Marta Sanjuán de Caso
320586666R

Práctica 2: Limpieza y validación de los datos

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

El conjunto de datos se ha obtenido del enlace facilitado en la propia tarea denominado Red WineQuality, obtenido de la dirección web https://archive.ics.uci.edu/ml/datasets/wine+quality . El conjunto de datos recoge una serie de características químicas y sensoriales de las variantes de vino tinto del denominado "Vinho Verde" Portugués, perteneciente a la zona norte del país.
Las clases del conjunto de datos se encuentran balanceadas pero no equilibradas según se recoge en la descripción del conjunto de datos realizada por el autor, además se indica que sería interesante utilizar algoritmos de detección de extremos para detectar tanto los vinos excelentes como pobres. Además, según indica el autor, no está seguro de si todas las variantes de entradas son variables relevantes, por lo que recomienda utilizar métodos de selección de características para calcular el nivel de interés de los mismos.

Las variables obtenidas en las pruebas físico químicas realizadas sobre el conjunto se dividen en dos grupos, variables de entrada:

1 - Acidez fija
2 - Acidez volátil
3 - Ácido cítrico
4 - Azúcar residual
5 - Cloruros
6 – Libre de dióxido de azufre
7 – Total de dióxido de azufre
8 - Densidad
9 - PH
10 - Sulfatos
11 – Alcohol

Y variables de salida (basada en datos sensoriales):

12 - Calidad (Puntuación entre 0 y 10)

A partir del conjunto de variables se pretende averiguar cuáles son las características que influyen más en la calidad de los vinos tintos de la variedad "Vinho Verde" portuguesa. Además, gracias al estudio del conjunto de datos actual, se podrá realizar predicciones de nuevas cosechas o de cosechas adquiridas a terceros para su distribución, por lo que atendiendo a las características del caldo, podremos predecir la calidad de los mismos.

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

Vamos a trabajar con lenguaje R mediante la herramienta RStudio.
Para comenzar, debemos de establecer el directorio de trabajo:

#Establecemos la carpeta en la que vamos a trabajar y tenemos el conjunto de datos
setwd("C:/Users/editrea/Desktop/Master 1819/TCVD/Practica 2")

Y a continuación vamos a trabajar con el conjunto de datos almacenado en el fichero winequality-red.csv. Leemos el fichero csvy lo convertimos en un objeto de tipo data.frame.

#Procedemos a la lectura del fichero csv, obteniendo el data.frame denominado vinos.
vinos <- read.csv("winequality-red.csv", header = TRUE)

#Procedemos a ver las dimensiones del fichero
dim(vinos)

```
> dim(vinos)
[1] 1599   12
```

El fichero contiene un total de 1599 observaciones con 12 variables cada una.

Veamos los primeros cinco registros del conjunto de datos.

#Veamos el contenido de los primeros cinco registros de la cabecera
head(vinos[,1:12])

```
> head(vinos[,1:12])
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density
1           7.4             0.70        0.00            1.9     0.076                  11                   34  0.9978
2           7.8             0.88        0.00            2.6     0.098                  25                   67  0.9968
3           7.8             0.76        0.04            2.3     0.092                  15                   54  0.9970
4          11.2             0.28        0.56            1.9     0.075                  17                   60  0.9980
5           7.4             0.70        0.00            1.9     0.076                  11                   34  0.9978
6           7.4             0.66        0.00            1.8     0.075                  13                   40  0.9978
    pH sulphates alcohol quality
1 3.51      0.56     9.4       5
2 3.20      0.68     9.8       5
3 3.26      0.65     9.8       5
4 3.16      0.58     9.8       6
5 3.51      0.56     9.4       5
6 3.51      0.56     9.4       5
```

A simple vista se visualiza que todos los campos son numéricos, pero veamos el tipo de dato que realmente almacenan:

 # Tipo de dato asignado a cada campo
library(knitr)
t <- sapply(vinos, function(x) class(x))
kable(data.frame(variables=names(t),clase=as.vector(t)))

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

Como vemos, la mayoría exceptuando el campo quality son de tipo numérico. El campo quality es de tipo entero, es decir, la lectura del fichero mediante la función read.csv le ha asignado correctamente los tipo de las variables con las que vamos a trabajar.

En el caso en que nos encontramos, la gran mayoría de atributos que se recogen en el conjunto de datos son susceptibles de estudio, es decir, hay que tenerlos en consideración a la hora de realizar el análisis y no debemos de prescindir de ninguno de ellos.

Para una mayor usabilidad, se va a renombrar las cabeceras de las columnas. Para ello:

```
#Procedemos a renombrar las columnas
names(vinos)
names(vinos)[names(vinos) == 'fixed.acidity'] <- 'acidez_fija'
names(vinos)[names(vinos) == 'volatile.acidity'] <- 'acidez_volátil'
names(vinos)[names(vinos) == 'citric.acid'] <- 'ácido_cítrico'
names(vinos)[names(vinos) == 'residual.sugar'] <- 'azúcar_residual'
names(vinos)[names(vinos) == 'chlorides'] <- 'cloruros'
names(vinos)[names(vinos) == 'free.sulfur.dioxide'] <- 'libre_dióxido_sulfurico'
names(vinos)[names(vinos) == 'total.sulfur.dioxide'] <- 'total_dióxido_sulfurico'
names(vinos)[names(vinos) == 'density'] <- 'densidad'
names(vinos)[names(vinos) == 'pH'] <- 'pH'
names(vinos)[names(vinos) == 'sulphates'] <- 'sulfatos'
names(vinos)[names(vinos) == 'alcohol'] <- 'alcohol'
names(vinos)[names(vinos) == 'quality'] <- 'calidad'
names(vinos)
```

```
 names(vinos)
[1] "acidez_fija"             "acidez_volátil"
[3] "ácido_cítrico"           "azúcar_residual"
[5] "cloruros"                "libre_dióxido_sulfurico"
[7] "total_dióxido_sulfurico" "densidad"
[9] "pH"                      "sulfatos"
[11] "alcohol"                "calidad"
```

3. Limpieza de los datos.

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

A continuación, vamos a analizar e identificar si existen campos con valores vacíos para cada uno de los atributos del juego de datos.

```
#Verificamos si existen valores vacíos
library(knitr)
t <- sapply(vinos, function(x) sum(is.na(x)))
kable(data.frame(variables=names(t),vacíos=as.vector(t)))
```

variables	vacíos
acidez_fija	0
acidez_volátil	0
ácido_cítrico	0
azúcar_residual	0
cloruros	0
libre_dióxido_sulfurico	0
total_dióxido_sulfurico	0
densidad	0
pH	0
sulfatos	0
alcohol	0
calidad	0

De la imagen anterior se recoge que no existen valores incompletos en el conjunto de datos con el que nos encontramos trabajando para ninguno de los atributos en estudio.

Si se hubiera dado el caso de que en algún atributo dispuesto a análisis existiera algún caso de valor vacío, existen una serie de técnica que nos pueden ayudar a completarlos, como por ejemplo ignorar el registro completo(poco recomendable, ya que perdemos información), rellenar el valor perdido manualmente (si son pocos valores afectados), rellenar con una constante globalmente, usar la media o mediana y la técnica más recomendable y más usada, la de usar el valor más probable calculado mediante regresión o inferencia.

A continuación vamos a identificar el número de ceros que existen en nuestro conjunto de datos:

```
#Verificamos cuántos ceros existen
library(knitr)
t <- sapply(vinos, function(x) sum(x==0))
kable(data.frame(variables=names(t),ceros=as.vector(t)))
```

variables	ceros
acidez_fija	0
acidez_volátil	0
ácido_cítrico	132
azúcar_residual	0
cloruros	0
libre_dióxido_sulfurico	0
total_dióxido_sulfurico	0
densidad	0
pH	0
sulfatos	0
alcohol	0
calidad	0

Como se puede ver en la anterior imagen, el único atributo que tiene valores con cero es el ácido_cítrico, con un total de 132 registros. Teniendo en cuenta que el número total de registros es de 1599, el porcentaje de registros con valor 0 en la variable ácido_cítrico es sumamente bajo, de alrededor de un 8%. Por lo tanto, teniendo en cuenta el dominio en el que nos encontramos, es totalmente viable que existan valores ceros, ya que se trata de un tipo de campo numérico y cuyo valor máximo y mínimo oscila en el intervalo de 0 a 1.

#Vemos un resumen del conjunto de datos. Dónde se indica los valores máximo y mínimo de la variable ácido_cítrico
summary(vinos)

```
  acidez_fija    acidez_volátil   ácido_citrico   azúcar_residual     cloruros       libre_dióxido_sulfurico
Min.    : 4.60   Min.    :0.1200   Min.    :0.000   Min.    : 0.900   Min.    :0.01200   Min.    : 1.00
1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00
Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :15.87
3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :72.00
total_dióxido_sulfurico    densidad           pH          sulfatos         alcohol         calidad
Min.    :  6.00          Min.    :0.9901   Min.    :2.740   Min.    :0.3300   Min.    : 8.40   Min.    :3.000
1st Qu.: 22.00          1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
Median : 38.00          Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   Median :6.000
Mean   : 46.47          Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
3rd Qu.: 62.00          3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
Max.   :289.00          Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

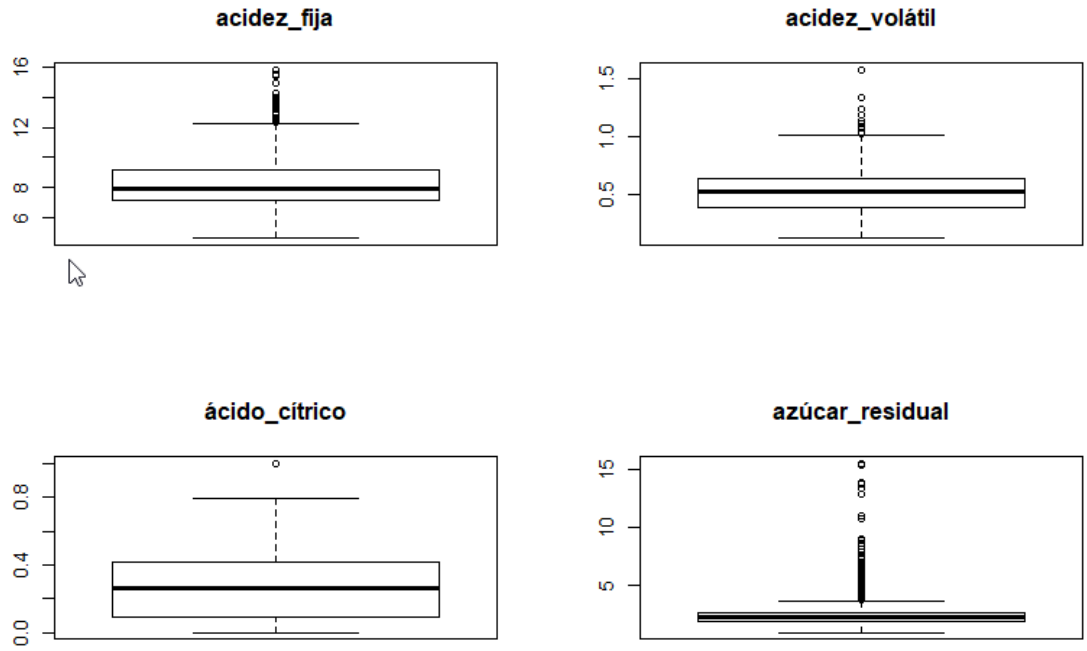
El tratamiento de los datos cuyo contenido es cero es individualizado a cada caso y entorno, es decir, hay que tratarlo de manera individualizada, por ejemplo, habría que conocer si la variable es medible o no, si el cero tiene significado para esa variable y si es susceptible de aplicarle cálculos matemáticos. En nuestro caso, como ya hemos mencionado, el cero tiene total sentido para la variable ácido_cítrico.

Por lo tanto, podemos resumir que los datos de este conjunto parece que han sido recogidos muy exhaustivamente o han sido transformados antes de ser publicados.

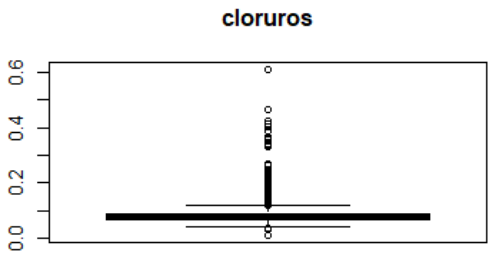
3.2. Identificación y tratamiento de valores extremos.

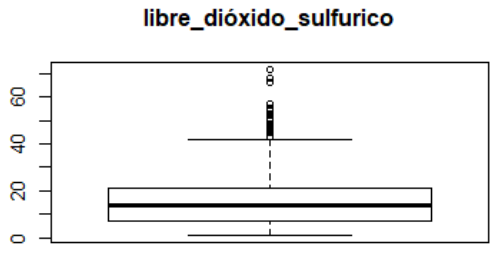
Para identificar los valores extremos, outliers o también denominados atípicos, vamos a hacer uso de los diagramas de cajas para cada una de las variables. Mediante la representación de las cajas, podremos visualizar los valores que distan mucho del rango intercuartílico, es decir, de la caja.

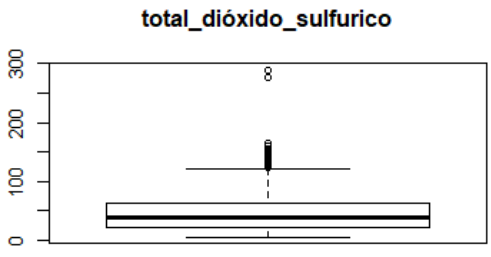
```
#Generamos las primeras cuatro cajas
 par(mfrow=c(2,2))
 for(i in 1:4) {
{ boxplot(vinos[,i], main = colnames(vinos)[i], width = 100)}
}
```

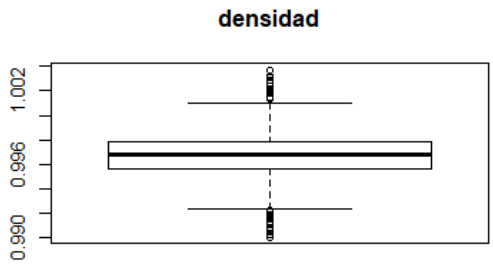



```
#Generamos las siguientes cuatro cajas
 par(mfrow=c(2,2))
 for(i in 5:8) {
{ boxplot(vinos[,i], main = colnames(vinos)[i], width = 100)}
}
```

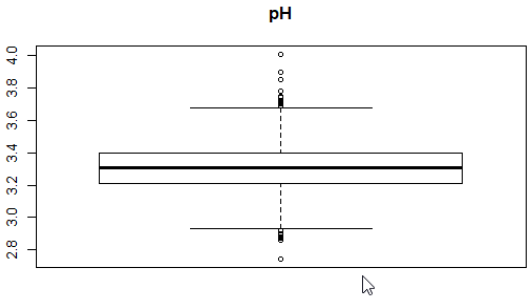


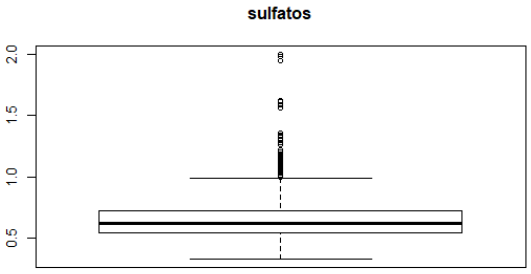


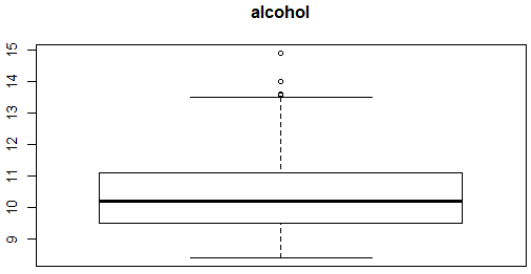


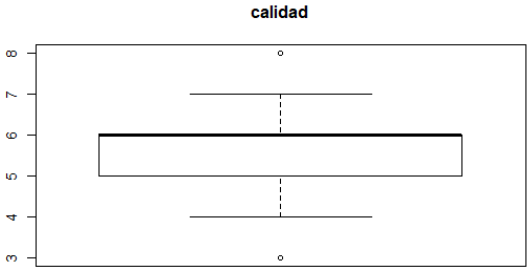


#Generamos las últimas cuatro cajas
par(mfrow=c(2,2))
for(i in 9:12) {
{ boxplot(vinos[,i], main = colnames(vinos)[i], width = 100)}
}









Mediante las cajas anteriores, podemos apreciar la presencia de valores atípicos en cada una de las variables.

Para obtener una mayor aproximación de esos valores atípicos, vamos a utilizar la función de boxplots.stats de R:

#Generamos los valores atipicos para cada una de las variables)
boxplot.stats(vinos$acidez_fija)$out
boxplot.stats(vinos$acidez_volátil)$out
boxplot.stats(vinos$ácido_cítrico)$out
boxplot.stats(vinos$azúcar_residual)$out
boxplot.stats(vinos$cloruros)$out
boxplot.stats(vinos$libre_dióxido_sulfurico)$out
boxplot.stats(vinos$total_dióxido_sulfurico)$out
boxplot.stats(vinos$densidad)$out
boxplot.stats(vinos$pH)$out
boxplot.stats(vinos$sulfatos)$out
boxplot.stats(vinos$alcohol)$out
boxplot.stats(vinos$calidad)$out

```
> names(vinos)
 [1] "acidez_fija"            "acidez_volátil"         "ácido_cítrico"          "azúcar_residual"
 [5] "cloruros"               "libre_dióxido_sulfurico" "total_dióxido_sulfurico" "densidad"
 [9] "pH"                     "sulfatos"               "alcohol"                "calidad"
> boxplot.stats(vinos$acidez_fija)$out
 [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 12.8 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5
[25] 13.0 12.5 13.3 12.4 12.5 12.9 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2 13.2 13.2 15.9 13.3 12.9 12.6
[49] 12.6
> boxplot.stats(vinos$acidez_volátil)$out
 [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040
> boxplot.stats(vinos$ácido_cítrico)$out
[1] 1
> boxplot.stats(vinos$azúcar_residual)$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  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80
 [21]  5.60  4.00  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00  4.50  4.80  5.80  5.80  3.80  4.40  6.20
 [41]  4.20  7.90  7.90  3.70  4.50  6.70  6.60  3.70  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  6.00  8.60  7.50  4.40  4.25  6.00  3.90  4.20
 [81]  4.00  4.00  4.00  6.60  6.00  6.00  3.80  9.00  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00  3.90
[101]  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  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  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80
[141]  5.20  5.20  3.75 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10  7.80
> boxplot.stats(vinos$cloruros)$out
  [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178 0.146 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611
 [21] 0.358 0.343 0.186 0.213 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122 0.122 0.174 0.121 0.127 0.413 0.152 0.152
 [41] 0.125 0.122 0.200 0.171 0.226 0.226 0.250 0.148 0.122 0.124 0.124 0.143 0.222 0.039 0.157 0.422 0.034 0.387 0.415 0.157
 [61] 0.157 0.243 0.241 0.190 0.132 0.126 0.038 0.165 0.145 0.147 0.012 0.012 0.039 0.194 0.132 0.161 0.120 0.120 0.123 0.123
 [81] 0.414 0.216 0.171 0.178 0.369 0.166 0.166 0.136 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414 0.166 0.168 0.415 0.153
[101] 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205 0.039 0.235 0.230 0.038
> boxplot.stats(vinos$libre_dióxido_sulfurico)$out
 [1] 52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51 51 52 55 55 48 48 66
> boxplot.stats(vinos$total_dióxido_sulfurico)$out
  [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127 126 145 144 135 165 124 124 134 124 129 151 133 142 149
 [32] 147 145 148 155 151 152 125 127 139 143 144 130 278 289 135 160 141 141 133 147 147 131 131 131
> boxplot.stats(vinos$densidad)$out
 [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220 1.00220 1.00140 1.00140 1.00140 1.00140 1.00320 1.00260
[16] 1.00140 1.00315 1.00315 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260 0.99210 0.99154 0.99064 0.99064 1.00289 0.99162
[31] 0.99007 0.99007 0.99020 0.99220 0.99150 0.99157 0.99080 0.99084 0.99191 1.00369 1.00369 1.00242 0.99182 1.00242 0.99182
> boxplot.stats(vinos$pH)$out
 [1] 3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87 2.89 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71 2.89
[25] 2.89 3.78 3.70 3.78 4.01 2.90 4.01 3.71 2.88 3.72 3.72
> boxplot.stats(vinos$sulfatos)$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 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18
[25] 1.13 1.04 1.11 1.13 1.07 1.06 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 1.07 1.34 1.16
[49] 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01
> boxplot.stats(vinos$alcohol)$out
 [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000 13.60000 14.00000 14.00000 13.56667 13.60000
> boxplot.stats(vinos$calidad)$out
 [1] 8 8 8 8 8 3 8 8 8 3 8 3 8 3 3 8 8 8 8 8 3 3 8 8 3 3 3 8
```

Revisando cada una de las variables del conjunto de datos, podemos ver que no se tratan en sí de valores extremos, sino de valores que se dan en una menor proporción que el resto y que no por ellos puedan darse, es decir, son susceptibles de que existan dentro de la muestra y por lo tanto debemos dejarlos tan y como se encuentran.

Llegados a este punto no se ha tenido que realizar ningún tipo de actuación sobre el conjunto de datos original, por lo que proseguimos el estudio con el mismo conjunto de datos.

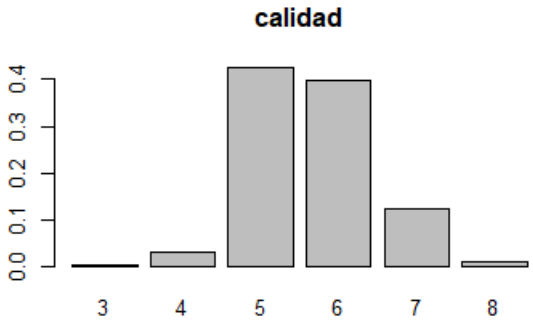
4. Análisis de los datos.

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

A continuación vamos a realizar un análisis previo del atributo objetivo, el atributo calidad.

Veamos su diagrama de cajas:

#Generamos el diagrama de barras de la variable calidad
barplot(prop.table(table(vinos$calidad)), main='calidad')



Podemos observar, que la variable calidad tiene sus valores distribuidos en un rango discreto comprendido entre 3 y 8, es decir, se trata de una variable discreta, donde la mayoría de las observaciones recaen en los valores 5, 6 y 7.

Veamos la distribución de valores dentro de la variable:

#Vemos la distribución que existe de los valores dentro de la variable calidad
table(vinos$calidad)

```
  3   4   5   6   7   8
 10  53 681 638 199  18
```

Podemos indicar que la gran mayoría de la muestra recae en los valores 5 o 6, en menor medida en 7 y reducidas ocasiones en 3, 4 y 7.

Vamos a guardar una copia del actual conjunto de datos en vinos2 .

#Copia del conjunto de datos con nombre de columnas modificadas
vinos_clean <- vinos

Finalmente exportamos el conjunto de datos:

 # Exportación de los datos limpios a .csv
write.csv(vinos_clean, "vinos_clean.csv")

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

Para comprobar la normalidad de las variables del conjunto de datos denominado *vinos_clean*, vamos a aplicar el test de Anderson-Darling. Este test comprueba si los datos de la muestra provienen de una distribución específica, es decir, de una distribución de probabilidad contínua específica, en este caso, vamos a comprobar si provienen de una distribución normal. Esta prueba se basa en la comparación de la distribución de probabilidades acumulada empírica con la distribución de probabilidades acumulada teórica (H0).

La hipótesis:

H0: La variable sigue una distribución Normal (μ - σ^2)
H1: La variable no sigue una distribución Normal (μ - σ^2)

Estadístico de prueba:

$$A^2 = -n - S$$

$$S = \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F(Yi) + \ln\left(1 - F\left(Y_{n+1-i}\right)\right)\right]$$

Donde:

n es el número de observaciones.
F(Y) es la distribución de probabilidades acumulada normal con media y varianzas específicadas a partir de la muestra.
Y_i Son los datos obtenidos en la muestra ordenados de menor a mayor.

Regla de decisión:
La hipótesis nula se rechaza con un nivel de significancia de α si A^2 es mayor que el valor crítico A_T^2.

Para un $\alpha = 0{,}05$, vamos a comprobar que para cada prueba por variable se obtiene un p-valor superior al nivel de significación prefijado. Si la variable lo cumple, se considera que sigue una distribución normal, veámoslo para cada una de ellas:

Teniendo en cuenta las variables de la muestra:
names(vinos_clean)

```
"acidez_fija"             "acidez_volátil"          "ácido_cítrico"
"azúcar_residual"         "cloruros"                "libre_dióxido_sulfurico"
"total_dióxido_sulfurico" "densidad"               "pH"
"sulfatos"                "alcohol"                 "calidad"
```

Vamos a comprobar si siguen una distribución normal aplicando el test de Anderson-Darling:

```
#Probamos si las variables siguen una distribución normal
library(nortest)
alpha = 0.05
col.names = colnames(vinos_clean)
for (i in 1:ncol(vinos_selected)) {
        if (i == 1) cat("Variables que no siguen una distribución normal:\n")
                if (is.integer(vinos_selected[,i]) | is.numeric(vinos_selected[,i])) {
                        p_val = ad.test(vinos_selected[,i])$p.value
                                if (p_val < alpha) {
                                        cat(col.names[i])
                                        # Formato de salida
                                        if (i <= ncol(vinos_selected) - 1) cat(", ")
                                                #if (i %% 3 == 0) cat("\n")
                                }
                }
}
```

```
Variables que no siguen una distribución normal:
acidez_fija, acidez_volátil, ácido_cítrico,
azúcar_residual, cloruros, libre_dióxido_sulfurico,
total_dióxido_sulfurico, densidad, pH,
sulfatos, alcohol, calidad
```

El resultado del test indica que ninguna de las variables del conjunto de datos sigue una distribución normal.

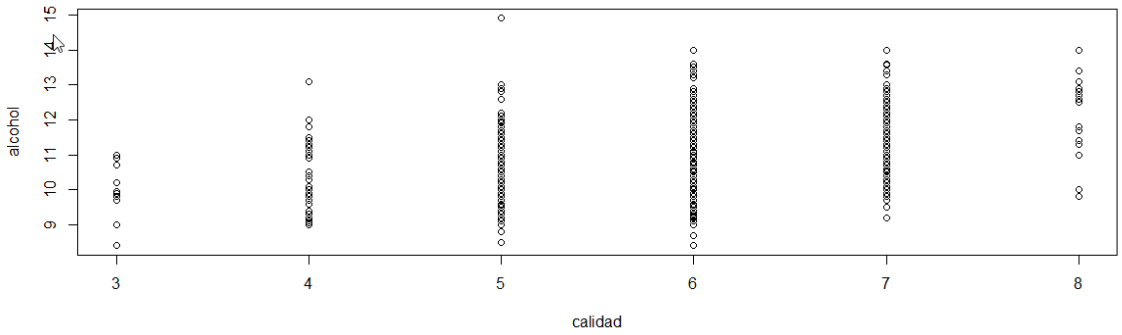
A continuación vamos a comprobar la homogeneidad de la varianza. Para comprobarlo vamos a aplicar el test de Fligner-Killeen.

Estadístico de prueba:

$$x = \frac{\sum_{i=1}^{k} n_i(\overline{a}_i - \overline{a})^2}{\sum_{j=1}^{N}(a_{N,j} - \overline{a})^2/(n-1)}$$

Vamos a estudiar la homogeneidad de varianzas entre el grupo que presenta un mayor porcentaje de alcohol con respecto a la calidad de los mismos.

#Test de Fligner-Killeen
plot(alcohol ~ calidad, data = vinos_clean)
fligner.test(alcohol ~ calidad, data = vinos_clean)




```
        Fligner-Killeen test of homogeneity of variances

data:  alcohol by calidad
Fligner-Killeen:med chi-squared = 135.61, df = 5, p-value < 2.2e-16
```

El p_valor obtenido es sumamente inferior a 0,05, por lo tanto, no podemos aceptar la hipótesis de que las varianzas de ambas muestras son homogéneas.

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.

4.3.1 Correlación

La primera prueba estadística que se va a realizar consiste en calcular la correlación existente entre el atributo calidad y el resto de atributos, para así hallar la variable que determine más la calidad del vino.

En primer lugar generamos la matriz de correlación.

#Calculamos la matriz de correlación
matrizcorrelación <- round(cor(vinos_clean),2)
head(matrizcorrelación)

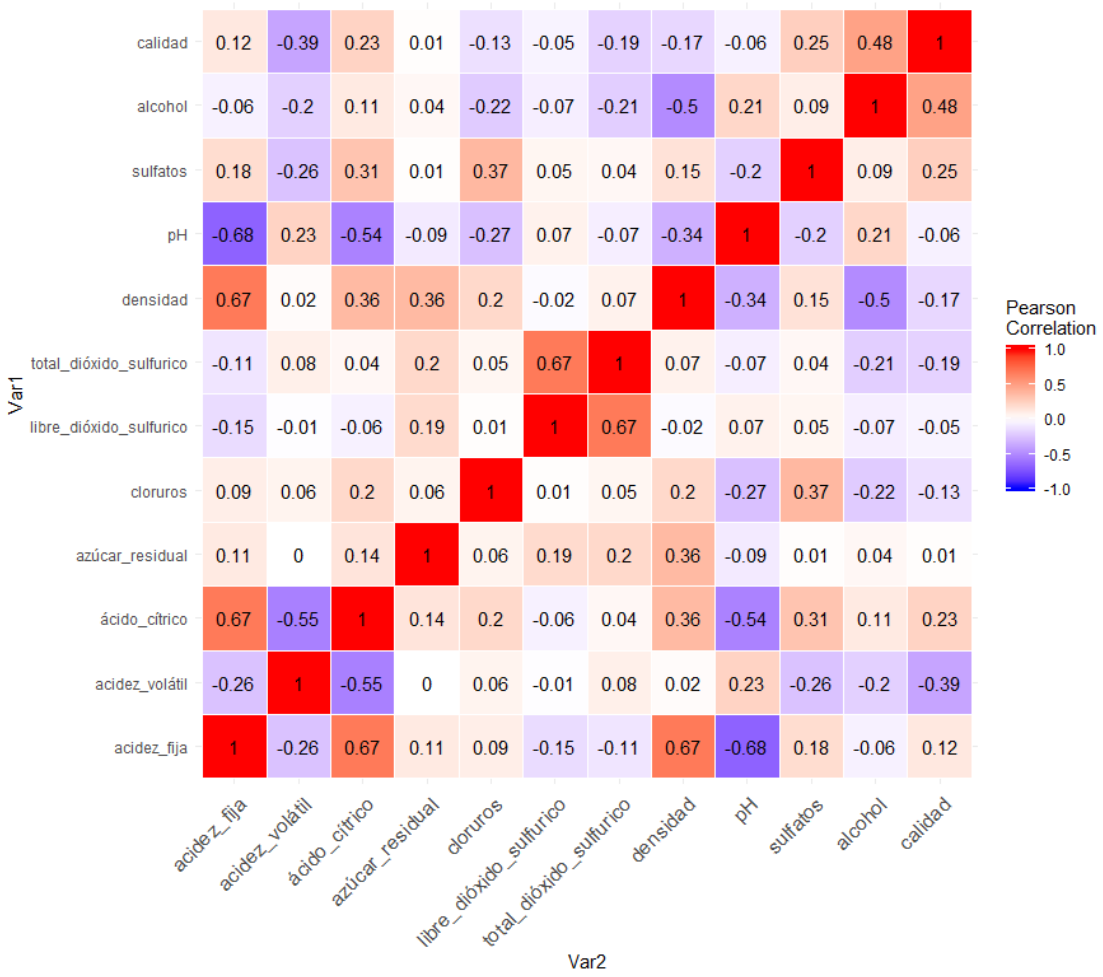
#Convertimos a formato largo de datos.
library(reshape2)
matcor <- melt(matrizcorrelación)
head(matcor)

```
                    Var1        Var2 value
           acidez_fija acidez_fija  1.00
        acidez_volátil acidez_fija -0.26
         ácido_cítrico acidez_fija  0.67
       azúcar_residual acidez_fija  0.11
             cloruros acidez_fija  0.09
libre_dióxido_sulfurico acidez_fija -0.15
```

Y a continuación la dibujamos:

```
#Visualizamos la matriz de correlación
library(ggplot2)
ggheatmap  <- ggplot(data = matcor, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +
 theme_minimal()+
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1))+
 coord_fixed()

#Añadimos los coeficientes de correlación por cuadrado
ggheatmap +
geom_text(aes(Var2, Var1, label = value), color = "black", size = 4)
```



Como vemos, las variables que tiene más correlación con respecto a la variable calidad es el Alcohol, seguido del nivel de sulfato, ácido cítrico y acidez_fija.

Vamos a guardar una copia denominada vinos_analisis que contendrá únicamente las variables con correlación más fuerte con respecto a la calidad.

```
#Copia del conjunto de datos con variables seleccionadas para el análisis posterior.ç
vinos_selected <- data.frame("alcohol"=vinos_clean$alcohol, "sulfatos"=vinos_clean$sulfatos,
"ácido_cítrico"=vinos_clean$ácido_cítrico, "acidez_fija"=vinos_clean$acidez_fija,
"calidad"=vinos_clean$calidad  )
head(vinos_selected)
```

Finalmente exportamos el conjunto de datos:

```
# Exportación de los datos con variables seleccionadas para su análisis
write.csv(vinos_selected, "vinos_selected.csv")
```

4.3.2 Contraste de hipótesis de dos muestras sobre la diferencia de medias

La segunda prueba estadística que se va a realizar consiste en averiguar si el nivel de alcohol en el vino en superior dependiendo de la calidad del vino.

Para poder realizar la prueba estadística, vamos a proceder a categorizar la variable discreta calidad de acuerdo a las siguientes condiciones:

Si la calidad es 3 o 4, se va a categorizar como malo
Si la calidad es 5 o 6, se va a categorizar como medio
Si la calidad es 7 u 8, se va a categorizar como bueno

```
#creamos un vector de tipo carácter
valoración <- as.vector(vinos_clean$calidad, mode="character")
valoración <- vinos_clean $calidad

#Lo completamos con los valores malo, bueno, medio y excelente.
for(i in 1:nrow(vinos_clean))
{
        if (valoración[i] <= 4)
        {valoración[i] <- 'malo'}
        else if (valoración[i] == 5 || valoración [i] == 6)
        {valoración[i] <- 'medio'}
        else if (valoración[i] == 7  || valoración [i] == 8)
        {valoración[i] <- 'bueno'}
        else if (valoración[i] >= 9)
        {valoración[i] <- 'excelente'}
}
```

#Vemos si cuadra el número de cada valor con lo obtenido anteriormente.
*table(*vinos_clean*$calidad)*

```
  3   4   5   6   7   8
 10  53 681 638 199  18
```

table(valoración)

```
bueno  malo
  855   744
```

Efectivamente cuadra el número total por tipo malo, medio y bueno con el sumatorio de su desglose, que hacen un total de 1599 registros.

Añadimos la columna al conjunto de datos de vinos.

#Añadimos la nueva columna al conjunto de datos
vinos_extended<- data.frame(vinos_clean, valoración)
head(vinos_extended)

Vamos a crear las dos muestras, una de ellas corresponderá al alcohol contenido en los vinos malos y una segunda muestra que contendrá el alcohol contenido en los bueno categorizado como bueno.
El test paramétrico que vamos a aplicar necesita que los datos se encuentren normalizados siempre y cuando la muestra sea inferior a 30, en nuestro caso, la muestra es muy superior a ese número, por lo que podemos proceder a aplicar el test de contraste de hipótesis de dos muestras sobre la diferencia de medias.

Generamos en primer lugar las dos muestras:

#Obtenemos ambas muestras
vinos_extended_malo <-vinos_extended[vinos_extended$valoración == "malo",]$alcohol
vinos_extended_bueno <-vinos_extended[vinos_extended$valoración == "bueno",]$alcohol

Vemos las medias de cada muestra:

#Obtenemos las estadísticas de cada muestras de alcohol
summary(vinos_extended_malo)
summary(vinos_extended_bueno)

```
summary(vinos_extended_malo)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
8.400   9.400   9.700   9.926  10.300  14.900
```

```
summary(vinos_extended_bueno)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 8.40   10.00   10.80   10.86   11.70   14.00
```

Planteamos el siguiente contraste de hipótesis sobre dos muestras con respecto a la diferencia de medias, atendiendo a:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 < 0$$

Donde,

μ_1 es la media de la población de la muestra de vinos malos
μ_2 es la media de la población de la muestra de vinos buenos

Tomando un $\alpha = 0{,}05$

Ejecutamos el test:

#Ejecutamos el test sobre diferencia de medias
t.test(vinos_extended_malo, vinos_extended_bueno, alternative = "less")

Obtenemos como resultado del test:

```
        Welch Two Sample t-test

data:  vinos_extended_malo and vinos_extended_bueno
t = -19.782, df = 1516.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.8512962
sample estimates:
mean of x mean of y
 9.926478 10.855029
```

Obtenemos un p-value mucho menor que el valor de significación fijado, por lo tanto, debemos rechazar la hipótesis nula y concluimos que el nivel de alcohol de un vino bueno es superior que el de un vino malo.

5. Resolución del problema. A partir de los resultados obtenidos, .cuales son las conclusiones? Los resultados permiten responder al problema?

Los resultados obtenido a través de las diferentes pruebas estadísticas realizadas sobre el conjunto de datos de vinos tintos de la variedad “Vinho Verde” portuguesa nos permite responder de forma clara al problema planteado en la primera sección, donde se planteó averiguar cuáles eran las características que más influían en la calidad del vino de esta variedad, y como hemos visto en el punto anterior, las características que más influyen en su calidad es el grado de alcohol y en segundo lugar el nivel de sulfato seguido del nivel de ácido cítrico que presente y de la acidez fija que tenga el vino.
El alcohol es sin lugar a dudas la característica que más influye en la calidad del vino, dónde se ha testimoniado a través del análisis de correlación y el contraste de hipótesis.