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

Tipologia i cicle de vida de les dades - Pràctica 2

Héctor Gutiérrez Muñoz

5 de juny, 2021

Sumari

1 Càrrega dels fitxers

En primer lloc, carreguem els dos fitxers.

```
blancs_data <- read.csv("wineQualityWhites.csv")
negres_data <- read.csv("wineQualityReds.csv")
```

Comprovem que les dades es llegeixen amb els tipus correctes i que no hi ha valors estranys o errors de codificació:

```
# Comencem pels vins blancs

str(blancs_data)
```

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
summary(blancs_data)
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median : 5.200   Median :0.04300   Median : 34.00      Median :134.0
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

```
# A continuació els de negres

str(negres_data)
```

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
```

```
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
summary(negres_data)
```

```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median : 2.200   Median :0.07900   Median :14.00       Median : 38.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

Veiem que tots els tipus són els esperats i que no hi ha cap valor estrany o que no s'hagi llegit correctament. Veiem que tampoc no hi ha cap valor desconegut o NA.

2 Integració dels datasets

A continuació hem d'integrar els dos fitxers en un sol dataset. Crearem una nova variable que indiqui el color del vi i eliminarem la columna autoincremental que no ens aporta cap informació.

```
blancs_data$X <- NULL
blancs_data$color <- "blanc"

negres_data$X <- NULL
negres_data$color <- "negre"

vins_data <- rbind(blancs_data, negres_data)
```

El dataset final sobre el qual treballarem és el següent:

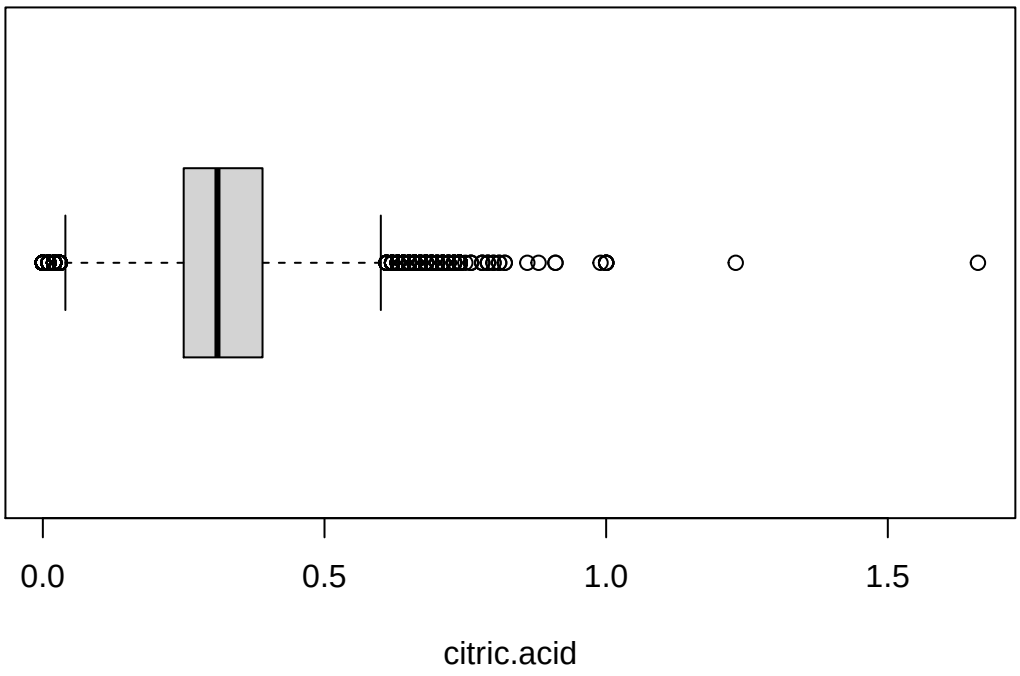
```
kable(rbind(
  head(vins_data[vins_data$color == "blanc",],3),
  head(vins_data[vins_data$color == "negre",],3)
),
  format = 'latex',
  booktabs = TRUE,
  row.names = FALSE) %>%
 kableExtra::kable_styling(latex_options = c("scale_down", "hold_position"))
```

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

3 Neteja de les dades

Pel que fa als valors desconeguts o NA no hem de fer res, ja que abans, amb els "summary" ja hem vist que totes les dades estan omplertes. A la variable del àcid cítric hi ha registres amb valor 0, però són observacions vàlides: com podem veure al boxplot, els valors típics de àcid cítric a un vi són baixos i així, no és rar que en alguns casos no n'hi hagi res i l'observació sigui 0.

```
boxplot(vins_data$citric.acid, horizontal = TRUE, xlab=c("citric.acid"))
```

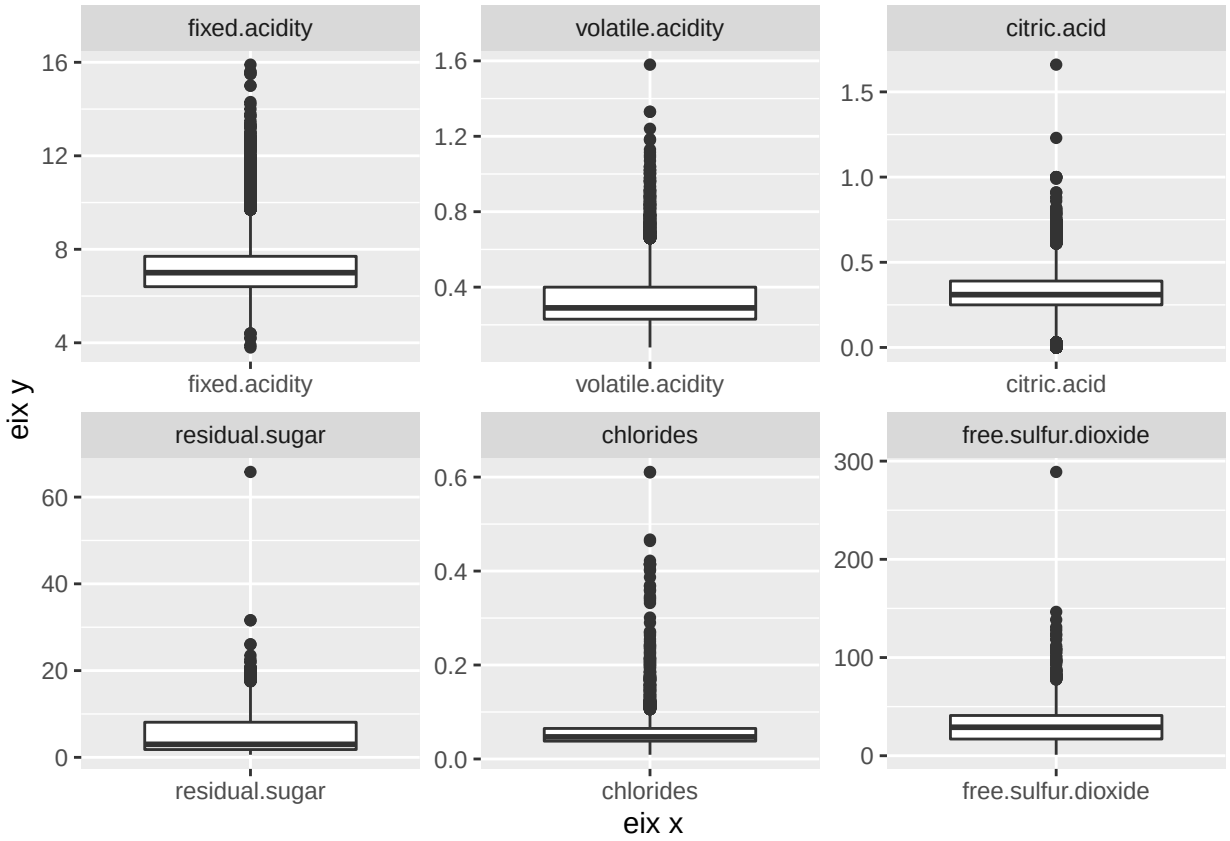


Així doncs, aquesta etapa de neteja es reduirà a estudiar els valors extrems o outliers, que ja podem apreciar que sí existeixen al boxplot anterior. Es presenten a continuació els boxplots de les variables explicatives:

```
p <- ggplot(data = melt(vins_data[,c(1:6)]), aes(x=variable, y=value)) +
  geom_boxplot() +
  xlab("eix x") +
  ylab("eix y")
```

```
## No id variables; using all as measure variables
```

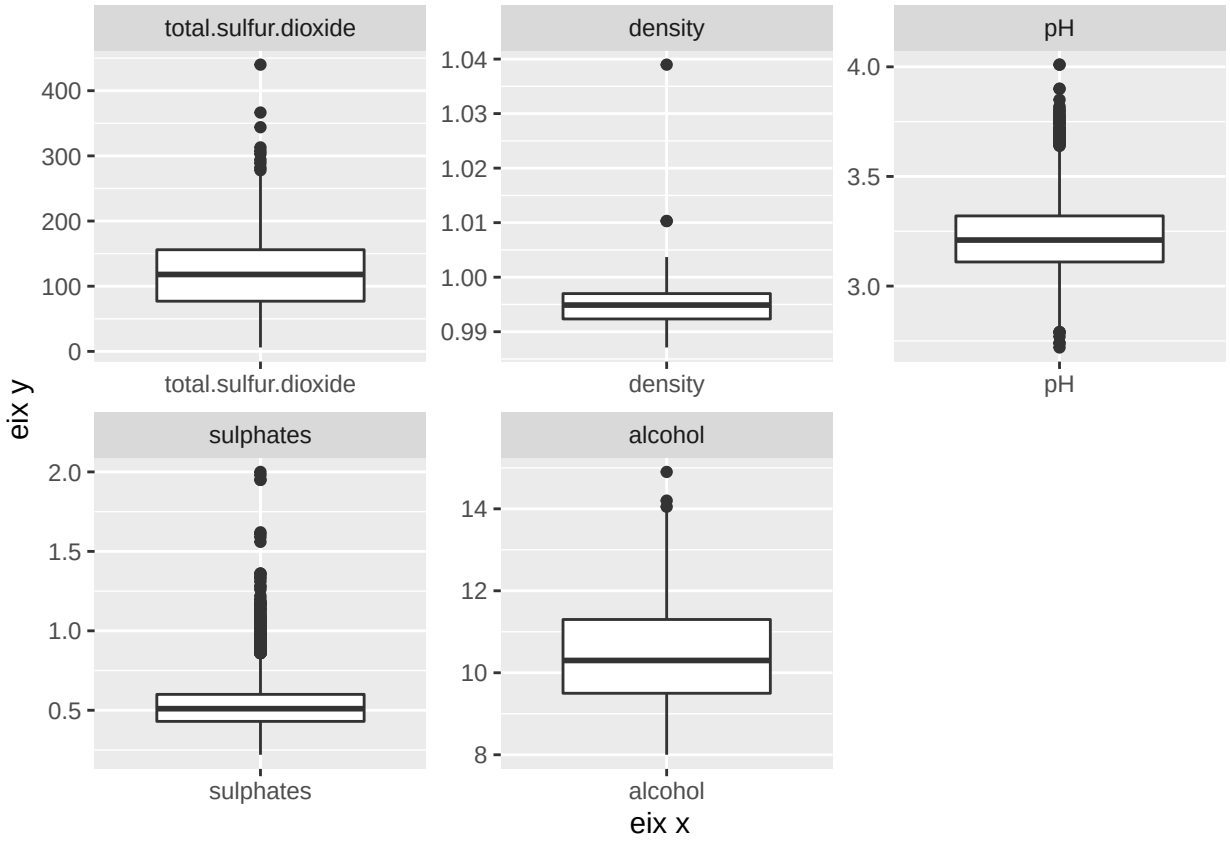
```
p + facet_wrap( ~ variable, scales="free")
```




```
p <- ggplot(data = melt(vins_data[,c(7:11)]), aes(x=variable, y=value)) +
  geom_boxplot() +
  xlab("eix x") +
  ylab("eix y")
```

```
## No id variables; using all as measure variables
```

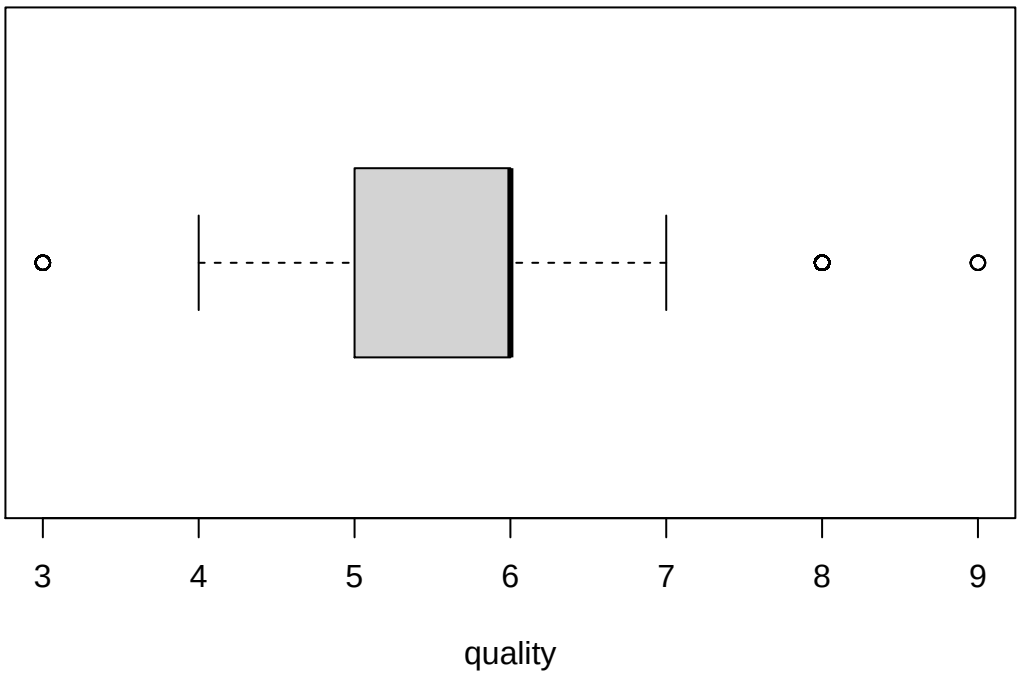
```
p + facet_wrap( ~ variable, scales="free")
```



Podem apreciar que en tots els casos sí que hi ha valors extrems (els punts que hi ha a sobre o a baix de les caixes i els bigotis), però que tots aquests valors extrems són observacions vàlides que no difereixen gaire de la resta, i que per tant, hem de considerar a l'hora de construir els nostres models o analitzar les dades. Fins i tot en els casos en què hi ha un punt més lluny de la resta, com ara la densitat, el sucre residual o el sulfur lliure, no és una diferència prou gran com per haver d'excloure'ls.

A la variable qualitat, la variable que més endavant volem predir en funció de la resta, passa el mateix:

```
boxplot(vins_data$quality, horizontal = TRUE, xlab=c("quality"))
```




```
kable(t(table(vins_data$quality)),
  format = 'latex',
  booktabs = TRUE) %>%
  kableExtra::kable_styling(latex_options = c("hold_position"))
```

3	4	5	6	7	8	9
30	216	2138	2836	1079	193	5

Sí que hi ha valors extrems, 30 vins puntuats amb un 3, 193 amb un 8 i 5 amb un 9. Però és clar que aquestes dades també les hem de fer servir per construir el model, ja que són completament vàlides i treure-les-en no aportaria cap guany a la qualitat del model.

4 Anàlisi de les dades

4.1 Separació de la mostra pels contrastos d'hipòtesis

Començarem la part d'anàlisi comprovant si alguns mites sobre les diferències entre els vins blancs i els de negres són veritat o són només llegendes urbanes. Per fer això, en primer lloc haurem de separar la mostra en dues, segons el color del vi:

```
vins_blancs_data <- vins_data[vins_data$color == "blanc", ]
vins_negres_data <- vins_data[vins_data$color == "negre", ]
```

4.2 Normalitat i homoscedasticitat

Les dues submostres tenen més de 30 observacions, com ja hem vist abans. Així, pel teorema del límit central, podem suposar la normalitat en totes dues.

Pel que fa a l'homoscedasticitat (homogeneïtat de variàncies), haurem de comprovar-ho amb un test d'igualtat de variàncies. Concretament, necessitarem saber si la distribució de la variable alcohol presenta homoscedasticitat en els dos grups per triar el test correcte més endavant.

```
var.test(vins_blancs_data[,"alcohol"], vins_negres_data[,"alcohol"], conf.level = 0.95)
```

```
##
##  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
```

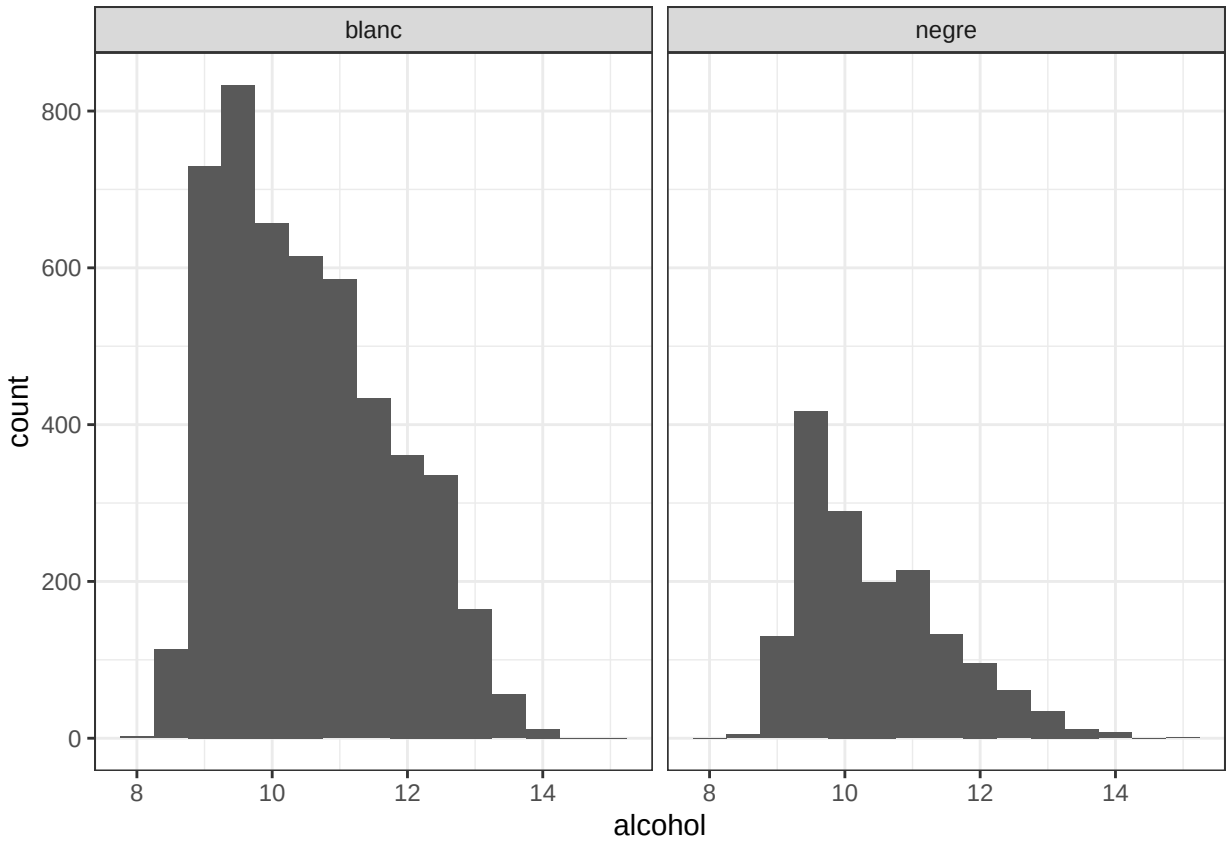
Veiem que el p-valor està tocant el 0 i per això hem de rebutjar la hipòtesi nul · la que les dues variàncies són iguals: ens trobem a una situació d'heteroscedasticitat. No necessitarem fer aquesta comparació per a cap variable més.

4.3 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. Per comprovar-ho farem servir un test d'igualtat de mitjanes, però tenint en compte que les variàncies no són iguals.

En primer lloc, representem gràficament la situació:

```
ggplot(vins_data,aes(x=alcohol)) +
  geom_histogram(binwidth = 0.5) +
  facet_grid(~color) +
  theme_bw()
```



Les dues distribucions dels continguts d'alcohol són molt semblants excepte per uns valors molt alts en els vins negres, cosa que ens pot fer pensar que realment no hi ha cap diferència.

Un cop comprovat això, fem el test:

```
t.test(vins_negres_data$alcohol,
       vins_blancs_data$alcohol,
       alternative="greater",
       var.equal=FALSE)
```

```
##
##  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
```

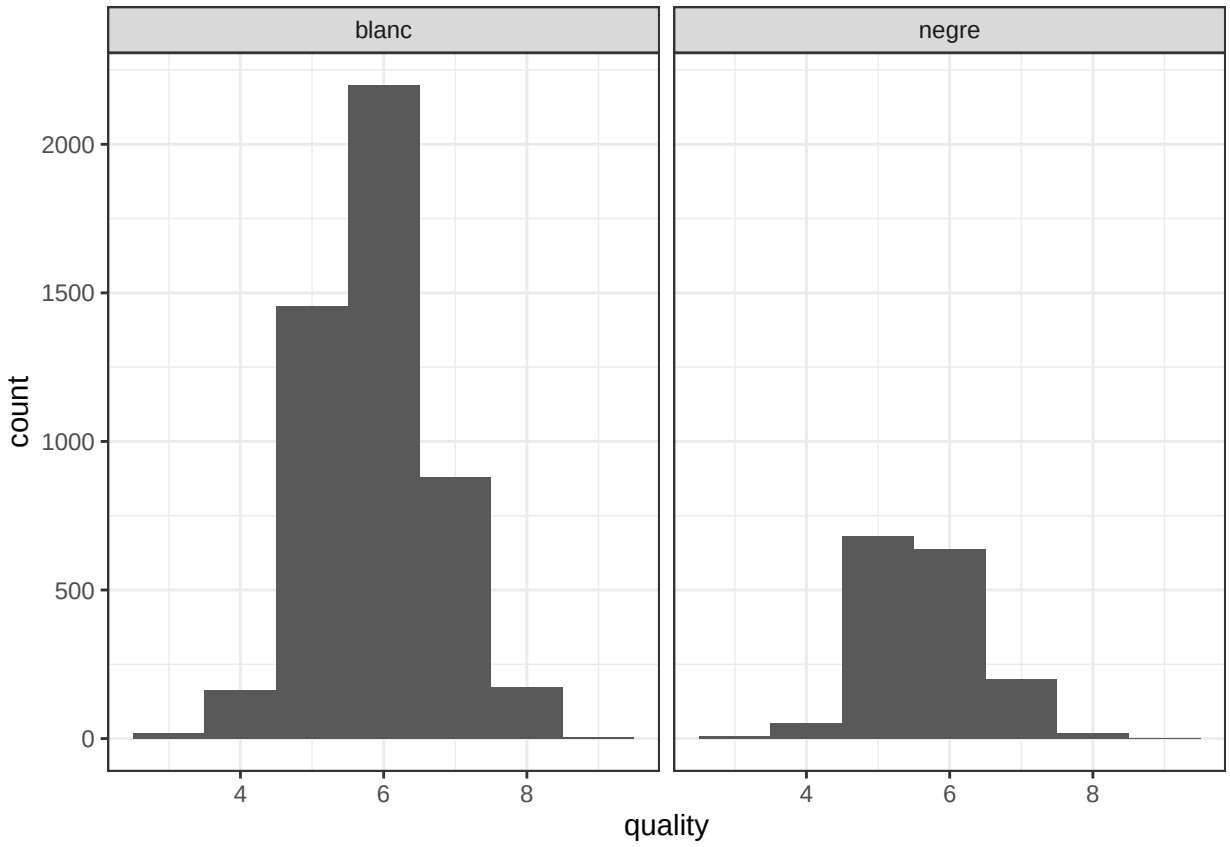
En aquest test, la hipòtesi nul · la és que les mitjanes són les mateixes i l'alternativa és que la mitjana d'alcohol dels vins negres és més gran que la dels de blancs. Com que el p-valor està tocant l'1, no podem rebutjar la hipòtesi nul · la, així que hem de concloure que això que els vins negres tenen més alcohol és només una llegenda urbana, almenys en els vins "Vinho Verde".

4.4 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:

```
ggplot(vins_data,aes(x=quality)) +
  geom_histogram(binwidth = 1) +
  facet_grid(~color) +
  theme_bw()
```



En tots dos casos el nombre de vins amb puntuacions elevades és molt petit, sembla que hi ha més casos als vins blancs.

Com que no hi ha vins puntuats amb un 10 a la mostra, definirem que un vi és excel · lent si ha estat puntuat amb un 8 o un 9. Desarem això a una nova variable:

```
vins_negres_data$excelent <- vins_negres_data$quality >= 8
vins_blancs_data$excelent <- vins_blancs_data$quality >= 8
```

En aquest cas, no importa si ens trobem a una situació d'heteroscedasticitat o d'homoscedasticitat, que les dues mostres siguin grans, com és el cas, és suficient.

Veiem ara quines són les proporcions mostrals:

```r
p_negres <- sum(vins_negres_data$excelent)/length(vins_negres_data$excelent)
p_blancs <- sum(vins_blancs_data$excelent)/length(vins_blancs_data$excelent)

kable(cbind(p_negres, p_blancs),
  format = 'latex',
  booktabs = TRUE) %>%
  kableExtra::kable_styling(latex_options = c("hold_position"))
```

p_negres	p_blancs
0.011257	0.0367497

La proporció de vins excel · lents veiem que és superior en els vins blancs: analitzarem si aquesta diferència és significativa o no. Així, el test sobre les proporcions tindrà com a hipòtesi nul · la que la proporció de vins excel · lents de cada color és la mateixa, i com a hipòtesi alternativa que la proporció en els vins blancs sigui superior.

```r
prop.test(c(sum(vins_blancs_data$excelent), sum(vins_negres_data$excelent)),
          c(length(vins_blancs_data$excelent), length(vins_negres_data$excelent)),
          alternative = "greater",
          correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(sum(vins_blancs_data$excelent), sum(vins_negres_data$excelent)) out of c(length(vins_blancs
## 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
```

Com que el p-valor està tocant el 0, rebutjem la hipòtesi nul · la i podem concloure que la proporció de vins blancs excel · lents és significativament superior a la de vins excel · lents negres en els vins "Vinho Verde".

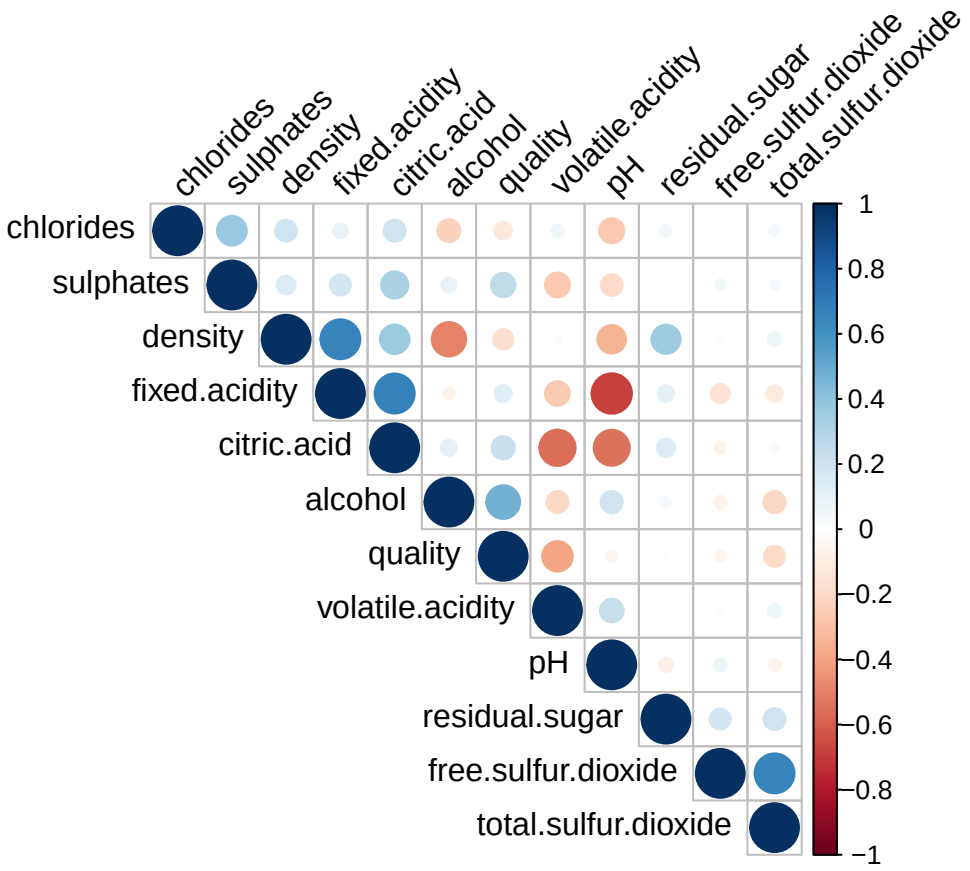
4.5 Model predictiu per la qualitat dels vins

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.

Comencem per comprovar la correlació de les variables amb la qualitat:

```r
# Vins negres

corrplot(cor(vins_negres_data[,c(1:12)]),
         type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```




```
cor(vins_negres_data[,c(1:12)])[, "quality"]
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.12405165          -0.39055778           0.22637251
##       residual.sugar            chlorides  free.sulfur.dioxide
##           0.01373164          -0.12890656          -0.05065606
## total.sulfur.dioxide              density                   pH
##          -0.18510029          -0.17491923          -0.05773139
##            sulphates              alcohol              quality
##           0.25139708           0.47616632           1.00000000
```

En els 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

corrplot(cor(vins_blancs_data[,c(1:12)]),
         type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```


```
cor(vins_blancs_data[,c(1:12)])[, "quality"]
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##         -0.113662831         -0.194722969         -0.009209091
##       residual.sugar            chlorides  free.sulfur.dioxide
##         -0.097576829         -0.209934411          0.008158067
## total.sulfur.dioxide              density                   pH
##         -0.174737218         -0.307123313          0.099427246
##            sulphates              alcohol              quality
##          0.053677877          0.435574715          1.000000000
```

Pel que fa als 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 desenvolupar els models, comencem per dividir els dos datasets en dues parts: una per construir el model i una segona de test. Ho farem de la forma estàndard, dos terços per desenvolupar-lo i un terç per provar-lo, amb la divisió aleatòria.

```
indexes_blanc <- sample(1:nrow(vins_blancs_data),
                        size = floor((2/3)*nrow(vins_blancs_data)))
indexes_negre <- sample(1:nrow(vins_negres_data),
                        size = floor((2/3)*nrow(vins_negres_data)))

vins_blancs_data_train <- vins_blancs_data[indexes_blanc, ]
```

```r
vins_blancs_data_test <- vins_blancs_data[-indexes_blanc, ]

vins_negres_data_train <- vins_negres_data[indexes_negre, ]
vins_negres_data_test <- vins_negres_data[-indexes_negre, ]
```

En primer lloc, fem models amb aquestes variables amb més correlació:

```r
# Vins negres

mod1n <- lm(quality ~ alcohol + volatile.acidity + sulphates + citric.acid,
   data = vins_negres_data_train)

summary(mod1n)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     citric.acid, data = vins_negres_data_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.69624 -0.37933 -0.07784  0.46620  2.20084
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.5808548  0.2492995  10.352  < 2e-16 ***
## alcohol           0.3205267  0.0195845  16.366  < 2e-16 ***
## volatile.acidity -1.2090098  0.1391386  -8.689  < 2e-16 ***
## sulphates         0.5515518  0.1223485   4.508 7.27e-06 ***
## citric.acid      -0.0004732  0.1267297  -0.004    0.997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6606 on 1061 degrees of freedom
## Multiple R-squared:  0.333,  Adjusted R-squared:  0.3305
## F-statistic: 132.5 on 4 and 1061 DF,  p-value: < 2.2e-16
```

```r
# Vins blancs

mod1b <- lm(quality ~ alcohol + density + chlorides + volatile.acidity + total.sulfur.dioxide,
   data = vins_blancs_data_train)

summary(mod1b)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + density + chlorides + volatile.acidity +
##     total.sulfur.dioxide, data = vins_blancs_data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3618 -0.4935 -0.0328  0.4797  3.0370
```

```
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -3.357e+01  8.304e+00  -4.042 5.41e-05 ***
## alcohol               3.951e-01  1.915e-02  20.629  < 2e-16 ***
## density               3.604e+01  8.221e+00   4.384 1.20e-05 ***
## chlorides            -1.886e+00  6.415e-01  -2.940   0.0033 **
## volatile.acidity     -1.938e+00  1.367e-01 -14.172  < 2e-16 ***
## total.sulfur.dioxide  6.128e-04  3.825e-04   1.602   0.1092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7683 on 3259 degrees of freedom
## Multiple R-squared:  0.2503, Adjusted R-squared:  0.2491
## F-statistic: 217.6 on 5 and 3259 DF,  p-value: < 2.2e-16
```

Veiem que tots dos models tenen un valor R^2 molt baix, però hem de tenir en compte que estem treballant amb una variable com és la puntuació d'un vi que també té un component subjectiu molt gran i per tant una variabilitat alta que és difícil d'explicar. Això es confirma si fem la prova d'afegir la resta de variables al model:

```r
# Vins negres

mod2n <- lm(quality ~ alcohol + density + chlorides + volatile.acidity +
             total.sulfur.dioxide + fixed.acidity + volatile.acidity +
             citric.acid + residual.sugar + free.sulfur.dioxide + pH + sulphates,
   data = vins_negres_data_train)

summary(mod2n)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + density + chlorides + volatile.acidity +
##     total.sulfur.dioxide + fixed.acidity + volatile.acidity +
##     citric.acid + residual.sugar + free.sulfur.dioxide + pH +
##     sulphates, data = vins_negres_data_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64801 -0.37111 -0.04244  0.45357  1.88593
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.157e+01  2.543e+01   1.241 0.214724
## alcohol               2.776e-01  3.127e-02   8.879  < 2e-16 ***
## density              -2.768e+01  2.599e+01  -1.065 0.287103
## chlorides            -2.295e+00  4.982e-01  -4.606 4.62e-06 ***
## volatile.acidity     -1.004e+00  1.479e-01  -6.788 1.90e-11 ***
## total.sulfur.dioxide -3.069e-03  8.729e-04  -3.516 0.000457 ***
## fixed.acidity         4.372e-02  3.176e-02   1.377 0.168954
## citric.acid          -1.574e-01  1.797e-01  -0.876 0.381349
## residual.sugar        2.466e-02  1.854e-02   1.330 0.183768
## free.sulfur.dioxide   3.089e-03  2.626e-03   1.176 0.239900
```

```
## pH                   -4.029e-01  2.322e-01  -1.735 0.082992 .
## sulphates             8.287e-01  1.340e-01   6.185 8.86e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6459 on 1054 degrees of freedom
## Multiple R-squared:  0.3666, Adjusted R-squared:   0.36
## F-statistic: 55.45 on 11 and 1054 DF,  p-value: < 2.2e-16
```

```r
# Vins blancs

mod2b <- lm(quality ~ alcohol + density + chlorides + volatile.acidity +
             total.sulfur.dioxide + fixed.acidity + volatile.acidity +
             citric.acid + residual.sugar + free.sulfur.dioxide + pH + sulphates,
   data = vins_blancs_data_train)

summary(mod2b)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + density + chlorides + volatile.acidity +
##     total.sulfur.dioxide + fixed.acidity + volatile.acidity +
##     citric.acid + residual.sugar + free.sulfur.dioxide + pH +
##     sulphates, data = vins_blancs_data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6085 -0.4913 -0.0431  0.4639  3.0934
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.193e+02  2.913e+01   7.529 6.58e-14 ***
## alcohol               1.101e-01  3.655e-02   3.014   0.0026 **
## density              -2.203e+02  2.952e+01  -7.465 1.06e-13 ***
## chlorides            -5.056e-01  6.457e-01  -0.783   0.4336
## volatile.acidity     -1.725e+00  1.392e-01 -12.393  < 2e-16 ***
## total.sulfur.dioxide -4.027e-04  4.709e-04  -0.855   0.3925
## fixed.acidity         1.244e-01  2.901e-02   4.288 1.85e-05 ***
## citric.acid           8.294e-02  1.140e-01   0.727   0.4671
## residual.sugar        1.038e-01  1.093e-02   9.492  < 2e-16 ***
## free.sulfur.dioxide   4.939e-03  1.040e-03   4.747 2.15e-06 ***
## pH                    9.151e-01  1.397e-01   6.548 6.75e-11 ***
## sulphates             7.354e-01  1.257e-01   5.849 5.42e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7487 on 3253 degrees of freedom
## Multiple R-squared:  0.2893, Adjusted R-squared:  0.2869
## F-statistic: 120.4 on 11 and 3253 DF,  p-value: < 2.2e-16
```

Veiem que la millora en R^2 no és significativa tot i incorporar totes les variables, així que ens quedarem amb els primers models, ja que produeixen un resultat semblant i és molt més senzill.

Una cosa del primer model que no hem comentat, però, és que per a cada color hi ha una variable que no és significativa a un nivell del 95%: l'àcid cítric pels vins negres i els clorurs pels vins blancs. Així doncs, les en traurem i aquests seran els nostres dos models finals que testejarem.

```
# Vins negres

mod3n <- lm(quality ~ alcohol + volatile.acidity + sulphates,
   data = vins_negres_data_train)

summary(mod3n)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates,
##     data = vins_negres_data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6963 -0.3794 -0.0779  0.4662  2.2007
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.58065    0.24333  10.605  < 2e-16 ***
## alcohol            0.32053    0.01957  16.375  < 2e-16 ***
## volatile.acidity -1.20875    0.11977 -10.092  < 2e-16 ***
## sulphates          0.55146    0.12005   4.594 4.88e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6603 on 1062 degrees of freedom
## Multiple R-squared:  0.333,  Adjusted R-squared:  0.3312
## F-statistic: 176.8 on 3 and 1062 DF,  p-value: < 2.2e-16
```

```
# Vins blancs

mod3b <- lm(quality ~ alcohol + density + volatile.acidity + total.sulfur.dioxide,
   data = vins_blancs_data_train)

summary(mod3b)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + density + volatile.acidity +
##     total.sulfur.dioxide, data = vins_blancs_data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3490 -0.4911 -0.0349  0.4827  3.0522
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -3.562e+01  8.284e+00  -4.300 1.76e-05 ***
## alcohol               4.106e-01  1.844e-02  22.267  < 2e-16 ***
```

```
## density               3.788e+01  8.207e+00   4.616 4.06e-06 ***
## volatile.acidity     -1.985e+00  1.360e-01 -14.599  < 2e-16 ***
## total.sulfur.dioxide  5.296e-04  3.819e-04   1.387    0.166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7692 on 3260 degrees of freedom
## Multiple R-squared:  0.2483, Adjusted R-squared:  0.2474
## F-statistic: 269.2 on 4 and 3260 DF,  p-value: < 2.2e-16
```

Veurem com d'acurats són aquests models fent una predicció sobre el conjunt de test.

```
pred_blancs <- predict(mod3b, vins_blancs_data_test)
pred_negres <- predict(mod3n, vins_negres_data_test)
```

La mitjana de l'error relatiu que tenen aquests models és:

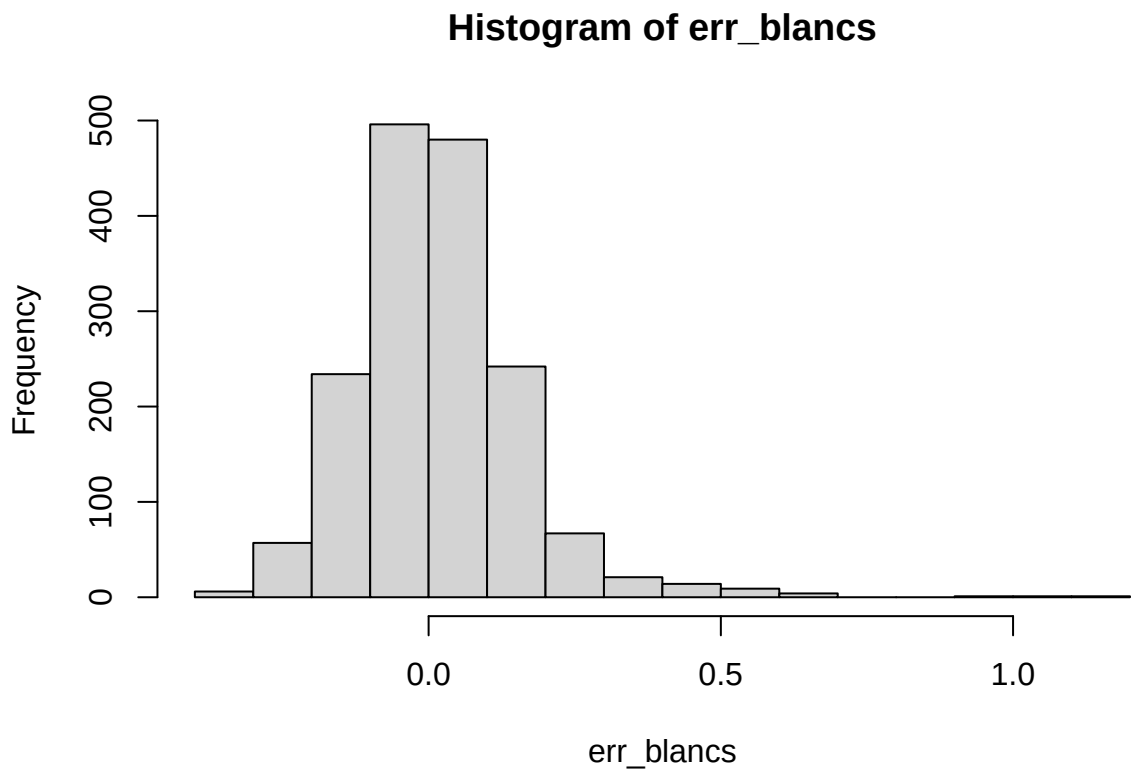
```
err_blancs <- (pred_blancs-vins_blancs_data_test$quality)/vins_blancs_data_test$quality
err_negres <- (pred_negres-vins_negres_data_test$quality)/vins_negres_data_test$quality

kable(cbind(mean(abs(err_blancs)), mean(abs(err_negres))),
  format = 'latex',
  booktabs = TRUE,
  col.names = c("Error relatiu mitjà vins blancs",
                "Error relatiu mitjà vins negres")) %>%
 kableExtra::kable_styling(latex_options = c("hold_position"))
```

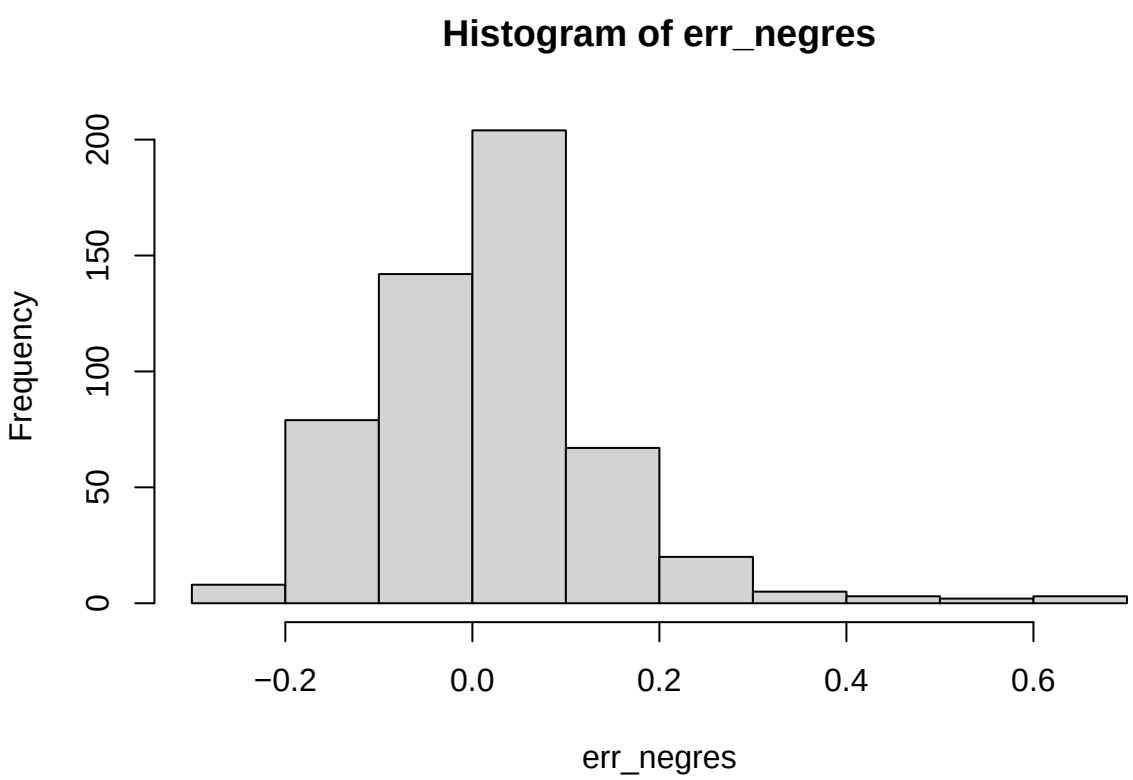
Error relatiu mitjà vins blancs	Error relatiu mitjà vins negres
0.1052168	0.0952343

Tot i que les R^2 no van sortir gaire bé, les prediccions sí que són bones, amb aproximadament un 10% d'error mitjà. També podem estudiar la distribució dels errors:

```
hist(err_blancs)
```




```
hist(err_negres)
```



Veiem que en el cas dels vins blancs, els errors s'acosten a una normal, excepte per uns pocs casos que estan molt allunyats del valor real, cosa que indica que realment les prediccions són bones en general. En el cas dels vins negres els errors no s'acosten tant a la normal, però no n'hi ha tan d'allunyats del valor real. Aquest model, en general, també té una bona capacitat predictiva.