M2.851 – Tipología y ciclo de vida de los datos. Aula 2. PRA 2: Limpieza y análisis de datos

Aleix Salvador Barrera y Víctor Miranda Hernández

19/5/2021

Índice de contenido

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com). Algunos ejemplos de dataset con los que podéis trabajar son:

- **Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)**
- **Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)**

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición. Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

Ejercicio 1 [0.5 puntos]. Descripción del dataset.

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

El dataset que hemos escogido para la realización de esta práctica se llama "Red Wine Quality" y se puede encontrar en la página web de Kaggle en la siguiente URL: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009. El conjunto de datos contiene un total de 1.599 registros (los cuales representan 1.599 vinos diferentes) y 12 campos (los cuales representan 12 características diferentes de cada vino). El nombre de cada uno de los campos (y su descripción) es el siguiente:

- **Fixed acidity:** cantidad de los ácidos involucrados con el vino (fijos o no volátiles) que no se evaporan fácilmente.
- **Volatile acidity:** cantidad de ácido acético en el vino, que en niveles demasiado altos puede provocar un sabor desagradable a vinagre.
- **Citric acid:** cantidad de ácido cítrico en el vino, el cual encontrado en pequeñas cantidades, puede agregar frescura y sabor a los vinos.
- **Residual sugar:** cantidad de azúcar que queda después de detenerse la fermentación.
- **Chlorides:** cantidad de sal en el vino.
- **Free sulfur dioxide:** Cantidad de formas libres de SO2.
- **Total sulfur dioxide:** Cantidad de formas libres y unidas de SO2.
- **Density:** Densidad del vino.
- **PH:** Describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico). La mayoría de los vinos están entre 3 y 4.
- **Sulphates:** Cantidad de sulfatos que contiene el vino.
- **Alcohol:** Porcentaje de alcohol que contiene el vino.
- **Quality:** Variable de salida que mide la calidad del vino entre 0 y 10.

Se debe destacar que todos los campos son de tipo numérico y contienen características (11 primeros campos) mediante las cuales se puede poner una nota al vino, la cual se registra en el último campo "Quality". Por lo tanto, estamos delante de un conjunto de datos interesante porque nos puede permitir la construcción de modelos predictivos (por ejemplo, de regresión) con el fin de predecir la calidad de un vino (campo

Quality) a partir de los valores de las variables explicativas anteriores (las 11 primeras variables expuestas antes). Finalmente, la pregunta o problema que pretende responder este dataset es: ¿Qué características son más importantes en los vinos para que reciban una mayor puntuación total de calidad? Por último, me gustaría comentar que los conjuntos de datos de este tipo son muy importantes/útiles para las empresas que produzcan vinos porque a partir de ellos, se pueden extraer conclusiones que permitan mejorar la calidad de sus productos y así, maximizar sus beneficios.

Ejercicio 2 [0.5 puntos]. Integración y selección de los datos de interés a analizar.

Integración y selección de los datos de interés a analizar.

La integración de los orígenes de los datos consistirá en realizar la importación del conjunto de datos ya que únicamente trabajaremos con los datos provenientes de un origen. Un aspecto importante a tener en cuenta es que el formato del archivo origen es "csv", por lo que se utilizará la función *read.csv* de R para importarlo. A continuación se muestra este paso, y una pequeña muestra del data frame "dd", es decir, el conjunto de datos importado:

```
carpeta <- "C:/Users/aleix.salvador/Desktop/Master Data Science/4 - Tipología y ciclo de vida de los da

pec <- "PRA 2/"

file_source <- "winequality-red.csv"

ruta_s <- paste0(carpeta,pec,file_source)

dd <- read.csv(ruta_s,header = TRUE,sep = ",")

head(dd,5)
```

```
##   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
##   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
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
```

Después de realizar la importación de los datos y de visualizar un fragmento del data frame obtenido, se analizarán las dimensiones del conjunto de datos importado y el tipo de dato contenido en cada columna (campo) del mismo:

```
nrow(dd);ncol(dd)
```

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

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

Se puede observar que, tal como se había descrito en el primer apartado de la práctica, el dataset está formado por 1.599 filas y 12 columnas. A continuación se analizará el tipo de dato que contiene cada columna mediante las funciones *sapply* y *class* de R:

```
sapply(dd,class)
```

```
##        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"
```

A partir de la salida de la función *sapply* podemos observar que todas las variables son numéricas, y que la variable "quality" es de tipo entero. Recordar que esta variable es la que contiene la nota del vino entre 0 y 10 (por lo tanto, esta nota estará definida con números enteros).

Finalmente, para terminar este ejercicio, se realizará un resúmen exploratorio de cada una de las variables del dataset (mediante las funciones *sapply* y *class* de R) con el fin de conocer las distribuciones de las mismas y tener una primera impresión sobre el número de nulos y valores extremos existentes en el dataset.

```
sapply(dd,summary)
```

```
##         fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides
## Min.         4.600000        0.1200000   0.0000000       0.900000 0.01200000
## 1st Qu.      7.100000        0.3900000   0.0900000       1.900000 0.07000000
## Median       7.900000        0.5200000   0.2600000       2.200000 0.07900000
## Mean         8.319637        0.5278205   0.2709756       2.538806 0.08746654
## 3rd Qu.      9.200000        0.6400000   0.4200000       2.600000 0.09000000
## Max.        15.900000        1.5800000   1.0000000      15.500000 0.61100000
##         free.sulfur.dioxide total.sulfur.dioxide   density       pH sulphates
## Min.                1.00000              6.00000 0.9900700 2.740000 0.3300000
## 1st Qu.             7.00000             22.00000 0.9956000 3.210000 0.5500000
## Median             14.00000             38.00000 0.9967500 3.310000 0.6200000
## Mean               15.87492             46.46779 0.9967467 3.311113 0.6581488
## 3rd Qu.            21.00000             62.00000 0.9978350 3.400000 0.7300000
## Max.               72.00000            289.00000 1.0036900 4.010000 2.0000000
##          alcohol  quality
## Min.     8.40000 3.000000
## 1st Qu.  9.50000 5.000000
## Median  10.20000 6.000000
## Mean    10.42298 5.636023
## 3rd Qu. 11.10000 6.000000
## Max.    14.90000 8.000000
```

Después de visualizar la salida de los resúmenes de cada una de las variables del dataset, podemos observar que aparentemente no existe ningún valor nulo y que pueden existir varios valores extremos ya que en las variables free.sulfur.dioxide y total.sulfur.dioxide el máximo de las dos variables es 72 y 289 respectivamente, los cuales están bastante alejados de los valores medios de sus distribuciones.

Antes de terminar, nos gustaría remarcar que en este dataset no se obviará ningún campo ya que todas las variables existentes contienen características del vino y serán importantes (en mayor o menor medida)

para detallar la nota de la calidad del vino. Por lo tanto, se mantendrán en el data frame todos los campos actuales. También nos parece interesante la construcción de una nueva variable cualitativa a partir de la variable *quality*, ya que en ejercicios posteriores se podrán realizar estudios de clasificación utilizando esta variable como variable salida. Esta variable tomará los valores "Malo" (si la variable quality es menor a 5), "Normal" (si la variable quality está entre 5 y 6 ambos incluidos) y "Bueno" (si la variable quality es superior a 6). A continuación se construye dicha variable mediante la función *case_when* de la librería *dplyr*:

```
library(dplyr)

tipo_vino <- case_when(dd$quality < 5 ~ "Malo",
                       dd$quality >= 5 & dd$quality <= 6 ~ "Normal",
                       TRUE ~ "Bueno")

dd$tipo_vino <- tipo_vino
dd$tipo_vino <- as.factor(dd$tipo_vino)
```

A continuación se muestra la distribución de esta nueva variable construida:

```
table(dd$tipo_vino)
```

```
##
##  Bueno   Malo Normal
##    217     63   1319
```

Ejercicio 3 [2 puntos]. Limpieza de los datos.

Limpieza de los datos.

Ejercicio 3.1. Valores perdidos.

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

En este apartado se realizarán dos comprobaciones: en primer lugar se comprobará si existe algún valor perdido (*NA*) en el dataset, y en segundo lugar se analizará la variable *citric.acid*, ya que en el resúmen mostrado en el ejercicio anterior se ha podido visualizar que su valor mínimo es 0, por lo tanto se analizará si este 0 es un valor correcto o si, por su defecto, es un valor perdido que se ha rellenado con el valor 0.

En primer lugar, se analiza si existen valores perdidos en el dataset mediante la función *is.na* de R:

```
sum(is.na(dd))
```

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

A partir de la función *is.na* de R podemos visualizar que no existe ningún valor nulo en el conjunto de datos, por lo tanto, no tiene sentido aplicar esta función para cada variable, ya que no encontraremos ningún valor perdido en ningun campo. A continuación se procederá a analizar los valores 0 de la variable *citric.acid*:

```
nrow(dd[dd$citric.acid ==0,])
```

```
## [1] 132
```

Vemos que hay 132 registros con el valor de la variable *citric.acid* igual a 0. Ahora se mostrará una pequeña muestra de estos registros:

```
head(dd[dd$citric.acid==0,])
```

```
##    fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1            7.4            0.700           0            1.9     0.076
## 2            7.8            0.880           0            2.6     0.098
## 5            7.4            0.700           0            1.9     0.076
## 6            7.4            0.660           0            1.8     0.075
## 8            7.3            0.650           0            1.2     0.065
## 13           5.6            0.615           0            1.6     0.089
##    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
## 5                   11                   34  0.9978 3.51      0.56     9.4
## 6                   13                   40  0.9978 3.51      0.56     9.4
## 8                   15                   21  0.9946 3.39      0.47    10.0
## 13                  16                   59  0.9943 3.58      0.52     9.9
##    quality tipo_vino
## 1        5    Normal
## 2        5    Normal
## 5        5    Normal
## 6        5    Normal
## 8        7     Bueno
## 13       5    Normal
```

Los registros con la variable *citric.acid* igual a 0 parecen totalmente normales. En las descripciones de las variables se detalla que este elemento, en pequeñas cantidades, puede mejorar algunos aspectos del vino como el sabor o la frescura. Además, en los resúmenes de cada variable del dataset se puede visualizar que el máximo de esta variable es 1, por lo tanto, parece que el número 0 no es un valor perdido, sino que es totalmente normal.

A partir del análisis realizado en este apartado, se puede concluir que no existe ningún valor perdido en el conjunto de datos, y que los valores ceros que aparecen en la variable *citric.acid* son totalmente normales ya que esta variable toma valores entre 0 y 1.

Ejercicio 3.2. Valores extremos.

Identificación y tratamiento de valores extremos.

En este segundo apartado del tercer ejercicio de la práctica se analizarán e identificarán los valores extremos de las variables del dataset.Para ello se utilizará la función *boxplot.stats* de R, la cual mostrará para cada variable, sus valores extremos:

```
boxplot.stats(dd$fixed.acidity)$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
## [16] 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4 12.5 12.9
## [31] 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
## [46] 13.3 12.9 12.6 12.6
```

```
boxplot.stats(dd$volatile.acidity)$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
## [13] 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

```
boxplot.stats(dd$citric.acid)$out
```

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

```
boxplot.stats(dd$residual.sugar)$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
```

```
boxplot.stats(dd$chlorides)$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
##  [13] 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343 0.186 0.213
##  [25] 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122 0.122 0.174 0.121
##  [37] 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171 0.226 0.226 0.250 0.148
##  [49] 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
##  [73] 0.039 0.194 0.132 0.161 0.120 0.120 0.123 0.123 0.414 0.216 0.171 0.178
##  [85] 0.369 0.166 0.166 0.136 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414
##  [97] 0.166 0.168 0.415 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205
## [109] 0.039 0.235 0.230 0.038
```

```
boxplot.stats(dd$free.sulfur.dioxide)$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
## [26] 55 55 48 48 66
```

```
boxplot.stats(dd$total.sulfur.dioxide)$out
```

```
##  [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127 126 145
## [20] 144 135 165 124 124 134 124 129 151 133 142 149 147 145 148 155 151 152 125
## [39] 127 139 143 144 130 278 289 135 160 141 141 133 147 147 131 131 131
```

```
boxplot.stats(dd$density)$out
```

```
##  [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220 1.00220
## [10] 1.00140 1.00140 1.00140 1.00140 1.00320 1.00260 1.00140 1.00315 1.00315
## [19] 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260 0.99210 0.99154 0.99064
## [28] 0.99064 1.00289 0.99162 0.99007 0.99007 0.99020 0.99220 0.99150 0.99157
## [37] 0.99080 0.99084 0.99191 1.00369 1.00369 1.00242 0.99182 1.00242 0.99182
```

```
boxplot.stats(dd$pH)$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
## [16] 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71 2.89 2.89 3.78 3.70 3.78 4.01 2.90
## [31] 4.01 3.71 2.88 3.72 3.72
```

```
boxplot.stats(dd$sulphates)$out
```

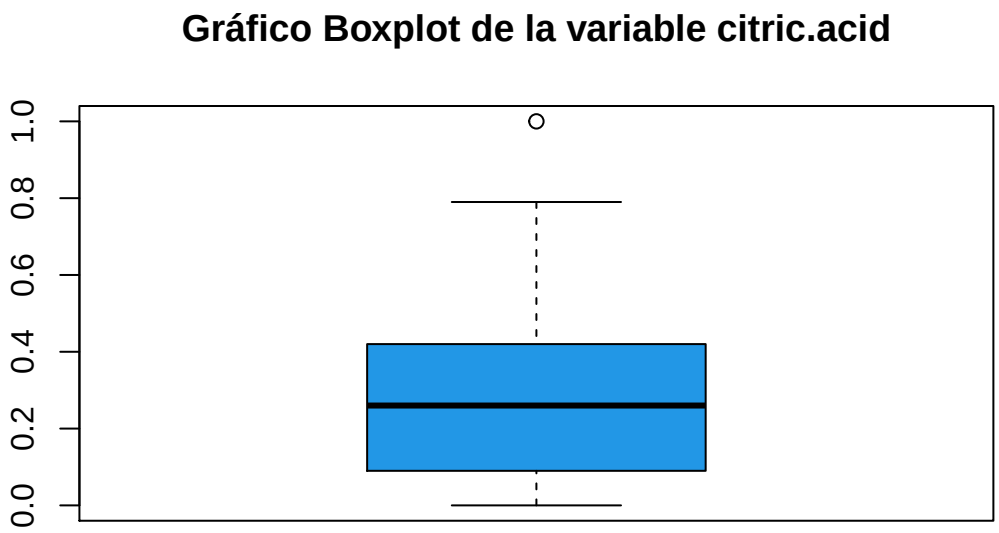
```
##  [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08 1.59
## [16] 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13 1.07 1.06
## [31] 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17 1.62 1.06 1.18
## [46] 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01
```

```
boxplot.stats(dd$alcohol)$out
```

```
##  [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000
##  [9] 13.60000 14.00000 14.00000 13.56667 13.60000
```

Después de visualizar los valores extremos de cada una de las variables hemos podido observar que en todas ellas existen valores alejados del centro de sus distribuciones, pero hay una variable en concreto que nos ha resultado peculiar, ya que únicamente aparece un valor extremo y según los resultados de los resúmenes de cada una de las variables realizados anteriormente, parece que está bastante alejado del resto. Esta variable comentada es *citric.acid*. A continuación, para analizar visualmente esta variable se construirá un gráfico de boxplot:

```
boxplot(dd$citric.acid,
        main="Gráfico Boxplot de la variable citric.acid",
        col=44)
```



Al visualizar el boxplot de la variable *citric.acid* podemos observar que el valor extremo 1 está muy alejado del resto de valores de dicha variable y por lo tanto se sitúa muy lejos en una cola de la distribución. Para mejorar los resultados posteriores del estudio se procederá a eliminar este registro del conjunto de datos ya que distorsionaría notablemente los resultados obtenidos. Nos gustaría comentar que del resto de variables no se eliminará ninguna observación porque no sus valores extremos no se tratan de casos aislados como el de la variable *citric.acid*, sino que existen muchos más valores extremos y no se encuentran tan alejados de la distribución. Por lo tanto, finalmente, se construirá el nuevo dataset sin esta observación y se creará el archivo "winequality-red_clean.csv" mediante la función *write.csv* de R:

```
dd_clean <- dd[dd$citric.acid != 1,]


file_target <- "winequality-red_clean.csv"

ruta_t <- paste0(carpeta,pec,file_target)

write.csv(dd_clean,ruta_t)
```

Ejercicio 4 [2.5 puntos]. Análisis de los datos.

Análisis de los datos.

Ejercicio 4.1. Planificación de los análisis a aplicar.

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

El conjunto de datos seleccionado para la realización de esta práctica contiene información relevante en todos sus campos, ya que no tiene ningún campo ID con el identificador único de cada registro ni ningún otro campo con información que no sea necesaria. En nuestro caso, todos los campos contienen alguna característica del vino que puede ser interesante para modelar la calidad. A parte de todos los campos de origen, se ha creado una variable cualitativa para definir cada vino como Malo, Normal o Bueno a partir del resultado obtenido en la varible numérica Quality (que se debe recordar que recoge la calidad del vino en una escala del 0 al 10). Por lo tanto, en esta parte de la práctica en la que se analizarán los datos del dataset seleccionado, se utilizarán todos los campos del conjunto de datos (y también el nuevo construido en apartados anteriores) y todos los registros del mismo.

Los análisis que se aplicarán en esta práctica son el cálculo de la correlación entre las variables del conjunto de datos y la variable "quality" para comprobar qué variables están más relacionadas con esta variable "dependiente", un contraste de hipótesis para comprobar si los valores de las variables más correlacionadas con la variable "quality" difieren para los diferentes tipos de vino, y en último lugar, un modelo de regresión para asegurar cuáles son las variables que más afectan a la calidad del vino y en qué medida.

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

Comprobación de la normalidad y homogeneidad de la varianza.

En este segundo apartado del cuarto ejercicio de la práctica se comprobará, mediante el test de Shapiro, la normalidad de las variables cuantitativas del conjunto de datos (es decir, de todas excepto la que se ha creado manualmente en ejercicios anteriores de la práctica a partir de la variable "Quality"). A parte, se comprobará mediante la aplicación del test XXXX, si los diferentes grupos de la nueva variable "tipo_vino" creada manualmente en ejercicios anteriores, tienen una varianza igual en los valores de la variable quality.

A continuación, en primer lugar, se contrasta la normalidad de las variables del conjunto de datos y se mostrará por pantalla qué variables siguen una distrubición Normal y cuales no:

```
alpha <- 0.05
vars  <- colnames(dd_clean)[-ncol(dd_clean)] #Variables cuantitativas

v_Norm   <- c()
v_noNorm <- c()

for(i in 1:length(vars)){
  if(shapiro.test(dd_clean[,i])$p.value < alpha){
    v_noNorm <- c(v_noNorm,vars[i])
  } else {
    v_Norm <- c(v_Norm,vars[i])
  }
}
```

Las variables que siguen una distribución Normal son:

```
v_Norm
```

```
## NULL
```

Las variables que no siguen una distribución Normal son:

```
v_noNorm
```

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
```

Después de analizar la normalidad de los datos contenidos en las variables cuantitativas del conjunto de datos se he podido observar que ninguna de estas variables sigue una distribución Normal. Por lo tanto, para contrastar la homogeneidad de la varianza en poblaciones no Normales, los tests más recomendados son el de Leven (utilizando la mediana) o el test no paramétrico de Fligner-Killeen (también basado en la varianza). En este apartado se contrastará, mediante estos dos tests comentados, la homogeneidad de la varianza de los valores del campo "quality" en los distintos tipos de vino (Malo, Normal y Bueno).

A continuación, en primer lugar, se contrasta la homogeneidad de la varianza mediante el test de Leven:

```
library(car)
leveneTest(y = dd_clean$quality, group = dd_clean$tipo_vino, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value    Pr(>F)
## group    2  76.841 < 2.2e-16 ***
##       1595
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A partir del test de Leven se puede observar que no existe homogeneidad de varianza de los valores de la variable "quality" en los diferentes tipos de vino. A continuación se realizará el mismo contraste pero con el test de Fligner-Killeen:

```
a <- dd_clean[dd_clean$tipo_vino == "Malo","quality"]
b <- dd_clean[dd_clean$tipo_vino == "Normal","quality"]
c <- dd_clean[dd_clean$tipo_vino == "Bueno","quality"]
fligner.test(x = list(a,b,c))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(a, b, c)
## Fligner-Killeen:med chi-squared = 140.35, df = 2, p-value < 2.2e-16
```

A partir del test de Fligner-Killeen se obtiene el mismo resultado que en el test de Leven: la varianza de la variable "quality" no es igual en los distintos tipos de vino.

Por lo tanto, para finalizar con este apartado, se puede afirmar que los datos contenidos en las variables de este dataset no siguen una distribución Normal y que la varianza de la variable "quality" en los diferentes grupos de vino (Malo, Normal y Bueno) no es homogenia.

Ejercicio 4.3. Aplicación de pruebas estadísticas.

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

En este tercer apartado del ejercicio 4 de la práctica se aplicarán las pruebas estadísticas comentadas en el primer apartado del ejercicio 4.

4.3.1. Cálculo de la matriz de correlaciones entre las variables del conjunto de datos

En primer lugar se calcula la correlación entre las variables cuantitativas (todas las originales, obviamos la variable "tipo_vino" creada manualmente porque es una variable cualitativa) del conjunto de datos y la variable "quality" para medir la asociación existente entre ellas y conocer qué variable tiene más relación con esta variable salida y por lo tanto, conocer qué aspectos afectan más al hecho de obtener un bueno vino o uno malo. Se debe destacar que el coeficiente de correlación de Pearson no es válido en nuestro caso porque requiere que la distribución de las variables sea Normal (recordar que ninguna variable sigue una distribución Normal), por lo tanto se utilizará el coeficiente de correlación de Spearman, ya que es una alternativa no paramétrica que no necesita ninguna suposición sobre la distribución de los datos.

```
library(knitr)

m_cor <- cor(dd_clean[,1:12],method = "spearman")

cor_test_value <- function(x){
  return(cor.test(dd_clean[,"quality"],x,method = "spearman")$p.value)
}

p_value <- sapply(dd_clean[1:11],cor_test_value)

kable(data.frame(m_cor[1:11,12],p_value),
      col.names=c("cor(quality)","p.value"),
      digits = c(100,100))
```

	cor(quality)	p.value
fixed.acidity	0.11513335	3.946029e-06
volatile.acidity	-0.38103908	2.235728e-56
citric.acid	0.21582280	2.690849e-18
residual.sugar	0.03362676	1.790915e-01
chlorides	-0.18836678	3.146927e-14
free.sulfur.dioxide	-0.05540195	2.678339e-02
total.sulfur.dioxide	-0.19583411	2.812781e-15
density	-0.17575938	1.487720e-12
pH	-0.04567159	6.796406e-02
sulphates	0.37972985	5.681786e-56
alcohol	0.47797904	5.373947e-92

Después de calcular la correlación de Spearman entre las variables cuantitativas del conjunto de datos y la variable "dependiente" **quality**, se ha podido observar que las variables que más relacionadas están con la variable **quality** son "alcohol", "sulphates" y "volatile.acidity" con coeficientes de 0.47, 0.37 y -0.38 respectivamente. Estos coeficientes se interpretan de tal forma que a mayores valores de las variables "sulphates" y "alcohol", mayor será la calidad (nota) del vino, y a mayores valores de la variable "volatile.acidity",

menor será la calidad del vino. Finalmente, destacar que la relación entre estas variables no es muy fuerte ya que en ningún caso supera el 50% de relación (para definir una relación como fuerte, debería superar el 80 o 90% de relación). A parte, queríamos comentar también que se ha calculado el p.valor asociado a la correlación para comprobar si la correlación es significativa o no. Vamos que en los tres casos que hemos comentado la correlación es muy significativa ya que el p.valor asociado es muy pequeño (muy inferior al nivel de significación del 5%).

4.3.2. Contrastes de hipótesis en las variables más correlacionadas con la calidad del vino

El segundo método que se aplicará es un contraste de hipótesis no paramétrico para contrastar si realmente los valores de las tres variables más correlacionadas con la variable "quality" son diferentes en los distintos niveles del factor "tipo_vino" calculado a partir de los valores de la variable "quality" (se debe recordar que se utilizan tests no paramétricos porque las variables de este conjunto de datos no siguen una distribución Normal). Para ello se realizarán tres tests de Kruskal-Wallis, uno para cada una de las tres variables comentadas ("alcohol","sulphates" y "volatile.acidity"):

```
kruskal.test(alcohol ~ tipo_vino, data = dd_clean)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  alcohol by tipo_vino
## Kruskal-Wallis chi-squared = 234.15, df = 2, p-value < 2.2e-16
```

Vemos que el porcentaje de alcohol es diferente en los vinos Buenos, Malos y Normales ya que el p.valor asociado al test no paramétrico de Kruskal-Wallis es muy inferior al nivel de significación del 5%.

```
kruskal.test(sulphates ~ tipo_vino, data = dd_clean)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  sulphates by tipo_vino
## Kruskal-Wallis chi-squared = 150.38, df = 2, p-value < 2.2e-16
```

Por otro lado, el nivel de sulfatos también es diferente en los vinos Buenos, Malos y Normales, porque el p.valor asociado al test no paramétrico de Kruskal-Wallis es muy inferior al nivel de significación del 5%.

```
kruskal.test(volatile.acidity ~ tipo_vino, data = dd_clean)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  volatile.acidity by tipo_vino
## Kruskal-Wallis chi-squared = 169.06, df = 2, p-value < 2.2e-16
```

En este tercer caso, la cantidad de ácido acético también es diferente en los vinos Buenos, Malos y Normales, porque el p.valor asociado al test no paramétrico de Kruskal-Wallis es muy inferior al nivel de significación del 5%.

Por lo tanto, después de realizar estos tests de hipótesis podemos confirmar que estas tres variables parecen ser relevantes para obtener un vino bueno, malo o normal.

4.3.3. Modelo de regresión cuantílica

Finalmente, el tercer método que se realizará será una regresión cuantílica. Este método es una alternativa robusta al método de los mínimos cuadrados ordinarios (regresión lineal) que se puede utilizar cuando algunas de las suposiciones básicas de la regresión lineal no se cumplen (en este caso, los datos no siguen una distribución Normal y la variabilidad no es constante). La peculiaridad de este método es que trata de predecir un cuantil de la variable dependiente en lugar de la media aritmética. Este cuantil puede ser el que nosotros deseeemos, pero para simplificar los resultados y las interpretaciones definiremos el cuantil 50 como objetivo de la predicción, es decir, la mediana. A continuación se construirá el modelo de regresión cuantílica mediante la función "rq" del paquete "quantreg":

```
library(quantreg)

dd_q <- dd_clean[,-13]

lm_q <- rq(quality ~.,
           data = dd_q,
           tau = 0.5)
summary(lm_q)
```

```
##
## Call: rq(formula = quality ~ ., tau = 0.5, data = dd_q)
##
## tau: [1] 0.5
##
## Coefficients:
##                       Value     Std. Error t value   Pr(>|t|)
## (Intercept)           60.07408  21.98980    2.73191   0.00637
## fixed.acidity          0.07990   0.02878    2.77580   0.00557
## volatile.acidity      -0.82318   0.10425   -7.89635   0.00000
## citric.acid           -0.20815   0.13911   -1.49637   0.13476
## residual.sugar         0.05972   0.01582    3.77435   0.00017
## chlorides             -1.63821   0.39501   -4.14732   0.00004
## free.sulfur.dioxide    0.00201   0.00158    1.27505   0.20248
## total.sulfur.dioxide  -0.00276   0.00038   -7.19375   0.00000
## density              -58.39849  22.46837   -2.59914   0.00943
## pH                    -0.05905   0.18815   -0.31383   0.75369
## sulphates              1.09303   0.15220    7.18162   0.00000
## alcohol                0.30032   0.02986   10.05787   0.00000
```

Antes de entrar en detalle con la salida del modelo se debe destacar que se ha creado un nuevo conjunto de datos llamado "dd_q" excluyendo la variable "tipo_vino" ya que en este caso no entrará en el estudio. Después de visualizar la salida del modelo de regresión cuantílica construido, se puede observar que las variables más significativas son "volatile.acidity", "total.sulfur.dioxide", "sulphates" y "alcohol" ya que son las que tienen un p.valor asociado al test t Student de significación de parámetros del modelo más pequeño. Por lo tanto, se puede concluir que estas variables son las que más afectarán a la calidad del vino. Estos resultados están muy relacionados con los resultados obtenidos anteriormente con el cálculo de las correlaciones entre las variables del conjunto de datos y la variable "quality" ya que las variables que han resultado más significativas en el modelo de regresión cuantílica son las que estaban más correlacionadas con la variable "quality". Por otro lado, para finalizar, se debe destacar que la variable "citric.acid" era de las variables que más correlación tenían con la variable "quality" (con una correlación del 21%) pero en el modelo ha resultado ser no significativa ya que tiene un p.valor asociado al test de significación de parámetros superior al nivel de significación del 5% (en concreto 0.13476).

Ejercicio 5 [2 puntos]. Representación de los resultados.

Representación de los resultados a partir de tablas y gráficas.

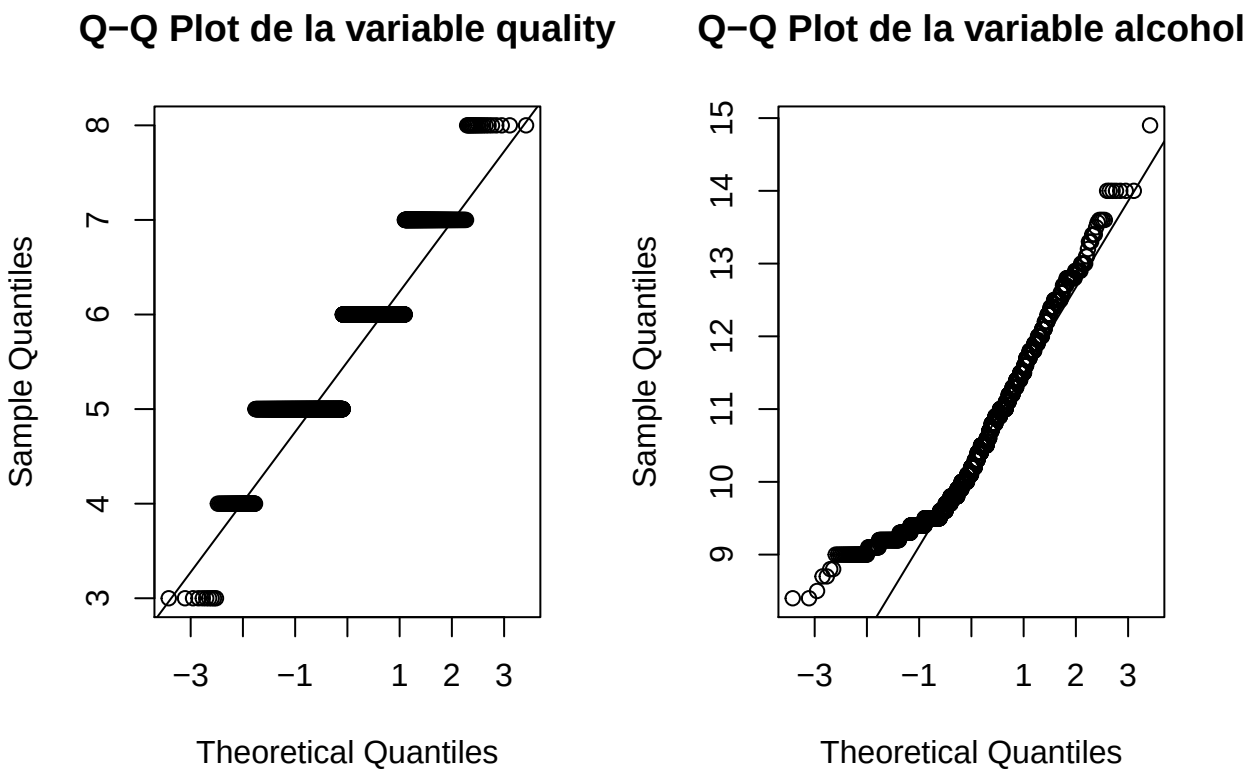
En este quinto ejercicio de la práctica se mostrarán de forma gráfica algunos de los resultados obtenidos en el ejercicio anterior.

En primer lugar, se mostrará un gráfico Q-Q de la variable "quality" y de la variable "alcohol" para observar que evidentemente no sigues una distribución Normal:

```
par(mfrow=c(1,2))

qqnorm(dd_clean$quality,main="Q-Q Plot de la variable quality")
qqline(dd_clean$quality)

qqnorm(dd_clean$alcohol,main="Q-Q Plot de la variable alcohol")
qqline(dd_clean$alcohol)
```



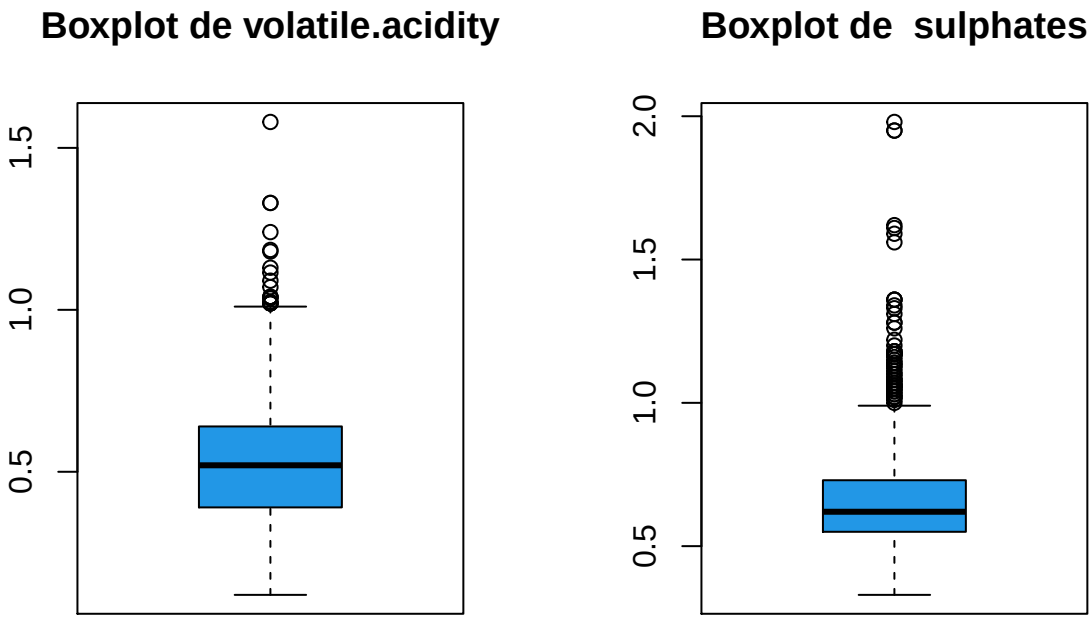
La salida de este gráfico Q-Q es muy interesante porque se puede apreciar a simple vista que ninguna de las dos variables siguen una distribución Normal ya que sus valores no se ajustan a la línea de cuantiles teóricos de una distribución Normal. Este ejercicio se podría repetir para todas las demás variables pero se obtendría el mismo resultado, ya que ninguna de ellas sigue una distribución Normal.

A continuación, en segundo lugar, se adjuntan dos gráficos de caja o boxplot para mostrar los valores extremos de las variables “volatile.acidity” y “sulphates”:

```
par(mfrow=c(1,2))

boxplot(dd_clean$volatile.acidity,col=44,main="Boxplot de volatile.acidity")

boxplot(dd_clean$sulphates,col=44,main="Boxplot de  sulphates")
```

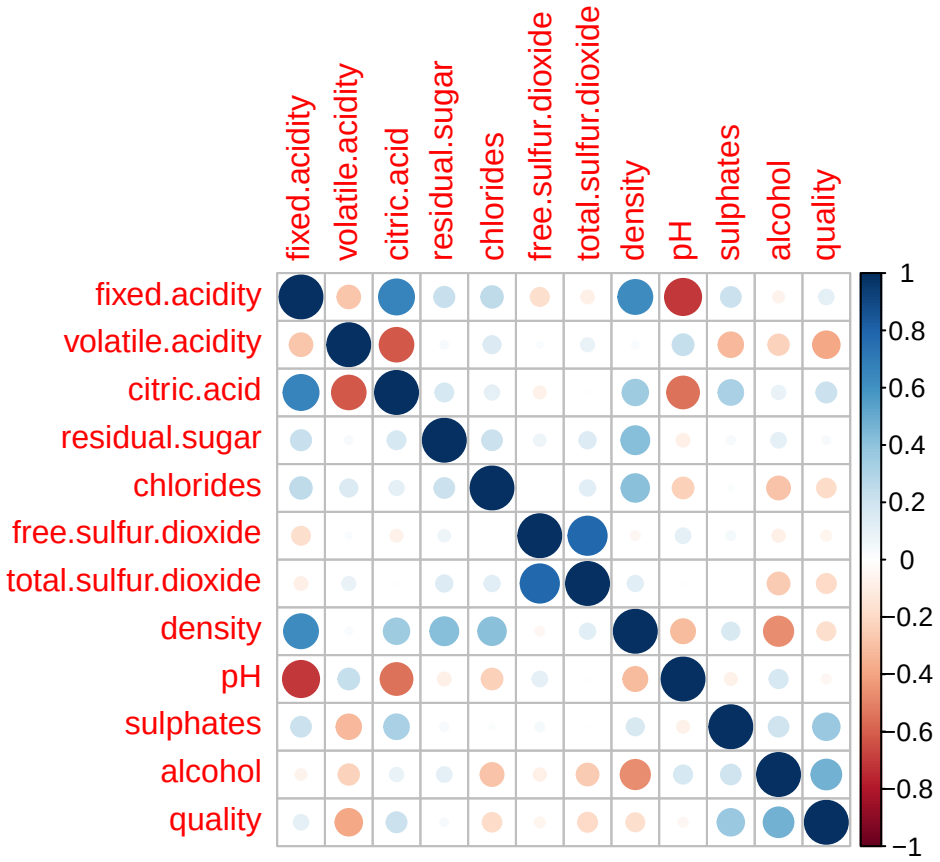


Después de visualizar los boxplots de estas dos variables se puede observar que aparecen bastantes valores extremos pero que no parecen ser tan influyentes como el de la variable “citric.acid”,y por eso no se han eliminado de la muestra de estudio, ya que el valor extremo que se ha analizado en dicha variable estaba muy alejado del resto de la distribución.

En tercer lugar se construirá un gráfico de correlaciones (mediante la función *corrplot* del paquete *corrplot*) entre las variables del conjunto de datos. Este gráfico es análogo a la matriz de correlaciones calculada en el ejercicio enterior pero permite visualizar de forma más intuitiva y rápida que variables están más relacionadas entre ellas.

```
library(corrplot)

corrplot(m_cor,method="circle")
```



En nuestro caso, como tenemos una variable de salida y el resto son variables explicativas, únicamente nos interesaría visualizar la información de la última fila o de la última columna, ya que son las que muestran las correlaciones entre la variable "quality" y el resto de variables. Sin embargo, esta matriz es muy útil para conocer relaciones entre las variables pH y fixed.acidity, pH y citric.acid y entre las variables citric.acid y volatile.acidity, las cuales no se habían tenido en cuenta anteriormente y podrían servir para tomar decisiones importantes a futuro.

Finalmente, se muestra una tabla con los parámetros estimados mediante el modelo de regresión cuantílica:

```
lm_q
```

```
## Call:
## rq(formula = quality ~ ., tau = 0.5, data = dd_q)
##
## Coefficients:
##          (Intercept)        fixed.acidity     volatile.acidity
##         60.074076413          0.079900443         -0.823181906
##          citric.acid       residual.sugar            chlorides
##         -0.208154333          0.059719638         -1.638213766
##  free.sulfur.dioxide total.sulfur.dioxide              density
##          0.002008549         -0.002758908        -58.398489525
##                   pH            sulphates              alcohol
##         -0.059047321          1.093032771          0.300319748
##
## Degrees of freedom: 1598 total; 1586 residual
```

A continuación, para finalizar con este quinto ejercicio de la práctica se interpretarán los valores de los parámetros asociados a las variables "alcohol" y "sulphates" ya que son dos de las que resultaron más significativas para modelar la calidad de un vino:

- El parámetro asociado a la variable "alcohol" es 0.30. Esto se interpreta de tal forma que si el porcentaje de alcohol en un vino se incrementa en un 1%, la calidad del vino aumenta en 0.30 puntos.
- El parámetro asociado a la variable "sulphates" es 1.09. Esto se interpreta de tal forma que si la cantidad de sulfatos en un vino se incrementan en una unidad, la calidad del vino aumenta en 1.09 puntos.

Por lo que nos muestran los parámetros del modelo, a mayores niveles de alcohol y sulfatos, mejores serán los vinos.

Ejercicio 6 [0.5 puntos]. Resolución del problema.

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

A partir de los resultados obtenidos en el análisis realizado en esta práctica se han obtenido las siguientes conclusiones:

- La calidad del vino se puede modelar a partir de algunas características recogidas en el conjunto de datos utilizado en esta práctica.
- Los aspectos más importantes para definir la calidad del vino son el porcentaje de alcohol, la cantidad de sulfatos, la cantidad de formas libres y unidas de SO2 y la cantidad de ácido acético.
- Dentro de estas características relevantes que se han comentado en el punto anterior, se debe destacar a mayores niveles de alcohol y de sulfatos y a menores niveles de formas libres y unidas de SO2 y de ácido acético, mayor es la calidad del vino.
- Hay algunas características de los vinos que no son relevantes para obtener una buena calidad. Estas son el ácido cítrico, la cnatidad de formas libres de SO2 y el pH.

Finalmente, es importante destacar que los resultados permiten responder al problema o pregunta que se ha planteado al inicio de esta práctica ya que después de realizar el análisis exaustivo en esta PRA 2, se han obtenido conclusiones que nos permitirían tomar decisiones importantes para mejorar la calidad del vino.

Ejercicio 7 [2 puntos]. Código.

Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

En este último ejercicio de la práctica se debe destacar que el documento PDF se ha generado mediante el software R Markdown. Por lo tanto se adjunta en el Github, juntamente con el documento PDF, el archivo ".Rmd" con el código R que se ha utilizado para solucionar la práctica y realizar el análisis de datos que se ha ido detallando en este documento.

Contribuciones

Las contribuciones de los miembros del equipo en las tareas de la práctica son las siguientes:

Contribuciones	**Firma**
Investigación previa	ASB, VMH
Redacción de las respuestas	ASB, VMH
Desarrollo código	ASB, VMH