Práctica 2: Limpieza y validación de los datos

Lorenzo Mesa Morales

07/01/2019

Índice

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. Descripción del dataset

¿Por qué es importante y qué pregunta/problema pretende responder?

El conjunto de datos objeto de análisis se ha obtenido a partir de este enlace en Kaggle (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009) y está constituido por 12 características (columnas) que presentan 1599 muestras de vinos tintos (filas o registros) de la región del norte de Portugal "Vinho Verde".

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

- **fixed acidity**: Por acidez fija entendemos la suma, un valor de concentración generalmente expresado en gramos por litro, de todos los ácidos presentes en el vino que presentan una característica común, son poco volátiles. Si en el laboratorio sometemos el vino a una destilación estos ácidos no pasan al destilado ya que no se volatilizan, permanecen en el vino. Esta característica hace que su concentración sea fácilmente medible.
- **volatile acidity**: Durante los procesos bioquímicos de ambas fermentaciones, tanto alcohólica como maloláctica, la actividad microbiana genera otros ácidos que presentan una característica diferenciadora de los anteriores, son volátiles. Cuando en el laboratorio sometemos el vino a destilación estos ácidos se volatilizan y pasan al destilado. Esta característica hace que su concentración sea más difícil de medir que en el caso de los no volátiles. La suma de estos ácidos volátiles, también expresada en gramos por litro, se denomina acidez volátil. El principal ácido volátil del vino es el ácido acético, procedente de la oxidación del alcohol, aunque existen otros presentes en menores cantidades como el fórmico, el butírico y el propiónico. Estos ácidos volátiles, en determinadas concentraciones, provocan defectos en los vinos por lo que se procura mantenerlos en los niveles más bajos que sea posible durante la fermentación.
- **citric acid**: El ácido cítrico da frescura al vino, puede ser utilizado para la acidificación química de los vinos o por su acción estabilizante particularmente para limitar los riesgos de quiebras férricas o para el prelavado de placas filtrantes. El contenido máximo en los vinos puede estar sometido a límites reglamentarios.
- **residual sugar**: cantidad de azúcar que permanece después de la fermentación. No es común encontrar vinos con menos de un gramo por litro y aquellos que tienen más de 45 gramos por litros se consideran dulces.

- **chlorides**: cantidad de sal en el vino.
- **free sulfur dioxide**: el SO2 Libre se divide a su vez en tres estados posibles, dependientes directamente del pH: Molecular (es el que cumple la acción antiséptica, antimicrobiana, y cierta función antioxidante), Bisulfito (responsable de la acción antioxidásica. También capaz de formar sales ácidas) y Sulfito (su presencia es despreciable y su influencia mínima). Para poder plasmar la influencia del pH en el SO2 Libre, bastaría con citar que a un pH igual a 4,0, la concentración de SO2 Molecular, es diez veces menor que a un pH igual a 3,0, por lo tanto lo es también su acción.
- **total sulfur dioxide**: el SO2 Total es la suma de SO2 Libre + SO2 Combinado. Es de destacar, que con las técnicas reinantes de vinificación y el avance de los estudios, el uso del anhídrido sulfuroso es menor a lo que era varios años atrás, lográndose con dosis inferiores, resultados superiores. Y en lo que respecta a los distintos tipos de vinos, se utilizan en mayor cantidad en los dulces (a causa de la alta cantidad de azúcar remanente), seguidos por los blancos (por ser muy oxidables), y finalmente los que menor tasa de SO2 requieren son los tintos, por poseer antioxidantes propios (polifenoles).
- **density**: la densidad relativa a 20ºC o la densidad 20ºC/20ºC es la relación entre la masa volúmica de un vino o mosto y la del agua a la temperatura de 20ºC.
- **pH**: La determinación del pH en el mosto y el vino es una medida complementaria de la acidez total porque nos permite medir la fuerza de los ácidos que contienen. El pH usual de un vino puede variar entre 2,7 y 3,8 dependiendo si es blanco o tinto.
- **sulphates**: la aplicación del sulfuroso en enología se limita a usos antimicrobianos, actuando contra microbios (mohos, bacterias y levaduras negativas); y usos antioxidativos o antioxidásicos, actuando contra oxidasas que son enzimas que van en las uvas y que pueden deteriorar el color y el sabor del vino.
- **alcohol**: porcentaje de alcohol que contiene el vino.
- **quality**: variable de salida (basada en los datos, otorga una puntuación entre 0 y 10).

A partir de este conjunto de datos se plantea la problemática de determinar qué variables (propiedades fisicoquímicas) influyen más sobre la calidad de un vino.

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

Carga de los datos

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

```
# Lectura de datos
 datos <- read.csv( "winequality-red.csv")
 head(datos)
```

```
##   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
```

```
# Tipo de dato asignado a cada campo
 sapply( datos, class)
```

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

Observamos cómo los tipos de datos asignados automáticamente por R a las variables se corresponden con el dominio de estas y que todas las variables son cuantitativas.

```
# Para tener una primera idea, mostramos un resumen de cada una de las variables
 summary(datos)
```

```
##  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
```

Selección de los datos de interés

La gran mayoría de los atributos presentes en el conjunto de datos se corresponden con características que reúnen los diversos vinos recogidos en forma de registros, por lo que será conveniente tenerlos en consideración durante la realización de los análisis.

3. Limpieza de los datos.

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

Comúnmente, se utilizan los ceros como centinela para indicar la ausencia de ciertos valores. Vamos a proceder a conocer a continuación qué campos contienen elementos con valores ceros o elementos vacíos:

```
# Números de valores cero por campo
sapply(datos, function(x) sum(x==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
```

Como podemos observar solo el campo citric.acid tiene valores cero. Según podemos ver en el siguiente enlace(http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity) la concentración de ácido cítrico puede ser entre 0 y 500 mg/L, por lo tanto, podemos dar esos valores como buenos.

```
# Números de valores vacíos por campo
sapply(datos, 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
```

Como podemos observar no existen valores vacíos

En el caso en que para hubieramos encontrado casos de valores cero que no fueran admisibles para las variables o valores vacíos, podríamos haber empleado un método de imputación de valores basado en la similitud o diferencia entre los registros: la imputación basada en k vecinos más próximos (en inglés, kNN-imputation). La elección de esta alternativa se realiza bajo la hipótesis de que nuestros registros guardan cierta relación. No obstante, es mejor trabajar con datos "aproximados" que con los propios elementos vacíos, ya que obtendremos análisis con menor margen de error.

3.2. Identificación y tratamiento de valores extremos.

Los valores extremos o outliers son aquellos que parecen no ser congruentes si los comparamos con el resto de los datos. Para identificarlos, vamos a utilizar la función boxplots.stats(). Así, se mostrarán sólo los valores

atípicos para aquellas variables que los contienen:

```
# Identificación de outliers
boxplot.stats(datos$fixed.acidity)$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
```

```
boxplot.stats(datos$volatile.acidity)$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(datos$citric.acid)$out
```

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

```
boxplot.stats(datos$residual.sugar)$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(datos$chlorides)$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(datos$free.sulfur.dioxide)$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(datos$total.sulfur.dioxide)$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(datos$density)$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(datos$pH)$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(datos$sulphates)$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(datos$alcohol)$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(datos$quality)$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
```

Tras revisar los valores comprobamos que pueden darse perfectamente ya que se encuentran dentro de los rangos normales para cada uno de ellos. Es por ello que el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están recogidos.

4. Análisis de los datos.

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

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar. No obstante, como se verá en el apartado consistente en la realización de pruebas estadísticas, no todos se utilizarán.

```
# Agrupación por nivel de pH
datos$pHfact[datos$pH <= 3.4] <- "normal"
datos$pHfact[datos$pH > 3.4] <- "alto"

datos$pHfact <- as.factor(datos$pHfact)
```

```
# Agrupación por nivel de acidez fija
datos$fixacidfact[datos$fixed.acidity <= 9.2] <- "normal"
```

```r
datos$fixacidfact[datos$fixed.acidity > 9.2] <- "alto"

datos$fixacidfact <- as.factor(datos$fixacidfact)
```

```r
# Agrupación por nivel de azúcar residual
datos$resisugfact[datos$residual.sugar <= 2.6] <- "normal"
datos$resisugfact[datos$residual.sugar > 2.6] <- "alto"

datos$resisugfact <- as.factor(datos$resisugfact)
```

```r
# Exportación de los datos finales en .csv

write.csv(datos, "winequality-red_final.csv")
```

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

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, 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 significación prefijado alpha = 0, 05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

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

Seguidamente, pasamos a estudiar 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 los test que presentan un pH alto (>3.4) frente a un pH normal (<=3.4). Para ello utilizamos la variable pHfact que representa ambos grupos. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```r
fligner.test(quality ~ pHfact, data = datos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  quality by pHfact
## Fligner-Killeen:med chi-squared = 0.45087, df = 1, p-value =
```

```
## 0.5019
```

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

Procedemos igualmente con la acidez fija.

```
fligner.test(quality ~ fixacidfact, data = datos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  quality by fixacidfact
## Fligner-Killeen:med chi-squared = 1.7227, df = 1, p-value = 0.1894
```

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

Finalmente analizaremos el caso del azúcar residual.

```
fligner.test(quality ~ resisugfact, data = datos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  quality by resisugfact
## Fligner-Killeen:med chi-squared = 2.9088, df = 1, p-value = 0.0881
```

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

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

¿Qué variables cuantitativas influyen más 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 el precio final del vehículo. Para ello, se utilizará el coeficiente de correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
# Calcular el coeficiente de correlación para cada variable cuantitativa
# con respecto al campo "precio"
for (i in 1:(ncol(datos) - 4)) {
  if (is.integer(datos[,i]) | is.numeric(datos[,i])) {
    spearman_test = cor.test(datos[,i],
                             datos[,length(datos)-3],
                             method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value
    # Añadimos una fila a la matriz
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(datos)[i]
  }
}
```

```r
print(corr_matrix)
```

```
##                         estimate      p-value
## fixed.acidity         0.11408367 4.801220e-06
## volatile.acidity     -0.38064651 2.734944e-56
## citric.acid           0.21348091 6.158952e-18
## residual.sugar        0.03204817 2.002454e-01
## chlorides            -0.18992234 1.882858e-14
## free.sulfur.dioxide  -0.05690065 2.288322e-02
## total.sulfur.dioxide -0.19673508 2.046488e-15
## density              -0.17707407 9.918139e-13
## pH                   -0.04367193 8.084594e-02
## sulphates             0.37706020 3.477695e-55
## alcohol               0.47853169 2.726838e-92
```

Así, identificamos cuáles son las variables más correlacionadas con la calidad en función de su proximidad con los valores -1 y +1. Teniendo esto en cuenta, queda patente que no existe ninguna variable relevante, la que más se aproxima a los valores -1 y +1 es alcohol pero se queda lejos.

Nota. Para cada coeficiente de correlación se muestra también su p-valor asociado, puesto que éste puede dar información acerca del peso estadístico de la correlación obtenida.

¿La calidad del vino es mayor en caso de tener un pH alto? La segunda prueba estadística que se aplicará consistirá en un contraste de hipótesis sobre dos muestras para determinar si la calidad del vino es superior dependiendo del nivel de pH (normal o alto). Para ello, tendremos dos muestras: la primera de ellas se corresponderá a la calidad de las muestras con pH normal y, la segunda, con aquellas que presentan un pH alto. Se debe destacar que un test paramétrico como el que a continuación se utiliza necesita que los datos sean normales, si la muestra es de tamaño inferior a 30. Como en nuestro caso, n > 30, el contraste de hipótesis siguiente es válido.

```r
# Agrupación por nivel de pH
datos.pHnormal.calidad <- datos[datos$pHfact == "normal",]$quality
datos.pHalto.calidad <- datos[datos$pHfact == "alto",]$quality
```

Así, se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias, el cual es unilateral atendiendo a la formulación de la hipótesis alternativa:

H0 : u1 ??? u2 = 0 H1 : u1 ??? u2 < 0

donde u1 es la media de la población de la que se extrae la primera muestra y u2 es la media de la población de la que extrae la segunda. Así, tomaremos alpha = 0, 05.

```r
t.test(datos.pHnormal.calidad,datos.pHalto.calidad, alternative="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  datos.pHnormal.calidad and datos.pHalto.calidad
## t = 2.0257, df = 645.11, p-value = 0.9784
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.1739411
## sample estimates:
## mean of x mean of y
##  5.659241  5.563307
```

Puesto que no hemos obtenido un p-valor menor que el valor de significación fijado, aceptamos la hipótesis nula. Por tanto, podemos concluir que la calidad del vino no es mayor si el pH es alto.

Modelo de regresión lineal

Tal y como se planteó en los objetivos de la actividad, resultará de mucho interés poder realizar predicciones sobre el calidad de las muestras dadas sus características. Así, se calculará un modelo de regresión lineal utilizando regresores cuantitativos con el que poder realizar las predicciones de la calidad. Para obtener un modelo de regresión lineal considerablemente eficiente, lo que haremos será obtener varios modelos de regresión utilizando las variables que estén más correladas con respecto a la calidad, según la tabla obtenida anteriormente. Así, de entre todos los modelos que tengamos, escogeremos el mejor utilizando como criterio aquel que presente un mayor coeficiente de determinación (R2).

```
# Regresores cuantitativos con mayor coeficiente
# de correlación con respecto a la calidad
alcohol = datos$alcohol
acido.volatil = datos$volatile.acidity
sulfuroso = datos$sulphates
acido.citrico = datos$citric.acid
so2.total = datos$total.sulfur.dioxide
sal = datos$chlorides
densidad = datos$density

# Variable a predecir
calidad = datos$quality

# Generación de varios modelos
modelo1 <- lm(calidad ~ alcohol + acido.volatil + sulfuroso +
                  acido.citrico + so2.total + sal + densidad, data = datos)
modelo2 <- lm(calidad ~ alcohol + acido.volatil + sulfuroso +
                  acido.citrico + so2.total, data = datos)
modelo3 <- lm(calidad ~ alcohol + sulfuroso + sal + densidad +
                  so2.total, data = datos)
modelo4 <- lm(calidad ~ acido.citrico + acido.volatil + sulfuroso +
                  so2.total, data = datos)
modelo5 <- lm(calidad ~ alcohol + so2.total + sal + densidad, data = datos)
```

Para los anteriores modelos de regresión lineal múltiple obtenidos, podemos utilizar el coeficiente de determinación para medir la bondad de los ajustes y quedarnos con aquel modelo que mejor coeficiente presente.

```
# Tabla con los coeficientes de determinación de cada modelo
tabla.coeficientes <- matrix(c(1, summary(modelo1)$r.squared,
                           2, summary(modelo2)$r.squared,
                           3, summary(modelo3)$r.squared,
                           4, summary(modelo4)$r.squared,
                           5, summary(modelo5)$r.squared),
                           ncol = 2, byrow = TRUE)
colnames(tabla.coeficientes) <- c("Modelo", "R^2")
tabla.coeficientes
```

```
##      Modelo       R^2
## [1,]      1 0.3516932
## [2,]      2 0.3438525
## [3,]      3 0.2954556
## [4,]      4 0.2041627
## [5,]      5 0.2402121
```

En este caso, tenemos que el primer modelo es el más conveniente dado que tiene un mayor coeficiente de determinación. Ahora, empleando este modelo, podemos proceder a realizar predicciones de calidad de

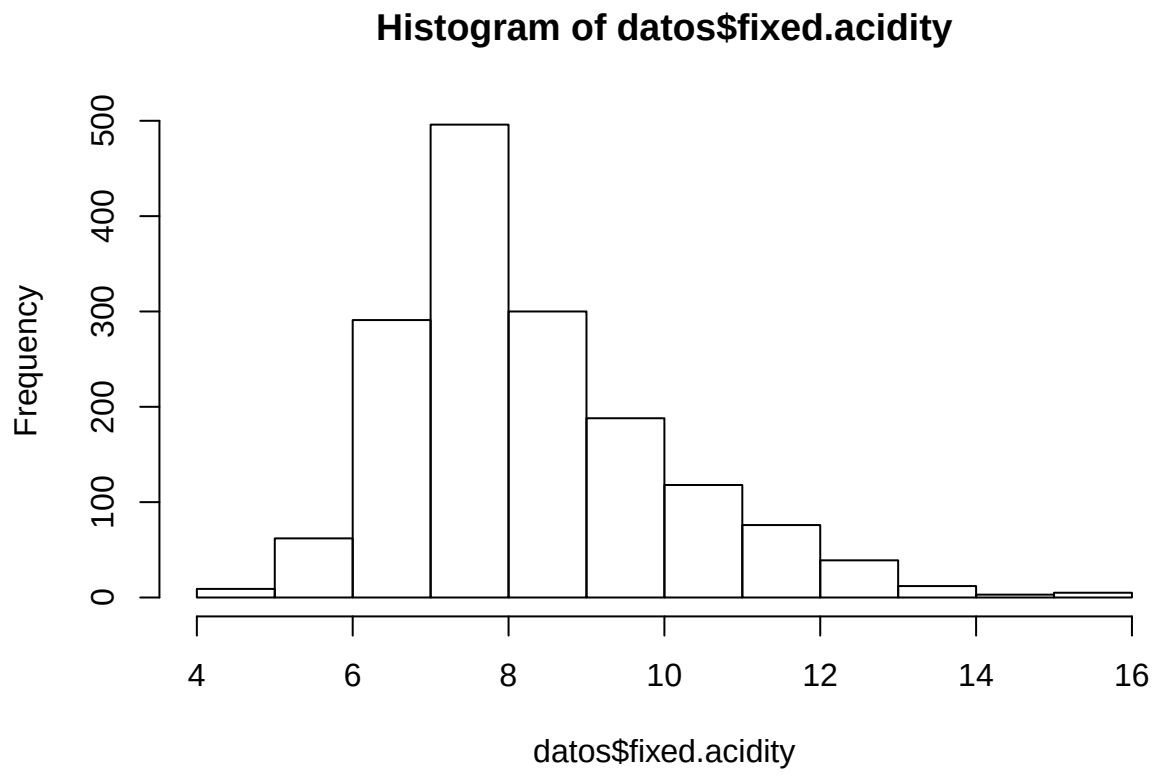
muestras como la siguiente:

```
newdata <- data.frame(
  alcohol = 9,
  acido.volatil = 0.54,
  sulfuroso = 0.59,
  acido.citrico = 0.18,
  so2.total = 35,
  sal = 0.08,
  densidad = 0.9972
)

# Predecir la calidad
predict(modelo1, newdata)
```

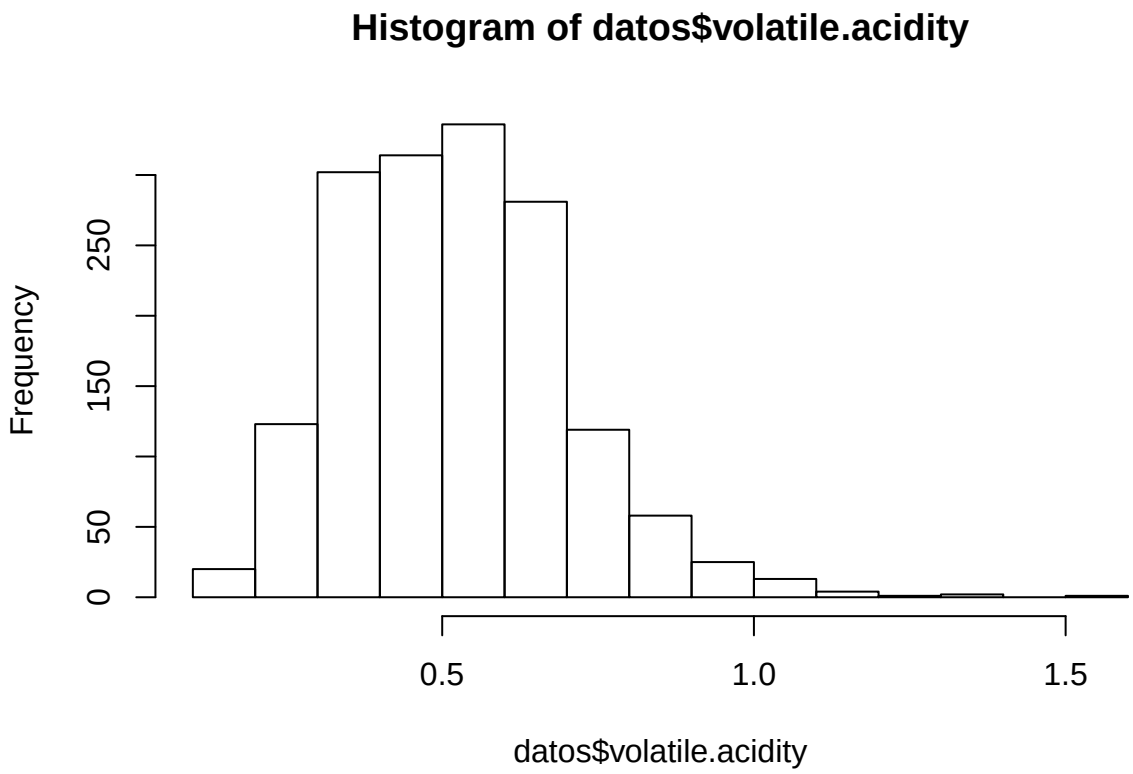
```
##        1
## 5.199482
```

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

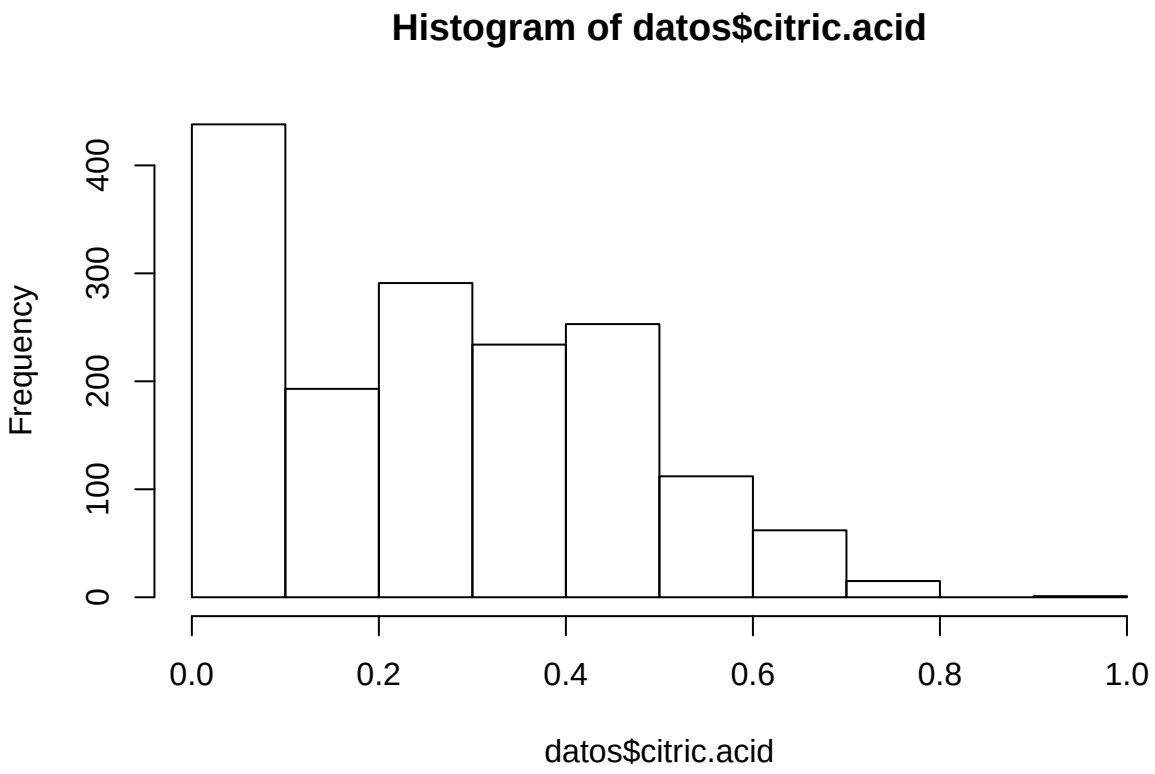
```
# Histograma de cada una de las variables
hist(datos$fixed.acidity)
```



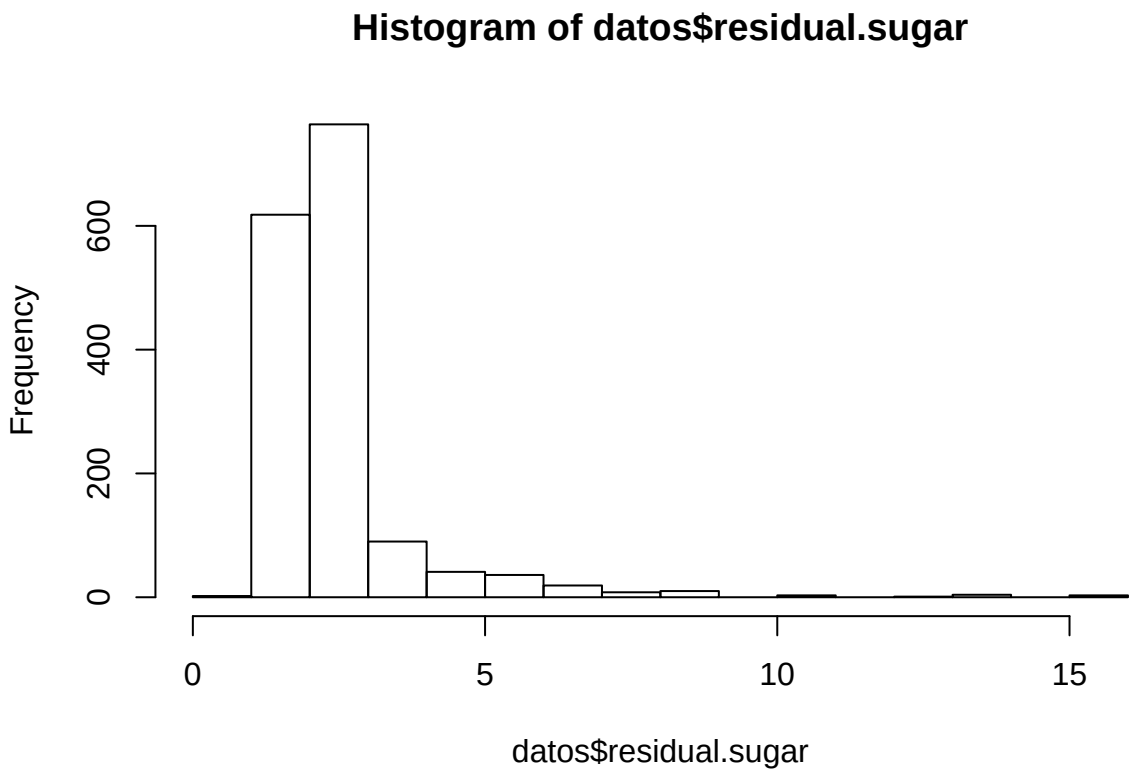

```
hist(datos$volatile.acidity)
```



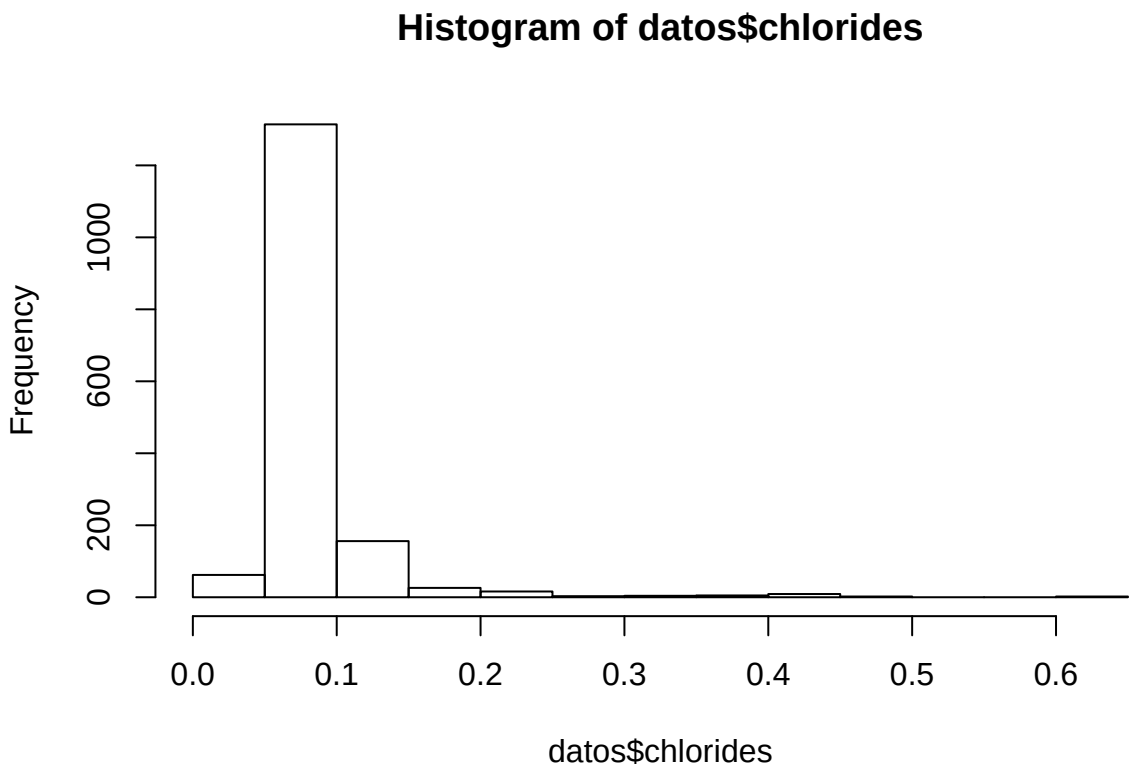

```
hist(datos$citric.acid)
```



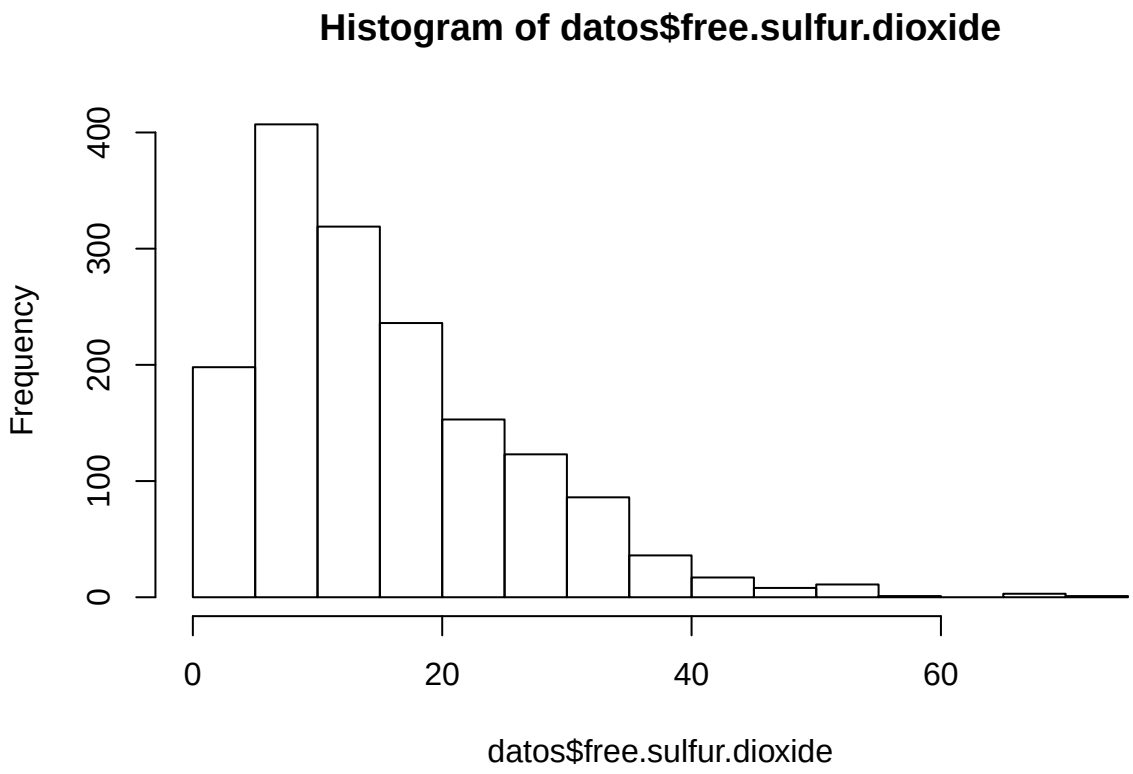

```r
hist(datos$residual.sugar)
```



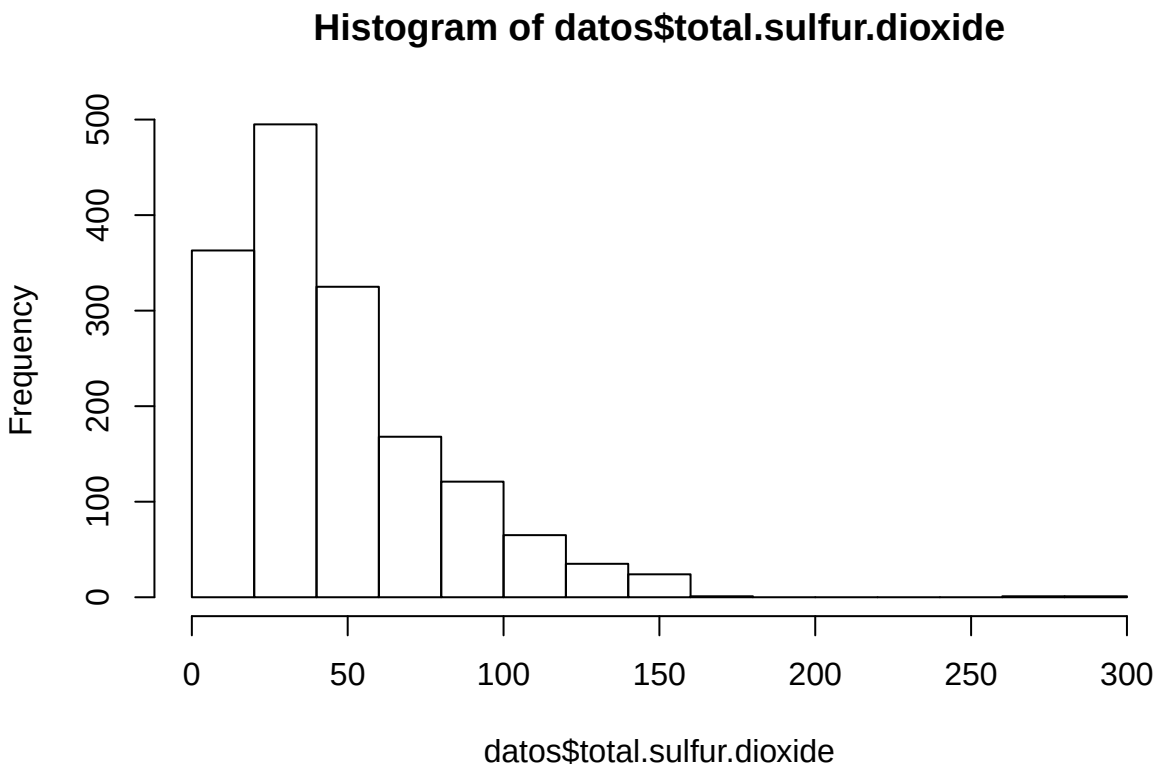

```
hist(datos$chlorides)
```



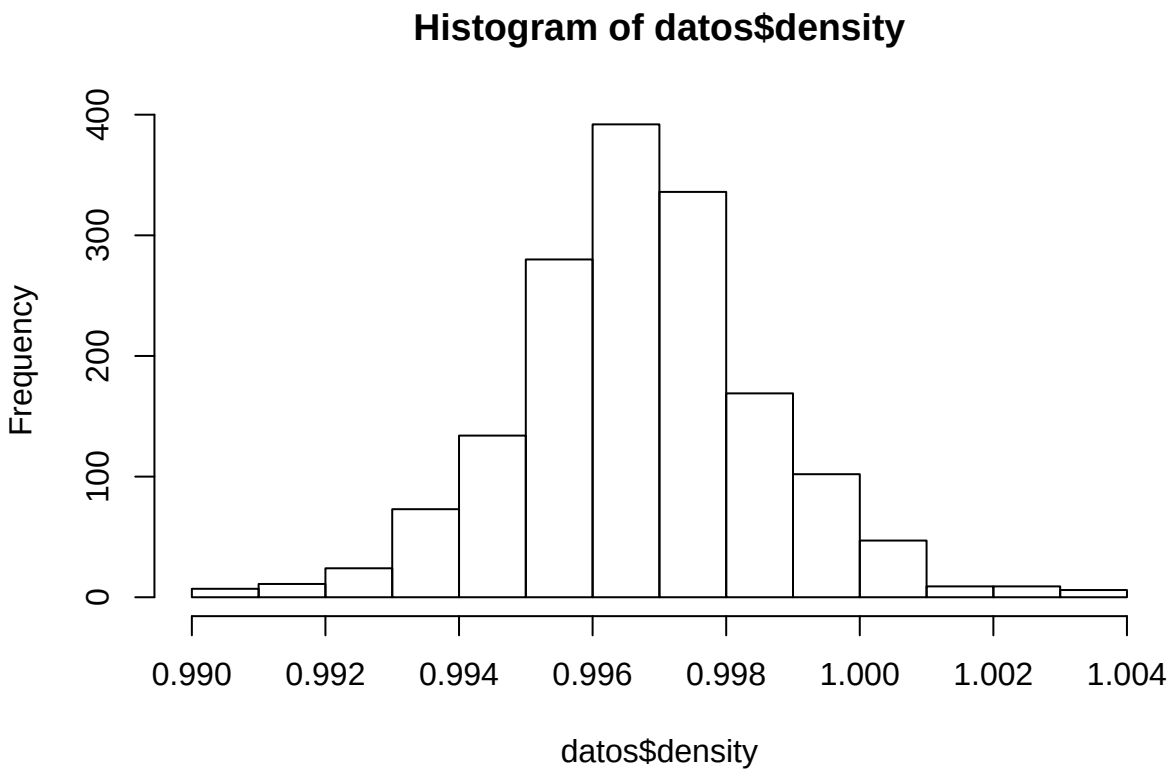

```r
hist(datos$free.sulfur.dioxide)
```



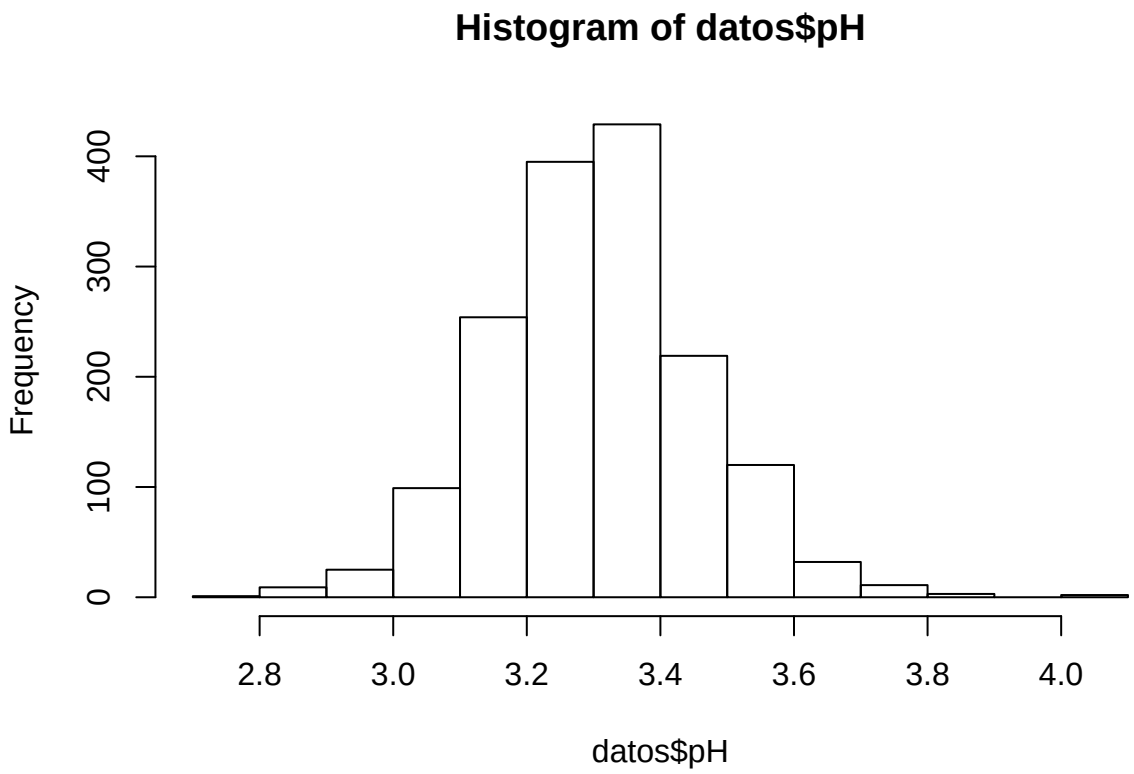

```r
hist(datos$total.sulfur.dioxide)
```



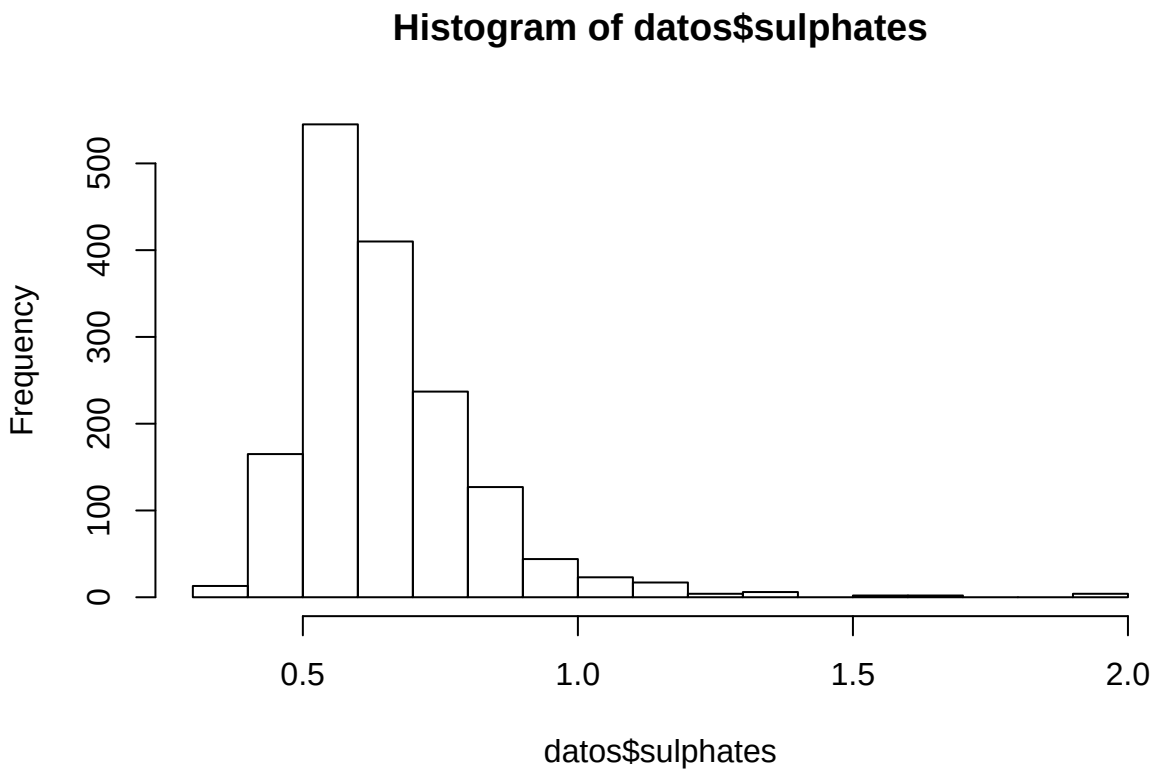

```r
hist(datos$density)
```



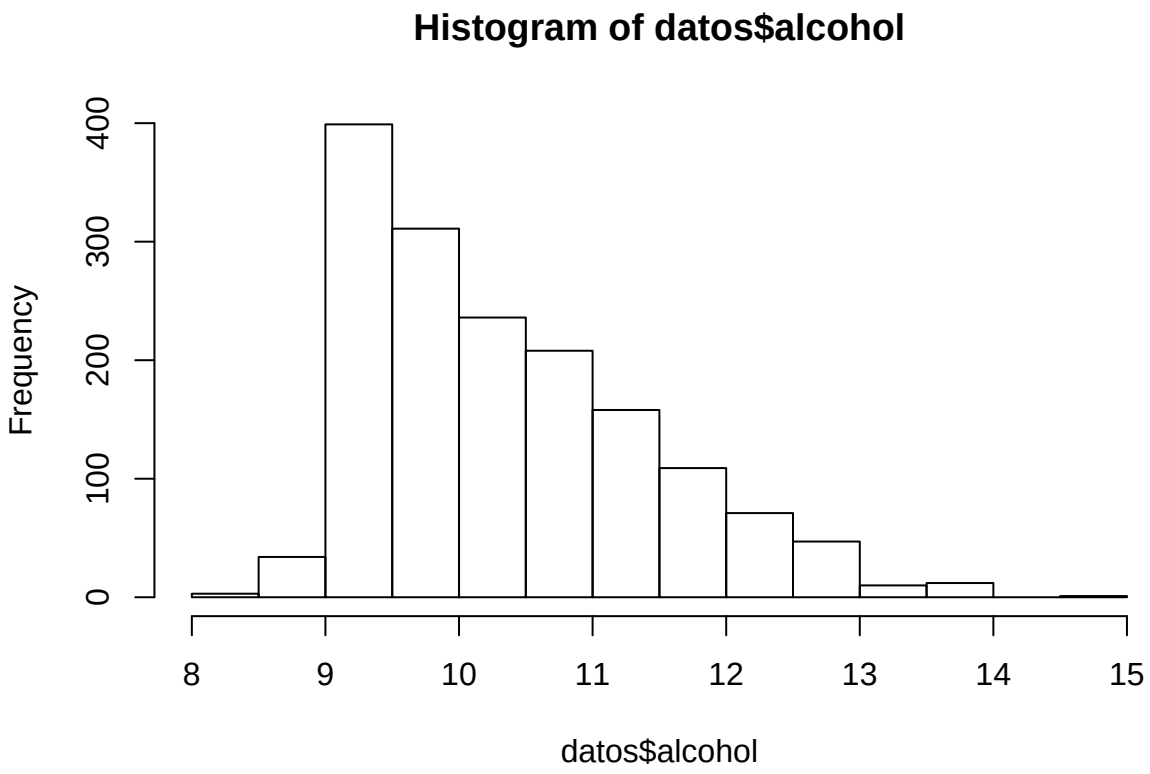

```
hist(datos$pH)
```



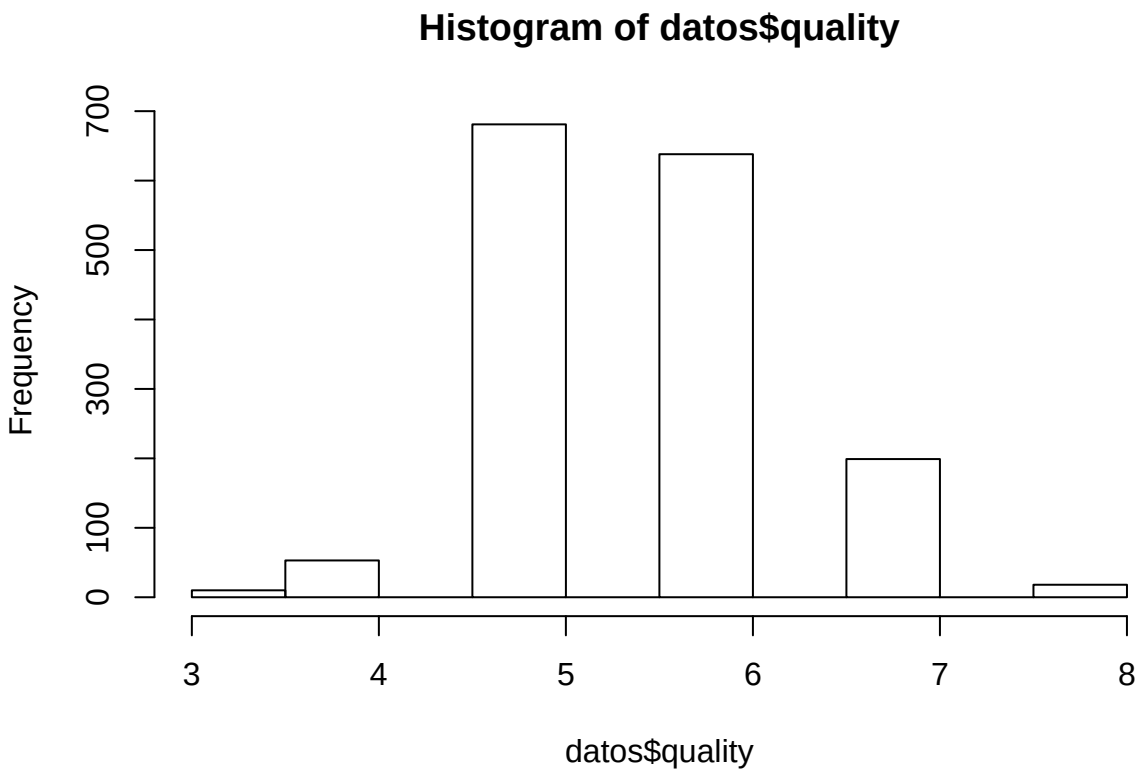

```r
hist(datos$sulphates)
```



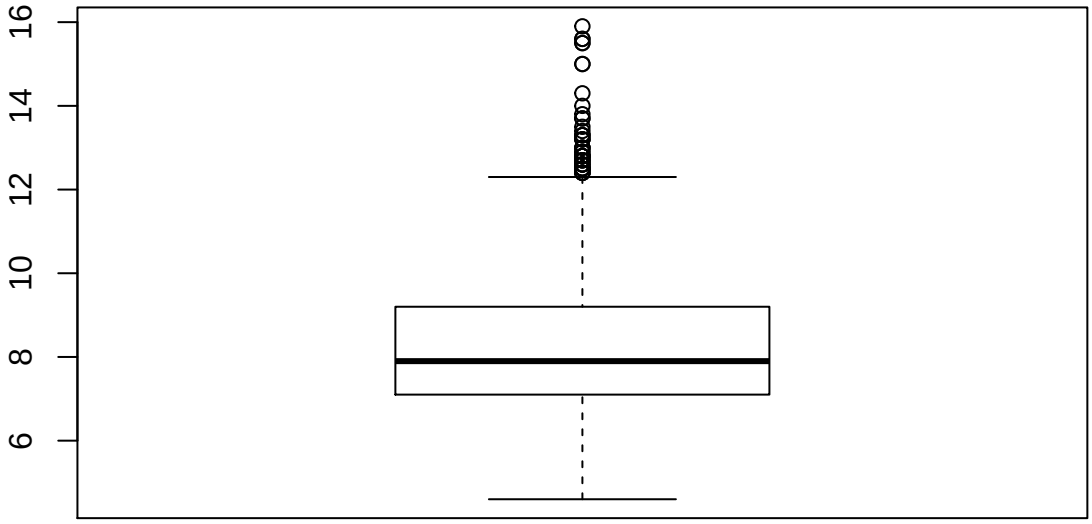

```
hist(datos$alcohol)
```



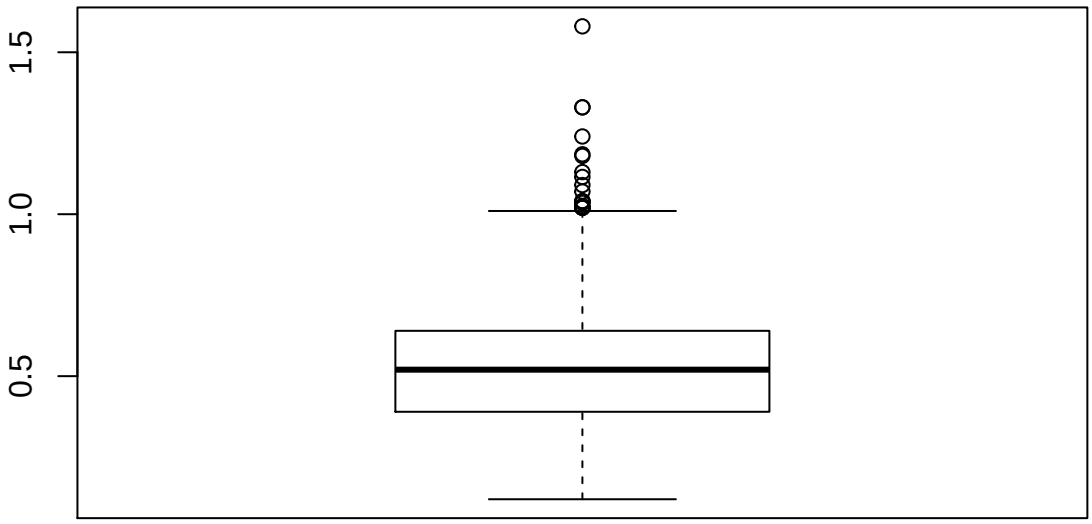

```
hist(datos$quality)
```



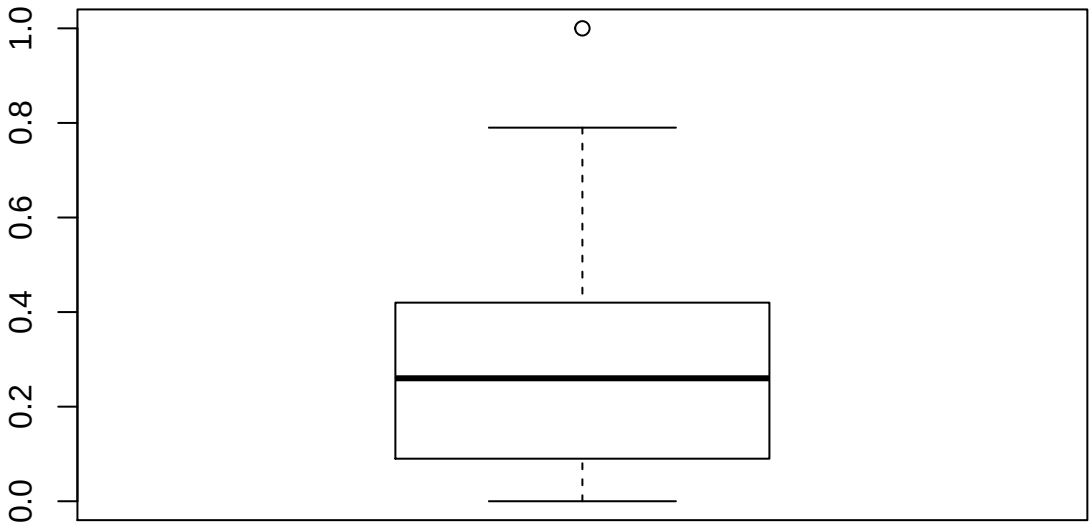

```r
# Boxplot con la representación de los outliers
boxplot(datos$fixed.acidity)
```



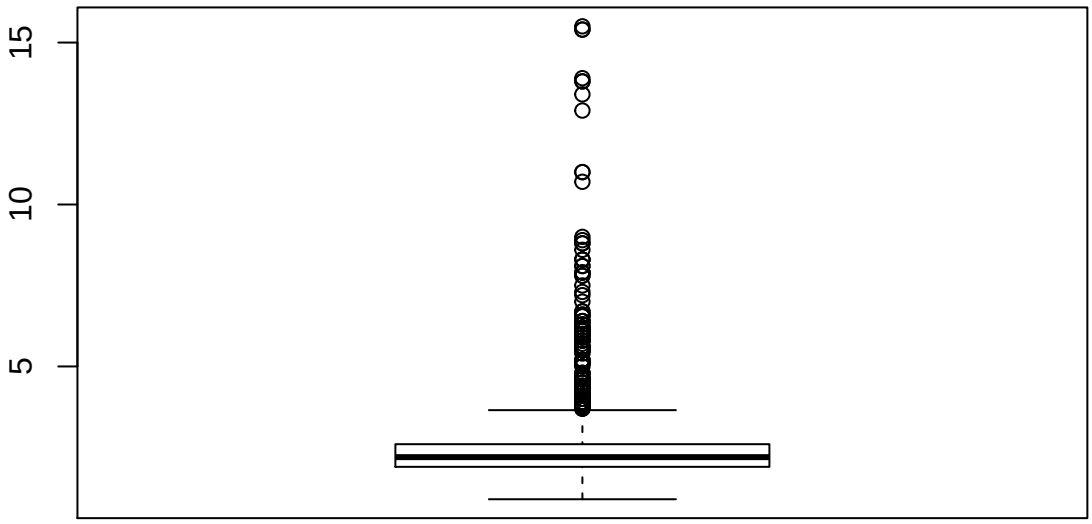

```
boxplot(datos$volatile.acidity)
```



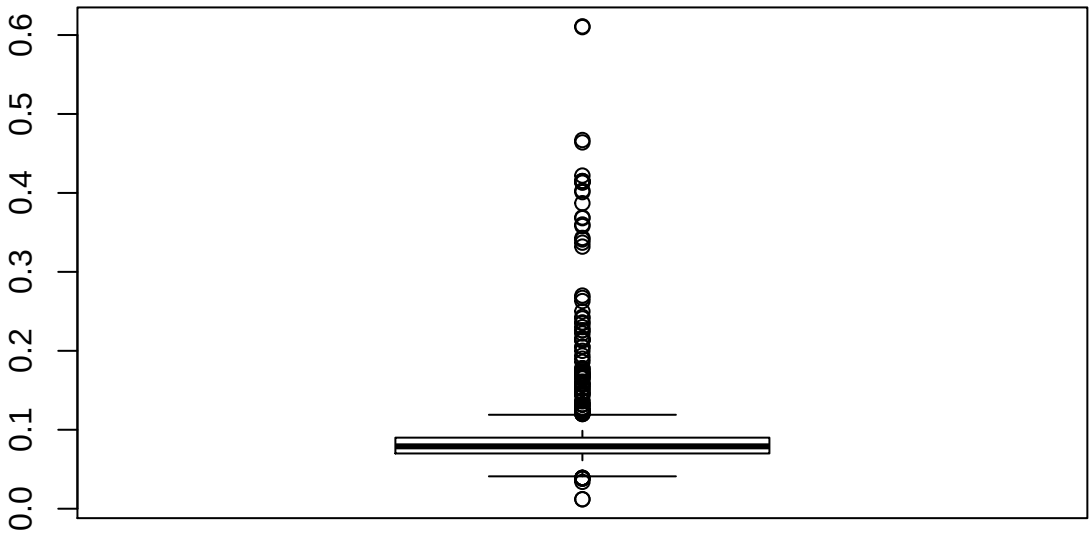

```
boxplot(datos$citric.acid)
```



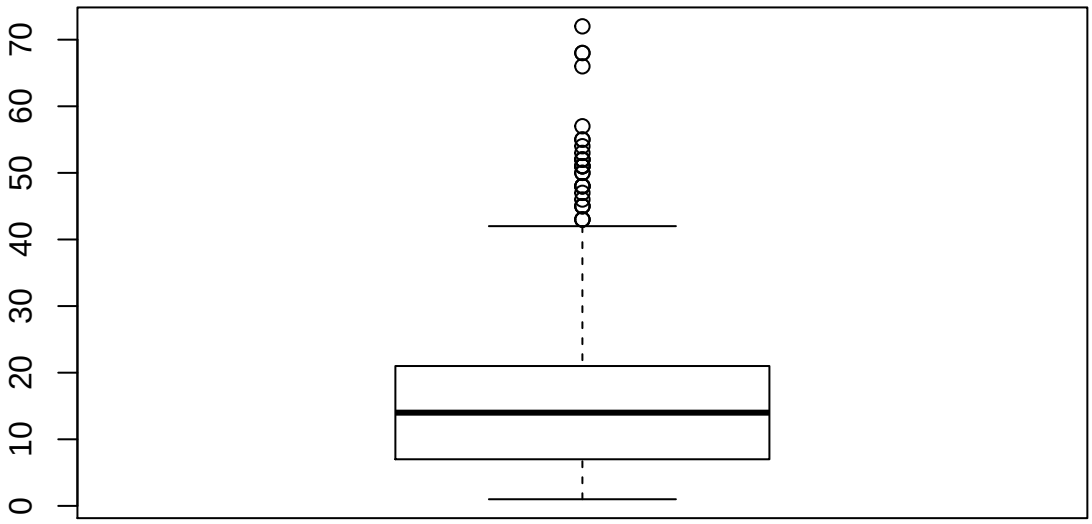

```
boxplot(datos$residual.sugar)
```



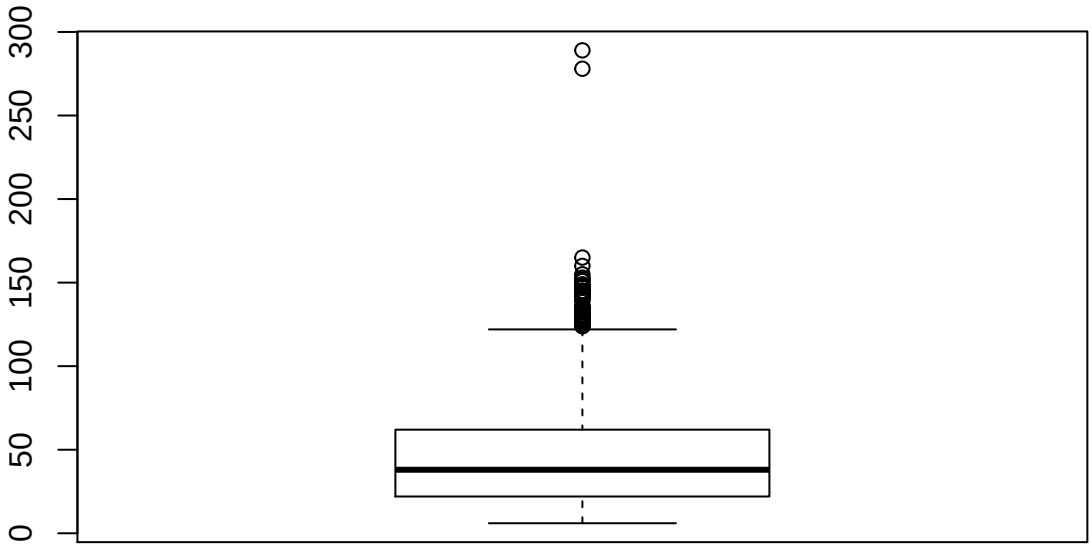

```
boxplot(datos$chlorides)
```



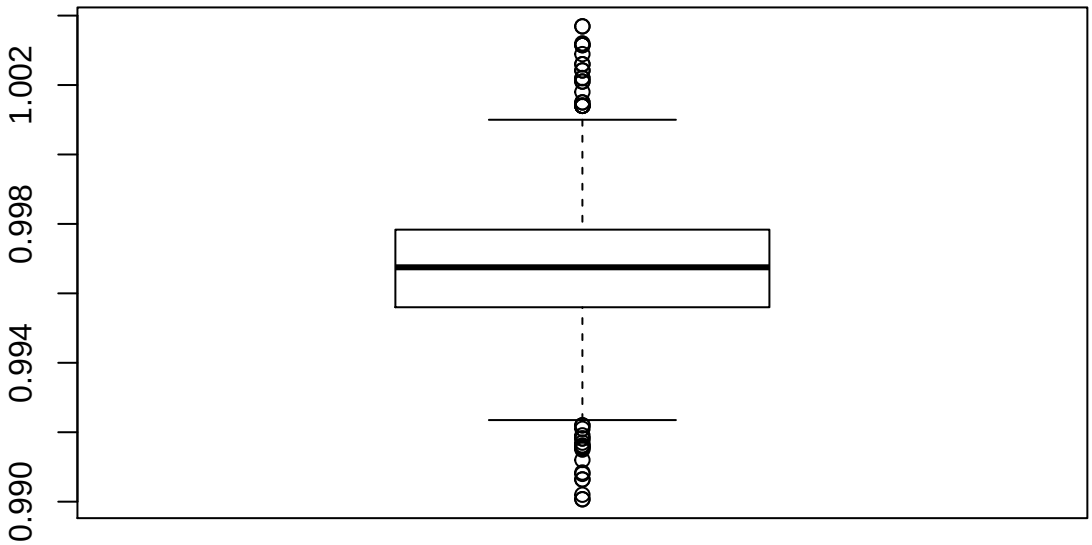

```
boxplot(datos$free.sulfur.dioxide)
```



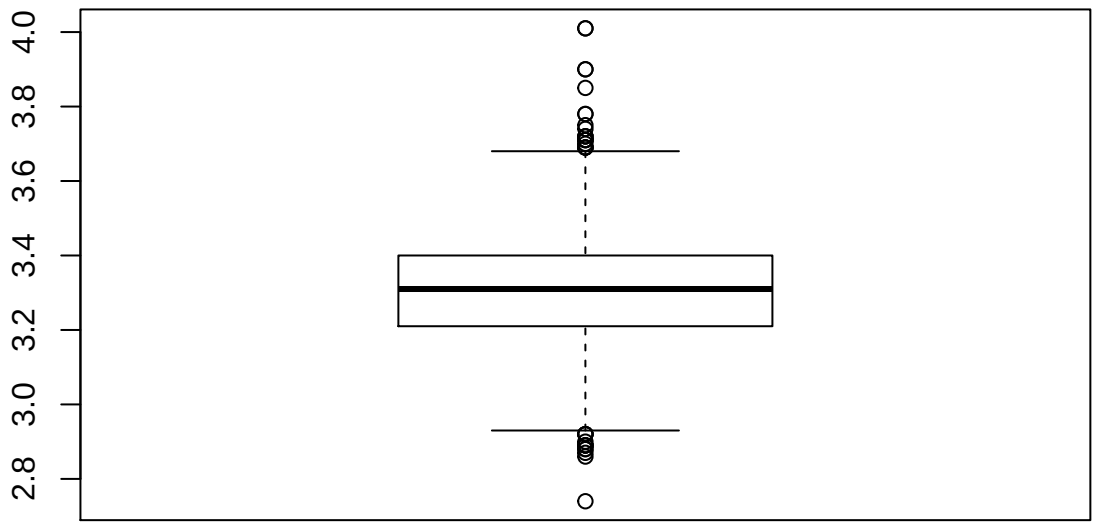

```
boxplot(datos$total.sulfur.dioxide)
```



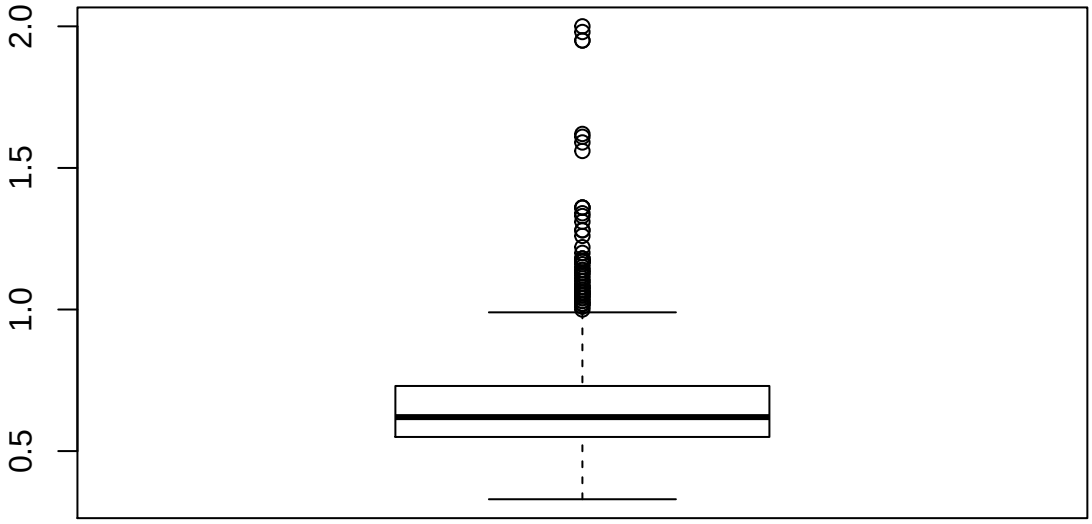

```
boxplot(datos$density)
```



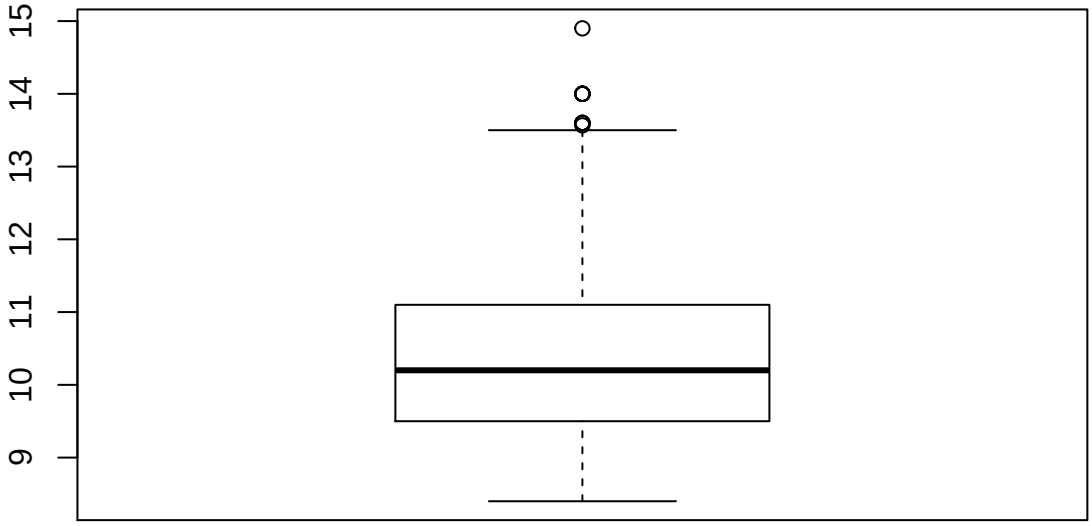

```
boxplot(datos$pH)
```



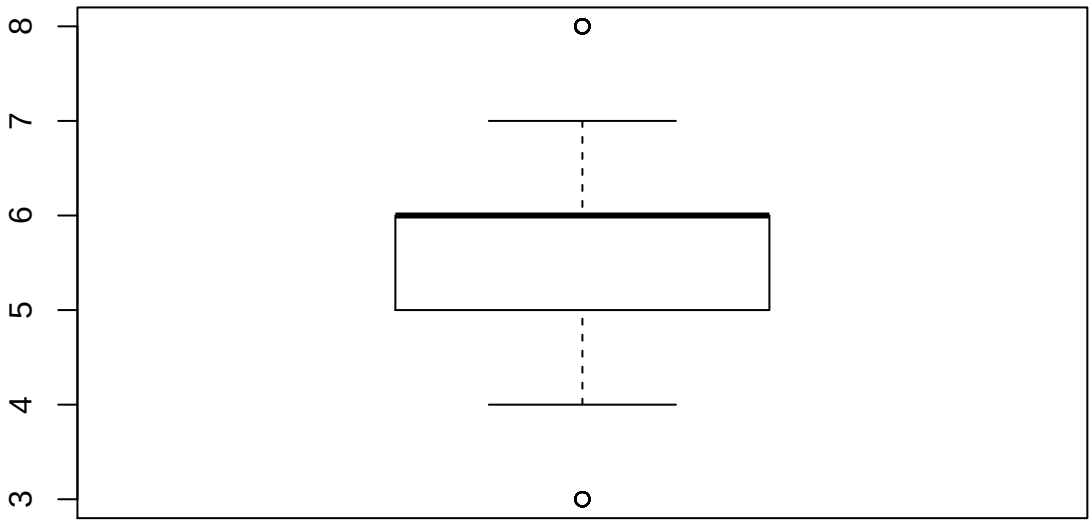

```
boxplot(datos$sulphates)
```



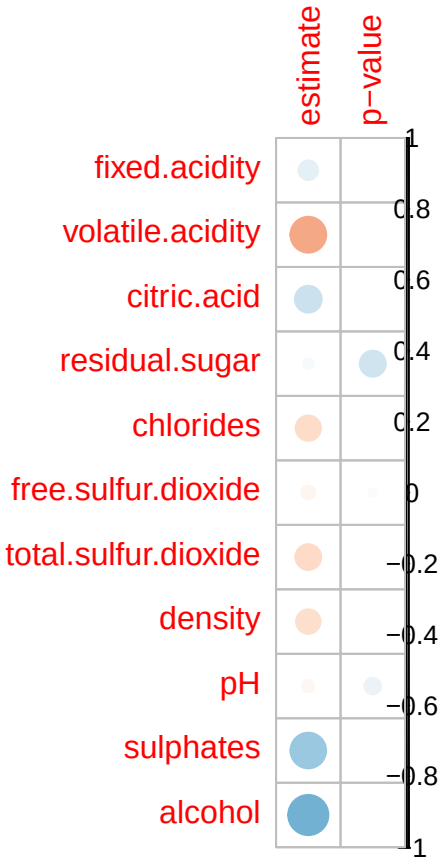

```
boxplot(datos$alcohol)
```



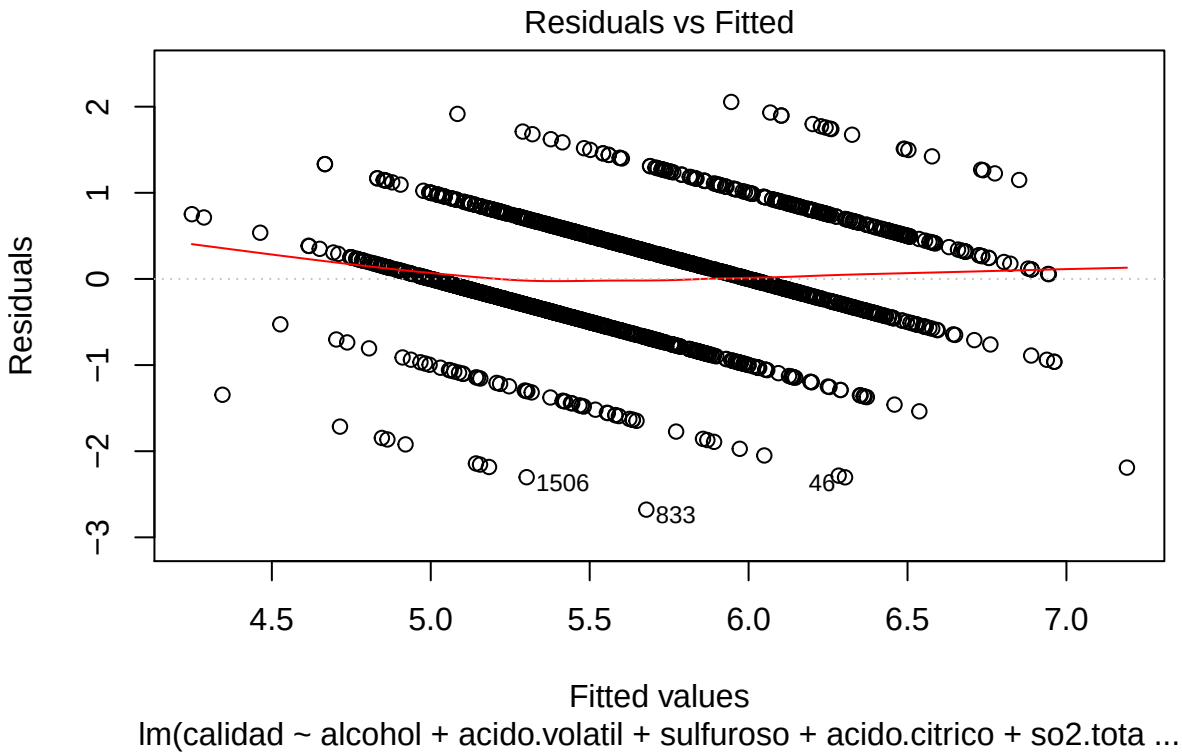

```
boxplot(datos$quality)
```

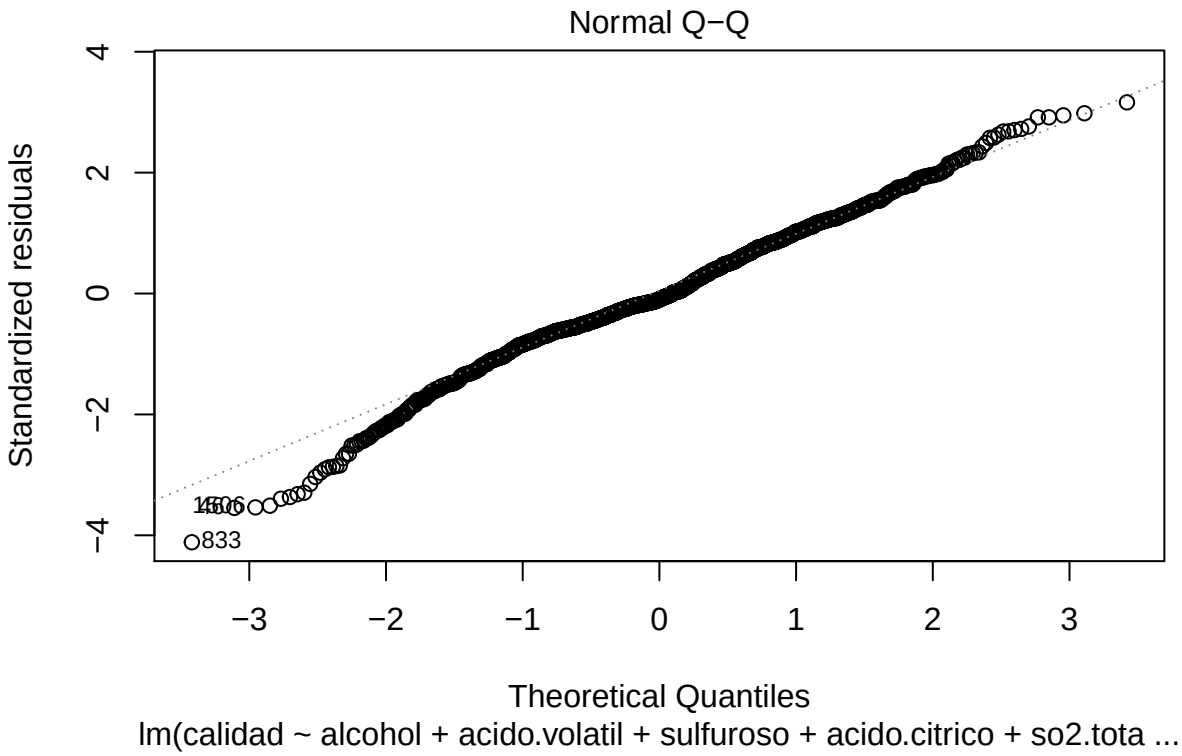


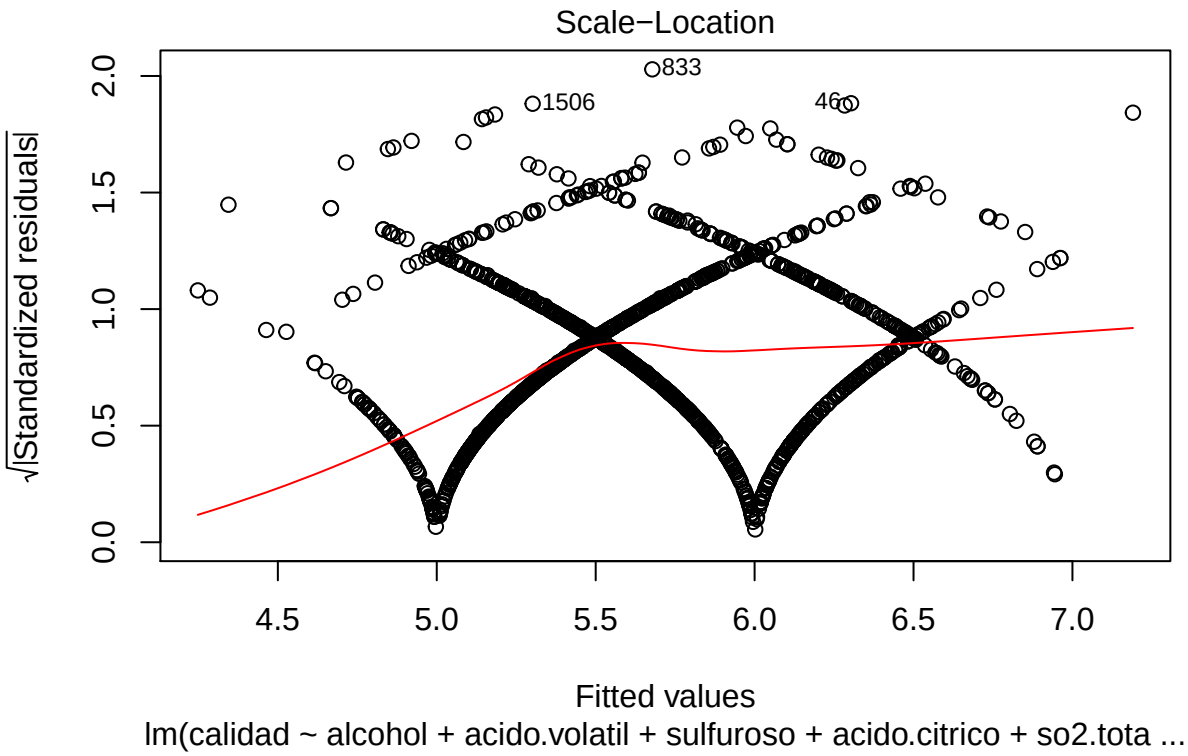

```
# Representación del matriz de correlación

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

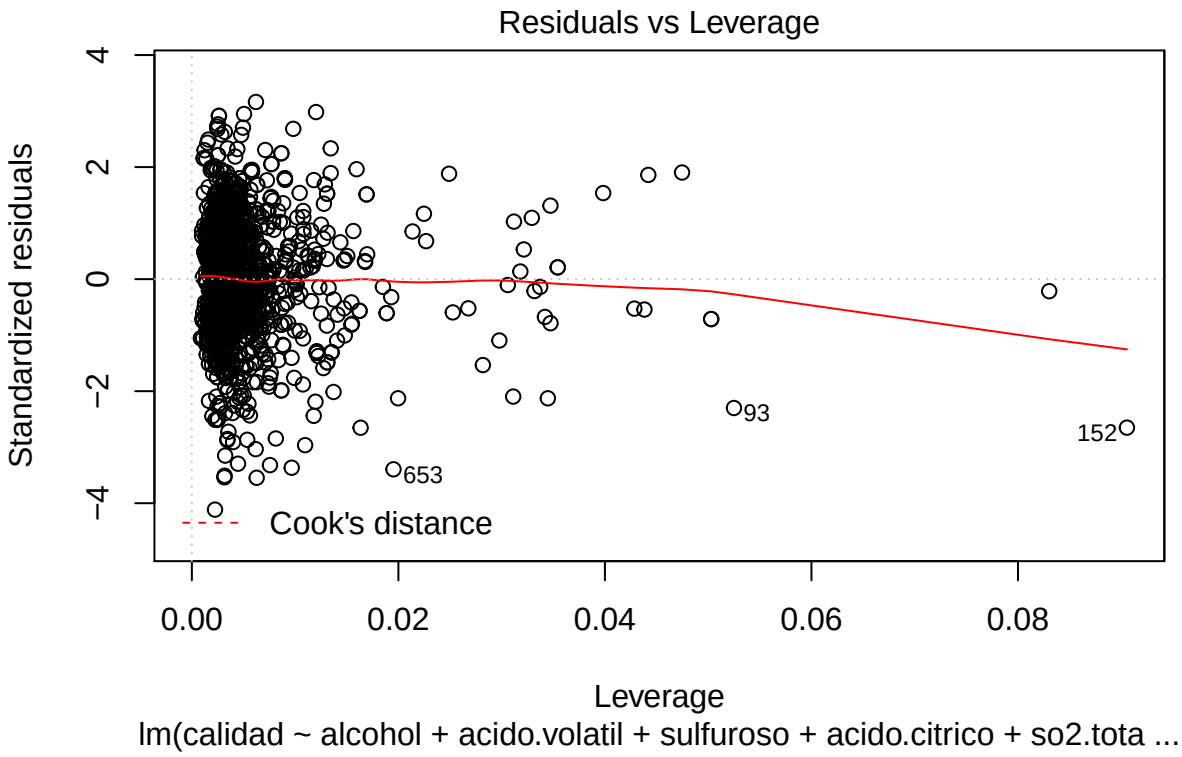



```
# Represtación del modelo 1

plot(modelo1)
```


Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
1506
833
46
4.5
5.0
5.5
6.0
6.5
7.0
Fitted values
lm(calidad ~ alcohol + acido.volatil + sulfuroso + acido.citrico + so2.tota ...


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(calidad ~ alcohol + acido.volatil + sulfuroso + acido.citrico + so2.tota ...
833
4 2 0 −2 −4
−3 −2 −1 0 1 2 3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
833
1506
46
4.5
5.0
5.5
6.0
6.5
7.0
Fitted values
lm(calidad ~ alcohol + acido.volatil + sulfuroso + acido.citrico + so2.tota ...



6. Resolución del problema.

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

Como se ha visto, se han realizado tres tipos de pruebas estadísticas sobre un conjunto de datos que se correspondía con diferentes variables relativas a test de muestras de 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.

Así, 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, mientras que el modelo de regresión lineal obtenido resulta de utilidad a la hora de realizar predicciones para esta variable dadas unas características concretas.

Previamente, se han sometido los datos a un preprocesamiento para manejar los casos de ceros o elementos vacíos y valores extremos (outliers). Para el caso del primero, se ha hecho uso de un método de imputación de valores de tal forma que no tengamos que eliminar registros del conjunto de datos inicial y que la ausencia de valores no implique llegar a resultados poco certeros en los análisis. Para el caso del segundo, el cual constituye un punto delicado a tratar, 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 test sobre muestras que se realizan normalmente.

7. Código

El código está disponible en el siguiente enlace de GitHub https://github.com/lmesamo/data-analysis

8. Referencias

Vinos diferentes - Acidez del vino

Blog Urbina Vinos - Práctica: Determinación de la Ácidez Total y pH de un Vino o Mosto

Infoagro - Métodos oficiales de análisis de vinos

Vinopack - Los 6 criterios que determinan la calidad del vino

Vinetur - Las propiedades del anhídrido sulfuroso en la elaboración del vino

Blog Bodegas Comenge - El SO2 en los vinos

El vino y su análisis - Departamento de Nutrición y Bromatología II - Facultad de Farmacia - Universidad Complutense de Madrid

Squire, Megan (2015). Clean Data. Packt Publishing Ltd.

Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.

Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp.1527-3369.

Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.

Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.