Pràctica 2: Neteja i validació de dades - Projecte analític sobre el preu i maridatge de vins D.O. Penedès

Joan Morral Ventura

Semestre 2021.1

Contents

1. Descripció del dataset

Per a consumidors de vi ocasionals, la compra d'una botella de vi acostuma a realitzar-se per recomanacions del venedor o per fixació de preu. Per altra banda, si s'acostuma a portar en un àpat social, sempre resulta un detall que maridi amb l'àpat.

La idea d'aquesta pràctica és la de poder ajudar a compradors no experts a obtenir un preu de referència, per veure si el que s'ofereix a la bodega és car o barat. Per altra banda, es presentarà una carta de maridatges amb el tipus de cupatge de forma que siguem nosaltres els que escollim el tipus de vi abans d'anar a la tenda. Per tal de complementar els punts anteriors també es realitzarán d'altres analítiques amb la voluntat de completar els punts sol · licitats en la pràctica.

La estatègia per per a obtenir la informació passa per capturar les dades dels cellers i bodegues com a fitxes de tast i compra dels propis vins, processar-les en el procés de neteja i tenir un dataset que permeti obtenir informació del vi.

El dataset d'entrada consta de 24 atributs:

Variables	Descripció
D.O	Denominació d'origen.
NOM DEL VI	Nom del producte
CELLER	Celler productor del vi
DIRECCIÓ	Adreça del celler
VOLUM	Volum de l'envàs en el que es presenta el vi
ENVELLIMENT	Tipus d'envelliment, pot ser: jove, semi criança, criança, reserva, gran reserva
ESTIL	Estil del vi, permet granular el tipus però és poc emprat
ANYADA	Any de producció de la botella
TERRENY	Característiques de posicionament, composició del terreny, alçada, etc.
ELABORACIÓ	Procés d'elaboració del vi. Inoxidable o bota per assignar si és jove o criança
TIPOLOGIA	Indica si és un vi negre, blanc, rosat, cava, escumós o dolç
CUPATGE	Varietat i proporció de tipus de raïm emprat pel vi
VISTA	Descripció de color i aspecte del vi
AROMA	Descripció de les aromes primàries, secundàries i terciàries del vi
PALADAR	Descripció del gust i postgust
MARIDATGE	Descripció de maridatge
SERVEI	Temperatura de servei (dóna informació sobre el tipus de vi)
GRAU ALC.	Grau d'alcohol en percentatge
ACIDESA	Acidesa tartàrica del vi
SULFURÓS	Valor de sulfits, s'obliga a posar la quantitat atès que hi ha gent al · lèrgica
T.	
SUCRES	Sucres residuals en la botella
RES.	
PH	Valoració àcid-base del vi
PREU	PVP en euros
PREMIS	Premis dels quals disposa el vi

Les variables que estructuren el dataset són la denominació d'origen, la tipologia del vi i el camp nom del vi (és la referència del producte). Les files que no continguin aquesta informació es podràn considerar nules.

La complexitat del dataset és que el nombre de camps buits resulta elevat perquè cada celler posa la informació que ell creu rellevant. No obstant, a l'emprar un llenguatge molt específic, es poden generar sentències que recequin mots concrets i permetin determinar el cupatge, la vista, el aroma, el paladar i el maridatge. Per a que s'entengui, una descripció pot dir que un vi porta un cupatge de Merlot i Ull de llebre, mentre que un altra pot dir que porta Ull de llebre i Merlot. Per tant, la neteja caldrà que identifiqui els dos datasets amb la mateixa informació. El tractament de les codificacions de maridatge i **coupage** és el que justifica l'extensió del document.

Per tal de reduir-ne les codificacions i fer-lo més simple, les descripcions del tast s'han convertit en columnes que engloben la descripció en les tres fases de tast del vi: fase visual, olfactiva i gustativa. Les noves columnes aspecte, color, aromes primàries, aromes secundàries, aromes terciàries, gust i post gust s'ha dissenyat per a que siguin factoritzables.

Els camps grau d'alcohol, acidesa, sulfurós total, sucres residuals i pH són els camps numèrics que donen els cellers en funció del preu del vi, ja que es reserva a vins de qualitat. Per altra banda, l'altra variable numeral a destacar és el preu que indica el PVP.

Finalment, també s'han afegit d'altres camps amb la finalitat d'ajudar a complementar la informació dels camps buits o aportar més informació que justifiqui el preu, no obstant, no s'han aplicat per falta de temps.

2. Lectura del fitxer i preparació de les dades

En aquest apartat: - es carrega el contingut del fitxer a l'objecte vins i validació de camps (2.1) - s'estandaritzen els noms de les columnes (2.2) - es converteixen els tipus de les dades (2.3) - es filtra la D.O. Penedès essent la única amb la que treballarem (2.4)

Carreguem les llibreries necessàries per a executar la pràctica.

```
# Carreguem els paquets R que utilitzarem al llarg de la pràctica
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.0.5
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr) # Permet poder filtrar els camps amb determinades condicions
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.6     v purrr   0.3.4
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(rmarkdown)
```

```
## Warning: package 'rmarkdown' was built under R version 4.0.5
```

```
library(ResourceSelection) # Llibreria pel test Hosmer-Lemeshow
```

```
## Warning: package 'ResourceSelection' was built under R version 4.0.5
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
library(nortest) # Llibreria per a realitzar lillie.test - contrast de normalitat
library(corrgram) #Llibreria per aplicarr correlació
```

```
## Warning: package 'corrgram' was built under R version 4.0.5
```

```
# Ocultació missatges de llibreries
suppressWarnings(suppressMessages(library(VIM)))

# Definim els colors segons llibre d'estil UOC
uoc_palette <- c("#000078", "#73EDFF", "#38FF90", "#FFE000", "#FF7D87", "#BD9EFF", "#FF87FF")
```

2.1 Lectura de dades, creació d'objecte vins i validació de camps

Llegim el fitxer vins.csv i guardem les dades en un objecte amb identificador vins

```
# Carreguem el fitxer de dades dels vins
vins<-read.csv("..\\csv\\vins2.csv", header=TRUE, sep=";", encoding = "UTF-8")

# Verifiquem l'estructura del joc de dades
str(vins)
```

```
## 'data.frame':    486 obs. of  24 variables:
##  $ X.U.FEFF.DO          : chr  "D.O Alella" "D.O Alella" "D.O Alella" "D.O Alella" ...
##  $ NOM.DEL.VI           : chr  "Alta Alella PB" "Marfil Blanc Clàssic" "Marfil Blanc sec" "Ivori Blan
##  $ CELLER               : chr  "Alta Alella" "Alella Vinícola" "Alella Vinícola" "Alella Vinícola" .
##  $ DIRECCIÓ             : chr  "Camí Baix de Tiana, s/n, 08328 Alella" "Rambla de Ángel Guimerá, 62,
##  $ VOLUM                : chr  "75cl" "75cl" "75cl" "75cl" ...
##  $ ENVELLIMENT          : chr  "" "" "" "" ...
##  $ ESTIL.DE.VI          : chr  "" "" "" "" ...
##  $ ANYADA               : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TERRENY.VINYA.TERROIR: chr  "" "" "" "Granodiorites arenoses ("sauló"), Mediterrani marítim" ...
##  $ ELABORACIÓ..AL.CELLER: chr  "El vi blanc Alta Alella PB prové d'agricultura ecològica. Els raïms
##  $ TIPOLOGIA            : chr  "Vi blanc jove eco" "Vi blanc jove eco" "Vi blanc jove eco" "Vi blanc
##  $ RAÏM.CUPATGE         : chr  "Pansa Blanca" "70 % Pansa blanca, 30% Garnatxa Blanca" "100% Pansa Bl
##  $ VISTA                : chr  "" "" "" "Color groc pàl.lid brillant amb un to daurat clar" ...
##  $ AROMA                : chr  "Intensitat mitjana a causa de les característiques de la anyada. Prec
```

```
##  $ PALADAR              : chr  "Entrada voluminosa i amable, amb un pas en boca fresc recordant les a
##  $ MARIDATGE            : chr  "Aperitiu, peix blau, arrossos de peix" "Peix blanc a la graella, amar
##  $ SERVEI               : chr  "" "" "" "" ...
##  $ GRAU.ALC.            : chr  "12%" "12.5%" "12.5%" "12.5%" ...
##  $ ACIDESA              : chr  "" "" "" "" ...
##  $ SULFURÓS.TOTAL       : chr  "" "" "" "" ...
##  $ SUCRES.RESIDUALS     : chr  "" "" "" "" ...
##  $ PH                   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ PREU                 : chr  "9.09\200" "9 \200" "8,51 \200" "9.99\200" ...
##  $ PREMIS               : chr  "" "" "" "" ...
```

Observacions: - Emprant la codificació utf-8, podem veure que s'han carregat les 24 variables amb les seves 486 observacions dins l'objecte vins. - El tipus de contigut majoritàriament són char degut a que són captures de les fitxes de tast i pàgines web dels cellers i bodegues, tota aquesta informació es processarà en l'apartat 3. El seu contingut està en català, de forma que les paraules clau caldrà treballar-les en aquest aspecte. - Els camps qualitatius es presenten amb unitats, de forma que caldrà tenir-ho present en la conversió de tipus de dada. - La taula es presenta forces camps buits degut al mecanisme emprat i a la informació disponible que varia en funció del celler. Aquest aspecte farà que els resultats no siguin del tot fiables ja que només es disposa de 486 camps i caldria ampliar-ne la quantitat, per exemple, a través d'enquesta als cellers. El webscrapping podria resultar resolutiu en els camps qualitatius que és el que s'acostuma a mostrar a la xarxa.

2.2 Neteja dels noms de les columnes

La columna denominació d'origen apareix com X.U.FEEF.DO, també tenim accents (direcció), dièresis (raïm), se separen els espais en punts, tot en majúscules, etc. Es treballa per posició de columnes per si es modifiquen els noms dels camps de les columnes, no obstant, el sistema és susceptible de noves incorporacions. També es decideix integrar les unitats en les columnes.

```
# S'ha plantejat un tractament individual atès que és més gràfic que treballar en llistat
colnames(vins)[1] <- 'do'
colnames(vins)[2] <- 'nom'
colnames(vins)[3] <- 'celler'
colnames(vins)[4] <- 'poblacio'
colnames(vins)[5] <- 'volum_cl'
colnames(vins)[6] <- 'envelliment'
colnames(vins)[7] <- 'estil'
colnames(vins)[8] <- 'anyada'
colnames(vins)[9] <- 'vinya'
colnames(vins)[10] <- 'elaboracio'
colnames(vins)[11] <- 'tipologia'
colnames(vins)[12] <- 'cupatge'
colnames(vins)[13] <- 'vista'
colnames(vins)[14] <- 'aroma'
colnames(vins)[15] <- 'paladar'
colnames(vins)[16] <- 'maridatge'
colnames(vins)[17] <- 'tservei'
colnames(vins)[18] <- 'graduacio_alcoholica_per'
colnames(vins)[19] <- 'acidesa_g_l_tartaric'
colnames(vins)[20] <- 'sulfuros_total_mg_l'
colnames(vins)[21] <- 'sucres_residuals_g_l'
colnames(vins)[22] <- 'ph'
colnames(vins)[23] <- 'preu_euros'
colnames(vins)[24] <- 'premis'

# Modifiquem el ordre de tipologia
```

```
vins<-vins[,c(1,2,11,6,8,7,3,4,5,9,10,12,13,14,15,16,17,18,19,20,21,22,23,24)]

# Verifiquem els canvis en les columnes del set de dades
str(vins)
```

```
## 'data.frame':    486 obs. of  24 variables:
##  $ do                       : chr  "D.O Alella" "D.O Alella" "D.O Alella" "D.O Alella" ...
##  $ nom                      : chr  "Alta Alella PB" "Marfil Blanc Clàssic" "Marfil Blanc sec" "Ivori
##  $ tipologia                : chr  "Vi blanc jove eco" "Vi blanc jove eco" "Vi blanc jove eco" "Vi bla
##  $ envelliment              : chr  "" "" "" "" ...
##  $ anyada                   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ estil                    : chr  "" "" "" "" ...
##  $ celler                   : chr  "Alta Alella" "Alella Vinícola" "Alella Vinícola" "Alella Vinícola
##  $ poblacio                 : chr  "Camí Baix de Tiana, s/n, 08328 Alella" "Rambla de Ángel Guimerá,
##  $ volum_cl                 : chr  "75cl" "75cl" "75cl" "75cl" ...
##  $ vinya                    : chr  "" "" "" "Granodiorites arenoses ("sauló"), Mediterrani marítim" .
##  $ elaboracio               : chr  "El vi blanc Alta Alella PB prové d'agricultura ecològica. Els raï
##  $ cupatge                  : chr  "Pansa Blanca" "70 % Pansa blanca, 30% Garnatxa Blanca" "100% Pansa
##  $ vista                    : chr  "" "" "" "Color groc pàl.lid brillant amb un to daurat clar" ...
##  $ aroma                    : chr  "Intensitat mitjana a causa de les característiques de la anyada.
##  $ paladar                  : chr  "Entrada voluminosa i amable, amb un pas en boca fresc recordant l
##  $ maridatge                : chr  "Aperitiu, peix blau, arrossos de peix" "Peix blanc a la graella,
##  $ tservei                  : chr  "" "" "" "" ...
##  $ graduacio_alcoholica_per: chr  "12%" "12.5%" "12.5%" "12.5%" ...
##  $ acidesa_g_l_tartaric     : chr  "" "" "" "" ...
##  $ sulfuros_total_mg_l      : chr  "" "" "" "" ...
##  $ sucres_residuals_g_l     : chr  "" "" "" "" ...
##  $ ph                       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ preu_euros               : chr  "9.09\200" "9 \200" "8,51 \200" "9.99\200" ...
##  $ premis                   : chr  "" "" "" "" ...
```

2.3 Tractament i conversions de dades quantitatives

Algunes dades numèriques es capturen amb les unitats, de forma que no són interpretades com a xifres. Per a veure si ens aporten alguna informació a simple vista, cal tractar-les abans del propi tractament de dades.

El preu s'ha llegit com un char, no com un double. A més, incorpora les unitats de euro com \200 i hi ha falta de discordància entre el punt i la coma.

```
# Eliminem les unitats de euro presents com a \200
vins$preu_euros <- gsub('\200', '', vins$preu_euros)

# Modifiquem les comes per punts
vins$preu_euros <- gsub(',', '.', vins$preu_euros)

# Convertim el tipus de variable de la columna de char a double
vins$preu_euros <- as.numeric(vins$preu_euros)
```

El volum s'ha llegit com un char, no com un integer degut a que incorpora les unitats de centilitres.

```
# Eliminem les unitats de centilitres de volum
vins$volum_cl <- gsub('cl', '', vins$volum_cl)

# Modifiquem les comes per punts
vins$volum_cl <- gsub(',', '.', vins$volum_cl)
```

```r
# Convertim el tipus de variable de la columna de char a numèric
vins$volum_cl <- as.numeric(vins$volum_cl)
```

La graduació alcohòlica també s'ha identificat com un char per la interpretació del símbol de %.

```r
# Eliminem el símbol de %
vins$graduacio_alcoholica_per <- gsub('%', '', vins$graduacio_alcoholica_per)

# Modifiquem les comes per punts
vins$graduacio_alcoholica_per <- gsub(',', '.', vins$graduacio_alcoholica_per)

# Convertim el tipus de variable de la columna de char a numèric
vins$graduacio_alcoholica_per <- as.numeric(vins$graduacio_alcoholica_per)
```

```
## Warning: NAs introduced by coercion
```

La acidesa també s'ha identificat com un char per la interpretació del símbol g/l tartàric.

```r
# Eliminem el símbol de %
vins$acidesa_g_l_tartaric <- gsub('g/L tartàric', '', vins$acidesa_g_l_tartaric)

# Modifiquem les comes per punts
vins$acidesa_g_l_tartaric <- gsub(',', '.', vins$acidesa_g_l_tartaric)

# Convertim el tipus de variable de la columna de char a numèric
vins$acidesa_g_l_tartaric <- as.numeric(vins$acidesa_g_l_tartaric)
```

```
## Warning: NAs introduced by coercion
```

El paràmetre de sulfits també s'ha identificat com un char per la interpretació del símbol de mg/l.

```r
# Eliminem el símbol de mg/l
vins$sulfuros_total_mg_l <- gsub('mg/l', '', vins$sulfuros_total_mg_l)

# Modifiquem les comes per punts
vins$sulfuros_total_mg_l <- gsub(',', '.', vins$sulfuros_total_mg_l)

# Convertim el tipus de variable de la columna de char a numèric
vins$sulfuros_total_mg_l <- as.numeric(vins$sulfuros_total_mg_l)
```

```
## Warning: NAs introduced by coercion
```

Finalment, també cal convertir els sucres residuals, passant-los de char a numeric.

```r
# Eliminem el símbol de g/l
vins$sucres_residuals_g_l <- gsub('g/l', '', vins$sucres_residuals_g_l)

# Modifiquem les comes per punts
vins$sucres_residuals_g_l <- gsub(',', '.', vins$sucres_residuals_g_l)

# Convertim el tipus de variable de la columna de char a numèric
vins$sucres_residuals_g_l <- as.numeric(vins$sucres_residuals_g_l)
```

```
## Warning: NAs introduced by coercion
```

```r
# Per tornar a mostrar les classes de les columnes
#str(vins)
```

```
# Mostar de forma més elegant els valors numèrics amb una estadística descriptiva bàsica
#summary(vins)
```

Observacions: - Podem veure que les conversions s'han realitzat correctament, ja que els camps presenten estadística descriptiva. - Resalta el nombre de camps buits, NA que es presenten. Acabaràn essent un presagi de que la correlació que obtinguem serà poc representativa.

2.4 Factorització D.O. i selecció de la D.O. Penedès

La manca de dades numèriques i l'alt volum de possibilitats de la pràctica fa que reduïm el scope a la D.O. Penedès. Les DO ens permeten granularitat i aïllament entre els diferents tipus de vins, també comparar-los si fes falta.

```
# Factoritzem les denominacions de origen dels vins inclosos a la base de dades
vins$do <- as.factor(vins$do)

# Mostrem el comptatge de la totalitat de vins inclosos en la base de dades
summary(vins$do)
```

```
##       D. O. Terra Alta             D.O Alella         D.O. Catalunya
##                     15                     30                     26
##  D.O. Conca de Barberà D.O. Costers del Segre           D.O. Empordà
##                      5                      5                      2
##          D.O. Montsant           D.O. Penedès      D.O. Pla de Bages
##                     19                    368                      1
##           D.O. Priorat         D.O. Tarragona
##                     13                      2
```

2.5 Selecció dels camps amb els que es treballarà: només D.O. Penedès

En un principi es volia realitzar una comparació entre vins, però el schedule no ha permès l'activitat, de forma que només ens em focalitzat en la D.O. Penedès.

```
vins <- vins[ which(vins$do=='D.O. Penedès'), ]

# Mostar de forma més elegant els valors numèrics amb una estadística descriptiva bàsica
#summary(vins)
```

Observacions: - Ens quedem en 368 vins de D.O. Penedès

3. Neteja de dades

En aquest apartat: - s'eliminen les columnes innecessàries (3.1) - s'eliminen files que no continguin els camps clau (3.2) - s'eliminen duplicats (3.3) - tractament de l'atribut "tipologia" (3.4) - tractament de l'atribut "envelliment" (3.5) - tractament de l'atribut "cupatge" (3.6) - tractament de l'atribut "celler" (3.7) - tractament de l'atribut "poblacio" (3.8) - tractament de l'atribut "tservei" (3.9) - extracció descriptiva visual, atributs "aspecte" i "color" (3.10) - extracció descriptiva olfactiva, atributs "aromes_1", "aromes_2" i "aromes_3" (3.11) - extracció descriptiva gustativa, atributs "gust" i "post_gust" (3.12) - tractament de l'atribut "maridatge" (3.13) - tractament de camps buits (3.14) - comprovació de valors extrems (3.15) - s'eliminen les columnes no treballades per falta de scope (3.16) - exportació del CLEAN dataset: vins_DO_Penedes.csv (3.17)

3.1 Eliminació de columnes innecessàries per contingut

Es considera que la columna de volum_cl no aporta informació, atès que les botelles presenten un format homogeni. D'aquest punt s'en treu la proposta de que els vinaters potser podrien innovar en format o volums.

```
# Es mostra com el valor mínim i màxim presenten el mateix valor.
summary(vins$volum_cl)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      75      75      75      75      75      75       4
```

```
# Eliminem la columna volum_cl
vins$volum_cl <- NULL

# Eliminem la columna estil de vi
vins$estil <- NULL
```

3.2 Eliminació de files que no tinguin el nom del vi (clau de traçabilitat)

Les columnes que considerem d'estructura no poden contenir files amb valors NA, de forma que són eliminades.

```
# Eliminem les files que no presentin noms
vins <- filter(vins, nom != "")

# Eliminem les files que no presentin la tipologia ja que és l'element clau per a trobar relacions
vins <- filter(vins, tipologia != "")
```

3.3 Eliminació de duplicats

El cas de tenir vins duplicats s'eliminen per evitar que tinguin una major presència.

```
# Eliminem les files duplicades
vins <- vins[!duplicated(vins), ]
```

3.4 Conversió de la tipologia del vi

En tipologia de vi es desitja categoritzar el vi segons si és negre, blanc o rosat. Les varietats de cava, escumosos i dolços no es consideraràn. Com que el camp de tipologia del vi incorpora altres elements descriptius, s'ha considerat crear una nova columna per identificar vins ecològics i els joves que afectin els camp d'estil.

```
# Mostra dels noms com exemple
#str(vins$tipologia)

# Generem una columna categòrica ecològica
vins$ecologic <- ifelse(grepl("eco", vins$tipologia), "Si", "No")

# Treballem amb l'envelliment del vi
vins$envelliment <- ifelse(grepl("jove", vins$tipologia), "vi jove", vins$envelliment)
```

Treballem la neteja de la tipologia del vi. En aquests cas, netegem el les descripcions per a quedar-nos només amb els tres grups amb els que es vol treballar.

```
#vins$TIPOLOGIA <- ifelse(vins$TIPOLOGIA %in% c('Vi blanc', 'blanc'), 'blanc', vins$TIPOLOGIA)

# Converteix les descripcions en blanc, negre i rosat
vins$tipologia <- ifelse(grepl('blanc', vins$tipologia, ignore.case = T), 'blanc', vins$tipologia)
vins$tipologia <- ifelse(grepl('negre', vins$tipologia, ignore.case = T), 'negre', vins$tipologia)
vins$tipologia <- ifelse(grepl('rosat', vins$tipologia, ignore.case = T), 'rosat', vins$tipologia)
```

```
# Elimina vi dolços i escumosos ja que surten del scope
#vins<-vins[- grep("Cava", vins$tipologia, ignore.case = T),]
#vins<-vins[- grep("dolç", vins$tipologia, ignore.case = T),]

# Correccions MANUALS detectades de noms mal escrits i arreglats per inserció
#vins$tipologia <- ifelse(grepl('nregre', vins$tipologia, ignore.case = T), 'negre', vins$tipologia)

# Mostrar taula final
#vins$tipologia
```

El tractament de valors nuls pel tipus de vi Analitzem els valors nuls del tipus de vi

```
# Identifiquem i analitzem els valors nuls en topologia
vins[which(vins$tipologia==""), ]
```

```
##  [1] do                       nom                     tipologia
##  [4] envelliment              anyada                  celler
##  [7] poblacio                 vinya                   elaboracio
## [10] cupatge                  vista                   aroma
## [13] paladar                  maridatge               tservei
## [16] graduacio_alcoholica_per acidesa_g_l_tartaric    sulfuros_total_mg_l
## [19] sucres_residuals_g_l     ph                      preu_euros
## [22] premis                   ecologic
## <0 rows> (or 0-length row.names)
```

```
# Converteix les descripcions en blanc, negre i rosat
vins$tipologia <- ifelse(grepl('blanc', vins$nom, ignore.case = T), 'blanc', vins$tipologia)
vins$tipologia <- ifelse(grepl('negre', vins$nom, ignore.case = T), 'negre', vins$tipologia)
vins$tipologia <- ifelse(grepl('rosat', vins$nom, ignore.case = T), 'rosat', vins$tipologia)

# Elimina els que no s'hagin pogut identificar
vins <- filter(vins, tipologia != "")

# Imputació de valors mitjançant la función kNN() del paquet VIM
vins$tipologia <- kNN(vins)$tipologia
```

Pasem a visualitzar-se els rangs per veure si realment tenim valors coherents

```
# Factoritzem les denominacions de origen dels vins inclosos a la base de dades
vins$tipologia <- as.factor(vins$tipologia)

# Mostrem el comptatge de la totalitat de vins inclosos en la base de dades
summary(vins$tipologia)
```

```
## blanc negre rosat
##   132   118    30
```

3.5 Conversió de l'envelliment/maduració del vi

A continuació treballem l'envelliment/maduració segons la Llei 24/2003, de 10 de juliol, de la Vinya i del Vi, que estableix els períodes mínims d'envelliment per a cada classificació. - Vi jove: no pasen per bóta (elaboració) + anyada actual - Vi amb semi criança: pasen per bóta però amb menys temps que un criança - Criança: 2 anys/24 mesos per vi negre - 1.5 anys/18 mesos per vi blanc - Reserva: 3 anys/36 mesos per vi negre - 2 anys/24 mesos per vi blanc - Gran Reserva: 5 anys/60 mesos per vi negre - 3 anys/36 mesos per vi blanc

```r
# Definim l'any actual
any_actual<-2021

# Definicions de Vi jove
vins$envelliment <- ifelse(grepl('Sense criança|inoxidable', vins$envelliment, ignore.case = T), 'Vi jo
vins$envelliment <- ifelse(vins$anyada > (any_actual-1) & (grepl('bóta', vins$envelliment, ignore.case =

# Definicions de Criança
vins$envelliment <- ifelse((any_actual-2) >= vins$anyada & vins$anyada < (any_actual-3) & vins$tipologia
vins$envelliment <- ifelse((any_actual-1) >= vins$anyada & vins$anyada < (any_actual-2) & vins$tipologia

# Definicions de Reserva
vins$envelliment <- ifelse((any_actual-3) >= vins$anyada & vins$anyada < (any_actual-5) & vins$tipologia
vins$envelliment <- ifelse((any_actual-2) >= vins$anyada & vins$anyada < (any_actual-3) & vins$tipologia

# Definicions de Gran Reserva
vins$envelliment <- ifelse(vins$anyada <= (any_actual-5) & vins$tipologia == 'negre', 'Gran Reserva', vi
vins$envelliment <- ifelse(vins$anyada <= (any_actual-3) & vins$tipologia == 'blanc', 'Gran Reserva', vi

vins$envelliment <- ifelse(grepl('bot', vins$elaboracio, ignore.case = T), 'Criança', vins$envelliment)
vins$envelliment <- ifelse(grepl('bót', vins$elaboracio, ignore.case = T), 'Criança', vins$envelliment)

# Imputació de valors mitjançant la función kNN() del paquet VIM
vins$envelliment <- kNN(vins)$envelliment

# Mostrar taula final
#vins$envelliment

# Factoritzem les 5 categories existents
vins$envelliment <- as.factor(vins$envelliment)

# Mostrem la repartició dels vins segons les categories d'envelliment del vi
summary(vins$envelliment)
```

```
##      Criança Gran Reserva      Vi jove
##          163           88           29
```

3.6 Conversió dels coupage

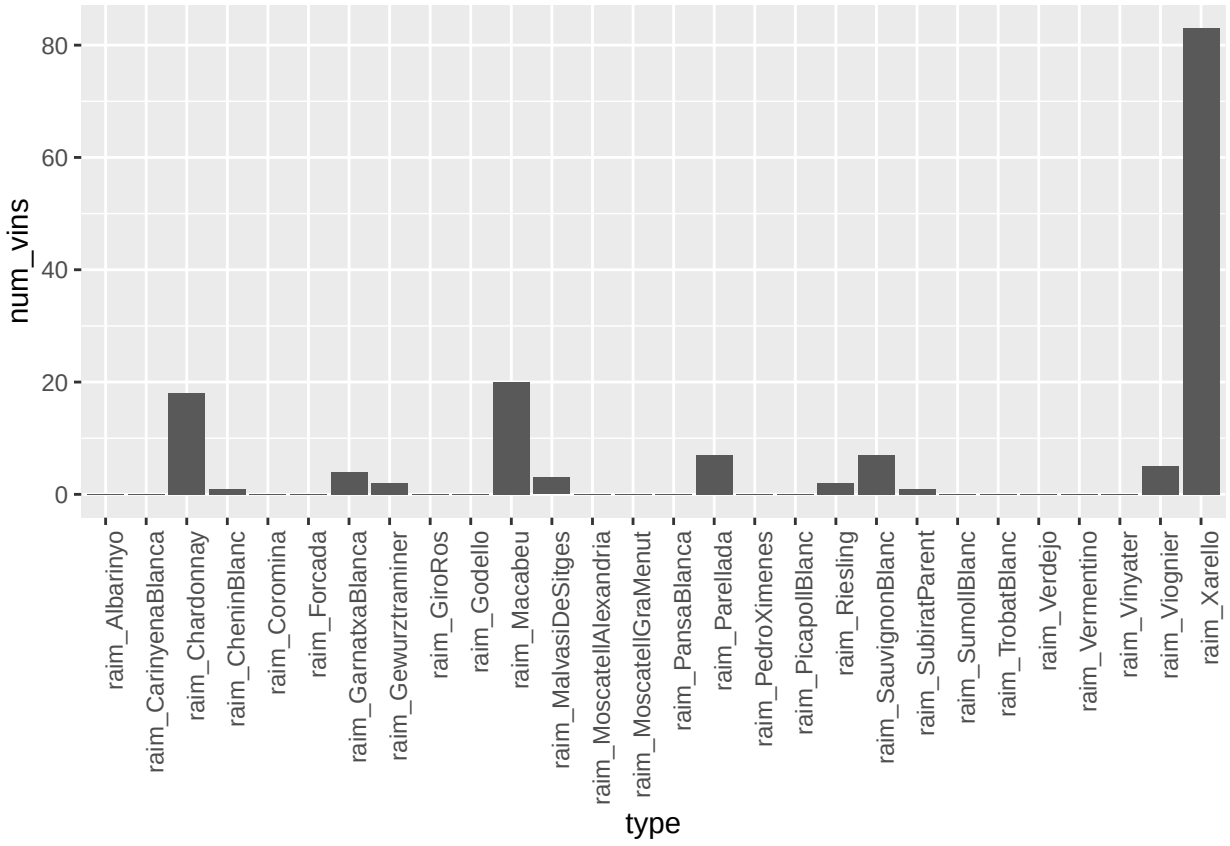
La mescla de les varietats de raïm Com que tornem a partir de descripcions, ens cal primer de tot detectar les paraules, després estandaritzar la estructura i finalment discretitzar els valors.

```r
raim_blanc_varietat <- c("Airén", "Albarinyo", "Carinyena blanca", "Chardonnay", "Chenin blanc", "Corom

raim_negre_varietat <- c("Boval", "Cabernet franc", "Cabernet sauvignon", "Garnatxa peluda", "Garnatxa

vins$raim_Airen<- ifelse(grepl('Airén', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_PansaBlanca<- ifelse(grepl('Pansa Blanca', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Albarinyo<- ifelse(grepl('Albarinyo', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_CarinyenaBlanca<- ifelse(grepl('Carinyena blanca', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Chardonnay<- ifelse(grepl('Chardonnay', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_CheninBlanc<- ifelse(grepl('Chenin blanc', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Coromina<- ifelse(grepl('Coromina', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Forcada<- ifelse(grepl('Forcada', vins$cupatge, ignore.case = T), 1, 0)
```

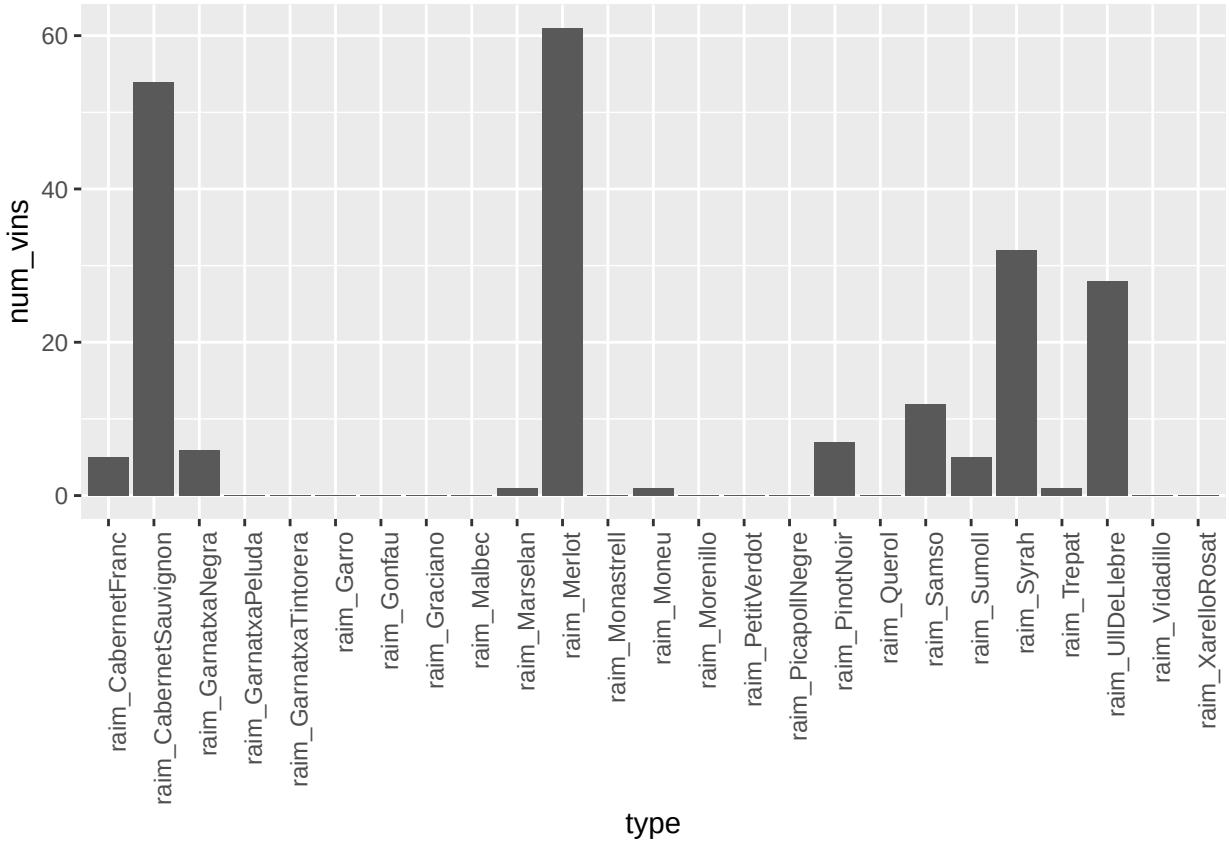
```
vins$raim_GarnatxaBlanca<- ifelse(grepl('Garnatxa blanca', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Gewurztraminer<- ifelse(grepl('Gewürztraminer', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_GiroRos<- ifelse(grepl('Giró ros', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Godello<- ifelse(grepl('Godello', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Macabeu<- ifelse(grepl('Macabeu', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_MalvasiDeSitges<- ifelse(grepl('Malvasia de Sitges', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_MoscatellAlexandria<- ifelse(grepl("Moscatell d'Alexandria", vins$cupatge, ignore.case = T),
vins$raim_MoscatellGraMenut<- ifelse(grepl('Moscatell de gra menut', vins$cupatge, ignore.case = T), 1,
vins$raim_Parellada<- ifelse(grepl('Parellada', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_PedroXimenes<- ifelse(grepl('Pedro ximenes', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_PicapollBlanc<- ifelse(grepl('Picapoll blanc', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Riesling<- ifelse(grepl('Riesling', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_SauvignonBlanc<- ifelse(grepl('Sauvignon blanc', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_SubiratParent<- ifelse(grepl('Subirat parent', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_SumollBlanc<- ifelse(grepl('Sumoll blanc', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_TrobatBlanc<- ifelse(grepl('Trobat blanc',vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Verdejo<- ifelse(grepl('Verdejo', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Vermentino<- ifelse(grepl('Vermentino', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Vinyater<- ifelse(grepl('Vinyater', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Viognier<- ifelse(grepl('Viognier', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Xarello<- ifelse(grepl('Xarel·lo', vins$cupatge, ignore.case = T), 1, 0)

vins$raim_Boval<- ifelse(grepl('Boval', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_CabernetFranc<- ifelse(grepl('Cabernet franc', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_CabernetSauvignon<- ifelse(grepl('Cabernet sauvignon', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_GarnatxaPeluda<- ifelse(grepl('Garnatxa peluda', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_GarnatxaNegra<- ifelse(grepl('Garnatxa negra', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_GarnatxaTintorera<- ifelse(grepl('Garnatxa tintorera', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Garro<- ifelse(grepl('Garró', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Gonfau<- ifelse(grepl('Gonfaus', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Graciano<- ifelse(grepl('Graciano', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Malbec<- ifelse(grepl('Malbec', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Marselan<- ifelse(grepl('Marselan', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Samso<- ifelse(grepl('Samsó', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Samso<- ifelse(grepl('carinyena', vins$cupatge, ignore.case = T), 1, vins$raim_Samso)
vins$raim_Merlot<- ifelse(grepl('Merlot', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Monastrell<- ifelse(grepl('Monastrell', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Moneu<- ifelse(grepl('Moneu', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Morenillo<- ifelse(grepl('Morenillo', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_PetitVerdot<- ifelse(grepl('Petit verdot', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_PicapollNegre<- ifelse(grepl('Picapoll negre', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_PinotNoir<- ifelse(grepl('Pinot noir', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Querol<- ifelse(grepl('Querol', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Solana<- ifelse(grepl('Solana', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Sumoll<- ifelse(grepl('Sumoll', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Syrah<- ifelse(grepl('Syrah', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_UllDeLlebre<- ifelse(grepl('Ull de llebre', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Trepat<- ifelse(grepl('Trepat', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_Vidadillo<- ifelse(grepl('Vidadillo', vins$cupatge, ignore.case = T), 1, 0)
vins$raim_XarelloRosat<- ifelse(grepl('Xarel·lo rosat', vins$cupatge, ignore.case = T), 1, 0)
```

```
# plotem el subset per veure quins són els aspectes que més potencien
raim <- select(vins, raim_Airen, raim_PansaBlanca, raim_Albarinyo, raim_CarinyenaBlanca, raim_Chardonna
```

```r
# Barplot de les varietats de raïm blanc
n_yes2 <- data.frame(type = names(raim[, -1]),
                     num_vins = colSums(raim[, -1] == 1))

ggplot(n_yes2, aes(x = type, y = num_vins)) +
  geom_bar(stat = "identity") + theme(axis.text.x=element_text(angle=90, hjust=1)) + scale_fill_manual(
```




```r
# plotem el subset per veure quins són els aspectes que més potencien
raim <- select(vins, raim_Boval, raim_CabernetFranc, raim_CabernetSauvignon, raim_GarnatxaPeluda, raim_

# Barplot de les varietats de raïm negre
n_yes2 <- data.frame(type = names(raim[, -1]),
                     num_vins = colSums(raim[, -1] == 1))

ggplot(n_yes2, aes(x = type, y = num_vins)) +
  geom_bar(stat = "identity") + theme(axis.text.x=element_text(angle=90, hjust=1)) + scale_fill_manual(
```



En les gràfiques es mostren les varietats de raïm i el percentatge d'explotació a través dels cellers seguits.

###3.6.1 Control de varietats de raïm de la DO del Penedès

```
raim_DO_Penedes_21 <- c("Chardonnay","Gewürztraminer","Macabeu","Malvasia de Sitges","Moscatell d'Alexa

# Calcular el nombre total de varietats de raïm del Penedès
n_varietats_Penedes<-length(raim_DO_Penedes_21)

# Relació de les varietats respecte el total de varietats catalogades de Catalunya
n_varietats_raim_negre<-length(raim_negre_varietat)
n_varietats_raim_blanc<-length(raim_blanc_varietat)
n_varietats_total<-n_varietats_raim_negre+n_varietats_raim_blanc
percentatge_varietats<-(n_varietats_Penedes/(n_varietats_raim_negre+n_varietats_raim_blanc))*100
sprintf("La demominació d'origen Penedès accepta un total de %i varietats de raïm, el que suposa un %#.
```

```
## [1] "La demominació d'origen Penedès accepta un total de 16 varietats de raïm, el que suposa un 29.0
```

```
# Creem una tira binària estandarditzada per a codificar els coupage
vins$vector_raim <- paste(vins$raim_Airen, vins$raim_PansaBlanca, vins$raim_Albarinyo, vins$raim_Carinyo

# Factoritzem
#vins$facto_raim <- as.numeric(factor(vins$vector_raim))
vins$codi_coupage <- factor(vins$vector_raim)

# Eliminem columnes afegides per a trobar les composicions
vins$vector_raim <-NULL
vins$raim_Airen<- NULL
```

```
vins$raim_PansaBlanca<- NULL
vins$raim_Albarinyo<- NULL
vins$raim_CarinyenaBlanca<- NULL
vins$raim_Chardonnay<- NULL
vins$raim_CheninBlanc<- NULL
vins$raim_Coromina<- NULL
vins$raim_Forcada<- NULL
vins$raim_GarnatxaBlanca<- NULL
vins$raim_Gewurztraminer<- NULL
vins$raim_GiroRos<- NULL
vins$raim_Godello<- NULL
vins$raim_Macabeu<- NULL
vins$raim_MalvasiDeSitges<- NULL
vins$raim_MoscatellAlexandria<- NULL
vins$raim_MoscatellGraMenut<- NULL
vins$raim_Parellada<- NULL
vins$raim_PedroXimenes<- NULL
vins$raim_PicapollBlanc<- NULL
vins$raim_Riesling<- NULL
vins$raim_SauvignonBlanc<- NULL
vins$raim_SubiratParent<- NULL
vins$raim_SumollBlanc<- NULL
vins$raim_TrobatBlanc<- NULL
vins$raim_Verdejo<- NULL
vins$raim_Vermentino<- NULL
vins$raim_Vinyater<- NULL
vins$raim_Viognier<- NULL
vins$raim_Xarello<- NULL

vins$raim_Boval<- NULL
vins$raim_CabernetFranc<- NULL
vins$raim_CabernetSauvignon<- NULL
vins$raim_GarnatxaPeluda<- NULL
vins$raim_GarnatxaNegra<- NULL
vins$raim_GarnatxaTintorera<- NULL
vins$raim_Garro<- NULL
vins$raim_Gonfau<- NULL
vins$raim_Graciano<- NULL
vins$raim_Malbec<- NULL
vins$raim_Marselan<- NULL
vins$raim_Samso<- NULL
vins$raim_Merlot<- NULL
vins$raim_Monastrell<- NULL
vins$raim_Moneu<- NULL
vins$raim_Morenillo<- NULL
vins$raim_PetitVerdot<- NULL
vins$raim_PicapollNegre<- NULL
vins$raim_PinotNoir<- NULL
vins$raim_Querol<- NULL
vins$raim_Solana<- NULL
vins$raim_Sumoll<- NULL
vins$raim_Syrah<- NULL
vins$raim_UllDeLlebre<- NULL
```

```
vins$raim_Trepat<- NULL
vins$raim_Vidadillo<- NULL
vins$raim_XarelloRosat<- NULL

summary(vins)
```

```
##                       do          nom            tipologia         envelliment
##  D.O. Penedès          :280   Length:280         blanc:132   Criança     :163
##  D. O. Terra Alta      :  0   Class :character   negre:118   Gran Reserva: 88
##  D.O Alella            :  0   Mode  :character   rosat: 30   Vi jove     : 29
##  D.O. Catalunya        :  0
##  D.O. Conca de Barberà :  0
##  D.O. Costers del Segre:  0
##  (Other)               :  0
##      anyada        celler            poblacio            vinya
##  Min.   :2003   Length:280         Length:280         Length:280
##  1st Qu.:2017   Class :character   Class :character   Class :character
##  Median :2019   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2018
##  3rd Qu.:2020
##  Max.   :2021
##  NA's   :34
##   elaboracio          cupatge             vista              aroma
##  Length:280         Length:280         Length:280         Length:280
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    paladar           maridatge           tservei
##  Length:280         Length:280         Length:280
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  graduacio_alcoholica_per acidesa_g_l_tartaric sulfuros_total_mg_l
##  Min.   :10.50            Min.   :2.90         Min.   : 10.00
##  1st Qu.:12.50            1st Qu.:4.20         1st Qu.: 38.00
##  Median :13.00            Median :5.37         Median : 60.00
##  Mean   :12.92            Mean   :5.22         Mean   : 61.38
##  3rd Qu.:13.50            3rd Qu.:6.05         3rd Qu.: 88.25
##  Max.   :14.70            Max.   :8.21         Max.   :123.00
##  NA's   :65               NA's   :169          NA's   :240
##  sucres_residuals_g_l       ph          preu_euros        premis
##  Min.   : 0.200       Min.   :2.900   Min.   :  3.33   Length:280
##  1st Qu.: 0.500       1st Qu.:3.150   1st Qu.:  7.80   Class :character
##  Median : 1.150       Median :3.250   Median :  9.98   Mode  :character
##  Mean   : 1.612       Mean   :3.277   Mean   : 14.60
##  3rd Qu.: 2.375       3rd Qu.:3.420   3rd Qu.: 14.65
##  Max.   :10.300       Max.   :3.690   Max.   :250.00
##  NA's   :210          NA's   :195     NA's   :83
```

```
##    ecologic
##  Length:280
##  Class :character
##  Mode  :character
##
##
##
##
##                                                                codi_coupage
##  00000000000000000000000000001000000000000000000000000000: 62
##  00000000000000000000000000000000000000000000000000000000: 36
##  00000000000000000000000000000000000000001000000000000000: 23
##  00000000000000000000000000001000000000100000000000000000: 15
##  00000000000000000000000000001000000000000000000000000000: 14
##  00000000000000000000000000000000000000000000000000010000: 10
##  (Other)                                                   :120
```

Existeixen un total de 41 cupatges dels quals el més emprat amb diferència és el 100% xarel · lo.

```
# Mostra tots els cupatges
#sort(summary(vins$codi_coupage), decreasing = TRUE, na.last = TRUE)


# Per a identificar un cupatge, per exemple, el més convinat
#ifelse(grepl('00000000000000000000000000001000000000000000000000000000', vins$codi_coupage, ignore.ca
#sprintf("La demominació d'origen Penedès accepta un total de %i varietats de raïm, el que suposa un %#
#sprintf("El cupatge més emprat és %s .", n_varietats_Penedes, percentatge_varietats)
```

3.7 Conversió de cellers D.O.

Els cellers productors dels vins D.O. es poden considerar com a categories, ja que cadascuna d'elles oferta una gran varietat de productes

```
# Factoritzem els cellers, el que ens permetrà englobar diferents productes per un mateix productor
vins$celler <- as.factor(vins$celler)

# Mostrem la repartició dels vins segons les categories d'envelliment del vi
#summary(vins$celler)
```

3.8 Tractament de l'adreça del celler a la població

Per ajudar a definir el terroir en funció de la població

```
# Correcció manual d'accents
vins$poblacio <- ifelse(grepl('Vilobi del Penedès', vins$poblacio, ignore.case = T), 'Vilobí del Penedès


# Definició localitats del Penedès
localitats_penedes <- c("Avionyonet del Penedès","Les Cabanyes","Castellet i la Gornal","Castellví de la
#str(vins$poblacio)

# Creem la iteració per substituir la direcció per la població
for (localitat in localitats_penedes) {
  vins$poblacio<-ifelse(grepl(localitat, vins$poblacio, ignore.case = T),localitat,vins$poblacio)
}
```

```r
# Factoritzem els cellers, el que ens permetrà englobar diferents productes per un mateix productor
vins$poblacio <- as.factor(vins$poblacio)

# Mostrem la repartició dels vins segons les categories d'envelliment del vi
#summary(vins$poblacio)
```

3.9 Tractament de la temperatura de servei

La temperatura de servei va intimament lligada al tipus de vi

```r
tipus_vi_pertemperaturaservei <- list("14-16ºC" = "Negres estructurat", "14ºC" = "Negra amb poca criança

# Filtre de temperatures
vins$tservei<-paste(sapply(str_extract_all(vins$tservei, "\\d+"), "[", 1),sapply(str_extract_all(vins$ts

# Tractament de casos NA-NA en funció del tipus de vi negre
vins$tservei<-ifelse(grepl('NA-NA', vins$tservei, ignore.case = T) & vins$tipologia=='negre', '14-16', 
# Tractament de casos NA-NA en funció del tipus de vi blanc
vins$tservei<-ifelse(grepl('NA-NA', vins$tservei, ignore.case = T) & vins$tipologia=='blanc|rosat', '6-8


# Tractament de factorització
vins$tservei <- as.factor(vins$tservei)

# Mostra el comptatge
#summary(vins$tservei)
```

3.10 Fase visual - Extracció descriptiva d'aspecte i color

```r
## Fase visual
# rànquing aspecte
ranq_aspecte<-list("net","brillant","cristalí","transparent","brut","apagat","tacat","tèrbol","velat")
# rànquing color
ranq_color<-list("Groc pàl·lid", "verdós", "palla", "daurat", "ambarí", "Rosat pàl·lid", "gerds", "madu

# Converteix les descripcions de fase visual de l'aspecte en característiques descriptives
vins$aspecte <- ifelse(grepl("net|nítid", vins$vista, ignore.case = T), 'net', NA)
vins$aspecte <- ifelse(grepl('brillant', vins$vista, ignore.case = T), 'brillant', vins$aspecte)
vins$aspecte <- ifelse(grepl('cristalí', vins$vista, ignore.case = T), 'cristalí', vins$aspecte)
vins$aspecte <- ifelse(grepl('transparent', vins$vista, ignore.case = T), 'transparent', vins$aspecte)
vins$aspecte <- ifelse(grepl('brut', vins$vista, ignore.case = T), 'brut', vins$aspecte)
vins$aspecte <- ifelse(grepl('apagat', vins$vista, ignore.case = T), 'apagat', vins$aspecte)
vins$aspecte <- ifelse(grepl('tacat', vins$vista, ignore.case = T), 'tacat', vins$aspecte)
vins$aspecte <- ifelse(grepl('tèrbol', vins$vista, ignore.case = T), 'tèrbol', vins$aspecte)
vins$aspecte <- ifelse(grepl('velat', vins$vista, ignore.case = T), 'velat', vins$aspecte)

# Factoritzem l'aspecte
vins$aspecte <- as.factor(vins$aspecte)

# Converteix les descripcions de fase visual de color en característiques descriptives
vins$color <- ifelse(grepl('Groc pàl·lid', vins$vista, ignore.case = T), 'groc pàl·lid', NA)
vins$color <- ifelse(grepl('verdós', vins$vista, ignore.case = T), 'verdós', vins$color)
vins$color <- ifelse(grepl('palla', vins$vista, ignore.case = T), 'palla', vins$color)
```

```
vins$color <- ifelse(grepl('daurat', vins$vista, ignore.case = T), 'daurat', vins$color)
vins$color <- ifelse(grepl('ambarí', vins$vista, ignore.case = T), 'ambarí', vins$color)
vins$color <- ifelse(grepl('rosat', vins$vista, ignore.case = T), 'rosat', vins$color)
vins$color <- ifelse(grepl('gerds', vins$vista, ignore.case = T), 'gerds', vins$color)
vins$color <- ifelse(grepl('maduixa', vins$vista, ignore.case = T), 'maduixa', vins$color)
vins$color <- ifelse(grepl('ceba', vins$vista, ignore.case = T), 'pell de ceba', vins$color)
vins$color <- ifelse(grepl('ataronjat', vins$vista, ignore.case = T), 'rosa ataronjat', vins$color)
vins$color <- ifelse(grepl('porpra', vins$vista, ignore.case = T), 'porpra', vins$color)
vins$color <- ifelse(grepl('violeta|violaci', vins$vista, ignore.case = T), 'violeta', vins$color)
vins$color <- ifelse(grepl('robí', vins$vista, ignore.case = T), 'robí', vins$color)
vins$color <- ifelse(grepl('cirera|picota', vins$vista, ignore.case = T), 'cirera', vins$color)
vins$color <- ifelse(grepl('teula', vins$vista, ignore.case = T), 'teula', vins$color)
vins$color <- ifelse(grepl('marró|cafè', vins$vista, ignore.case = T), 'marró', vins$color)

# Trobem el color dels vins dels uqe no s'ha identificat posant especial èmfasis el tipus de vi, ja que
vins.negres <- vins[vins$tipologia == "negre",]
vins.blancs <- vins[vins$tipologia == "blanc",]
vins.rosats <- vins[vins$tipologia == "rosat",]
vins$color <- ifelse(vins$tipologia == "negre", kNN(vins.negres)$color, vins$color)
vins$color <- ifelse(vins$tipologia == "blanc", kNN(vins.blancs)$color, vins$color)
vins$color <- ifelse(vins$tipologia == "rosat", kNN(vins.rosats)$color, vins$color)

# Factoritzem els colors
vins$color <- as.factor(vins$color)

# Mostrem la repartició dels vins segons les categories d'envelliment del vi
summary(vins$aspecte)
```

```
##      apagat    brillant         net      tèrbol transparent        NA's
##           1          76           5           2           3         193
```

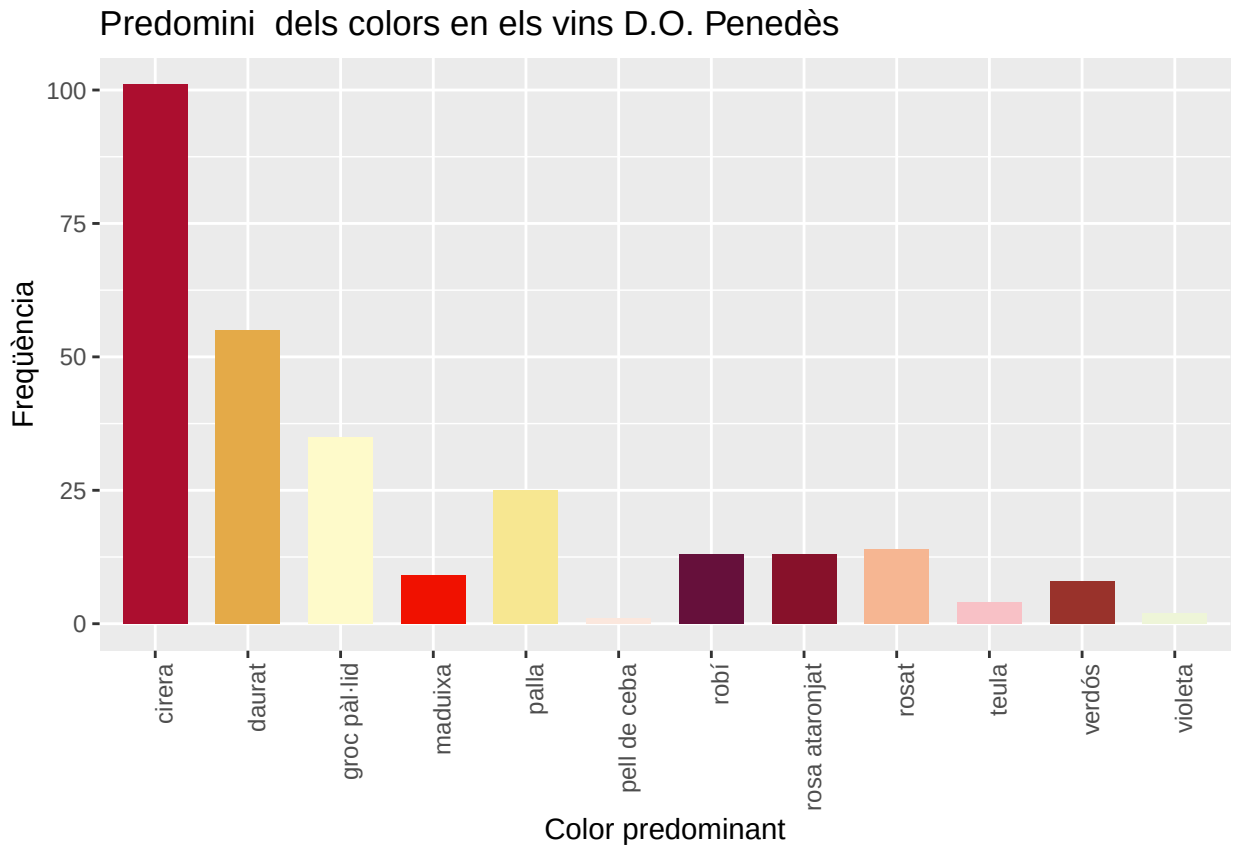
```
summary(vins$color)
```

```
##         cirera         daurat   groc pàl·lid        maduixa          palla
##            101             55             35              9             25
##   pell de ceba           robí rosa ataronjat          rosat          teula
##              1             13             13             14              4
##         verdós        violeta
##              8              2
```

```
# Eliminemla columna vins$vista
vins$vista <- NULL
```

Plotem la distribució de colors del vi pròpis

```
vins_palette_name <- c("color_groc_palid"="#FEFACA", "color_verdos"="#EEF5D8", "color_palla"="#F7E791",

vins_palette <- c("#AC0E2F","#E4AA48","#FEFACA","#F01100","#F7E791","#FBE7DD","#66103B", "#87112A","#F6

ggplot(vins, aes(x=color, fill=color)) + geom_bar(stat="count", width=0.7)+ scale_fill_manual(values=vi
```



Com es pot obervar en el gràfic, la característica de color dels vins DO és de color cirera picota pels vins negres i color daurat pels vins blancs. Seguit pel color robí en els negres i el color palla en els blancs. Tot el que surti d'aquesta gamma de colors es pot considerar exòtic i rellevant.

3.11 Fase olfactiva - Extracció descriptiva d'aromes primàries, secundàries i terciàries

```
## Fase olfactiva
# rànquing intensistat
ranq_intensitat<-list("baixa","suficient","mitja","alta")

vins$intensitat <- ifelse(grepl('baixa', vins$aroma, ignore.case = T), "baixa", NA)
vins$intensitat <- ifelse(grepl('suficient', vins$aroma, ignore.case = T), "suficient", vins$intensitat)
vins$intensitat <- ifelse(grepl('mitja', vins$aroma, ignore.case = T), "mitja", vins$intensitat)
vins$intensitat <- ifelse(grepl('alta', vins$aroma, ignore.case = T), "alta", vins$intensitat)

# Factoritzem la intensitat
vins$intensitat <- as.factor(vins$intensitat)

# aromes primàries (sense agitar la copa)
vins$aromes_1<- ifelse(grepl('cítrics|pera|meló|grosella|maduixa', vins$aroma, ignore.case = T), 'afruit
vins$aromes_1<- ifelse(grepl('romaní|eucalipturs|menta', vins$aroma, ignore.case = T), "vegetal", vins$a
vins$aromes_1<- ifelse(grepl('rosa|espígol|violeta', vins$aroma, ignore.case = T), "floral", vins$aromes
vins$aromes_1<- ifelse(grepl('pissarra|querosè|granit', vins$aroma, ignore.case = T), "mineral", vins$ar
vins$aromes_1<- ifelse(grepl('nou moscada|regalèssia|pebre|vainilla', vins$aroma, ignore.case = T), "esp
```

```
# aromes secundàries (voltes suaus)
vins$aromes_2<- ifelse(grepl('nous', vins$aroma, ignore.case = T), "fruits secs", NA)
vins$aromes_2<- ifelse(grepl('iogurt|mantega', vins$aroma, ignore.case = T), "làctics", 0)
vins$aromes_2<- ifelse(grepl('galetes|caramel', vins$aroma, ignore.case = T), "dolces", 0)

# aromes terciàries ()
vins$aromes_3<- ifelse(grepl('cafè|xocolata', vins$aroma, ignore.case = T), "aroma de bota", "sense aro

# Factoritzem les aromes primàries, secundàries i terciàries
vins$aromes_1 <- as.factor(vins$aromes_1)
vins$aromes_2 <- as.factor(vins$aromes_2)
vins$aromes_3 <- as.factor(vins$aromes_3)

# Mostrem la repartició dels vins segons les aromes del vi
summary(vins$intensitat)
```

```
##  alta baixa mitja  NA's
##    10     1     6   263
```

```
summary(vins$aromes_1)
```

```
## afruitat espècies   floral  vegetal     NA's
##       39       33       12        6      190
```

```
summary(vins$aromes_2)
```

```
##      0 dolces
##    278      2
```

```
summary(vins$aromes_3)
```

```
##       aroma de bota sense aroma de bota
##                  12                 268
```

```
# Eliminemla columna vins$aroma
vins$aroma <- NULL
```

3.12 Fase gustativa - Extracció descriptiva de gust

```
## Fase gustativa
# rànquing gust
vins$gust<- ifelse(grepl('Fresc', vins$paladar, ignore.case = T), "Fresc", NA)
vins$gust<- ifelse(grepl('Afruitat', vins$paladar, ignore.case = T), "Afruitat", vins$gust)
vins$gust<- ifelse(grepl('Àcid', vins$paladar, ignore.case = T), "Àcid", vins$gust)
vins$gust<- ifelse(grepl('Lleuger', vins$paladar, ignore.case = T), "Lleuger", vins$gust)
vins$gust<- ifelse(grepl('Calent', vins$paladar, ignore.case = T), "Calent", vins$gust)
vins$gust<- ifelse(grepl('Amarg', vins$paladar, ignore.case = T), "Amarg", vins$gust)
vins$gust<- ifelse(grepl('Saborós', vins$paladar, ignore.case = T), "Saborós", vins$gust)
vins$gust<- ifelse(grepl('Carnós', vins$paladar, ignore.case = T), "Carnós", vins$gust)
vins$gust<- ifelse(grepl('Amable', vins$paladar, ignore.case = T), "Amable", vins$gust)
vins$gust<- ifelse(grepl('Equilibrat', vins$paladar, ignore.case = T), "Equilibrat", vins$gust)
vins$gust<- ifelse(grepl('Franc', vins$paladar, ignore.case = T), "Franc", vins$gust)
vins$gust<- ifelse(grepl('Aspre', vins$paladar, ignore.case = T), "Aspre", vins$gust)
vins$gust<- ifelse(grepl('Rodó', vins$paladar, ignore.case = T), "Rodó", vins$gust)
vins$gust<- ifelse(grepl('amb cos', vins$paladar, ignore.case = T), "amb cos", vins$gust)
vins$gust<- ifelse(grepl('pla', vins$paladar, ignore.case = T), "pla", vins$gust)
```

```r
vins$gust<- ifelse(grepl('fragant', vins$paladar, ignore.case = T), "Fragant", vins$gust)
vins$gust<- ifelse(grepl('suau', vins$paladar, ignore.case = T), "Suau", vins$gust)
vins$gust<- ifelse(grepl('amb personalitat', vins$paladar, ignore.case = T), "amb personalitat", vins$gu
vins$gust<- ifelse(grepl('càlid', vins$paladar, ignore.case = T), "càlid", vins$gust)
vins$gust<- ifelse(grepl('astringent', vins$paladar, ignore.case = T), "astringent", vins$gust)
vins$gust<- ifelse(grepl('sedós', vins$paladar, ignore.case = T), "sedós", vins$gust)
vins$gust<- ifelse(grepl('pastós', vins$paladar, ignore.case = T), "pastós", vins$gust)
vins$gust<- ifelse(grepl('ardent', vins$paladar, ignore.case = T), "ardent", vins$gust)

# rànquing post gust
post_gust_per_defecte<-"curt"
vins$post_gust<- ifelse(grepl('mig', vins$paladar, ignore.case = T), "mig", post_gust_per_defecte)
vins$post_gust<- ifelse(grepl('molt curt', vins$paladar, ignore.case = T), "molt curt", vins$post_gust)
vins$post_gust<- ifelse(grepl('llarg', vins$paladar, ignore.case = T), "llarg", vins$post_gust)
vins$post_gust<- ifelse(grepl('molt llarg', vins$paladar, ignore.case = T), "molt llarg", vins$post_gus

# Factoritzemel gust i postgust
vins$gust <- as.factor(vins$gust)
vins$post_gust <- as.factor(vins$post_gust)


# Mostrem la repartició dels vins segons les categories d'envelliment del vi
summary(vins$gust)
```

```
##             Àcid         Afruitat           Amable            Amarg
##                1               12                7                7
##          amb cos amb personalitat       astringent           càlid
##                6                2                1                6
##           Carnós       Equilibrat          Fragant           Franc
##                9               28                1                2
##            Fresc          Lleuger             pla            Rodó
##               43               11                2               10
##          Saborós            sedós            Suau            NA's
##                5               15               32               80
```

```r
summary(vins$post_gust)
```

```
##       curt      llarg        mig molt llarg
##        206         71          1          2
```

```r
# Eliminemla columna vins$paladar
vins$paladar <- NULL
```

3.13 Generació del codi de maridatge

```r
## Maridatge

# Converteix les descripcions de fase visual de l'aspecte en característiques descriptives
vins$maridatge_arrossos <- ifelse(grepl('arrossos|paelles', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_pastes <- ifelse(grepl('pastes|pasta|fideus|fideuà', vins$maridatge, ignore.case = T), 1
vins$maridatge_aperitius <- ifelse(grepl('aperitius|tapes|entrants|entrant|pica-pica', vins$maridatge,
vins$maridatge_canalons <- ifelse(grepl('canalons', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_amanida <- ifelse(grepl('amanida|amanides', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_carn_v <- ifelse(grepl('carns|carn vermella|carns vermelles|vedella|bou|cavall', vins$ma
```

```r
vins$maridatge_carn_b <- ifelse(grepl('carns|carn blanca|ànec|pollastre|aus|porc|conill|indi|xai', vins$
vins$maridatge_carn_c <- ifelse(grepl('carns|caça|senglar|llebre', vins$maridatge, ignore.case = T), 1,
vins$maridatge_pates <- ifelse(grepl('patés', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_guisats <- ifelse(grepl('plats amb salses|salses|guisats|guisades|fricandó|estofat|estufa
vins$maridatge_rostits <- ifelse(grepl('rostits', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_embotits <- ifelse(grepl('embotits|pernils ibèrics', vins$maridatge, ignore.case = T), 1
vins$maridatge_formatges_frescos <- ifelse(grepl('formatges frescos|formatge fresc', vins$maridatge, ig
vins$maridatge_formatges_tendres <- ifelse(grepl('formatges de pasta tova|tendres|formatges cremosos', v
vins$maridatge_formatges_semicurats <- ifelse(grepl('formatges semi curats| semicurat', vins$maridatge,
vins$maridatge_formatges_curats <- ifelse(grepl('formatges curats', vins$maridatge, ignore.case = T), 1
vins$maridatge_formatges_potents <- ifelse(grepl('formatges blaus', vins$maridatge, ignore.case = T), 1
vins$maridatge_peix_blau <- ifelse(grepl('peix|peix blau|peixos blaus|peixos grassos|salmó', vins$marida
vins$maridatge_peix_blanc <- ifelse(grepl('peix|peix blanc|peixos blancs', vins$maridatge, ignore.case =
vins$maridatge_peix_cru <- ifelse(grepl('carpaccio|peix cru| salmó|tonyina|carpaccions', vins$maridatge
vins$maridatge_peix_fregit <- ifelse(grepl('peix fregit', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_peix_forn <- ifelse(grepl('peix forn', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_peix_brasa <- ifelse(grepl('peix brasa', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_peix_fumat <- ifelse(grepl('fumats', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_peix_salaons <- ifelse(grepl('anxoves', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_peix_escabetx <- ifelse(grepl('escabetxats', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_marisc <- ifelse(grepl('marisc|crustàcis|ostres|gambes|galeres', vins$maridatge, ignore.
vins$maridatge_pop <- ifelse(grepl('pop', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_bolets <- ifelse(grepl('bolets', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_cuina_asiatica <- ifelse(grepl('cuina asiàtica|japonesa|cuina de fusió', vins$maridatge,
vins$maridatge_xocolata <- ifelse(grepl('xocolata', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_pizza <- ifelse(grepl('pizza|pizzes', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_paambtomaquet <- ifelse(grepl('pa amb tomàquet', vins$maridatge, ignore.case = T), 1, 0)

# Creem una tira binària estandarditzada per a codificar els maridatge
vins$vector_maridatge<- paste(vins$maridatge_arrossos,vins$maridatge_pastes,vins$maridatge_aperitius,vi

# Factoritzem
#vins$facto_raim <- as.numeric(factor(vins$vector_raim))
vins$vector_maridatge <- factor(vins$vector_maridatge)

# Adició de les possibilitats dels vins
vins$maridatge_possibilitats <- str_count(vins$vector_maridatge, "1")

# Eliminem columnes afegides per a trobar les composicions + columna original
#vins$maridatge<-NULL
vins$maridatge_arrossos<-NULL
vins$maridatge_pastes<-NULL
vins$maridatge_aperitius<-NULL
vins$maridatge_canalons<-NULL
vins$maridatge_amanida<-NULL
vins$maridatge_carn_v<-NULL
vins$maridatge_carn_b<-NULL
vins$maridatge_carn_c<-NULL
vins$maridatge_pates<-NULL
vins$maridatge_guisats<-NULL
vins$maridatge_rostits<-NULL
vins$maridatge_embotits<-NULL
vins$maridatge_formatges_frescos<-NULL
```

```
vins$maridatge_formatges_tendres<-NULL
vins$maridatge_formatges_semicurats<-NULL
vins$maridatge_formatges_curats<-NULL
vins$maridatge_formatges_potents<-NULL
vins$maridatge_peix_blau<-NULL
vins$maridatge_peix_blanc<-NULL
vins$maridatge_peix_cru<-NULL
vins$maridatge_peix_fregit<-NULL
vins$maridatge_peix_forn<-NULL
vins$maridatge_peix_brasa<-NULL
vins$maridatge_peix_fumat<-NULL
vins$maridatge_peix_salaons<-NULL
vins$maridatge_peix_escabetx<-NULL
vins$maridatge_marisc<-NULL
vins$maridatge_pop<-NULL
vins$maridatge_bolets<-NULL
vins$maridatge_cuina_asiatica<-NULL
vins$maridatge_xocolata<-NULL
vins$maridatge_pizza<-NULL
vins$maridatge_paambtomaquet<-NULL

summary(vins)
```

```
##                      do          nom            tipologia         envelliment
##  D.O. Penedès          :280   Length:280         blanc:132   Criança     :163
##  D. O. Terra Alta      :  0   Class :character   negre:118   Gran Reserva: 88
##  D.O Alella            :  0   Mode  :character   rosat: 30   Vi jove     : 29
##  D.O. Catalunya        :  0
##  D.O. Conca de Barberà :  0
##  D.O. Costers del Segre:  0
##  (Other)               :  0
##      anyada                        celler                    poblacio
##  Min.   :2003   Albet i Noya         : 17   Sant Sadurní d'Anoia: 41
##  1st Qu.:2017   Sumarroca            : 16   Vilobí del Penedès  : 27
##  Median :2019   Loxarel              : 14   Subirats            : 19
##  Mean   :2018   Alsina i Sardà       : 12   El Pla del Penedès  : 18
##  3rd Qu.:2020   Torre del Veguer     : 11   Font-Rubí           : 18
##  Max.   :2021   Bodegues Ca N'Estella:  8   Sant Pau d'Ordal    : 17
##  NA's   :34     (Other)              :202   (Other)             :140
##     vinya            elaboracio          cupatge           maridatge
##  Length:280         Length:280         Length:280         Length:280
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     tservei   graduacio_alcoholica_per acidesa_g_l_tartaric sulfuros_total_mg_l
##  NA-NA  :92   Min.   :10.50            Min.   :2.90         Min.   : 10.00
##  14-16  :67   1st Qu.:12.50            1st Qu.:4.20         1st Qu.: 38.00
##  6-8    :26   Median :13.00            Median :5.37         Median : 60.00
##  16-18  :18   Mean   :12.92            Mean   :5.22         Mean   : 61.38
##  4-6    : 8   3rd Qu.:13.50            3rd Qu.:6.05         3rd Qu.: 88.25
##  7-10   : 8   Max.   :14.70            Max.   :8.21         Max.   :123.00
```

```
##  (Other):61   NA's   :65              NA's   :169          NA's   :240
##  sucres_residuals_g_l       ph          preu_euros        premis
##  Min.   : 0.200       Min.   :2.900   Min.   :  3.33   Length:280
##  1st Qu.: 0.500       1st Qu.:3.150   1st Qu.:  7.80   Class :character
##  Median : 1.150       Median :3.250   Median :  9.98   Mode  :character
##  Mean   : 1.612       Mean   :3.277   Mean   : 14.60
##  3rd Qu.: 2.375       3rd Qu.:3.420   3rd Qu.: 14.65
##  Max.   :10.300       Max.   :3.690   Max.   :250.00
##  NA's   :210          NA's   :195     NA's   :83
##    ecologic
##  Length:280
##  Class :character
##  Mode  :character
##
##
##
##
##                                                              codi_coupage
##  00000000000000000000000000001000000000000000000000000000: 62
##  00000000000000000000000000000000000000000000000000000000: 36
##  00000000000000000000000000000000000000000100000000000000: 23
##  00000000000000000000000000000001000000001000000000000000: 15
##  00000000000000000000000000001000000000000000000000000000: 14
##  00000000000000000000000000000000000000000000000010000000: 10
##  (Other)                                                     :120
##         aspecte             color     intensitat      aromes_1     aromes_2
##  apagat     :  1   cirera      :101   alta : 10   afruitat: 39   0     :278
##  brillant   : 76   daurat      : 55   baixa:  1   espècies: 33   dolces:  2
##  net        :  5   groc pàl·lid: 35   mitja:  6   floral  : 12
##  tèrbol     :  2   palla       : 25   NA's :263   vegetal :  6
##  transparent:  3   rosat       : 14               NA's    :190
##  NA's       :193   robí        : 13
##                    (Other)     : 37
##                 aromes_3           gust         post_gust
##  aroma de bota      : 12   Fresc     :43   curt      :206
##  sense aroma de bota:268   Suau      :32   llarg     : 71
##                            Equilibrat:28   mig       :  1
##                            sedós     :15   molt llarg:  2
##                            Afruitat  :12
##                            (Other)   :70
##                            NA's      :80
##                           vector_maridatge maridatge_possibilitats
##  00000000000000000000000000000000:119     Min.   : 0.00
##  00000111000000000000000000000000:  5     1st Qu.: 0.00
##  00000111000000000110000000000000:  5     Median : 2.00
##  01000111000000000000000000000000:  5     Mean   : 2.55
##  00000111000100000000000000000000:  4     3rd Qu.: 5.00
##  00000111010000000000000000000000:  4     Max.   :10.00
##  (Other)                          :138
```

3.14 Generació de camps buits automàtic per KNN

Finalment, realitzem un últim check per a validar que no existeixen camps buits

```r
# Nombre de valors desconeguts per camp
#sapply(vins, function(x) sum(is.na(x)))

# Agrupació per varietats de vi
vins.negres <- vins[vins$tipologia == "negre",]
vins.blancs <- vins[vins$tipologia == "blanc",]
vins.rosats <- vins[vins$tipologia == "rosat",]

vins$preu_euros <- ifelse(vins$tipologia == "negre", kNN(vins.negres)$preu_euros, vins$preu_euros)
vins$preu_euros <- ifelse(vins$tipologia == "blanc", kNN(vins.blancs)$preu_euros, vins$preu_euros)
vins$preu_euros <- ifelse(vins$tipologia == "rosat", kNN(vins.rosats)$preu_euros, vins$preu_euros)
vins$anyada <- kNN(vins)$anyada
vins$graduacio_alcoholica_per <- kNN(vins)$graduacio_alcoholica_per
vins$acidesa_g_l_tartaric <- kNN(vins)$acidesa_g_l_tartaric
vins$sulfuros_total_mg_l <- kNN(vins)$sulfuros_total_mg_l
vins$sucres_residuals_g_l <- kNN(vins)$sucres_residuals_g_l
vins$ph <- kNN(vins)$ph
vins$aspecte <- kNN(vins)$aspecte
vins$intensitat <- kNN(vins)$intensitat
vins$aromes_1 <- kNN(vins)$aromes_1
vins$gust <- kNN(vins)$gust


# Nombre de valors desconeguts per camp
sapply(vins, function(x) sum(is.na(x)))
```

```
##                       do                      nom                tipologia
##                        0                        0                        0
##              envelliment                   anyada                   celler
##                        0                        0                        0
##                 poblacio                   vinya               elaboracio
##                        0                        0                        0
##                  cupatge                maridatge                  tservei
##                        0                        0                        0
## graduacio_alcoholica_per     acidesa_g_l_tartaric      sulfuros_total_mg_l
##                        0                        0                        0
##     sucres_residuals_g_l                       ph              preu_euros
##                        0                        0                        0
##                   premis                 ecologic            codi_coupage
##                        0                        0                        0
##                  aspecte                    color               intensitat
##                        0                        0                        0
##                aromes_1                 aromes_2                 aromes_3
##                        0                        0                        0
##                    gust                post_gust         vector_maridatge
##                        0                        0                        0
##  maridatge_possibilitats
##                        0
```

```r
# Per manca de temps s'han eliminat camps
vins$estil<-NULL
vins$premis<-NULL

# Mostrem el tipus de classe associat a cada camp de l'objecte vins
```

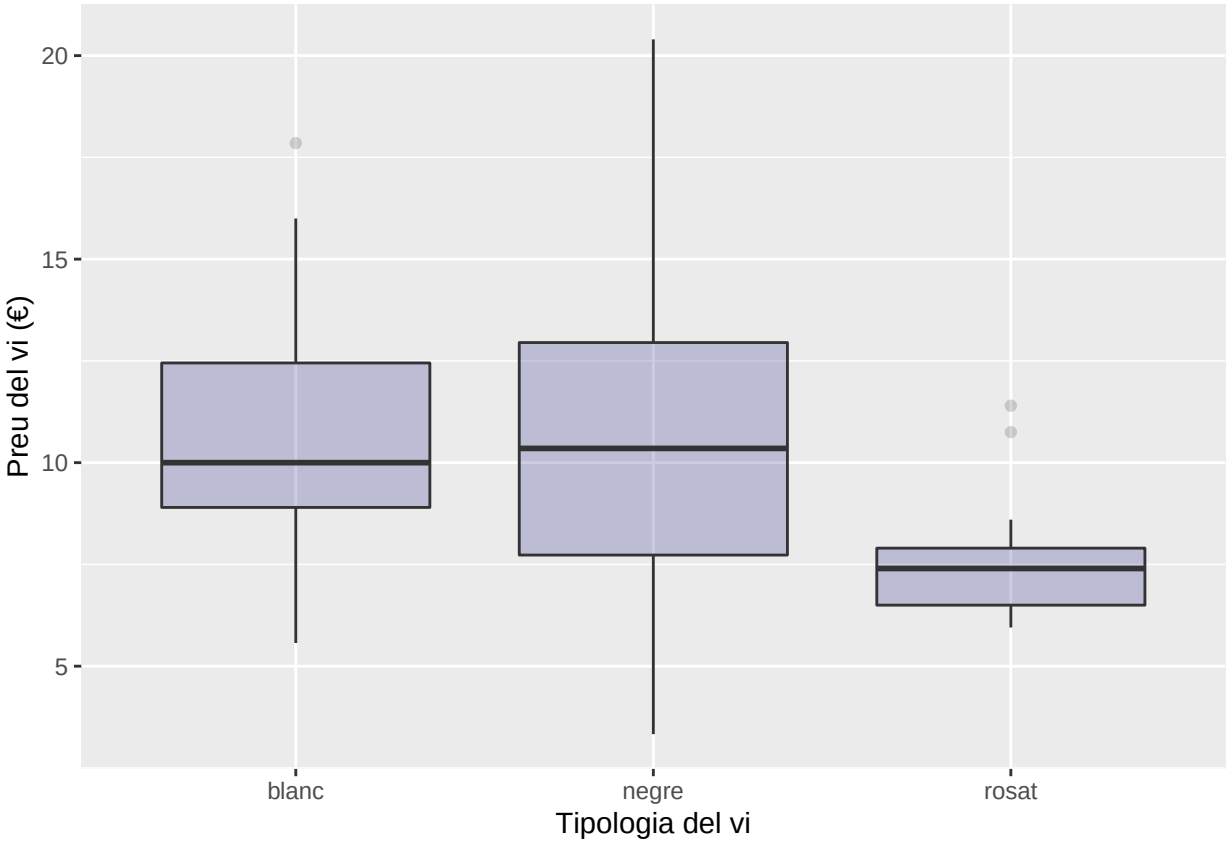
```
sapply(vins, function(x) class(x))
```

```
##                       do                      nom                tipologia
##                 "factor"              "character"                 "factor"
##              envelliment                   anyada                   celler
##                 "factor"                "integer"                 "factor"
##                 poblacio                    vinya               elaboracio
##                 "factor"              "character"              "character"
##                  cupatge                maridatge                   tservei
##              "character"              "character"                 "factor"
## graduacio_alcoholica_per     acidesa_g_l_tartaric      sulfuros_total_mg_l
##                "numeric"                "numeric"                "numeric"
##     sucres_residuals_g_l                       ph               preu_euros
##                "numeric"                "numeric"                "numeric"
##                 ecologic             codi_coupage                  aspecte
##              "character"                 "factor"                 "factor"
##                    color               intensitat                 aromes_1
##                 "factor"                 "factor"                 "factor"
##                 aromes_2                 aromes_3                     gust
##                 "factor"                 "factor"                 "factor"
##                post_gust         vector_maridatge  maridatge_possibilitats
##                 "factor"                 "factor"                "integer"
```

3.15 Comprovació de valors extrems

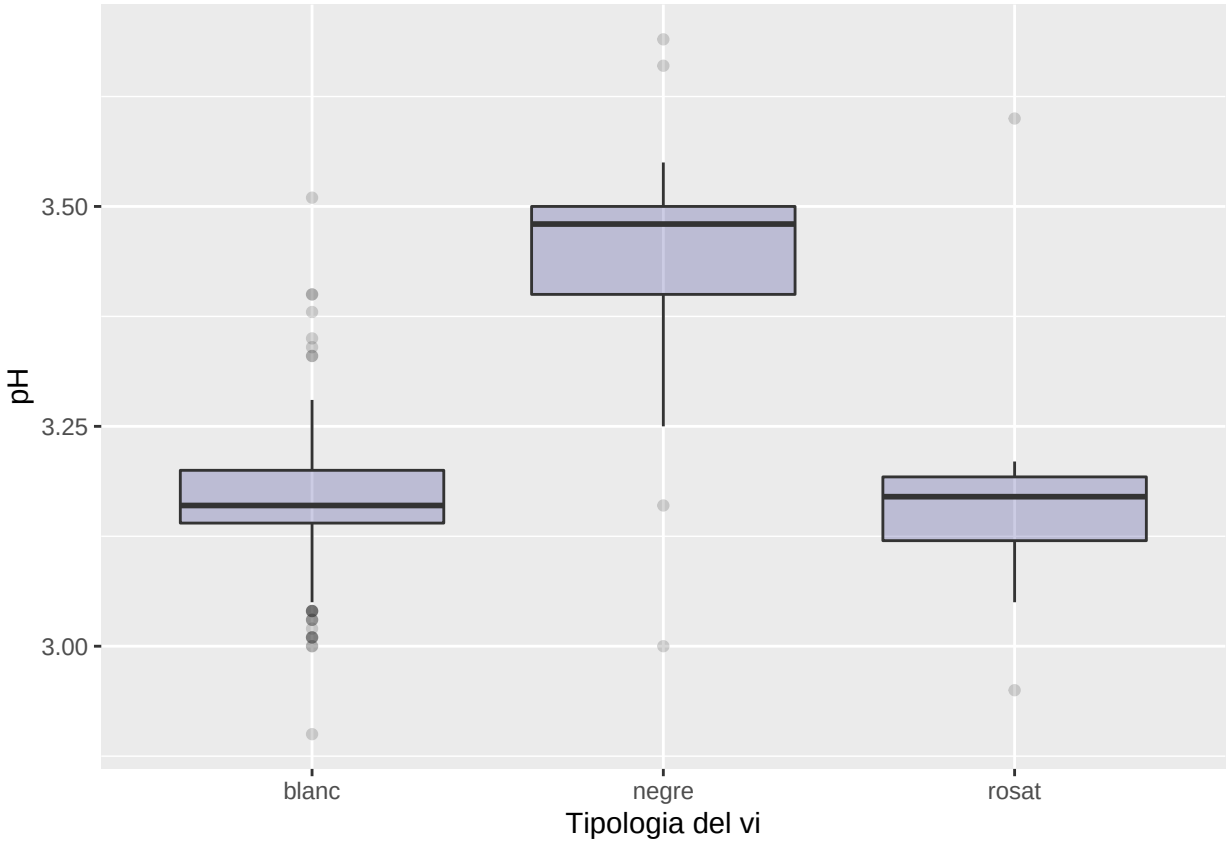
Tot i emprar tècniques dels veïns més propers, la generació automàtica de valors requereix de revalidar que els valors siguin propers als delimitats en els diagrames de bigotis amb els que s'ha investigat el contingut.

```
# Identifiquem els preus fora de sèrie
outliers <- boxplot(vins$preu_euros, plot = FALSE)$out

# Eliminem els preus extrems
vins<-vins[!(vins$preu_euros %in% outliers), ]

# Creem un diagrama de bigotis per a poder identificar que no hi han outliers derivats de l'error de co
ggplot(vins, aes(x=tipologia, y=preu_euros)) +
    geom_boxplot(fill=uoc_palette[1], alpha=0.2) +
    xlab("Tipologia del vi") + ylab("Preu del vi (€)")
```



Els valors de pH poden anar de un 4 (menys àcid) a un 2.8 (més àcid).

```
# Creem un diagrama de bigotis per a poder identificar que no hi han outliers derivats de l'error de co
ggplot(vins, aes(x=tipologia, y=ph)) +
    geom_boxplot(fill=uoc_palette[1], alpha=0.2) +
    xlab("Tipologia del vi") + ylab("pH")
```



Com podem veure,els paràmetres de pH es troben dins dels rangs establerts.

Visualitzem de forma més ràpida la resta de columnes numèriques.

```
# Identifiquem anyades fora de sèrie
outliers <- boxplot(vins$anyada, plot = FALSE)$out

# Mostrem les anyades extremes detectades
outliers
```

```
##  [1] 2012 2012 2012 2010 2011 2011 2003 2012 2010 2003 2010
```

```
# Eliminem els preus extrems
vins<-vins[!(vins$anyada %in% outliers), ]
```

```
# Identifiquem acidesa tartàrica fora de sèrie
outliers <- boxplot(vins$acidesa_g_l_tartaric, plot = FALSE)$out

# Mostrem les acideses que s'eliminaràn
outliers
```

```
##  [1] 7.05 7.11 8.21 8.00 3.20 7.00 7.40 3.00 2.90 7.30
```

```
# Eliminem els les acideses extremes
vins<-vins[!(vins$acidesa_g_l_tartaric %in% outliers), ]
```

```
# Identifiquem els sulfits total fora de sèrie
outliers <- boxplot(vins$sulfuros_total_mg_l, plot = FALSE)$out

# Mostrem els sulfits que eliminem
```

```
outliers
```

```
## [1] 19 19 25 19 18 10 10 10 10
```

```
# Eliminem els sulfits extrems
vins<-vins[!(vins$sulfuros_total_mg_l %in% outliers), ]
```

```
# Identifiquem els sucres residuals fora de sèrie
outliers <- boxplot(vins$sucres_residuals_g_l, plot = FALSE)$out

# Mostrem els sucres residuals que eliminem
outliers
```

```
##  [1]  2.60  2.50  3.30  2.90  0.20  3.00  2.50  2.50  4.20  2.50  3.35  2.50
## [13]  3.35  2.90  3.35  3.35  2.90  3.35  3.35  2.50  2.90  3.10  2.90  2.80
## [25]  2.80  2.80  2.50 10.30  3.40  4.65  4.20  3.35  3.90
```

```
# Eliminem els sucres residuals extrems
vins<-vins[!(vins$sucres_residuals_g_l %in% outliers), ]
```

3.16 Eliminació de columnes innecessàries per falta de scope

Per temes de schedule no s'han pogut portar a terme les correccions que es volien fer en el terroir i en la elaboració.

```
# Eliminació de la variable vinya
#vins<-vins$vinya <- NULL

# Eliminació de la variable elaboració
#vins<-vins$elaboracio <- NULL
```

3.17 Exportació del CLEAN dataset: vins_DO_Penedes.csv

```
# Exportación de les dades en .csv
write.csv(vins, "vins_DO_Penedes.csv")
```

4. Anàlisi de les dades

En aquest apartat: - selecció dels grups de dades que es volen analitzar/comparar (4.1) - Comprovació de la normalitat de les dades qualitatives i homogeneïtat de la variància (4.2) - Proves estadístiques per comparar els grups de dades (4.3) - Anàlisis de correlació quantitativa de variables que influencien més en el preu del vi (4.3.1) - Anàlisis de correlació qualitativa per determinar si el color del vi és suficient per a especificar el tipus de vi que es és? (4.3.2) - Hipòtesis de contrast per entendre si el preu del vi és superior en funció de si és ecològic? (4.3.3) - Model de regressió lineal múltiple per predir el preu d'una botella de vi. (4.3.4)

4.1 Selecció dels grups de dades que es volen analitzar/comparar

Es manté el conjunt de dades de l'apartat anterior.

4.2 Comprovació de la normalitat i homogeneïtat de la variància

Per a poder veure si la població de valors està distribuida normalment, s'aplica el el test de Anderson-Darling.

```
col.names = colnames(vins)

# Recorrem per totes les columnes
```

```
for (i in 1:ncol(vins)) {
  if (i == 1) cat("Variables que no segueixen una distribución normal:\n")
    # Identifiquem si la columna és numèrica
    if (is.integer(vins[,i]) | is.numeric(vins[,i])) {
      # apliquem el test de Anderson-Darling
      p_val = ad.test(vins[,i])$p.value
      if (p_val < 0.05) {
        # mostrem el valor del test amb un cat per evitar mostrar el valor per defecte [1]
        cat(col.names[i])
        # format per a permetre concatenar valors
        if (i < ncol(vins) - 1) cat(", ")
        if (i %% 3 == 0) cat("\n")
      }
  }
}
```

```
## Variables que no segueixen una distribución normal:
## anyada, graduacio_alcoholica_per, acidesa_g_l_tartaric, sulfuros_total_mg_l,
## sucres_residuals_g_l, ph, preu_euros,
## maridatge_possibilitats
```

Per estudiar la homogeneitat de variàncies entre el preu del vi i l'anyada la realitzem mitjançant l'aplicació d'un test de Fligner-Killeen

```
fligner.test(preu_euros ~ anyada, data = vins)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  preu_euros by anyada
## Fligner-Killeen:med chi-squared = 17.569, df = 8, p-value = 0.0247
```

Com que el p-value és superior a 0.05, acceptem la hipòtesis de que les variànces entre mostres són homogènes.

```
fligner.test(graduacio_alcoholica_per ~ ph, data = vins)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  graduacio_alcoholica_per by ph
## Fligner-Killeen:med chi-squared = 29.271, df = 33, p-value = 0.6534
```

Al obtenir un p-value superior a 0.05 acceptem la hipòtesis de que les variances de les dues mostres són homogènies.

4.3 Aplicació de proves estadístiques per comparar els grups de dades

4.3.1 Anàlisis de correlació quantitativa de variables que influencien més en el preu del vi

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")

# Calcular el coeficiente de correlació per cada variable quantitativa
for (i in 1:(ncol(vins) - 1)) {
  if (is.integer(vins[,i]) | is.numeric(vins[,i])) {
    spearman_test = cor.test(vins[,i],
                             vins[,length(vins)],
```

```
                              method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value

    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(vins)[i]
  }
}
```

```
## Warning in cor.test.default(vins[, i], vins[, length(vins)], method =
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(vins[, i], vins[, length(vins)], method =
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(vins[, i], vins[, length(vins)], method =
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(vins[, i], vins[, length(vins)], method =
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(vins[, i], vins[, length(vins)], method =
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(vins[, i], vins[, length(vins)], method =
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(vins[, i], vins[, length(vins)], method =
## "spearman"): Cannot compute exact p-value with ties
```

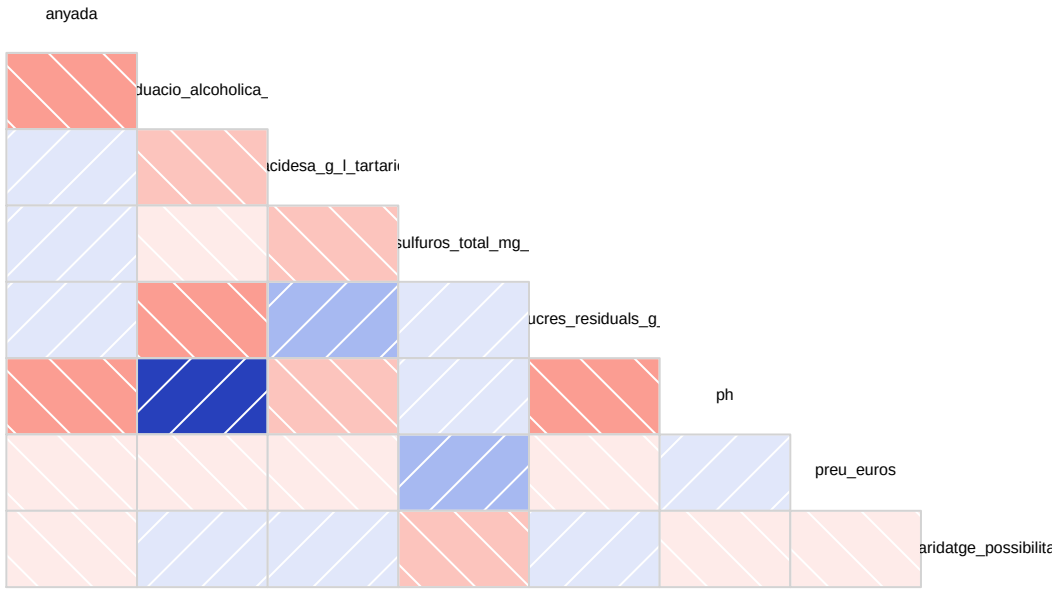
```
print(corr_matrix)
```

```
##                               estimate      p-value
## anyada                     -0.03688395 0.6124546847
## graduacio_alcoholica_per    0.07152519 0.3254750923
## acidesa_g_l_tartaric        0.04768186 0.5124526601
## sulfuros_total_mg_l        -0.25369060 0.0003983583
## sucres_residuals_g_l        0.02405374 0.7411771030
## ph                         -0.07582272 0.2971735660
## preu_euros                 -0.03196615 0.6606664345
```

La variable més correlada amb el preu és el sulfuros_total_mg_l i té coherència. Doncs els vins joves acostumen a tenir una menor quantitat de sulfits (60 mgs./l) que els criança o reserva que tenen una major relació (80 mgs./l). Obviament, el preu d'un criança sempre serà superior.

També podem graficar la correlació de forma gràfica on podem trobar la resta de relacions de forma més general:

```
corrgram(vins, order=NULL, lower.panel=panel.shade, upper.panel=NULL, text.panel=panel.txt, main="Correl
```



Observacions: La correlació de variables numèriques presenta una forta correlació entre la graduació alcohòlica i el pH. En efecte, els dos paràmetres estan molt polaritzats depenent de si tenim un vi blanc o un vi negre.

També es poden copsar un parell de relacions quelcom més tènues. La primera és la de la acidesa i els sucres residuals. Aquesta relació és verídica, ja que conforme avança el estat de maduració del vi, el raïm acumula sucre mentre redueix la acidesa. La segona presenta la relació del preu trobada en l'apartat anterior, on el preu en euros és relaciona amb el número de sulfits.

```
#table(vins$tipologia)
#prop.table(table(vins$tipologia))

chisq.test(vins$tipologia, vins$gust)
```

```
## Warning in chisq.test(vins$tipologia, vins$gust): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  vins$tipologia and vins$gust
## X-squared = 90.903, df = 36, p-value = 1.211e-06
```

4.3.2 Anàlisis de correlació qualitativa per determinar si el color del vi és suficient per a especificar el seu tipus

```
CODI_CUPATGE<-as.factor(as.numeric(vins$codi_coupage))
#codi_cup_Rel=relevel(CODI_CUPATGE, ref = '1')

# apliquem el model de relació
logit_model_1 <- glm(formula=codi_coupage~factor(color), data=vins, family=binomial)
summary(logit_model_1)
```

```
##
## Call:
## glm(formula = codi_coupage ~ factor(color), family = binomial,
##     data = vins)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2815   0.3923   0.4033   0.5553   0.7938
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   2.46810    0.46578   5.299 1.17e-07 ***
## factor(color)daurat          -1.47485    0.59497  -2.479   0.0132 *
## factor(color)groc pàl·lid     0.05763    0.87003   0.066   0.9472
## factor(color)maduixa         -0.67634    1.17627  -0.575   0.5653
## factor(color)palla           -1.14634    0.73049  -1.569   0.1166
## factor(color)robí            -0.52219    1.16611  -0.448   0.6543
## factor(color)rosa ataronjat  -0.85866    1.19036  -0.721   0.4707
## factor(color)rosat           16.09797 1966.64959   0.008   0.9935
## factor(color)teula           16.09797 3765.84718   0.004   0.9966
## factor(color)verdós          16.09797 2465.32571   0.007   0.9948
## factor(color)violeta         16.09797 4612.20201   0.003   0.9972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.41  on 190  degrees of freedom
## Residual deviance: 129.27  on 180  degrees of freedom
## AIC: 151.27
##
## Number of Fisher Scoring iterations: 17
```

Les variables explicatives que permeten identificar amb certa claredat el cupatge o mescla de vins amb el color són el daurat i el color palla. La resta de variables explicatives estàn llunt de l'aproximació. Per tant, no és suficient el color del vi per a poder encertar el tipus de vi que porta. En el cas del daurat es pot relacionar per la gran quantitat de Chardonnay.

4.3.3 Hipòtesis de contrast per entendre si el preu del vi és superior en funció de si és ecològic?

Es realitza una hipòtesis de contrast sobre la diferència de preu en funció de si és ecològic o no ho és.

```
# Separació de mostres de vi ecològic i no ecològic
vins.eco.preus <- vins[vins$ecologic == "Si",]$preu_euros
vins.noeco.preus <- vins[vins$ecologic == "No",]$preu_euros
```

Com es mostra a continuació, la hipòtesis nula conclou que el preu no és superior. La hipòtesis alternativa considerarà que el preu mig del vi ecològic és superior, de forma que la diferència de mitges serà inferior a 0, el que fa que estiguem davant d'un cas unilateral.

$H_0 : \mu_1 - \mu_2 = 0$ $H_1 : \mu_1 - \mu_2 < 0$ on μ_1 és la mitja de la població de la primera mostra i μ_2 és la mitja de la població de la segona mostra. $\alpha = 0.05$

```
t.test(vins.noeco.preus,vins.eco.preus, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  vins.noeco.preus and vins.eco.preus
## t = -0.22856, df = 166.57, p-value = 0.4097
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.6011337
## sample estimates:
## mean of x mean of y
##  10.09518  10.19156
```

Amb un p-value superior a la significança fixada, acceptem la hipòtesis nula. Podem refutar la hipòtesis de que un vi ecològic és més car que un vi no ecològic. En part té sentit perquè la viticultura ecològica competeix amb els gran reserva i reserva que són molt més cars.

4.3.4 Model de regressió lineal múltiple per a identificar el preu d'una botella de vi

Es cerca poder modelar el preu del vi, de forma que quan es vagi a buscar una botella es tingui una preu de referència. La forma òptima de realitzar les regressions lineal múltiples, a través de les variables que estan més correlades amb el preu i considerar-les en funció de la seva màxima correlació. No obstant, com que la variable no s'ha pogut correlar, s'ha optat per a implementar regressors qualitatius amb la finalitat d'escollir aquella que presenti un major coeficient de determinació.

```
# Generació de models variant els regressors
# Model algorismic 1 per trobar el preu del vi
model1 <- lm(preu_euros ~ tipologia + color + codi_coupage + vector_maridatge + gust, data = vins)
# Model fitxa de cata: Amb la informació de cata és suficient per establir el preu del vi?
model2 <- lm(preu_euros ~ tipologia + aspecte + color + intensitat + aromes_1 + aromes_2 + aromes_3 + gu
# Model etiqueta: La informació de la etiqueta és suficient per establir el preu del vi?
model3 <- lm(preu_euros ~ tipologia + envelliment + codi_coupage + anyada + ecologic + celler + graduaci
# Model cupatge, maridatge, celler
model4 <- lm(preu_euros ~ codi_coupage + vector_maridatge + celler, data = vins)

# Calculem la bondat d'ajust per cada model
tabla.coeficientes <- matrix(c(1, summary(model1)$r.squared,
2, summary(model2)$r.squared,
3, summary(model3)$r.squared,
4, summary(model4)$r.squared),
ncol = 2, byrow = TRUE)
colnames(tabla.coeficientes) <- c("Model", "R^2")

# mostra taula
tabla.coeficientes
```

```
##      Model       R^2
## [1,]     1 0.8989359
## [2,]     2 0.2919655
```

```
## [3,]     3 0.6180809
## [4,]     4 0.9306741
```

En la taula comparativa de bondats d'ajust, el model més òptim és el 4t, tot i que no ajusta correctament a les dades.

Apliquem un exemple segons el model 4, en el que predirem el preu d'un vi Cabernet Sauvignon, que maridi amb carns blanques del celler Torelló:

```
# Exemple 1: Busquem un Cabernet Sauvignon,que maridi amb carns blanques de Torelló

# Exemple individual de la funció
#codicoupage_a_raims("00000000000000000000000000000001000000000000000000000000")


newdata <- data.frame(
codi_coupage = "00000000000000000000000000000001000000000000000000000000",
vector_maridatge = "00000111000000000000000000000000",
celler="Torelló"
)
# Predicció del preu
preu_estimat <- predict(model4, newdata)
```

```
## Warning in predict.lm(model4, newdata): prediction from a rank-deficient fit may
## be misleading
```

```
sprintf("El preu per a la botella de vi que busques hauria d'aproximar-se al preu de %#.2f €.", preu_es
```

```
## [1] "El preu per a la botella de vi que busques hauria d'aproximar-se al preu de 16.38 \200."
```

```
#library(pROC)
#prob_low=predict(model4, vins, type="response")
#r=roc(preu_estimat,prob_low, data=vins)
```

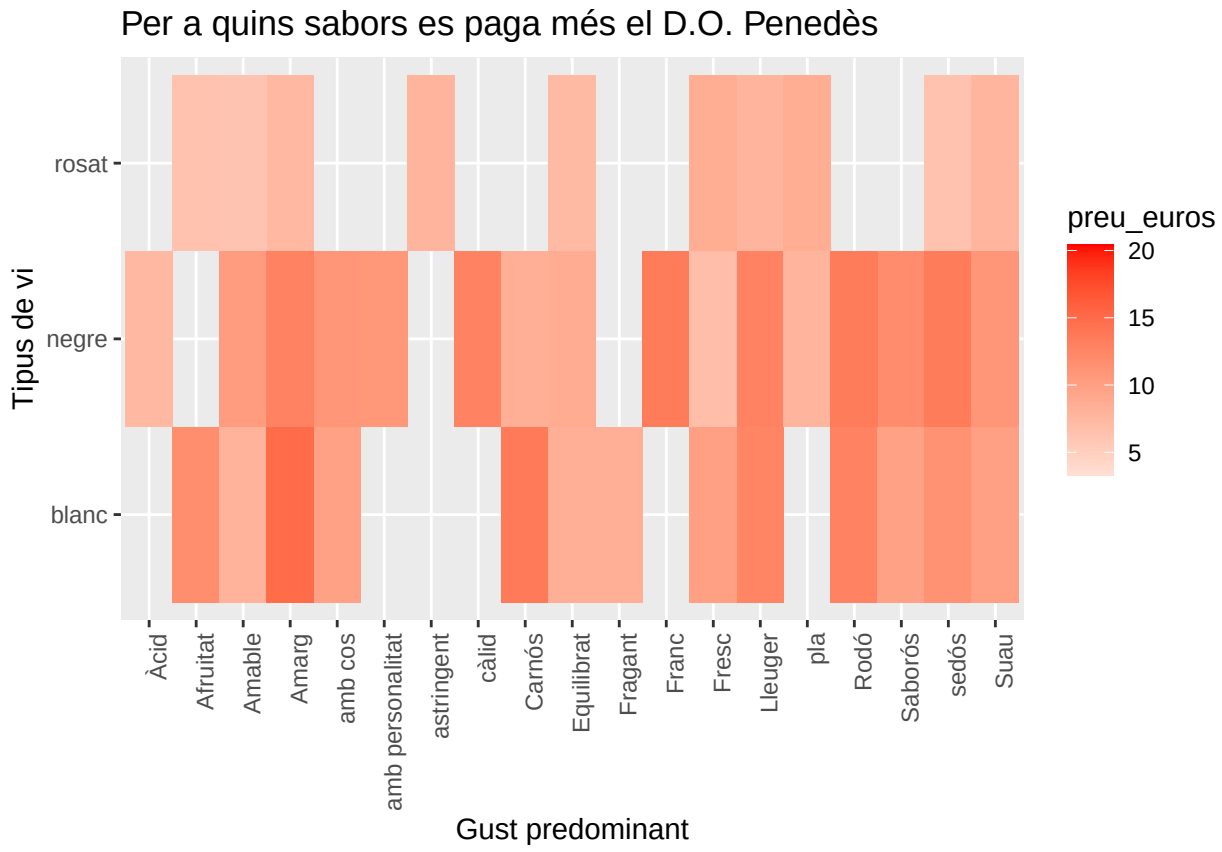
5. Representació dels resultats a partir de taules i gràfiques

Com que l'apartat anterior no disposa de gràfiques, en aquest apartat s'han desenvolupat unes quantes preguntes amb resposta visual. S'ha entès que, al fer èmfasis per part del professor en el desenvolupament de gràfiques elaborades, s'esperaria dins el contingut de la pràctica.

En aquest apartat: - Quins són els sabors del vi pels quals estem disposats a pagar més? (5.1) - Quins són els cupatges per anyada més rendibles en D.O. Penedès? (5.2) - Com podem aconseguir un gust en funció de la varietat de vi D.O. Penedès? (5.3) - Quines varietats de vins D.O. Penedès mariden millor? (5.4)

5.1 Quins són els sabors pels que paguem més el vi?

```
ggplot(vins, aes(x = gust, y = tipologia, fill = preu_euros)) + geom_tile() +
  scale_fill_gradient2(low = "green", high = "red")+ theme(axis.text.x=element_text(angle=90, hjust=1))
```



5.2 Quins cupatges són els més rendibles en la D.O. Penedès?

```
# Funció codicoupage a conversió de raïms emprats en el cupatge
codicoupage_a_raims <- function(x) {
  # Conversor del codi de coupatge a tipus de raïm
  cupatge_str<-""

  raim <-c("Airen", "PansaBlanca", "Albarinyo", "Carinyena Blanca", "Chardonnay", "Chenin Blanc", "Coron
  n <- gregexpr(pattern ='1',x)[[1]]
  if ( n != "-1") {
    for (p in gregexpr(pattern ='1',x)[[1]]) {
      cupatge_str<-paste(cupatge_str,raim[p], sep="-")
    }
    cupatge_str<-substring(cupatge_str, 2)
    } else {
    cupatge_str<-NA
  }
  return(cupatge_str)
}

# Exemple individual de la funció
#codicoupage_a_raims("00000000000000000001000000000000000000000001000000000000000")

# Apliquem la conversió de 1 del codi al format de interpretable
vins$codi_coupage<-apply(vins['codi_coupage'],1,codicoupage_a_raims)
```

```
## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used
```

```
## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used
```

```
## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used

## Warning in if (n != "-1") {: the condition has length > 1 and only the first
## element will be used
```

```r
ggplot(vins, aes(x = anyada, y = codi_coupage, fill = preu_euros)) + geom_tile() +
  scale_fill_gradient2(low = "#73EDFF", high = "#000078")+ theme(axis.text.x=element_text(angle=90, hju
```

5.3 Identificar els gustos amb els cupatges

```
library("scales")
```

```
##
## Attaching package: 'scales'
## The following object is masked from 'package:purrr':
##
##     discard
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
vins$ph2 <- rescale(vins$preu_euros, to = c(0, 1))
ggplot(vins, aes(x = gust, y = codi_coupage, fill = ph2)) + geom_tile() +
  scale_fill_gradient2(low = "white", high = "red", mid="black")+ theme(axis.text.x=element_text(angle=
```

5.4 Entendre els gustos amb els maridatges: quines varietats de vins mariden millor

```
# Converteix les descripcions de fase visual de l'aspecte en característiques descriptives
vins$maridatge_arrossos <- ifelse(grepl('arrossos|paelles', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_pastes <- ifelse(grepl('pastes|pasta|fideus|fideuà', vins$maridatge, ignore.case = T), 1
vins$maridatge_aperitius <- ifelse(grepl('aperitius|tapes|entrants|entrant|pica-pica', vins$maridatge,
vins$maridatge_canalons <- ifelse(grepl('canalons', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_amanida <- ifelse(grepl('amanida|amanides', vins$maridatge, ignore.case = T), 1, 0)
```

```r
vins$maridatge_carn_v <- ifelse(grepl('carns|carn vermella|carns vermelles|vedella|bou|cavall', vins$mar
vins$maridatge_carn_b <- ifelse(grepl('carns|carn blanca|ànec|pollastre|aus|porc|conill|indi|xai', vins$
vins$maridatge_carn_c <- ifelse(grepl('carns|caça|senglar|llebre', vins$maridatge, ignore.case = T), 1,
vins$maridatge_pates <- ifelse(grepl('patés', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_guisats <- ifelse(grepl('plats amb salses|salses|guisats|guisades|fricandó|estofat|estufa
vins$maridatge_rostits <- ifelse(grepl('rostits', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_embotits <- ifelse(grepl('embotits|pernils ibèrics', vins$maridatge, ignore.case = T), 1
vins$maridatge_formatges_frescos <- ifelse(grepl('formatges frescos|formatge fresc', vins$maridatge, ig
vins$maridatge_formatges_tendres <- ifelse(grepl('formatges de pasta tova|tendres|formatges cremosos', v
vins$maridatge_formatges_semicurats <- ifelse(grepl('formatges semi curats| semicurat', vins$maridatge,
vins$maridatge_formatges_curats <- ifelse(grepl('formatges curats', vins$maridatge, ignore.case = T), 1
vins$maridatge_formatges_potents <- ifelse(grepl('formatges blaus', vins$maridatge, ignore.case = T), 1
vins$maridatge_peix_blau <- ifelse(grepl('peix|peix blau|peixos blaus|peixos grassos|salmó', vins$marida
vins$maridatge_peix_blanc <- ifelse(grepl('peix|peix blanc|peixos blancs', vins$maridatge, ignore.case =
vins$maridatge_peix_cru <- ifelse(grepl('carpaccio|peix cru| salmó|tonyina|carpaccions', vins$maridatge
vins$maridatge_peix_fregit <- ifelse(grepl('peix fregit', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_peix_forn <- ifelse(grepl('peix forn', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_peix_brasa <- ifelse(grepl('peix brasa', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_peix_fumat <- ifelse(grepl('fumats', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_peix_salaons <- ifelse(grepl('anxoves', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_peix_escabetx <- ifelse(grepl('escabetxats', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_marisc <- ifelse(grepl('marisc|crustàcis|ostres|gambes|galeres', vins$maridatge, ignore.
vins$maridatge_pop <- ifelse(grepl('pop', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_bolets <- ifelse(grepl('bolets', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_cuina_asiatica <- ifelse(grepl('cuina asiàtica|japonesa|cuina de fusió', vins$maridatge,
vins$maridatge_xocolata <- ifelse(grepl('xocolata', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_pizza <- ifelse(grepl('pizza|pizzes', vins$maridatge, ignore.case = T), 1, 0)
vins$maridatge_paambtomaquet <- ifelse(grepl('pa amb tomàquet', vins$maridatge, ignore.case = T), 1, 0)

subset_x<-subset(vins, select = c('codi_coupage', 'maridatge_arrossos', 'maridatge_pastes', 'maridatge_a
#y1<-table(subset_x)

subset_x$codi_coupage <- factor(subset_x$codi_coupage) # make year a factor.

t1<-aggregate(x = subset_x$maridatge_arrossos,by = list(subset_x$codi_coupage),FUN = sum)
t2<-aggregate(x = subset_x$maridatge_pastes,by = list(subset_x$codi_coupage),FUN = sum)
t3<-aggregate(x = subset_x$maridatge_aperitius,by = list(subset_x$codi_coupage),FUN = sum)
t4<-aggregate(x = subset_x$maridatge_canalons,by = list(subset_x$codi_coupage),FUN = sum)
t5<-aggregate(x = subset_x$maridatge_amanida,by = list(subset_x$codi_coupage),FUN = sum)
t6<-aggregate(x = subset_x$maridatge_carn_v,by = list(subset_x$codi_coupage),FUN = sum)
t7<-aggregate(x = subset_x$maridatge_carn_b,by = list(subset_x$codi_coupage),FUN = sum)
t8<-aggregate(x = subset_x$maridatge_carn_c,by = list(subset_x$codi_coupage),FUN = sum)
t9<-aggregate(x = subset_x$maridatge_pates,by = list(subset_x$codi_coupage),FUN = sum)
t10<-aggregate(x = subset_x$maridatge_guisats,by = list(subset_x$codi_coupage),FUN = sum)
t11<-aggregate(x = subset_x$maridatge_rostits,by = list(subset_x$codi_coupage),FUN = sum)
t12<-aggregate(x = subset_x$maridatge_embotits,by = list(subset_x$codi_coupage),FUN = sum)
t13<-aggregate(x = subset_x$maridatge_formatges_frescos,by = list(subset_x$codi_coupage),FUN = sum)
t14<-aggregate(x = subset_x$maridatge_formatges_tendres,by = list(subset_x$codi_coupage),FUN = sum)
t15<-aggregate(x = subset_x$maridatge_formatges_semicurats,by = list(subset_x$codi_coupage),FUN = sum)
t16<-aggregate(x = subset_x$maridatge_formatges_curats,by = list(subset_x$codi_coupage),FUN = sum)
t17<-aggregate(x = subset_x$maridatge_formatges_potents,by = list(subset_x$codi_coupage),FUN = sum)
t18<-aggregate(x = subset_x$maridatge_peix_blau,by = list(subset_x$codi_coupage),FUN = sum)
t19<-aggregate(x = subset_x$maridatge_peix_blanc,by = list(subset_x$codi_coupage),FUN = sum)
```

```r
t20<-aggregate(x = subset_x$maridatge_peix_cru,by = list(subset_x$codi_coupage),FUN = sum)
t21<-aggregate(x = subset_x$maridatge_peix_fregit,by = list(subset_x$codi_coupage),FUN = sum)
t22<-aggregate(x = subset_x$maridatge_peix_forn,by = list(subset_x$codi_coupage),FUN = sum)
t23<-aggregate(x = subset_x$maridatge_peix_brasa,by = list(subset_x$codi_coupage),FUN = sum)
t24<-aggregate(x = subset_x$maridatge_peix_fumat,by = list(subset_x$codi_coupage),FUN = sum)
t25<-aggregate(x = subset_x$maridatge_peix_salaons,by = list(subset_x$codi_coupage),FUN = sum)
t26<-aggregate(x = subset_x$maridatge_peix_escabetx,by = list(subset_x$codi_coupage),FUN = sum)
t27<-aggregate(x = subset_x$maridatge_marisc,by = list(subset_x$codi_coupage),FUN = sum)
t28<-aggregate(x = subset_x$maridatge_pop,by = list(subset_x$codi_coupage),FUN = sum)
t29<-aggregate(x = subset_x$maridatge_bolets,by = list(subset_x$codi_coupage),FUN = sum)
t30<-aggregate(x = subset_x$maridatge_cuina_asiatica,by = list(subset_x$codi_coupage),FUN = sum)
t31<-aggregate(x = subset_x$maridatge_xocolata,by = list(subset_x$codi_coupage),FUN = sum)
t32<-aggregate(x = subset_x$maridatge_pizza,by = list(subset_x$codi_coupage),FUN = sum)
t33<-aggregate(x = subset_x$maridatge_paambtomaquet,by = list(subset_x$codi_coupage),FUN = sum)

df3 <- data.frame(t1,t2[2],t3[2],t4[2],t5[2],t6[2],t7[2],t8[2],t9[2],t10[2],t11[2],t12[2],t13[2],t14[2]

colnames(df3) <- c('codi_coupage', 'arrossos', 'pastes', 'aperitius','canalons','amanida','carns vermel

# Guardem la informació com rownames
df4 <- df3;
rownames(df4) <- df3[, 1];

# Eliminem la columna no numèrica
df4 <- df4[, -1];

# Emprem as.matrix per a convertir data.frame a una matriu
logmap <- heatmap(as.matrix(df4), scale = "column", Colv=NA, Rowv=NA, margins = c(5, 10), main = "Relac
```

6. Conclusions

6.1 Resultats

Els paràmetres de correlació i contrast d'hipòtesis no han sigut concluents per a poder identificar quines variables exerceixen una major influència sobre el preu final de la botella de vi. S'ha trobat que l'execució del markdown tenia una forta variació provinent de les dades generades per KNN, també que el nombre de NA en la taula llegida era considerable. Ambdos indicadors determinen que no hi havia suficient informació en la taula inicial com per a poder donar uns resultats concluents.

En un àmbit professional caldria reunir-se amb els stakeholders per reavaluar les dades o les preguntes. No obstant, com que aquesta pràctica requereix exposar els resultats, a continuació s'exposen els resultats de les slides, atès que són els únics que romandràn estàtics. Comento aquest aspecte perquè cada vegada que s'executa Knit, el programa s'executa de nou i hi ha risc de que es modifiquin.

Com ja s'ha avançat en el paràgraf anterior, s'ha realitzat un anàlisis de correlació quantitiva amb la finalitat de saber quins paràmetres numèrics poden influenciar en el preu del vi. Els paràmetres que es tenien en comtpe era la anyada de la botella (any de producció), la graduació alcohòlica, la acidesa , el número de sulfits, el ph i el preu en euros. La variable que presentava una major correlació eren els sulfits. De fet, els vins amb criança tenen molts més sulfits que els vins joves, de forma que té certa lògica.

Pel que fa a l'anàlisis de correlació qualitativa per a determinar si el color del vi era suficient per a especifar de quin tipus de vi es tractava, s'ha obtingut una lleugera resposta en vins blancs, on hi hauria certa probabilitat de detectar correctament els blancs de color daurat i color palla.

Pel que fa a la hipòtesis de contraast per entendre si el preu d'un vi ecològic és major a un vi convencional ha donat que resulta fals, en bona mesura perquè s'ha tingut en compte els reserva i criança que no ho són, mentre que la majoria de ecològics són vins joves.

Finalment, s'ha realitzat un model de regressió lineal múltiples per predir el preu d'una botella de vi on en el millor dels casos s'ha obtingut una bondat d'ajust de 0.92 introduint el codi del cupatge de vins (per exemple un Merlot), el maridatge (per prendre amb carn) i el nom del celler (del celler Torres).

Pel que fa a les gràfiques, destaca en temes de sabors que en els vins vermells paguem més per sedosos, càlids i amables, mentre que en blancs per equilibrats i afruitats. Pel que fa al preu segons varietat i l'anyada es presenta un patró individual. Faltaria un major nombre de dades per a poder identificar grans anyades derivades de la climatologia i meteorologia de l'any. Es considera no concluent. Pel que fa als gustos que més defineixen els vins D.O. Penedès són equilibrat, suau i fresc. El Xarel · lo, Merlot, Cabernet-Sauvignon són els vins tot terreny per tenir a la cuina, ja que mariden amb infinitat de plats.

Per a solucionar el problema de la falta d'informació, s'ha fet ús abusiu del mètode d'imputació de valors. Aquest mètode té sentit si tenim un gran volum de dades i un petit percentatge de camps buits. No obstant, s'ha près aquesta decisió per a poder entregar la pràctica a temps. Un altra punt interessant al mancar dades és que entre elles es poden ajudar per extreure informació. Aquest aspecte fa que sigui necessari un ordre, no es pot començar a netejar un camp sense abans estar segur de que no pot ajudar-ne un altra. Per exemple, hi ha tipologies que apareix "vi negre jove"", el valor jove es correspont a la criança i cal implementar-lo en el camp corresponent (si no existeix).

En conclusió, per a donar uns resultats estadísticament acceptables, *caldria demanar als nostres stakeholders més dades quantitatives* per a poder aproximar els resultats.

6.2 Pràctica

En aquesta ocasió, s'ha escollit treballar amb un dataset real per a experimentar les dificultats amb les que es troba un analista de dades. Lluny dels datasets preparats de Kaggle, UCI Machine Learning Repository, etc, aquest ha permès veure la importància de la qualitat en les dades. El que s'ha destacat és el gran volum de treball en la primera fase de neteja de dades per intentar estabilitzar-les, especialment en identificar aquelles columnes troncals a partir de les quals pots netejar la resta.

El problema més important amb el que m'he trobat és que les fitxes de cata són molt qualitatives i molt poc quantitatives. S'acostuma a donar el grau d'alcohol i el preu. Però la resta de parametres científics com el pH, sulfats, sucres residuals, etc. només s'acostuma a donar per a botelles on cal justificar el preu i això altera la informació.

La motivació de la pràctica 2 és la d'ampliar el porfolio dels meus treballs per a poder demostrar les habilitats en el camp de la ciència de dades. En aquesta ocasió, també em tenia encuriosit després d'una participació en una cata a cegues sense tenir-ne cap coneixement.

7. Taula de contribucions

Contribucions	Signatura
Investigació prèvia	JMV

Contribucions	Signatura
Redacció de les respostes	JMV
Desenvolupament del codi	JMV

8. Recursos

1. Nofuentes O. INCAVI (2020) A taula! Vins Catalans. Manual per a la restauració. Editorial SOM.
2. Bodegas Sierra Norte. Ficha de cata. http://www.bodegasierranorte.com/wp-content/uploads/2016/01/FICHA-DE-CATA_BODEGA-SIERRA-NORTE_CATA-A-CIEGAS.pdf
3. Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
4. Recurs online sobre test en R: https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/
5. Recurs online sobre visualitzacions en R: https://www.data-to-viz.com/
6. Teguayco Gutiérrez González. Ejemplo Práctica 2: Limpieza y validación de los datos