PR2-Tipologia

Octavi Castro Nuez

27 de desembre de 2017

Contents

L'objectiu d'aquesta activitat serà el tractament d'un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (https://www.kaggle.com). Les diferents tasques a realitzar (i justificar) són les següents:

Per aquesta activitat triem un dataset de kaggle, concretament, un sobre vins que es pot trobar a https://www.kaggle.com/zynicide/wine-reviews

1 Descripció del dataset. Perquè és important i quina pregunta/problema pretèn respondre?

A l'adreça anterior trobarem dos datasets, enlloc d'un, i treballarem amb els dos, per poder tenir un major nombre de mostres de vins. El primer dataset que trobem conté més de 150 mil vins i no conté tots els camps. El segon dataset conté una mica menys de 130 mil mostres i disposa de tres camps més que l'anterior.

Aquests datasets contenen informació sobre vins que han obtingut una puntuació entre 80 i 100 punts (el màxim és 100 punts). Aquestes dades van ser obtingudes mitjançant scraping de WineEnthusiast (http://www.winemag.com/?s=&drink_type=wine) durant la setmana del 15 de juny de 2017.

Passem a veure la llista d'atributs:

- Points: el nombre de punts obtinguts pel vi, va de 1 fins a 100, però aquí només hi ha vins amb una puntuació de 80 o més.
- Variety: el tipus de raïm que s'utilitza per elaborar el vi

- Descripction: unes poques frases del tastador del vi descrivint el tast.
- Country: el país d'on prové el vi.
- Province: la província o estat d'on prové el vi. (Comentar que Province es refereix més aviat a la zona on es produeix el vi o a la seva denominació d'origen, ja que si la revisem podrem veure que per Country = Spain tenim una provincia anomenada Northern Spain que correspondria a les tres comunitats automes que conformen la D.O. Rioja.)
- Region 1: l'àrea vinícola d'una província o estat.
- Region 2: de vegades hi ha una regió més específica de l'àrea vinícola, però aquest camp pot estar en blanc.
- Winery: el celler que ha fet el vi.
- Designation: la vinya dins del celler d'on procedeixen els raïms que han fet el vi.
- Price: el cost per una ampolla del vi (en dollars).
- Taster Name: el nom de la persona que va fer el tast i la ressenya del vi.
- Taster Twitter Handle: compte a Twitter del tastador del vi.
- Title: El títol del vi i en molts casos la data de la verema.

Els tres últims camps només es troben presents en el segon dataset.

En aquests datasets trobem força informació sobre vins amb una bona puntuació, i del qual podem veure alguns estudis fets. En el nostre cas pretendrem respondre a la pregunta següent:

Quina zona m'ofereix la millor relació qualitat-preu per a una varietat concreta?

```
# els valors absents venen indicats per un camp en blanc.
# llegim el primer dataset.
wine.150 <- read.csv("winemag-data_first150k.csv", na.strings = "")
# llegim el segon dataset.
wine.130 <- read.csv("winemag-data-130k-v2.csv", na.strings = "")
```

2 Neteja de les dades.

Examinem les dades dels datasets.

En el primer dataset tenim 150930 mostres i un total de 11 camps.

Amb els factors següents:

```
str(wine.150)
```

```
## 'data.frame':    150930 obs. of  11 variables:
##  $ X          : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ country    : Factor w/ 48 levels "Albania","Argentina",..: 47 41 47 47 16 41 41 41 47 47 ...
##  $ description: Factor w/ 97821 levels ". Big, lively and very intense, this powerful Amarone opens 
##  $ designation: Factor w/ 30621 levels "¡Adentro! Red",..: 17369 4413 25554 22403 14344 19205 23925 4
##  $ points     : int  96 96 96 96 95 95 95 95 95 95 ...
##  $ price      : num  235 110 90 65 66 73 65 110 65 60 ...
##  $ province   : Factor w/ 455 levels "Achaia","Aconcagua Costa",..: 52 275 52 283 315 275 275 275 283
##  $ region_1   : Factor w/ 1236 levels "Abruzzo","Adelaida District",..: 739 1071 529 1223 67 1071 107
##  $ region_2   : Factor w/ 18 levels "California Other",..: 8 NA 14 18 NA NA NA NA 18 14 ...
##  $ variety    : Factor w/ 632 levels "Agiorgitiko",..: 71 550 470 403 423 550 550 550 403 403 ...
##  $ winery     : Factor w/ 14810 levels ":Nota Bene","'37 Cellars",..: 7305 1240 9050 11038 5106 10203
```

En el segon dataset tenim 129971 mostres i un total de 14 camps.

Amb els factors següents:

```
str(wine.130)
```

```
## 'data.frame':    129971 obs. of  14 variables:
##  $ X                     : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ country               : Factor w/ 43 levels "Argentina","Armenia",..: 23 32 43 43 43 38 23 16 18 1
##  $ description           : Factor w/ 119955 levels ". A delightfully intriguing "White Burgundy" blen
##  $ designation           : Factor w/ 37979 levels "??? Vineyard",..: 36976 2352 NA 28123 36715 1996 3
##  $ points                : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ price                 : num  NA 15 14 13 65 15 16 24 12 27 ...
##  $ province              : Factor w/ 425 levels "Achaia","Aconcagua Costa",..: 334 110 269 220 269 26
##  $ region_1              : Factor w/ 1229 levels "Abruzzo","Adelaida District",..: 425 NA 1218 550 12
##  $ region_2              : Factor w/ 17 levels "California Other",..: NA NA 17 NA 17 NA NA NA NA NA .
##  $ taster_name           : Factor w/ 19 levels "Alexander Peartree",..: 10 16 15 1 15 13 10 16 2 16 .
##  $ taster_twitter_handle: Factor w/ 15 levels "@AnneInVino",..: 5 11 8 NA 8 13 5 11 NA 11 ...
##  $ title                 : Factor w/ 118840 levels ":Nota Bene 2005 Una Notte Red (Washington)",..: 7
##  $ variety               : Factor w/ 707 levels "Abouriou","Agiorgitiko",..: 692 452 438 481 442 593
##  $ winery                : Factor w/ 16757 levels ":Nota Bene","1+1=3",..: 11641 12988 13054 14432 14
```

Com ja havíem comentat el segon dataset conté un major nombre de camps, per tant, haurem d'igualar-los per a poder unir-los.

```
# aprofitem per eliminar el primer camp que son les row.names
wine.t <- rbind(wine.150[,-1], wine.130[, -c(1,10,11,12)])
str(wine.t)
```

```
## 'data.frame':    280901 obs. of  10 variables:
##  $ country    : Factor w/ 50 levels "Albania","Argentina",..: 47 41 47 47 16 41 41 41 47 47 ...
##  $ description: Factor w/ 169430 levels ". Big, lively and very intense, this powerful Amarone opens
##  $ designation: Factor w/ 47239 levels "¡Adentro! Red",..: 17369 4413 25554 22403 14344 19205 23925 4
##  $ points     : int  96 96 96 96 95 95 95 95 95 95 ...
##  $ price      : num  235 110 90 65 66 73 65 110 65 60 ...
##  $ province   : Factor w/ 490 levels "Achaia","Aconcagua Costa",..: 52 275 52 283 315 275 275 275 283
##  $ region_1   : Factor w/ 1332 levels "Abruzzo","Adelaida District",..: 739 1071 529 1223 67 1071 107
##  $ region_2   : Factor w/ 18 levels "California Other",..: 8 NA 14 18 NA NA NA NA 18 14 ...
##  $ variety    : Factor w/ 756 levels "Agiorgitiko",..: 71 550 470 403 423 550 550 550 403 403 ...
##  $ winery     : Factor w/ 19186 levels ":Nota Bene","'37 Cellars",..: 7305 1240 9050 11038 5106 10203
```

2.1 Selecció de les dades d'interès a analitzar. Quins són els camps més rellevants per tal de respondre al problema?

Abans de procedir amb aquest apartat passarem a eliminar els elements repetits que pugui contenir el nostre dataset final.

```
wine.t <- wine.t[!duplicated(wine.t), ]
str(wine.t)
```

```
## 'data.frame':    170531 obs. of  10 variables:
##  $ country    : Factor w/ 50 levels "Albania","Argentina",..: 47 41 47 47 16 41 41 41 47 47 ...
##  $ description: Factor w/ 169430 levels ". Big, lively and very intense, this powerful Amarone opens
##  $ designation: Factor w/ 47239 levels "¡Adentro! Red",..: 17369 4413 25554 22403 14344 19205 23925 4
##  $ points     : int  96 96 96 96 95 95 95 95 95 95 ...
##  $ price      : num  235 110 90 65 66 73 65 110 65 60 ...
##  $ province   : Factor w/ 490 levels "Achaia","Aconcagua Costa",..: 52 275 52 283 315 275 275 275 283
##  $ region_1   : Factor w/ 1332 levels "Abruzzo","Adelaida District",..: 739 1071 529 1223 67 1071 107
##  $ region_2   : Factor w/ 18 levels "California Other",..: 8 NA 14 18 NA NA NA NA 18 14 ...
```

```
##  $ variety    : Factor w/ 756 levels "Agiorgitiko",..: 71 550 470 403 423 550 550 550 403 403 ...
##  $ winery     : Factor w/ 19186 levels ":Nota Bene","'37 Cellars",..: 7305 1240 9050 11038 5106 1020
```

Per a respondre la pregunta que plantegem en l'apartat 1 considerem que els camps rellevants són els següents:

country, points, price, province, variety, winery

```
# establim els índexs de les columnes a eliminar
indexs <- c(2,3,7,8)
wine.a <- wine.t[ ,-indexs]
dim(wine.a)
```

```
## [1] 170531      6
```

```
names(wine.a)
```

```
## [1] "country"  "points"   "price"    "province" "variety"  "winery"
```

2.2 Les dades contenen zeros o elements buits? I valors extrems? Com gestionaries cadascun d'aquests casos?

```
# mostrem les variables que contenen buits i la quantitat d'elements buits que tenen
vbles.buits <- names(wine.a)[!complete.cases(t(wine.a))]
sapply(wine.a[vbles.buits], function(x) sum(is.na(x)))
```

```
##  country    price province  variety
##       60    12841       60        1
```

Veiem que tant country com province contenen el mateix nombre d'elements buits i que winery no en conté cap, per tant, podem intentar completar les files a partir d'aquest camp complert. Tot i que, primer haurem de normalitzar els camps per homogeneitzar-los i evitar, així, errors d'escritura.

Pel que fa a preu tenim diverses opcions:

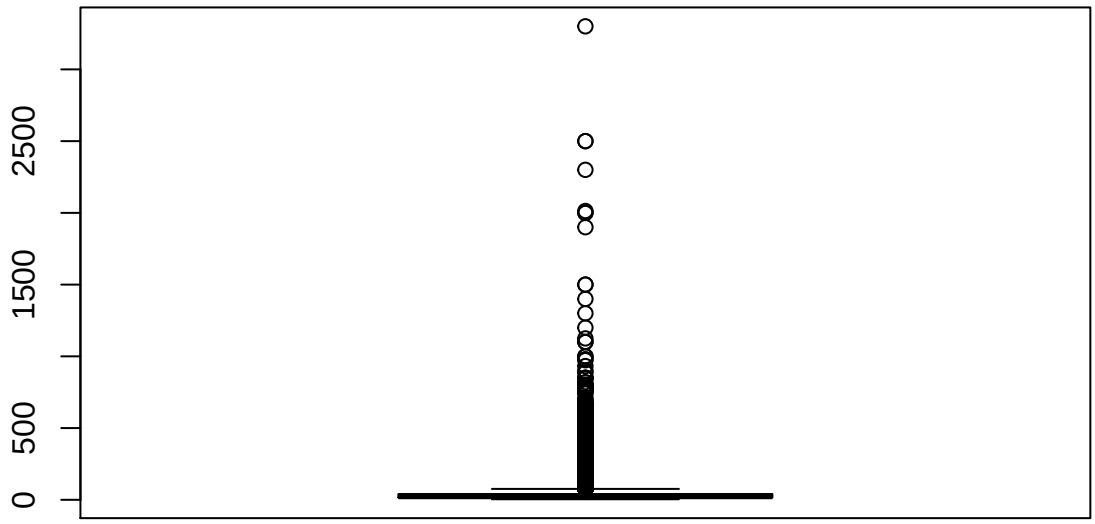
1. Intentar aconseguir els preus originals de la web original.
2. Mirar d'assignar valor a aquests camps per mitjà d'algun algorisme com per exemple kNN o kmeans.
3. Eliminar les files amb camps buits.

Tot i que, la primera opció seria la idònea tenim un nombre massa elevat de valors faltants, per la qual cosa optarem per fer una anàlisi per cada una de les altres dues opcions i compararem el resultat.

Per últim, veiem que només tenim una mostra sense variety, mirem la web original i la web de la bodega, però no obtenim més informació al respecte. Per sort al camp description d'aquest vi se'ns indica que es tracta d'un Petite Syrah, per tant, assignarem aquesta varietat al camp variety.

Pel que fa als valors extrems comprovarem si price, l'únic camp numèric que tenim, en té.

```
boxplot(wine.a$price)
```




```
#indices <- which(wine.a$price  %in%  boxplot.stats(wine.a$price)$out)
#length(boxplot.stats(wine.a$price)$out)
#length(unique(boxplot.stats(wine.a$price)$out))
#min(boxplot.stats(wine.a$price)$out)
#max(boxplot.stats(wine.a$price)$out)
```

En el boxplot veiem que aquest camp té un gran nombre de valors extrems, però no ens dona massa informació al respecte.

Per tant, anem a veure'ls numèricament.

Tenim un total de 9568 mostres catalogades com a valors extrems amb un total de 342 valors diferents, que van des de'l 77 fins al 3300.

Tots aquests valors entren dins del rang de preus del vi. De totes formes donarem un cop d'ull a aquells que tinguin preus de 4 xifres per si hi hagués hagut un error a l'hora de ficar el preu amb els decimals.

```
wine.a[which(wine.a[,3] > 1000),]
```

```
##        country points price   province                 variety
## 10652  Austria     94  1100     Wachau         Grüner Veltliner
## 13319       US     91  2013 California               Chardonnay
## 26297   France    100  1400  Champagne               Chardonnay
## 34921   France     99  2300   Bordeaux Bordeaux-style Red Blend
## 34923   France     98  1900   Bordeaux Bordeaux-style Red Blend
## 34928   France     97  1100   Bordeaux Bordeaux-style Red Blend
## 34940   France     96  1300   Bordeaux Bordeaux-style Red Blend
## 34943   France     96  1200   Bordeaux Bordeaux-style Red Blend
## 166771  France     96  2500   Bordeaux Bordeaux-style Red Blend
```

```
## 216283  France     97  2000   Bordeaux Bordeaux-style Red Blend
## 231221  France     88  3300   Bordeaux Bordeaux-style Red Blend
## 249311  France     96  2500   Burgundy               Pinot Noir
## 262684  France    100  1500   Bordeaux Bordeaux-style Red Blend
## 262686  France    100  1500   Bordeaux Bordeaux-style Red Blend
## 264495  France     96  2000   Burgundy               Pinot Noir
## 264512  France     94  1125   Burgundy               Pinot Noir
##                               winery
## 10652                 Emmerich Knoll
## 13319                          Blair
## 26297                           Krug
## 34921                 Château Latour
## 34923                Château Margaux
## 34928  Château La Mission Haut-Brion
## 34940      Château Mouton Rothschild
## 34943             Château Haut-Brion
## 166771                Château Pétrus
## 216283                Château Pétrus
## 231221      Château les Ormes Sorbet
## 249311 Domaine du Comte Liger-Belair
## 262684     Château Lafite Rothschild
## 262686          Château Cheval Blanc
## 264495 Domaine du Comte Liger-Belair
## 264512 Domaine du Comte Liger-Belair
```

Podem destacar dues coses d'aquest llistat.

La primera seria que molts dels vins més cars provenen de Bordeaux a França que sabem és una regió amb molta fama i, per tant, és habitual veure vins amb preus elevats.

La segona cosa a destacar és que tots aquests vins tenen més de 90 punts, l'excepció és el vi amb el preu més elevat.

Després de comprovar els preus a internet veiem que el preu del vi més car és una errada ja que podem trobar-lo per uns 30$, com podem comprovar a http://www.hachette-vins.com/guide-vins/les-vins/ch-les-ormes-sorbet-2013-2017/201706208/ o a https://www.chateau.fr/chateau-les-ormes-sorbet-2013-cbo-12x75cl-rouge.html.

Per tant, procedirem a arranjar el preu i a deixar-lo en 33 dollar, enlloc dels 3300$ que actualment té.

```
wine.a[which(wine.a[,"price"] == 3300), "price"] <- 33.0
```

Per a la resta de vins comprovem que el preu és correcte i donarem aquest punt per finalitzat.

Abans de continuar és convenient comprovar que les dades siguin del tipus corresponent i normalitzar/estandarditzar.

```
res <- sapply(wine.a, class)
kable(data.frame(variables=names(res), classe=as.vector(res)))
```

variables	classe
country	factor
points	integer
price	numeric
province	factor
variety	factor
winery	factor

L'únic tipus que haurem de canviar és el de points, ja que és una variable quantitativa, encara que estigui representada per valors numérics sencers.

```
wine.a$points <- as.factor(wine.a$points)
res <- sapply(wine.a, class)
kable(data.frame(variables=names(res), classe=as.vector(res)))
```

variables	classe
country	factor
points	factor
price	numeric
province	factor
variety	factor
winery	factor

Ara que ja tenim els tipus de variable correctament assignats procedim a normalitzar/estandarditzar.

```
#wine.a <- winet #ESBORRAR AQUESTA LINEA ABANS D'ENTREGAR!!!!!
txtvar <- c("country", "province", "variety", "winery")
accents <- c("áéíóúàèìòùâêîôûäëïöüãõű")
noaccents <- c("aeiouaeiouaeiouaeiouaou")
puntua <- c("-_")
nopuntua <- ("  ")
f.origin = f.blancs = f.minus = f.accents = f.puntua = 0
j <- 1
f.puntua <- 1
for(i in txtvar) {
  f.origin[j] <- nlevels(wine.a[,i])
  # traiem espais en blanc al principi i final del text
  wine.a[,i] <- as.factor(trimws(wine.a[,i], "both"))
  f.blancs[j] <- nlevels(wine.a[,i])
  # possem tot el text en minúscula
  wine.a[,i] <- as.factor(tolower(wine.a[,i]))
  f.minus[j] <- nlevels(wine.a[,i])
  # eliminem accents
    wine.a[,i] <- as.factor(chartr(accents, noaccents, wine.a[,i]))
  f.accents[j] <- nlevels(wine.a[,i])
  wine.a[,i] <- as.factor((chartr(puntua, nopuntua, wine.a[,i])))
  wine.a[,i] <- as.factor(gsub("\\.", "", wine.a[,i]))
  wine.a[,i] <- as.factor(gsub("\\,", "", wine.a[,i]))
  wine.a[,i] <- as.factor(gsub("\\:", "", wine.a[,i]))
  wine.a[,i] <- as.factor(gsub("\\;", "", wine.a[,i]))
  wine.a[,i] <- as.factor(gsub("\\'", "", wine.a[,i]))
  f.puntua[j] <- nlevels(wine.a[,i])
  j <- j + 1
}
kable(data.frame(variables=txtvar, original=f.origin, sense.blancs=f.blancs, en.minuscules=f.minus, sens
```

variables	original	sense.blancs	en.minuscules	sense.accents	sense.puntuacions
country	50	50	50	50	50
province	490	490	490	490	490
variety	756	756	756	756	756
winery	19186	19186	19158	19119	19086

Un cop normalitzades les dades passarem a assignar el valor "petite shyrah" a l'exemple sense variety. Però, abans comprovarem que aquest valor existeixi per no crear un nou factor.

```
varietats<-grep("s[i|y]rah",wine.a$variety)
sort(unique(wine.a[varietats, "variety"]))
```

```
##  [1] cabernet sauvignon syrah cabernet syrah
##  [3] carignan syrah           carmenere syrah
##  [5] garnacha syrah           grenache syrah
##  [7] malbec syrah             merlot syrah
##  [9] monastrell syrah         mourvedre syrah
## [11] petite sirah             petite syrah
## [13] pinot noir syrah         sangiovese syrah
## [15] syrah                    syrah bonarda
## [17] syrah cabernet           syrah cabernet franc
## [19] syrah cabernet sauvignon syrah carignan
## [21] syrah grenache           syrah grenache viognier
## [23] syrah malbec             syrah merlot
## [25] syrah mourvedre          syrah petit verdot
## [27] syrah petite sirah       syrah tempranillo
## [29] syrah viognier           tannat syrah
## [31] tempranillo syrah
## 756 Levels: abouriou agiorgitiko aglianico aidani airen ... zweigelt
```

Veiem que aquesta varietat es presenta amb diferents noms. El mateix ens passarà amb altres varietats com podem comprovar en els següents enllaços:

https://vivancoculturadevino.es/blog/2015/07/17/variedades-de-uva/ https://turismodevino.com/saber-de-vino/tipos-de-uva-en-el-vino/

Reassignarem algunes varietat, tot i que, per no allargar més la neteja (i la pràctica) només juntarem les que veiem són formes diferents d'escriure una mateixa varietat, com per exemple shirah i shyrah. Però deixarem aquelles que tot i ser la mateixa varietat rebin diferents noms en diferents Denominacions d'Origen (DO), com per exemple shiraz, que és el nom australià de la varietat syrah com podem veure a https://www.leaf.tv/articles/what-is-a-shiraz-wine/.

Pel que fa als vins formats per més d'una varietat mantindrem l'ordre, és a dir, si tenim les varietats syrah tempranillo (mostra [28]) la considerarem diferent a tempranillo syrah (mostra [31]), ja que indica que la varietat dominant en el vi és la primera i, per tant, el vi tindrà propietats diferents.

Per tant, després d'examinar les varietats actuals durem a terme els canvis següents:

aragones, aragonez = aragones assyrtico, assyrtiko = assyrtiko carignan, carignane, carignano = carignan chardonel, chardonelle = chardonel durella, durello = durella insolia, inzolia = inzolia malagousia, malagouzia = malagouzia malvasia, mavazija = malvasia moscatel, muscatel = moscatel moschofilero, moscofilero = moschofilero muscadel, muscadelle = muscadel muscat blanc a petits grains, muscat blanc a petit grain = muscat blanc a petit grain muscat, muskat = muskat petit verdot, petite verdot = petite verdot pinot bianco, pinot blanc = pinot blanc pinot nero, pinot noir = pinto noir pinot grigio, pinot gris = pinot gris sirah, syrah = syrah tinta de toro, tinta del toro = tinta de toro tinta fina, tinto fino = tinta fina tinta del pais, tinto del pais = tinta del pais tocai, tokay = tokay vranac, vranec = vranac

```
o.variety <- c("aragonez", "assyrtico", "chardonelle", "durello", "insolia", "malagousia", "malvazija",
n.variety <- c("aragones", "assyrtiko", "chardonel", "durella", "inzolia", "malagouzia", "malvasia", "mu
#wine.a <- w.net # ELIMINAR ABANS D'ENTREGAR!!!!!!
# ho farem en dos vegades
# primer els que no presenten modficacions
for(n in 1:length(o.variety)) {
  wine.a[which(wine.a[,"variety"] == o.variety[n]), "variety"] <- as.factor(n.variety[n])
```

```
}
# segon els que sí en presenten
om.variety <- c("carignan[e|o]", "muscat", "pinot grigio", "sirah", "tocai")
nm.variety <- c("carignan", "muskat", "pinot gris", "syrah", "tokay")
for(n in 1:length(om.variety)) {
  indexs<-grep(om.variety[n],wine.a$variety)
  wine.a[indexs, "variety"] <- as.factor(nm.variety[n])
}
#nlevels(factor(wine.a$variety))
```

Per a variety el nombre de factors actual és 717. Ha disminuit en 39.

```
#winet <- wine.a
#sort(factor(unique(winet[,5])))
```

Passem a obtenir les mostres incompletes de country i province i veure si són les mateixes.

```
#wine.nacountry <- subset(wine.a, is.na(wine.a$country))
#wine.nacountry[!duplicated(wine.nacountry$winery), "winery"]
#wine.naprovince <- subset(wine.a, is.na(wine.a$province))
#identical(wine.nacountry, wine.naprovince)
#wine.naprice <- subset(wine.a, is.na(wine.a$price))
#wine.navariety <- subset(wine.t, is.na(wine.a$variety))
#wine.navariety
#nawinery <- wine.a[wine.a[,"winery"] == wine.navariety$winery & wine.a[,"province"] == wine.navariety$
#nawinery
```

Veiem que són iguals, per tant, només haurem de cercar les dades un cop. Agrupem per winery.

3 Anàlisi de les dades.

3.1 Selecció dels grups de dades que es volen analitzar/comparar.

3.2 Comprovació de la normalitat i homogeneïtat de la variància. Si és necessari (i possible), aplicar transformacions que normalitzin les dades.

3.3 Aplicació de proves estadístiques (tantes com sigui possible) per comparar els grups de dades.

4 Representació dels resultats a partir de taules i gràfiques.

5 Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

6 Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

```
#sort(unique(wine$variety))
#winet <- winet[!duplicated(winet), ]
#dim(winet)
#summary(winet)
#wp <- as.matrix(sort(table(winet$price), decreasing = TRUE))
#length(wp)
#kable(wp[1:50,], format="markdown", col.names = "total botellas de este importe")
```