Pràctica 2 – Tipologia i cicle de vida de les dades

COMPONENTS DEL GRUP: **Arnau Santos Ribelles** i **Ferran Pintó Haro**.

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

El dataset conté informació relacionada amb les variants negre i blanca de la varietat de vi portuguès "Vinho Verde". Segons els autors del conjunt de dades, per motius de privacitat i logística només hi ha variables fisicoquímiques (entrades) i sensorials (sortida), i no hi ha, per exemple, dades sobre tipus de raïm, marca de vi, preu de venta, etc.

Comptem amb dos datasets (un per la variant blanca i un per la negre) que ajuntem posteriorment. Les 14 variables que conté el dataset són:
"**type**": valor categòric corresponent al tipus de varietat (white / red).
"**fixed acidity**": valor numèric corresponent al la quantitat d'àcids implicats (que no s'evaporen fàcilment).
"**volatile acidity**": valor numèric corresponent a la quantitat d'àcid acètic en el vi.
"**citric acid**": valor numèric corresponent a la quantitat d'àcid cítric. Entre 0 i 1.
"**residual sugar**": valor numèric corresponent a la quantitat (grams/litre) de sucre que queda després de la parada de la fermentació.
"**chlorides**": valor numèric corresponent a la quantitat de sal al vi.
"**free sulfur dioxide**": valor numèric corresponent a la quantitat de la forma lliure del SO2.
"**total sulfur dioxide**": valor numèric corresponent a la quantitat de formes lliures i lligades de SO2.
"**density**": valor numèric corresponent a la densitat del vi.
"**pH**": valor numèric corresponent al PH del vi. Escala de 0 (molt àcid) al 14 (molt bàsic).
"**sulphates**": valor numèric corresponent a la quantitat de sulfats.
"**alcohol**": valor numèric corresponent als graus d'alcohol que té el vi (percentatge d'alcohol).
"**quality**": qualitat percebuda del vi. Valor numèric entre 0 i 10.
"quality_d": variable categòrica creada segons la qualitat del vi ("bo" si la qualitat és 7 o superior i "no bo" si és inferior a 7)

Mostrem el resultat de la funció str() en R pel dataset de vi blanc i en el de vi negre, per veure de forma més visual les dades. En el codi de R també s'utilitza la funció summary().

```
summary(red_wine_df)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Dataset vi negre

```
summary(white_wine_df)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

Dataset vi blanc

Mostrem també el resultat de la funció str() pel dataset total un cop unificats els dos.

```
summary(wine)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923
##  Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949
##  Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970
##  Max.   :0.61100   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.210   Median :0.5100   Median :10.30   Median :6.000
##  Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
##  3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
##      type
##  Length:6497
##  Class :character
##  Mode  :character
```

És un dataset rellevant per poder fer tasques de classificació i regressió, al tenir diferents variables numèriques corresponents a les propietats del vi i una variable corresponent a la qualitat percebuda. Així podrem observar quines característiques del vi estan més relacionades amb la qualitat.

És pretén respondre la pregunta següent:
Quines són les propietats fisicoquímiques que fan que un vi sigui bo?

Alhora que observar:
Les propietats que fan que un vi sigui bo són les mateixes per vins negres i blancs?

2. Integració i selecció de les dades d'interès a analitzar. Pot ser el resultat d'addicionar diferents datasets o una subselecció útil de les dades originals, en base a l'objectiu que es vulgui aconseguir.

Com s'ha comentat, originàriament tenim dos datasets, un corresponent a vi negre i un a vi blanc. En aquest pas ajuntem els dos datasets per tenir un únic dataset corresponent a vins.
Abans de fer-ho s'ha creat una nova columna als dos datasets anomenada "type" corresponent al tipus de vi que és: "White" si és vi blanc i "Red" si és negre. Es fa ja que posteriorment ens interessarà saber la varietat de vi que correspon a cada registre, com per exemple, per fer anàlisis per cada tipus de vi.

S'han integrat els dos datasets en un de sol mitjançant una fusió vertical, ja que tenim el mateix format de les bases de dades: les mateixes columnes amb els mateixos tipus de variables. Així, tenim un dataset que inclou els dos tipus de vins, per poder fer anàlisis conjunts de "vins", però permetent fer anàlisis per tipus de vi, segons la variable creada "type".

Ara tenim 6497 files, corresponent a la suma de les files dels dos datasets originals.

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos.

S'ha comprovat que el dataset no té valors absents (NA) en cap variable (es pot veure com en el codi).
Al comprovar si les dades contenen zeros, s'ha vist que la columna "citric acid" té 151 zeros, però els considerem vàlids ja que està al rang de valors possibles de la variable: pot no haver-hi àcid cítric als vins.

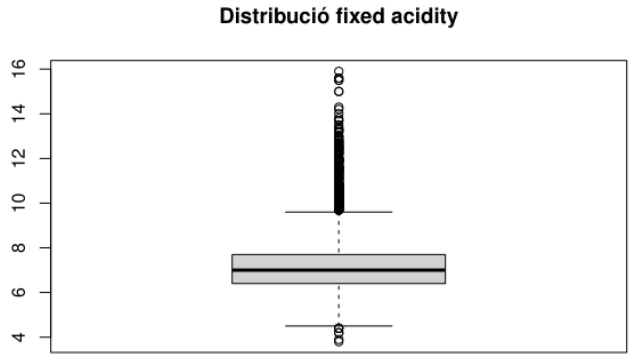
3.2. Identifica i gestiona els valors extrems.

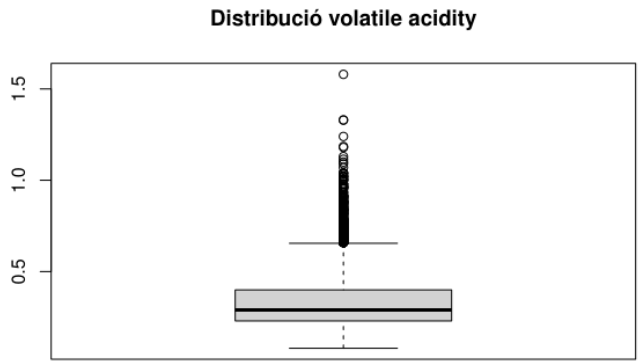
Per identificar els valors extrems, s'ha creat un boxplot per cada variable per poder observar la distribució i els valors extrems, veient també en algun cas la distribució amb l'histograma. En algun cas que s'ha considerat, s'han observat els valors numèrics dels valors extrems.

Hem vist que en gairebé totes les variables hi ha valors extrems (o atípics), i s'han gestionat els que han semblat més atípics i s'han considerat anòmals, així com els valors que no són possibles per alguna raó. Anem a veure-ho variable a variable, mostrant els gràfics de distribució generats (abans de l'eliminació dels outliers):

- Pel que fa a la variable "**fixed acidity**", s'observen molts valors atípics però no els considerem anòmals ja que no s'ha trobat cercant a Internet el màxim/mínim permesos d'acidesa fixe en els vins.

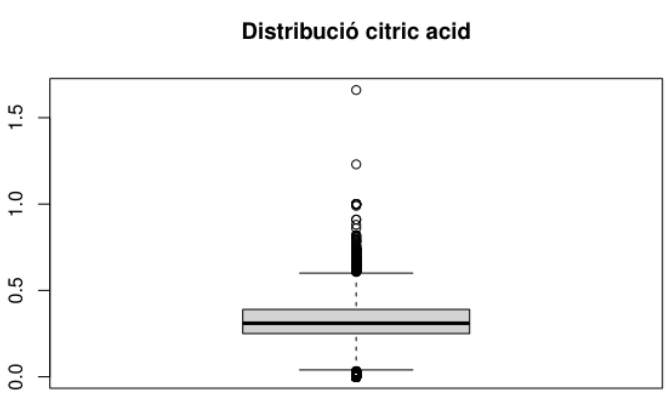
- Pel que fa a la variable "**volatile acidity**", s'ha trobat fent recerca que els vins blancs i rosats poden tenir com a màxim 1.5 g/l. Per tant, considerem anòmals els valors que superiors a 1.5, que només afecta a un registre que eliminem.

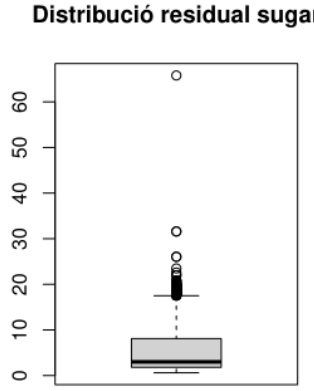


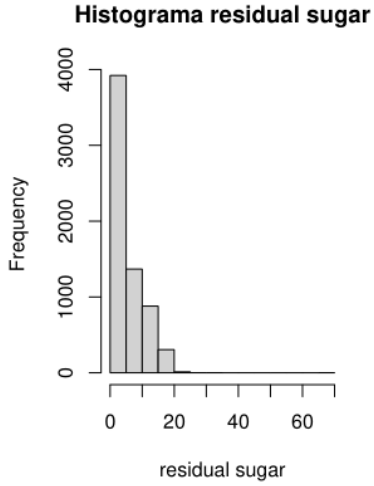


- Per la variable "**citric acid**", eliminem els valors atípics que són superiors a un valor de 1 gram per litre, ja que s'ha trobat en una cerca a Internet que la quantitat legal màxima d'àcid cítric en el vi és de 1 gram per litre, i per tant, s'han considerat anòmals els valors atípics a partir d'aquest valor. Dos registres han complert aquesta condició i hem eliminat les files corresponents, al ser un percentatge ínfim del total.

- Pel que fa a la variable "**residual sugar**", hem considerat anòmals els valors per sobre de 30 ja que estan força allunyats de la resta (malgrat s'ha trobat que el màxim de sucre residual permès és de 45 g/l). Hem eliminat 3 registres que complien aquesta condició.

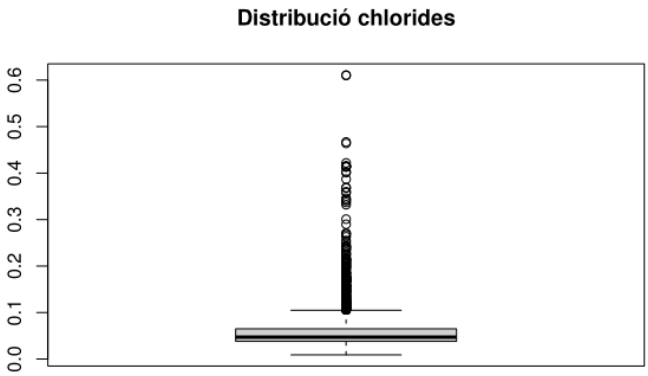


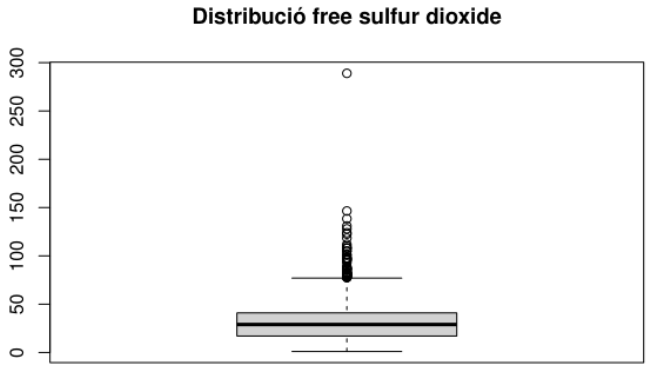




- Per la variable "**chlorides**", no s'han considerat anòmals cap dels valors atípics detectats. El valor màxim permès que s'ha trobat en diferents països ronda el 0.6, per tant no apuntem cap valor com anòmal.

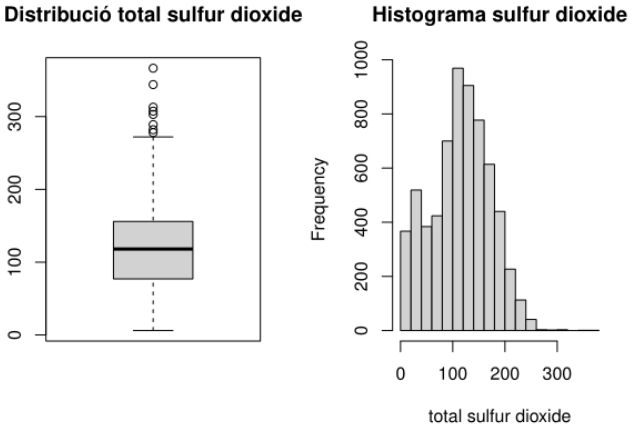
- Per la variable "**free sulfur dioxide**", malgrat veure que sembla que cap valor supera el límit màxim permès per la llei (de 350 aproximadament depenent del país), s'han considerat atípics els valors a partir del 150, eliminant una fila que complia la condició, ja que el valor es trobava molt allunyat de la resta.

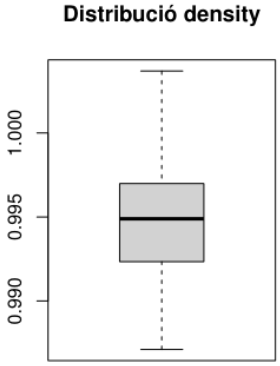


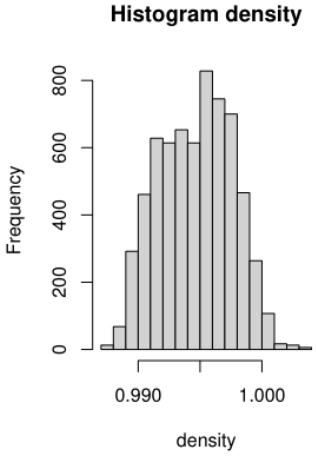


- Per la variable "**total sulfur dioxide**" s'han considerat anòmals els valors superiors a 350, ja que és el límit permès. S'ha eliminat una fila. Veiem en la imatge els valors numèrics dels outliers.

- Per la variable "**density**", no s'ha observat cap valor extrem.



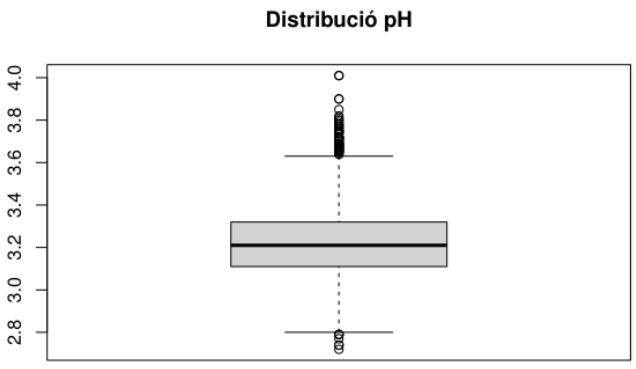


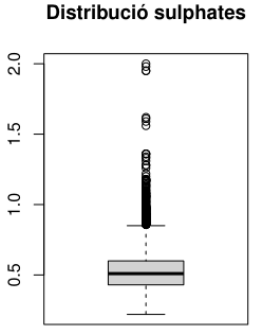


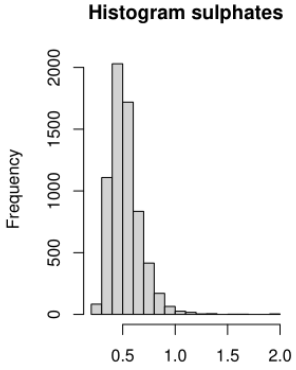

```
boxplot.stats(wine$total.sulfur.dioxide)$out
```

```
## [1] 278.0 289.0 313.0 366.5 307.5 344.0 282.0 303.0
```

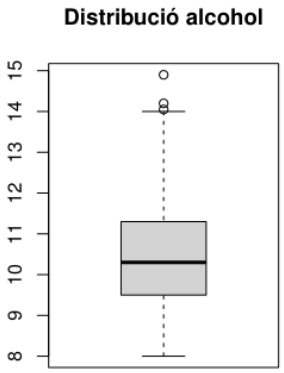
- Per la variable "**pH**" no s'han considerat anòmals cap dels valors atípics observats, al ser valors lògics dins l'escala del pH.
- Per la variable "**sulphates**", no s'han considerat anòmals cap dels valors extrems detectats, al trobar-se que és possible tenir fins a 2 grams per litre de sulfats; i cap valor supera aquest límit.

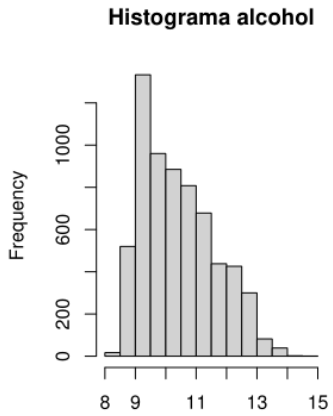


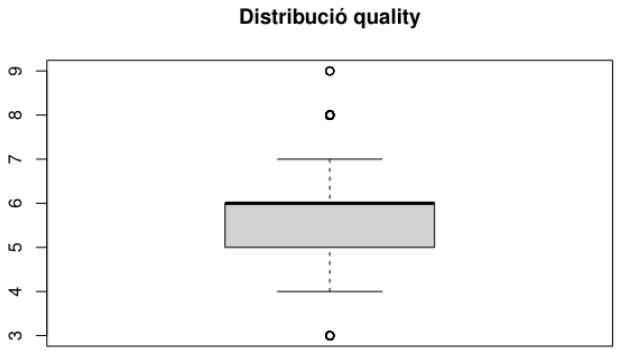




- Pel que fa a "**alcohol**", no es consideren anòmals cap dels 3 valors extrems observats, ja que el vi pot tenir aquestes quantitats d'alcohol.
- Pel que fa a "**quality**", malgrat observar valors extrems, es troben dins el rang possible de la variable (el rang de la variable va del 0 al 10). Els donem com a valors vàlids.







Cal dir que a l'observar els valors extrems i anar eliminant els registres amb valors considerats anòmals a mesura d'anar-los veient en cada variable, pot ser que en alguna variable no hagin aparegut valors extrems o valors que haguéssim considerat extrems perquè en eliminar files amb el criteri d'una altra variable, hagin estat eliminades files que tenien valors extrems en altres columnes.

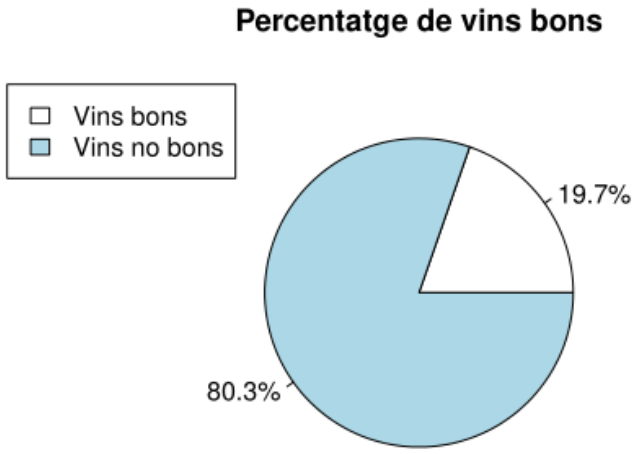
Ara el dataset s'ha reduït a 6489 files. Només hem eliminat 8 files del dataset per valors extrems anòmals, i per aquest motiu hem pogut eliminar els registres: 8 files és un percentatge molt petit del total.

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?).

Abans de començar amb l'anàlisi de dades volem discretitzar la variable qualitat del vi ("quality").

Per fer-ho, substituïm els valors numèrics per etiquetes (bo/no bo). Amb la variable nova podrem interpretar i comparar resultats. Classifiquem els 7 o superior com a "bo" i la resta com a "no bo".



Veiem que el 19.7% dels vins han estat classificats com a vins bons (puntuació 7 o més en qualitat) i el 80.3% com a vins no bons (puntuació inferior a 7).

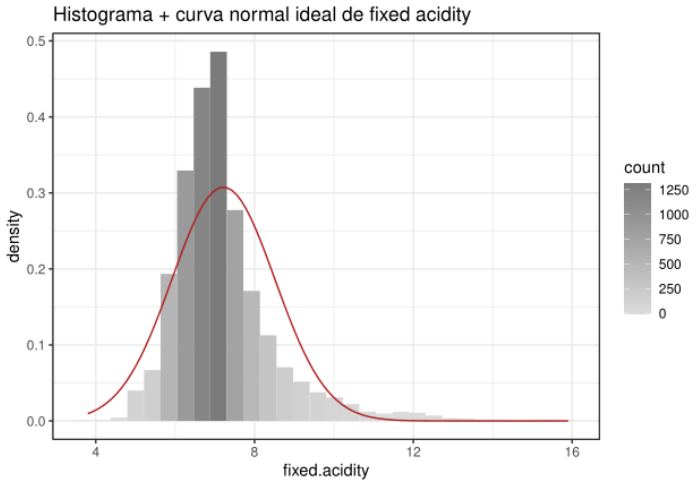
En l'anàlisi de dades ens interessarà comparar aquests grups de dades: els vins bons i els no bons. Volem comparar la diferència en les característiques entre els vins bons i els vins no bons, per veure si quines característiques fan que un vi sigui bo. Per fer-ho farem alguns test d'hipòtesis, una regressió logística i un arbre de decisió.

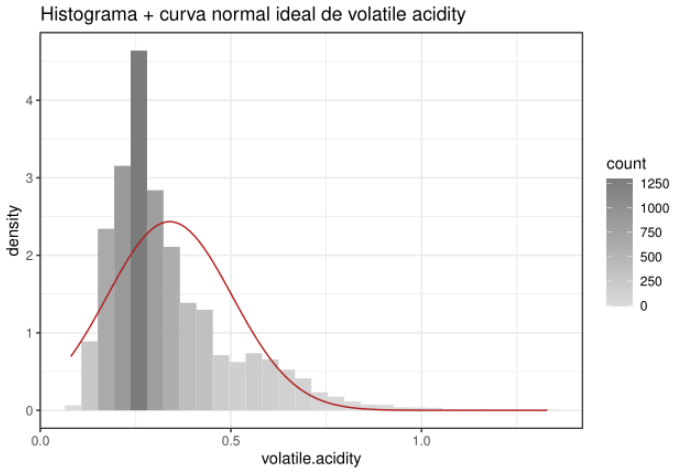
L'altra variable categòrica que tenim és "type", que indica si el vi és negre o blanc. També farem algunes comparacions entre aquests grups, estudiant si la qualitat del vi blanc i negre són percebudes diferents amb un test d'hipòtesis. També farem alguns anàlisis diferenciats segons el tipus de vi: farem una matriu de correlació per cada tipus de vi, i buscarem per cada tipus de vi el millor model de regressió lineal per explicar la qualitat amb les característiques del vi.

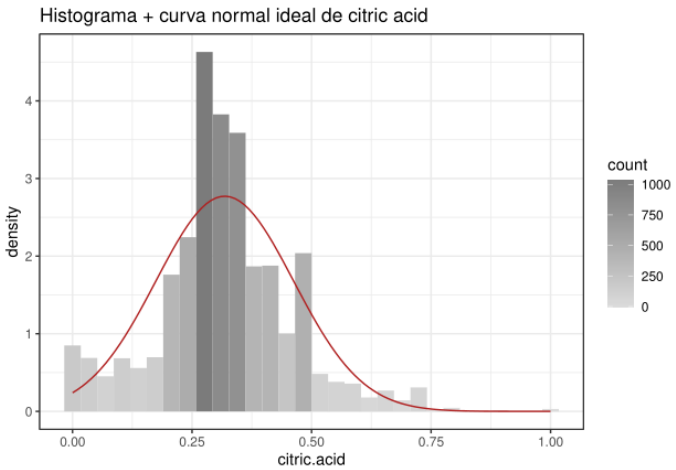
4.2. Comprovació de la normalitat i homogeneïtat de la variància.

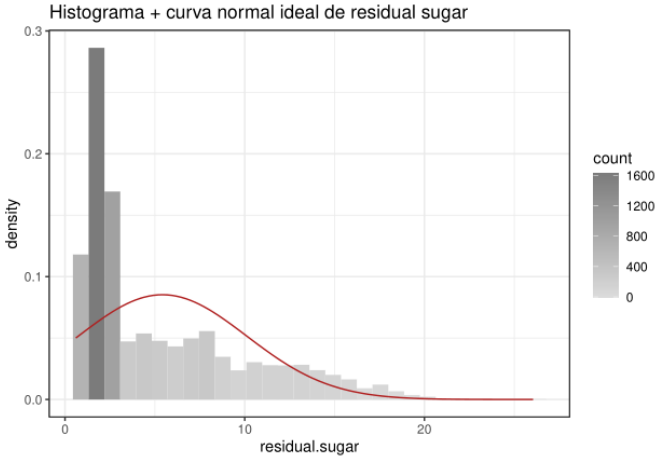
A l'estudiar la **normalitat** de cada variable amb el test Kolmogorov-Smirnov, s'ha vist que la normalitat de totes les variables (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol i quality), tenen un valor p inferior a 0.05, que ens porta a acceptar que les dades no provenen d'una distribució normal ja que rebutgem la hipòtesi nul·la, en totes les variables.

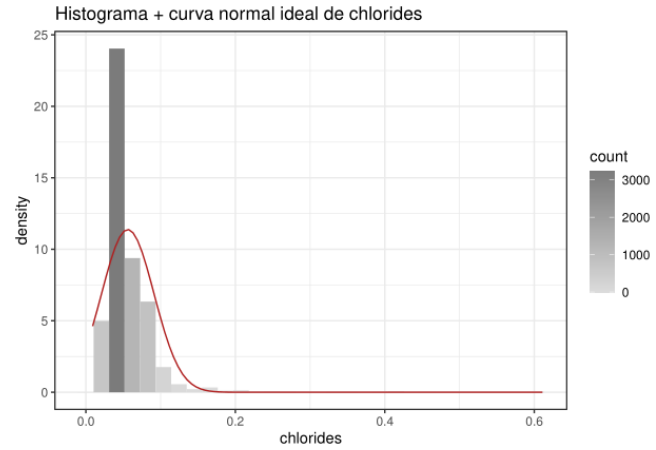
Observem els histogrames de cada variable comparant-ho amb la seva distribució normal ideal, per veure-ho de manera visual.

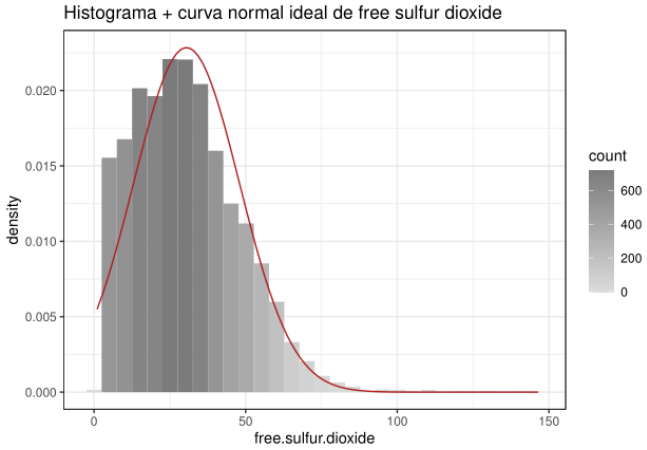


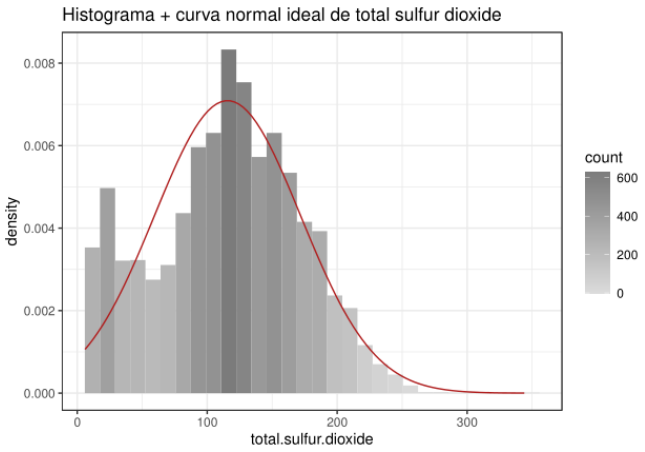


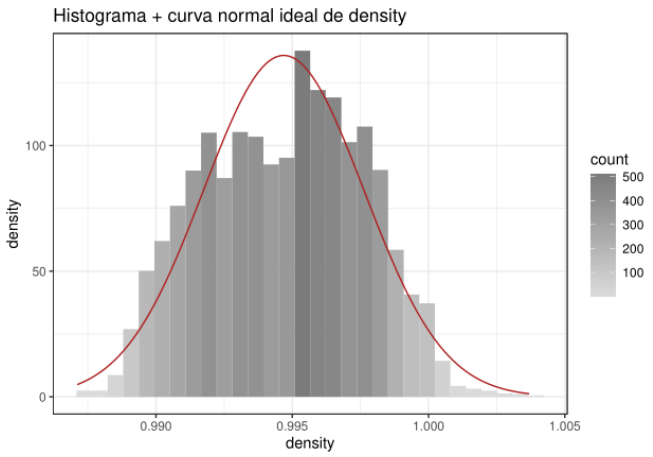


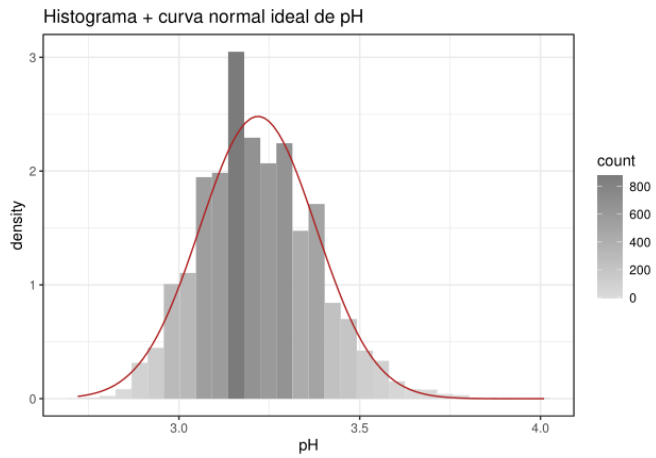


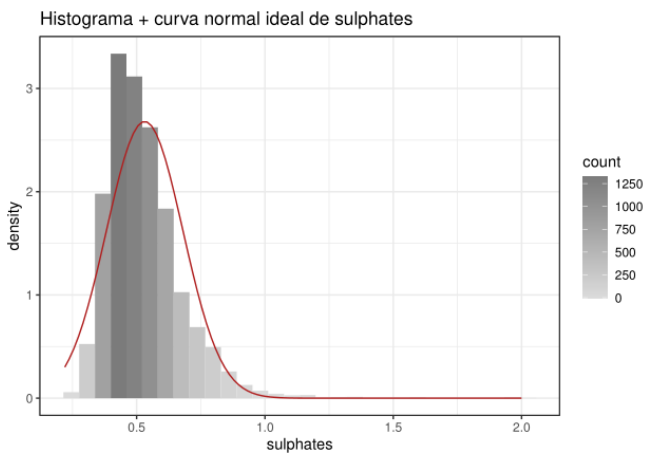

Histograma + curva normal ideal de chlorides
density
chlorides
count

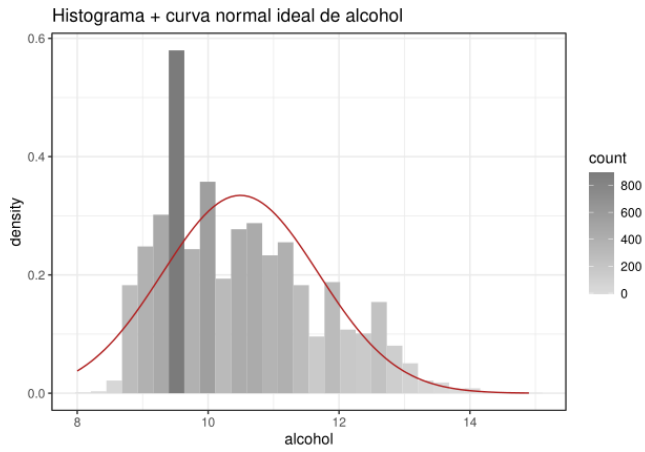

Histograma + curva normal ideal de free sulfur dioxide
density
free.sulfur.dioxide
count

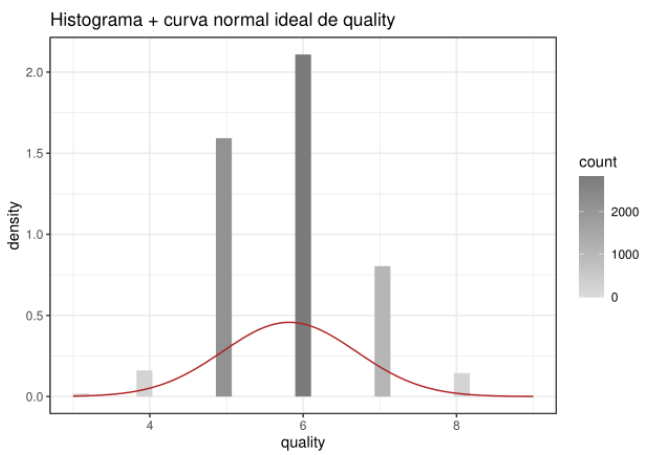

Histograma + curva normal ideal de total sulfur dioxide
density
total.sulfur.dioxide
count


Histograma + curva normal ideal de density
density
density
count


Histograma + curva normal ideal de pH
density
pH
count


Histograma + curva normal ideal de sulphates
density
sulphates
count


Histograma + curva normal ideal de alcohol
density
alcohol
count


Histograma + curva normal ideal de quality
density
quality
count

Pel que fa a la **homoscedasticitat**, hem contrastat la igualtat de variàncies entre grups que necessitarem saber posteriorment amb la prova de Levene. Per això hem contrastat la homogeneïtat de la variància entre els grups de vi bo i no bo en totes les variables, obtenint que entre ambdós grups les variàncies són diferents pel que fa a totes les variables excepte pel pH, on tenen variàncies iguals.

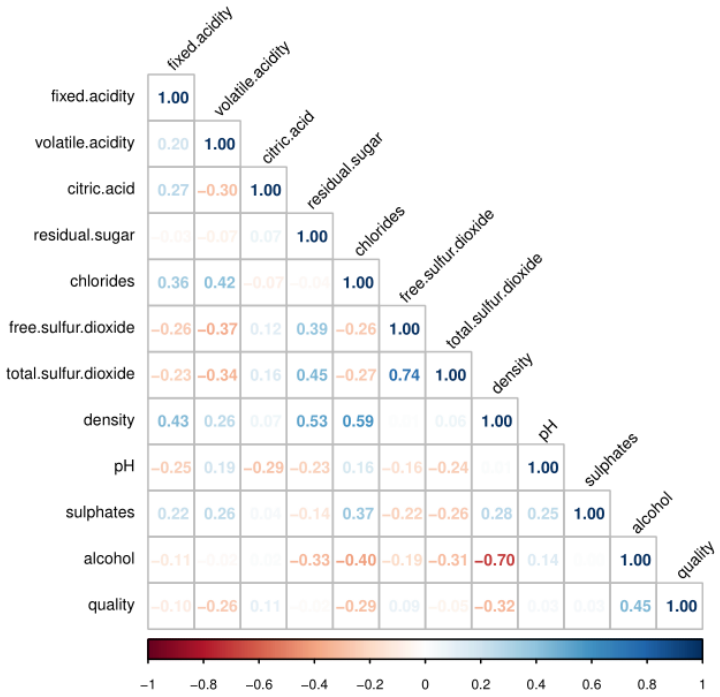
També s'ha contrastat la igualtat de variàncies entre el vi blanc i el negre pel que fa a la qualitat, i s'ha vist que les variàncies són iguals ($p>0.05$).

Tots els tests es poden veure al codi de la Pràctica.

4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

CORRELACIONS

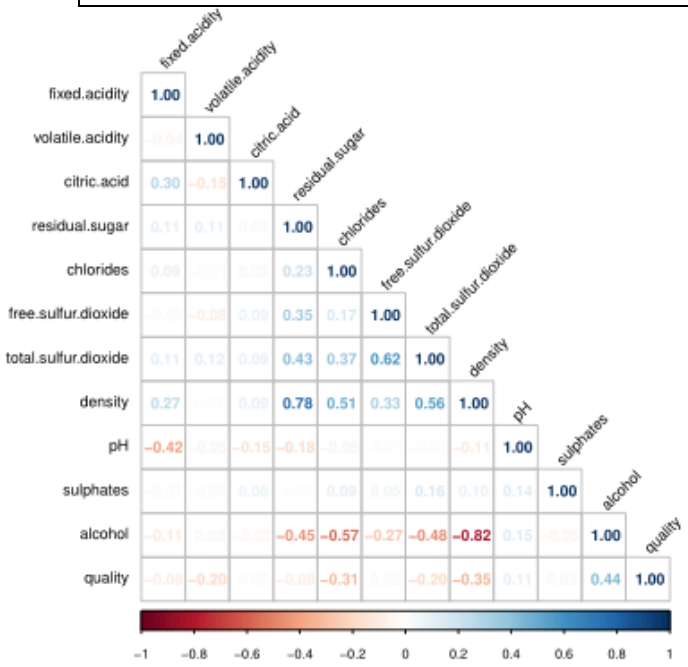
Comencem l'anàlisi realitzant una matriu de correlació entre les variables corresponents a les característiques i la variable quality. La correlació s'ha fet amb Spearman, ja que s'ha vist que cap variable segueix una distribució normal. Aquesta alternativa no paramètrica mesura el grau de dependència entre dues variables i no comporta cap suposició sobre la distribució de les dades.

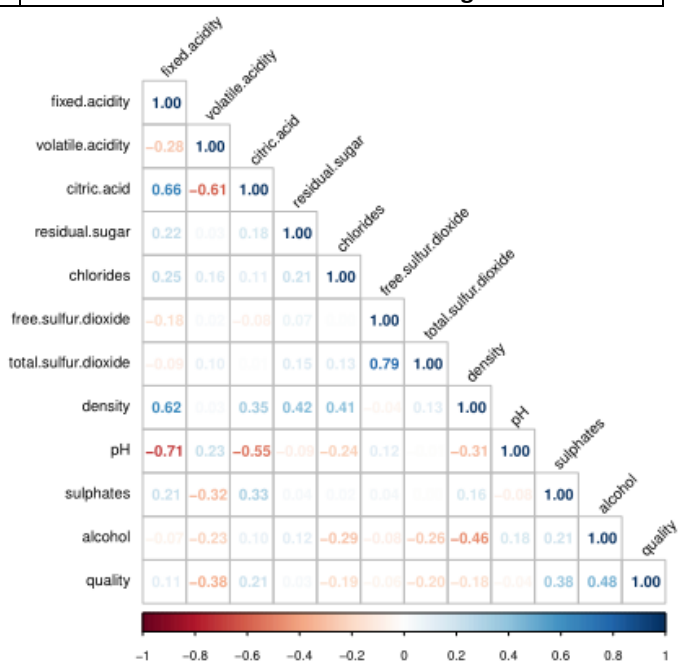


Atenent als coeficients de correlació (method = "spearman"), observem que les relacions lineals més destacables amb la qualitat són amb alcohol (0.45) i density (-0.32). També veiem una relació negativa moderada entre chlorides i la qualitat (-0.29) i entre la acidesa volàtil i la qualitat (-0.26). La resta són relacions dèbils/baixes. Les quatre correlacions anomenades són correlacions moderades: la primera implica que a mesura que en certa mesura, a mesura que augmenta l'alcohol, augmenta la qualitat percebuda del vi; la segona implica que a mesura que augmenta la densitat, disminueix la qualitat; la tercera que a mesura que augmenten els chlorides, disminueix en certa manera la qualitat; i la quarta que a mesura que augmenta l'acidesa volàtil, disminueix la qualitat.
Cal destacar també la relació entre les dues variables següents: la densitat (density) i l'alcohol (alcohol), amb un coeficient de -0.70. Aquesta forta correlació negativa ens pot fer pensar en una col·linealitat entre variables. En una cerca a Internet s'ha trobat que degut a la química, la quantitat d'alcohol redueix la densitat, i aquesta relació trobada podria afectar els anàlisis. La quantitat d'alcohol serà la millor opció com a predictor de la qualitat del vi. En anàlisis posteriors, possiblement no sigui útil introduir ambdós predictors als models, i s'haurà d'anar en compte amb la col·linealitat.
El SO2 lliure (free sulfur dioxide) i el SO2 total (free sulfur dioxide) també estan altament correlacionats entre si, com podíem esperar.

Ara fem la matriu de correlacions pels vins blancs i pels vins negres, per observar si aquestes correlacions es veuen accentuades en algun dels dos tipus de vins.

Matriu correlacions vi blanc	Matriu correlacions vi negre





Observem que en el vi blanc les correlacions relacionades amb la qualitat són molt similars, disminuint la força de la relació lineal de la "volatile acidity" amb la qualitat.
En el cas dels vins negres, la correlació de la densitat amb la qualitat és de -0.18 (molt dèbil), la relació de l'alcohol amb la qualitat augmenta una mica respecte el total de vins (0.48) i respecte el total augmenta la correlació entre la volatilitat de l'acidesa (volatile acidity) i la qualitat (-0.38), en el sentit que a més acidesa volàtil, menys qualitat.

CORRELACIONS DE LES VARIABLES AMB LA QUALITAT DEL VI

Variables	**TOTAL Vi**	**Vi blanc**	**Vi negre**
fixed acidity	-0.10	-0.08	0.11
volatile acidity	-0.26	-0.20	-0.38
citric acid	0.11	0.02	0.21
residual sugar	-0.02	-0.08	0.03
chlorides	-0.29	-0.31	-0.19
free sulfur dioxide	0.09	0.02	-0.06
total sulfur dioxide	-0.05	-0.20	-0.20
density	-0.32	-0.35	-0.18
pH	0.03	0.11	-0.04
sulphates	0.03	0.03	0.38
alcohol	0.45	0.44	0.48

En aquesta taula veiem les correlacions més notables amb la qualitat del vi, segons la seva modalitat (marcades com a destacables les superiors a 0.2).
Pel vi blanc les correlacions més notables són la correlació positiva entre l'alcohol i la qualitat, i les negatives de la densitat i la quantitat de sal en vi (chlorides) amb la qualitat. Pel que fa al vi negre, la correlació més forta amb la qualitat també és amb l'alcohol (correlació positiva), seguit de amb els sulfats (positiva), amb l'acidesa volàtil (negativa) i l'àcid cítric.

TEST D'HIPÒTESIS

Com que s'ha observat que sembla que l'alcohol és la variable més correlacionada amb la qualitat del vi (tant en els blancs com en els negres), estudiarem ara si la mitjana d'alcohol és diferent pels vins bons i pels vins no bons.

Veient la distribució sembla que els vins bons tenen més alcohol que els vins no bons.

Cal analitzar estadísticament si la mitjana d'alcohol és diferent pels vins bons i els vins no bons.

Farem un contrast d'hipòtesi per la diferència de mitjanes de alcohol.

La pregunta de recerca és: "La quantitat d'alcohol és diferent en els vins bons i els vins no bons?"

- La *hipòtesi nul·la (H0)* és que la mitjana d'alcohol és iguals entre vins bons i dolents.
- La *hipòtesi alternativa (H1)* és que la mitjana d'alcohol és diferents entre vins bons i dolents.

Pel teorema del límit central podem assumir normalitat. Com s'ha vist al punt 4.2 en la igualtat de variàncies, les variàncies són diferents amb un nivell de confiança del 95%. El test aplicat és un test de dues mostres independents sobre la mitjana amb variàncies desconegudes diferents.

```
t.test(vibo$alcohol, vinobo$alcohol, alternative = "two.sided", var.equal=FALSE, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  vibo$alcohol and vinobo$alcohol
## t = 31.624, df = 1787.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.100102 1.245577
## sample estimates:
## mean of x mean of y
##  11.43336  10.26052
```

Exemple del test d'igualtat de mitjanes amb variàncies desconegudes amb un nivell de confiança del 95%.

Rebutgem la hipòtesi nul·la del test d'igualtat de mitjanes (p-valor inferior a 0.05), afirmant que amb una confiança del 95%, la mitjana de alcohol és estadísticament diferent entre els vins bons i els no bons. Si veiem les mitjanes, veiem que **la mitjana d'alcohol és més alta en els vins bons (11.43 graus) que en els vins no bons (10.26 graus),** amb un nivell de confiança del 95%.

Realitzant test d'igualtat de mitjanes per la resta de variables, estudiant si la quantitat de cada variable és diferent entre els vins bons i els vins no bons s'ha trobat que totes les característiques excepte el "free sulfur dioxide" tenen quantitats estadísticament diferents en els vins bons respecte els vins no bons. S'ha pogut aplicar la prova tenint en compte el teorema central del límit, considerant que segueixen una distribució normal al ser mides de les mostres grans, i apuntant que l'homogeneïtat de la variància és diferents entre grup, excepte en el pH. S'ha fet com en l'exemple de la variable alcohol mostrat.

La quantitat de sal (chlorides) és, en mitjana, més baixa en els vins bons (0.045) que en els dolents (0.059). L'acidesa fixe és, en mitjana, més baixa en els vins bons (7.09) que en els no bons (7.25). L'acidesa volàtil és més baixa en els vins bons (0.29) que en els no bons (0.35). L'àcid cítric és significativament més alt en

els vins bons (0.33) que en els no bons (0.31). El sucre residual és més baix en els vins bons (4.83) que en els no bons (5.57). El "total sulfur dioxide" és més baix en els vins bons (109.89) respecte els no bons (117.05). La densitat és més baixa en els vins bons (0.993) que en els no bons (0.995). El pH és més alt en els vins bons (3.227) que en els no bons (3.216). Els sulfats són més alts en els vins bons (0.541) que en els no bons (0.529).

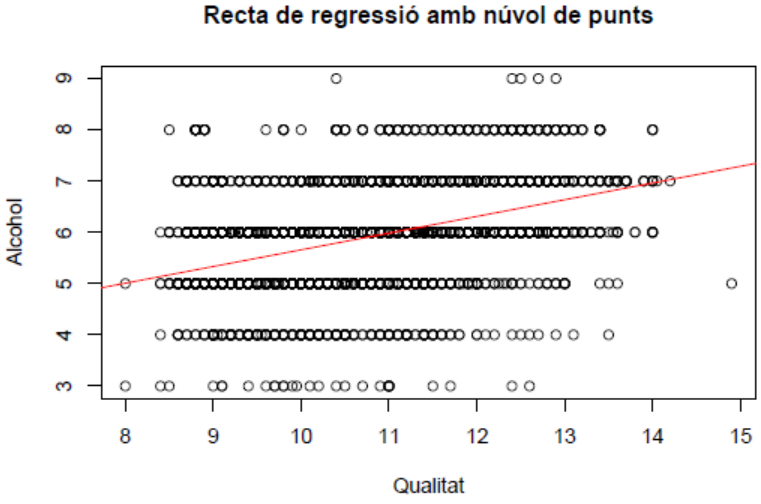
Tots els resultats són coincidents amb la direcció de la correlació, per molt petita que sigui. Malgrat la correlació lineal entre cada variable i la qualitat és diferent al test d'hipòtesis per veure si els vins bons una quantitat diferent de cada variable que els dolents.
En les correlacions s'ha vist que no n'hi ha gaires de "fortes" amb la qualitat, i en canvi hem observat que totes les variables excepte el "free sulfur dioxide" tenen valors estadísticament diferents en els vins bons i els vins dolents, i per tant, un valor més alt o més baix en aquestes variables podria influir en la qualitat (malgrat la correlació lineal sigui molt dèbil).

Més per curiositat que per resoldre preguntes, ens interessa saber si la qualitat és la mateixa en els vins blancs i els vins negres. Per comprovar-ho, comparem les mitjanes de qualitat entre ambdós tipus de vins. Apliquem el teorema central del límit i considerem que les dades segueixen una distribució normal al tenir mides de les mostres grans (n>30). Pel que fa a l'homoscedasticitat, a l'apartat 4.2 hem vist que les variàncies entre grups (tipus de vi) són iguals pel que fa a la variable qualitat.

De la prova es conclou que existeixen diferències estadísticament significatives entre el vi blanc i el negre en la qualitat percebuda. El vi blanc es percep amb una millor qualitat (5.88 vs 5.64).

REGRESSIÓ

Volem veure amb una regressió lineal simple la relació entre la variable explicativa alcohol i la variable qualitat (variable resposta). El model de regressió resulta significatiu: com havíem vist, a mesura que augmenta l'alcohol, augmenta la qualitat percebuda del vi.



El coeficient de determinació apunta que la variable alcohol és capaç d'explicar el 19.9% de la variabilitat present en la variable de resposta (quality).

S'ha intentat buscar el millor model de regressió lineal múltiple per trobar les variables que millor expliquen la qualitat. Per fer-ho, primerament s'ha observat el VIF per tal d'observar si existeix col·linealitat entre variables i s'ha trobat un valor molt alt per la variable "density" (VIF=17.4), que com ja hem vist presentava correlacions molt altes, sobretot amb "alcohol". Hem eliminat, doncs, aquesta variable del model ja que exhibeix col·linealitat i hem fet servir la tècnica de "stepwise mixte" i el *Akaike (AIC)* per seleccionar els millors predictors. S'han exclòs del model les variables "*citric acid*" i "*fixed acidity*" i obtenim un model on totes les variables influeixen significativament, amb uns valors pel VIF correctes i explicant el 29.08% de la variabilitat de la qualitat. Abans d'eliminar les tres variables eliminades

s'explicava el 29.34% de la variabilitat R^2; per tant les variables excloses no ajudaven en gran mesura a explicar la variabilitat de la variable qualitat.
Així doncs, el millor model de regressió per explicar la qualitat del vi està format amb les variables:

- alcohol
- volatile acidity
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- pH
- sulphates

Al comparar fer models de regressió lineal en els vins blancs i en els vins negres, hem vist que hi ha diferents variables que prediuen millor la qualitat del vi segons el tipus.

En el vi blanc, segons el model generat les variables que millor prediuen la qualitat del vi són: l'alcohol, l'acidesa volàtil, el sucre residual, el "free sulfur dioxide", la densitat, el pH, els sulfats i l'acidesa fixe.
En canvi, en el vi negre les variables que millors prediuen la qualitat del vi són: l'alcohol, l'acidesa volàtil, els sulfats, el "total sulfur dioxide", la quantitat de sal, el pH i el "free sulfur dioxide", en aquest ordre.

Segons el model, la qualitat del vi blanc és ben predita per varies variables, tres de les quals no prediuen bé la qualitat del vi negre: el sucre residual, la densitat i l'acidesa fixe. En canvi, la quantitat de sal i el "total sulfur dioxide" prediuen bé la qualitat del vi negre però no del blanc. L'alcohol, l'acidesa volàtil, el "free sulfur dioxide", el pH i els sulfats ajuden a predir la qualitat dels vins blancs i negres segons els millors models de predicció amb regressió lineal.

Cal apuntar que el model pel vi negre explica un percentatge més elevat de la variabilitat de la qualitat del vi (35.7%) que el del vi blanc (28.9%).

S'ha generat un arbre de decisions per predir si el vi és bo o no mitjançant les variables predictores. El model presenta una precisió del 89.49%, i ens indica que les variables més usades pel model per predir si el vi és bo o no són: l'alcohol, l'acidesa volàtil i el sucre residual.

Al fer un arbre de decisió per cada tipus de vi, veiem que les variables que més s'utilitzen per predir si el vi és bo o no també són diferents entre el vi blanc i el vi negre. Les que s'usen més pel model del vi blanc (amb una precisió en el conjunt d'entrenament de 89.5%) són: alcohol, pH i l'acidesa volàtil; i les que més s'usen pel model del vi negre (amb una precisió del 94.8%) són l'alcohol i l'acidesa fixe.

La regressió logística mostra que totes les variables excepte l'àcid cítric són significatives per predir la probabilitat que un vi sigui bo o no.

Veiem la importància de les variables en el model de regressió logística un cop eliminada la variable "*cítric acid*".

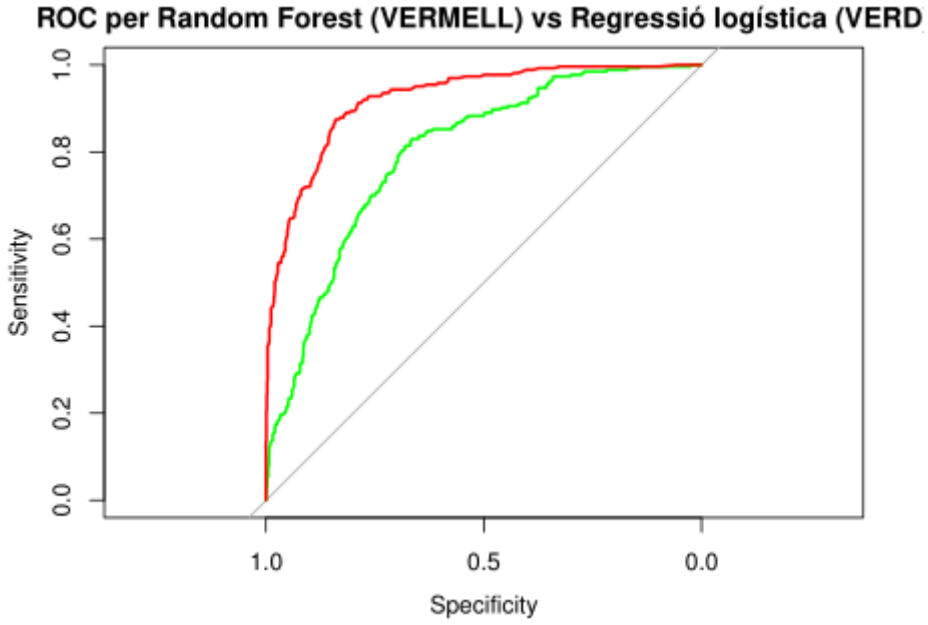
```
##                       Overall
## alcohol              8.358627
## volatile.acidity     9.232610
## sulphates            8.672380
## residual.sugar       7.906811
## free.sulfur.dioxide  4.569190
## density              5.621184
## total.sulfur.dioxide 5.163399
## chlorides            2.697436
## pH                   6.785464
## fixed.acidity        6.789324
```

L'acidesa volàtil, els sulfats i l'alcohol són les variables amb més importància pel model de regressió logística per predir la probabilitat de que un vi sigui bo o no.

Hem fet un anàlisis extra per veure les corbes ROC del model de regressió logística i d'un model d'arbre de decisió "Random forest". Ens permet observar la sensibilitat enfront la especificitat en cada model. Les dades són partides en conjunts d'entrenament (train, 80%) i prova (test, 20%) i es construeixen els dos models que usaran el conjunt d'entrenament per aprendre i llavors fer prediccions al conjunt de test.

El model de regressió logística es construeix com el millor model de regressió logística trobat (amb totes les variables excepte l'àcid cítric, al no ser significatiu), i el de Random Forest amb totes les variables.

Observem les corbes ROC:



L'àrea sota la corba pel model de Random Forest està al voltant de 0.92 i pel model de regressió logística al voltant de 0.80. Els models prediuen bastant bé, mitjançant les característiques del vi, si aquest serà bo o no. El model de "Random Forest" discrimina millor les dades, al ser la seva àrea sota la corba més gran que la del model de regressió logística.

Veiem ara les variables que són més importants en una execució exemple dels models. Com que el conjunt d'entrenament i de test són escollits a l'atzar, en cada execució obtenim dades diferents.

Importància variab. model lineal

```
##                       Overall
## alcohol              7.205251
## volatile.acidity     8.039250
## sulphates            7.869940
## residual.sugar       7.745661
## free.sulfur.dioxide  3.885392
## density              5.613084
## total.sulfur.dioxide 4.378804
## chlorides            1.890850
## pH                   6.622944
## fixed.acidity        6.624248
```

Importància variab. regr. logística

```
##                       Overall
## alcohol              270.7610
## volatile.acidity     145.3333
## sulphates            133.3384
## residual.sugar       139.2924
## free.sulfur.dioxide  125.7011
## density              193.5416
## total.sulfur.dioxide 133.0494
## chlorides            141.6013
## pH                   130.1192
## fixed.acidity        111.6224
## citric.acid          124.7977
```

En l'exemple d'execució, les variables més importants pel model de regressió logística són l'acidesa volàtil, els sulfats, el sucre residual i l'alcohol, sent "chlorides" la menys important. En l'arbre de decisió, ho són l'alcohol i la densitat, sent les menys importants l'acidesa fixe i l'acid citric.

S'ha observat amb varies execucions que les variables més importants es mantenen estables.

5. Representació dels resultats a partir de taules i gràfiques. Aquest apartat es pot respondre al llarg de la pràctica, sense la necessitat de concentrar totes les representacions en aquest punt de la pràctica.

S'han anat incloent taules i gràfiques al llarg de la pràctica.

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

Les variables més correlacionades amb la qualitat són l'alcohol, els chlorides i l'acidesa volàtil. La densitat també té una correlació alta però presenta molta correlació amb l'alcohol i s'ha trobat col·linealitat observant el FIV. A més alcohol es perceb més qualitat, a més chlorides se'n percep menys i a més acidesa volàtil menys.
Pel que fa al vi blanc, trobem les mateixes relacions amb la qualitat, malgrat la correlació amb l'acidesa volàtil és més baixa. Pel que fa al vi negre, l'acidesa volàtil es troba força correlacionada de manera negativa amb la qualitat, els sulfats tenen una correlació positiva amb la qualitat, i l'alcohol també hi té una correlació positiva.
Pel que fa a les correlacions ja veiem que les variables relacionades linealment amb la qualitat són diferents entre els vins blancs i negre: en els blancs l'alcohol i la quantitat de sal són les que tenen més relació, i en els negres, l'alcohol, els sulfats i l'acidesa volàtil.

A l'estudiar en més profunditat la relació entre l'alcohol i ser un vi bo o no, s'ha observat que els vins considerats bons tenen, de mitjana, més graus d'alcohol que els vins que no són bons. En estudiar la igualtat de mitjanes de cada variable pels vins bons i negres per cada variable, s'ha vist que totes les variables excepte el "free sulfur dioxide" tenen valors estadísticament diferents en els vins bons i els vins dolents; el que significa que un valor més alt o més baix en aquestes variables podria influir en la qualitat (malgrat la correlació lineal amb la qualitat sigui, en algunes, molt dèbil).

Els vins bons tenen menys quantitat de sal, menys acidesa fixe, acidesa volàtil, sucre residual, "total sulfur dioxide" i densitat; i tenen més alcohol, més àcid cítric, "free sulfur dioxide", pH i sulfats.

Com a dada curiosa, s'ha determinat que els vins blancs són percebuts com de més qualitat que els vins negres; malgrat la diferència no sigui molta (valor de 5.64 de mitjana de qualitat en el els vins blancs i 5.88 en vins negres).

Segons el millor model de regressió lineal generat, totes les variables explicatives excepte l'àcid cítric, l'acidesa fixe i la densitat ajuden significativament a predir la qualitat d'un vi.
Al comparar entre els vins blancs i negre, hem vist que hi ha diferents variables que prediuen millor la qualitat del vi (segons els models generats).

En el vi blanc, les variables que millor prediuen la qualitat del vi són segons la regressió lineal:

- Alcohol
- Acidesa volàtil
- Sucre residual
- Free sulfur dioxide
- Densitat
- pH
- Sulfats
- Acidesa fixe

En canvi, en el vi negre les variables que millors prediuen la qualitat del vi són segons la regressió lineal:

- Alcohol
- Acidesa volàtil
- Free sulfur dioxide
- pH
- Sulfats
- Total sulfur dioxide
- Quantitat de sal

En l'arbre de decisions generat per predir si un vi és bo o no (variable dicotòmica) indica que les variables que més importants en l'arbre de precisió 89.5 són l'alcohol, l'acidesa volàtil i el sucre residual.

Comparant els arbres de decisions pel vi blanc i el vi negre s'observa com les variables amb més importància a l'hora de predir si el vi és bo o no, són diferents entre el tipus de vi. Alcohol, pH i acidesa volàtil són les més usades en el model pel vi blanc i alcohol i acidesa fixe són les més usades en el model pel vi negre.

La regressió logística per predir la probabilitat que un vi sigui bo o no, indica que totes les variables excepte l'àcid cítric ajuden significativament a predir la predicció.

Variables més importants en cada anàlisis (més correlacionades o de més importància en cada model):

Variables	**Correlació** Amb la variable qualitat	**Test d'hipòtesis** (mitjana variable diferent en vins bons I no bons)	**Regressió lineal** (Variables del millor model per predir la qualitat del vi)	**Arbre de decisions** (Variables més rellevants)	**Regressió logística** (Variables més importants)	**Random Forest** (Variables més importants)
fixed acidity						
volatile acidity						
citric acid						
residual sugar						
chlorides						
free sulfur dioxide						
total sulfur dioxide						
density	Colinealitat					
pH						
sulphates						
alcohol						

Tornem a la pregunta/problema que es pretenia resoldre: les característiques (propietats fisicoquímiques) més relacionades amb que un vi sigui bo. L'alcohol és la característica que en tots els anàlisis s'observa com la més relacionada amb la qualitat del vi, i més determinant per a que un vi sigui bo. L'acidesa volàtil sembla ser la segona variable més relacionada amb la qualitat del vi i més determinant per a que un vi sigui bo.
La quantitat de sal (chlorides) sembla tenir correlació amb la qualitat i ajuda a predir la qualitat del vi, però no sembla que sigui determinant per predir si un vi és bo o no.

La característica menys relacionada amb que un vi sigui bo o no i la seva qualitat semblen ser l'àcid cítric, l'acidesa fixe i el "free sulfur dioxide". La primera té més quantitat en els vins bons que en els dolents i la segona en té menys de manera significativa, però no són significatius per predir la qualitat del vi ni per determinar si un vi és bo o no. Les variables "total sulfur dioxide" i el "pH" tampoc determinen si un vi és bo o no.

Comparant les variables que expliquen la qualitat del vi segons el tipus de vi, hem vist que segons els models generats, les variables predictores són diferents segons si el vi és blanc o és negre.

Les variables més relacionades linealment amb el vi blanc semblen ser l'alcohol, la densitat i els chlorides. Les que millor prediuen la qualitat són l'alcohol, l'acidesa volàtil, el sucre resiudal, el free sulfur dioxide, la densitat, el pH, els sulfats i l'acidesa fixe. Les que millors prediuen si un vi és bo o no són l'alcohol, el pH i l'acidesa volàtil.

Pel que fa al vi negre, les variables alcohol, acidesa volàtil, acid cítric i sulfats són les més correlacionades amb la qualitat del vi negre. Les que millor prediuen la qualitat són l'alcohol, l'acidesa volàtil, el free sulfur dioxide, el pH, els sulfats, el total sulfur dioxide i els chlorides. L'alcohol i l'acidesa fixe són les millors per predir si un vi és bo o no.

Així doncs, les propietats fisicoquímiques que determinen la qualitat del vi són diferents en el vi blanc i el vi negre, malgrat hi ha propietats comunes rellevants per ambdues tipologies de vins.

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

S'ha adjuntat el codi en R al github, amb el nom de **codi/Prac2 - Qualitat del vi.Rmd,** amb el que s'ha realitzat la neteja, anàlisi i representació de les dades

TAULA DE CONTRIBUCIONS

Contribucions	Signatura
Investigació prèvia	FPH, ASR
Redacció de les respostes	FPH, ASR
Desenvolupament del codi	FPH, ASR

Recursos

1. Calvo, M., Pérez, D., Subirats, L. (2019). Introducció a la neteja i anàlisi de dades. Editorial UOC.