# Respuestas de la Práctica 2

## 1. Descripción del dataset

El conjunto de datos escogido es el de los datos sobre la calidad del vino tinto (Red wine quality) de Kaggle.

Esta constituido de 12 características (columnas) que presentan 1599 vinos diferentes (filas).

Entre los campos de este conjunto de datos, encontramos los siguientes:

1. fixed acidity: Cantidad de acido del vino.
2. volatile acidity: Acido acético, niveles elevados pueden resultar en un sabor avinagrado.
3. citric acid: Se suele encontrar en pequeñas dosis, da frescura al sabor del vino.
4. residual sugar: La cantidad de azúcar que queda después de la fermentación. Es raro encontrar vinos con menos de 1 gramo por litro o más de 45 gramos por litro.
5. Chlorides: Cantidad de sal en el vino.
6. free sulfur dioxide: La existencia de este previene el crecimiento de microbios y la oxidación del vino.
7. total sulfur dioxide:
8. density: La densidad depende del porcentaje de alcohol y azúcar del vino.
9. PH: Describe la acidez del vino ( 0 muy acido, 14 muy básico). En general los vinos tienen un ph de 3-4.
10. sulphates: Aditivo que contribuye a los niveles de dióxido de sulfuro. Es anti-microbios y antioxidante.
11. Alcohol: Porcentaje de alcohol.
12. Quality: Puntuación de 0 a 10.

## 2. Objetivos del análisis

A partir de este conjunto de datos se plantea la problemática de determinar qué variables influyen más sobre la calidad (quality) del vino tinto.

También se podrá proceder a crear modelos de regresión que permitan predecir la calidad de un vino en concreto en función de sus características.

Este análisis es importante para el sector vinícola para categorizar si un vino es de alta calidad o no.

## 3. Limpieza de datos

Antes de empezar con la limpieza de datos, vamos a realizar la lectura del fichero en formato CSV mediante la llamada read.csv(), esta nos devuelve un objeto data.frame con los datos:

```
# ABRIR FICHERO CON LOS DATOS RED WINE
wine_data <- read.csv("data/winequalityred.csv", header=TRUE, sep=",")
```

Veamos las primeras líneas de los datos:

```
> head(wine_data[,1:12])
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol quality
1           7.4             0.70        0.00            1.9     0.076                 11                  34  0.9978 3.51      0.56     9.4       5
2           7.8             0.88        0.00            2.6     0.098                 25                  67  0.9968 3.20      0.68     9.8       5
3           7.8             0.76        0.04            2.3     0.092                 15                  54  0.9970 3.26      0.65     9.8       5
4          11.2             0.28        0.56            1.9     0.075                 17                  60  0.9980 3.16      0.58     9.8       6
5           7.4             0.70        0.00            1.9     0.076                 11                  34  0.9978 3.51      0.56     9.4       5
6           7.4             0.66        0.00            1.8     0.075                 13                  40  0.9978 3.51      0.56     9.4       5
```

*(imagen en el fichero images/head.png)*

El tipo de dato de cada campo:

```
> sapply(wine_data, function(x) class(x))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide total.sulfur.dioxide              density                   pH            sulphates              alcohol              quality
           "numeric"            "numeric"            "numeric"            "numeric"            "numeric"            "numeric"            "numeric"            "numeric"            "numeric"            "numeric"            "numeric"            "integer"
```

*(imagen en el fichero images/types.png)*

Vemos que todos los campos son de tipo numérico.

Y un resumen de las dimensiones de cada campo (min, max, mean, median, etc):

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

*(imagen en el fichero images/summary.png)*

Aquí podemos hacernos a la idea de las dimensiones de los valores de cada campo. Por ejemplo, vemos que la calidad (quality) de los vinos van de 3 a 8, o que el indice de alcohol toma valores desde 8.4 a 14.9 grados.

### 3.1 Selección de los datos de interés

Todos los atributos del conjunto de datos corresponden a características químicas de los vinos y se considera interesante tenerlos todos en consideración.

Si tuviéramos campos como marca o precio en el conjunto de datos, se eliminarían ya que no son atributos que deberían influir en la calidad del vino.

## 3.2 Ceros y elementos vacíos

Según el origen de datos original (link), no existen valores vacíos en el dataset: *"8. Missing Attribute Values: None".*

De todos modos vamos a comprobarlo:

```
> sapply(wine_data, function(x) sum(is.na(x)))
      fixed.acidity    volatile.acidity         citric.acid      residual.sugar           chlorides  free.sulfur.dioxide total.sulfur.dioxide 
                  0                   0                   0                   0                   0                   0                   0 
            density                  pH           sulphates             alcohol             quality 
                  0                   0                   0                   0                   0 
```

*(imagen en el fichero images/na.png)*

Como vemos no existe ningún elemento vacío en el conjunto de datos.

En el caso de tener elementos vacíos tendríamos de decidir que hacer con las filas que los contienen. En ese caso tenemos varias opciones:

- **Eliminar** la fila en cuestión, pero esto también supone perder información del conjunto de datos que podría ser relevante.
- **Rellenar** ese elemento vacío por algún valor por defecto. En este caso asignarle el valor cero a un elemento vacío podría alterar los resultados, así que optaría por asignarle un valor medio de ese campo.

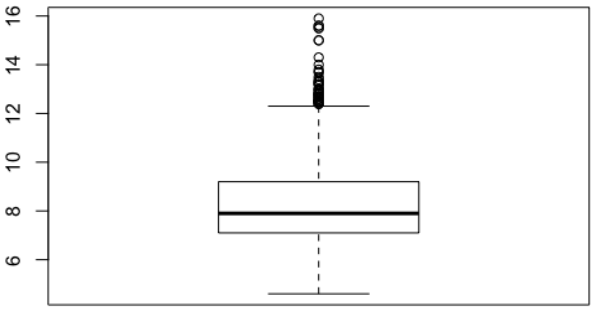
## 3.3 Valores extremos

Los valores extremos (*outliners*) son aquellos que parecen no ser congruentes si los comparamos con el resto de los datos. Para identificarlos, podemos hacer uso de:

- Diagrama de caja para cada variable y ver qué valores distan mucho del rango intercuartílico.les.
- Utilizar la función boxplots.stats() que muestra los valores atípicos en un listado.

Vamos a ver uno a uno:

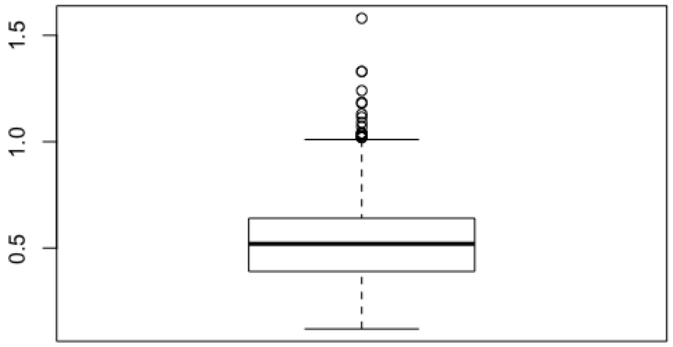
Fixed acidity:



```
> boxplot.stats(wine_data$fixed.acidity)$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 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 12.5 12.9 14.3 12.4 15.5 15.5
[35] 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2 13.2 13.2 15.9 13.3 12.9 12.6 12.6
```

*(imagen en el fichero images/fixedacidity_plot_out.png)*

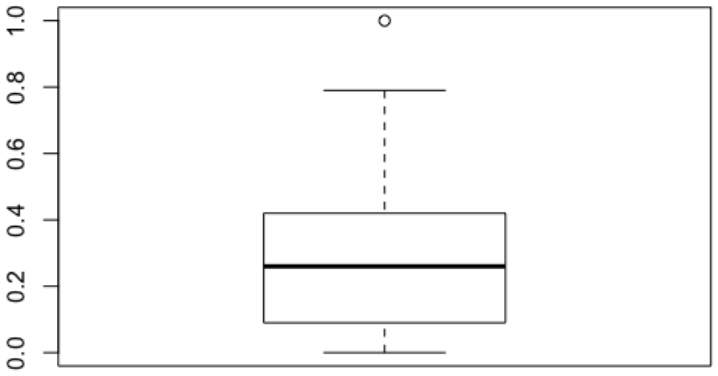
## Volatile acidity:




```
> boxplot.stats(wine_data$volatile.acidity)$out
 [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

*(imagen en el fichero images/volatileacidity_plot_out.png)*

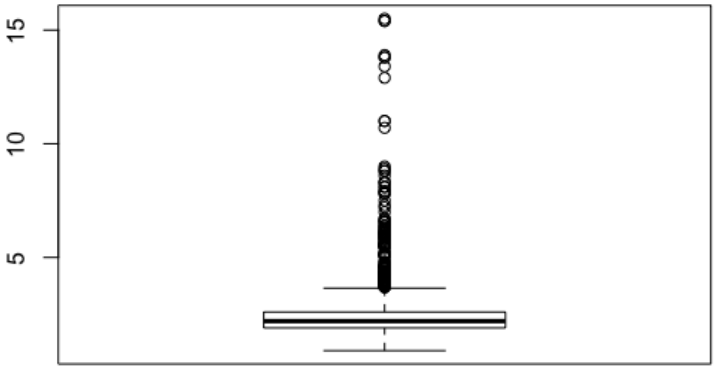
## Citric acid:




```
> boxplot.stats(wine_data$citric.acid)$out
[1] 1
```

*(imagen en el fichero images/citricacid_plot_out.png)*

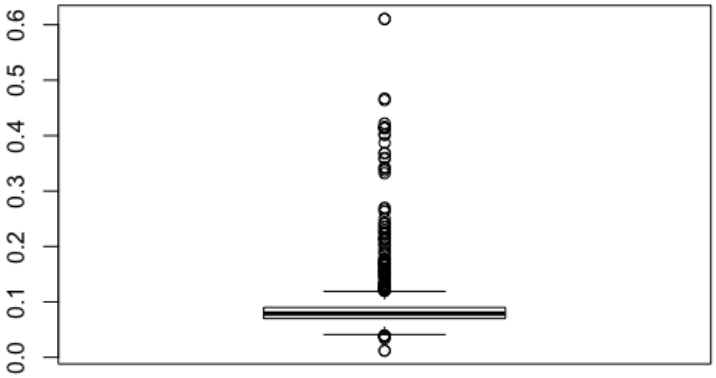
## Residual sugar:




```
> boxplot.stats(wine_data$residual.sugar)$out
  [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00  4.00  4.00  4.00  7.00  4.00  4.00
 [29]  6.40  5.60  5.60 11.00 11.00  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60  6.10
 [57]  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40  4.25  6.00  3.90  4.20  4.00  4.00  4.00  6.60
 [85]  6.00  6.00  3.80  9.00  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50  4.30  5.50
[113]  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90  4.30 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80
[141]  5.20  5.20  3.75 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10  7.80
```

*(imagen en el fichero images/residualsugar_plot_out.png)*

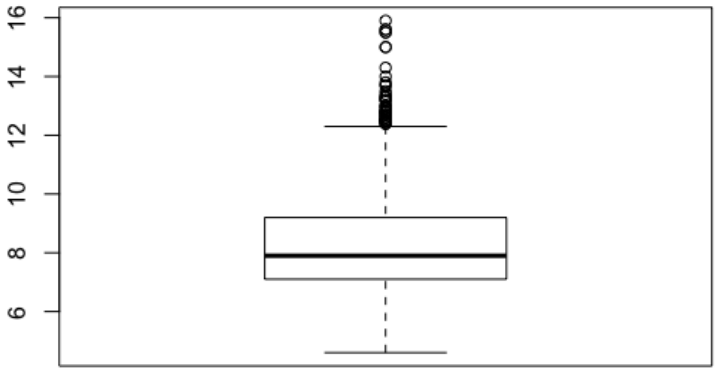
## Chlorides:




```
> boxplot.stats(wine_data$chlorides)$out
  [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178 0.146 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343 0.186 0.213 0.214 0.121 0.122 0.122
 [29] 0.128 0.120 0.159 0.124 0.122 0.122 0.174 0.121 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171 0.226 0.226 0.250 0.148 0.122 0.124 0.124 0.143 0.222 0.039 0.157 0.422
 [57] 0.034 0.387 0.415 0.157 0.157 0.243 0.241 0.190 0.132 0.126 0.038 0.165 0.145 0.147 0.012 0.012 0.039 0.194 0.132 0.161 0.120 0.120 0.123 0.123 0.414 0.216 0.171 0.178
 [85] 0.369 0.166 0.166 0.136 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414 0.166 0.168 0.415 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205 0.039 0.235 0.230 0.038
```

*(imagen en el fichero images/chlorides_plot_out.png)*

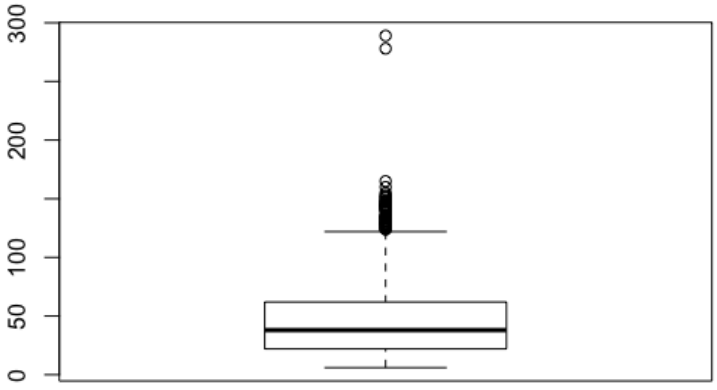
## Free sulfur dioxide:




```
> boxplot.stats(wine_data$free.sulfur.dioxide)$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 51 52 55 55 48 48 66
```

*(imagen en el fichero images/freesulfurdioxide_plot_out.png)*

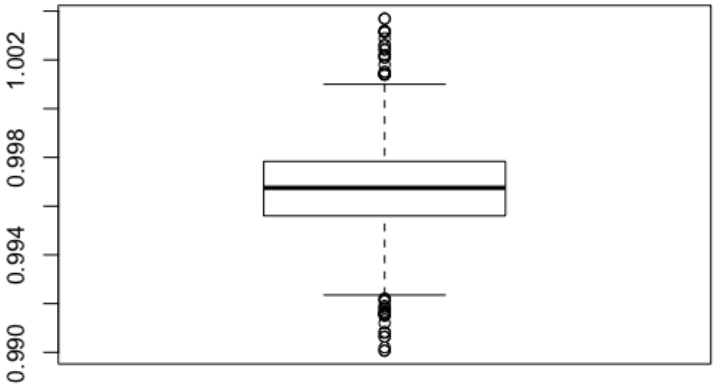
## Total sulfur dioxide:




```
> boxplot.stats(wine_data$total.sulfur.dioxide)$out
  [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127 126 145 144 135 165 124 124 134 124 129 151 133 142 149 147 145 148 155 151 152 125 127 139 143 144 130
 [44] 278 289 135 160 141 141 133 147 147 131 131 131
```

*(imagen en el fichero images/totalsulfurdioxide_plot_out.png)*

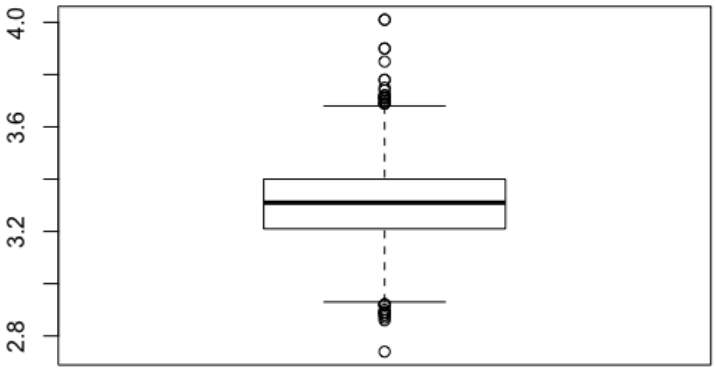
## Density:




```
> boxplot.stats(wine_data$density)$out
 [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220 1.00220 1.00140 1.00140 1.00140 1.00140 1.00320 1.00260 1.00140 1.00315 1.00315 1.00315 1.00210 1.00210
[22] 0.99170 0.99220 1.00260 0.99210 0.99154 0.99064 0.99064 1.00289 0.99162 0.99007 0.99007 0.99020 0.99220 0.99150 0.99157 0.99080 0.99084 0.99191 1.00369 1.00369 1.00242
[43] 0.99182 1.00242 0.99182
```

*(imagen en fichero images/density_plot_out.png)*

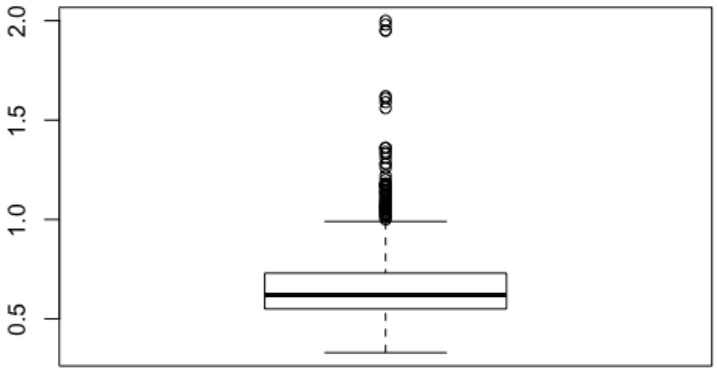
## pH:




```
> boxplot.stats(wine_data$pH)$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 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 4.01 2.90 4.01 3.71 2.88 3.72
[35] 3.72
```

*(imagen en el fichero images/ph_plot_out.png)*

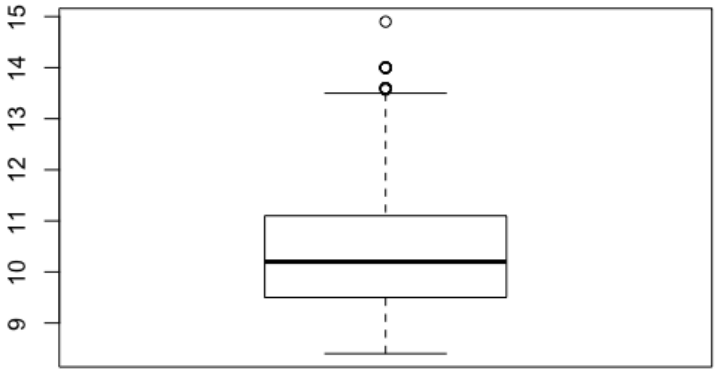
## Sulphates:




```
> boxplot.stats(wine_data$sulphates)$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 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 1.07 1.06 1.06 1.05 1.06 1.04
[35] 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17 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 1.17 1.10 1.01
```

*(imagen en el fichero images/sulphates_plot_out.png)*

Alcohol:




```
> boxplot.stats(wine_data$alcohol)$out
 [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000 13.60000 14.00000 14.00000 13.56667 13.60000
```

*(imagen en el fichero images/alcohol_plot_out.png)*

Después de analizar cada campo, vemos que en algunos casos (como chlorides o density) aparecen bastantes valores outliners, pero mirando al conjunto de datos, no parecen erróneos.

En el resto de campos, e investigando un poco en el sector vinícola, no parecen valores suficientemente dispersos como para considerarlos outliners. Por ejemplo en el caso de alcohol, considera outliners valores superiores a 13.5º, pero 15º no parece un valor extraño.

En cambio, en el caso de la característica total sulfur dioxide, hay dos valores que aparecen muy alejados del resto: 278 y 289 cuando el valor del tercer cuartil es de 62.

En el caso de residual sugar, aunque aparentemente parezca un outliner, sabemos que esa característica puede llegar al valor 45 en el caso de vinos dulces, así que no se considera un valor extremo.

Vamos a eliminar todas esas filas con total sulful dioxide superior a 200:

```
#eliminar outliners:
wine_data<-wine_data[!(wine_data$total.sulfur.dioxide>200),]
```

Si vemos ahora los valores de este campo, vemos que el máximo es de 165:

```
> summary(wine_data$total.sulfur.dioxide)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   22.00   38.00   46.17   62.00  165.00
```

Y que se han eliminado 2 filas:

```
> dim(wine_data)
[1] 1597   12
```

### 3.4 Exportación de los datos a analizar

Ahora ya podemos proceder a guardar los los datos en un nuevo fichero:

```
# Guardar en csv:
write.csv(wine_data, file = "data/winedata_output.csv")
```

*(fichero data/winedata_output.csv)*

## 4. Análisis de los datos

### 4.1 Selección de los grupos de datos a analizar

A continuación se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar. Cómo se trata de valores numéricos y no categóricos, para cada uno de ellos se va a definir rangos.

- **Agrupación según pH**: ¿Influye el pH en la calidad? Sabemos que los vinos normalmente toman valores de pH de entre 3 y 4. y que el valor medio en el conjunto de datos es de 3.3, vamos a definir dos grupos:
  - <3.5 → pH bajo.
  - >= 3.5 → pH alto.
- **Agrupación según citric acid**: ¿Influye el sabor cítrico en la calidad del vino?

- < 0.5 citric acid bajo.
- >= 0.5 → citric acid alto

- **Agrupación según residual sugar:** ¿Influye el azúcar en la calidad? Se crean la siguientes agrupaciones:
  - <= 2.5 → azúcar residual bajo.
  - > 2.5→ azúcar residual alto.

El código R para las agrupaciones:

```
#grupos de datos:
wine_data.high_ph<-subset(wine_data, pH>=3.5)
wine_data.low_ph<-subset(wine_data, pH<3.5)
wine_data.high_citric<-subset(wine_data, citric.acid>=0.5)
wine_data.low_citric<-subset(wine_data, citric.acid<0.5)
wine_data.high_sugar<-subset(wine_data, residual.sugar>2.5)
wine_data.low_sugar<-subset(wine_data, residual.sugar<=2.5)
```

## 4.2 Comprobación normalidad y homogeneidad de la varianza

Para comprobar que los valores de nuestras variables siguen una distribución normal, se puede utilizar la prueba de normalidad de Anderson-Darling. Esto comprueba que para cada prueba se obtiene un p-valor superior a 0.05, si se cumple se considera que la variable sigue una distribución normal.

```
library(nortest)
alpha= 0.05
col.names = colnames(wine_data)
for(i in 1:ncol(wine_data)){
  if (i == 1) cat("Variables que no siguen distribución normal:\n")
  if (is.integer(wine_data[,i]) | is.numeric(wine_data[,i])){
    p_val = ad.test(wine_data[,i])$p.value
    if (p_val < alpha){
      cat(paste(col.names[i],"\n"))
    }
  }
}
```

Obtenemos el siguiente resultado:

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

Para ver la homogeneidad de la varianza lo haremos mediante la aplicación del test de Fligner-Killeen.

Vamos a comprobarlo por las 3 características que hemos escogido para las agrupaciones, pero antes, se añade una columna al dataset con la clasificación según la agrupación, asignando valores 0 o 1 según si tiene valor bajo o alto:

```
#añadir columnas según agrupación
wine_data$ph.class = ifelse (wine_data$pH<3.5,0,1)
wine_data$citric.class = ifelse (wine_data$citric.acid<0.5,0,1)
wine_data$sugar.class = ifelse (wine_data$residual.sugar<=2.5,0,1)
```

Utilizaremos estos campos para el test:

```
> fligner.test(quality ~ ph.class, data=wine_data)

        Fligner-Killeen test of homogeneity of variances

data:  quality by ph.class
Fligner-Killeen:med chi-squared = 6.605, df = 1, p-value = 0.01017
```

En el caso del pH, vemos que p-value es inferior a 0.05, por lo que no se cumple la hipótesis de que las varianzas de las muestras son homogéneas.

```
> fligner.test(quality ~ citric.class, data=wine_data)

        Fligner-Killeen test of homogeneity of variances

data:  quality by citric.class
Fligner-Killeen:med chi-squared = 0.071702, df = 1, p-value = 0.7889
```

```
> fligner.test(quality ~ sugar.class, data=wine_data)

        Fligner-Killeen test of homogeneity of variances

data:  quality by sugar.class
Fligner-Killeen:med chi-squared = 2.8749, df = 1, p-value = 0.08997
```

En el caso de citric.class y sugar.class, obtenemos valores de p-value superiores a 0,05 por lo que se cumple la hipótesis de que las varianzas son homogéneas.

## 4.3 Pruebas estadísticas

### 4.3.1. Variables que influyen más en la calidad del vino

Vamos a realizar un análisis de correlación entre las distintas variables para determinar cuáles influyen más sobre la calidad del vino utilizando el coeficiente de correlación de Spearman.

```
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 "quality"
for (i in 1:(ncol(wine_data) - 1)) {
  if (is.integer(wine_data[,i]) | is.numeric(wine_data[,i])) {
    spearman_test = cor.test(wine_data[,i],
                             wine_data[,length(wine_data)],
                             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(wine_data)[i]
  }
}
```

```
> print(corr_matrix)
                       estimate      p-value
fixed.acidity         0.11422748 4.734418e-06
volatile.acidity     -0.37877752 1.207062e-55
citric.acid           0.21076778 1.709573e-17
residual.sugar        0.02871777 2.513934e-01
chlorides            -0.18725332 4.555188e-14
free.sulfur.dioxide  -0.06024822 1.604147e-02
total.sulfur.dioxide -0.20084662 5.373192e-16
density              -0.17427344 2.338051e-12
pH                   -0.04047972 1.058639e-01
sulphates             0.38064590 3.201752e-56
alcohol               0.47685436 1.859983e-91
```

Para identificar cuáles son las variables más correlacionadas con el precio hay que mirar que su valor este próximo a -1 y 1. En este caso la variable más relevante es alcohol, seguida de sulphates.

## 4.4 ¿Residual sugar influye en la calidad del vino?

Vamos a contrastar dos muestras para determinar si la calidad es superior dependiendo de si el vino tiene la característica “residual sugar” alta o baja.

Para ello vamos a tener dos muestras: la primera con las calidades de esos vinos con bajo azúcar residual y la segunda con las calidades de los vinos con niveles altos.

```
#subset de la calidad segun residual.sugar
wine_data.high_sugar.quality <- wine_data[wine_data$residual.sugar>2.5,]$quality
wine_data.low_sugar.quality <- wine_data[wine_data$residual.sugar<=2.5,]$quality
```

```
> t.test(wine_data.low_sugar.quality, wine_data.high_sugar.quality, alternative="less")

        Welch Two Sample t-test

data:  wine_data.low_sugar.quality and wine_data.high_sugar.quality
t = 0.11274, df = 736.75, p-value = 0.5449
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.08200385
sample estimates:
mean of x mean of y
 5.635739  5.630485
```

Nos fijaremos en el valor de p-value, como este es superior a 0,05 **no** podemos demostrar los vinos tienen mejor calidad a mayor cantidad de azúcar residual.

## 4.5 Modelo de regresión lineal

Resulta útil realizar predicciones sobre la calidad del vino dadas sus características.

Vamos a calcular un modelo de regresión lineal utilizando regresores cuantitativos para hacer predicciones de la calidad. Vamos a crear varios modelos utilizando las variables más correlacionadas con la calidad y escogeremos el mejor mirando cual tiene el mayor coeficiente de determinación.

```
alcohol = wine_data$alcohol
sulphates = wine_data$sulphates
volatile = wine_data$volatile.acidity
citric = wine_data$citric.acid
total.sulfur = wine_data$total.sulfur.dioxide
chlorides = wine_data$chlorides
density = wine_data$density

# Variable a predecir
calidad = wine_data$quality
```

```r
# Generación de varios modelos
modelo1 <- lm(calidad ~  alcohol + sulphates + volatile + citric , data = wine_data)
modelo2 <- lm(calidad ~ sulphates + volatile + citric + chlorides, data = wine_data)
modelo3 <- lm(calidad ~  volatile + citric + chlorides + density, data = wine_data)
modelo4 <- lm(calidad ~  citric + chlorides + density + alcohol, data = wine_data)
modelo5 <- lm(calidad ~ chlorides + density + alcohol + sulphates, data = wine_data)
modelo6 <- lm(calidad ~  density + alcohol + sulphates + volatile, data = wine_data)
tabla.coeficientes <- matrix(c(1, summary(modelo1)$r.squared,
                               2, summary(modelo2)$r.squared,
                               3, summary(modelo3)$r.squared,
                               4, summary(modelo3)$r.squared,
                               5, summary(modelo3)$r.squared,
                               6, summary(modelo3)$r.squared),
                             ncol = 2, byrow = TRUE)


colnames(tabla.coeficientes) <- c("Modelo", "R^2")
```

```
> tabla.coeficientes
     Modelo       R^2
[1,]      1 0.3345349
[2,]      2 0.2096944
[3,]      3 0.1969935
[4,]      4 0.1969935
[5,]      5 0.1969935
[6,]      6 0.1969935
```

Vemos que el primer modelo es el más conveniente dado que tiene un mayor coeficiente de determinación. Ahora vamos a hacer una predicción de la calidad de un vino con este modelo:

```r
newdata <- data.frame(
  alcohol = 9.4,
  sulphates = 0.56,
  volatile = 0.7,
  citric = 0
)
```

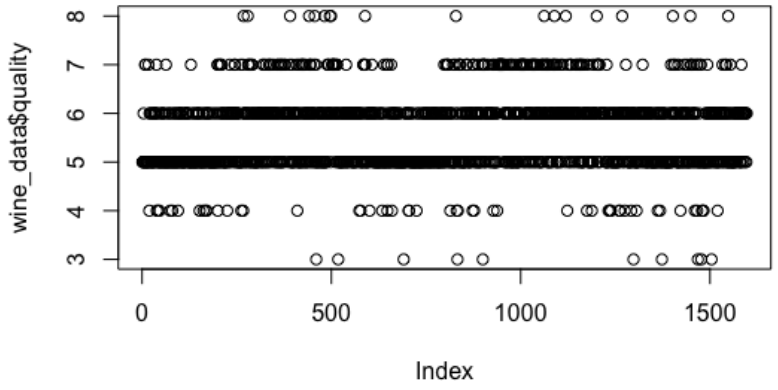
```
> predict(modelo1, newdata)
       1
5.057604
```

Se ha escogido los datos se la primera fila del dataset (con calidad = 5) y efectivamente vemos que ha hecho una predicción correcta.

## 5. Representación

Vamos a analizar los datos en torno a la variable quality (que es la que nos interesa) de forma visual:

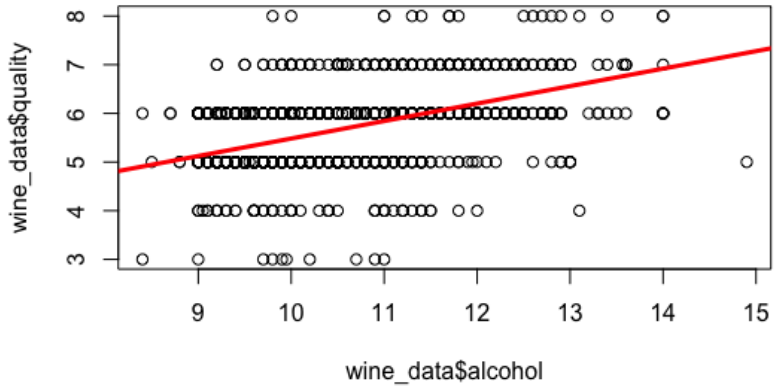
```
plot(wine_data$quality)
```




Vemos que hay muchos más vinos con calidad 5-6, menos de calidad 7, y aún menos de calidad 8.

Ahora, vamos a ver la relación de la calidad con las características que hemos detectado tenían más y menos correlación de forma gráfica:

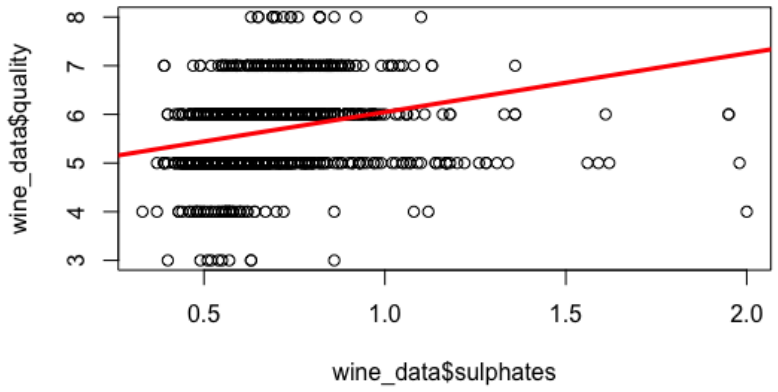
```
plot(wine_data$alcohol, wine_data$quality)
abline(lm(wine_data$quality~wine_data$alcohol),col="red",lwd=3)
```



En esta gráfica vemos la relación entre la calidad y la característica alcohol. Con la linea vemos que existe dicha relación (aunque ligera), ya que coincide que los vinos de menor calidad tienen menor graduación en alcohol y los de mayor calidad tienen una graduación media-alta.

Vamos a ver en el caso de los sulphates:

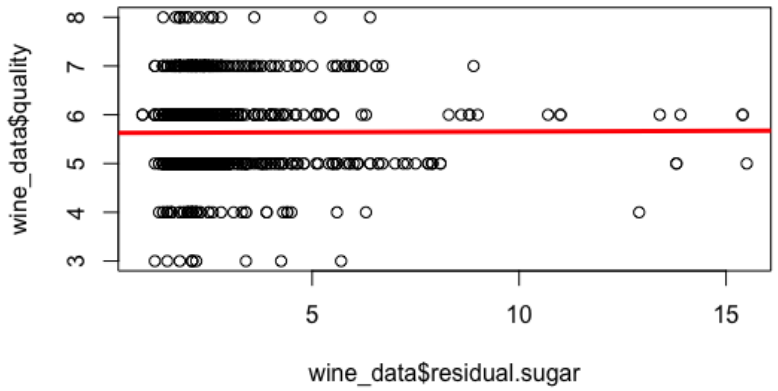
```
plot(wine_data$sulphates, wine_data$quality)
abline(lm(wine_data$quality~wine_data$sulphates),col="red",lwd=3)
```




Vemos que ocurre algo parecido que con el alcohol. Los vinos de menor calidad solo tienen valores de sulphates inferiores a 1, en cambio a mayor calidad los valores se encuentran más cercanos a 1.

Y por último, en el caso de residual sugar, que antes hemos visto que tenía muy poca correlación con la calidad del vino:

```
plot(wine_data$residual.sugar, wine_data$quality)
abline(lm(wine_data$quality~wine_data$residual.sugar),col="red",lwd=3)
```



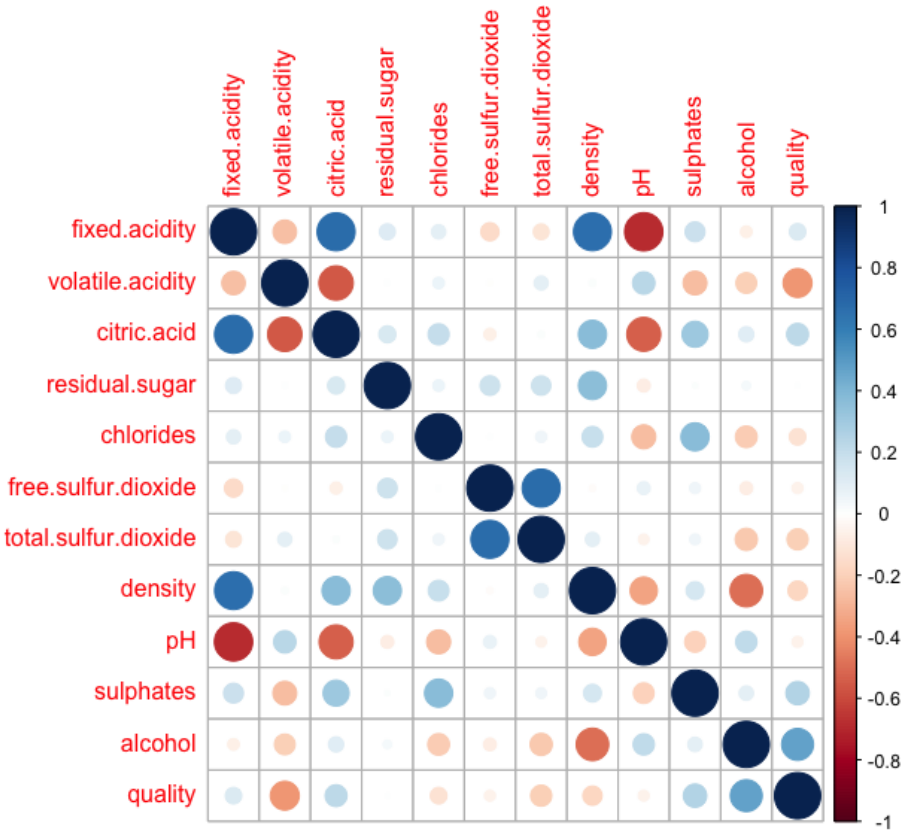

En cambio, si nos fijamos en la relación con residual sugar, vemos que no hay apenas variación en la linea, lo que indica que no se ha encontrado relación. Si nos fijamos en el gráfico, tanto vinos con baja y alta calidad tienen valores de residual sugar < 5, y los que tienen valores más elevados de azúcar son de calidad media.

Aprovechando que tenemos la matriz de correlación, vamos a ver la gráfica de la correlación que hay entre las características.

Primera tendremos que instalar la librería: install.packages("corrplot")

```
library(corrplot)

M <- cor(wine_data)
corrplot(M, method = "circle")
```




En esta tabla podemos que, por ejemplo, fixed acidity esta correlacionada con density y pH o, lo ya comentado anteriormente, que las más correlacionadas con la calidad son alcohol y volatile acidity.

## 6. Conclusiones

Mediante el análisis de correlación y otros contrastes hemos podido ver que características ejercen mayor influencia sobre la calidad del vino tinto. También de forma visual en las gráficas.

Hemos concluido que las características que más influyen en la predicción de la calidad son:

- Alcohol
- Sulphates
- Volatile acidity

También hemos probado diferentes modelos de regresión con distintos campos para ver cual resultaba más efectivo. Se ha decidido que el primero, el cuál tiene como parámetros las características: alcohol, sulphates, volatile acidity y citric acid.

Por último se hecho la predicción con datos de un vino existente y la calidad ha dado el resultado que esperábamos.