Práctica 2: Red Wine Quality

Autoras: Sandra Campos Suárez y Mª de los Ángeles García Carrión

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

El objetivo general de este estudio es conocer qué variables influyen en la calidad de un vino, es decir, de qué forma influye el pH, ácido cítrico y densidad, entre otros, para que un vino tenga Alta o Baja calidad. Para ello, estudiaremos y analizaremos la relación entre las variables mediante análisis de correlación, regresión lineal y múltiple, etc.

Las variables de nuestro conjunto de datos son: **acidez fija**, **acidez volátil**, **ácido cítrico**, **azúcar residual**, **cloruros**, **anhídrido sulfuroso libre**, **anhídrido sulfuroso total**, **densidad**, **pHv,** sulfatos** y **alcohol**. La variable de salidad que está basada en los datos anteriores es la **calidad** (puntuación entre 0 y 10).

2 . 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 primer lugar, leemos el fichero de datos

```
# Conexión a la fuente de datos
wine <- read.csv('csv/winequality-red.csv',stringsAsFactors = FALSE)
```

A través de la siguiente función observamos que nuestro dataset está compuesto por 1599 registros y 12 columnas

```
# Dimensiones del dataset
dim(wine)
```

```
## [1] 1599   12
```

A continuación se muestra el detalle de las variables y por el momento vamos a utilizar todas las variables. En los apartados siguientes analizaremos la correlación entre las variables y donde concluiremos si eliminamos alguna variable o mantenemos la selección actual.

```
# Detalle del dataset. Todas las variables son numéricas excepto la variable quality siendo un integer.
str(wine)
```

```
## '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 ...
```

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos. Tal y como se ha indicado en el apartado anterior, no tenemos presencia de valores nulos en las variables.Otra forma de conocer si existe valores nulos a lo largo de la tabla es mediante la siguiente función, donde observamos que hay 0 valores nulos.

```
colSums(is.na(wine))
```

```
##        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
```

```
# Resumen de las variables
summary(wine)
```

3.2. Identifica y gestiona los valores extremos.

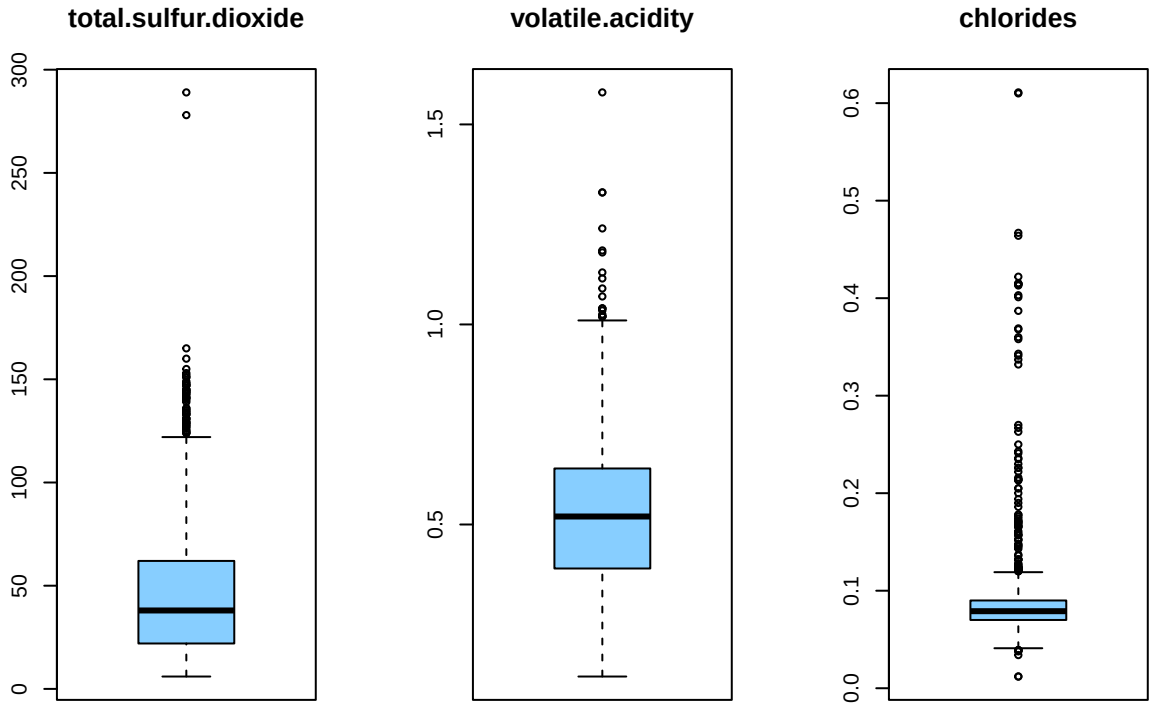
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :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
```

La Comunidad Europea establece en el Reglamento (CE) nº 643/2006 de la Comisión, de 27 de abril de 2006 los máximos admitidos en el vino (https://www.boe.es/doue/2006/115/L00006-00009.pdf). Por ende, el

citric.acid (1 g/l), sulphates (2 g/l) y chlorides (<1 g/l) están dentro de la normativa así que no consideramos presencia de oultiers al igual que en la calidad, el alcohol (<15 grados), pH (entre 3-4), density (<1 g/l) y residual.sugar (<18), fixed.acidity.

Sin embargo, tendremos que tratar las siguientes variables que no están dentro de la normativa como total.sulfur.dioxide (<150 mg), volatile.acidity (<1.2 g/l), chlorides (<0.5g/l) así que esta variable tendremos que tratar sus outliers. Excluiremos estos registros del análisis ya que no cumple la normativa vinos con estos valores. Dejaremos los outliers que están por debajo del valor de la normativa, ya que si son valores válidos.

```
# Boxplot de las variables que superan el máximo de la Normativa (CE)
par(mfrow = c(1, 3))
boxplot(wine$total.sulfur.dioxide, main = "total.sulfur.dioxide", col="skyblue1")
boxplot(wine$volatile.acidity, main = "volatile.acidity", col="skyblue1")
boxplot(wine$chlorides, main = "chlorides", col="skyblue1")
```




```
# Total outliers
sum(boxplot.stats(wine$total.sulfur.dioxide)$out>150)
```

```
## [1] 9
```

```
sum(boxplot.stats(wine$volatile.acidity)$out>1.2)
```

```
## [1] 4
```

```
sum(boxplot.stats(wine$chlorides)$out>0.5)
```

```
## [1] 2
```

```
# Filtramos el dataset excluyendo los outliers ya que no cumplen la normativa
wine_df<- subset(wine, total.sulfur.dioxide <= 150 & chlorides <=0.5 & volatile.acidity <=1.2)
```

```r
# Comprobamos que no existen outliers que no cumplan la normativa
sum(boxplot.stats(wine_df$total.sulfur.dioxide)$out>150)
```

```
## [1] 0
```

```r
sum(boxplot.stats(wine_df$volatile.acidity)$out>1.2)
```

```
## [1] 0
```

```r
sum(boxplot.stats(wine_df$chlorides)$out>0.5)
```

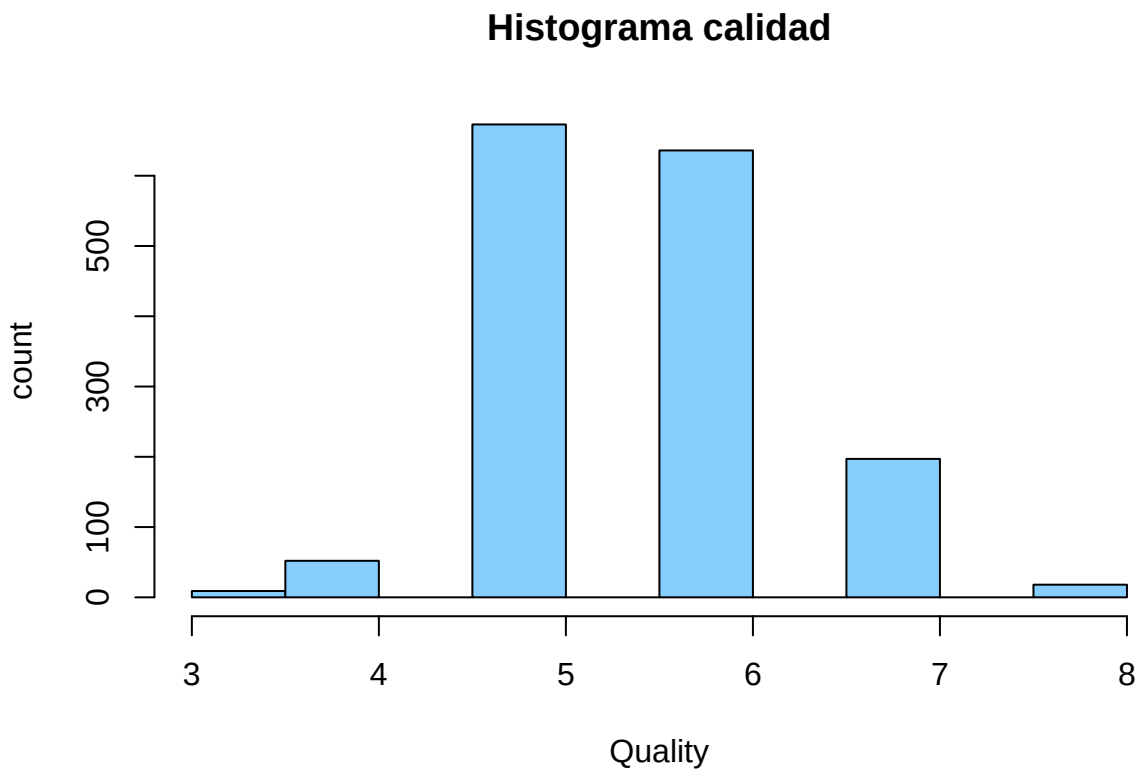
```
## [1] 0
```

4. Análisis de los datos.

4.1. 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?)

Variable dependiente: Quality Antes de continuar con el análisis estudiamos la distribución de la calidad con el fin de saber si disponemos información sobre los vinos de baja, media o elevada calidad. Para ello, utilizamos el histográma donde efectivamente tenemos vinos con calidad inferior a 5, de 5 y superior a 6 .

```r
# Histograma variable calidad del vino
hist(wine_df$quality, main = "Histograma calidad", ylab = "count", xlab = "Quality", col="skyblue1")
```




```r
# Convertimos la variable quality en categorica
breakPoints <- c(0, 4, 6, Inf)
```

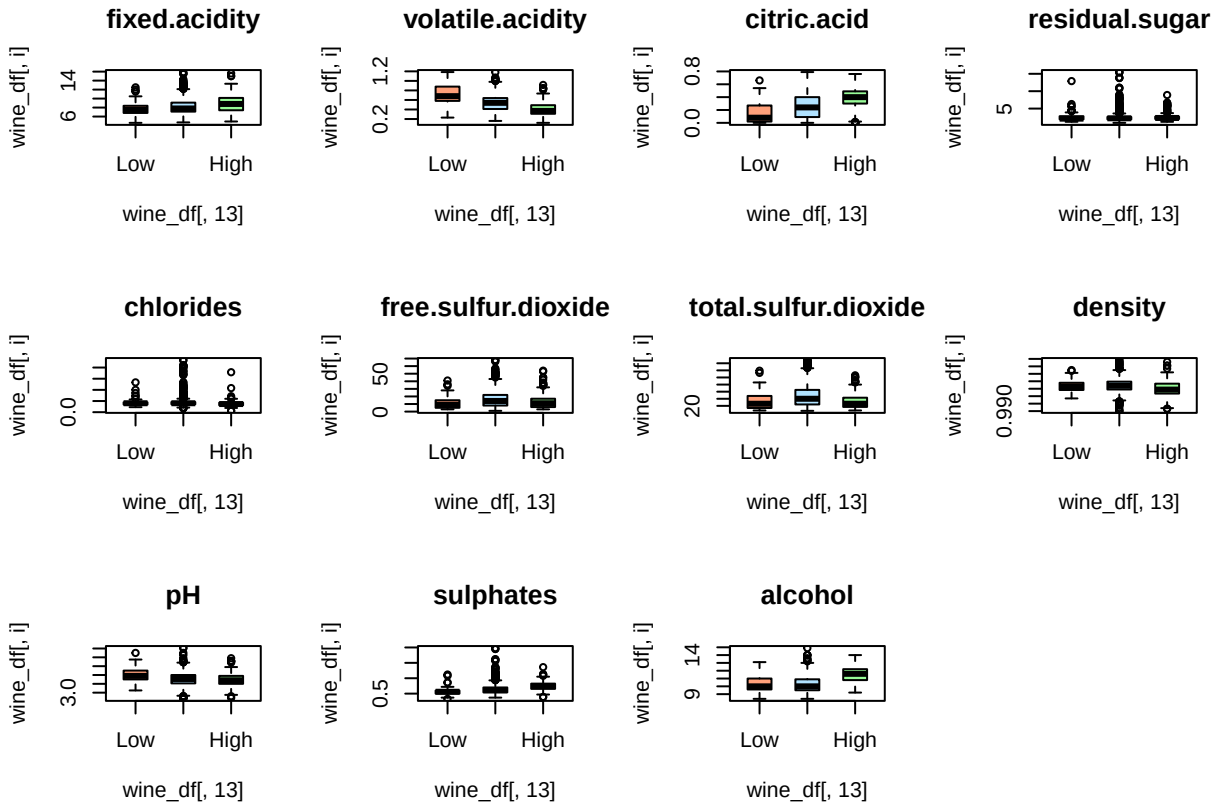
```r
categories <- c("Low", "Medium", "High")
wine_df$quality_group <- cut(wine_df$quality, breaks = breakPoints, labels = categories)

# Tabla datos
table(wine_df$quality, wine_df$quality_group)
```

```
##
##     Low Medium High
##   3   9      0    0
##   4  52      0    0
##   5   0    673    0
##   6   0    636    0
##   7   0      0  197
##   8   0      0   18
```

Variables independientes Vamos a visualizar todas las variables independientes en función de la calidad.

```r
# Boxplot de todas las variables por grupo de calidad
par(mfrow = c(3, 4))
for(i in 1:11) {
  boxplot(wine_df[, i] ~ wine_df[, 13],
          col = c("#FFA07A", "#B0E2FF", "#98fb98"),
          boxwex = 0.5)
  nombre.de.variable <- names(wine_df)[i]
  title(main = nombre.de.variable)
}
```



A priori, todas las variables excepto residual.sugar y pH presentan relación tanto positiva como negativa con la calidad del vino. Sin embargo, nos vamos a apoyar en los análisis de correlaciones, contraste de hipótesis y regresión múltiple.

```
# Conocer si existe relación entre el alcohol y la calidad del vino
high_alcohol<-quantile(wine_df$alcohol, probs =0.75)
wine_high_alcohol<-wine_df[wine_df$alcohol>=high_alcohol,]$quality
wine_low_alcohol<-wine_df[wine_df$alcohol<high_alcohol,]$quality
t.test(wine_high_alcohol, wine_df$quality, alternative = "greater")
```

Relación entre la calidad y sus variables independientes

```
##
##  Welch Two Sample t-test
##
## data:  wine_high_alcohol and wine_df$quality
## t = 12.413, df = 620.51, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4846099       Inf
## sample estimates:
## mean of x mean of y
##  6.198511  5.639748
```

```
t.test(wine_high_alcohol, wine_df$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_high_alcohol and wine_df$quality
## t = 12.413, df = 620.51, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.6329171
## sample estimates:
## mean of x mean of y
##  6.198511  5.639748
```

Rechazamos hipótesis nula ya que el p-valor es menor que el nivel de significación, por tanto, la variable alcohol es significativa, es decir, cuanto más alcohol la calidad será mayor.

```
# Conocer si existe relación entre el azúcar y la calidad del vino
high_sugar<-quantile(wine_df$residual.sugar, probs =0.75)
wine_high_sugar<-wine_df[wine_df$residual.sugar>=high_sugar,]$quality
wine_low_sugar<-wine_df[wine_df$residual.sugar<high_sugar,]$quality
t.test(wine_high_sugar, wine_df$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_high_sugar and wine_df$quality
## t = -0.074755, df = 656.57, p-value = 0.5298
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.07794844         Inf
## sample estimates:
```

```
## mean of x mean of y
##  5.636364  5.639748
```

```
t.test(wine_high_sugar, wine_df$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_high_sugar and wine_df$quality
## t = -0.074755, df = 656.57, p-value = 0.4702
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.07118045
## sample estimates:
## mean of x mean of y
##  5.636364  5.639748
```

En este caso, aceptamos la hipótesis nula ya que el p-valor (0.5298) es mayor que el nivel de significación (0.05), por tanto, la variable azúcar no es significativa, es decir, no influye la cantidad de azúcar en la calidad de un vino.

```
# Conocer si existe relación entre fixed.acidity y la calidad del vino
high_fixed.acidity<-quantile(wine_df$fixed.acidity, probs = 0.75)
wine_high_fixed.acidity<-wine_df[wine_df$fixed.acidity>=high_fixed.acidity,]$quality
wine_low_fixed.acidity<-wine_df[wine_df$fixed.acidity<high_fixed.acidity,]$quality
t.test(wine_high_fixed.acidity, wine_df$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_high_fixed.acidity and wine_df$quality
## t = 4.0196, df = 612.03, p-value = 3.279e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1099085       Inf
## sample estimates:
## mean of x mean of y
##  5.825980  5.639748
```

```
t.test(wine_high_fixed.acidity, wine_df$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_high_fixed.acidity and wine_df$quality
## t = 4.0196, df = 612.03, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 0.262557
## sample estimates:
## mean of x mean of y
##  5.825980  5.639748
```

En este caso, rechazamos la hipótesis nula ya que el p-valor (3.279e-05) es menor que el nivel de significación (0.05), por tanto, la variable fixed.acidity es significativa, es decir, influye en la calidad de un vino.

```r
# Conocer si existe relación entre volatile.acidity y la calidad del vino
high_volatile.acidity<-quantile(wine_df$volatile.acidity, probs =0.75)
wine_df.high_Volatile.acidity<-wine_df[wine_df$volatile.acidity>=high_volatile.acidity,]$quality
wine_df.low_Volatile.acidity<-wine_df[wine_df$volatile.acidity<high_volatile.acidity,]$quality
t.test(wine_df.high_Volatile.acidity, wine_df$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_Volatile.acidity and wine_df$quality
## t = -8.2298, df = 667.95, p-value = 4.916e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.2761054
## sample estimates:
## mean of x mean of y
##  5.294554  5.639748
```

```r
t.test(wine_df.high_Volatile.acidity, wine_df$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_Volatile.acidity and wine_df$quality
## t = -8.2298, df = 667.95, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.414281       Inf
## sample estimates:
## mean of x mean of y
##  5.294554  5.639748
```

En este caso, rechazamos la hipótesis nula ya que el p-valor (4.916e-16) es menor que el nivel de significación (0.05), por tanto, la variable volatile.acidity es significativa, es decir, influye en la calidad de un vino.

```r
# Conocer si existe relación entre citric.acid y la calidad del vino
high_citric.acid<-quantile(wine_df$citric.acid, probs = 0.75)
wine_high_citric.acid<-wine_df[wine_df$citric.acid>=high_volatile.acidity,]$quality
wine_low_citric.acid<-wine_df[wine_df$citric.acid<high_volatile.acidity,]$quality
t.test(wine_high_citric.acid, wine_df$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_high_citric.acid and wine_df$quality
## t = 3.3157, df = 63.644, p-value = 0.0007565
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1951786       Inf
## sample estimates:
## mean of x mean of y
##  6.032787  5.639748
```

```r
t.test(wine_high_citric.acid, wine_df$quality, alternative = "less")
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  wine_high_citric.acid and wine_df$quality
## t = 3.3157, df = 63.644, p-value = 0.9992
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.5908999
## sample estimates:
## mean of x mean of y
##  6.032787  5.639748
```

En este caso, rechazamos la hipótesis nula ya que el p-valor (0.0007565) es menor que el nivel de significación (0.05), por tanto, la variable citric.acid es significativa, es decir, influye en la calidad de un vino.

```
# Conocer si existe relación entre chlorides y la calidad del vino
high_chlorides<-quantile(wine_df$chlorides, probs =0.75)
wine_df.high_Chlorides<-wine_df[wine_df$chlorides>=high_chlorides,]$quality
wine_df.low_Chlorides<-wine_df[wine_df$chlorides<high_chlorides,]$quality
t.test(wine_df.high_Chlorides, wine_df$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_Chlorides and wine_df$quality
## t = -2.7399, df = 668.63, p-value = 0.003155
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.04651793
## sample estimates:
## mean of x mean of y
##  5.523114  5.639748
```

```
t.test(wine_df.high_Chlorides, wine_df$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_Chlorides and wine_df$quality
## t = -2.7399, df = 668.63, p-value = 0.9968
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.1867486        Inf
## sample estimates:
## mean of x mean of y
##  5.523114  5.639748
```

En este caso, rechazamos la hipótesis nula ya que el p-valor (0.003155) es menor que el nivel de significación (0.05), por tanto, la variable chlorides es significativa, es decir, influye en la calidad de un vino.

```
# Conocer si existe relación entre free.sulfur.dioxide y la calidad del vino
high_free.sulfur.dioxide<-quantile(wine_df$free.sulfur.dioxide, probs =0.75)
wine_df.high_free.sulfur.dioxide<-wine_df[wine_df$free.sulfur.dioxide>=high_free.sulfur.dioxide,]$quali
wine_df.low_free.sulfur.dioxide<-wine_df[wine_df$free.sulfur.dioxide<high_free.sulfur.dioxide,]$quality
t.test(wine_df.high_free.sulfur.dioxide, wine_df$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  wine_df.high_free.sulfur.dioxide and wine_df$quality
## t = -1.641, df = 738.73, p-value = 0.05062
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf 0.0002407037
## sample estimates:
## mean of x mean of y
##  5.573427  5.639748
```

```
t.test(wine_df.high_free.sulfur.dioxide, wine_df$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_free.sulfur.dioxide and wine_df$quality
## t = -1.641, df = 738.73, p-value = 0.9494
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.1328828        Inf
## sample estimates:
## mean of x mean of y
##  5.573427  5.639748
```

En este caso, rechazamos la hipótesis nula ya que el p-valor (0.05062) es menor que el nivel de significación (0.05), por tanto, la variable free.sulfur.dioxide es significativa, es decir, influye en la calidad de un vino.

```
# Conocer si existe relación entre total.sulfur.dioxide y la calidad del vino
high_total.sulfur.dioxide<-quantile(wine_df$total.sulfur.dioxide, probs =0.75)
wine_df.high_total.sulfur.dioxide<-wine_df[wine_df$total.sulfur.dioxide>=high_total.sulfur.dioxide,]$qua
wine_df.low_total.sulfur.dioxide<-wine_df[wine_df$total.sulfur.dioxide<high_total.sulfur.dioxide,]$quali
t.test(wine_df.high_total.sulfur.dioxide, wine_df$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_total.sulfur.dioxide and wine_df$quality
## t = -6.8546, df = 735.55, p-value = 7.572e-12
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1989852
## sample estimates:
## mean of x mean of y
##  5.377834  5.639748
```

```
t.test(wine_df.high_total.sulfur.dioxide, wine_df$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_total.sulfur.dioxide and wine_df$quality
## t = -6.8546, df = 735.55, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.3248426        Inf
## sample estimates:
## mean of x mean of y
```

```
##  5.377834  5.639748
```

En este caso, rechazamos la hipótesis nula ya que el p-valor (7.572e-12) es menor que el nivel de significación (0.05), por tanto, la variable total.sulfur.dioxide es significativa, es decir, influye en la calidad de un vino.

```
# Conocer si existe relación entre density y la calidad del vino
high_density<-quantile(wine_df$density, probs =0.75)
wine_df.high_density<-wine_df[wine_df$density>=high_density,]$quality
wine_df.low_density<-wine_df[wine_df$density<high_density,]$quality
t.test(wine_df.high_density, wine_df$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_density and wine_df$quality
## t = -1.9336, df = 640.07, p-value = 0.0268
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01230325
## sample estimates:
## mean of x mean of y
##  5.556675  5.639748
```

```
t.test(wine_df.high_density, wine_df$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_density and wine_df$quality
## t = -1.9336, df = 640.07, p-value = 0.9732
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.1538419        Inf
## sample estimates:
## mean of x mean of y
##  5.556675  5.639748
```

En este caso, rechazamos la hipótesis nula ya que el p-valor (0.0268) es menor que el nivel de significación (0.05), por tanto, la variable density es significativa, es decir, influye en la calidad de un vino.

```
# Conocer si existe relación entre pH y la calidad del vino
high_pH<-quantile(wine_df$pH, probs =0.75)
wine_df.high_pH<-wine_df[wine_df$pH>=high_pH,]$quality
wine_df.low_pH<-wine_df[wine_df$pH<high_pH,]$quality
t.test(wine_df.high_pH, wine_df$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_pH and wine_df$quality
## t = -1.1556, df = 650.58, p-value = 0.1241
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.02197566
## sample estimates:
## mean of x mean of y
##  5.588095  5.639748
```

```r
t.test(wine_df.high_pH, wine_df$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.high_pH and wine_df$quality
## t = -1.1556, df = 650.58, p-value = 0.8759
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.1252805        Inf
## sample estimates:
## mean of x mean of y
##  5.588095  5.639748
```

En este caso, rechazamos la hipótesis nula ya que el p-valor (0.1241) es menor que el nivel de significación (0.05), por tanto, la variable pH es significativa, es decir, influye en la calidad de un vino.

```r
# Conocer si existe relación entre sulphates y la calidad del vino
high_sulphates<-quantile(wine_df$sulphates, probs =0.75)
wine_df.altoSulphates<-wine_df[wine_df$sulphates>=high_sulphates,]$quality
wine_df.bajoSulphates<-wine_df[wine_df$sulphates<high_sulphates,]$quality
t.test(wine_df.altoSulphates, wine_df$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.altoSulphates and wine_df$quality
## t = 9.0909, df = 626.28, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3370149       Inf
## sample estimates:
## mean of x mean of y
##  6.051345  5.639748
```

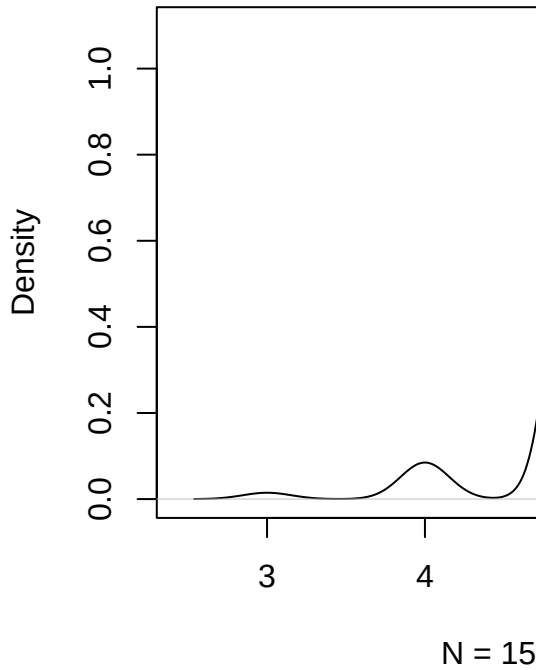
```r
t.test(wine_df.altoSulphates, wine_df$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_df.altoSulphates and wine_df$quality
## t = 9.0909, df = 626.28, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.4861793
## sample estimates:
## mean of x mean of y
##  6.051345  5.639748
```

En este caso, rechazamos la hipótesis nula ya que el p-valor (2.2e-16) es menor que el nivel de significación (0.05), por tanto, la variable sulphates es significativa, es decir, influye en la calidad de un vino.

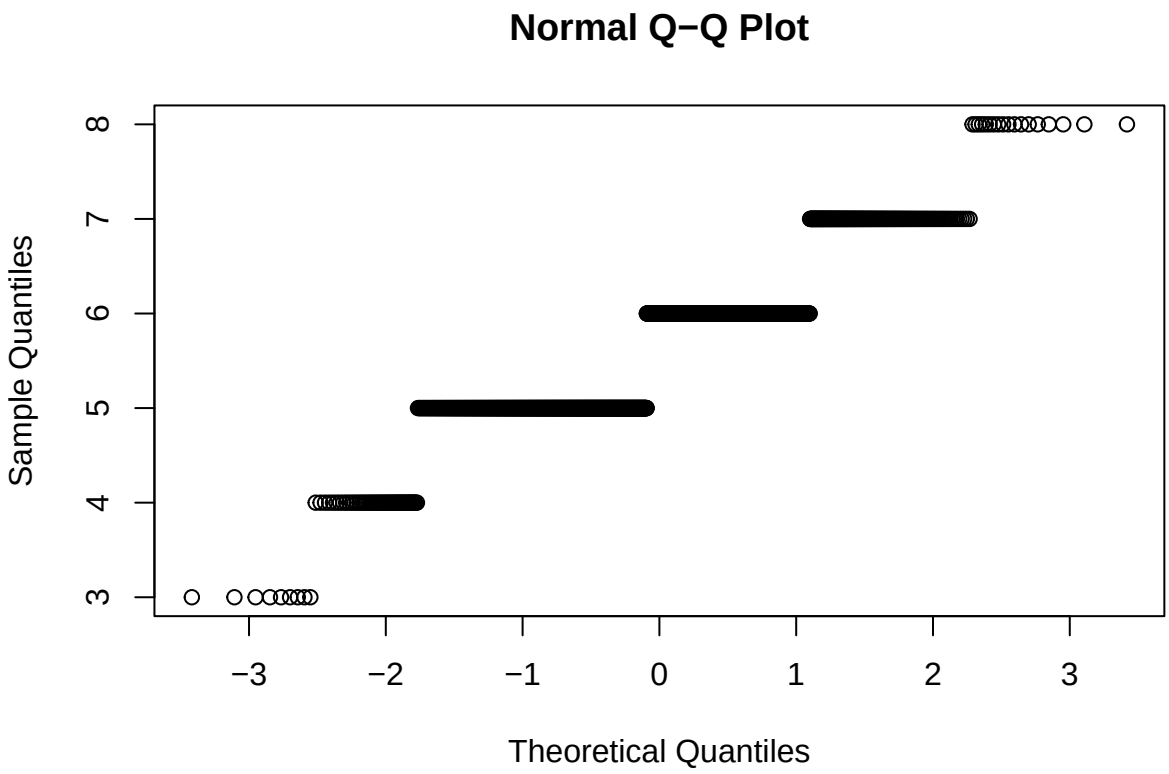
No influyen: residual.sugar. Las variables que son significativas son: alcohol, citric.acid, fixed.acidity, volatile.acidity, citric.acid, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH y sulphates

```r
# Comprobación de la normalidad de la variable quality
plot(density(wine_df$quality),main="Density")
```



4.2. Comprobación de la normalidad y homogeneidad de la varianza.

```r
qqnorm(wine_df$quality)
```



El 'density plot' muestra una distribución asimétrica. El 'qqplot' nos indica una distribución que no es normal de la variable Weight. Por tanto, los gráficos parecen indicar que la variable Weight sigue una distribución normal.

```
#Contaste de normalidad. Test de Shapiro-Wilk
shapiro.test (wine_df$quality)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_df$quality
## W = 0.85692, p-value < 2.2e-16
```

Con un p-value (<2.2e-16) menor de 0.05 no podemos rechazar la hipótesis nula. Esto indica que la variable quality no cumple el supuesto de normalidad.

```
# Comprobación de la normalidad para el resto de variables
library('nortest')
alpha = 0.05
wine_nor <- wine_df[, c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)]
col.names = colnames(wine_nor)
for (i in 1:ncol(wine_nor)) {
  if (i==1) cat("Las variables que no siguen una distribución normal:\n")
  if (is.integer(wine_nor[,i]) | is.numeric(wine_nor[,i])) {
      p_val = ad.test(wine_nor[,i])$p.value
        if (p_val < alpha) {
            cat(col.names[i])
                  # Format output
```

```
            if (i < ncol(wine_nor)) cat(", ")
        }
      }
  }
```

```
## Las variables que no siguen una distribución normal:
## fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.s
```

```
# correlación entre las variables sin las variables no significativas
library(corrplot)
```

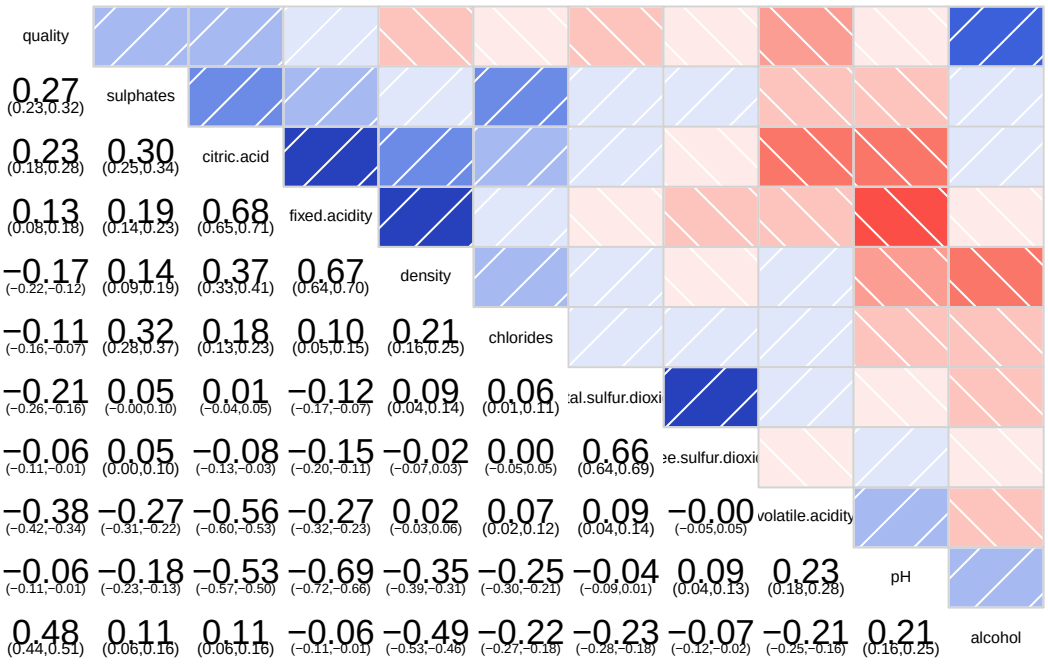
4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

```
## Warning: package 'corrplot' was built under R version 4.1.2
```

```
## corrplot 0.92 loaded
```

```
wine_df2 <- wine_df[, c(1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12)]

corrgram (wine_df2, order = TRUE , lower.panel=panel.conf)
```




```
# Modelo de regresión lineal quality-alcohol
Model_quality_alcohol <- lm(quality~alcohol, data=wine_df2 )
summary(Model_quality_alcohol)
```

```
##
## Call:
## lm(formula = quality ~ alcohol, data = wine_df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8475 -0.3871 -0.1641  0.5121  2.5841
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.89107    0.17486   10.81   <2e-16 ***
## alcohol      0.35968    0.01669   21.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7077 on 1583 degrees of freedom
## Multiple R-squared:  0.2268, Adjusted R-squared:  0.2263
## F-statistic: 464.4 on 1 and 1583 DF,  p-value: < 2.2e-16
```

```
# Modelo de regresión lineal quality-density
Model_quality_density <- lm(quality~density, data=wine_df2 )
summary(Model_quality_density)
```

```
##
## Call:
## lm(formula = quality ~ density, data = wine_df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7887 -0.6220  0.1652  0.4226  2.5103
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    78.52      10.58   7.423 1.86e-13 ***
## density       -73.12      10.61  -6.890 8.02e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.793 on 1583 degrees of freedom
## Multiple R-squared:  0.02912,    Adjusted R-squared:  0.0285
## F-statistic: 47.47 on 1 and 1583 DF,  p-value: 8.023e-12
```

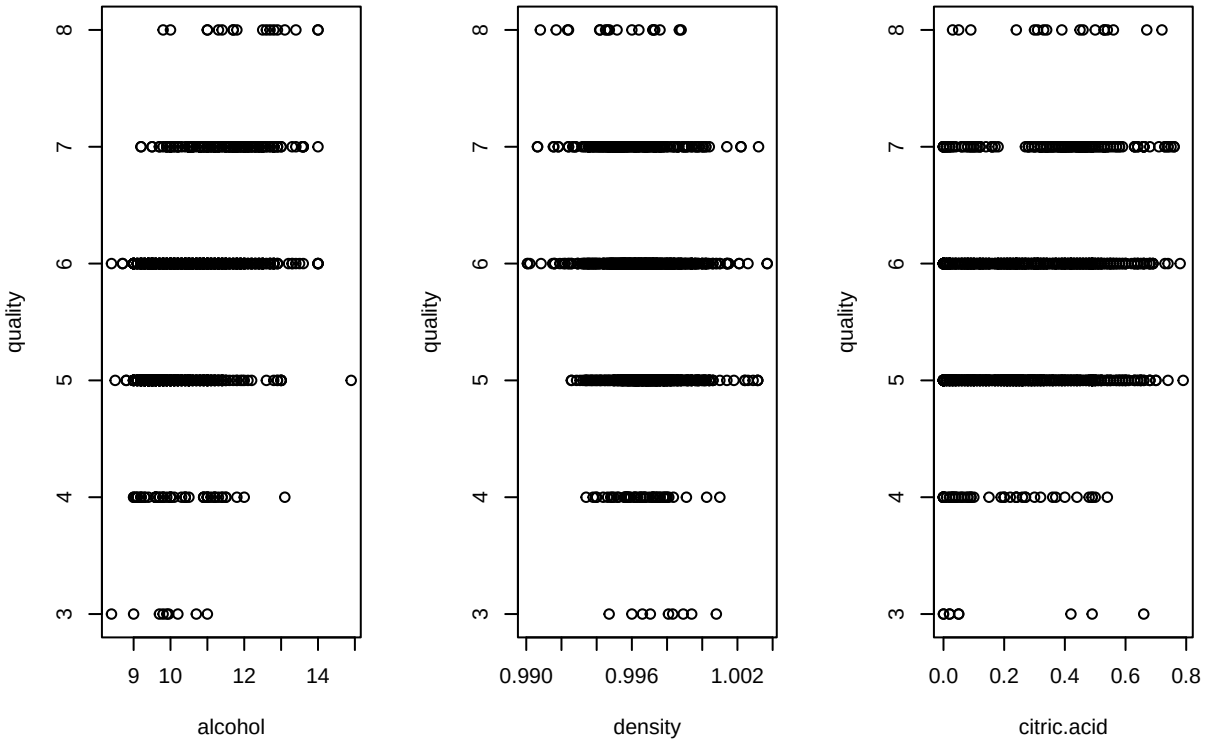
```
# Modelo de regresión lineal quality-citric.acid
Model_citric_acid <- lm(quality~citric.acid, data=wine_df2 )
summary(Model_citric_acid)
```

```
##
## Call:
## lm(formula = quality ~ citric.acid, data = wine_df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0127 -0.6017  0.1020  0.5034  2.5894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  5.38188    0.03379 159.282   <2e-16 ***
## citric.acid  0.95583    0.10182   9.387   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7833 on 1583 degrees of freedom
## Multiple R-squared:  0.05273,    Adjusted R-squared:  0.05213
## F-statistic: 88.12 on 1 and 1583 DF,  p-value: < 2.2e-16
```

El p-valor es menor al nivel de confianza en los tres modelos. La variable alcohol y ácido cítrico tienen un efecto positivo incremental mientras que la densidad afecta negativamente. Otra diferencia entre los modelos es que el modelo basado en el alcohol y ácido cítrico ajusta mejor que el modelo densidad ya que el R2 es mayor.

```
par(mfrow=c(1,3))
plot(quality~alcohol, data=wine_df2)
plot(quality~density, data=wine_df2)
plot(quality~citric.acid, data=wine_df2)
```



Las tendencias que se observan en los gráficos resultantes concuerdan con la interpretación que hemos efectuado con los modelos anteriores de la relación de las variables regresoras con la calidad. La variable calidad aumenta con el alcohol y ácido cítrico (primer y tercer gráfico) y decrece con la densidad (segundo gráfico)

```
# Modelo Regresión Multiple
Model_quality_multiple <- lm(quality~alcohol+citric.acid+fixed.acidity+chlorides+free.sulfur.dioxide+to
summary(Model_quality_multiple)
```

```
##
```

```
## Call:
## lm(formula = quality ~ alcohol + citric.acid + fixed.acidity +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates, data = wine_df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73763 -0.37221 -0.04749  0.46128  1.90307
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.858e+01  1.736e+01   1.647  0.09977 .
## alcohol               2.708e-01  2.343e-02  11.554  < 2e-16 ***
## citric.acid           5.123e-01  1.271e-01   4.029 5.86e-05 ***
## fixed.acidity         1.235e-04  2.406e-02   0.005  0.99591
## chlorides            -2.479e+00  4.373e-01  -5.668 1.71e-08 ***
## free.sulfur.dioxide   8.755e-03  2.200e-03   3.980 7.21e-05 ***
## total.sulfur.dioxide -5.334e-03  7.685e-04  -6.942 5.64e-12 ***
## density              -2.476e+01  1.772e+01  -1.397  0.16259
## pH                   -5.009e-01  1.857e-01  -2.698  0.00705 **
## sulphates             1.144e+00  1.127e-01  10.153  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.66 on 1575 degrees of freedom
## Multiple R-squared:  0.3308, Adjusted R-squared:  0.327
## F-statistic: 86.51 on 9 and 1575 DF,  p-value: < 2.2e-16
```

Las variables fixed.acidity y density no son significativas. El modelo de regresión es significativo (p-value: < 2.2e-16), con un R2 ajustado de 32.05 %. Las variables alcohol, citric.acid, fixed.acidity, free.sulfur.dioxide y sulphates influyen de forma positiva mientras que el resto de variables negativamente.

```
# Modelo Regresión Logística
high_wine <-ifelse(test=wine_df2$quality>=7,yes=1,no=0)
wine_df2$high_wine=high_wine
quality<- wine_df2$quality

logit_model_quality <- glm(formula=high_wine~alcohol+citric.acid+fixed.acidity+chlorides+free.sulfur.di
summary(logit_model_quality)
```

```
##
## Call:
## glm(formula = high_wine ~ alcohol + citric.acid + fixed.acidity +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates, family = binomial(logit), data = wine_df2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4023  -0.4338  -0.2327  -0.1106   2.8104
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.254e+02  8.829e+01   1.420  0.15549
## alcohol               8.613e-01  1.173e-01   7.340 2.14e-13 ***
## citric.acid           2.143e+00  6.778e-01   3.162  0.00157 **
```

```
## fixed.acidity         1.204e-01  1.169e-01   1.030  0.30298
## chlorides            -1.100e+01  3.505e+00  -3.140  0.00169 **
## free.sulfur.dioxide   3.015e-02  1.307e-02   2.306  0.02111 *
## total.sulfur.dioxide -2.960e-02  5.939e-03  -4.984 6.22e-07 ***
## density              -1.393e+02  9.014e+01  -1.545  0.12224
## pH                   -2.734e-01  9.623e-01  -0.284  0.77630
## sulphates             4.238e+00  5.379e-01   7.879 3.30e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1258.4  on 1584  degrees of freedom
## Residual deviance:  873.0  on 1575  degrees of freedom
## AIC: 893
##
## Number of Fisher Scoring iterations: 6
```

```
exp(cbind(coef(logit_model_quality),confint(logit_model_quality)))
```

```
## Waiting for profiling to be done...
```

```
##                                          2.5 %        97.5 %
## (Intercept)          2.909576e+54  3.010652e-21 8.739498e+129
## alcohol              2.366251e+00  1.884938e+00  2.987855e+00
## citric.acid          8.524033e+00  2.261142e+00  3.232897e+01
## fixed.acidity        1.127996e+00  8.966755e-01  1.418610e+00
## chlorides            1.662790e-05  8.485760e-09  7.631418e-03
## free.sulfur.dioxide  1.030605e+00  1.004543e+00  1.057483e+00
## total.sulfur.dioxide 9.708302e-01  9.591453e-01  9.817859e-01
## density              3.165256e-61 2.762615e-138  1.131058e+16
## pH                   7.607642e-01  1.130656e-01  4.938721e+00
## sulphates            6.928366e+01  2.420590e+01  2.011760e+02
```

Se observa que: - Un incremento de una unidad en alcohol (ajustando por las otras regresoras) aumenta el odds de calidad en 2.366251e+00. - Un incremento de una unidad en citric.acid (ajustando por las otras regresoras) aumenta el odds de calidad en 8.524033e+00. - Un incremento de una unidad en fixed.acidity (ajustando por las otras regresoras) aumenta el odds de calidad en 1.127996e+00. - Un incremento de una unidad en chlorides (ajustando por las otras regresoras) reduce el odds de calidad en 1.662790e-05. - Un incremento de una unidad en free.sulfur.dioxide (ajustando por las otras regresoras) aumenta el odds de calidad en 1.030605e+00. - Un incremento de una unidad en total.sulfur.dioxide (ajustando por las otras regresoras) reduce el odds de calidad en 9.708302e-01. - Un incremento de una unidad en density (ajustando por las otras regresoras) reduce el odds de calidad en 3.165256e-61. - Un incremento de una unidad en pH (ajustando por las otras regresoras) reduce el odds de calidad en 7.607642e-01. - Un incremento de una unidad en sulphates (ajustando por las otras regresoras) aumenta el odds de calidad en 6.928366e+01.

5. Representación de los resultados a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.

Se han ido generando los gráficos a lo largo de la práctica

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

Tras el modelo de regresión lineal múltiple así como el análsis de correlación y contrastes de hipótesis determinamos que las variables que influyen en la calidad del vino son el alcohol, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, pH y sulphates. El modelo de regresión, siendo un modelo significativo (p-value: < 2.2e-16), con un R2 ajustado de 32.05 %, indica que las variables alcohol, fixed.acidity, free.sulfur.dioxide y sulphates influyen de forma positiva mientras que el resto de variables negativamente. Esto también lo hemos comprobado a través del modelo logístico.

Contribuciones

- **Investigación previa:** Sandra Campos Suárez y Mª de los Ángeles García Carrión
- **Redacción de las respuestas:** Sandra Campos Suárez y Mª de los Ángeles García Carrión
- **Desarrollo del código:** Sandra Campos Suárez y Mª de los Ángeles García Carrión