Tipología y ciclo de vida de los datos: Práctica2

Antonio Guzmán Martín & Joaquín Fernández León

Diciembre 2018

Contents

---

1. Descripción del dataset

Comenzamos con la lectura del dataset.

```
library(readr)
redwine <- read.csv(file="winequality-red.csv", header = TRUE)
#numero de filas por dataset
nrow(redwine)
```

```
## [1] 1599
```

```
ncol<-ncol(redwine)
#sacamos 5 primeras filas
head(redwine[,1:ncol])
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
```

```
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
## 6                  13                  40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```
sapply(redwine,class)
```

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

Variables del dataset

- **fixed acidity:** Conjunto de los ácidos naturales procedentes de la uva (tartárico, málico, cítrico y succínico) o formados en la fermentación maloláctica (láctico). En general, los ácidos (acidez fija) son preservante naturales del vino y ayuda a mantener el color y cualidades aromáticas.
- **volatile acidity:** Conjunto de ácidos formados durante la fermentación o como consecuencia de alteraciones microbianas. Estos ácidos son, principalmente: ácido Acético, ácido Propionico, ácido Butírico y ácido Sulfúrico. Si la acidez volátil, presente en todos los vinos, es muy elevada el vino se picará y avigranará con el paso del tiempo. Es conveniente que la acidez volatil de un vino sea lo más baja posible.

El contenido en acidez volátil no puede ser superior a: a) 18 miliequivalentes por litro para los mostos de uva parcialmente fermentados, b) 18 miliequivalentes por litro para los vinos blancos y rosados, c) 20 miliequivalentes por litro para los vinos tintos.

- **citric acid:** En pequeñas cantidades este hacido puede añadir frescor y sabor a los vinos (dentro de acido fijo).
- **residual sugar:** Azúcar que queda en el vino después de la fermentación. Es raro encontrar vinos con menos de 1 g/l y vinos con más de 45 g/l son considerados dulces.
- **chlorides:** cantidad de sal en el vino.
- **free sulfur dioxide :** Previene del crecimiento microbial y de la oxidación del vino.La oxidación enturbia sus colores característicos (tornándolos en amarillos intensos e, incluso, marrones).Por lo que respecta al gusto, al beberlo notaremos sabores más secos y ásperos, incluso amargos en algunos casos.
- **total sulfur dioxide:** suma de concentraciones libres y amarradas de S02; concentraciones de dioxodo de sulfuro libres superiores a 50 ppm se vuelven evidentes en el sabor y olor.
- **density:** densidad del vino, suele ser similar al del agua dependiendo de la concentración de azucar y alcohol.

- **pH:** Describe como de ácido o básico es el vino 0 (very acidic) to 14 (very basic); mayoria vinos en escala 3-4.(principalmente 3,55 a 4).
- **sulphates:** Actua como un antimicrobial and antioxidante. Los sulfatos de sodio y calcio aparecen en el agua y por lo tanto la uva y el vino pueden contenerlos. Un agua con una cantidad de sulfatos inferior a 250mg/l se considera en este aspecto un agua de calidad y con valores superiores a 400mg/l insalubre.
- **alcohol:** cantidad de alcohol del vino. No es muy útil para hallar la calidad.
- **quality:** calidad del vino entre 0 y 10.

1.1 ¿Por qué es importante y qué pregunta/problema pretende responder?

Pregunta: ¿Qué componentes fisico-químicos influyen en que un vino sea bueno?. Obtener un modelo cuya combinación de variables permita determinar si es un buen vino.

---

2. Integración y selección de los datos de interés a analizar.

La mayoría de los atributos corresponden con características necesarias para determinar la calidad del vino.

Sin embargo, a priori, podemos prescindir de la variable `total sulfur dioxide` (indica elsuma de concentraciones libres y amarradas, solo nos interesan las libres) y `density` (indica proporción de alcohol y esta no es interesante para determinar la calidad) (Según el estudio de: https://www.vinopack.es/criterios-que-determinan-la-calidad-en-el-vino).

No obstante, en los siguientes apartados comprobaremos si esto es cierto, o por el contrario si que afecta en la calidad.

---

3 Limpieza de datos

3.1 Elementos vacíos

```
# Números de valores desconocidos por campo
sapply(redwine, function(x) sum(is.na(x)))
```

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

```r
#No se han encontrado valores vacíos o NAs.
# Resumen de las variables
summary(redwine)
```

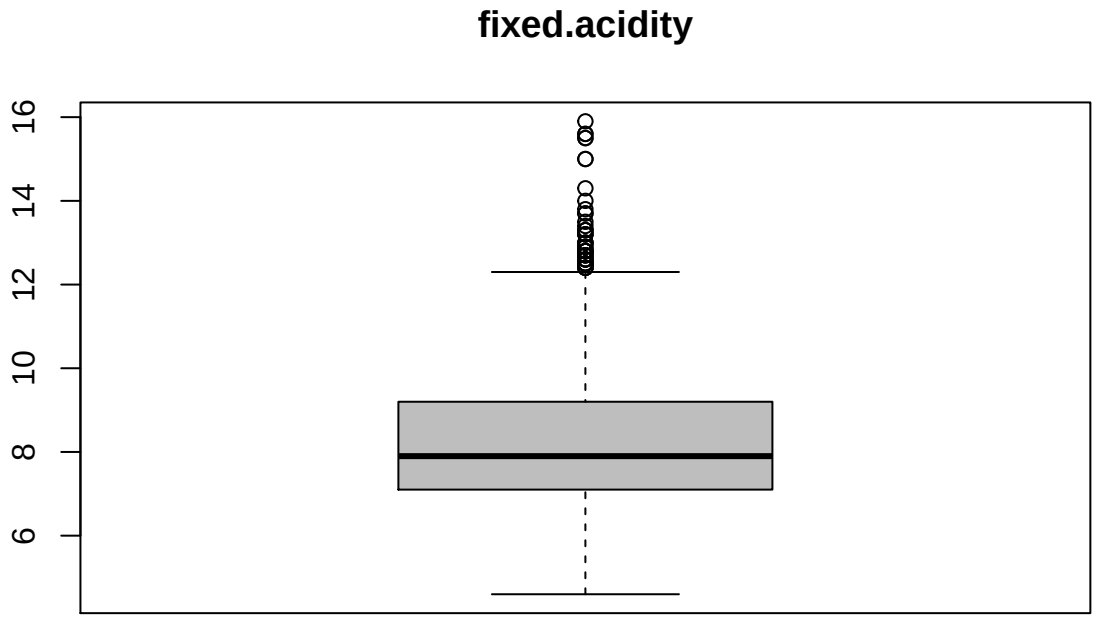
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean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

```r
#ph: correcto (entre 2 y 4)
```

3.2 Valores extremos

Para cada una de las variables observemos si existen valores atípicos:

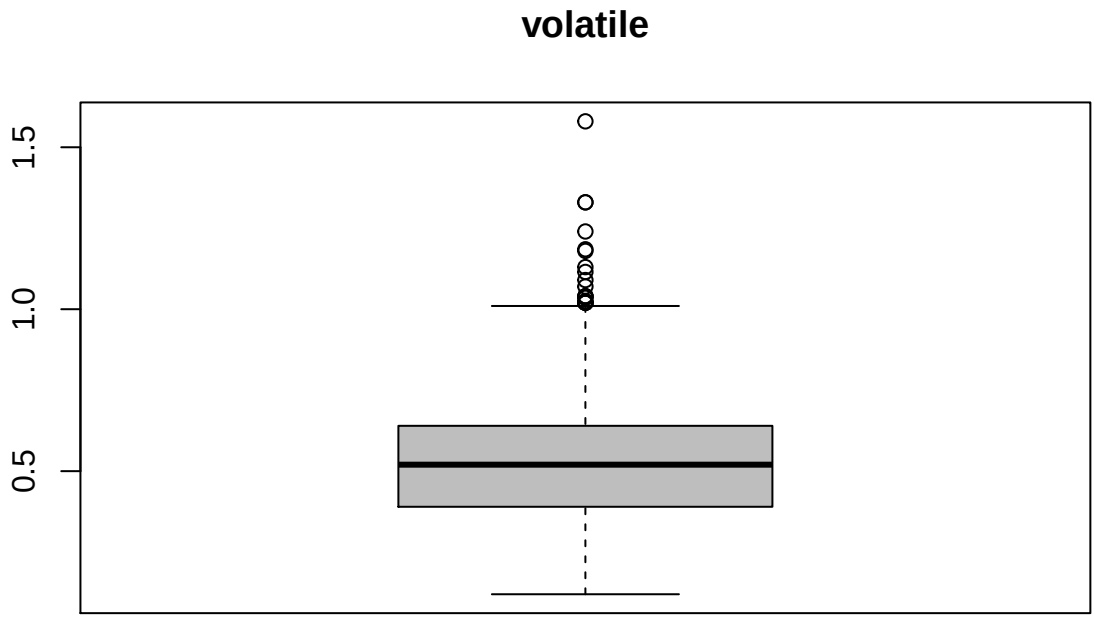
```r
boxplot(redwine$fixed.acidity,main = "fixed.acidity",col="gray")
```




```
boxplot.stats(redwine$fixed.acidity)$out
```

```
##  [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8
## [15] 12.8 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4
## [29] 12.5 12.9 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2
## [43] 13.2 13.2 15.9 13.3 12.9 12.6 12.6
```

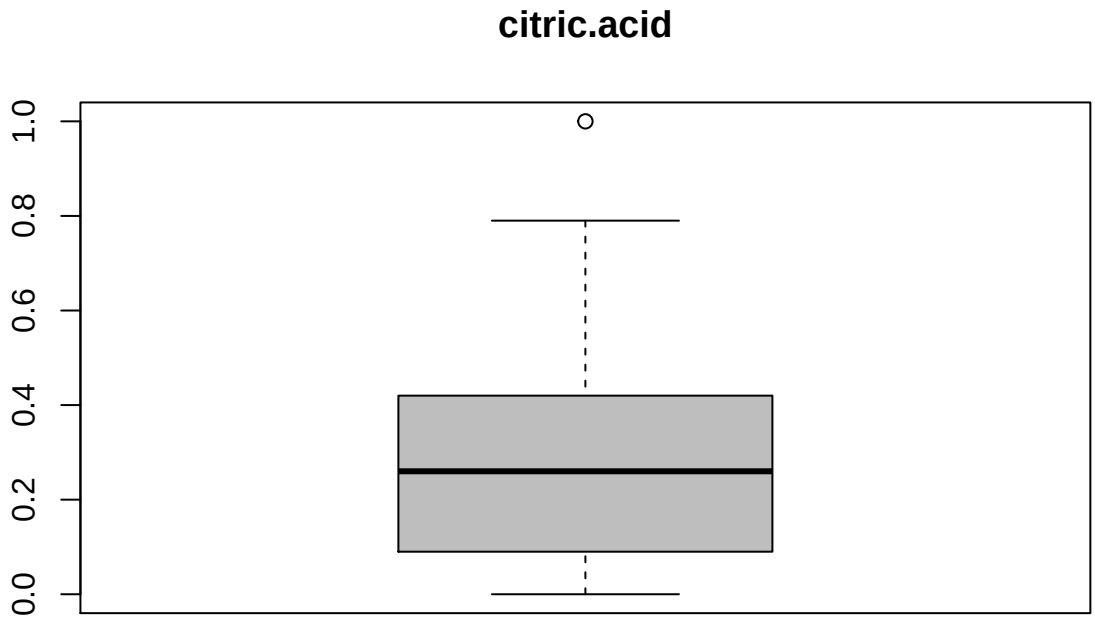
```
boxplot(redwine$volatile.acidity,main = "volatile",col="gray")
```




```
boxplot.stats(redwine$volatile.acidity)$out
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020
## [12] 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

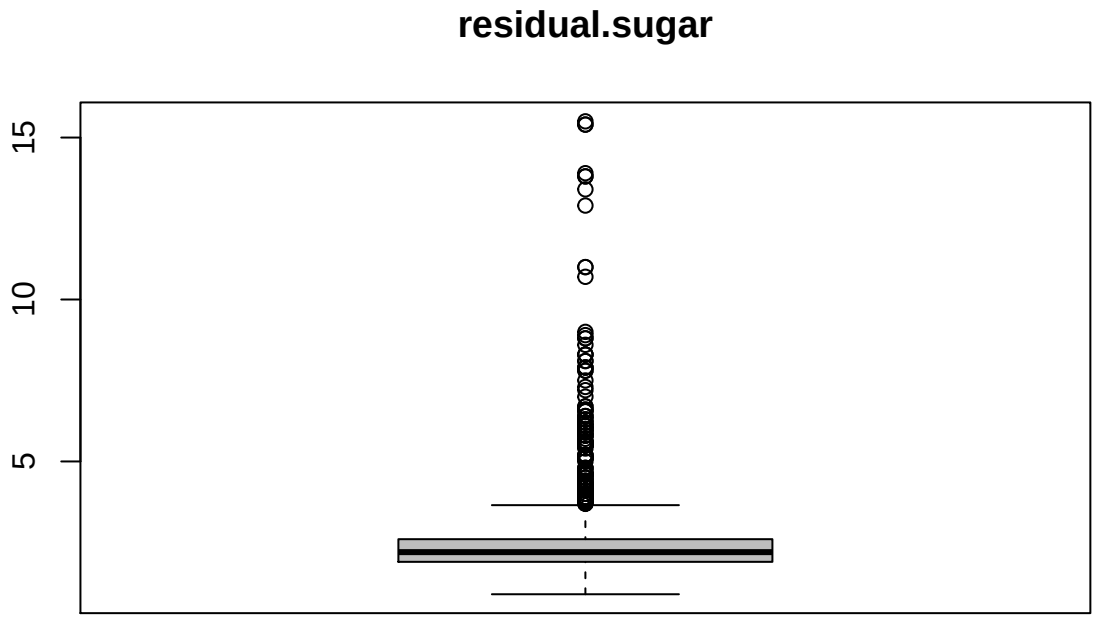
```
boxplot(redwine$citric.acid,main = "citric.acid",col="gray")
```




```
boxplot.stats(redwine$citric.acid)$out
```

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

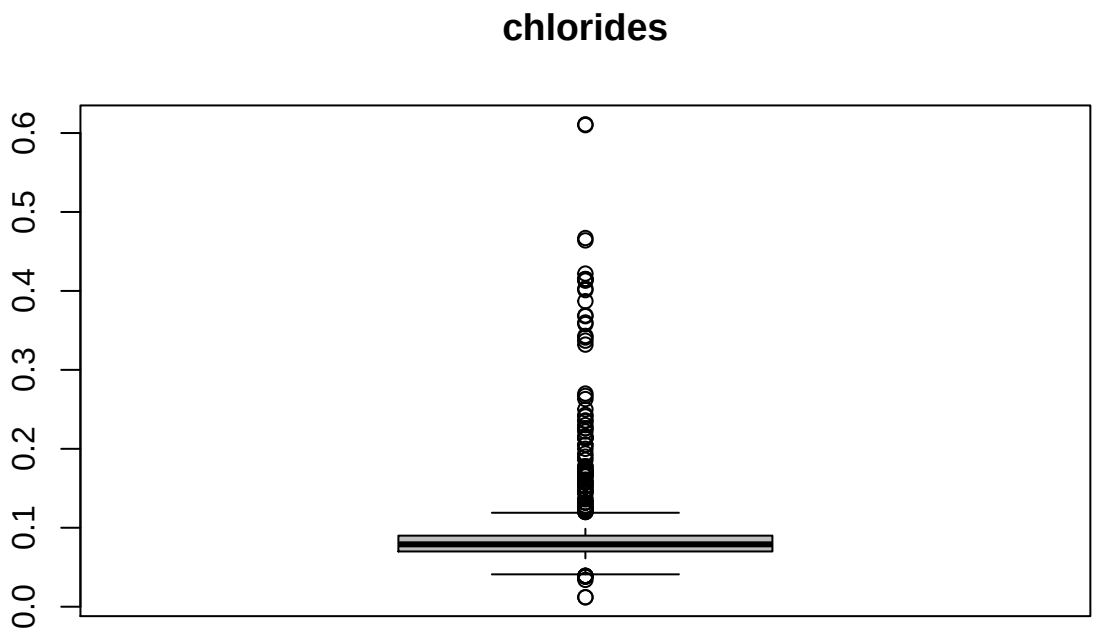
```
boxplot(redwine$residual.sugar,main = "residual.sugar",col="gray")
```




```
boxplot.stats(redwine$residual.sugar)$out
```

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10
##  [12]  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00
##  [23]  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00
##  [34]  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70
##  [45]  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60
##  [56]  6.10  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30
##  [67]  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40  4.25
##  [78]  6.00  3.90  4.20  4.00  4.00  4.00  6.60  6.00  6.00  3.80  9.00
##  [89]  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00
## [100]  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50
## [111]  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90  4.30
## [122] 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80
## [133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75
## [144] 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10
## [155]  7.80
```

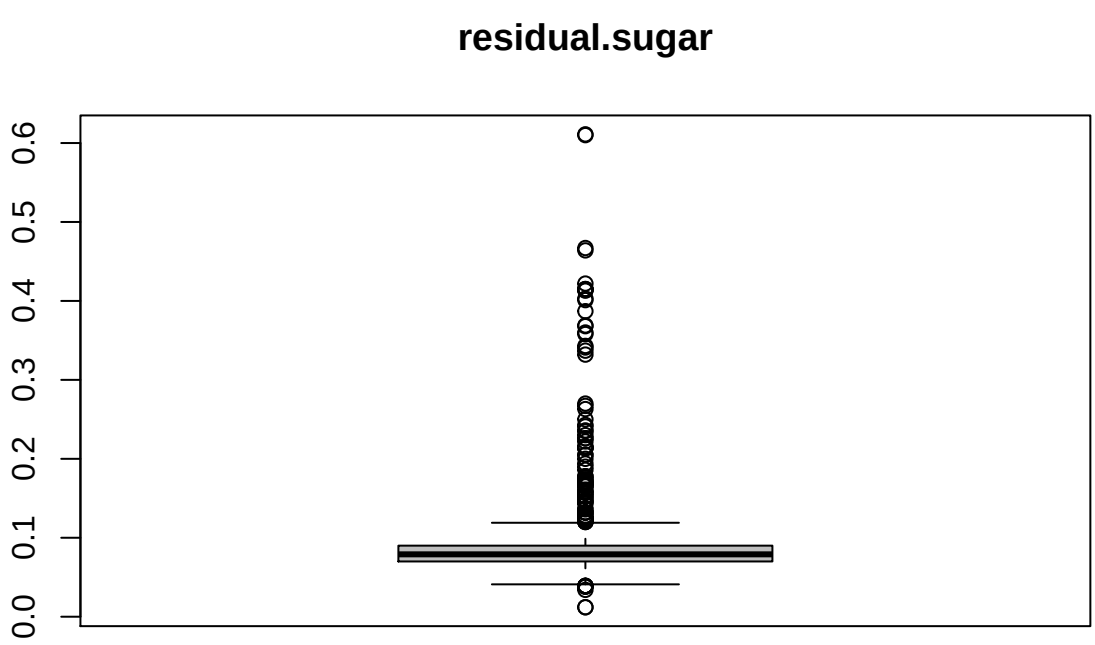
```
boxplot(redwine$chlorides,main = "chlorides",col="gray")
```




```
boxplot.stats(redwine$residual.sugar)$out
```

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10
##  [12]  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00
##  [23]  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00
##  [34]  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70
##  [45]  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60
##  [56]  6.10  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30
##  [67]  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40  4.25
##  [78]  6.00  3.90  4.20  4.00  4.00  4.00  6.60  6.00  6.00  3.80  9.00
##  [89]  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00
## [100]  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50
## [111]  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90  4.30
## [122] 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80
## [133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75
## [144] 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10
## [155]  7.80
```

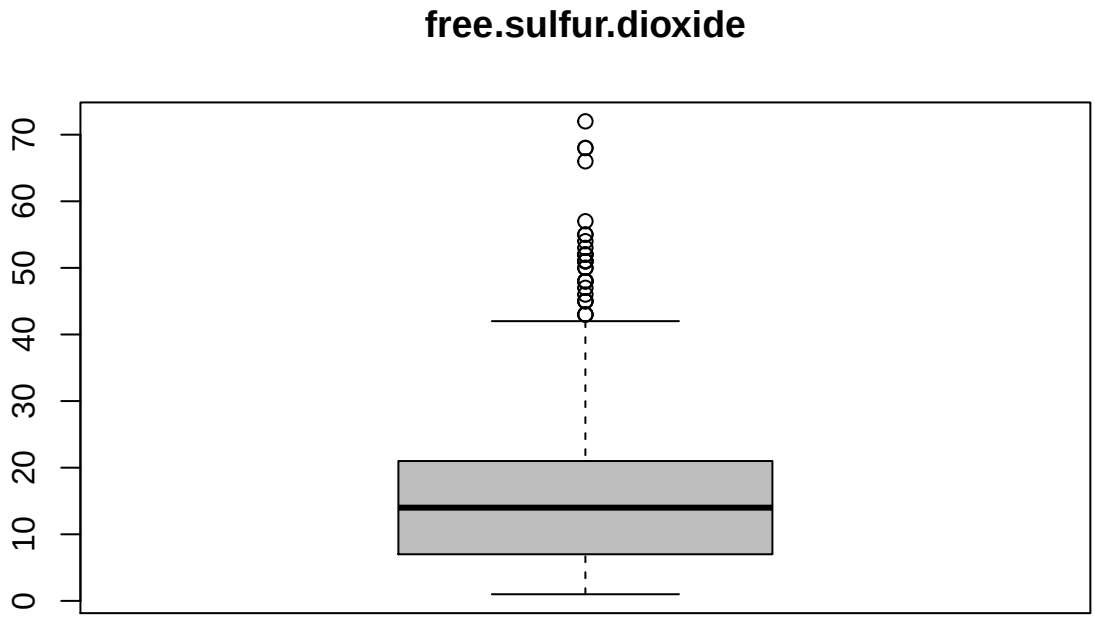
```
boxplot(redwine$chlorides,main = "residual.sugar",col="gray")
```




```
boxplot.stats(redwine$residual.sugar)$out
```

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10
##  [12]  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00
##  [23]  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00
##  [34]  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70
##  [45]  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60
##  [56]  6.10  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30
##  [67]  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40  4.25
##  [78]  6.00  3.90  4.20  4.00  4.00  4.00  6.60  6.00  6.00  3.80  9.00
##  [89]  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00
## [100]  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50
## [111]  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90  4.30
## [122] 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80
## [133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75
## [144] 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10
## [155]  7.80
```

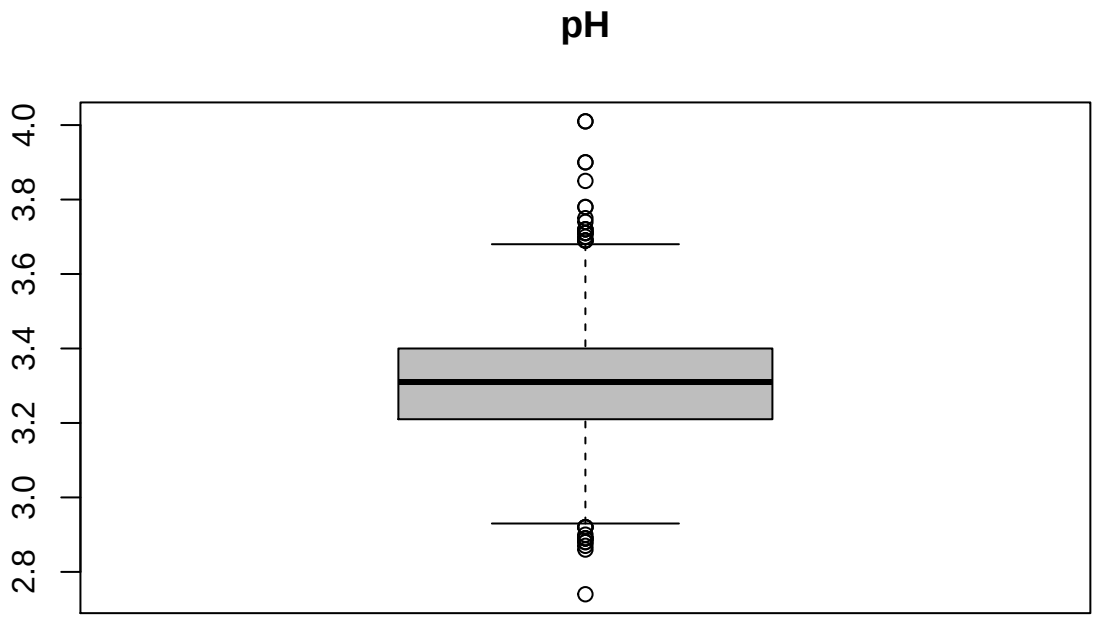
```
boxplot(redwine$free.sulfur.dioxide,main = "free.sulfur.dioxide",col="gray")
```




```
boxplot.stats(redwine$free.sulfur.dioxide)$out
```

```
##  [1] 52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51
## [24] 51 52 55 55 48 48 66
```

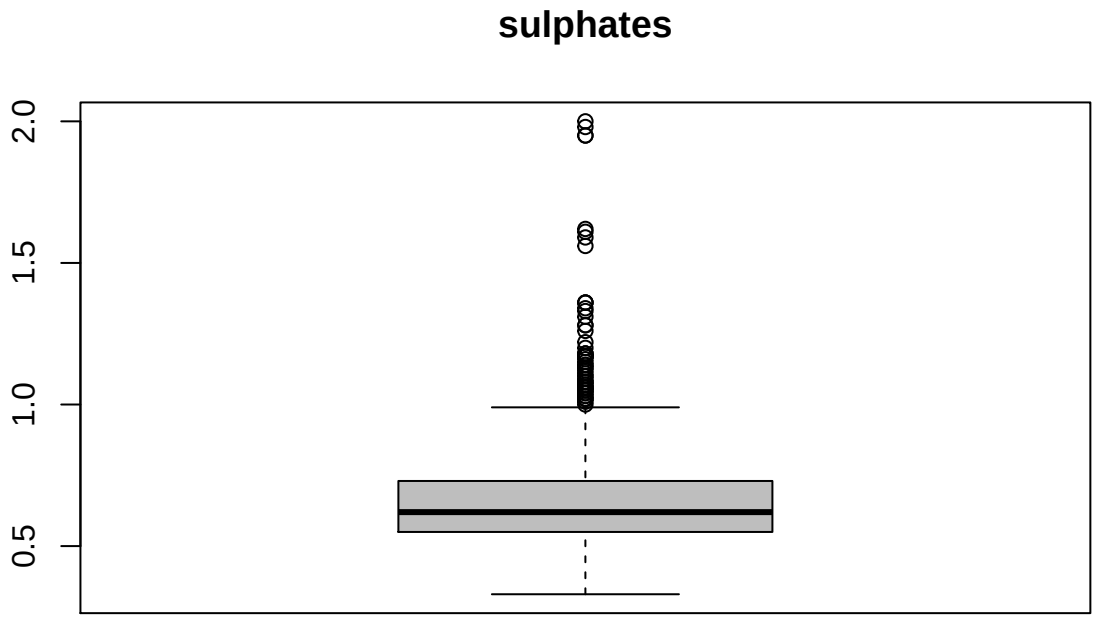
```
boxplot(redwine$pH,main = "pH",col="gray")
```




```
boxplot.stats(redwine$pH)$out
```

```
##  [1] 3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87
## [15] 2.89 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71 2.89 2.89 3.78 3.70 3.78
## [29] 4.01 2.90 4.01 3.71 2.88 3.72 3.72
```

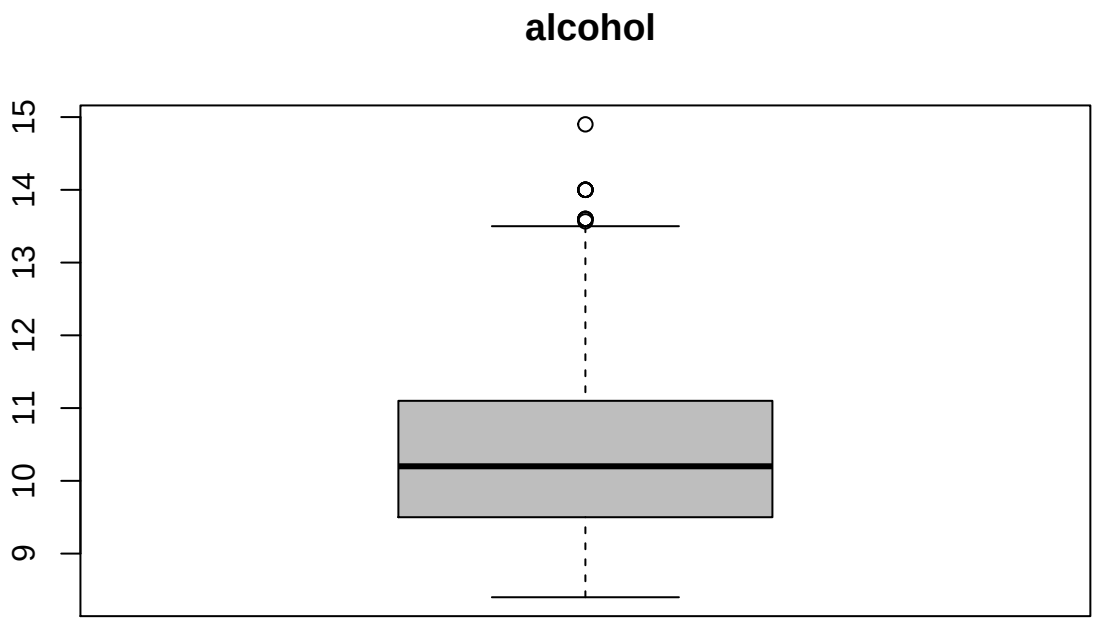
```
#datos correctos porque los valores de pH estan entre 2 y 7
boxplot(redwine$sulphates,main = "sulphates",col="gray")
```




```
boxplot.stats(redwine$sulphates)$out
```

```
##  [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08
## [15] 1.59 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13
## [29] 1.07 1.06 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17
## [43] 1.62 1.06 1.18 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03
## [57] 1.17 1.10 1.01
```

```
boxplot(redwine$ alcohol,main = " alcohol",col="gray")
```




```
boxplot.stats(redwine$ alcohol)$out
```

```
##  [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000
##  [8] 13.60000 13.60000 14.00000 14.00000 13.56667 13.60000
```

```
#alcohol está dentro de unos rangos adecuados
```

Todas las gráficas presentan outliers pero dado a que, la creación de un vino es meramente una reacción química, se puede decir que los valores que se presentan son posibles y simplemente podemos decir que dichos outliers pertenecen a vinos que tienen características muy diferentes al resto.

---

4. Análisis de datos.

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Vamos a analizar si son significativas cada variable del dataset para ello haremos un test t con un nivel de significación=0.05.

Diferentes hipótesis se pueden establecer cambiando el valor de parámetro alternative en la función t.test

1. **Alternative='two.sided':** Hipótesis nula H0: $\mu 1 = \mu 2$ Hipótesis alternativa H1: $\mu 1 != \mu 2$

2. **Alternative='greater':** Hipótesis nula H0: $\mu 1 = \mu 2$ Hipótesis alternativa H1: $\mu 1 > \mu 2$
3. **Alternative='less':** Hipótesis nula H0: $\mu 1 = \mu 2$ Hipótesis alternativa H1: $\mu 1 < \mu 2$

```
 # A más alcohol, hay más calidad, significativo
high_alcohol<-quantile(redwine$alcohol, probs =0.75)
redwine.altoAlcohol<-redwine[redwine$alcohol>=high_alcohol,]$quality
redwine.bajoAlcohol<-redwine[redwine$alcohol<high_alcohol,]$quality
t.test(redwine.altoAlcohol, redwine$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altoAlcohol and redwine$quality
## t = 12.555, df = 628.55, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4891266       Inf
## sample estimates:
## mean of x mean of y
##  6.199017  5.636023
```

```
t.test(redwine.bajoAlcohol, redwine$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.bajoAlcohol and redwine$quality
## t = -6.6585, df = 2711.5, p-value = 1.668e-11
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1447275
## sample estimates:
## mean of x mean of y
##  5.443792  5.636023
```

El `p-valor` es menor que el nivel de significación por lo que rechazamos la hipótesis nula y podemos decir que a más cantidad de alcohol hay más calidad y por tanto que la variable `alcohol` es significativa.

```
# A más azucar no hay más calidad y a menos azúcar tampoco
high_sugar<-quantile(redwine$residual.sugar, probs =0.75)
redwine.altoGradoAzucar<-redwine[redwine$residual.sugar>=high_sugar,]$quality
redwine.bajoGradoAzucar<-redwine[redwine$residual.sugar<high_sugar,]$quality
t.test(redwine.altoGradoAzucar, redwine$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altoGradoAzucar and redwine$quality
## t = 0.049196, df = 664.01, p-value = 0.4804
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.07231395         Inf
## sample estimates:
## mean of x mean of y
##  5.638249  5.636023
```

```
t.test(redwine.bajoGradoAzucar, redwine$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.bajoGradoAzucar and redwine$quality
## t = -0.026921, df = 2532, p-value = 0.4893
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.04986301
## sample estimates:
## mean of x mean of y
##  5.635193  5.636023
```

El `p-valor` no es menor que el nivel de significación por lo que no podemos rechazar la hipótesis nula y por tanto que la variable `residual.sugar` no es significativa.

```
#  Volatile.acidity, significativo
high_volatile.acidity<-quantile(redwine$volatile.acidity, probs =0.75)
redwine.altoVolatile.acidity<-redwine[redwine$volatile.acidity>=high_volatile.acidity,]$quality
redwine.bajoVolatile.acidity<-redwine[redwine$volatile.acidity<high_volatile.acidity,]$quality

t.test(redwine.altoVolatile.acidity, redwine$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altoVolatile.acidity and redwine$quality
## t = -8.6351, df = 669.24, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.2918063
## sample estimates:
## mean of x mean of y
##  5.275434  5.636023
```

```
t.test(redwine.bajoVolatile.acidity, redwine$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.bajoVolatile.acidity and redwine$quality
## t = 3.9702, df = 2595, p-value = 3.689e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.07114568        Inf
## sample estimates:
## mean of x mean of y
##  5.757525  5.636023
```

El `p-valor` es menor que el nivel de significación por lo que rechazamos la hipótesis nula y podemos decir que la variable `volatile.acidity` es significativa.

```
#sulphates, significativo
high_sulphates<-quantile(redwine$sulphates, probs =0.75)
redwine.altoSulphates<-redwine[redwine$sulphates>=high_sulphates,]$quality
redwine.bajoSulphates<-redwine[redwine$sulphates<high_sulphates,]$quality
```

```r
t.test(redwine.altoSulphates, redwine$quality, alternative = "greater") # significativo sulfato alto
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altoSulphates and redwine$quality
## t = 8.9216, df = 628.67, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3304164       Inf
## sample estimates:
## mean of x mean of y
##  6.041262  5.636023
```

```r
t.test(redwine.bajoSulphates, redwine$quality, alternative = "less") #
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.bajoSulphates and redwine$quality
## t = -4.73, df = 2646.4, p-value = 1.182e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.09172558
## sample estimates:
## mean of x mean of y
##  5.495366  5.636023
```

El p-valor es menor que el nivel de significación por lo que rechazamos la hipótesis nula y podemos decir que la variable sulphates es significativa.

```r
#PH, no significativo
high_pH<-quantile(redwine$pH, probs =0.75)
redwine.altopH<-redwine[redwine$pH>=high_pH,]$quality
redwine.bajopH<-redwine[redwine$pH<high_pH,]$quality
t.test(redwine.altopH, redwine$quality, alternative = "greater") # NO significativo  PH
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altopH and redwine$quality
## t = -1.2978, df = 654.48, p-value = 0.9026
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.1320493        Inf
## sample estimates:
## mean of x mean of y
##  5.577830  5.636023
```

```r
t.test(redwine.bajopH, redwine$quality, alternative = "less") # NO significativo  PH
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.bajopH and redwine$quality
## t = 0.68001, df = 2541.6, p-value = 0.7517
```

```
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.07181025
## sample estimates:
## mean of x mean of y
##  5.657021  5.636023
```

El p-valor no es menor que el nivel de significación por lo que no podemos rechazar la hipótesis nula y por tanto que la variable pH no es significativa.

```
# Citric acid es significativo
high_citric_acid<-quantile(redwine$citric.acid, probs =0.75)
redwine.altoCitricAcid<-redwine[redwine$citric.acid>=high_citric_acid,]$quality
redwine.bajoCitricAcid<-redwine[redwine$citric.acid<high_citric_acid,]$quality
t.test(redwine.altoCitricAcid, redwine$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altoCitricAcid and redwine$quality
## t = 5.371, df = 644.36, p-value = 5.475e-08
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1711653       Inf
## sample estimates:
## mean of x mean of y
##  5.882904  5.636023
```

```
t.test(redwine.bajoCitricAcid, redwine$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.bajoCitricAcid and redwine$quality
## t = -2.973, df = 2585.7, p-value = 0.001488
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.04016442
## sample estimates:
## mean of x mean of y
##  5.546075  5.636023
```

El p-valor es menor que el nivel de significación por lo que rechazamos la hipótesis nula y podemos decir que la variable citric.acid es significativa.

```
# fixed acidity es significativo
high_fixed_acidity<-quantile(redwine$fixed.acidity, probs =0.75)
redwine.altoFixedAcidity<-redwine[redwine$fixed.acidity>=high_fixed_acidity,]$quality
redwine.bajoFixedAcidity<-redwine[redwine$fixed.acidity<high_fixed_acidity,]$quality
t.test(redwine.altoFixedAcidity, redwine$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altoFixedAcidity and redwine$quality
## t = 3.8769, df = 618.02, p-value = 5.857e-05
## alternative hypothesis: true difference in means is greater than 0
```

```
## 95 percent confidence interval:
##  0.1032353       Inf
## sample estimates:
## mean of x mean of y
##  5.815534  5.636023
```

```
t.test(redwine.bajoFixedAcidity, redwine$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.bajoFixedAcidity and redwine$quality
## t = -2.0469, df = 2594, p-value = 0.02039
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01221934
## sample estimates:
## mean of x mean of y
##  5.573715  5.636023
```

El `p-valor` es menor que el nivel de significación por lo que rechazamos la hipótesis nula y podemos decir que la variable `fixed.acidity` es significativa.

```
#Chlorides, significativo, a más chlorides peor calidad
high_chlorides<-quantile(redwine$chlorides, probs =0.75)
redwine.altoChlorides<-redwine[redwine$chlorides>=high_chlorides,]$quality
redwine.bajoChlorides<-redwine[redwine$chlorides<high_chlorides,]$quality
t.test(redwine.altoChlorides, redwine$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altoChlorides and redwine$quality
## t = -2.9858, df = 674.92, p-value = 0.001466
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.05722153
## sample estimates:
## mean of x mean of y
##  5.508393  5.636023
```

```
t.test(redwine.bajoChlorides, redwine$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.bajoChlorides and redwine$quality
## t = 1.4446, df = 2529.7, p-value = 0.07434
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.006259493          Inf
## sample estimates:
## mean of x mean of y
##  5.681049  5.636023
```

El `p-valor` es menor que el nivel de significación por lo que rechazamos la hipótesis nula y podemos decir que la variable `chlorides` es significativa.

```r
#free sulfur dioxide, no significativo
high_free_sulfur<-quantile(redwine$free.sulfur.dioxide, probs =0.75)
redwine.altoFreeSulfur<-redwine[redwine$free.sulfur.dioxide>=high_free_sulfur,]$quality
redwine.bajoFreeSulfur<-redwine[redwine$free.sulfur.dioxide<high_free_sulfur,]$quality
t.test(redwine.altoFreeSulfur, redwine$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altoFreeSulfur and redwine$quality
## t = -1.5957, df = 758.2, p-value = 0.9445
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.130598       Inf
## sample estimates:
## mean of x mean of y
##  5.571754  5.636023
```

```r
t.test(redwine.bajoFreeSulfur, redwine$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.bajoFreeSulfur and redwine$quality
## t = 0.76627, df = 2449.8, p-value = 0.7782
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.07655152
## sample estimates:
## mean of x mean of y
##  5.660345  5.636023
```

El `p-valor` no es menor que el nivel de significación por lo que no podemos rechazar la hipótesis nula y por tanto que la variable `free.sulfur.dioxide` no es significativa.

```r
# total sulfur dioxide,  significativo, a más cantidad de  total sulfur peor calidad
high_total_sulfur<-quantile(redwine$total.sulfur.dioxide, probs =0.75)
redwine.altoTotalSulfur<-redwine[redwine$total.sulfur.dioxide>=high_total_sulfur,]$quality
redwine.bajoTotalSulfur<-redwine[redwine$total.sulfur.dioxide<high_total_sulfur,]$quality
t.test(redwine.altoTotalSulfur, redwine$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altoTotalSulfur and redwine$quality
## t = -6.677, df = 752.79, p-value = 2.368e-11
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1922452
## sample estimates:
## mean of x mean of y
##  5.380835  5.636023
```

```r
t.test(redwine.bajoTotalSulfur, redwine$quality, alternative = "greater")
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  redwine.bajoTotalSulfur and redwine$quality
## t = 2.7632, df = 2516.4, p-value = 0.002883
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.03524473        Inf
## sample estimates:
## mean of x mean of y
##  5.723154  5.636023
```

El p-valor es menor que el nivel de significación por lo que rechazamos la hipótesis nula y podemos decir que la variable total.sulfur.dioxide es significativa.

```
# Density, significativo a mayor densidad peor calidad
high_density<-quantile(redwine$density, probs =0.75)
redwine.altoDensity<-redwine[redwine$density>=high_density,]$quality
redwine.bajoDensity<-redwine[redwine$density<high_density,]$quality
t.test(redwine.altoDensity, redwine$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.altoDensity and redwine$quality
## t = -2.0039, df = 645.06, p-value = 0.02275
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01531291
## sample estimates:
## mean of x mean of y
##  5.550000  5.636023
```

```
t.test(redwine.bajoDensity, redwine$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  redwine.bajoDensity and redwine$quality
## t = 0.92086, df = 2556.1, p-value = 0.1786
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.0225815        Inf
## sample estimates:
## mean of x mean of y
##  5.664721  5.636023
```

El p-valor es menor que el nivel de significación por lo que rechazamos la hipótesis nula y podemos decir que la variable density es significativa. Sólo se acepta una de las colas, la cola derecha.

Finalmente realizamos una recopilación de las variables significativas que serán: alcohol , volatile.acidity , sulphates , citric.acid , fixed.acidity , chlorides , total.sulfur.dioxide , density

4.2 Comprobación de la normalidad y homogeneidad de la varianza

Comprobar que variables siguen distribución normal

Aplicando la siguiente función podemos comprobar si siguen una distribución normal o si por el contrario no lo hacen.

```
#install.packages("nortest")
library(nortest)
```

```
## Warning: package 'nortest' was built under R version 3.5.2
```

```
alpha = 0.05
col.names = colnames(redwine)
for (i in 1:ncol(redwine)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(redwine[,i]) | is.numeric(redwine[,i])) {
      p_val = ad.test(redwine[,i])$p.value
        if (p_val < alpha) {
            cat(col.names[i])
                  # Format output
            if (i < ncol(redwine) - 1) cat(", ")
            if (i %% 3 == 0) cat("\n")
        }
      }
  }
```

```
## Variables que no siguen una distribución normal:
## fixed.acidity, volatile.acidity, citric.acid,
## residual.sugar, chlorides, free.sulfur.dioxide,
## total.sulfur.dioxide, density, pH,
## sulphates, alcoholquality
```

De esta manera podemos afirmar que las variables no siguen una distrubución normal.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

En primer lugar, procedemos a realizar un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre la calidad del vino:

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
# Calcular el coeficiente de correlación para cada variable cuantitativa # con respecto al campo "preci
for (i in 1:(ncol(redwine) - 1)) {
if (is.integer(redwine[,i]) | is.numeric(redwine[,i]))
    {
    spearman_test = cor.test(redwine[,i],redwine[,length(redwine)],method = "spearman")
      corr_coef = spearman_test$estimate
      p_val = spearman_test$p.value
      # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(redwine)[i]
  }
}
```

```
## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(redwine[, i], redwine[, length(redwine)],
## method = "spearman"): Cannot compute exact p-value with ties
```

```r
a <- corr_matrix[,'p-value']
corr_matrix[order(a),]
```

```
##                         estimate      p-value
## alcohol               0.47853169 2.726838e-92
## volatile.acidity     -0.38064651 2.734944e-56
## sulphates             0.37706020 3.477695e-55
## citric.acid           0.21348091 6.158952e-18
## total.sulfur.dioxide -0.19673508 2.046488e-15
## chlorides            -0.18992234 1.882858e-14
## density              -0.17707407 9.918139e-13
## fixed.acidity         0.11408367 4.801220e-06
## free.sulfur.dioxide  -0.05690065 2.288322e-02
## pH                   -0.04367193 8.084594e-02
## residual.sugar        0.03204817 2.002454e-01
```

Así, identificamos cuáles son las variables más correlacionadas con la calidad según su proximidad con los valores -1 y +1.

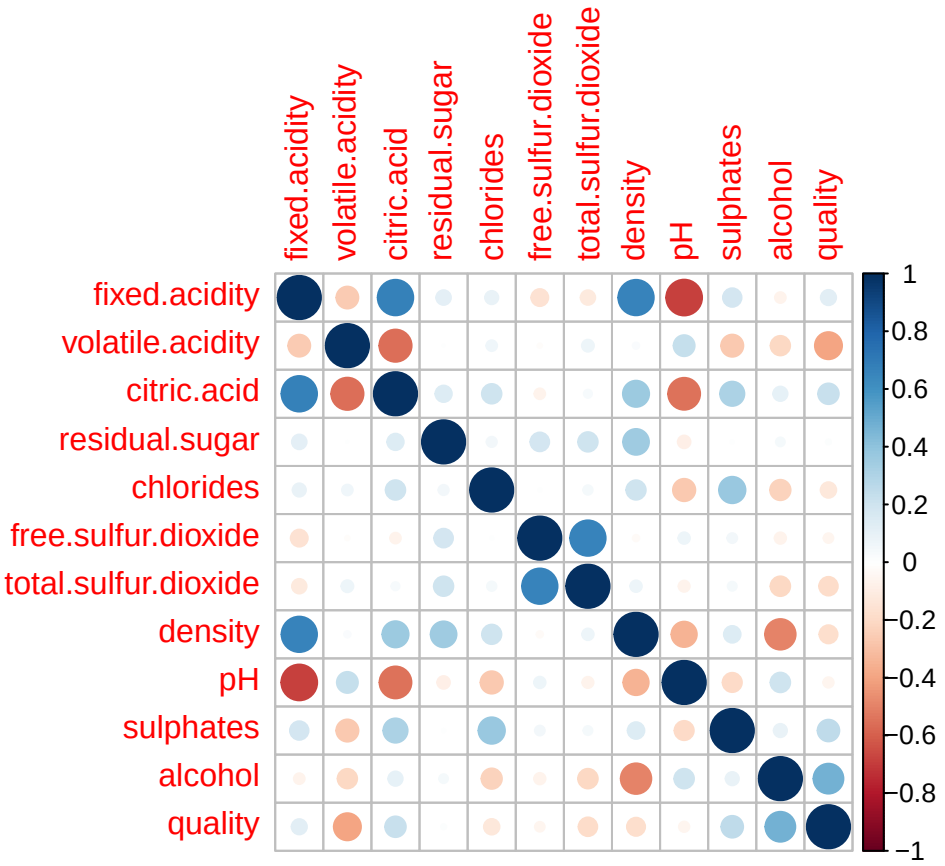
Teniendo esto en cuenta, queda patente cómo la variable más relevante para **la calidad** es la variable `alcohol`. Pero en términos generales podemos decir que los valores que obtenemos son bastante modestos y no sería adecuado obtener ninguna conclusión por ahora. Lo único que podemos hacer es utilizar estos valores como tendencias.

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

```
## Warning: package 'corrplot' was built under R version 3.5.2
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor(redwine))
```



Nota: Para cada coeficiente de correlación se muestra también su p-valor asociado, puesto que éste puede dar información acerca del peso estadístico de la correlación obtenida.

Modelo de regresión lineal logística

Para realizar predicciones en la calidad del vino vamos a plantear un modelo de regresión logística en dónde solamente tendremos regresores cuantitativos. La variable a predecir será una variable dicotómica transformada previamente (`good_wine`esta tomará valores de calidad *0* si es menor que 7 ó *1* en caso contrario).

```
good_wine <-ifelse(test=redwine$quality>=7,yes=1,no=0)
redwine$good_wine=good_wine
quality<- redwine$quality
```

Se usa regresión logística ya que nuestro modelo predicirá si el vino es de calidad o no (variable dicotómica).

Para obtener un modelo de regresión logística considerablemente eficiente, lo que haremos será obtener varios modelos de regresión.

En primer lugar partiremos del modelo que utiliza las variables que hemos detectado como significativas según el test t para un valor significación=0.05.

```
GLM.1 <- glm( redwine$good_wine ~ alcohol + volatile.acidity + sulphates + citric.acid + fixed.acidity +

summary(GLM.1)$aic
```

```
## [1] 898.8508
```

En segundo lugar realizamos una comparación con la selección de todas las variables de nuestro dataset.

```
GLM.2 <- glm( redwine$good_wine ~ . -quality , family=binomial(logit),data=redwine)
summary(GLM.2)$aic
```

```
## [1] 894.8644
```

Finalmente, vamos a realizar una última prueba con la función `step` que os proporcionará la mejor combinación de atributos para la obtención de un mejor modelo.

```
info = step(object = GLM.2, direction = "both", trace = 1)
```

```
## Start:  AIC=894.86
## redwine$good_wine ~ (fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + quality) - quality
##
##                        Df Deviance    AIC
## - pH                    1   870.91 892.91
## - citric.acid           1   871.32 893.32
## - free.sulfur.dioxide   1   871.64 893.64
## <none>                      870.86 894.86
## - fixed.acidity         1   875.67 897.67
## - density               1   876.34 898.34
## - residual.sugar        1   880.02 902.02
## - chlorides             1   880.85 902.85
## - volatile.acidity      1   882.52 904.52
## - total.sulfur.dioxide  1   884.49 906.49
## - alcohol               1   904.51 926.51
## - sulphates             1   915.26 937.26
##
## Step:  AIC=892.91
## redwine$good_wine ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + sulphates + alcohol
##
##                        Df Deviance    AIC
## - citric.acid           1   871.33 891.33
## - free.sulfur.dioxide   1   871.78 891.78
## <none>                      870.91 892.91
## + pH                    1   870.86 894.86
## - density               1   877.94 897.94
## - fixed.acidity         1   878.81 898.81
## - residual.sugar        1   880.40 900.40
## - chlorides             1   881.53 901.53
## - volatile.acidity      1   882.72 902.72
## - total.sulfur.dioxide  1   885.36 905.36
## - sulphates             1   915.50 935.50
```

```
## - alcohol               1   916.76 936.76
##
## Step:  AIC=891.33
## redwine$good_wine ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + sulphates + alcohol
##
##                        Df Deviance    AIC
## - free.sulfur.dioxide   1   872.08 890.08
## <none>                      871.33 891.33
## + citric.acid           1   870.91 892.91
## + pH                    1   871.32 893.32
## - density               1   878.15 896.15
## - residual.sugar        1   881.27 899.27
## - chlorides             1   881.76 899.76
## - fixed.acidity         1   883.87 901.87
## - total.sulfur.dioxide  1   885.36 903.36
## - volatile.acidity      1   892.88 910.88
## - sulphates             1   915.82 933.82
## - alcohol               1   921.78 939.78
##
## Step:  AIC=890.08
## redwine$good_wine ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + total.sulfur.dioxide + density + sulphates +
##     alcohol
##
##                        Df Deviance    AIC
## <none>                      872.08 890.08
## + free.sulfur.dioxide   1   871.33 891.33
## + citric.acid           1   871.78 891.78
## + pH                    1   872.01 892.01
## - density               1   878.99 894.99
## - residual.sugar        1   881.60 897.60
## - chlorides             1   882.47 898.47
## - fixed.acidity         1   884.45 900.45
## - total.sulfur.dioxide  1   890.34 906.34
## - volatile.acidity      1   894.16 910.16
## - sulphates             1   917.01 933.01
## - alcohol               1   922.50 938.50
```

```
GLM.Mejor = glm(formula = redwine$good_wine ~ fixed.acidity + volatile.acidity +
    residual.sugar + chlorides + total.sulfur.dioxide + density +
    sulphates + alcohol, family = binomial(logit), data = redwine)

summary(GLM.Mejor)$aic
```

```
## [1] 890.076
```

Como se puede observar, el mejor modelo es el último modelo ya que proporciona un mayor valor de `AIC` y la función `step` en sí te lo proporciona. Sin embargo, creemos que para nuevas entradas de datos, este modelo proporcionado por la función `step` está sobreajustado al conjunto de datos y no será muy bueno para nuevos conjuntos de datos. Por este motivo, vamos a seleccionar el modelo *GLM.1* que parte de las variables que hemos estudiado como significativas.

```
redwine$prob_qualityM=predict(GLM.1, redwine, type="response")
newdatarisk=subset(redwine, prob_qualityM>0.7)
Q3 <-quantile(redwine$alcohol)[4]
alcohol <- which(newdatarisk$alcohol>Q3)
```

5. Representación de los resultados a partir de tablas y gráficas.

Calculo curva ROC

El cálculo de la curva ROC será interesante para poder evaluar cómo de bueno es nuestro modelo. Esta medida está comprendida en tanto por 1 de manera que cuanto mayor sea mejor será la predicción del modelo.

Es interesante utilizar este tipo de métrica ya que podemos detectar cuando, para clases desbalanceadas, el modelo arrastra los datos a la clase mayoritaria y obtiene una tasa de acierto bastante alta como consecuencia del desbalanceo pero una curva ROC muy mala ya que para la otra clase no tiene ningún acierto.

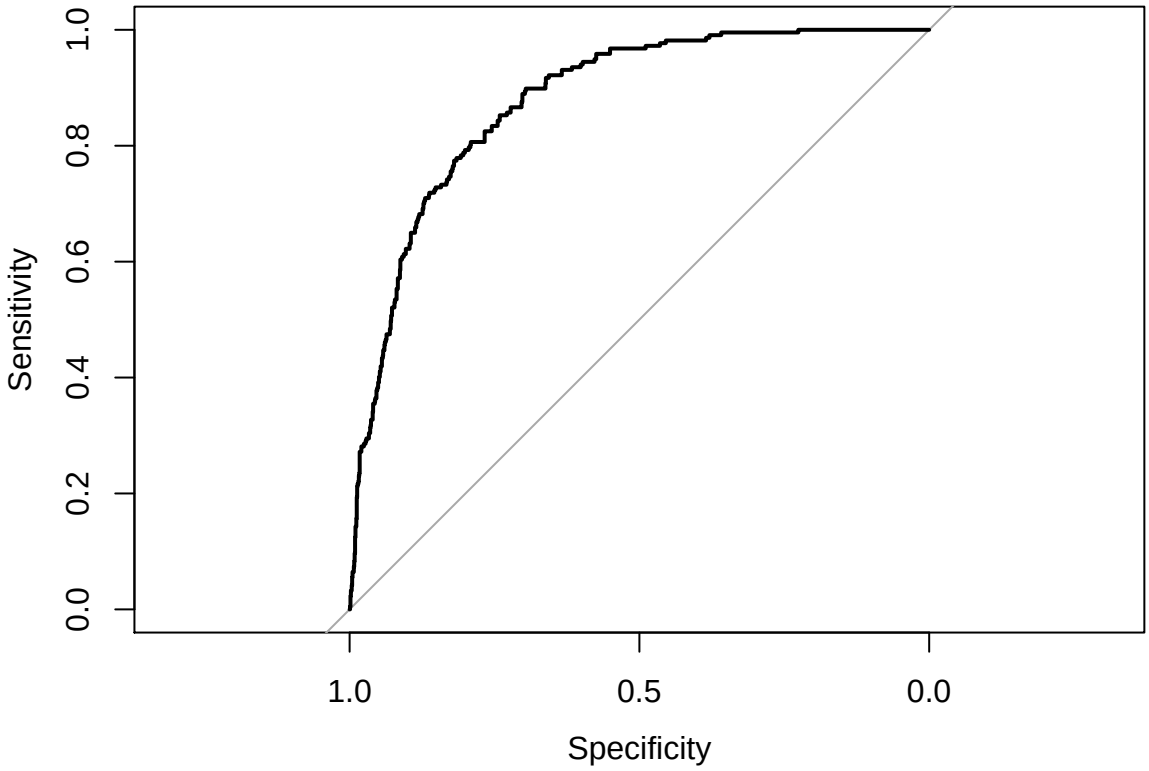
```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
g=roc(redwine$good_wine,redwine$prob_qualityM, data=redwine)
plot(g)
```




```
auc(g)
```

```
## Area under the curve: 0.8779
```

Como vemos, obtenemos un área bajo la curva de 0.8779 lo cual supone un modelo bastante bueno.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Ahora podemos responder a nuestra pregunta del primer apartado *¿Qué componentes fisico-químicos influyen en que un vino sea bueno?* diciendo que los componentes más influyentes son: *alcohol + volatile.acidity + sulphates + citric.acid + fixed.acidity + chlorides + total.sulfur.dioxide + density.*

Todos ellos en conjunto permiten montar un modelo capaz de predecir la calidad de un vino.