Analysis of data set: properties of wines

Statistical Analysis for Applied Physics

Joana Moura
February 1st, 2022

OBJECTIVE

QUESTIONS:

1. Do red and white wines have the same alcohol percentage?
2. How does the pH of the wines depend on the volatile acidity and the fixed acidity?
3. Which parameters will be more important on the final quality rate of a wine?

ABOUT THE DATASET...

number	Type	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
1	Red	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5

Nobs = 3198

There are 1599 observations for both types of wines (equal length).

DO RED AND WHITE WINES HAVE THE SAME ALCOHOL PERCENTAGE?

Let's start by...

H_0: mean red = mean white
H_1: mean red ≠ mean white

INITIAL METRICS ABOUT THE ALCOHOL(%) PER TYPE OF WINE

Type	Mean μ (%)	Standard Deviation σ (%)	Variance σ^2 (%)
Red	10,423	1,066	1,136
White	10,285	1,124	1,263

Correlation = 0,0142

DATA DISTRIBUTIONS: HISTOGRAMS

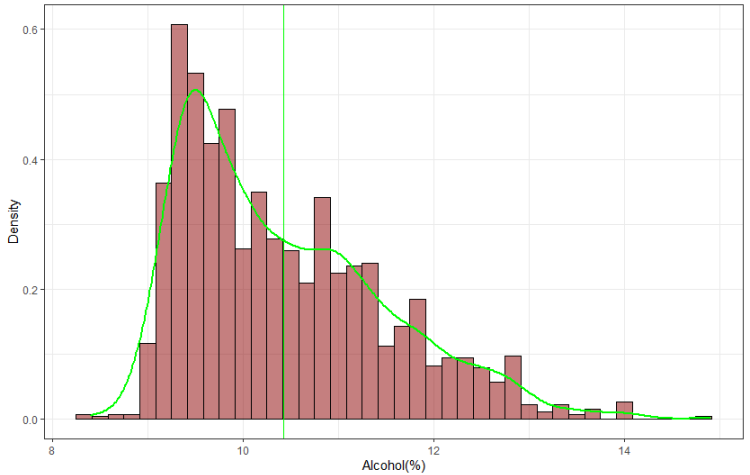

Fig.1 – Histogram of alcohol percentage in red wines

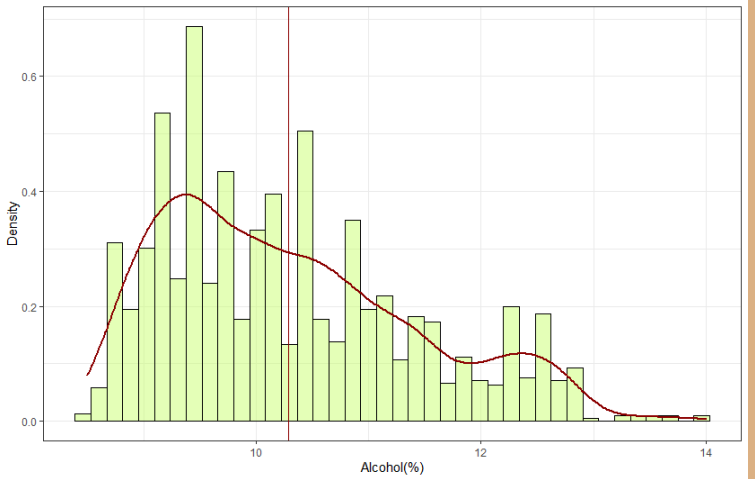

Fig.2 – Histogram of alcohol percentage in white wines

T-TEST

Requisites:

- Equal Variance
- Independent
- Normally Distributed

Q-Q PLOT

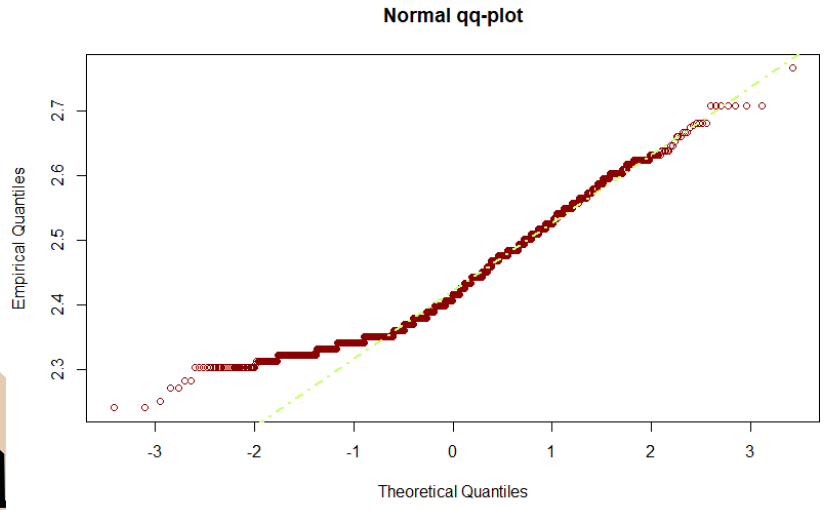

Fig.5 – Normal Q-Q plot of alcohol percentage in red wines.

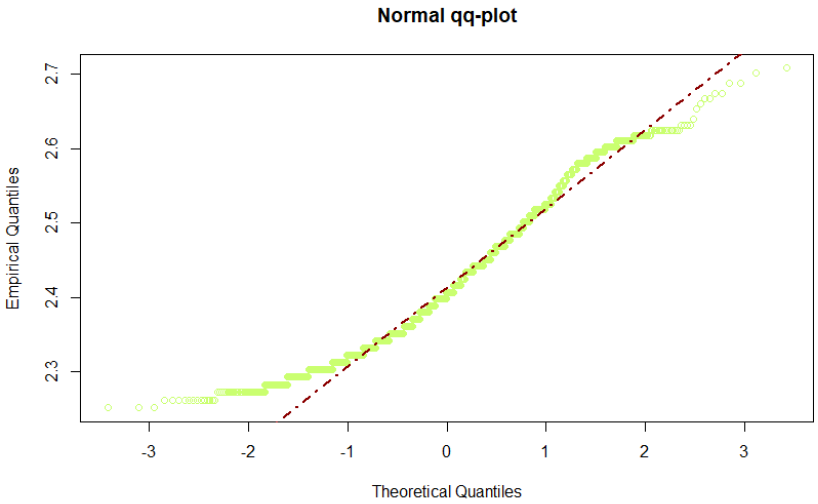

Fig.6 – Normal Q-Q plot of alcohol percentage in white wines.

T-TEST

```
        Welch Two Sample t-test

data:  alcohol_red and alcohol_white
t = 3.5562, df = 3187, p-value = 0.0003818
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.06179736 0.21368732
sample estimates:
mean of x mean of y
 10.42298  10.28524
```

Fig.5 – Welch's two sample t-test for both types of wine.

POWER OF THE T-TEST

```
     Two-sample t test power calculation

              n = 1599
          delta = 0.1377423
             sd = 1.094817
      sig.level = 0.05
          power = 0.9447989
    alternative = two.sided

NOTE: n is number in *each* group
```

Fig.5 – Power of Welch's two sample t-test for both types of wine.

BOOTSTRAP

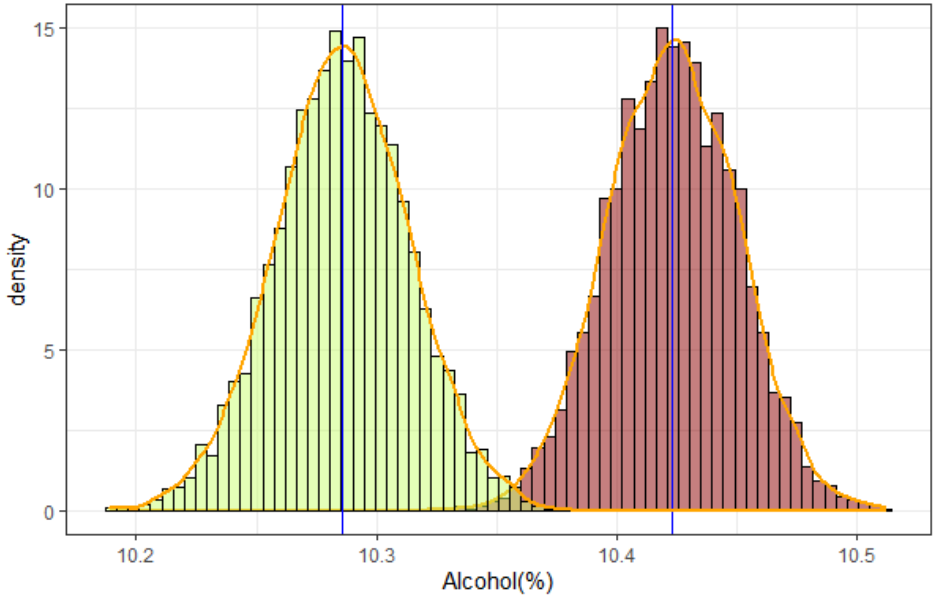

Fig.2 – Bootstrap of alcohol percentage for wines

CONCLUSION

Confidence Interval: (0,13+-0,10) at 95%

Dependance of the pH on the fixed acidity and the volatile acidity

CORRELATION

Type	Correlation (fixed acidity vs pH)	Correlation (volatile acidity vs pH)
Red	-0,683	0,235
White	-0,504	-0,097

Tab.2 – Correlation values of for fixed acidity vs pH and volatile acidity vs pH

Dependance of the pH on the fixed acidity

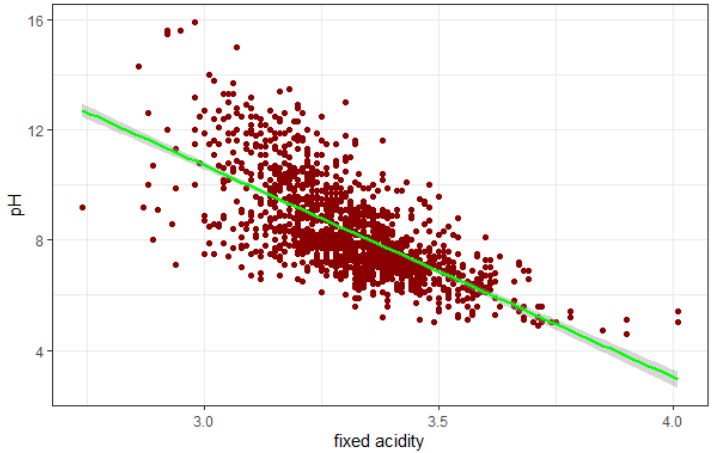

Fig.2 – Fixed acidity in red wines for values of pH

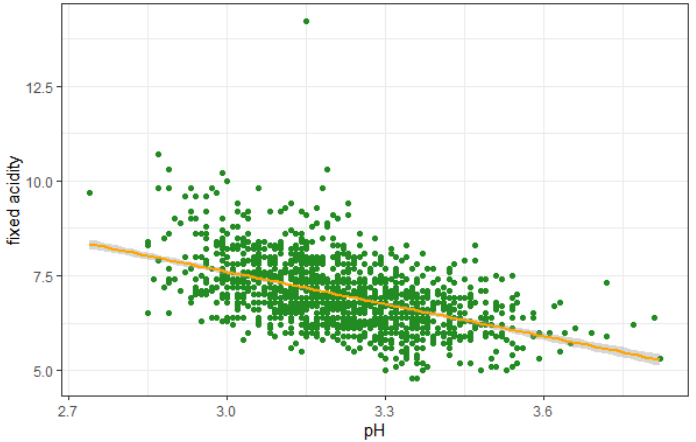

Fig.2 – Fixed acidity in white wines for values of pH

Dependance of the pH on the volatile acidity

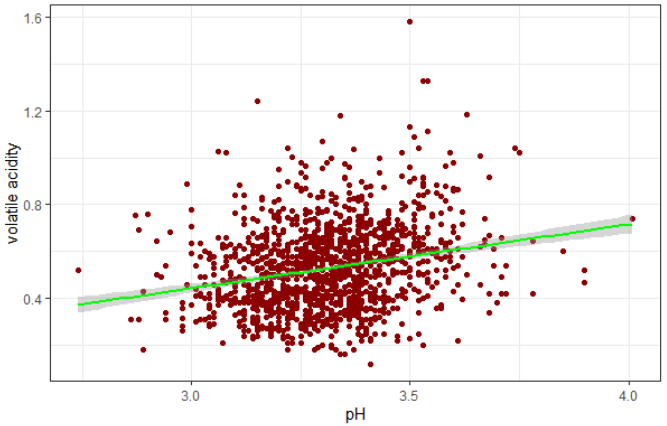

Fig.2 – Volatile acidity in red wines for values of pH

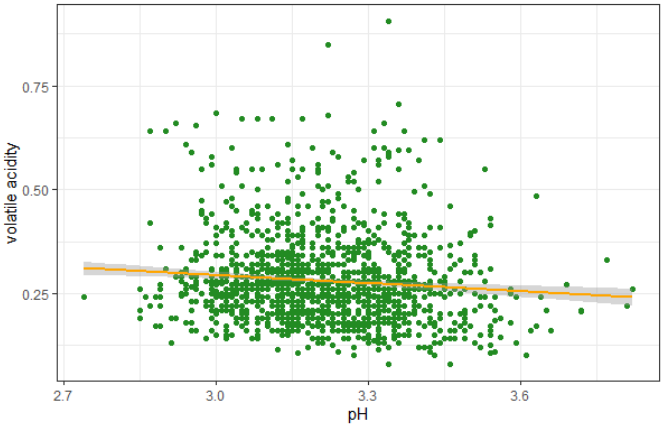

Fig.2 – Volatile acidity in white wines for values of pH

Dependance of the pH on the fixed acidity

Type	a	δa	Intercept b	δb	R^2	P-value	F-value
Red	-7,702	0,206	33,823	0,683	0,466	$2.200x10^{-16}$	1396
White	-2,828	0,121	16,086	0,390	0,253	$2.200x10^{-16}$	543,1

Tab.2 – Regression coefficients for the fixed acidity in terms of the values of pH

Dependance of the pH on the volatile acidity

Type	a	δa	Intercept b	δb	R^2	P-value	F-value
Red	0.272	0.028	-0.374	0.0935	0.0546	$2.200x10^{-16}$	93.3
White	-0.0644	0.0164	0.487	0.0527	0,00891	$9,225x10^{-5}$	15,37

Tab.2 – Regression coefficients for the volatile acidity in terms of the values of pH

What parameter has the greatest influence the global quality of the wine?

A. LM Regression

Standard Least Squares Method

B. Elastic Net Regression:

1. Look for best parameter alpha through K-fold cross validation;
2. Use optimized parameter to finally find the coefficients of the regressions.

LM REGRESSION

```
Residuals:
      Min        1Q    Median        3Q       Max
-2.68911 -0.36652 -0.04699  0.45202  2.02498

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.197e+01  2.119e+01   1.036   0.3002
fixed_acidity_red     2.499e-02  2.595e-02   0.963   0.3357
volatile_acidity_red -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
citric_acid_red      -1.826e-01  1.472e-01  -1.240   0.2150
chlorides_red        -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
residual_sugar_red    1.633e-02  1.500e-02   1.089   0.2765
free_sulfur_dioxide_red  4.361e-03  2.171e-03   2.009   0.0447 *
total_sulfur_dioxide_red -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
density_red          -1.788e+01  2.163e+01  -0.827   0.4086
pH_red               -4.137e-01  1.916e-01  -2.159   0.0310 *
sulphates_red         9.163e-01  1.143e-01   8.014 2.13e-15 ***
alcohol_red           2.762e-01  2.648e-02  10.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.648 on 1587 degrees of freedom
Multiple R-squared:  0.3606,    Adjusted R-squared:  0.3561
F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Fig.2 – Regression coefficients for the quality of red wines (lm regression)

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.6088 -0.5130 -0.0316  0.4988  3.0250

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.613e+02  4.076e+01   6.410 1.91e-10 ***
fixed_acidity_white      1.714e-01  4.153e-02   4.128 3.85e-05 ***
volatile_acidity_white  -1.705e+00  2.053e-01  -8.308  < 2e-16 ***
citric_acid_white       -9.521e-02  1.605e-01  -0.593   0.5532
chlorides_white         -9.925e-01  9.347e-01  -1.062   0.2885
residual_sugar_white     1.066e-01  1.531e-02   6.963 4.86e-12 ***
free_sulfur_dioxide_white  7.406e-03  1.713e-03   4.323 1.64e-05 ***
total_sulfur_dioxide_white -7.132e-05  6.758e-04  -0.106   0.9160
density_white           -2.644e+02  4.131e+01  -6.400 2.04e-10 ***
pH_white                 1.435e+00  2.035e-01   7.055 2.58e-12 ***
sulphates_white          8.578e-01  1.894e-01   4.530 6.35e-06 ***
alcohol_white            8.678e-02  5.043e-02   1.721   0.0855 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7888 on 1587 degrees of freedom
Multiple R-squared:  0.2941,    Adjusted R-squared:  0.2892
F-statistic: 60.12 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Fig.2 – Regression coefficients for the quality of white wines (lm regression)

ELASTIC NET REGRESSION

For α = 0,1:

	s0
(Intercept)	3.679814e+01
fixed_acidity_red	2.657825e-02
volatile_acidity_red	-7.475593e-01
citric_acid_red	1.927443e-01
chlorides_red	-1.245722e+00
residual_sugar_red	1.724735e-02
free_sulfur_dioxide_red	2.761038e-04
total_sulfur_dioxide_red	-2.092591e-03
density_red	-3.273324e+01
pH_red	-1.334303e-01
sulphates_red	5.734417e-01
alcohol_red	1.738083e-01

>

Tab.2 – Regression coefficients for the quality of red wines (elastic net regression)

For α = 0,9:

	s0
(Intercept)	0.812864741
fixed_acidity_white	-0.001180688
volatile_acidity_white	-1.335573342
citric_acid_white	.
chlorides_white	-0.270708187
residual_sugar_white	0.001314539
free_sulfur_dioxide_white	0.005592644
total_sulfur_dioxide_white	.
density_white	.
pH_white	0.634324712
sulphates_white	0.063973151
alcohol_white	0.308991252

>

Tab.2 – Regression coefficients for the quality of white wines (elastic net regression)

ELASTIC NET REGRESSION

Parameters that influence the most:

- **Sulphates (+)**
- **Density (-)**
- **Volatile Acidity (-)**
- **pH (for the white wine) (+)**
- **Chlorides (-)**

Conclusion