70	0.00	1.9	0.076	11	3
7.8	0.88	0.00	2.6	0.098	25	6
7.8	0.76	0.04	2.3	0.092	15	5
11.2	0.28	0.56	1.9	0.075	17	6
7.4	0.70	0.00	1.9	0.076	11	3
7.4	0.66	0.00	1.8	0.075	13	4

Resumen

```
 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    densi
ty
 Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :
0.9901
 1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:
0.9956
 Median :0.07900   Median :14.00       Median : 38.00       Median :
0.9968
 Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :
0.9967
 3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:
0.9978
 Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :
1.0037
       pH           sulphates         alcohol         quality
 Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
 1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
 Median :3.310   Median :0.6200   Median :10.20   Median :6.000
 Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
 3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

2. Integración y selección de los datos de interés

2.1. Integración de los dos conjuntos de datos

Se van a unir ambos *datasets* en un único *data.frame*, añadiendo un campo **colour** que indique el tipo (color) del vino (0 para los blancos, 1 para los tintos).

Se podría haber añadido un código distinto o una cadena de texto para diferenciar el tipo de vino.

A continuación se puede ver el código para unificar los *datasets*, junto con un resumen de las características del nuevo conjunto.

En el nuevo conjunto quedan los vinos blancos al principio y los tintos, al final.

In [2]:

```
# Se añade el campo color (colour) a los dos dataframes: 0 para los vinos blanco
s, 1 para los tintos

cat("VINOS BLANCOS \n")
whitewines = cbind(colour = rep(0L,nrow(whitewines)), whitewines)
cat("Número de registros, número de campos \n")
dim(whitewines)
cat("Nombres de los campos \n")
names (whitewines)
cat("Tipos de los campos \n")
sapply(whitewines, function(x) class(x))

cat("VINOS TINTOS \n")
redwines = cbind(colour = rep(1L,nrow(redwines)), redwines)

cat("Número de registros, número de campos \n")
dim(redwines)
cat("Nombres de los campos \n")
names (redwines)
cat("Nombres de los campos \n")
names (redwines)

# Integrar en un dataframe los dos anterioress
wines = rbind(whitewines, redwines)

# Resumen del nuevo conjunto
cat("VINOS \n")
cat("Número de registros, número de campos \n")
dim(wines)
cat("Nombres de los campos \n")
names(wines)
cat("Tipos de los campos \n")
sapply(wines, function(x) class(x))

cat("Primeras filas  \n")
head(wines)
cat("Últimas filas  \n")
tail(wines)
cat("Resumen  \n")
summary(wines)
```

VINOS BLANCOS
Número de registros, número de campos

```
    4898  13
```

Nombres de los campos

```
    'colour'  'fixed.acidity'  'volatile.acidity'  'citric.acid'  'residual.sugar'
    'chlorides'  'free.sulfur.dioxide'  'total.sulfur.dioxide'  'density'  'pH'
    'sulphates'  'alcohol'  'quality'
```

Tipos de los campos

colour
'integer'
fixed.acidity
'numeric'
volatile.acidity
'numeric'
citric.acid
'numeric'
residual.sugar
'numeric'
chlorides
'numeric'
free.sulfur.dioxide
'numeric'
total.sulfur.dioxide
'numeric'
density
'numeric'
pH
'numeric'
sulphates
'numeric'
alcohol
'numeric'
quality
'integer'

VINOS TINTOS
Número de registros, número de campos

```
    1599  13
```

Nombres de los campos

```
    'colour'  'fixed.acidity'  'volatile.acidity'  'citric.acid'  'residual.sugar'
    'chlorides'  'free.sulfur.dioxide'  'total.sulfur.dioxide'  'density'  'pH'
    'sulphates'  'alcohol'  'quality'
```

Nombres de los campos

```
    'colour'  'fixed.acidity'  'volatile.acidity'  'citric.acid'  'residual.sugar'
    'chlorides'  'free.sulfur.dioxide'  'total.sulfur.dioxide'  'density'  'pH'
    'sulphates'  'alcohol'  'quality'
```

VINOS
Número de registros, número de campos

6497 13

Nombres de los campos

'colour' 'fixed.acidity' 'volatile.acidity' 'citric.acid' 'residual.sugar' 'chlorides' 'free.sulfur.dioxide' 'total.sulfur.dioxide' 'density' 'pH' 'sulphates' 'alcohol' 'quality'

Tipos de los campos

colour
'integer'
fixed.acidity
'numeric'
volatile.acidity
'numeric'
citric.acid
'numeric'
residual.sugar
'numeric'
chlorides
'numeric'
free.sulfur.dioxide
'numeric'
total.sulfur.dioxide
'numeric'
density
'numeric'
pH
'numeric'
sulphates
'numeric'
alcohol
'numeric'
quality
'integer'

Primeras filas

colour	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.di
0	7.0	0.27	0.36	20.7	0.045	45
0	6.3	0.30	0.34	1.6	0.049	14
0	8.1	0.28	0.40	6.9	0.050	30
0	7.2	0.23	0.32	8.5	0.058	47
0	7.2	0.23	0.32	8.5	0.058	47
0	8.1	0.28	0.40	6.9	0.050	30

Últimas filas

	colour	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.su
6492	1	6.8	0.620	0.08	1.9	0.068	28
6493	1	6.2	0.600	0.08	2.0	0.090	32
6494	1	5.9	0.550	0.10	2.2	0.062	39
6495	1	6.3	0.510	0.13	2.3	0.076	29
6496	1	5.9	0.645	0.12	2.0	0.075	32
6497	1	6.0	0.310	0.47	3.6	0.067	18

Resumen

```
     colour          fixed.acidity     volatile.acidity  citric.acid
 Min.   :0.0000   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500
 Median :0.0000   Median : 7.000   Median :0.2900   Median :0.3100
 Mean   :0.2461   Mean   : 7.215   Mean   :0.3397   Mean   :0.3186
 3rd Qu.:0.0000   3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900
 Max.   :1.0000   Max.   :15.900   Max.   :1.5800   Max.   :1.6600
 residual.sugar     chlorides       free.sulfur.dioxide total.sulfu
r.dioxide
 Min.   : 0.600   Min.   :0.00900   Min.   :  1.00      Min.   :  6.
0
 1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.
0
 Median : 3.000   Median :0.04700   Median : 29.00      Median :118.
0
 Mean   : 5.443   Mean   :0.05603   Mean   : 30.53      Mean   :115.
7
 3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.
0
 Max.   :65.800   Max.   :0.61100   Max.   :289.00      Max.   :440.
0
    density             pH          sulphates         alcohol
 Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
 1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50
 Median :0.9949   Median :3.210   Median :0.5100   Median :10.30
 Mean   :0.9947   Mean   :3.219   Mean   :0.5313   Mean   :10.49
 3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30
 Max.   :1.0390   Max.   :4.010   Max.   :2.0000   Max.   :14.90
    quality
 Min.   :3.000
 1st Qu.:5.000
 Median :6.000
 Mean   :5.818
 3rd Qu.:6.000
 Max.   :9.000
```

2.2. Datos repetidos

Vamos a comprobar si hay datos repetidos en el nuevo *dataset*. Esta tarea también se podría haber hecho individualmente en cada uno de los *datasets* originales.

In [3]:

```
### Duplicated da los duplicados, no la primera aparición, y por eso no los veo
 en repetidos
cat("Numero de registros originales: ", nrow(wines), "\n")
sinrep = wines[!duplicated(wines),]
cat("Numero de registro sin repetir: ", nrow(sinrep), "\n")

repetidos = wines[duplicated(wines),]

cat("Numero de registro repetidos: ", nrow(repetidos), "\n")

x = nrow(repetidos)+nrow(sinrep)

cat("Total (igual al número de registros originales): ", x, "\n")

cat("Porcentaje de repetidos: ", nrow(repetidos)/nrow(wines), "\n")
```

```
Numero de registros originales:  6497
Numero de registro sin repetir:  5320
Numero de registro repetidos:  1177
Total (igual al número de registros originales):  6497
Porcentaje de repetidos:  0.1811605
```

Los valores repetidos indican que hay vinos que tienen las mismas características químicas.

Hay bastantes valores repetidos (un 18%) pero no los vamos a eliminar.

Se supone que no son errores del *dataset*.

2.3. Campo identificador

Vamos a añadir un campo identificador numérico. Esto podría ser útil para recuperar toda la información de algún dato concreto (por ejemplo, en la detección de *outliars*).

Además del *data.frame* **wines**, tendremos los *data.frame* **white_wines** y **red_wines** que agrupan a los vinos blancos y tintos.

In [4]:

```
# Se añade un campo id

wines = cbind(id=rep(1:nrow(wines)), wines)
# Resumen del nuevo conjunto
cat("VINOS \n")
cat("Número de registros, número de campos \n")
dim(wines)
cat("Nombres de los campos \n")
names(wines)
cat("Tipos de los campos \n")
sapply(wines, function(x) class(x))

cat("Primeras filas  \n")
head(wines)
cat("Últimas filas  \n")
tail(wines)
cat("Resumen  \n")
summary(wines)


# para simplificar cuando tengamos que usar solo vinos blancos o tintos, definim
os los data frame siguientes:
white_wines = wines[wines$colour==0,]
red_wines = wines[,][wines$colour==1,]
```

VINOS
Número de registros, número de campos

6497 14

Nombres de los campos

'id' 'colour' 'fixed.acidity' 'volatile.acidity' 'citric.acid' 'residual.sugar'
'chlorides' 'free.sulfur.dioxide' 'total.sulfur.dioxide' 'density' 'pH'
'sulphates' 'alcohol' 'quality'

Tipos de los campos

id
'integer'
colour
'integer'
fixed.acidity
'numeric'
volatile.acidity
'numeric'
citric.acid
'numeric'
residual.sugar
'numeric'
chlorides
'numeric'
free.sulfur.dioxide
'numeric'
total.sulfur.dioxide
'numeric'
density
'numeric'
pH
'numeric'
sulphates
'numeric'
alcohol
'numeric'
quality
'integer'

Primeras filas

id	colour	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfu
1	0	7.0	0.27	0.36	20.7	0.045	45
2	0	6.3	0.30	0.34	1.6	0.049	14
3	0	8.1	0.28	0.40	6.9	0.050	30
4	0	7.2	0.23	0.32	8.5	0.058	47
5	0	7.2	0.23	0.32	8.5	0.058	47
6	0	8.1	0.28	0.40	6.9	0.050	30

Últimas filas

	id	colour	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	f
6492	6492	1	6.8	0.620	0.08	1.9	0.068	2
6493	6493	1	6.2	0.600	0.08	2.0	0.090	3
6494	6494	1	5.9	0.550	0.10	2.2	0.062	3
6495	6495	1	6.3	0.510	0.13	2.3	0.076	2
6496	6496	1	5.9	0.645	0.12	2.0	0.075	3
6497	6497	1	6.0	0.310	0.47	3.6	0.067	1

Resumen

```
       id            colour       fixed.acidity    volatile.acidity
 Min.   :   1   Min.   :0.0000   Min.   : 3.800   Min.   :0.0800
 1st Qu.:1625   1st Qu.:0.0000   1st Qu.: 6.400   1st Qu.:0.2300
 Median :3249   Median :0.0000   Median : 7.000   Median :0.2900
 Mean   :3249   Mean   :0.2461   Mean   : 7.215   Mean   :0.3397
 3rd Qu.:4873   3rd Qu.:0.0000   3rd Qu.: 7.700   3rd Qu.:0.4000
 Max.   :6497   Max.   :1.0000   Max.   :15.900   Max.   :1.5800
  citric.acid     residual.sugar     chlorides       free.sulfur.dio
xide
 Min.   :0.0000   Min.   : 0.600   Min.   :0.00900   Min.   :  1.00

 1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00

 Median :0.3100   Median : 3.000   Median :0.04700   Median : 29.00

 Mean   :0.3186   Mean   : 5.443   Mean   :0.05603   Mean   : 30.53

 3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00

 Max.   :1.6600   Max.   :65.800   Max.   :0.61100   Max.   :289.00

 total.sulfur.dioxide    density             pH          sulphates

 Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.220
0
 1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.430
0
 Median :118.0        Median :0.9949   Median :3.210   Median :0.510
0
 Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.531
3
 3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.600
0
 Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.000
0
    alcohol         quality
 Min.   : 8.00   Min.   :3.000
 1st Qu.: 9.50   1st Qu.:5.000
 Median :10.30   Median :6.000
 Mean   :10.49   Mean   :5.818
 3rd Qu.:11.30   3rd Qu.:6.000
 Max.   :14.90   Max.   :9.000
```

2.4. Selección de variables

En principio, todas las variables que se incluyen en el dataset son potencialmente interesantes, así que, en principio, se mantienen todas.

Posteriormente, en el análisis de los datos, veremos si están correlacionadas y se pueden eliminar algunas de ellas.

Tampoco tengo conocimientos sobre vinos para poder seleccionar un subconjunto de características más importantes.

3. Limpieza de los datos

Los conjuntos de datos originales presentan información completa en todos los registros, así que la tarea de limpieza es bastante sencilla.

3.1. Datos vacíos y nulos

En los resúmenes de los conjuntos de datos hemos visto que no hay datos vacíos. Si los hubiera, habría salido el número de NA en cada campo.

Además, en el cuadro siguiente contamos, de otra forma, el número de datos vacíos (que es 0 en todos los campos).

Hay valores nulos (0) en el campo **citric.acid**, como se puede ver en el resumen, pero es un valor posible para ese componente.

Si hubiera habido datos vacíos (NA) se podría haber asignados valores según los repetidos iguales, medias, etc. Si hubiera habido pocos casos, también se podrían haber eliminado, al ser los conjuntos bastante numerosos.

In [5]:

```
# Número de valores nulos en cada variable:
sapply(wines, function(x) sum(is.na(x)))
```

id
0
colour
0
fixed.acidity
0
volatile.acidity
0
citric.acid
0
residual.sugar
0
chlorides
0
free.sulfur.dioxide
0
total.sulfur.dioxide
0
density
0
pH
0
sulphates
0
alcohol
0
quality
0

3.2. Identificación y tratamiento de valores extremos

Vamos a buscar valores atípicos en el *dataset*.

He buscado información por Internet de los rangos de valores habituales de estas características y, para los que he encontrado, los datos del resumen entran dentro de los rangos posibles.

Hay que tener en cuenta que los autores del *dataset* ya indicaban que hay muchos vinos de calidad media y pocos de calidades extremas (muy malos o muy buenos). Si aparecen *outliers* relacionados con esas calidades, quizás no sean valores atípicos, sino valores adecuados para esas calidades extremas.

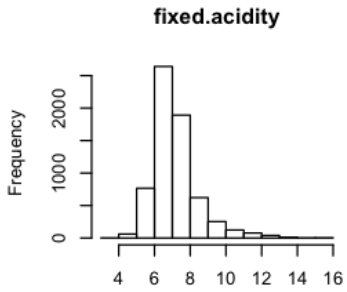
En principio, solo vamos a identificar los valores extremos, sin eliminarlos.

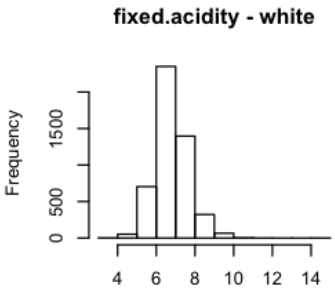
Primero vamos a representar, para cada variable, su histograma según las calidades y su *boxplot*, para ver cómo se distribuyen los datos.

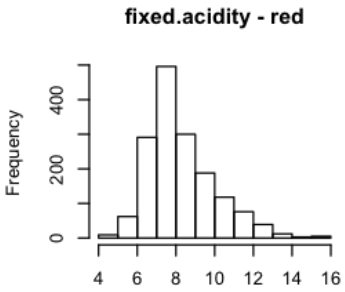
Como es muy probable que los valores sean distintos para vinos blancos y tintos, se van a representar también la misma información para cada uno de esos subconjuntos.

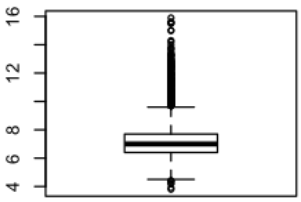
In [6]:

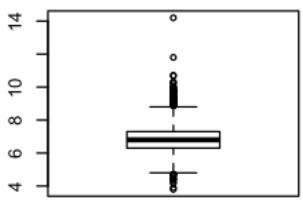
```
# Código
# Para que no salgan demasiado grandes, se divide el tamaño normal de un gráfico
 en una matriz de 3 por 3
# Representamos en una fila los histogramas del conjunto completo, solo de blanc
os y solo de tintos,
# y en la fila siguiente, el boxpolot correspondiente
par(mfrow=c(3,3))
for (i in seq(3, 14)) {
    hist(wines[,i], xlab="", main=names(wines[i]))
    hist(white_wines[,i], xlab="", main=paste(names(wines[i]),"- white"))
    hist(red_wines[,i], xlab="", main=paste(names(wines[i]),"- red"))
    boxplot(wines[,i])
    boxplot(white_wines[,i])
    boxplot(red_wines[,i])
}
```

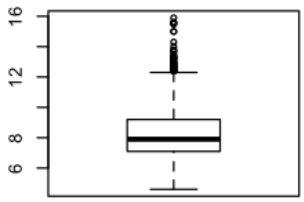

fixed.acidity
Frequency
0
1000
2000
4
6
8
10
12
14
16

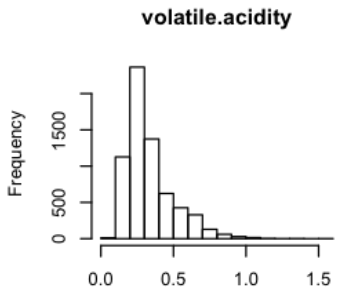

fixed.acidity - white
Frequency
0
500
1500
4
6
8
10
12
14

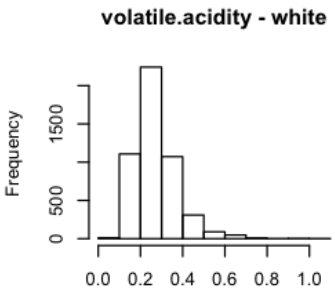

fixed.acidity - red
Frequency
0
200
400
4
6
8
10
12
14
16

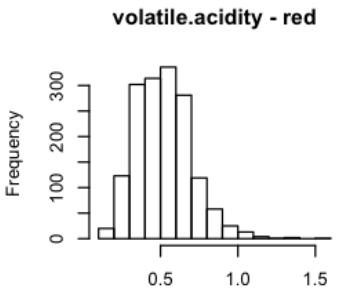

4
6
8
12
16

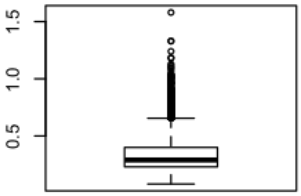

4
6
8
10
14

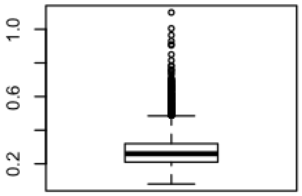

6
8
12
16

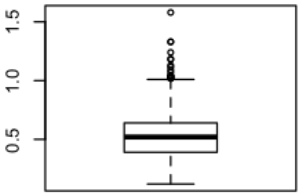

volatile.acidity
Frequency
0
500
1500
0.0
0.5
1.0
1.5

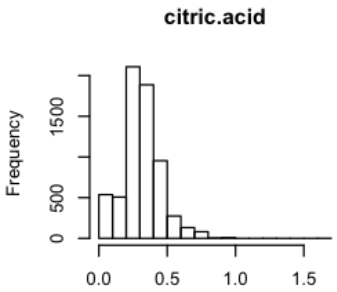

volatile.acidity - white
Frequency
0
500
1500
0.0
0.2
0.4
0.6
0.8
1.0

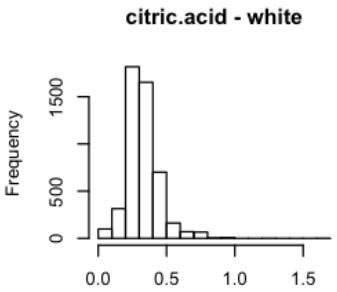

volatile.acidity - red
Frequency
0
100
200
300
0.5
1.0
1.5

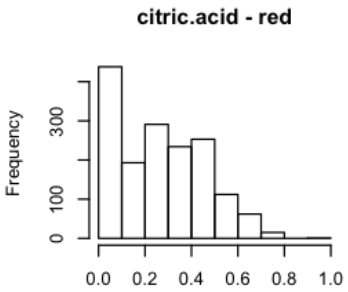

0.5
1.0
1.5

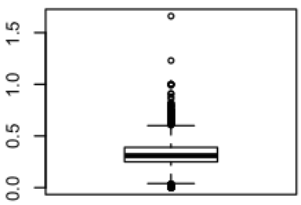

0.2
0.6
1.0

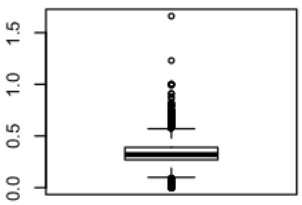

0.5
1.0
1.5

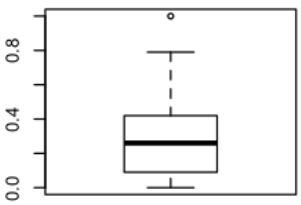

citric.acid
Frequency
0
500
1500
0.0
0.5
1.0
1.5

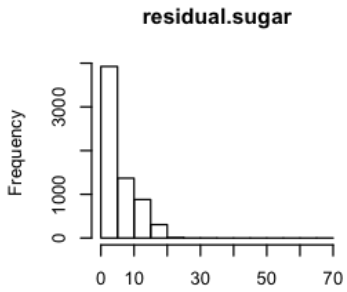

citric.acid - white
Frequency
0
500
1500
0.0
0.5
1.0
1.5

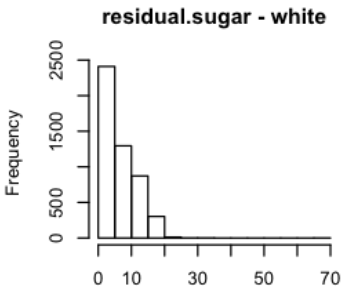

citric.acid - red
Frequency
0
100
300
0.0
0.2
0.4
0.6
0.8
1.0

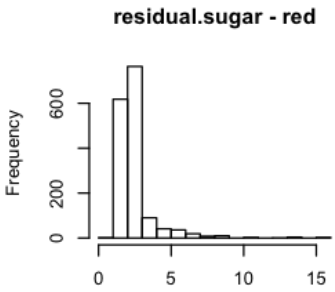

0.0
0.5
1.0
1.5

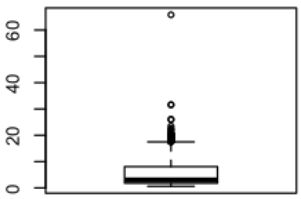

0.0
0.5
1.0
1.5

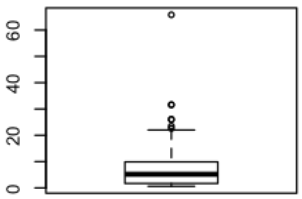

0.0
0.4
0.8

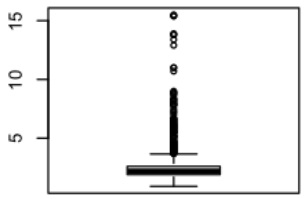

residual.sugar
Frequency
0
1000
3000
0
10
30
50
70

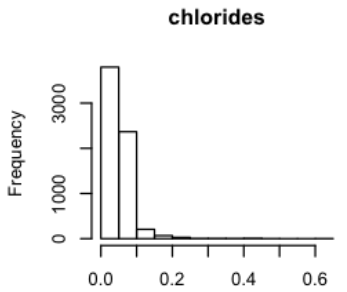

residual.sugar - white
Frequency
0
500
1500
2500
0
10
30
50
70

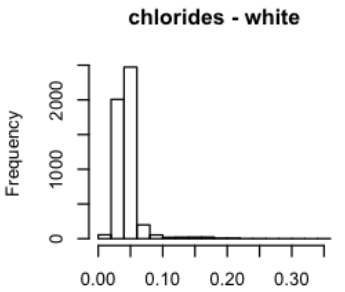

residual.sugar - red
Frequency
0
200
600
0
5
10
15

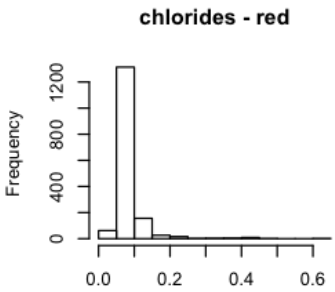

0
20
40
60

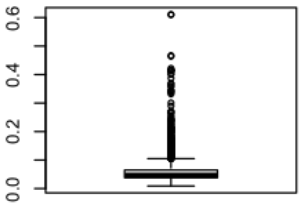

0
20
40
60

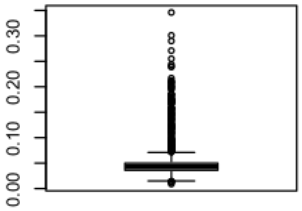

5
10
15

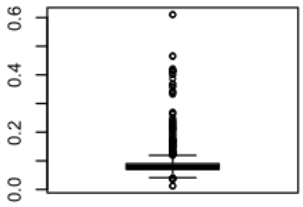

chlorides
Frequency
0
1000
3000
0.0
0.2
0.4
0.6

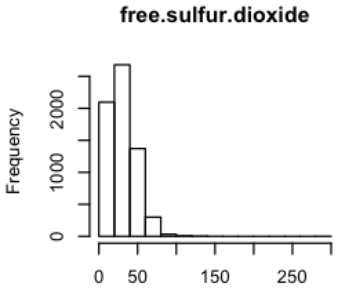

chlorides - white
Frequency
0
1000
2000
0.00
0.10
0.20
0.30

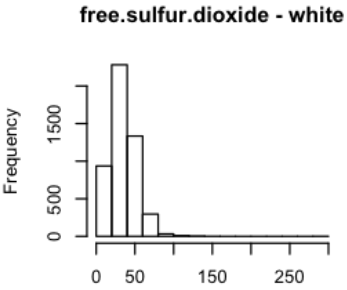

chlorides - red
Frequency
0
400
800
1200
0.0
0.2
0.4
0.6

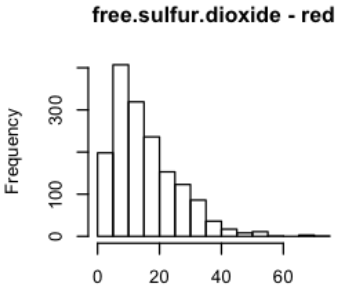

0.0
0.2
0.4
0.6

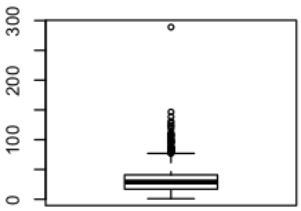

0.00
0.10
0.20
0.30

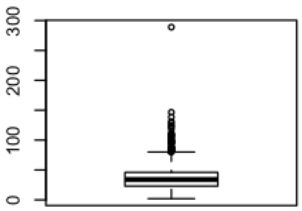

0.0
0.2
0.4
0.6

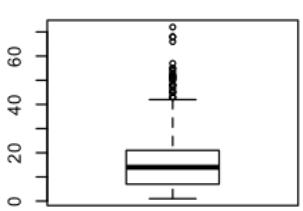

free.sulfur.dioxide
Frequency
0
1000
2000
0
50
150
250

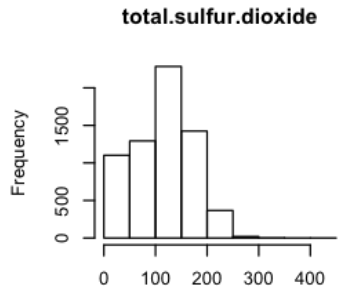

free.sulfur.dioxide - white
Frequency
0
500
1500
0
50
150
250

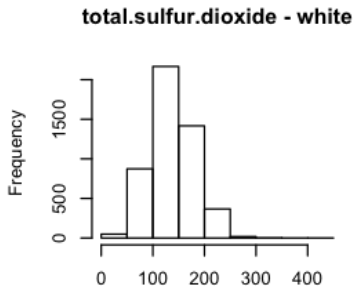

free.sulfur.dioxide - red
Frequency
0
100
300
0
20
40
60

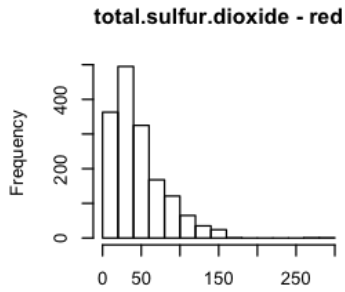

0
100
200
300

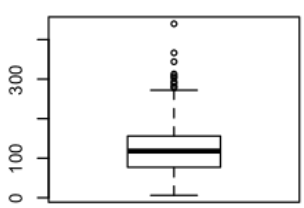

0
100
200
300

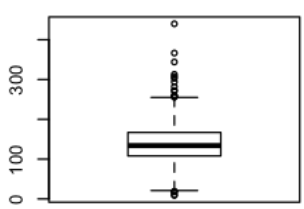

0
20
40
60

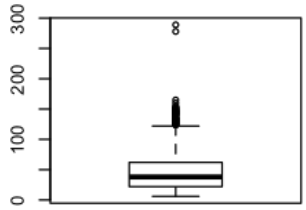

total.sulfur.dioxide
Frequency
0
500
1500
0
100
200
300
400

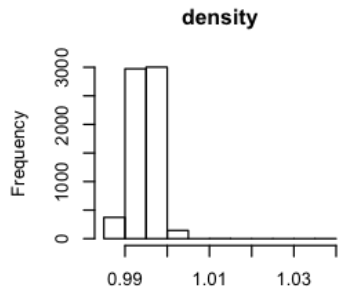

total.sulfur.dioxide - white
Frequency
0
500
1500
0
100
200
300
400

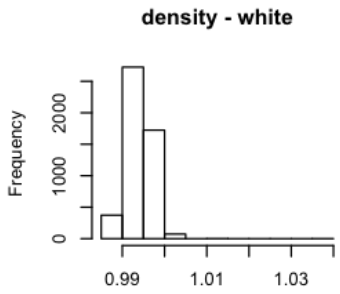

total.sulfur.dioxide - red
Frequency
0
200
400
0
50
150
250

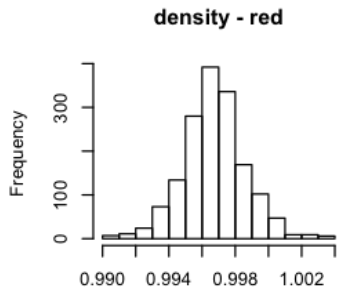

0
100
300

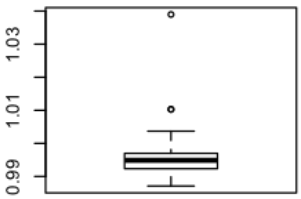

0
100
300

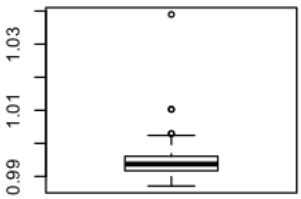

0
100
200
300

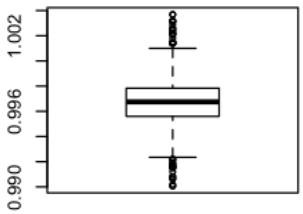

density
Frequency
0
1000
2000
3000
0.99
1.01
1.03

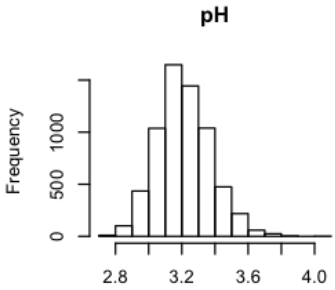

density - white
Frequency
0
1000
2000
0.99
1.01
1.03

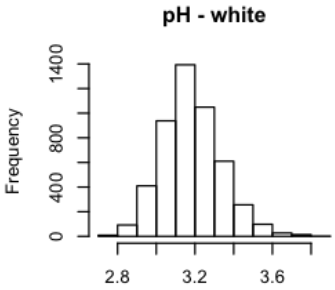

density - red
Frequency
0
100
300
0.990
0.994
0.998
1.002

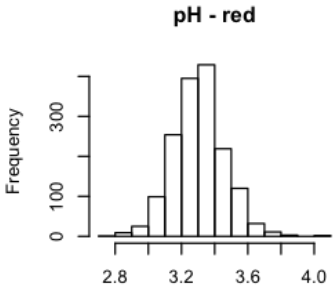

0.99
1.01
1.03

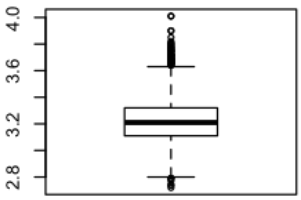

0.99
1.01
1.03

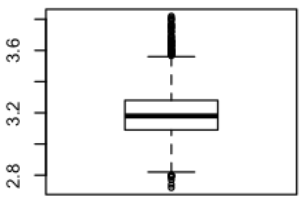

0.990
0.996
1.002

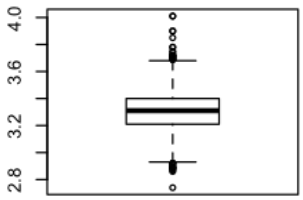

pH
Frequency
0
500
1000
2.8
3.2
3.6
4.0

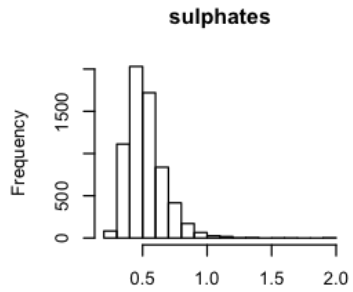

pH - white
Frequency
0
400
800
1400
2.8
3.2
3.6

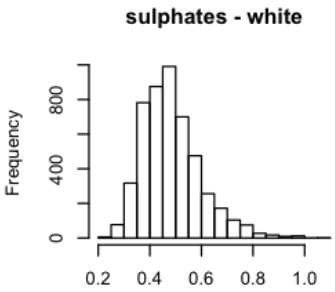

pH - red
Frequency
0
100
300
2.8
3.2
3.6
4.0

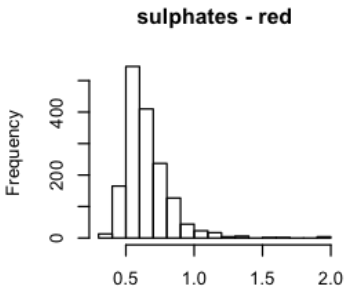

2.8
3.2
3.6
4.0

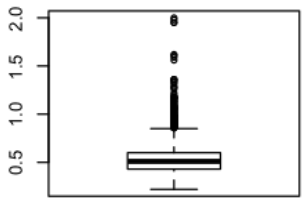

2.8
3.2
3.6

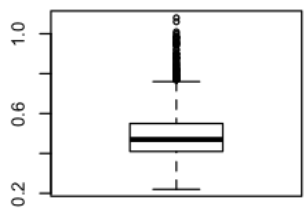

2.8
3.2
3.6
4.0

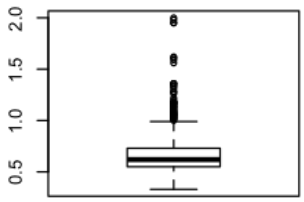

sulphates
Frequency
0
500
1500
0.5
1.0
1.5
2.0

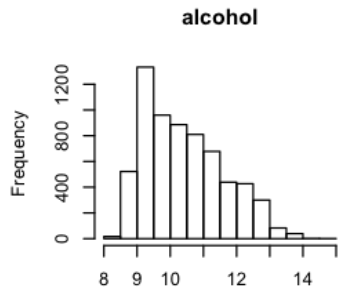

sulphates - white
Frequency
0
400
800
0.2
0.4
0.6
0.8
1.0

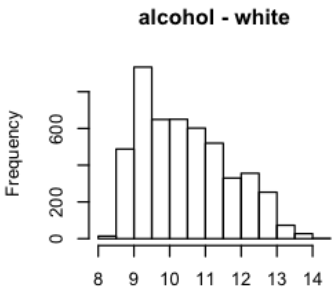

sulphates - red
Frequency
0
200
400
0.5
1.0
1.5
2.0

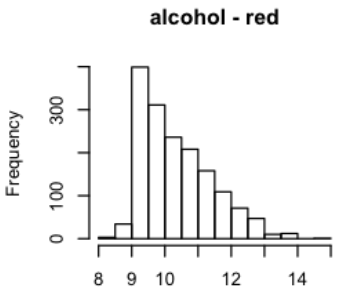

0.5
1.0
1.5
2.0

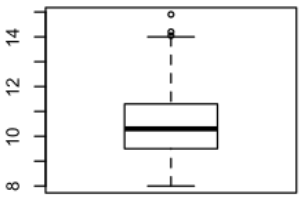

0.2
0.6
1.0

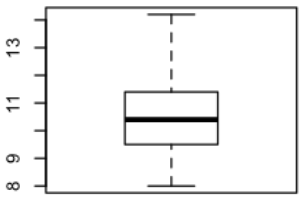

0.5
1.0
1.5
2.0

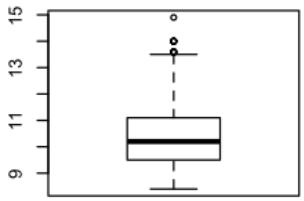

alcohol
Frequency
0
400
800
1200
8
9
10
12
14

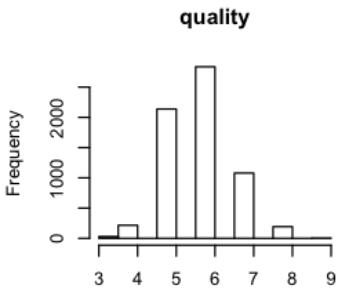

alcohol - white
Frequency
0
200
600
8
9
10
11
12
13
14

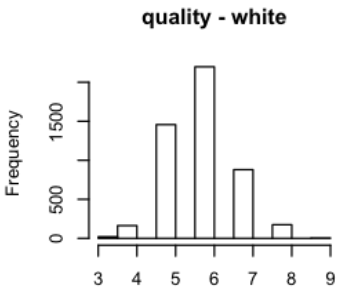

alcohol - red
Frequency
0
100
300
8
9
10
12
14

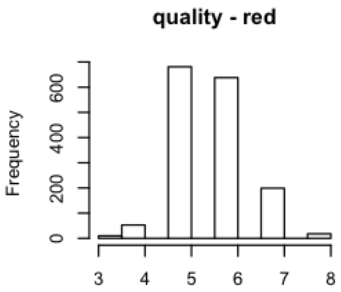

8
10
12
14

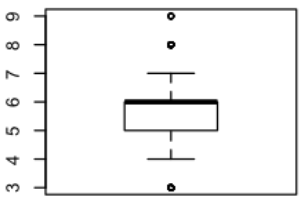

8
9
11
13

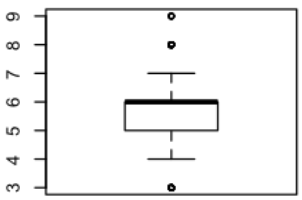

9
11
13
15

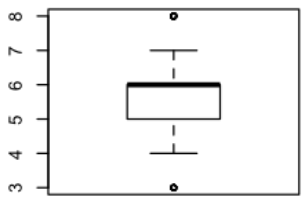

quality
Frequency
0
1000
2000
3
4
5
6
7
8
9


quality - white
Frequency
0
500
1500
3
4
5
6
7
8
9


quality - red
Frequency
0
200
400
600
3
4
5
6
7
8


3
4
5
6
7
8
9


3
4
5
6
7
8
9


3
4
5
6
7
8

In [7]:

```
cat ("Número de valores atípicos: \n")
for(i in seq(3, 13)) {
    cat("  Variable:",names(wines[i]), "\n")
    cat("     En total  : ", length(boxplot.stats(wines[,i])$out), "\n")
    cat("     En blancos: ", length(boxplot.stats(white_wines[,i])$out), "\n")
    cat("     En tintos : ", length(boxplot.stats(red_wines[,i])$out), "\n")
}
```

```
Número de valores atípicos:
  Variable: fixed.acidity
     En total  :  357
     En blancos:  119
     En tintos :  49
  Variable: volatile.acidity
     En total  :  377
     En blancos:  186
     En tintos :  19
  Variable: citric.acid
     En total  :  509
     En blancos:  270
     En tintos :  1
  Variable: residual.sugar
     En total  :  118
     En blancos:  7
     En tintos :  155
  Variable: chlorides
     En total  :  286
     En blancos:  208
     En tintos :  112
  Variable: free.sulfur.dioxide
     En total  :  62
     En blancos:  50
     En tintos :  30
  Variable: total.sulfur.dioxide
     En total  :  10
     En blancos:  19
     En tintos :  55
  Variable: density
     En total  :  3
     En blancos:  5
     En tintos :  45
  Variable: pH
     En total  :  73
     En blancos:  75
     En tintos :  35
  Variable: sulphates
     En total  :  191
     En blancos:  124
     En tintos :  59
  Variable: alcohol
     En total  :  3
     En blancos:  0
     En tintos :  13
```

En los gráficos se ve claramente que los valores de todas las variables son distintas para los vinos blancos y tintos (cambia la media, el rango de valores, etc.)

Excepto el pH (para todos los casos) y la densidad (de los vinos tintos), los histogramas no hacen pensar que las dsitribuciones de las varaibles sean normales.

Como se puede ver en el listado anterior, aparecen muchos valores atípicos.

Casi todos los valores atípicos son valores altos. Pero hay tantos que no hay un salto brusco entre los últimos valores reguales y los siguientes atípicos.

Hay alguna excepción, como el alcohol, donde hay 13 valores atípicos en los tintos y ninguno en los blancos.

En la tabla anterior se ve fácilmente que la suma de valores atípicos en blancos y tintos no coincide con el número de ellos en el *dataset* completo. Esto se debe a que los valores calculados en el conjunto completo no coinciden con los de los valores de cada subconjunto. Como decíamos antes, las características son distintas para cada tipo de vino.

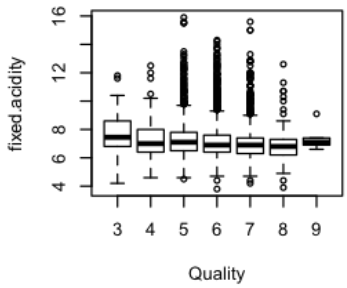
Los vinos blancos son los que suelen presentar valores atípicos más alejados del resto (por ejemplo en fixed.acidity, citric.acid, residual.sugar). Por ejemplo, en free.sulphur.dioxide solo hay un vino blanco que se aleja mucho del resto de valores.

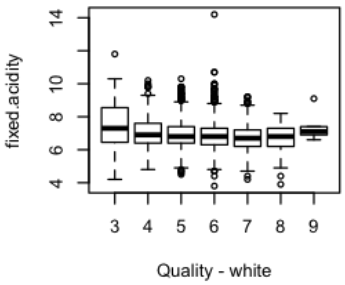
Por ejemplo, en total.sulfur.dioxide, hay un valor muy separado en los vinos tintos, pero con valor inferior a otros blancos.

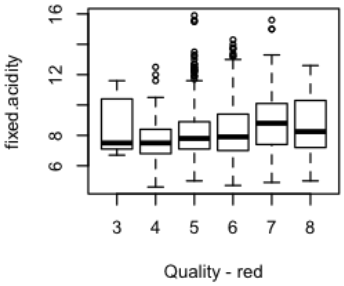
Vamos a ver si existe diferencias en los valores atípicos si separamos por calidades. Vamos a representar solo los *boxplot* para cada característica.

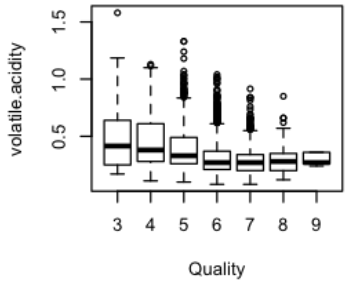
In [8]:

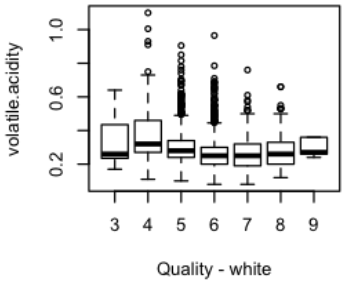
```
par(mfrow=c(3,3))
for (i in seq(3, 13)) {

    boxplot(wines[,i]~wines$quality, xlab = "Quality", ylab=names(wines[i]))
    boxplot(white_wines[,i]~white_wines$quality, xlab = "Quality - white", ylab=
names(wines[i]))
    boxplot(red_wines[,i]~red_wines$quality, xlab = "Quality - red", ylab=names(
wines[i]))

}
```

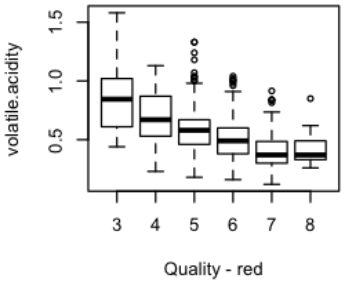

fixed.acidity
4 6 8 10 12 16
3 4 5 6 7 8 9
Quality

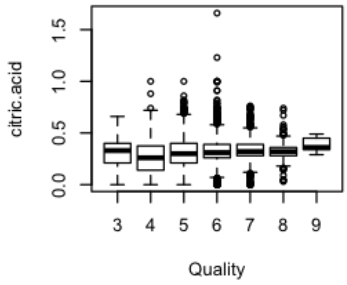

fixed.acidity
4 6 8 10 14
3 4 5 6 7 8 9
Quality - white

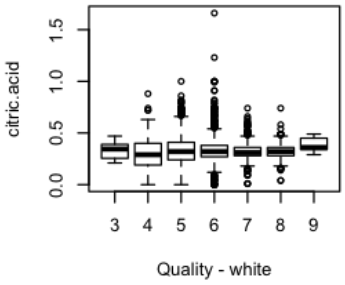

fixed.acidity
6 8 10 12 14 16
3 4 5 6 7 8
Quality - red

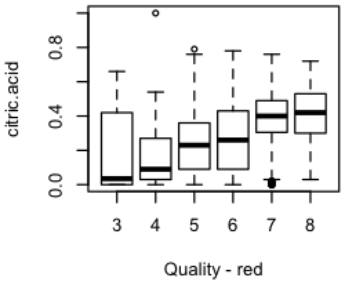

volatile.acidity
0.5 1.0 1.5
3 4 5 6 7 8 9
Quality

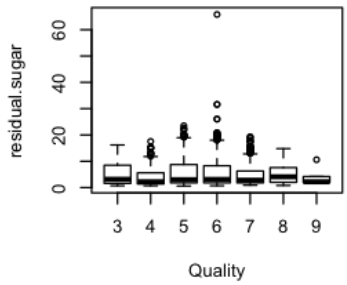

volatile.acidity
0.2 0.6 1.0
3 4 5 6 7 8 9
Quality - white

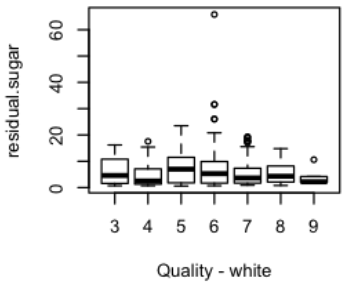

volatile.acidity
0.5 1.0 1.5
3 4 5 6 7 8
Quality - red

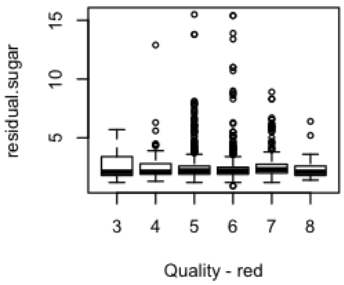

citric.acid
0.0 0.5 1.0 1.5
3 4 5 6 7 8 9
Quality

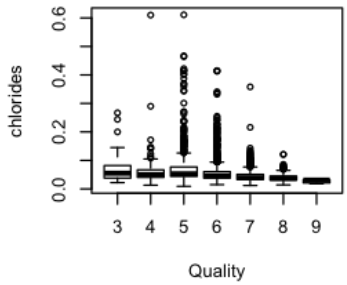

citric.acid
0.0 0.5 1.0 1.5
3 4 5 6 7 8 9
Quality - white

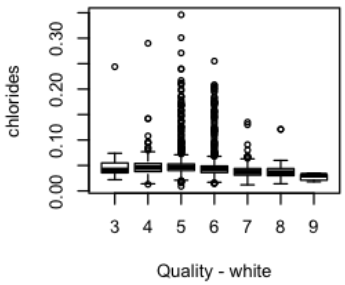

citric.acid
0.0 0.4 0.8
3 4 5 6 7 8
Quality - red

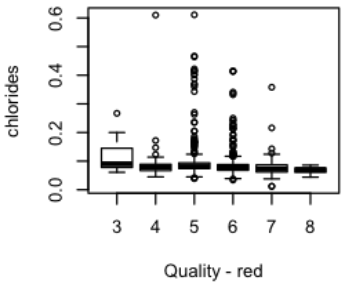

residual.sugar
0
20
40
60
3
4
5
6
7
8
9
Quality

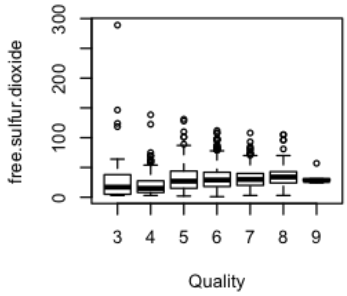

residual.sugar
0
20
40
60
3
4
5
6
7
8
9
Quality - white

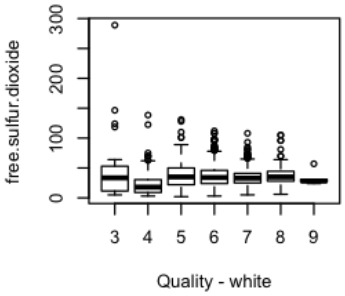

residual.sugar
5
10
15
3
4
5
6
7
8
Quality - red

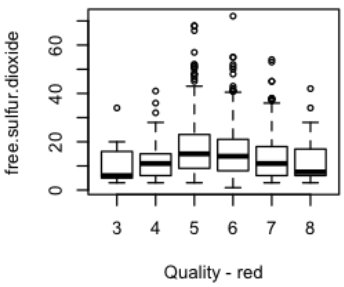

chlorides
0.0
0.2
0.4
0.6
3
4
5
6
7
8
9
Quality

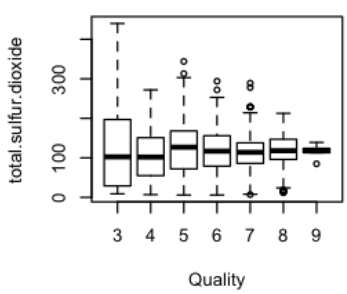

chlorides
0.00
0.10
0.20
0.30
3
4
5
6
7
8
9
Quality - white

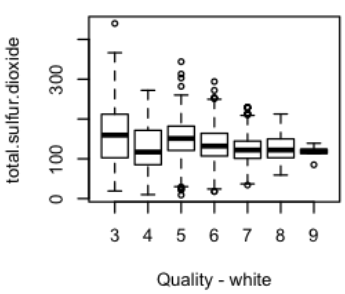

chlorides
0.0
0.2
0.4
0.6
3
4
5
6
7
8
Quality - red

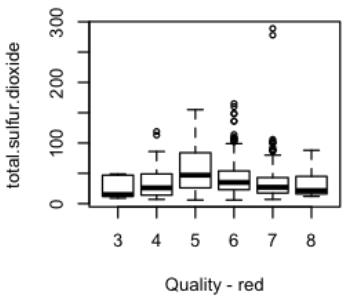

free.sulfur.dioxide
0
100
200
300
3
4
5
6
7
8
9
Quality

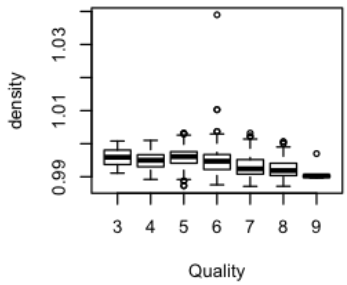

free.sulfur.dioxide
0
100
200
300
3
4
5
6
7
8
9
Quality - white

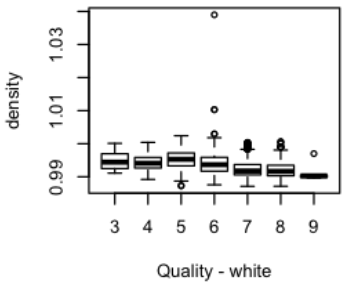

free.sulfur.dioxide
0
20
40
60
3
4
5
6
7
8
Quality - red

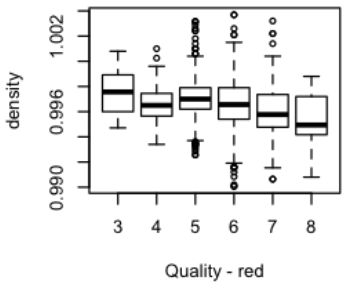

total.sulfur.dioxide
0
100
300
3
4
5
6
7
8
9
Quality

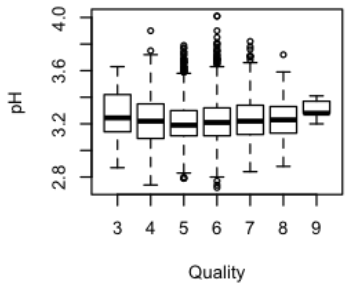

total.sulfur.dioxide
0
100
300
3
4
5
6
7
8
9
Quality - white

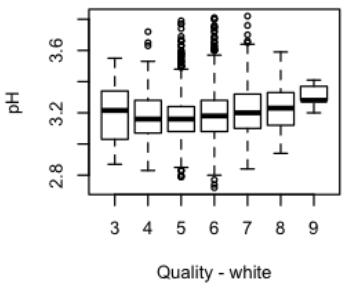

total.sulfur.dioxide
0
100
200
300
3
4
5
6
7
8
Quality - red

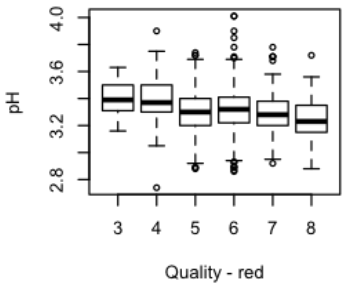

density
0.99
1.01
1.03
3
4
5
6
7
8
9
Quality

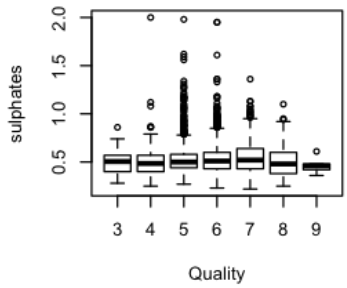

density
0.99
1.01
1.03
3
4
5
6
7
8
9
Quality - white

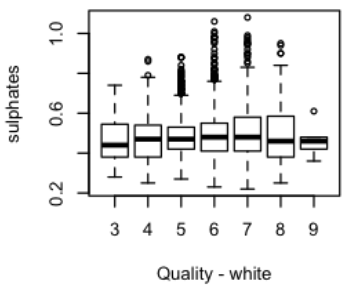

density
0.990
0.996
1.002
3
4
5
6
7
8
Quality - red

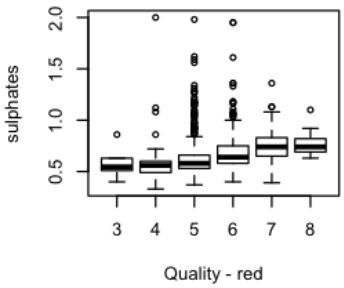

pH
2.8
3.2
3.6
4.0
3
4
5
6
7
8
9
Quality

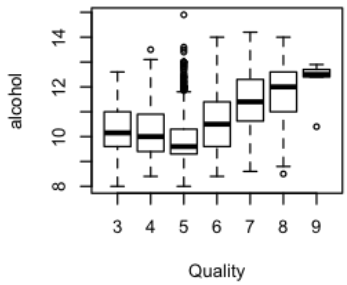

pH
2.8
3.2
3.6
3
4
5
6
7
8
9
Quality - white

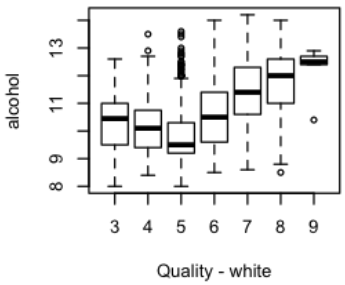

pH
2.8
3.2
3.6
4.0
3
4
5
6
7
8
Quality - red

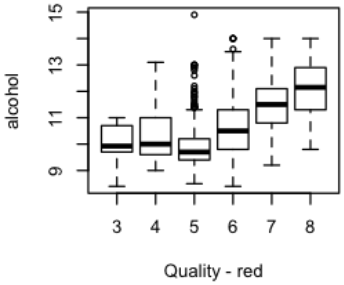

sulphates
0.5
1.0
1.5
2.0
3
4
5
6
7
8
9
Quality


sulphates
0.2
0.6
1.0
3
4
5
6
7
8
9
Quality - white


sulphates
0.5
1.0
1.5
2.0
3
4
5
6
7
8
Quality - red


alcohol
8
10
12
14
3
4
5
6
7
8
9
Quality


alcohol
8
9
11
13
3
4
5
6
7
8
9
Quality - white


alcohol
9
11
13
15
3
4
5
6
7
8
Quality - red

Se ven más valores extremos en las calidades intermedias, que sabemos que tienen más elementos. Esos outliers no están relacionados con calidades extremas.

No parece adecuado, por tanto, eliminar los vinos que presenten valores atípicos.

3.3. Exportación del conjunto de datos

Vamos a guardar en formato CSV el conjunto de datos completo en el fichero "winequality.csv".

Hay que tener en cuenta que estan ordenados por color primero, y luego según cómo estaban ordenados en los originales

En el fichero CSV generado, los valores están separados por comas, a diferencia de los puntos y comas que separaban los valores en los ficheros originales.

Para comprobar que se exportan bien los datos, se lee de nuevo el fichero y se muestra un resumen de los datos.

In [9]:

```
# Exportación del conjunto de datos a un .csv

write.table(wines, file = "winequality.csv",col.names=TRUE, row.names=FALSE, sep
=",")
vinos_tmp = read.csv("winequality.csv", header = TRUE)
cat("Número de registros, número de campos \n")
dim(vinos_tmp)
summary(vinos_tmp)
```

```
Número de registros, número de campos

    6497  14

       id           colour       fixed.acidity    volatile.acidity
 Min.   :   1   Min.   :0.0000   Min.   : 3.800   Min.   :0.0800
 1st Qu.:1625   1st Qu.:0.0000   1st Qu.: 6.400   1st Qu.:0.2300
 Median :3249   Median :0.0000   Median : 7.000   Median :0.2900
 Mean   :3249   Mean   :0.2461   Mean   : 7.215   Mean   :0.3397
 3rd Qu.:4873   3rd Qu.:0.0000   3rd Qu.: 7.700   3rd Qu.:0.4000
 Max.   :6497   Max.   :1.0000   Max.   :15.900   Max.   :1.5800
  citric.acid     residual.sugar     chlorides       free.sulfur.dio
xide
 Min.   :0.0000   Min.   : 0.600   Min.   :0.00900   Min.   :  1.00

 1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00

 Median :0.3100   Median : 3.000   Median :0.04700   Median : 29.00

 Mean   :0.3186   Mean   : 5.443   Mean   :0.05603   Mean   : 30.53

 3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00

 Max.   :1.6600   Max.   :65.800   Max.   :0.61100   Max.   :289.00

 total.sulfur.dioxide    density             pH          sulphates

 Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.220
0
 1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.430
0
 Median :118.0        Median :0.9949   Median :3.210   Median :0.510
0
 Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.531
3
 3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.600
0
 Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.000
0
    alcohol         quality
 Min.   : 8.00   Min.   :3.000
 1st Qu.: 9.50   1st Qu.:5.000
 Median :10.30   Median :6.000
 Mean   :10.49   Mean   :5.818
 3rd Qu.:11.30   3rd Qu.:6.000
 Max.   :14.90   Max.   :9.000
```

4. Análisis de los datos

Se va a empezar realizando un análisis descriptivo de los datos del *dataset*, que complete el análisis anterior.

Después, se estudiará la normalidad y homogeneidad de la varianza de las variables.

Posteriormente se analizará la correlación entre las variables y se realizarán algunos modelos de regresión lineal.

4.1. Análisis descriptivo de los datos

Para comprender mejor los datos, además de los resúmenes anteriores, vamos a ver, para cada variable:

- un analisis descriptivo (summary),
- la desviación típica, que no aparece en el resumen,
- un *barplot* por categoría

Este análisis lo aplicaremos al conjunto completo y a los *data.frames* de vinos blancos y tintos.

Empezaremos por un *bar plot* para hacernos una idea visual de la cantidad de vinos por categorías.

In [10]:

```
cat("Todos los vinos \n")
summary(wines[,3:14])
cat("Desviación típica:\n")
for (i in 3:14){
   cat(names(wines[i]),"  ", sd(wines[,i]), "\n")
}
cat("\n\n")
cat("Frecuencias absolutas por calidad:\n")
table(wines$quality)
bpt = barplot(table(wines$quality), main = "Distribución de vinos por calidad",
 xlab = "Quality")
text(bpt, table(wines$quality)+50, format(table(wines$quality)))
```

Todos los vinos

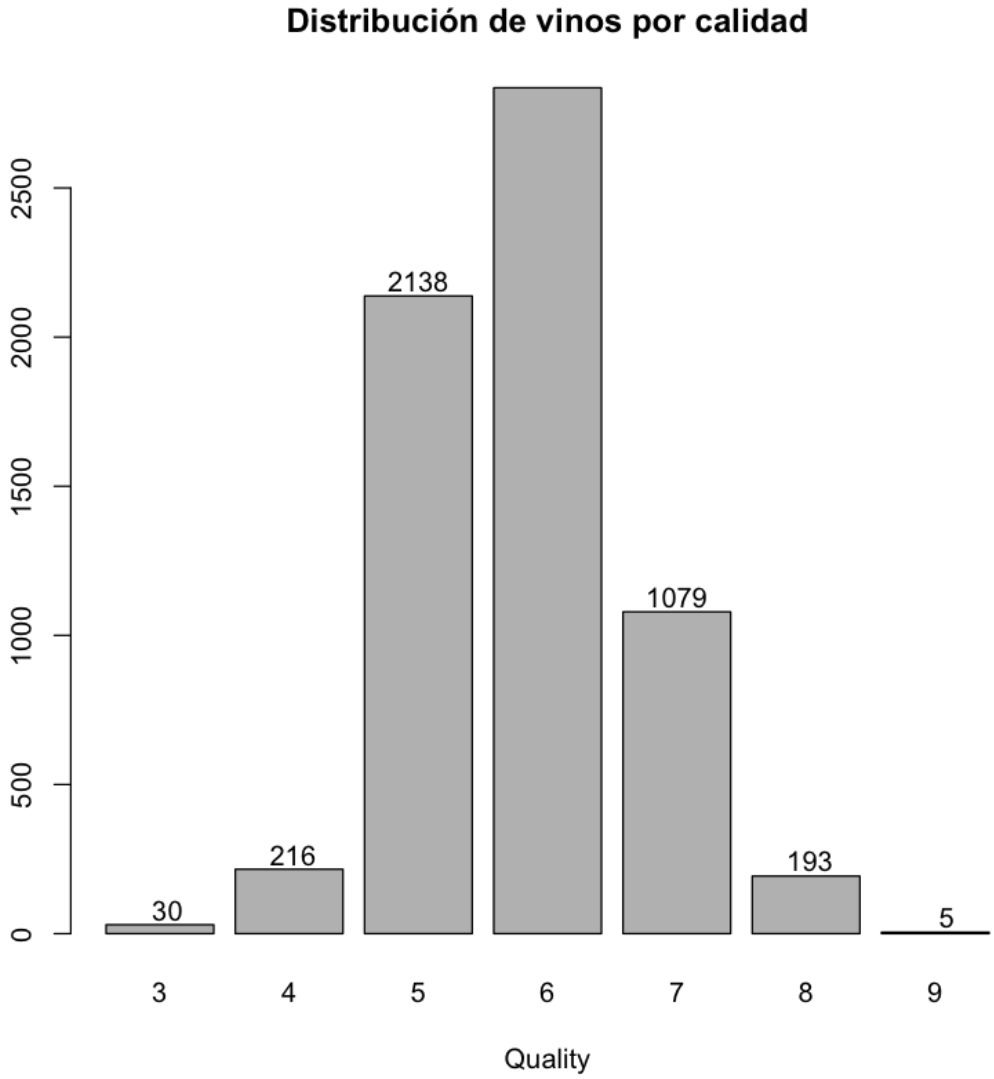
```
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
 1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
 Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
 Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
 3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
 Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    densi
ty
 Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :
0.9871
 1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:
0.9923
 Median :0.04700   Median : 29.00      Median :118.0        Median :
0.9949
 Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :
0.9947
 3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:
0.9970
 Max.   :0.61100   Max.   :289.00      Max.   :440.0        Max.   :
1.0390
       pH           sulphates         alcohol         quality
 Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
 1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000
 Median :3.210   Median :0.5100   Median :10.30   Median :6.000
 Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
 3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
 Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
```

Desviación típica:

```
fixed.acidity    1.296434
volatile.acidity    0.1646365
citric.acid    0.1453179
residual.sugar    4.757804
chlorides    0.0350336
free.sulfur.dioxide    17.7494
total.sulfur.dioxide    56.52185
density    0.002998673
pH    0.1607872
sulphates    0.1488059
alcohol    1.192712
quality    0.8732553
```

Frecuencias absolutas por calidad:

```
   3    4    5    6    7    8    9
  30  216 2138 2836 1079  193    5
```


Distribución de vinos por calidad
0
500
1000
1500
2000
2500
30
216
2138
1079
193
5
3
4
5
6
7
8
9
Quality

In [11]:

```
cat("Vinos blancos \n")
summary(white_wines[,3:14])
cat("Desviación típica:\n")
for (i in 3:14){
   cat(names(white_wines[i]),"  ", sd(white_wines[,i]), "\n")
}
cat("\n\n")
cat("Frecuencias absolutas por calidad:\n")
table(white_wines$quality)
bpt = barplot(table(white_wines$quality), col = "lightyellow",
              main = "Distribución de vinos blancos por calidad", xlab = "Qualit
y")
text(bpt, table(white_wines$quality)+50, format(table(white_wines$quality)))
```

Vinos blancos

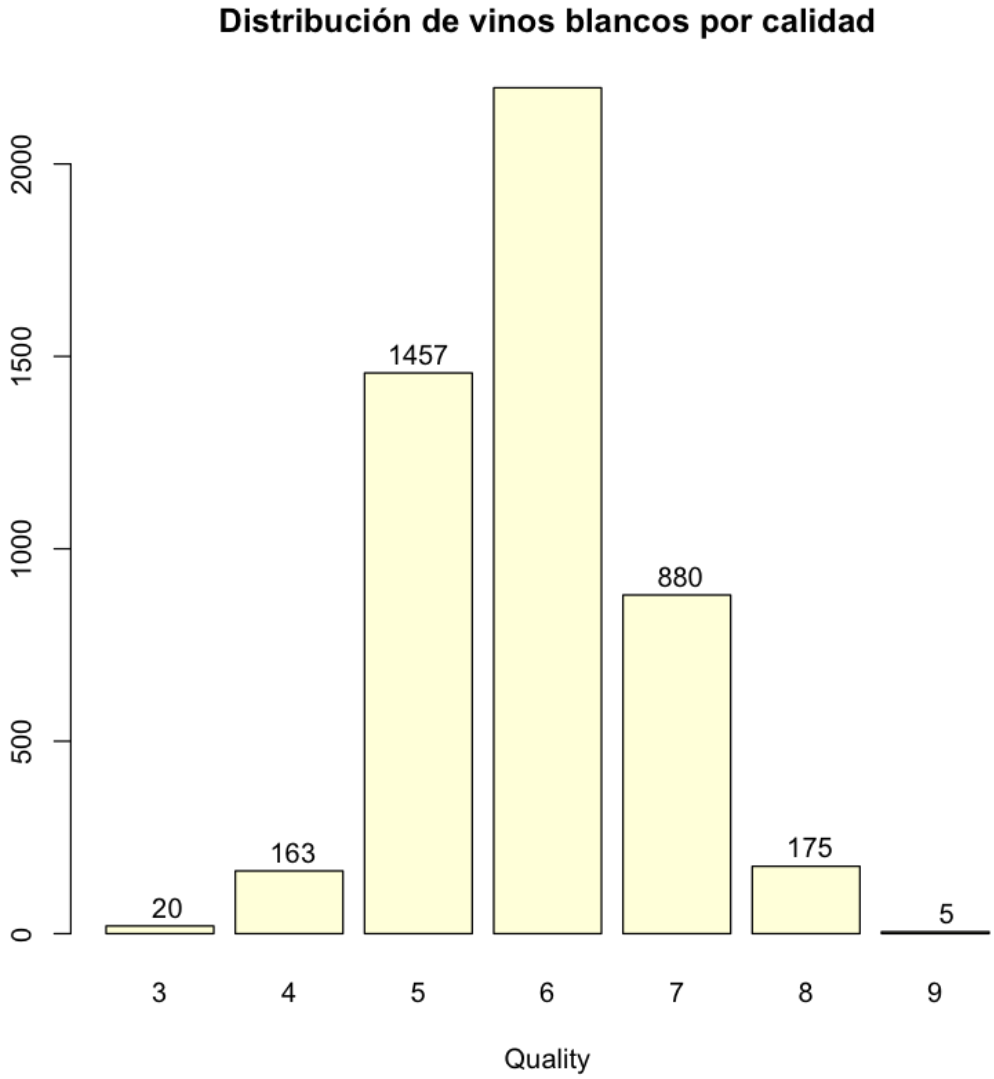
```
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
 1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
 Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
 Mean   : 6.855   Mean   :0.2782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

Desviación típica:

```
fixed.acidity    0.8438682
volatile.acidity    0.1007945
citric.acid    0.1210198
residual.sugar    5.072058
chlorides    0.02184797
free.sulfur.dioxide    17.00714
total.sulfur.dioxide    42.49806
density    0.002990907
pH    0.1510006
sulphates    0.1141258
alcohol    1.230621
quality    0.8856386
```

Frecuencias absolutas por calidad:

```
   3    4    5    6    7    8    9
  20  163 1457 2198  880  175    5
```


Distribución de vinos blancos por calidad
0
500
1000
1500
2000
20
163
1457
880
175
5
3
4
5
6
7
8
9
Quality

In [12]:

```
cat("Vinos tintos \n")
summary(red_wines[,3:14])
cat("Desviación típica:\n")
for (i in 3:14){
   cat(names(red_wines[i]),"  ", sd(red_wines[,i]), "\n")
}
cat("\n\n")
cat("Frecuencias absolutas por calidad:\n")
table(red_wines$quality)
bpt = barplot(table(red_wines$quality), col = "red4",
              main = "Distribución de vinos tintos por calidad", xlab = "Qualit
y")
text(bpt, table(red_wines$quality)+30, format(table(red_wines$quality)))
```

Vinos tintos

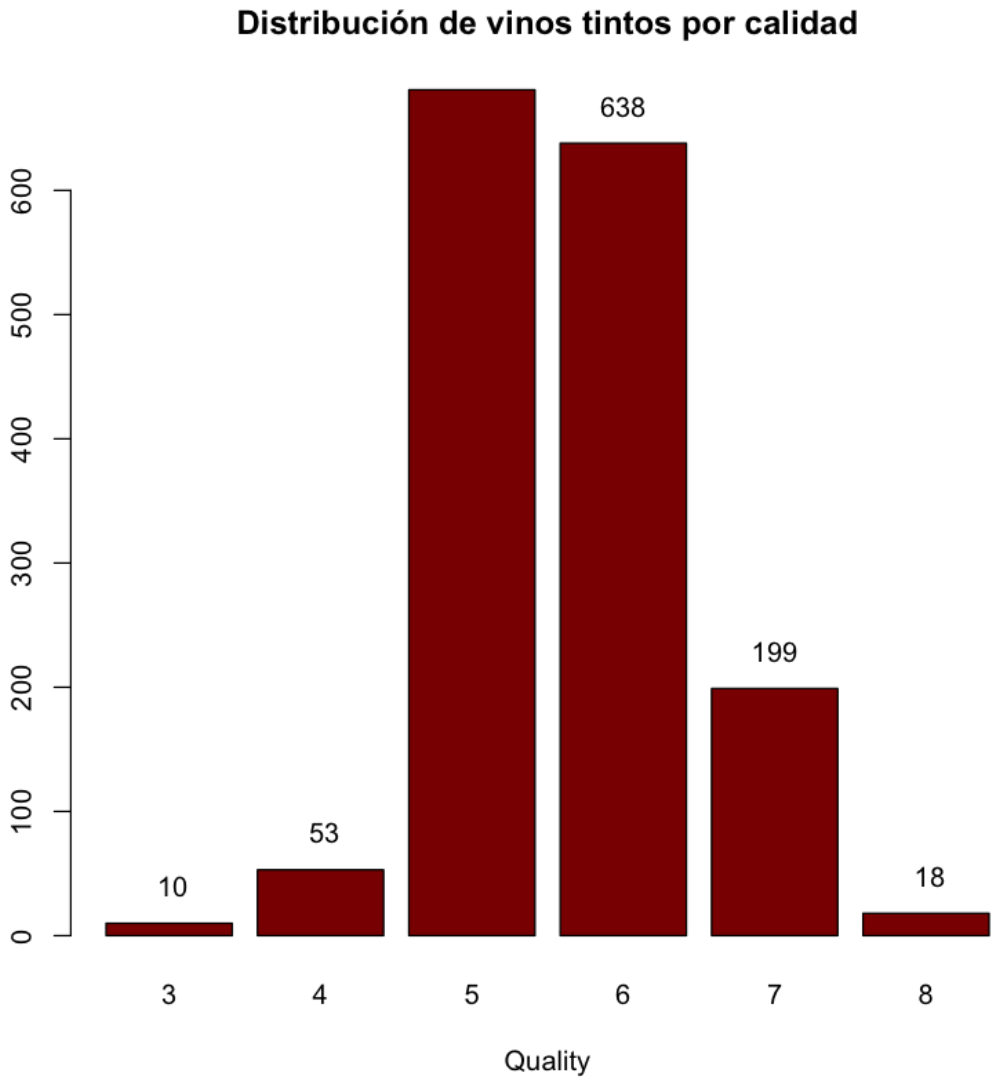
```
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
 Mean   : 8.32   Mean   :0.5278   Mean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

Desviación típica:

```
fixed.acidity    1.741096
volatile.acidity    0.1790597
citric.acid    0.1948011
residual.sugar    1.409928
chlorides    0.0470653
free.sulfur.dioxide    10.46016
total.sulfur.dioxide    32.89532
density    0.001887334
pH    0.1543865
sulphates    0.169507
alcohol    1.065668
quality    0.8075694
```

Frecuencias absolutas por calidad:

```
  3   4   5   6   7   8
 10  53 681 638 199  18
```

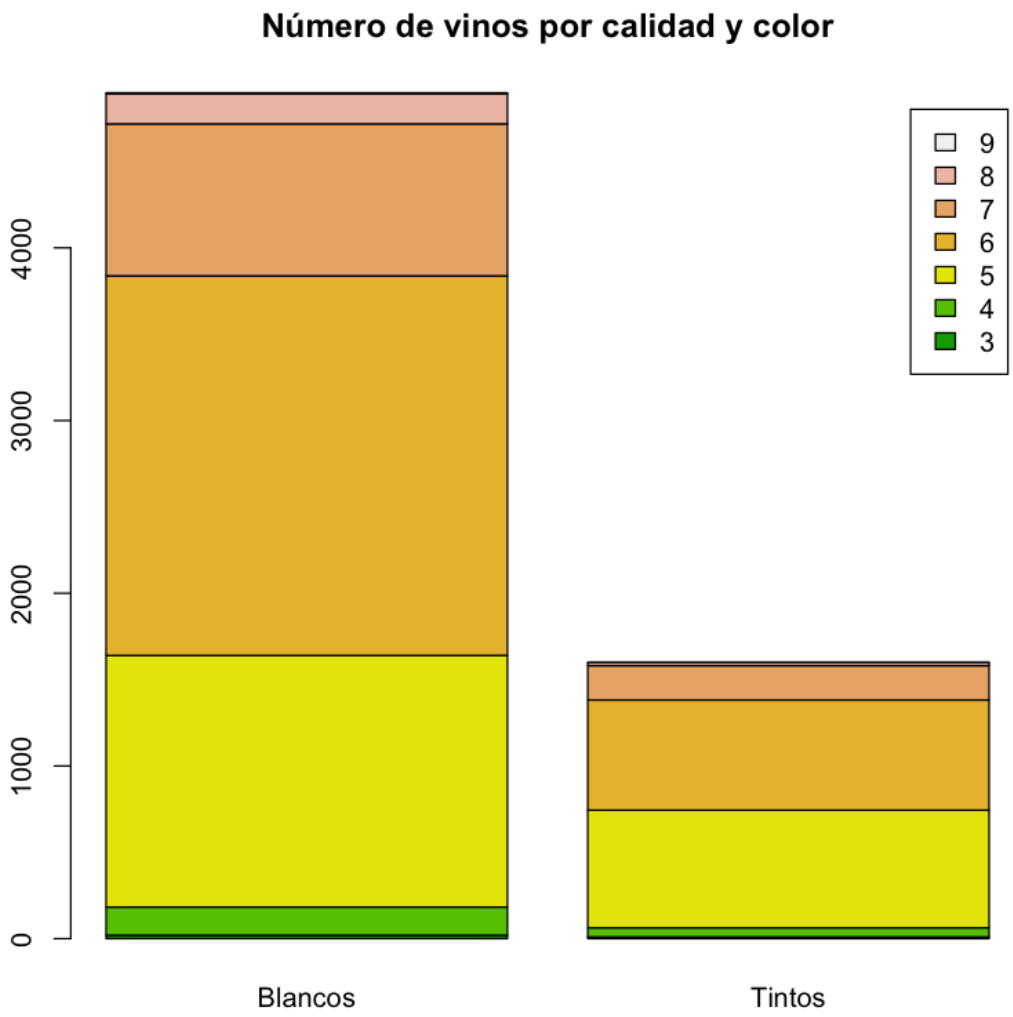

Distribución de vinos tintos por calidad
638
199
53
10
18
0
100
200
300
400
500
600
3
4
5
6
7
8
Quality

In [13]:

```
cat("Frecuencias absolutas por calidad y tipo de vino (0: blanco, 1: tinto):\n")
table(wines$quality, wines$colour)

barplot(table(wines$quality, wines$colour), col = terrain.colors(7), beside = FALSE, main = "Número de vinos por calidad y color", names.arg = c("Blancos", "Tintos"), legend = seq(3,9))
```

```
Frecuencias absolutas por calidad y tipo de vino (0: blanco, 1: tinto):
```

	0	1
3	20	10
4	163	53
5	1457	681
6	2198	638
7	880	199
8	175	18
9	5	0



4.2. Normalidad y homogeneidad de la varianza

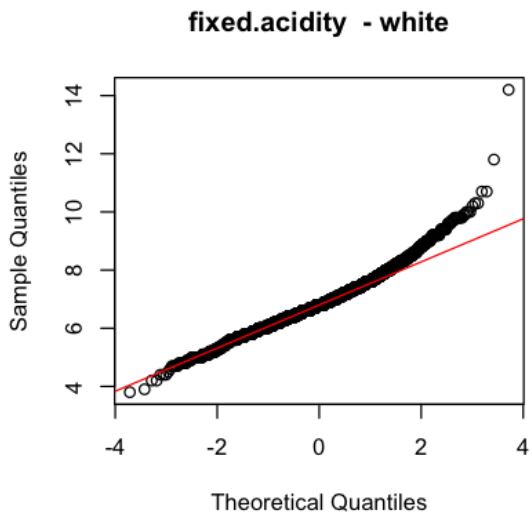
A la vista de los histogramas, no parece que las variables sigan una normal. De todas formas, vamos a realizar el siguiente análisis:

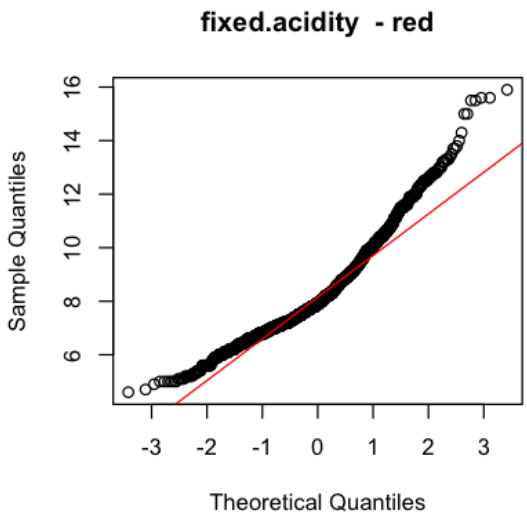
- Representar el gráfico **qqnorm** con cada variable para ver si se aproxima a una normal. Aquí se rerpesentan los cuantiles de la distribución observada con los cuantiles teóricos de una normal con la misma media y desviación típica. Los valores de la normal están representados con una línea roja.
- Aplicar el test de Shapiro-Wilk para determinar la normalidad o no de la distribución.
- Aplicar el test de Lillifors.
- Test de homocedasticidad.

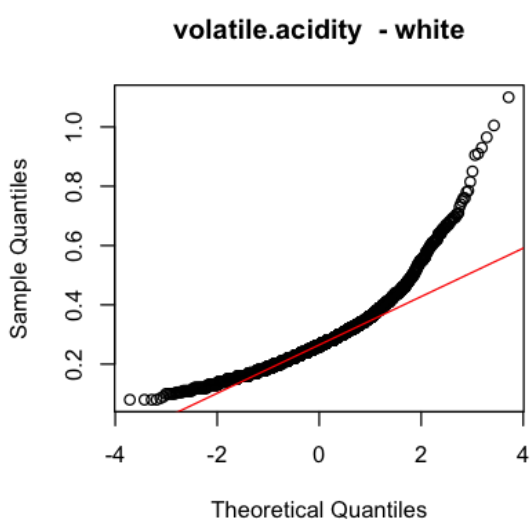
Vamos a hacer este análisis a los dos subconjuntos (vinos blancos y tintos) por separado.

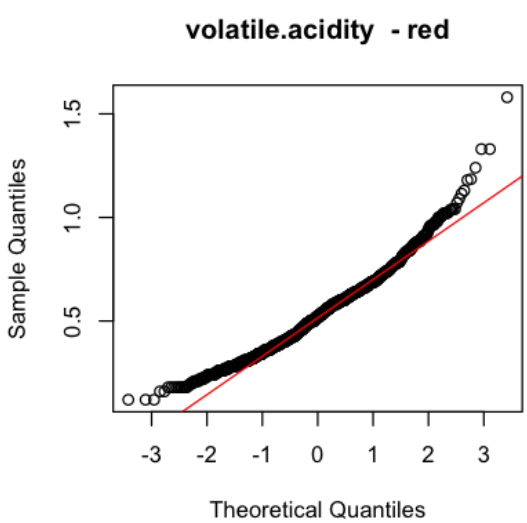
In [14]:

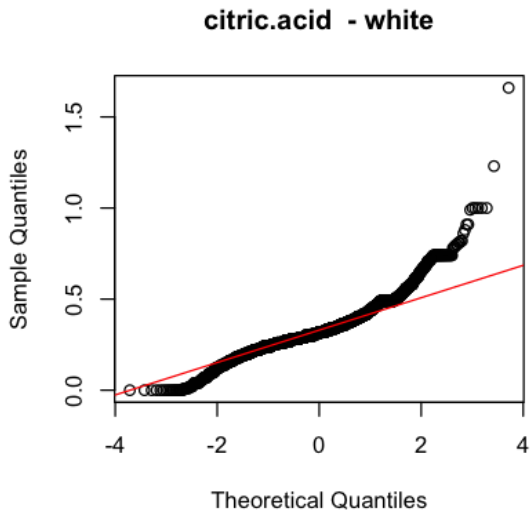
```
par(mfrow=c(2,2))
for (i in 3:13) {
    qqnorm(white_wines[,i],main = paste(colnames(white_wines)[i], " - white"))
    qqline(white_wines[,i],col="red")
    qqnorm(red_wines[,i],main = paste(colnames(red_wines)[i], " - red"))
    qqline(red_wines[,i],col="red")
}
```

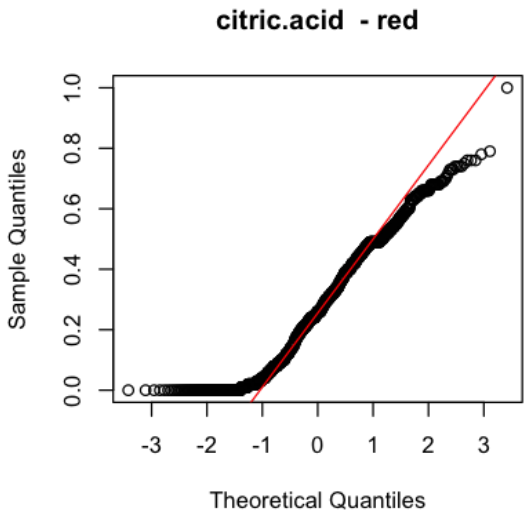

fixed.acidity - white
Sample Quantiles
Theoretical Quantiles
4
6
8
10
12
14
-4
-2
0
2
4

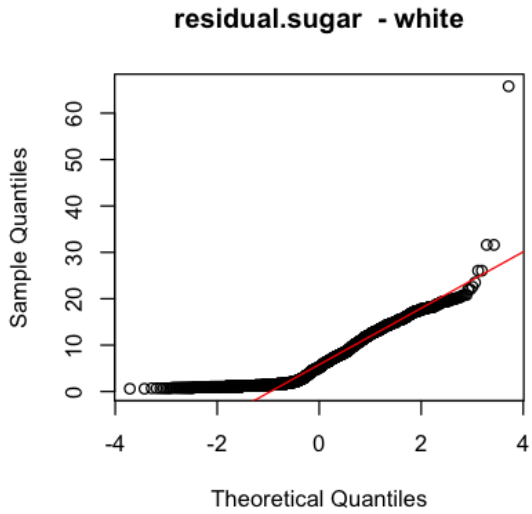

fixed.acidity - red
Sample Quantiles
Theoretical Quantiles
6
8
10
12
14
16
-3
-2
-1
0
1
2
3

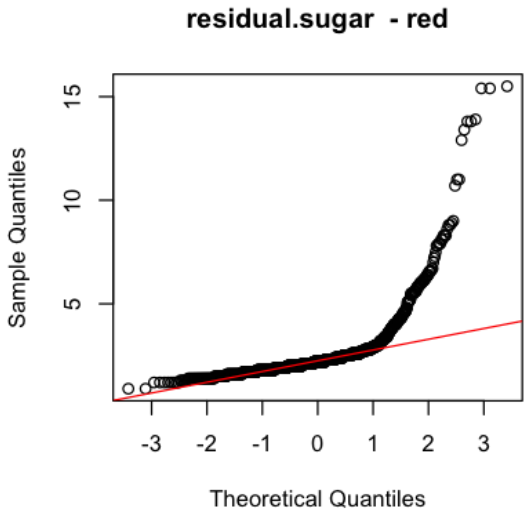

volatile.acidity - white
Sample Quantiles
Theoretical Quantiles
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4

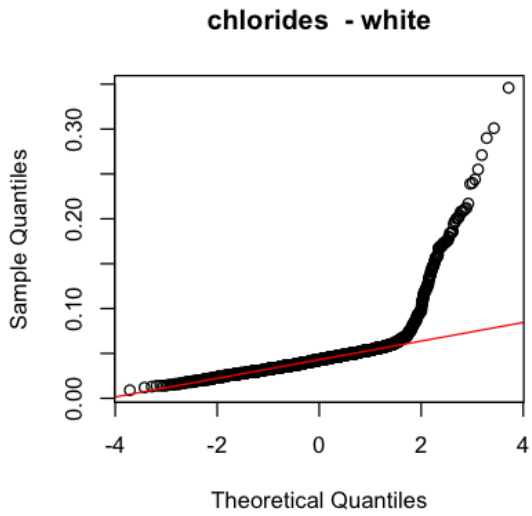

volatile.acidity - red
Sample Quantiles
Theoretical Quantiles
0.5
1.0
1.5
-3
-2
-1
0
1
2
3

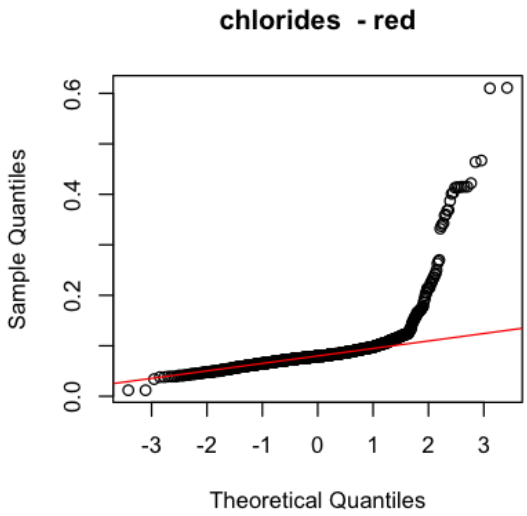

citric.acid - white
Sample Quantiles
Theoretical Quantiles
0.0
0.5
1.0
1.5
-4
-2
0
2
4

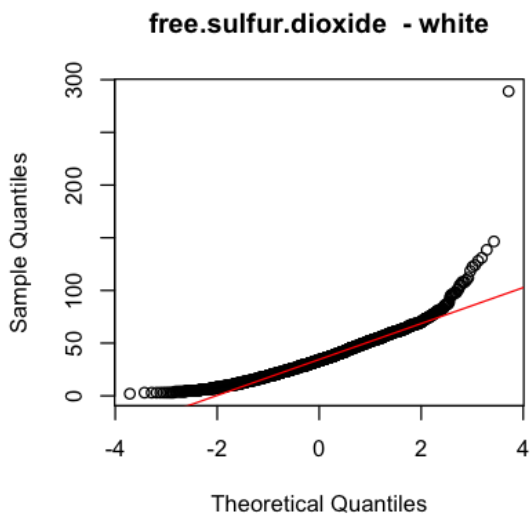

citric.acid - red
Sample Quantiles
Theoretical Quantiles
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

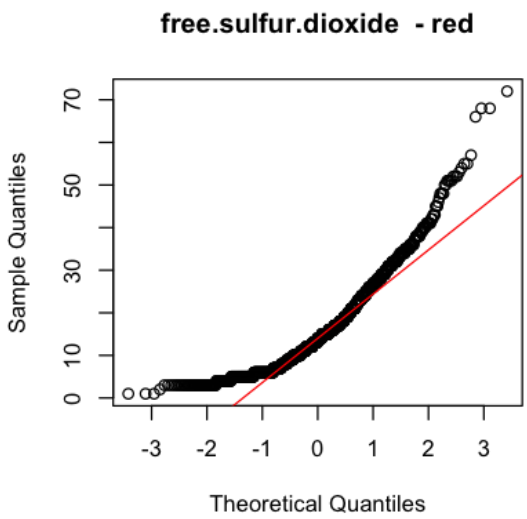

residual.sugar - white
Sample Quantiles
Theoretical Quantiles
0
10
20
30
40
50
60
-4
-2
0
2
4

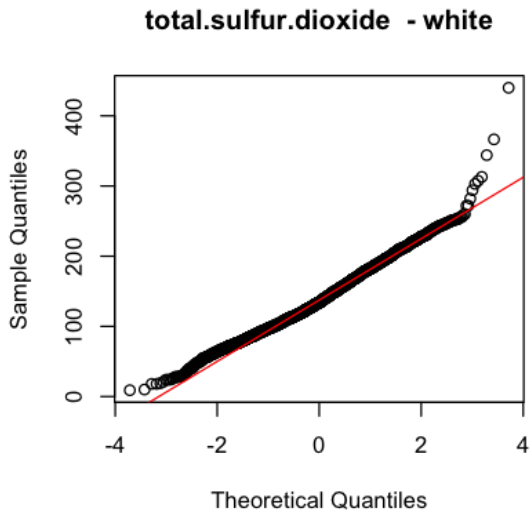

residual.sugar - red
Sample Quantiles
Theoretical Quantiles
5
10
15
-3
-2
-1
0
1
2
3

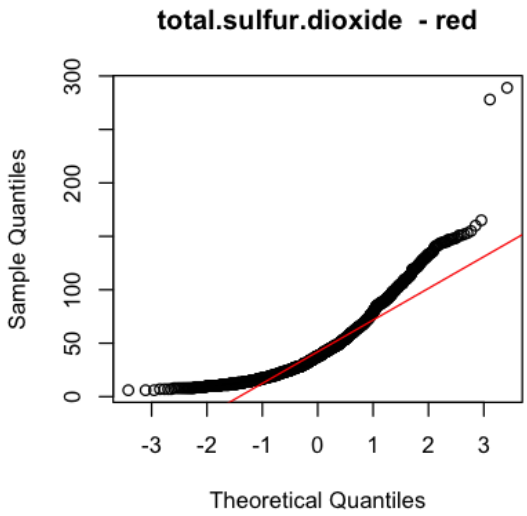

chlorides - white
Sample Quantiles
0.00
0.10
0.20
0.30
-4
-2
0
2
4
Theoretical Quantiles

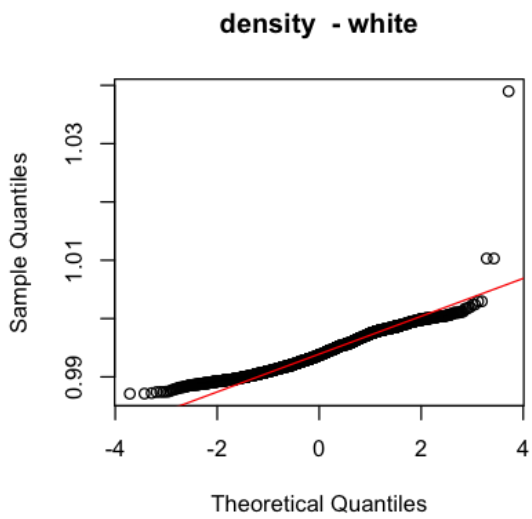

chlorides - red
Sample Quantiles
0.0
0.2
0.4
0.6
-3
-2
-1
0
1
2
3
Theoretical Quantiles

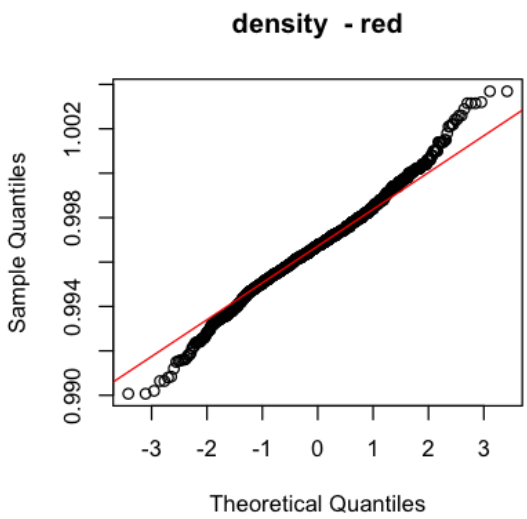

free.sulfur.dioxide - white
Sample Quantiles
0
50
100
200
300
-4
-2
0
2
4
Theoretical Quantiles

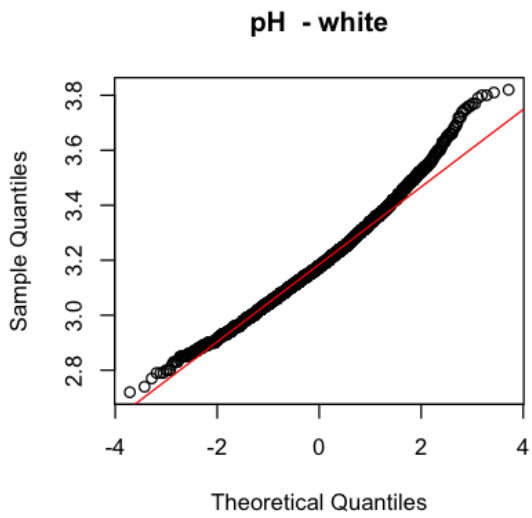

free.sulfur.dioxide - red
Sample Quantiles
0
10
30
50
70
-3
-2
-1
0
1
2
3
Theoretical Quantiles

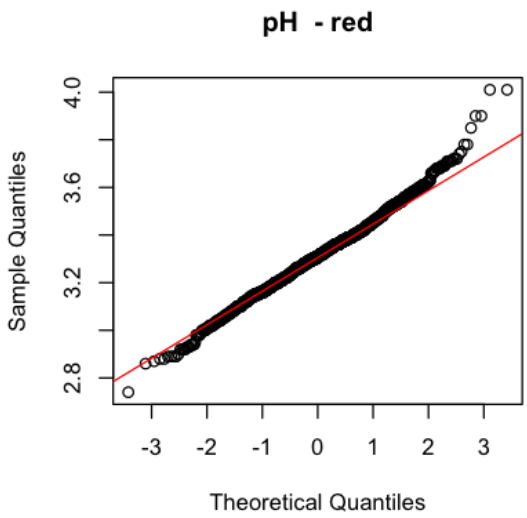

total.sulfur.dioxide - white
Sample Quantiles
0
100
200
300
400
-4
-2
0
2
4
Theoretical Quantiles

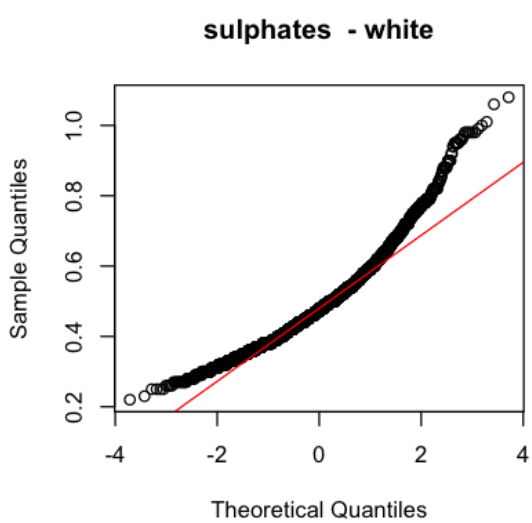

total.sulfur.dioxide - red
Sample Quantiles
0
50
100
200
300
-3
-2
-1
0
1
2
3
Theoretical Quantiles

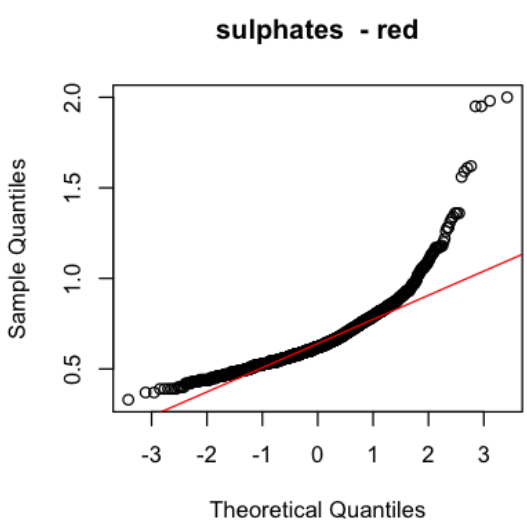

density - white
Sample Quantiles
0.99
1.01
1.03
-4
-2
0
2
4
Theoretical Quantiles

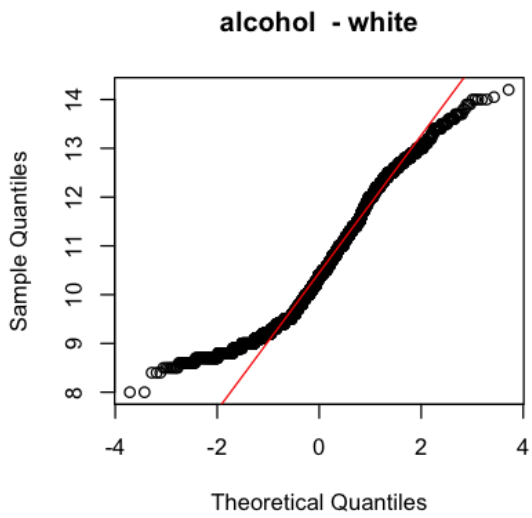

density - red
Sample Quantiles
0.990
0.994
0.998
1.002
-3
-2
-1
0
1
2
3
Theoretical Quantiles

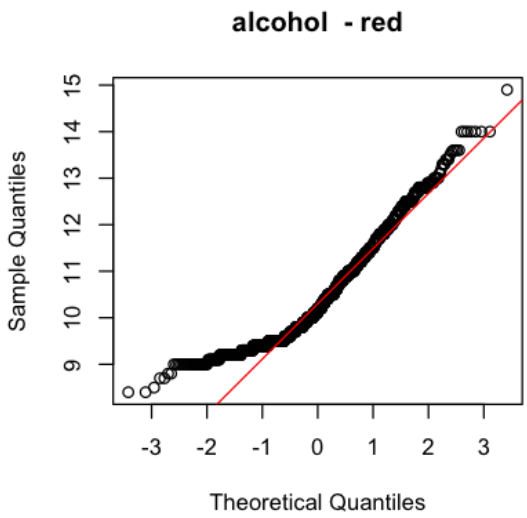

pH - white
Sample Quantiles
Theoretical Quantiles
2.8
3.0
3.2
3.4
3.6
3.8
-4
-2
0
2
4


pH - red
Sample Quantiles
Theoretical Quantiles
2.8
3.2
3.6
4.0
-3
-2
-1
0
1
2
3


sulphates - white
Sample Quantiles
Theoretical Quantiles
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4


sulphates - red
Sample Quantiles
Theoretical Quantiles
0.5
1.0
1.5
2.0
-3
-2
-1
0
1
2
3


alcohol - white
Sample Quantiles
Theoretical Quantiles
8
9
10
11
12
13
14
-4
-2
0
2
4


alcohol - red
Sample Quantiles
Theoretical Quantiles
9
10
11
12
13
14
15
-3
-2
-1
0
1
2
3

En los gráficos Q-Q vemos que hay muchos valores intermedios que se ajustan bien a la normal, pero los extremos se separan bastante.

Como los conjuntos de datos son grandes, se pueden considerar que se aproximan a distribuciones normales, aplicando el teorema central del límite.

Aplicamos a continuación el test de Shapiro-Wilkins a las variables para estudiar su normalidad.

Este test está pensado sobre todo para muestras pequeñas. La hipótesis nula es que los valores estén distribuidos siguiente una normal. Si el p-valor es menor que el nivel de significación **alfa**, se rechaza la hipótesis nula y se se concluye que los datos no vienen de una distribución normal).

Vamos a usar un valor de alfa de 0.05 (95%).

In [15]:

```
alfa = 0.05

for (i in 3:14) {
    cat("Variable: ", colnames(wines)[i],"\n")
    test = shapiro.test(white_wines[,i])
    cat("   Vinos blancos: \t")
    cat("   W: ", test[["statistic"]], "\t")
    cat("   p-value: ", test[["p.value"]], "\t")
    if (test[["p.value"]] < alfa)
        cat("      Los datos no siguen una normal \n")
    else
        cat("\n")
    test = shapiro.test(red_wines[,i])
    cat("   Vinos tintos: \t")
    cat("   W: ", test[["statistic"]], "\t")
    cat("   p-value: ", test[["p.value"]], "\t")
    if (test[["p.value"]] < alfa)
        cat("      Los datos no siguen una normal \n")
    else
        cat("*** \n")

}
```

```
Variable:  fixed.acidity
   Vinos blancos:          W:  0.9765615          p-value:  1.15015
1e-27         Los datos no siguen una normal
   Vinos tintos:           W:  0.9420298          p-value:  1.52501
2e-24         Los datos no siguen una normal
Variable:  volatile.acidity
   Vinos blancos:          W:  0.9045497          p-value:  4.58679
7e-48         Los datos no siguen una normal
   Vinos tintos:           W:  0.9743369          p-value:  2.69293
5e-16         Los datos no siguen una normal
Variable:  citric.acid
   Vinos blancos:          W:  0.9222473          p-value:  1.01317
9e-44         Los datos no siguen una normal
   Vinos tintos:           W:  0.955292           p-value:  1.02193
2e-21         Los datos no siguen una normal
Variable:  residual.sugar
   Vinos blancos:          W:  0.8845686          p-value:  2.82071
e-51          Los datos no siguen una normal
   Vinos tintos:           W:  0.5660771          p-value:  1.02016
2e-52         Los datos no siguen una normal
Variable:  chlorides
   Vinos blancos:          W:  0.5908084          p-value:  2.14058
4e-75         Los datos no siguen una normal
   Vinos tintos:           W:  0.4842466          p-value:  1.17905
6e-55         Los datos no siguen una normal
Variable:  free.sulfur.dioxide
   Vinos blancos:          W:  0.9420691          p-value:  3.85784
5e-40         Los datos no siguen una normal
   Vinos tintos:           W:  0.9018395          p-value:  7.69459
7e-31         Los datos no siguen una normal
Variable:  total.sulfur.dioxide
   Vinos blancos:          W:  0.9890146          p-value:  4.38345
3e-19         Los datos no siguen una normal
   Vinos tintos:           W:  0.8732246          p-value:  3.57345
1e-34         Los datos no siguen una normal
Variable:  density
   Vinos blancos:          W:  0.9548048          p-value:  1.78089
5e-36         Los datos no siguen una normal
   Vinos tintos:           W:  0.9908655          p-value:  1.93605
3e-08         Los datos no siguen una normal
Variable:  pH
   Vinos blancos:          W:  0.9880965          p-value:  6.50552
1e-20         Los datos no siguen una normal
   Vinos tintos:           W:  0.9934863          p-value:  1.71223
7e-06         Los datos no siguen una normal
Variable:  sulphates
   Vinos blancos:          W:  0.9516094          p-value:  1.82197
9e-37         Los datos no siguen una normal
   Vinos tintos:           W:  0.8330438          p-value:  5.82314
e-38          Los datos no siguen una normal
Variable:  alcohol
   Vinos blancos:          W:  0.9553024          p-value:  2.56901
4e-36         Los datos no siguen una normal
   Vinos tintos:           W:  0.9288391          p-value:  6.64405
7e-27         Los datos no siguen una normal
Variable:  quality
   Vinos blancos:          W:  0.8890432          p-value:  1.34011
1e-50         Los datos no siguen una normal
   Vinos tintos:           W:  0.8575895          p-value:  9.51508
5e-36         Los datos no siguen una normal
```

Como vemos, según este test, ninguna de las variables en cada uno de los *datasets* sigue una normal.

Para conjuntos de valores grandes se usa el test de Lillifors, que es una modificación del test de Kolmogorov-Smirnov para contrastar la normalidad cuando no se conoce la media ni la varianza.

Para poder calcular este test debe estár instalada la librería **nortest**.

In [16]:

```
library("nortest")

for (i in 3:14) {
    cat("Variable: ", colnames(wines)[i],"\n")
    test = lillie.test(white_wines[,i])
    cat("   Vinos blancos: \t")
    cat("   D: ", test[["statistic"]], "\t")
    cat("   p-value: ", test[["p.value"]], "\t")
    if (test[["p.value"]] < alfa)
        cat("      Los datos no siguen una normal \n")
    else
        cat("*** \n")
    test = lillie.test(red_wines[,i])
    cat("   Vinos tintos: \t")
    cat("   D: ", test[["statistic"]], "\t")
    cat("   p-value: ", test[["p.value"]], "\t")
    if (test[["p.value"]] < alfa)
        cat("      Los datos no siguen una normal \n")
    else
        cat("*** \n")

}
```

```
Variable:  fixed.acidity
   Vinos blancos:          D:  0.06623227          p-value:  4.54710
8e-57         Los datos no siguen una normal
   Vinos tintos:           D:  0.1105032           p-value:  6.98245
6e-53         Los datos no siguen una normal
Variable:  volatile.acidity
   Vinos blancos:          D:  0.1045128           p-value:  2.23447
5e-146        Los datos no siguen una normal
   Vinos tintos:           D:  0.05466244          p-value:  4.48908
4e-12         Los datos no siguen una normal
Variable:  citric.acid
   Vinos blancos:          D:  0.1127502           p-value:  4.97030
6e-171        Los datos no siguen una normal
   Vinos tintos:           D:  0.08386605          p-value:  9.85942
9e-30         Los datos no siguen una normal
Variable:  residual.sugar
   Vinos blancos:          D:  0.1366236           p-value:  2.37158
6e-253        Los datos no siguen una normal
   Vinos tintos:           D:  0.2606766           p-value:  3.98171
2e-309        Los datos no siguen una normal
Variable:  chlorides
   Vinos blancos:          D:  0.2072626           p-value:  0
  Los datos no siguen una normal
   Vinos tintos:           D:  0.2596402           p-value:  1.26010
7e-306        Los datos no siguen una normal
Variable:  free.sulfur.dioxide
   Vinos blancos:          D:  0.0576817           p-value:  8.4452e
-43           Los datos no siguen una normal
   Vinos tintos:           D:  0.1112397           p-value:  1.28359
9e-53         Los datos no siguen una normal
Variable:  total.sulfur.dioxide
   Vinos blancos:          D:  0.04465003          p-value:  4.83286
8e-25         Los datos no siguen una normal
   Vinos tintos:           D:  0.1209779           p-value:  7.94099
6e-64         Los datos no siguen una normal
Variable:  density
   Vinos blancos:          D:  0.05221027          p-value:  9.31027
7e-35         Los datos no siguen una normal
   Vinos tintos:           D:  0.04478707          p-value:  6.25170
7e-08         Los datos no siguen una normal
Variable:  pH
   Vinos blancos:          D:  0.04926856          p-value:  8.74764
4e-31         Los datos no siguen una normal
   Vinos tintos:           D:  0.04036845          p-value:  2.24404
8e-06         Los datos no siguen una normal
Variable:  sulphates
   Vinos blancos:          D:  0.08684996          p-value:  5.27074
2e-100        Los datos no siguen una normal
   Vinos tintos:           D:  0.1247865           p-value:  4.60248
8e-68         Los datos no siguen una normal
Variable:  alcohol
   Vinos blancos:          D:  0.09157332          p-value:  1.54124
8e-111        Los datos no siguen una normal
   Vinos tintos:           D:  0.1214532           p-value:  2.39150
1e-64         Los datos no siguen una normal
Variable:  quality
   Vinos blancos:          D:  0.2287622           p-value:  0
  Los datos no siguen una normal
   Vinos tintos:           D:  0.2498185           p-value:  1.95145
5e-283        Los datos no siguen una normal
```

Con este test obtenemos el mismo resultado: no siguen una normal ninguna de las variables de los *datasets*.

Vamos a aplicar ahora un test para evaluar la homocedasticidad (la homogeneidad de las varianzas) de todas las variables contraponiendo los grupos de vinos blancos y vinos tintos.

Usaremos el test de Fligner-Killen. Es un test no paramétrico que compara las varianzas basándose en la mediana y se usa cuando las muestras no cumplen la condición de normalidad.

In [17]:

```
for (i in 3:14) {
    cat("Variable: ", colnames(wines)[i],"\n")
    test = fligner.test(list(white_wines[,i], red_wines[,i]))
    cat("   med chi-2: ", test[["statistic"]], "\t")
    cat("   p-value: ", test[["p.value"]], "\t")
    if (test[["p.value"]] < alfa)
        cat("      No homogeneidad de las varianzas \n")
    else
        cat("*** \n")
}
```

```
Variable:  fixed.acidity
   med chi-2:  751.5082            p-value:  1.885649e-165
  No homogeneidad de las varianzas
Variable:  volatile.acidity
   med chi-2:  852.7903            p-value:  1.798635e-187
  No homogeneidad de las varianzas
Variable:  citric.acid
   med chi-2:  839.1307            p-value:  1.677256e-184
  No homogeneidad de las varianzas
Variable:  residual.sugar
   med chi-2:  1768.915            p-value:  0        No homogeneida
d de las varianzas
Variable:  chlorides
   med chi-2:  200.5291            p-value:  1.600938e-45
  No homogeneidad de las varianzas
Variable:  free.sulfur.dioxide
   med chi-2:  348.2164            p-value:  1.036427e-77
  No homogeneidad de las varianzas
Variable:  total.sulfur.dioxide
   med chi-2:  198.2962            p-value:  4.916348e-45
  No homogeneidad de las varianzas
Variable:  density
   med chi-2:  448.9288            p-value:  1.23377e-99
  No homogeneidad de las varianzas
Variable:  pH
   med chi-2:  0.4121955           p-value:  0.5208573  ***
Variable:  sulphates
   med chi-2:  83.07863            p-value:  7.885344e-20
  No homogeneidad de las varianzas
Variable:  alcohol
   med chi-2:  78.15042            p-value:  9.548546e-19
  No homogeneidad de las varianzas
Variable:  quality
   med chi-2:  0.617752            p-value:  0.4318839  ***
```

En este caso, pasan el test e homogeneidad de la varianza el pH y la calidad de los vinos.

(Se han aplicado a calidad también aunque es la variable dependiente.)

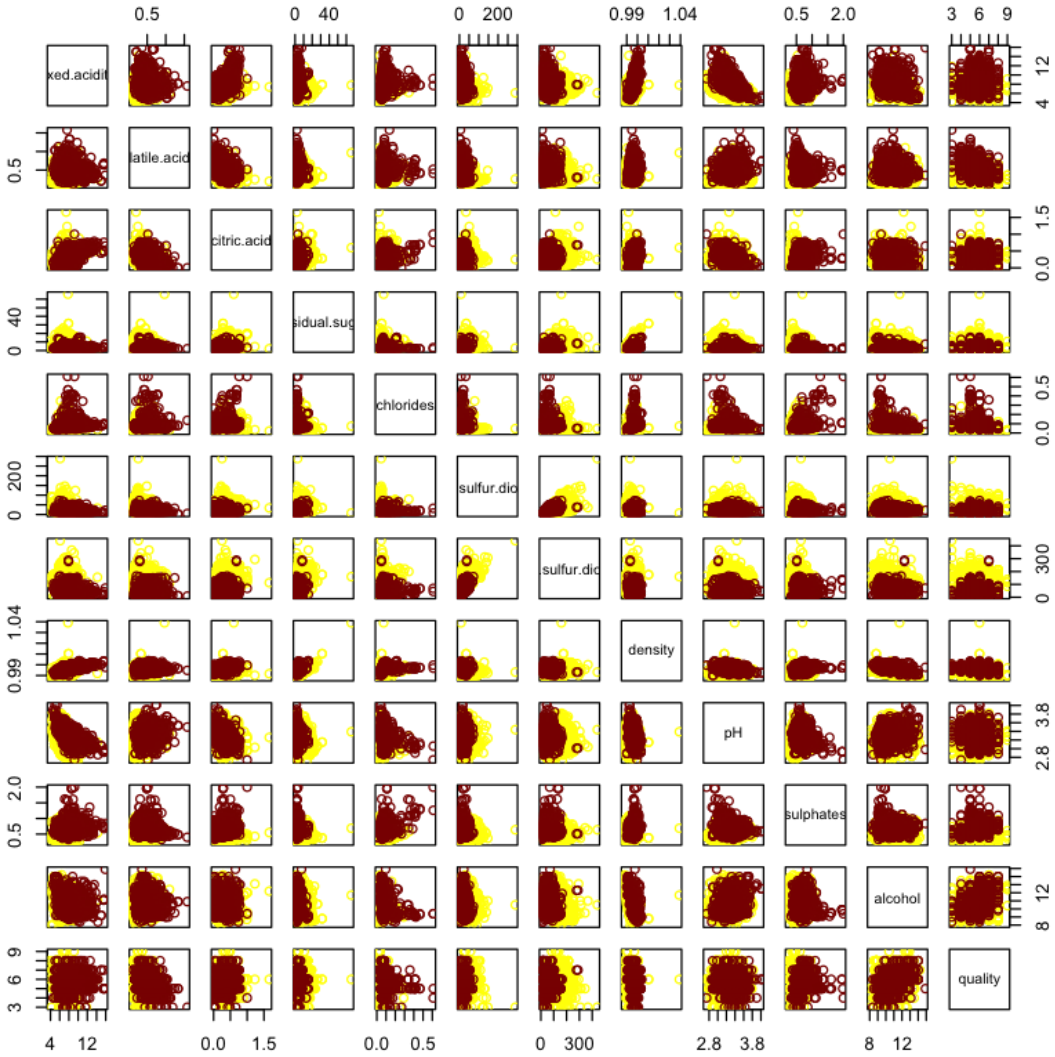
4.3. Correlación entre las variables

Nos interesa comprobar si hay variables correlacionadas entre sí para ver si se pueden eliminar algunas de ellas y para determinar cuáles influyen más en la calidad

Empezamos por un plot de todos los pares de variables. Representamos todos los vinos (distinguiendo por colores los blancos y los tintos) y luego, por separado, blancos y tintos.

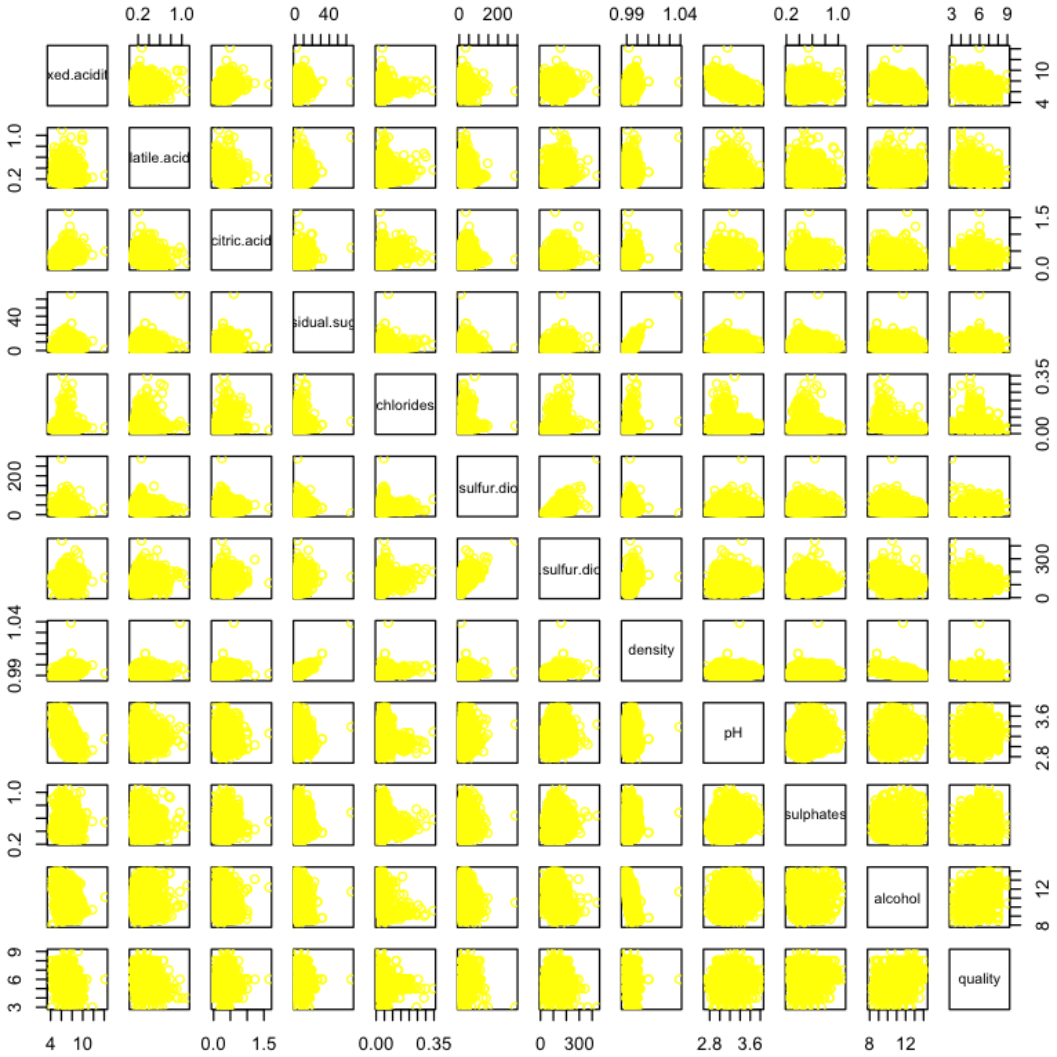
In [18]:

```
colores = ifelse(wines$colour==0, "yellow", "red4")

plot(wines[3:14],col = colores)
```



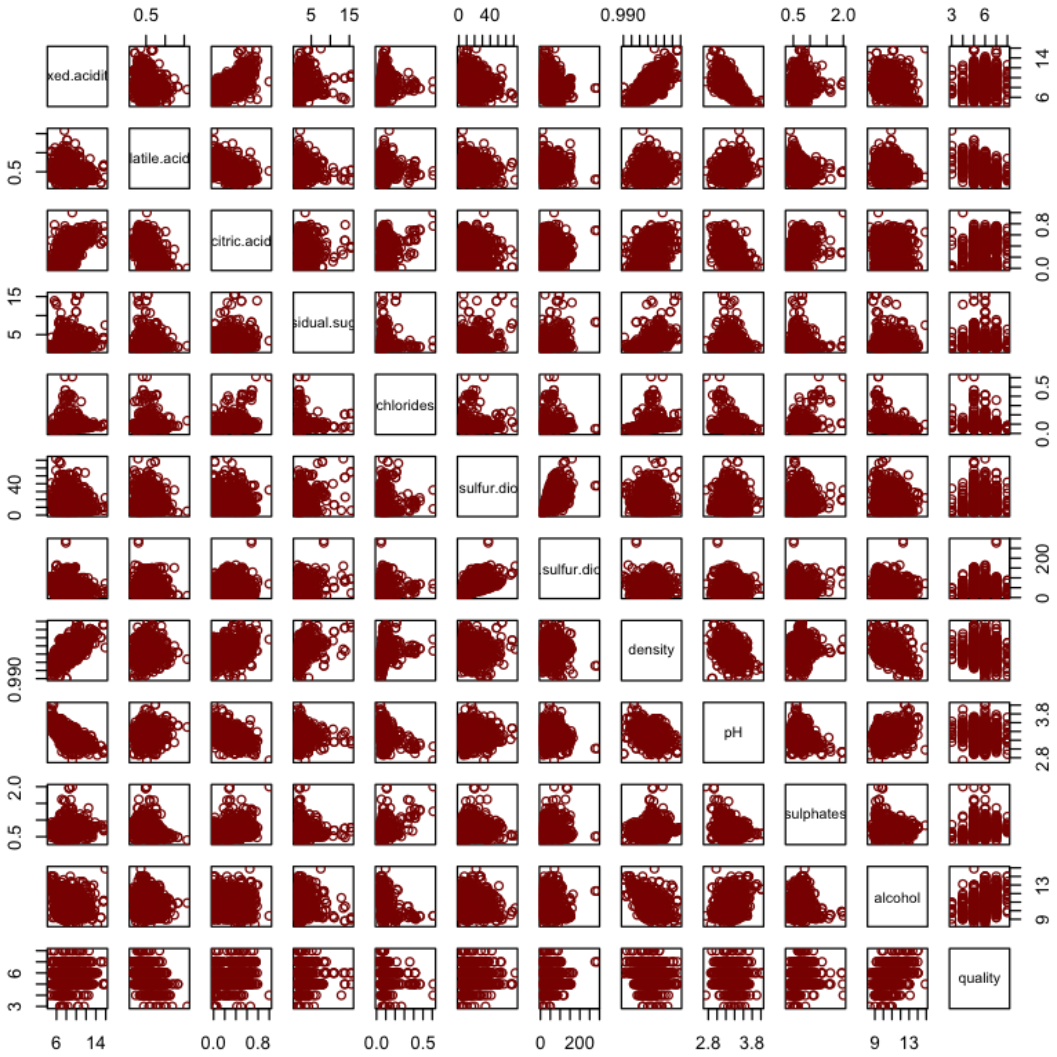
In [19]:

```
# vinos blancos

plot(white_wines[3:14], col = "yellow")
```



In [20]:

```
# vinos tintos

plot(red_wines[3:14], col = "red4")
```



Estas visualizaciones no ayudan demasiado porque son muy pequeñas y no se aprecian los detalles.

Si nos interesara algún plot en concreto, deberíamos representarlo por sepaado.

Sin embargo, no parece que en ningún caso haya un algo índice de correlación.

Vamos a ver los datos con un ***correlogram***, que ayuda a visualizar las matrices de correlación.

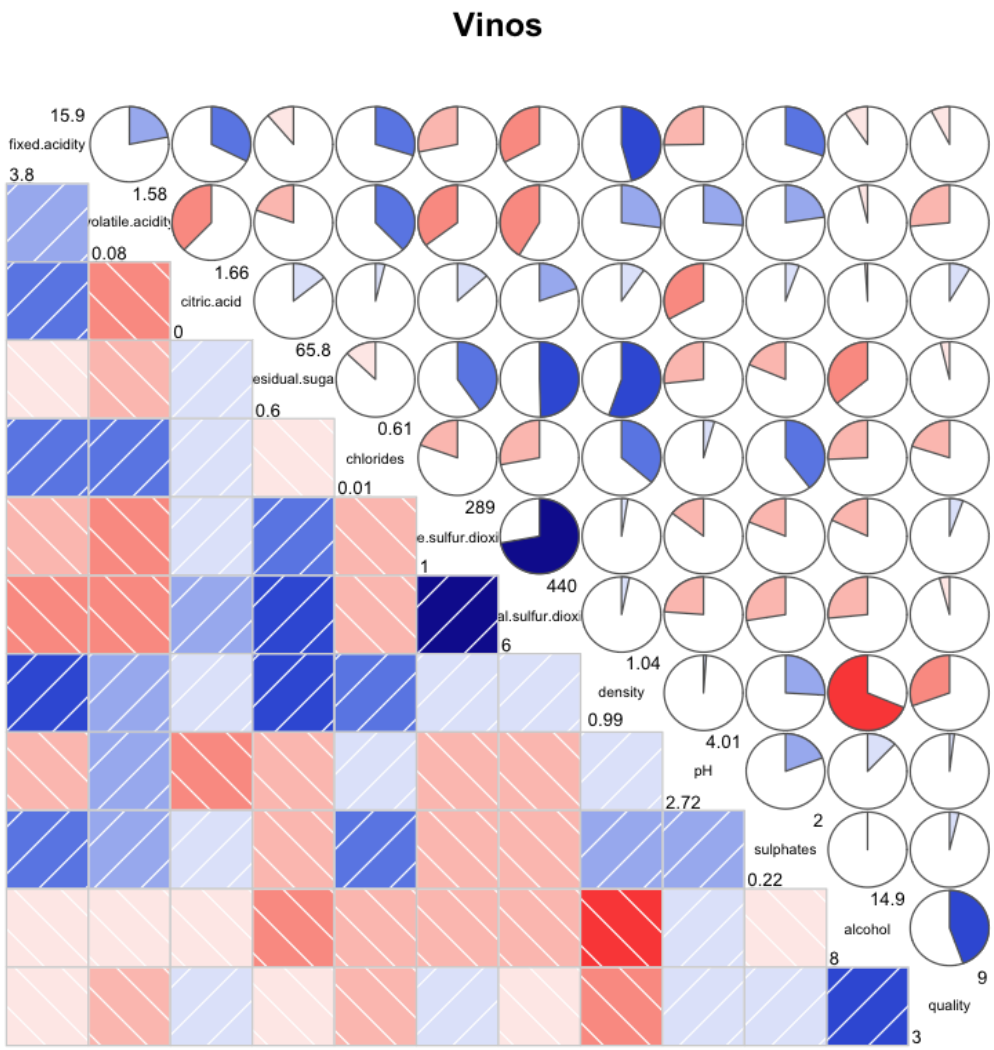
Sobre la diagonal principal se puede ver, mediante un diagrama de tarta, si la correlación es mayor o menor. En la parte inferior, el tono del color indica también la correlación (son dos formas de visualizar la misma información).

Tiene que estar instalada la librería corrgram.

Se hará un correlogram por *dataset*.

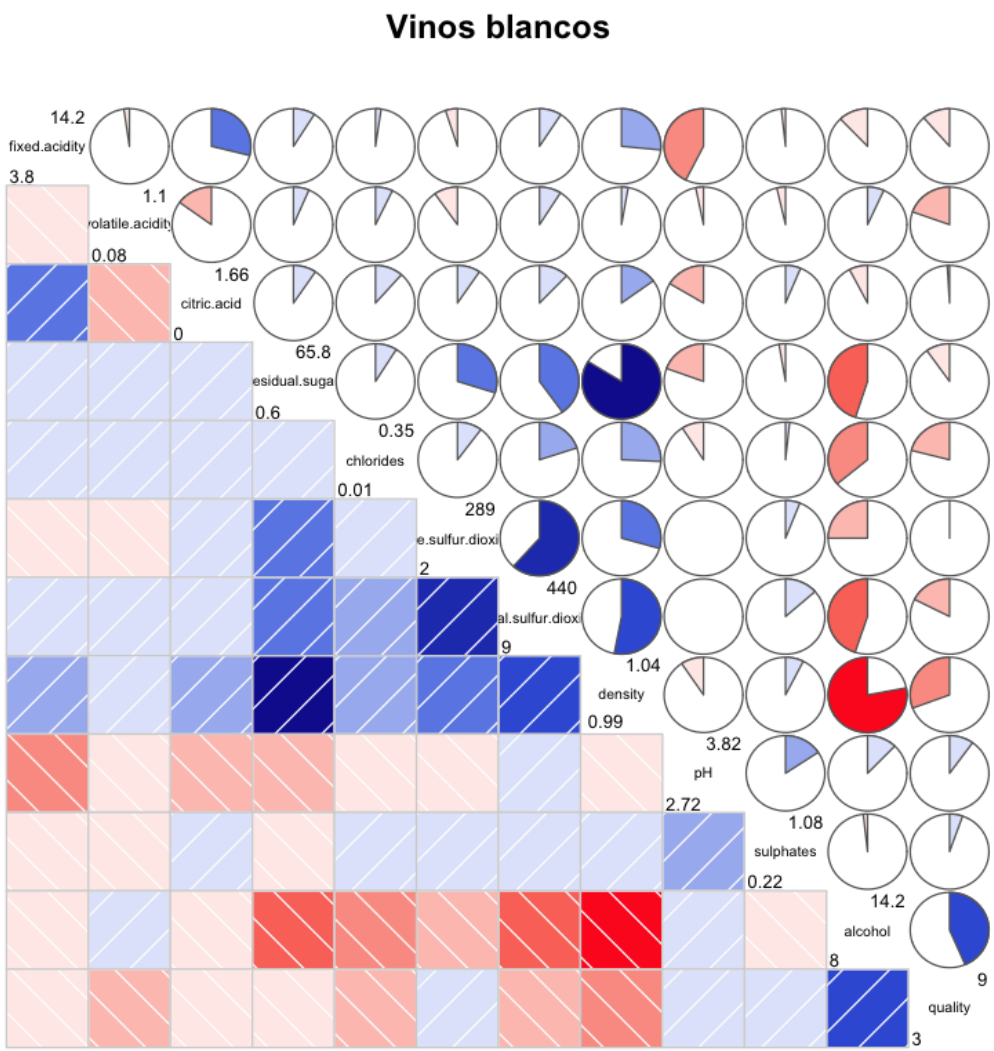
In [21]:

```
library(corrgram)
corrgram(wines[3:14], order=FALSE, lower.panel=panel.shade,
  upper.panel=panel.pie, text.panel=panel.txt,
  diag.panel=panel.minmax,
  main="Vinos")
```



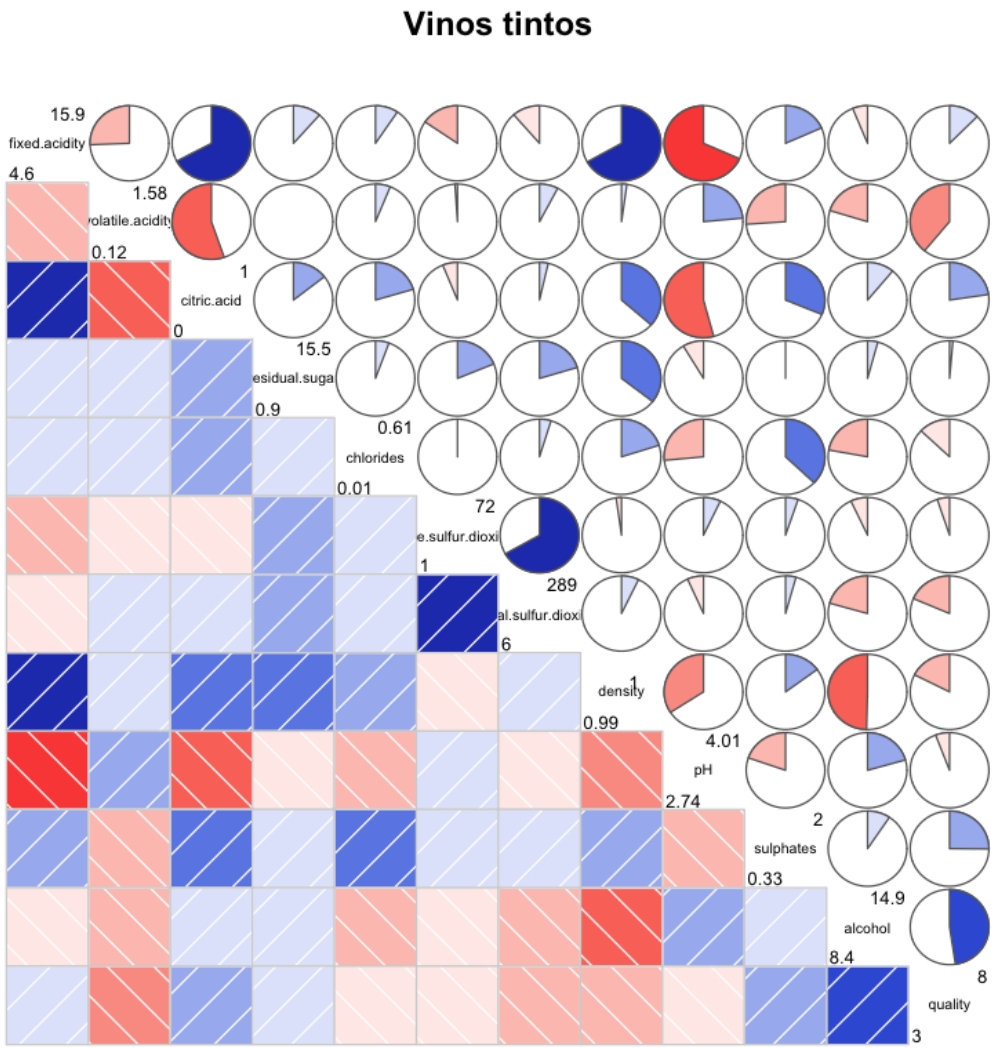
In [22]:

```
corrgram(white_wines[3:14], order=FALSE, lower.panel=panel.shade,
  upper.panel=panel.pie, text.panel=panel.txt,
  diag.panel=panel.minmax,
  main="Vinos blancos")
```



In [23]:

```
corrgram(red_wines[3:14], order=FALSE, lower.panel=panel.shade,
  upper.panel=panel.pie, text.panel=panel.txt,
  diag.panel=panel.minmax,
  main="Vinos tintos")
```



Hay algunas variables correlacionadas, pero no demasiado.

Por ejemplo, con la calidad, solo el porcentaje de alcohol alcanza un valor próximo a la mitad (juntos o separados por clase)

Es alta la correlación positiva entre residual.sugar y density en vinos blancos. Hay correlación negativa entre density y alcohol (menos en los tintos que en los blancos).

'free.sulfur.dioxide' 'total.sulfur.dioxide' están bastante correlacionadas.

En los tintos: 'fixed.acidity'y 'citric.acid', 'fixed.acidity' y 'density' (positiva). Con correlación negativa encontramos 'fixed.acidity' y 'pH', 'volatile.acidity'y 'citric.acid', 'citric.acid'y 'pH'

Este tipo de gráficos proporciona bastante información de un vistazo. (También se puede incluir un plot.)

Vamos a analizar la correlación entre las variables con el coeficiente de correlación de Spearman.

Como las variables que tenemos no siguen distribuciones normales no se debe usar el coeficiente de correlación de Pearson.

Lo aplicaremos al conjunto completo y también separado en blancos y tintos.

En este test no debe haber elementos repetidos, aunque lo puede solucionar internamente y muestra un mensaje de aviso. Se han desactivado en este caso para que la salida sea más clara.

Solo se muestran los resultados del test para aquellos casos en los que el p-value es superior a 0.05 y, por tanto, hay correlación significativa entre las variables.

In [24]:

```
cat("Conjunto completo de vinos \n")
for (i in 3:13) {
    for (j in (i+1):14) {

        test = cor.test(wines[,i], wines[,j], method="spearman", exact = FALSE)

        if (test[["p.value"]] >= alfa) {
            cat(names(wines[i]), " - ", names(wines[j]) )
            print(test)
            cat("     ****** \n")
        }

    }
}
```

```
Conjunto completo de vinos
volatile.acidity  -  alcohol
        Spearman's rank correlation rho

data:  wines[, i] and wines[, j]
S = 4.6801e+10, p-value = 0.05382
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.0239242

     ******
citric.acid  -  alcohol
        Spearman's rank correlation rho

data:  wines[, i] and wines[, j]
S = 4.4809e+10, p-value = 0.1132
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01965297

     ******
residual.sugar  -  quality
        Spearman's rank correlation rho

data:  wines[, i] and wines[, j]
S = 4.648e+10, p-value = 0.1734
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01689059

     ******
free.sulfur.dioxide  -  density
        Spearman's rank correlation rho

data:  wines[, i] and wines[, j]
S = 4.5441e+10, p-value = 0.6379
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
0.005840651

     ******
density  -  pH
        Spearman's rank correlation rho

data:  wines[, i] and wines[, j]
S = 4.5169e+10, p-value = 0.3425
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01177733

     ******
sulphates  -  alcohol
        Spearman's rank correlation rho

data:  wines[, i] and wines[, j]
S = 4.5498e+10, p-value = 0.7119
```

```
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
0.004583412

     ******
```

In [25]:

```
Según los resultados del test,  la correlación de las siguientes variables es si
gnificativa:
volatile.acidity  -  alcohol
citric.acid  -  alcohol
residual.sugar  -  quality
free.sulfur.dioxide  -  density
density  -  pH
sulphates  -  alcohol

Sin embargo, los valores de rho son bastante bajos en todos los casos, muy aleja
dos de los valores -1 o 1.

Vamos a ver los resultados con los conjuntos de datos separados (blancos y tinto
s).
```

```
Error in parse(text = x, srcfile = src): <text>:1:7: unexpected symb
ol
1: Según los
          ^
Traceback:
```

In [ ]:

```
cat("Conjunto de vinos blancos \n")
for (i in 3:13) {
    for (j in (i+1):14) {

        test = cor.test(white_wines[,i], white_wines[,j], method="spearman", exa
ct = FALSE)

        if (test[["p.value"]] >= alfa) {
            print(test)
            cat(names(white_wines[i]), " - ", names(white_wines[j]) , "\n")

        }

    }
}
```

En este *dataset* hay 11 pares de variables cuya correlación es significativa, pero con valores muy bajos, próximos a 0, que indican que no están correlacionados.

In [26]:

```
for (i in 3:13) {
    for (j in (i+1):14) {

        test = cor.test(red_wines[,i], red_wines[,j], method="spearman", exact =
 FALSE)

        if (test[["p.value"]] >= alfa) {
            cat(names(red_wines[i]), " - ", names(red_wines[j]), "\n" )
            print(test)

        }

    }
}
```

```
volatile.acidity  -  residual.sugar

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 659320000, p-value = 0.1955
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.0323856

volatile.acidity  -  free.sulfur.dioxide

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 666970000, p-value = 0.3977
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02116264

volatile.acidity  -  density

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 664340000, p-value = 0.3175
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02501412

citric.acid  -  total.sulfur.dioxide

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 674980000, p-value = 0.7072
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
0.009399602

residual.sugar  -  sulphates

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 655270000, p-value = 0.1255
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.038332

residual.sugar  -  quality

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 659550000, p-value = 0.2002
```

```
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.03204817

chlorides  -  free.sulfur.dioxide

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 680840000, p-value = 0.9743
alternative hypothesis: true rho is not equal to 0
sample estimates:
         rho
0.0008051686

chlorides  -  sulphates

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 667200000, p-value = 0.4053
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02082548

free.sulfur.dioxide  -  density

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 709450000, p-value = 0.09976
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.04117768

free.sulfur.dioxide  -  sulphates

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 650140000, p-value = 0.06674
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.04586235

total.sulfur.dioxide  -  pH

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 688090000, p-value = 0.6941
alternative hypothesis: true rho is not equal to 0
sample estimates:
         rho
-0.009841438

total.sulfur.dioxide  -  sulphates
```

```
        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 681730000, p-value = 0.9839
alternative hypothesis: true rho is not equal to 0
sample estimates:
          rho
-0.0005038194

pH  -  quality

        Spearman's rank correlation rho

data:  red_wines[, i] and red_wines[, j]
S = 711140000, p-value = 0.08085
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.04367193
```

El el conjunto de vinos tintos hay 13 pares de variables cuyos valores de correlación son significativos estadísticamente, pero con valores muy bajos, cercanos a 0, lo que indica que son variables poco correlacionadas.

4.4. Modelo de regresión lineal

Vamos a proponer algunos modelos de regresión lineal para intentar explicar la calidad de un vino a partir de algunos subconjuntos de características químicas.

En el siguiente artículo se indican 6 criterios que determinan la calidad del vino: http://www.vinopack.es/criterios-que-determinan-la-calidad-en-el-vino (http://www.vinopack.es/criterios-que-determinan-la-calidad-en-el-vino)

De los seis, en nuestro *dataset* se incluyen cuatro: densidad, alcohol, ph, acidez volatil. No hay información sobre el color (no blanco o tinto, sino otros matices) o el hierro presente.

Vamos a ver si existe un modelo de regresión lineal para explicar la calidad a partir de esas 4 características.

Se hace la prueba para el conjunto de vinos tintos.

In [27]:

```
fit = lm(red_wines$quality ~ red_wines$density + red_wines$alcohol + red_wines$p
H
         + red_wines$volatile.acidity, data=redwines)
summary(fit)
```

```
Call:
lm(formula = red_wines$quality ~ red_wines$density + red_wines$alcoh
ol +
    red_wines$pH + red_wines$volatile.acidity, data = redwines)

Residuals:
     Min       1Q   Median       3Q      Max
-2.56329 -0.39869 -0.07628  0.45888  2.23283

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -8.12905   10.75099  -0.756  0.44969
red_wines$density          12.21377   10.58512   1.154  0.24873
red_wines$alcohol           0.33962    0.01854  18.314  < 2e-16 ***
red_wines$pH               -0.38495    0.11939  -3.224  0.00129 **
red_wines$volatile.acidity -1.27742    0.09911 -12.889  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6652 on 1594 degrees of freedom
Multiple R-squared:  0.3233,    Adjusted R-squared:  0.3216
F-statistic: 190.4 on 4 and 1594 DF,  p-value: < 2.2e-16
```

Podemos ver que los resultados del ajuste no es bueno y que no es significativo estadísticamente.

Si probamos ahora con el vino blanco y con el conjunto de vinos, obtenemos los mismos resultados.

In [28]:

```
fit = lm(white_wines$quality ~ white_wines$density + white_wines$alcohol + white
_wines$pH
         + white_wines$volatile.acidity, data=whitewines)
summary(fit)
```

```
Call:
lm(formula = white_wines$quality ~ white_wines$density + white_wines
$alcohol +
    white_wines$pH + white_wines$volatile.acidity, data = whitewine
s)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4056 -0.4870 -0.0434  0.4841  3.0322

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)

(Intercept)                 -37.04119    5.99835  -6.175 7.14e-10 *
**
white_wines$density          38.86584    5.91549   6.570 5.55e-11 *
**
white_wines$alcohol           0.39530    0.01445  27.349  < 2e-16 *
**
white_wines$pH                0.22014    0.07330   3.003  0.00268 *
*
white_wines$volatile.acidity  -2.05852    0.11016 -18.687  < 2e-16 *
**
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7682 on 4893 degrees of freedom
Multiple R-squared:  0.2483,    Adjusted R-squared:  0.2477
F-statistic:   404 on 4 and 4893 DF,  p-value: < 2.2e-16
```

In [29]:

```
fit = lm(wines$quality ~ wines$density + wines$alcohol + wines$pH + wines$volati
le.acidity, data=wines)
summary(fit)
```

```
Call:
lm(formula = wines$quality ~ wines$density + wines$alcohol +
    wines$pH + wines$volatile.acidity, data = wines)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4363 -0.4765 -0.0389  0.4723  3.0274

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -35.03370    4.57019  -7.666 2.04e-14 ***
wines$density           37.05778    4.52401   8.191 3.08e-16 ***
wines$alcohol            0.37860    0.01104  34.308  < 2e-16 ***
wines$pH                 0.16747    0.06036   2.774  0.00555 **
wines$volatile.acidity  -1.53195    0.06168 -24.838  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7473 on 6492 degrees of freedom
Multiple R-squared:  0.2682,    Adjusted R-squared:  0.2677
F-statistic: 594.7 on 4 and 6492 DF,  p-value: < 2.2e-16
```

Vamos a intentar otro modelo con dos características (azucar y pH) sobre el conjunto de los vinos tintos.

In [30]:

```
fit = lm(redwines$quality ~ redwines$residual.sugar + redwines$pH, data = redwin
es)
summary(fit)
```

```
Call:
lm(formula = redwines$quality ~ redwines$residual.sugar + redwines$p
H,
    data = redwines)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6788 -0.6401  0.3027  0.3902  2.4886

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.60992    0.43949  15.040   <2e-16 ***
redwines$residual.sugar 0.00507    0.01437   0.353   0.7242
redwines$pH            -0.29802    0.13119  -2.272   0.0232 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8067 on 1596 degrees of freedom
Multiple R-squared:  0.003411,  Adjusted R-squared:  0.002162
F-statistic: 2.731 on 2 and 1596 DF,  p-value: 0.06546
```

Aunque el p-valor es superior a 0.05, el ajuste proporcionado es muy bajo.

Parece claro que, en el caso que nos ocupa, estas variables no presentan un ajuste lineal.

5. Representación de resultados

Las representaciones gráficas se han ido haciendo a lo largo de los puntos anteriores.

6. Conclusiones

- Los conjuntos de datos de partida estaban bien preparados y no ha habido que realizar trabajos de limpieza y preparación complejos.
- Aunque hay muchos valores atípicos, no se han eliminado porque pueden representar valores posibles.
- En general, las características analizadas no seguían distribuciones normales, ni se ha visto correlación significativas entre ellas.
- Hay demasiadas medias y creo que no he conseguido determinar aquellas de las que podíamos prescindir por no servir para explicar la calidad del vino.
- Sería interesante usar este *dataset* para intentar realizar predicciones de la calidad del vino con árboles de regresión o redes neuronales, por ejemplo.

Referencias

- Dataset (vinos tintos): https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009 (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)
- Dataset completo: https://archive.ics.uci.edu/ml/datasets/wine+quality (https://archive.ics.uci.edu/ml/datasets/wine+quality)
- P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
- Dalgaard, Peter. Introductory statistics with R (Second Edition). New York : Springer, 2002. ISBN 038722632X
- https://www.statmethods.net/advgraphs/correlograms.html (https://www.statmethods.net/advgraphs/correlograms.html)