Red Wine Quality

Descripción del dataset

Este Dataset contiene los datos fisicoquimicos y sensoriales correspondientes a la variación roja del vino portugués "Vinho Verde". Las variables son las siguientes:

1. fixed acidity: ácidos no volatiles que no se evaporan fácilmente.
2. volatile acidity: indica el contenido de ácido acético en el vino, un alto contenido produce un desagradable sabor a vinagre.
3. citric acid: actúa como conservante para aumentar la acidez. En pequeñas cantidades aporta frescura y sabor al vino.
4. residual sugar: es la cantidad de azucar restante despues de la fermantación.Los vinos con más de 45g/l son dulces.
5. chlorides: La cantidad de sal que tiene el vino.
6. free sulfur dioxide: previene el crecimiento de microbios y la oxidación del vino.
7. total sulfur dioxide: Es la cantidad total (libres + ligadas) de SO2 del vino.
8. density: densidad del vino. Los vinos dulces tienen una mayor densidad.
9. pH: el nivel de acidez.
10. sulphates: un aditivo del vino que contribuye a los niveles de SO2 y actia como antimicrobiano y antioxidante.
11. alcohol: la cantidad de alcohol que contiene el vino.
12. quality: calidad del vino (puntuación entre 0 y 10).

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

El objetivo de este análisis es observar la relación entre las variables fisicoquímicas y la calidad del vino y ver cual de estas variables tiene más peso a la hora de determinar la calidad. También se crearán modelos de regresión que permitan predecir la calidad a partir del resto de variables.

Este estudio puede ser de gran utilidad en cualquier sector relacionado con el vino, especialmente en bodegas, que pueden usar estos datos para mejorar la calidad de sus vinos.

Integración y selección de los datos de interés a analizar.

Para este análisis trabajaremos con el dataset winequality-red disponible en la siguiente URL: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Lo primero que haremos es cargar el dataset:

```
wineData <- read.csv('winequality-red.csv', sep=",")
head(wineData)
```

```
##   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
```

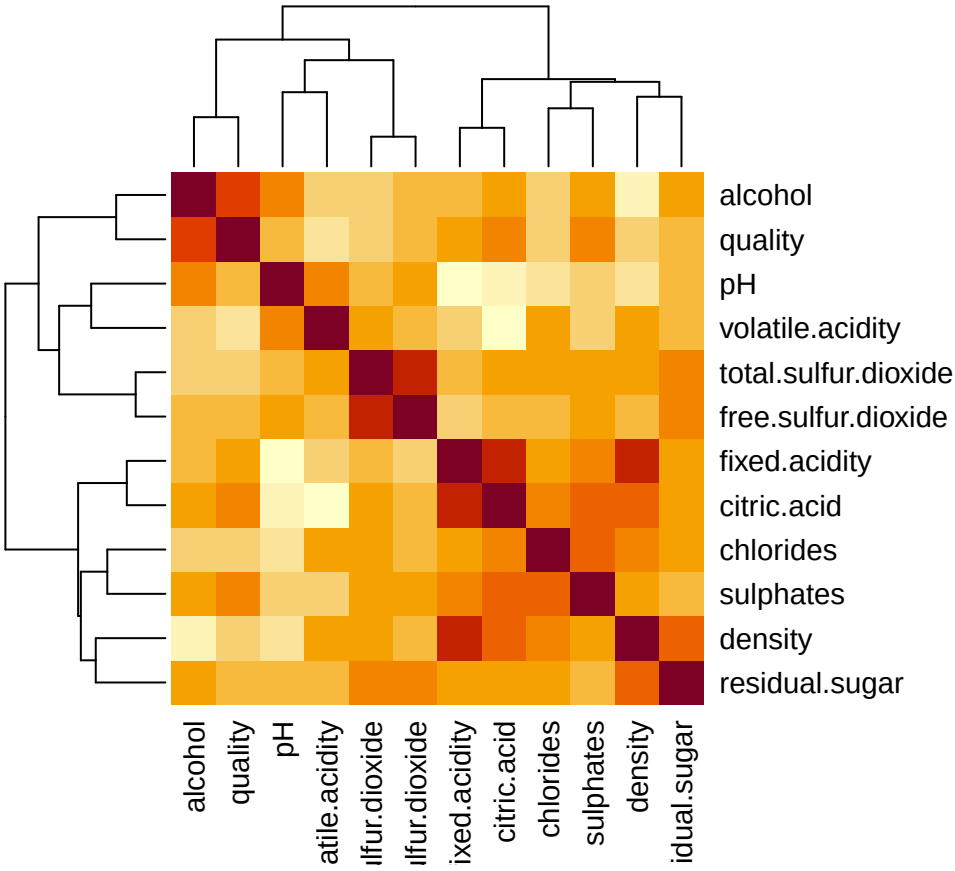
Observamos la correlación entre las diferentes variables para decidir si podemos prescindir de alguna variable redundante.

```
cor(wineData)
```

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity           1.00000000     -0.256130895  0.67170343    0.114776724
## volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882
## citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162
## residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000
## chlorides               0.09370519      0.061297772  0.20382291    0.055609535
## free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995
## total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882
## density                 0.66804729      0.022026232  0.36494718    0.355283371
## pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422
## sulphates               0.18300566     -0.260986685  0.31277004    0.005527121
## alcohol                -0.06166827     -0.202288027  0.10990325    0.042075437
## quality                 0.12405165     -0.390557780  0.22637251    0.013731637
##                         chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.093705186        -0.153794193          -0.11318144
## volatile.acidity      0.061297772        -0.010503827           0.07647000
## citric.acid           0.203822914        -0.060978129           0.03553302
## residual.sugar        0.055609535         0.187048995           0.20302788
## chlorides             1.000000000         0.005562147           0.04740047
## free.sulfur.dioxide   0.005562147         1.000000000           0.66766645
## total.sulfur.dioxide  0.047400468         0.667666450           1.00000000
## density               0.200632327        -0.021945831           0.07126948
## pH                   -0.265026131         0.070377499          -0.06649456
## sulphates             0.371260481         0.051657572           0.04294684
## alcohol              -0.221140545        -0.069408354          -0.20565394
## quality              -0.128906560        -0.050656057          -0.18510029
##                           density          pH    sulphates     alcohol
## fixed.acidity         0.66804729 -0.68297819  0.183005664 -0.06166827
## volatile.acidity      0.02202623  0.23493729 -0.260986685 -0.20228803
## citric.acid           0.36494718 -0.54190414  0.312770044  0.10990325
## residual.sugar        0.35528337 -0.08565242  0.005527121  0.04207544
```

```
## chlorides             0.20063233 -0.26502613  0.371260481 -0.22114054
## free.sulfur.dioxide  -0.02194583  0.07037750  0.051657572 -0.06940835
## total.sulfur.dioxide  0.07126948 -0.06649456  0.042946836 -0.20565394
## density               1.00000000 -0.34169933  0.148506412 -0.49617977
## pH                   -0.34169933  1.00000000 -0.196647602  0.20563251
## sulphates             0.14850641 -0.19664760  1.000000000  0.09359475
## alcohol              -0.49617977  0.20563251  0.093594750  1.00000000
## quality              -0.17491923 -0.05773139  0.251397079  0.47616632
##                          quality
## fixed.acidity         0.12405165
## volatile.acidity     -0.39055778
## citric.acid           0.22637251
## residual.sugar        0.01373164
## chlorides            -0.12890656
## free.sulfur.dioxide  -0.05065606
## total.sulfur.dioxide -0.18510029
## density              -0.17491923
## pH                   -0.05773139
## sulphates             0.25139708
## alcohol               0.47616632
## quality               1.00000000
```

```
heatmap(x = cor(wineData), symm = TRUE)
```



Observamos que las variables más correlacionadas son: fixed.acidity y density: tienen una correlación de 0.66804729 fixed.acidity y citric.acid: tienen una correlación de 0.67170343 total.sulfur.dioxide y free.sulfur.dioxide: tienen una correlación de 0.667666450

Vemos que ninguna de las variables están altamente correlacionadas (>80) por lo que decidimos trabajar con todas ellas.

Limpieza de los datos.

¿Los datos contienen ceros o elementos vacíos?

Una vez cargado el dataset comprobamos si este contiene elementos nulos o vacíos.

```
colSums(is.na(wineData))
```

```
##        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
```

Observamos que no tiene elementos vacíos.

Comprobamos si el dataset contiene ceros.

```
colSums(wineData==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
```

Vemos que la única columna que contiene zeros es "citric.acid" con 132 registros.

¿Cómo gestionarías cada uno de estos casos?

Hemos visto que para la única variable que tenemos ceros es "citric.acid". Sospechamos que posiblemente, y más habiendo 132 registros con valor zero, sea relativamente frecuente encontrar algunos vinos sin ácido cítrico.

Procedemos a ver más detalladamente las estadísticas de esta variable.

```
summary(wineData[,3])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

Observamos que el primer cuartil tiene un valor de 0.090, la media de 0.260 y el valor máximo es 1. Viendo que hay valores tan cercanos a 0, asumimos que los 132 registros no son campos vacíos y simplemente significan que ese vino no contiene ácido cítrico.

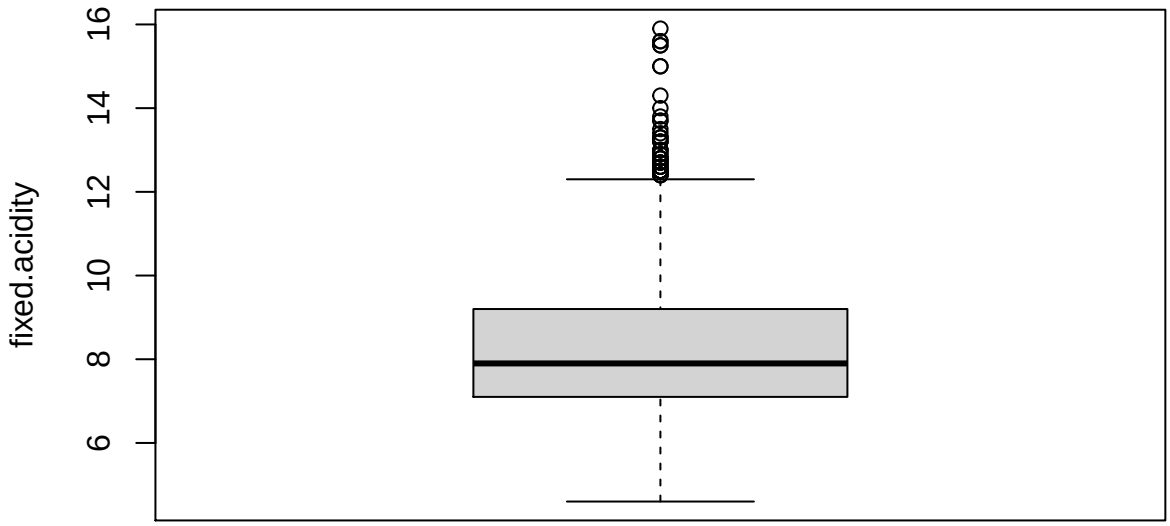
Identificación y tratamiento de valores extremos.

Para la identificación de outliers utilizaremos los boxplots, que permite recuperar los valores de los outliers además de representarlos gráficamente.

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

```
##  [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 12.8
## [16] 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4 12.5 12.9
## [31] 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2 13.2 13.2 15.9
## [46] 13.3 12.9 12.6 12.6
```

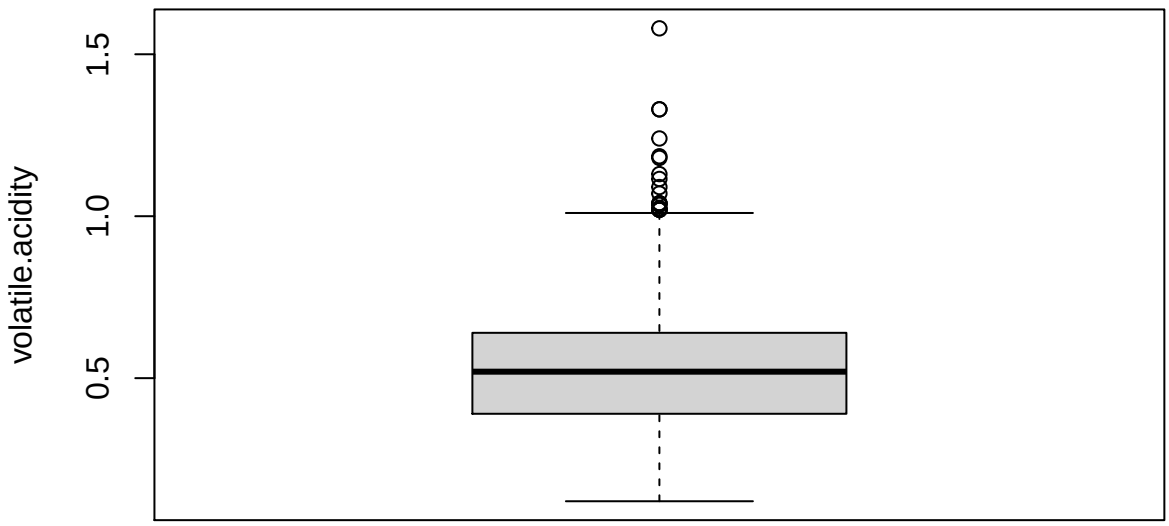
```
boxplot(wineData$fixed.acidity,
  ylab = "fixed.acidity"
)
```




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

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035
## [13] 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

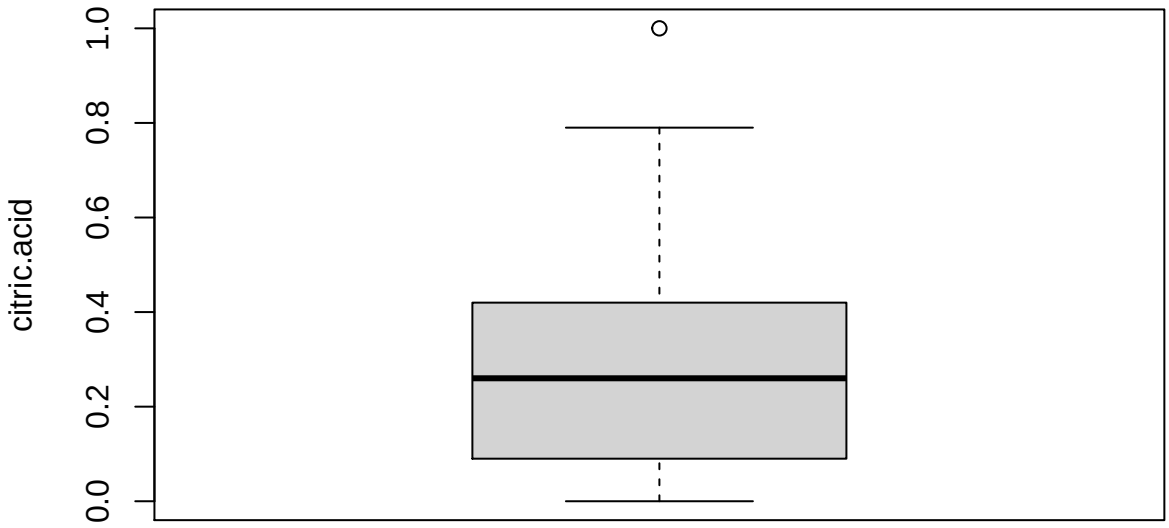
```
boxplot(wineData$volatile.acidity,
  ylab = "volatile.acidity"
)
```




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

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

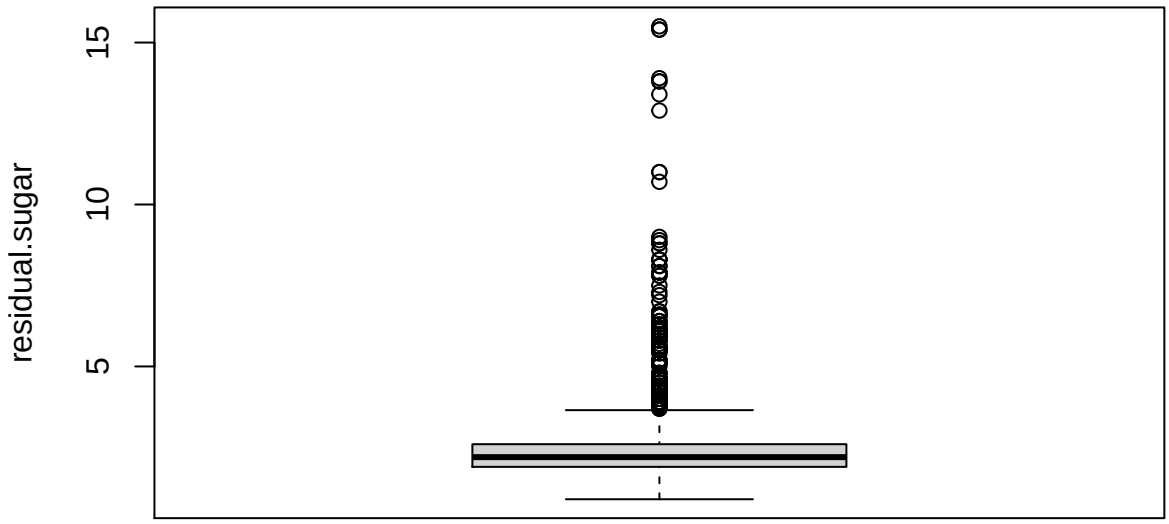
```
boxplot(wineData$citric.acid,
  ylab = "citric.acid"
)
```




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

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10  4.65
##  [13]  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00  4.00  4.00
##  [25]  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00  4.50  4.80  5.80
##  [37]  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70  4.50  6.70  6.60  3.70
##  [49]  5.20 15.50  4.10  8.30  6.55  6.55  4.60  6.10  4.30  5.80  5.15  6.30
##  [61]  4.20  4.20  4.60  4.20  4.60  4.30  4.30  7.90  4.60  5.10  5.60  5.60
##  [73]  6.00  8.60  7.50  4.40  4.25  6.00  3.90  4.20  4.00  4.00  4.00  6.60
##  [85]  6.00  6.00  3.80  9.00  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10
##  [97]  6.20  8.90  4.00  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70
## [109]  5.50  5.50  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90
## [121]  4.30 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80
## [133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75 13.80
## [145] 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10  7.80
```

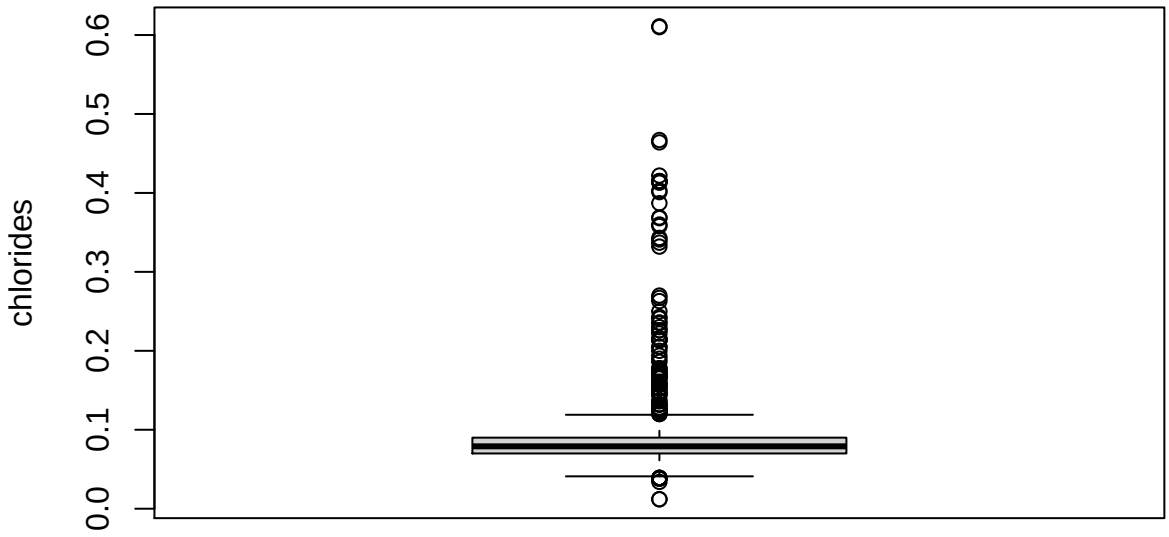
```
boxplot(wineData$residual.sugar,
  ylab = "residual.sugar"
)
```




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

```
##   [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178 0.146
##  [13] 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343 0.186 0.213
##  [25] 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122 0.122 0.174 0.121
##  [37] 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171 0.226 0.226 0.250 0.148
##  [49] 0.122 0.124 0.124 0.143 0.222 0.039 0.157 0.422 0.034 0.387 0.415 0.157
##  [61] 0.157 0.243 0.241 0.190 0.132 0.126 0.038 0.165 0.145 0.147 0.012 0.012
##  [73] 0.039 0.194 0.132 0.161 0.120 0.120 0.123 0.123 0.414 0.216 0.171 0.178
##  [85] 0.369 0.166 0.166 0.136 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414
##  [97] 0.166 0.168 0.415 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205
## [109] 0.039 0.235 0.230 0.038
```

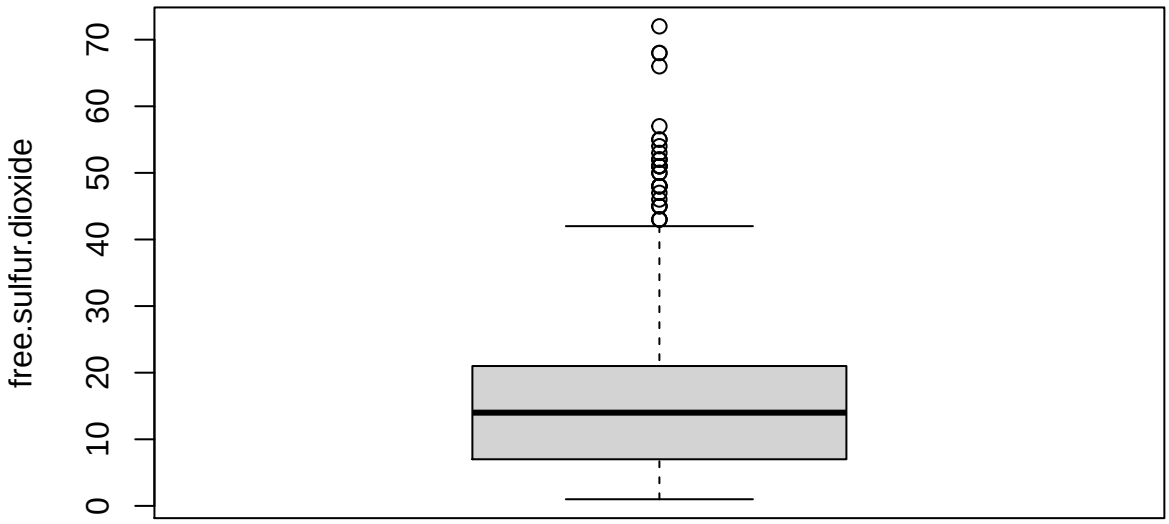
```
boxplot(wineData$chlorides,
  ylab = "chlorides"
)
```




```
boxplot.stats(wineData$free.sulfur.dioxide)$out
```

```
##  [1] 52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51 51 52
## [26] 55 55 48 48 66
```

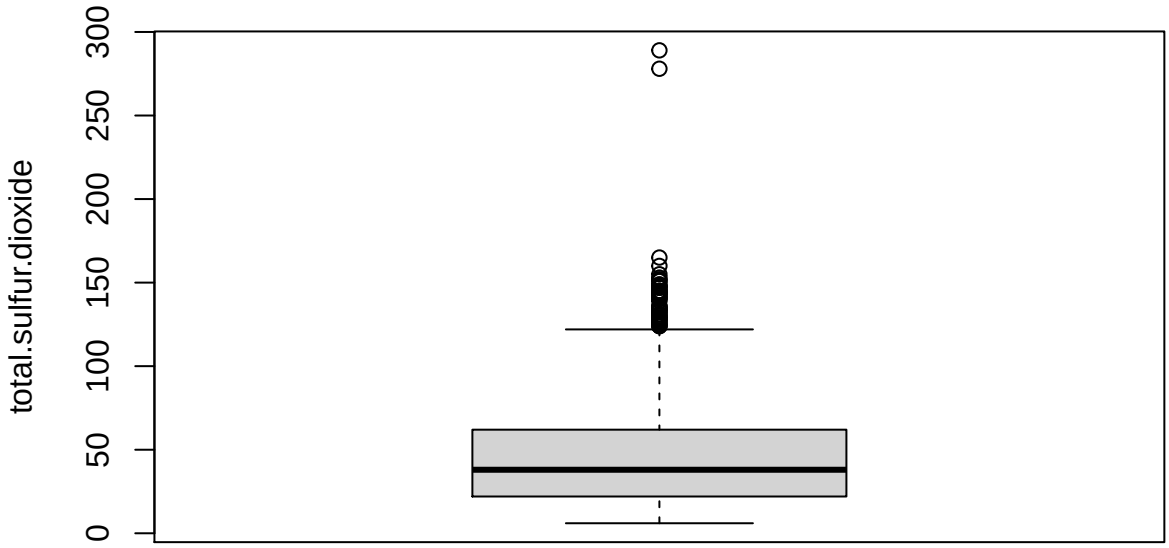
```
boxplot(wineData$free.sulfur.dioxide,
  ylab = "free.sulfur.dioxide"
)
```




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

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

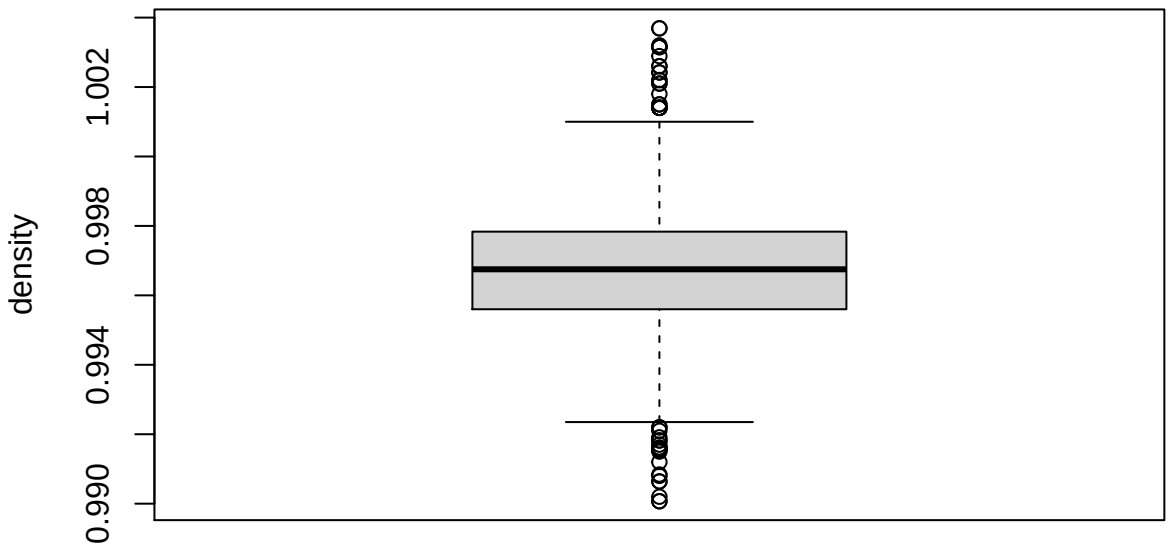
```
boxplot(wineData$total.sulfur.dioxide,
  ylab = "total.sulfur.dioxide"
)
```




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

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

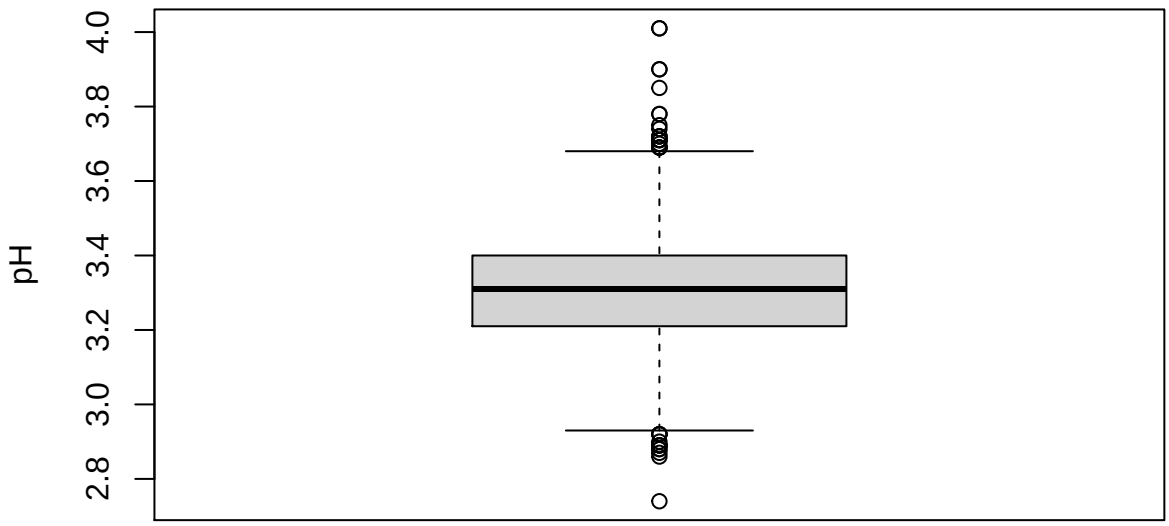
```
boxplot(wineData$density,
  ylab = "density"
)
```




```
boxplot.stats(wineData$pH)$out
```

```
##  [1] 3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87 2.89
## [16] 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71 2.89 2.89 3.78 3.70 3.78 4.01 2.90
## [31] 4.01 3.71 2.88 3.72 3.72
```

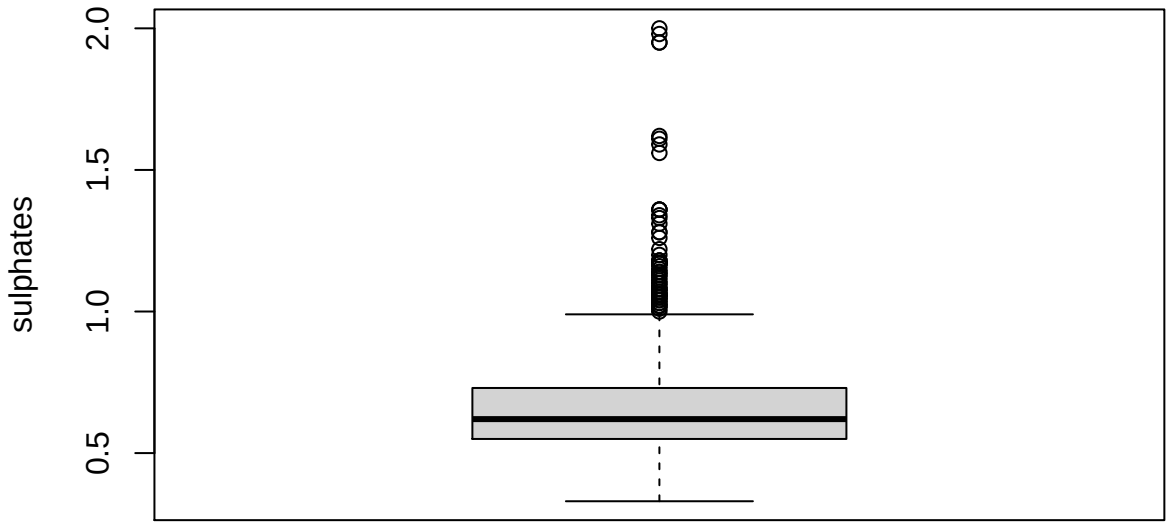
```
boxplot(wineData$pH,
  ylab = "pH"
)
```




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

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

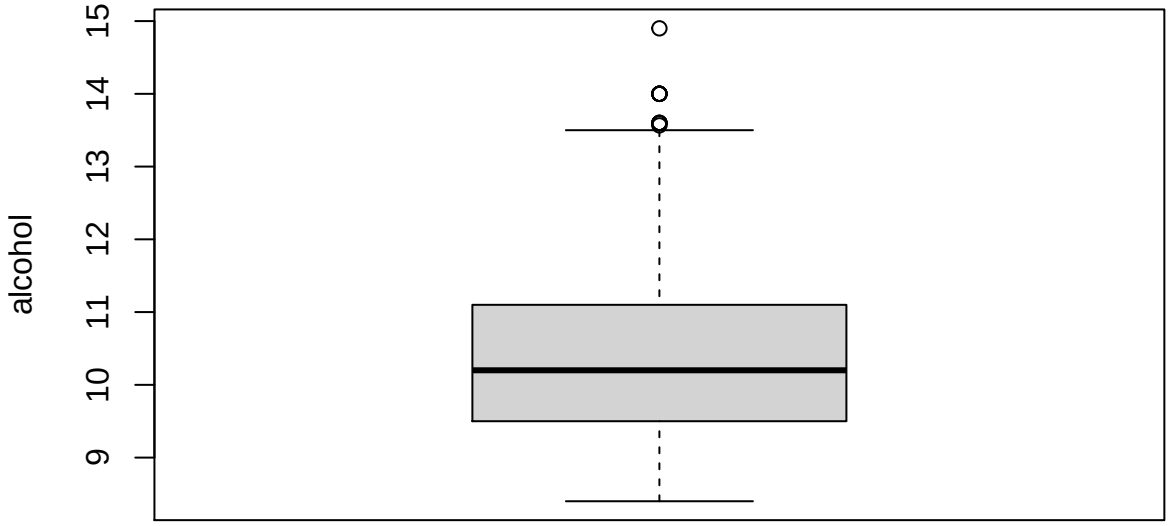
```
boxplot(wineData$sulphates,
  ylab = "sulphates"
)
```




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

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

```
boxplot(wineData$alcohol,
  ylab = "alcohol"
)
```



A continuación mostramos algunas estadísticas del dataset como la media y el valor máximo de las variables que pueden resultarnos útiles a la hora de identificar mejor los outliers

```
summary(wineData)
```

```
##  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    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Observamos que todas las variables presentan outliers.

Los outliers más alejados de la media son:

residual.sugar free.sulfur.dioxide total.sulfur.dioxide

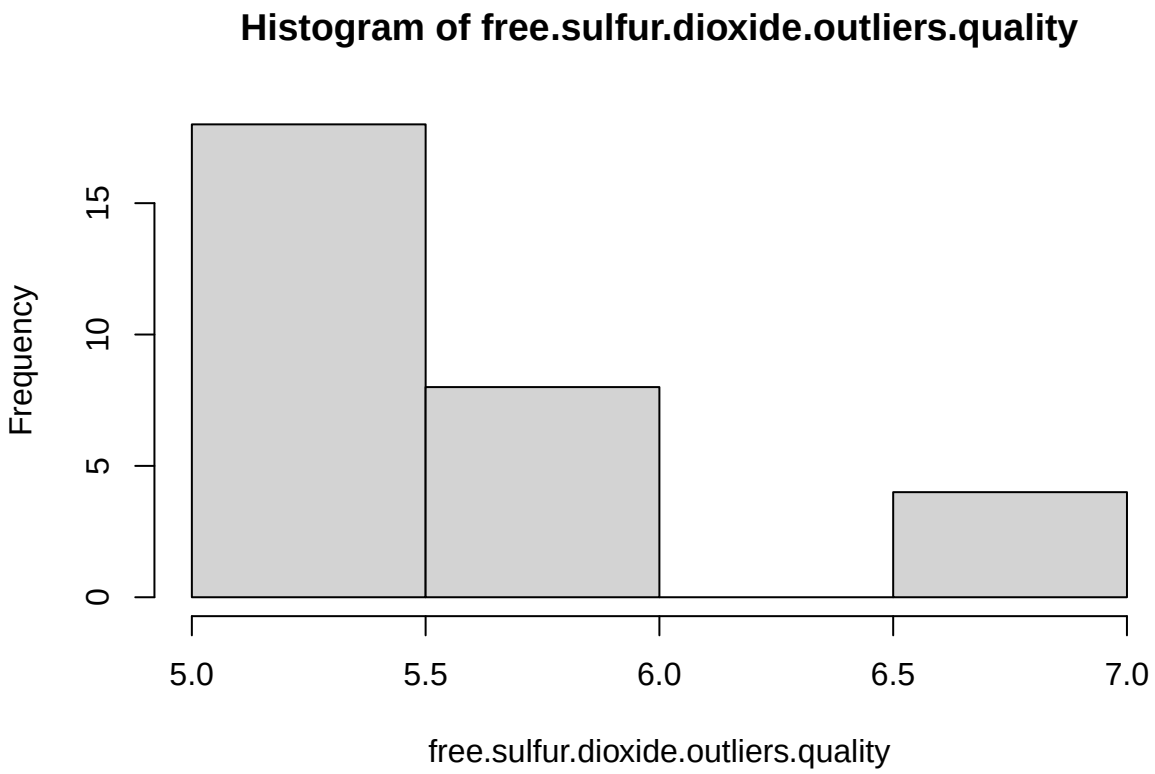
Observamos la distribución de la calidad de los vinos de estos outliers.

```
residual.sugar.outliers.quality = wineData$quality[wineData$residual.sugar %in% boxplot.stats(wineData$
```

```
hist(residual.sugar.outliers.quality)
```

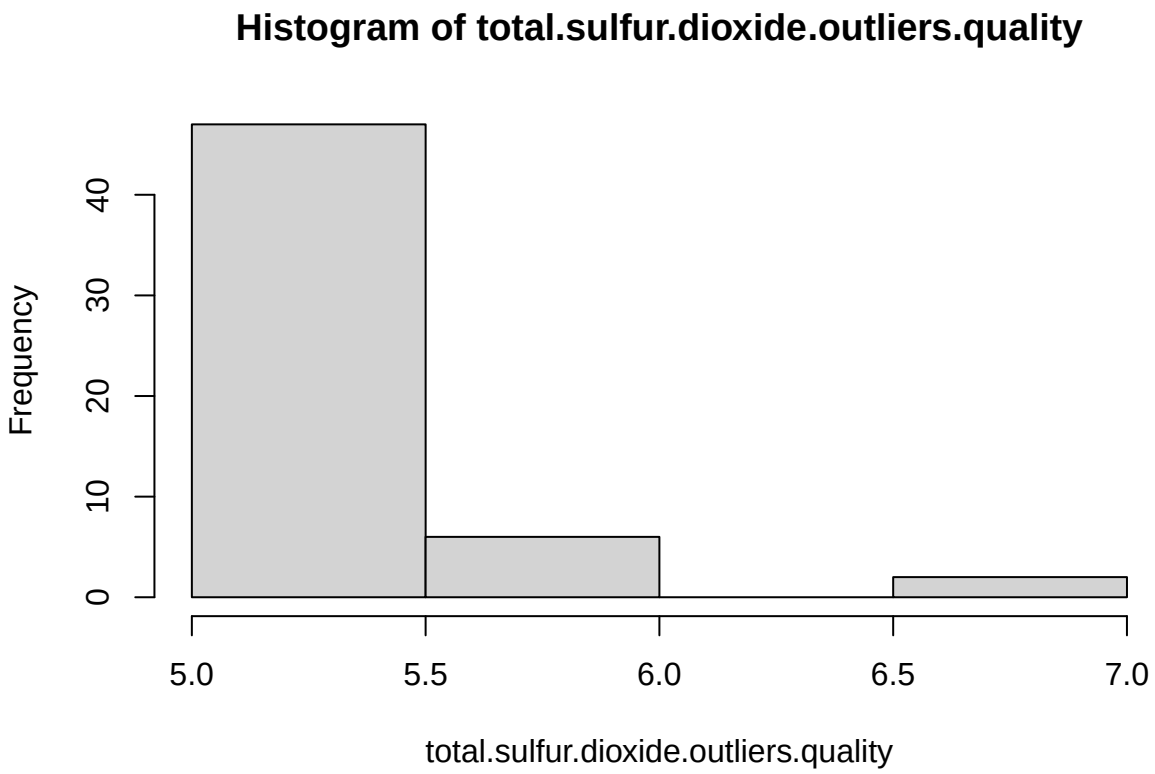

```
free.sulfur.dioxide.outliers.quality = wineData$quality[wineData$free.sulfur.dioxide %in% boxplot.stats
```

```
hist(free.sulfur.dioxide.outliers.quality)
```




```
total.sulfur.dioxide.outliers.quality = wineData$quality[wineData$total.sulfur.dioxide %in% boxplot.sta

hist(total.sulfur.dioxide.outliers.quality)
```



Observamos que los outliers de residual.sugar siguen una distribución normal, mientras que los outliers tanto de free.sulfur.dioxide como de total.sulfur.dioxide adquieren principalmente valores de calidad de 5.

Previamente hemos visto que la correlación entre residual.sugar y quality era muy baja (0.013731637) por lo que tiene sentido que en los outliers de residual.sugar observemos una distribución normal para la variable quality.

La correlación entre quality y free.sulfur.dioxide, y quality y total.sulfur.dioxide también es relativamente baja, -0.050656057 y -0.18510029 respectivamente. En estos casos podemos observar que estos outliers coinciden con una calidad mediocre del vino (5 en la mayoría de casos).

Una vez analizados los outliers asumimos que estos valores no son datos erróneos, por lo que no los eliminamos.

Análisis de los datos.

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

Para este analisis separaremos los grupos de datos en 2, los que tienen alcohol < 12 y los que tienen alcohol $>= 12$

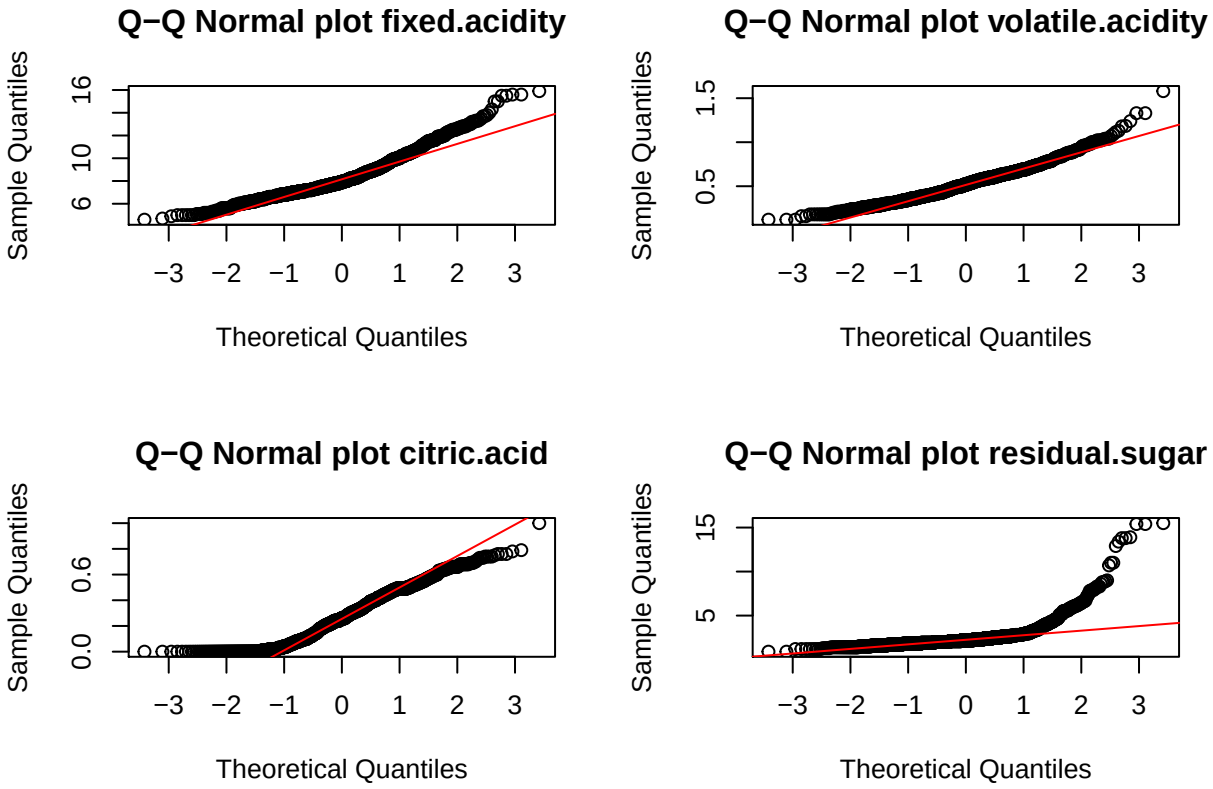
```
low.alcohol = wineData[wineData$alcohol < 12,]
high.alcohol = wineData[wineData$alcohol >= 12,]
```

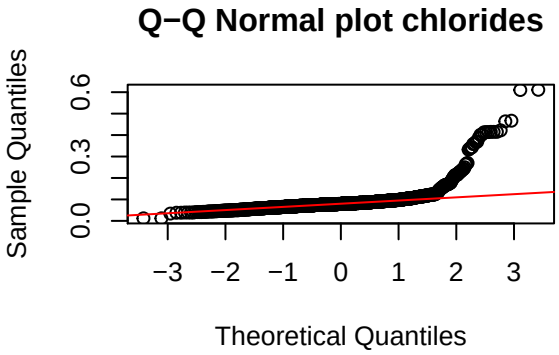
Comprobación de la normalidad y homogeneidad de la varianza.

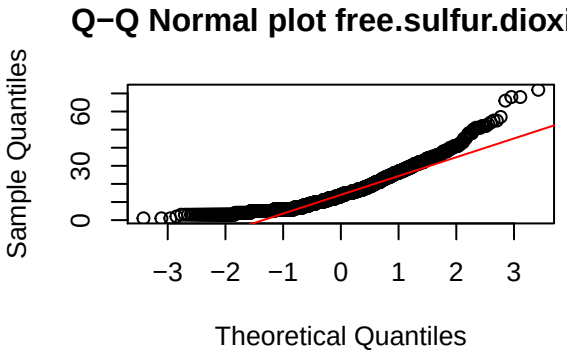
Comprobaremos la nornmalidad y homogeneidad de la varianza para cada una de las variables del dataset.

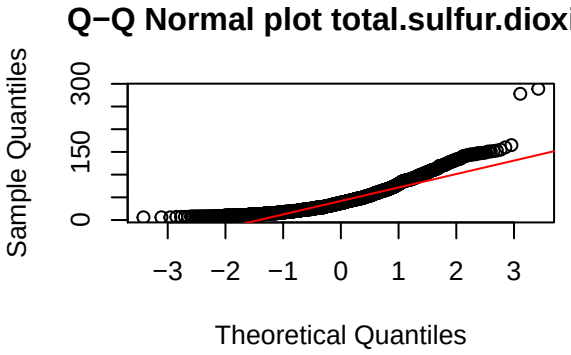
- Primero visualizaremos los Q-Q Plots de cada uno de los campos del dataset:

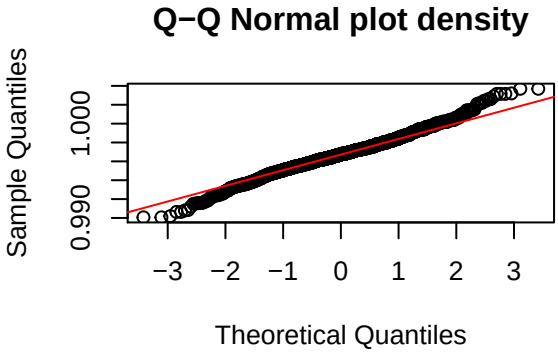
```r
par(mfrow=c(2,2))
for (i in 1:ncol(wineData)){
    qqnorm(wineData[,i],main = paste ("Q-Q Normal plot", colnames(wineData)[i]))
    qqline(wineData[,i],col="red")
}
```

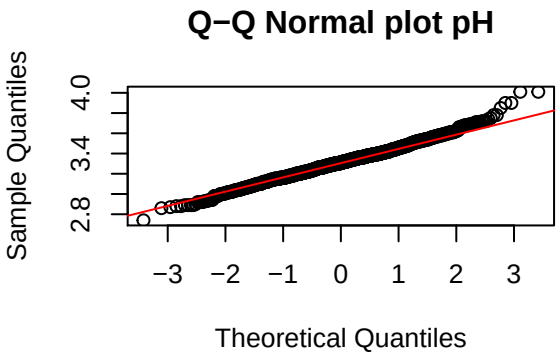


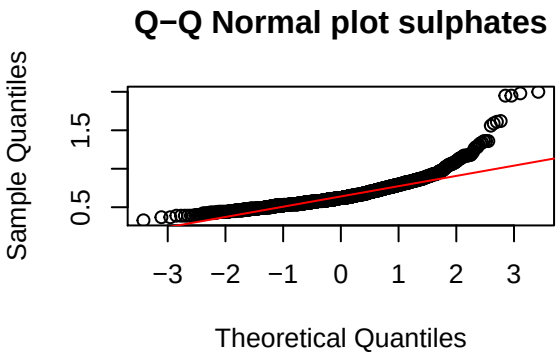

Q−Q Normal plot chlorides
Sample Quantiles
0.0
0.3
0.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

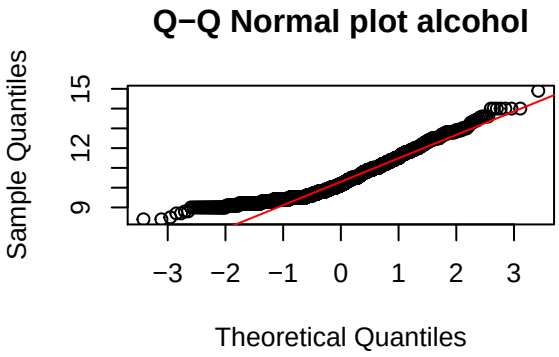

Q−Q Normal plot free.sulfur.dioxide
Sample Quantiles
0
30
60
−3
−2
−1
0
1
2
3
Theoretical Quantiles

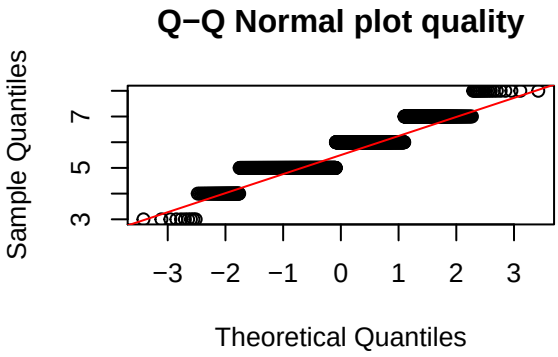

Q−Q Normal plot total.sulfur.dioxide
Sample Quantiles
0
150
300
−3
−2
−1
0
1
2
3
Theoretical Quantiles


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









Después de realizar los Q-Q Plots y observar que en principio las variable serían candidatas a normalidad, realizamos también el test de Shapiro para cada una de ellas. El test se Shapiro-Wilk asume que toda la población sigue una distribución normal como hipótesis nula. Si el p-value es menor que 0.05 entonces esta hipotesis queda rechazada y se determina que no se sigue una distribución normal.

```
shapiro.test(wineData$quality)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wineData$quality
## W = 0.85759, p-value < 2.2e-16
```

```
shapiro.test(wineData$fixed.acidity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wineData$fixed.acidity
## W = 0.94203, p-value < 2.2e-16
```

```
shapiro.test(wineData$volatile.acidity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wineData$volatile.acidity
## W = 0.97434, p-value = 2.693e-16
```

```r
shapiro.test(wineData$citric.acid)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wineData$citric.acid
## W = 0.95529, p-value < 2.2e-16
```

```r
shapiro.test(wineData$residual.sugar)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wineData$residual.sugar
## W = 0.56608, p-value < 2.2e-16
```

```r
shapiro.test(wineData$chlorides)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wineData$chlorides
## W = 0.48425, p-value < 2.2e-16
```

```r
shapiro.test(wineData$free.sulfur.dioxide)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wineData$free.sulfur.dioxide
## W = 0.90184, p-value < 2.2e-16
```

```r
shapiro.test(wineData$total.sulfur.dioxide)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wineData$total.sulfur.dioxide
## W = 0.87322, p-value < 2.2e-16
```

```r
shapiro.test(wineData$density)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wineData$density
## W = 0.99087, p-value = 1.936e-08
```

```r
shapiro.test(wineData$pH)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wineData$pH
## W = 0.99349, p-value = 1.712e-06
```

```r
shapiro.test(wineData$sulphates)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  wineData$sulphates
## W = 0.83304, p-value < 2.2e-16
```

```
shapiro.test(wineData$alcohol)
```

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

Observamos que ninguna de las variables sigue una distribución normal según el test de Shapiro-Wilk.

Pasamos a comprobar la homogeneidad de la varianza con el test de Fligner-Killeen dado que tal y como hemos visto, las variables no siguen una distribución normal.

```
fligner.test(fixed.acidity ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  fixed.acidity by quality
## Fligner-Killeen:med chi-squared = 34.635, df = 5, p-value = 1.779e-06
```

```
fligner.test(volatile.acidity ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  volatile.acidity by quality
## Fligner-Killeen:med chi-squared = 30.826, df = 5, p-value = 1.014e-05
```

```
fligner.test(citric.acid ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  citric.acid by quality
## Fligner-Killeen:med chi-squared = 10.916, df = 5, p-value = 0.05307
```

```
fligner.test(residual.sugar ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  residual.sugar by quality
## Fligner-Killeen:med chi-squared = 7.9984, df = 5, p-value = 0.1563
```

```
fligner.test(chlorides ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  chlorides by quality
## Fligner-Killeen:med chi-squared = 15.24, df = 5, p-value = 0.009383
```

```
fligner.test(free.sulfur.dioxide ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  free.sulfur.dioxide by quality
## Fligner-Killeen:med chi-squared = 14.137, df = 5, p-value = 0.01476
```

```
fligner.test(total.sulfur.dioxide ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  total.sulfur.dioxide by quality
## Fligner-Killeen:med chi-squared = 137.27, df = 5, p-value < 2.2e-16
```

```
fligner.test(density ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  density by quality
## Fligner-Killeen:med chi-squared = 49.468, df = 5, p-value = 1.78e-09
```

```
fligner.test(pH ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  pH by quality
## Fligner-Killeen:med chi-squared = 1.2419, df = 5, p-value = 0.9408
```

```
fligner.test(sulphates ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  sulphates by quality
## Fligner-Killeen:med chi-squared = 9.4044, df = 5, p-value = 0.09398
```

```
fligner.test(alcohol ~ quality, data = wineData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  alcohol by quality
## Fligner-Killeen:med chi-squared = 135.61, df = 5, p-value < 2.2e-16
```

Observamos que las variables citric.acid, residual.sugar, pH y sulphates presentan un p-value >= 0.05 para el test de Fligner-Killeen por lo que podemos asumir una igualdad de varianza en los diferentes grupos de quality.

Aplicación de pruebas estadísticas para comparar los grupos de datos.

Hipotesis A continuación realizaremos un contraste de hipótesis para determinar si la calidad del vino es más alta según si el vino tiene poco alcohol (<12) o mucho (>=12).

Para ello haremos la prueba de Wilcoxon (ya que los datos no siguen una distribución norma) que plantea las siguientes hipotesis

- H0: las distribuciones de los grupos de datos son las mismas

- H1: la distribución de los vinos con más alcohol es mayor que la distribución de vinos con poco alcohol

```
wilcox.test(high.alcohol$quality, low.alcohol$quality, alternative = 'greater')
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  high.alcohol$quality and low.alcohol$quality
## W = 181481, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

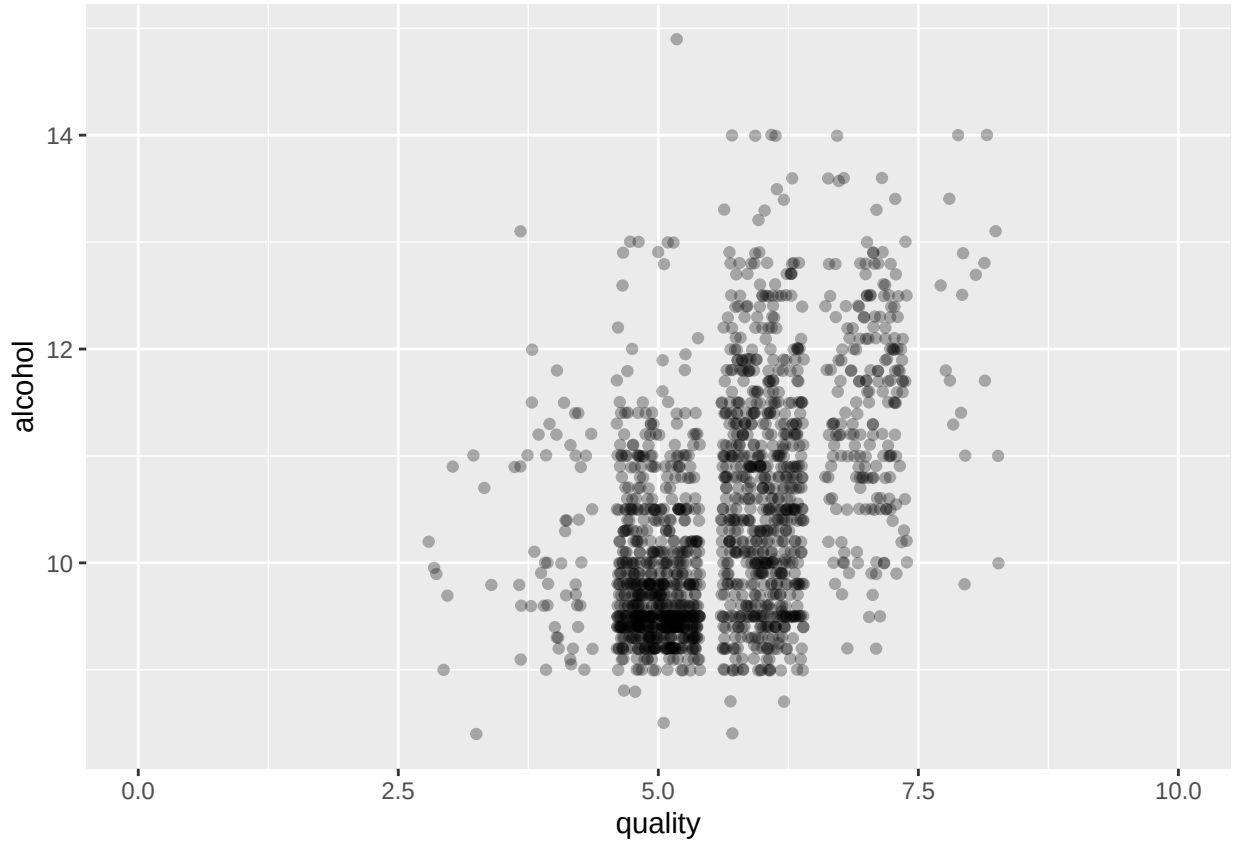
Observamos que el p-value es $<$ que 0.05 por lo que rechazamos la hipotesis nula. Confirmando asi, que los vinos con más alcohol tienen una mejor calidad.

Correlaciones Empezaremos visualizando las correlaciones entre quality y las tres variables que tienen una correlación más alta (alcohol, citric.acid y sulphates).

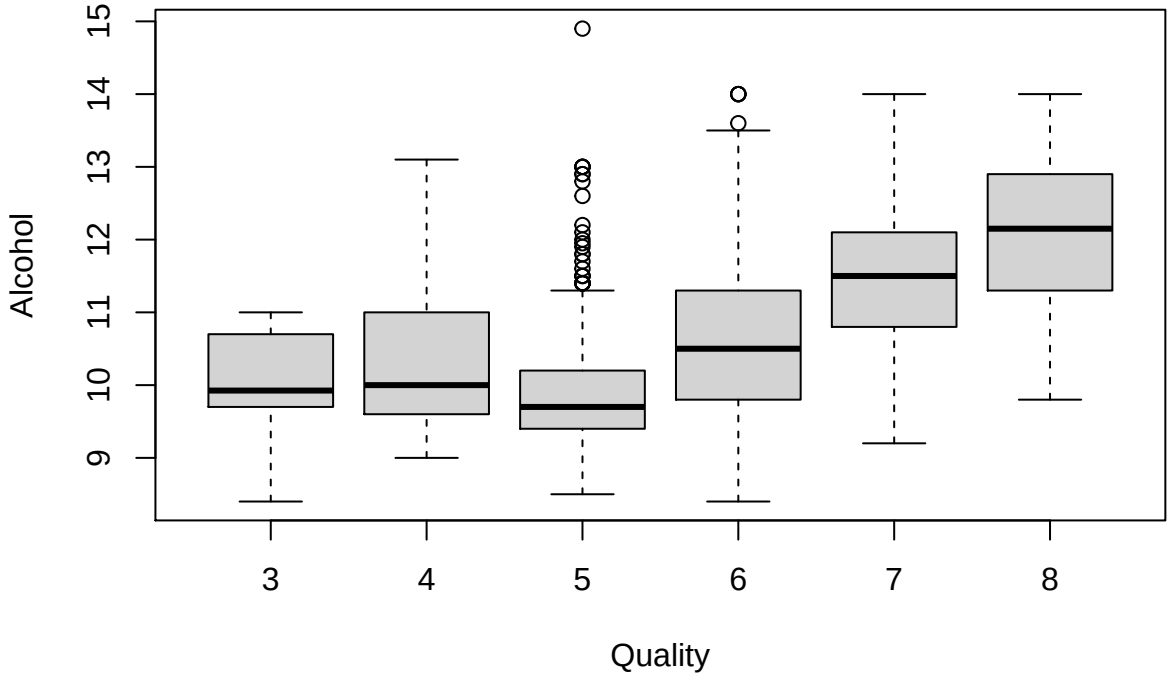
Primero vamos a visualizar como se relacionan entre ellas uno a uno, es decir, la calidad del vino con cada una de las variables. Visualizamos dos gráficas, una de correlaciones típicas y otra de correlación por distribución de boxplot.

Visualizamos la correlación entre quality y alcohol. Observamos como a mayor cantidad de alcohol más calidad en el vino.

```
library(ggplot2)
ggplot(aes(x = quality, y = alcohol), data = wineData) + geom_point(alpha = 0.30, position = 'jitter') +
  coord_cartesian(xlim = c(0,10))
```

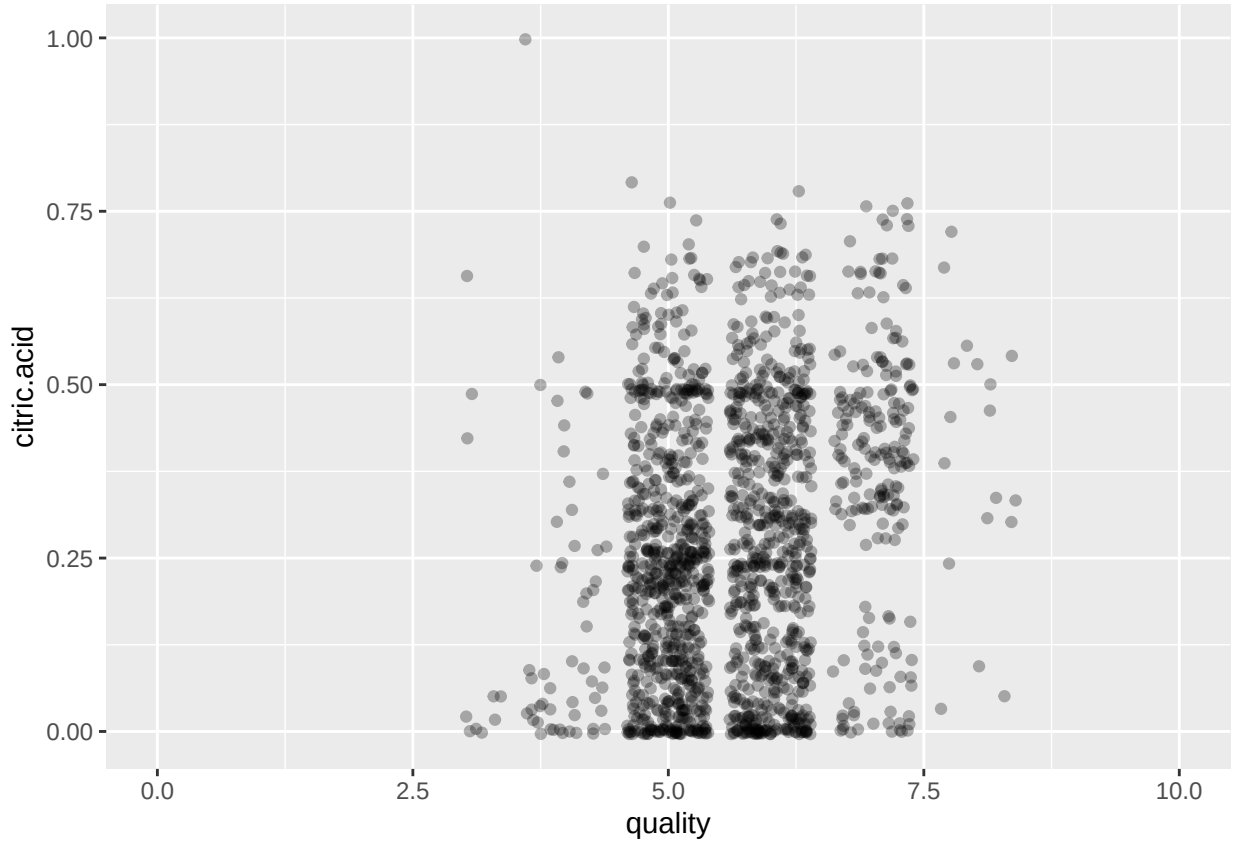


```
boxplot(alcohol~quality, data = wineData,
    xlab = "Quality",
    ylab = "Alcohol")
```

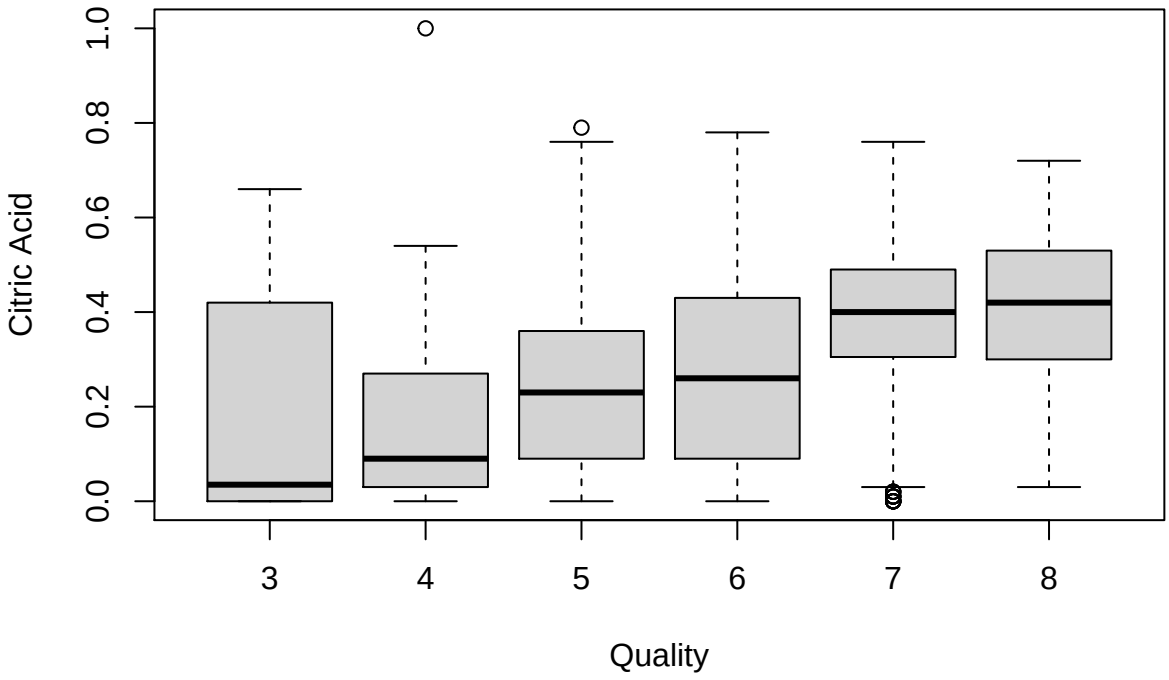


Visualizamos la correlación entre quality y citric.acid. Observamos como a mayor cantidad de citric.acid más calidad en el vino.

```
library(ggplot2)
ggplot(aes(x = quality, y = citric.acid), data = wineData) + geom_point(alpha = 0.30, position = 'jitte
  coord_cartesian(xlim = c(0,10))
```

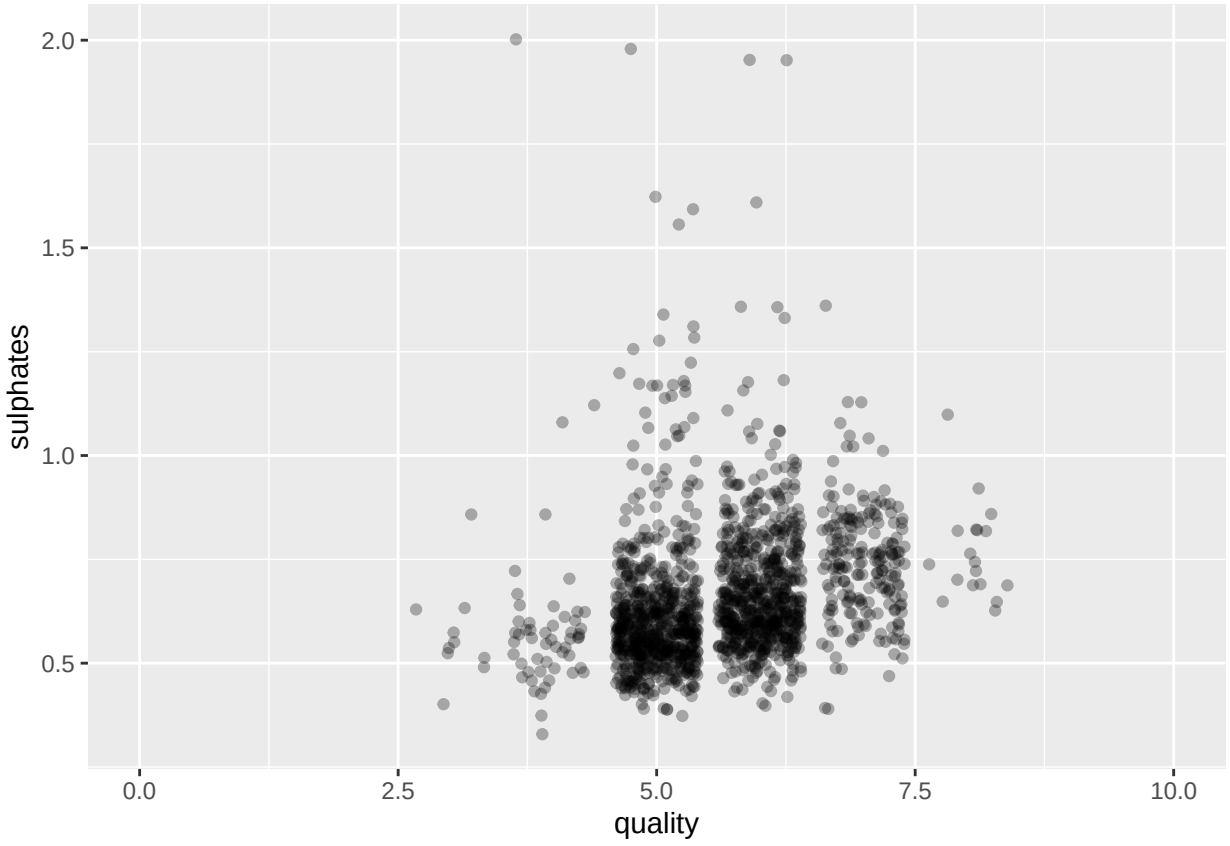



```
boxplot(citric.acid~quality, data = wineData,
    xlab = "Quality",
    ylab = "Citric Acid")
```

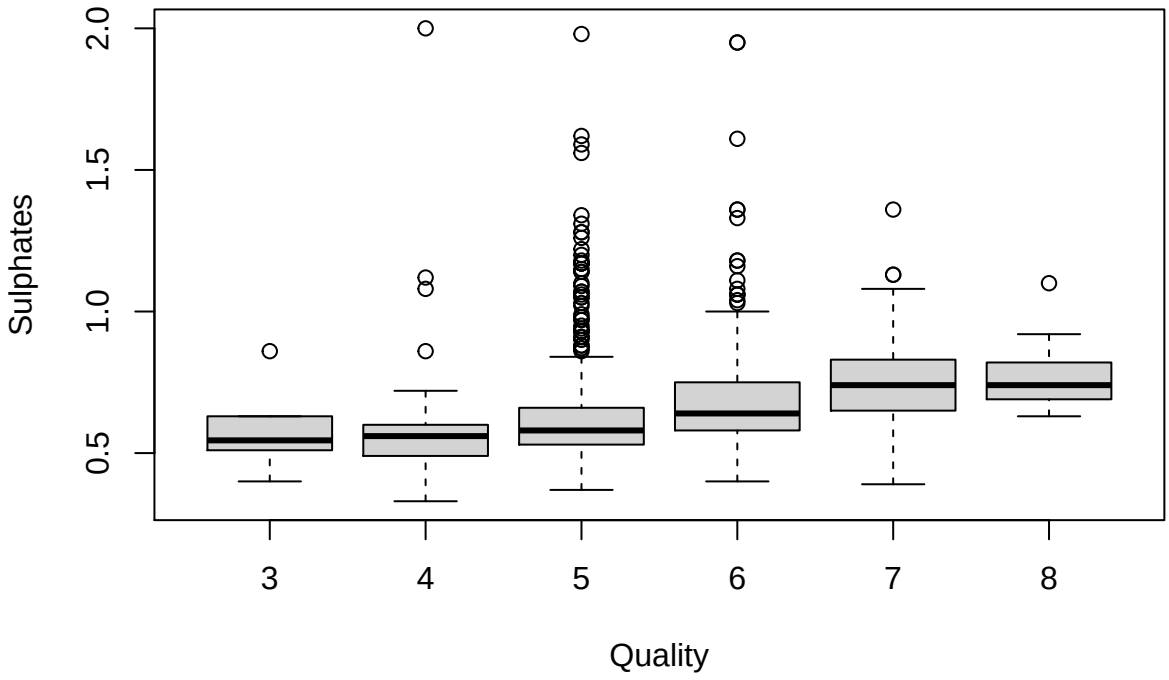


Visualizamos la correlación entre quality y sulphates. Observamos como a mayor cantidad de sulphates más calidad en el vino, aunque la correlación tiene menos fuerza que las dos observadas anteriormente.

```
library(ggplot2)
ggplot(aes(x = quality, y = sulphates), data = wineData) + geom_point(alpha = 0.30, position = 'jitter')
  coord_cartesian(xlim = c(0,10))
```



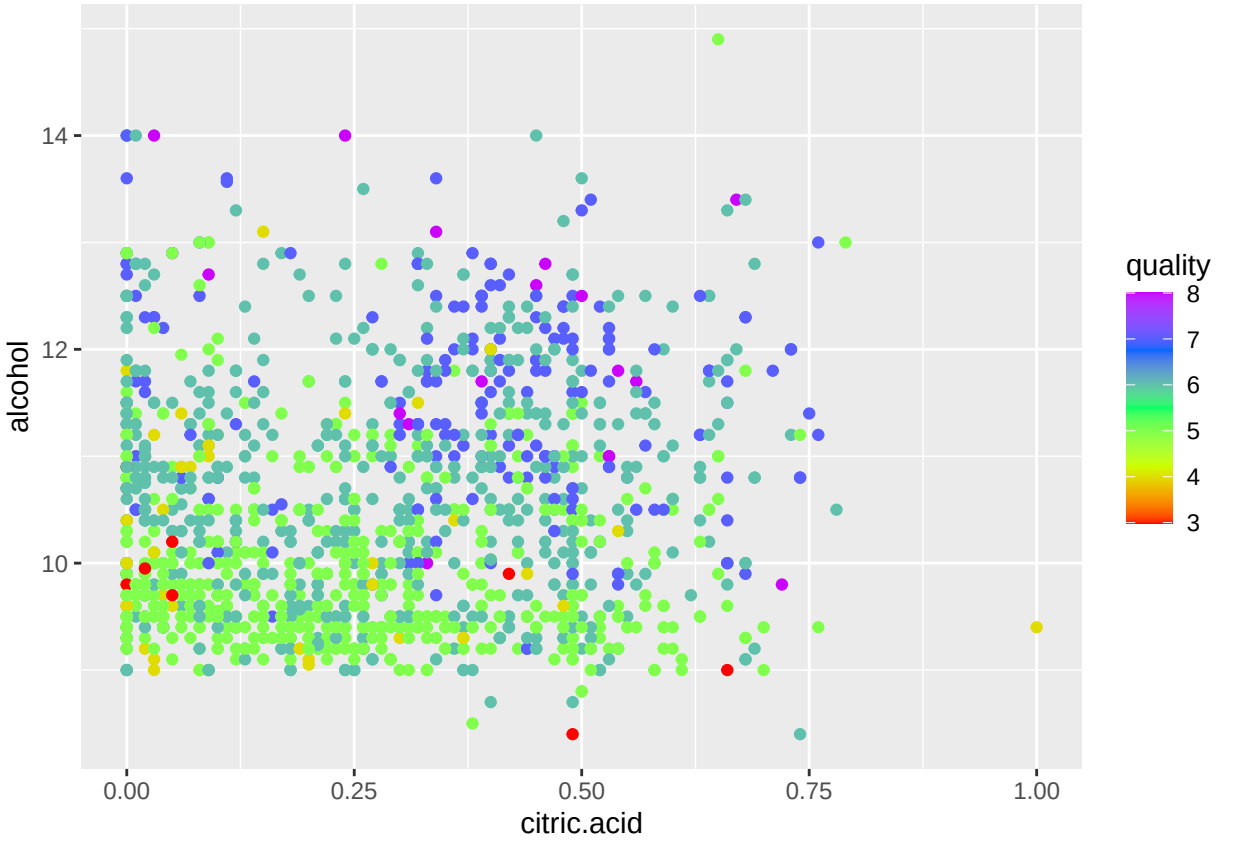

```
boxplot(sulphates~quality, data = wineData,
    xlab = "Quality", ylab = "Sulphates")
```



Ahora vamos a pasar a visualizar la correlación de la calidad del vino con parejas de los tres elementos que hemos visto.

Empezamos con la visualización de la correlación de quality con citric.acid y alcohol.

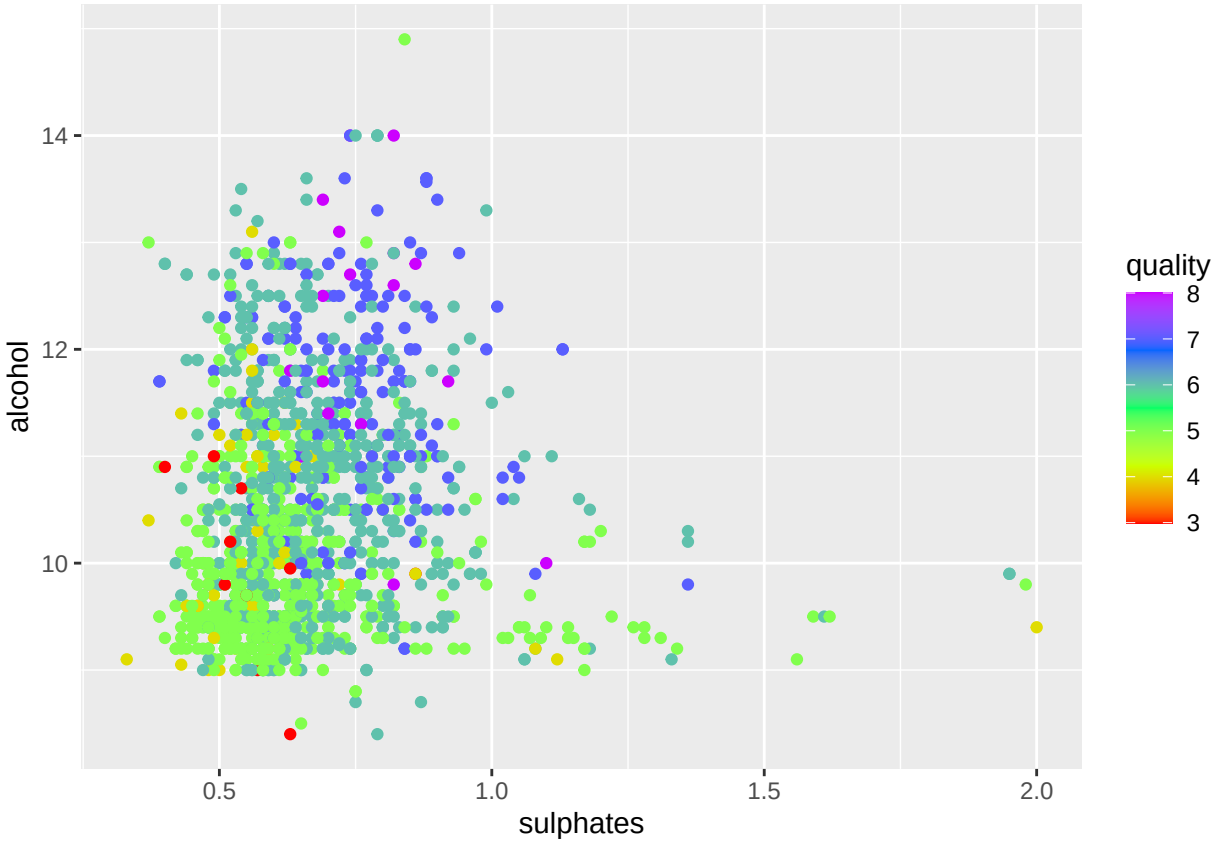
```
ggplot(aes(x = citric.acid, y = alcohol, color = quality),
  data = wineData) + geom_point() +
  scale_color_gradientn(colors = rainbow(5))
```



Observamos como a mayor cantidad de ambos, la calidad tiende a ser superior. Aunque observamos que estos dos factores no explican toda la variabilidad de quality.

Continuamos con la visualización de la correlación de quality con sulphates y alcohol.

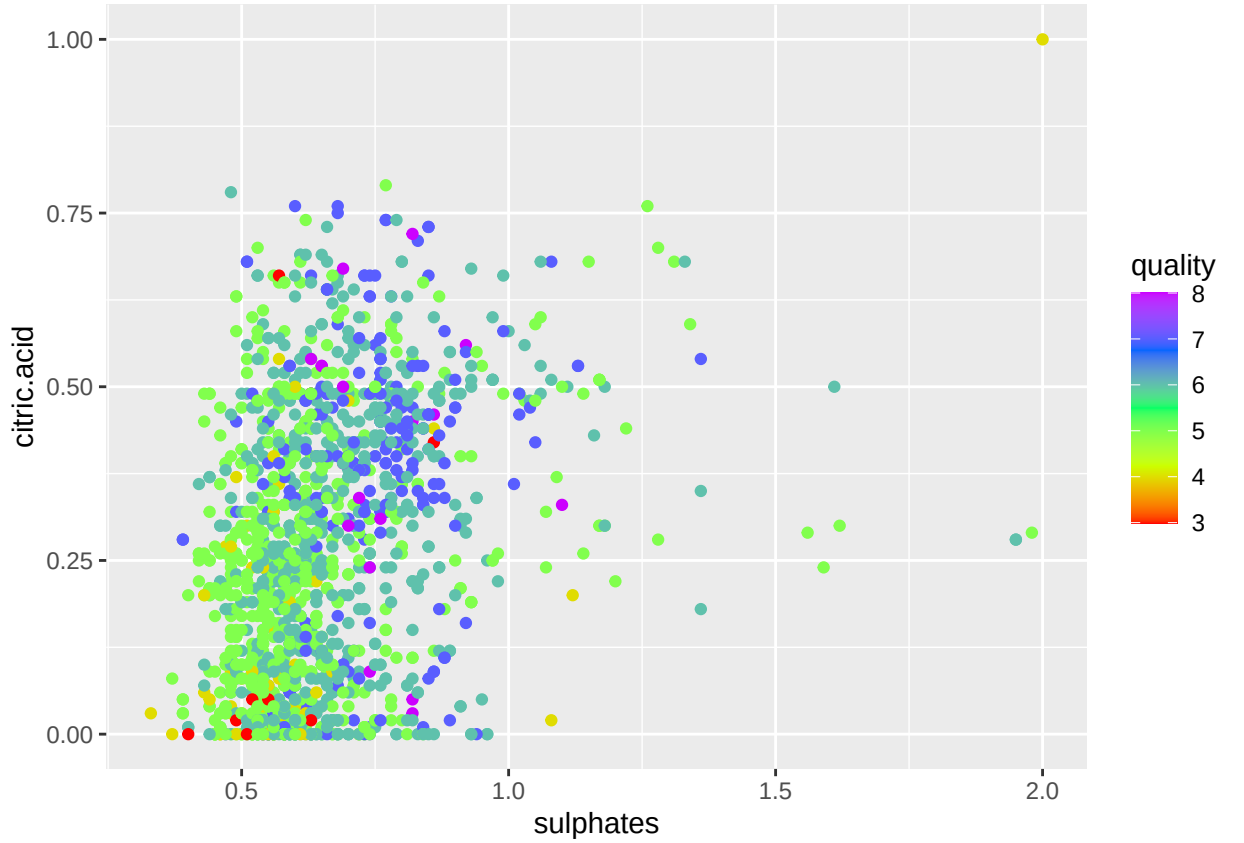
```
ggplot(aes(x = sulphates, y = alcohol, color = quality),
  data = wineData) + geom_point() +
  scale_color_gradientn(colors = rainbow(5))
```



De forma similar a la gráfica anterior observamos como a mayor cantidad de ambos, la calidad tiende a ser superior. Igual que en el caso anterior observamos que estos dos factores no explican toda la variabilidad de quality.

Por último visualizamos la correlación de quality con sulphates y citric.acid.

```
ggplot(aes(x = sulphates, y = citric.acid, color = quality),
  data = wineData) + geom_point() +
  scale_color_gradientn(colors = rainbow(5))
```



Observamos que en este caso la correlación no es tan clara como en las otras dos parejas. Si bien es verdad que la mayotia de vinos con alta calidad se presentan para valores altos de citric.acid y sulphates, también encontramos bastantes vinos de baja calidad para estos mismos valores.

Regresiones Ahora realizaremos dos análisis con la creación de modelos de regresión lineal, concretamente utilizaremos dos modelos:

- Modelo de regresión lineal con todas las variables del dataset.
- Modelo de regresión lineal con solo las variables con mayor correlación

Realizamos modelo de regresión lineal con todas las variables del dataset.

```
model_linear <- lm(quality ~., data = wineData)
summary(model_linear)
```

```
##
## Call:
## lm(formula = quality ~ ., data = wineData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.197e+01  2.119e+01   1.036   0.3002
## fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357
## volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
```

```
## citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150
## residual.sugar        1.633e-02  1.500e-02   1.089   0.2765
## chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
## total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## density              -1.788e+01  2.163e+01  -0.827   0.4086
## pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
## sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Observamos como la mayoría de variables tienen un impacto nulo o muy bajo en relación con la calidad. Las variables que tienen un impacto mayor son alcohol y sulphates.

Realizamos modelo de regresión lineal con solo las variables con mayor correlación.

```
model_linear_sig <- lm(quality ~ citric.acid + sulphates + alcohol, data = wineData)
summary(model_linear_sig)
```

```
##
## Call:
## lm(formula = quality ~ citric.acid + sulphates + alcohol, data = wineData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7565 -0.3535 -0.1007  0.5067  2.2125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.43392    0.17615   8.140 7.86e-16 ***
## citric.acid  0.51345    0.09284   5.531 3.72e-08 ***
## sulphates    0.81403    0.10651   7.643 3.65e-14 ***
## alcohol      0.33841    0.01619  20.903  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6842 on 1595 degrees of freedom
## Multiple R-squared:  0.2836, Adjusted R-squared:  0.2823
## F-statistic: 210.5 on 3 and 1595 DF,  p-value: < 2.2e-16
```

Observamos como la variable con mayor influencia en quality es alcohol, seguida a distancia por sulphates y citric.acid.

Para determinar cual de los dos modelos es mas preciso podemos mirar cual tiene el coeficiente de determinación R-squared más elevado. Podemos determinar así que el modelo con todas las variables (R-squared = 0.3606) es mejor que el modelo con las tres variables más significativas (R-squared = 0.2836).

Resolución del problema.

Conclusiones

Después del análisis de los datos y de haber aplicado tres metodos de análisis estadístico distintos hemos podido entender mejor que elementos fisicoquimicos tienen más influencia a la hora de determinar la calidad del vino.

La prueba de Wilcoxon ha determinado que los vinos con 12º de alcohol o más tienen por lo general una calidad del vino más alta que los que tienen menos de 12º de alcohol.

Más adelante, observando las correlaciones entre la calidad del vino y diferentes variables hemos visto que:

- A mayor cantidad de alcohol mayor calidad. Coincidiendo con la prueba de Wilcoxon.
- A mayor cantidad de ácido cítrico mayor calidad.
- A mayor cantidad de sulfatos mayor calidad.
- A mayor cantidad de alcohol y ácido cítrico mayor calidad.
- A mayor cantidad de alcohol y sulfatos mayor calidad.
- A mayor cantidad de ácido cítrico y sulfatos no necesariamente mayor calidad.

Realizando diferentes modelos de regresión lineal hemos visto que el que mejor funciona es el que utiliza todas las variables, aún así solo presenta un 0.08 más de coeficiente de determinación respecto al que utiliza las 3 variables con más peso.

Este analisis ha servido para comprender que el alcohol es el componente con más peso a la hora de determinar la calidad del vino y que, en general, a mayor alcohol mayor calidad. Hay que tener en cuenta que el vino que contiene más alcohol del dataset no alcanza los 15º, por lo que no sabemos como se comporta para cantidades superiores a esta. Aún así, la correlación entre alcohol y calidad sigue siendo relativamente baja, por lo que para determinar con exactitud la calidad del vino hacen falta o bien más datos, o bien más variables, o bien usar otros modelos que no sean de regresión.

Contribuciones

Contribuciones	Firma
Investigación previa	GGC, OAA
Redacción de las respuestas	GGC, OAA
Desarrollo código	GGC, OAA