Daniel Fernández Martínez

Práctica 2

1. Descripción del dataset

El dataset que se va a limpiar y analizar registra una serie de vinos blancos asociados a sus propiedades obtenidas tras la realización de pruebas fisicoquímicas (acidez, densidad, dióxido de azufre presente, etc.). Además de esto, cada vino tiene una puntuación de calidad asignada, representada con un número entero del 0 al 10, basada en datos sensoriales.

El dataset ha sido obtenido mediante el siguiente enlace:

https://archive.ics.uci.edu/ml/datasets/wine+quality

Mediante el análisis de este conjunto de datos, se pretende averiguar qué variables fisicoquímicas afectan directamente a la puntuación de calidad, tanto positiva como negativamente.

2. Integración y selección

El dataset cuenta con 4898 registros totales. Para cada uno de estos registros, se incluyen las siguientes 12 variables:

- **fixed acidity**: ácido tartárico, en g/dm³.
- **volatile acidity**: ácido acético, en g/dm³.
- **citric acid**: ácido cítrico, en g/dm³.
- **residual sugar**: azúcar residual, en g/dm³.
- **chlorides**: cloruros, en g/dm³.
- **free sulfur dioxide**: dióxido de sulfuro libre en el vino, en mg/dm³.
- **total sulfur dioxide**: dióxido de sulfuro total, en mg/dm³.
- **density**: densidad, en g/cm³.
- **pH**: medida de acidez o alcalinidad, sin unidad.
- **sulphates**: sulfato de potasio, en g/dm³.
- **alcohol**: graduación alcohólica, en %.
- **quality**: puntuación del 0 al 10, basada en datos sensoriales.

Inicialmente se utilizarán todas las variables en el análisis, y se identificarán cuáles son las relevantes durante el proceso.

3. Limpieza de datos

Inicialmente, se carga y analiza el dataset para comprobar que no contiene datos vacíos o no válidos y conocer los rangos de cada uno de sus parámetros. Además de ofrecer esta información, la opción *summary* permite obtener conocimiento sobre los rangos de los datos a tratar y otros cálculos para realizar un análisis descriptivo previo:

```
wine_data <- read.csv("./winequality-white.csv", sep = ";")
summary(wine_data)
```

```
 fixed.acidity     volatile.acidity     citric.acid         residual.sugar
 Min.   : 3.800    Min.   :0.0800      Min.   :0.0000      Min.   : 0.600
 1st Qu.: 6.300    1st Qu.:0.2100      1st Qu.:0.2700      1st Qu.: 1.700
 Median : 6.800    Median :0.2600      Median :0.3200      Median : 5.200
 Mean   : 6.855    Mean   :0.2782      Mean   :0.3342      Mean   : 6.391
 3rd Qu.: 7.300    3rd Qu.:0.3200      3rd Qu.:0.3900      3rd Qu.: 9.900
 Max.   :14.200    Max.   :1.1000      Max.   :1.6600      Max.   :65.800

   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
 Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
 1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
 Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
 Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
 3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
 Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390

       pH            sulphates           alcohol             quality
 Min.   :2.720     Min.   :0.2200      Min.   : 8.00        Min.   :3.000
 1st Qu.:3.090     1st Qu.:0.4100      1st Qu.: 9.50        1st Qu.:5.000
 Median :3.180     Median :0.4700      Median :10.40        Median :6.000
 Mean   :3.188     Mean   :0.4898      Mean   :10.51        Mean   :5.878
 3rd Qu.:3.280     3rd Qu.:0.5500      3rd Qu.:11.40        3rd Qu.:6.000
 Max.   :3.820     Max.   :1.0800      Max.   :14.20        Max.   :9.000
```

Ningún campo contiene valores NA, y tan solo el campo citric.acid contiene ceros, los cuales son válidos dada la naturaleza de los datos (la cantidad de ácido cítrico contenida en un vino puede ser igual a 0).

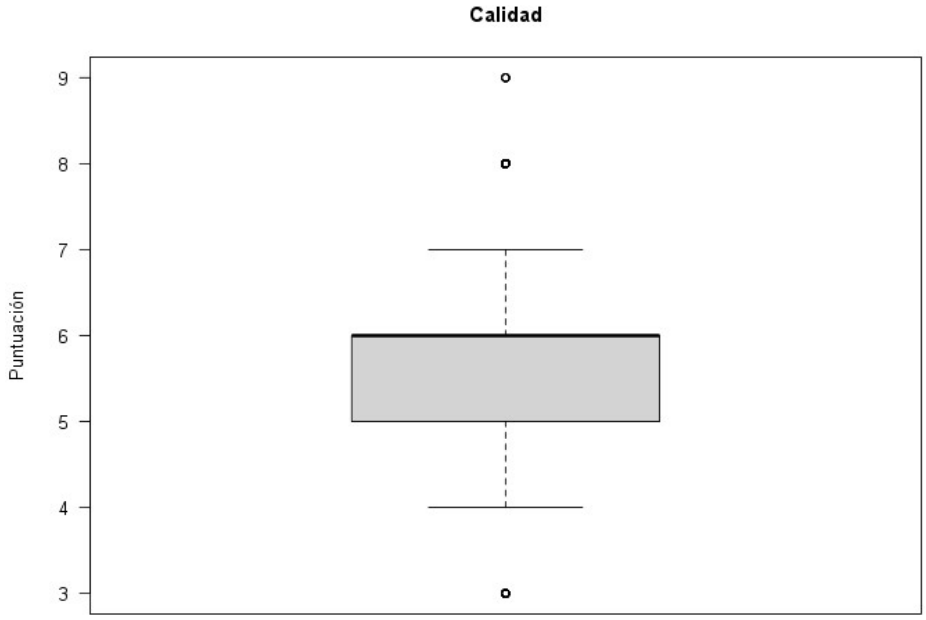
Tras la importación, se comprueban que los tipos de datos asignados sean los correctos. En este caso lo son, todas las variables son numéricas y la calidad se ha interpretado como valor entero:

```
sapply(wine_data, class)
```

```
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar
           "numeric"            "numeric"            "numeric"            "numeric"

           chlorides  free.sulfur.dioxide total.sulfur.dioxide              density
           "numeric"            "numeric"            "numeric"            "numeric"

                  pH            sulphates              alcohol              quality
           "numeric"            "numeric"            "numeric"            "integer"
```

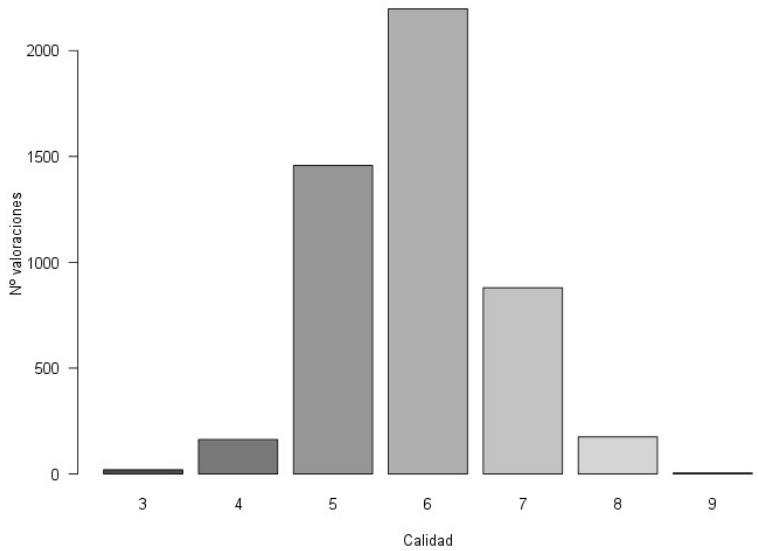
En cuanto a los valores extremos, se pueden identificar principalmente en la variable de calidad mediante un diagrama de caja:

```
boxplot(wine_data$quality, main = "Calidad", ylab = "Puntuación", las=1)
```



Mediante un diagrama de barras, se pueden observar mejor la cantidad de valoraciones para las puntuaciones 3, 8 y 9:

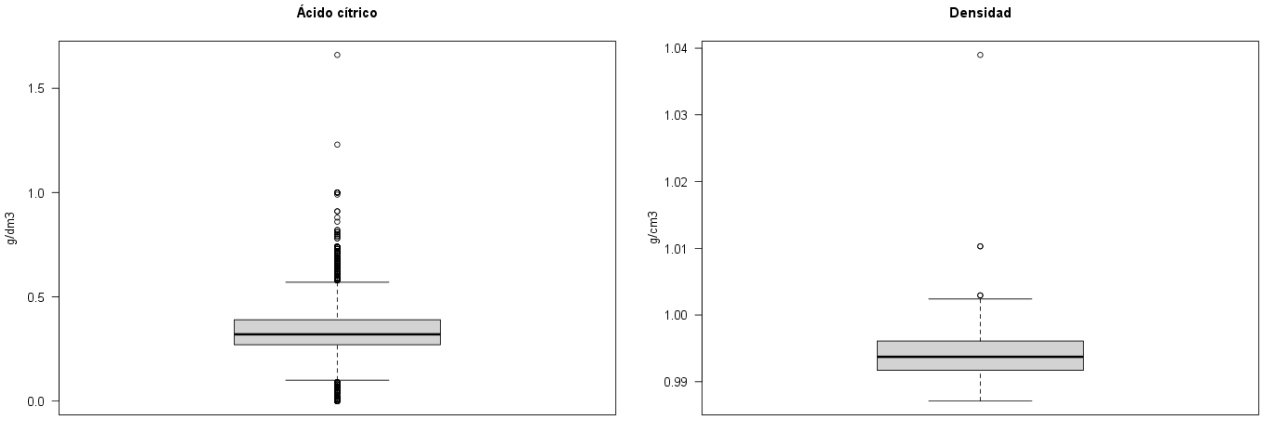
```
barplot(table(wine_data$quality),
        xlab = "Calidad",
        ylab = "Nº valoraciones",
        las = 1,
        col = gray.colors(7))
```

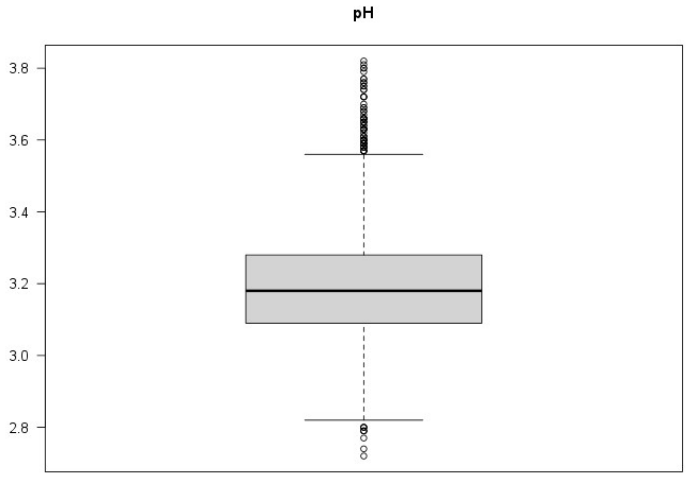


Especialmente las valoraciones 3 y 9 parecen ser muy poco frecuentes. Pese a esto, todos los valores se conservarán, al no surgir de errores en los datos o problemas relacionados.

En otras variables se puedan observar *outliers*, como se puede ver en las siguientes imágenes, pero estos no se tratarán ni excluirán al ser resultados válidos obtenidos tras la realización de las pruebas fisicoquímicas. Tampoco parecen haber registros con valores extremos que pudiesen indicar un error en los datos o magnitudes.

```
# Comprobación inicial de outliers en cada variable
for(var in seq_len(ncol(wine_data))) {
  boxplot(wine_data[,var], xlab = names(wine_data)[var])
}

# Ejemplos de columnas con outliers a destacar
boxplot(wine_data$citric.acid, main = "Ácido cítrico", ylab = "g/dm3", las=1)
boxplot(wine_data$density, main = "Densidad", ylab = "g/cm3", las=1)
boxplot(wine_data$pH, main = "pH", las=1)
```





4. Análisis de datos

4.1. Grupos a analizar

El análisis tiene como objetivo principal encontrar cualquier factor que sea relevante en la determinación de la calidad de un vino blanco. Además de esto, se extraerá cualquier relación o conclusión adicional que pueda resultar de interés.

4.2. Normalidad y homogeneidad

Inicialmente, se comprueba la normalidad de todas las variables de las que se dispone (exceptuando la calidad, al tratarse de una variable discreta):

```
apply(wine_data, 2, shapiro.test)
```

```
$fixed.acidity

      Shapiro-Wilk normality test

data:  newX[, i]
W = 0.97656, p-value < 2.2e-16
```

```
$volatile.acidity

      Shapiro-Wilk normality test

data:  newX[, i]
W = 0.90455, p-value < 2.2e-16
```

```
$citric.acid

      Shapiro-Wilk normality test

data:  newX[, i]
W = 0.92225, p-value < 2.2e-16
```

```
$residual.sugar

      Shapiro-Wilk normality test

data:  newX[, i]
W = 0.88457, p-value < 2.2e-16
```

```
$chlorides

      Shapiro-Wilk normality test

data:  newX[, i]
W = 0.59081, p-value < 2.2e-16
```

```
$free.sulfur.dioxide

      Shapiro-Wilk normality test

data:  newX[, i]
W = 0.94207, p-value < 2.2e-16
```

```
$total.sulfur.dioxide

      Shapiro-Wilk normality test
```

```
data:  newX[, i]
W = 0.98901, p-value < 2.2e-16
```

```
$density

      Shapiro-Wilk normality test

data:  newX[, i]
W = 0.9548, p-value < 2.2e-16
```

```
$pH

      Shapiro-Wilk normality test

data:  newX[, i]
W = 0.9881, p-value < 2.2e-16
```

```
$sulphates

      Shapiro-Wilk normality test

data:  newX[, i]
W = 0.95161, p-value < 2.2e-16
```

```
$alcohol

      Shapiro-Wilk normality test

data:  newX[, i]
W = 0.9553, p-value < 2.2e-16
```

Como se puede observar, en ningún caso el p-value supera el nivel de significancia de 0.05, por lo que se rechaza la hipótesis nula y se concluye que ninguno de los conjuntos de datos cuentan con una distribución normal.

Tras saber esto, se recurrirán a tests no paramétricos para obtener información adicional en la fase de análisis. Además, se utilizará a continuación el test de Fligner-Killeen para comprobar la homocedasticidad:

```
wine_data_hc <- apply(wine_data,2,function(x) fligner.test(x~quality, data=wine_data))
sapply(wine_data_hc, function (x) x$p.value)
```

```
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar
        1.609730e-07         5.738311e-19         2.806926e-63         5.979175e-28
           chlorides  free.sulfur.dioxide total.sulfur.dioxide              density
        6.561525e-06         7.064941e-36         2.697433e-35         9.079164e-07
                  pH            sulphates              alcohol
        1.965494e-09         7.900076e-19         9.162004e-49
```

Todos los tests resultan en p-valores menores al nivel de significancia 0.05, por lo que todas las variables tienen varianzas estadísticamente distintas para las diferentes evaluaciones de calidad.

4.3. Análisis

Dado que se está tratando varias variables numéricas que no cuentan con una distribución normal, y una variable numérica discreta, las pruebas que se realizarán con el objetivo de obtener información sobre los factores que determinan la calidad del vino serán las siguientes:

- Contraste de hipótesis
- Regresión
- Correlación

Contraste de hipótesis

Dado que se cuenta con más de dos grupos, en el caso de la calidad del vino, se recurre a un test de Kruskal-Wallis. Se aplicará este test para comparar la calidad con el resto de variables:

```
apply(wine_data, 2, function (x) kruskal.test(x ~ quality, data=wine_data))
```

```
$fixed.acidity

      Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 40.868, df = 6, p-value = 3.075e-07
```

```
$volatile.acidity

      Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 286.13, df = 6, p-value < 2.2e-16
```

```
$citric.acid

      Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 13.12, df = 6, p-value = 0.04117
```

```
$residual.sugar

      Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 94.519, df = 6, p-value < 2.2e-16
```

```
$chlorides

      Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 512.99, df = 6, p-value < 2.2e-16
```

```
$free.sulfur.dioxide

      Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 115.07, df = 6, p-value < 2.2e-16
```

```
$total.sulfur.dioxide
```

```
	Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 266.67, df = 6, p-value < 2.2e-16
```

```
$density

	Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 652.61, df = 6, p-value < 2.2e-16
```

```
$pH

	Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 65.473, df = 6, p-value = 3.454e-12
```

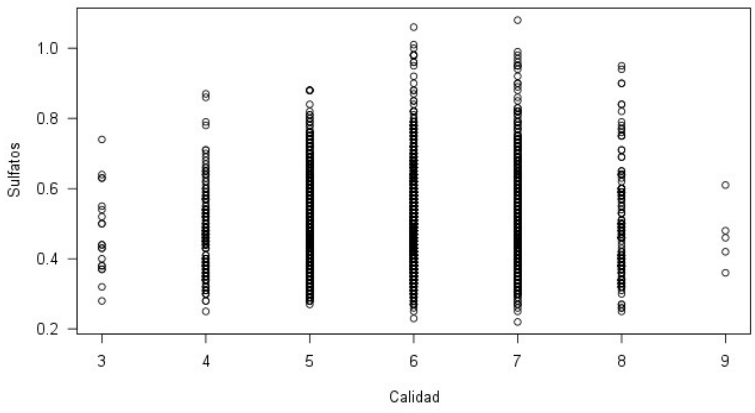
```
$sulphates

	Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 13.78, df = 6, p-value = 0.03219
```

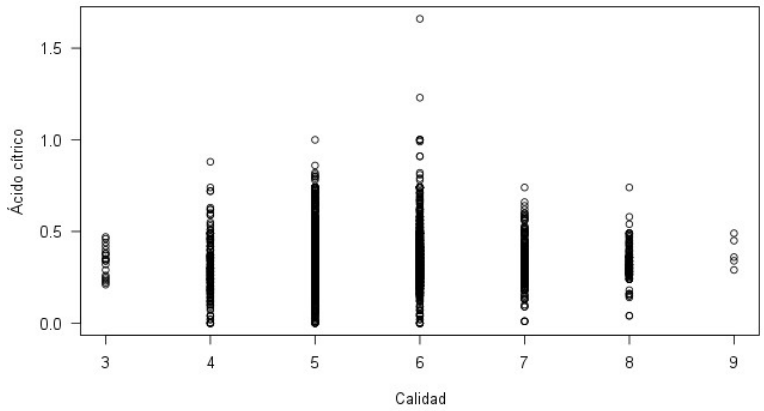
```
$alcohol

	Kruskal-Wallis rank sum test

data:  x by quality
Kruskal-Wallis chi-squared = 1014.1, df = 6, p-value < 2.2e-16
```

Pese a que todos los p-valores están por debajo del nivel de significancia de 0.05, la calidad parece guardar cierta relación con los sulfatos (p-valor de 0.03219) y la cantidad de ácido cítrico (p-valor de 0.04117). A lo largo de los dos siguientes tests se comprobará si estas relaciones pueden resultar de interés.

Regresión

Se realiza un test de regresión lineal para comprobar si las dos variables mencionadas anteriormente tienen una relación con la calidad. Antes de realizar el test en sí y comprobar los coeficientes de determinación, se puede buscar una relación a simple vista:





Se puede observar a simple vista que no hay una relación lineal entre las variables seleccionadas. De todas formas, se realiza el test de regresión para corroborar lo observado:

```
summary(lm(quality ~ sulphates, data=wine_data))
```

```
Call:
lm(formula = quality ~ sulphates, data = wine_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9821 -0.8488  0.1137  0.1803  3.1762

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6739     0.0557 101.863  < 2e-16 ***
sulphates     0.4165     0.1108   3.761 0.000171 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8845 on 4896 degrees of freedom
Multiple R-squared:  0.002881,  Adjusted R-squared:  0.002678
F-statistic: 14.15 on 1 and 4896 DF,  p-value: 0.000171
```

```
summary(lm(quality ~ citric.acid, data=wine_data))
```

```
Call:
lm(formula = quality ~ citric.acid, data = wine_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8863 -0.8735  0.1191  0.1326  3.1326

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.90043    0.03717 158.736   <2e-16 ***
citric.acid -0.06739    0.10458  -0.644    0.519
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8857 on 4896 degrees of freedom
Multiple R-squared:  8.481e-05, Adjusted R-squared:  -0.0001194
F-statistic: 0.4153 on 1 and 4896 DF,  p-value: 0.5193
```

Se confirma que la relación lineal entre la calidad y ambas variables es prácticamente nula.

Correlación

Dado que los datos no siguen una distribución normal, se recurre directamente a la realización de un test de Spearman:

```
cor_data <- cor(wine_data, method = "spearman")
sort(cor_data["quality",])
```

```
            density           chlorides total.sulfur.dioxide     volatile.acidity
        -0.34835102         -0.31448848          -0.19668029          -0.19656168
      fixed.acidity      residual.sugar          citric.acid  free.sulfur.dioxide
        -0.08448545         -0.08206979           0.01833273           0.02371338
          sulphates                  pH              alcohol
          0.03331897          0.10936208           0.44036918
```

Se comprueba que los valores p sean significativos:

```
sort(apply(wine_data[,-12], 2, function (x) {
  cor.test(wine_data$quality, x, method = "spearman", exact = FALSE)[["p.value"]]
}))
```

```
             alcohol              density            chlorides total.sulfur.dioxide
      1.659196e-231        9.497078e-140        6.907550e-113         6.582657e-44
   volatile.acidity                   pH        fixed.acidity       residual.sugar
       7.417221e-44         1.656016e-14         3.183308e-09         8.821911e-09
          sulphates  free.sulfur.dioxide          citric.acid
       1.970574e-02         9.703376e-02         1.995589e-01
```

El valor p es significativo para todas las variables exceptuando la correspondiente al ácido cítrico y la correspondiente al dióxido de sulfuro libre, por lo que se ignorará la correlación de ambas con la calidad del vino. Teniendo en cuenta esto, como se puede observar en el resultado, las variables que se correlacionan positivamente con la calidad de un vino blanco, de mayor a menor influencia, son las siguientes:

1. Alcohol
2. pH
3. Sulfatos

Y las variables que se correlacionan negativamente con la calidad de un vino blanco son:

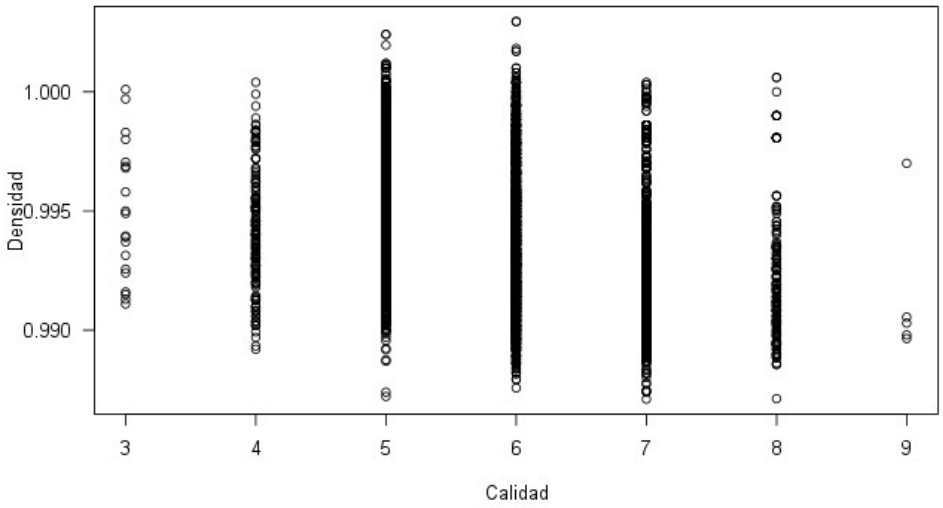
1. Densidad
2. Cloruros
3. Dióxido de sulfuro total
4. Ácido acético
5. Ácido tartárico
6. Azúcar residual

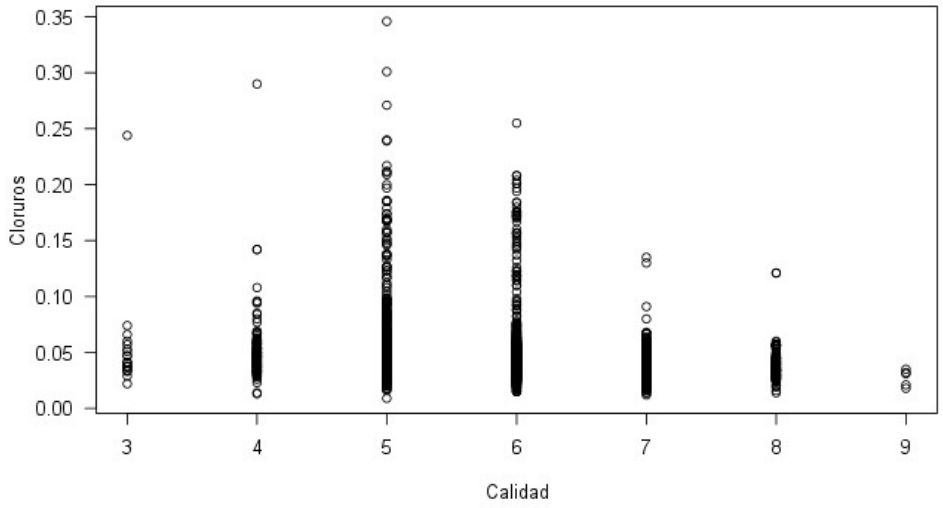
Aún teniendo en cuenta esto, todas las correlaciones tienen coeficientes relativamente bajos, por lo que no hay factores concretos que tengan un gran peso a destacar con respecto al resto a la hora de determinar la calidad del vino.

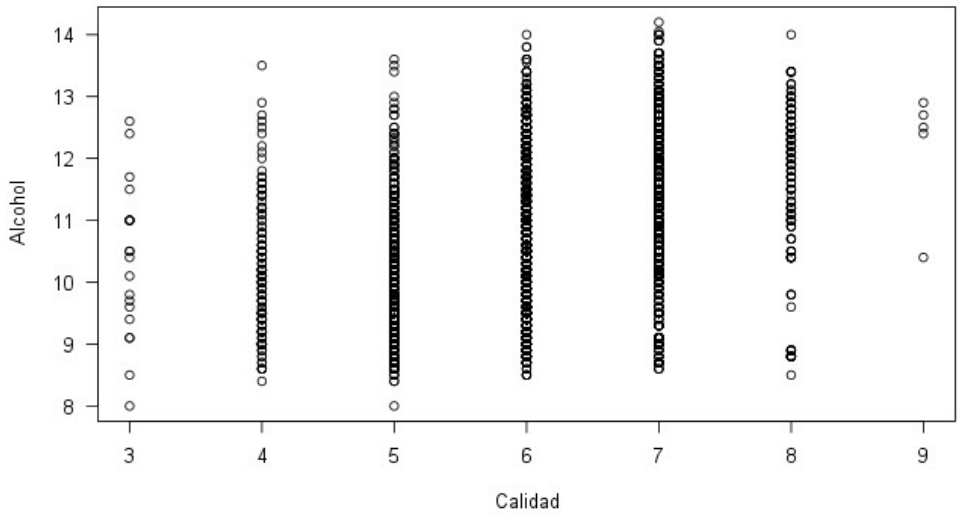
5. Visualización de los resultados

Se destacan los resultados del test de correlación. Las variables más significativas y con más impacto sobre la calidad son el alcohol, la densidad y los cloruros (se ignoran 3 outliers con densidad igual o superior a 1.01 simplemente para facilitar la visualización de los datos):

```
wine_data_d <- wine_data[wine_data$density<1.01,]
plot(wine_data_d$quality, wine_data_d$density, xlab="Calidad", ylab="Densidad", las=1)
plot(wine_data$quality, wine_data$density, xlab="Calidad", ylab="Cloruros", las=1)
plot(wine_data$quality, wine_data$alcohol, xlab="Calidad", ylab="Alcohol", las=1)
```







6. Conclusiones

No se han encontrado valores fisicoquímicos concretos que afecten en gran medida a la calidad del vino blanco basada en datos sensoriales. La parte más relevante a destacar ha sido la última fase del análisis, en la que se ha hecho el test de correlación para ordenar finalmente las variables según su impacto, aunque este fuese bajo.

De esta forma, los vinos blancos con las siguientes cualidades han sido ligeramente favorecidos durante la realización de las evaluaciones mediante información sensorial:

- Una graduación alcohólica alta.
- Una densidad baja.
- Una proporción de cloruros baja.