Práctica 2: Limpieza y validación de datos

Irene Rodríguez Merchán

7 de enero 2019

Contents

1 Detalles de de la actividad

1.1 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.

1.2 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.

1.3 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.

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

Este conjunto de datos está relacionado con las variantes rojas del vino portugués "Vinho Verde". Gracias a este dataset se puede relacionar la calidad de los vinos con los distintos componentes del mismo. Así, se puede determinar de una forma objetiva qué variables influyen más en la eleboración de vinos excelentes o de mala calidad.

Sin duda se trata de una conjunto de gran importancia que permitirá a la industria del vino Portugués mejorar la calidad en su producción de vino tinto y aumentar las ventas a nivel nacional en internacional.

El dataset se ha obtenido a partir de Kaggle y está compuesto por las siguientes variables (columnas):

- **Acidez fija**: la mayoría de los ácidos relacionados con el vino que son fijos o no volátiles no se evaporan fácilmente.
- **Acidez volátil**: cuando la cantidad de estos acidos es demasiado alta puede conducir a un sabor desagradable parecido al vinagre
- ** Ácido cítrico **: se encuentra en pequeñas cantidades y puede agregar 'frescura' y sabor a los vinos.
- **Azúcar residual**: se trata de la cantidad de azúcar restante después de que se detenga la fermentación.
- **Cloruros**: la cantidad de sal en el vino.
- **Dióxido de azufre libre**: la forma libre de SO2 existe en equilibrio entre SO2 molecular (como gas disuelto) e ión bisulfito; previene microbios Crecimiento y oxidación del vino.
- **Dióxido de azufre total**: en baja concentraciones, el SO2 es mayormente indetectable en el vino, pero en concentraciones superiores a 50 ppm, el SO2 se hace evidente en la nariz y en el sabor del vino.
- **Densidad**: la densidad del agua es cercana a la del agua dependiendo de la porcentaje de alcohol y contenido de azúcar.
- ** pH **: describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); La mayoría de los vinos están entre 3-4 en la escala de pH.
- **Sulfatos**: un aditivo de vino que actúa como antimicrobiano y antioxidante.
- **Alcohol**: el porcentaje de alcohol del vino.
- **Calidad**: variable de salida (basada en datos sensoriales, puntuación entre 0 y 10)

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

Primeramente se cargarán los datos y se realizará una descripción de los mismos.

```
# Carga del dataframe
wineoriginal <- read.csv("winequality_red.csv", header = TRUE, sep = ",", dec = ".")
```

```
# Copia del fichero para trabajar en wine
wine <- wineoriginal
```

Se muestra a continuación los primeros y últimos valores del fichero wine para ver si se ha cargado correctamente.

```
# Primeros seis valores
head(wine)
```

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

```
# Últimos seis valores
tail(wine)
```

```
##      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1594           6.8            0.620        0.08            1.9     0.068
## 1595           6.2            0.600        0.08            2.0     0.090
## 1596           5.9            0.550        0.10            2.2     0.062
## 1597           6.3            0.510        0.13            2.3     0.076
## 1598           5.9            0.645        0.12            2.0     0.075
## 1599           6.0            0.310        0.47            3.6     0.067
##      free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
## 1594                  28                   38 0.99651 3.42      0.82
## 1595                  32                   44 0.99490 3.45      0.58
## 1596                  39                   51 0.99512 3.52      0.76
## 1597                  29                   40 0.99574 3.42      0.75
## 1598                  32                   44 0.99547 3.57      0.71
## 1599                  18                   42 0.99549 3.39      0.66
##      alcohol quality
```

```
## 1594     9.5       6
## 1595    10.5       5
## 1596    11.2       6
## 1597    11.0       6
## 1598    10.2       5
## 1599    11.0       6
```

Veamos si el tipo de cada variable es el correcto

```
# tipo de dato de cada variable
sapply(wine, function(x) class(x))
```

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

Todas las variables tienen clase numeric lo cual es correcto excepto quality que debería ser un factor y aparece como "integer".

```
#Tranformación de Quality de Integer  a Factor
wine$quality <- factor(wine$quality, ordered = T)

print(class(wine$quality))
```

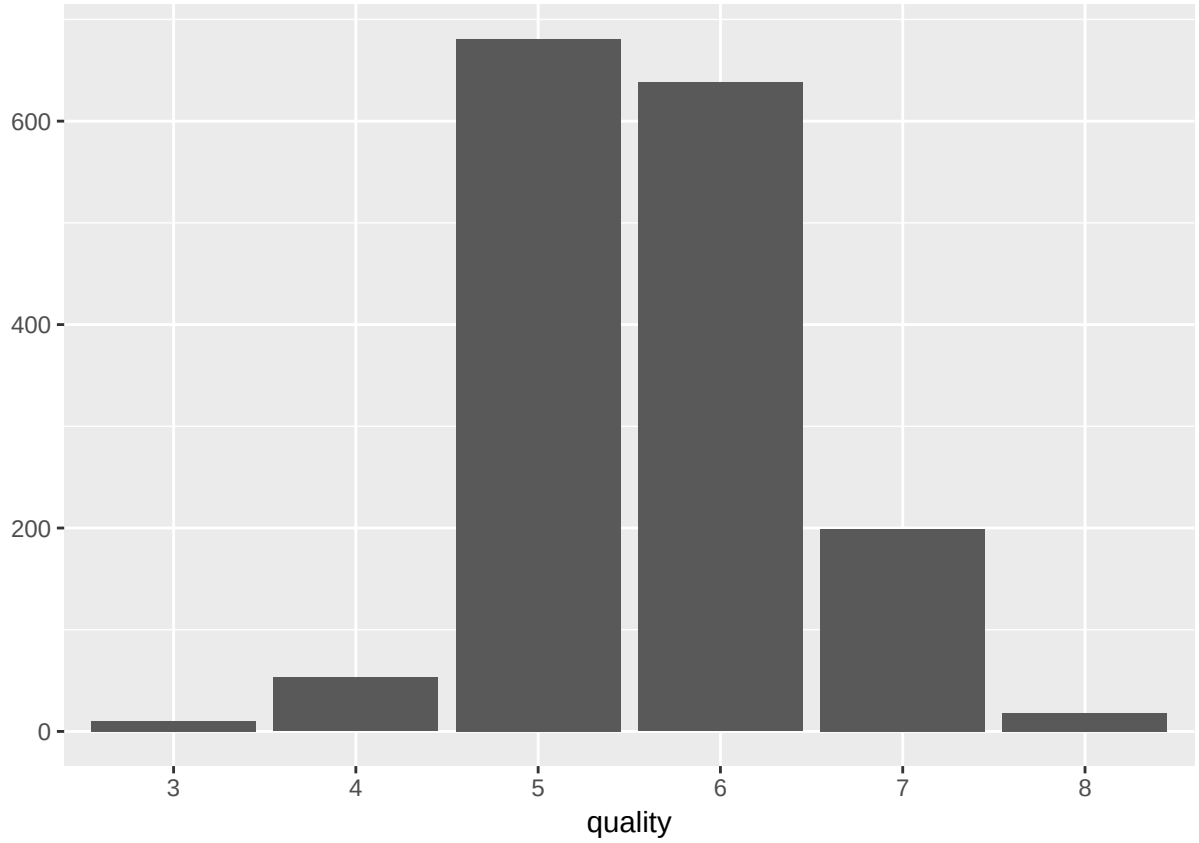
```
## [1] "ordered" "factor"
```

Veamos a continuación un resumen de los datos

```
summary(wine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  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      quality
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40   3: 10
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   4: 53
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   5:681
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42   6:638
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   7:199
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90   8: 18
```

```
qplot(quality, data = wine)
```



4. Limpieza de datos

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

Veamos a continuación si el dataset tiene elementos vaciós

```
sapply(wine, function(x) sum(is.na(x)))
```

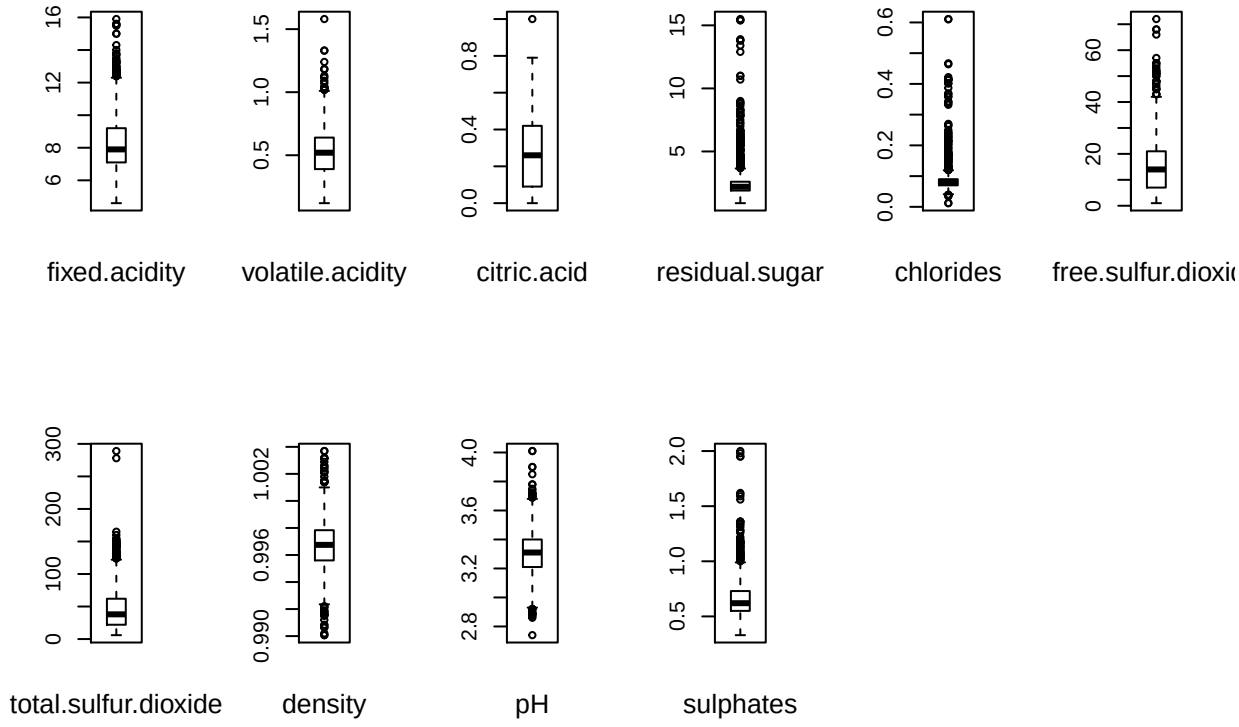
```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide         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

Por lo tanto, no se observan valores vacios.

4.2. Identificación y tratamiento de valores extremos

Primero vamos a realizar un boxplot para cada variable.

```
oldpar = par(mfrow = c(2,6))
for ( i in 1:10 ) {
  boxplot(wine[[i]])
  mtext(names(wine)[i], cex = 0.8, side = 1, line = 2)
}
par(oldpar)
```



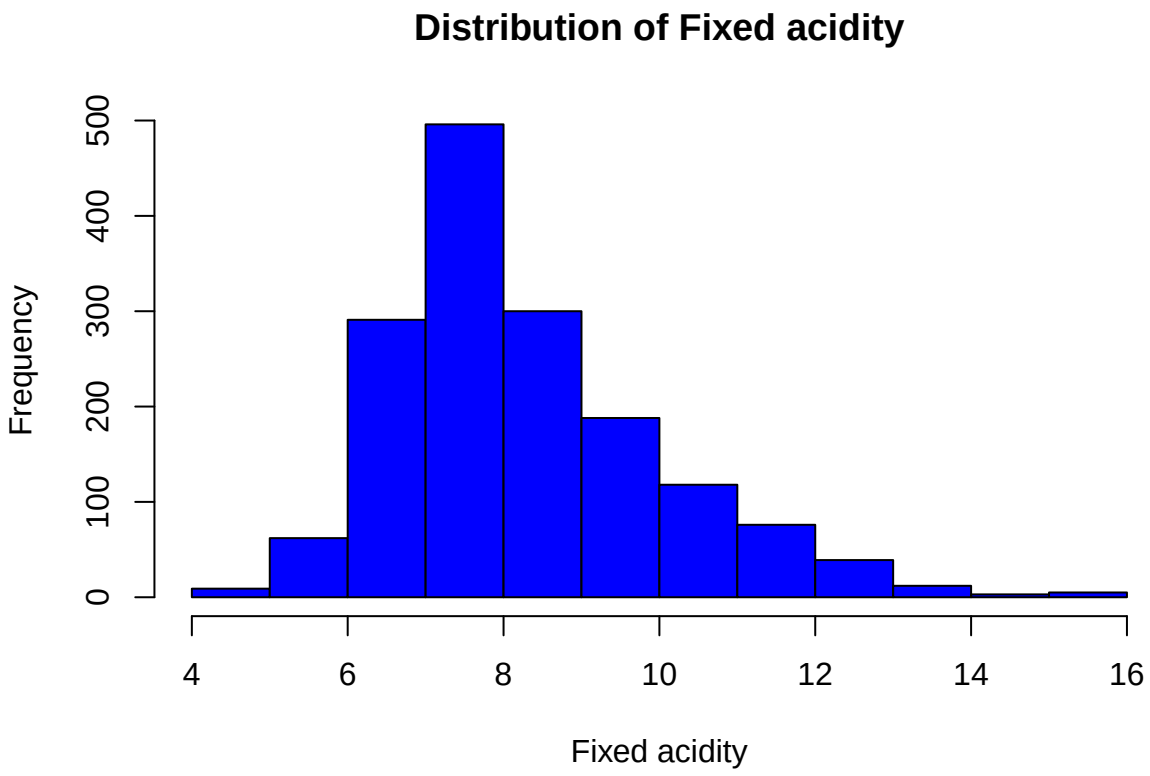
En la imagen más arriba se observan valores extremos en cada una delas variables del dataset.

Detectamos para cada variables cuales son los valores extremos.

```
boxplot.stats(wine$fixed.acidity)$out
```

```
##  [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8
## [15] 12.8 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4
## [29] 12.5 12.9 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2
## [43] 13.2 13.2 15.9 13.3 12.9 12.6 12.6
```

```
hist(wine$fixed.acidity,main = "Distribution of Fixed acidity",xlab = "Fixed acidity",ylab = "Frequency
```

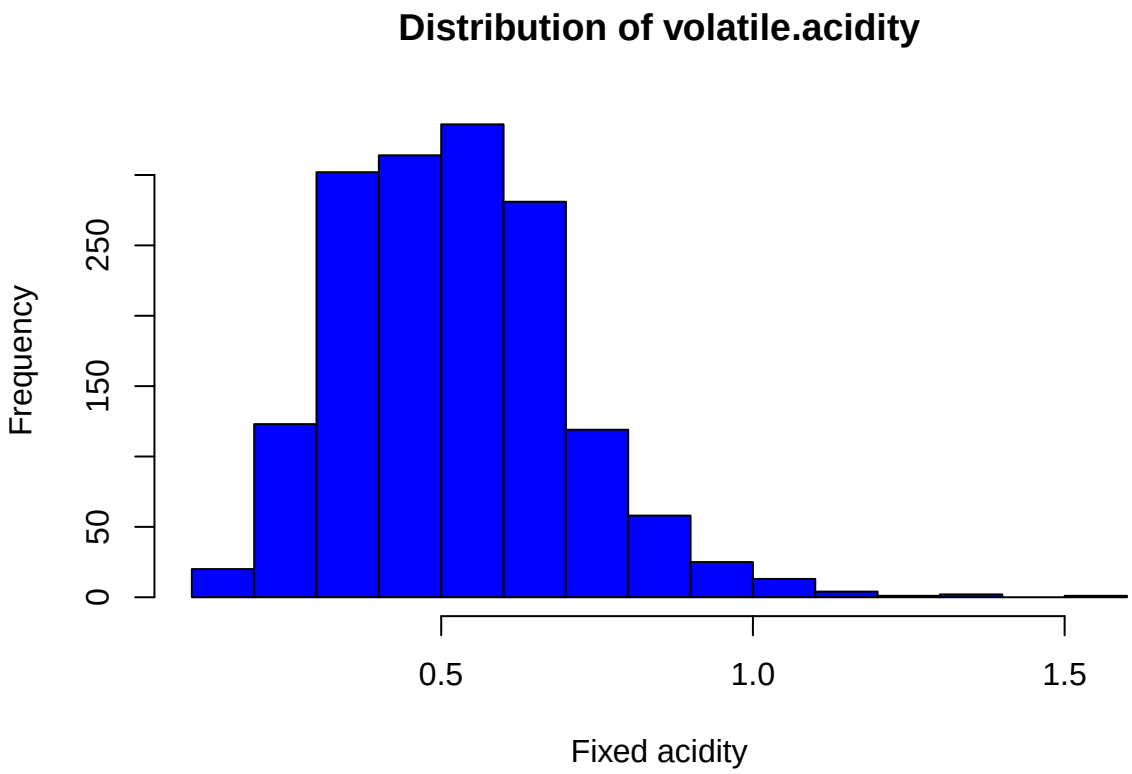


La mayoría de los vinos tienen entre 6 y 10 de acidez fija. A partir de 8 ya se empieza a considerar alta. Se toma la decisión de no borrar estos valores extremos ya que explorando el dataset se observa que algunos vinos con acidez alta (15) son calificados con calidad 7. Por lo tanto, se decide dejarlos.

```
boxplot.stats(wine$volatile.acidity)$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
```

```
hist(wine$volatile.acidity,main = "Distribution of volatile.acidity",xlab = "Fixed acidity",ylab = "Fre
```

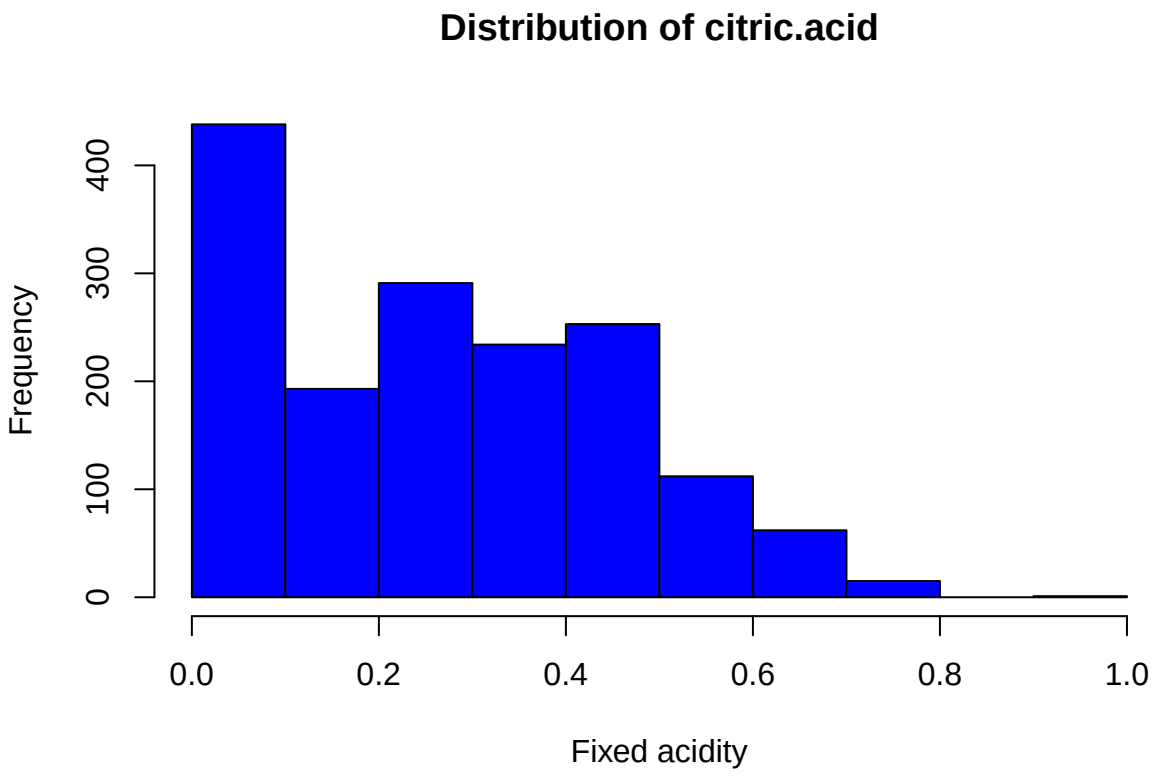


El máximo permitido por la organización internacional de la viña y el vino es es 20 por lo que aceptamos los valores extremos.

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

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

```
hist(wine$citric.acid,main = "Distribution of citric.acid",xlab = "Fixed acidity",ylab = "Frequency",col
```

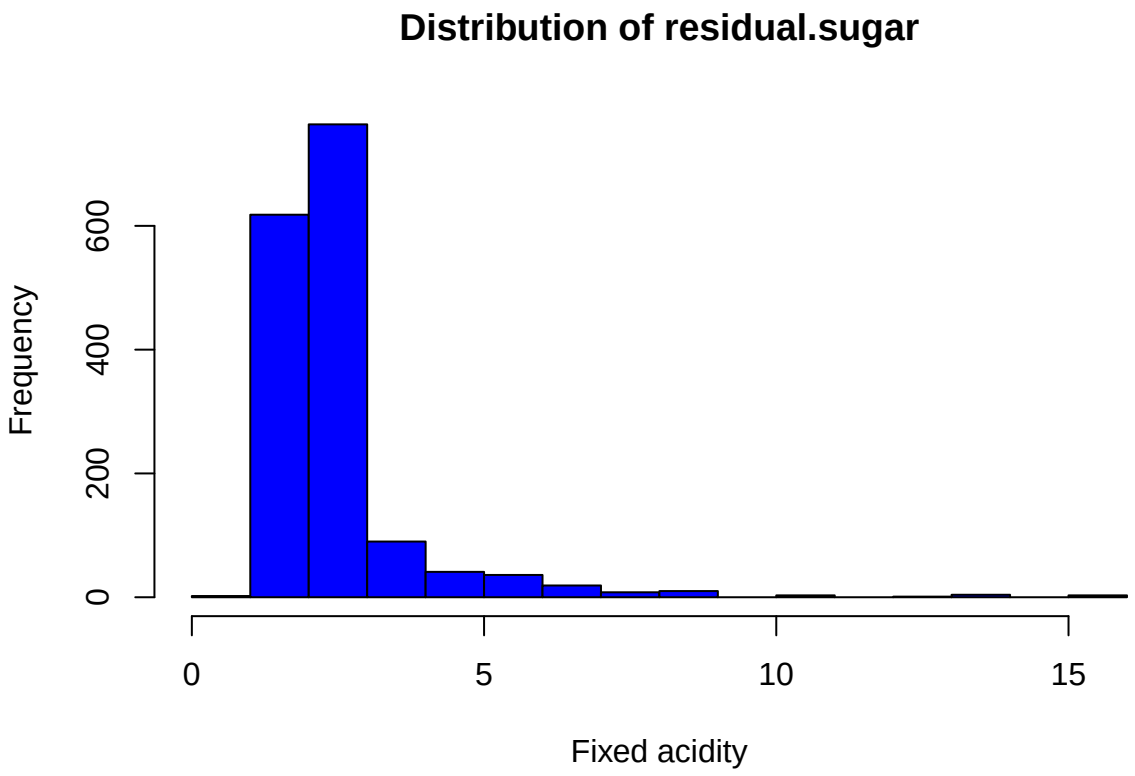


El máximo permitido por la organización internacional de la viña y el vino es es 1 por lo que aceptamos este único valor extremo.

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

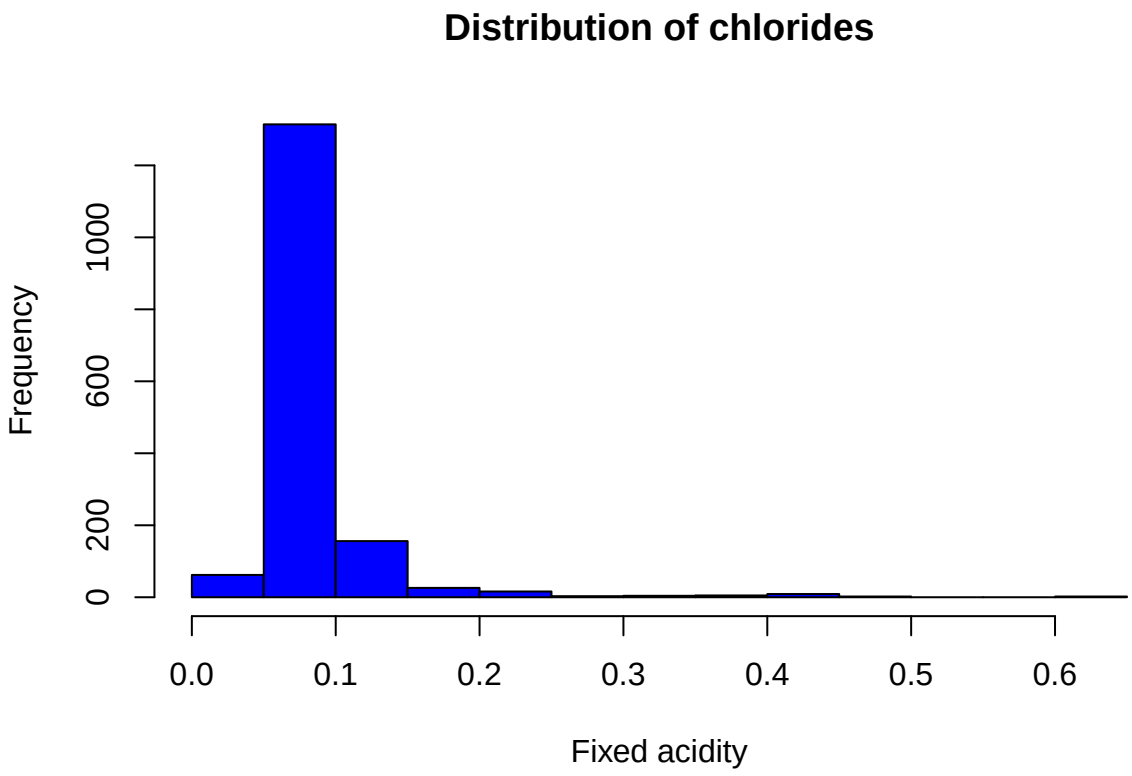
```
hist(wine$residual.sugar,main = "Distribution of residual.sugar",xlab = "Fixed acidity",ylab = "Frequen
```




```
boxplot.stats(wine$chlorides)$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
```

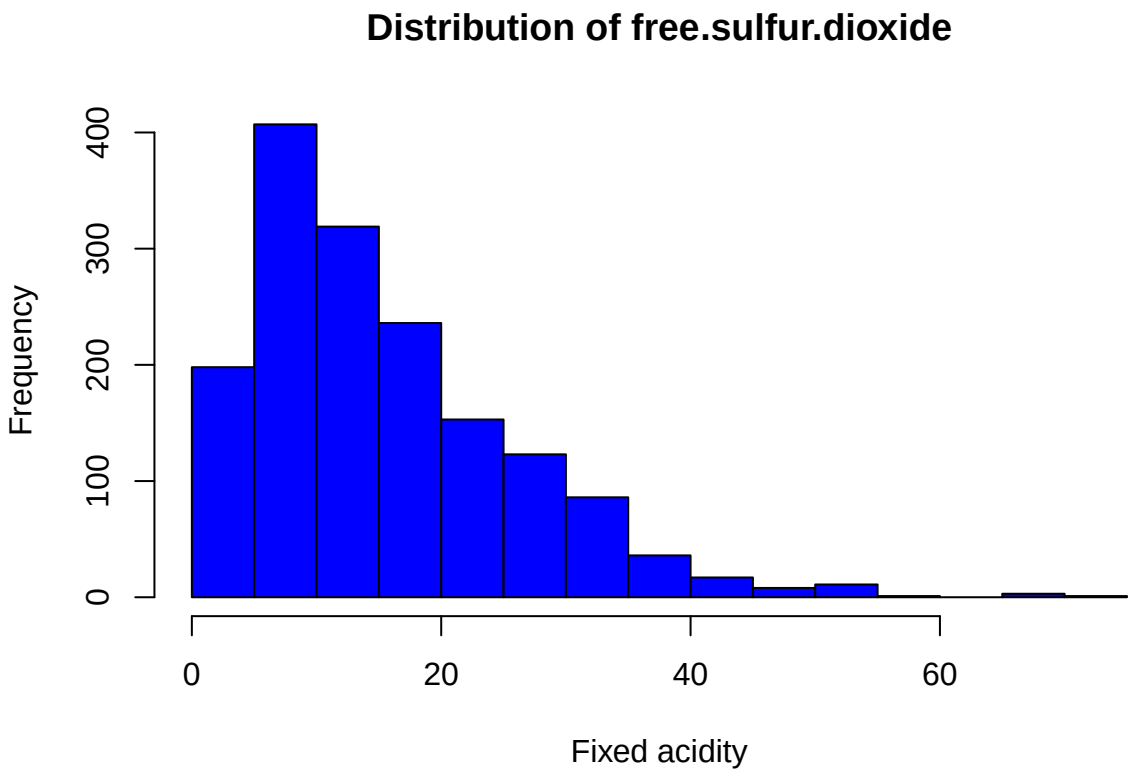
```
hist(wine$chlorides,main = "Distribution of chlorides",xlab = "Fixed acidity",ylab = "Frequency",col =
```




```
boxplot.stats(wine$free.sulfur.dioxide)$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
```

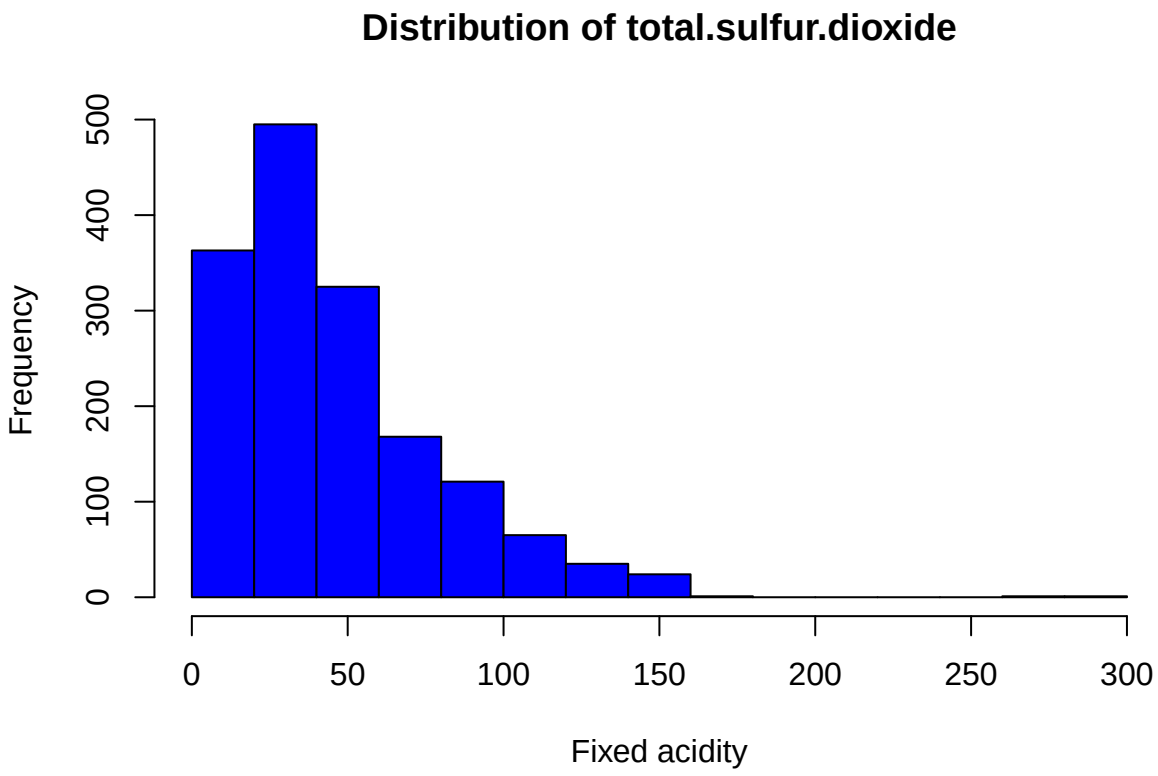
```
hist(wine$free.sulfur.dioxide,main = "Distribution of free.sulfur.dioxide",xlab = "Fixed acidity",ylab =
```




```
boxplot.stats(wine$total.sulfur.dioxide)$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
```

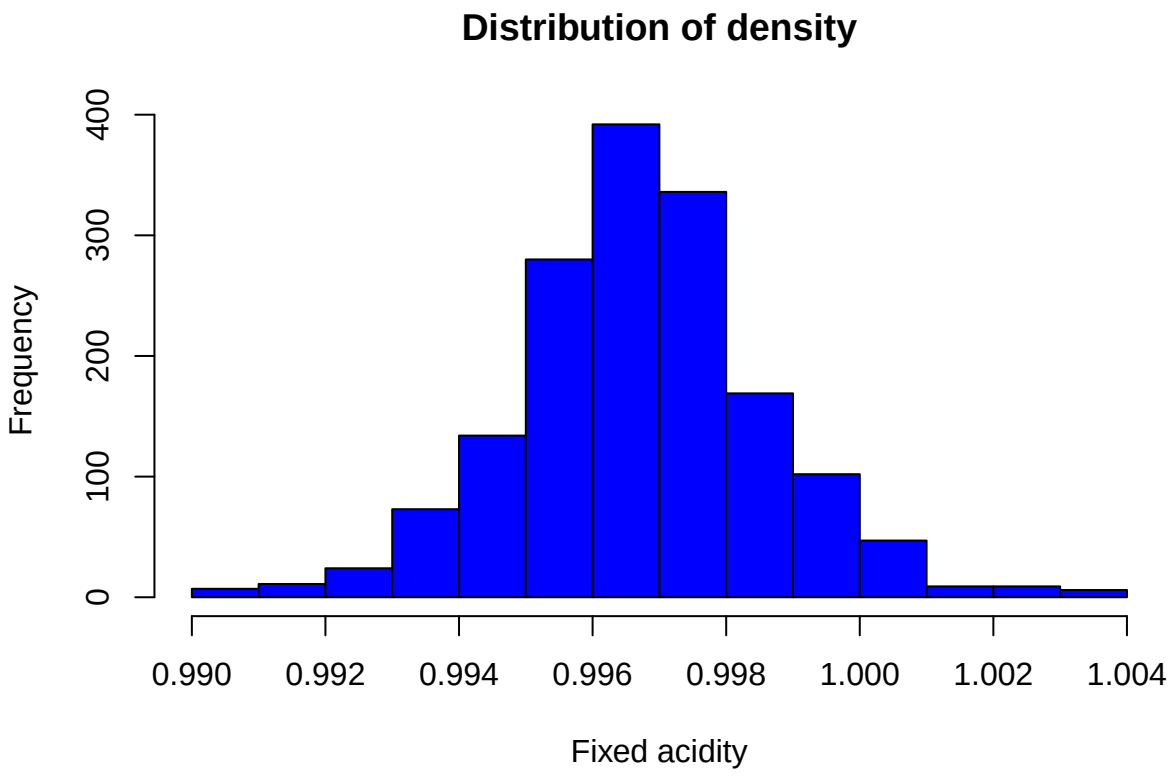
```
hist(wine$total.sulfur.dioxide,main = "Distribution of total.sulfur.dioxide",xlab = "Fixed acidity",ylab
```




```
boxplot.stats(wine$density)$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
```

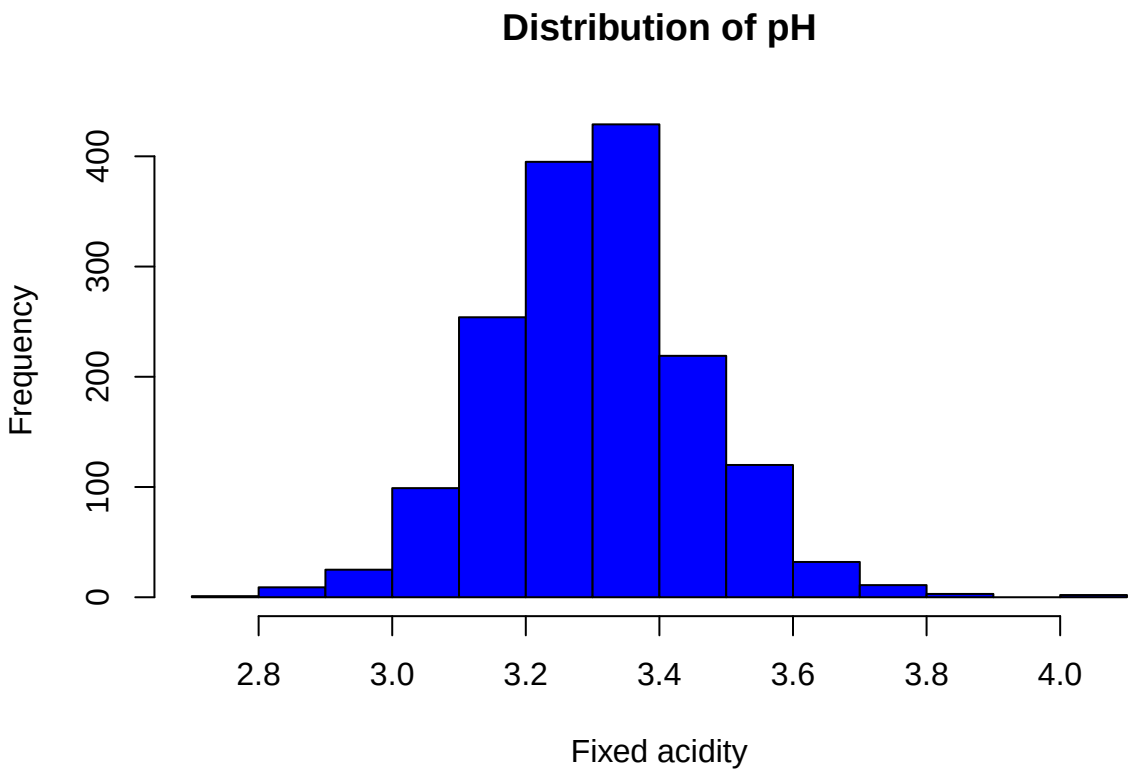
```
hist(wine$density,main = "Distribution of density",xlab = "Fixed acidity",ylab = "Frequency",col = "blu
```




```
boxplot.stats(wine$pH)$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
```

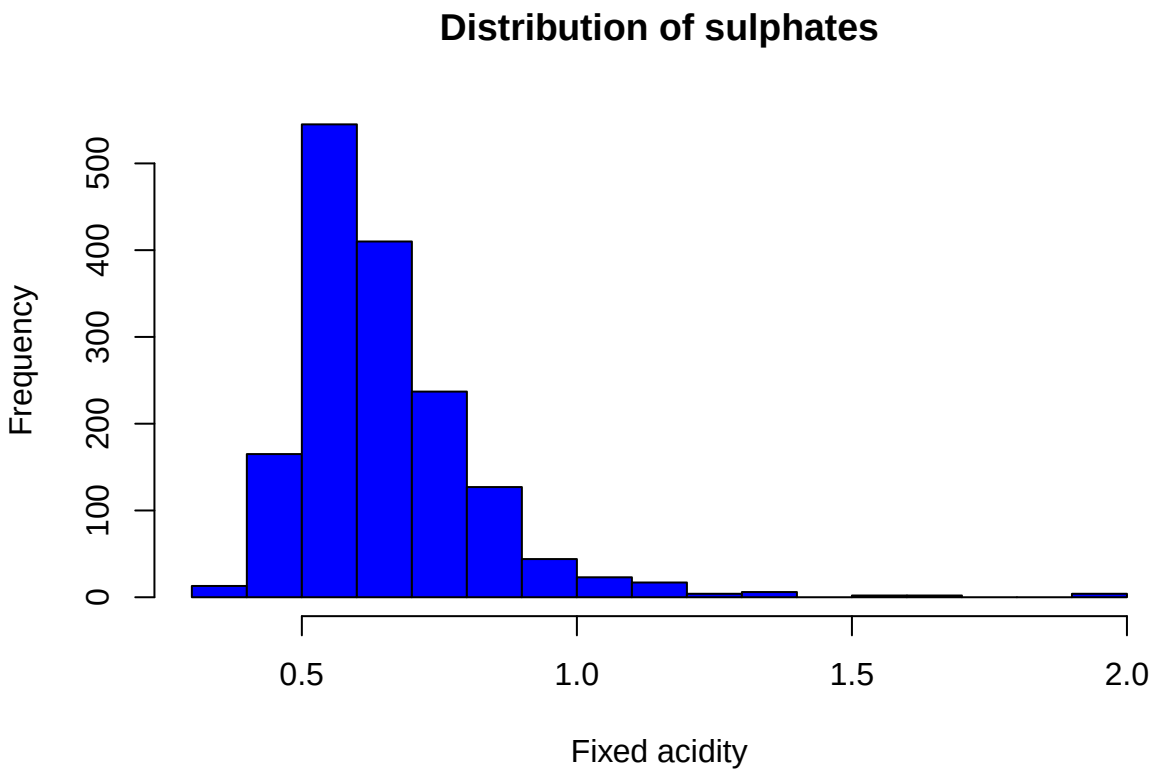
```
hist(wine$pH,main = "Distribution of pH",xlab = "Fixed acidity",ylab = "Frequency",col = "blue")
```




```
boxplot.stats(wine$sulphates)$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
```

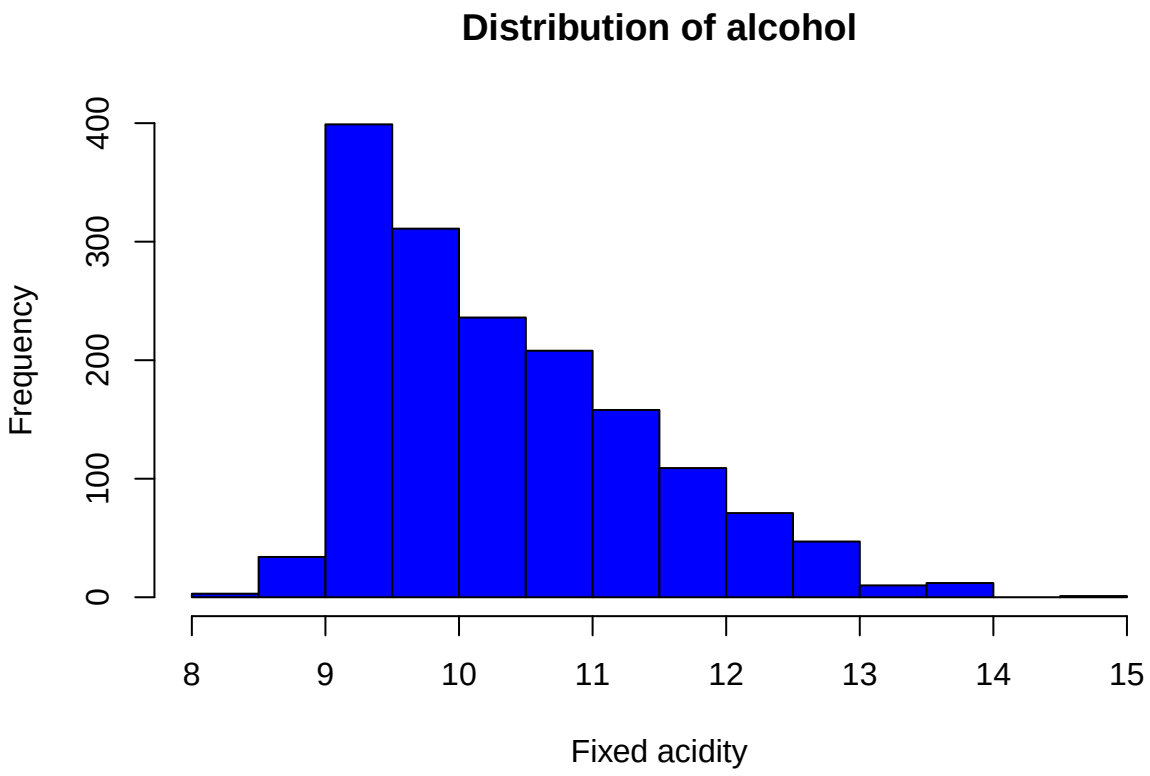
```
hist(wine$sulphates,main = "Distribution of sulphates",xlab = "Fixed acidity",ylab = "Frequency",col =
```




```
boxplot.stats(wine$alcohol)$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
```

```
hist(wine$alcohol,main = "Distribution of alcohol",xlab = "Fixed acidity",ylab = "Frequency",col = "blu
```

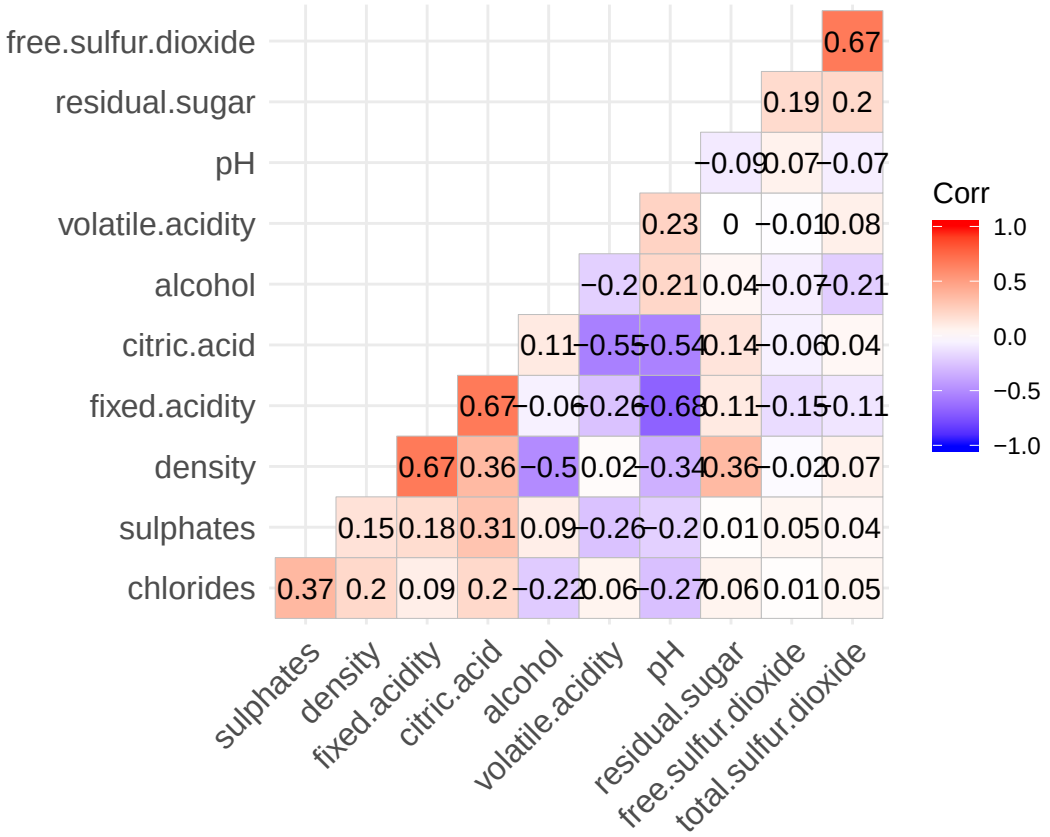


Se han contrastado los valores extremos con los limites proporcionados por la Organización Internaiconal de la viña y el vino (OIV). Aunque en alguna variable el valor extremo se encuentra muy al límite (por ejemplo el de ácido cítrico) se ha decido dejarlo y no considerarlos errores. Consideremos importante ver si algunos de esos valores afectan luego a la calificación proporcionada.

5 Análisis de los datos.

Para realizar el análisis de los datos primeramente estudiaremos la correlación entre las variables.

```
ggcorrplot(cor(wine[1:11]), hc.order = TRUE, type = "lower", lab = TRUE, insig = "blank")
```



Según la tabla de correlación más arriba. No se observan correlaciones muy altas por lo que no existe gran dependencia en las mismas.

La correlación positiva más alta se encuentra entre las variables:

- free.sulfur.dioxide y total.sulfur.dioxide
- fixed.acidity y density
- fixed.acidity y citric.acid

La correlación negativa más alta se encuentra entre las variables:

- fixed.acidity y pH
- citric.acid y volatile.acidity
- citric.acid y pH
- density y alcohol

Veamos ahora que variables está más relacionada con la calidad:

```
# function to return correlation
cor_test <- function(x, y) {
  return(cor(as.numeric(x), as.numeric(y)))
}
# calculate normal correlations
correlations <- c(
  cor_test(wine$fixed.acidity, wine$quality),
  cor_test(wine$volatile.acidity, wine$quality),
  cor_test(wine$citric.acid, wine$quality),
```

```
  cor_test(wine$residual.sugar, wine$quality),
  cor_test(wine$chlorides, wine$quality),
  cor_test(wine$free.sulfur.dioxide, wine$quality),
  cor_test(wine$total.sulfur.dioxide, wine$quality),
  cor_test(wine$density, wine$quality),
  cor_test(wine$pH, wine$quality),
  cor_test(wine$sulphates, wine$quality),
  cor_test(wine$alcohol, wine$quality))
names(correlations) <- c('fixed.acidity', 'volatile.acidity', 'citric.acid',
                         'residual.sugar', 'chlordies', 'free.sulfur.dioxide',
                         'total.sulfur.dioxide', 'density', 'pH',
                         'sulphates', 'alcohol')
print(correlations)
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.12405165          -0.39055778           0.22637251
##       residual.sugar            chlordies  free.sulfur.dioxide
##           0.01373164          -0.12890656          -0.05065606
## total.sulfur.dioxide              density                   pH
##          -0.18510029          -0.17491923          -0.05773139
##            sulphates              alcohol
##           0.25139708           0.47616632
```

La variable con una mayor correlación con la calidad es alcohol.

- alcohol (0.47)
- volatile acidity (-0.39)
- sulphates (log10) (0.25)
- citric acid (0.23)

Comprobemos si las variables cuantitativas provienen de una población distribuida normalmente. Para ello, utilizaremos la prueba de normalidad de Anderson Darling. Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de signi???cación pre???jado ?? = 0,05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

A continuación se va a crear la variable rating que clasificará a los vinos según sean malos (calidad menor de 5), normales o excelentes (calidad de 7 o 8).

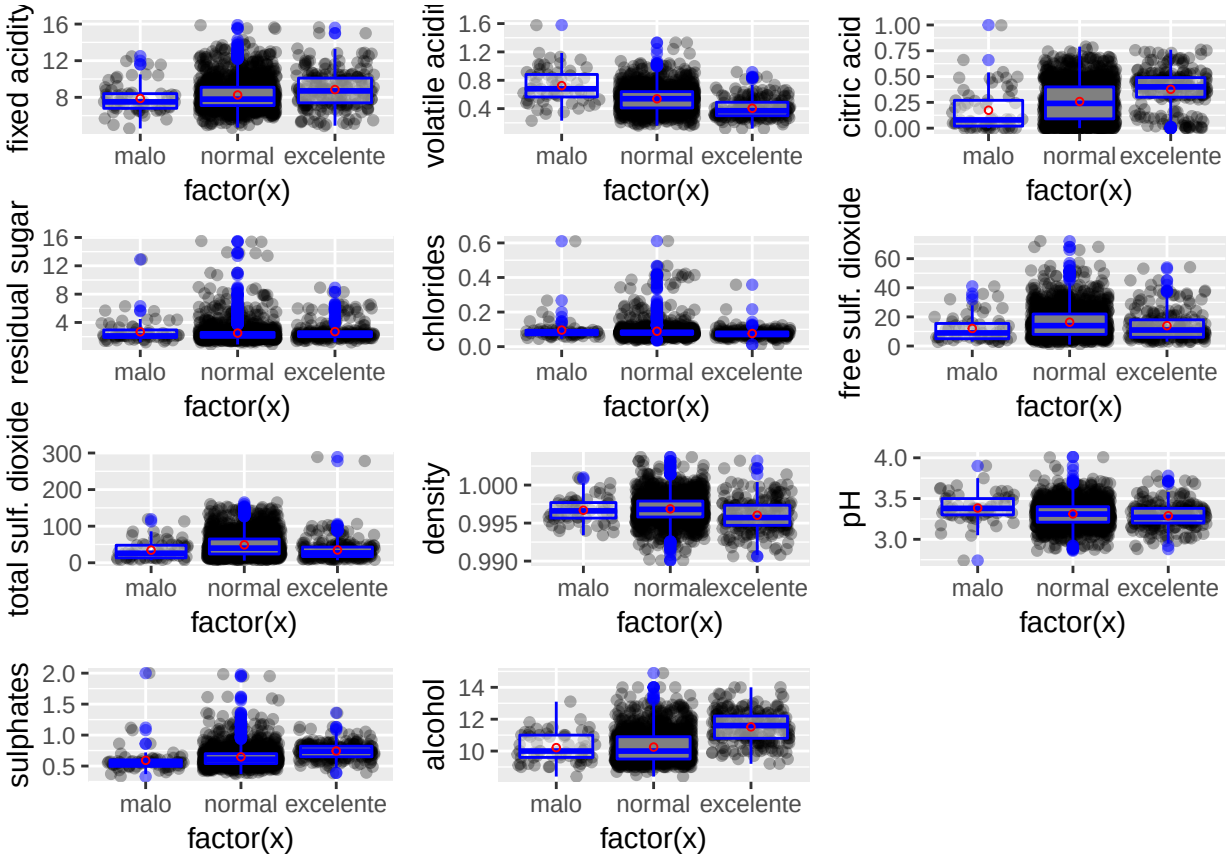
```
wine$rating <- ifelse(wine$quality < 5, 'malo', ifelse(
  wine$quality < 7, 'normal', 'excelente'))
wine$rating <- ordered(wine$rating,
                       levels = c('malo', 'normal', 'excelente'))
```

Vamos a analizar los boxplot según la variable rating

```
# Create a function to generate boxplots
get_bivariate_boxplot <- function(x, y, ylab) {
  return(ggplot(aes(factor(x), y), data = wine) +
           geom_jitter( alpha = .3)  +
           geom_boxplot( alpha = .5,color = 'blue')+
           stat_summary(fun.y=mean, shape=1, col = 'red',
                        geom = 'point', size = 1) +
           ylab(ylab))
}


grid.arrange(get_bivariate_boxplot(wine$rating, wine$fixed.acidity,
```

```
                          'fixed acidity'),
              get_bivariate_boxplot(wine$rating, wine$volatile.acidity,
                          'volatile acidity'),
              get_bivariate_boxplot(wine$rating, wine$citric.acid,
                          'citric acid'),
              get_bivariate_boxplot(wine$rating, wine$residual.sugar,
                          'residual sugar'),
              get_bivariate_boxplot(wine$rating, wine$chlorides,
                          'chlorides'),
              get_bivariate_boxplot(wine$rating, wine$free.sulfur.dioxide,
                          'free sulf. dioxide'),
              get_bivariate_boxplot(wine$rating,
                          wine$total.sulfur.dioxide,
                          'total sulf. dioxide'),
              get_bivariate_boxplot(wine$rating, wine$density,
                          'density'),
              get_bivariate_boxplot(wine$rating, wine$pH,
                          'pH'),
              get_bivariate_boxplot(wine$rating, wine$sulphates,
                          'sulphates'),
              get_bivariate_boxplot(wine$rating, wine$alcohol,
                          'alcohol'),
              ncol = 3)
```



Según los gráficos más arriba, se observa que los vinos excelentes tienen las siguientes características:

- Más alto el % de alcohol

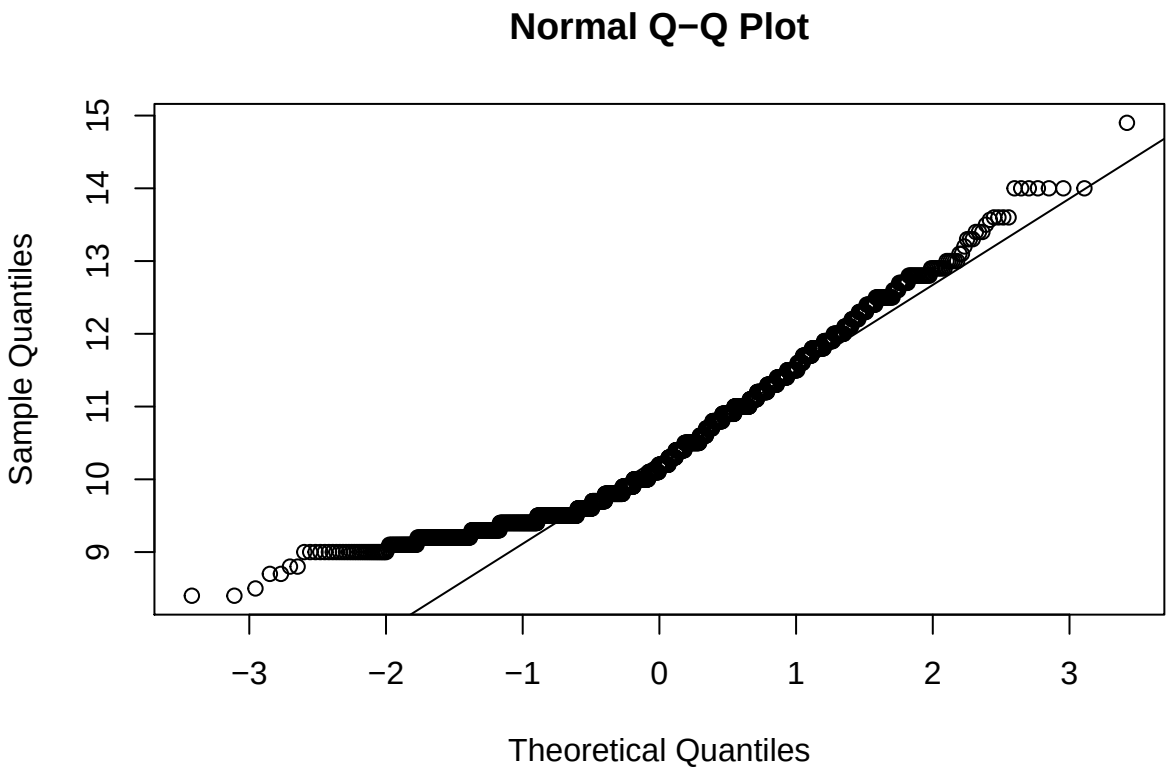
- Más alto la cantidad de sulfatos
- Más alta la cantidad de acidez fija, acidez cítrica y más baja la acidez volátil

Como la variable alcohol es una de las que más afecta a la calidad vamos a realizarel test de Shapiro para comprobar la normalidad

```
# Shapiro Test para comprobar normalidad
shapiro.test(wine$alcohol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine$alcohol
## W = 0.92884, p-value < 2.2e-16
```

```
qqnorm(wine$alcohol)
qqline(wine$alcohol)
```



En el test de Shapiro-Wilk, cuando Pr(D)es mayor o igual a ?? entonces se acepta la hipótesis nula, existe normalidad. El valor p del test de Shapiro ha dado 2.2e-16. Por tanto, se rechaza la hipótesis nula de normalidad. Asumimos que la muestra sigue una que no es normal. No obstante, la condición de normalidad se debe cumplir para cada grupo. Por ello, se debe aplicar la prueba de normalidad a cada grupo.

```
# Shapiro Test para comprobar normalidad
DS <- summarize( group_by(wine, rating), n=length(alcohol), p.shapiro=shapiro.test(alcohol)[[2]])
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
DS
```

```
## # A tibble: 3 x 3
```

```
##   rating        n p.shapiro
##   <ord>     <int>     <dbl>
## 1 malo         63  3.20e- 2
## 2 normal     1319  3.89e-27
## 3 excelente   217  2.32e- 1
```

Tampoco se cumple la condición de normalidad para cada uno de los grupos. Veamos la igualdad de varianzas entre grupos

```
pairwise.t.test(wine$alcohol, wine$rating, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  wine$alcohol and wine$rating
##
##           malo   normal
## normal    0.77   -
## excelente <2e-16 <2e-16
##
## P value adjustment method: none
```

Por lo tanto la variable alcohol de los vinos excelentes presentan diferencias significativas con respecto al resto de categorías.

Una vez que hemos realizado sobre el conjunto de datos inicial los procedimientos de integración, validación y limpieza anteriores, procedemos a guardar estos en un nuevo ???chero denominado wine_data_clean.csv:

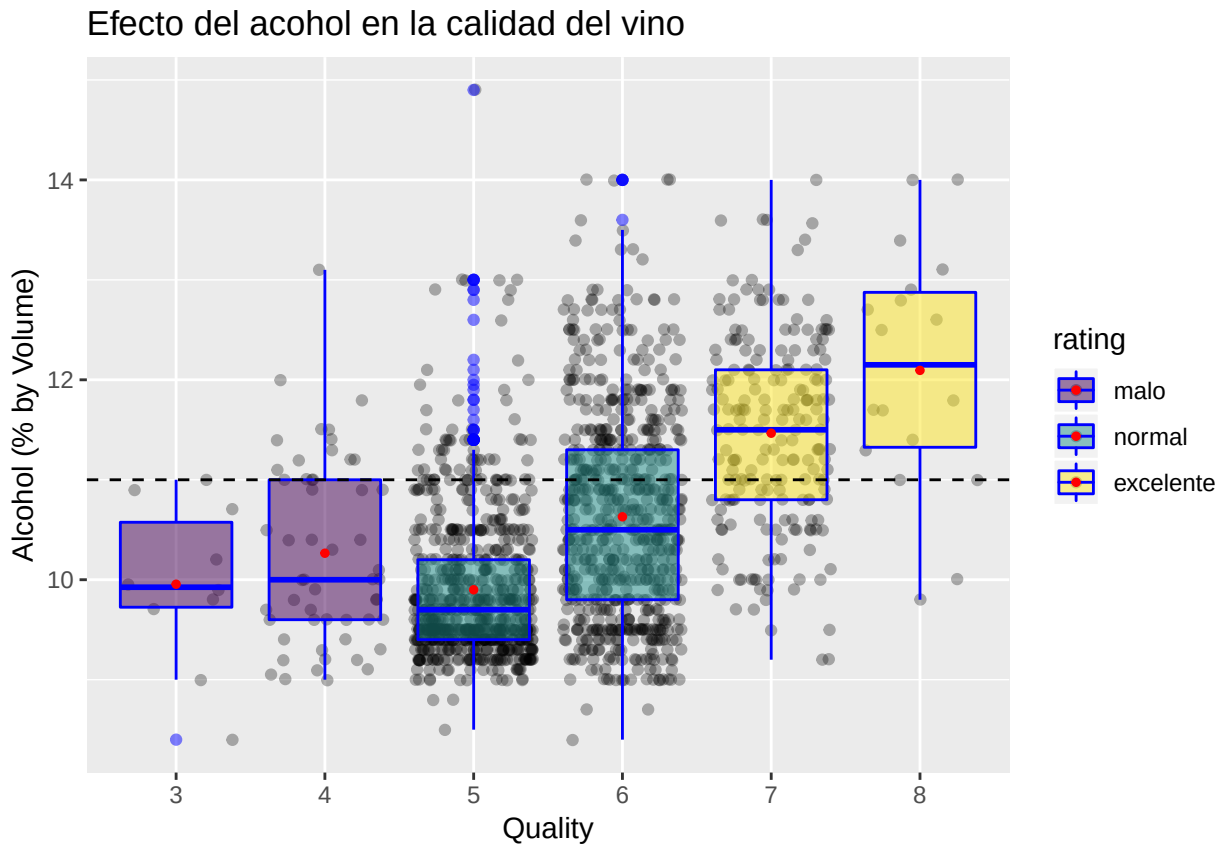
```
# Exportación de los datos limpios en .csv
write.csv(wine, "Wine_data_clean.csv")
```

5 Conclusiones

A través del estudio se ha concluido que altos niveles de alcohol en el vino contribuye a una buena calidad del vinos. Además añadiendo sulfatos o ácido cítrico contribuyen positivamente también a una buena calidad. Por el contrario, añadir acidos volátiles influyen negativamente.

6 Representación gráfica de los resutlados

A lo largo del presente trabajo se han realizado distintas represetaciones que han ayudado a identificar los resultados obtenidos. Sin embargo, el siguiente gráfico resume de una forma muy clara, como la variabel acohol influye en la calidad de los vinos portugueses.




```
##
##  Pearson's product-moment correlation
##
## data:  wine$alcohol and as.numeric(wine$quality)
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

```
## wine$rating: malo
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.60   10.00   10.22   11.00   13.10
## ------------------------------------------------------------
## wine$rating: normal
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.00   10.25   10.90   14.90
## ------------------------------------------------------------
## wine$rating: excelente
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.60   11.52   12.20   14.00
```