Pràctica 2: Neteja i validació de les dades

Natalia Gutierrez Navarro

8 de junio, 2018

Contents

1 Descripció del dataset

El conjunt de dades en qüestió té relació amb la variant de vi negre portuguès "Vinho Verde" i ha sigut extret directament de *Kaggle* (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). Està constituït per 1599 registres amb 12 variables, de les quals 11 fan referència a propietats fisicoquímiques i una a la valoració sensorial de la qualitat. Aquestes són (P.Cortez 2009 , Atenea (2012)):

- **fixed acidity**: és l'acidesa d'un vi relativa a la suma de la quantitat d'àcids fixos que es troben en la seva composició, majoritàriament el tartàric. Una quantitat òptima és de 7.5 gr/litre [gr/l].
- **volatile acidity**: és la concentracíó d'àcid acètic i derivats [gr/l].
- **citric acid**: concentració d'àcid cítric. Al vi es troba en quantitats que normalment oscil · len entre 100 i 300 mg/litre [gr/l].
- **residual sugar**: principalment glucosa i fructosa que no s'han fermentat. Depenent de la seva quantitat, es categoritza el vi com sec (4-9g/l), semisec (12-18g/l), semidolç (<45 g/l) o dolç (>45g/l). Els valors dins dels rangs depenen de l'acidesa total [gr/l].
- **chlorides**: concentració de clorurs, algunes de les sals minerals que hi han al vi en petites quantitats (el total de totes les sals és 2-4 g/l) [g/l].
- **free sulfur dioxide**: anhídrid sulfuròs que no es troba combinat amb altres components, i per tant, és el que proporciona les propietats antisèptiques [mg/l].

- **total sulfur dioxide**: el total d' anhídrid sulfuròs, lliure i combinat, és un conservant afegit . El límit legal és 250mg/l; per vins de qualitat, 200 mg/l [mg/l].
- **density**: en g/cm3.
- **pH**: mesura la concentració d'ions H+.
- **sulphates**: concentració de sulfats, algunes de les sals minerals que hi han al vi en petites quantitats (el total de totes les sals és 2-4 g/l) [g/l].
- **alcohol**: quantitat d'alcohol, el qual serà etanol principalment. Es mesura en graus i el seu valor indica el volum en que es troba, expressat en percentatge, com en aquest cas.
- **quality**: puntuació mitja proporcionada per catadors dins del rang [0,10].

2 Objectius de l'anàlisi

A partir del conjunt de dades disponible, en aquesta anàlisi es planteja la possibilitat de poder predir la qualitat del vi des del punt de vista sensorial fent servir les seves propietats fisicoquímiques.

Amb un model d'aquestes caracterítiques es podria tenir una valoració objectiva per conèixer la qualitat sense haver de recórrer a catadors. Tanmateix, es tindria informació de com modificar les seves propietats per millorar el vi.

Aquest conjunt de dades constitueix una mostra de vins de denominació d'origen *Vinho Verde*. Per tant, de totes les observacions s'esperen una tendència semblant en les seves característiques; si més no, podem assumir que la qualitat s'haurà avaluat considerant els mateixos aspectes, propietats que s'esperaria trobar en el tast. Així doncs, la mostra pot ser una bona candidata per construir un model de predicció.

3 Integració i selecció de les dades d'interès a analitzar

Les dades que disposem estan ja integrades en un únic fitxer en format `csv`. Per procedir amb la neteja i anàlisi de les dades, les carreguem amb la funció `read.csv` de manera que les tindrem totes elles incloses en un dataframe.

```
idata <- read.csv("../data/winequality-red.csv",header=TRUE)
summary(idata)
```

```
##  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
```

Cridant la funció `summary` per obtenir un resum de les columnes que constitueixen el dataframe, podem observar totes les variables que esperavem tenir al conjunt de dades: 11 variables quantitatives indicant propietats fisicoquímiques de les observacions de la mostra, i una altra variable quantitativa amb la valoració sensorial. A més a més, comprovem que totes elles estan identificades adequadament amb els noms assignats.

Excepte *quality*, que com hem vist és la variable objectiu en el nostre estudi, la resta són atributs que caracteritzen diferents aspectes dels vins avaluats, i a priori tots ells són susceptibles a ser útils en l'anàlisi. Per altra banda, el tamany de la mostra (1599 observacions) no és desmesurat i es podrà procesar sense dificultats. Així que inicialment seleccionem totes les dades que tenim ja carregades en el dataframe.

4 Neteja de les dades

4.1 Tractament de zeros o elements buits

```
kable(sapply(idata, class))
```

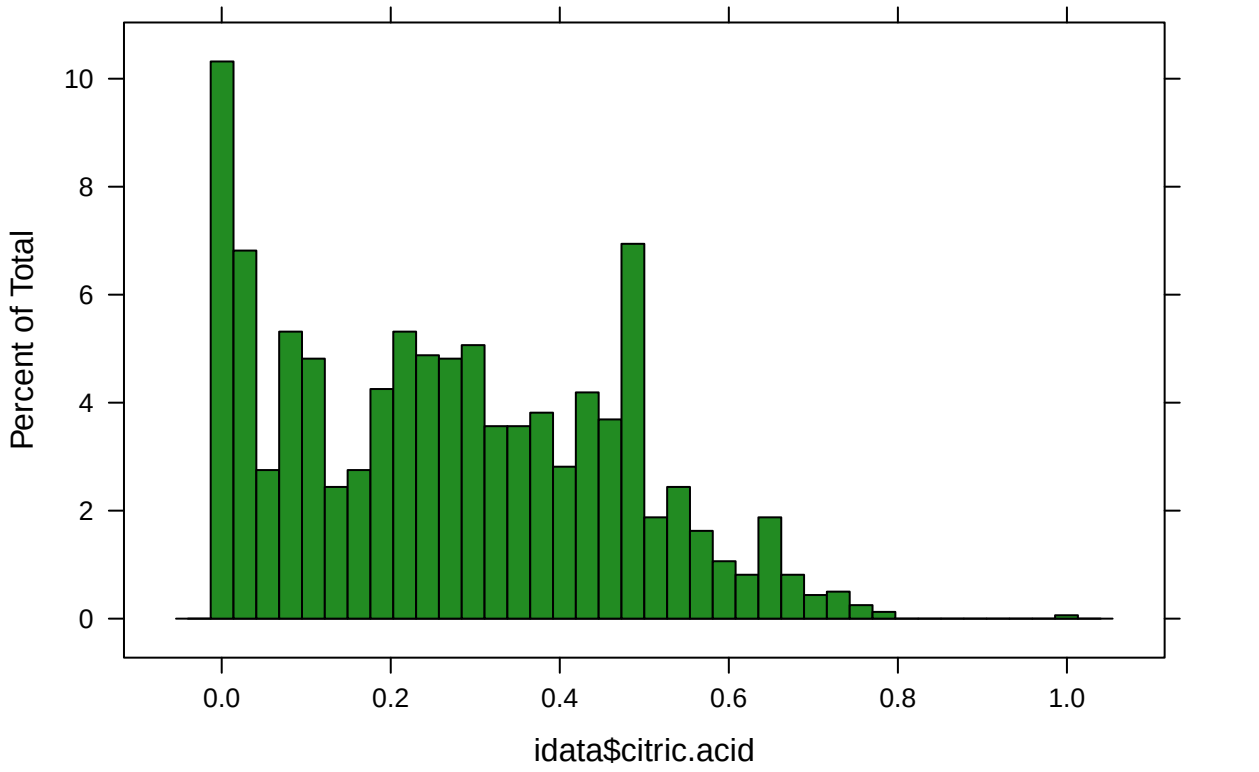
	x
fixed.acidity	numeric
volatile.acidity	numeric
citric.acid	numeric
residual.sugar	numeric
chlorides	numeric
free.sulfur.dioxide	numeric
total.sulfur.dioxide	numeric
density	numeric
pH	numeric
sulphates	numeric
alcohol	numeric
quality	integer

Podem comprovar que totes les variables carregades del `csv` s'han reconegut com a numèriques (o integer). Això vol dir que no hi ha cap valor de tipus caràcter que s'hagi establert com a centinela per indicar element buit, perque llavors s'hauria fixat el tipus de dades com factor. També podriem haver arribat a aquesta mateixa conclusió directament observant el resum que vam mostrar anteriorment: en totes les variables es mostra un resum format per mitja i el sumari dels cinc números de Tukey, propi de les dades del tipus numèric.

Per altra banda, si els elements buits s'haguèssin indicat amb espais buits ("") en el fitxer `csv`, s'haurien traduït com "NA" (valors perduts) i aquests estarien registrats en el sumari anterior. Tampoc és el cas.

Per últim, hi ha la possibilitat que s'haguès fixat com a centinela el zero o algun valor sense sentit com podria ser en aquests atributs un número negatiu. En el sumari observem que els rangs dinàmics de les variables són inicialment coherents i només *citric.acid* conté zeros. A continuació mostrem l'histograma d'aquesta variable, on podem comprovar que aquest zeros no semblen ser valors extrems. Tanmateix, és factible que un vi tingui una concentració d'àcid cítric de 0 g/l (C.Catania 2007), així que descartem que aquests zeros signifiquin ausència de valor.

```
histogram(idata$citric.acid,nint=40,col='forestgreen')
```



4.2 Valors extrems

En el sumari mostrat anteriorment es pot apreciar una distància entre el tercer quartil i el màxim més gran que en la resta de mesures en la majoria de variables. Això ja ens indica la presència de valors extrems o *outliers* en les nostres dades.

Basant-nos en el mètode de Turkey, podem identificar possibles outliers en el diagrama de caixa o directament amb la funció boxplot.stats de R. A continuació analitzem en aquest aspecte cadascuna de les variables quantitatives.

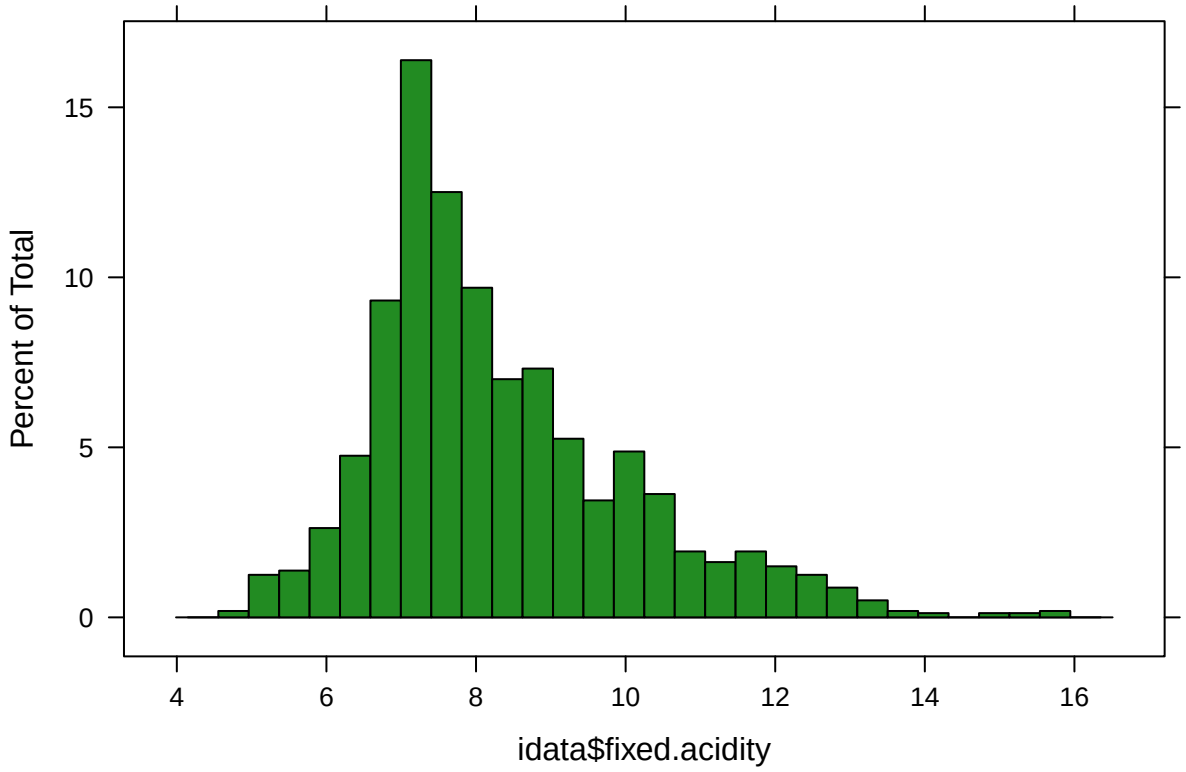
fixed acidity

```
boxplot.stats(idata$fixed.acidity)$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
```

Resulta que 49 valors són detectats com valors atípics. Però si representen la distribució de la variable observem que té una cua que s'extèn cap a aquests valors de manera que no semblen ser incongruents amb la resta de valors.

```
histogram(idata$fixed.acidity,nint=30,col='forestgreen')
```



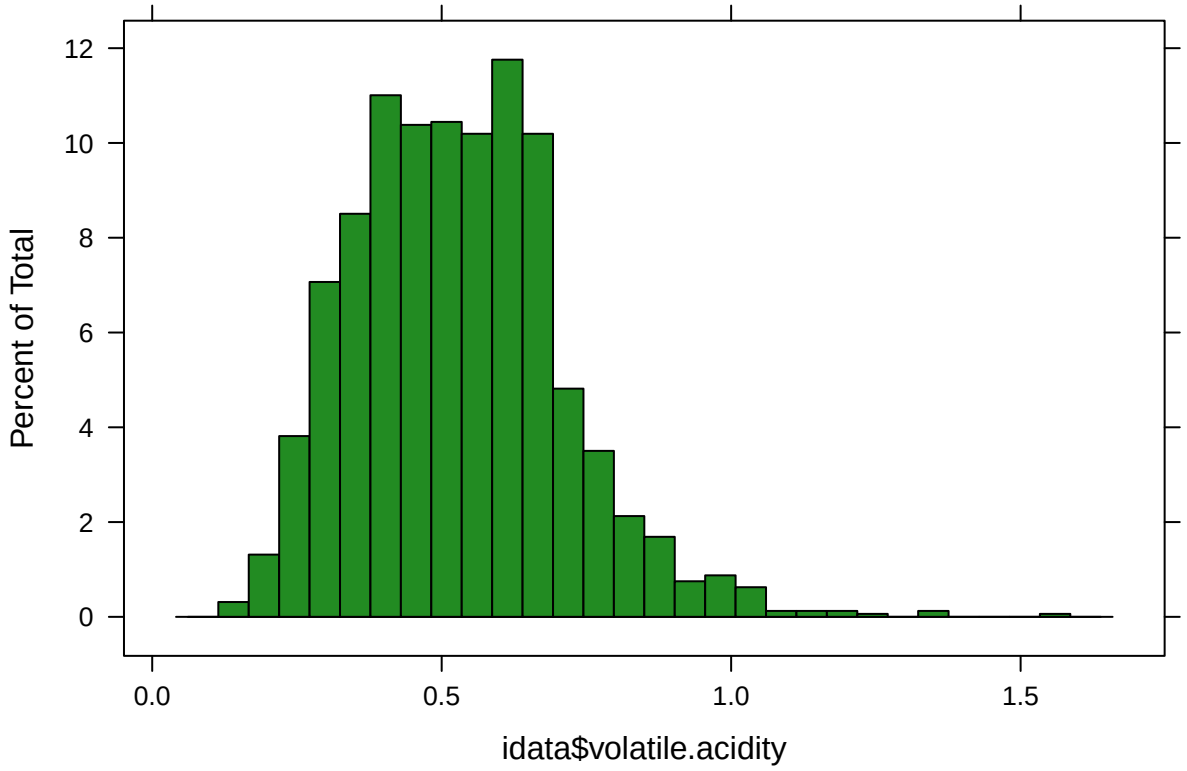
Considerem que són valors atípics propis de la distribució.

volatile acidity

```
length(boxplot.stats(idata$volatile.acidity)$out)
```

```
## [1] 19
```

```
#boxplot(idata$volatile.acidity~idata$quality,col="forestgreen",cex.axis=0.7)
histogram(idata$volatile.acidity,nint=30,col='forestgreen')
```



Anàlogament al cas de l'acidesa fixa, considerem aquests outliers propis de la distribució.

citric acid

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

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

En l'histograma representat en l'apartat de "Tractament de zeros" comprovem que la distribució es concentra en els valors esperats per la concentració d'àcid cítric inclosos en el rang [0.1,0.3] gr/l. Tanmateix es veu una cua que s'allarga cap a valors més alts de manera natural, però no s'oberva el valor detectat com outlier integrat en aquesta cua. Podria ser un valor atípic propi de la distribució, però la seva localització aïllada ens fa sospitar que no sigui legítim. El marquem com valor perdut per ser tractat posteriorment.

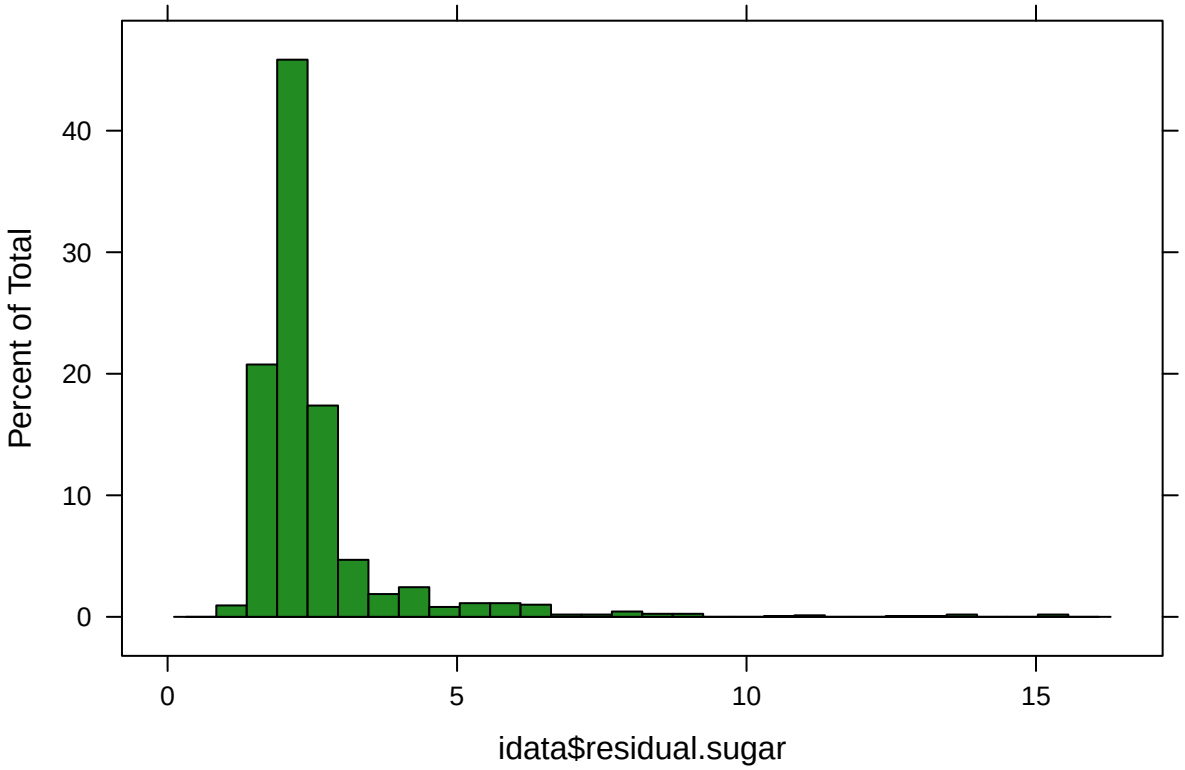
```
idata$citric.acid[which(idata$citric.acid==max(idata$citric.acid))] <- NA
```

residual sugar

```
length(boxplot.stats(idata$residual.sugar)$out)
```

```
## [1] 155
```

```
histogram(idata$residual.sugar,nint=30,col='forestgreen')
```



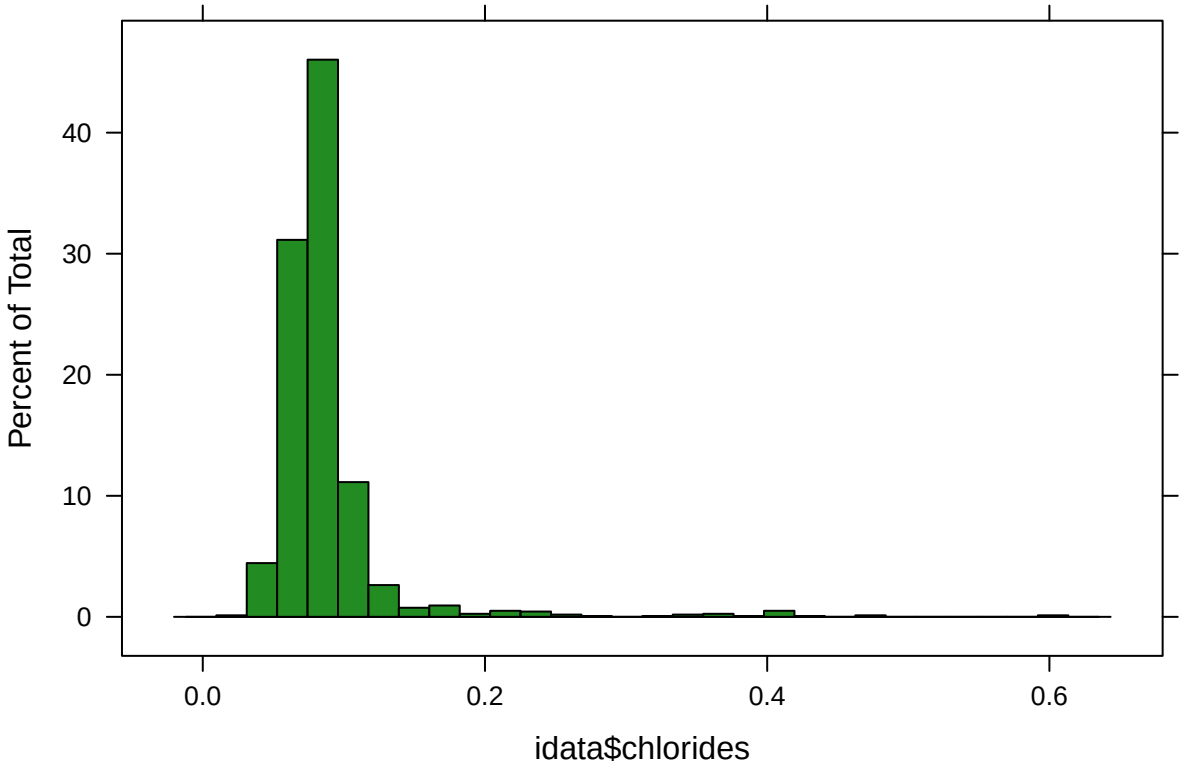
En el cas de *residual sugar* ens trobem inicialment amb un número gens despreciable d'outliers: 155. En l'histograma es veuen com valors molt allunyats de la resta de observacions, i per tant són candidats a considerar-los valors atípics contaminants i ser tractats com a tal. Però com vam descriure en el primer apartat, el vi pot tenir concentracions de sucre molt divers, trobant concentracions fins a 18 g/l pel tipus semisec. És a dir, els valors que tenim aqui com candidats a valors extrems il · legítims són valors totalment pausibles. Que apareguin en la distribució com outliers pot ser degut a que aquesta mostra pertany a la població de vins d'una denominació d'origen específica, de la qual probablement la producció s'especialitza més en vins més secs. Llavors concluïm que aquests valors atípics són propis de la distribució de la variable.

chlorides

```
length(boxplot.stats(idata$chlorides)$out)
```

```
## [1] 112
```

```
histogram(idata$chlorides,nint=30,col='forestgreen')
```



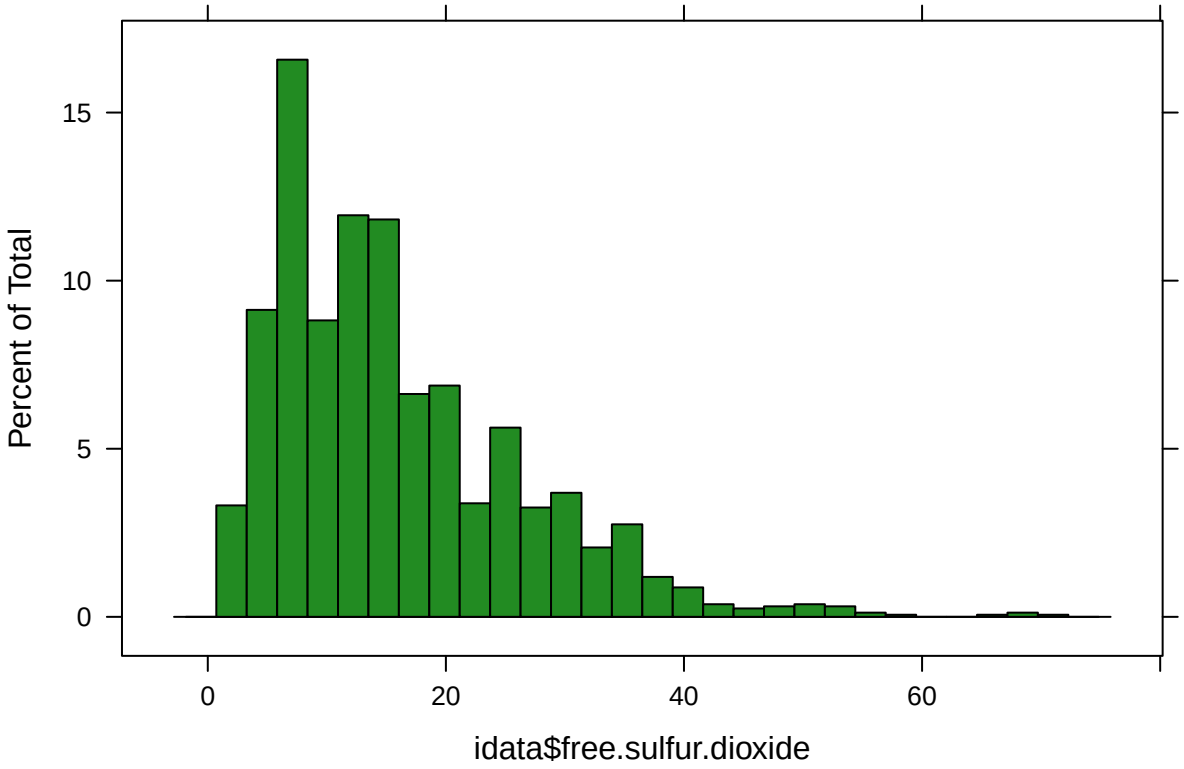
En el cas dels *chlorides* també tenim un número important d'outliers: 112. Observem que aquests estam molt dispersats arribant a un valor de ~0.6 gr/l, però hi ha certa continuïtat en la distribució. Tanmateix, els valors de concentració de clorurs en el vi poden ser generalment fins a valors de 0.5 g/l, arribant a valors de 1 g/l per vinyes situades ubicades a prop del mar (Enrique, n.d.). La mostra que estem analitzant inclou vins de D.O. "Vinho Verde", i aquests es cultiven en una regió amb costa (Wikipedia, n.d.); per tant, pot haver-hi perfectament en la mostra individus amb aquestes concentracions.

free sulfur dioxide

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

```
histogram(idata$free.sulfur.dioxide,nint=30,col='forestgreen')
```



En la variable *free sulfur dioxide* es detecten 30 outliers. En l'histograma veiem que la gran majoria d'aquests segueixen la cua allargada de la distribució, però hi ha un petit conjunt aïllat en els valors més alts. Procedim a considerar-los com valors atípcs contaminants i els marquem com valors perduts.

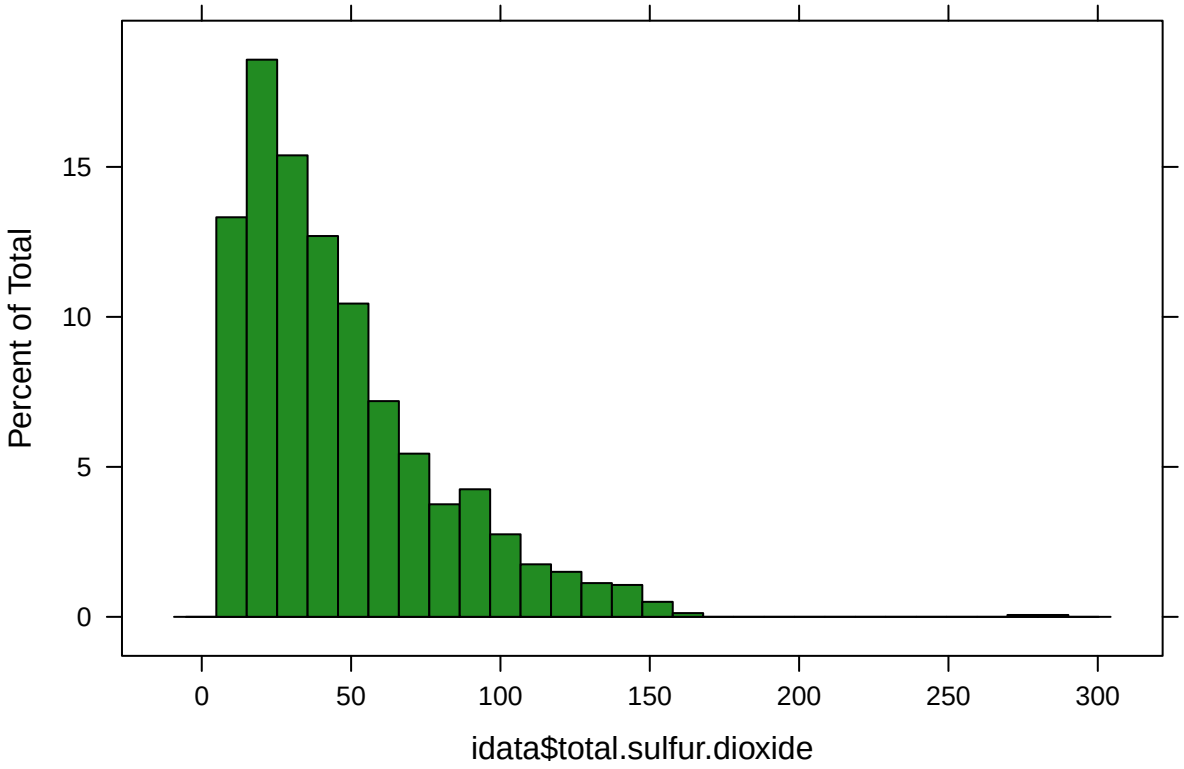
```
idata$free.sulfur.dioxide[which(idata$free.sulfur.dioxide > 60)] <- NA
```

total sulfur dioxide

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

```
histogram(idata$total.sulfur.dioxide,nint=30,col='forestgreen')
```



La gran majoria d'outliers detectats es poden considerar valors propis de la distribució que tenim tal com s'observa en l'histograma. Però hi ha alguns valors allunyats els quals sobre passen el límit que es considera legal en concentracions d'anhídrid sulfuròs. Això ens fa pensar que es tractin de valors atípics contaminants, i per tant, els marquem com valors perduts.

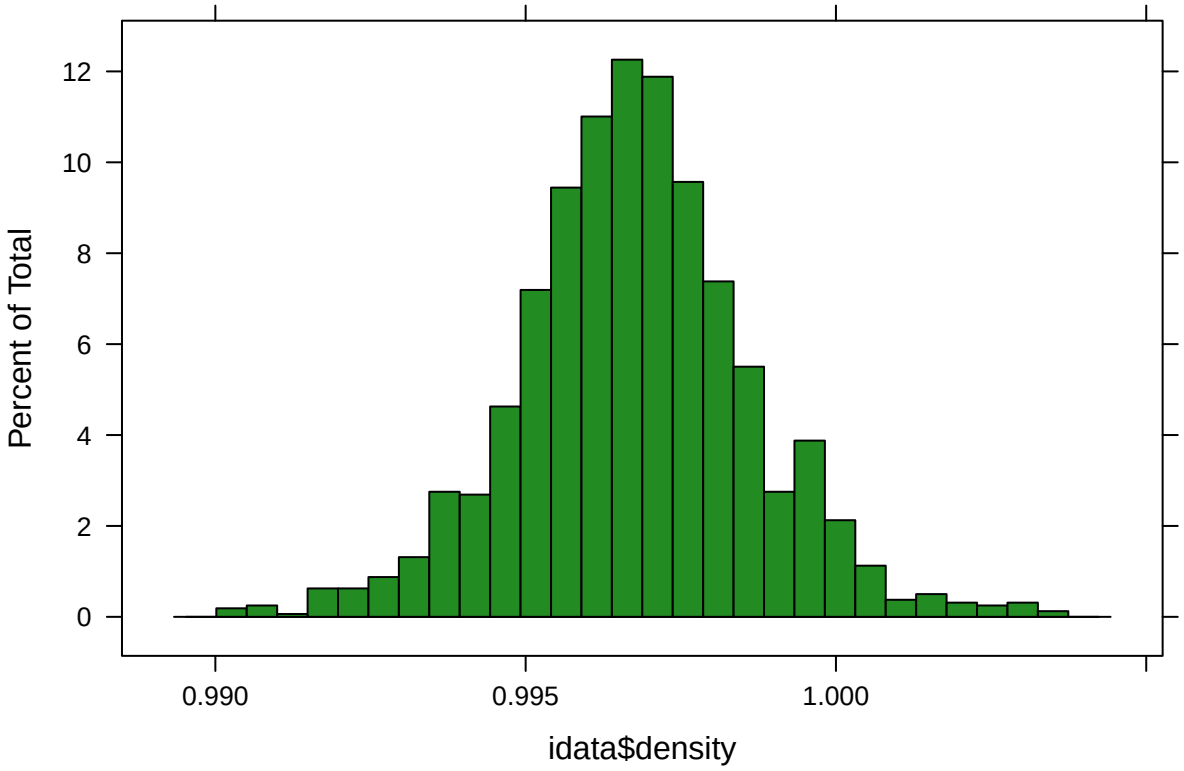
```
idata$total.sulfur.dioxide[which(idata$total.sulfur.dioxide > 200)] <- NA
```

density

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

```
histogram(idata$density,nint=30,col='forestgreen')
```



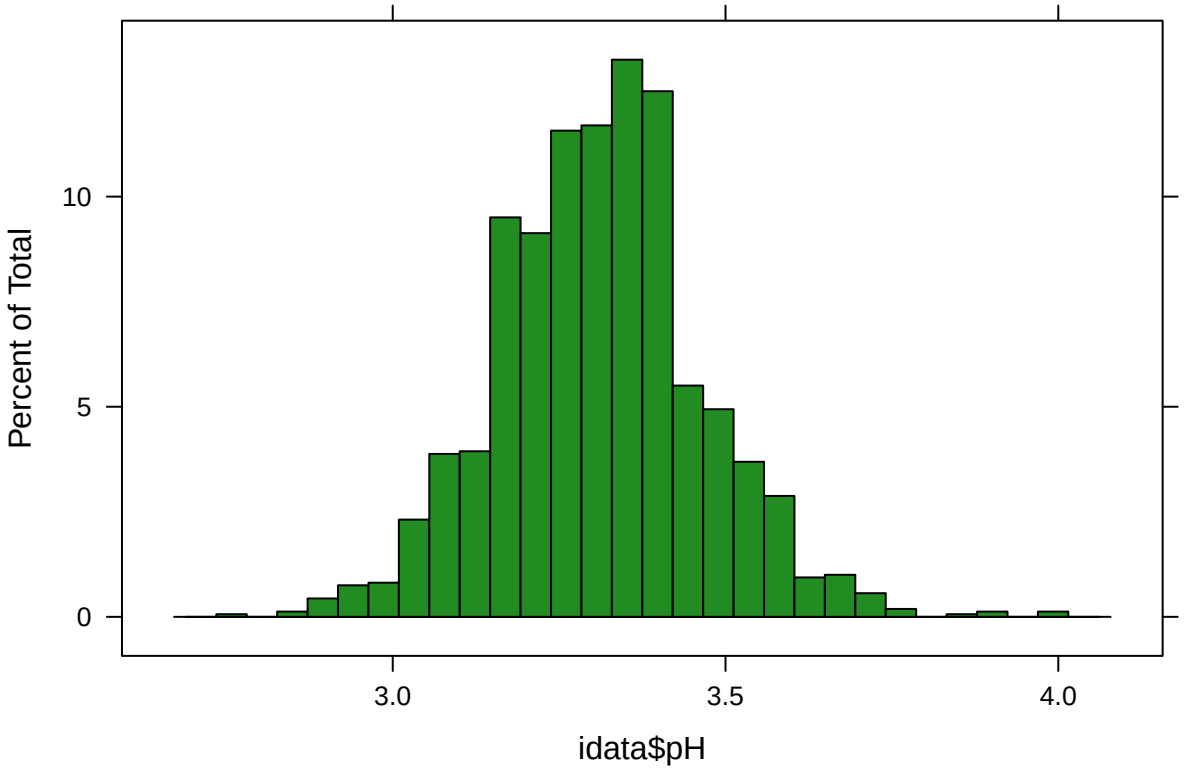
Malgrat que inicialment es detecten outliers, amb la visualització de l'histograma podem considerar que aquests formen part de la distribució coherentment i els considerarem legítims. No prendrem mesures per corregir-los.

pH

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

```
histogram(idata$pH,nint=30,col='forestgreen')
```



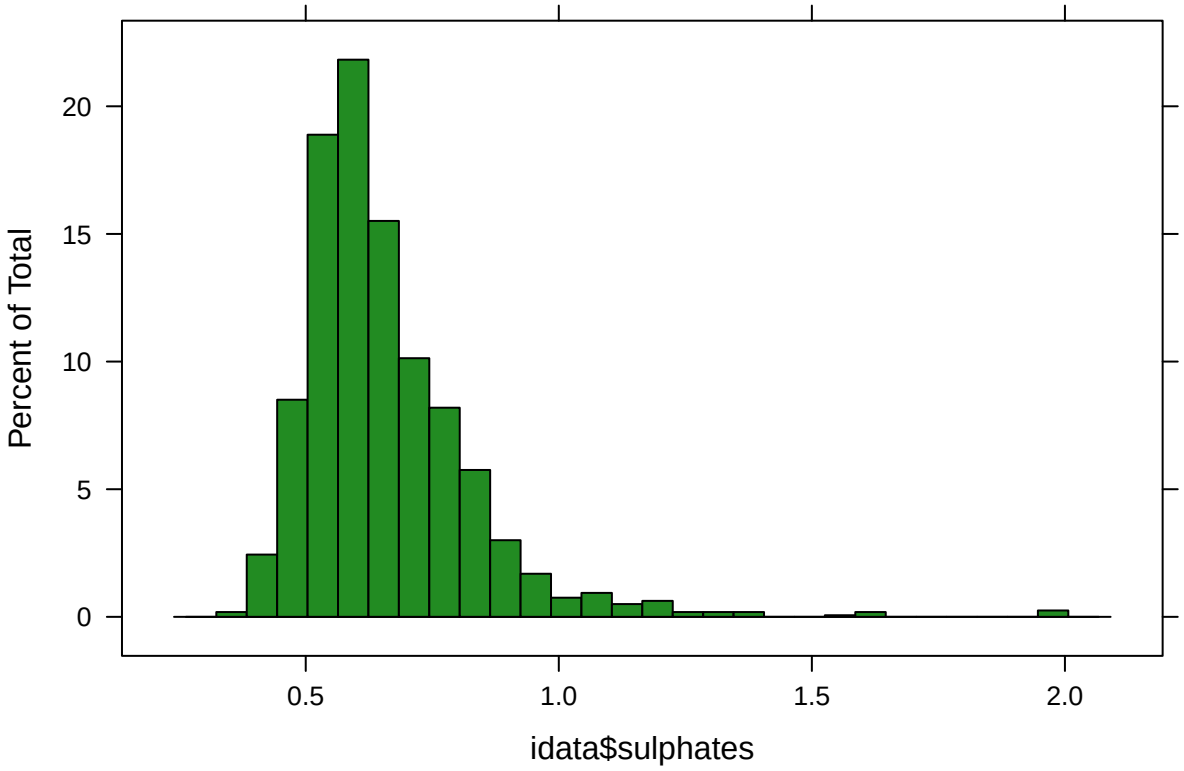
En el cas de *pH*, del conjunt d'outliers podriem sospitar dels localitzats a valors més grans, però són valors que es troben normalment en el vi (Enrique, n.d.), així que suposarem que són tots legítims.

sulphates

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

```
histogram(idata$sulphates,nint=30,col='forestgreen')
```



La variable *sulphates* presenta una cua en la distribució que s'allunya a valors alts. Els primers valors mostren continuïtat en l'histograma, però hi ha alguns >1.5 que són sospitosos. La quantitat de sulfats en vins normals és de l'ordre de 0.6-0.7 g/l, arribant a ser 2 g/l. en vins envellits (Enrique, n.d.). Però els de D.O. *Vinho Verde* es caracteritzen per ser joves (Wikipedia, n.d.); per tant, es confirma que aquests valors són probablement contaminants. Considerarem valors perduts a partir de 1.5.

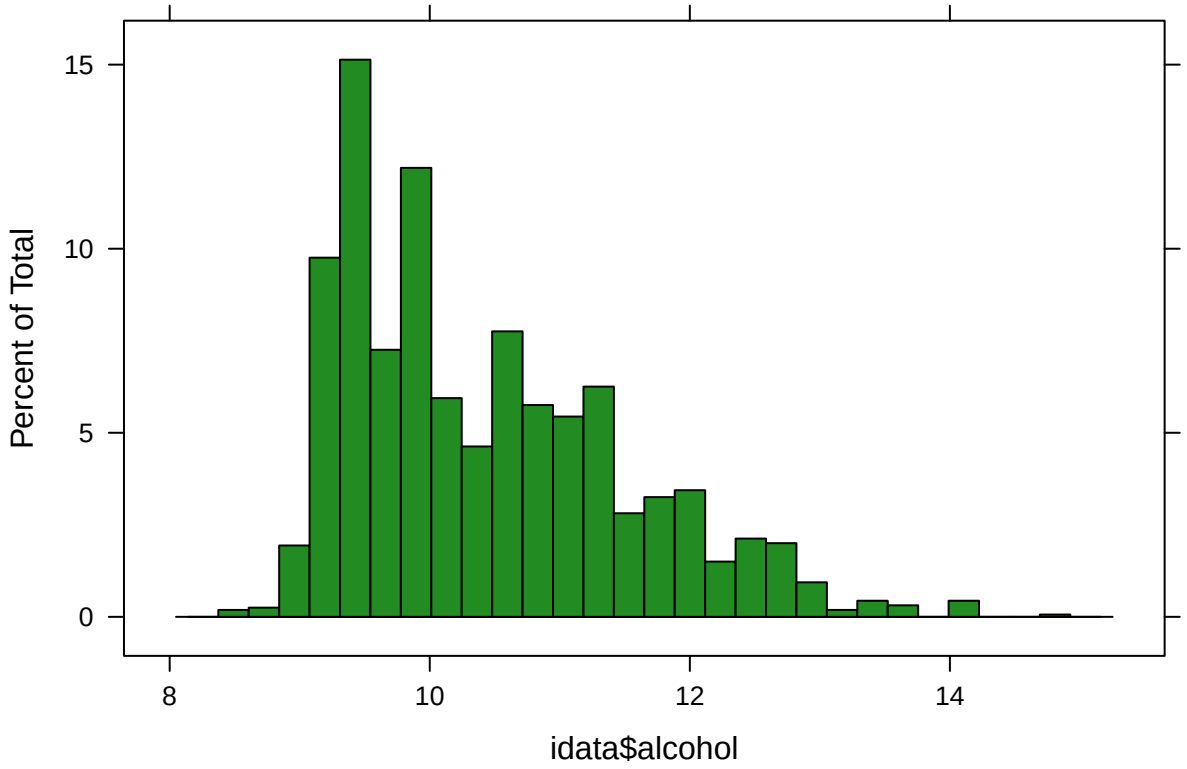
```
idata$sulphates[which(idata$sulphates > 1.5)] <- NA
```

alcohol

```
length(boxplot.stats(idata$alcohol)$out)
```

```
## [1] 13
```

```
histogram(idata$alcohol,nint=30,col='forestgreen')
```



Es podrien considerar els últims valors de la cua que presenta la distribucíó de *alcohol* com outliers. Però el vi pot presentar ua graduació fins del 15%, així que assumirem que són outliers legítims.

quality

Finalment, pel que fa a la variable de qualitat, tots els valors estan dins del rang com vam veure en el sumari de l'apartat de "Integració", així que no hi ha motius per considerar que hi hagin valors atípics.

** Tractament de valors perduts generats **

```
# Numero d'outliers per observació
nacount<-apply(idata, 1, function(x) sum(is.na(x)))
# Número total de outliers
sum(nacount)
```

```
## [1] 15
```

```
# Número màxim d'outliers per observació.
max(nacount)
```

```
## [1] 2
```

Una vegada determinats els valors atípics contaminants, i per tant marcats con valors perduts, hem de tractar-los. Les opcions que tenim són:

- Descartar les observacions que tenen algun valor perdut. En el nostre cas, tenim un total de 15 registres afectats, amb un màxim de 2 NA per registre. Malgrat que només suposaria <1% de les dades, perderiem informació ja que els registres tenen molt pocs valors perduts.
- Fer servir un mètode d'imputació de valors. Podem assignar valors de tendència central de la variable en qüestió, dins de la mostra o considerant només les observacions de la mateixa categoria (qualificació de qualitat). O bé, fer servir que mesuri similitud entre observacions de la mostra, que degut als objectius fixats en l'anàlisi i les característiques de les dades, es pot considerar més adient.

Decidim aplicar aquesta segona opció, i per aquesta finalitat farem servir la imputació kNN. Fem servir la comanda `kNN` de la llibreria `VIM`, amb els paràmetres per defecte.

```
odata<-kNN(idata,imp_var=FALSE)
```

5 Anàlisis de les dades

5.1 Selecció dels grups de dades

Inicialment, seleccionem totes les variables per la nostra anàlisi de dades. No obstant, pot ser que en alguna fase es descarti alguna d'elles per no ser significativa per al model de predicció final. Tanmateix, hem vist que aquestes variables tenen uns rangs de valors molt heterogenis; sobre tot destaca el cas extrem de *total.sulfur.dioxide*, com podem veure en la següent taula.

```
kable(t(rbind(sapply(colnames(odata),function(x) fivenum(odata[,x])))),
      col.names=c("Min","1Q","Median","3Q","Max"),align='c',digits=2)
```

	Min	1Q	Median	3Q	Max
fixed.acidity	4.60	7.10	7.90	9.20	15.90
volatile.acidity	0.12	0.39	0.52	0.64	1.58
citric.acid	0.00	0.09	0.26	0.42	0.79
residual.sugar	0.90	1.90	2.20	2.60	15.50
chlorides	0.01	0.07	0.08	0.09	0.61
free.sulfur.dioxide	1.00	7.00	14.00	21.00	57.00
total.sulfur.dioxide	6.00	22.00	38.00	62.00	165.00
density	0.99	1.00	1.00	1.00	1.00
pH	2.74	3.21	3.31	3.40	4.01
sulphates	0.33	0.55	0.62	0.73	1.36
alcohol	8.40	9.50	10.20	11.10	14.90
quality	3.00	5.00	6.00	6.00	8.00

Aquesta diferència de rangs pot arribar a ser perjudicial per la regressió que volem realitzar, ja que les variables amb valors més grans tindran més impacte que la resta alhora de construir el model. Per evitar aquest efecte, apliquem estandardització i així corregim les diferències de rangs normalitzant-los mantenint les distribucions.

Per altra banda, formem dos conjunts de dades: un conjunt d'entrenament per construir el model de regressió, i un conjunt de prova per avaluar-lo. Considerem reservar el 20% de les dades per la fase d'avaluació i fem una extracció aleatòria per obtenir unes mostres representatives de la població respectant les distribucions de les variables.

```
# Estandardització de les variables excepte quality
odata[,-12]<-scale(odata[,-12])

# Construïm el set d'entrenament i el set de prova
set.seed(2018)
in_train <- sample(1:nrow(odata), size = round(nrow(odata)*0.8))

edata<-odata[in_train,]
tdata<-odata[-in_train,]

# Generem un fitxer amb les dades que farem servir en l'anàlisi
```

```r
write.csv(odata, file = "../data/winequality-red_clean.csv")
```

5.2 Comprovació de la normalitat i homogeneïtat de la variància

Per cada variable, comprovem la seva condició de normalitat mitjançant el test de Shapiro-Wilk (P.López-Roldán 2015).

```r
# Funció per avaluar si una variable segueix distribució
# normal
isnormal<-function(x){
  # nivell de significació
  alpha<-0.05
  # Llancem el test
  res<-shapiro.test(x)
  # Avaluem el p-valor
  return (res$p.value >= alpha)
}
# Apliquem el test de normalitat a cadascuna de les variables
kable(apply(edata,2,isnormal))
```

	x
fixed.acidity	FALSE
volatile.acidity	FALSE
citric.acid	FALSE
residual.sugar	FALSE
chlorides	FALSE
free.sulfur.dioxide	FALSE
total.sulfur.dioxide	FALSE
density	FALSE
pH	FALSE
sulphates	FALSE
alcohol	FALSE
quality	FALSE

Ens trobem que cap de les variables segueix una distribució normal, com ja havíem sospitat per la majoria degut als seus histogrames.

A continuació, comprovem l'homogeneïtat de les variàncies entre els diferents grups que tenim definits per la qualificació que han rebut les observacions. Per la falta de normalitat que hem observat, aplicarem el test de Fligner-Killeen (cookbook-r, n.d.).

Abans de procedir, un detall. La variable *quality* és una variable numèrica definida entre [0,10], però hem observat que en la nostra mostra prèn només valors discrets. És possible tractar-la com una variable qualitativa amb un numero finit de categories, útil per fer alguns dels anàlisis que posteriorment descriurem. Tanmateix, recuperarem la seva representació numèrica per fer la regressió.

En el test en qüestió es planteja per cadascuna de les variables la hipòtesi nul · la H_0: homogeneïtat de variàncies contra la hipòtesi alternativa H_A: heterogeneïtat de variàncies. Si obtenim un p-valor més gran que el nivell de significació que imposem de 0.05, acceptem la hipòtesi nul · la, i per tant, podem concloure que hi ha homogeneïtat de variàncies. En canvi, si surt un valor més petit, la rebutgem a favor de la hipòtesi alternativa i concluïm que hi ha heterogeneïtat de variàncies.

```r
# Transformem la variable "quality" en factor
edata$quality<-factor(as.character(edata$quality))

# Funció per testejar si hi ha homogeneïtat de variables en un
# dataset segons una variable categòrica
homoscedasticitat<-function(data,mifac){
  # nivell de significació
  alpha<-0.05
  results<-list()
  # Loop de variables
  for (field in colnames(data)){
    # Saltem la variable categòrica
    if (field ==mifac)
      next
    # Construïm la fòrmula que necessita la comanda
    fmla<-as.formula(paste(field,"~",mifac))
    # Apliquem el test
    res<-fligner.test(fmla,data=edata)
    # Avaluem el p-valor resultant
    results[field]<-(res$p.value >= alpha)
  }
  return (results)
}
# Avaluem el nostre conjunt de dades
res<-homoscedasticitat(edata,"quality")
# Mostrem les variable que sí presenten homogeneïtat
kable(names(which(res==TRUE)),col.names = "Homoscedasticitat en:")
```

Homoscedasticitat en:
residual.sugar
pH
sulphates

Obtenim que només 3 variables presenten homogeneïtat de variància respecte *quality*.

5.3 Correlació entre variables

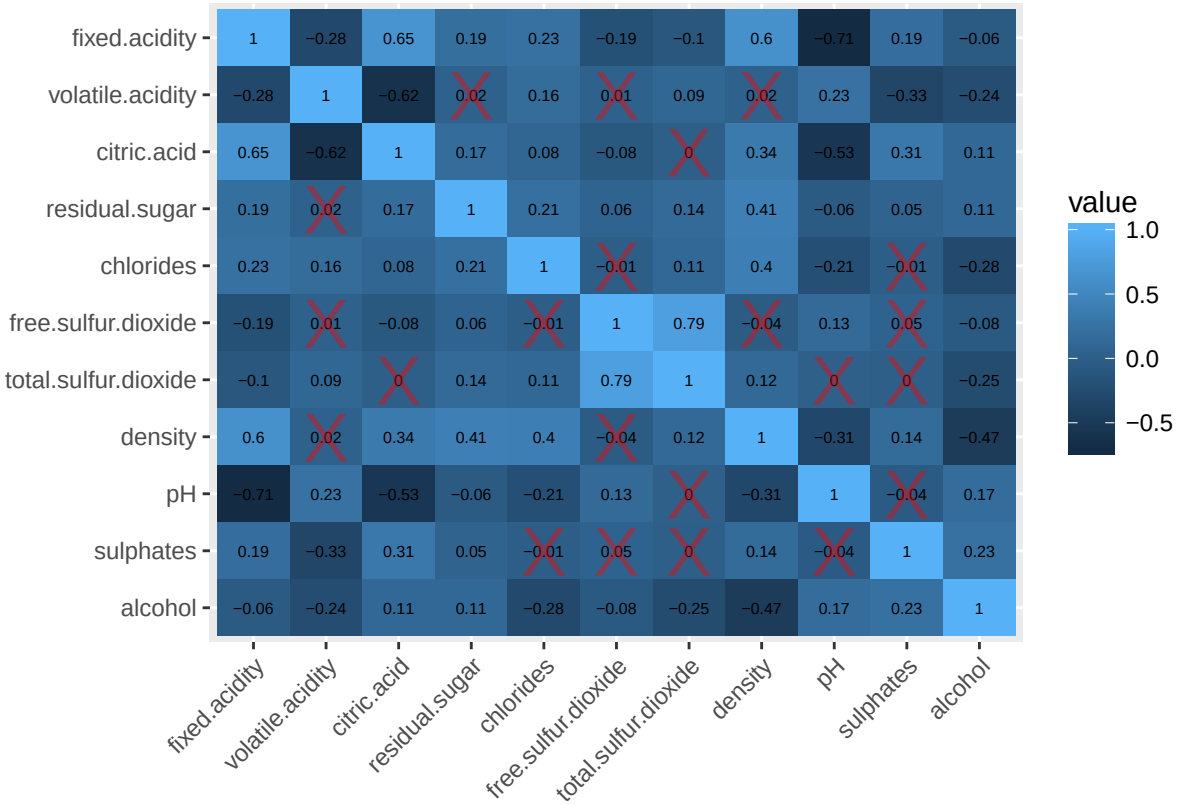
Per la regressió que volem realitzar ens plantegem trobar el conjunt de variables regressives que són útils i necessàries per aquest propòsit entre les que disposem. Per això, una anàlisi que podem fer és avaluar la correlació entre elles per evitar redundància. Ja que les variables no presenten normalitat, fem servir el mètode no paramètric de Spearman (P.Dalgaard 2008).

A continuació, mostrem la matriu de correlació que obtenim en un heatmap, incloent-hi els valors numèrics. A més a més, s'han marcat amb una X vermella els casos en els que la ρ obtinguda no és estadísticament significativa (s'ha fixat un nivell de significació de 0.05.

```r
# Calculem la matriu de correlació per les variables regressives
vars<-colnames(edata)[1:11]
cormatrix = rcorr(as.matrix(edata[,1:11]), type='spearman')
# preparem la matriu de rhos per poder visualitzar-la
cordata = melt(cormatrix$r)
# Reordenem les categories del factor on hi ha els noms de les variables,
```

```r
# així la matriu es veurà amb la diagonal principal ben posicionada.
cordata$Var2<-factor(as.character(cordata$Var2),levels=rev(vars))

txtsize <- par('din')[2] / 2
# Aproximem els valors de rho per poder visualitzar-los
rhos<-0.01*as.numeric(as.integer(cordata$value*100))
# Marcar amb X els constrasts que no són siginificatius
cordata$strike = ""
cordata$strike[cormatrix$P > 0.05] = "X"
# Plotegem la matriu en un heatmap
ggplot(cordata, aes(x=Var1, y=Var2, fill=value)) +
  theme(axis.text.x = element_text(angle=45, hjust=TRUE)) +
  geom_tile() + xlab("") + ylab("") +
  geom_text(label=rhos, size=txtsize) +
  geom_text(label=cordata$strike, size=txtsize * 4, color="red", alpha=0.4)
```



Fixem un umbral de $|\rho| > 0.7$ per considerar que dues variables estan prou correlades. En aquestes condicions resulta que el parell *fixed.acidity* - *pH* estan correlades negativament i *free.sulfur.dioxide* - *total.sulfur.dioxide* positivament. Descartem *fixed.acidity* i *free.sulfur.dioxide* del nostre conjunt de variables regressives.

```r
edata<-select(edata,-fixed.acidity,-free.sulfur.dioxide)
```

5.4 Diferències significatives entre els vins de diferent qualitat

Per acabar de determinar el conjunt de variables regressives, ens hem de plantejar que aquestes han de presentar diferències significatives segons la variable que volem explicar, és a dir, la qualitat. Per poder determinar-ho, realitzem una anàlisi de variància per cadascuna de les variables. Ja que ni tan sols segueixen

una distribució normal apliquem el test no paramètric de Kruskal-Wallis (P.López-Roldán 2015).

```
# Funció per testejar si hi ha homogeneïtat de variables en un
# dataset segons una variable categòrica
differentpergroup<-function(data,mifac){
  # nivell de significació
  alpha<-0.05
  results<-list()
  # Loop de variables
  for (field in colnames(data)){
    # Saltem la variable categòrica
    if (field ==mifac)
      next
    # Construïm la fòrmula que necessita la comanda
    fmla<-as.formula(paste(field,"~",mifac))
    # Apliquem el test
    res<-kruskal.test(fmla,data=edata)
    # Avaluem el p-valor resultant
    results[field]<-(res$p.value < alpha)
  }
  return (results)
}
# Avaluem el nostre conjunt de dades
res<-differentpergroup(edata,"quality")
# Mostrem les variable que si son diferents per quality
names(which(res==TRUE))
```

```
## [1] "volatile.acidity"     "citric.acid"          "chlorides"
## [4] "total.sulfur.dioxide" "density"              "pH"
## [7] "sulphates"            "alcohol"
```

Obtenim que totes les variables són significativament diferents depenent de la qualificació excepte *residual.sugar*. Així que aquesta no aporta informació per la regressió i per tant podem descartar-la.

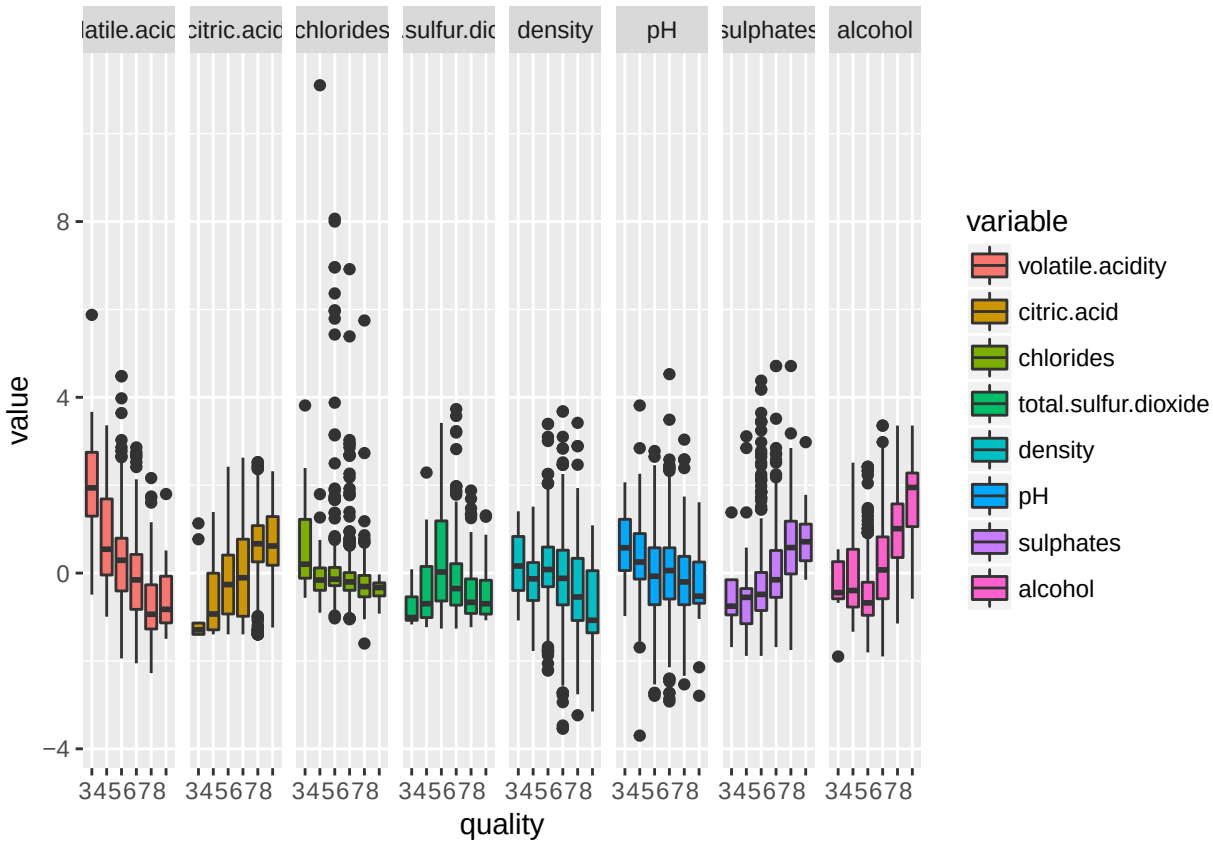
```
edata<-select(edata,-residual.sugar)
```

5.5 Predicció de la qualitat del vi

Una vegada que hem construït el conjunt de dades descartant les variables que no aportarien informació al nostre model, ens plantegem trobar un model de predicció de la qualitat del vi a partir d'aquest.

Amb aquest propòsit, primer visualitzem les distribucions de les variables per qualificació de qualitat.

```
dfmelt<-melt(edata,measure.vars=1:8)
ggplot(dfmelt, aes(x=quality, y=value,fill=variable))+
  geom_boxplot()+
  facet_grid(.~variable)+
  labs(x="quality")+
  theme(axis.text.x=element_text(angle=0, vjust=0.4,hjust=1))
```



Si ens fixem en les medianes de les variables dels diagrames de caixa observem que no hi ha una dependència lineal amb la qualitat, i sembla que una relació més precisa seria polinòmica, però de les variables explicatives amb l'explicada, no a l'inrevés. Així que decidim fer servir aquesta aproximació i apliquem regressió múltiple lineal de primer ordre (Gibergans 2018).

```
# Recuperem la variable independent en format numèric
edata$quality<-as.numeric(as.character(edata$quality))

# Construïm la fòrmula
predictors<-colnames(edata)[1:8]
fmla<-as.formula(paste("quality","~",paste(predictors,collapse="+")))

# Construïm el model lineal
model<-lm(formula=fmla, edata)
```

Obtenim un coeficient de determinació de tan sols $R^2 = 0.3726682$; és a dir, el model només explica el 37.3% de la variabilitat de *quality* i per tant la bondat de l'ajust és a priori molt pobra.

Però recordem la particularitat dels valors discrets que té *quality* mentres que hem assumit un rang de valors continus per la variable explicada: són puntuacions arrodonides dins del rang especificat. Anem a avaluar les prediccions que realitza sobre el mateix conjunt de dades si quantifiquem de la mateixa manera els resultats obtinguts del model.

```
prediccio<-round(predict(model,edata))
sum(edata$quality==prediccio)/nrow(edata)
```

```
## [1] 0.5926505
```

Tenim que el 59.2 % de les prediccions de la qualitat del vi sobre el conjunt d'entrenament són correctes.

6 Resultats

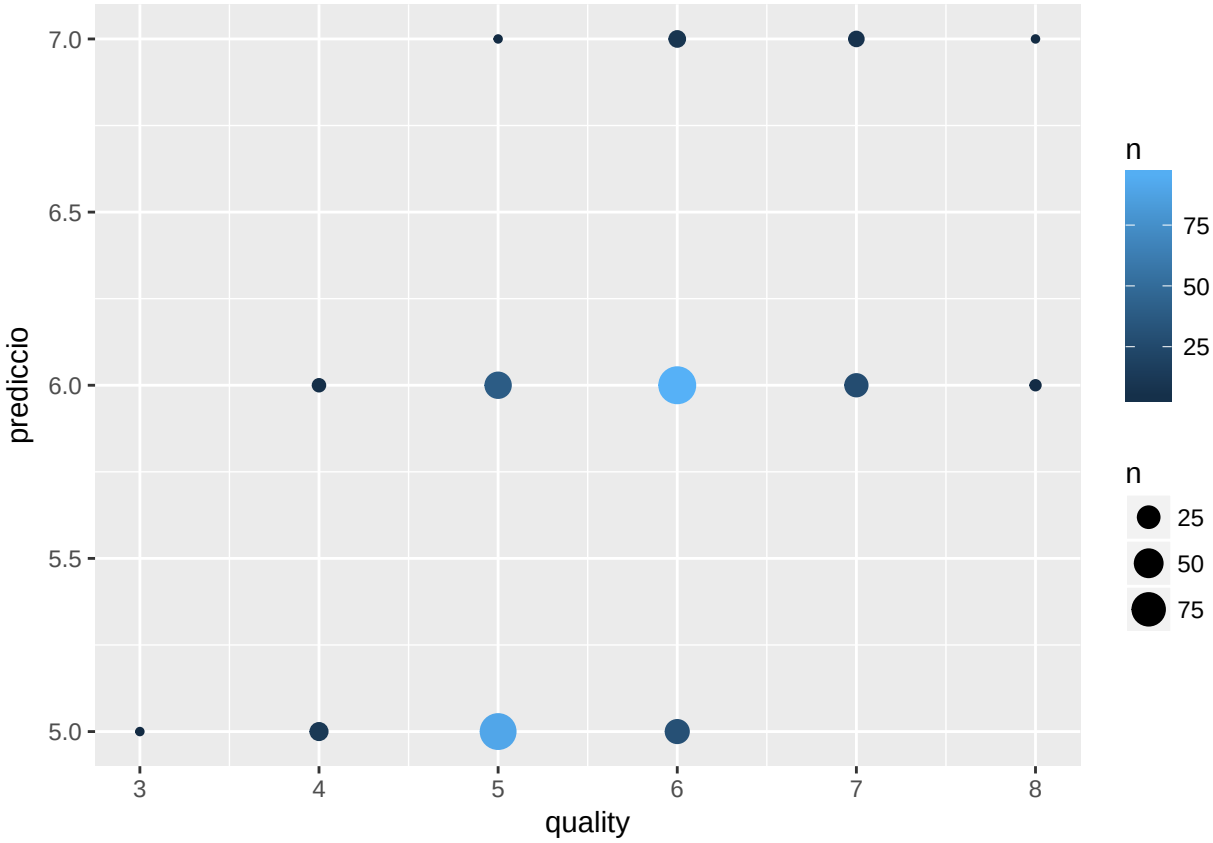
En aquest punt, avaluem el model que hem obtingut sobre el conjunt de prova considerant també en aquest cas l'arrodoniment de les qualificacions.

```
tdata$prediccio<-round(predict(model,tdata))
sum(tdata$quality==tdata$prediccio)/nrow(tdata)
```

```
## [1] 0.60625
```

Obtenim una proporció similar de prediccions correctes: 60.6%. En la següent gràfica podem observar la relació de les qualificacions predites en relació a les reals.

```
ggplot(tdata,aes(x=quality,y=prediccio)) + geom_count(aes(color=..n..))
```



No obstant, s'ha avaluat la capacitat de predicció de la qualificació exacta dels vins en qüestió. Suposem que podem relaxar l'exactitud d'aquestes puntuacions en una categorització més qualitativa simplificant a una classificació binaria: distingim un bon vi d'un de dolent a partir d'una puntuació de 6. Veiem que passa en aquest cas.

```
# Construïm les noves variables
tdata$realbo<-"dolent"
tdata$realbo[tdata$quality >=6]<-"bo"
tdata$realbo<-factor(tdata$realbo)
tdata$predbo<-"dolent"
tdata$predbo[tdata$prediccio >=6]<-"bo"
tdata$predbo<-factor(tdata$predbo)
# Mostrem la matriu de confusió
table(real=tdata$realbo,predit=tdata$predbo)
```

```
##         predit
## real      bo dolent
##   bo     143     30
##   dolent  44    103
```

En la matriu de confusió observem que obtenim una exactitud del 76.9%, amb un ràtio de falsos positius del 29.9% i un ràtio de falsos negatius del 17.3%.

7 Conclusions

En aquesta anàlisi s'ha plantejat la possibilitat de poder predir la qualitat del vi des del punt de vista sensorial fent servir les seves propietats fisicoquímiques, i per aquest propòsit el mètode escollit ha sigut una regressió múltiple lineal.

Per escollir les variables regressives s'ha avaluat la correlació de les variables i les diferències siginificatives que presentaven respecte la qualitat del vi. Tot això, prèviament s'ha netejat les dades tractant els valors perduts i els outliers, si ha sigut necessari.

Hem pogut comprovar que un model regressiu lineal no és adequat per predir la qualitat del vi. Ha resultat que el model trobat només prediu al voltant del 60% de les puntuacions de la mostra, per tant no es pot garantir la qualificació que tindrà un vi a partir de les seves propietats amb aquest tipus de model. No obstant, es pot valorar positivament la possibilitat d'utilitzar les variables considerades per distingir si el vi serà bó o dolent en linies generals, amb una probabilitat de fals positiu del 30%.

References

Atenea. 2012. "La Química Del Vi." http://www.edubcn.cat/rcs_gene/treballs_recerca/2011-2012-08-1-TR.pdf.

C.Catania, S.Avagnina. 2007. *Curso de Degustación de Vinos*. EEAMendoza, INTA. https://inta.gob.ar/sites/default/files/script-tmp-2___los_estmulos_cidos_del_vino.pdf.

cookbook-r. n.d. "Homogeneity of Variance." http://www.cookbook-r.com/Statistical_analysis/Homogeneity_of_variance/.

Enrique. n.d. "Análisis de Vinos." https://www.scribd.com/doc/18175580/Analisis-de-Vinos.

Gibergans, J. 2018. *Regressió Lineal Múltiple*. Edicions UOC.

P.Cortez, F.Almeida, A.Cerdeira. 2009. *Modeling Wine Preferences by Data Mining from Physicochemical Properties*. Decision Support Systems, Elsevier, 47(4):547-553.

P.Dalgaard. 2008. *Introductory Statistics with R*. Springer Science & Business Media.

P.López-Roldán, S.Fachelli. 2015. *Metodologia de La Investigación Social Cuantitativa*. Edicions Universitat Autònoma de Barcelona.

Wikipedia. n.d. "Vinho Verde." https://es.wikipedia.org/wiki/Vinho_verde.