Tipología y ciclo de vida de los datos: Práctica 2

Autor: Miguel Tablado

Junio 2022

Contents

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

He decidido trabajar con el dataset del vino porque fue un ejemplo parecido el que, a finales de 2020, en el periodo de vacaciones por Navidad, un caso parecido me ayudó a dar el paso de matricularme en el máster. Por aquel entonces, parte del tiempo de toma de decisión lo dediqué a explorar casos de uso expuestos en internet y recuerdo un caso de la plataforma de Google donde se exponía como precedir el precio de un vino en función del texto de la etiqueta, encontrando los tipos de uva y varias palabras clave. Este caso me recuerda a ese momento, y quiero ver qué similitudes y diferencias podré encontrar con aquel caso de demostración.

En un análisis rápido de, me parecen interesantes poder tomar para la clasificación y predicción la calidad del vino a partir de las diferentes propiedades del mismo. Sin duda, el objetivo del dataset es poder predecir la calidad del vino a partir de sus propiedades, seguramente a partir de regresiones, pero también podrían hacerse agrupaciones y otros métodos como los árboles de decisión.

Podemos ver que es un *dataset* potente con un número elevado de dimensiones y datos suficientes para cumplir con los requisitos de la práctica. Comprobaremos la calidad de los datos aquí, pero en un análisis rápido de kaggle, en la sección "Column" podemos ver que no hay datos perdidos ni erróneos, por tanto, a priori estamos ante un dataset con un conjunto de datos de calidad.

```
# Cargamos las librerías necesarias para el ejercicio.
if (!require('ggplot2')) install.packages('ggplot2')
library('ggplot2')

if (!require('corrplot')) install.packages('corrplot')
library('corrplot')

# Carga del fichero y muestra inicial de campos.
df_red<-read.csv("winequality-red.csv", header=TRUE, sep=",")
str(df_red)
```

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

```
summary(df_red)
```

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

Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir

En la página de Kaggle se cita:

> The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult the reference [Cortez et al., 2009].

Por tanto, he buscado en Kaggle el dataset del vino blanco (https://www.kaggle.com/datasets/piyushagni5/white-wine-quality), que se muestra a continuación:

```
# Carga del fichero y muestra inicial de campos.
df_white<-read.csv("winequality-white.csv", header=TRUE, sep=";")
str(df_white)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
```

```
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

Ambos datasets son extractos de un dataset original que agrupa los registros los diferencia mediante una variable `type` con valores `red` y `white` para los vinos tintos y blancos respectivamente.

En este ejercicico, en lugar de descargar el dataset original, fusionaremos ambos datasets buscando la misma estructura que en el dataset original, con el objetivo de comprobar las tareas de fusión.

Es relevante la importancia de hacer este procedimiento para obtener una información más amplía que nos permita verificar diferencias entre los tipos de vino y cómo sus variables ofrecen o no diferencias de calidad según el tipo de vino, es decir, analizar la relación entre el tipo y el resto de variables y la calidad final.

```
df_red$type <- "red"
df_white$type <- "white"
df = rbind(df_red, df_white)
df$type <- as.factor(df$type)
summary(df)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923
##  Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949
##  Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970
##  Max.   :0.61100   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality         type
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000   red  :1599
##  1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000   white:4898
##  Median :3.210   Median :0.5100   Median :10.30   Median :6.000
##  Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
##  3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
```

En este caso, es obvio que, en lugar de seleccionar un subconjunto de vinos, lo que hemos hecho es fusionar porque nuestro objetivo es determinar la calidad de cualquier vino. Una vez fusionados vinos blancos y vinos tintos, podríamos seleccionar los vinos de un determinado rango de calidad, por ejemplo, los vinos excelentes. Sin embargo, se entiende que dejar fuera un conjunto de vinos nos llevaría a conclusiones que podrían no ser verdad.

Antes de continuar, lo primero que haremos será estandarizar el nombre de las columnas, sustituyendo los

puntos por guiones bajos, para facilitar el trabajo posterior con estas variables.

```
colnames(df) <- c("fixed_acidity", "volatile_acidity", "citric_acid",
                  "residual_sugar", "chlorides", "free_sulfur_dioxide",
                  "total_sulfur_dioxide", "density", "pH", "sulphates",
                  "alcohol", "quality", "type")
```

Limpieza de los datos

¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

Estadísticas de valores vacíos:

```
colSums(is.na(df))
```

```
##        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
##                 type
##                    0
```

```
colSums(df=="")
```

```
##        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
##                 type
##                    0
```

Podemos apreciar que no hay ningún dato nulo o vacío con lo que estamos ante un conjunto de datos de muy buena calidad.

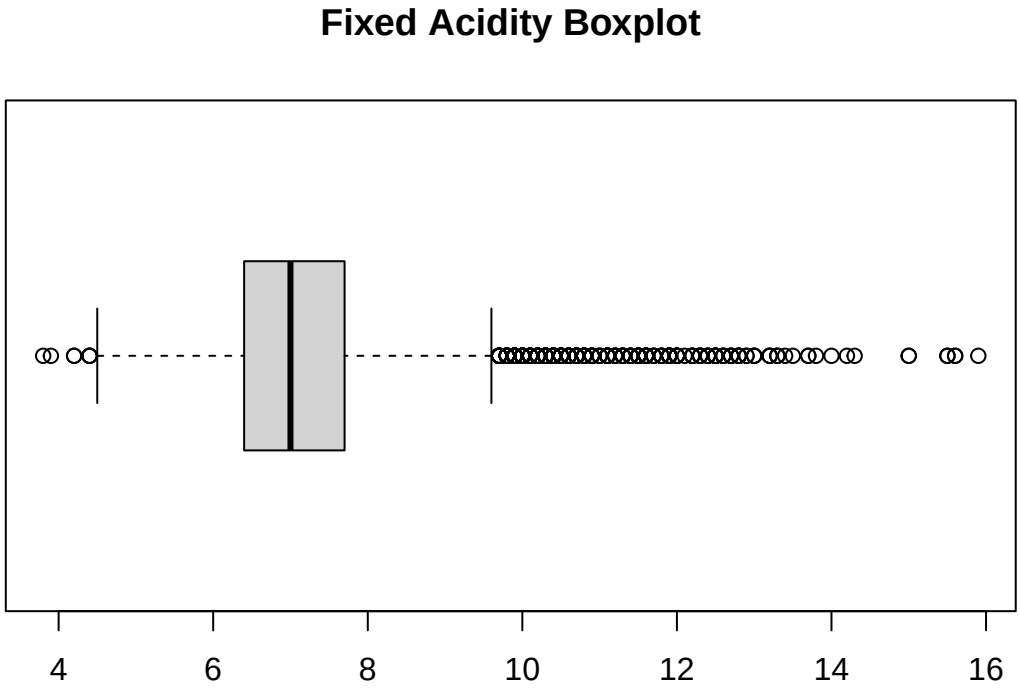
Identifica y gestiona los valores extremos

Buscando información para poder interpretar los diferentes *outliers*, he encontrado un artículo interesante que habla de los diferentes tipos de cítricos que pueden existir en un vino y cómo las diferentes condiciones de cultivo y tipo de uva influyen en los valores de los mismos.

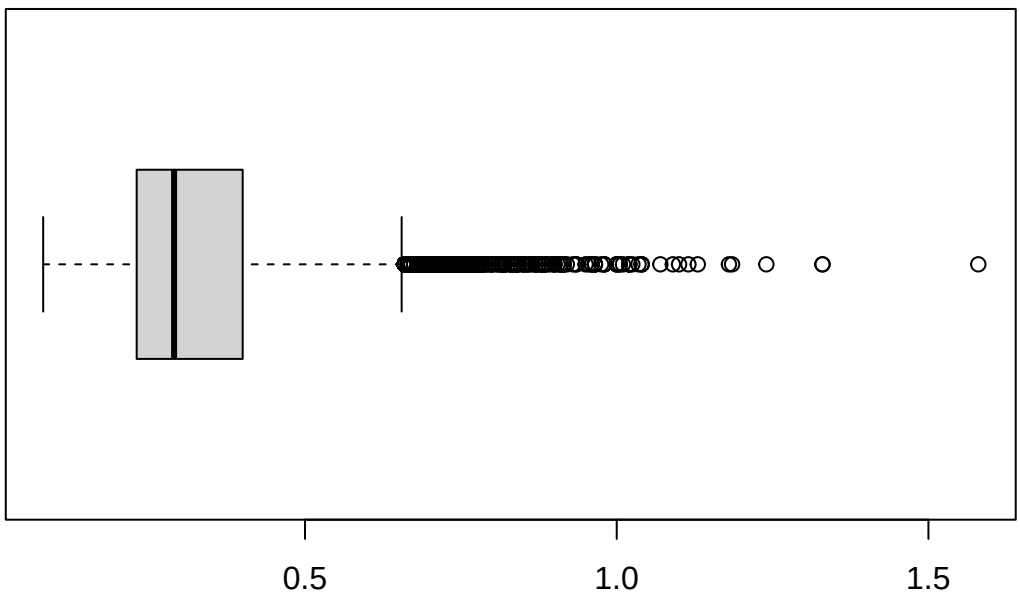
Es por eso, que no soy capaz de determinar si estos valores deben ser corregidos (ningún valor digamos químico). Por tanto, en lugar de corregir los valores, simplemente mostraré un diagrama de cajas para cada uno de ellos y un histograma en aquellos donde la interpretación de los cuartiles nos lleve a pensar que no hay una distribución normal.

Ref: https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity

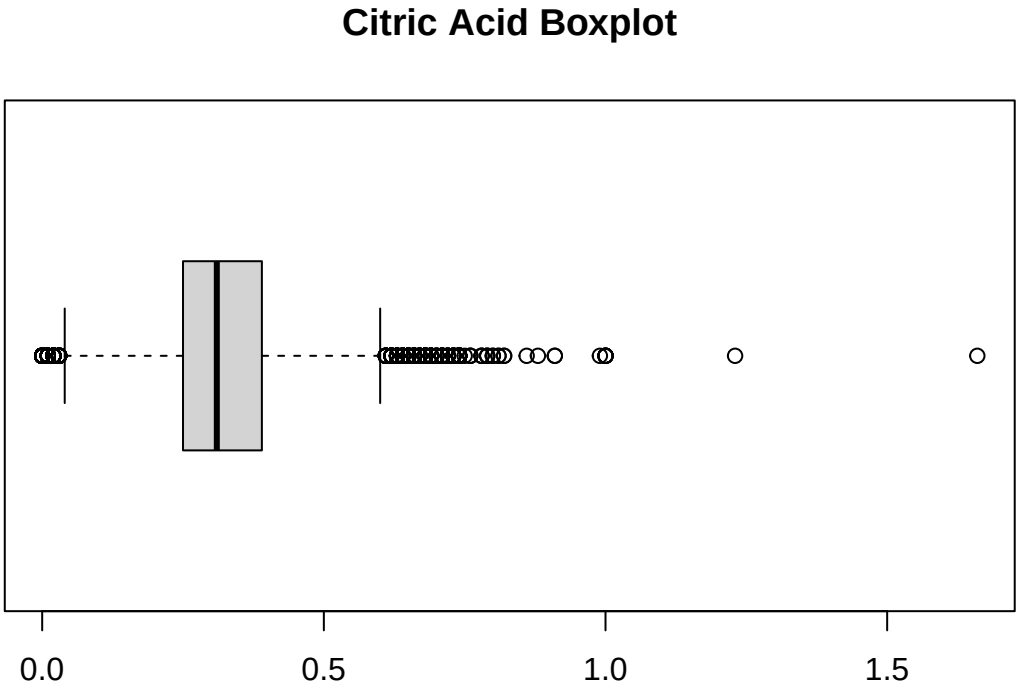
```
boxplot(df$fixed_acidity, main="Fixed Acidity Boxplot", horizontal = TRUE)
```



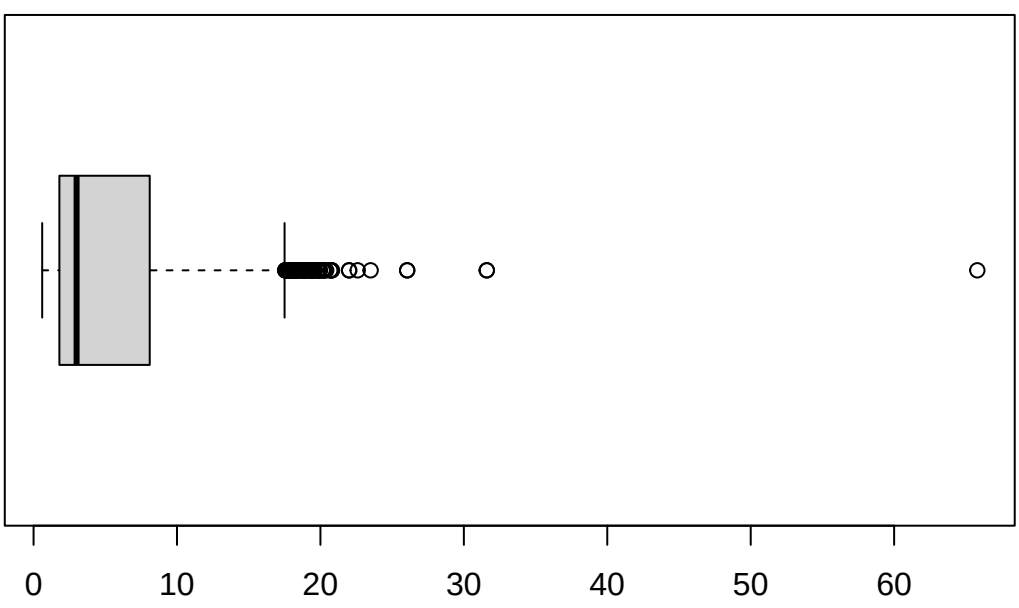

```
boxplot(df$volatile_acidity, main="Volatile Acidity Boxplot", horizontal = TRUE)
```



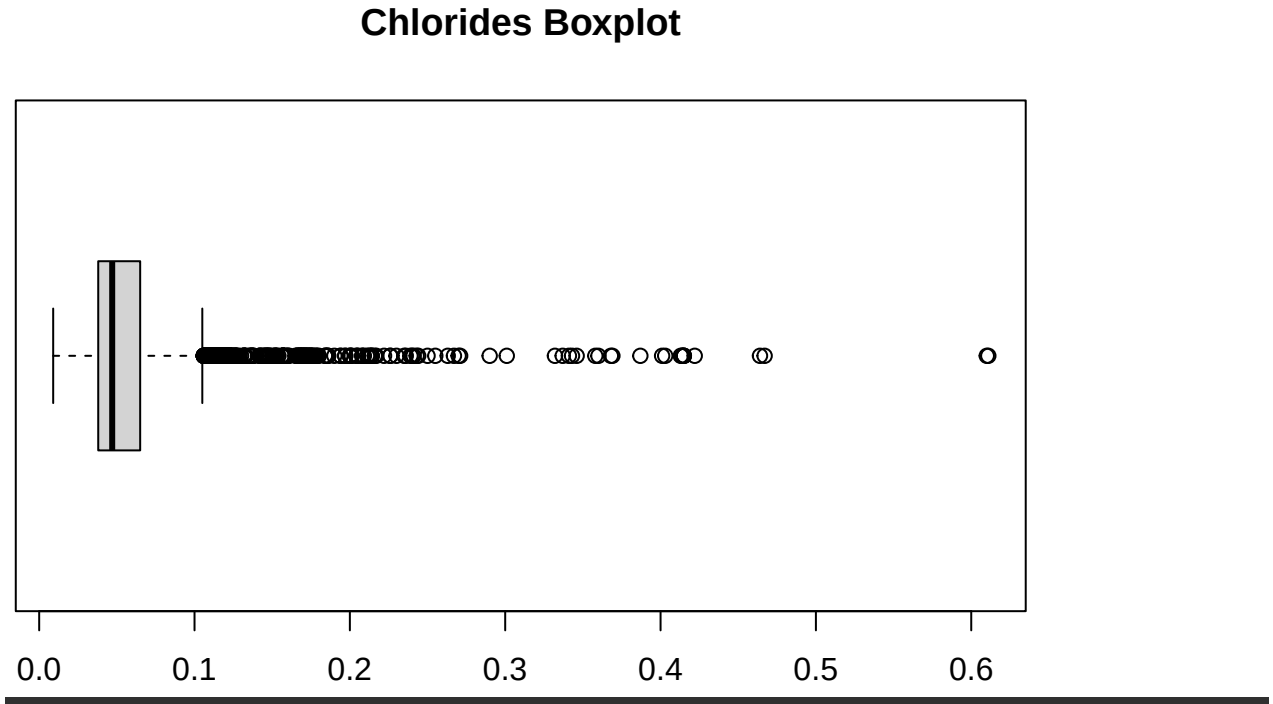

```
boxplot(df$citric_acid, main="Citric Acid Boxplot", horizontal = TRUE)
```



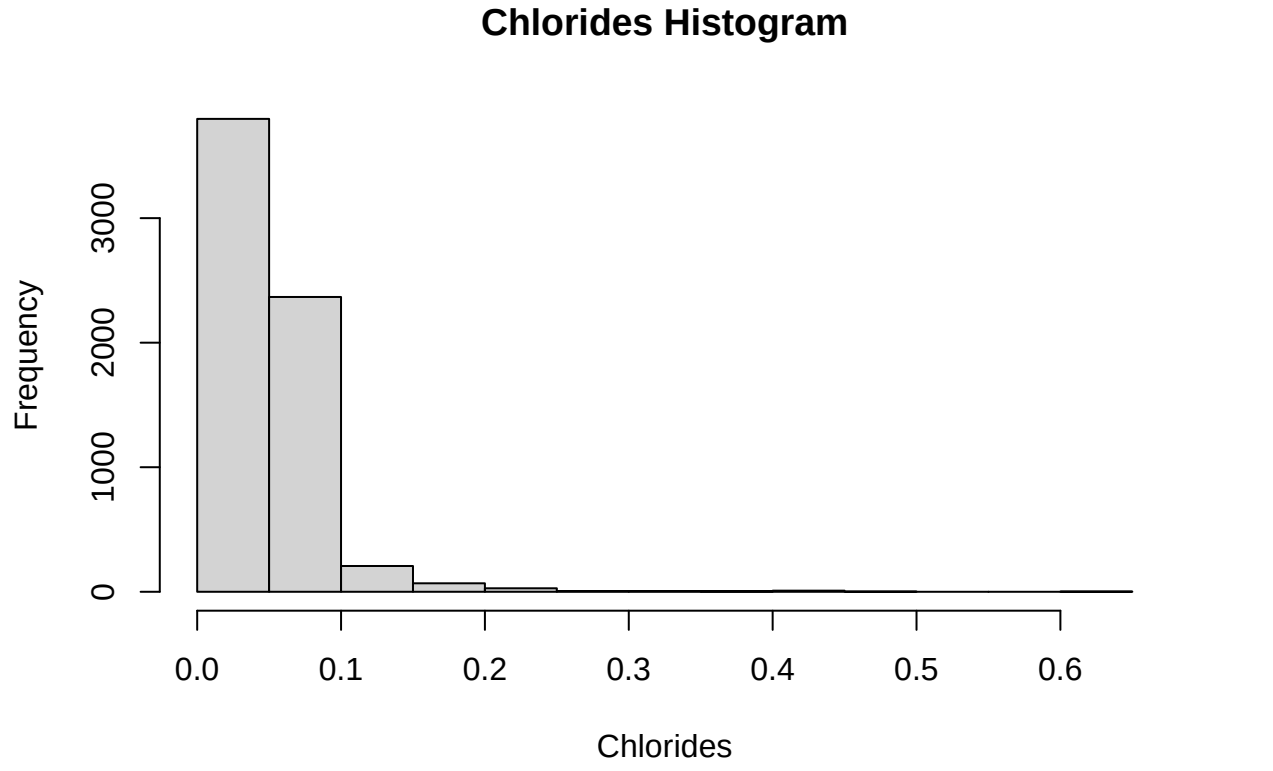

```
boxplot(df$residual_sugar, main="Residual Sugar Boxplot", horizontal = TRUE)
```



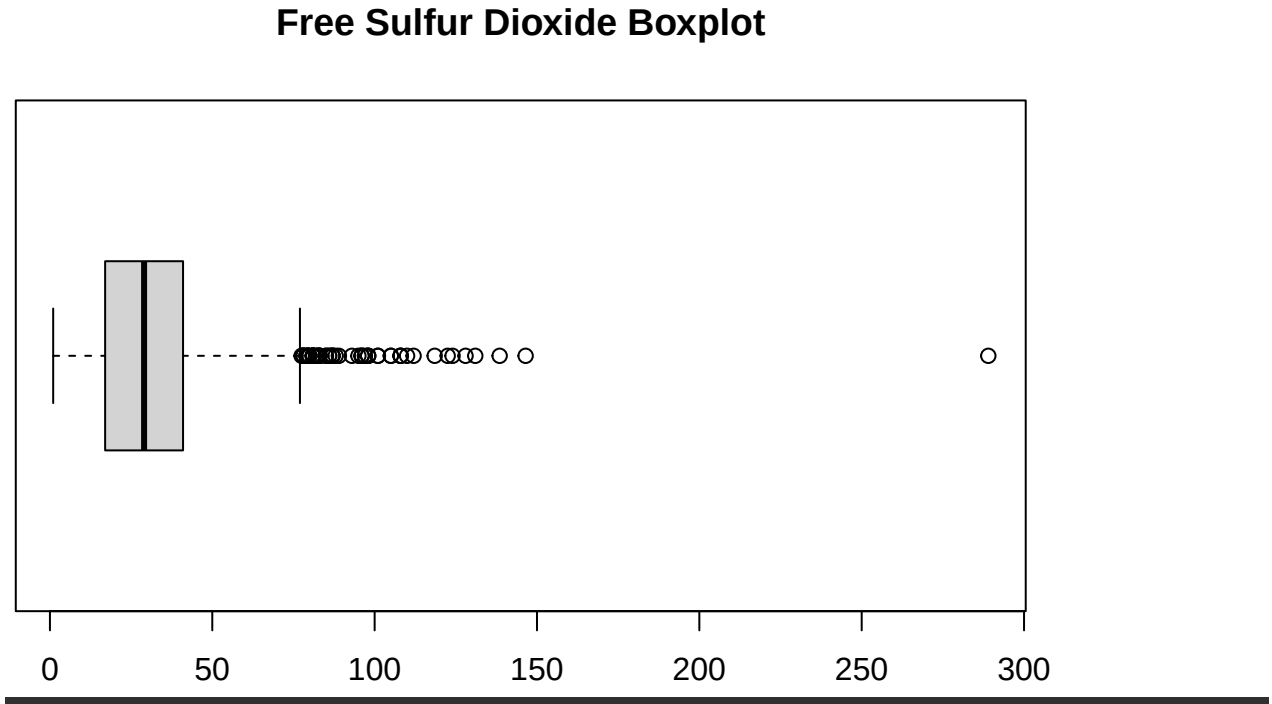

```
boxplot(df$chlorides, main="Chlorides Boxplot", horizontal = TRUE)
```



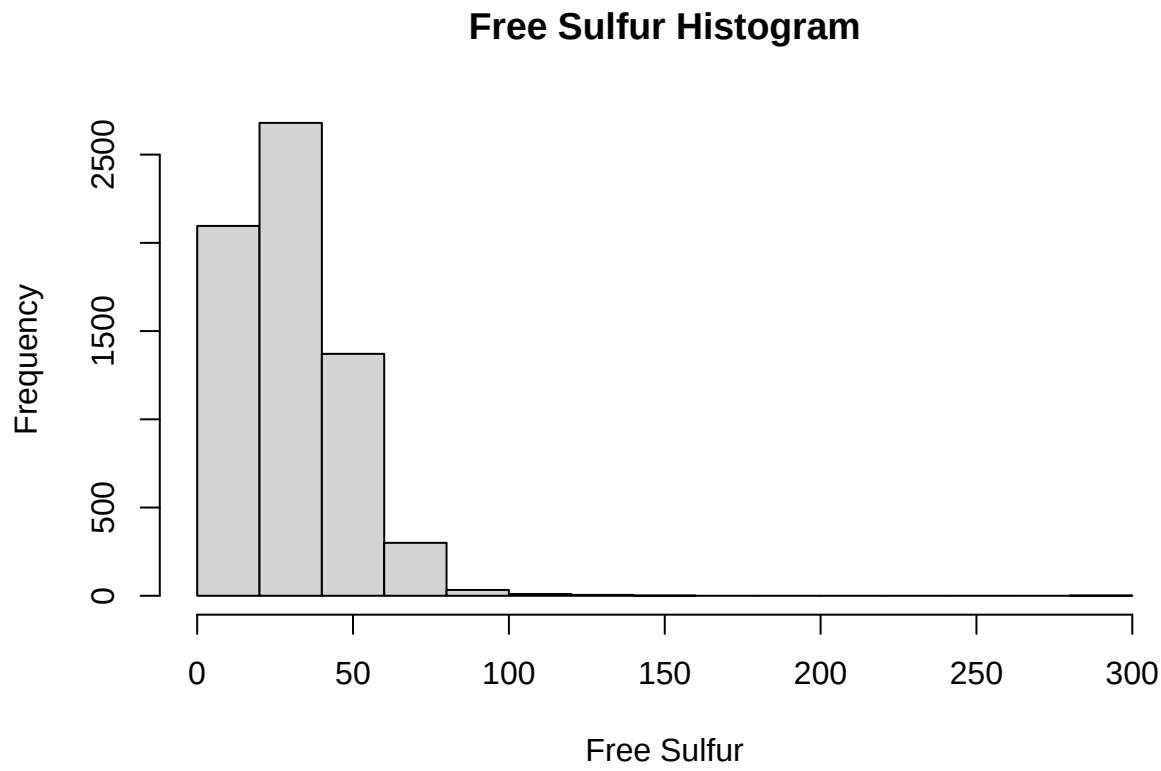

```
hist(df$chlorides, xlab = "Chlorides", main = "Chlorides Histogram")
```



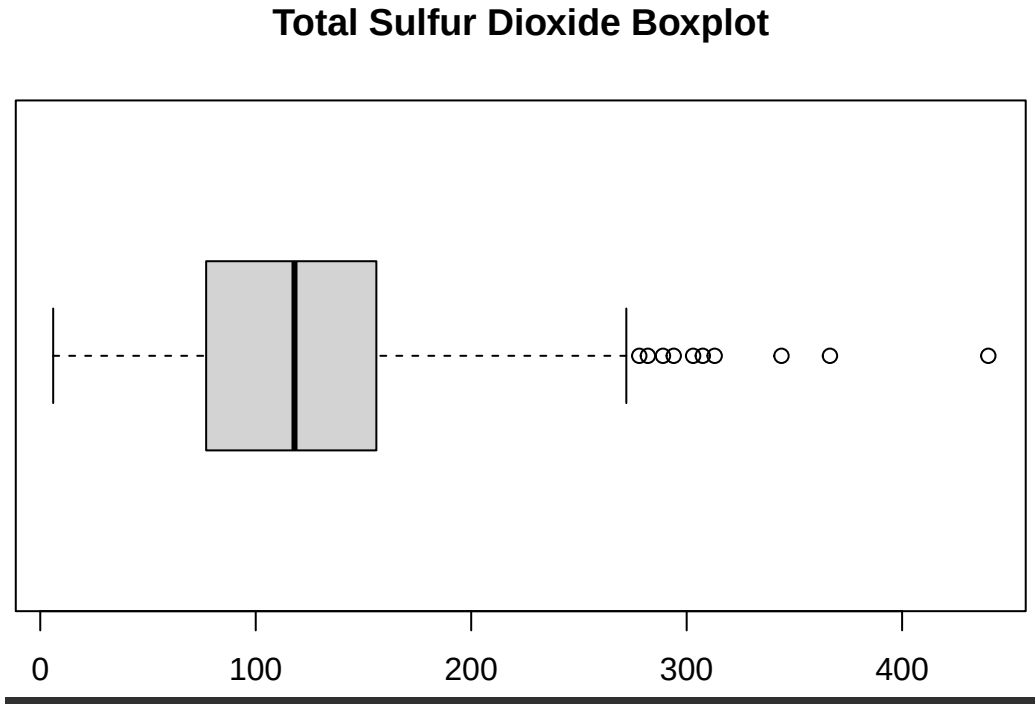

```
boxplot(df$free_sulfur_dioxide, main="Free Sulfur Dioxide Boxplot", horizontal = TRUE)
```



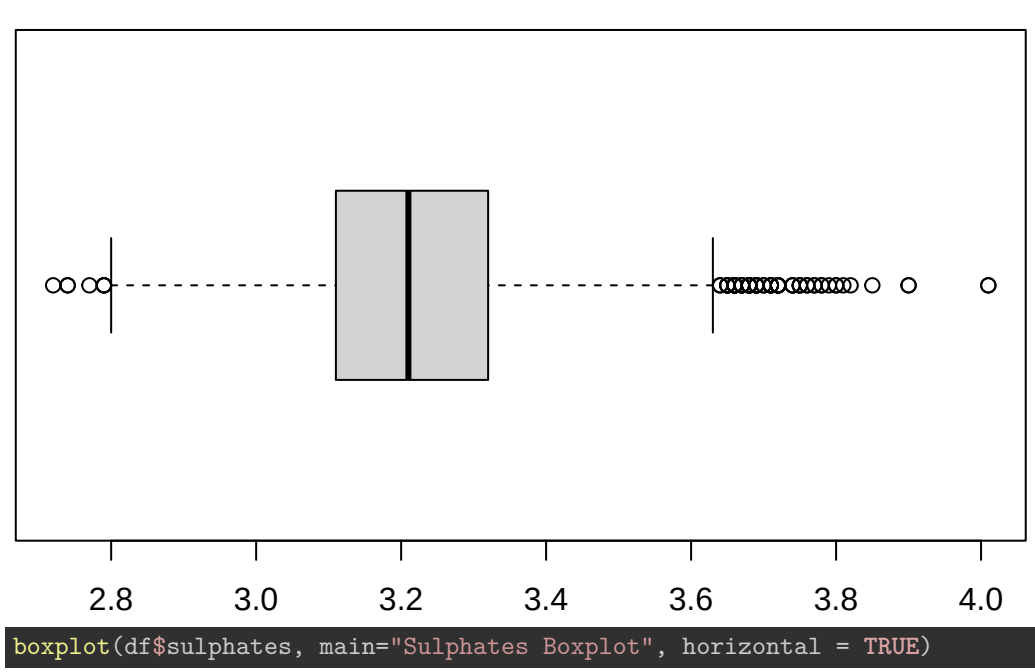

```
hist(df$free_sulfur_dioxide, xlab = "Free Sulfur", main = "Free Sulfur Histogram")
```



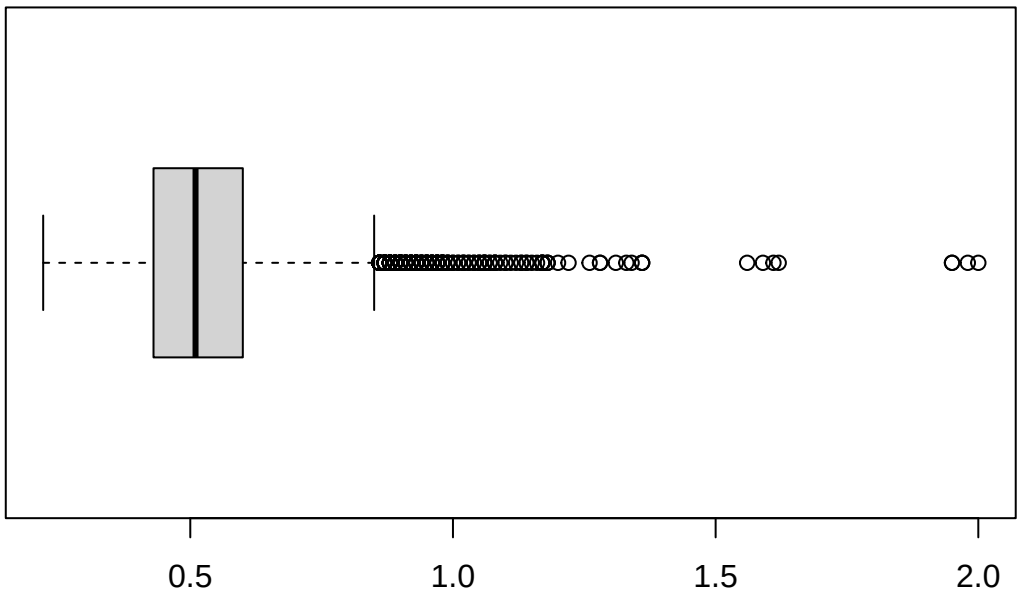

```
boxplot(df$total_sulfur_dioxide, main="Total Sulfur Dioxide Boxplot", horizontal = TRUE)
```



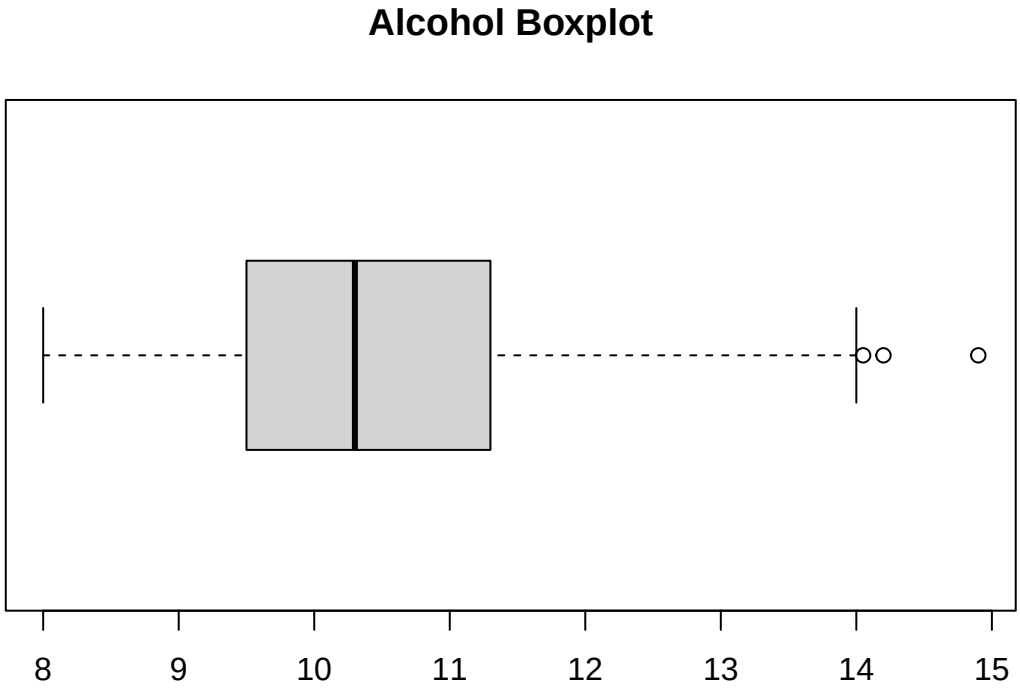

```r
boxplot(df$pH, main="PH Boxplot", horizontal = TRUE)
```




```r
boxplot(df$sulphates, main="Sulphates Boxplot", horizontal = TRUE)
```

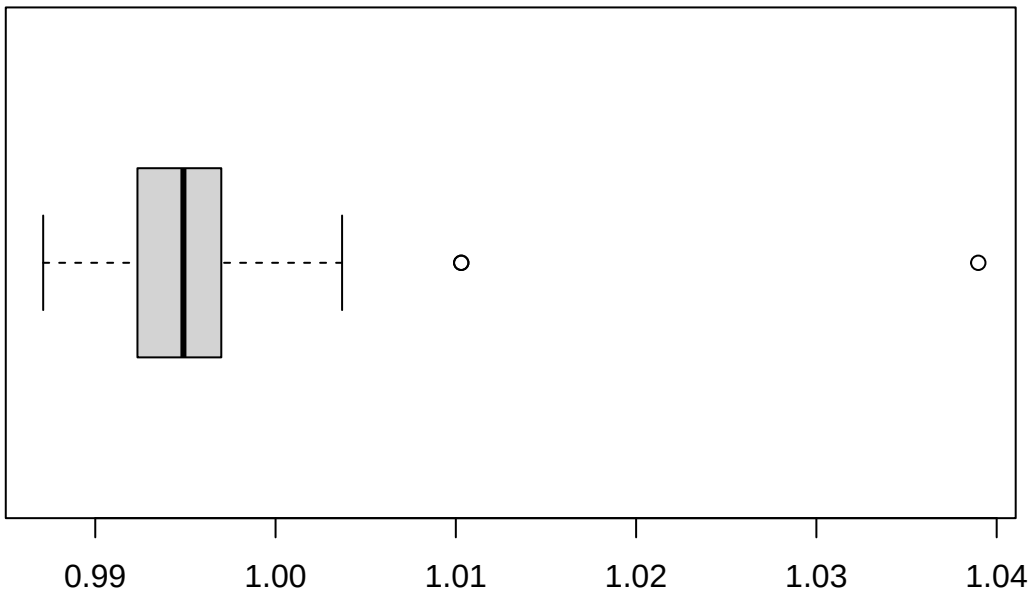



```
boxplot(df$alcohol, main="Alcohol Boxplot", horizontal = TRUE)
```



En el caso del Alcohol, sí podemos saber que hay vinos que pueden dar porcentajes como los valores que tenemos en el dataset, por lo que no corregimos estos datos.

```
boxplot(df$density, main="Density Boxplot", horizontal = TRUE)
```

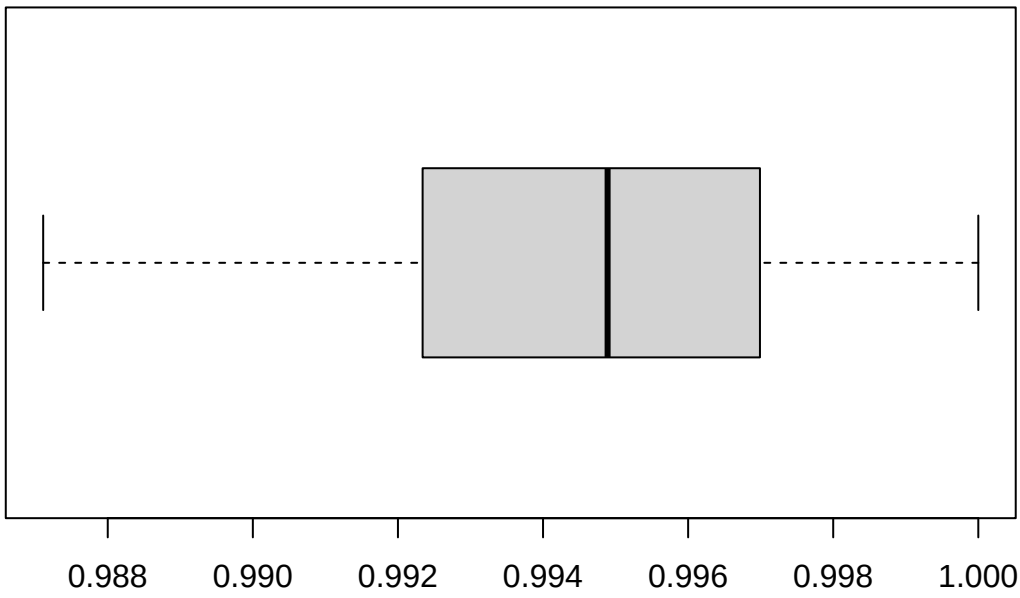


La densidad del vino debe ser un poco inferior a 1, que es la densidad del agua. Aunque el zumo de uva es claramente superior en densidad al agua, el proceso de fermentación lo reduce, debiendo este ser algo inferior a 1. Por tanto, en este caso, corregiremos los valores de densidad superior a 1, marcando la media como valor del vino. Aunque esos valores pudieran ser válidos por un mal proceso de fermentación, claro está, haremos la conversión igualmente.

```
df$density[df$density > 1] <- 1
summary(df$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9923  0.9949  0.9947  0.9970  1.0000
```

```
boxplot(df$density, main="Density Boxplot", horizontal = TRUE)
```



Análisis de los datos.

Selección de los grupos de datos que se quieren analizar/comparar (p. e., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)

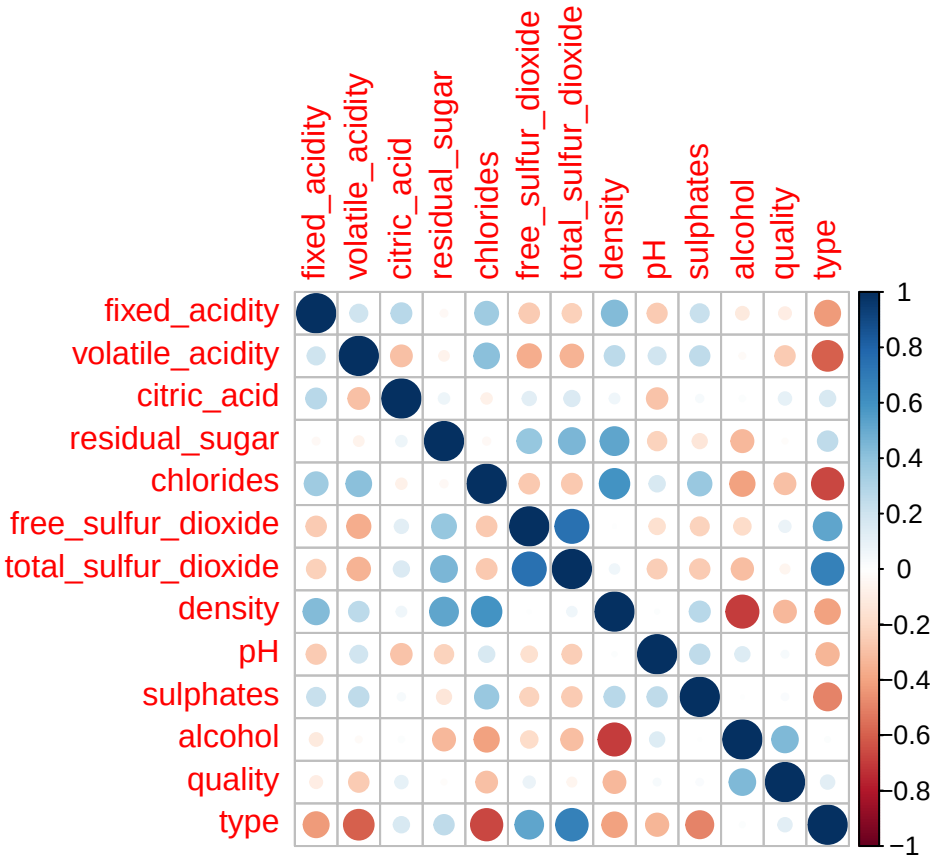
En primer lugar, dado el número de variables que tiene nuestro conjunto de datos, vamos a realizar un análisis de la correlación entre estas variables, a ver si podemos determinar que alguna sea explicable a partir de otra.

Correlaciones con mapas de calor Trabajaremos con un nuevo data frame para utilizar las funciones cor y corrplot que nos permitirán mostrar un mapa de calor de las correlaciones existentes en los datos. Este df necesitará tener los datos del tipo de vino en modo numérico, con lo que deberemos transformarlo y comprobar sus valores.

```
cor_df <- data.frame(df)
cor_df$type <- as.numeric(cor_df$type)
```

Hemos comprobado que la función *as.numeric* ha asignado los valores 1 y 2 para los vinos tintos y blancos respectivamente. El gráfico de barras es en este caso mejor que un resumen de los datos que nos mostraría valores estadísticos que no nos dejarían verificar que solamente existiesen valores discretos.

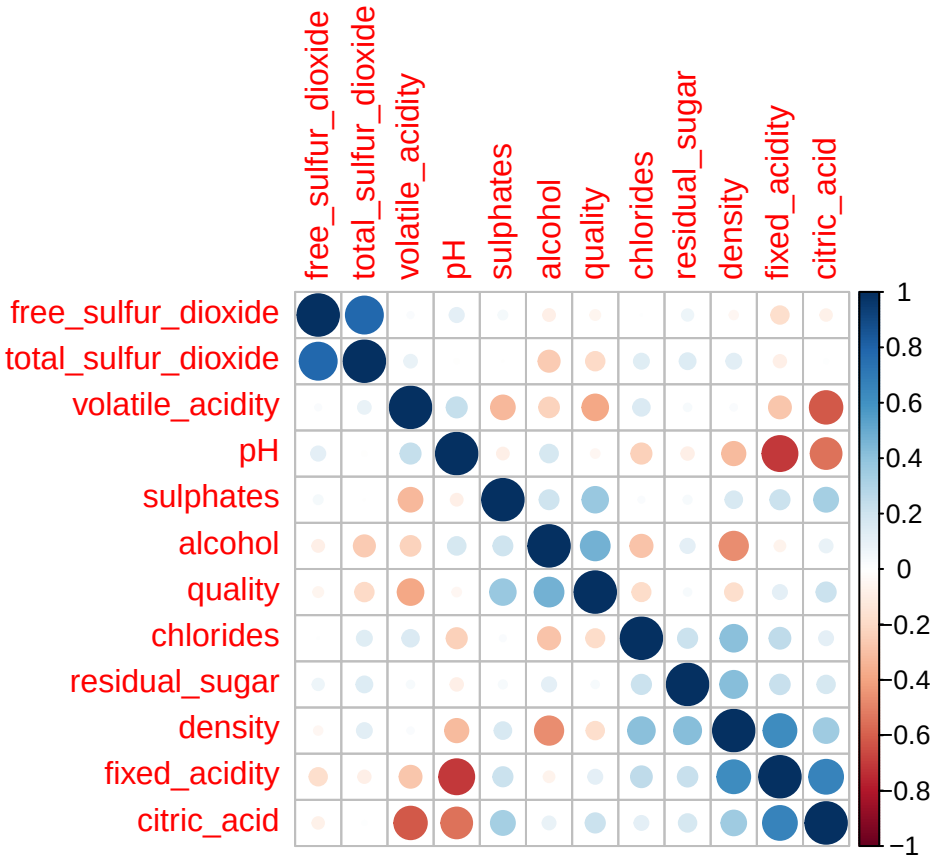
```
cc = cor(cor_df, method = "spearman")
corrplot(cc)
```



En el gráfico podemos ver relaciones fuertes entre múltiples valores como por ejemplo entre la densidad y el alcohol, pero, es seguramente más relevante encontrar las relaciones para los tipos de vino. Empecemos con el vino tinto.

```
dfw <- data.frame(
  cor_df[cor_df$type==1, ])
dfw <- subset(dfw, select = -c(type))

cc = cor(dfw, method = "spearman")
corrplot(cc, order = "hclust", hclust.method = "average")
```

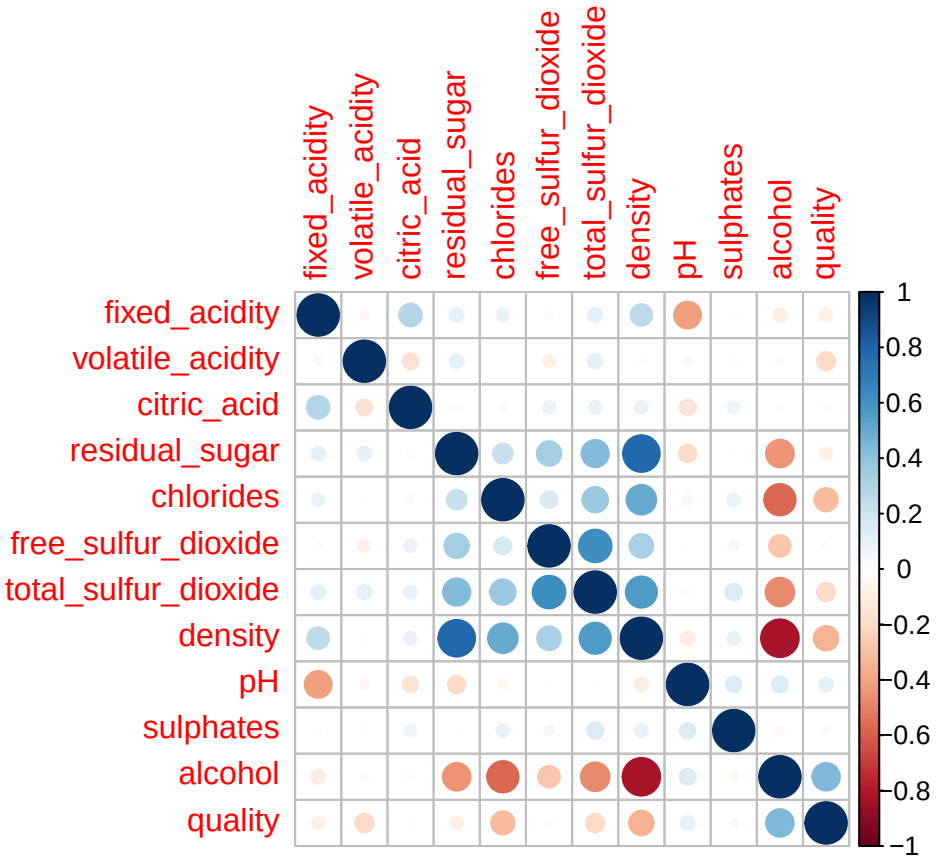


Las relaciones encontradas en el mapa de calor para el **vino tinto** son:

- Ácido y pH, a medida que uno aumenta, el otro disminuye dado que tienen correlaciones negativas. En concreto, el pH y el ácido fijo tienen una relación muy alta y permiten una predicción perfecta el uno en base al otro
- Ácido cítrico y ácido fijo, están relacionados positivamente, lo que significa que cuando aumenta uno, también lo hace el otro y tienen una correlación bastante alta, permitiendo un nivel de predictibilidad interesante.
- Los valores que más influyen en la calidad son el ácido, los sulfatos y el alcohol, donde el primero la reduce, los otros la mejoran.

```
dfw <- data.frame(
  cor_df[cor_df$type==2, ])
dfw <- subset(dfw, select = -c(type))

cc = cor(dfw, method = "spearman")
corrplot(cc)
```



Las relaciones encontradas en el mapa de calor para el **vino blanco** son:

- Densidad y alcohol: correlación negativa y con muy buenas opciones de predicción.
- Azúcar y densidad: correlación positiva y valor cercano a 1
- Azúcar y alcohol: correlación negativa, como es lógico por la regla de tres que se puede establecer dadas las relaciones anteriores.
- También se puede ver que, en cuanto al vino blanco se refiere, a mayor alcohol mayor calidad y que este es el elemento más destacado en cuanto a la calidad se refiere

Dado que los datos a analizar varían claramente entre los vinos tintos y los vinos blancos, para este ejercicio, utilizaremos los datos de los vinos tintos (dado que son los datos originales del ejercicio):

- Ácido Fijo
- Ácido Cítrico
- Alcohol
- Calidad del vino

Comprobación de la normalidad y homogeneidad de la varianza.

Anteriormente hemos analizado con *boxplots* los valores *outliers* y hemos podido ver cómo algunos de los valores no se repartían de forma normal, ya que la línea divisoria dentro de la caja, se ve claramente decantada hacia algún cuartil. Adicionalmente, utilizaremos métodos paramétricos para verificar la normalidad de los atributos seleccionados. Para ello utilizarmos la prueba de Shapiro-Wilk.

Comprobación de normalidad

```
shapiro.test(df_red$fixed.acidity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_red$fixed.acidity
## W = 0.94203, p-value < 2.2e-16
```

```
shapiro.test(df_red$volatile.acidity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_red$volatile.acidity
## W = 0.97434, p-value = 2.693e-16
```

```
shapiro.test(df_red$alcohol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_red$alcohol
## W = 0.92884, p-value < 2.2e-16
```

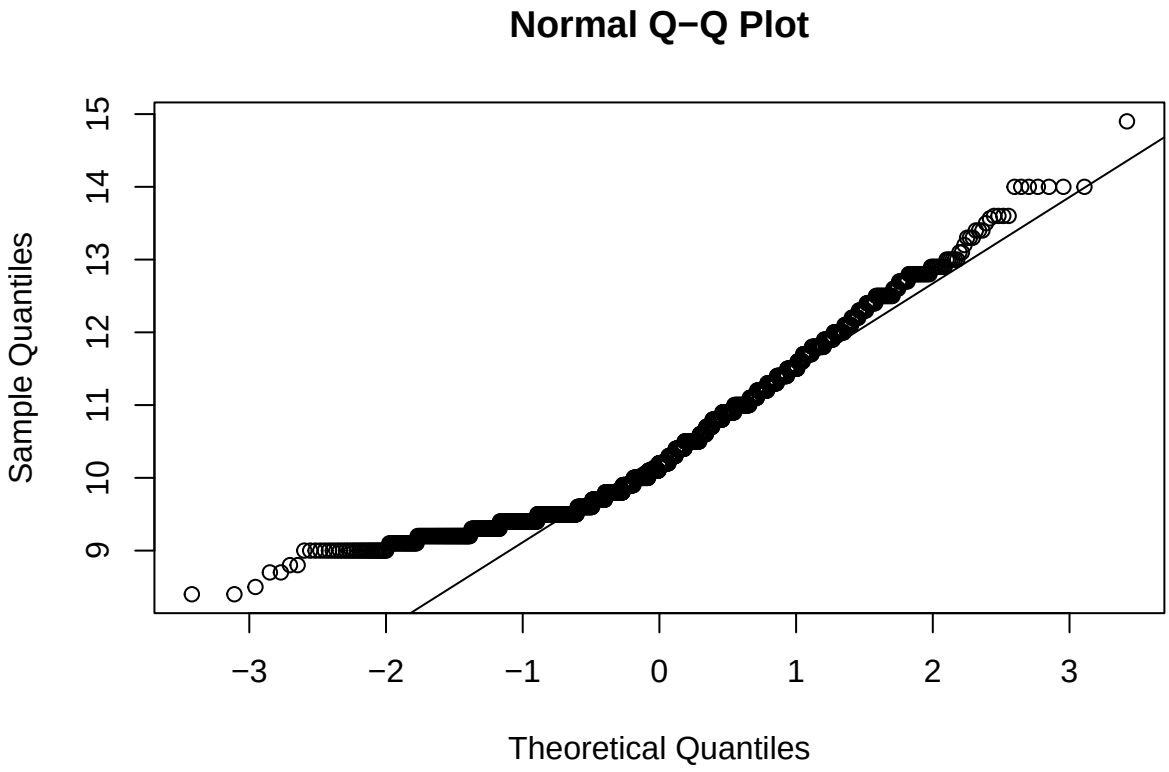
```
shapiro.test(df_red$quality)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_red$quality
## W = 0.85759, p-value < 2.2e-16
```

Podemos observar que, en todos los casos el valor p es menor al valor de significancia 0.05, con lo que se rechaza la hipotesis nula que supone que los datos están distribuidos normalmente.

Veámos un ejemplo gráfico que ayudará visualmente a contrastar la prueba de shapiro.

```
qqnorm(df_red$alcohol)
qqline(df_red$alcohol)
```



Comprobación de homocedasticidad

Dado que los atributos a revisar no siguen distribución normal, utilizaremos la prueba de Fligner-Killeen.

```
fligner.test(fixed.acidity ~ volatile.acidity, data = df_red)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  fixed.acidity by volatile.acidity
## Fligner-Killeen:med chi-squared = 245.16, df = 142, p-value = 1.709e-07
```

Dado que el p-valor es menor que el valor de significancia, se rechaza la hipotesis nula de homocedasticidad y se concluye las variables `fixed.acidity` y `volatile.acidity` presentan varianzas estadísiticas entre sus diferentes grupos.

```
fligner.test(alcohol ~ quality, data = df_red)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  alcohol by quality
## Fligner-Killeen:med chi-squared = 135.61, df = 5, p-value < 2.2e-16
```

Dado que el p-valor es menor que el valor de significancia, se rechaza la hipotesis nula de homocedasticidad y se concluye las variables `alcohol` presenta varianzas estadísiticas para las diferentes calidades `quality`.

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.

Dado que las pruebas de los datos han demostrado que las variables no siguen una distribución normal y que presentan heterocedasticidad, aplicaremos pruebas de correlacion de Spearman.

Aplicaremos también regresiones entre los ácidos y el alcohol y la calidad del vino.

```
lm(fixed.acidity ~ volatile.acidity, data = df_red)
```

```
##
## Call:
## lm(formula = fixed.acidity ~ volatile.acidity, data = df_red)
##
## Coefficients:
##      (Intercept)  volatile.acidity
##            9.634            -2.491
```

```
cor.test(df_red$fixed.acidity, df_red$volatile.acidity, method="spearman")
```

```
## Warning in cor.test.default(df_red$fixed.acidity, df_red$volatile.acidity, :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  df_red$fixed.acidity and df_red$volatile.acidity
## S = 871005139, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.2782822
```

```
lm(alcohol ~ quality, data = df_red)
```

```
##
## Call:
## lm(formula = alcohol ~ quality, data = df_red)
##
## Coefficients:
## (Intercept)      quality
##      6.8816       0.6283
```

```
cor.test(df_red$alcohol, df_red$quality, method="spearman")
```

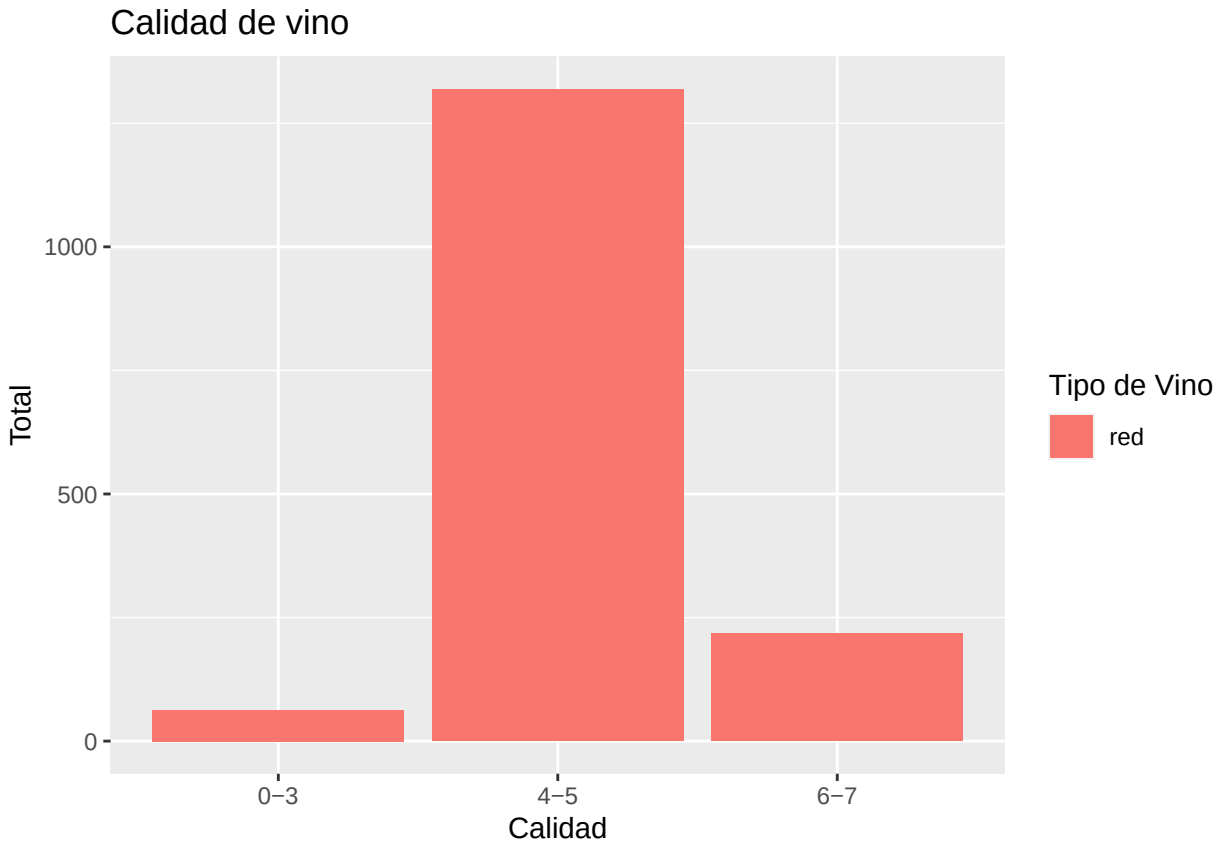
```
## Warning in cor.test.default(df_red$alcohol, df_red$quality, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  df_red$alcohol and df_red$quality
## S = 355321833, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4785317
```

Por último, realizaremos la comprobación entre los grupos de Kruskal, que es una alternativa no paramétrica. Para hacerlo más visual, lo haremos solamente con las variables alcohol y calidad del vino, pero antes, crearemos grupos de calidad de vino de la siguiente manera:

- Malo: 0 a 3
- Regular: 4 a 5
- Bueno: 6 a 7
- Muy bueno: 8 a 10

```
quality_df <- data.frame(df_red)
quality_df["quality_segment"] <- cut(quality_df$quality, breaks = c(0,4,6,8,10), labels = c("0-3", "4-5
ggplot(quality_df, aes(x=quality_segment, fill=type)) +
  geom_bar() +
  labs(x = "Calidad", y = "Total", fill = "Tipo de Vino", title = "Calidad de vino")
```



Ahora ejecutamos la prueba.

```
kruskal.test(alcohol ~ quality, data = quality_df )
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  alcohol by quality
## Kruskal-Wallis chi-squared = 412.38, df = 5, p-value < 2.2e-16
```

Dado que el p-valor obtenido es menor al nivel de significancia, se puede concluir que el alcohol muestra diferencias significativas para los diferentes rangos de calidad del vino.

Resolución

Analizando los datos y las comprobaciones llevadas a cabo, no podemos resolver el problema de saber cuál será la calidad del vino a partir de sus propiedades. Los motivos principales son:

- Falta de conocimiento de los valores químicos presentes en un vino por mi parte.
- Falta de profundidad en el análisis. Hemos analalizados todas las variables y trabajado en determinar la relación de la calidad solamente con el alcohol, lo cual es insuficiente para resolver el problema. En los mapas de calor de las correlaciones, no se apreció una relación fuerte/directa para trabajar en la comparación con solamente 2 grupos y se requiere de un análisis más exhaustivo y completo
- Las muestras de vinos no están balanceadas y presentan mucho vinos normales, muy pocos peores y ningún vino excelente.

Por otra parte, hemos podido observar la dificultad de trabajar con este tipo de datos ya que no siguen distribuciones normales, si bien, los datos son de calidad.

Video

El video explicativo puede descargarse de en este enlace: https://drive.google.com/file/d/1yFRLmDb-I3fdxxE-dpiU4up2tkHrnlcY/view?usp=sharing

Contribuciones

Contribuciones	Firma
Descripción del dataset.	MTL
Integración y selección de los datos	MTL
Limpieza de los datos	MTL
Análisis de los datos	MTL
Representación	MTL
Resolución	MTL
Código	MTL