Pràctica 2: Neteja i validació de les dades

Natalia Gutierrez Navarro

20 de mayo, 2018

Contents

1 Descripció del dataset **1**

2 Objectius de l'anàlisi **2**

3 Integració i selecció de les dades d'interès a analitzar **2**

4 Neteja de les dades **3**
4.1 Tractament de zeros o elements buits 3
4.2 Valors extrems 4

5 Anàlisis de les dades **8**
5.1 Selecció dels grups de dades 8
5.2 Comprovació de la normalitat i homogeneïtat de la variància 8
5.3 Proves estadístiques 8

6 Representació dels resultats **9**

7 Conclusions **9**

References **9**

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:

http://www.edubcn.cat/rcs_gene/treballs_recerca/2011-2012-08-1-TR.pdf

https://es.wikipedia.org/wiki/Anh%C3%ADdrido_sulfuroso_en_vinos

https://inta.gob.ar/sites/default/files/script-tmp-2___los_estmulos_cidos_del_vino.pdf

- **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 concentració 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**: és al que es 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**
- **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 aquest 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 característiques 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 aspected 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

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

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

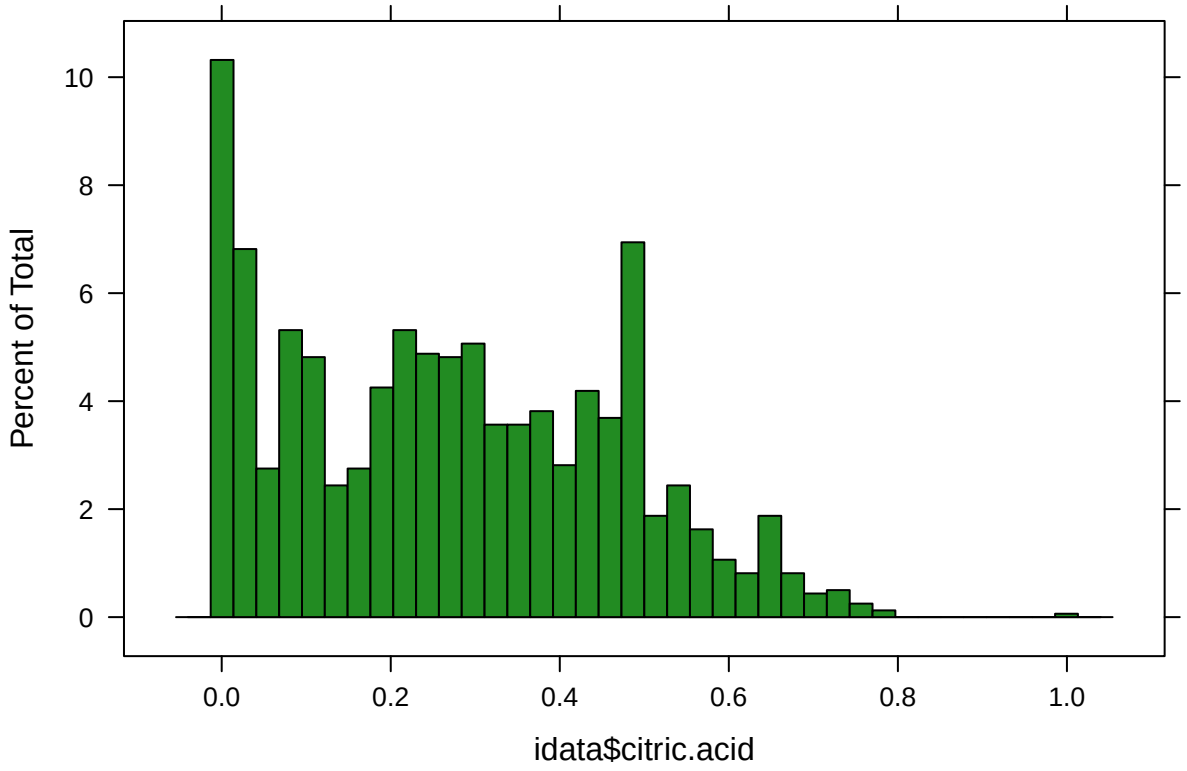
Podem comprovar que totes les variables carregades del *CSV* s'han reconegut com a numèriques (o integer). Això vol dir que no hi ha cap valor de tipus caràcter que s'hagi establert com a centinela per indicar element buit, perque llavors s'hauria fixat el tipus de dades com factor. També podriem haver 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.

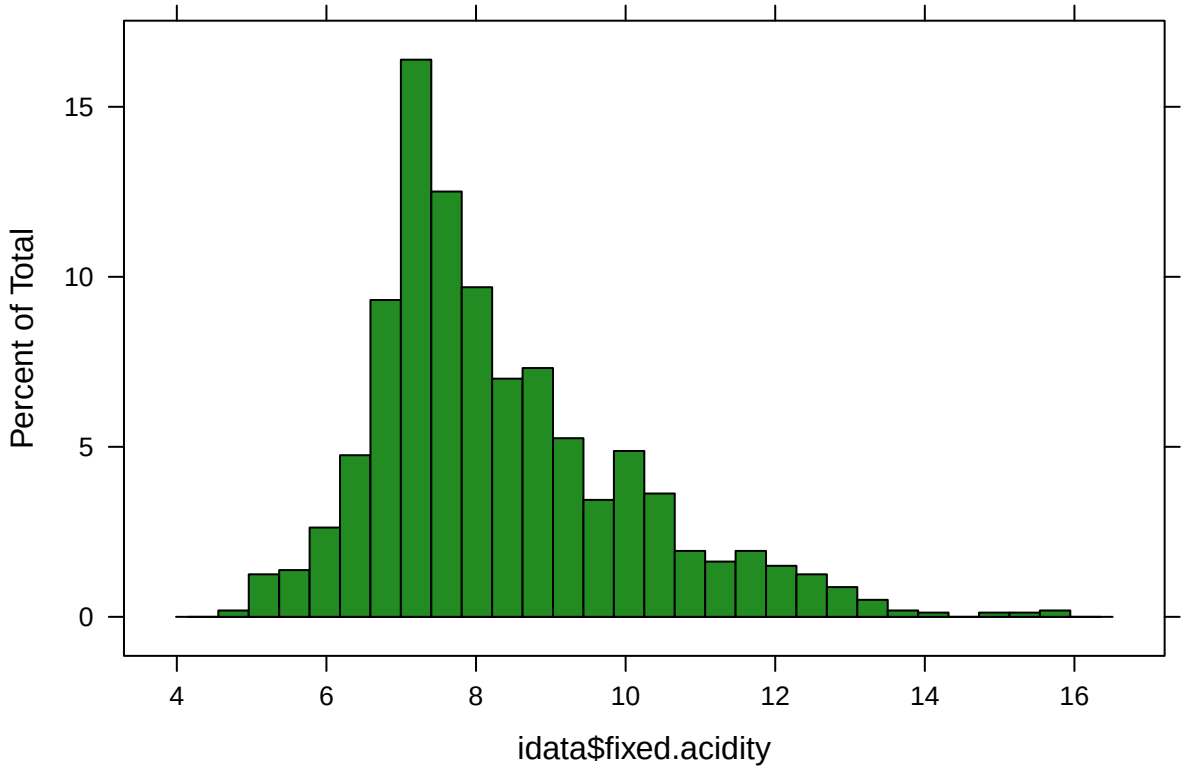
fixed acidity

```
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')
```



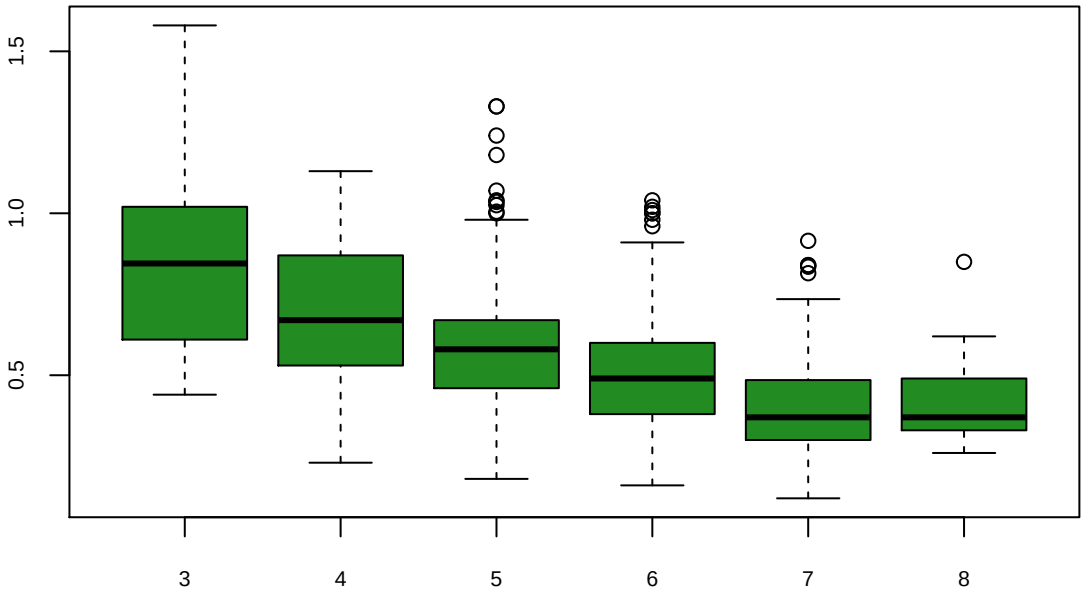
Considerem que són valors atípics propis de la distribució.

volatile acidity

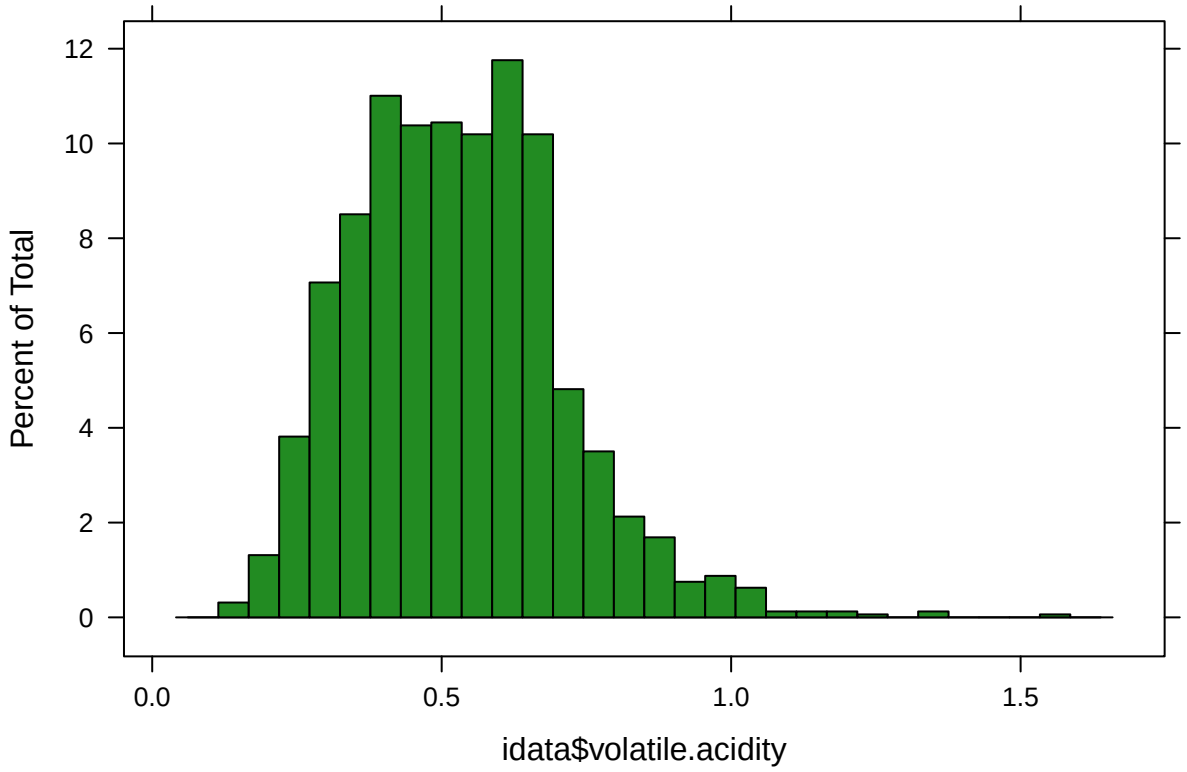
```
boxplot.stats(idata$volatile.acidity)$out
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020
## [12] 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

```
boxplot(idata$volatile.acidity~idata$quality,col="forestgreen",cex.axis=0.7)
```



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



citric acid

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

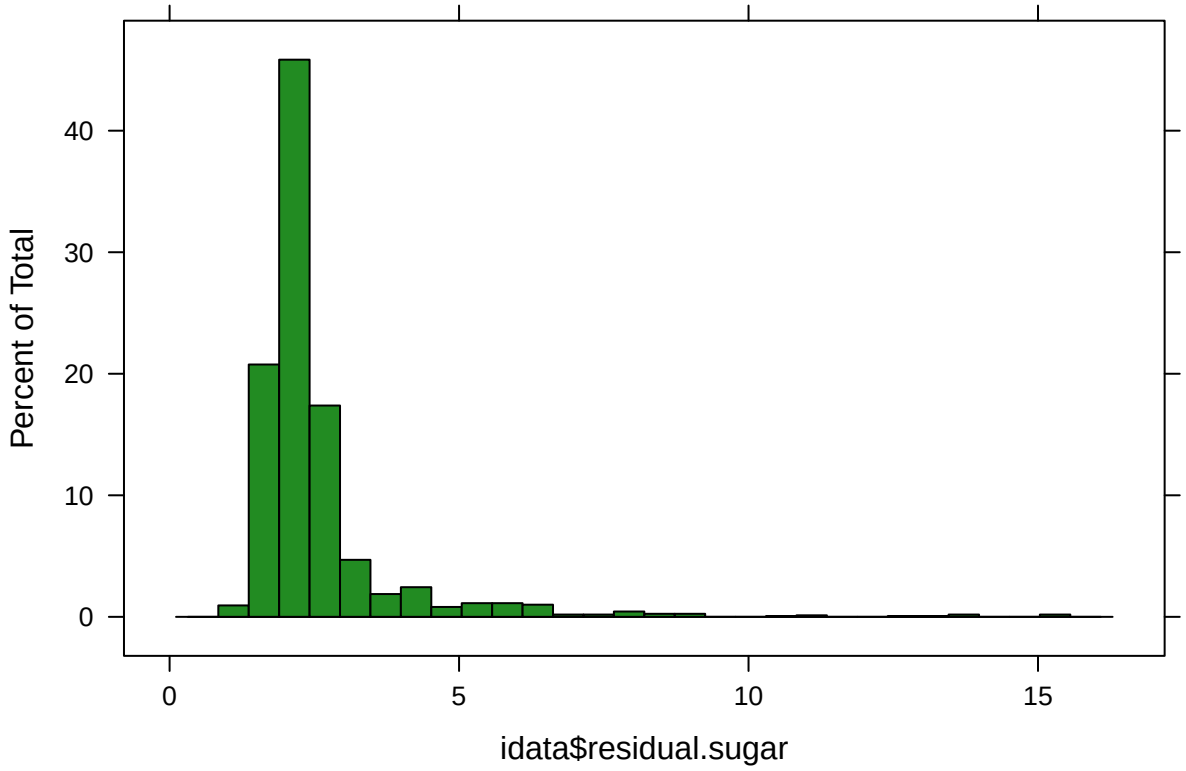
```
## [1] 1
```

residual sugar

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

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

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




```
#boxplot(idata$residual.sugar~idata$quality,col="forestgreen",cex.axis=0.7)
```

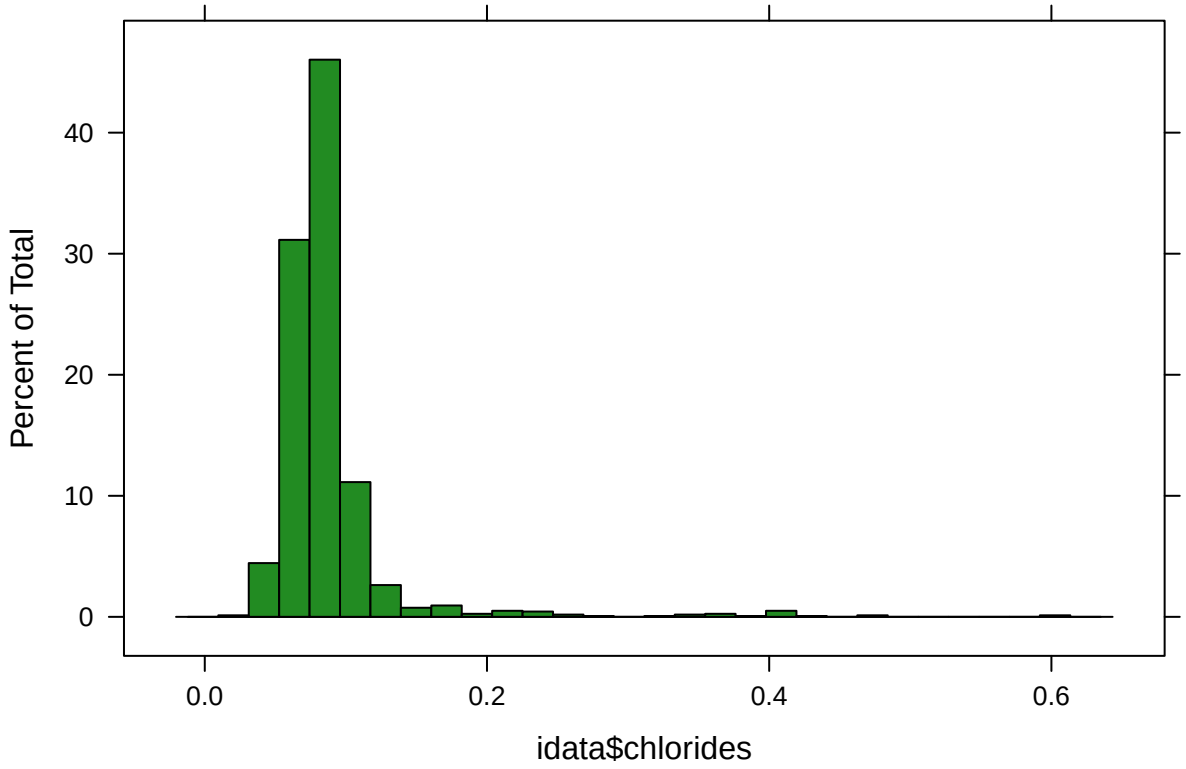
En el cas de *residual sugar* ens trobem inicialment amb un número gens despreciable de outliers: 155. En l'histograma es veuen com valors molt allunyats de la resta de observacions, i per tant són candidats a considerar-los valors atípics contaminants i ser tractats com a tal. Però tal com vam descriure 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

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

```
##   [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178
##  [12] 0.146 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343
##  [23] 0.186 0.213 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122
##  [34] 0.122 0.174 0.121 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171
##  [45] 0.226 0.226 0.250 0.148 0.122 0.124 0.124 0.143 0.222 0.039 0.157
##  [56] 0.422 0.034 0.387 0.415 0.157 0.157 0.243 0.241 0.190 0.132 0.126
##  [67] 0.038 0.165 0.145 0.147 0.012 0.012 0.039 0.194 0.132 0.161 0.120
##  [78] 0.120 0.123 0.123 0.414 0.216 0.171 0.178 0.369 0.166 0.166 0.136
##  [89] 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414 0.166 0.168 0.415
## [100] 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205 0.039 0.235
## [111] 0.230 0.038
```

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



https://www.scribd.com/doc/18175580/Analisis-de-Vinos

5 Anàlisis de les dades

5.1 Selecció dels grups de dades

(Formar conjunts de test i prova)

5.2 Comprovació de la normalitat i homogeneïtat de la variància

http://www.dummies.com/programming/r/how-to-test-data-normality-in-a-formal-way-in-r/

http://www.cookbook-r.com/Statistical_analysis/Homogeneity_of_variance/

5.3 Proves estadístiques

(Regressió lineal multivariant)

6 Representació dels resultats

7 Conclusions

References

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.