PRÁCTICA 2 - TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Manuel Cubertorer Gumbau y Francisco Javier Corrales Estrella

11/12/2021

Contents

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Los conjuntos de datos corresponden a una serie de registros de tipos de vino, obtenidos a partir de:

https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

El dataset está formado por un total de 1599 registros de vino rojo y 4898 registros de vino blanco, y por 12 variables fisicoquímicas. Se definen una serie atributos como la acidez o la graduación, y una variable target con la calidad del vino. Extraeremos los dos dataset disponibles, uno para vinos blancos y otros para

vinos tintos, y los fusionaremos en uno solo creando una variable categórica para el tipo de vino, el resto de variables son numéricas.

Los campos de los que se compone el dataset son los siguientes:

- **fixed acidity**: La mayoría de los ácidos involucrados con el vino son fijos o no volátiles (no se evaporan fácilmente)
- **volatile acidity**: Cantidad de ácido acético en el vino, que en niveles demasiado altos puede llevar a un sabor desagradable a vinagre.
- **citric acid**: En pequeñas cantidades, el ácido cítrico puede agregar "frescura" y sabor a los vinos.
- **residual sugar**: Cantidad de azúcar residual después de la fermentación. Es raro encontrar vinos con menos de 1 g/l y los vinos con más de 45 g/l se consideran dulces.
- **chlorides**: Cantidad de sal en el vino.
- **free sulfur dioxide**: En estado natural, el SO2 presenta un equilibrio entre el SO2 molecular (como un gas disuelto) y el ion bisulfito. Previene el crecimiento microbiano y la oxidación del vino.
- **total sulfur dioxide**: Cantidad de formas libres y ligadas de SO2. En bajas concentraciones, el SO2 es mayormente indetectable en el vino, pero a concentraciones de SO2 libres superiores a 50 ppm, el SO2 se hace evidente en el olfato y también en el sabor del vino.
- **density**: Densidad del agua según el porcentaje de alcohol y contenido en azúcar.
- **pH**: Describe el grado de acidez o basicidad del vino en una escala de 0 (muy ácido) a 14 (muy básico). La mayoría de los vinos están entre 3 y 4 en la escala de pH.
- **sulphates**: Aditivo para vinos que puede contribuir a los niveles de gas de SO2, que actúa como antimicrobiano y antioxidante.
- **alcohol**: Porcentaje de alcohol en el vino.
- **quality**: Indica la calidad del vino en una escala del 1 al 10.
- **tipo_vino**: Variable categórica que distingue entre vinos blancos y vinos tintos.

Nuestro análisis, tratará de determinar que variable/s son más determinantes en la calidad del vino, y compararemos cómo influye en algunas de ellas el tipo de vino (blanco o tinto).

Este tipo de análisis son muy relevantes en el mundo de las bodegas y los vinos donde se utilizan estos datos para realizar investigaciones sobre la calidad de los vinos, las uvas y sus cualidades fisicoquímicas.

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

Primero cargamos los datos desde el repositorio de datatsets UCI Machine Learning. Luego creamos la variable "tipo" que nos indique el tipo de vino (blanco o tinto) y juntamos los dos datasets en uno.

Una **consideración importante**, es que no fusionaremos realmente los dos datasets hasta que hayamos completado las tareas de limpieza y preparación de datos pues no queremos que las distribuciones de los datos se mezclen, por ejemplo, los valores extremos los queremos tratar separados por cada tipo de vino.

```
red_wine_data <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequa
white_wine_data <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winec

colnames(red_wine_data) <- c("fixed_acidity","volatile_acidity", "citric_acid", "residual_sugar", "chlo

colnames(white_wine_data) <- c("fixed_acidity","volatile_acidity", "citric_acid", "residual_sugar", "ch

red_wine_data$tipo <- 'tinto'
white_wine_data$tipo <- 'blanco'

wine_data = rbind(white_wine_data, red_wine_data)
```

Ahora vamos a mostrar las primeras líneas del dataset para comprobar que se ha cargado correctamente.

```
head(wine_data, 10)
```

```
##    fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
## 1            7.0             0.27        0.36           20.7     0.045
## 2            6.3             0.30        0.34            1.6     0.049
## 3            8.1             0.28        0.40            6.9     0.050
## 4            7.2             0.23        0.32            8.5     0.058
## 5            7.2             0.23        0.32            8.5     0.058
## 6            8.1             0.28        0.40            6.9     0.050
## 7            6.2             0.32        0.16            7.0     0.045
## 8            7.0             0.27        0.36           20.7     0.045
## 9            6.3             0.30        0.34            1.6     0.049
## 10           8.1             0.22        0.43            1.5     0.044
##    free_sulfur_dioxide total_sulfur_dioxide density   ph sulphates alcohol
## 1                   45                  170  1.0010 3.00      0.45     8.8
## 2                   14                  132  0.9940 3.30      0.49     9.5
## 3                   30                   97  0.9951 3.26      0.44    10.1
## 4                   47                  186  0.9956 3.19      0.40     9.9
## 5                   47                  186  0.9956 3.19      0.40     9.9
## 6                   30                   97  0.9951 3.26      0.44    10.1
## 7                   30                  136  0.9949 3.18      0.47     9.6
## 8                   45                  170  1.0010 3.00      0.45     8.8
## 9                   14                  132  0.9940 3.30      0.49     9.5
## 10                  28                  129  0.9938 3.22      0.45    11.0
##    quality   tipo
## 1        6 blanco
## 2        6 blanco
## 3        6 blanco
## 4        6 blanco
## 5        6 blanco
## 6        6 blanco
## 7        6 blanco
## 8        6 blanco
## 9        6 blanco
## 10       6 blanco
```

A continuación mostraremos la estructura de los datos. Donde comprobamos que todas las variables son numéricas, excepto la variable categórica **tipo** que hemos creado.

```
str(wine_data)
```

```
## 'data.frame':    6497 obs. of  13 variables:
##  $ fixed_acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile_acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric_acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual_sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free_sulfur_dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total_sulfur_dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ ph                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
```

```
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ tipo                : chr  "blanco" "blanco" "blanco" "blanco" ...
```

También vemos que la variable **quality** es del tipo integer, así que la transformaremos a tipo numeric para que sea completamente compatible con el resto de variables y evitar posibles conflictos.

```
# Convertimos la columna "quality" a tipo numeric
wine_data$quality<-as.numeric(wine_data$quality)
class(wine_data$quality)
```

```
## [1] "numeric"
```

Por otra parte, también vemos que existen dos variables para definir la acidez (fixed_acidity y volatile_acidity). Así pues podemos crear una variable nueva llamada **acidity** que recoja la suma de estas dos y por tanto indique la acidez total del vino.

```
# Creamos la nueva columna acidity
wine_data$acidity<-wine_data$fixed_acidity + wine_data$volatile_acidity
# Eliminamos las columnas fixed_acidity y volatile_acidity
wine_data <- wine_data[, -(1:2)]
wine_data <- subset(wine_data, select=c(12,1:11))
head(wine_data)
```

```
##   acidity citric_acid residual_sugar chlorides free_sulfur_dioxide
## 1    7.27        0.36           20.7     0.045                  45
## 2    6.60        0.34            1.6     0.049                  14
## 3    8.38        0.40            6.9     0.050                  30
## 4    7.43        0.32            8.5     0.058                  47
## 5    7.43        0.32            8.5     0.058                  47
## 6    8.38        0.40            6.9     0.050                  30
##   total_sulfur_dioxide density   ph sulphates alcohol quality   tipo
## 1                  170  1.0010 3.00      0.45     8.8       6 blanco
## 2                  132  0.9940 3.30      0.49     9.5       6 blanco
## 3                   97  0.9951 3.26      0.44    10.1       6 blanco
## 4                  186  0.9956 3.19      0.40     9.9       6 blanco
## 5                  186  0.9956 3.19      0.40     9.9       6 blanco
## 6                   97  0.9951 3.26      0.44    10.1       6 blanco
```

Estadísticas principales de los datos:

```
summary(wine_data)
```

```
##     acidity        citric_acid     residual_sugar     chlorides
##  Min.   : 4.110   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900
##  1st Qu.: 6.710   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800
##  Median : 7.300   Median :0.3100   Median : 3.000   Median :0.04700
##  Mean   : 7.555   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603
##  3rd Qu.: 8.050   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500
##  Max.   :16.285   Max.   :1.6600   Max.   :65.800   Max.   :0.61100
```

```
##  free_sulfur_dioxide total_sulfur_dioxide    density             ph
##  Min.   :  1.00      Min.   :  6.0        Min.   :0.9871   Min.   :2.720
##  1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110
##  Median : 29.00      Median :118.0        Median :0.9949   Median :3.210
##  Mean   : 30.53      Mean   :115.7        Mean   :0.9947   Mean   :3.219
##  3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320
##  Max.   :289.00      Max.   :440.0        Max.   :1.0390   Max.   :4.010
##    sulphates         alcohol         quality          tipo
##  Min.   :0.2200   Min.   : 8.00   Min.   :3.000   Length:6497
##  1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000   Class :character
##  Median :0.5100   Median :10.30   Median :6.000   Mode  :character
##  Mean   :0.5313   Mean   :10.49   Mean   :5.818
##  3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :2.0000   Max.   :14.90   Max.   :9.000
```

3. Limpieza de los datos

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Mostramos las estadísticas de valores vacíos o nulos.

```
# Estadísticas de valores vacíos
colSums(is.na(wine_data))
```

```
##              acidity          citric_acid       residual_sugar
##                    0                    0                    0
##            chlorides  free_sulfur_dioxide total_sulfur_dioxide
##                    0                    0                    0
##              density                   ph            sulphates
##                    0                    0                    0
##              alcohol              quality                tipo
##                    0                    0                    0
```

```
colSums(wine_data=="")
```

```
##              acidity          citric_acid       residual_sugar
##                    0                    0                    0
##            chlorides  free_sulfur_dioxide total_sulfur_dioxide
##                    0                    0                    0
##              density                   ph            sulphates
##                    0                    0                    0
##              alcohol              quality                tipo
##                    0                    0                    0
```

En nuestro caso no existen valores nulos, si los hubiese la alternativa mas sencilla es setear el valor de la media para todo el conjunto de datos. Esto lo podemos mejorar tomando alguna medida de tendencia central dependiendo de la distribución de los datos, esto se puede hacer para toda la muestra o en función de alguna variable categórica, en nuestro caso el tipo de vino.

Existen otros métodos, como kNN que se basa en la similitud, básicamente se fija en que valores tiene esa variable en los "vecinos" mas cercanos, donde definimos cuantos vecinos queremos tomar.´

Otro análisis que vamos a realizar para la limpieza de los datos es detectar si existen valores duplicados. Los valores duplicados no aportan ninguna información adicional y se deberían eliminar para mayor integridad de los datos.

```
library("dplyr")
# Detección y eliminación de valores duplicados
sum(duplicated(wine_data))
```

```
## [1] 1177
```

```
wine_data <- distinct(wine_data)
```

Con estos cambios el total de registros que tenemos ahora es de:

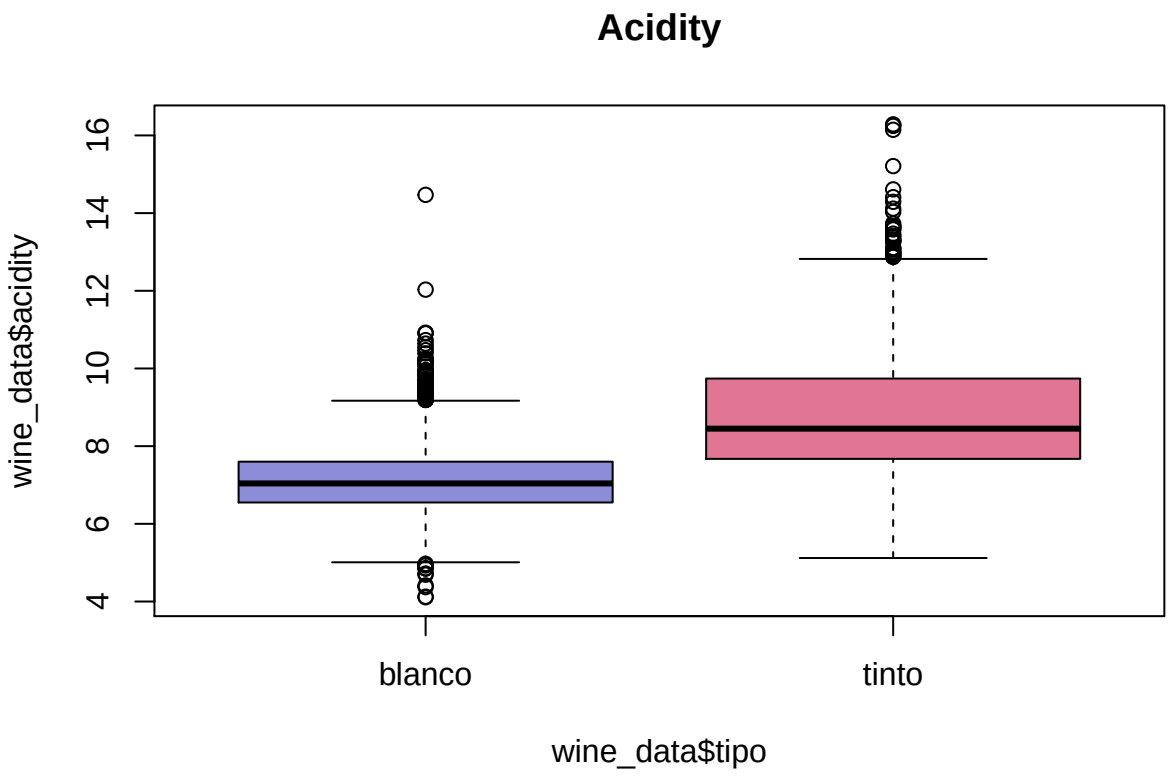
```
dim(wine_data)
```

```
## [1] 5320   12
```

3.2 Identificación y tratamiento de valores extremos

Consideramos valores extremos como aquellos valores que son sospechosos por alejarse demasiado del resto de datos, esto en términos numéricos quiere decir, que están demasiado alejados de la media teniendo en cuenta la desviación típica. Podemos hacer esta aproximación de una forma visual mediante un gráfico de caja o calculando los valores fuera del rango intercuartílico, podemos usar la función boxplots.stats() para esto.

```
myColors <- c(rgb(0.1,0.1,0.7,0.5) ,rgb(0.8,0.1,0.3,0.6))

tintos <- subset(wine_data, wine_data$tipo == "tinto")
blancos <- subset(wine_data, wine_data$tipo == "blanco")

boxplot(wine_data$acidity ~ wine_data$tipo, main="Acidity", col =myColors )
```



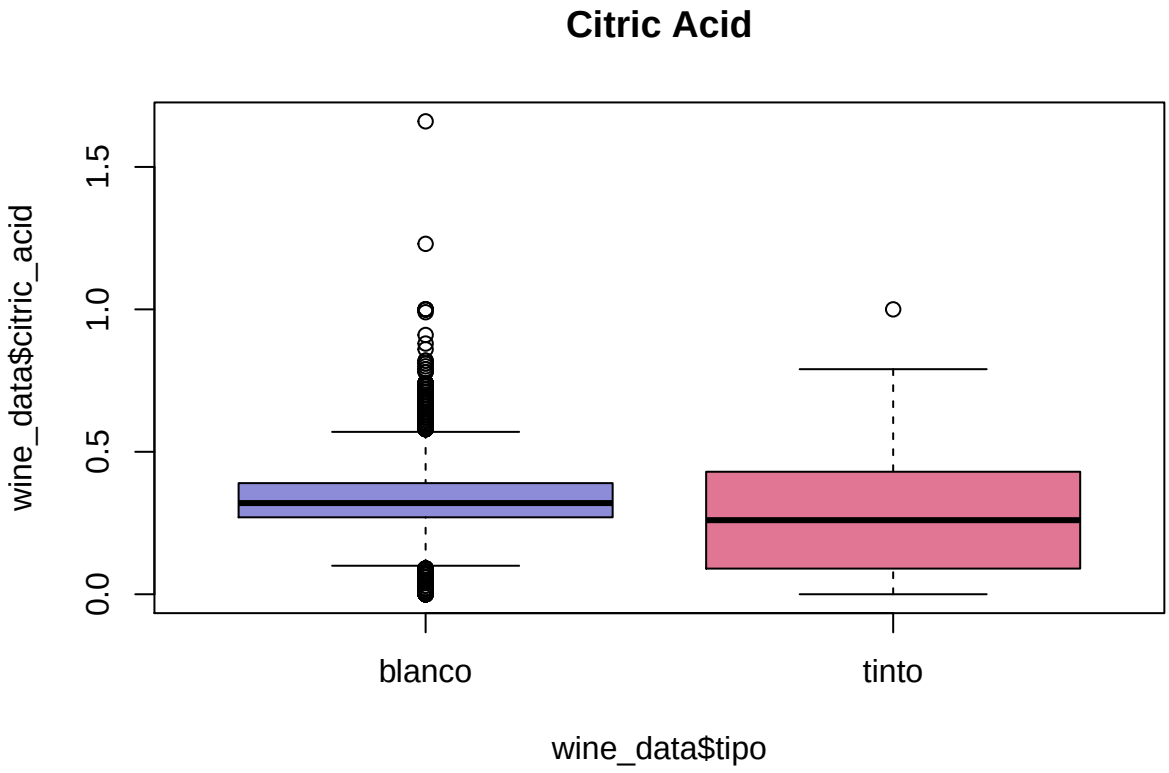

```
out_ac_tinto <- boxplot(tintos$acidity, plot=FALSE)$out
out_ac_blanco <- boxplot(blancos$acidity, plot=FALSE)$out
out_ac_tinto
```

```
##  [1] 13.100 15.210 13.060 13.640 13.670 12.890 14.290 14.030 12.980 12.960
## [11] 13.415 14.410 14.115 13.300 12.960 13.640 12.910 16.285 12.880 13.320
## [21] 12.870 13.590 13.100 13.250 14.610 16.145 16.245 13.470 13.300 13.290
## [31] 13.660 13.580 16.260 13.730 13.400 13.010 12.990
```

```
out_ac_blanco
```

```
##  [1] 10.160 10.220 10.640  9.690 10.200  9.450  9.270  9.370  9.540 10.470
## [11]  9.690  9.420 10.160  9.830  9.480  9.310  9.640  9.370  9.240 10.240
## [21]  9.230  9.480  9.940  9.660 10.550  9.890  9.245  9.840  9.350  9.570
## [31]  9.850  9.520  9.240  9.550 10.920  9.310  9.480 10.050  9.380 14.470
## [41]  9.380 10.110  9.300  9.610  9.220  9.220  9.190  9.430  9.500  9.430
## [51]  9.810  9.560  9.560  9.190 10.910  9.240  9.230 10.230  9.470  9.910
## [61] 10.905  9.710  9.550  9.920 12.030  9.700  9.430 10.730 10.390  9.430
## [71]  9.540  9.820  9.680  9.630  9.390  9.360  9.960  9.260  9.290  4.690
## [81]  9.340 10.255  4.370  9.940  9.380  4.415  9.640  9.190  9.200  4.970
## [91]  4.110  4.860  4.845  9.200  4.930  4.720  4.125  4.940
```

```
boxplot(wine_data$citric_acid ~ wine_data$tipo, main="Citric Acid", col =myColors )
```



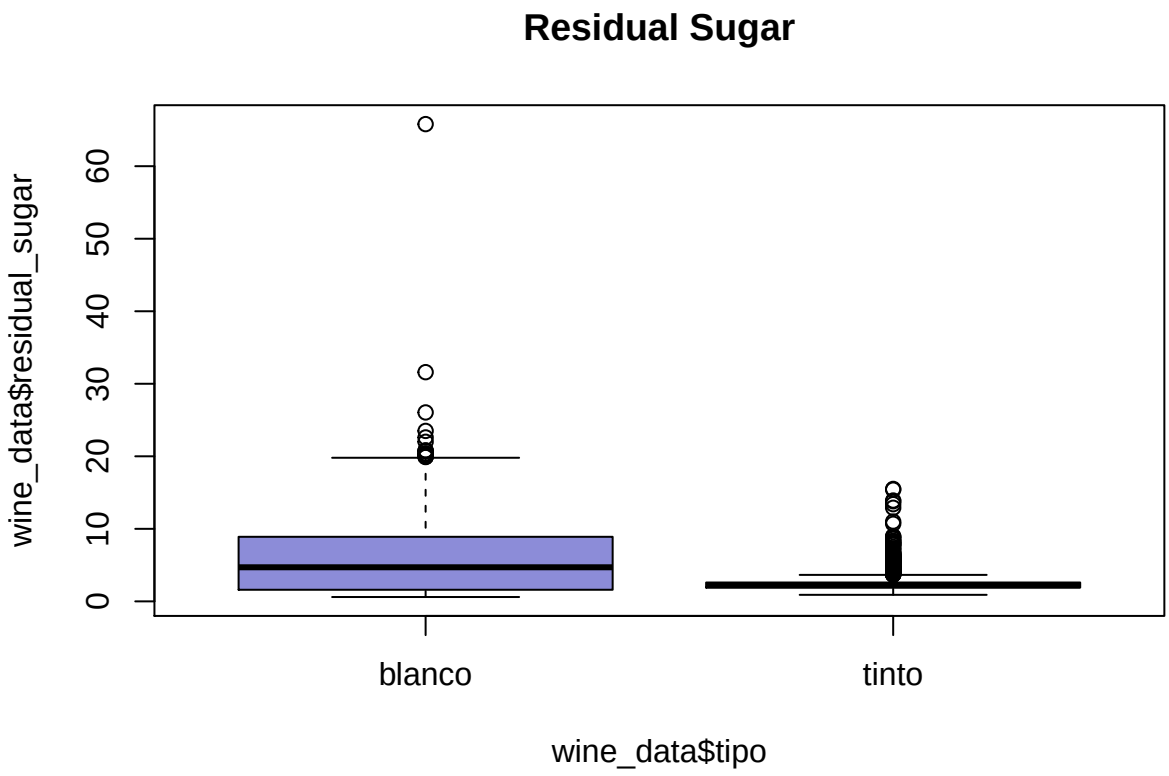

```
out_ca_tinto <- boxplot(tintos$citric_acid, plot=FALSE)$out
out_ca_blanco <- boxplot(blancos$citric_acid, plot=FALSE)$out
out_ca_tinto
```

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

```
out_ca_blanco
```

```
##   [1] 0.62 0.04 0.59 0.07 0.03 0.61 0.62 0.63 0.66 0.00 0.04 0.67 0.04 0.07 0.88
##  [16] 0.08 0.59 0.07 0.07 0.07 0.07 0.58 0.70 0.00 0.60 0.07 0.09 0.04 0.62 0.58
##  [31] 0.70 0.62 0.58 0.02 0.65 0.71 0.66 0.07 0.06 0.68 0.68 0.06 0.72 0.69 0.58
##  [46] 0.70 1.66 0.04 0.63 0.60 0.00 0.08 0.58 0.05 0.58 0.00 0.00 0.65 0.00 0.05
##  [61] 0.05 0.62 0.58 1.00 0.09 0.01 0.71 0.71 0.60 0.06 0.74 0.81 0.69 0.58 0.00
##  [76] 0.07 0.64 0.72 0.73 0.65 0.68 0.74 0.71 0.59 0.68 0.08 0.72 0.64 0.02 0.74
##  [91] 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74
## [106] 0.74 0.74 0.99 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.01 0.74
## [121] 0.01 0.74 0.74 1.00 0.04 0.58 0.07 1.00 0.00 0.58 0.61 0.02 0.67 0.58 0.65
## [136] 0.58 0.09 0.08 0.71 0.04 0.03 0.05 0.64 0.58 0.81 0.58 0.61 0.62 0.59 0.00
## [151] 0.04 0.63 0.73 0.68 0.09 0.78 0.79 0.09 0.64 0.65 0.00 0.73 0.64 0.60 0.71
## [166] 0.72 0.82 0.07 0.58 1.00 0.66 0.80 1.23 0.59 0.02 0.00 1.00 0.62 0.00 0.71
## [181] 0.61 0.00 0.60 0.58 0.09 0.09 0.72 0.62 0.62 0.79 0.82 0.67 0.01 0.86 0.61
## [196] 0.02 0.05 0.00 0.69 0.59 0.01 0.66 0.78 0.00 0.04 0.91 0.06 0.06 0.04 0.74
## [211] 0.09 0.60 0.62 0.73 0.00 0.09 0.00 0.09 0.67 0.01 0.09 0.00 0.02
```

```
boxplot(wine_data$residual_sugar ~ wine_data$tipo, main="Residual Sugar", col =myColors )
```



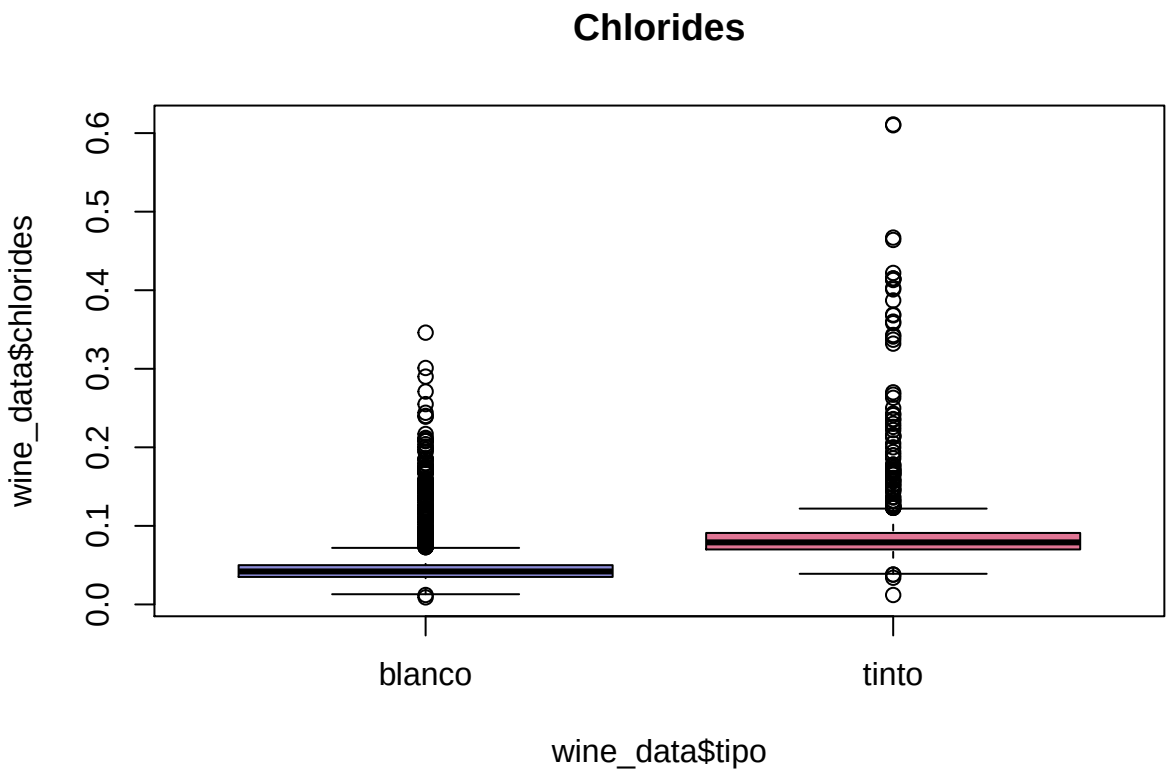

```
out_rs_tinto <- boxplot(tintos$residual_sugar, plot=FALSE)$out
out_rs_blanco <- boxplot(blancos$residual_sugar, plot=FALSE)$out
out_rs_tinto
```

```
##   [1]  6.10  3.80  3.90  4.40 10.70  5.50  5.90  3.80  5.10  4.65  5.50  5.50
##  [13]  7.30  7.20  3.80  5.60  4.00  4.00  4.00  7.00  6.40  5.60 11.00  4.50
##  [25]  4.80  5.80  3.80  4.40  6.20  4.20  7.90  3.70  4.50  6.70  6.60  3.70
##  [37]  5.20 15.50  4.10  8.30  6.55  4.60  6.10  4.30  5.80  5.15  6.30  4.20
##  [49]  4.60  4.20  4.30  7.90  4.60  5.10  5.60  6.00  8.60  7.50  4.40  4.25
##  [61]  6.00  3.90  4.20  4.00  4.00  6.60  6.00  3.80  9.00  4.60  8.80  5.00
##  [73]  3.80  4.10  5.90  4.10  6.20  8.90  4.00  3.90  8.10  6.40  8.30  8.30
##  [85]  4.70  5.50  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90
##  [97]  4.30 13.40  4.80  6.30  4.50  4.30  3.90  3.80  5.40  3.80  6.10  3.90
## [109]  5.10  3.90 15.40  4.80  5.20  5.20  3.75 13.80  5.70  4.30  4.10  4.10
## [121]  4.40  3.70  6.70 13.90  5.10  7.80
```

```
out_rs_blanco
```

```
##  [1] 20.70 22.00 20.80 23.50 31.60 19.95 20.40 65.80 20.20 20.15 19.95 19.90
## [13] 26.05 20.80 20.30 22.60
```

```r
boxplot(wine_data$chlorides ~ wine_data$tipo, main="Chlorides", col =myColors )
```




```r
out_ch_tinto <- boxplot(tintos$chlorides, plot=FALSE)$out
out_ch_blanco <- boxplot(blancos$chlorides, plot=FALSE)$out
out_ch_tinto
```

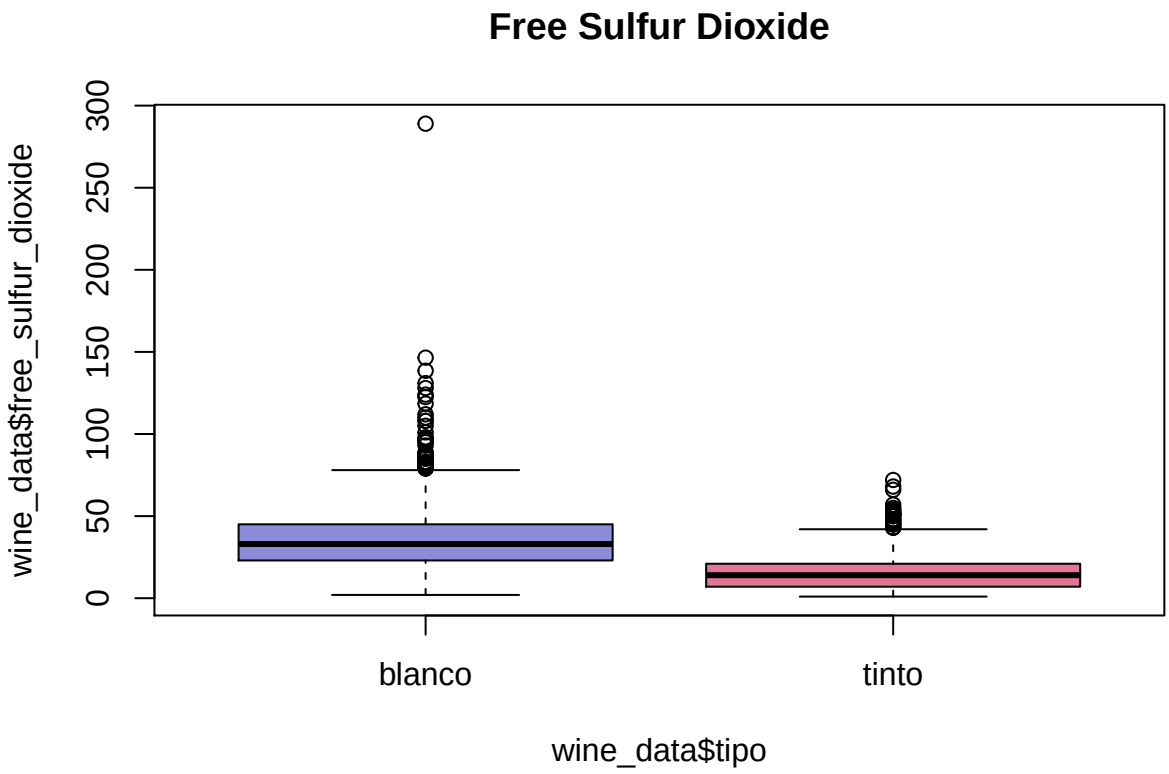
```
##  [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.178 0.146 0.236
## [13] 0.610 0.360 0.270 0.337 0.263 0.611 0.358 0.343 0.186 0.213 0.214 0.128
## [25] 0.159 0.124 0.174 0.127 0.413 0.152 0.152 0.125 0.200 0.171 0.226 0.250
## [37] 0.148 0.124 0.143 0.222 0.157 0.422 0.034 0.387 0.415 0.157 0.157 0.243
## [49] 0.241 0.190 0.132 0.126 0.038 0.165 0.145 0.147 0.012 0.194 0.132 0.161
## [61] 0.123 0.414 0.216 0.171 0.178 0.369 0.166 0.166 0.136 0.132 0.123 0.123
## [73] 0.403 0.137 0.414 0.166 0.168 0.415 0.153 0.267 0.123 0.214 0.169 0.205
## [85] 0.235 0.230 0.038
```

```r
out_ch_blanco
```

```
##   [1] 0.074 0.080 0.172 0.173 0.147 0.092 0.082 0.092 0.200 0.197 0.197 0.074
##  [13] 0.132 0.089 0.108 0.081 0.073 0.346 0.090 0.114 0.186 0.180 0.084 0.083
##  [25] 0.096 0.094 0.240 0.290 0.185 0.110 0.078 0.130 0.135 0.115 0.170 0.080
##  [37] 0.119 0.126 0.150 0.152 0.088 0.244 0.137 0.093 0.077 0.079 0.073 0.076
##  [49] 0.201 0.074 0.301 0.138 0.169 0.083 0.093 0.168 0.122 0.172 0.167 0.239
##  [61] 0.076 0.138 0.137 0.123 0.133 0.073 0.211 0.123 0.255 0.204 0.208 0.083
##  [73] 0.080 0.076 0.086 0.084 0.168 0.160 0.179 0.076 0.087 0.217 0.094 0.157
```

```
##  [85] 0.148 0.158 0.157 0.168 0.157 0.092 0.099 0.084 0.085 0.091 0.093 0.080
##  [97] 0.095 0.096 0.147 0.142 0.079 0.074 0.075 0.121 0.079 0.156 0.012 0.119
## [109] 0.081 0.170 0.171 0.082 0.074 0.083 0.152 0.169 0.073 0.078 0.112 0.154
## [121] 0.126 0.104 0.142 0.102 0.184 0.096 0.076 0.146 0.117 0.117 0.118 0.085
## [133] 0.087 0.076 0.088 0.160 0.167 0.009 0.098 0.086 0.194 0.094 0.144 0.149
## [145] 0.185 0.084 0.175 0.090 0.098 0.110 0.095 0.174 0.097 0.142 0.145 0.208
## [157] 0.209 0.105 0.086 0.176 0.108 0.096 0.271 0.120 0.212 0.094 0.094 0.117
## [169] 0.173 0.074 0.076 0.076 0.175 0.174 0.075 0.127 0.096 0.136
```

```
boxplot(wine_data$free_sulfur_dioxide ~ wine_data$tipo, main="Free Sulfur Dioxide", col =myColors )
```



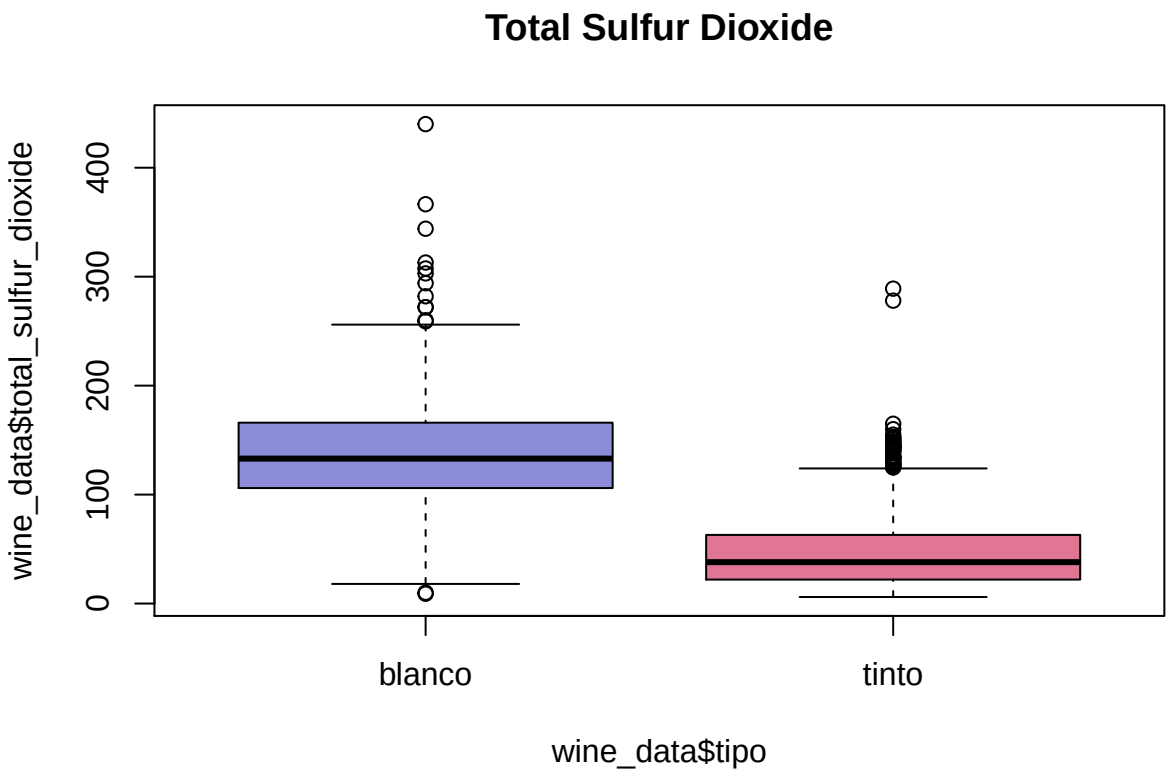

```
out_fsd_tinto <- boxplot(tintos$free_sulfur_dioxide, plot=FALSE)$out
out_fsd_blanco <- boxplot(blancos$free_sulfur_dioxide, plot=FALSE)$out
out_fsd_tinto
```

```
##  [1] 52 51 50 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51 52 55 48
## [26] 66
```

```
out_fsd_blanco
```

```
##  [1]  81.0  82.0 131.0  82.5  87.0  83.0  79.0 122.5  83.0  81.0  80.0  88.0
## [13]  82.0 118.5  81.0  96.0  83.0 146.5 128.0 110.0  85.0  89.0  86.0  96.0
## [25]  93.0  85.0  81.0 138.5  95.0 124.0  87.0 105.0 101.0 108.0  79.5  79.5
## [37] 108.0  98.0 112.0  81.0  81.0  79.0 289.0  97.0
```

```
boxplot(wine_data$total_sulfur_dioxide ~ wine_data$tipo, main="Total Sulfur Dioxide", col =myColors )
```



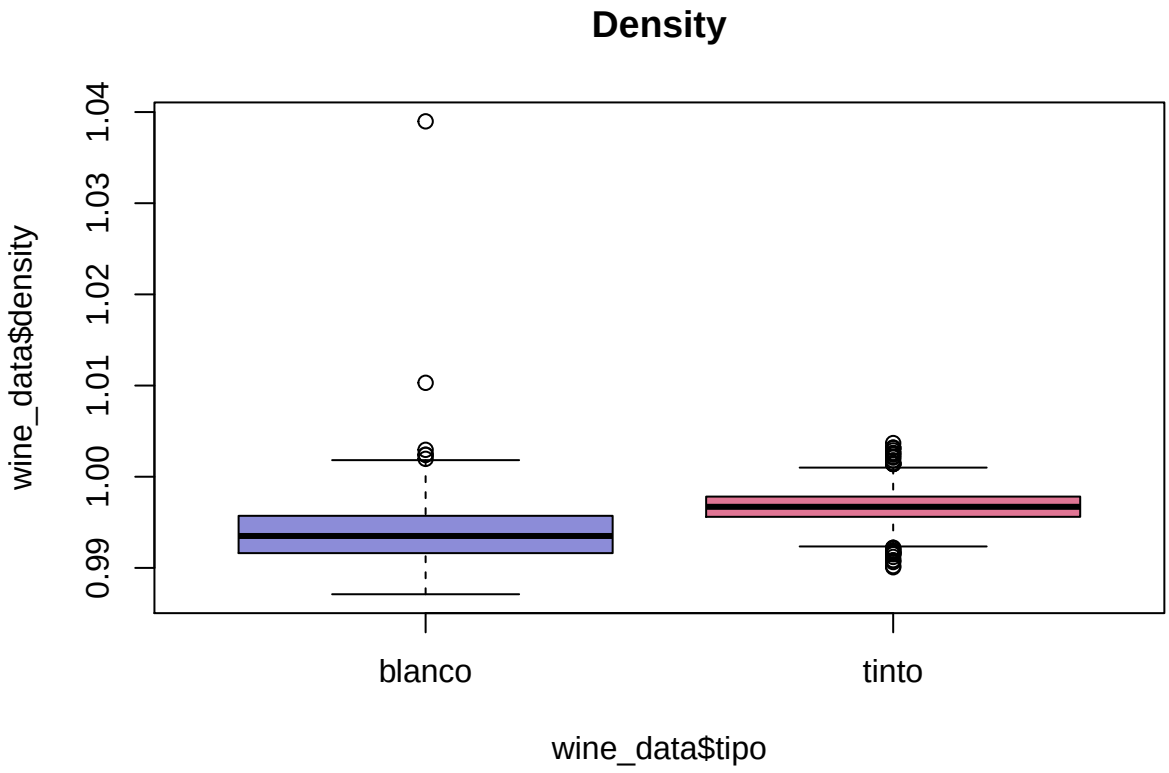

```
out_tsd_tinto <- boxplot(tintos$total_sulfur_dioxide, plot=FALSE)$out
out_tsd_blanco <- boxplot(blancos$total_sulfur_dioxide, plot=FALSE)$out
out_tsd_tinto
```

```
##  [1] 145 148 136 125 140 133 153 134 141 129 128 143 144 127 126 145 144 135 165
## [20] 134 129 151 133 142 149 147 145 148 155 151 152 125 127 139 143 144 130 278
## [39] 289 135 160 141 133 147 131
```

```
out_tsd_blanco
```

```
##  [1] 272.0 313.0 260.0 366.5 307.5 344.0 282.0 303.0 272.0 294.0   9.0  10.0
## [13] 259.0 440.0
```

```
boxplot(wine_data$density ~ wine_data$tipo, main="Density", col =myColors )
```



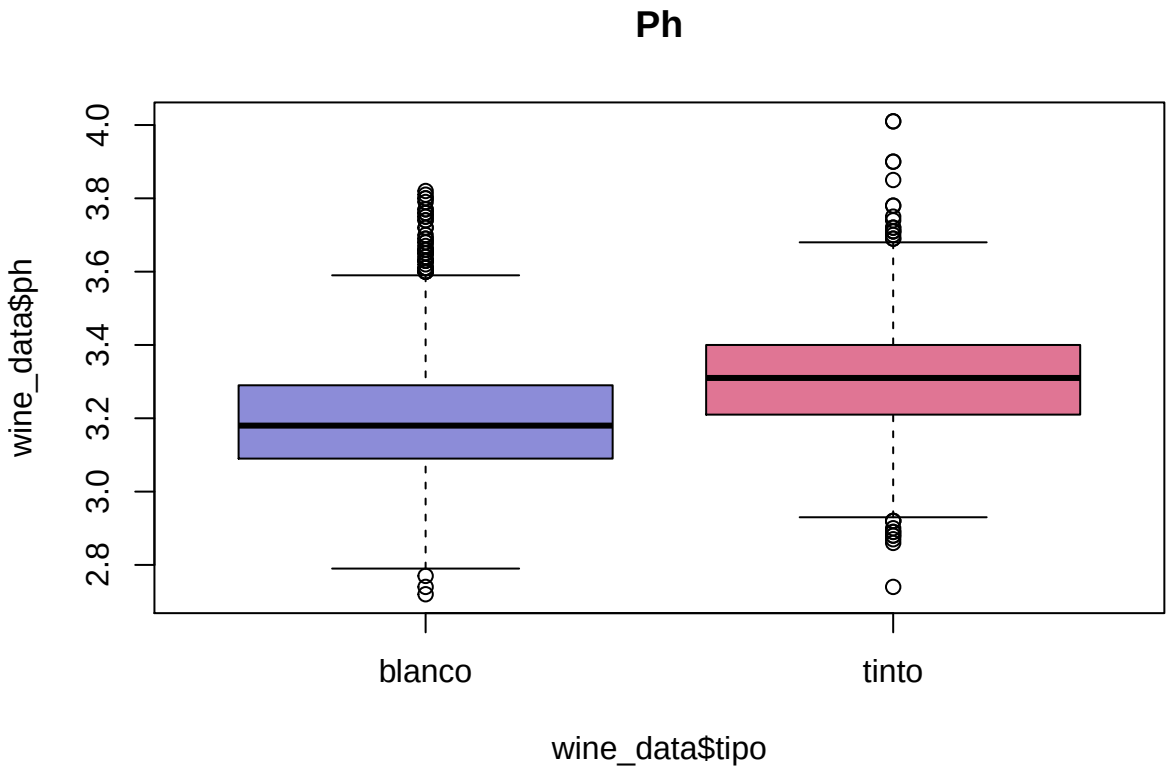

```
out_de_tinto <- boxplot(tintos$density, plot=FALSE)$out
out_de_blanco <- boxplot(blancos$density, plot=FALSE)$out
out_de_tinto
```

```
##  [1] 0.99160 1.00140 1.00150 1.00180 0.99120 1.00220 1.00140 1.00140 1.00140
## [10] 1.00320 1.00260 1.00140 1.00315 1.00315 1.00210 0.99170 0.99220 1.00260
## [19] 0.99210 0.99154 0.99064 1.00289 0.99162 0.99007 0.99020 0.99220 0.99150
## [28] 0.99157 0.99080 0.99084 0.99191 1.00369 1.00242 0.99182 0.99182
```

```
out_de_blanco
```

```
## [1] 1.00240 1.01030 1.00241 1.03898 1.00196 1.00295
```

```
boxplot(wine_data$ph ~ wine_data$tipo, main="Ph", col =myColors )
```



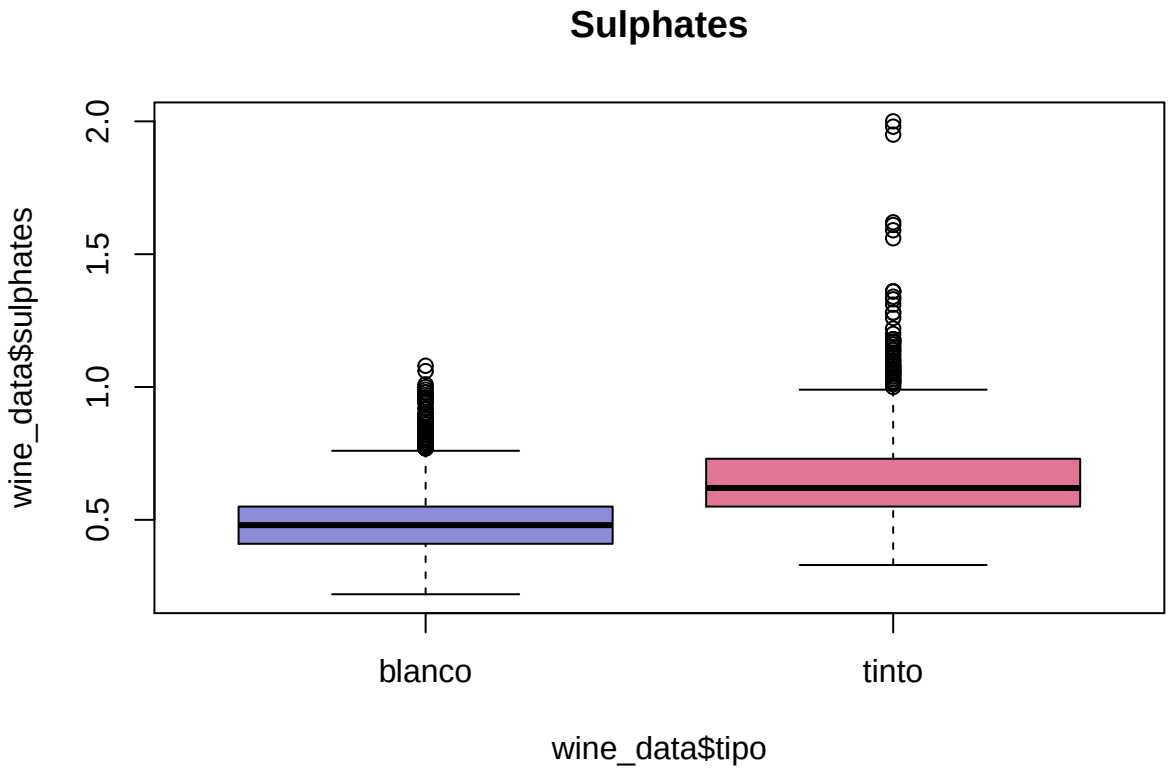

```
out_ph_tinto <- boxplot(tintos$ph, plot=FALSE)$out
out_ph_blanco <- boxplot(blancos$ph, plot=FALSE)$out
out_ph_tinto
```

```
##  [1] 3.90 3.75 3.85 2.74 3.69 2.88 2.86 3.74 2.92 2.92 3.72 2.87 2.89 2.92 3.90
## [16] 3.71 3.69 3.71 2.89 3.78 3.70 3.78 4.01 2.90 4.01 3.71 2.88 3.72
```

```
out_ph_blanco
```

```
##  [1] 3.69 3.63 3.72 3.61 3.64 3.72 3.66 2.74 3.82 3.81 3.65 3.77 3.62 3.63 3.65
## [16] 3.74 3.60 3.60 2.72 3.60 3.80 3.60 3.68 3.63 2.77 3.63 3.60 3.61 3.61 3.79
## [31] 3.68 3.66 3.70 3.74 3.80 3.65 3.77 3.76 3.69 3.66 3.75 3.63 3.75 3.76 3.66
## [46] 3.67
```

```
boxplot(wine_data$sulphates ~ wine_data$tipo, main="Sulphates", col =myColors )
```



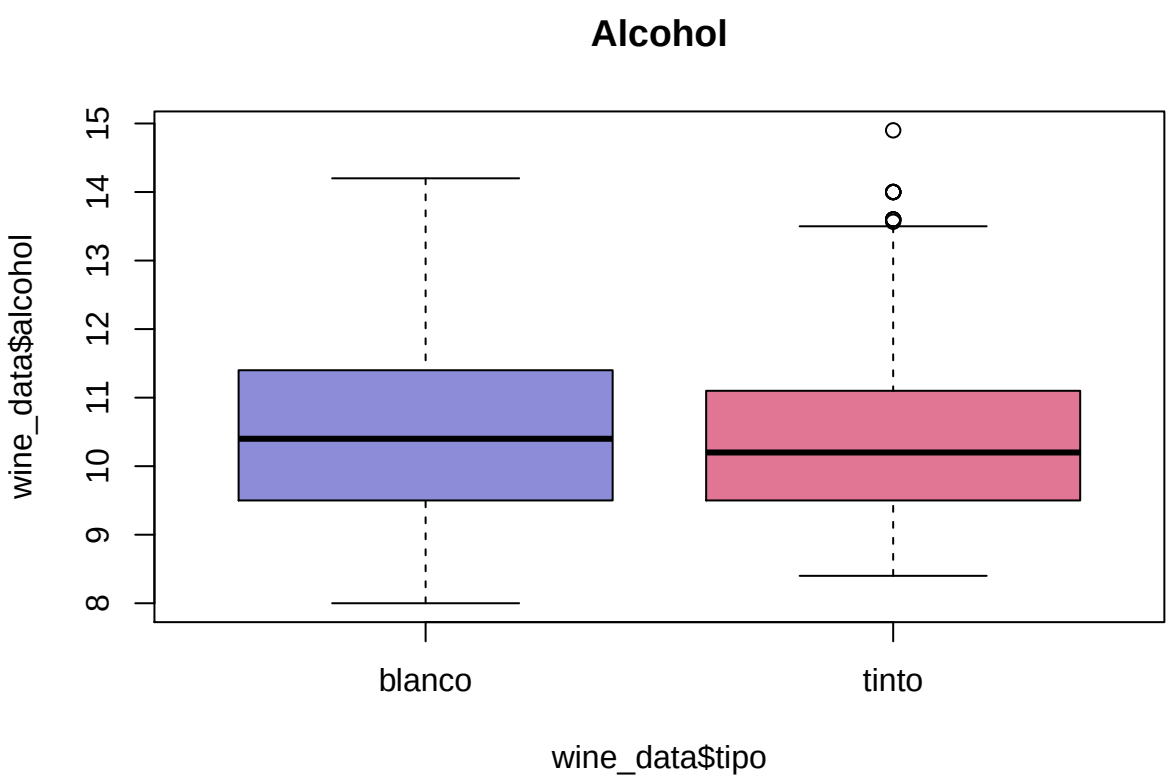

```
out_su_tinto <- boxplot(tintos$sulphates, plot=FALSE)$out
out_su_blanco <- boxplot(blancos$sulphates, plot=FALSE)$out
out_su_tinto
```

```
##  [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.98 1.31 2.00 1.08 1.59 1.02
## [16] 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.07 1.06 1.06 1.05
## [31] 1.06 1.04 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
## [46] 1.16 1.10 1.15 1.17 1.33 1.18 1.17 1.03 1.10 1.01
```

```
out_su_blanco
```

```
##  [1] 0.77 0.84 0.77 0.79 0.85 0.78 0.79 0.79 0.77 0.78 0.85 0.96 0.97 0.82 0.77
## [16] 0.95 0.77 0.82 0.82 0.90 0.88 0.88 0.79 0.80 0.78 0.87 0.86 0.90 0.78 0.79
## [31] 0.81 0.81 0.77 0.82 0.79 0.77 0.82 0.92 0.79 0.82 0.82 0.79 0.78 0.79 0.77
## [46] 0.77 0.98 1.06 0.88 0.88 0.80 0.78 1.00 0.80 0.90 0.89 0.94 0.99 0.86 0.84
## [61] 0.95 0.84 0.81 0.80 0.87 0.82 0.78 0.78 0.77 0.85 0.78 0.78 0.88 0.78 0.78
## [76] 0.79 0.77 0.83 0.83 0.81 0.98 0.79 0.78 0.82 0.98 0.77 0.96 1.01 0.77 0.96
## [91] 0.77 0.92 0.94 0.95 1.08 0.79
```

```
boxplot(wine_data$alcohol ~ wine_data$tipo, main="Alcohol", col =myColors )
```




```
out_al_tinto <- boxplot(tintos$alcohol, plot=FALSE)$out
out_al_blanco <- boxplot(blancos$alcohol, plot=FALSE)$out
out_al_tinto
```

```
##  [1] 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000 13.60000
##  [9] 14.00000 14.00000 13.56667 13.60000
```

```
out_al_blanco
```

```
## numeric(0)
```

Como podemos observar en los resultados anteriores, en todas las variables hay algunos valores atípicos. Aunque sí que es cierto que algunos de estos valores seguramente sean válidos y se correspondan con la realidad, dado que no tenemos el conocimiento suficiente para saberlo al cien por cien, hemos decidido eliminarlos para así asegurarnos que no causan problemas en los análisis estadísticos.

```
# Eliminamos valores outliers de cada una de las variables

tintos <- tintos[-which(tintos$acidity %in% out_ac_tinto),]
blancos <- blancos[-which(blancos$acidity %in% out_ac_blanco),]

tintos <- tintos[-which(tintos$citric_acid %in% out_ca_tinto),]
blancos <- blancos[-which(blancos$citric_acid %in% out_ca_blanco),]
```

```
tintos <- tintos[-which(tintos$residual_sugar %in% out_rs_tinto),]
blancos <- blancos[-which(blancos$residual_sugar %in% out_rs_blanco),]

tintos <- tintos[-which(tintos$chlorides %in% out_ch_tinto),]
blancos <- blancos[-which(blancos$chlorides %in% out_ch_blanco),]

tintos <- tintos[-which(tintos$free_sulfur_dioxide %in% out_fsd_tinto),]
blancos <- blancos[-which(blancos$free_sulfur_dioxide %in% out_fsd_blanco),]

tintos <- tintos[-which(tintos$total_sulfur_dioxide %in% out_tsd_tinto),]
blancos <- blancos[-which(blancos$total_sulfur_dioxide %in% out_tsd_blanco),]

tintos <- tintos[-which(tintos$density %in% out_de_tinto),]

tintos <- tintos[-which(tintos$ph %in% out_ph_tinto),]
blancos <- blancos[-which(blancos$ph %in% out_ph_blanco),]

tintos <- tintos[-which(tintos$sulphates %in% out_su_tinto),]
blancos <- blancos[-which(blancos$sulphates %in% out_su_blanco),]

tintos <- tintos[-which(tintos$alcohol %in% out_al_tinto),]

wine_data = rbind(tintos, blancos)

ncol(wine_data)
```

```
## [1] 12
```

```
nrow(wine_data)
```

```
## [1] 4365
```

4. Análisis de los datos

Nuestro análisis de datos está orientado a saber si hay diferencias en la calidad entre los vinos blancos y los vinos tintos, conocer que variables influyen mas en la calidad del vino y saber si podemos predecir y con que garantia que calidad tendrá un vino en función de sus atributos.

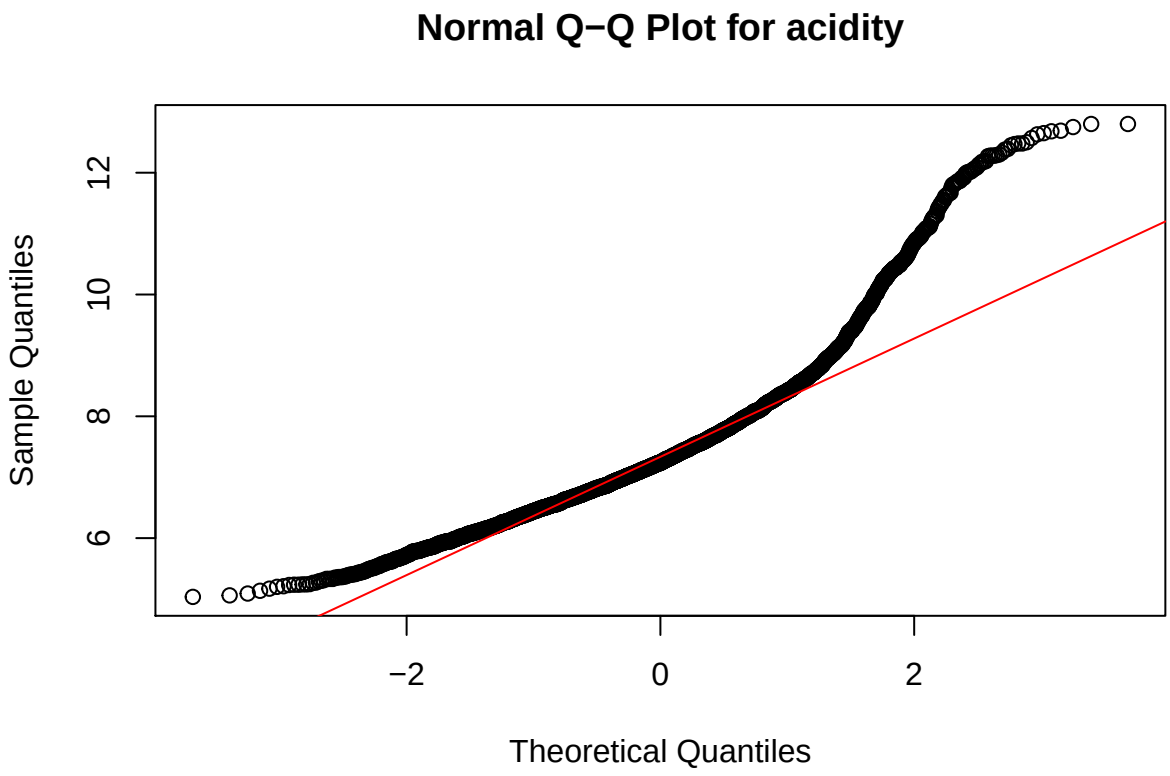
4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Los grupos de datos que vamos a seleccionar para su análisis son la calidad con el tipo de vino, la calidad con el alcohol, el tipo de vino con el alcohol.

```
blancos2 <- blancos[which(blancos$density %in% out_de_blanco),]
blancos3 <- blancos[-which(blancos$density %in% blancos2$density),]
#blancos <- blancos3
```

4.2 Comprobación de la normalidad y homogenidad de la varianza

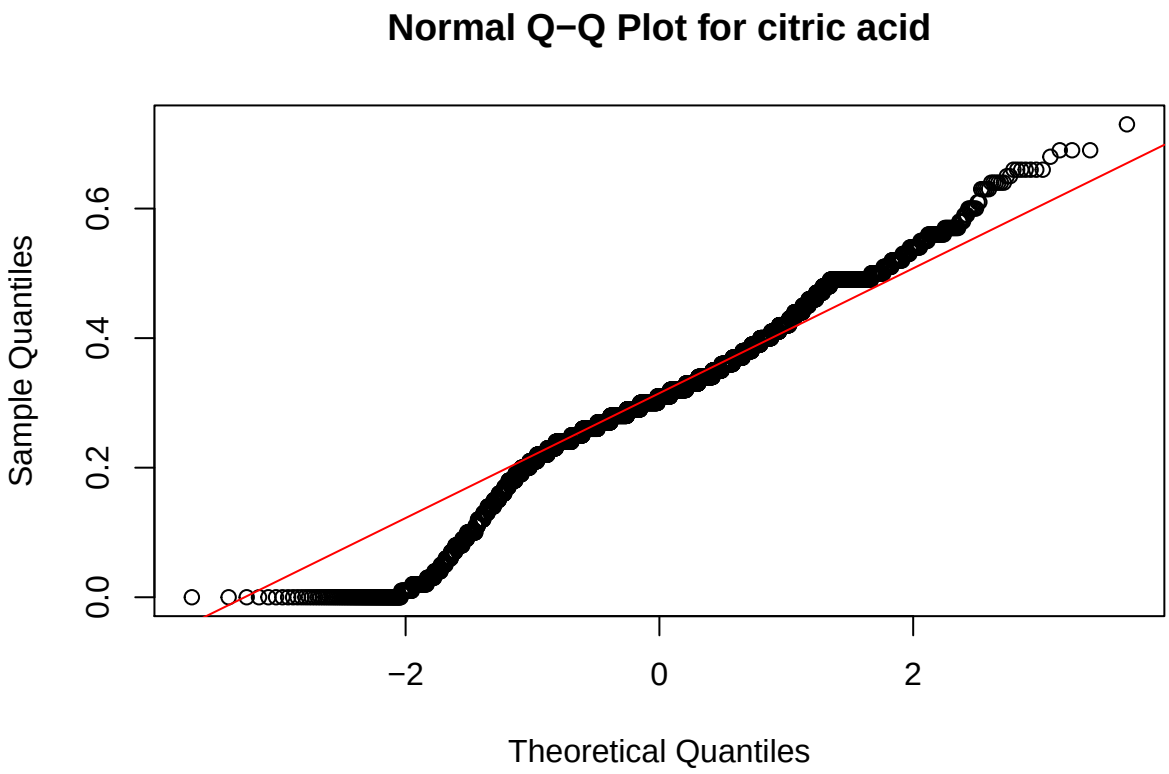
```
library(nortest)

qqnorm(wine_data$acidity, main = "Normal Q-Q Plot for acidity")
qqline(wine_data$acidity, col = "red")
```




```
lillie.test(tintos$acidity)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  tintos$acidity
## D = 0.088243, p-value < 2.2e-16
```

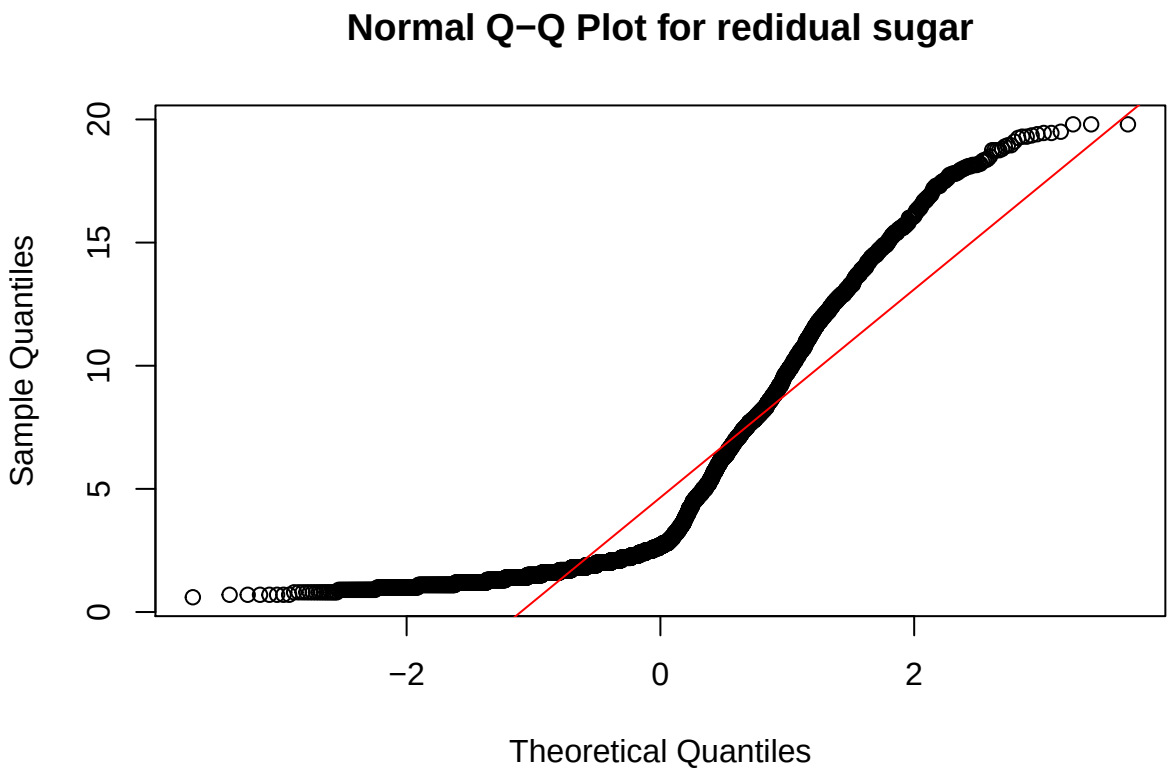
```
qqnorm(wine_data$citric_acid, main = "Normal Q-Q Plot for citric acid")
qqline(wine_data$citric_acid, col = "red")
```




```
lillie.test(wine_data$citric_acid)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wine_data$citric_acid
## D = 0.083197, p-value < 2.2e-16
```

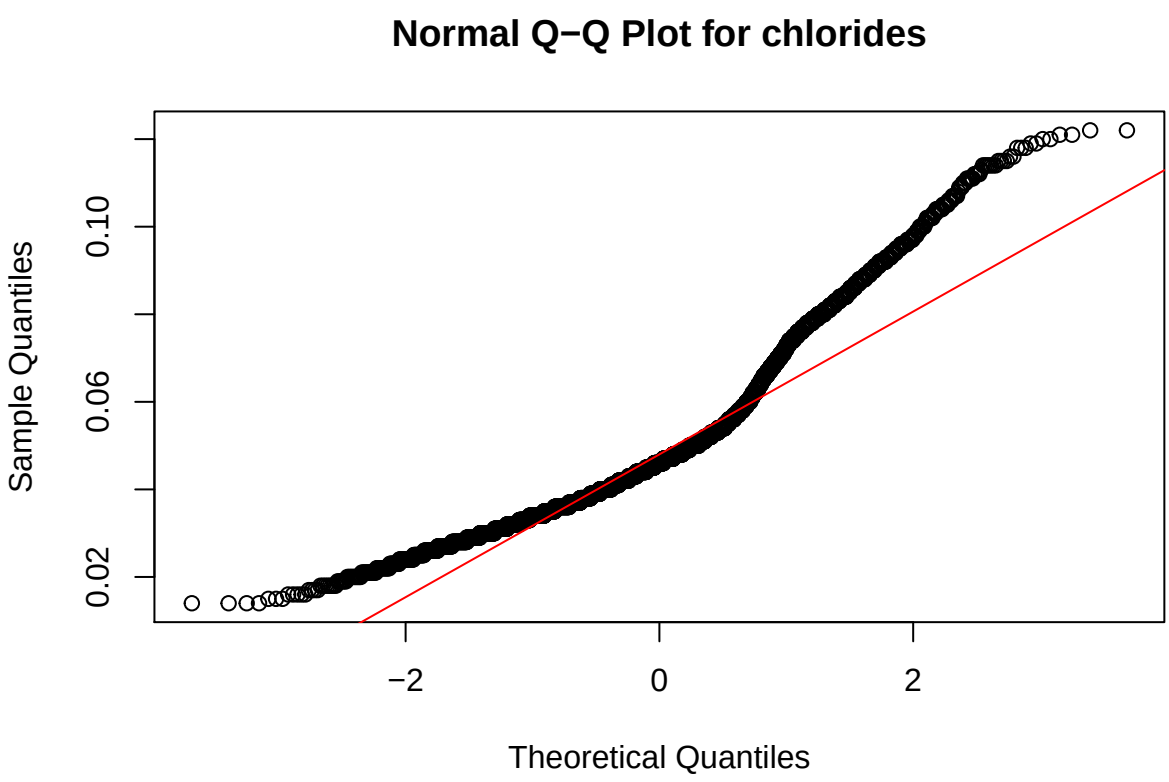
```
qqnorm(wine_data$residual_sugar, main = "Normal Q-Q Plot for redidual sugar")
qqline(wine_data$residual_sugar, col = "red")
```




```
lillie.test(wine_data$residual_sugar)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wine_data$residual_sugar
## D = 0.21633, p-value < 2.2e-16
```

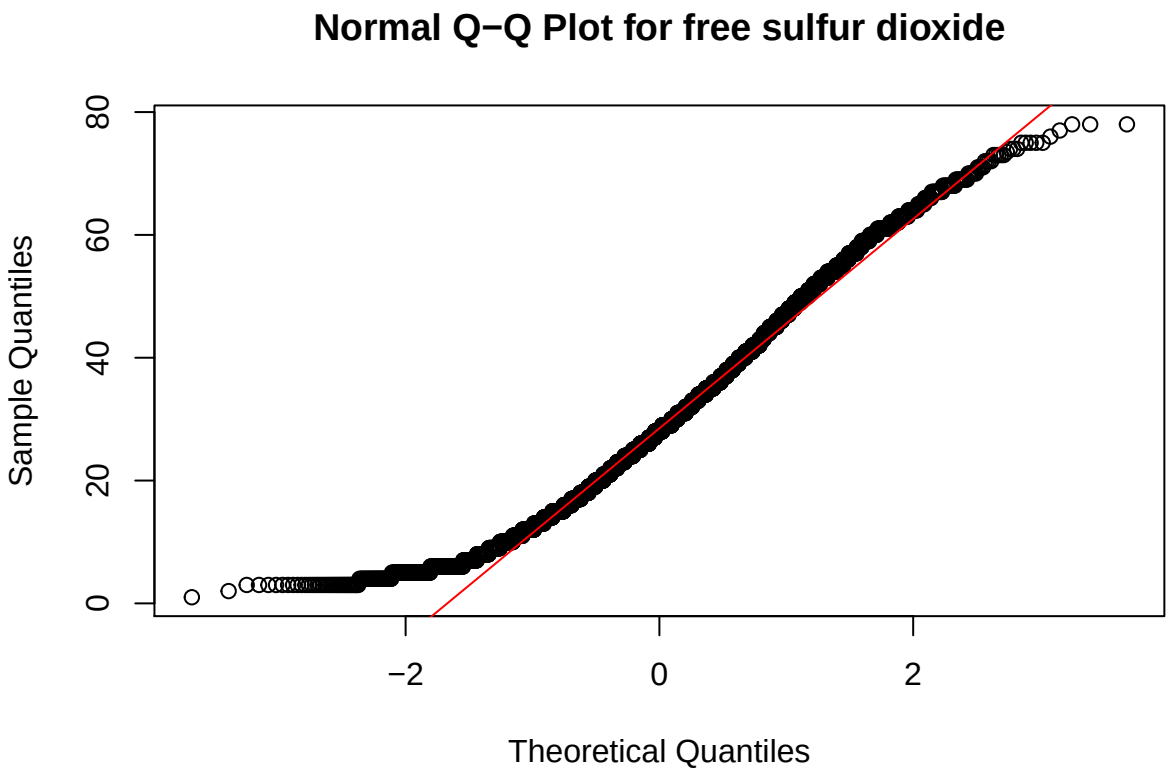
```
qqnorm(wine_data$chlorides, main = "Normal Q-Q Plot for chlorides")
qqline(wine_data$chlorides, col = "red")
```




```
lillie.test(wine_data$chlorides)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wine_data$chlorides
## D = 0.1339, p-value < 2.2e-16
```

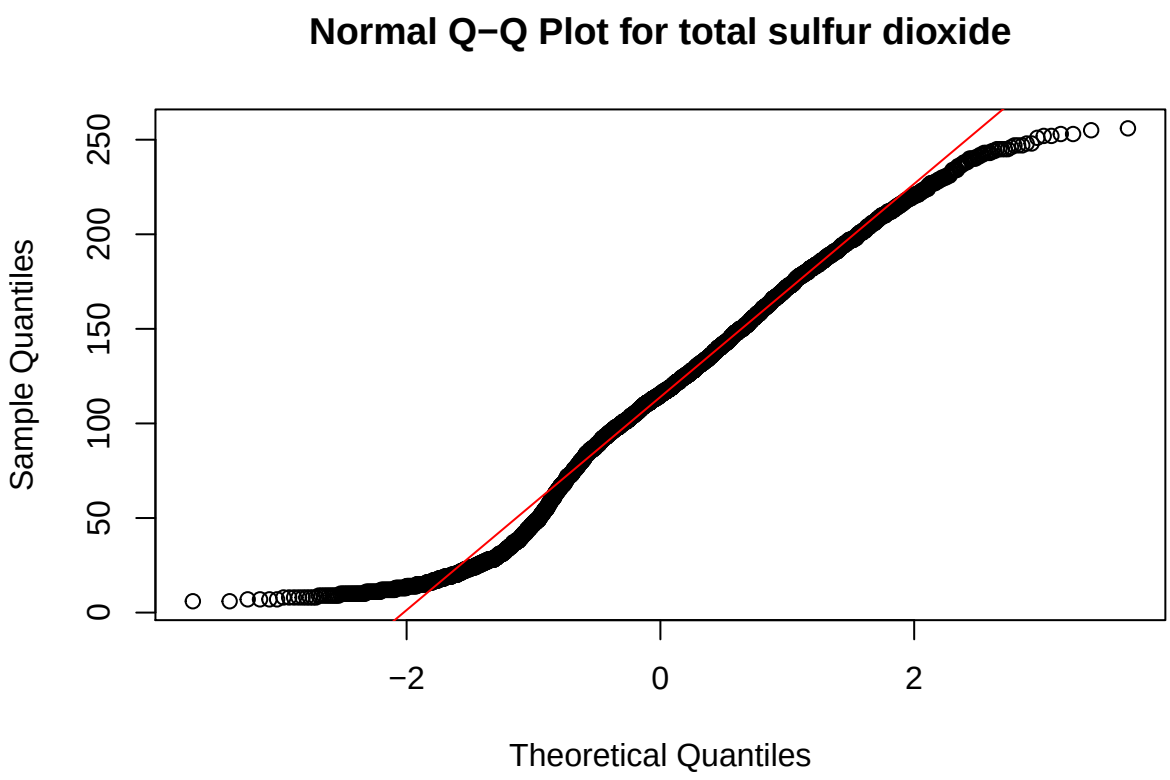
```
qqnorm(wine_data$free_sulfur_dioxide, main = "Normal Q-Q Plot for free sulfur dioxide")
qqline(wine_data$free_sulfur_dioxide, col = "red")
```




```
lillie.test(wine_data$free_sulfur_dioxide)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wine_data$free_sulfur_dioxide
## D = 0.055312, p-value < 2.2e-16
```

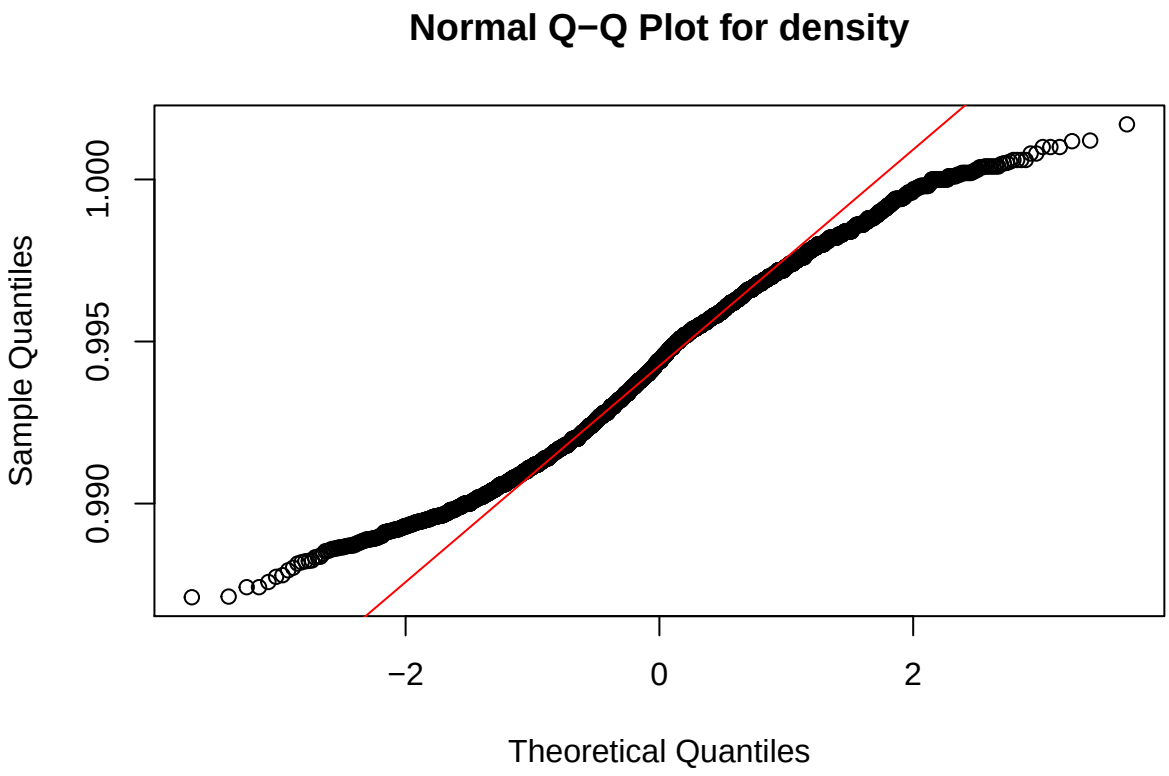
```
qqnorm(wine_data$total_sulfur_dioxide, main = "Normal Q-Q Plot for total sulfur dioxide")
qqline(wine_data$total_sulfur_dioxide, col = "red")
```




```
lillie.test(wine_data$total_sulfur_dioxide)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wine_data$total_sulfur_dioxide
## D = 0.046751, p-value < 2.2e-16
```

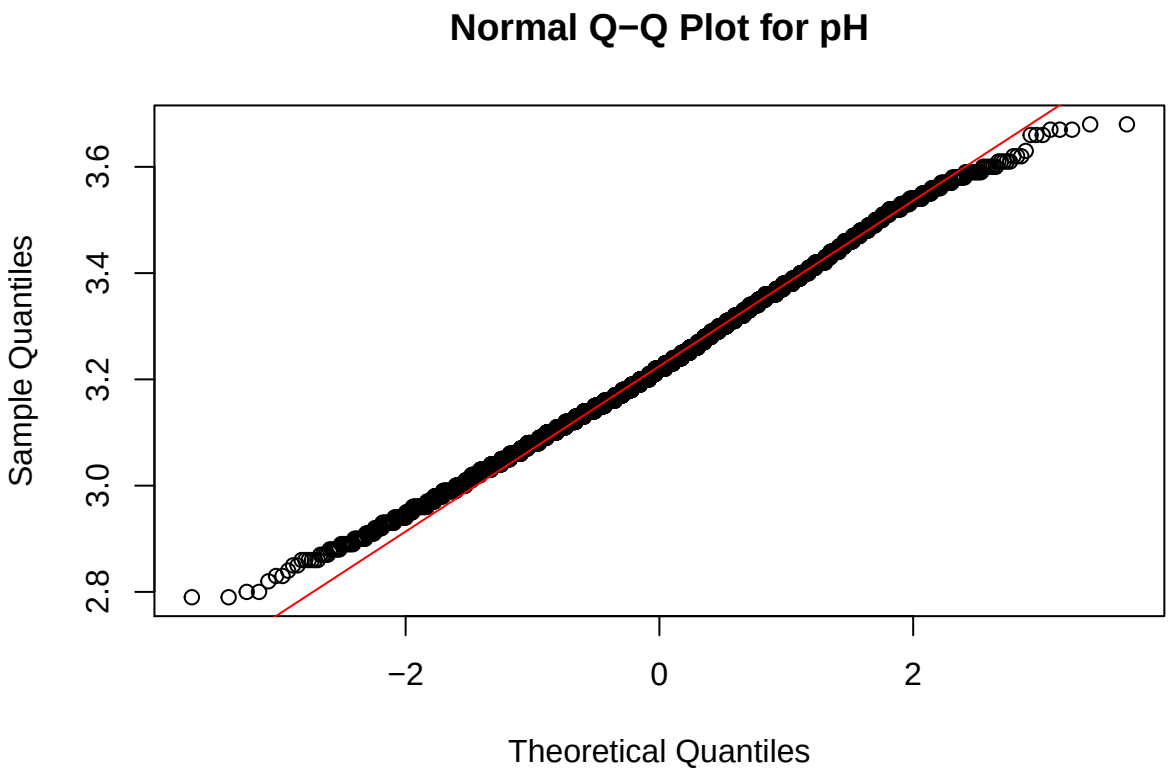
```
qqnorm(wine_data$density, main = "Normal Q-Q Plot for density")
qqline(wine_data$density, col = "red")
```




```
lillie.test(wine_data$density)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wine_data$density
## D = 0.052313, p-value < 2.2e-16
```

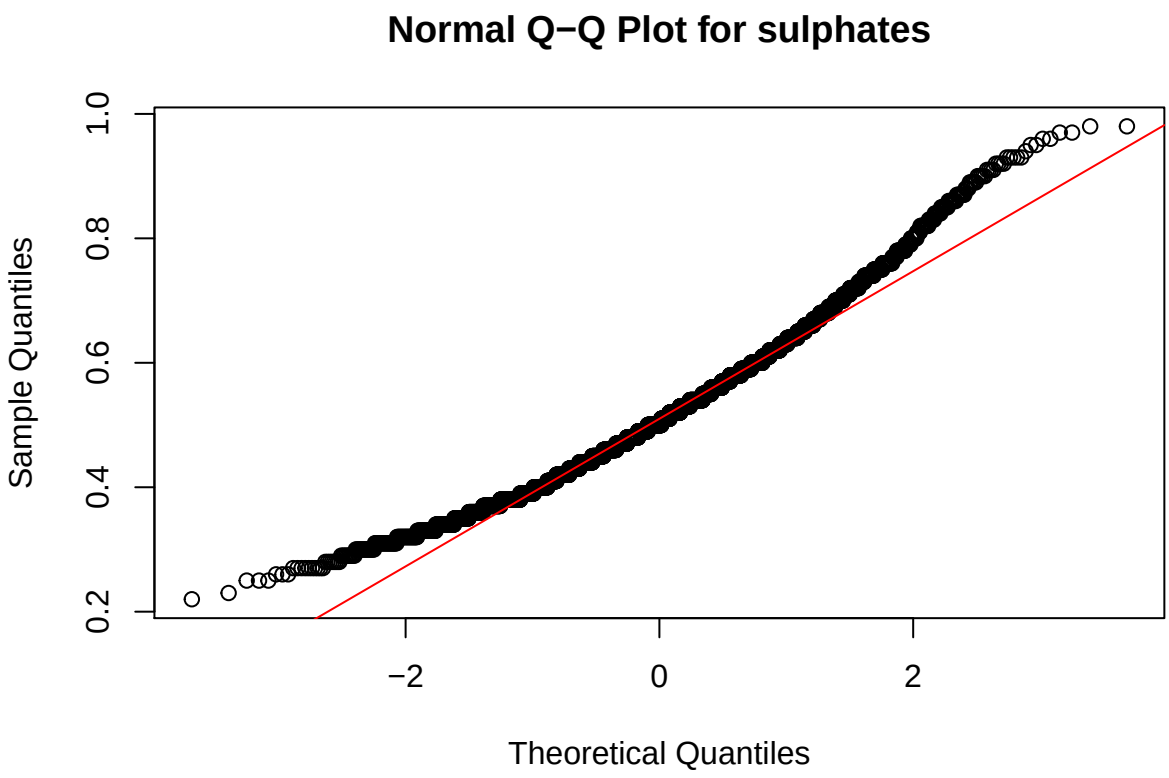
```
qqnorm(wine_data$ph, main = "Normal Q-Q Plot for pH")
qqline(wine_data$ph, col = "red")
```




```
lillie.test(wine_data$ph)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wine_data$ph
## D = 0.038171, p-value = 5.493e-16
```

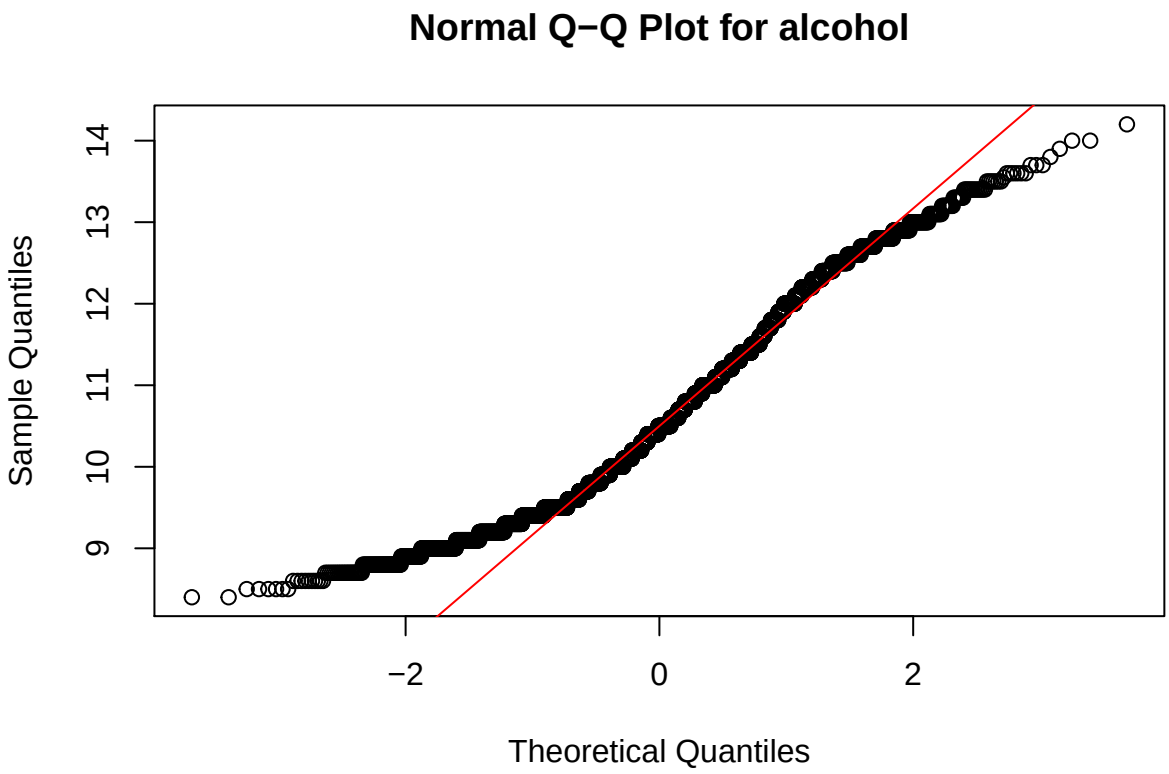
```
qqnorm(wine_data$sulphates, main = "Normal Q-Q Plot for sulphates")
qqline(wine_data$sulphates, col = "red")
```




```
lillie.test(wine_data$sulphates)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wine_data$sulphates
## D = 0.060933, p-value < 2.2e-16
```

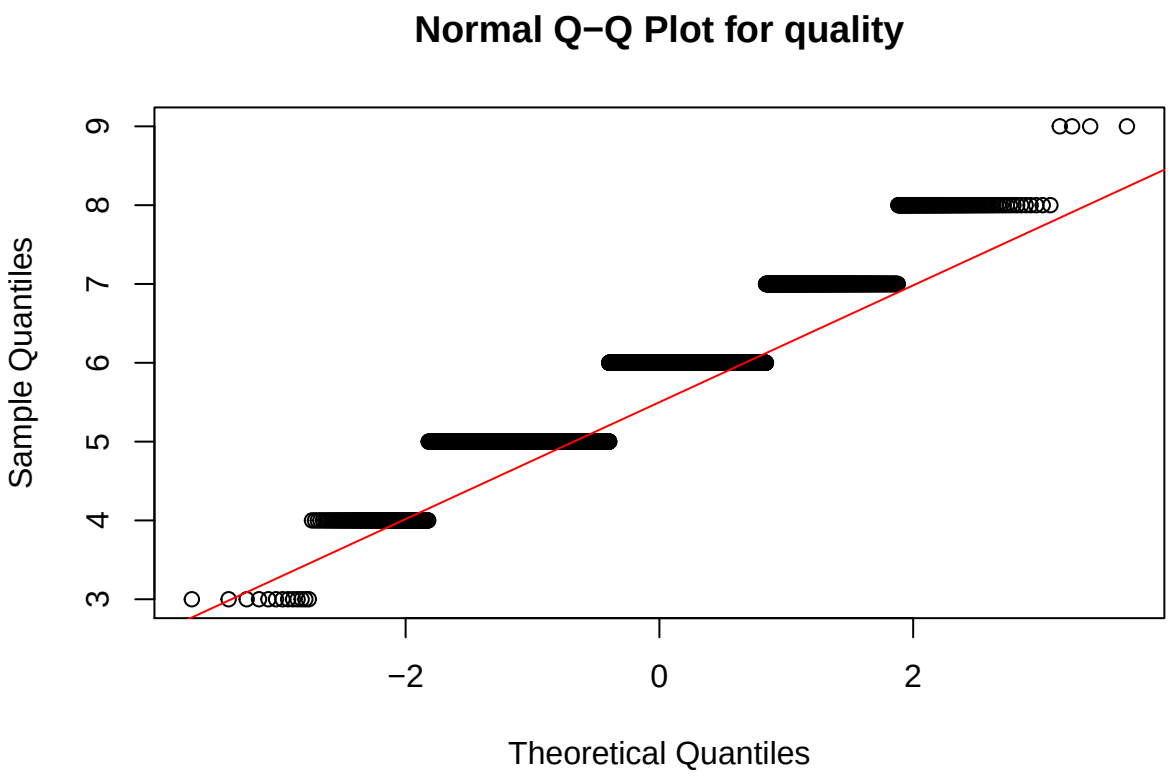
```
qqnorm(wine_data$alcohol, main = "Normal Q-Q Plot for alcohol")
qqline(wine_data$alcohol, col = "red")
```




```
lillie.test(wine_data$alcohol)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wine_data$alcohol
## D = 0.086583, p-value < 2.2e-16
```

```
qqnorm(wine_data$quality, main = "Normal Q-Q Plot for quality")
qqline(wine_data$quality, col = "red")
```




```
lillie.test(wine_data$quality)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wine_data$quality
## D = 0.22918, p-value < 2.2e-16
```

Lo que podemos concluir de los test de normalidad de lillie es que ninguna de las variables está normalizada, ya que los p-values son muy inferiores a 0.05, y por tanto, no podemos rechazar la hipótesis nula y aceptar que existe normalidad.

No obstante sí que es cierto que todas las veriables se pueden normalizar, pues cumplen con el teorema del límite central.

Ahora vamos a estudiar la homogeneidad de varianzas

```
library(psych)

fligner.test(wine_data$quality, wine_data$tipo)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  wine_data$quality and wine_data$tipo
## Fligner-Killeen:med chi-squared = 0.97439, df = 1, p-value = 0.3236
```

```
fligner.test(wine_data$quality, wine_data$alcohol)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  wine_data$quality and wine_data$alcohol
## Fligner-Killeen:med chi-squared = 113.11, df = 101, p-value = 0.193
```

```
fligner.test(wine_data$alcohol, wine_data$tipo)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  wine_data$alcohol and wine_data$tipo
## Fligner-Killeen:med chi-squared = 60.511, df = 1, p-value = 7.317e-15
```

```
fligner.test(wine_data$quality, wine_data$ph)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  wine_data$quality and wine_data$ph
## Fligner-Killeen:med chi-squared = 80.171, df = 86, p-value = 0.6568
```

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. Aplicar al menos tres métodos de análisis diferentes.

(COMPLETAR) Vamos a aplicar las siguientes pruebas estadísticas.

4.3.1 Comparación de dos grupos. Queremos hacer un contraste de hipótesis sobre si los vinos blancos tienen la misma caildad que los vinos tintos.

4.3.2 Comparación de mas de dos grupos (ANOVA). Tomando la calidad del vino como una variable categórica, comprobaremos si algún valor como la acidez del vino son iguales o distitnos

4.3.3 Correlación

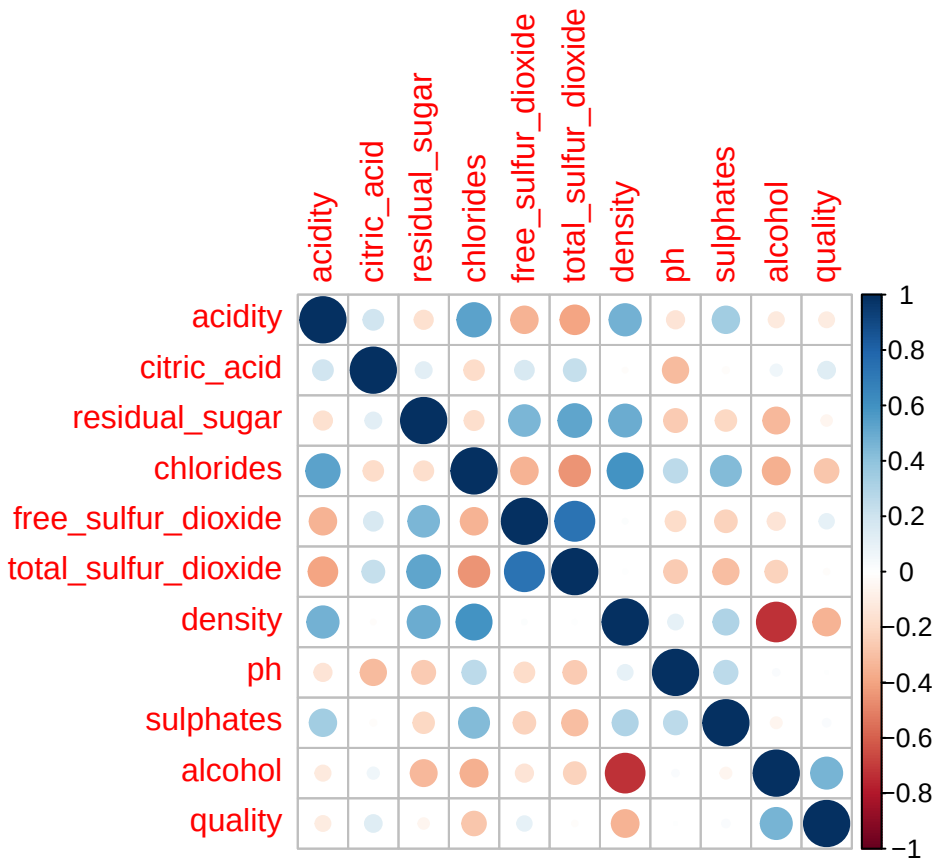
En este punto vamos a visualizar cuál es la correlación entre todas las variables.

```
data_corr <- cor(wine_data[,-12])
data_corr
```

```
##                       acidity citric_acid residual_sugar  chlorides
## acidity             1.0000000  0.19016901    -0.16088793  0.5396525
## citric_acid         0.1901690  1.00000000     0.12335553 -0.1864718
## residual_sugar     -0.1608879  0.12335553     1.00000000 -0.1758152
## chlorides           0.5396525 -0.18647177    -0.17581524  1.0000000
```

```
## free_sulfur_dioxide  -0.3413216  0.16972872     0.45699031 -0.3430248
## total_sulfur_dioxide -0.3976279  0.23228929     0.52154267 -0.4491200
## density               0.4752670 -0.01128960     0.49682281  0.5947832
## ph                   -0.1441042 -0.31156961    -0.25579263  0.2633242
## sulphates             0.3400080 -0.01801474    -0.20535748  0.4373251
## alcohol              -0.1110930  0.06205528    -0.32760827 -0.3516163
## quality              -0.1079237  0.13615920    -0.05417556 -0.2701777
##                     free_sulfur_dioxide total_sulfur_dioxide      density
## acidity                     -0.34132156         -0.397627933  0.475267026
## citric_acid                  0.16972872          0.232289291 -0.011289598
## residual_sugar               0.45699031          0.521542671  0.496822807
## chlorides                   -0.34302484         -0.449120004  0.594783172
## free_sulfur_dioxide          1.00000000          0.738785891  0.011047787
## total_sulfur_dioxide         0.73878589          1.000000000  0.008613818
## density                      0.01104779          0.008613818  1.000000000
## ph                          -0.18886954         -0.252284938  0.107404588
## sulphates                   -0.22837051         -0.309839707  0.305284704
## alcohol                     -0.14428671         -0.229912458 -0.729059550
## quality                      0.10167249         -0.014586628 -0.348582877
##                                   ph   sulphates     alcohol      quality
## acidity              -0.144104222  0.34000796 -0.11109299 -0.107923695
## citric_acid          -0.311569608 -0.01801474  0.06205528  0.136159201
## residual_sugar       -0.255792628 -0.20535748 -0.32760827 -0.054175563
## chlorides             0.263324179  0.43732510 -0.35161632 -0.270177712
## free_sulfur_dioxide  -0.188869538 -0.22837051 -0.14428671  0.101672489
## total_sulfur_dioxide -0.252284938 -0.30983971 -0.22991246 -0.014586628
## density               0.107404588  0.30528470 -0.72905955 -0.348582877
## ph                    1.000000000  0.26008632  0.02136302  0.003319324
## sulphates             0.260086321  1.00000000 -0.05834224  0.028370828
## alcohol               0.021363019 -0.05834224  1.00000000  0.469684041
## quality               0.003319324  0.02837083  0.46968404  1.000000000
```

```
library(corrplot)

corrplot(data_corr)
```



4.3.4 Regresión lineal

```
modelo <- lm(quality ~ acidity + citric_acid + residual_sugar + chlorides + free_sulfur_dioxide + total_

summary(modelo)
```

```
##
## Call:
## lm(formula = quality ~ acidity + citric_acid + residual_sugar +
##     chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
##     alcohol + ph + density + sulphates, data = wine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5244 -0.4623 -0.0331  0.4641  2.8421
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.138e+02  1.806e+01   6.302 3.24e-10 ***
## acidity               9.638e-02  2.127e-02   4.531 6.03e-06 ***
## citric_acid           6.243e-01  1.036e-01   6.028 1.80e-09 ***
## residual_sugar        5.499e-02  7.351e-03   7.481 8.85e-14 ***
## chlorides            -2.779e+00  1.050e+00  -2.647  0.00815 **
## free_sulfur_dioxide   1.097e-02  1.027e-03  10.678  < 2e-16 ***
```

```
## total_sulfur_dioxide -2.430e-03  3.525e-04  -6.893 6.25e-12 ***
## alcohol               2.025e-01  2.320e-02   8.727  < 2e-16 ***
## ph                    7.442e-01  1.173e-01   6.345 2.45e-10 ***
## density              -1.148e+02  1.840e+01  -6.242 4.72e-10 ***
## sulphates             1.172e+00  1.143e-01  10.252  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.725 on 4354 degrees of freedom
## Multiple R-squared:  0.2876, Adjusted R-squared:  0.286
## F-statistic: 175.8 on 10 and 4354 DF,  p-value: < 2.2e-16
```

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

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