Pràctica 2 - Tipologia i cicle de vida de les dades

Raúl Morcillo López

19/05/2022

Contents

Descripció del dataset

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

Podem observar que la base de dades conté 1599 observacions i 12 variables.

Aquestes variabels son:

- **fixed.acidity** variable numèrica
- **volatile.acidity** variable numèrica
- **citric.acid** variable numèrica
- **residual.sugar** variable numèrica
- **chlorides** variable numèrica
- **free.sulfur.dioxide** variable numèrica
- **total.sulfur.dioxide** variable numèrica
- **density** variable numèrica
- **pH** variable numèrica
- **sulphates** variable numèrica
- **alcohol** variable numèrica
- **quality** variable sencera

Totes les variables numèriques fan referència a les característiques químiques dels vins i la variable sencera fa referència a la qualitat del vi obtinguda amb la mitjana d'almenys 3 experts en vins amb una escala de 0 a 10 on 0 és molt dolent i 10 és molt excel · lent. (font: (**https://archive.ics.uci.edu/ml/datasets/wine+quality**)).

Context del dataset

Aquest fitxer de dades ens serveix per saber si un vi és bo o no. Per això disposem d'una variable, `quality`, que fa referència a la qualitat del vi en una escala de 0 a 10 proporcionada per experts en el món del vi. Tractarem de trobar un model que ens indiqui en funció de les variables de característiques químiques si un vi es pot considerar bo o no.

Càlcul dels estadístics bàsics

```
 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
```

Tractament dels valors perduts

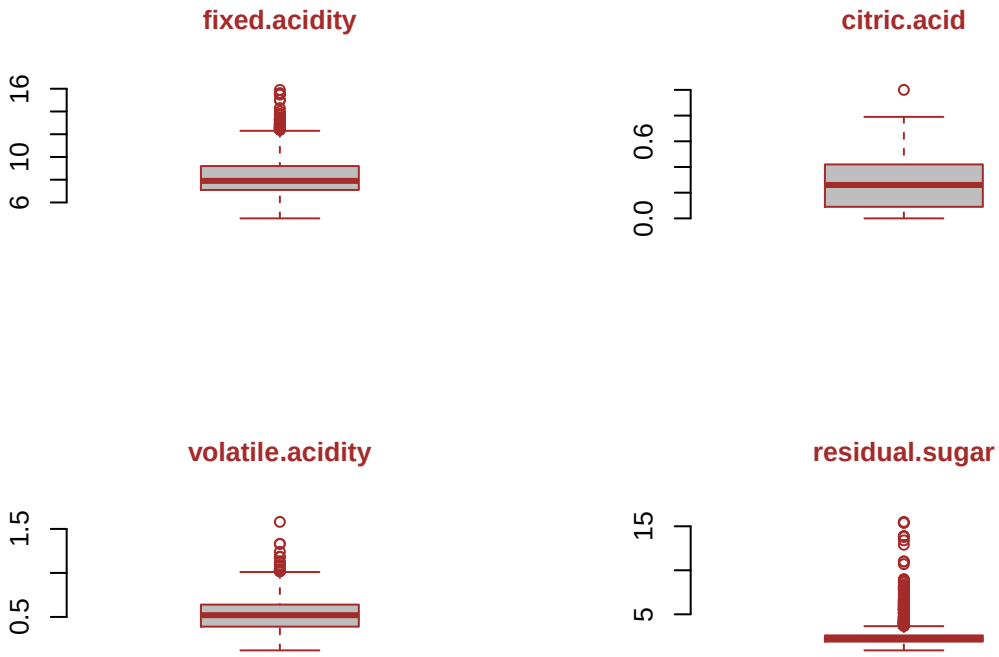
```
Valors missing per alcohol : 0 %
Valors missing per chlorides : 0 %
Valors missing per citric.acid : 0 %
Valors missing per density : 0 %
Valors missing per fixed.acidity : 0 %
Valors missing per free.sulfur.dioxide : 0 %
Valors missing per pH : 0 %
Valors missing per quality : 0 %
Valors missing per residual.sugar : 0 %
Valors missing per sulphates : 0 %
Valors missing per total.sulfur.dioxide : 0 %
Valors missing per volatile.acidity : 0 %
```

Com podem veure aquest fitxer no presenta cap valor perdut així que podrem utilitzar el fitxer sense necessitat d'imputar cap valor o eliminar observacions.

Tractament dels valors extrems

Què és un valor extrem? Es considera un valor extrem aquell valor que es troba allunyat 3 desviacions estàndard de la mitjana. Per poder detectar aquests valors extrems utilitzem el gràfic `boxplot` que ens indica amb un cercle aquells valors que compleixen aquesta condició. Però realment es pot considerar a aquests valors extrems? Doncs depèn de com sigui la seva distribució a la variable. Podem tenir una variable que té un rang entre el valor màxim i el mínim molt gran, però que tots els valors presenten una distribució molt uniforme. Entre tots aquests valors considerats extrems podem considerar verdaderament extrems aquells que s'allunyen de la distribució dels valors. Per exemple si tenim de valors extrems 8,9,10,25 podem considerar un valor extrem pur el 25, ja que s'allunya molt de la distribució i no considerar com a valors extrems el 8,9,10.

Tenint això en compte, amen a veure les variables del fitxer.

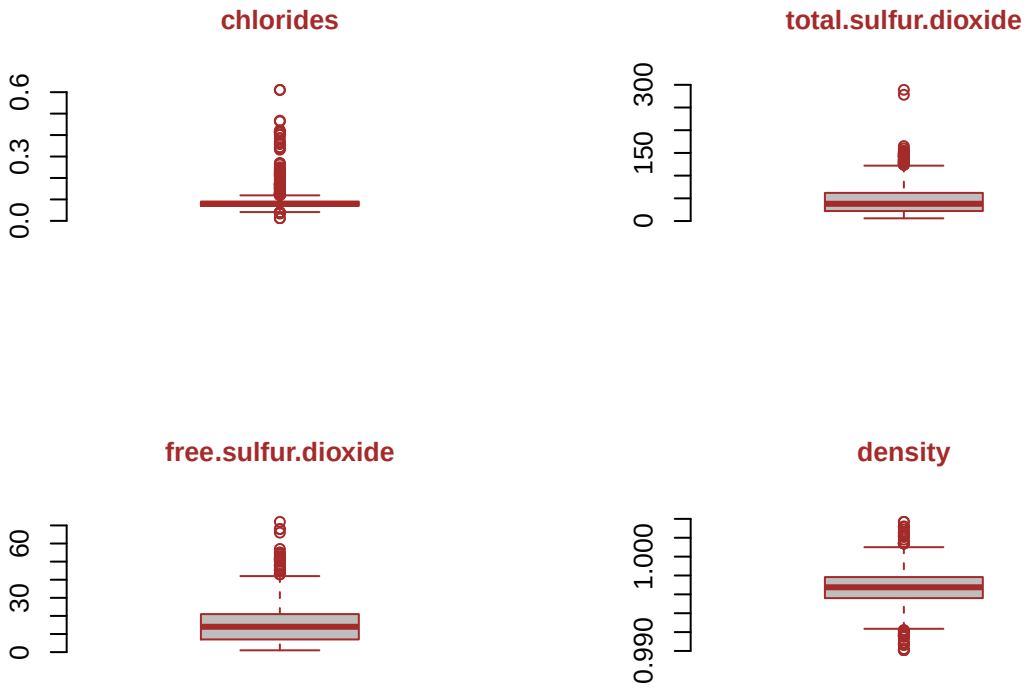



```
Variable fixed.acidity

 num [1:49] 12.8 12.8 15 15 12.5 13.3 13.4 12.4 12.5 13.8 ...

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


Variable volatile.acidity
```

```
 num [1:19] 1.13 1.02 1.07 1.33 1.33 ...

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


Variable citric.acid

 num 1

[1] 1


   0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09  0.1 0.11 0.12 0.13 0.14 0.15
 132   33   50   30   29   20   24   22   33   30   35   15   27   18   21   19
0.16 0.17 0.18 0.19  0.2 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29  0.3 0.31
   9   16   22   21   25   33   27   25   51   27   38   20   19   21   30   30
0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39  0.4 0.41 0.42 0.43 0.44 0.45 0.46 0.47
  32   25   24   13   20   19   14   28   29   16   29   15   23   22   19   18
0.48 0.49  0.5 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59  0.6 0.61 0.62 0.63
  23   68   20   13   17   14   13   12    8    9    9    8    9    2    1   10
0.64 0.65 0.66 0.67 0.68 0.69  0.7 0.71 0.72 0.73 0.74 0.75 0.76 0.78 0.79    1
   9    7   14    2   11    4    2    1    1    3    4    1    3    1    1    1


Variable residual.sugar

 num [1:155] 6.1 6.1 3.8 3.9 4.4 10.7 5.5 5.9 5.9 3.8 ...

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

En aquest set de quatre variables, veient el boxplot i els valors que dona com a valors extrems podem dir que:

- Variable fixed.acidity no presenta valors extrems
- Variable volatile.acidity presenta 1 valor extrem (1.580)
- Variable citric.acid presenta 1 valor extrem (1) Aquesta variable només presenta un valor extrem, en veure la freqüència de la variable veiem que passem de 0.79 a 1. És un salt molt gran i per això el considero valor extrem.
- Variable residual.sugar presenta 11 valors extrems (valors majors de 9.0)

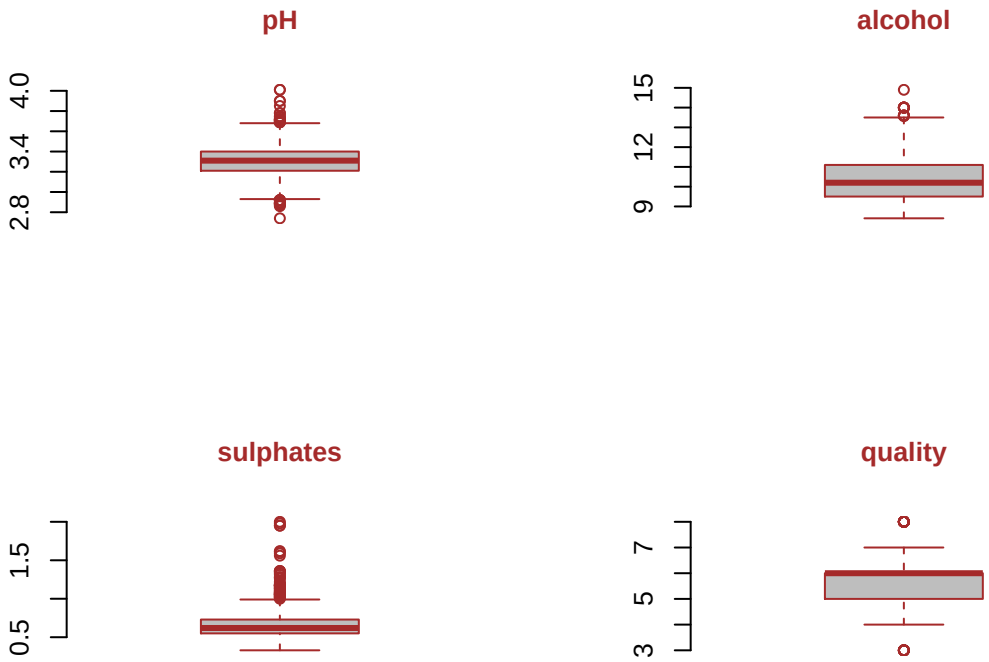



```
Variable chlorides

 num [1:112] 0.176 0.17 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 ...

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


Variable free.sulfur.dioxide

 num [1:30] 52 51 50 68 68 43 47 54 46 45 ...

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

```
Variable total.sulfur.dioxide

 num [1:55] 145 148 136 125 140 136 133 153 134 141 ...

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


Variable density

 num [1:45] 0.992 0.992 1.001 1.002 1.002 ...

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

En aquest set de quatre variables, veient el boxplot i els valors que dona com a valors extrems podem dir que:

- Variable chlorides presenta 9 valors extrems per abaix (<0.04) i 22 valors extrems per dalt (>0.3)
- Variable free.sulfur.dioxide presenta 4 valors extrems (66,68,68,72)
- Variable total.sulfur.dioxide presenta 4 valors extrems (160,165,278,289)
- Variable density no presenta valors extrems




```
Variable pH

 num [1:35] 3.9 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 ...

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


Variable sulphates

 num [1:59] 1.56 1.28 1.08 1.2 1.12 1.28 1.14 1.95 1.22 1.95 ...

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


Variable alcohol

 num [1:13] 14 14 14 14 14.9 14 13.6 13.6 13.6 14 ...
```

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


Variable quality

 num [1:28] 8 8 8 8 8 3 8 8 8 3 ...

 [1] 3 3 3 3 3 3 3 3 3 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8


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

En aquest set de quatre variables, veient el boxplot i els valors que dona com a valors extrems podem dir que:

- Variable pH presenta no presenta valors extrems
- Variable sulphates presenta 8 valors extrems (valors més grans de 1.36)
- Variable alcohol no presenta valors extrems
- Variable quality no presenta valors extrems. Aquesta variable només presenta 6 codis i és cert que del 3 i del 8 hi ha poca mostra però no els consideraria valors extrems.

Un cop tenim localitzats els valors extrems de les variables, els passarem a missing i imputarem la mitjana de la variable. Farem un set de variables nou per tal de no embolicar-nos amb les variables a l'hora de fer anàlisis.

```
Valors missing per alcohol : 0 %
Valors missing per chlorides : 1.94 %
Valors missing per citric.acid : 0.06 %
Valors missing per density : 0 %
Valors missing per fixed.acidity : 0 %
Valors missing per free.sulfur.dioxide : 0.25 %
Valors missing per pH : 0 %
Valors missing per quality : 0 %
Valors missing per residual.sugar : 0.69 %
Valors missing per sulphates : 0.5 %
Valors missing per total.sulfur.dioxide : 0.25 %
Valors missing per volatile.acidity : 0.06 %
```

Un cop hem eliminat els valors extrems veiem que hi ha molt pocs. Ara els imputarem la mitjana de la variable per tal de no tenir valors perduts.

```
Valors missing per alcohol : 0 %
Valors missing per chlorides : 0 %
Valors missing per citric.acid : 0 %
Valors missing per density : 0 %
Valors missing per fixed.acidity : 0 %
Valors missing per free.sulfur.dioxide : 0 %
Valors missing per pH : 0 %
Valors missing per quality : 0 %
Valors missing per residual.sugar : 0 %
Valors missing per sulphates : 0 %
Valors missing per total.sulfur.dioxide : 0 %
Valors missing per volatile.acidity : 0 %
```

Després de fer la imputació disposem d'un fitxer sense els valors extrems considerats.

Variable qualitat del vi

Per tal de poder saber si un vi és bo o no recodificarem la variable `quality` en dos talls. Aquesta variable tot i només tenir valors del 3 al 8 està mesurada en una escala de 0 a 10. Això és un tema una mica subjectiu, però per tema de mostra podem dir que un vi serà bo quan tingui una qualitat superior a 7.

```
Variable quality
```

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

```
Comprovació creació variable
```

```
            3   4   5   6   7   8
  No Vi Bo  10  53 681 638   0   0
  Vi Bo      0   0   0   0 199  18
```

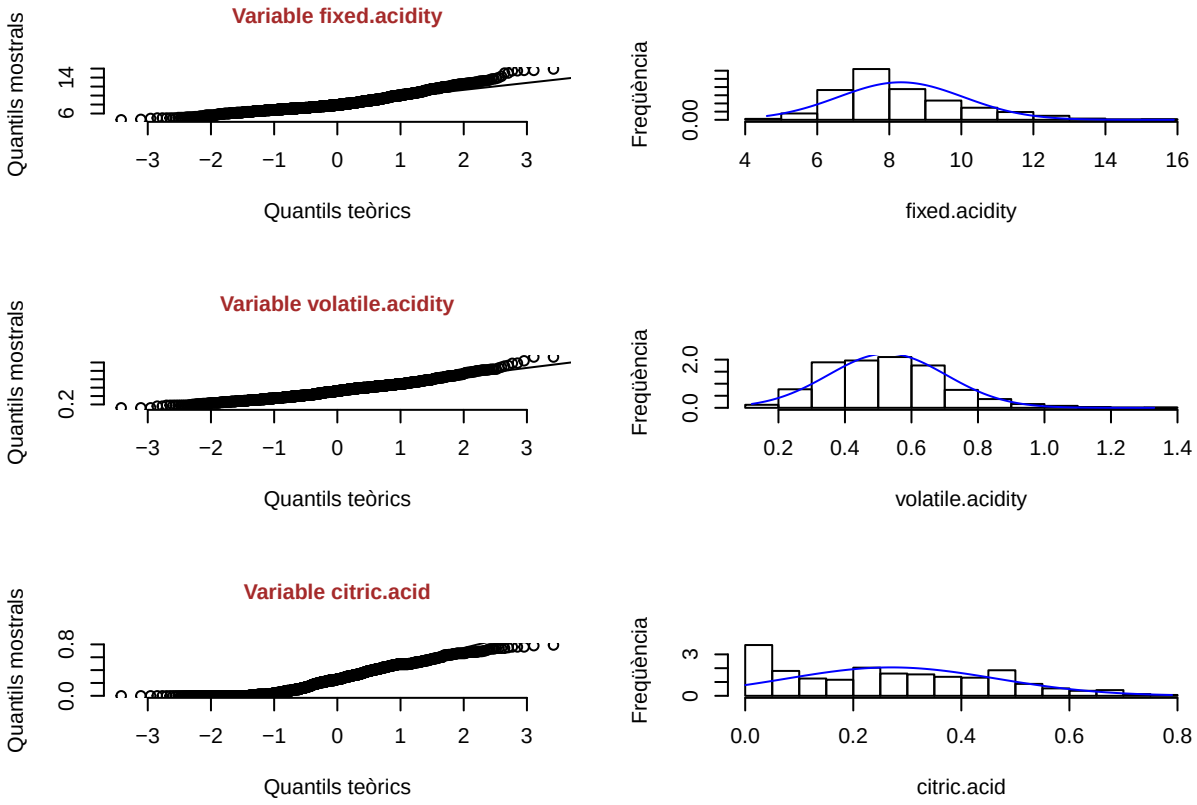
```
'data.frame':   1599 obs. of  13 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
 $ qualityr            : Factor w/ 2 levels "No Vi Bo","Vi Bo": 1 1 1 1 1 1 1 2 2 1 ...
```

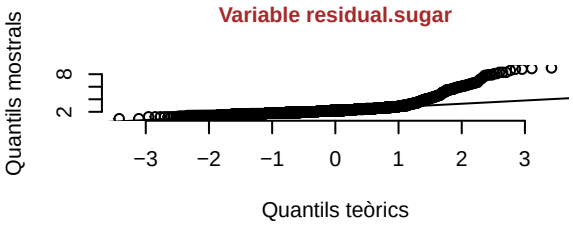
Comprovació de la normalitat

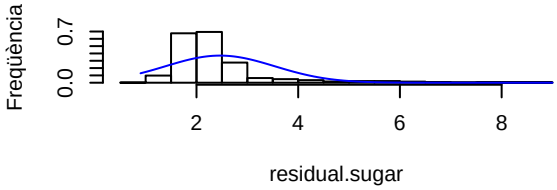
Ara que tenim un fitxer de dades net, comprovarem la normalitat de les variables.

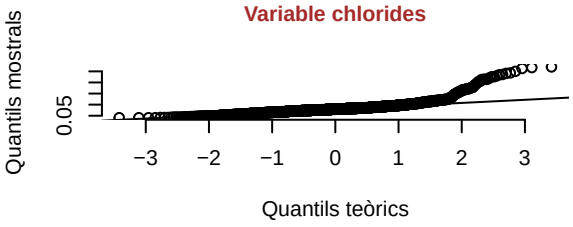
Per comprovar la normalitat de les variables podem utilitzar el test de Shapiro-Wilk. Per aquest test s'assumeix com a hipòtesi nul · la que la distribució és normal, llavors si el p-valor del contrast és inferior al nivell de significació situat en 0.05, rebutjarem la hipòtesi nul · la i tindrem que no segueixen una distribució normal.

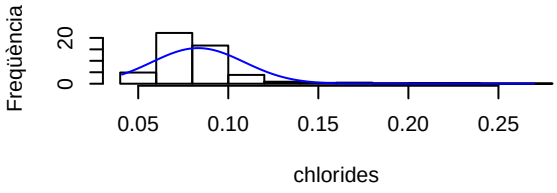
Una altra manera més visual de veure si una variable segueix una distribució normal és amb els gràfics qqnorm (compara els quantils mostrals amb els teòrics) i l'histograma.

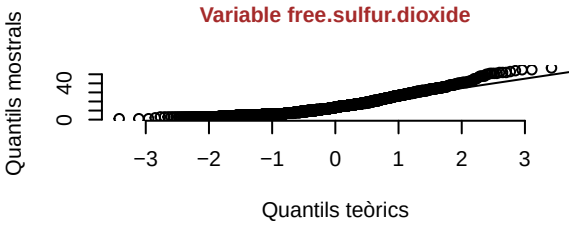


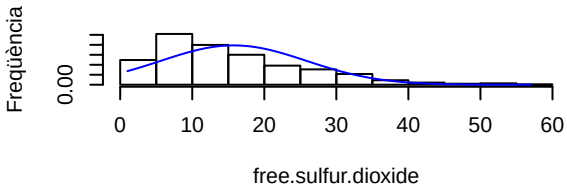

Variable residual.sugar
Quantils mostrals
Quantils teòrics

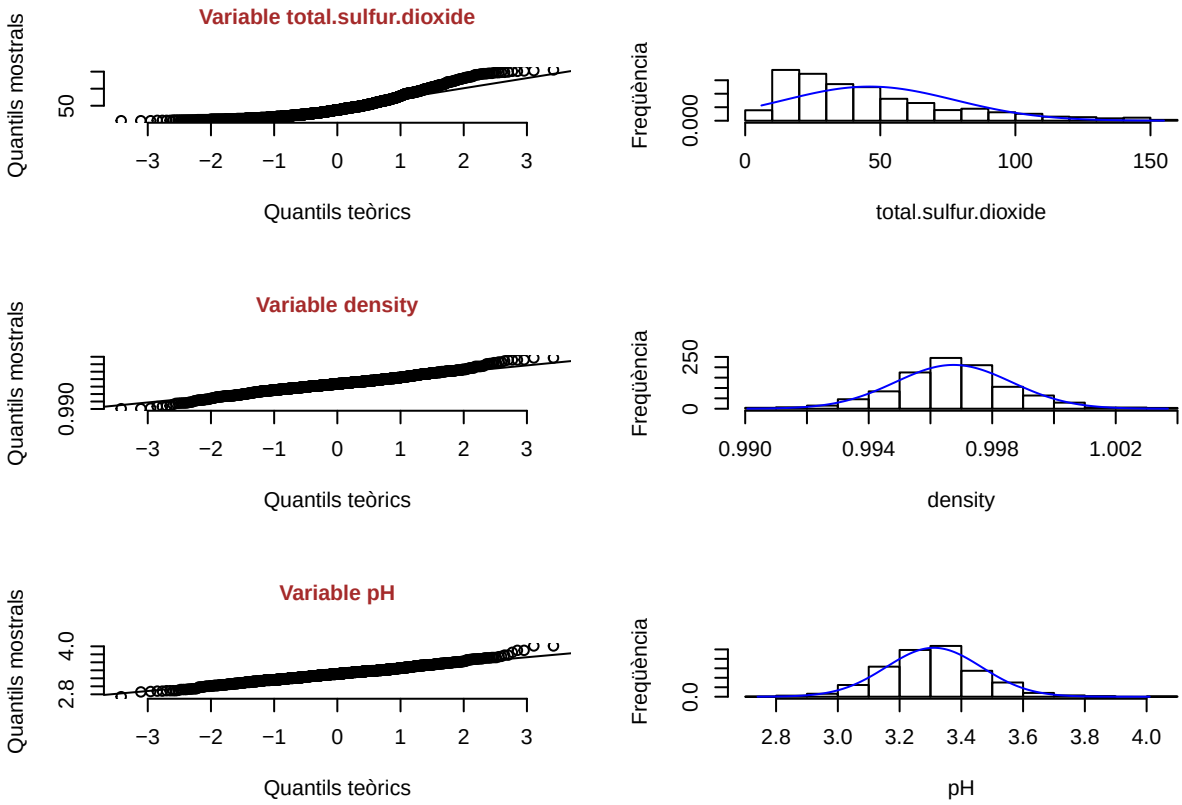

Freqüència
residual.sugar

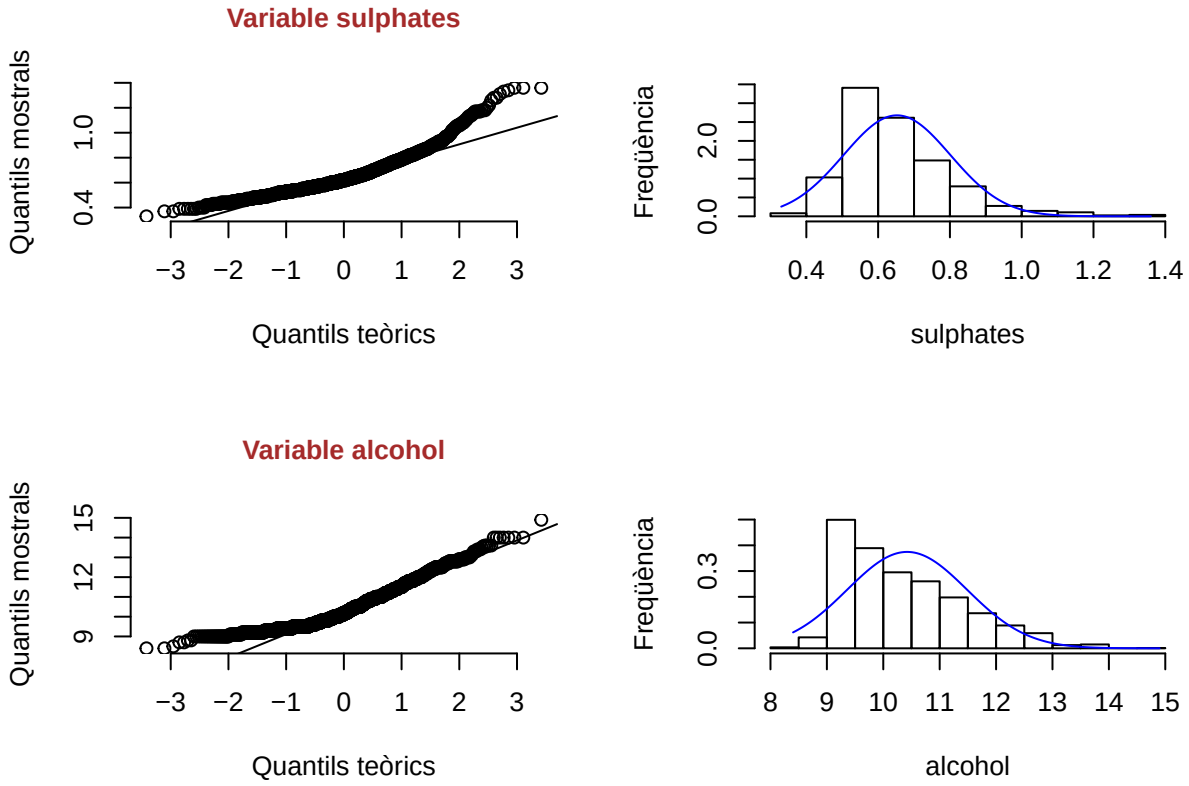

Variable chlorides
Quantils mostrals
Quantils teòrics


Freqüència
chlorides


Variable free.sulfur.dioxide
Quantils mostrals
Quantils teòrics


Freqüència
free.sulfur.dioxide


Variable total.sulfur.dioxide
Quantils mostrals
Quantils teòrics
Freqüència
total.sulfur.dioxide
Variable density
density
Variable pH
pH




```
Test per comprovar la normalitat de les variables

p-value del Shapiro-Wilk Test per fixed.acidity : 1.525012e-24
p-value del Shapiro-Wilk Test per volatile.acidity : 1.746552e-14
p-value del Shapiro-Wilk Test per citric.acid : 7.173338e-22
p-value del Shapiro-Wilk Test per residual.sugar : 2.472359e-47
p-value del Shapiro-Wilk Test per chlorides : 5.040096e-44
p-value del Shapiro-Wilk Test per free.sulfur.dioxide : 2.781797e-29
p-value del Shapiro-Wilk Test per total.sulfur.dioxide : 4.09305e-32
p-value del Shapiro-Wilk Test per density : 1.936053e-08
p-value del Shapiro-Wilk Test per pH : 1.712237e-06
p-value del Shapiro-Wilk Test per sulphates : 2.242804e-29
p-value del Shapiro-Wilk Test per alcohol : 6.644057e-27
```

Donant un cop d'ull als gràfics podem veure que hi ha variables que poden semblar que segueixen una distribució normal, ja que en el gràfic qqnorm sembla que les dades s'ajusten a la recta de referència i l'histograma sembla una campana com per exemple `volatile.acidity`, `density` o `pH`. Quan el resultat del test de Shapiro-Wilk podem veure que tots els p-valors son inferiors al nivell de significació que està situat en 0.05. Això vol dir que hem de rebutjar la hipòtesi nul · la i, per tant, arribar a la conclusió que les variables no segueixen una distribució normal.

De totes maneres com tenim una mostra prou gran (n = 1599) pel teorema central del límit podem considerar que les dades del fitxer segueixen una distribució normal.

Comprovació de la homogeneïtat de la variància

L'homogeneïtat de la variància consisteix a veure si hi ha igualtat de variàncies entre grups de dades. Com tenim la variable 'quality' en dos grups mirarem això entre aquests dos grups, ja que aquesta variable és la que utilitzarem per fer els models estadístics. Com les variables no segueixen una distribució normal farem servir el test de Fligner-Killeen on la hipòtesi nul · la indica igualtat de variàncies.

```
Test per comprovar l'homogeneïtat de la variància de les variables enfront qualityr
```

```
p-value del Fligner-Killen Test per fixed.acidity : 4.785198e-05
p-value del Fligner-Killen Test per volatile.acidity : 0.0001090342
p-value del Fligner-Killen Test per citric.acid : 0.935506
p-value del Fligner-Killen Test per residual.sugar : 0.01352106
p-value del Fligner-Killen Test per chlorides : 0.2868306
p-value del Fligner-Killen Test per free.sulfur.dioxide : 0.03916107
p-value del Fligner-Killen Test per total.sulfur.dioxide : 1.483642e-10
p-value del Fligner-Killen Test per density : 3.615554e-05
p-value del Fligner-Killen Test per pH : 0.770414
p-value del Fligner-Killen Test per sulphates : 0.4929434
p-value del Fligner-Killen Test per alcohol : 0.08011055
```

Mirant els resultats del test de Fligner-Killen podem veure que les variables amb un p-valor inferior al nivell de significació (0.05) rebutgen la hipòtesi nul · la i, per tant, presenten variàncies estadísticament diferents per als diferents grups de la variable `qualityr`. Aquestes variables són `fixed.acidity`, `volatile.acidity`, `residual.sugar`, `chlorides`, `free.sulfur.dioxide`, `total.sulfur.dioxide` i `density`. Per contra les variables `citric.acid`, `pH`, `sulphates` i `alcohol` presenten un p-valor superior al nivell de significació i, per tant, estadísticament les variàncies són iguals.

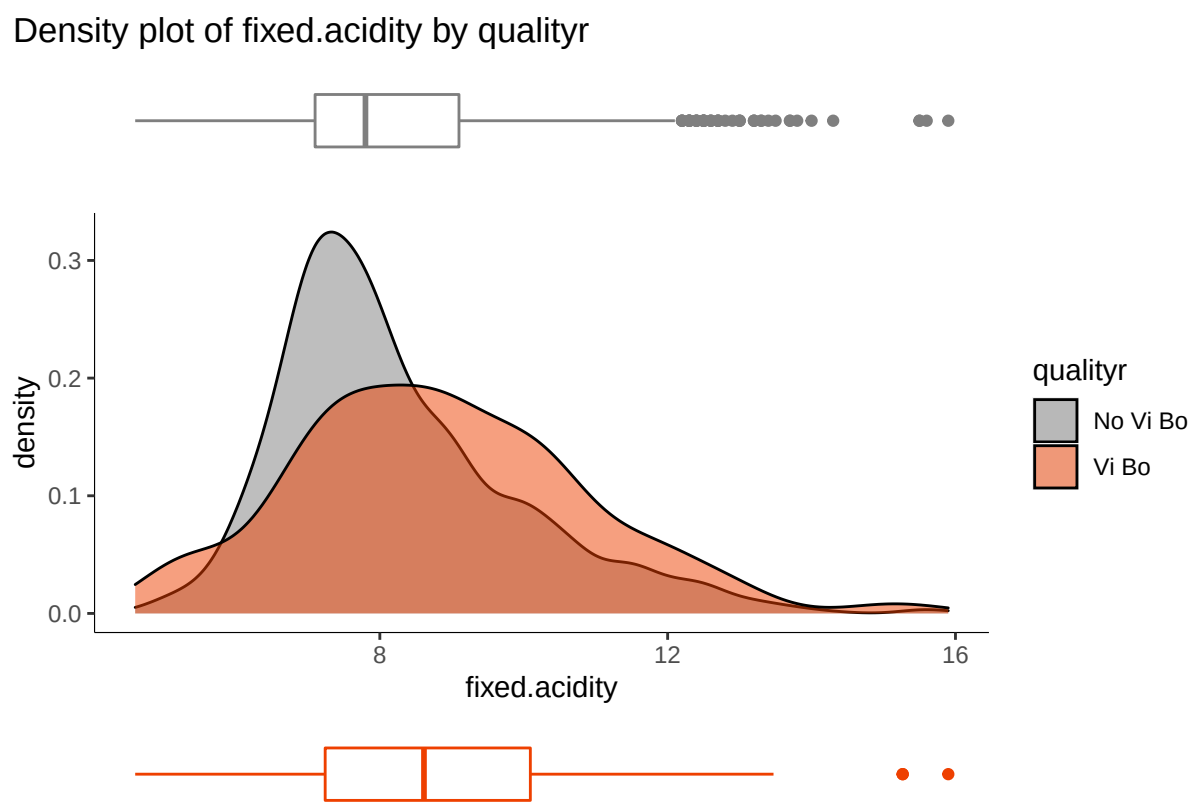
Contrast d'hipòtesi

Com volem saber si un vi es considerat bo o no, podem analitzar les variables en funció de la qualitat per veure si aquestes influeixen a l'hora de sapiguer si un vi es bo o no. Per fer això utilitzarem un contrast d'hipòtesi on volem veure si hi ha diferencies en funció de la variable qualityr (Vi bo, No vi bo). Aquest contrast tindrà una hipòtesi nul · la on la mitjana de la variable és igual en funció de qualityr.

Test del contrat d'hipòtesi Mann-Whitney-Wilcoxon

Per a aquestes variables utilitzen el test de `Mann-Whitney-Wilcoxon`, ja que hem vist abans que les variàncies són estadísticament diferents. Primer veurem aquestes variables gràficament i després aplicarem el test.

Variable fixed.acidity



Gràficament, podem veure que les mitjanes de la variable són semblants, però hi ha concentració de valors baixos quant el vi no és bo. Per tant, es pot afirmar que la mitjana entre el vi bo és més gran que entre el vi no bo?

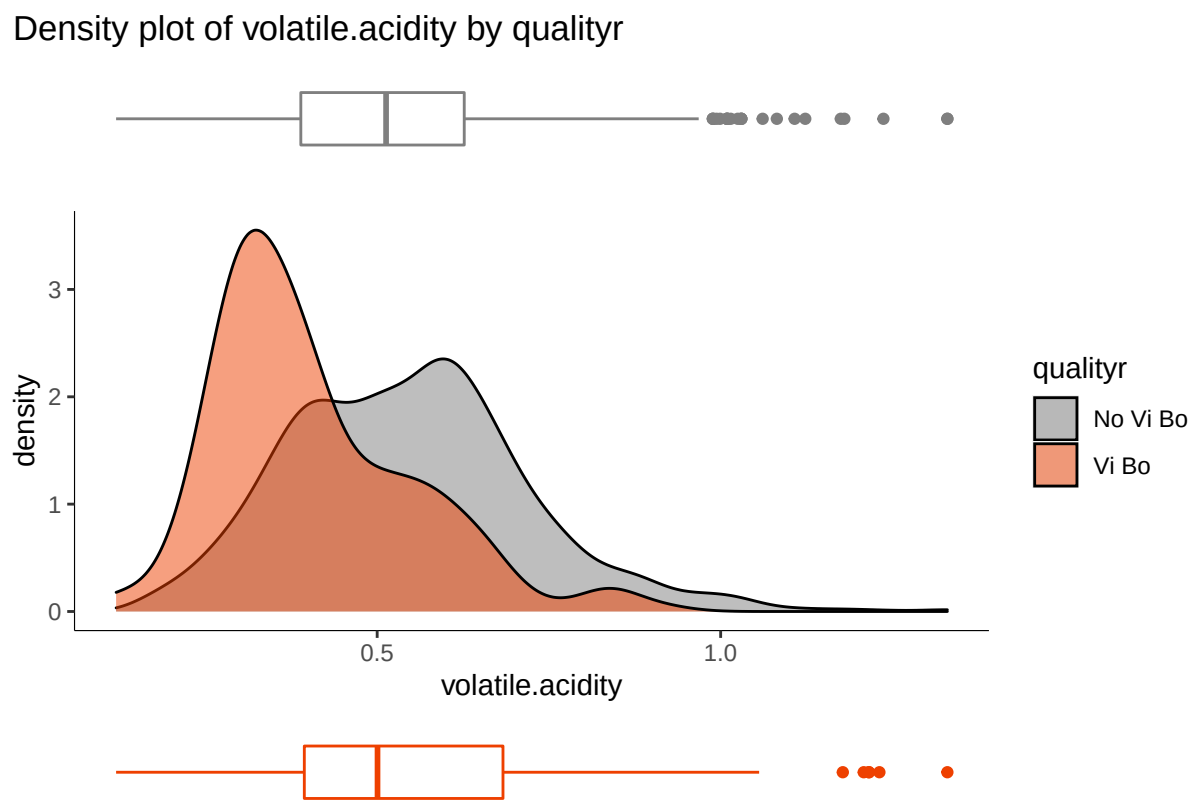
$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

$$H_0 : \mu_{bo} - \mu_{no\,bo} > 0$$

```
    Wilcoxon rank sum test with continuity correction

data:  df2$fixed.acidity[df2$qualityr == "Vi Bo"] and df2$fixed.acidity[df2$qualityr == "No Vi Bo"]
W = 181422, p-value = 3.2e-07
alternative hypothesis: true location shift is greater than 0
```

Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que el Vi bo té un nivell de fixed.acidity més alt que el vi no bo.

Variable volatile.acidity



Gràficament, podem veure que les mitjanes de la variable són semblants, però hi ha concentració de valors baixos quant el vi és bo. Per tant, es pot afirmar que la mitjana entre el vi bo és més petita que entre el vi no bo?

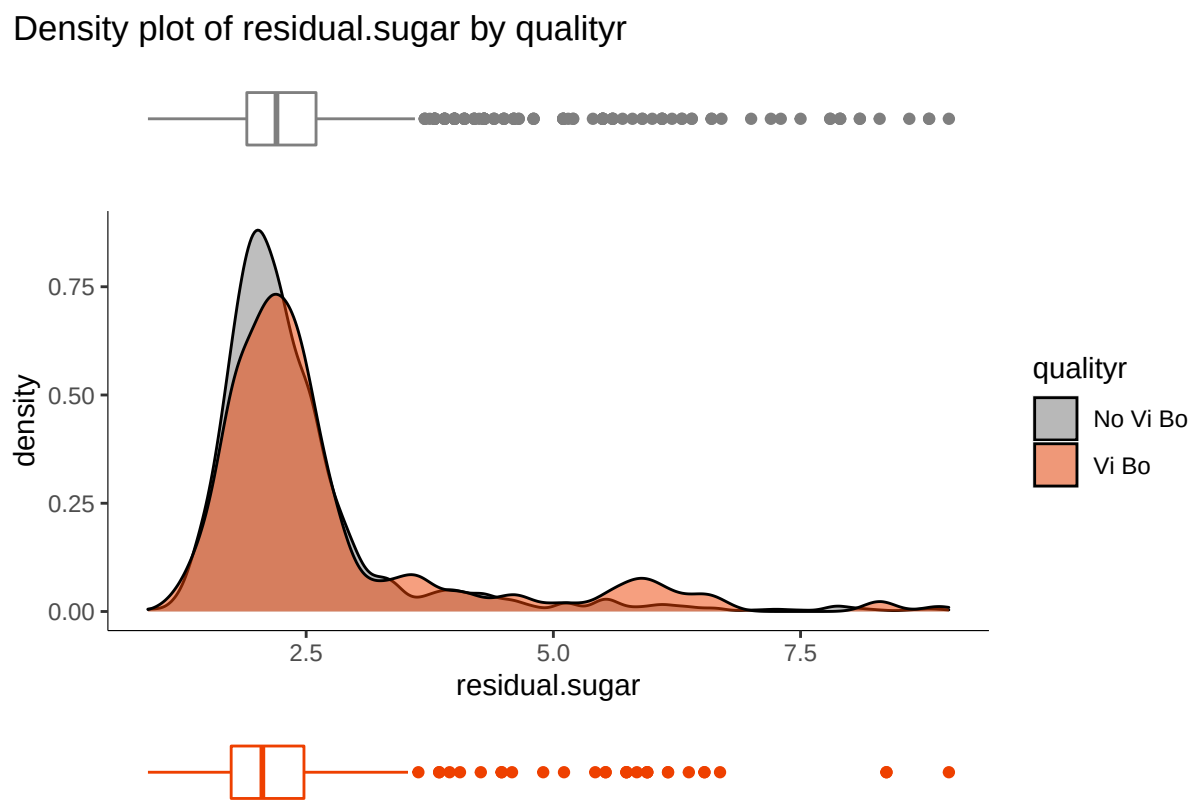
$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

$$H_0 : \mu_{bo} - \mu_{no\,bo} < 0$$

```
    Wilcoxon rank sum test with continuity correction

data:  df2$volatile.acidity[df2$qualityr == "Vi Bo"] and df2$volatile.acidity[df2$qualityr == "No Vi Bo
W = 76489, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que el Vi bo té un nivell de volatile.acidity més baix que el vi no bo.

Variable residual.sugar



Gràficament, podem veure que les mitjanes de la variable són semblants i la distribució de la variable també.

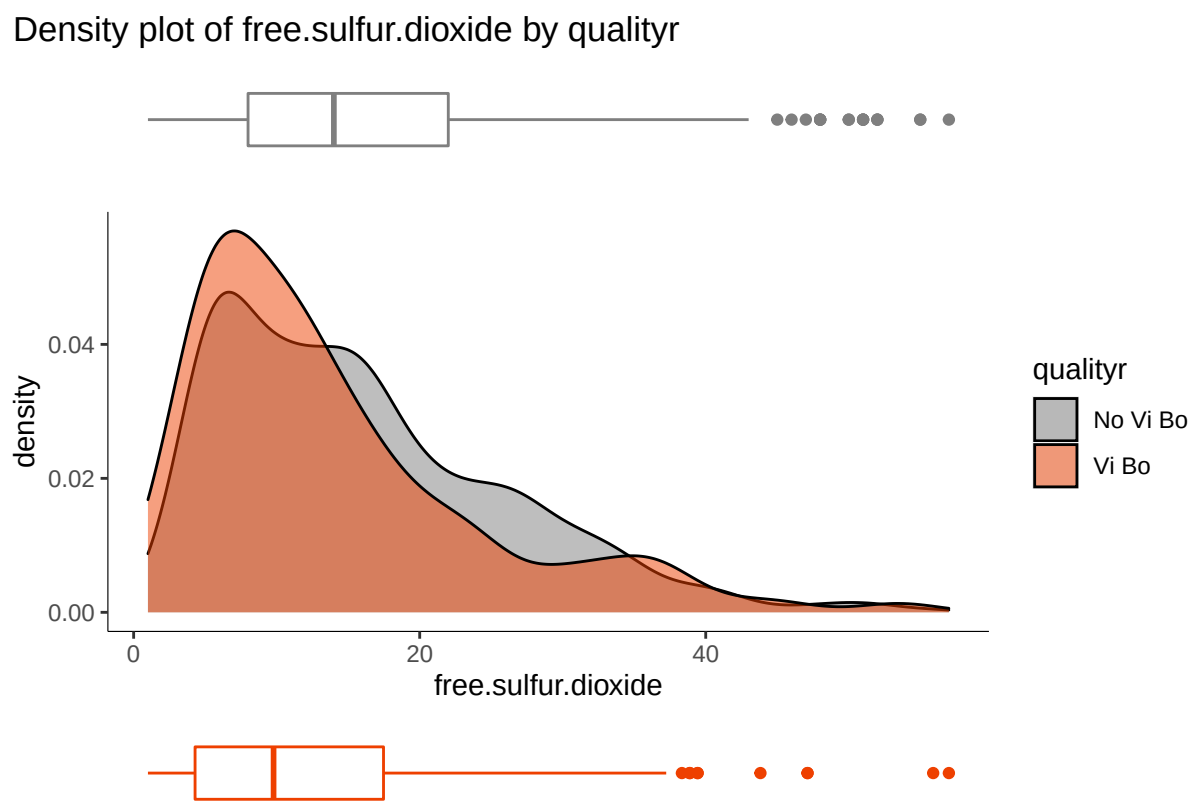
$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

$$H_0 : \mu_{bo} - \mu_{no\,bo} <> 0$$

```
    Wilcoxon rank sum test with continuity correction

data:  df2$residual.sugar[df2$qualityr == "Vi Bo"] and df2$residual.sugar[df2$qualityr == "No Vi Bo"]
W = 166050, p-value = 0.01073
alternative hypothesis: true location shift is not equal to 0
```

Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que les mitjanes de residual.sugar són diferents per a cada tipus de vi.

Variable free.sulfur.dioxide



Gràficament, podem veure que la distribució de la variable es semblant però s'aprecia que a valors mes petits el Vi és considerat bo. Per tant, es pot afirmar que la mitjana entre el vi bo és més petita que entre el vi no bo?

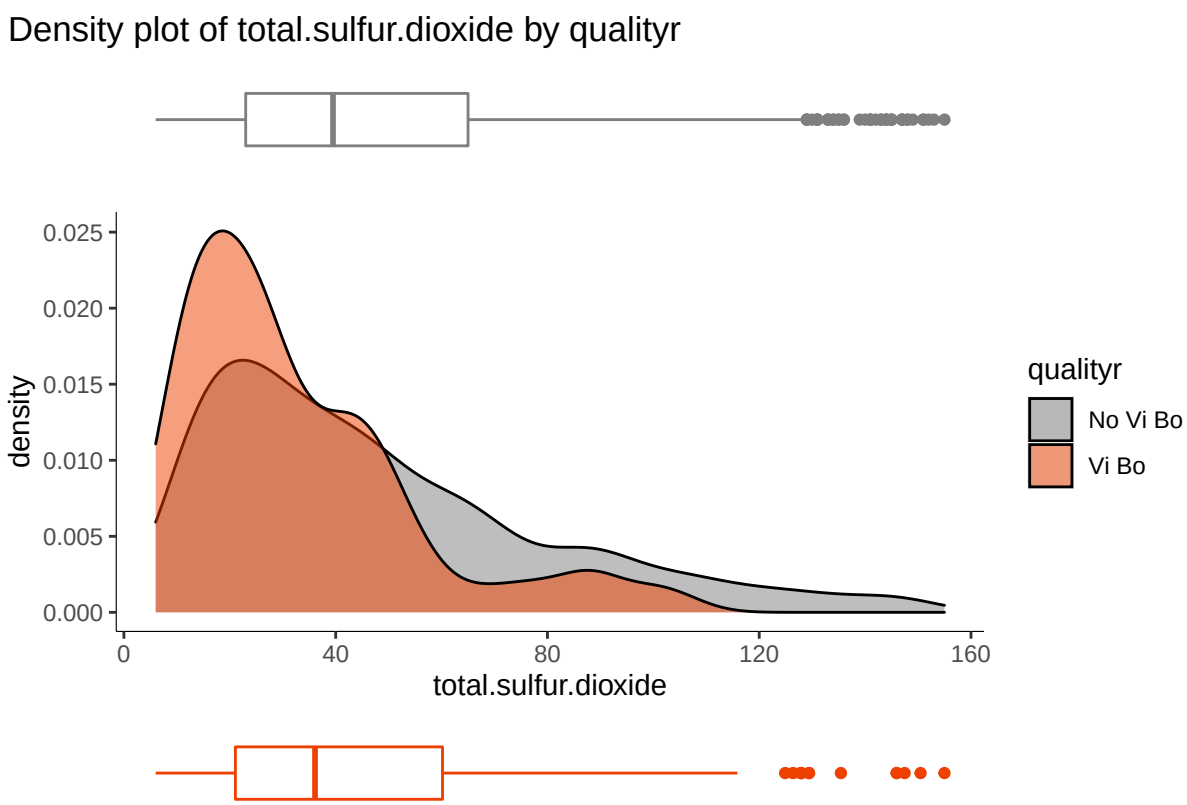
$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

$$H_0 : \mu_{bo} - \mu_{no\,bo} < 0$$

```
    Wilcoxon rank sum test with continuity correction

data:  df2$free.sulfur.dioxide[df2$qualityr == "Vi Bo"] and df2$free.sulfur.dioxide[df2$qualityr == "No
W = 127734, p-value = 0.0002191
alternative hypothesis: true location shift is less than 0
```

Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que a menor valor de free.sulfur.dioxide el vi és considerat bo.

Variable total.sulfur.dioxide



Gràficament, podem veure que la distribució de la variable es semblant però s'aprecia que a valors mes petits el Vi és considerat bo. Per tant, es pot afirmar que la mitjana entre el vi bo és més petita que entre el vi no bo?

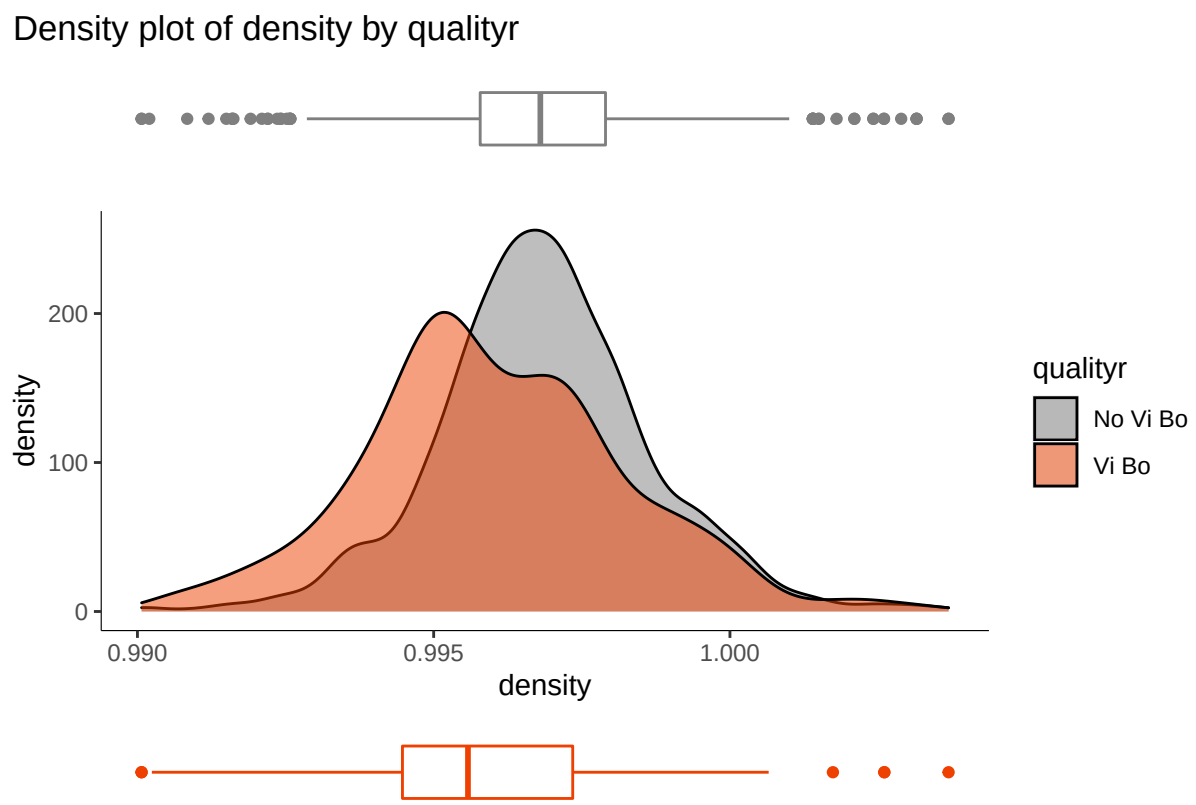
$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

$$H_0 : \mu_{bo} - \mu_{no\,bo} < 0$$

```
    Wilcoxon rank sum test with continuity correction

data:  df2$total.sulfur.dioxide[df2$qualityr == "Vi Bo"] and df2$total.sulfur.dioxide[df2$qualityr == "
W = 105426, p-value = 9.519e-13
alternative hypothesis: true location shift is less than 0
```

Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que a menor valor de total.sulfur.dioxide el vi és considerat bo.

Variable density



Gràficament, podem veure que la distribució de la variable es semblant però s'aprecia que a valors mes petits el Vi és considerat bo. Per tant, es pot afirmar que la mitjana entre el vi bo és més petita que entre el vi no bo?

$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

$$H_0 : \mu_{bo} - \mu_{no\,bo} < 0$$

```
    Wilcoxon rank sum test with continuity correction

data:  df2$density[df2$qualityr == "Vi Bo"] and df2$density[df2$qualityr == "No Vi Bo"]
W = 111741, p-value = 7.6e-10
alternative hypothesis: true location shift is less than 0
```

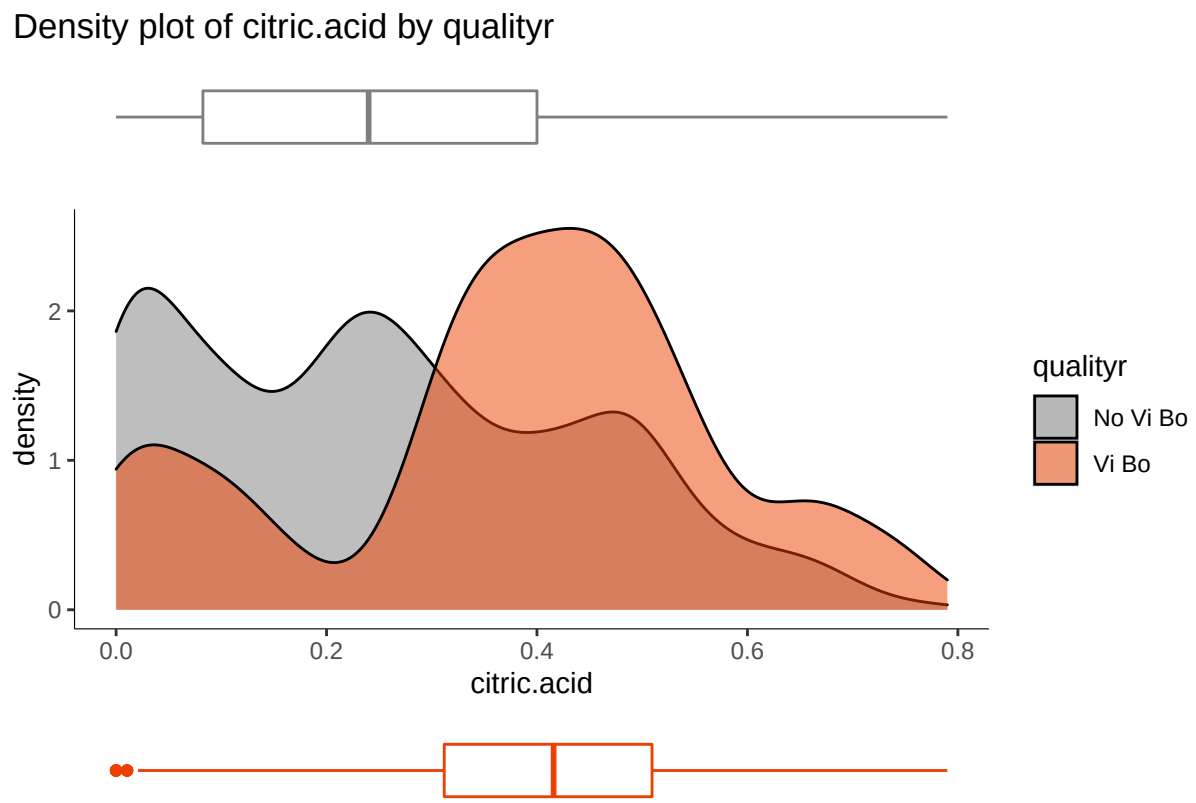
Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que a menor valor de density el vi és considerat bo.

Test del contrat d'hipòtesi t-Student

Per a aquestes variables utilitzem el test `t-Student`, ja que hem vist abans que les variàncies són estadísticament iguals. Tot i que hem vist que els variables no segueixen una distribució normal, pel teorema central del límit podem fer aquest supòsit. Per tant, amb distribució normal i variàncies iguals podem fer servir aquest test.

Primer veurem aquestes variables gràficament i després aplicarem el test.

Variable citric.acid



Gràficament es pot veure que per valors mes alts el vi es considerat bo. Per tant, es pot afirmar que la mitjana entre el vi bo és més gran que entre el vi no bo?

$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

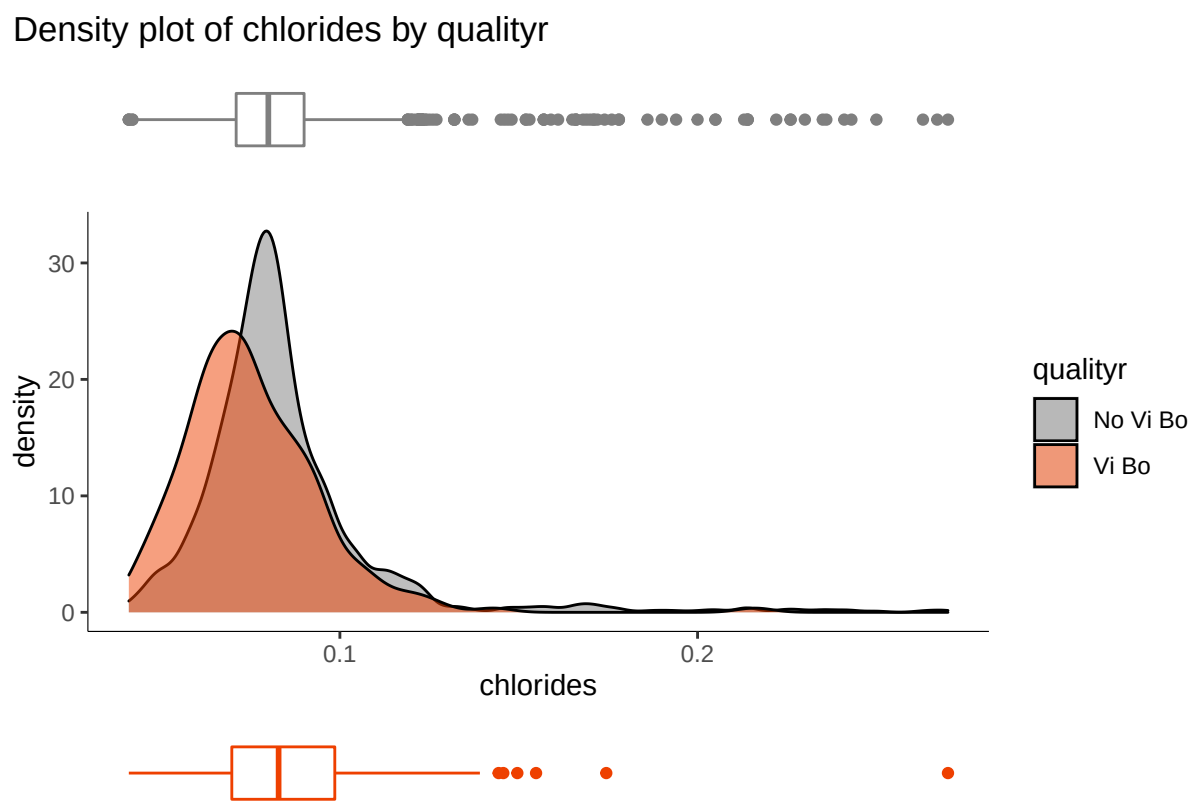
$$H_0 : \mu_{bo} - \mu_{no\,bo} > 0$$

```
    Welch Two Sample t-test

data:  df2$citric.acid[df2$qualityr == "Vi Bo"] and df2$citric.acid[df2$qualityr == "No Vi Bo"]
t = 8.6713, df = 283.57, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
```

```
95 percent confidence interval:
 0.09928322        Inf
sample estimates:
mean of x mean of y
0.3764977 0.2538788
```

Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que a major valor de citric.acid el vi és considerat bo.

Variable chlorides



Gràficament, podem veure que la distribució de la variable es semblant. Per tant, es pot afirmar que la mitjana entre el vi bo igual que entre el vi no bo?

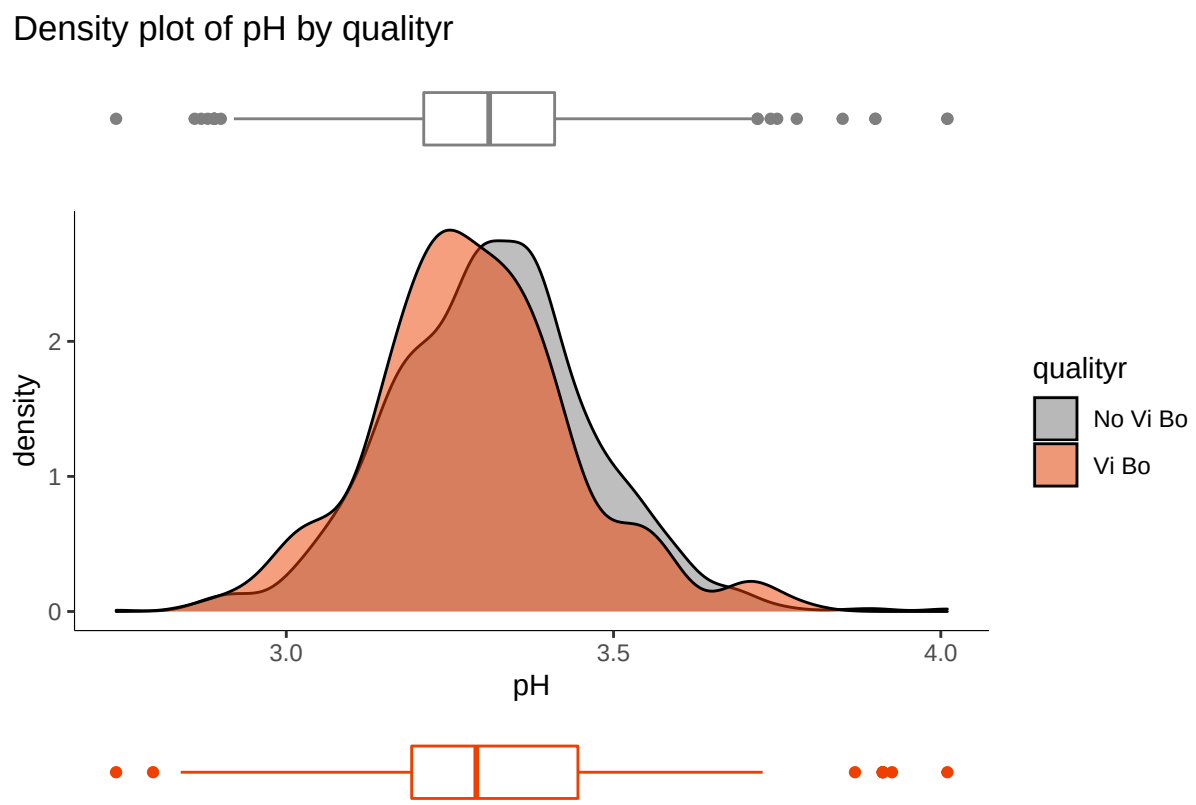
$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

$$H_0 : \mu_{bo} - \mu_{no\,bo} <> 0$$

```
    Welch Two Sample t-test

data:  df2$chlorides[df2$qualityr == "Vi Bo"] and df2$chlorides[df2$qualityr == "No Vi Bo"]
t = -5.8438, df = 345, p-value = 1.182e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.011941920 -0.005927549
sample estimates:
 mean of x  mean of y
0.07551122 0.08444595
```

Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que les mitjanes de chlorides són diferents per grup.

Variable pH



Gràficament, podem veure que la distribució de la variable es semblant però es podria dir que amb un pH mes petit el vi es considerat bo. Per tant, es pot afirmar que la mitjana entre el vi bo és més petita que entre el vi no bo?

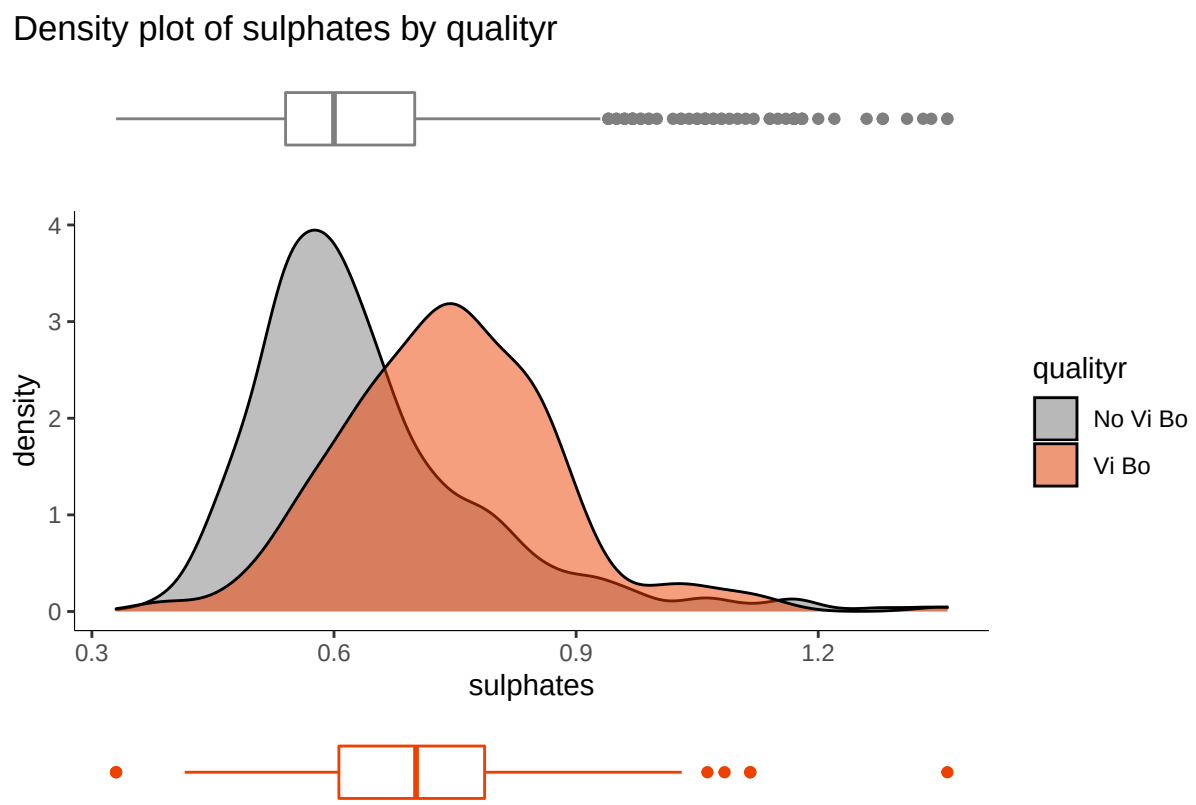
$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

$$H_0 : \mu_{bo} - \mu_{no\,bo} < 0$$

```
    Welch Two Sample t-test

data:  df2$pH[df2$qualityr == "Vi Bo"] and df2$pH[df2$qualityr == "No Vi Bo"]
t = -2.2892, df = 287.71, p-value = 0.01139
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
         -Inf -0.007206495
sample estimates:
mean of x mean of y
 3.288802  3.314616
```

Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que a menor valor de pH es considera que el vi és bo.

Variable sulphates



Gràficament, podem veure que hi ha diferencies en la distribució de la variable, veient-se que a major valor de sulphates el vi és considerat bo. Per tant, es pot afirmar que la mitjana entre el vi bo és més gran que entre el vi no bo?

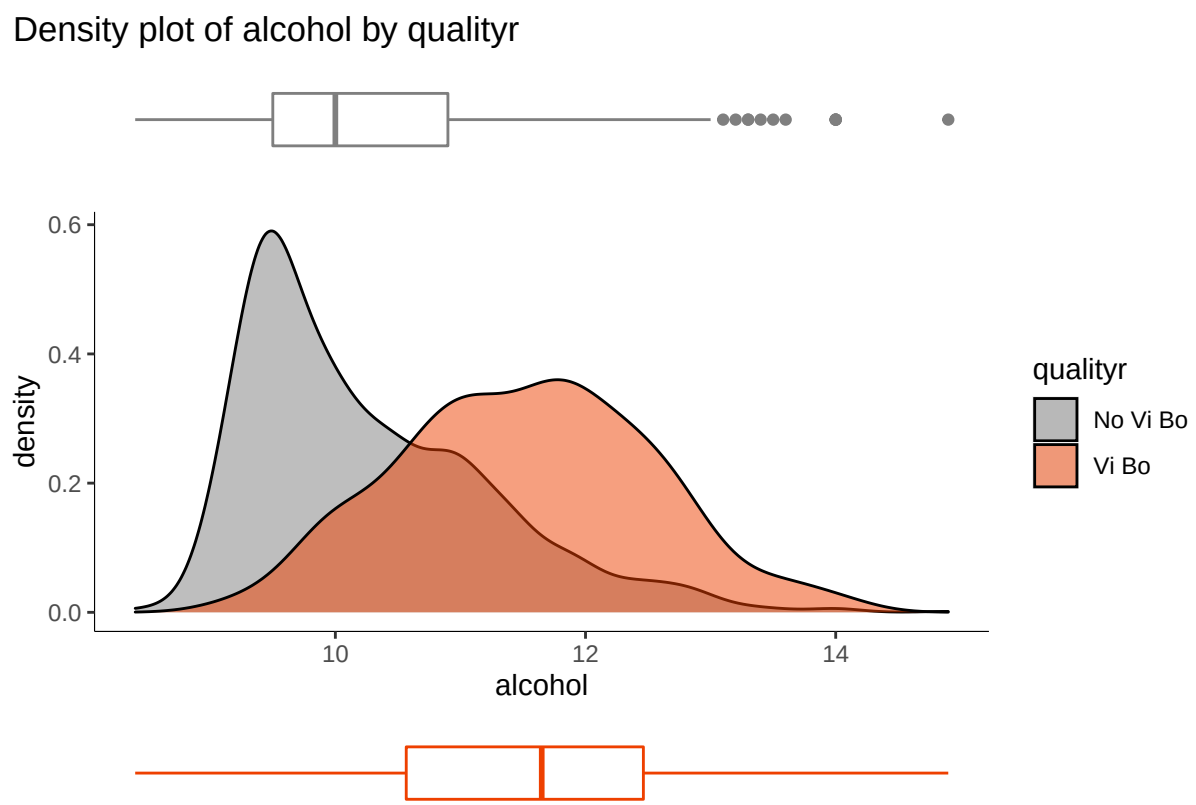
$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

$$H_0 : \mu_{bo} - \mu_{no\,bo} > 0$$

```
    Welch Two Sample t-test

data:  df2$sulphates[df2$qualityr == "Vi Bo"] and df2$sulphates[df2$qualityr == "No Vi Bo"]
t = 10.617, df = 302.55, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.0888891       Inf
sample estimates:
mean of x mean of y
0.7434562 0.6382127
```

Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que a major valor de sulphates es considera que el vi és bo.

Variable alcohol



Gràficament, podem veure que hi ha diferencies en la distribució de la variable, veient-se que a major valor de alcohol el vi és considerat bo. Per tant, es pot afirmar que la mitjana entre el vi bo és més gran que entre el vi no bo?

$$H_0 : \mu_{bo} - \mu_{no\,bo} = 0$$

$$H_0 : \mu_{bo} - \mu_{no\,bo} > 0$$

```
    Welch Two Sample t-test

data:  df2$alcohol[df2$qualityr == "Vi Bo"] and df2$alcohol[df2$qualityr == "No Vi Bo"]
t = 17.45, df = 283.78, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.147195      Inf
sample estimates:
mean of x mean of y
 11.51805  10.25104
```

Com el p-value és inferior al nivell de significació (0.05) hem de rebutjar la hipòtesi nul · la d'igualtat i concloure que a major valor de alcohol es considera que el vi és bo.

Conclusió

Després de fer els contrasts d'hipòtesi saber si hi ha diferencies en les variables a l'hora de considerar un vi bo es pot dir que un vi és considerat bo pels experts (qualificació de 7,8,9,10) si conté valors alts de fixed.acidity, citric.acid, sulphates i alcohol. En canvi, un vi serà considerat dolent pels experts (qualificació 0,1,2,3,4,5,6) si conté valors alts de volatile.acidity, free.sulfur.dioxide, total.sulfur.dioxide, density i pH.

Abans de passar a la següent secció, guardarem les dades processades que utilitzarem per modelar en el fitxer `winequality-red_net.csv`.

Model de regressió logística

Per tal de poder fer un model de regressió logística que ens pugui dir si un vi és bo o no generarem dos fitxers a partir del fitxer gran. Un fitxer per a entrenar el model (train - 70% de la mostra) i un fitxer per a testejar-lo (test - 30% de la mostra).

Fitxer train - 70% de la mostra

```
'data.frame':   1119 obs. of  13 variables:
 $ fixed.acidity       : num  7 8 8 7.9 9.8 9.5 9.6 7.8 7 8.2 ...
 $ volatile.acidity    : num  0.43 0.59 0.43 0.4 0.5 0.59 0.77 0.34 0.5 0.6 ...
 $ citric.acid         : num  0.02 0.05 0.36 0.29 0.34 0.44 0.12 0.37 0.14 0.17 ...
 $ residual.sugar      : num  1.9 2 2.3 1.8 2.3 2.3 2.9 2 1.8 2.3 ...
 $ chlorides           : num  0.08 0.089 0.075 0.157 0.094 0.071 0.082 0.082 0.078 0.072 ...
 $ free.sulfur.dioxide : num  15 12 10 1 10 21 30 24 10 11 ...
 $ total.sulfur.dioxide: num  28 32 48 44 45 68 74 58 23 73 ...
 $ density             : num  0.995 0.997 0.998 0.997 0.999 ...
 $ pH                  : num  3.35 3.36 3.34 3.3 3.24 3.46 3.3 3.34 3.53 3.2 ...
 $ sulphates           : num  0.81 0.61 0.46 0.92 0.6 0.63 0.64 0.59 0.61 0.45 ...
 $ alcohol             : num  10.6 10 9.4 9.5 9.7 9.5 10.4 9.4 10.4 9.3 ...
 $ quality             : int  6 5 5 6 7 5 6 6 5 5 ...
 $ qualityr            : Factor w/ 2 levels "No Vi Bo","Vi Bo": 1 1 1 1 2 1 1 1 1 1 ...
```

Fitxer test - 30% de la mostra

```
'data.frame':   480 obs. of  13 variables:
 $ fixed.acidity       : num  7.9 7.8 5.6 8.5 8.5 7.8 6.9 8.1 7.3 8.8 ...
 $ volatile.acidity    : num  0.6 0.58 0.615 0.28 0.49 0.645 0.605 0.38 0.45 0.61 ...
 $ citric.acid         : num  0.06 0.02 0 0.56 0.11 0 0.12 0.28 0.36 0.3 ...
 $ residual.sugar      : num  1.6 2 1.6 1.8 2.3 ...
 $ chlorides           : num  0.069 0.073 0.089 0.092 0.084 0.082 0.073 0.066 0.074 0.088 ...
 $ free.sulfur.dioxide : num  15 9 16 35 9 8 40 13 12 17 ...
 $ total.sulfur.dioxide: num  59 18 59 103 67 16 83 30 87 46 ...
 $ density             : num  0.996 0.997 0.994 0.997 0.997 ...
 $ pH                  : num  3.3 3.36 3.58 3.3 3.17 3.38 3.45 3.23 3.33 3.26 ...
 $ sulphates           : num  0.46 0.57 0.52 0.75 0.53 0.59 0.52 0.73 0.83 0.51 ...
 $ alcohol             : num  9.4 9.5 9.9 10.5 9.4 9.8 9.4 9.7 10.5 9.3 ...
 $ quality             : int  5 7 5 7 5 6 6 7 5 4 ...
 $ qualityr            : Factor w/ 2 levels "No Vi Bo","Vi Bo": 1 2 1 2 1 1 1 2 1 1 ...
```

Un cop fets els fitxers comprovarem que la variable dependent (qualityr) no es trobi esbiaixada, és a dir, tingui una proporció similar.

```
Fitxer train - 70% de la mostra


 No Vi Bo     Vi Bo
0.8686327 0.1313673



Fitxer test - 30% de la mostra


 No Vi Bo     Vi Bo
0.8541667 0.1458333
```

Com es pot observar la variable independent es troba en una proporció similar de vi bo (13% vs 14%).

Entrenament del model

```
 Entrenament del model


Call:
glm(formula = as.factor(qualityr2) ~ fixed.acidity + volatile.acidity +
    citric.acid + residual.sugar + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + density + pH + sulphates + alcohol,
    family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2220  -0.3804  -0.1919  -0.1008   2.9424

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           3.714e+02  1.406e+02   2.641 0.008256 **
fixed.acidity         3.915e-01  1.563e-01   2.505 0.012251 *
volatile.acidity     -1.656e+00  9.059e-01  -1.828 0.067505 .
citric.acid           8.199e-01  1.020e+00   0.804 0.421411
residual.sugar        4.211e-01  1.128e-01   3.734 0.000189 ***
chlorides            -1.049e+01  5.162e+00  -2.032 0.042175 *
free.sulfur.dioxide   1.874e-03  1.649e-02   0.114 0.909510
total.sulfur.dioxide -1.994e-02  7.018e-03  -2.841 0.004495 **
density              -3.911e+02  1.434e+02  -2.727 0.006387 **
pH                    8.105e-01  1.207e+00   0.671 0.502058
sulphates             4.597e+00  7.647e-01   6.011 1.84e-09 ***
alcohol               7.554e-01  1.682e-01   4.491 7.08e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 870.53  on 1118  degrees of freedom
Residual deviance: 560.11  on 1107  degrees of freedom
AIC: 584.11

Number of Fisher Scoring iterations: 6
```

Aquest model presenta variables que no són significatives enfront del nivell de significació del 5%. volatile.acidity, citric.acid, free.sulfur.dioxide i pH presenten valors de `Pr(>|z|)` majors que 0.05. La resta de variables sí que són significatives respecte al nivell de significació.

Una manera de mesurar la qualitat del model és comparant la null deviance amb la residual deviance., fent la resta de les dues i dividint per la primera. Valors propers a la unitat indiquen bons ajustos, i per contra, valors propers a zero n'indiquen una mala qualitat.

Qualitat del model = 0.3565921

Podem veure que la qualitat del model no és gaire bona, ja que aquest valor es troba més a prop del 0 que de l'1.

Farem un altre prova traient les variables que no són significatives a veure si el model millora una mica.

```
 Entrenament del model sense variables no significatives


Call:
glm(formula = as.factor(qualityr2) ~ fixed.acidity + residual.sugar +
    chlorides + total.sulfur.dioxide + density + sulphates +
    alcohol, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2830  -0.4085  -0.2060  -0.1044   2.8775

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.343e+02  1.139e+02   3.814 0.000137 ***
fixed.acidity         4.786e-01  9.536e-02   5.019 5.19e-07 ***
residual.sugar        4.393e-01  1.072e-01   4.098 4.16e-05 ***
chlorides            -1.077e+01  4.969e+00  -2.167 0.030254 *
total.sulfur.dioxide -2.039e-02  5.031e-03  -4.053 5.06e-05 ***
density              -4.534e+02  1.142e+02  -3.970 7.20e-05 ***
sulphates             5.113e+00  7.131e-01   7.171 7.47e-13 ***
alcohol               7.820e-01  1.432e-01   5.460 4.76e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 870.53  on 1118  degrees of freedom
Residual deviance: 568.26  on 1111  degrees of freedom
AIC: 584.26

Number of Fisher Scoring iterations: 6
```

Després de treure les variables no significatives veiem que la qualitat del model baixa una mica (0.3472275 vs 0.3565921) però són molt semblants així que utilitzarem aquest últim model amb menys variables per veure la seva capacitat predictiva.

```
 Odds ratios del model

        (Intercept)        fixed.acidity       residual.sugar
      4.045821e+188         1.613853e+00         1.551677e+00
         chlorides total.sulfur.dioxide              density
       2.109610e-05         9.798151e-01        1.192912e-197
          sulphates              alcohol
       1.661883e+02         2.185938e+00
```

Dintre del model de regressió logística podem veure l'odds ratio que indiquen les variables més importants a l'hora de predir si un vi és bo o no. Aquest valor com més allunyat de la unitat indica que la variable és més rellevant. En aquest cas les variables més rellevants per predir són sulphates i alcohol.

Predicció del model

```
$‘Matriu de Confusió‘
                           prediccio
modelpred2$model$qualityr2 No Vi Bo Vi Bo
                  No Vi Bo      399    11
                  Vi Bo          55    15

$‘Matriu de Confusió - %row‘
                           prediccio
modelpred2$model$qualityr2 No Vi Bo Vi Bo
                  No Vi Bo     97.3   2.7
                  Vi Bo        78.6  21.4

$Ajust
[1] 86.25

$Sensibilitat
[1] 21.42857

$Especificitat
[1] 97.31707
```

Com podem veure, aquest model té un ajust molt bo (86%) però pel que podem observar això bé donat perquè troba molt bé quan un vi no és bo (especificitat 97%) però no troba gaire bé quan un vi és bo (sensibilitat 21%).

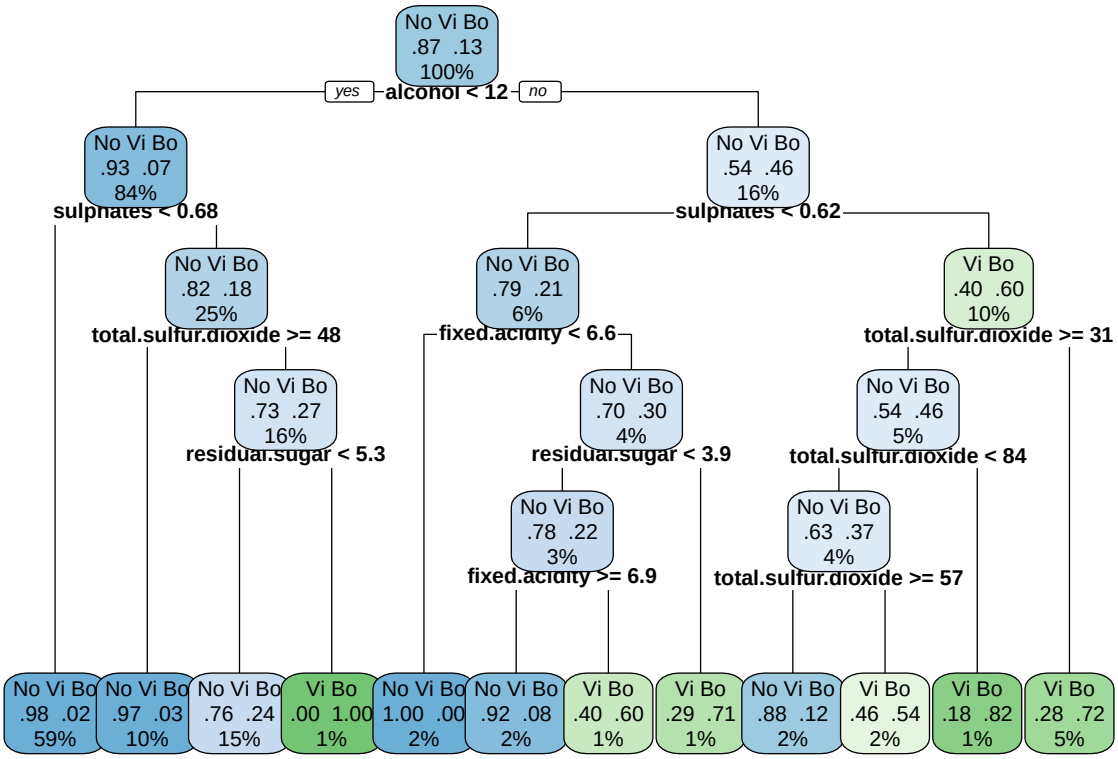
Llavors, tot i que podríem considerar que és un model bo, no ajuda a poder dir si un vi és considerat bo pels experts.

Arbre de decisió

Un altre model que podem utilitzar són els arbres de decisió. Aquest model és una tècnica predictiva que dona com a resultat unes regles de classificació molt intuïtives i amb la seva representació gràfica és molt fàcil d'interpretar.

Entrenament del model

Per tal de poder comparar amb el model de regressió logística, utilitzarem les variables que van sortir significatives a la regressió (fixed.acidity, residual.sugar, chlorides, total.sulfur.dioxide, density, sulphates i alcohol).



Les caixes de la filera de sota ens indiques el final dels nodes i, per tant, completen una regla de classificació. A la caixa hi surt l'etiqueta de la probabilitat més alta, les probabilitats dels dos supòsits i el percentatge d'individus a cada caixa.

Per exemple el node de més a la dreta indica que hi ha un 72% de probabilitats que sigui un Vi Bo aquell que alcohol>=12, sulphates >=0.62 i total.sulfur.dioxide < 31.

Predicció del model

```
$`Matriu de Confusió`
              arbre_pred
test$qualityr2 No Vi Bo Vi Bo
      No Vi Bo      386    24
      Vi Bo          43    27

$`Matriu de Confusió - %row`
              arbre_pred
test$qualityr2 No Vi Bo Vi Bo
      No Vi Bo     94.1   5.9
      Vi Bo        61.4  38.6

$Ajust
[1] 86.04167

$Sensibilitat
[1] 38.57143

$Especificitat
[1] 94.14634
```

Com podem veure, aquest model té un ajust molt bo (87%) però pel que podem observar això bé donat perquè troba molt bé quan un vi no és bo (especificitat 93%) però no troba gaire bé quan un vi és bo (sensibilitat 47%).

Si el comparem amb el model de regressió logística, tenim que l'ajust és el mateix (87% vs 86%), l'especificitat si fa no fa també (93% vs 97%) però pel que fa a la sensibilitat hi ha una millora substancial (47% vs 21%). No és que sigui cap meravella, però el model basat en arbres de decisió prediu un Vi Bo molt millor que la regressió logística.

Conclusió

Volíem poder saber quan un vi és bo a través de les seves característiques químiques, poder predir com serà el vi sense haver d'anar a preguntar als experts. Per fer això hem utilitzat dos models diferents, regressió logística i arbres de decisió.

Després de realitzar els dos models hem arribat a la conclusió que, tot i que no son cap meravella perquè és més probable saber que un vi és dolent que no pas un vi és bo, el model basat en arbres de decisió és capaç de predir si un vi és bo millor que el model basat en la regressió logística.