Tipologia i Cicle de vida de les dades

Pràctica 2 – Neteja i anàlisi de les dades

Autor: Héctor Gutiérrez Muñoz

Estudi dels vins portuguesos *Vinho Verde* segons les seves característiques físico-químiques

Resum

En aquesta pràctica farem un estudi de diverses variants dels vins portuguesos *Vinho Verde*. La principal pregunta que s'intentarà respondre és si es pot predir la qualitat d'un vi en funció de les seves característiques físico-químiques, però també farem altres anàlisis que ens permetin comprovar si certes coses que normalment se'n diuen dels vins, com que els vins negres tenen més alcohol que els de blanc són veritat o només una llegenda urbana.

L'anàlisi que fem aquí és força interessant, sobretot perquè si podem concloure que hi ha algunes característiques físico-químiques que influeixen significativament en la qualitat d'un vi, els productors podrien fer servir aquesta informació per millorar els seus vins.

Tot el treball fet s'hi pot trobar de forma oberta a GitHub, al següent repositori:

`https://github.com/hectorgut47/tipologia-prac2`

Descripció dels datasets

Per assolir els objectius plantejat, es faran servir dos datasets. Tots dos es poden trobar a: https://www.kaggle.com/danielpanizzo/wine-quality i van estar recollits per un grup d'investigació de la Universitat do Minho portuguesa, els quals van publicar els seus resultats a:

- Cortez, P., et. al. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547-553. Recuperat a: https://doi.org/10.1016/j.dss.2009.05.016

Les dades estan estan sota una llicència oberta i són unes dades força conegudes al món del *data science* pel seu interès i les possibilitats d'anàlisi que ofereixen.

Com ja s'ha dit, les dades estan a dos datasets: un amb els vins negres i d'altre amb els de blancs. Tots dos tenen la mateixa estructura i les mateixes columnes, però hi ha una diferència gran al nombre segons el color: 1599 negres i 4898 blancs. Les columnes dels dos datasets són:

- ***(id)***: autoincremental (int).
- **fixed.acidity**: acidesa que persisteix al vi (decimal).
- **volatile.acidity**: acidesa volàtil al vi, donada pel àcid acètic, que pot resultar en un sabor a vinagre (decimal).
- **citric.acid**: quantitat d'àcid cítric, que dona frescor al vi (decimal).
- **residual.sugar**: quantitat de sucre que resta al vi després de la fermentació (decimal).
- **chlorides**: quantitat de clorurs al vi (decimal).
- **free.sulfur.dioxide**: quantitat de sulfurs en forma lliure que prevé l'oxidació (decimal).
- **total.sulfur.dioxide**: quantitat total de sulfurs al vi que pot resultar en una afectació al sabor (decimal).
- **density**: densitat del vi, depèn del percentatge d'alcohol i el contingut en sucre (decimal).
- **pH**: pH del vi, normalment entre 3 i 4 (decimal).
- **sulphates**: quantitat de sulfats al vi, afegits com a additiu per prevenir microbis (decimal).
- **alcohol**: percentatge d'alcohol al vi (decimal).
- **qualitat**: puntuació del 0 al 10 del vi, donada per un jurat d'experts (int).

Integració i selecció de les dades

D'aquí endavant, es treballarà amb R. El codi desenvolupat amb breus explicacions es pot trobar a:

`https://github.com/hectorgut47/tipologia-prac2/raw/main/src/tipolpr2.pdf`

En aquest cas, cal començar per una etapa de integració dels dos datasets que tenim, ja que a l'origen els vins negres i els de blancs estan separats en dos fitxers. Per diferenciar-los, crearem una nova variable que indiqui el color: "blanc" o "negre". Finalment, eliminarem el ID autoincremental de tots dos datasets ja que no aporta cap informació rellevant i conservarem la resta de variables, tot i que després haurem d'estudiar si realment totes són rellevants per predir la qualitat dels vins. A la imatge es poden trobar uns exemples del dataset amb el qual treballarem:

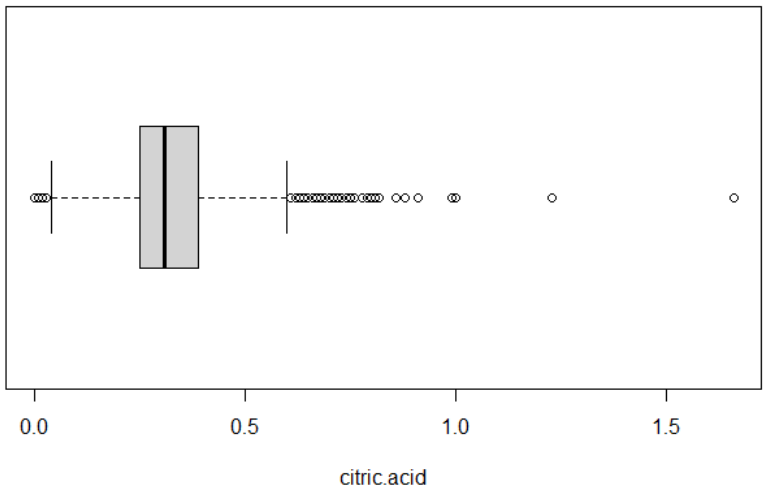
fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality	color
7.0	0.27	0.36	20.7	0.045	45	170	1.0010	3.00	0.45	8.8	6	blanc
6.3	0.30	0.34	1.6	0.049	14	132	0.9940	3.30	0.49	9.5	6	blanc
8.1	0.28	0.40	6.9	0.050	30	97	0.9951	3.26	0.44	10.1	6	blanc
7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	negre
7.8	0.88	0.00	2.6	0.098	25	67	0.9968	3.20	0.68	9.8	5	negre
7.8	0.76	0.04	2.3	0.092	15	54	0.9970	3.26	0.65	9.8	5	negre

Neteja de les dades

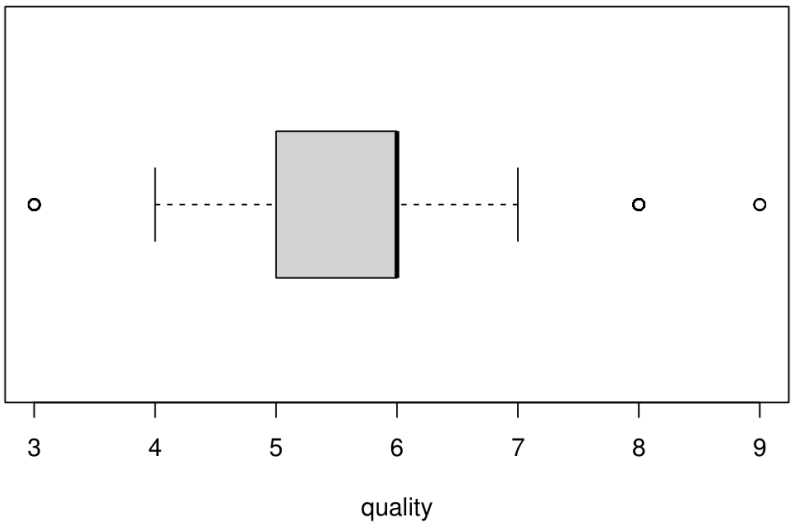
[El codi referent a aquest punt de la pràctica es pot trobar al mateix document RMarkdown indicat abans.]

Pel que fa a la neteja de les dades no cal fer res en el cas d'aquest dataset. Les dades ja estan tractades i sense valors desconeguts o zeros (a la variable del àcid cítric sí que hi ha zeros, però són observacions vàlides).

Per tant, només ens queda tractar els valors extrems, que sí que hi són presents, tant a les variables explicatives com a la qualitat. Però amb els diagrames de caixa i bigotis es pot apreciar que aquests outliers no estan molt allunyats de la resta de dades, que estan a dins del rang esperat per a l'observació i que per tant, són dades vàlides que hem de tractar com la resta i que no les hem de treure. Això es pot veure millor amb exemples: en primer lloc, presentem el diagrama de caixa i bigotis de una variable explicativa, com ara el àcid cítric.



Es pot veure com els valors extrems no estan allunyats de la resta de valors, i tots formen un rang en el qual les observacions són vàlides. Això es pot veure encara millor amb la variable que volem predir, la qualitat:



És clar que no té sentit treure'n la informació dels vins puntuats amb 3, 8 i 9, ja que són puntuacions completament vàlides que haurem de tenir en compte si fem un model o qualsevol tipus d'anàlisi.

Així doncs, podem concloure que les dades ja estan preparats per passar a la següent fase d'anàlisi sense haver de fer cap feina extra.

Anàlisi de les dades i representacions visuals

[El codi referent a aquest punt de la pràctica es pot trobar al mateix document RMarkdown indicat abans.]

En aquest punt fem els anàlisis estadístics que permeten donar respostes a les preguntes que hem comentat a l'inici de la pràctica. En aquesta memòria només recollim un breu comentari i les representacions gràfiques adients, ja que al document RMarkdown es troben totes les explicacions i càlculs fet i no cal repetir-les aquí.

Consideracions prèvies

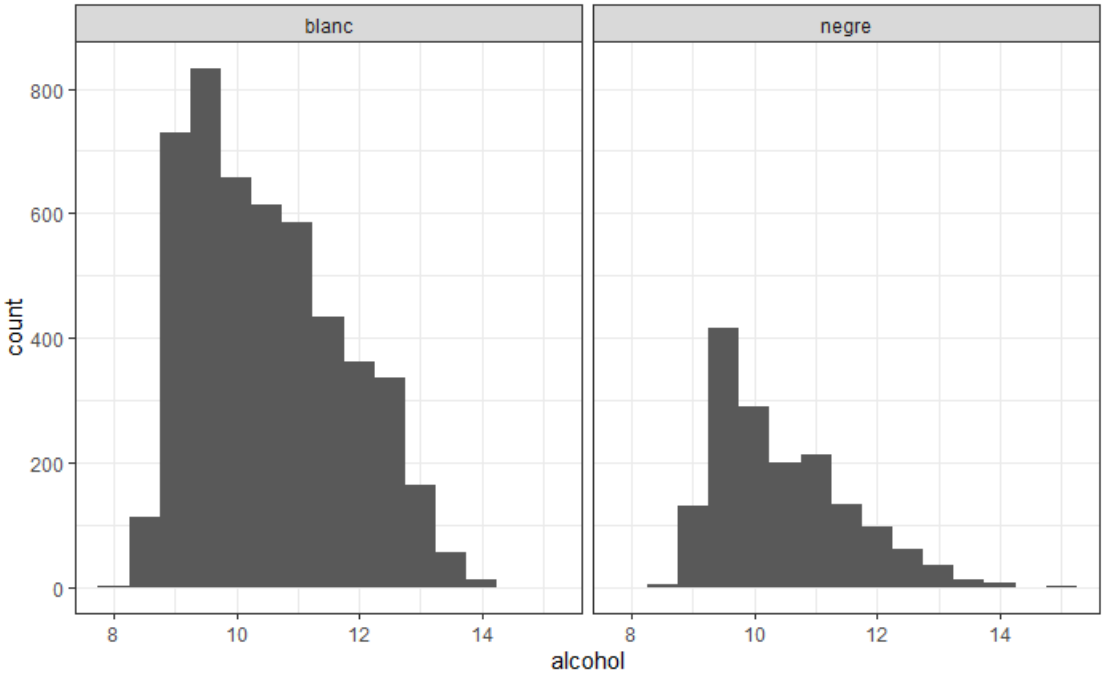
Com que les proves que farem després comparen vins blancs i negres, en primer lloc s'ha de separar el dataset en dos, segons el color del vi. Un cop fet això, cal comprovar la normalitat i l'homogeneïtat de les variàncies pel que fa a l'alcohol, ja que més endavant serà rellevant en l'elecció del test. Pel que fa a la normalitat, es pot suposar en les dues mostres, ja que són suficientment grans, però no podem dir el mateix de l'homogeneïtat de les variàncies: ens trobem a una situació d'heteroscedasticitat. Aquest resultat se dona per qualsevol nivell de confiança, ja que el p-valor corresponent és gairebé 0:

```
##
##  F test to compare two variances
##
## data:  vins_blancs_data[, "alcohol"] and vins_negres_data[, "alcohol"]
## F = 1.3335, num df = 4897, denom df = 1598, p-value = 5.947e-12
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.230150 1.443239
## sample estimates:
## ratio of variances
##           1.333536
```

Comparació del percentatge d'alcohol entre els vins blancs i els vins negres

Hi ha una creença molt estesa: els vins negres tenen més alcohol que els vins blancs, la qual cosa s'associa a que són més forts, tenen més anys de fermentació, etc. Veurem si aquesta creença és veritat o no amb un test d'igualtat de mitjanes, però tenint en compte

Amb la representació gràfica dels graus d'alcohol dels dos tipus de vi, es pot veure que tenim una situació molt semblant tant als negres com als blancs, potser l'única excepció són uns vins negres amb un percentatge d'alcohol molt alt (gairebé 15%):



El test d'igualtat de mitjanes sense homogeneïtat de variàncies ens confirma això que acabem de dir: no hi ha diferència significativa entre la mitjana del percentatge d'alcohol dels dos tipus de vins, amb un p-valor de gairebé 1.

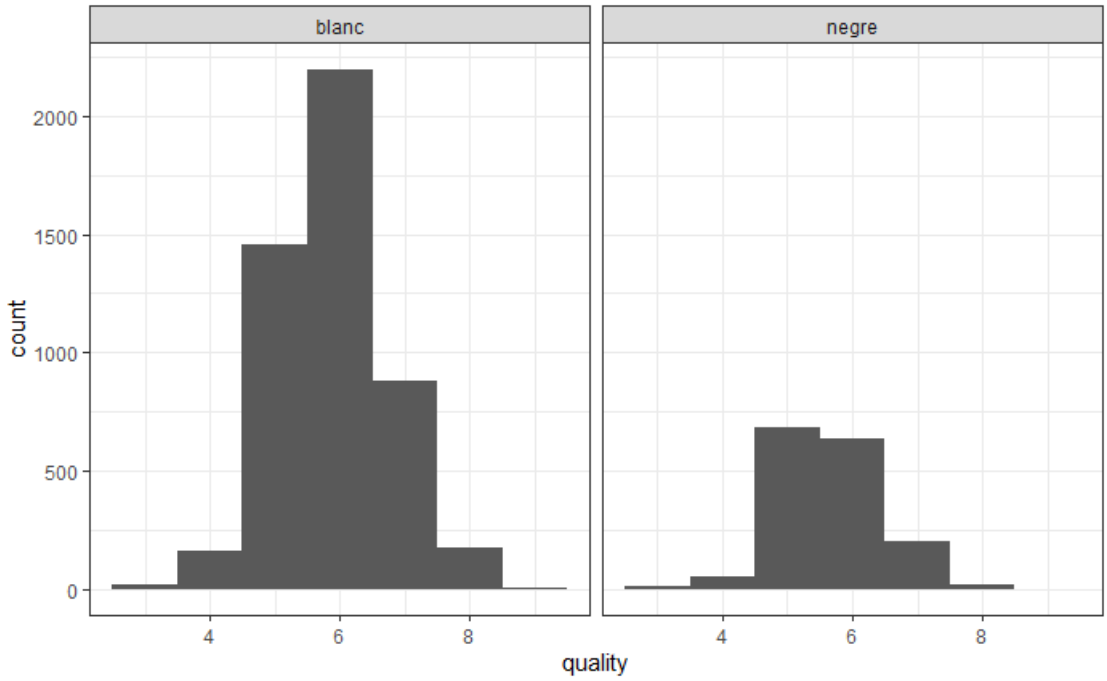
```
##
##  Welch Two Sample t-test
##
## data:  vins_negres_data$alcohol and vins_blancs_data$alcohol
## t = -2.859, df = 3100.5, p-value = 0.9979
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.143817       Inf
## sample estimates:
## mean of x mean of y
##  10.42298  10.51427
```

Així doncs, hem de concloure que allò de que els vins negres són més forts és més aviat una llegenda urbana pel cas dels vins *Vinho Verde*.

Comparació de la excel·lència dels vins blancs i els vins negres

També és interessant comprovar si hi ha diferències entre la qualitat dels vins blancs i negres. En aquest cas, un test d'igualtat de mitjanes no aporta molta informació, ja que hi ha vins bons i dolents de tots dos colors. Serà molt més interessant comprovar si hi ha una diferència en la proporció de vins excel·lents sobre el total de vins de cada color.

Comencem veient la distribució de les puntuacions dels vins de cada color:



En tots dos casos el nombre de vins amb puntuacions elevades és molt petit, tot i que sembla que la proporció podria ser una mica més elevada als vins blancs. Veiem també que no hi ha vins puntuats amb un 10 a la mostra, així que definirem que un vi és excel·lent si ha estat puntuat amb un 8 o un 9 (notar que amb un 9 només hi ha vins blancs).

Un cop feta aquesta comprovació podem fer un test d'igualtat de proporcions en el qual prendrem com a hipòtesi alternativa que els vins blancs tenen una proporció de vins excel·lents més alta, ja que és lo que sembla a la mostra:

p_negres	p_blancs
0.011257	0.0367497

A la taula hi són les proporcions de vins excel·lents blancs i negres, clarament superior en el cas dels blancs. El test comprovarà si aquesta diferència és significativa:

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(sum(vins_blancs_data$excelent), sum(vins_negres_data$excelent))
## X-squared = 26.514, df = 1, p-value = 1.308e-07
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.01929698 1.00000000
## sample estimates:
##     prop 1     prop 2
## 0.03674969 0.01125704
```

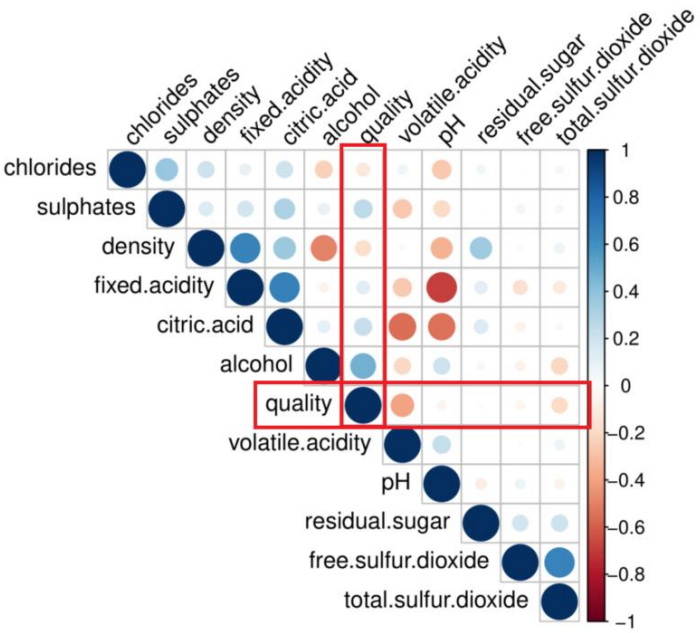
Efectivament, com que tenim un p-valor de gairebé 0, acceptem que hi ha una proporció significativament superior de vins blancs excel·lents respecte als negres.

Model predictiu per a la qualitat dels vins

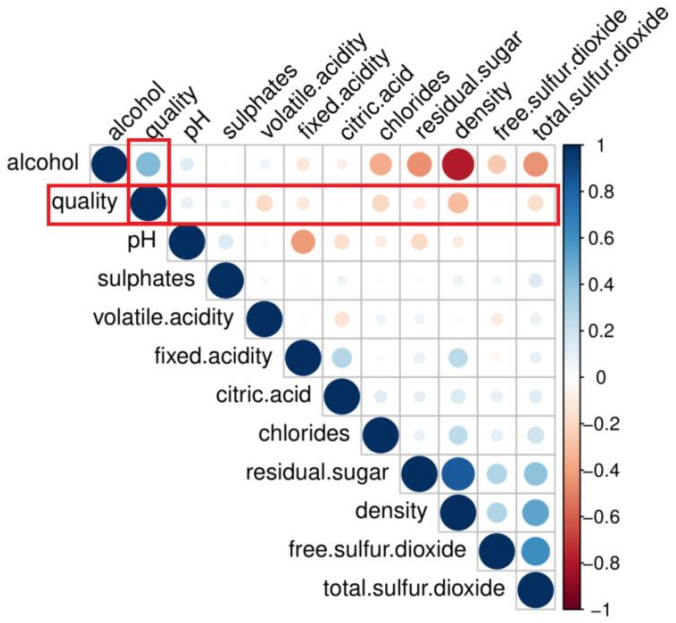
Finalment, intentarem desenvolupar un model de regressió lineal múltiple que ens permeti calcular la qualitat dels vins. Com que ja hem vist que hi ha diferències entre els vins blancs i negres, sobretot pel que fa a la qualitat, en realitat desenvoluparem dos models, un per a cada color de vi.

En primer lloc, s'han comprovat les correlacions de les variables amb la qualitat per cada tipus de vi. El resultat es pot veure a les dues imatges següents:

Vins negres: trobem les correlacions més fortes amb l'alcohol, l'acidesa volàtil, els sulfats i l'àcid cítric, de les quals totes són positives menys l'acidesa.



Vins blancs: la correlació més forta és també amb l'alcohol i positiva, però la següent és ara la densitat, negativa. Menys rellevància tenen els clorurs, l'acidesa volàtil i el diòxid de sulfur total, totes tres negatives.



Per construir el model, es comença agafant aquestes variables que semblen tenir més correlació amb la qualitat. Després de unes proves, que s'hi poden trobar al document RMarkdown, hem arribat als següents models:

Vins negres:

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates,
##     data = vins_negres_data_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73818 -0.38741 -0.04552  0.47084  1.97330
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.5449     0.2387  10.661  < 2e-16 ***
## alcohol            0.3100     0.0194  15.978  < 2e-16 ***
## volatile.acidity  -1.2644     0.1176 -10.755  < 2e-16 ***
## sulphates          0.7917     0.1213   6.529 1.03e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6576 on 1062 degrees of freedom
## Multiple R-squared:  0.3539, Adjusted R-squared:  0.3521
## F-statistic: 193.9 on 3 and 1062 DF,  p-value: < 2.2e-16
```

El model té en compte l'alcohol i els sulfats amb una influència positiva i l'acidesa volàtil amb influència negativa. El valor de R^2 no és molt bo, menys de 0.5, però ja veurem més endavant

com les prediccions que fem no són dolentes, en general. Un valor tan petit es pot deure a la variabilitat subjectiva que hi ha sempre a una variable com la puntuació donada per uns jutges.

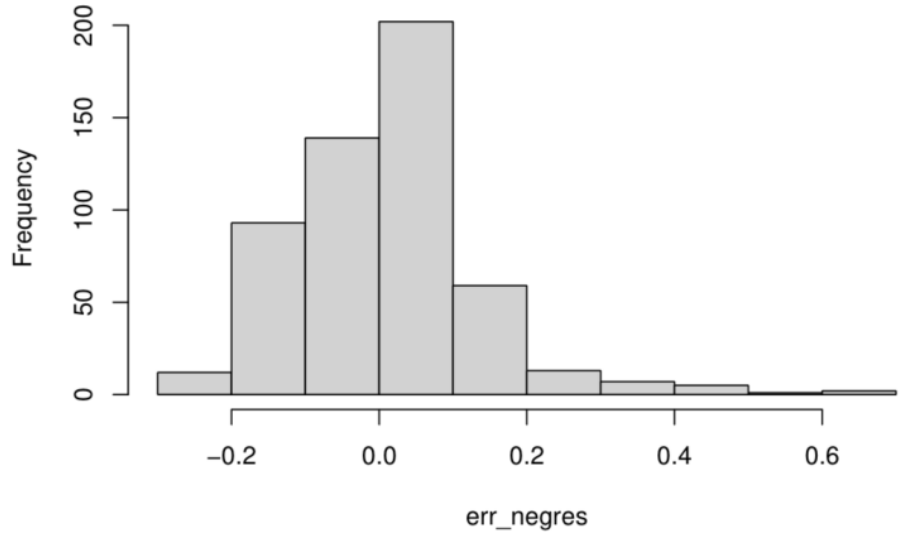
Vins blancs:

```
##
## Call:
## lm(formula = quality ~ alcohol + density + volatile.acidity +
##     total.sulfur.dioxide, data = vins_blancs_data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3035 -0.4916 -0.0324  0.4876  3.0691
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -3.187e+01  7.435e+00  -4.287 1.87e-05 ***
## alcohol               3.984e-01  1.751e-02  22.747  < 2e-16 ***
## density               3.422e+01  7.366e+00   4.646 3.52e-06 ***
## volatile.acidity     -2.127e+00  1.349e-01 -15.762  < 2e-16 ***
## total.sulfur.dioxide  9.412e-04  3.690e-04   2.551   0.0108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7691 on 3260 degrees of freedom
## Multiple R-squared:  0.242,  Adjusted R-squared:  0.2411
## F-statistic: 260.3 on 4 and 3260 DF,  p-value: < 2.2e-16
```

El model té en compte l'alcohol i la densitat amb una influència positiva i l'acidesa volàtil amb influència negativa. També té en compte el diòxid de sulfur total, significatiu a un nivell de confiança estàndard del 95%, tot i que la seva influència és molt petita. El comentari al valor de R^2 és el mateix que als vins negres.

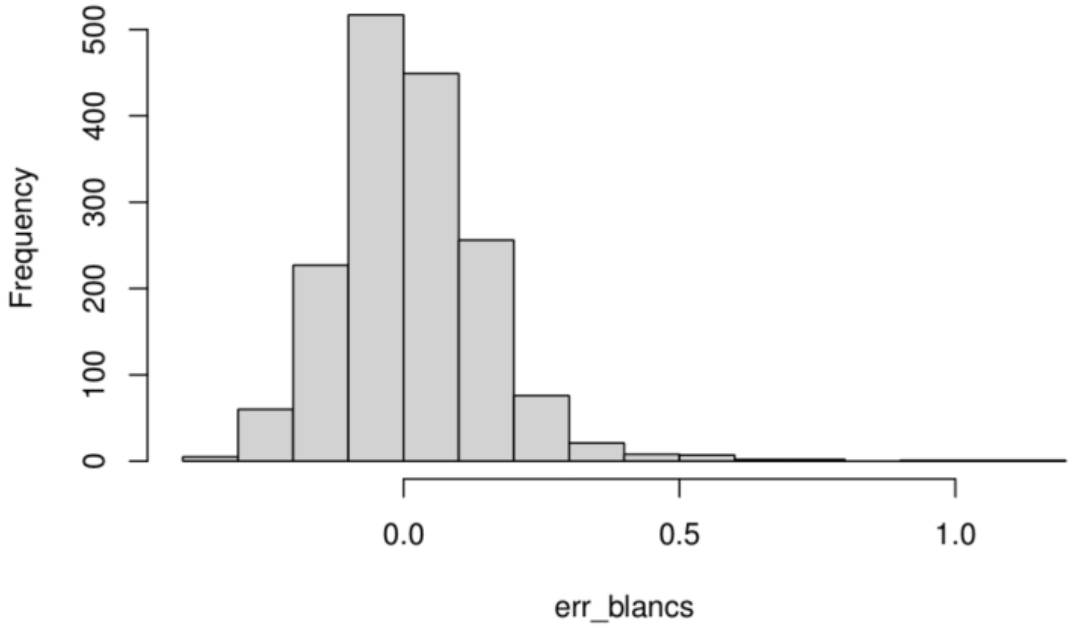
Com és normal per construir aquests tipus de models, hem fet servir només dos terços de la mostra, reservant un terç (aleatòriament) per fer proves. Si fem prediccions sobre aquests registres de testeig, podem fer-nos una idea de lo bones que són estudiant els errors. En primer lloc, veiem la distribució dels errors relatius.

Vins negres:



La distribució no és gaire simètrica, amb una cua a dretes. Tot i això, la majoria de les observacions tenen un error petit.

Vins blancs:



En aquest cas, tot i que hi ha uns pocs valors amb un error alt, la resta d'errors es distribueixen d'una forma semblant a una normal, que seria el comportament esperat.

La mitjana dels errors relatius (ara en valor absolut) és la següent:

Error relatiu mitjà vins blancs	Error relatiu mitjà vins negres
0.1041107	0.092238

En general, podem concloure que, tot i que els valors de R^2 dels models no són els millors, les prediccions que fem en els dos casos sí que son acurades, amb un 10% d'error mitjà i una distribució força simètrica dels errors, sobretot en el cas dels vins blancs.

Conclusions

En aquesta pràctica, amb les dades físico-químiques i de qualitat dels vins portuguesos de la varietat *Vinho Verde* hem aconseguit donar uns insights força interessants sobre aquests vins, contestant les preguntes plantejades a l'inici de la pràctica. En resum, hem vist que la creença que els vins negres tenen més alcohol és només una llegenda urbana, que hi ha una diferència significativa en la proporció de vins amb puntuacions més altes favorable als vins blancs i finalment hem desenvolupat dos models que prediuen la qualitat dels vins de cada color segons a partir de les característiques físico-químiques més significatives en cada cas. Aquest models, si bé semblaven a l'inici força dolents, ja que la bondat de l'ajustament no era molt alta, sí que ens permeten fer unes prediccions bastant acurades.