# Práctica 2 - Limpieza y análisis de datos

Autores: Alberto Sarabia y Esther González

Junio 2021

## Contents

# 1. Descripción del dataset

## 1.1. Objetivo de la práctica

El objetivo principal de esta práctica consiste en el desarrollo de un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto, así como usar todas las herramientas de integración, limpieza, validación y análisis estudiadas durante el último bloque teórico.

## 1.2. Descripción y carga del dataset

Para el desarrollo de la práctica, se ha realizado a partir de la base de datos "winequality-red.csv" [1]. Los datos recopilados se corresponden a un conjunto de propiedades relacionada con una variedad de un vino tinto portugués muy concreto, llamado "vinho verde" [2]. A continuación, se carga el dataset:

```
# Cargar el dataset
data_wine<-read.csv("./winequality-red.csv",header=T,sep=",",
                    stringsAsFactors = T)

# Presentación 6 primeras filas
head(data_wine)
```

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

El conjunto de datos está compuesto por 12 variables distintas (atributos) y 1599 vinos tintos estudiados (muestras). Estos atributos se corresponden a las propiedades de los vinos y, además, hay una variable con la calificación de calidad de este vino concreto. Además para el estudio realizado durante esta práctica, se ejecutará todo el código con el lenguaje R.

## 1.3. Descripción de los atributos disponibles

Para poder realizar un estudio completo, es necesario conocer con mayor precisión la información obtenida de cada uno de los atributos disponibles. Por lo tanto, obtenemos el tipo de datos que nos ofrece cada una de las variables:

```r
# Variables y tipo de datos disponibles
sapply(data_wine, function(x) class(x))
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##            "numeric"            "numeric"            "numeric"
##       residual.sugar            chlorides  free.sulfur.dioxide
##            "numeric"            "numeric"            "numeric"
## total.sulfur.dioxide              density                   pH
##            "numeric"            "numeric"            "numeric"
##            sulphates              alcohol              quality
##            "numeric"            "numeric"            "integer"
```

Como podemos observar, todos los atributos son variables numéricas a excepción de la calidad, que es un valor entero. No obstante, necesitamos conocer, a nivel teórico, que es cada una de estas variables, así como sus valores típicos. Además, para realizar el estudio debemos conocer la unidad de medida de cada una de las muestras, por lo que también hay que detallar este punto en concreto.

Por lo tanto, se detallan todos los atributos disponibles:

- **fixed.acidity:** (numeric) todos los vinos contienen múltiples ácidos fijos (no volátiles), los cuales no se evaporan con facilidad. En este conjunto de datos, concretamente, el ácido fijo estudiado es el ácido tartártico, ya que se corresponde al ácido fijo con mayor aparición en los vinos (entre un 25% y un 30%). La unidad de medida de este atributo es el gramo por litro (g/l).
- **volatile.acidity:** (numeric) el ácido volátil de un vino es el ácido acético. La acidez volátil suele obtener un valor máximo de 0.6g/litro, ya que si superase este valor el vino podría obtener un sabor a vinagre y ser desagradable. Por ello, idealmente, el valor de este ácido será bajo. La unidad de medida de este atributo es el gramo por litro (g/l).
- **citric-acid:** (numeric) el ácido cítrico es un tipo de ácido fijo poco abundante en las uvas, pero que en cantidades moderadas puede añadir "frescura" en los vinos. En general, los valores correspondientes al ácido cítrico más comunes de los vinos están entre los 0.15g/l y 0.3g/l. La unidad de medida de este atributo es el gramo por litro (g/l).
- **residual.sugar:** (numeric) cantidad de azúcar residual en el vino después del proceso de fermentación. La clasificación de los vinos, acorde a la cantidad de azúcar residual, se mide acorde al dulzor, es decir que los vinos estarían subdivididos en los siguientes grupos: seco (menos de 4g/l), semiseco (entre 4 y 18g/l), semidulce (entre 18g/l y 45g/l) y dulce (más de 45g/l). La unidad de medida de este atributo es el gramo por litro (g/l).
- **free.sulfur.dioxide:** (numeric) cantidad de dióxido de azufre libre que contiene el vino. Esta característica impide el crecimiento de microbiano y la oxidación del vino. La unidad de medida de este atributo es el miligramo por litro (mg/l).
- **total.sulfur.dioxide:** (numeric) cantidad de dióxido de azufre libre y ligado. Cuanto mayor sea este valor, se acentúa en su aroma y su sabor. La unidad de medida de este atributo es el miligramo por litro (mg/l).
- **density:** (numeric) densidad del vino. Este valor se aproxima a la densidad del agua (997kg/mˆ3) en función del porcentaje de alcohol y del contenido de azúcar. La unidad de medida de este atributo es el gramo por centímetro cúbico (g/cmˆ3).
- **pH:** (numeric) grado de acidez o base del vino. Estos valores van de 0, siendo muy ácido, a 14, siendo muy básico. La mayoría de los vinos están entre los valores 3 y 4 de pH. Este parámetro no tiene unidades.

- **sulphates:** (numeric) cantidad de aditivos que contiene el vino que contribuye al nivel de gas de dióxido de azufre, que actúa como antimicrobiano y antioxidante. La unidad de medida de este atributo es el miligramo por litro (mg/l).
- **alcohol:** (numeric) cantidad de alcohol (grados) que tienen el vino.
- **quality:** (integer) valoración de la calidad del vino, en una escala de 1 a 10.

### 1.4. Importancia del dataset y problemas/preguntas a responder

Teniendo en cuenta la información disponible del dataset, tanto a nivel genérico con la descripción general (apartado 1.2) y el estudio detallado de los atributos disponibles (apartado 1.3), podemos comentar la transcendencia de este dataset.

La importancia del dataset partiría del interés que quiera aportar cada una de las personas sobre el estudio con este conjunto de datos. Este podría llegar a utilizarse dentro de un concepto laboral, para externos enólogos, hasta para gente con desconocimiento total sobre las propiedades del vino.

Para remarcar este rango de personas interesadas en el conjunto de datos, podremos dar un par de ejemplos. En el caso de los profesionales del mundo de la enología, podría ser útil para verificar aquella propiedad que más implica un aumento de calidad en los vinos tintos. Por otro lado, podría darse el caso que un inexperto en vinos tintos quiera sorprender a sus familiares con un vino de alta calidad. La utilización de un algoritmo predictivo con estos datos podría ayudar en este caso.

Como hemos observado, el conjunto de datos puede ser útil para un conjunto de usuarios muy amplios y puede dar respuesta a múltiples problemáticas y cuestiones. A continuación, se presentan algunas de las preguntas a estudiar durante esta práctica:

- ¿Qué características están ligadas directamente a que el vino sea de buena calidad?
- ¿Cuáles de los métodos utilizados para analizar este dataset tiene una mejor respuesta?
- ¿Los datos son adecuados para realizar un análisis profesional y no profesional?
- ¿El dulzor de los vinos es determinante para la cadidad de este?

## 2. Integración y selección de datos de interés

Durante este apartado se justificará la necesidad o no necesidad de la realización de una integración y selección de los datos ofrecidos mediante el dataset.

### 2.1. Integración de los datos

El conjunto de datos seleccionado para esta práctica es un único dataset, conformado por un conjunto de datos que no esta subdividido en distintos ficheros. Por lo tanto, no se aplicará ningún tipo de integración de los datos disponibles.

Sin embargo, cuando estemos trabajando con mayor profundidad en distintos modelos de análisis sí que se requiere una selección más concreta se llevará a cabo. En este caso, se detallará el motivo que nos lleva a tener que realizar una selección, así como su propia justificación en el apartado correspondiente.

### 2.2. Selección de los datos

A nivel general, y dado nuestro desconocimiento de la importancia que tiene cada una de las propiedades sobre el vino, hemos decidido no prescindir de ninguno de los atributos sin un previo análisis. Por ello, durante este primer apartado de selección de datos, se escogerá todo el dataset completo.

Sin embargo, cuando estemos trabajando con mayor profundidad en distintos modelos de análisis sí que se requerirá una selección más concreta. En este caso, se detallará el motivo que nos lleva a tener que realizar una selección, así como su propia justificación.

# 3. Limpieza de datos

Durante este apartado se van a preparar los datos para poder realizar un posterior análisis con el conjunto obtenido al final. Además, se estudiará y validará la coherencia de los datos disponibles con la información teórica descrita en el primer apartado de este mismo documento.

## 3.1. Validación de ceros y elementos vacíos

Para la correcta realización de los análisis, el primer paso será validar los ceros y cualquier elemento vacío que pueda contener el conjunto de datos.

Como se ha estudiado anteriormente, todas las muestras disponibles son valores numéricos (decimales o enteros), por lo que los valores desconocidos estarían asignados como NA.

```
# Comprobamos si hay valores vacíos en alguna de las columnas
#(mediante el valor NA)
colSums(is.na(data_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
```

Se valida que nuestra base de datos no tiene valores vacíos. A pesar de ello, se debe realizar otra comprobación para determinar que los datos disponibles no contienen valores vacíos. Al tratarse de un conjunto de datos numéricos, debemos verificar que no haya números como el 999 o 0, que determine este comportamiento.

En concreto, el valor 0 podría ser un valor válido para el conjunto de datos, por lo que no se estudiará esta opción como valor de elemento vacío. En cambio, el valor 999 no estaría en un margen aceptable y, en caso de estar disponible como muestra, se correspondería a un elemento vacío.

```
# Comprobamos si hay valores vacíos en alguna de las columnas
#(mediante el valor 999)
colSums(data_wine == 999)
```

```
##        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
```

En este caso, también obtenemos que no tenemos valores vacíos mediante la utilización de este valor.

Los casos verificados anteriormente son los más comunes para remarcar elementos vacíos dentro de un conjunto de datos. No obstante, para validar que no se utiliza ningún otro valor numérico distinto a 999, observamos el resumen de los datos disponibles.

```
# Resumen datos disponibles
summary (data_wine)
```

```
##  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
```

Ante todo, podemos determinar que no existe un valor numérico concreto que se identifique como indicador de elemento vacío.

Por otro lado, pero, a partir de este resumen también podemos destacar diversos puntos relacionados con las explicaciones teóricas detalladas durante la descripción de las propiedades del vino. En concreto, tenemos cuatro puntos destacados:

- **Valores ácido cítrico (citrical.acid)**

El valor medio de este atributo en la base de datos disponible es de 0.27g/l. Acorde al razonamiento y estudio realizado, la mayoría de los vinos tintos tienen un nivel de ácido cítrico es determinado entre 0.15g/l y 0.3g/l.

Por lo tanto, es coherente que el valor medio de un conjunto de vinos tintos se encuentre dentro de este mismo rango. De este modo, se ratifica y valida la base teórica descrita previamente.

- **Valores pH (pH)**

Los vinos tintos suelen tener valores de pH entre los valores 3 y 4. Esto se debe a que los valores muy extremos pueden llevar al vino a contener un sabor muy ácido o común, por lo que interesa que cada vino tenga su propio "carácter".

La base de datos que estamos estudiando tiene, en valor medio, niveles de pH de 3.31. Detallando con mayor profundidad sobre los datos proporcionados, el valor mínimo registrado de pH será de 2.7, muy cercano a 3, y su valor máximo es 4. Por lo tanto, podemos destacar que todos los vinos registrados de este conjunto no tendrán un nivel de acidez alto y, todos ellos, tendrán un "carácter" personal.

- **Valores ácido volátil (volatile.acidity)**

Recordamos que, idealmente, el ácido acético debe estar en niveles inferiores a 0.6g/l. No obstante, cuanto mayor es la cantidad de ácido acético, mayor será la sensación de gusto avinagrado del vino. Además, hay que tener en cuenta que también hay otra propiedad, el azúcar residual, que podría contrarrestar este sabor ligeramente.

De este modo, estudiando simplemente este atributo, sin combinación con otros, observamos que la mayoría de las muestras cumplen esta condición de no superar el valor 0.6g/l, ya que el valor del tercer quintil es de 0.64g/l y la media es de 0.52g/l. Por lo tanto, en la mayoría de los casos se cumple la no superación de este umbral, para estar dentro de los valores idóneos.

- **Valores del azúcar residual (residual.sugar)**

El azúcar residual de los vinos indica el nivel de dulzor de estos. En este caso, obtenemos que la muestra cuyo valor de azúcar residual es superior es de 15.5g/l, acorde al valor máximo ofrecido en el resumen.

Por lo tanto, se puede concluir que todos los vinos detallados en este dataset se pueden clasificar como vinos secos o semisecos, ya que para formar parte de este conjunto deben tener un valor de azúcar residual inferior a 18g/l.

Sin embargo, debemos remarcar que el valor máximo proporcionado dista bastante de mayoría del conjunto de muestras, ya que el valor medio es de 2.54g/l y tercer quintil es de 2.6g/l. El análisis detallado de esta observación comentada se realizará en el próximo apartado.

## 3.2. Identificación y tratamiento de valores extremos (*outliers*)

Los valores extremos (*outliers*) son aquellos valores incongruentes en comparación al conjunto de muestras de un mismo atributo. Para la detección de estos, se puede realizar mediante una representación grafica a partir de un *boxplot* o, aprovechando que estamos realizando el estudio en R, listar los valores de todos aquellos *ouliers*.

Hasta el momento, se ha seguido el proceso de limpieza de la base de datos original, pero no se ha requerido ninguna modificación sobre ésta. Llegados a este punto, pero, observamos que, a partir del resumen presentado en el apartado anterior, alguna modificación tendrá el dataset. Para mantener los datos originales intactos, realizaremos una copia del dataset original y cualquier modificación requerida queda en un nuevo conjunto de datos que llamaremos *data_wine_clean*.

```
# Copia del dataset original donde se irán realizando las modificaciones que
# vayamos detectando durante el proceso de limpieza
data_wine_clean <- data_wine
```

Antes de comenzar con las representaciones y listas de valores extremos, se debe validar para que variables debemos aplicar un análisis detallado y que tipo de ajuste del dato es el más apropiado. Para simplificar el estudio y comprensión de la toma de decisiones, el estudio de valores extremos se realizará de manera individual para cada uno de los atributos.

- **Ácido fijo (fixed.acidity) y ácido volátil (volatile.acidity)** El primer paso será validar con mayor profundidad los registros de estas dos variables. Para ello, volvemos a partir del resumen de estas variables concretas:

```
# Resumen del atributo
summary(data_wine_clean[,1:2])
```

```
##  fixed.acidity   volatile.acidity
##  Min.   : 4.60   Min.   :0.1200
##  1st Qu.: 7.10   1st Qu.:0.3900
##  Median : 7.90   Median :0.5200
##  Mean   : 8.32   Mean   :0.5278
##  3rd Qu.: 9.20   3rd Qu.:0.6400
##  Max.   :15.90   Max.   :1.5800
```

En el caso del ácido tartártico, los valores obtenidos son coherentes y válidos. Acorde a la legislación vigente, no hay un valor máximo que pueda descatalogar el vino según este valor. Además, aunque los valores más elevados de los ácidos fijos pudrieran identificarse como outliers, siguen siendo valores válidos y necesarios para los posteriores estudios.

En el segundo caso, para el ácido acético, idealmente el valor de este parámetro no debería superar los 0.6g/l, para evitar un sabor desagradable del vino. No obstante, puede que este valor sea superior y seguir una de estas dos opciones: estar catalogado como un vino de baja calidad o saciar este sabor no agradable mediante una mayor cantidad de azúcar residual.

Por otro lado, acorde a la legislación actual, el nivel máximo de ácidos volátiles que puede tener un vino es de 1.2g/l. En el resumen observamos que el valor máximo es de 1.58 g/l, por lo que uno o varias muestras superan el valor máximo de ácido permitido por la normativa. Por ello, se eliminarán todas aquellas muestras cuyo valor de acidez volátil sea superior al indicado por la legislación.

```
# Eliminacion de mustras con valores superiores a 1.2g/l de ácido volátil
data_wine_clean = subset(data_wine_clean, data_wine_clean$volatile.acidity < 1.2 )

# Resumen del atributo
summary(data_wine_clean[,2])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5257  0.6375  1.1850
```

Una vez eliminadas las muestras de vino que no cumplen la normativa, el número de registros disponibles es de 1595, por lo que se han eliminado cuatro muestras del conjunto de datos. Para validar la acción, observamos que el valor máximo disponible después del borrado es de 1.18g/l.

- **citric.acid**

Siguiendo el mismo análisis que para los atributos anteriores, es primer paso es validar la legislación vigente acorde a este ácido en concreto. Acorde a lo investigado, el nivel máximo de ácido cítrico permitido para cualquier vino en España es de 1g/l. Repasando el resumen de esta variable:

```
# Resumen del atributo
summary(data_wine_clean[,3])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2600  0.2714  0.4200  1.0000
```

Aunque el valor deseado, como se ha presentado al inicio del documento, debería estar contenido entre el 0.15g/l y 0.3g/l, los vinos puedes ser vendidos y distribuidos hasta un valor de 1g/l. Las muestras que vamos a estudiar tienen todos los niveles entre 0g/l y 1g/l. Por lo tanto, los valores para este atributo son válidos y no requieren ningún ajuste sobre los mismos.

- **residual.sugar**

Los niveles de azúcar residual simplemente permiten la categorización del vino acorde a su dulzor. Como ya se ha observado anteriormente, los valores ofrecidos de estas variables clasifican a todo el conjunto de muestras como vinos secos o semisecos.

Recordemos que la división para la clasificación de los vinos se corresponde a: seco (menos de 4g/l), semiseco (entre 4 y 18g/l), semidulce (entre 18g/l y 45g/l) y dulce (más de 45g/l). Acorde a las muestras disponibles en el conjunto de los datos tendremos el siguiente resumen:

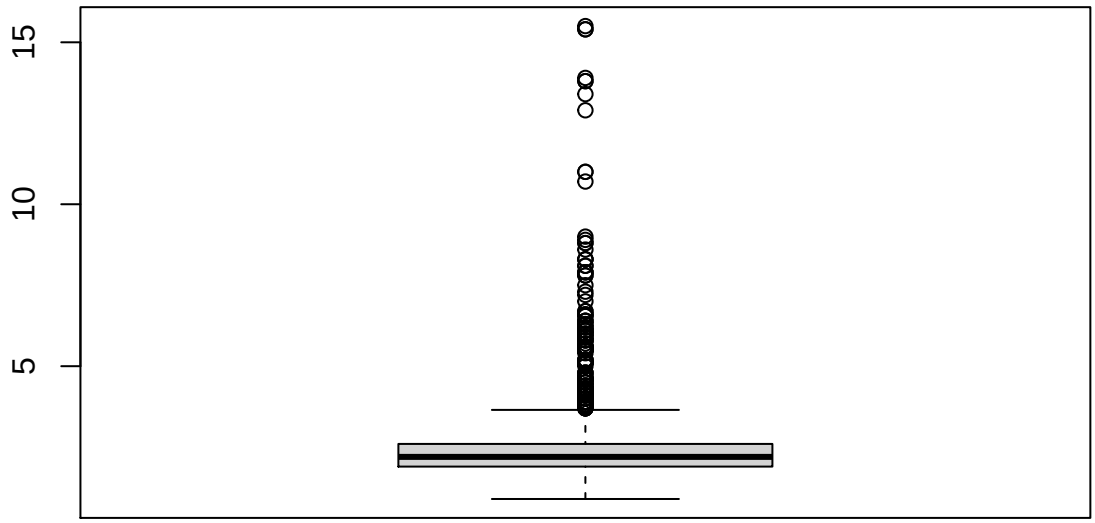
```
# Resumen del atributo
summary(data_wine_clean[,4])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.90    1.90    2.20    2.54    2.60   15.50
```

Cuando se observan los datos disponibles, se obtiene que algunas muestras de vino contienen valores extremos en este parámetro concreto. Eso se valida así porque entre valor máximo presentado y la media de la variable hay una diferencia superior a tres veces la desviación estándar de la variable a partir de la media del atributo.

Además, podemos observar mediante la representación y el listado de valores aquellos que serían valores extremos:

```
# Representación boxplot
boxplot(data_wine_clean[,4])
```



```
# Lista de números de valores extremos
boxplot.stats(data_wine_clean[,4])$out
```

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10  4.65
##  [13]  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00  4.00  4.00
##  [25]  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00  4.50  4.80  5.80
##  [37]  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70  4.50  6.70  6.60  3.70
##  [49]  5.20 15.50  4.10  8.30  6.55  6.55  4.60  6.10  4.30  5.80  5.15  6.30
##  [61]  4.20  4.20  4.60  4.20  4.60  4.30  4.30  7.90  4.60  5.10  5.60  5.60
##  [73]  6.00  8.60  7.50  4.40  4.25  6.00  3.90  4.20  4.00  4.00  4.00  6.60
##  [85]  6.00  6.00  3.80  9.00  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10
##  [97]  6.20  8.90  4.00  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70
## [109]  5.50  5.50  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90
## [121]  4.30 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80
## [133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75 13.80
## [145] 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10  7.80
```

Puesto que, como ya se ha comentado anteriormente, todas las muestras están dentro de valores coherentes y válidos como valores residuales de azúcar, no se va a tratar el listado de valores extremos.

En este caso, todos los datos son coherentes y necesarios para un futuro estudio, por lo que no se ajustará este atributo.

- **chlorides**

En el caso de la cantidad de sal no hay una legislación estricta que determine el máximo que puede tener un vino tinto. No obstante, raramente el nivel de sal supera los 3g/l.

```
# Resumen del atributo
summary(data_wine_clean[,5])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08745 0.09000 0.61100
```

Como podemos observar, en nuestro estudio los valores de sal de las muestras no superan la unidad de gramos por litro, por lo que todos los valores son validos para el estuido y, del mismo modo que hemos realizado anteriormente, no alteraremos los datos mediante ningún tipo de ajuste.

- **free.sulfur.dioxide y total.sulfur.dioxide**

A continuación, se realizará el estudio de las dos variables conjuntamente, ya que ambas están relacionadas directamente con las cantidades de dióxido de azufre, una en estado libre y la otra es su totalidad.

Acorde a la revista de enología científica y profesional, Acenología [7], la legislación actual indica que el nivel máximo de dióxido de azufre total del vino tinto es de 160mg/l, mientras que para los vinos blancos y rosados pueden llegar a niveles de 210mg/l.

```
# Resumen del atributo
summary(data_wine_clean[,6:7])
```

```
##  free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 1.00       Min.   :  6.00
##  1st Qu.: 7.00       1st Qu.: 22.00
##  Median :14.00       Median : 38.00
##  Mean   :15.89       Mean   : 46.47
##  3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :72.00       Max.   :289.00
```

De este modo, podemos observar que hay algunos valores de las muestras de vino que superan este nivel marcado por la legislación, en el total de dióxido de azufre. Por lo tanto, estas muestras que superan el umbral designado por las autorizaciones se eliminan del conjunto de datos.

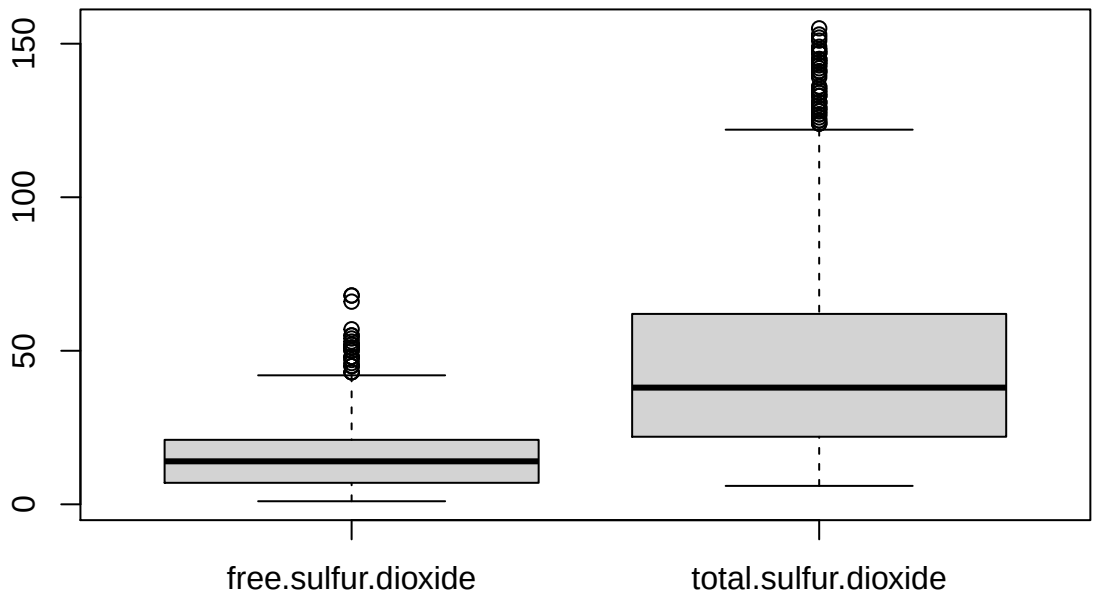
```
# Eliminacion de mustras con valores superiores a 1.2g/l de ácido volátil
data_wine_clean = subset(data_wine_clean,
                         data_wine_clean$total.sulfur.dioxide < 160 )

# Resumen del atributo
summary(data_wine_clean[,6:7])
```

```
##  free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 1.00       Min.   :  6.00
##  1st Qu.: 7.00       1st Qu.: 22.00
##  Median :14.00       Median : 38.00
##  Mean   :15.81       Mean   : 46.03
##  3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :68.00       Max.   :155.00
```

Una vez suprimida las muestras, observamos que la cantidad de muestras disponibles es de 1591, por lo que se han eliminado cuatro muestras. No obstante, repasando los datos disponibles, se tendría que todavía hay valores extremos en este conjunto de datos. Para visualizarlo gráficamente, se representa mediante los boxplots de las variables:

```
# Representación boxplot
boxplot(data_wine_clean[,6:7])
```




Teóricamente, se podrían tratar estos valores extremos detectados con múltiples técnicas; asignándoles el mayor valor, ajustándolo a la media, eliminando los registros, etc. Sin embargo, estos valores son coherentes con la legislación vigente, por lo que pueden darse estos casos.

Dada esta última observación, se concluye que no se suprimirán ninguno de los valores ya que estos representan una minoría de casos. Partiendo de los estudios teóricos investigados, la cantidad de vinos tintos que se encuentran con niveles de dióxido de azufre total superior a 100mg/l o de dióxido de azufre libre superior a 40mg/l son minoritarios, por lo que es coherente mantener este mismo comportamiento reflejado en la base de datos con la que se realizarán los análisis.

- **density y pH**

Primeramente, mostramos el resumen de estos dos atributos.

```
# Resumen del atributo
summary(data_wine_clean[,8:9])
```

```
##     density             pH
##  Min.   :0.9901   Min.   :2.740
##  1st Qu.:0.9956   1st Qu.:3.210
##  Median :0.9968   Median :3.310
##  Mean   :0.9968   Mean   :3.311
##  3rd Qu.:0.9978   3rd Qu.:3.400
##  Max.   :1.0037   Max.   :4.010
```

Idealmente los vinos deben tener una densidad próxima a la densidad del agua, es decir a 997kg/mˆ3 (0.997g/cmˆ3, acorde a las unidades de nuestra base de datos). Con los datos disponibles, tenemos que los

valores de las distintas muestras, en cuanto a la densidad, cumplen con esta aproximación a este valor y que no hay valores extremos en este parámetro.

Por otro lado, los niveles de pH de los vinos tintos tienen un rango válido entre 0 y 14. Por ello, podemos comprobar, mediante la observación de los valores mínimos y máximos del conjunto de datos, que todas las muestras están dentro de este conjunto. Es más, como ya se ha comentado en el apartado anterior, la mayoría de los calores están entre el rango de niveles de pH indicados como "ideales", es decir, entre 3 y 4.

- **sulphates**

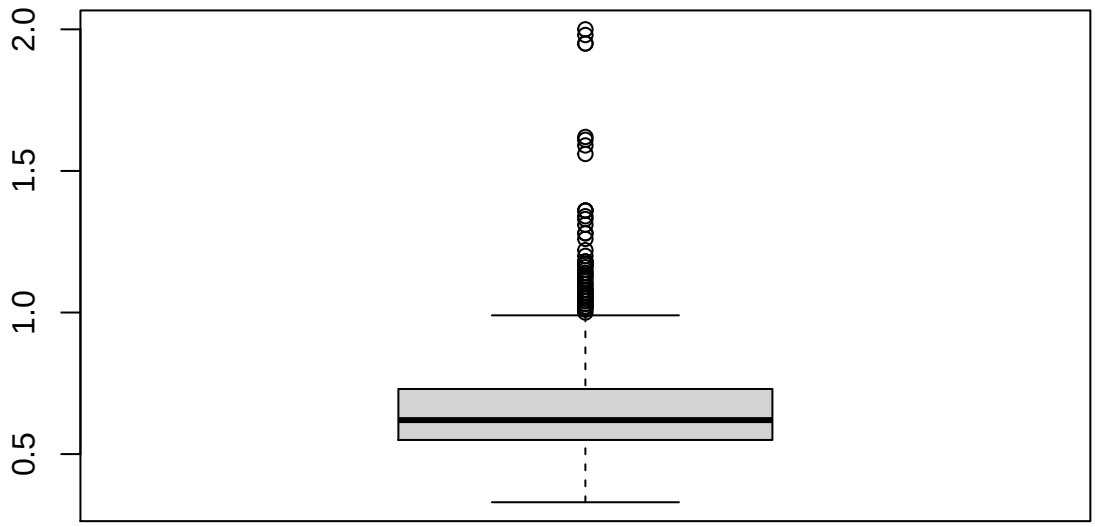
La cantidad de aditivos que contiene el vino queda determinada dentro de un margen fijo. Sin embargo, cuanto mayor sean estos niveles, podría llegar a influenciar sobre el sabor de los vinos, pero esto ya serán niveles bastante altos. Los datos disponibles serán:

```
# Resumen del atributo
summary(data_wine_clean[,10])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6589  0.7300  2.0000
```

En este caso concreto, la mayoría de las muestras son valores inferiores a 1mg/l, aproximadamente. No obstante, hay algunas muestras que tienen valores de sulfatos por encima, por lo que se tratarían de valores extremos. En concreto, tendremos que la representación de estos outliers, así como el listado de valores de las muestras de estos puntos serán:

```
# Representación boxplot
boxplot(data_wine_clean[,10])
```



```
# Listado valores extremos
boxplot.stats(data_wine_clean$sulphates)$out
```

```
##  [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08 1.59
## [16] 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13 1.07 1.06
## [31] 1.06 1.05 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
## [46] 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01
```

Con ambos resultados, gráficamente y con valores concretos, verificamos que la aproximación extraída en base al resumen de los datos es correcta, ya que los valores extremos son valores superiores a 1mg/l.

Por lo tanto, se considera que estas muestras deben ajustarse y tratarse, para que los vinos estudiados estén dentro de un rango determinado, acorde a un rango concreto en la cantidad de aditivos de las muestras.

El método utilizado para este tratamiento de este atributo es la eliminación de las muestras. La decisión de utilizar este método es porque la cantidad de muestras que requieren el ajuste es de 59, por lo que no estriamos tratando ni en 1% de los vinos disponibles y es preferible prescindir de ellos que modificarlos y después causar incongruencias con otras variables.

Para realizar este proceso de borrado realizamos las siguientes acciones:

```
# Eliminación de las muestras con valores extemos sobre el atributo
data_wine_clean = subset(data_wine_clean,
                         data_wine_clean$sulphates <
                           min(boxplot.stats(data_wine_clean$sulphates)$out))

# Resumen del atributo
summary(data_wine_clean[,10])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6371  0.7100  0.9900
```

Después de la eliminación de las muestras con valores extremos, nos quedamos con un conjunto de 1532 muestras y, como podemos corroborar, los datos que permanecen en el conjunto limpio son valores inferiores a 1mg/l.

- **alcohol**

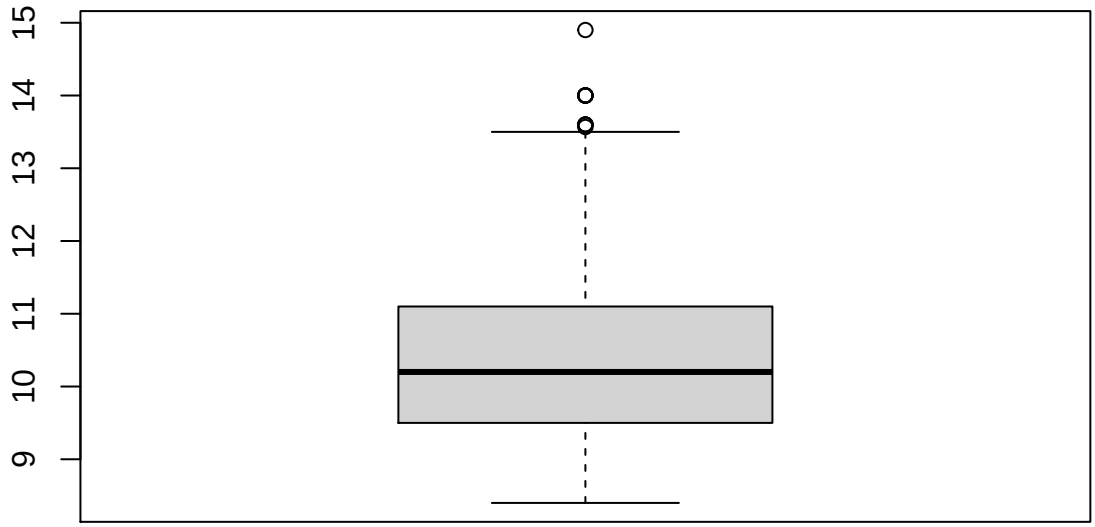
Los vinos, al tratarse de una bebida alcohólica, tiene unos grados de alcohol que depende de cada una de las muestras en concreto. Por lo general, los vinos no suelen tener grados superiores a 15, pero no hay un limite inferior, ya que también puede haber sin alcohol. En el conjunto del dataset tendremos que:

```
# Resumen del atributo
summary(data_wine_clean[,11])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.44   11.10   14.90
```

Todos los datos ofrecidos son niveles de alcohol inferiores a 15 grados y no destacan como valores muy alejados del margen, por lo que no se consideraran valores extremos a simple vista. Sin embargo, cuando realizamos el estudio matemáticamente, tenemos la siguiente representación:

```
# Representación boxplot
boxplot(data_wine_clean[,11])
```




Aunque gráficamente se representen "outliers", no se ajustarán ni trataran, porque tal y como se ha detallado previamente, son valores probables y realistas, dentro de un conjunto válido.

- **quality**

La validación de la calidad del vino es el valor numérico que le asignan profesionales. Idealmente, los vinos tendrían un 10, siendo esta la nota máxima, y en su minoría un 1, siendo la peor puntuación. Para las muestras, objeto de estudio, tenemos que:

```
# Resumen del atributo
summary(data_wine_clean[,11])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.44   11.10   14.90
```

Puesto que todos los valores están dentro del rango comentado, todas las muestras son óptimas para la realización de múltiples análisis, que se van a llevar a cabo durante el próximo apartado.

## 3.3. Creación del conjunto de datos limpio

Una vez finalizado el proceso de limpieza del dataset, este se guarda en un documento csv, bajo el nombre "winequality-red_clean.csv", para tenerlo disponible en futuras ocasiones. El proceso para guardarlo será:

```
# Guardar el CSV limpio
write.csv(data_wine_clean, "winequality-red_clean.csv")
```

# 4. Análisis de los datos

A lo largo de este apartado se van a realizar diferentes estudios sobre las variables para poder dar respuesta a los problemas identificados, así como extraer nuevas conclusiones a partir de los datos extraídos de cada método individualmente.

## 4.1. Selección grupos de datos

Para cada uno de los estudios que realizaremos con mayor profundidad en los próximos apartados, requeriremos tener una estructura sobre los datos y agrupaciones, ya que no siempre utilizaremos del mismo modo el dataset, ni todas las variables. Los objetivos concretos del análisis serán:

- Determinar que variable influye más sobre la calidad del vino tinto
- Validar la precisión de un modelo de clasificación sobre la variable calidad
- Verificar que propiedades del vino pueden predecir una mayor calidad

En primer lugar, se realizará un estudio de la correlación sobre la totalidad de las variables. Este proceso será posible, puesto que todas las variables son de tipo numérico. En caso de que el dataset hubiese tenido atributos categóricos y numéricos, deberíamos haber realizado un ejercicio de selección de grupo de datos, quedándonos con aquellas variables numéricas.

Para el modelo de clustering, se aplicará el modelo k-means. Idealmente, vamos a utilizar el conjunto completo, por lo que no requerirá ninguna selección de grupo. Además, no es necesario realizar una separación entre un conjunto de entreno y validación, como deberíamos realizar en el caso de un modelo supervisado.

Sin embargo, en este modelo sería interesante minimizar el rango de muestras donde agrupar el conjunto. Es decir, pasar la variable numérica calidad a un conjunto más acotado (por ejemplo, calidad baja y alta). El estudio se realizará en dos partes, y está división estará más detallada en el apartado correspondiente, junto el análisis del modelo.

Por último, se realizará la verificación descrita mediante el estudio de modelos de regresión lineal y logística. Del mismo modo que en el estudio anterior, se determinaran dos etiquetas en la variable calidad, para disminuir la variedad de valores en este conjunto a 2 y obtener conclusiones mas claras.

Aunque estos tres puntos son cuestiones muy concretas para los modelos que se van a desarrollar, durante el propio estudio también se darán respuesta a las problemáticas destacadas en el apartado 1.4

## 4.2. Normalidad y homogeneidad de la varianza

Para determinados estudios es muy útil cuando la distribución de las variables sigue una distribución normal o, en otras palabras, una distribución gaussiana. Para la realización de este estudio, lo más común es la utilización de dos tests: Kolmogorov-Smirnov y de Shapiro-Wilk.

Comparativamente hablando, el test de Shapiro-Wilk suele darse como es test más fiable, aunque también es posible la utilización del otro. Por este razonamiento comentado, para nuestro estudio aplicaremos el test de Shapiro-Wilk.

Primeramente, definiremos las hipótesis nula ($H_0$) y alternativa ($H_1$):

- $H_0$: Distribución normal variables

- $H_1$: Distribución variables no normal

El término que tendremos en cuenta es el valor p-value. Para dar por valida la hipótesis nula este valor debe superar el valor de la significancia, que sería $\alpha$= 0.05. Aplicando el test y mostrando los resultados sobre todos los atributos disponibles en nuestro conjunto de datos:

```
# Cálculo test de Shapiro-Wilk
test.results.fa <- shapiro.test(data_wine_clean$fixed.acidity)$p.value
test.results.va <- shapiro.test(data_wine_clean$volatile.acidity)$p.value
test.results.ca <- shapiro.test(data_wine_clean$citric.acid)$p.value
test.results.rs <- shapiro.test(data_wine_clean$residual.sugar)$p.value
test.results.chl <- shapiro.test(data_wine_clean$chlorides)$p.value
test.results.fsd <- shapiro.test(data_wine_clean$free.sulfur.dioxide)$p.value
test.results.tsd <- shapiro.test(data_wine_clean$total.sulfur.dioxide)$p.value
test.results.d <- shapiro.test(data_wine_clean$density)$p.value
test.results.ph <- shapiro.test(data_wine_clean$pH)$p.value
test.results.sul <- shapiro.test(data_wine_clean$sulphates)$p.value
test.results.al <- shapiro.test(data_wine_clean$alcohol)$p.value

# Guardar todos los resultados de la p.value en una misma lista
p.values <- list(test.results.fa, test.results.va, test.results.ca,
                 test.results.rs, test.results.chl, test.results.fsd,
                 test.results.tsd, test.results.d, test.results.ph,
                 test.results.sul, test.results.al)

# Mostrar listado de valores p-value para todas las variables independientes
cat ("La lista de valores p.value es: \n")
```

```
## La lista de valores p.value es:
```

```
for (i in p.values){
  cat ("\t", i, "\n")
}
```

```
##   1.983811e-24
##   5.678492e-12
##   8.446856e-22
##   9.028917e-52
##   6.909347e-46
##   4.52291e-30
##   2.455928e-31
##   7.006569e-08
##   1.219643e-07
##   1.565172e-17
##   4.230936e-26
```

Los resultados muestran que todos nuestros valores tienen un valor p-value bastante inferior al de la significancia. Así que, en base al resultado obtenido, deberíamos rechazar la hipótesis nula y quedarnos con la alternativa y afirmar que ninguna de las variables sigue una distribución normal. Sin embargo, podemos afirmar que todas las muestras disponibles siguen una distribución normal a partir de la aplicación del teorema central del límite.

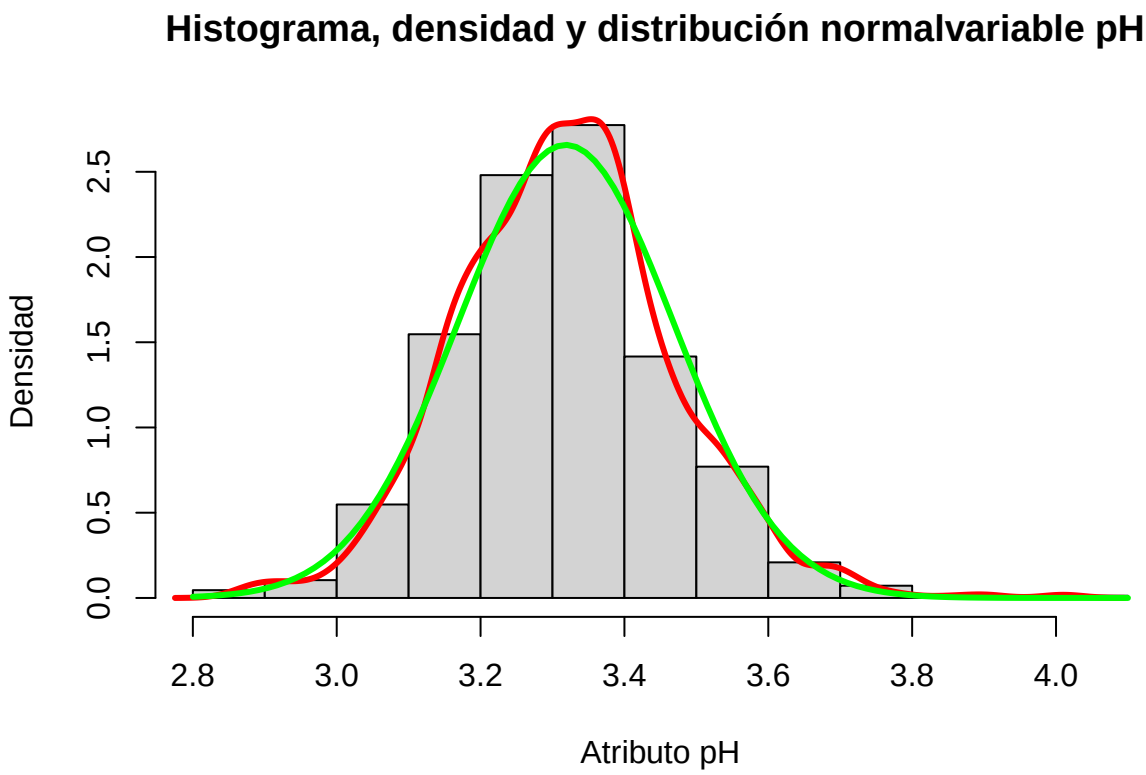
El teorema del límite central permite afirmar que un conjunto de muestras sigue una distribución normal, siempre que la cantidad de registros sea suficientemente grande (superior a los 30 elementos). Nuestra base

de datos, objeto de estudio, consta de 1532 muestras, por lo que cumple las condiciones y se aplica este teorema en todos los atributos.

Para dejar reflejado este proceso también de forma mas visual, hemos tomado la variable pH como ejemplo. Se representan sus datos en forma de histograma y dibujamos la línea de densidad de la variable, así como de la normal correspondiente.

```
# Ejemplo visual distribucion normal sobre la variable PH
ph.mean = mean(data_wine_clean$pH)
ds.mean = sd(data_wine_clean$pH)

hist(data_wine_clean$pH, freq = F,
     ylab = "Densidad", xlab = "Atributo pH",
     main = "Histograma, densidad y distribución normalvariable pH ")
lines(density(data_wine_clean$pH), col = "red", lwd = 3)
curve(dnorm(x, ph.mean, ds.mean), add = TRUE, col = "green", lwd = 3)
```




Como se observa, la densidad de los datos sigue, como ya hemos justificado anteriormente, esta distribución normal comentada.

Para el estudio sobre la homocedasticidad debemos comprender que estamos validando con mas profundidad. Esta prueba consiste en comprobar la igualdad de varianzas entre los grupos que se vayan a comparar. Puesto que no tendría sentido realizar un estudio de todos los atributos entre sí, realizaremos el estudio sobre las relaciones que se llevarán a cabo durante los modelos detallados en los próximos apartados. La justificación de la elección de estas variables en concretas estará detallada en su apartado correspondiente (apartado 4.3).

En todos los casos, al tratarse de variables numéricas, se tendrá que realizar la prueba mediante el test de Fligner-Killeen.

En concreto, para cada conjunto de variables a estudiar tendremos:

- **Alcohol ~Quality**

Para el estudio de homocedasticidad entre las variables calidad y alcohol, tendremos que las hipótesis definidas serán:

- $H_0$: Homocedasticidad de la variables
- $H_1$: No homocedasticidad de la variables

Aplicando el test de Fligner-Killeen:

```
library(car)

# Aplicación test de Fligner-Killeen
fligner.test(alcohol ~ quality, data = data_wine_clean)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  alcohol by quality
## Fligner-Killeen:med chi-squared = 122.45, df = 5, p-value < 2.2e-16
```

Dado que el valor obtenido del p-value es inferior a 0.05, valor de la significancia, podemos rechazar la hipótesis nula, por lo que la variable *alcohol* presenta varianzas estadísticamente diferentes para los diferentes grupos de calidad.

- **Sulphates ~Quality**

Siguiendo la misma estructura que en el apartado anterior, tendremos que:

Las hipotesis y el resultado del test serán:

- $H_0$: Homocedasticidad de la variables
- $H_1$: No homocedasticidad de la variables

```
library(car)

# Aplicación test de Fligner-Killeen
fligner.test(sulphates ~ quality, data = data_wine_clean)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  sulphates by quality
## Fligner-Killeen:med chi-squared = 19.424, df = 5, p-value = 0.001602
```

Del mismo modo que para la variable alcohol, valor obtenido del p-value es inferior a 0.05, valor de la significancia, y se rechaza la hipótesis nula. De este modo, se puede afirmar que la variable *sulphates* presenta varianzas estadísticamente diferentes para los diferentes grupos de calidad.

- **Volatile.acidity ~Quality**

Finalmente, aplicamos el mismo proceso para la variable del ácido volátil. La hipótesis y resultados de test serán:

- $H_0$: Homocedasticidad de la variables
- $H_1$: No homocedasticidad de la variables

```
library(car)

# Aplicación test de Fligner-Killeen
fligner.test(volatile.acidity ~ quality, data = data_wine_clean)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  volatile.acidity by quality
## Fligner-Killeen:med chi-squared = 27.651, df = 5, p-value = 4.259e-05
```

En este caso, valor obtenido del p-value también es inferior a 0.05, valor de la significancia, y se rechaza la hipótesis nula. De este modo, se puede afirmar que la variable *volatile.quality* presenta varianzas estadísticamente diferentes para los diferentes grupos de calidad.

## 4.3.Métodos de análisis

En este apartado se realizarán distintos análisis sobre el conjunto de datos. Entre ellos tendremos: correlaciones, regresiones y modelos no supervisados.

#### Análisis estadístico inferencial: Correlación

Mediante el estudio de los coeficientes de correlación, podremos determinar la asociación entre variables y, de este modo, poder concluir cuales son las propiedades del vino tinto que más influencia tienen sobre la clasificación en la calidad de estos mismos.

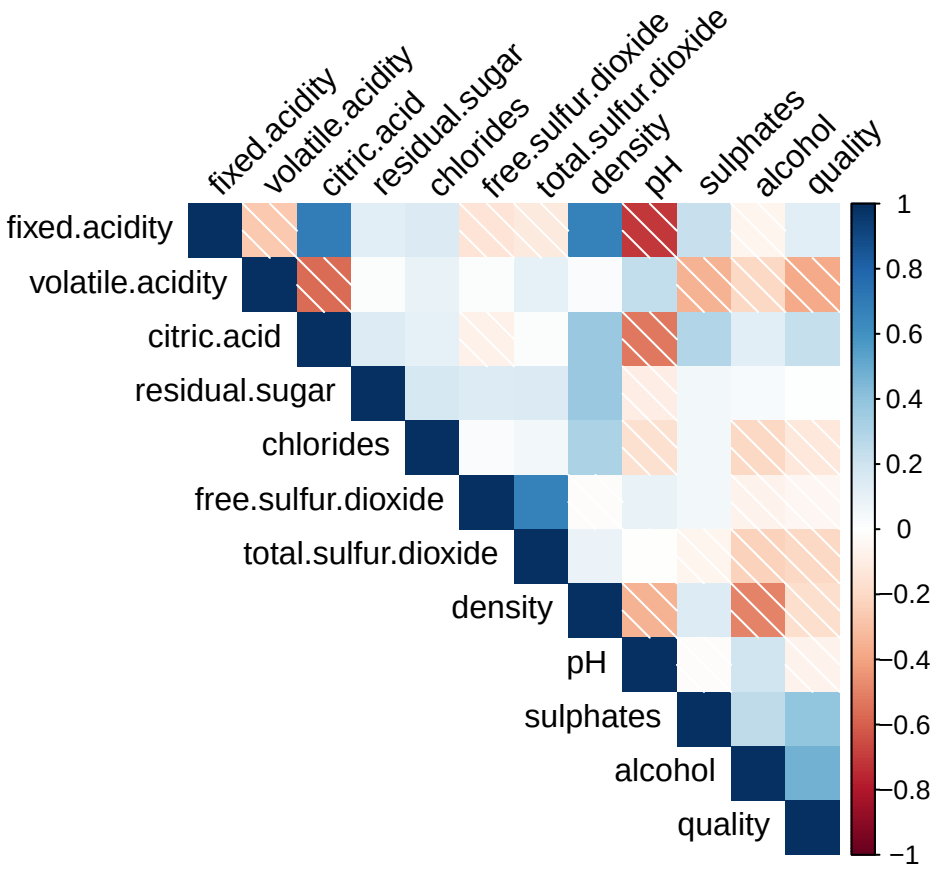
Para poder hacer el análisis de modo visual, se realizará una matriz de correlación de todas las variables disponibles. Esto es posible, ya que la totalidad de las muestras son de tipo numérico y, utilizando la base de datos original, después del proceso de limpieza, no se requiere la transformación de los atributos a rangos categóricos.

Además, los coeficientes de correlación extraídos serán de tipo Pearson. Este método es el más utilizado cuando las variables están relacionadas linealmente. Además, una de las condiciones a cumplir para su utilización es que la distribución entre las variables siga una distribución normal.

Como ya hemos verificado en el apartado 4.2, en nuestro dataset todos los atributos siguen esta distribución gaussiana, por lo que se cumple la condición y podemos aplicar el método Pearson para la creación de la matriz de correlación. Obtendremos la siguiente representación:

```
library (ggplot2)
library(corrplot)

wine_data.cor <- round (cor(data_wine_clean, method = "pearson"), digits = 2)
corrplot(wine_data.cor, method = "shade", tl.col = "black", tl.srt = 45, type = "upper")
```




En presencia de los resultados, y recordando que queremos determinar qué características o propiedades del vino tinto son las más influyentes cuando queremos extraer la calidad de este, repasamos aquellos atributos con mayor correlación a esta variable.

Destacaríamos que las variables *alcohol*, *sulphates* y *volatile.acidity* son aquellas mas correladas con la variable calidad. En los dos primeros casos estaríamos hablando de una correlación positiva, entre 0.5 y 0.7 ambas, y en el caso del ácido volátil el valor de la correlación seria de 0.8 con signo negatividad.

Por el contrario, también obtendríamos los atributos menos correlados y que, por lo tanto, son los menos influyen en la calidad del vino tinto. En concreto serían el azúcar residual y el dióxido de azufre libre. En ambos casos el color del cuadro en la representación es, prácticamente, blanco, por lo que la correlación de estos con la variable calidad es 0.

**Análisis mediante modelos no supervisados: k-mean**

El modelo k-mean se corresponde a un método de agrupamiento, cuyo objetivo es la partición de un conjunto en un determinado número de observaciones (n) en k subconjuntos (siendo k<n).

En este caso, este método se clasifica dentro de modelos no supervisados, ya que para el entreno del modelo no se requiere la participación de la variable dependiente sobre la que, finalmente, queremos realizar la agrupación.

En nuestros datos actuales queremos aplicar el método para tener una agrupación sobre la variable calidad. Puesto que la variable es de tipo entero, con un rango entre el 3 y el 8, la agrupación debería ser, idealmente, sobre 6 conjuntos, que dentro del modelo serían los 6 clusters.

No obstante, para realizar un análisis de agrupación parece más lógico, la utilización de clasificaciones más genéricas. En otras palabras, hemos variado la variable edad en 2 y 3 etiquetas y estudiado que sucede en cada caso.

- **División calidad en 2 etiquetas (Alta calidad y Baja calidad)**

En este primer caso vamos a suponer que la calidad del vino puede ser baja, cuando el valor numérico de la calidad se inferior a 5, o alta, en caso opuesto. Para ello, se ajusta este parámetro:

```
# Ajustar la variable calidad en 2 etiquetas posibles
wine_calidad2<- data_wine_clean
wine_calidad2$quality <- ifelse(wine_calidad2$quality < 5,
                                "Baja calidad","Alta Calidad")
```

De esta manera, idealmente nuestra agrupación se realizará en 2 clusters, alta y baja calidad. Y este será el parámetro k que asignaremos al modelo. Previamente, pero, recordamos que se trata de un modelo no supervisado, por lo que se eliminarán los valores de la calidad que acabamos de asignar y aplicamos el modelo para analizar su precisión:

```
# Aplicación del modelo k-means
data.cl<-wine_calidad2
data.cl$quality<-NULL
kmeans.res<-kmeans(data.cl,2)
table(wine_calidad2$quality,kmeans.res$cluster)
```

```
##
##                    1    2
##   Alta Calidad 1049  424
##   Baja calidad   50    9
```

Calculando la precisión del modelo:

```
correct  = (1049+9)
non_correct = 424+50
result1 = 100*(correct/(correct+non_correct))
cat("La precisión del modelo es del", round(result1,2), "%")
```

```
## La precisión del modelo es del 69.06 %
```

Puesto que estamos haciendo un estudio sobre un conjunto bastante amplio en cuanto a variables, es normal que la precisión no sea extremadamente alta. A excepción de que todos los atributos estén muy relacionados, que verificando el apartado anterior no estaríamos en este caso tampoco.

Por ello, puesto que el resultado de la precisión es del 70%, se puede concluir como un buen resultado.

Ahora, comprobemos que sucede si ampliamos el número de agrupaciones.

- **División calidad en 3 etiquetas (Alta calidad, Calidad Media y Baja calidad)**

Siguiendo el mismo proceso, ahora vamos a clasificar la calidad del vino en 3 opciones: baja calidad (3-4), calidad media (5-6) o alta calidad (7-8).

```
# Ajustar la variable calidad en 3 etiquetas posibles
wine_calidad3 <- data_wine_clean
wine_calidad3$quality <- ifelse(wine_calidad3$quality < 5,
                                "Baja calidad",
                                ifelse(wine_calidad3$quality < 7,
                                "Calidad Media","Alta Calidad"))
```

En este caso, la modificación que realizaremos es que ahora el número de clusters en los que queremos realizar la agrupación, idealmente, es de 3. Obtenemos la tabla y, también, la precisión del modelo:

```
# Aplicación del modelo k-means
data.cl<-wine_calidad3
data.cl$quality<-NULL
kmeans.res<-kmeans(data.cl,3)
table(wine_calidad3$quality,kmeans.res$cluster)
```

```
##
##                    1   2   3
##   Alta Calidad    15 135  55
##   Baja calidad     4  38  17
##   Calidad Media  218 611 439
```

```
correct  = (15+38+439)
non_correct = (135+55+4+17+218+611)
result2 = 100*(correct/(correct+non_correct))
cat("La precisión del modelo es del", round(result2,2), "%")
```

```
## La precisión del modelo es del 32.11 %
```

- **Sin aplicar divisiones**

Por último, comprobamos que sucede si no hubiésemos realizado una agrupación en la variable calidad. Es decir, mantenemos el valor numérico de 3 a 8, por lo que el número de clusters, idealmente, serían 6.

```
# Aplicación del modelo k-means
data.cl<-data_wine_clean
data.cl$quality<-NULL
kmeans.res<-kmeans(data.cl,6)
table(data_wine_clean$quality,kmeans.res$cluster)
```

```
##
##       1   2   3   4   5   6
##   3   0   0   5   1   3   0
##   4   4   6  20  12   8   0
##   5 105 105 128 135 113  63
##   6  45  99 151 177 144   3
##   7  11  15  72  50  40   0
##   8   2   1   8   3   3   0
```

```
correct  = (0+20+113+45+15+50)
non_correct = (5+3+1+8+4+6+12+63+128+105+105+135+3+151+144+99+177+50+11+40+72+8+3+2+1)
result3 = 100*(correct/(correct+non_correct))
cat("La precisión del modelo es del", round(result3,2), "%")
```

```
## La precisión del modelo es del 15.39 %
```

Volvemos a obtener el resultado esperado. Es decir, una peor precisión en el modelo. Desde la primera conclusión extraída, queda reflejado que la poca correlación entre los datos no beneficia cuando queremos realizar una agrupación por clusters, por ello, cuanto menor sean las agrupaciones requeridas mejor precisión tendrá nuestro modelo. Siempre y cuando estas agrupaciones comentadas tengan sentido, como seria en este caso presentado.

**Análisis estadístico inferencial: Regresión**

**Regresión lineal** Los modelos de regresión lineal aproximan la relación de dependencia entre una variable dependiente y un conjunto de variables independiente a partir de una recta. En este primer caso, pero, se estudiará mediante una única variable independiente, por lo que estaremos ante un caso de regresión lineal simple.

Para escoger las variables adecuadas a analizar con el modelo de regresión lineal nos fijamos en la tabla de correlaciones, seleccionando las variables que tengan una correlación mayor o igual a 0,5 o -0,5 debido a que esto significa que su correlación es fuerte. La variable que seleccionaremos para el análisis es la de quality, puesto que queremos obtener la relación entre esta y las variables volatile.acidity, alcohol y sulphates.

- **Análisis regresión lineal volatile.acidity**

```
# Modelo de regresión lineal
reg_lin_simp <- lm(  data_wine_clean$volatile.acidity ~ data_wine_clean$quality,
                     data = data_wine_clean, na.action = na.exclude)

# Resumen del modelo
summary(reg_lin_simp)
```

```
##
## Call:
## lm(formula = data_wine_clean$volatile.acidity ~ data_wine_clean$quality,
##     data = data_wine_clean, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43181 -0.11864 -0.00022  0.10136  0.59978
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.988149   0.029232   33.80   <2e-16 ***
## data_wine_clean$quality -0.081585   0.005133  -15.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1611 on 1530 degrees of freedom
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.1412
## F-statistic: 252.6 on 1 and 1530 DF,  p-value: < 2.2e-16
```

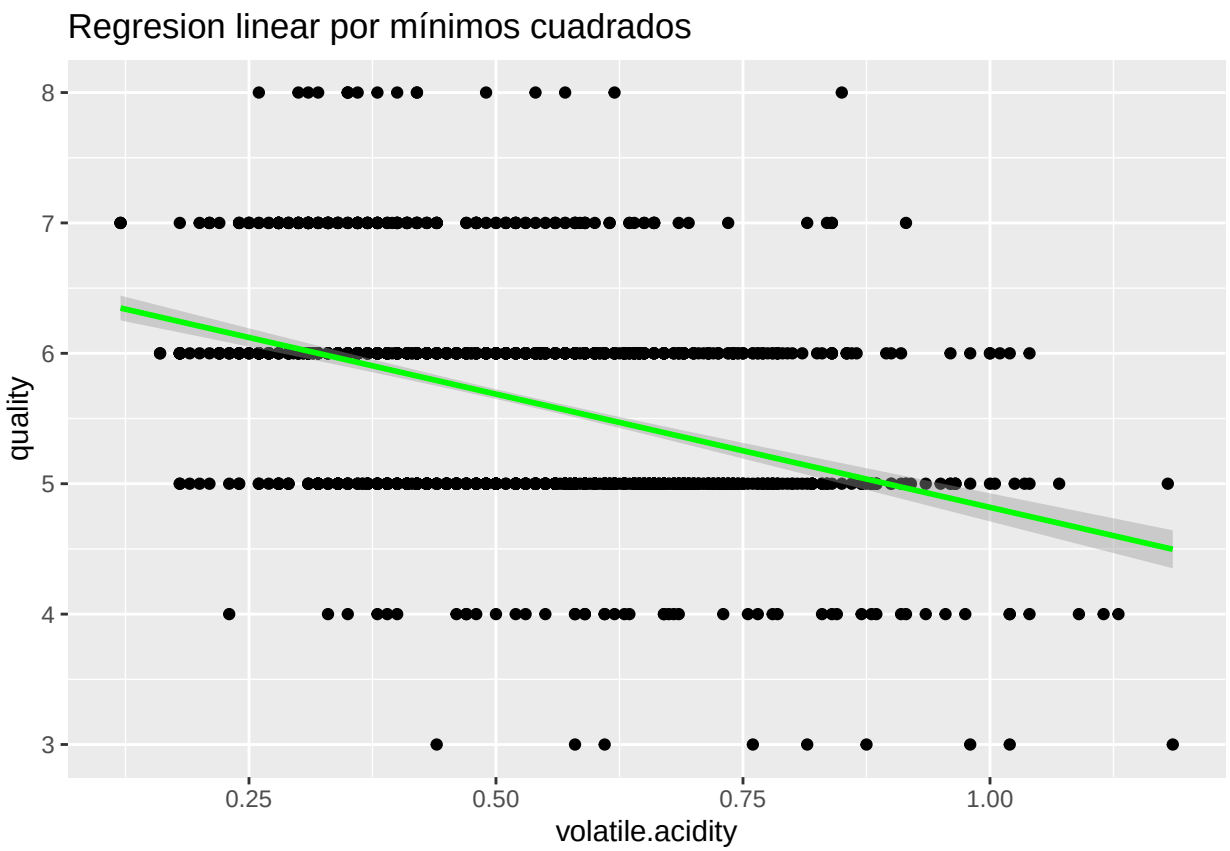
La relación de dependencia comentada, puesto que estamos ante un modelo de regresión lineal simple, estará compuesto por una sola variable independiente ($X_i$), por lo tanto, la aproximación de relación estará definida por recta, cuya fórmula seguirá la estructura:

- $Y_i = \beta_0 + \beta_1 * X_i$

data_wine_clean$quality i = 1.014395 + (−0.081585 ∗ $data_wine_clean$volatile.acidityi)

Como podemos observar la Multiple R-squared es el coeficiente que determina la calidad del modelo y toma valores entre 0 y 1, dando para este modelo un valor de 0.1417 podemos entender que las variables no están fuertemente correlacionadas relacionadas La representación visual del modelo obtenido sería la siguiente.

```
# Representacón regresión lineal
ggplot(data_wine_clean, aes(data_wine_clean$volatile.acidity,data_wine_clean$quality )) +
  geom_point() +
  geom_smooth (method ='lm', colour = "Green") +
  ggtitle("Regresion linear por mínimos cuadrados") +
  xlab('volatile.acidity') + ylab('quality')
```




En la gráfica vemos que las variables no están correlacionadas y esto nos indica que, al no superar la mayoría de las observaciones, el valor normal de 0,6g/l o esté pudiéndose contrarrestar con otros valores más altos como el azúcar residual este valor no es un buen valedor o clasificador de la calidad de los vinos.

- **Analisis regresión lineal alcohol**

```
# Modelo de regresión lineal
reg_lin_simp <- lm(  data_wine_clean$alcohol ~ data_wine_clean$quality ,
                     data = data_wine_clean, na.action = na.exclude)

# Resumen del modelo
summary(reg_lin_simp)
```

```
##
## Call:
## lm(formula = data_wine_clean$alcohol ~ data_wine_clean$quality,
##     data = data_wine_clean, na.action = na.exclude)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2679 -0.6363 -0.2047  0.5637  4.8637
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.87822    0.17062   40.31   <2e-16 ***
## data_wine_clean$quality   0.63162    0.02996   21.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9406 on 1530 degrees of freedom
## Multiple R-squared:  0.2251, Adjusted R-squared:  0.2246
## F-statistic: 444.5 on 1 and 1530 DF,  p-value: < 2.2e-16
```

La relación de dependencia comentada, puesto que estamos ante un modelo de regresión lineal simple, estará compuesto por una sola variable independiente (Xi), por lo tanto, la aproximación de relación estará definida por recta, cuya fórmula seguirá la estructura:

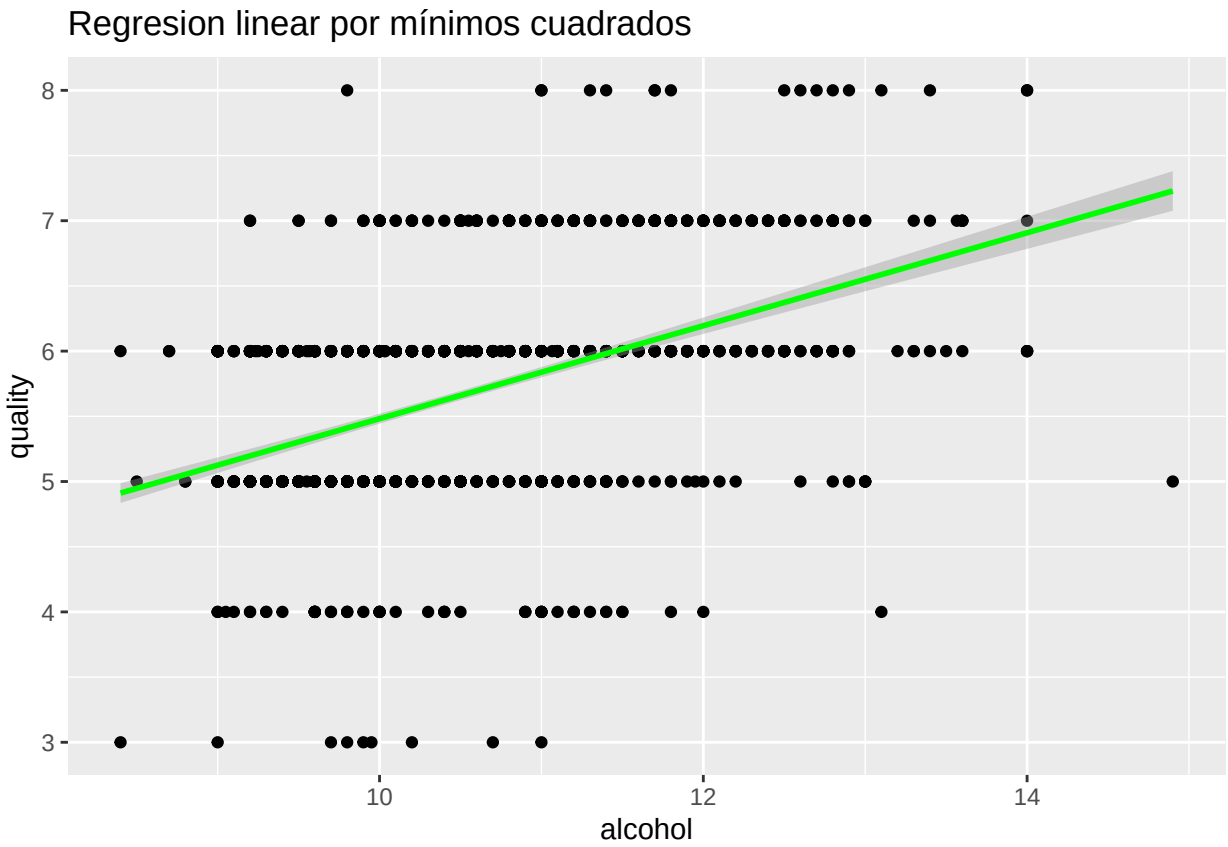
- $Y_i = \beta_0 + \beta_1 * X_i$

En concreto, para el casi que estamos estudiando y, gracias a los datos obtenidos, a partir del modelo:

data_wine_clean$quality i = 0.62567 + 0.62567 * data_wine_cleanalcoholi

Como podemos observar la Multiple R-squared es el coeficiente que determina la calidad del modelo y toma valores entre 0 y 1, dando para este modelo un valor de 0.2249 podemos entender que las variables no están fuertemente correlacionadas relacionadas La representación visual del modelo obtenido sería la siguiente.

```r
# Representacón regresión lineal
ggplot(data_wine_clean, aes(data_wine_clean$alcohol,data_wine_clean$quality )) +
  geom_point() +
  geom_smooth (method ='lm', colour = "Green") +
  ggtitle("Regresion linear por mínimos cuadrados") +
  xlab('alcohol') + ylab('quality')
```




En la gráfica vemos que las variables no están correlacionadas esto nos indica que a pesar de tener una alta correlación en la matriz por encima del 0,5 en realidad estas variables no están para nada relacionadas una vez realizado el estudio lo que nos indica que el alcohol o la cantidad de alcohol que tenga un vino no refleja si es bueno o malo.

- **Analisis regresion lineal sulphates**

```
# Modelo de regresión lineal
reg_lin_simp <- lm(  data_wine_clean$sulphates ~ data_wine_clean$quality ,
                     data = data_wine_clean, na.action = na.exclude)

# Resumen del modelo
summary(reg_lin_simp)
```

```
##
## Call:
## lm(formula = data_wine_clean$sulphates ~ data_wine_clean$quality,
##     data = data_wine_clean, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32802 -0.07859 -0.01859  0.06141  0.39085
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)             0.301990   0.020151   14.99   <2e-16 ***
## data_wine_clean$quality 0.059433   0.003538   16.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1111 on 1530 degrees of freedom
## Multiple R-squared:  0.1557, Adjusted R-squared:  0.1551
## F-statistic: 282.1 on 1 and 1530 DF,  p-value: < 2.2e-16
```

La relación de dependencia comentada, puesto que estamos ante un modelo de regresión lineal simple, estará compuesto por una sola variable independiente (Xi), por lo tanto, la aproximación de relación estará definida por recta, cuya fórmula seguirá la estructura:

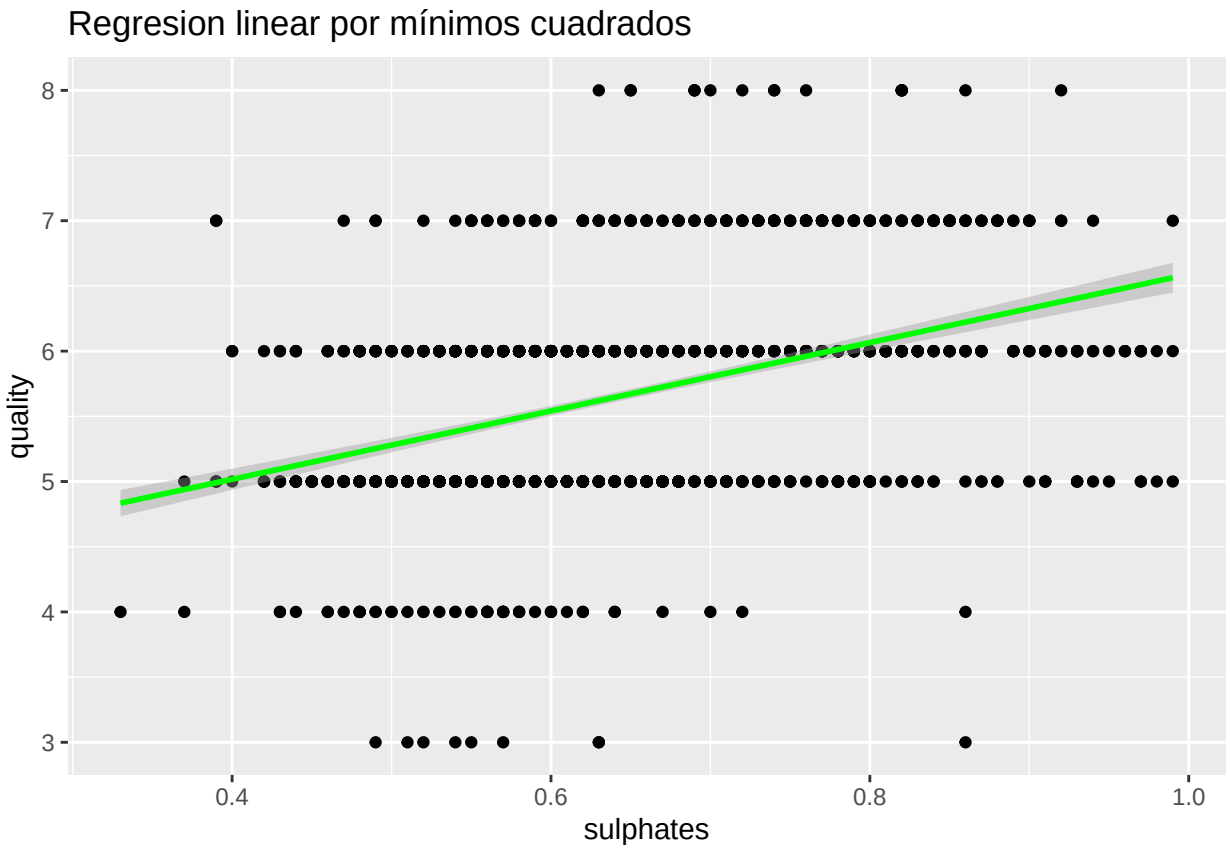
- $Y_i = \beta_0 + \beta_1 * X_i$

En concreto, para el casi que estamos estudiando y gracias a los datos obtenidos a partir del modelo:

data_wine_clean$quality i = 0.35776 + 0.05335 * data_wine_clean sulphates i

Como podemos observar la Multiple R-squared es el coeficiente que determina la calidad del modelo y toma valores entre 0 y 1, dando para este modelo un valor de 0.164 podemos entender que las variables no están fuertemente correlacionadas relacionadas, La representación visual del modelo obtenido sería la siguiente.

```
# Representacón regresión lineal
ggplot(data_wine_clean, aes(data_wine_clean$sulphates,data_wine_clean$quality )) +
  geom_point() +
  geom_smooth (method ='lm', colour = "Green") +
  ggtitle("Regresion linear por mínimos cuadrados") +
  xlab('sulphates') + ylab('quality')
```




Como podemos observar en la grafica las variables no están correlacionadas esto nos indica que a pesar de tener una alta correlación en la matriz por encima del 0,5 en realidad estas variables no están para nada relacionadas ,una vez realizado el estudio lo que nos indica que los sulphates elementos que determinan la cantidad de dióxido de azufre que contendrá el vino no tiene una relación directa con la calidad del vino.

No se realizará un análisis de regresión múltiple debido a que las variables estudiadas no están correlacionadas, lo que nos permite saber que no obtendremos ningún resultado significativo al realizar el análisis múltiple.

**Análisis de regresión logística.** La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable dicotómica dependiente, en función de una serie de variables independientes o predictoras. Dado que este modelo estima las probabilidades de ocurrencia, en lugar de utilizar un modelo aditivo que podría predecir valores fuera del rango (0,1) utiliza una escala transformada basada en una función logística.Así, el modelo lineal para probabilidades transformadas se define como:

Durante este apartado estudiaremos la relación entre la variable y la variable quality en comparación con las 3 variables antes estudiadas sulphates,alcohol y volatile.acidity .

Pero antes vamos a recodificar la variable Quality para que esta sea categórica de la siguiente forma.

- Si el valor de quality es $< 5$ es “Baja calidad”.
- Si el valor de quality es $=> 5$ es “Alta calidad”

```
# Clasificación variable calidad en 2 etiquetas
quality_cat <- cut(data_wine_clean$quality, breaks = c(-1,5,10),
                   labels = c("Baja calidad", "Alta calidad"))
```

- **Análisis de regresión logística para la variable sulphates**

```r
sulphates_quality <- table(data_wine_clean$sulphate, data_wine_clean$quality)
modelo_sulphates_quality = glm(formula = quality_cat ~ data_wine_clean$sulphate ,
                               data = data_wine_clean,
                               family = binomial(link = logit))

# Resumen del modelo
summary(modelo_sulphates_quality)
```

```
##
## Call:
## glm(formula = quality_cat ~ data_wine_clean$sulphate, family = binomial(link = logit),
##     data = data_wine_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3132  -1.0684   0.5569   1.0595   1.8404
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -4.1518     0.3319  -12.51   <2e-16 ***
## data_wine_clean$sulphate   6.8242     0.5256   12.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2115.0  on 1531  degrees of freedom
## Residual deviance: 1908.5  on 1530  degrees of freedom
## AIC: 1912.5
##
## Number of Fisher Scoring iterations: 4
```

Como podemos observar el modelo nos indica que al estar el p Valor por debajo de 0.05 (2e-16) esto nos indica que a través del método de regresión logarítmica esta variable tiene una alta significancia, seguramente sea debido al tratamiento de la variable quality que ha pasado a ser categórica , además el medidor AIC 2123.4 ( Akaike Information Criterion ) es una medida que mide la entropía del sistema que nos ofrece una estimación relativa de la información perdida, es bastante alto lo que nos puede indicar de que a pesar de tener una alta significancia el modelo no es fiable.

Ahora obtendremos el factor de riesgo del modelo anterior.

```r
exp(coefficients(modelo_sulphates_quality))
```

```
##              (Intercept) data_wine_clean$sulphate
##               0.01573638            919.82253911
```

Como podemos ver a pesar de tener una alta significancia el Valor OR nos indica que es altamente riesgoso usar este modelo para predecir la variable debido a que el valor está por encima de 1.

- **Análisis de regresión logística para la variable alcohol**

```
alcohol_quality <- table(data_wine_clean$alcohol, data_wine_clean$quality)
modelo_alcohol_quality = glm(formula = quality_cat ~ data_wine_clean$alcohol,
                             data = data_wine_clean, family = binomial(link = logit))

summary(modelo_alcohol_quality)
```

```
##
## Call:
## glm(formula = quality_cat ~ data_wine_clean$alcohol, family = binomial(link = logit),
##     data = data_wine_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1268  -0.9278   0.3984   0.9653   2.0047
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -10.58398    0.69367  -15.26   <2e-16 ***
## data_wine_clean$alcohol   1.03792    0.06754   15.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2115.0  on 1531  degrees of freedom
## Residual deviance: 1793.3  on 1530  degrees of freedom
## AIC: 1797.3
##
## Number of Fisher Scoring iterations: 4
```

Como podemos observar este modelo nos indica que al estar el p Valor por debajo de 0.05 (2e-16) también es altamente significante pero su medida AIC es más baja que la del modelo anterior 1868.6 vs 2123.4, esto nos indica que en este modelo se pierde menos información pero aun asi su valor es muy alto, veremos a continuación el resultado de los valores OR.

```
exp(coefficients(modelo_alcohol_quality))
```

```
##             (Intercept) data_wine_clean$alcohol
##            2.531832e-05            2.823332e+00
```

Como podemos ver este resultado es mejor que el anterior ya que se acerca a 1, pero sigue siendo un factor de riesgo considerable.

- **Análisis de regresión logística para la variable volatile.acidity.**

```
acidity_quality <- table(data_wine_clean$alcohol, data_wine_clean$quality)
modelo_acidity_quality = glm(formula = quality_cat ~
                               data_wine_clean$volatile.acidity,
                             data = data_wine_clean, family = binomial(link = logit))

summary(modelo_acidity_quality)
```

```
##
## Call:
## glm(formula = quality_cat ~ data_wine_clean$volatile.acidity,
##     family = binomial(link = logit), data = data_wine_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8650  -1.1237   0.7297   1.0327   2.0114
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       2.2630     0.1892   11.96   <2e-16 ***
## data_wine_clean$volatile.acidity  -3.9845     0.3422  -11.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2115.0  on 1531  degrees of freedom
## Residual deviance: 1957.9  on 1530  degrees of freedom
## AIC: 1961.9
##
## Number of Fisher Scoring iterations: 4
```

Como podemos observar el p valor de este modelo también nos indica una alta significancia pero su valor AIC (2036.2), también lo es lo que indica bastante pérdida de información del modelo, para tener una idea más clara vamos a estudiar los valores OR.

```
exp(coefficients(modelo_acidity_quality))
```

```
##                       (Intercept) data_wine_clean$volatile.acidity
##                         9.6120990                        0.0186012
```

Como podemos observar para esta variable está los valores OR dan una alta significancia, esto significa que el modelo numero 3 es completamente fiable para poder predecir si un vino será bueno o malo.

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

Los análisis realizados durante el apartado anterior nos han proporcionado múltiples conclusiones y, a continuación, se presenta un recopilatorio de los resultados más significativos.

Durante el estudio de correlación, hemos observado la relación de todas las variables entre sí. En concreto, destacando que la influencia de las variables *alcohol*, *sulphates* y *volatile.acidity* son las mas influyentes sobre la variable calidad. No obstante, gracias a la visualización, presentado durante su apartado correspondiente, también se observan otras variables están bastante relacionadas entre sí. Por ejemplo, las variables relativas al dióxido de azufre libre y total están muy correladas, con un valor aproximado al 0.8.

En cuanto a los resultados del análisis del modelo no supervisado, hemos obtenido que la precisión de los modelos estudiados es:

| | Agrupación en 2 etiquetas | Agrupación en 3 etiquetas | Agrupación en 6 etiquetas |
|---|---|---|---|
| **Precision del modelo** | 69.06 | 32.11 | 15.39 |

Rápidamente, podemos observar que la mejor precisión es cuando la cantidad de clusters sobre los que hemos realizado los estudios es de k=2. En este caso, la poca correlación entre las variables hace que esto se acentúe como uno pérdida de precisión a medida que tenemos que realizar una mayor cantidad de agrupaciones.

Finalmente, para el estudio de regresiones:

| | Rectas_lin | Pendientes_lin | R_squared | Familia | p_valor | Or_Values |
|---|---|---|---|---|---|---|
| **v.acidity** | qualityi = 0.2249 + -0.081585 * volatile.acidityi | Negativa | 0.1513 | binomial | 2.2e-16 | 0.01718273 |
| **Alcohol** | qualityi = 6.89544 + 0.62567 * alcoholi | Positiva | 0.2249 | binomial | 2e-16 | 2.868461 |
| **sulphates** | qualityi = 0.35776 + 0.05335 * sulphatesi | Positiva | 0.06443 | binomial | 2e-16 | 25.4539950 |

Los resultados presentados son de los distintos análisis de regresión y para cada una de las variables estudiadas. Podemos observar que se ha obtenido un mejor resultado en el análisis logístico, debido a que las variables lineales están débilmente correlacionadas. Además, destacar que el mejor resultados comparando todas las regresiones logísticas es la correspondiente a la variable *volatile.acidity*.

Estas conclusiones realizadas son un recopilatorio de los resultados estudiados. No obstante, durante el proceso de análisis, en sus apartados correspondientes, hay más concusiones detalladas, así como la presentación de algunas evidencias visuales en caso de requerirlo.

# 6. Resolución del problema

En términos genéricos, observamos que este conjunto de datos es lo suficientemente completo como para ofrecer un abanico de opciones para el análisis muy amplio. En concreto, el tener una variable dependiente, permite la ejecución de modelos tanto supervisados como no supervisado, así como otros análisis como contrastes de hipótesis, regresiones o correlaciones, entre otros.

Durante el proceso de limpieza hemos podido comprender la importancia de un buen entendimiento de los datos ofrecidos. Este punto ha sido altamente remarcado durante el estudio de valores extremos. El dataset tenía, a nivel de resultados de un cálculo, una gran cantidad de valores extremos. Sin embargo, cunado comprendías el significado de los datos veías que los valores eran válidos, pero muy poco comunes.

En cuanto a la aplicación de los distintos modelos, hemos detectado que, en algunas ocasiones, resulta complejo el realizar estudios exclusivamente con variables numéricas. Para determinados estudios, las variables categóricas simplifican la comprensión del proceso, así como es mas "representativo" el resultado obtenido.

Finalmente, retomando las cuestiones presentadas al inicio de la práctica y, ya terminado el análisis, vemos como podemos extraer respuestas y/o consideraciones sobre las mismas:

- ¿Qué características están ligadas directamente a que el vino sea de buena calidad?

Tal y como se ha ido reflexionando a lo largo de los distintos analisis, las variables *alcohol*, *sulphates* y *volatile.acidity* son aquellas que determinan, teóricamente, la calidad de un vino tinto. Esto es posible remarcarlo gracias al análisis de correlación realizado.

- ¿Cuáles de los métodos utilizados para analizar este dataset tiene una mejor respuesta?

Dando respuesta estricta a la pregunta realizada, debemos comentar que no se pueden comprar entre si los estudios realizados, ya que cada uno esta destinado a una u otra finalidad, por lo que se dará conclusión sobre si son buenos o malos dentro del objetivo propio del algoritmo.

Para el caso del clustering, cabe destacar que para este dataset no hemos obtenido resultados aceptables, hasta que hemos realizado un cambio en la variable calidad, limitando a que tenga, únicamente, dos etiquetas posibles. Esto, en la práctica, ha significado que hemos añadido conocimiento realizando una discriminación sobre la categoría, que en un principio tenía 6 valores posibles. Gracias a este paso, se han mejorado, significativamente, los resultados obtenidos.

Por otro lado, la regresión logística, teóricamente, nos indica que, si vamos a comparar modelos, deberíamos escoger el que tenga un menor índice AIC, debido a que este tiene en cuenta la bondad del modelo. No obstante, el índice AIC no puede decir nada acerca de la calidad del modelo en sentido absoluto ya que, si todos los modelos encajan mal con el problema, los datos proporcionados el índice AIC no darán ningún aviso sobre esto.

Al tener un índice AIC tan alto en los resultados obtenidos de los tres modelos, podríamos entender que no son adecuados para predecir la variable, reflejado en los modelos lineales que las variables eran prácticamente incorreladas. Por lo tanto, se concluye que el modelo comparado con la variable *volatile.acidity* da unos buenos resultados en los valores OR, y estos son muy próximo a 0. Por lo tanto, esto nos está indicando que es muy seguro para tomar este modelo y predecir los datos. Aun así, creemos que estos análisis estadísticos no son adecuados totalmente para predecir estas variables.

- ¿Los datos son adecuados para realizar un análisis profesional y no profesional?

Después del proceso de análisis realizado, podemos concluir que los datos disponibles para un análisis profesional de vinos son adecuados, debido a la gran cantidad de muestras y detalles. Gracias a ello, se pueden realizar cómodamente varios tipos de análisis con distintos métodos sin una mayor complicación.

En cambio, para el ámbito no profesional es posible que los datos deban tener menos detalle en algunos apartados, porque complica el entendimiento de estos. Consideramos que, una persona inexperta en la temática no requerirá un conocimiento tan profundo de los datos, y los resultados aceptables que obtenga serán conclusiones más genéricas.

- ¿El dulzor de los vinos es determinante para la calidad de este?

El azúcar residual, atributo que representa la clasificación de sabores (entre seco y dulce), da un resultado nulo en la matriz de correlación cuando vemos la relación entre este parámetro y la calidad del vino. Por lo tanto, no es un atributo significativo cuando hablamos de añadir un valor numérico de calidad.

Sn embargo, cabe notar que es una propiedad donde el nivel de subjetividad de cada una de las personas puede entrar en juego. A nivel números, para extraer un valor no tendrá sentido, pero es una de las propiedades más importantes para que un vino guste o no a las personas.

# 7. Código

A lo largo del desarrollo del documento, junto con las explicaciones y conclusiones necesarias, está el código desarrollado. Además, si se requiere validar el documento rmd individualmente, también está disponible en github [8] (https://github.com/ASarabiaSuarez/red_wine_practica_2).

# 8. Contribuciones y firmas

| Contribuciones | Frima |
|---|---|
| Investigacion previa teórica (lectura temaria y comprensión práctica a nivel conceptual) | ASS,EGB |
| Investigacion previa dataset (elección y comprension de los datos) | ASS,EGB |
| Redaccion de las respuestas | ASS,EGB |
| Desarrollo del código | ASS,EGB |

# 9. Recursos


[1] Información del dataset “winequality-red” en kaggle

[2] Información adicional del “Vinho verde”en su página web oficial

[3] Red and White Wine Quality

[4] Información de las propiedades del vino en La vinoteca

[5] Información vinos en CATAT

[6] Tabla de valores máximos admitidos de los vinos en España. Valores válidos para todos los países de la CEE.

[7][Acentología](http://www.acenologia.com/cienciaytecnologia/azufre_seguridad_vinos_ecologicos_cienc173_1219.htm#:~:text=El%20dióxido%20de%20azufre%20(SO,y%20bacterias%20del%20ácido%20acético) Revista de enología científica y profesional

[8] Github repositorio con los datos de la practica