Tipología y ciclo de vida de los datos
M2.851 - Aula 3

Práctica 2:
Limpieza y análisis de datos

Red Wine Quality Analysis

Rubén Gonzalo Coll Menéndez

Índice

Detalles de la actividad

Descripción

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas. Tendréis que entregar un solo archivo con el enlace Github (https://github.com) donde se encuentren las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github.

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos. Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Resolución

1. Descripción del dataset

El conjunto de datos "Red Wine Quality" objeto de análisis se ha obtenido a partir de Kaggle (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009) y está constituido por 12 características (columnas) que presentan un total de 1599 vinos (filas o registros).

Entre los campos de este conjunto de datos, encontramos los siguientes:

- quality: (calidad) Indica la calidad del vino, que va de 1 a 10. Aquí, cuanto mayor es el valor, mejor es el vino.
- fixed.acidity: Indica la cantidad de ácido tartárico en el vino y se mide en g/dm^3.
- volatile.acidity: Indica la cantidad de ácido acético en el vino. Se mide en g/dm^3.
- citric.acid: Indica la cantidad de ácido cítrico en el vino. También se mide en g/dm^3.
- residual.sugar: Indica la cantidad de azúcar que queda en el vino una vez finalizado el proceso de fermentación. Se mide en g/dm^3.
- chlorides: Indica la cantidad de sales en el vino.
- free.sulfur.dioxide: Mide la cantidad de dióxido de azufre (SO_2) en forma libre. Se mide en mg/dm^3.
- total.sulfur.dioxide: Mide la cantidad total de SO_2 en el vino ($SO_2\ libre + SO_2\ combinado$). Se mide en mg/dm^3.
- density: Indica la densidad del vino y depende en gran medida del porcentaje de alcohol y del contenido de azúcar. Se mide en g/dm^3
- pH: Indica cuan ácido o básico un vino es en una escala que va desde el 0 (muy ácido) hasta el 14 (muy básico). La mayoría de vinos están en entre 3 y 4 en dicha escala.
- sulphates: Indica la cantidad de sulfitos que contiene el vino. Puede contribuir a los niveles de dióxido de sulfuro (SO_2).
- alcohol: Indica el contenido de alcohol del vino. Se mide en porcentaje (%).

2. Objetivos del análisis

A partir del conjunto de datos descrito anteriormente se plantea la problemática de determinar qué variables influyen más sobre la calidad del vino. Además, se procederá a crear un modelo de regresión que permita predecir la calidad del vino dadas una serie de características.

Esto es especialmente importante ya que nos permitirá, a partir de las mediciones de las diferentes características del vino, determinar la calidad que tendrá frente al resto de vinos y/o competidores siendo un factor clave en la elección del precio.

Por otro lado, también nos permitirá influir en alguna de dichas características con el fin de aumentar la calidad del vino antes de que se envase y se comercialice. Esto es especialmente relevante ya que cambia totalmente la manera de producir vino y, en lugar de crear un vino para a posteriori medir la calidad, seríamos capaces de determinar la calidad del mismo de manera "artificial" en un punto temprano de la producción.

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

Antes de comenzar con la limpieza de los datos, procedemos a realizar la lectura del fichero en formato CSV en el que se encuentra. El resultado devuelto por la llamada a la función read.csv() será un objeto data.frame:

```
> # Set working directory
> setwd("C:/Users/RCOLL/Downloads/wine_data")
>
> # Lectura de los datos
> df <- read.csv("winequality-red.csv", header=TRUE,
+                sep=",", na.strings="NA", dec=".", strip.white=TRUE)
>
> class(df)
[1] "data.frame"
```

Todas las variables de los que disponemos son, a priori, relevantes y se corresponden con los atributos que normalmente se suelen medir de un vino por lo que será conveniente tenerlos en consideración para el análisis.

A continuación presentamos dichos atributos y el tipo de datos que contienen donde destaca que todos ellos son numéricos.

```
> # Primer vistazo
> str(df)
'data.frame':   1599 obs. of  12 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

4. Limpieza de los datos

4.1. Gestión de duplicados en los datos

En relación a la preparación y limpieza de los datos para el posterior análisis, lo primero que verificamos es la existencia de valores duplicados. Mediante el uso de la siguiente sentencia de R, comprobamos que hay 240 registros duplicados de un total de 1.599 de los que se compone el conjunto de datos.

```
> # Duplicados
> duplicates <- df[duplicated(df),]
> nrow(duplicates)
[1] 240
> duplicates
     fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol quality
5              7.4             0.700        0.00            1.90     0.076                11                   34 0.99780 3.51      0.56     9.4       5
12             7.5             0.500        0.36            6.10     0.071                17                  102 0.99780 3.35      0.80    10.5       5
28             7.9             0.430        0.21            1.60     0.106                10                   37 0.99660 3.17      0.91     9.5       5
41             7.3             0.450        0.36            5.90     0.074                12                   87 0.99780 3.33      0.83    10.5       5
66             7.2             0.725        0.05            4.65     0.086                 4                   11 0.99620 3.41      0.39    10.9       5
77             8.8             0.410        0.64            2.20     0.093                 9                   42 0.99860 3.54      0.66    10.5       5
92             8.6             0.490        0.28            1.90     0.110                20                  136 0.99720 2.93      1.95     9.9       6
94             7.7             0.490        0.26            1.90     0.062                 9                   31 0.99660 3.39      0.64     9.6       5
103            8.1             0.545        0.18            1.90     0.080                13                   35 0.99720 3.30      0.59     9.0       6
106            8.1             0.575        0.22            2.10     0.077                12                   65 0.99670 3.29      0.51     9.2       5
115            7.8             0.560        0.19            1.80     0.104                12                   47 0.99640 3.19      0.93     9.5       5
122            8.8             0.550        0.04            2.20     0.119                14                   56 0.99620 3.21      0.60    10.9       6
133            5.6             0.500        0.09            2.30     0.049                17                   99 0.99370 3.63      0.63    13.0       5
141            8.4             0.745        0.11            1.90     0.090                16                   63 0.99650 3.19      0.82     9.6       5
142            8.3             0.715        0.15            1.80     0.089                10                   52 0.99680 3.23      0.77     9.5       5
145            5.2             0.340        0.00            1.80     0.050                27                   63 0.99160 3.68      0.79    14.0       6
154            7.5             0.600        0.03            1.80     0.095                25                   99 0.99500 3.35      0.54    10.1       5
157            7.1             0.430        0.42            5.50     0.070                29                  129 0.99730 3.42      0.72    10.5       5
158            7.1             0.430        0.42            5.50     0.071                28                  128 0.99730 3.42      0.71    10.5       5
173            8.0             0.420        0.17            2.00     0.073                 6                   18 0.99720 3.29      0.61     9.2       6
177            7.3             0.380        0.21            2.00     0.080                 7                   35 0.99610 3.33      0.47     9.5       5
181            8.8             0.610        0.14            2.40     0.067                10                   42 0.99690 3.19      0.59     9.5       5
195            7.6             0.550        0.21            2.20     0.071                 7                   28 0.99640 3.28      0.55     9.7       5
207           12.8             0.300        0.74            2.60     0.095                 9                   28 0.99940 3.20      0.77    10.8       7
229            7.7             0.430        0.25            2.60     0.073                29                   63 0.99615 3.37      0.58    10.5       6
234            6.9             0.520        0.25            2.60     0.081                10                   37 0.99685 3.46      0.50    11.0       5
237            7.2             0.630        0.00            1.90     0.097                14                   38 0.99675 3.37      0.58     9.0       6
239            7.2             0.630        0.00            1.90     0.097                14                   38 0.99675 3.37      0.58     9.0       6
240            8.2             1.000        0.09            2.30     0.065                 7                   37 0.99685 3.32      0.55     9.0       6
```

Si bien es cierto que dichos registros con los mismos valores podrían pertenecer a diferentes vinos y por tanto no ser en realidad duplicados, al no tener más información sobre el vino calificado como podría ser un identificador o el nombre del mismo, hemos optado por considerar que son duplicados y hemos procedido a la eliminación de los mismos quedando el dataset con un total de 1.359 registros, a priori, válidos.

```
> df <- unique(df)
> nrow(df)
[1] 1359
```

4.2. Gestión de valores nulos

Una vez verificada la existencia de valores duplicados, procedemos a analizar la existencia de valores nulos. Para dicho análisis hemos considerado como valores nulos aquellos no imputados o con valores como puede ser NA o 0 dado que todas las características de las que disponemos son numéricas.

```
> # Número de ceros
> colSums(df == 0)
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides
                   0                    0                  118                   0                    0
 free.sulfur.dioxide total.sulfur.dioxide              density                  pH            sulphates
                   0                    0                    0                   0                    0
             alcohol              quality
                   0                    0
> # Número de vacios
> colSums(df == '')
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides
                   0                    0                    0                   0                    0
 free.sulfur.dioxide total.sulfur.dioxide              density                  pH            sulphates
                   0                    0                    0                   0                    0
             alcohol              quality
                   0                    0
> # Número de nulos
> colSums(is.na(df))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides
                   0                    0                    0                   0                    0
 free.sulfur.dioxide total.sulfur.dioxide              density                  pH            sulphates
                   0                    0                    0                   0                    0
             alcohol              quality
                   0                    0
```

Como observamos en la captura de pantalla anterior, si bien no se han encontrado valores nulos sí que existen 118 registros que tienen un valor de cero para el campo “citric.acid”. A pesar de que en según en qué circunstancias se podrían considerar dichos registros como erróneos siempre hay que contextualizar el análisis. En este caso hemos estado recabando información acerca del rango de valores de ácido cítrico en los que se suelen mover los vinos tintos y parece ser que el cero es un valor aceptado, por este motivo decidimos no eliminar o imputar dichos registros.

4.3. Gestión de valores extremos o atípicos

El último paso que realizaremos en la limpieza de los datos será el análisis de los valores extremos. Para ello vamos a proceder a extraer los principales estadísticos de cada una de las variables y a graficar los valores de las mismas lo que nos permitirá detectar la existencia de valores atípicos en cada uno de los atributos.

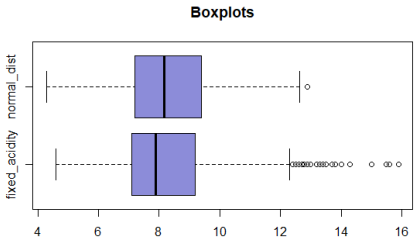
```
> summary(df)
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides       free.sulfur.dioxide
 Min.   : 4.600   Min.   :0.1200   Min.   :0.0000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
 1st Qu.: 7.100   1st Qu.:0.3900   1st Qu.:0.0900   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
 Median : 7.900   Median :0.5200   Median :0.2600   Median : 2.200   Median :0.07900   Median :14.00
 Mean   : 8.311   Mean   :0.5295   Mean   :0.2723   Mean   : 2.523   Mean   :0.08812   Mean   :15.89
 3rd Qu.: 9.200   3rd Qu.:0.6400   3rd Qu.:0.4300   3rd Qu.: 2.600   3rd Qu.:0.09100   3rd Qu.:21.00
 Max.   :15.900   Max.   :1.5800   Max.   :1.0000   Max.   :15.500   Max.   :0.61100   Max.   :72.00
 total.sulfur.dioxide    density             pH           sulphates         alcohol         quality
 Min.   :  6.00       Min.   :0.9901   Min.   :2.74   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
 1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.21   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
 Median : 38.00       Median :0.9967   Median :3.31   Median :0.6200   Median :10.20   Median :6.000
 Mean   : 46.83       Mean   :0.9967   Mean   :3.31   Mean   :0.6587   Mean   :10.43   Mean   :5.623
 3rd Qu.: 63.00       3rd Qu.:0.9978   3rd Qu.:3.40   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :289.00       Max.   :1.0037   Max.   :4.01   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

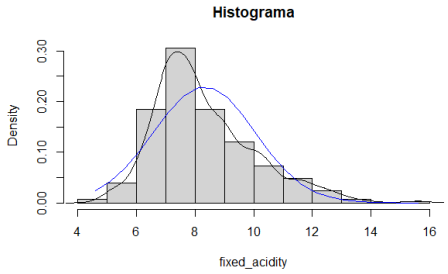
- fixed.acidity

Atendiendo a los principales estadísticos de la variable que representa el ácido tartárico del vino, observamos como, en el boxplot, hay una serie de observaciones que estadísticamente se podrían considerar “outliers” dado que superan en 1,5 veces el rango intercuartílico. En concreto se trata de 42 observaciones que superan el “bigote” superior del diagrama de caja situado en un valor de 12,3 g/dm^3.

```
> summary(df)
 fixed.acidity
 Min.   : 4.600
 1st Qu.: 7.100
 Median : 7.900
 Mean   : 8.311
 3rd Qu.: 9.200
 Max.   :15.900
```

```
> # Num de outliers
> length(b$out)
[1] 42
>
> # Outliers
> b$out
  [1] 12.80000 15.00000 12.50000 13.30000 13.40000 12.40000 12.50000 13.80000 13.50000 12.60000 12.50000
 [12] 12.80000 14.00000 13.70000 12.70000 12.50000 12.80000 12.60000 15.60000 12.50000 13.00000 12.50000
 [23] 13.30000 12.40000 12.50000 12.90000 14.30000 12.40000 15.50000 15.60000 13.00000 12.70000 12.40000
 [34] 12.70000 13.20000 13.20000 15.90000 13.30000 12.90000 12.60000 12.60000 12.87966
>
> # Rango intercuartilico
> b$stats
      [,1]     [,2]
[1,]  4.6 4.303648
[2,]  7.1 7.212500
[3,]  7.9 8.160911
[4,]  9.2 9.398671
[5,] 12.3 12.626259
```





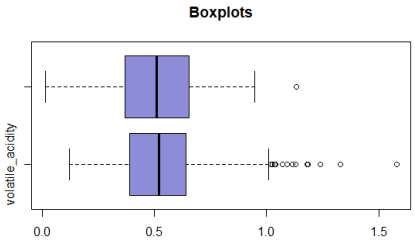
A pesar de que estadísticamente podríamos clasificarlos como valores atípicos, están dentro del rango funcional de la variable y por tanto, decidimos no vamos a excluir dichas observaciones.

- volatile.acidity

Si analizamos los principales estadísticos de la variable que representa el ácido acético del vino, observamos que se mueve en un rango de valores que van entre 0,12 y 1,58 g/dm^3 situándose la media de las observaciones en 0,5295 g/dm^3; muy próxima a la mediana que está en 0,52 g/dm^3.

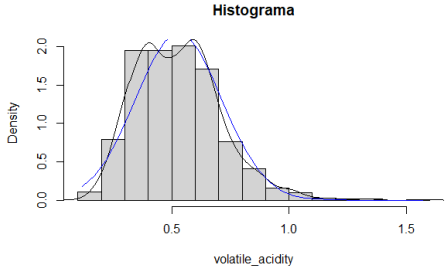
```
volatile.acidity
Min.   :0.1200
1st Qu.:0.3900
Median :0.5200
Mean   :0.5295
3rd Qu.:0.6400
Max.   :1.5800
```

En el diagrama de caja siguiente observamos como hay un total de 20 observaciones que quedan por encima del valor 1,01 g/dm^3 que representa el bigote superior de dicho gráfico y que por tanto podrían considerarse valores atípicos estadísticamente hablando.




```
> # Num de outliers
> length(b$out)
[1] 20
>
> # Outliers
> b$out
 [1] 1.130000 1.020000 1.070000 1.330000 1.330000 1.040000 1.090000 1.040000 1.240000 1.185000 1.020000
[12] 1.035000 1.025000 1.115000 1.020000 1.020000 1.580000 1.180000 1.040000 1.134046
>
> # Rango intercuartilico
> b$stats
     [,1]       [,2]
[1,] 0.12 0.01599747
[2,] 0.39 0.36764204
[3,] 0.52 0.51150655
[4,] 0.64 0.65556405
[5,] 1.01 0.94653412
```

Por otro lado, si representamos el número de observaciones existentes para cada rango de valores de la variable “volatile_acidity” observamos como no sigue exactamente una distribución normal existiendo una cierta asimetría hacía la derecha que queda representada por la cola que se puede ver en el siguiente gráfico.



Al contrario de lo que pasaba con la primera variable analizada en la que a pesar de existir valores estadísticamente atípicos, estaban dentro del rango de valores posibles de la variable considerada, en

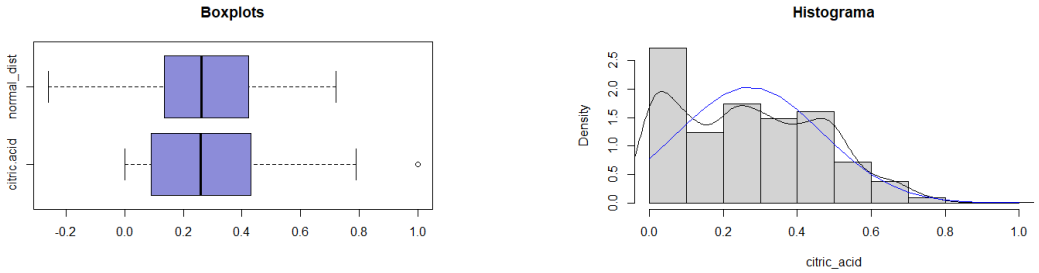
este caso dicho rango está entre los 0,3 y 0,7 g/l (g/dm^3)[1] por lo que procedemos a excluir dichos registros.

- citric.acid

Vemos como la variable "citric.acid" que indica la cantidad de ácido cítrico presente en el vino se mueve en el rango de valores comprendido entre 0 y 1 g/dm^3, situándose la media en 0,27 g/dm^3. El rango normal para vinos tintos y blancos se encuentra entre 0,2 y 0,6 g/dm^3, si bien el límite superior está marcado a nivel europeo por ley y no puede superar en ningún caso 1 g/dm^3 [2].

```
  citric.acid
Min.   :0.0000
1st Qu.:0.0900
Median :0.2600
Mean   :0.2723
3rd Qu.:0.4300
Max.   :1.0000
```

En nuestro conjunto de datos vemos como la mayoría de vinos se encuentran en el rango normal y tan solo 1 registro iguala el límite superior establecido por ley. Si bien podemos considerar dicho registros como un valor extremo, está dentro del dominio de valores que puede tomar esta variable y por tanto no vamos a proceder a tratarlo de ninguna manera dado que consideramos que no es un error.



- residual.sugar

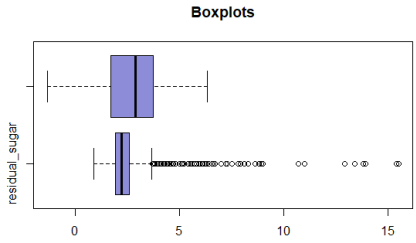
La variable "resigual.sugar" indica la cantidad de azúcar que queda en el vino una vez finalizado el proceso de fermentación. Como vemos en el resumen estadístico de las principales medidas, las observaciones del conjunto de datos se mueven en un rango que va entre 0,9 y 15,5 gramos de azúcar por litro g/l estando la media en 2,2 g/l. Atendiendo a la clasificación de los vinos según la cantidad de azúcar[3], estos pueden ir desde los 1 g/l en el caso de los más secos hasta más de 200 g/l de los clasificados como dulces.

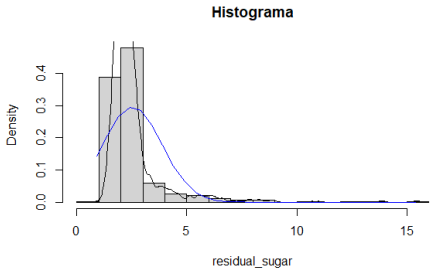
```
residual.sugar
Min.   : 0.900
1st Qu.: 1.900
Median : 2.200
Mean   : 2.523
3rd Qu.: 2.600
Max.   :15.500
```

A continuación podemos ver la distribución de las observaciones, en las que destaca principalmente la asimetría que presenta el histograma.

[1] https://quercuslab.es/blog/determinacion-acidez-volatil-en-vinos/
[2] https://www.vason.com/uploads/MediaGalleryArticoliDocumenti/%C3%81cido%20citrico%202_0%20es.pdf

[3] https://www.catadelvino.com/blog-cata-vino/que-son-los-azucares-residuales-en-el-vino



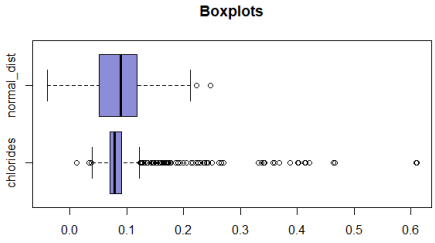


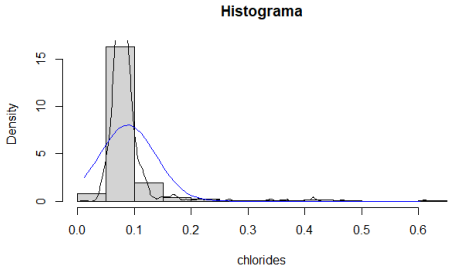
A pesar de dicha simetría y del número de observaciones que están por encima del bigote superior en el diagrama de caja, en principio todos los valores parecen estar en el rango que puede tomar la variable desde un punto de vista funcional y, por lo tanto, no vamos a excluirlos del análisis.

- chlorides

En cuanto a la variable "chlorides" que representa la cantidad de sales presente en el vino, observamos como se encuentra muy concentrada en un rango pequeño de valores. Si bien esto provoca que estadísticamente se detecten muchos ouliers al estar 1,5 veces por encima (o debajo) del rango intercuartílico, son valores normales[4] que se pueden hallar en el vino y no vamos tratar dichos registros.

```
   chlorides
Min.   :0.01200
1st Qu.:0.07000
Median :0.07900
Mean   :0.08812
3rd Qu.:0.09100
Max.   :0.61100
```





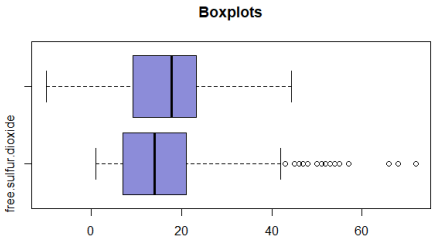
- free.sulfur.dioxide

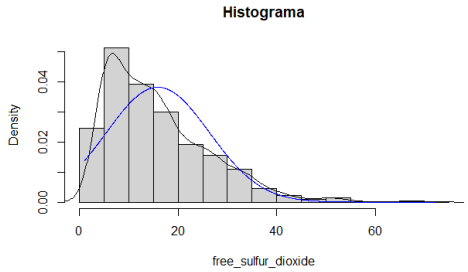
La variable "free.sulfur.dioxide" representa la cantidad de sulfitos libres presentes en el vino. Como observamos en el resumen estadístico, todas las observaciones se hallan en un rango que va entre 1 y 72 mg/dm^3.

```
free.sulfur.dioxide
Min.   : 1.00
1st Qu.: 7.00
Median :14.00
Mean   :15.89
3rd Qu.:21.00
Max.   :72.00
```

Si bien observamos como en histograma hay cierta asimetría, todas las observaciones se encuentran en un rango de valores que se podría considerar normal y por lo tanto, no vamos a excluir ningún registro.

[4] https://www.wineaustralia.com/labelling/wine-production-standard





- total.sulfur.dioxide

En la variable “total.sulfur.dioxide”, que representa la cantidad de sulfitos total que contiene el vino, hemos encontrado quizás el punto más delicado de este análisis. Teniendo en cuenta que los vinos que conforman el conjunto de datos son variedades el vino portugués "Vinho Verde" y por tanto al ser de una región europea quedan sujetos a la legislación de dicho ámbito geográfico, nos extraña encontrar valores máximos de 289 mg/dm^3 cuando por ley las cantidades máximas de sulfitos que puede contender un vino tinto para ser comercializado son de 160 mg/dm^3 [5].

```
total.sulfur.dioxide
Min.   :  6.00
1st Qu.: 22.00
Median : 38.00
Mean   : 46.83
3rd Qu.: 63.00
Max.   :289.00
```

Como vemos a continuación hay un total de 3 observaciones que superan dicho límite legal establecido y que por lo tanto no sería posible su comercialización.

```
> outliers <- filter(df, total.sulfur.dioxide > 160)
> outliers
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide
1           6.1             0.21        0.40            1.4     0.066                40.5
2           7.9             0.30        0.68            8.3     0.050                37.5
3           7.9             0.30        0.68            8.3     0.050                37.5
  total.sulfur.dioxide density   pH sulphates alcohol quality
1                  165 0.99120 3.25      0.59    11.9       6
2                  278 0.99316 3.01      0.51    12.3       7
3                  289 0.99316 3.01      0.51    12.3       7
```

Por lo comentado anteriormente, hemos decidido proceder a la exclusión de dichos registros.

- density

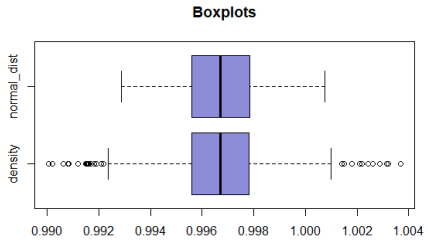
La variable “density” mide la densidad del vino en g/dm^3. Este atributo es muy sensible a la temperatura y para poder analizarlo es importante saber a que temperatura ha sido tomada dicha medición. Si bien el conjunto de datos no nos proporciona dicha información, vamos a hacer la suposición que ha sido a 20 °C.

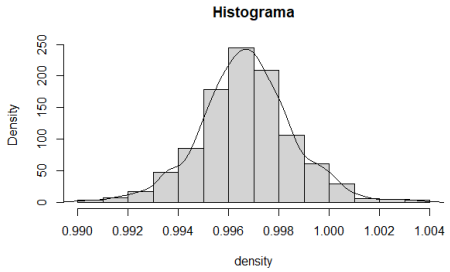
El rango que se podría considerar normal para la densidad de un vino tinto medido a una temperatura de 20 °C va desde los 0,9910 y los 1,0080 g/dm^3 [6]. Como vemos en el diagrama de caja y en el histograma, la distribución de esta variable se asemeja mucho a una distribución normal. En el diagrama de caja vemos como algunas observaciones quedan fuera de los bigotes superiores e inferiores pero como tanto el valor mínimo como máximo de nuestras observaciones se encuentran dentro del rango descrito anteriormente como normal no vamos a considerarlas valores extremos.

```
    density
Min.   :0.9901
1st Qu.:0.9956
Median :0.9967
Mean   :0.9967
3rd Qu.:0.9978
Max.   :1.0037
```

[5] http://www.acenologia.com/cienciaytecnologia/azufre_seguridad_vinos_ecologicos_cienc173_1219.htm

[6] http://www.usc.es/caa/MetAnalisisStgo1/enologia.pdf





- pH

Es pH es una unidad de medida de la acidez o alcalinidad de una disolución. Se mide en una escala del 0 al 14 siendo 14 muy básico o alcalino, y los pH cercanos a cero son disoluciones muy ácidas. El pH del vino suele oscilar entre 2,8 y 4, siendo 2,8 un vino extremadamente ácido y un vino con pH 4 es un vino plano sin acidez.[7]

Como vemos a continuación, el pH de nuestra muestra oscila entre un valor de 2,74 y de 4,01 estando en ambos casos justo por debajo y por encima, respectivamente de lo que en teoría se considera normal para el caso del vino.

```
       pH
Min.   :2.74
1st Qu.:3.21
Median :3.31
Mean   :3.31
3rd Qu.:3.40
Max.   :4.01
```

Dado que la diferencia es menor, no excluiremos ninguna de las observaciones que rompen dichos límites dado que también nos interesa comprobar si dichos vinos que no cumplen con lo habitual tienen notas más bajas debido a ello.

- sulphates

Para determinar los valores habituales de sulfatos que un vino puede tener nos apoyaremos en la regulación española que, mediante el reglamento de la Ley nº18.455 que fija las normas sobre producción, elaboración, comercialización de alcoholes etílicos, bebidas alcohólicas y vinagres establece unos valores máximos de g/l.

Si bien desconocemos a que regulación están sujetos los vinos incluidos en el conjunto de datos del que disponemos, vemos como se mueven en un rango que va desde los 0,33 g/l hasta los 2,00 g/l y por tanto, consideraremos que no existen valores atípicos en lo que a esta variable se refiere.

```
   sulphates
Min.   :0.3300
1st Qu.:0.5500
Median :0.6200
Mean   :0.6587
3rd Qu.:0.7300
Max.   :2.0000
```

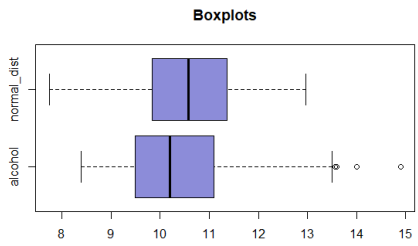
- alcohol

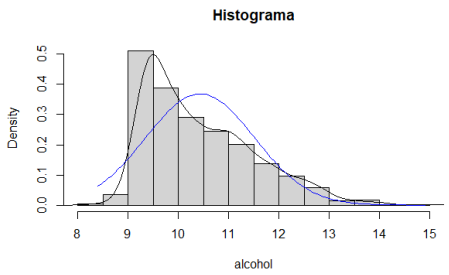
La variable “alcohol” representa el % de alcohol presente en el vino. Si observamos los principales estadísticos de este variable vemos como el rango de valores oscila entre un mínimo de 8,40% y un máximo de 14,90%.

```
    alcohol
Min.   : 8.40
1st Qu.: 9.50
Median :10.20
Mean   :10.43
3rd Qu.:11.10
Max.   :14.90
```

[7] https://catatu.es/blog/ph-vinos/

Después de haber analizado diferentes fuentes de información, hemos llegado a la conclusión que los valores de dicha variable están dentro de la graduación alcohólica habitual que va desde un 7% a un 16%[8] para un vino tinto y que por tanto no existen valores atípicos.



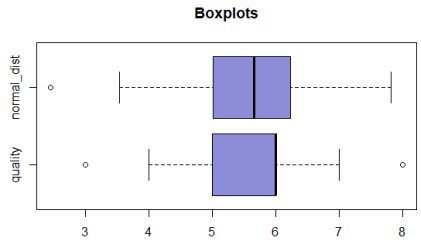


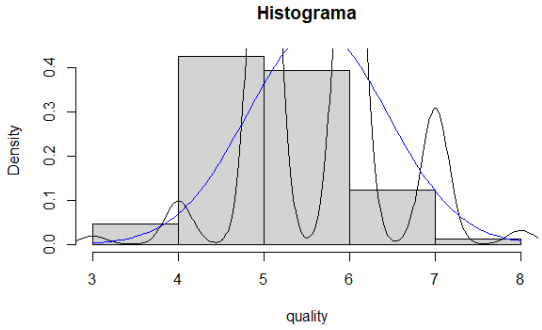
- quality

La variable “quality” indica la calidad del vino puntuándolo en una escala de 1 a 10 siendo está última la mayor puntuación posible. En el caso de nuestro conjunto de datos, se mueve en el rango comprendido entre una puntuación mínima de 3 y una máxima de 8, estando la media situada en 5,623.

```
    quality
 Min.   :3.000
 1st Qu.:5.000
 Median :6.000
 Mean   :5.623
 3rd Qu.:6.000
 Max.   :8.000
```

Dado que no hay valores fuera del domino de la variable no excluiremos ningún registro.





Una vez realizado el análisis de las principales características de las que disponemos, procedemos a la exclusión de aquellas observaciones que por algún motivo hemos considerado como atípicas. En este caso de estudio hemos decidido excluir únicamente aquellas observaciones que hemos marcado como outliers en relación a la variable “volatile.acidity” y la variable “total.sulfur.dioxide”.

```
> # Outliers
> outliers <- b$out
> df <- df[-which(df$volatile.acidity %in% outliers),]

> # Vinos con más de 160 mg/l
> outliers <- filter(df, total.sulfur.dioxide > 160)
> outliers <- outliers$total.sulfur.dioxide
>
> # Exclusion outliers - total_sulfur_dioxide
> df <- df[-which(df$total.sulfur.dioxide %in% outliers),]
```

Después de la exclusión de dichas observaciones el conjunto de datos queda configurado con un total de 1.337 registros y 12 columnas siendo el resumen de los principales estadísticos el siguiente:

[8] https://www.catadelvino.com/blog-cata-vino/como-varia-el-grado-de-alcohol-de-los-vinos

```
> nrow(df)
[1] 1337
> ncol(df)
[1] 12
```

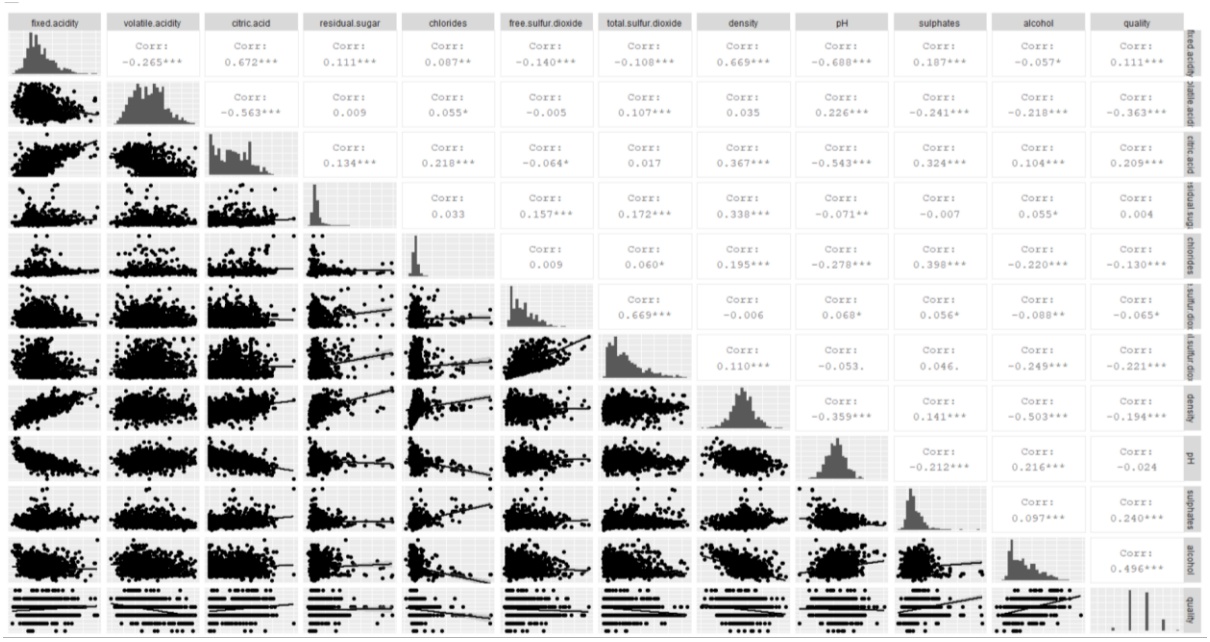
```
> summary(df)
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides       free.sulfur.dioxide
 Min.   : 4.600   Min.   :0.1200   Min.   :0.0000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
 1st Qu.: 7.100   1st Qu.:0.3900   1st Qu.:0.1000   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
 Median : 7.900   Median :0.5200   Median :0.2600   Median : 2.200   Median :0.07900   Median :14.00
 Mean   : 8.322   Mean   :0.5215   Mean   :0.2738   Mean   : 2.516   Mean   :0.08818   Mean   :15.86
 3rd Qu.: 9.200   3rd Qu.:0.6350   3rd Qu.:0.4300   3rd Qu.: 2.600   3rd Qu.:0.09100   3rd Qu.:21.00
 Max.   :15.900   Max.   :1.0100   Max.   :1.0000   Max.   :15.500   Max.   :0.61100   Max.   :72.00
 total.sulfur.dioxide    density             pH           sulphates         alcohol         quality
 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 :0.9967   Median :3.310   Median :0.6200   Median :10.10   Median :6.000
 Mean   : 46.38       Mean   :0.9967   Mean   :3.308   Mean   :0.6609   Mean   :10.43   Mean   :5.637
 3rd Qu.: 63.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :160.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

5. Análisis de los datos

Para la creación de un modelo que explique la calificación del vino contenida en la variable “quality” a partir de una serie de variables dadas, vamos a aplicar la técnica de análisis multivariante de dependencia conocida como regresión lineal múltiple. Esta técnica se basa en la estimación del peso que cada una de las variables (independientes) tiene en la explicación lineal de la variable dependiente “quality”.

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

Requisito fundamental antes de proceder con la creación del modelo de regresión, es la valoración de la existencia de relación y el grado de la misma entre cada una de las variables independientes y la variable dependiente. Para ello la matriz de correlaciones y el análisis visual de la distribución de los datos nos ayudará a detectar aquellas variables que de manera bivariante pueden tener un mayor impacto en la calidad del vino y el tipo de relación de las mismas (lineal, logarítmica, exponencial, etc).



Como vemos en la captura de pantalla anterior, en general podemos decir que todas las variables independientes de las que disponemos y utilizaremos para el análisis tienen una correlación baja con la variable dependiente “quality” algo que puede influir negativamente en la capacidad explicativa del modelo. Recordemos que la correlación es una medida que vive en el rango comprendido entre el -1 y el 1 donde cuan más próximo a uno de esos dos valores está se dice que tiene una correlación elevada (negativa o positiva respectivamente) y cuanto más se acerca a 0 se dice que la correlación es baja indicando una débil relación lineal entre las variables consideradas.

Aparte de las correlaciones de cada una de las variables con la variable dependiente, es importante analizar las relaciones existentes entre las variables independientes dado que correlaciones muy fuertes (multicolinealidad) nos podrían estar indicando que dichas variables explican lo mismo y que por tanto estaríamos aumentando dicho efecto. En este sentido destacan las variables que miden los ácidos del vino como son "citric.acid", "fixed.acidity", "volatile.acidity" y "pH" que tienen correlaciones relativamente altas (mayores a 0,5 en valor absoluto) entre ellas. Por otro lado, también observamos como las variables "total.sulphur.dioxide" y "free.sulphur.dioxide" tienen una correlación de 0,669. Esto es debido a que, como veíamos al inicio la variable "free.sulphur.dioxide" forma parte del cálculo de "total.sulphur.dioxide".

A pesar de los puntos comentados anteriormente y de tener variables con correlaciones relativamente altas, consideramos que dicho efecto no es lo suficientemente elevado como para excluirlas del análisis y por lo tanto procederemos a la elección del mejor modelo partiendo de todas las variables de las que disponemos.

Otro punto importante antes de proceder a realizar el análisis de regresión lineal múltiple es la normalización de los datos. Dado que tenemos variables con diferentes escalas de medida, es importante normalizar los datos de manera que los pesos finales de las variables en el modelo no se vean afectados por dichas diferencias en las escalas.

```
> # Escalamos los datos
> df <- scale(df)
> summary(df)
 fixed.acidity     volatile.acidity    citric.acid       residual.sugar      chlorides
 Min.   :-2.1406   Min.   :-2.37691   Min.   :-1.40456   Min.   :-1.19822   Min.   :-1.5358
 1st Qu.:-0.7020   1st Qu.:-0.77623   1st Qu.:-0.89280   1st Qu.:-0.46053   1st Qu.:-0.3654
 Median :-0.2417   Median :-0.00553   Median :-0.07398   Median :-0.23923   Median :-0.1838
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
 3rd Qu.: 0.5064   3rd Qu.: 0.67624   3rd Qu.: 0.79602   3rd Qu.: 0.05585   3rd Qu.: 0.0584
 Max.   : 4.3618   Max.   : 2.89941   Max.   : 3.71307   Max.   : 9.57203   Max.   :10.5519
 free.sulfur.dioxide total.sulfur.dioxide    density             pH           sulphates
 Min.   :-1.4278     Min.   :-1.2252      Min.   :-3.554624   Min.   :-3.69803   Min.   :-1.9345
 1st Qu.:-0.8532     1st Qu.:-0.7450      1st Qu.:-0.597217   1st Qu.:-0.63776   1st Qu.:-0.6475
 Median :-0.1828     Median :-0.2648      Median : 0.001752   Median : 0.01336   Median :-0.2379
 Mean   : 0.0000     Mean   : 0.0000      Mean   : 0.000000   Mean   : 0.00000   Mean   : 0.0000
 3rd Qu.: 0.5116     3rd Qu.: 0.4855      3rd Qu.: 0.590025   3rd Qu.: 0.59937   3rd Qu.: 0.4056
 Max.   : 5.3721     Max.   : 7.2683      Max.   : 3.729262   Max.   : 4.57121   Max.   : 7.8354
    alcohol           quality
 Min.   :-1.8689   Min.   :-3.2543
 1st Qu.:-0.8567   1st Qu.:-0.7885
 Median :-0.2126   Median : 0.4444
 Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.6156   3rd Qu.: 0.4444
 Max.   : 4.1124   Max.   : 2.9102
```

5.2. 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.

A continuación procedemos a la selección de los mejores predictores para el modelo de regresión lineal múltiple. Para ello existen diversos métodos: jerárquico, de entrada forzosa, paso a paso… pero en este caso utilizaremos este último método con una estrategia doble o mixta según la cual iremos introduciendo y extrayendo las variables independientes de las que disponemos e iremos evaluando la implicación que cada una de ellas tiene, basándonos en el criterio de información de Akaike (AIC) para determinar cuál es el mejor modelo. Cabe mencionar que del proceso anterior obtendremos el que sería el mejor modelo dentro de las posibilidades de las que disponemos (con las variables de las que disponemos) lo que no quiere decir que dicho modelo sea correcto.

```
# Modelo sin variables
rlm_0 <- lm(quality ~ 1,data = data.frame(df) )
summary(rlm_0)

# Modelo con todas las variables
rlm_1 <- lm(quality ~ .,data = data.frame(df) )
summary(rlm_1)

# Regresión stepwise - Selección mejores predictores
stw <- step(rlm_0, scope = list(lower=rlm_0, upper=rlm_1), direction = "both")
summary(stw)
```

De la selección anterior observamos como el mejor modelo de entre todos los que podemos crear con las variables de las que disponemos es el siguiente.

```
> summary(stw)

Call:
lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
    chlorides + total.sulfur.dioxide + pH + free.sulfur.dioxide,
    data = data.frame(df))
```

Procedemos por tanto a la ejecución de la regresión lineal múltiple utilizando los predictores anteriores con el fin de hallar los coeficientes del modelo.

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n + \varepsilon$$

Como vemos en la captura de pantalla siguiente todos los coeficientes hallados son significativos (p-valor < 0,05). La variable que más peso tiene en el modelo y por tanto que más influye en la calidad del vino es el porcentaje de alcohol (variable "alcohol") seguida de la cantidad de sulfitos ("sulphates") y del nivel de ácido cítrico.

```
> # Mejor modelo
> modelo <- lm(quality ~ alcohol + volatile.acidity + sulphates + chlorides +
+              total.sulfur.dioxide + pH + free.sulfur.dioxide, data = data.frame(df))
> summary(modelo)

Call:
lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
    chlorides + total.sulfur.dioxide + pH + free.sulfur.dioxide,
    data = data.frame(df))

Residuals:
    Min      1Q  Median      3Q     Max
-3.3166 -0.4500 -0.0584  0.5769  2.4652

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2.014e-15  2.198e-02   0.000   1.0000
alcohol               3.921e-01  2.474e-02  15.849  < 2e-16 ***
volatile.acidity     -1.921e-01  2.444e-02  -7.861 7.85e-15 ***
sulphates             1.930e-01  2.554e-02   7.557 7.65e-14 ***
chlorides            -1.212e-01  2.568e-02  -4.721 2.60e-06 ***
total.sulfur.dioxide -1.524e-01  3.100e-02  -4.917 9.90e-07 ***
pH                   -7.023e-02  2.468e-02  -2.846   0.0045 **
free.sulfur.dioxide   6.535e-02  3.023e-02   2.162   0.0308 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8035 on 1329 degrees of freedom
Multiple R-squared:  0.3577,    Adjusted R-squared:  0.3543
F-statistic: 105.7 on 7 and 1329 DF,  p-value: < 2.2e-16
```

Para comprobar los resultados del modelo utilizaremos el coeficiente de determinación R2 que nos indica que proporción de la variación total queda explicada por el modelo creado anteriormente.

$$R^2 = \frac{Varianza\ explicada\ por\ el\ modelo}{Total\ de\ varianza\ en\ los\ datos}$$

Como regla general se suele considerar que cuanto más se acerca a 1 el coeficiente anterior, mejor es el modelo ya que es capaz de explicar una mayor variabilidad, por el contrario cuanto más cerca de 0 está, significa que el modelo poco se aproxima a los datos.

En nuestro caso observamos como el coeficiente de determinación es de 0,3577 lo que se interpreta como que el modelo es capaz de explicar casi un 36% de la variabilidad de los datos siendo la parte residual o no explicada de un 64% ($1 - R^2$). Si cogemos como referencia los valores anteriores vemos como está más próximo a 0 que a 1 lo que podríamos clasificar como una baja capacidad explicativa.

5.3. Comprobación de los supuestos de aplicación.

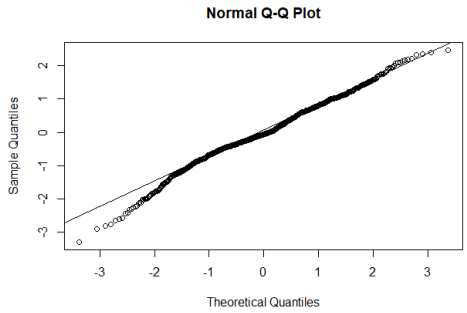
Antes de dar por válido del modelo anterior debemos comprobar si se cumplen o no los supuestos de aplicación de la regresión lineal múltiple. Al inicio de este apartado hemos analizado la

multicolinealidad con el fin de descartar aquellas variables que estuvieran fuertemente correlacionadas entre ellas y que nos pudieran llevar a conclusiones erróneas.

A continuación validaremos los supuestos de normalidad y de homocedasticidad.

Para el primer caso, procederemos a graficar los residuos con el fin de validar la normalidad del modelo.

En la captura de pantalla inferior vemos como los datos no son del todo normales ya que se desvían ligeramente de la diagonal. Para comprobar estadísticamente dicha sospecha, podemos utilizar el test de Shapiro-Wilk que Comprueba la hipótesis nula de que los datos son normales. Si rechazamos la hipótesis nula (p-valor < 0.05) podemos por tanto asumir que nuestro modelo no es normal.




```
> shapiro.test(df$quality)

        Shapiro-Wilk normality test

data:  df$quality
W = 0.85885, p-value < 2.2e-16
```

Como observamos en el resultado del test de Shapiro-Wilk rechazamos la hipótesis nula y por lo tanto confirmamos la no normalidad del modelo algo que podría llevar a la incorrecta estimación del modelo y a invalidar las conclusiones.

En segundo lugar procederemos a la validación del supuesto de homocedasticidad u homogeneidad de las varianzas. Para ello utilizaremos el test estadístico Breusch-Pagan que toma como hipótesis nula la homocedasticidad de los datos y como hipótesis alternativa la heterocedasticidad, es decir, si el p-valor es inferior a 0,05, rechazaríamos la hipótesis nula.

```
> bptest(modelo)

        studentized Breusch-Pagan test

data:  modelo
BP = 45.871, df = 7, p-value = 9.261e-08
```

Como vemos en la captura de pantalla anterior el p-valor es inferior a 0,05 y por tanto podríamos decir que no se cumple el supuesto de homocedasticidad de los datos necesario para el análisis de regresión lineal múltiple.

6. Resolución del problema

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

A lo largo de este análisis hemos intentado crear un modelo que nos permitiera, dadas una serie de características, estimar la calidad del vino tinto. Con ese fin hemos empezado analizando los valores no informados del conjunto de datos para seguir con un análisis de los principales estadísticos de cada una de las variables del set de datos con el fin de hallar posibles valores atípicos.

Una vez hemos tenido el conjunto de datos preparado, hemos procedido a realizar un análisis bivariante de los diferentes atributos con el fin de determinar si las relaciones existentes entre las variables independientes y la variable dependiente "quality" eran apropiadas para el tipo de análisis que queríamos llevar a cabo así como las relaciones entre las propias variables independientes. En este paso hemos hallado los primeros indicios de que el conjunto de datos del que disponemos podía no ser el más adecuado para el fin buscado debido principalmente a la falta de correlación entre las variables independientes y la variable dependiente.

A continuación hemos estimado cual era el modelo de regresión lineal múltiple que mejor podía explicar la variable dependiente de entre todos los posibles. Aquí hemos hallado un modelo compuesto por las siguientes variables.

```
modelo <- lm(quality ~ alcohol + volatile.acidity + sulphates + chlorides +
               total.sulfur.dioxide + pH + free.sulfur.dioxide, data = data.frame(df))
```

Como hemos visto en el análisis del resultado del modelo, el coeficiente de determinación era de alrededor de un 0,36 lo que se interpreta como que el modelo tiene una capacidad explicativa baja.

El último paso ha sido validar los supuestos de aplicación del modelo de regresión lineal múltiple donde hemos hallado que no se cumple ninguno de los dos supuestos planteados: normalidad y homocedasticidad lo que podría explicar la falta de capacidad explicativa del modelo.

Por todo lo comentado anteriormente determinamos que el modelo creado no sirve para estimar la calidad del vino a partir de las variables con las que hemos trabajado y, a pesar de que puede deberse a diversos factores, las principales causas que barajamos son las siguientes:

- Las variables incluidas en el análisis y/o de las que disponemos no son las más adecuadas para el fin perseguido ya que no son las que determinan la calidad del vino (variable dependiente).
- A pesar de que desconocemos la procedencia de la calificación de cada vino y/o el método de cálculo, creemos que viene de una valoración subjetiva y debido a ello las calificaciones de cada observación están poco relacionadas con las variables. Tal vez con una muestra mayor se podrían descubrir los patrones que relacionan dichas características con la puntuación dada a cada vino.

7. 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.

El código utilizado para el análisis anterior se puede encontrar en el repositorio de github siguiente, dentro de la carpeta "Código":

- https://github.com/rcollmenendezUOC/UOC_PRA2_Limpieza_y_Analisis_de_Datos.git