Práctica 2: Limpieza y validación de los datos

Carlos Garabatos Fernández - Jorge Ronchel Diaz-Leante

18 de junio, 2019

Índice general

1 Descripción del dataset. 1
1.1 Importancia y objetivos de los análisis 2

2 Integración y selección de los datos de interés a analizar. 2
2.1 Tipo de variable estadística de cada variable 4
2.2 Selección de datos de interes 5
2.3 Normalizacion de variables 5

3 Limpieza de los datos. 10
3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? . 10
3.2 Identificación y tratamiento de valores extremos. 12

4 Análisis de los datos. 20
4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). 20
4.2 Comprobación de la normalidad y homogeneidad de la varianza. 20

5 Aplicación de pruebas estadísticas para comparar los grupos de datos. 23

6 Representación de los resultados a partir de tablas y gráficas. 40

7 Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? 43

8 Exportación del código en R y de los datos producidos. 44

---

1 Descripción del dataset.

Descripción de la Práctica a realizar El objetivo de esta actividad será el tratamiento de un dataset, ??? Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)

Introducción:

Los dos conjuntos de datos están relacionados con las variantes rojas y blancas del vino portugués “Vinho Verde”. Para más detalles, consulte la referencia [Cortez et al., 2009]. Debido a problemas de privacidad y logística, solo están disponibles las variables fisicoquímicas (entradas) y sensoriales (la salida) (por ejemplo, no hay datos sobre tipos de uva, marca de vino, precio de venta del vino, etc.).

Attribute Information:

For more information, read [Cortez et al., 2009]. Input variables (based on physicochemical tests):

1 - fixed.acidity: La mayoría de los Ácidos presentes en el vino que no se evaporan fácilmente.

2 - volatile.acidity: La cantidad de Ácido acético en el vino, que en niveles demasiado altos puede llevar a un sabor desagradable a vinagre.

3 - citric.acid: Encontrado en pequeñas cantidades, el Ácido cítrico puede agregar 'frescura' y sabor a los vinos

4 - residual.sugar: La cantidad de azúcar restante después de que se detiene la fermentación

5 - chlorides: La cantidad de sal en el vino

6 - free.sulfur.dioxide: El SO2 de forma libre se presenta en equilibrio entre el SO2 molecular (como un gas disuelto) y el ión bisulfito; Previene el crecimiento microbiano y la oxidaciÃ³n del vino.

7 - total.sulfur.dioxide: Cantidad de formas libres y ligadas de S02. En bajas concentraciones, el SO2 es mayormente indetectable en el vino, pero a concentraciones de SO2 libres de más de 50 ppm, el SO2 se hace evidente en la nariz y el sabor del vino.

8 - density: Densidad

9 - pH: Describe como de Ácido o básico es un vino en una escala de 0 (muy Ácido) a 14 (muy básico). La mayoría de los vinos están entre 3-4 en la escala de pH.

10 - sulphates: Un aditivo del vino que actúa como antimicrobiano y antioxidante.

11 - alcohol: El porcentaje de alcohol del vino

12 - quality: (Variable de salida) - Puntuación de calidad del vino entre 0 y 10

Source:

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009

1.1 Importancia y objetivos de los análisis

A partir de este conjunto de datos se plantea la problemática de determinar qué variables influyen más sobre la calidad del vino.

Además, se podrá proceder a crear modelos de regresión que permitan predecir la calidad de un vino Portugues en función de sus características y contrastes de hipótesis que ayuden a identificar propiedades interesantes en las muestras que puedan ser inferidas con respecto a la población.

Estos análisis adquieren una gran relevancia en casi cualquier sector relacionado con la enología.Un ejemplo de ello se puede observar en la gestión de una bodega. En este caso, el enologo podra valorar cual es la espectativa de precio en funcion de su baremo de calidad, y tambien el mercado donde dichas caracterisiticas resultaran en una mejor venta.

Se aplicarán 3 pruebas estadisticas: Correlación, regresión y contraste de hipotesis.

---

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

---

1. Carga del archivo de datos en R y una breve descripción del archivo donde se indica el número de registros, el número de variables y los nombres de las variables.

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
```

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

El archivo se denomina *winequality-red.csv*, contiene 1599 registros y 12 variables. Estas variables son: fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, quality

Ejecutamos el comando summary en el que se aprecian los percentiles 25 y 75 de cada variable, su media, mediana y valores mínimo y máximo

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

2.1 Tipo de variable estadística de cada variable

El fichero de datos contiene 1599 registros y 12 variables.

Las variables son fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, quality.

```
# Visualizamos la tipologia de cada variable
str(data)
```

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

```
# read data
res <- sapply(data,class)
kable(data.frame(variables=names(res),clase=as.vector(res)))
```

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

```
table(data$quality)
```

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

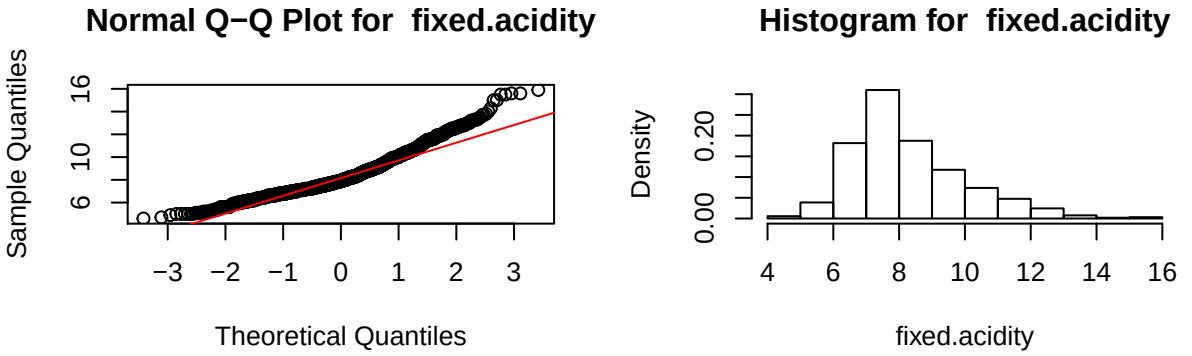
De la tabla anterior se desprende que existe un desequilibrio de clase. Hay 1599 muestras pero solo 10 son de la clase 3 y solo 18 son de la clase 8.

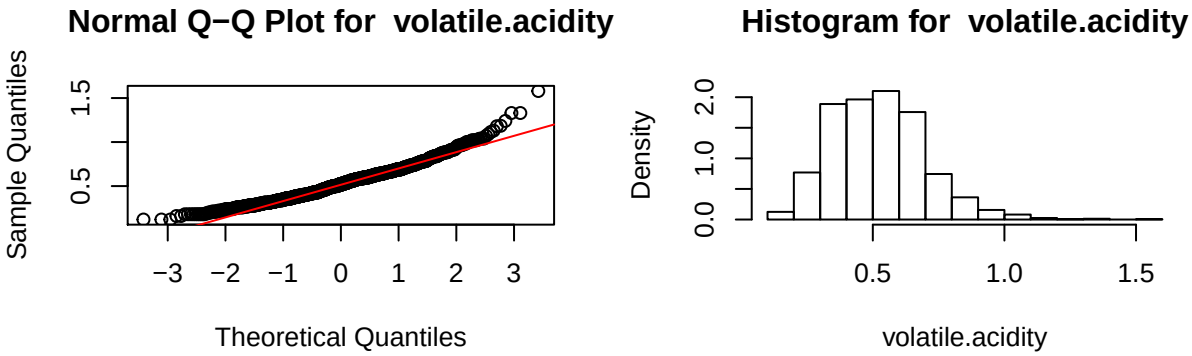
2.2 Selección de datos de interes

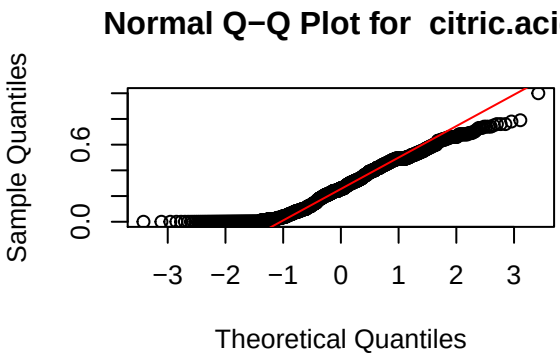
Podríamos convertir la variable quality a tipo factor pero de momento nos interesa consevarla para tratarla como número. Si es necesario mas adelante crearemos una variable factor equivalente. A priori todos los atributos presentes en el conjunto de datos se corresponden con características influyentes en la variable de salida del conjunto de datos, que es la calidad del vino.

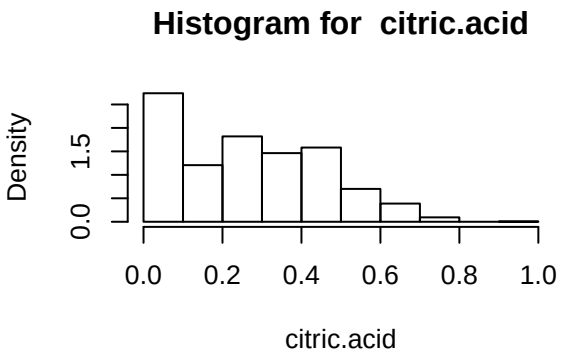
2.3 Normalizacion de variables

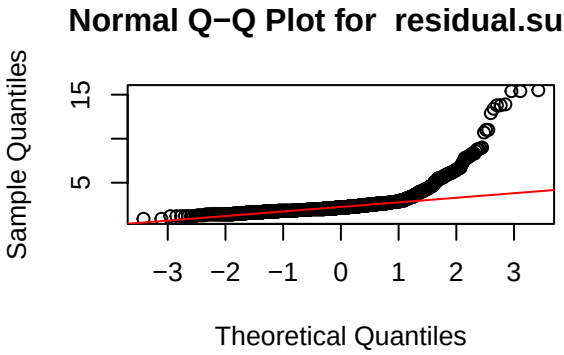
```
#Para revisar si las variables pueden ser candidatas a la normalización miramos las graficas de quantil

par(mfrow=c(2,2))
for(i in 1:ncol(data)) {
  if (is.numeric(data[,i])){
    qqnorm(data[,i],main = paste("Normal Q-Q Plot for ",colnames(data)[i]))
    qqline(data[,i],col="red")
    hist(data[,i],
    main=paste("Histogram for ", colnames(data)[i]),
    xlab=colnames(data)[i], freq = FALSE)
  }
}
```

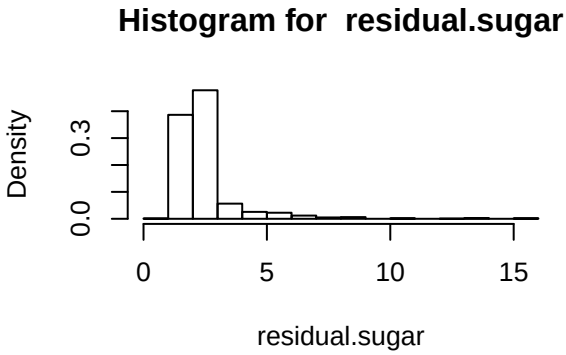


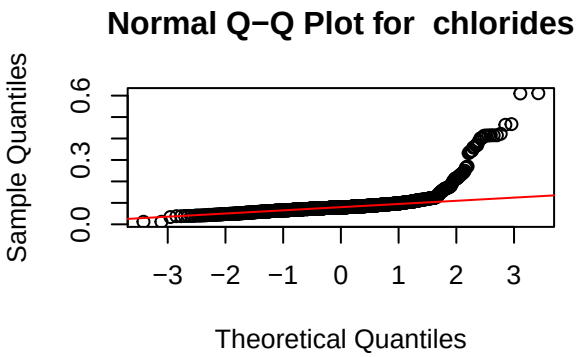


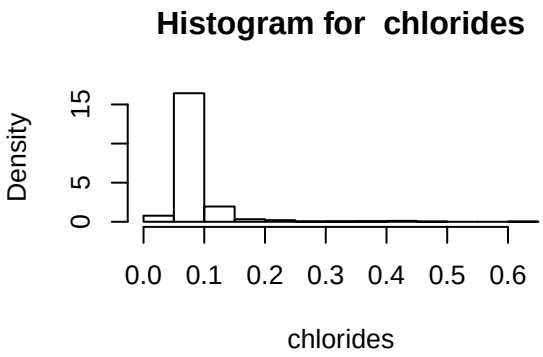

Normal Q−Q Plot for citric.acid
Sample Quantiles
0.0
0.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

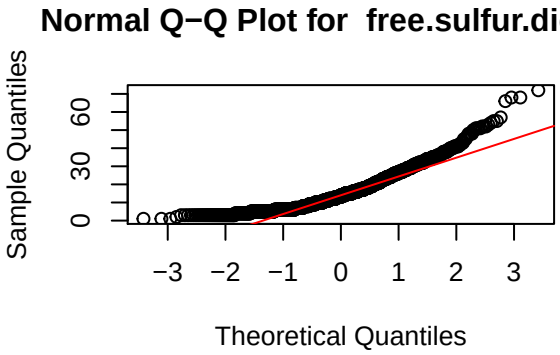

Histogram for citric.acid
Density
0.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
citric.acid

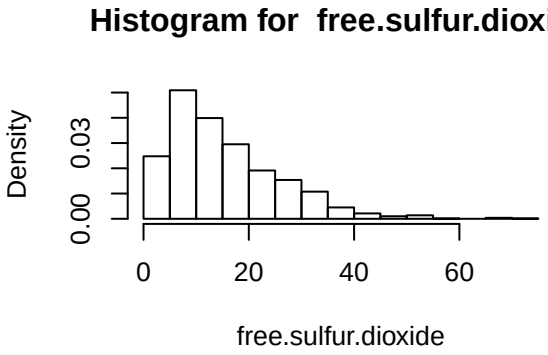

Normal Q−Q Plot for residual.sugar
Sample Quantiles
5
15
−3
−2
−1
0
1
2
3
Theoretical Quantiles

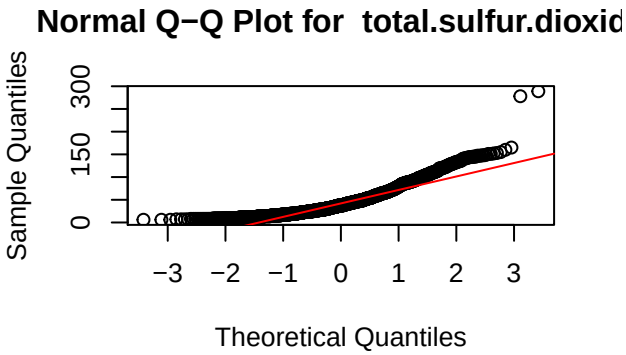

Histogram for residual.sugar
Density
0.0
0.3
0
5
10
15
residual.sugar

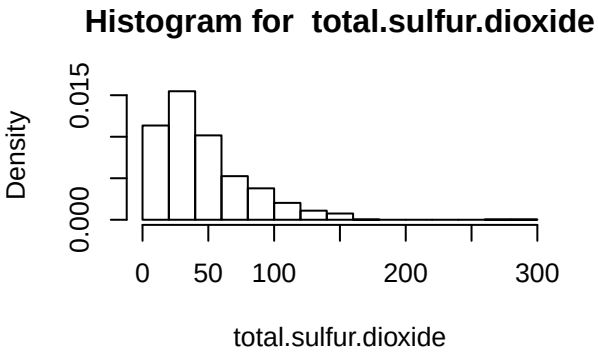

Normal Q−Q Plot for chlorides
Sample Quantiles
0.0
0.3
0.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

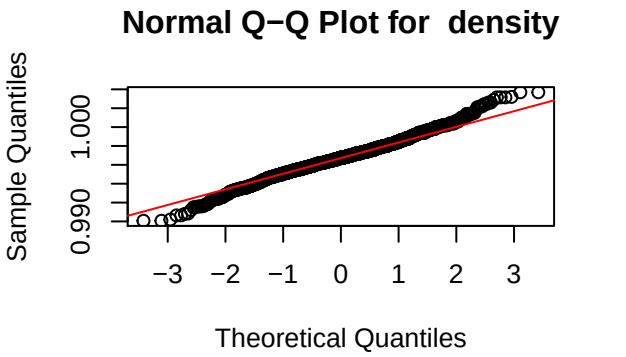

Histogram for chlorides
Density
0
5
15
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides

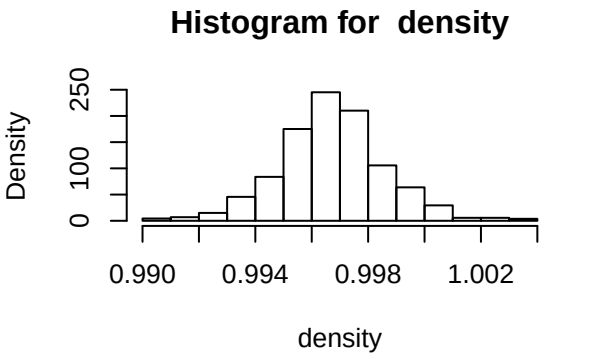

Normal Q−Q Plot for free.sulfur.dioxid
Sample Quantiles
0
30
60
−3
−2
−1
0
1
2
3
Theoretical Quantiles

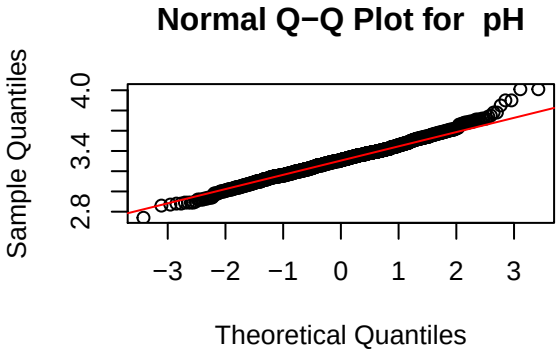

Histogram for free.sulfur.dioxide
Density
0.00
0.03
0
20
40
60
free.sulfur.dioxide

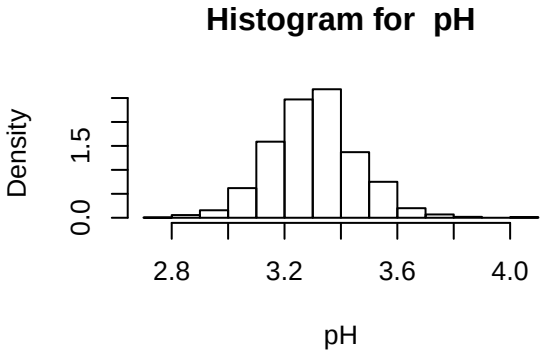

Normal Q−Q Plot for total.sulfur.dioxid
Sample Quantiles
0
150
300
−3
−2
−1
0
1
2
3
Theoretical Quantiles

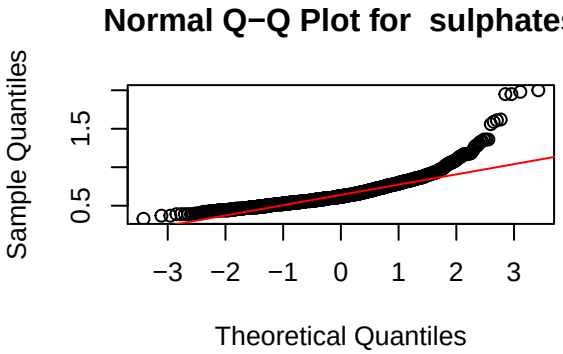

Histogram for total.sulfur.dioxide
Density
0.000
0.015
0
50
100
200
300
total.sulfur.dioxide

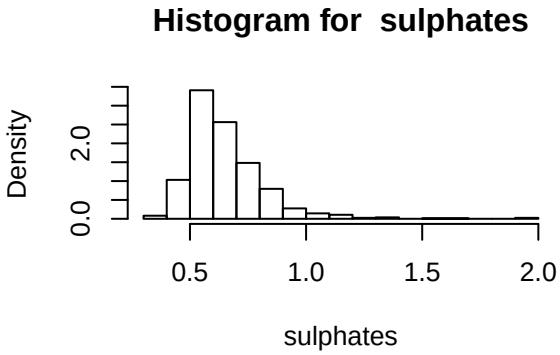

Normal Q−Q Plot for density
Sample Quantiles
0.990
1.000
−3
−2
−1
0
1
2
3
Theoretical Quantiles

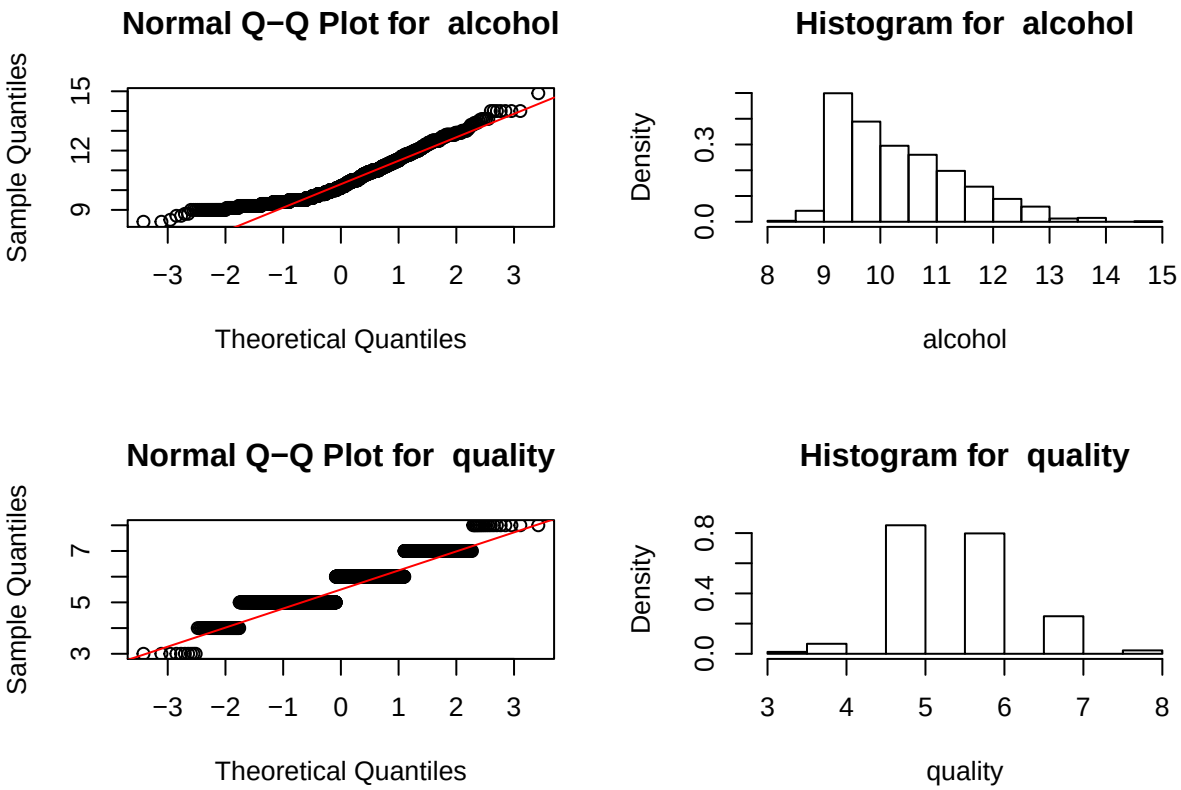

Histogram for density
Density
0
100
250
0.990
0.994
0.998
1.002
density


Normal Q−Q Plot for pH
Sample Quantiles
2.8
3.4
4.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Histogram for pH
Density
0.0
1.5
2.8
3.2
3.6
4.0
pH


Normal Q−Q Plot for sulphates
Sample Quantiles
0.5
1.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Histogram for sulphates
Density
0.0
2.0
0.5
1.0
1.5
2.0
sulphates



Los resultados del quantile-quantile plot nos indica que las variables pueden ser candidatas a la normalización si es necesario.

3 Limpieza de los datos.

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

Comprobamos la cantidad de valores nulos que existen por cada atributo.

```
#Contar número de nulos por columna
sapply(data, function(x) sum(is.na(x)))#
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

```
#Eliminar todas las filas que contengan algun valor nulo (forma simple)
#datos <- na.omit(datos)
#data[!is.na(data)]
#complete.cases(data)
```

Comprobamos la cantidad de valores menores que 0 que existen por cada atributo.

```
#corrplot(data)
# Valores menores que 0
colSums(data<0)
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

```
# Valores menores que 0
colSums(data=="")
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

No tenemos ningún valor vacío ni nulo. Por tanto no es necesaria ninguna modificación. En el caso de que hubiera habido alguno, deberímos optar por alguna de las tÃ©cnicas de tratamiento de elementos vacíos o ceros. Bien mediante la asignación de un valor cercano a la media, bien mediante la eliminación del registro completo, etc.

Comprobamos la cantidad de valores que 0 que existen por cada atributo.

```
#Valores vacios
colSums(data==0)
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                  132
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

Aparecen 132 valores a 0 del acido citrico. Pero, el acido citrico es poco abundante en la uva, de 150 a 300 mg/ litro de mosto. Después es fermentado por las bacterias lácticas y desaparece.Por lo que se refiere a vinos con una fermentación completa, y dichos valores no deben cambiarse.

Comprobamos si se puede discretizar alguna variable que tenga un rango de valores pequeño.

```
apply(data,2, function(x) length(unique(x)))
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                   96                  143                   80
##       residual.sugar            chlorides  free.sulfur.dioxide
##                   91                  153                   60
## total.sulfur.dioxide              density                   pH
```

```
##                  144                  436                   89
##            sulphates              alcohol              quality
##                   96                   65                    6
```

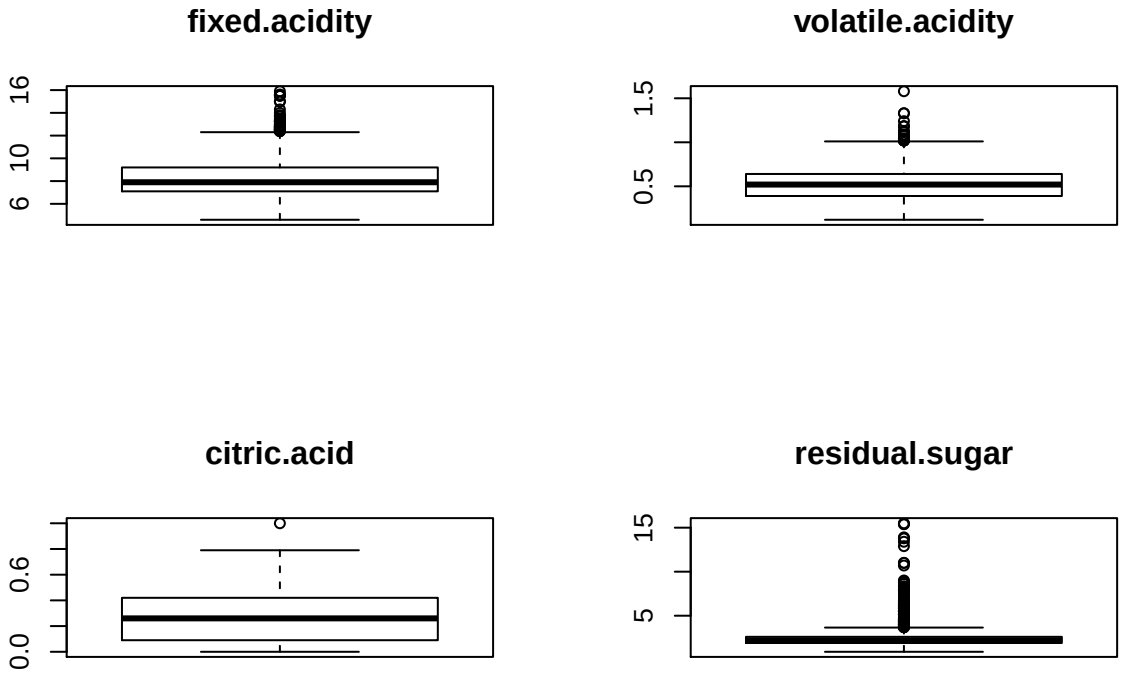
No hay ningún valor que ofrezca un rango pequeño de valores a excepción de la calidad, sin embargo no creemos óptimo discretizar dicha variable.

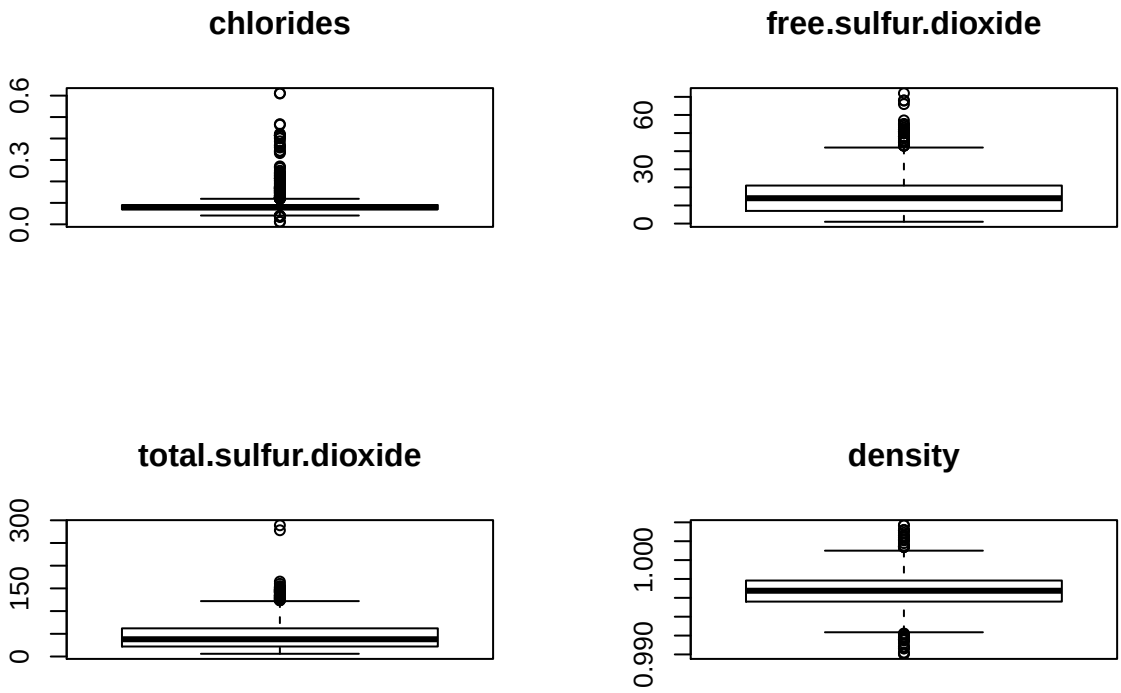
3.2 Identificación y tratamiento de valores extremos.

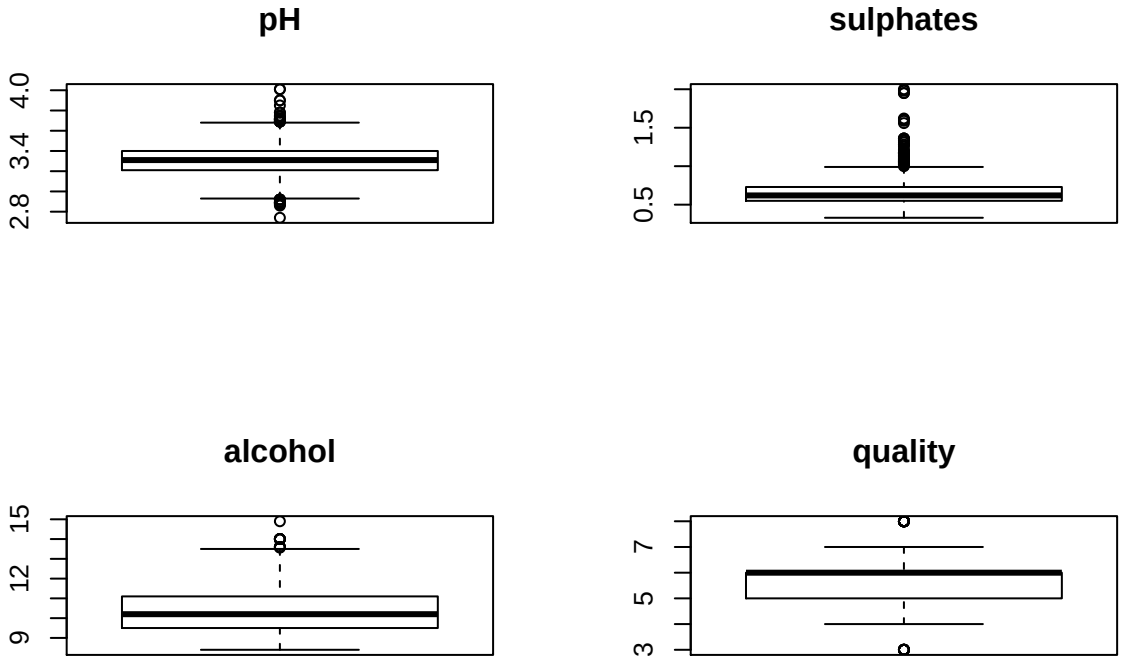
```
#Veamos una representación mediante boxplot de las variables numéricas:

describe(data)
```

```
##                       vars    n  mean    sd median trimmed   mad  min
## fixed.acidity            1 1599  8.32  1.74   7.90    8.15  1.48 4.60
## volatile.acidity         2 1599  0.53  0.18   0.52    0.52  0.18 0.12
## citric.acid              3 1599  0.27  0.19   0.26    0.26  0.25 0.00
## residual.sugar           4 1599  2.54  1.41   2.20    2.26  0.44 0.90
## chlorides                5 1599  0.09  0.05   0.08    0.08  0.01 0.01
## free.sulfur.dioxide      6 1599 15.87 10.46  14.00   14.58 10.38 1.00
## total.sulfur.dioxide     7 1599 46.47 32.90  38.00   41.84 26.69 6.00
## density                  8 1599  1.00  0.00   1.00    1.00  0.00 0.99
## pH                       9 1599  3.31  0.15   3.31    3.31  0.15 2.74
## sulphates               10 1599  0.66  0.17   0.62    0.64  0.12 0.33
## alcohol                 11 1599 10.42  1.07  10.20   10.31  1.04 8.40
## quality                 12 1599  5.64  0.81   6.00    5.59  1.48 3.00
##                          max  range skew kurtosis   se
## fixed.acidity          15.90  11.30 0.98     1.12 0.04
## volatile.acidity        1.58   1.46 0.67     1.21 0.00
## citric.acid             1.00   1.00 0.32    -0.79 0.00
## residual.sugar         15.50  14.60 4.53    28.49 0.04
## chlorides               0.61   0.60 5.67    41.53 0.00
## free.sulfur.dioxide    72.00  71.00 1.25     2.01 0.26
## total.sulfur.dioxide 289.00 283.00 1.51     3.79 0.82
## density                 1.00   0.01 0.07     0.92 0.00
## pH                      4.01   1.27 0.19     0.80 0.00
## sulphates               2.00   1.67 2.42    11.66 0.00
## alcohol                14.90   6.50 0.86     0.19 0.03
## quality                 8.00   5.00 0.22     0.29 0.02
```

```
res <- sapply(data,class)
resCont <- which(res=="numeric")

par(mfrow=c(2,2))
for(i in 1:ncol(data)) {
  if (is.numeric(data[,i])){
    boxplot(data[,i], main = colnames(data)[i], width = 100)
  }
}
```

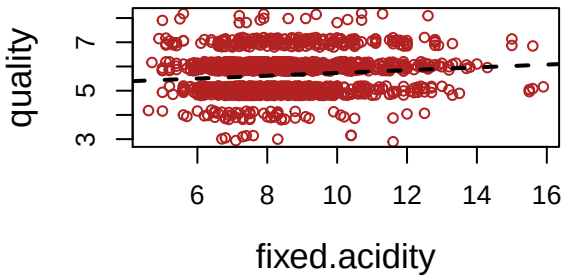

fixed.acidity
6
10
16
volatile.acidity
0.5
1.5
citric.acid
0.0
0.6
residual.sugar
5
15

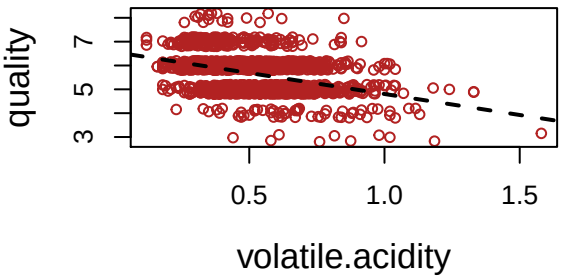

chlorides
0.0
0.3
0.6
free.sulfur.dioxide
0
30
60
total.sulfur.dioxide
0
150
300
density
0.990
1.000

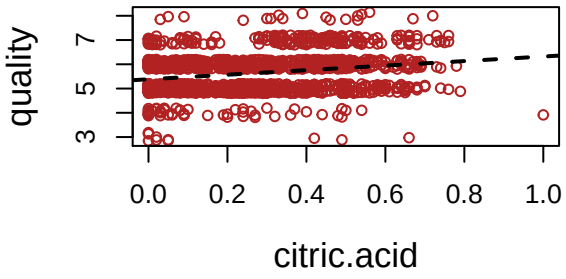


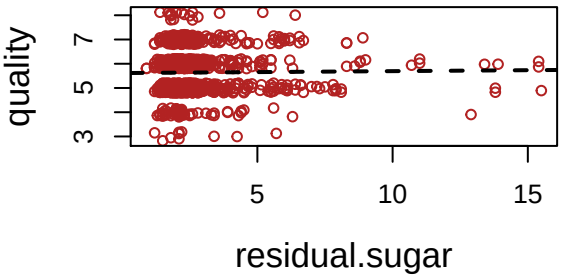

```
par(mfrow=c(1,1))
```

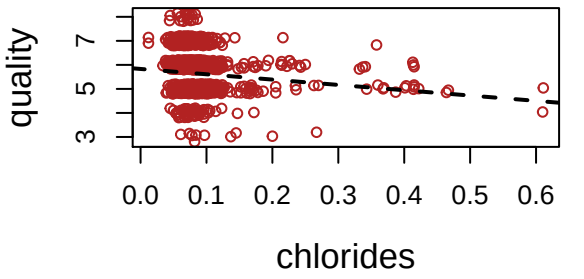
```
par(mfrow = c(2,2))
for (i in c(1:11)) {
    plot(data[, i], jitter(data[, "quality"]), xlab = names(data)[i],
         ylab = "quality", col = "firebrick", cex = 0.8, cex.lab = 1.3)
    abline(lm(data[, "quality"] ~ data[ ,i]), lty = 2, lwd = 2)
}
```

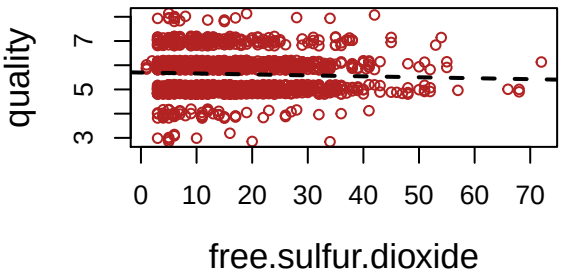

quality
3
5
7
6
8
10
12
14
16
fixed.acidity

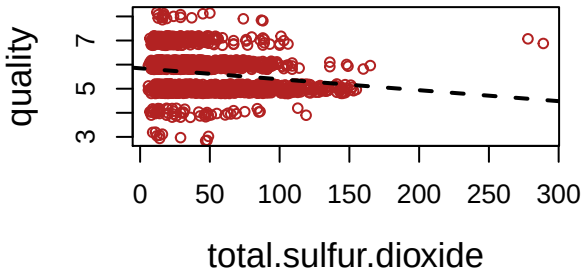

quality
3
5
7
0.5
1.0
1.5
volatile.acidity

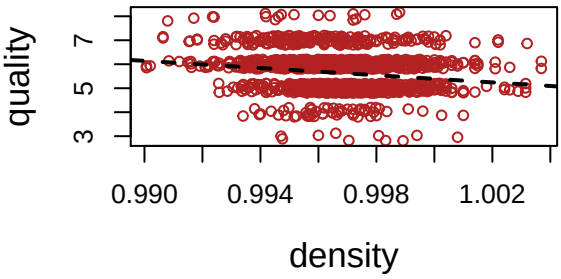

quality
3
5
7
0.0
0.2
0.4
0.6
0.8
1.0
citric.acid


quality
3
5
7
5
10
15
residual.sugar

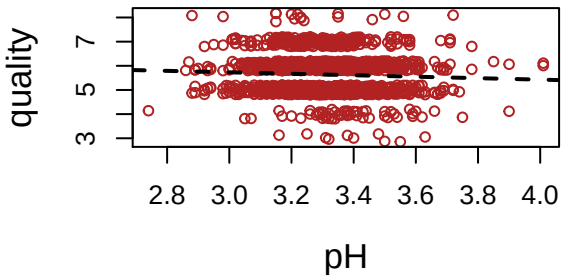


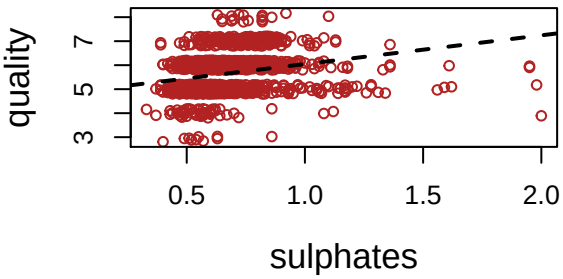


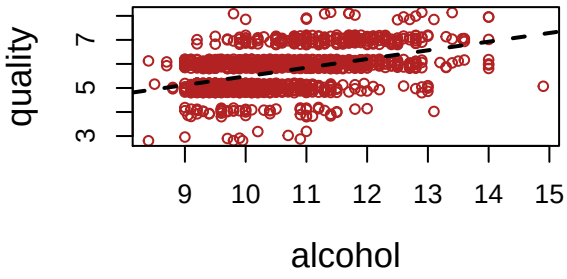





```
par(mfrow = c(1, 1))
```







Destaca la presencia de valores atípicos para la mayoría de las variables predictoras. El conjunto de datos de vino se limpió antes de su publicación, por lo que no se suponen que sean errores.

```
boxplot.stats(data$volatile.acidity)$out
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020
## [12] 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

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

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

```
boxplot.stats(data$residual.sugar)$out
```

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10
##  [12]  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00
##  [23]  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00
##  [34]  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70
##  [45]  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60
##  [56]  6.10  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30
##  [67]  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40  4.25
##  [78]  6.00  3.90  4.20  4.00  4.00  4.00  6.60  6.00  6.00  3.80  9.00
##  [89]  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00
## [100]  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50
## [111]  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90  4.30
## [122] 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
## [144] 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10
```

```
## [155]  7.80
```

```
boxplot.stats(data$chlorides)$out
```

```
##   [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178
##  [12] 0.146 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343
##  [23] 0.186 0.213 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122
##  [34] 0.122 0.174 0.121 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171
##  [45] 0.226 0.226 0.250 0.148 0.122 0.124 0.124 0.143 0.222 0.039 0.157
##  [56] 0.422 0.034 0.387 0.415 0.157 0.157 0.243 0.241 0.190 0.132 0.126
##  [67] 0.038 0.165 0.145 0.147 0.012 0.012 0.039 0.194 0.132 0.161 0.120
##  [78] 0.120 0.123 0.123 0.414 0.216 0.171 0.178 0.369 0.166 0.166 0.136
##  [89] 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414 0.166 0.168 0.415
## [100] 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205 0.039 0.235
## [111] 0.230 0.038
```

```
boxplot.stats(data$free.sulfur.dioxide)$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
## [24] 51 52 55 55 48 48 66
```

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

```
##  [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127
## [18] 126 145 144 135 165 124 124 134 124 129 151 133 142 149 147 145 148
## [35] 155 151 152 125 127 139 143 144 130 278 289 135 160 141 141 133 147
## [52] 147 131 131 131
```

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

```
##  [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220
##  [9] 1.00220 1.00140 1.00140 1.00140 1.00140 1.00320 1.00260 1.00140
## [17] 1.00315 1.00315 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260
## [25] 0.99210 0.99154 0.99064 0.99064 1.00289 0.99162 0.99007 0.99007
## [33] 0.99020 0.99220 0.99150 0.99157 0.99080 0.99084 0.99191 1.00369
## [41] 1.00369 1.00242 0.99182 1.00242 0.99182
```

```
boxplot.stats(data$pH)$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
## [15] 2.89 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
## [29] 4.01 2.90 4.01 3.71 2.88 3.72 3.72
```

```
boxplot.stats(data$sulphates)$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
## [15] 1.59 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
## [29] 1.07 1.06 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17
## [43] 1.62 1.06 1.18 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03
## [57] 1.17 1.10 1.01
```

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

```
##  [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000
##  [8] 13.60000 13.60000 14.00000 14.00000 13.56667 13.60000
```

```
boxplot.stats(data$quality)$out
```

```
##  [1] 8 8 8 8 8 3 8 8 8 3 8 3 8 3 3 8 8 8 8 8 3 3 8 8 3 3 3 8
```

Observamos valores extremos en todas las variables, pero son principalmente evidentes en azúcar residual, cloruros, densidad y sulfatos. Pero no los eliminaremos ya que los datos ya han sido evaluados y además estos valores son posibles por estar dentro de la escala de clasificación del vino tinto.

```
#max.Alc <- which(data$alcohol == max(data$alcohol))
#data <- data[-max.Alc, ]
```

4 Análisis de los datos.

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

```
# El acido citrico es de interes porque si aparece significa que la fermentación del vino no se ha comp
data.sinC=data[data$citric.acid==0,] # Sin acido citrico
data.conC=data[data$citric.acid>0,]# Con acido citrico

#El nivel de alcohol parece ser superior en los vinos de mejor calidad, por esa razon se divide en grup
tapply ( data$alcohol , data$ quality , mean )
```

```
##         3         4         5         6         7         8
##  9.955000 10.265094  9.899706 10.629519 11.465913 12.094444
```

```
data.muchoA=data[data$alcohol>=median(alcohol),] # Con mucho alcohol
data.pocoA=data[data$alcohol<median(alcohol),]# Con poco alcohol
#mean(data.pocoA$alcohol)
#mean(data.muchoA$alcohol)
```

Pueden apreciarse niveles de alcohol ascendentes respecto al nivel de calidad.De todas formas, más adelante se realizará un análisis de varianzas respecto a estas dos variables para probar su influencia en la calidad del vino.

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Para determinar si los datos no siguen una distribución normal, hay que comparar el p-valor con el nivel de significancia. Por lo general, un nivel de significancia (denotado como o alfa) de 0.05 funciona adecuadamente. Un nivel de significancia de 0.05 indica un riesgo de 5% de concluir que los datos no siguen una distribución normal, cuando los datos sí siguen una distribución normal. Para revisar si las variables siguen una distribución normal se aplica el test de Shapiro Wilk en cada variables numérica.

```
shapiro.test(data[,1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 1]
## W = 0.94203, p-value < 2.2e-16
```

```
shapiro.test(data[,2])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 2]
## W = 0.97434, p-value = 2.693e-16
```

```r
shapiro.test(data[,3])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 3]
## W = 0.95529, p-value < 2.2e-16
```

```r
shapiro.test(data[,4])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 4]
## W = 0.56608, p-value < 2.2e-16
```

```r
shapiro.test(data[,5])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 5]
## W = 0.48425, p-value < 2.2e-16
```

```r
shapiro.test(data[,6])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 6]
## W = 0.90184, p-value < 2.2e-16
```

```r
shapiro.test(data[,7])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 7]
## W = 0.87322, p-value < 2.2e-16
```

```r
shapiro.test(data[,8])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 8]
## W = 0.99087, p-value = 1.936e-08
```

```r
shapiro.test(data[,9])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 9]
## W = 0.99349, p-value = 1.712e-06
```

```r
shapiro.test(data[,10])
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  data[, 10]
## W = 0.83304, p-value < 2.2e-16
```

```
shapiro.test(data[,11])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 11]
## W = 0.92884, p-value < 2.2e-16
```

```
alpha = 0.05
col.names = colnames(data)
for (i in 1:ncol(data)) {
  if (i == 1) cat("\nVariables que no siguen una distribución normal:\n")
  if (is.integer(data[,i]) | is.numeric(data[,i])) {
    p_val = ad.test(data[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(data) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
##
## Variables que no siguen una distribución normal:
## fixed.acidity, volatile.acidity, citric.acid,
## residual.sugar, chlorides, free.sulfur.dioxide,
## total.sulfur.dioxide, density, pH,
## sulphates, alcoholquality
```

4.2.1 Homogeneidad de varianzas

Estudiamos la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. En este caso, estudiaremos esta homogeneidad en cuanto a los grupos conformados por vinos que presentan un nivel de alcohol mas alto que los otros.En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```
# Creamos la variable
data$NivA <- ifelse (data$alcohol>=median(alcohol), 'alto', 'bajo')
table(data$NivA)
```

```
##
## alto bajo
##  803  796
```

```
data$NivA=as.factor(data$NivA)

fligner.test(quality ~ NivA, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  quality by NivA
```

```
## Fligner-Killeen:med chi-squared = 24.049, df = 1, p-value =
## 9.391e-07
```

Puesto que obtenemos un p-valor superior a 0,05, aceptamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

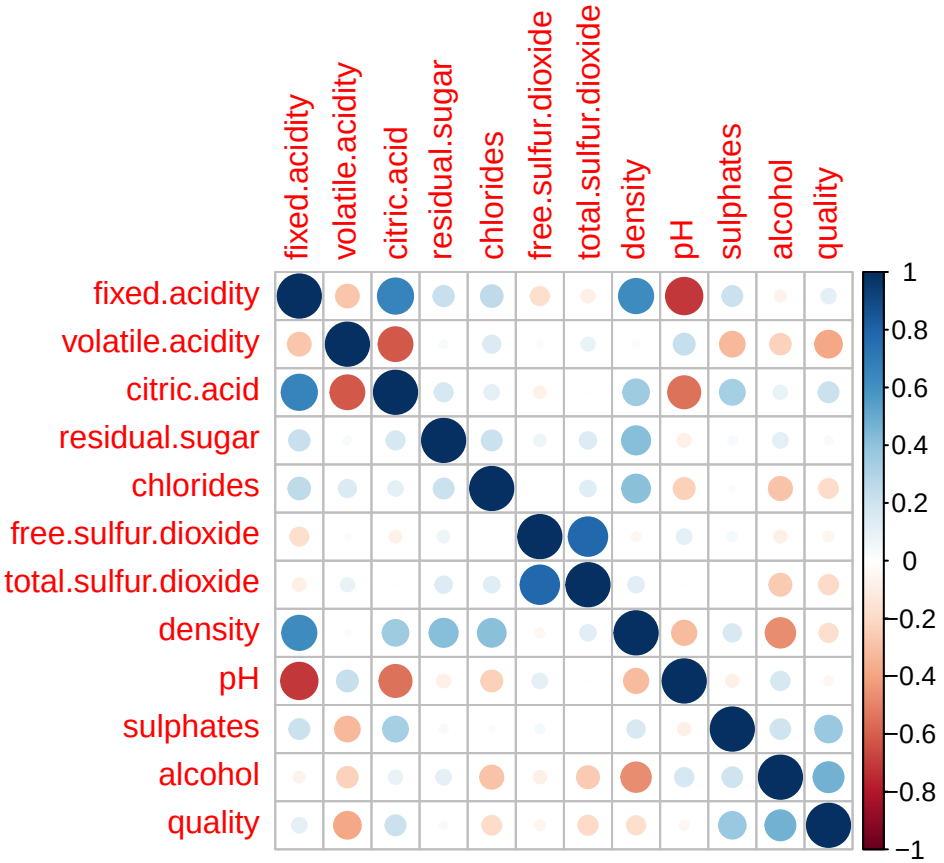
5 Aplicación de pruebas estadísticas para comparar los grupos de datos.

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.

5.0.1 ¿Variables cuantitativas que mas influyen en la calidad?

En primer lugar, procedemos a realizar un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre la calidad. Para ello, se utilizará el coeficiente de correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.

```
#corrplot(cor(data[, 1 : 12 ]), type = "lower", method = "number")
#cor ( x = data [, 1 : 12 ], y = data$quality )
data.cor <- cor(data [, 1 : 12 ], method = c("spearman"))
corrplot(data.cor)
```



Identificamos que las variables más correlacionadas con la calidad son:

Alcohol (+++)

Acidez volátil (—)

Ácido cítrico (++)

Acidez fija (+)

Sulfatos (+)

Dióxido de azufre total (-)

Densidad (-)

Cloruros (-)

Además, vamos a ver como influyen algunas de las variables entre sí, viendo la correlación existente entre ellas mismas, a parte de entre ellas y la calidad.

En cuanto a la acidez fija, vemos que tiene correlación positiva con el ácido cítrico e incluso podría haber algo de redundancia, ya que el ácido cítrico es uno de los ácidos fijos. Además, muestra correlación negativa con el ph y la acidez volátil.

La acidez volátil, muestra una correlación negativa fuerte con el ácido cítrico, así como con la calidad, cómo se ha comentado anteriormente.

En cuánto a la densidad, se observa una correlación negativa con el alcohol y positiva con la acidez fija y cítrica y el ph. Destacar que la correlación con la acidez fija es bastante fuerte.

5.0.2 Modelo de regresión lineal

Es interesante poder realizar predicciones sobre la calidad del vino dadas sus características.

```
#Todas las variables de entrada disponibles utilizadas.
mymodel2=lm(formula = quality~ ., data = data)
summary(mymodel2)
```

```
##
## Call:
## lm(formula = quality ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.69874 -0.36876 -0.04852  0.45350  2.01796
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.098e+01  2.131e+01   0.985   0.3250
## fixed.acidity         2.416e-02  2.602e-02   0.929   0.3531
## volatile.acidity     -1.083e+00  1.211e-01  -8.937  < 2e-16 ***
## citric.acid          -1.786e-01  1.475e-01  -1.211   0.2259
## residual.sugar        1.626e-02  1.501e-02   1.084   0.2787
## chlorides            -1.885e+00  4.201e-01  -4.488 7.71e-06 ***
## free.sulfur.dioxide   4.477e-03  2.186e-03   2.048   0.0408 *
## total.sulfur.dioxide -3.327e-03  7.415e-04  -4.487 7.74e-06 ***
## density              -1.699e+01  2.172e+01  -0.782   0.4342
## pH                   -4.167e-01  1.918e-01  -2.173   0.0299 *
## sulphates             9.188e-01  1.145e-01   8.025 1.96e-15 ***
## alcohol               2.863e-01  3.437e-02   8.329  < 2e-16 ***
## NivAbajo              2.583e-02  5.626e-02   0.459   0.6463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.6482 on 1586 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3558
## F-statistic: 74.55 on 12 and 1586 DF,  p-value: < 2.2e-16
```

El resultado del modelo lineal de la regresión lineal dice que la calidad de la variable de respuesta se puede explicar como: 21.965208 + 0.276198 (alcohol) -1.874225 (cloruros) -0.182564 (ácido cítrico) -17.881164 (densidad) + 0.024991 (acidez fija) + 0.004361 (dióxido de azufre libre) -0.413653 (pH pH) + 0.016331 (pH residual)) + 0.916334 (sulfatos) - 0.003265 (total.sulfuro.dióxido) -1.08359 (volatilidad.acidez)

El R2 ajustado no es alto en 0.3561 pero el valor de p de R2 es <0.05, por lo que estamos 95% seguros de que existe una relación entre al menos algunas de las variables de entrada y la clasificación de calidad.

Para obtener un modelo de regresión lineal considerablemente eficiente, lo que haremos mediante la selección de predictores empleando stepwise.

```
step(mymodel2, direction = "both", trace = 0)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, data = data)
##
## Coefficients:
##          (Intercept)      volatile.acidity             chlorides
##             4.430099            -1.012753            -2.017814
##  free.sulfur.dioxide  total.sulfur.dioxide                    pH
##             0.005077            -0.003482            -0.482661
##            sulphates              alcohol
##             0.882665             0.289303
```

```
#La selección de predictores empleando stepwise selection (hybrid/doble) ha identificado como mejor mod
mymodel2=lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol, data = data)
summary(mymodel2)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68918 -0.36757 -0.04653  0.46081  2.02954
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
## volatile.acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
## chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
## free.sulfur.dioxide   0.0050774  0.0021255   2.389    0.017 *
## total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
## pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
## sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
## alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.6477 on 1591 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3567
## F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

El resultado del modelo lineal de la regresión lineal dice que la calidad de la variable de respuesta se puede explicar como

4.430099 + 0.289303 (alcohol) - 2.017814 (cloruros) + 0.004361 (dióxido de azufre libre) -0.482661 (pH) + 0.882665 (sulfatos) -0.003482 (dióxido de azufre total) -1.012753 (volatilidad.acidez)

El resumen indica que los valores de p para la variable de entrada restante son mayores que 0.05. Como tal, rechazamos la hipótesis nula de que estas variables de entrada no hacen una contribución significativamente mayor que 0 a la varianza de la clasificación de calidad.

El R2 ajustado ha aumentado ligeramente, pero aún no es alto en 0.3567, pero el valor p del R2 sigue siendo <0.05, por lo que tenemos al menos un 95% de confianza de que existe una relación entre al menos algunas de las variables de entrada y la clasificación de calidad.

5.0.2.1 Predicciones

Se prueba el poder predictivo del modelo utilizando valores reales de variables de entrada

```
# Se espera rango 6
predict.lm(mymodel2, data.frame( alcohol=9.8, chlorides=0.075, free.sulfur.dioxide=17.0, pH=3.16, sulpha
```

```
##        1
## 5.694475
```

```
# Se espera rango 5
predict.lm(mymodel2, data.frame( alcohol=9.4, chlorides=0.076, free.sulfur.dioxide=11.0, pH=3.51, sulpha
```

```
##        1
## 5.024869
```

```
# Se espera rango 7
predict.lm(mymodel2, data.frame( alcohol=10, chlorides=0.065, free.sulfur.dioxide=15.0, pH=3.39, sulphat
```

```
##        1
## 5.315343
```

5.0.2.2 Resultados

Al redondear los resultados de las predicciones, podemos ver que el modelo predijo los valores correctos para los primeros conjuntos de variables de entrada, pero fue incorrecto cuando el valor esperado era 7.

Algunas sugerencias para explicaciones de los resultados son: La pérdida de información de relación al tratar la variable de respuesta ordinal como una variable continua La estrecha distribución de los valores de respuesta (Q1 = 5 Q3 = 6) proporcionó poca información para que la regresión lineal determine correctamente las relaciones para los valores fuera del primer y tercer cuartil. El R2 ajustado no es alto en 0.3567

5.0.3 Modelo de regresión logística sobre la calidad del vino

Vamos a separar la calidad del vino en dos tipos según un umbral de decisión e valor 6, basándonos en la media de la calidad. Por tanto, todo lo que quede por encima será de calidad ALTA mientras que lo que quede por debajo será de calidad BAJA. Para ello, vamos a crear una nueva variable binaria de nombre bad_quality y sobre ella aplicaremos una regresión logística. De tal manera, que si calidad es mala(menor que 6), bad_quality = 1 y sino bad_quality = 0(calidad buena). Utilizaremos las mismas variables para probar el modelo que las utilizadas en la regresión lineal y veremos su significancia en el mismo.

```r
# Creación variable binarias y modelo binario
data$bad_quality<- (data$quality < 6)*1
data$bad_quality<- as.factor(data$bad_quality)
modelo_bin <- glm(bad_quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol, family=binomial, data=data)
summary(modelo_bin)
```

```
##
## Call:
## glm(formula = bad_quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2857  -0.8358  -0.3163   0.8575   3.1850
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           5.941111   1.496558   3.970 7.19e-05 ***
## volatile.acidity      2.700549   0.386157   6.993 2.68e-12 ***
## chlorides             4.922388   1.471359   3.345 0.000821 ***
## free.sulfur.dioxide  -0.026283   0.008071  -3.256 0.001128 **
## total.sulfur.dioxide  0.018114   0.002713   6.676 2.46e-11 ***
## pH                    0.757145   0.435945   1.737 0.082424 .
## sulphates            -2.669704   0.431652  -6.185 6.22e-10 ***
## alcohol              -0.884124   0.072371 -12.217  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2209.0  on 1598  degrees of freedom
## Residual deviance: 1661.7  on 1591  degrees of freedom
## AIC: 1677.7
##
## Number of Fisher Scoring iterations: 4
```

Podemos determinar que todos los regresores tienen influencia significativa (Pr(>|z|) por debajo de 0.5), siendo el que menos influencia tiene el ph con un Pr de 0.4.

Además, podemos concluir que conforme aumenta "chlorides" y "volatile.acidity" desciende la calidad del vino considerablemente, siendo de manera más notoria en "chloridres". Además, vemos que cuanto menos sulfatos y menos alcohol, la calidad también tiende a empeorar, siendo más significativo en el caso de los sulfatos.

Veamos ahora la predicción con los mismos valores que para el caso de predicción anterior

```r
# Predicción con modelo de regresión logística

# Se espera bad_quality = 0
predict(modelo_bin, data.frame( alcohol=9.8, chlorides=0.075, free.sulfur.dioxide=17.0, pH=3.16, sulpha
```

```
##         1
## 0.4715828
```

```r
# Se espera bad_quality = 1
predict(modelo_bin, data.frame( alcohol=9.4, chlorides=0.076, free.sulfur.dioxide=11.0, pH=3.51, sulpha
```

```
##         1
## 0.7996547
```

```r
# Se espera bad_quality = 0
predict(modelo_bin, data.frame( alcohol=10, chlorides=0.065, free.sulfur.dioxide=15.0, pH=3.39, sulphat
```

```
##         1
## 0.6161615
```

En este caso, obtenemos resultados correctos para la predicción 1 y 2 pero incorrecto para la tercera . Comentar también que en el primer caso, la calidad era 6, que era el valor de calidad medio, por tanto, en este caso lo hemos incluido en el grupo de calidad buena, pero podríamos haberlo incluido en calidad mala, la cosa es que había que decidir al ser el valor medio. Lo demuestra el valor obtenido de 0.47 (cercano a 0.5)

De todas formas, para salir de dudas, vamos a realizar el mismo procedimiento incluyendo el 6 dentro de bad_quality.

```r
# Creación variable binarias y modelo binario
data$bad_quality<- (data$quality <= 6)*1
data$bad_quality<- as.factor(data$bad_quality)
modelo_bin2 <- glm(bad_quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol, family=binomial, data=data)
summary(modelo_bin2)
```

```
##
## Call:
## glm(formula = bad_quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0221   0.1238   0.2321   0.4341   2.4668
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           6.522230   2.128198   3.065 0.002179 **
## volatile.acidity      3.248292   0.629321   5.162 2.45e-07 ***
## chlorides             8.834449   3.194559   2.765 0.005684 **
## free.sulfur.dioxide  -0.010331   0.012441  -0.830 0.406293
## total.sulfur.dioxide  0.016517   0.004969   3.324 0.000888 ***
## pH                    1.742461   0.619724   2.812 0.004928 **
## sulphates            -3.328046   0.516739  -6.440 1.19e-10 ***
## alcohol              -0.998365   0.087355 -11.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  882.67  on 1591  degrees of freedom
## AIC: 898.67
##
```

```
## Number of Fisher Scoring iterations: 6
```

Nuevamente obtenemos que todos los regresores osn significativos, siendo esta vez el free.sulfur.dioxide el de menor influencia en el modelo. Las variables que más influyen en la mejora y empeoramiento de calidad son las mismas que las comentadas anteriormente en el anterior modelo de regresión lgística.

```
# Predicción con modelo de regresión logística

# Se espera bad_quality = 1
predict(modelo_bin2, data.frame( alcohol=9.8, chlorides=0.075, free.sulfur.dioxide=17.0, pH=3.16, sulpha
```

```
##         1
## 0.9371243
```

```
# Se espera bad_quality = 1
predict(modelo_bin2, data.frame( alcohol=9.4, chlorides=0.076, free.sulfur.dioxide=11.0, pH=3.51, sulpha
```

```
##         1
## 0.9916994
```

```
# Se espera bad_quality = 0
predict(modelo_bin2, data.frame( alcohol=10, chlorides=0.065, free.sulfur.dioxide=15.0, pH=3.39, sulphat
```

```
##         1
## 0.9772206
```

En este caso, acierta en los dos primeros valores y vuelve a fallar en el tercero. De todas formas, observamos como el valor de AIC es menor en este segundo modelo, siendo por tanto más fiable. Ahora, en comparación con el modelo linear planteado antes, no sabria asegurar cual funciona mejor, ya que en mi opinión harían falta más pruebas en profundidad para llegar a una conclusión óptima.

5.0.4 ¿La calidad del vino es superior si contiene más alcohol?

La segunda prueba estadística que se aplicará consistirá en un contraste de hipótesis sobre dos muestras para determinar si la calidad de vino es superior dependiendo de si contiene más alcohol. Para ello, tendremos dos muestras: la primera de ellas se corresponderá con la aproximadamente al mitad que tiene menos y la segunda el resto.

Aunque los datos se puedan aproximar por seguir una distribución normal, se ha demostrado que la calidad no presenta homogeneidad de varianzas cuando se divide en dos grupos de alcohol, por lo que la prueba de t-test no será del todo adecuada y será mejor aplicar un test no paramétrico como Mann-Whitney.

El test de Mann–Whitney–Wilcoxon (WMW), también conocido como Wilcoxon rank-sum test o u-test, es un test no paramétrico que contrasta si dos muestras proceden de poblaciones equidistribuidas.

La idea en la que se fundamenta este test es la siguiente: si las dos muestras comparadas proceden de la misma población, al juntar todas las observaciones y ordenarlas de menor a mayor, cabría esperar que las observaciones de una y otra muestra estuviesen intercaladas aleatoriamente. Por lo contrario, si una de las muestras pertenece a una población con valores mayores o menores que la otra población, al ordenar las observaciones, estas tenderán a agruparse de modo que las de una muestra queden por encima de las de la otra.

R contiene una función llamada wilcox.test() que realiza un test de Mann–Whitney–Wilcoxon entre dos muestras cuando se indica que paired = False y además genera el intervalo de confianza para la diferencia de localización.

```
#t.test(data.pocoA, data.muchoA, alternative = "less")
wilcox.test(x = data.pocoA$quality, y = data.muchoA$quality, alternative = "less", mu = 0,
            paired = FALSE, conf.int = 0.95)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data.pocoA$quality and data.muchoA$quality
## W = 174640, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
## 95 percent confidence interval:
##        -Inf -0.9999482
## sample estimates:
## difference in location
##             -0.9999819
```

Puesto que obtenemos un p-valor menor que el valor de significación fijado, rechazamos la hipótesis nula. Por tanto, podemos concluir que, efectivamente, la calidad del vino es superior si éste trae consigo más alcohol.

5.0.5 Análisis de varianzas ANOVA

Vamos a realizar un análisis de varianzas ANOVA para dos variables que consideramos significativas respecto a la calidad del vino para probar si deverdad influyen en la calidad final. Vamos a plantear un análisis de varianzas ANOVA con las variables de fixed.acidity y alcohol planteadas al inicio como grupos de valores a analizar para comprobar si influyen en la calidad del vino.

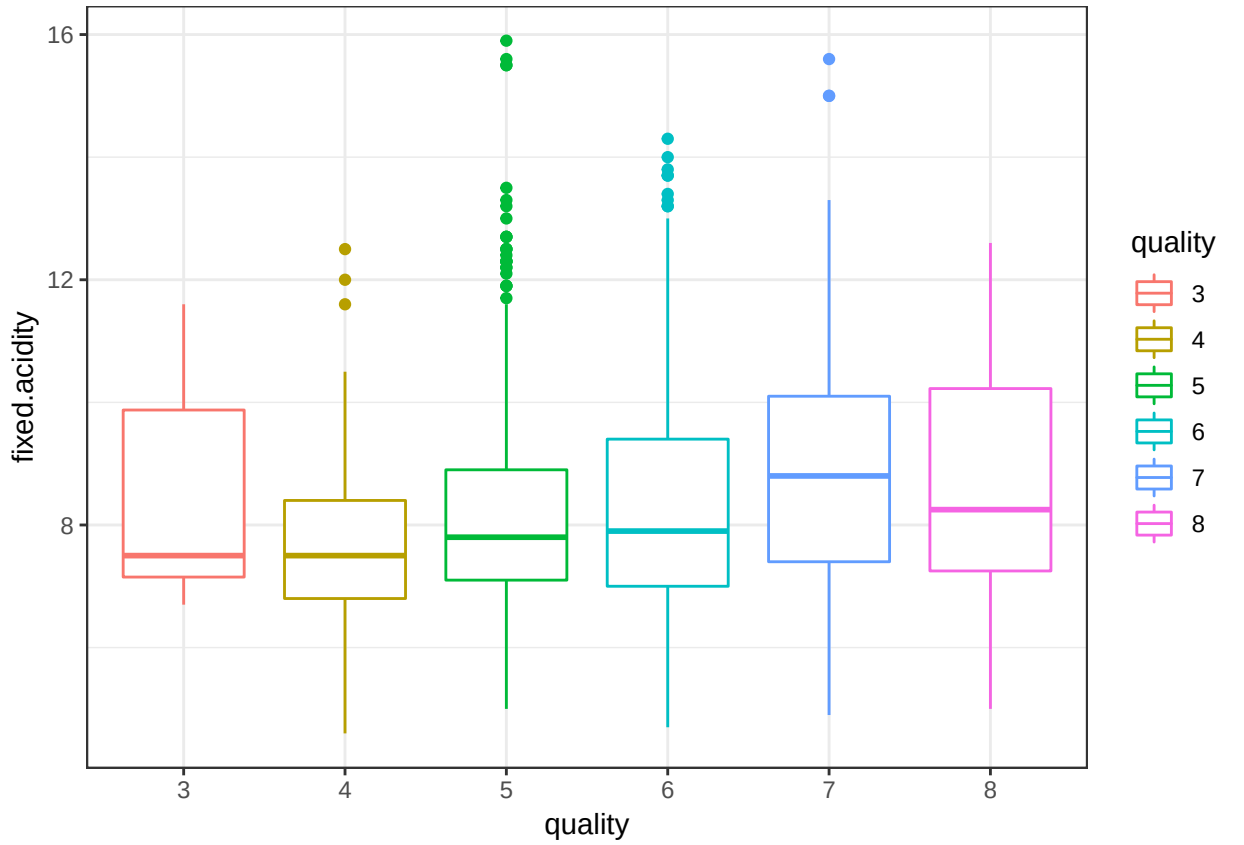
En el ANOVA de una vía la hipótesis nula H0 es que no hay diferencia entre las medias y la hipótesis alternativa H1 que al menos una de las medias difiere del resto. En nuestro caso, si el peso de cada individuo varía según el grupo de edad al que pertenece el sujeto.

Hipótesis nula es H0: 1= 2= 3

Hipótesis alternativa es al menos una es diferente H1: 1 2 3

El modelo ANOVA contrasta las diferencias en las medias en la calidad entre los vinos según el nivel de alcohol y la acidez fija. Para estimar el modelo ANOVA de una vía se usa la función aov( ), que sigue la estructura aov(variable dependiente ~ factor)

```
data$quality <- as.factor(data$quality)
ggplot(data = data, aes(x = quality, y = fixed.acidity, color = quality)) +
    geom_boxplot() +
    theme_bw()
```

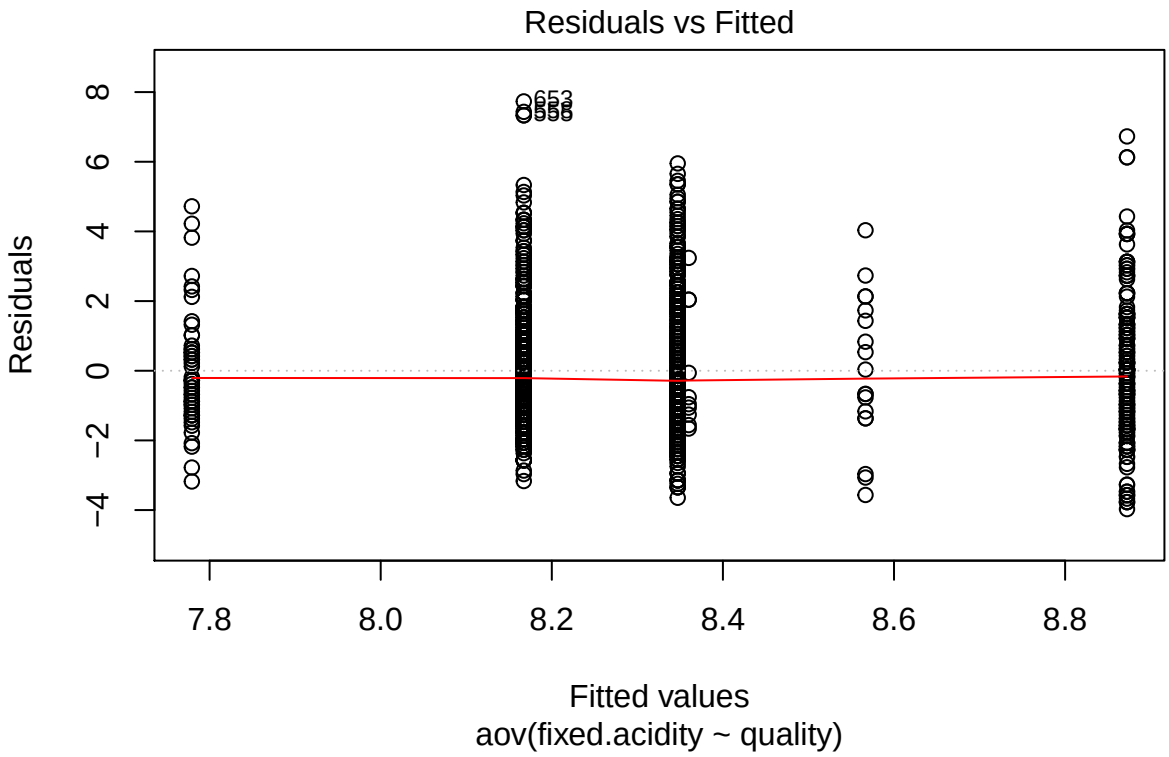


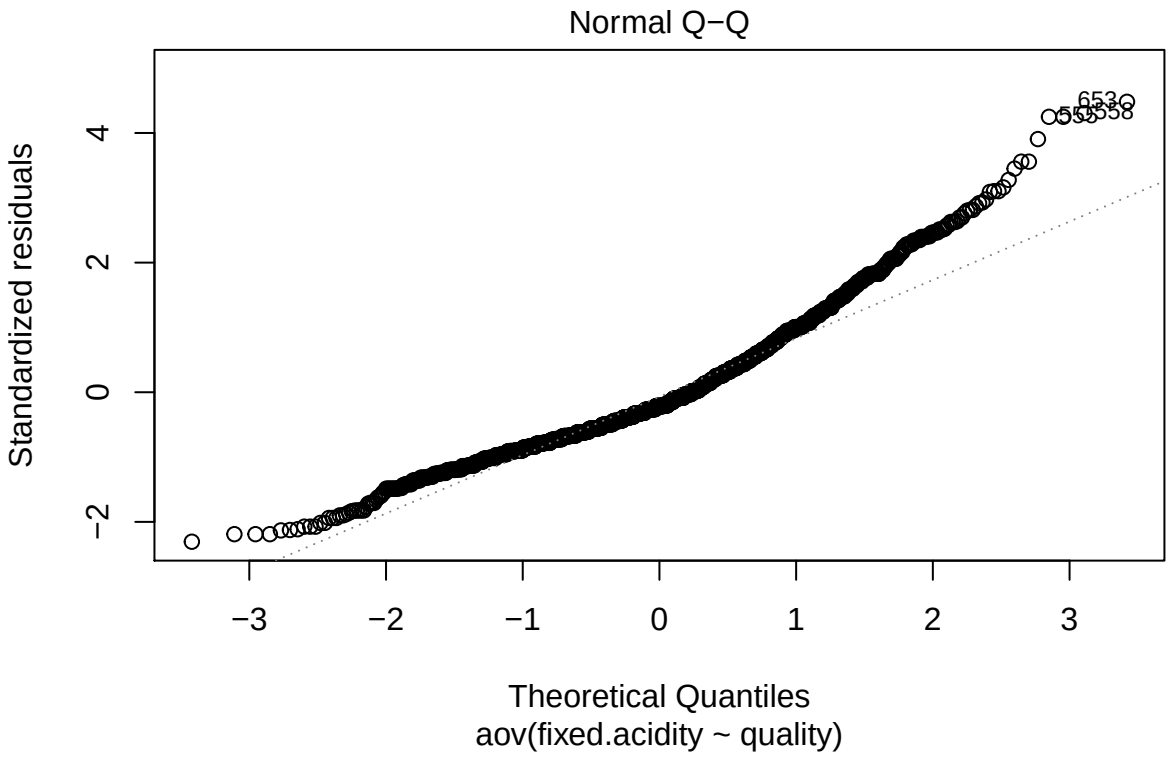

```
anova1 <- aov(fixed.acidity ~ quality, # var. cuantitativa con respecto al factor
              data = data )
summary( anova1 ) # para ver los rdos del anova
```

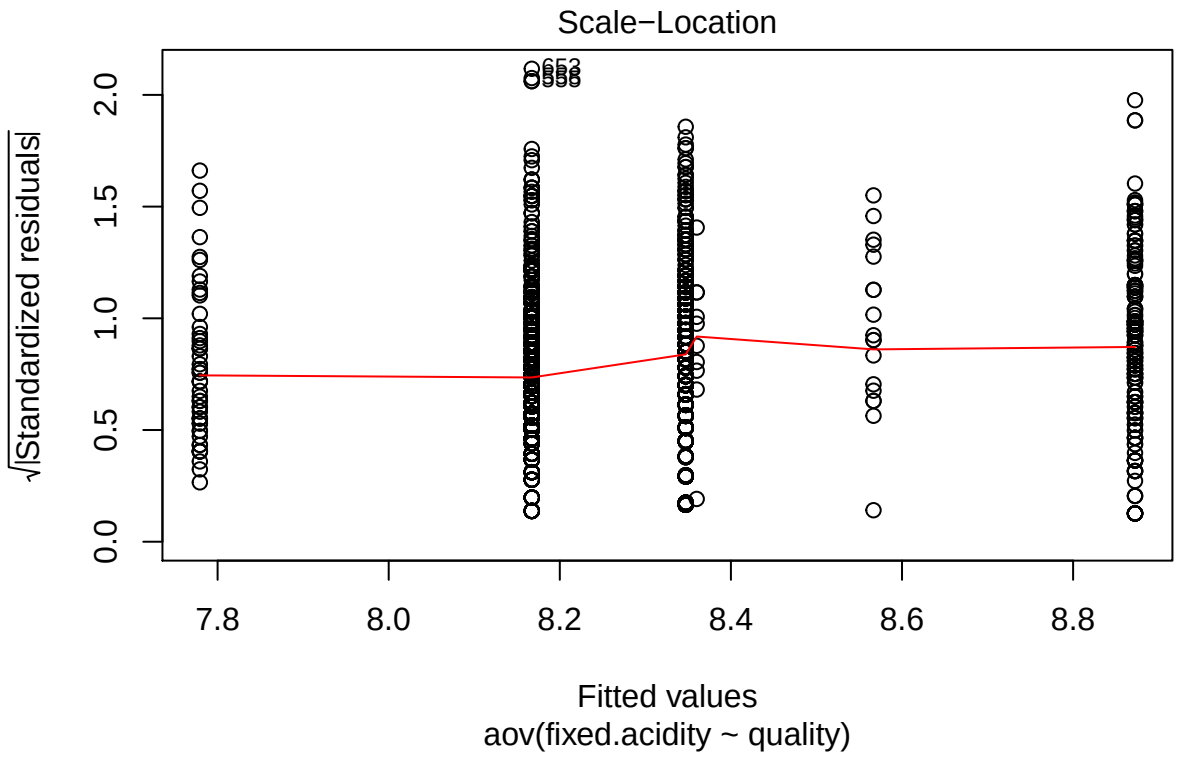
```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## quality        5     94  18.737   6.283 8.79e-06 ***
## Residuals   1593   4751   2.982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

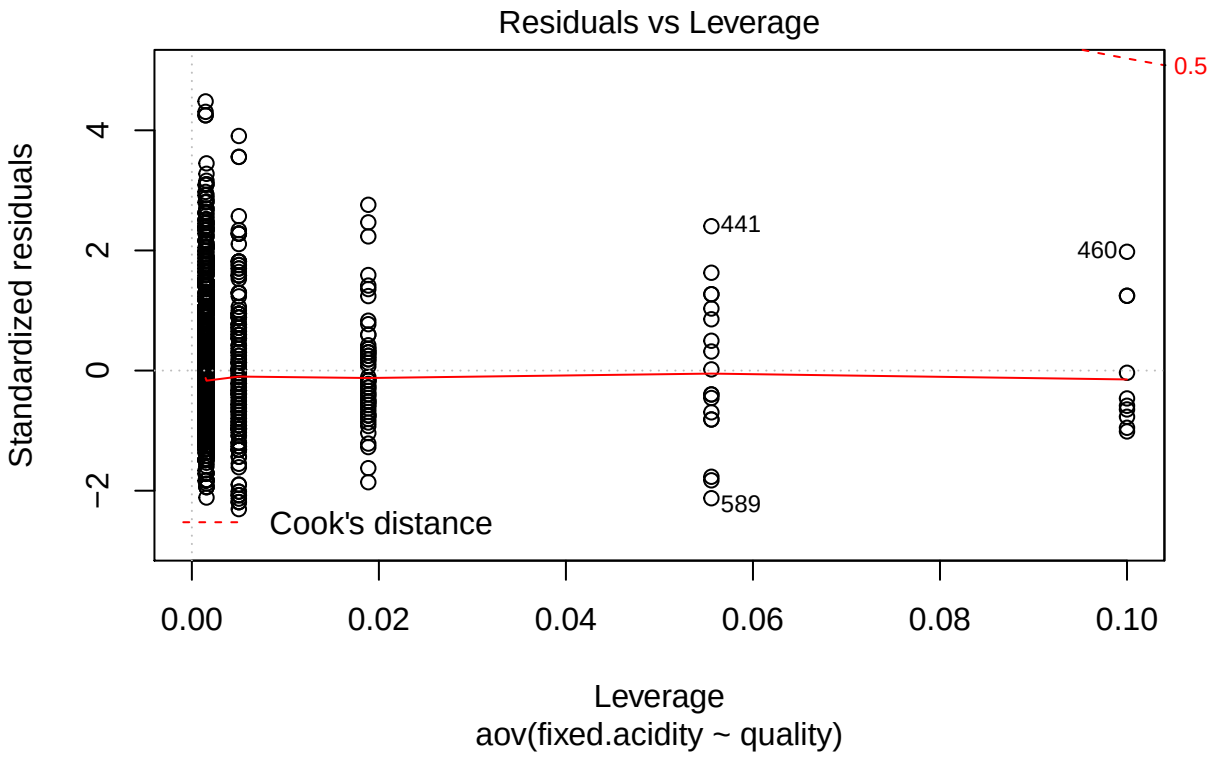
Con un p-valor de 8,79e-06 < 0.05 podemos decir que la acidez fija tiene un peso significativo en la calidad del vino. Además el tamaño de las cajas es similar en todos los niveles por lo que no hay indicios de falta de homocedasticidad. Los grupos de calidad parecen seguir una distribución simétrica.

```
plot( anova1 )
```


Residuals vs Fitted
Residuals
8
6
4
2
0
−2
−4
653
7.8
8.0
8.2
8.4
8.6
8.8
Fitted values
aov(fixed.acidity ~ quality)

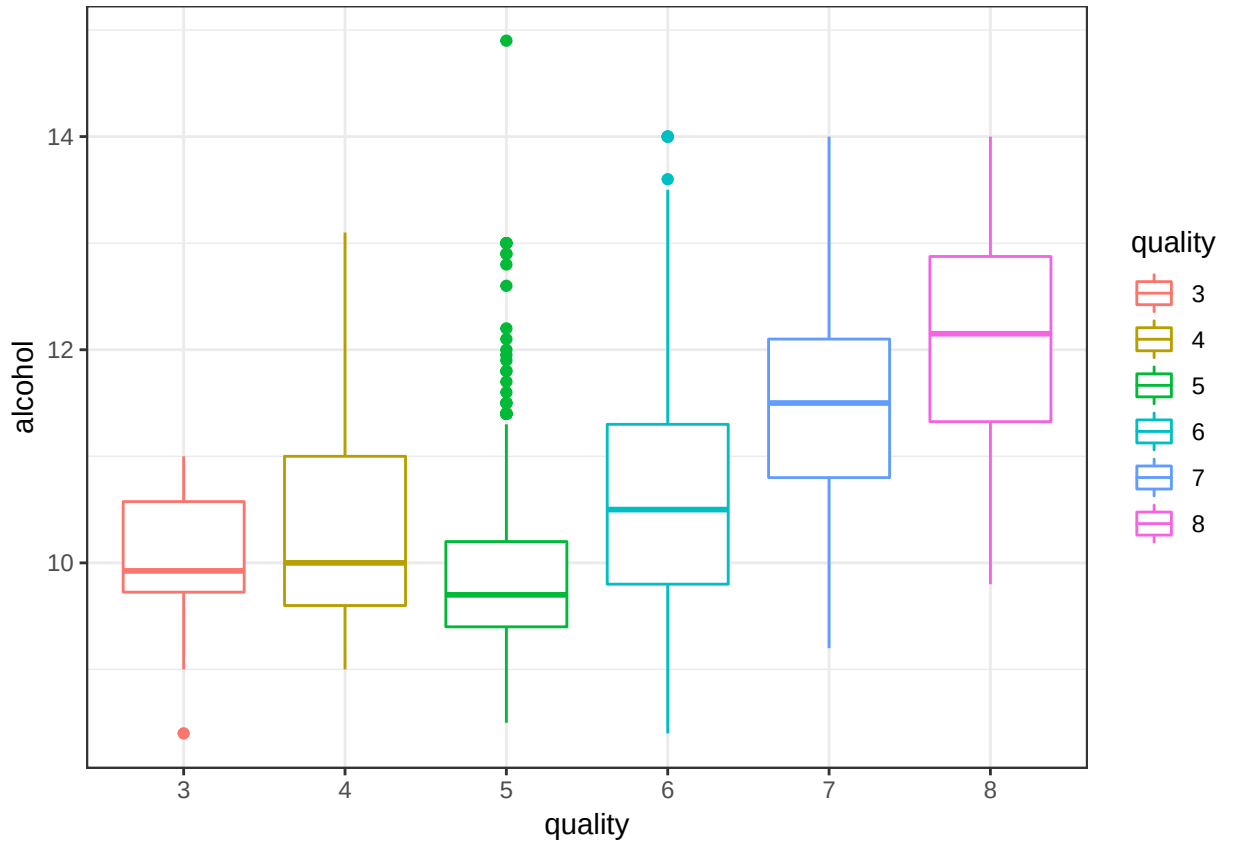

Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(fixed.acidity ~ quality)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
7.8
8.0
8.2
8.4
8.6
8.8
Fitted values
aov(fixed.acidity ~ quality)



La representación gráfica de los residuos no muestra falta de homocedasticidad (gráfico Residuals vs Fitted) y en el qqplot los residuos se distribuyen muy cercanos a la linea de la normal (gráfico Normal Q-Q).

```
ggplot(data = data, aes(x = quality, y = alcohol, color = quality)) +
    geom_boxplot() +
    theme_bw()
```

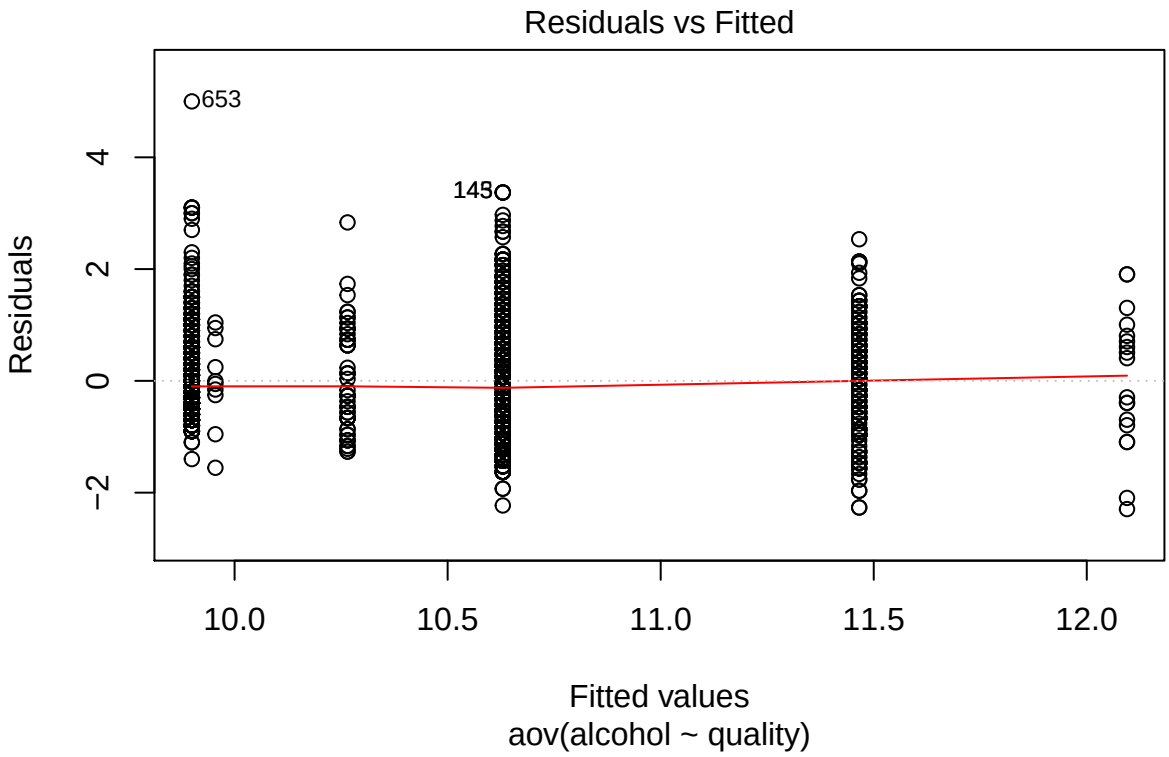


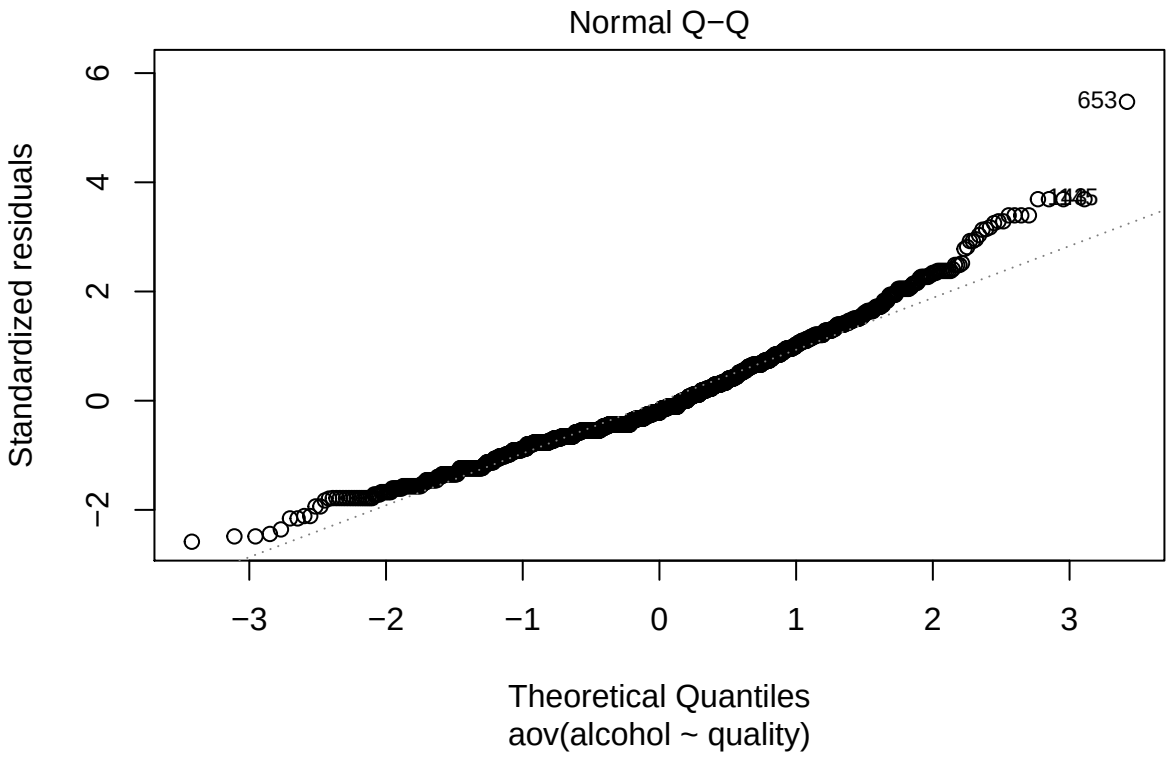

```
anova2 <- aov( alcohol ~ quality  , # var. cuantitativa con respecto al factor
               data = data )
summary( anova2 ) # para ver los rdos del anova
```

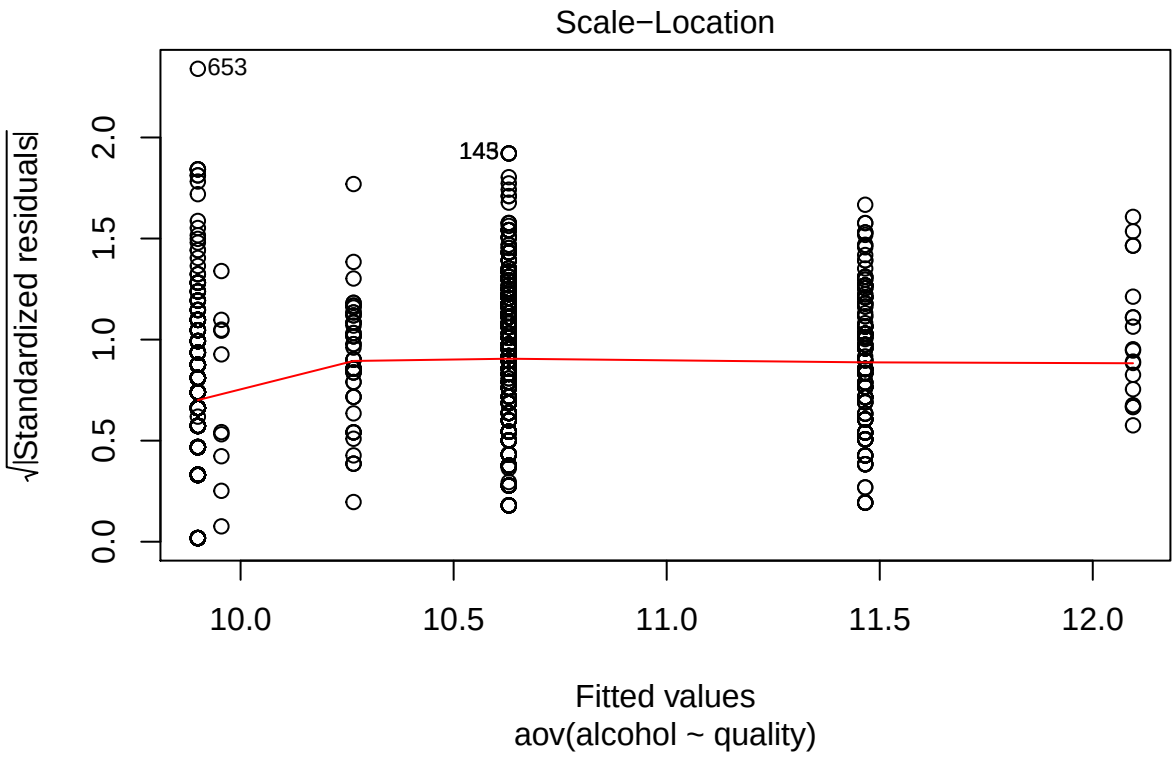
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## quality        5  483.9   96.79   115.9 <2e-16 ***
## Residuals   1593 1330.8    0.84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

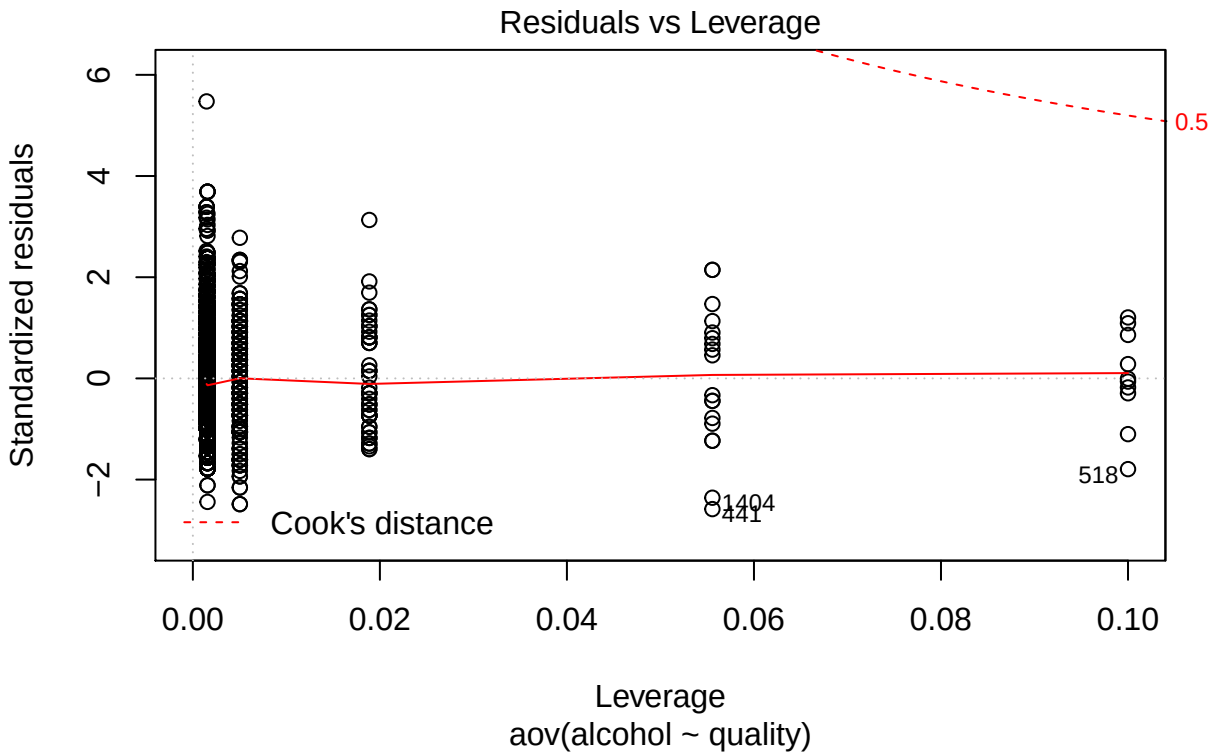
Con un p-valor de <2e-16 < 0.05 podemos decir que el alcohol tiene un peso significativo en la calidad del vino. Se observan claramente valores de calidad crecientes a partir de nivel 6 de calidad conforme aumenta el nivel de alcohol.

```
plot( anova2 ) # para ver los rdos del anova
```


Residuals vs Fitted
653
145
Residuals
4
2
0
−2
10.0
10.5
11.0
11.5
12.0
Fitted values
aov(alcohol ~ quality)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(alcohol ~ quality)
653

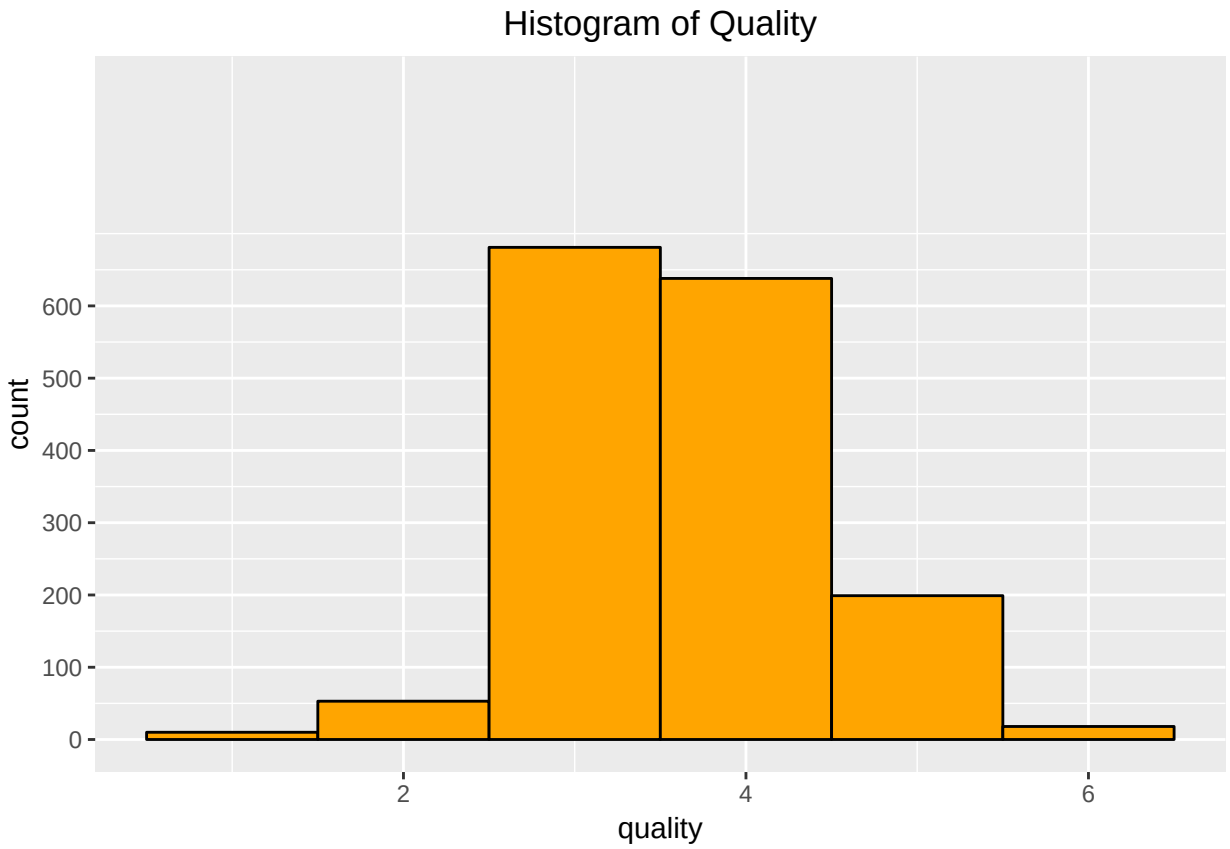

Scale−Location
653
145
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
10.0
10.5
11.0
11.5
12.0
Fitted values
aov(alcohol ~ quality)



La representación gráfica de los residuos no muestra falta de homocedasticidad (gráfico Residuals vs Fitted) y en el qqplot los residuos se distribuyen muy cercanos a la linea de la normal (gráfico Normal Q-Q).

6 Representación de los resultados a partir de tablas y gráficas.

```
data$quality <- as.integer(data$quality)
ggplot(aes(x=quality), data=data)+
  geom_histogram(aes(color=I('black'),fill=I('orange')),binwidth=1)+
  scale_y_continuous(lim=c(0,900), breaks=seq(0,600,100))+
  ggtitle('Histogram of Quality')+
  theme(plot.title=element_text(hjust=0.5))
```

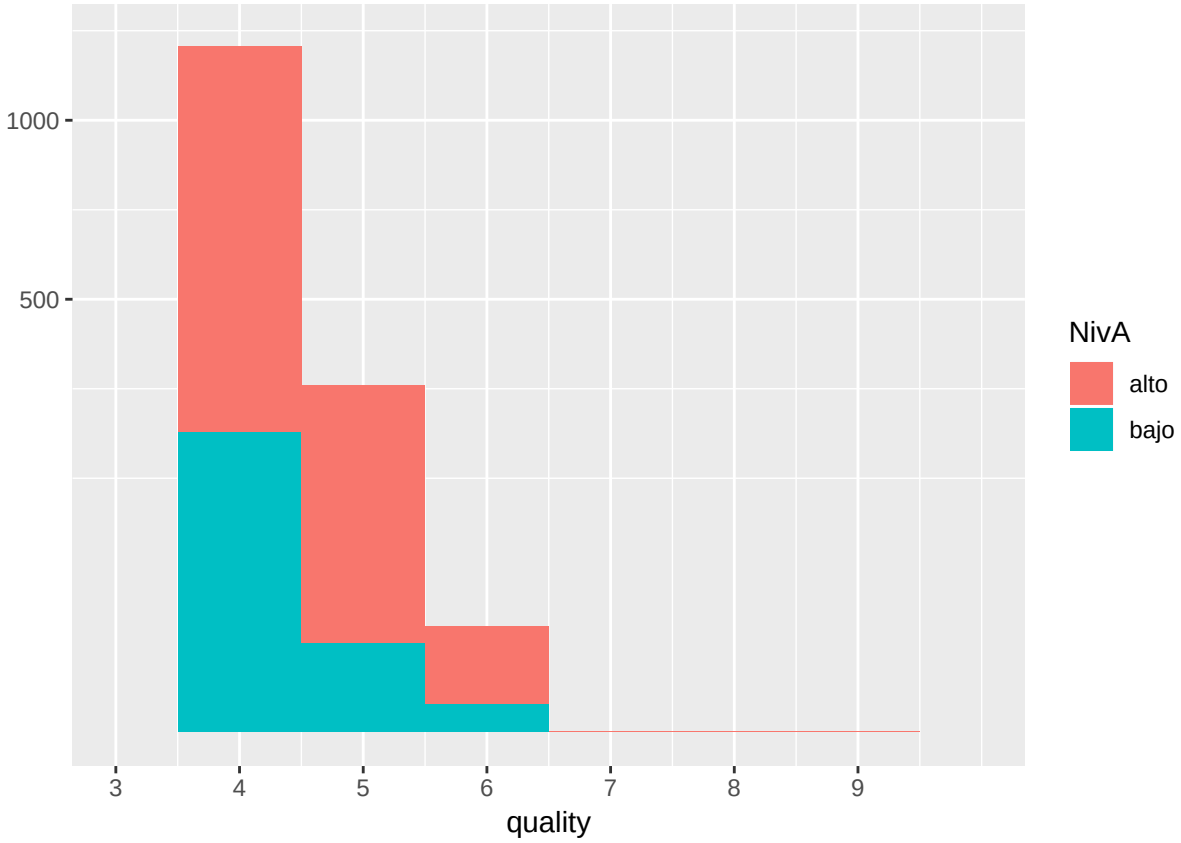


Como ya se habia visto la trama nos muestra que el recuento de la calidad más baja y más alta es muy pequeño. Esto puede indicar que las calificaciones de calidad más bajas y más altas se dieron solo en condiciones extremas. Como se espera, los valores de calidad promedio de 5 y 6 son los que más ocurren. Si lo dividimos en los 2 gupos de alcohol:

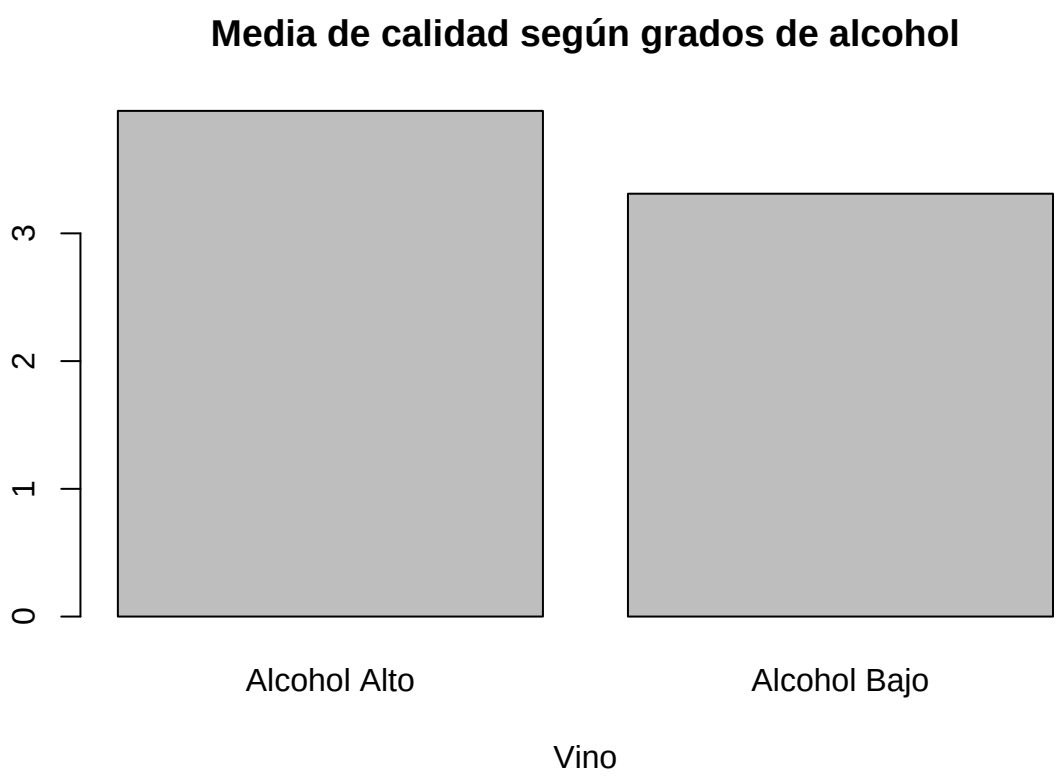
```
qplot(quality, data = data, fill = NivA, binwidth = 1) +
    scale_x_continuous(breaks = seq(3,9,1), lim = c(3,10)) +
    scale_y_sqrt()
```

```
## Warning: Removed 63 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```




```
DatamediaA=(data[data$NivA=="alto",])
DatamediaB=(data[data$NivA=="bajo",])

mediaA=mean(DatamediaA$quality)
mediaB=mean(DatamediaB$quality)

counts=c(mediaA, mediaB)
names(counts)=c("Alcohol Alto", "Alcohol Bajo")
barplot(counts, main="Media de calidad según grados de alcohol", xlab="Vino")
```



Es evidente que se las muestras analizadas, los vinos con mayor contenido de alcohol mostraron una mayor calidad.

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

En un principio se han sometido los datos a un preprocesamiento para manejar los casos de ceros o elementos vacíos y valores extremos (outliers). Se ha optado por incluir los valores extremos en los análisis dado que parecen no resultar del todo atípicos si los comparamos con los valores que toman las correspondientes variables para vinos que existen en el mercado actual.

Se han realizado tres tipos de pruebas estadísticas sobre un conjunto de datos que se correspondía con diferentes variables relativas al vino con motivo de cumplir en la medida de lo posible con el objetivo que se planteaba al comienzo. Para cada una de ellas, hemos podido ver cuáles son los resultados que arrojan (entre otros, mediante tablas) y qué conocimientos pueden extraerse a partir de ellas.

El análisis de correlación y el contraste de hipótesis nos ha permitido conocer cuáles de estas variables ejercen una mayor influencia sobre la calidad del vino El modelo de regresión lineal obtenido resulta de utilidad a la hora de realizar predicciones para esta variable dadas unas características concretas.Asimismo, se ha aplicado un modelo de regresión logística con la variable de calidad binaria entre calidad buena y mala. Sin embargo, siguen faltando pruebas para determinar cual de todos los modelos es más óptimo.

Finalmente se ha realizado un análisis ANOVA de las variables de fixed.acidity y alcohol llegando a la conclusión nuevamente de que influyen en la calidad del vino.

Una limitación del análisis es que los datos actuales consisten en muestras recopiladas de una región específica

de Portugal. Será interesante obtener conjuntos de datos en varias regiones vinícolas para eliminar cualquier sesgo creado por alguna cualidad específica del producto,

Hay muchos otros factores que están relacionados con los buenos vinos. Muchos de ellos están relacionados con olores y sabores y no con propiedades químicas y percepciones gustativas como las que tenemos en nuestro conjunto de datos. Aunque nuestras variables son un tanto explicativas de lo que tenemos, también hemos visto algunos casos en los que deben ser otras explicaciones para niveles de calidad altos o bajos.

8 Exportación del código en R y de los datos producidos.

Crear el archivo limpio

```
my.newfile <- "fichero_clean.csv"
write.csv(data, file=my.newfile, row.names = FALSE)
```

Referencias:

https://rpubs.com/Joaquin_AR/218456 (contraste de hipótesis)

https://rpubs.com/Joaquin_AR/218466 (homogeneidad de la varianza)

https://rpubs.com/Joaquin_AR/218465 (análisis normalidad)

https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/normality-test/interpret-the-results/key-results/

Recursos de la UOC de la asignatura estadística avanzada