Impact of Physicochemical Variables on Wine Quality

02.03.2023

Agenda

Topics Covered

1. About Wine Datasets
2. Findings & Observations
3. Conclusion and next step

Red Wine
&
White Wine

Vinho Verde and CVRVV

- The data sets contain information about the physicochemical variables of both red and white Vinho Verde wine samples.
- Vinho Verde is a unique and distinctive wine region located in the northwestern corner of Portugal. The region is known for producing light and fresh wines that are typically consumed young.
- Were tested at the official certification entity (CVRVV). The CVRVV is an inter-professional organization with the goal of improving the quality and marketing of Vinho Verde.

Red wine Dataset

Rows and Columns

1599, 12

Were collected

2004-2007

Missing Values

Zero

Outliers

Yes

New added column(s)

quality_label

Removed part(s)

total_sulfur_dioxide > 160

White wine Dataset

Rows and Columns

4898, 12

Were collected

2004-2007

Missing Values

Zero

Outliers

Yes

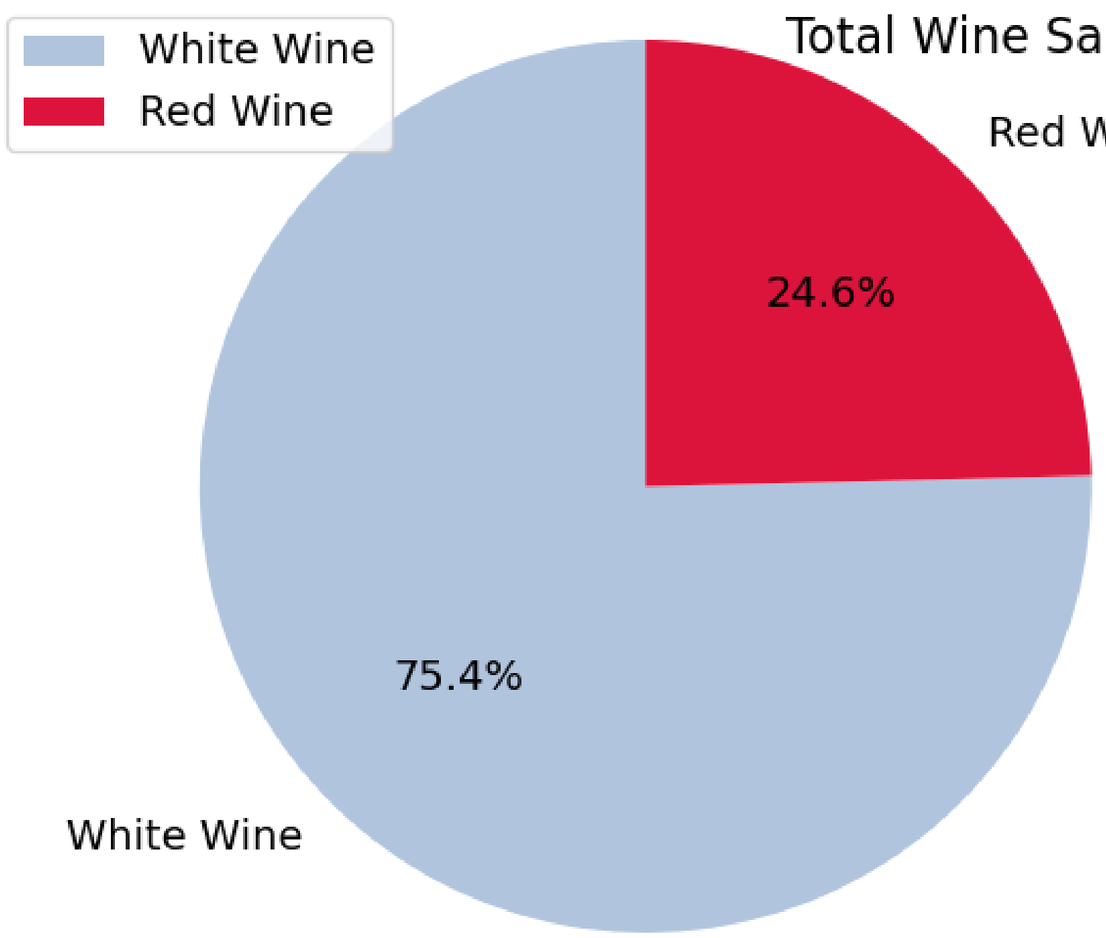
New added column(s)

quality_label

Removed part(s)

total_sulfur_dioxide > 210

Wine Types

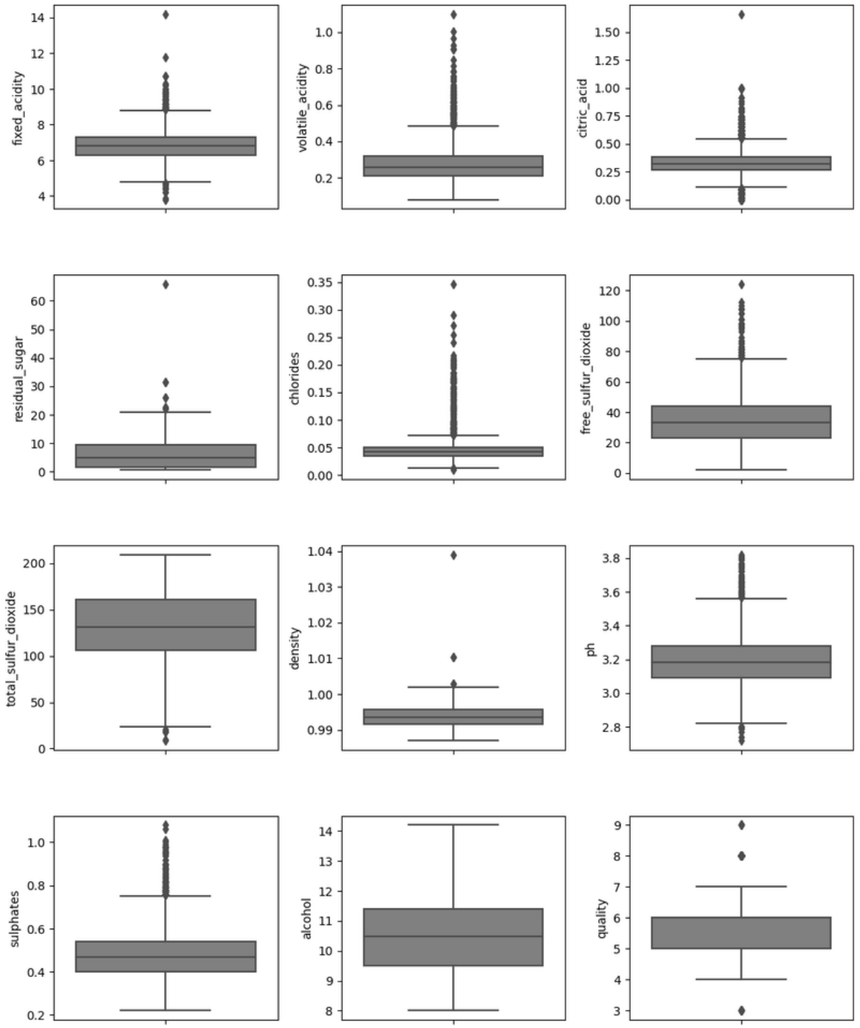


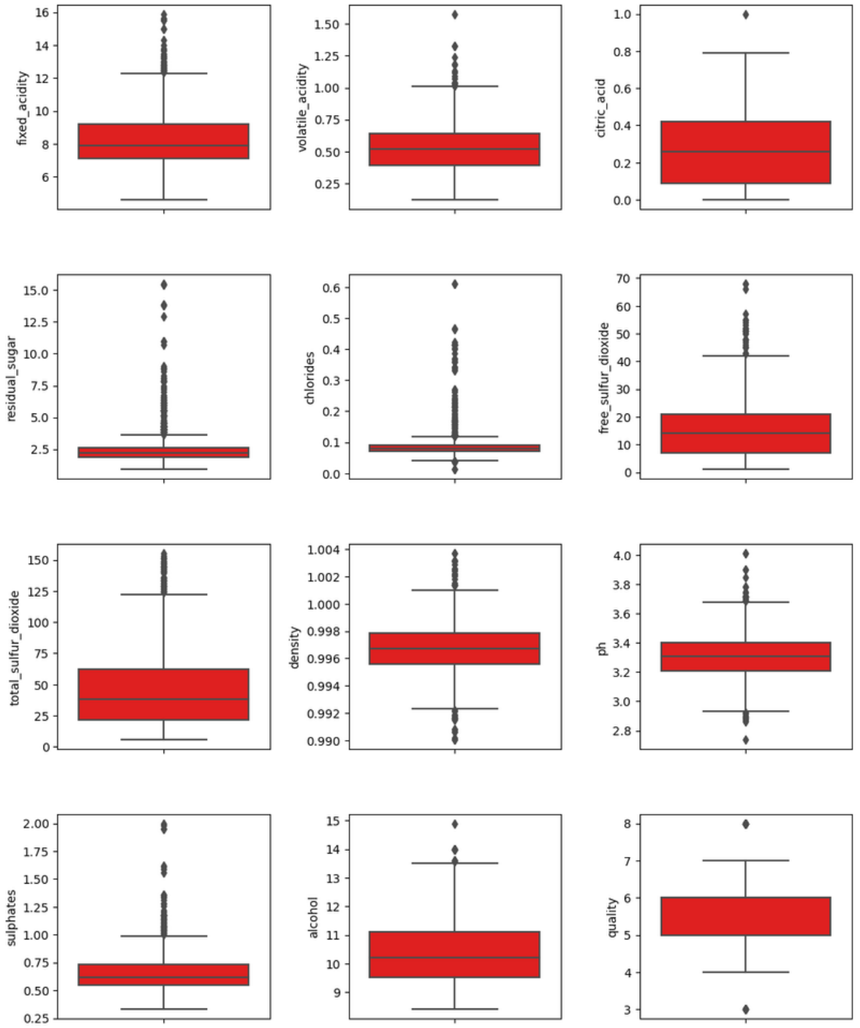
- Red wines are loved for their soft rich and velvety flavors and white wines are valued for their zesty acidic flavor.
- We have more white wines compared to red wines since the Minho region where Vinho Verde comes from is quite cool and often rainy making it hard to ripen red grapes.

Physicochemical Variables

1. Fixed Acidity
2. Volatile Acidity
3. Citric Acid
4. Residual Sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. Sulphates
10. PH
11. Alcohol
12. Quality

Outliers on both Datasets



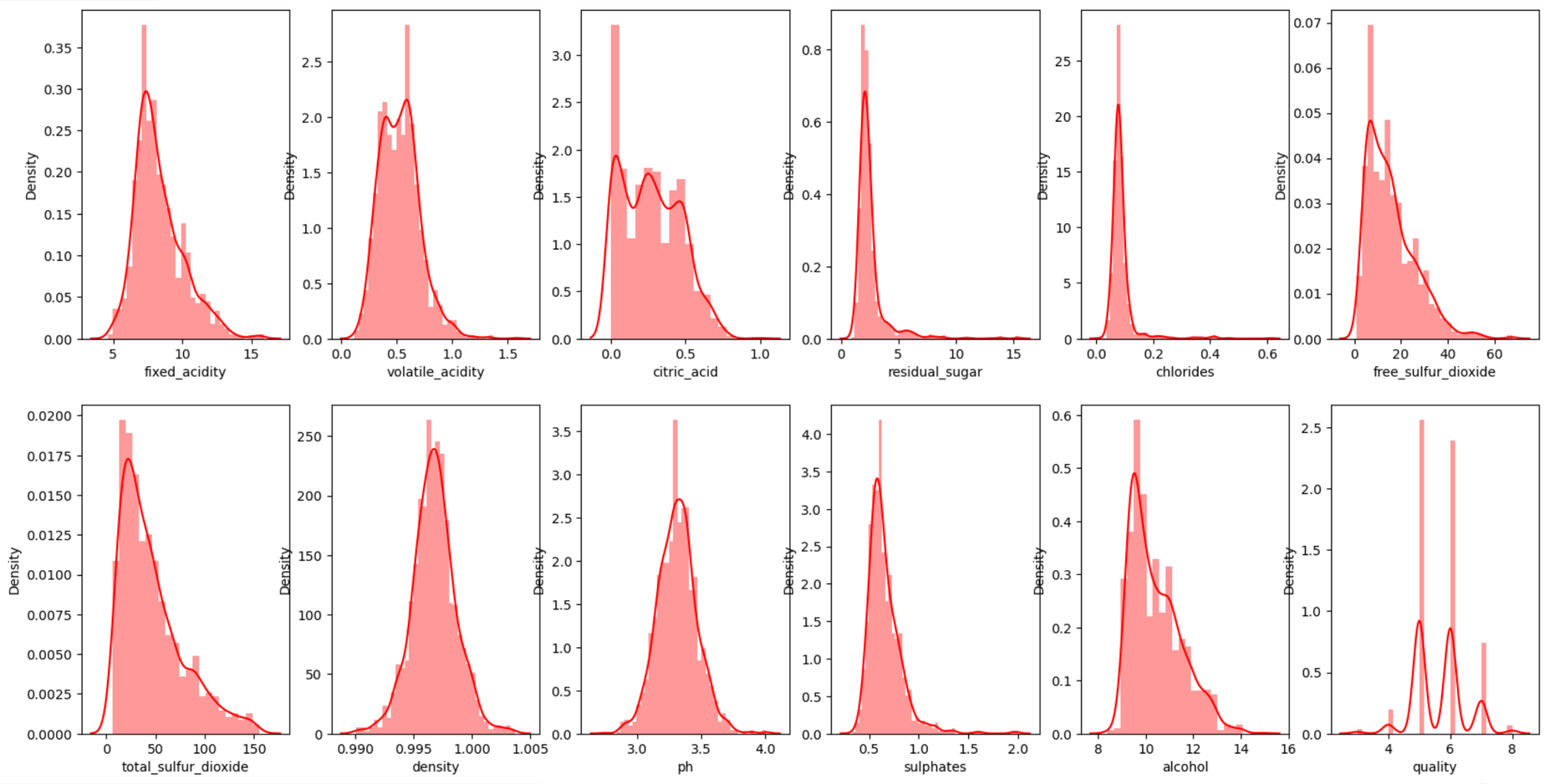


From the above box plots, we can clearly see that there are outliers in all features.

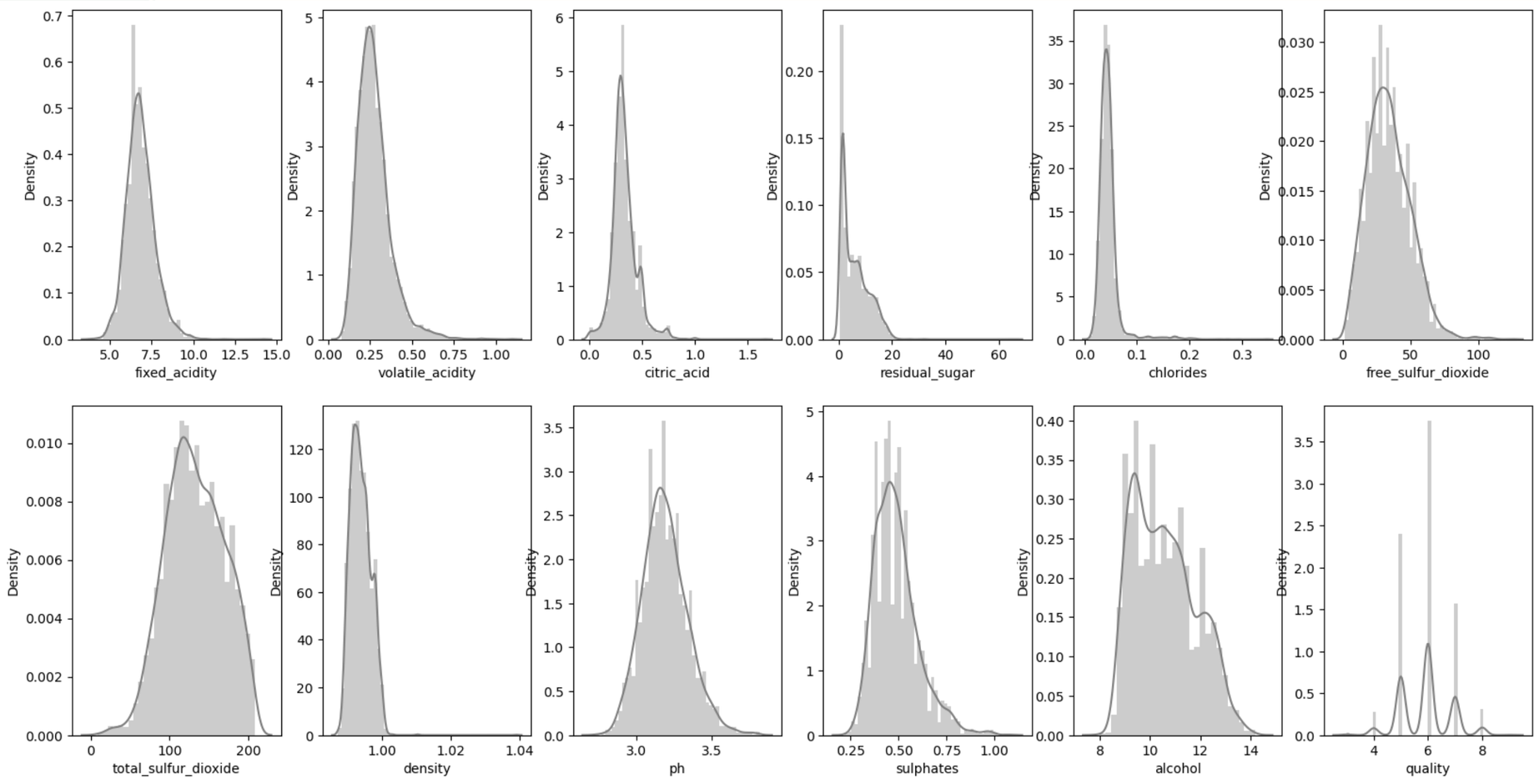
BUT

Here I am choosing not to remove/modify outliers as we are looking for accuracy to minute levels, not just some approximation — high-quality wine may have a very rare composition (hence outlier) from other average-quality wines, so we can not remove or modify outlier values in our dataset.

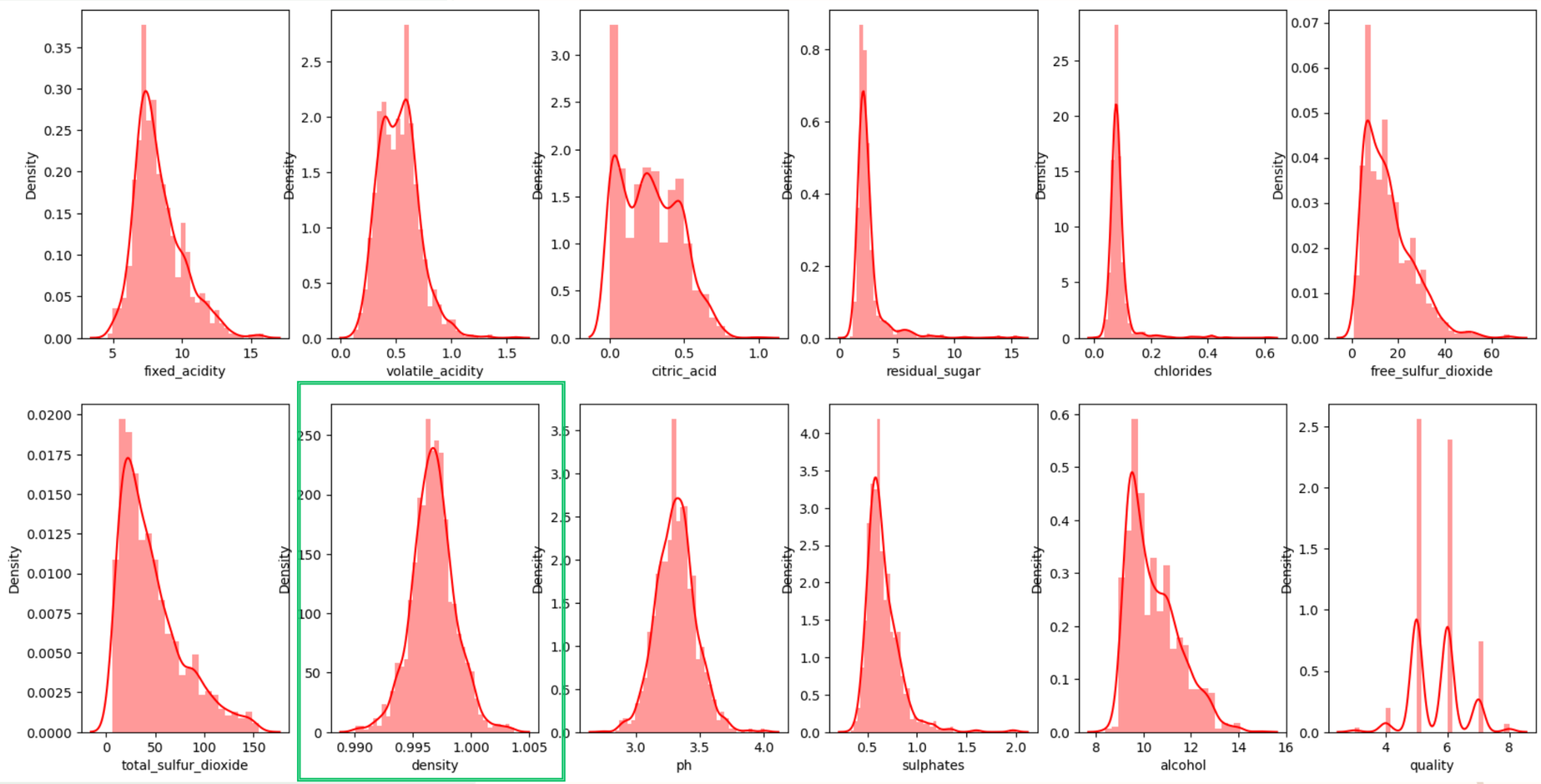
Data Distribution: Red wine



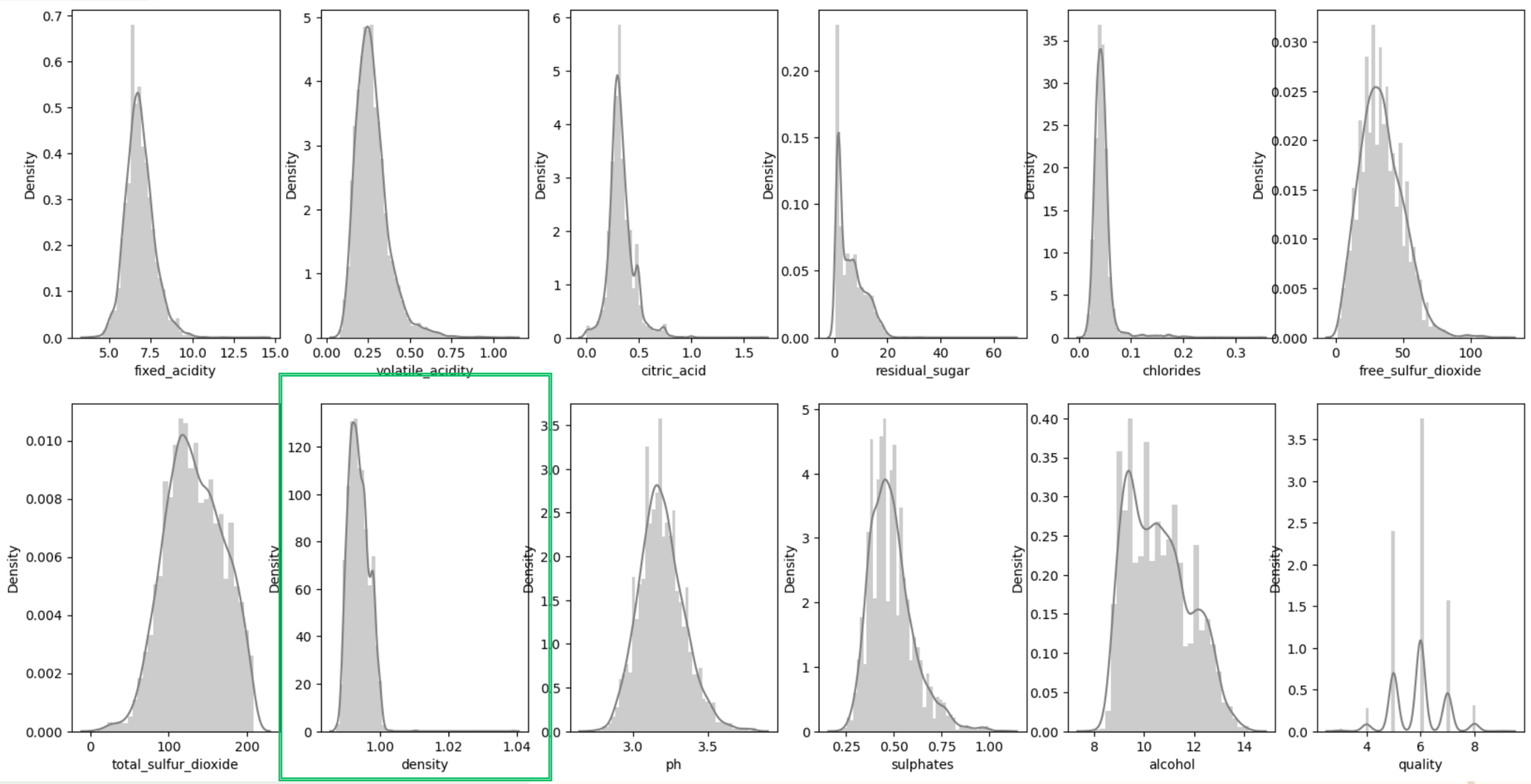
Data Distribution: White wine



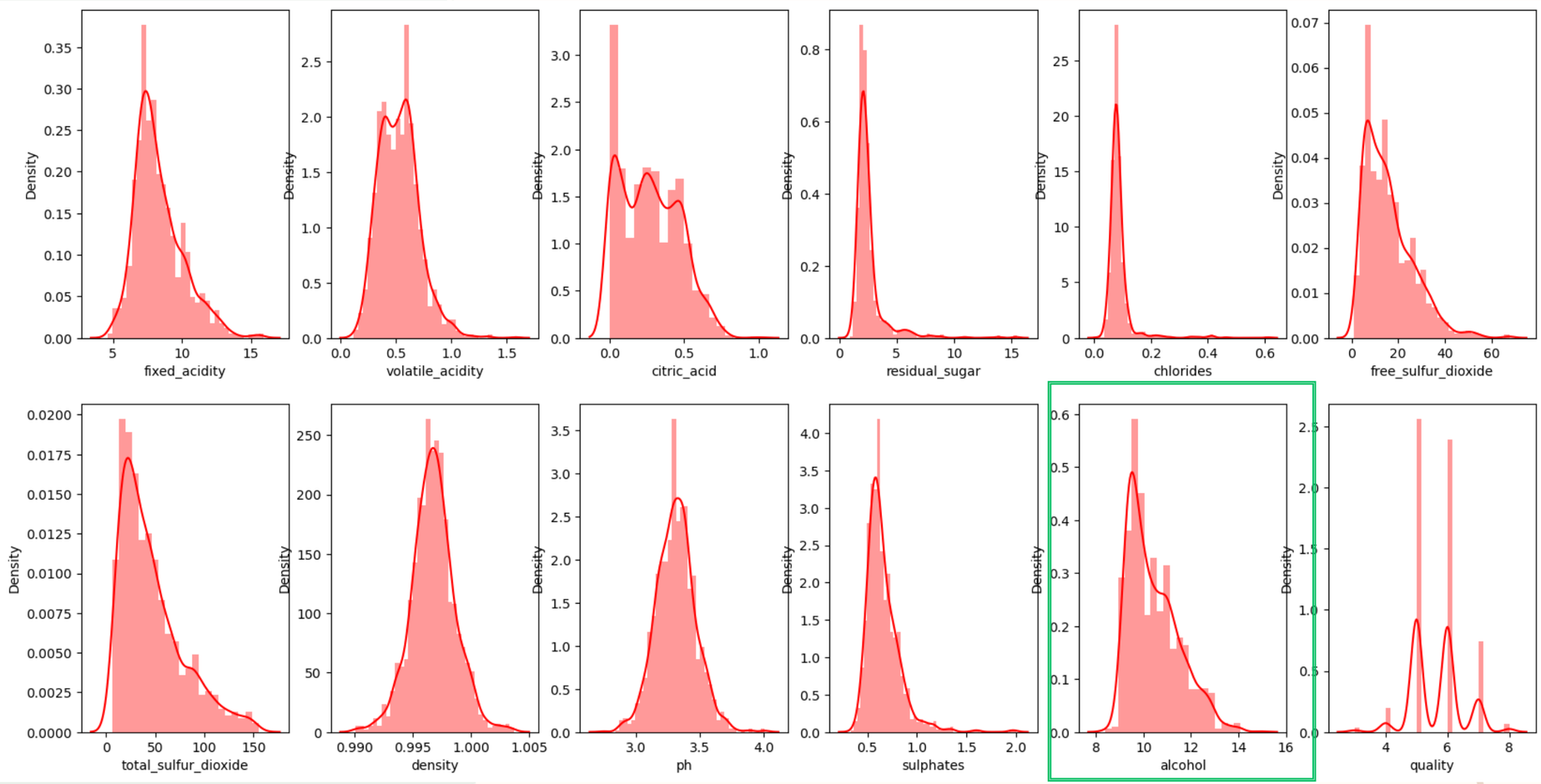
Data Distribution: Red wine



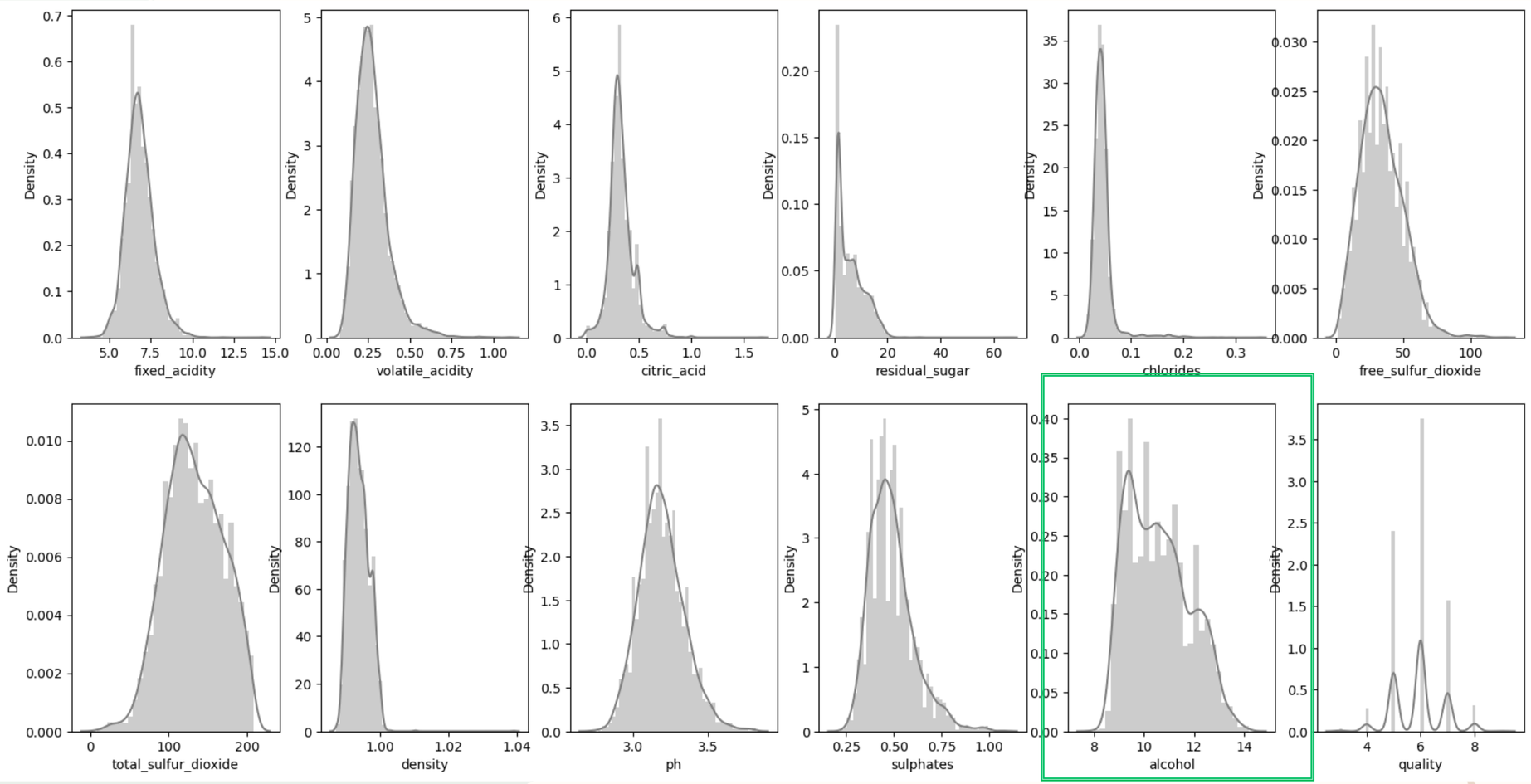
Data Distribution: White wine



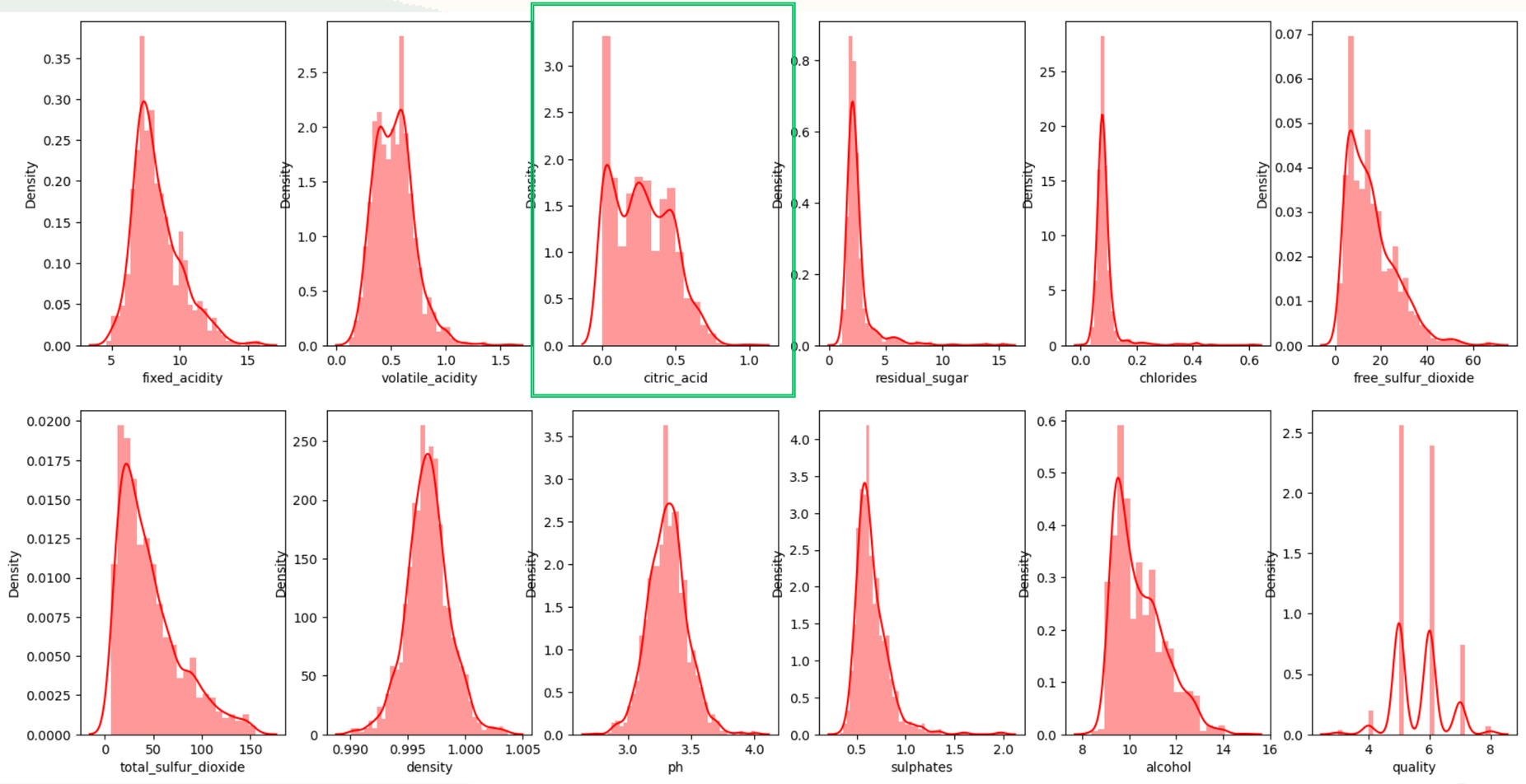
Data Distribution: Red wine



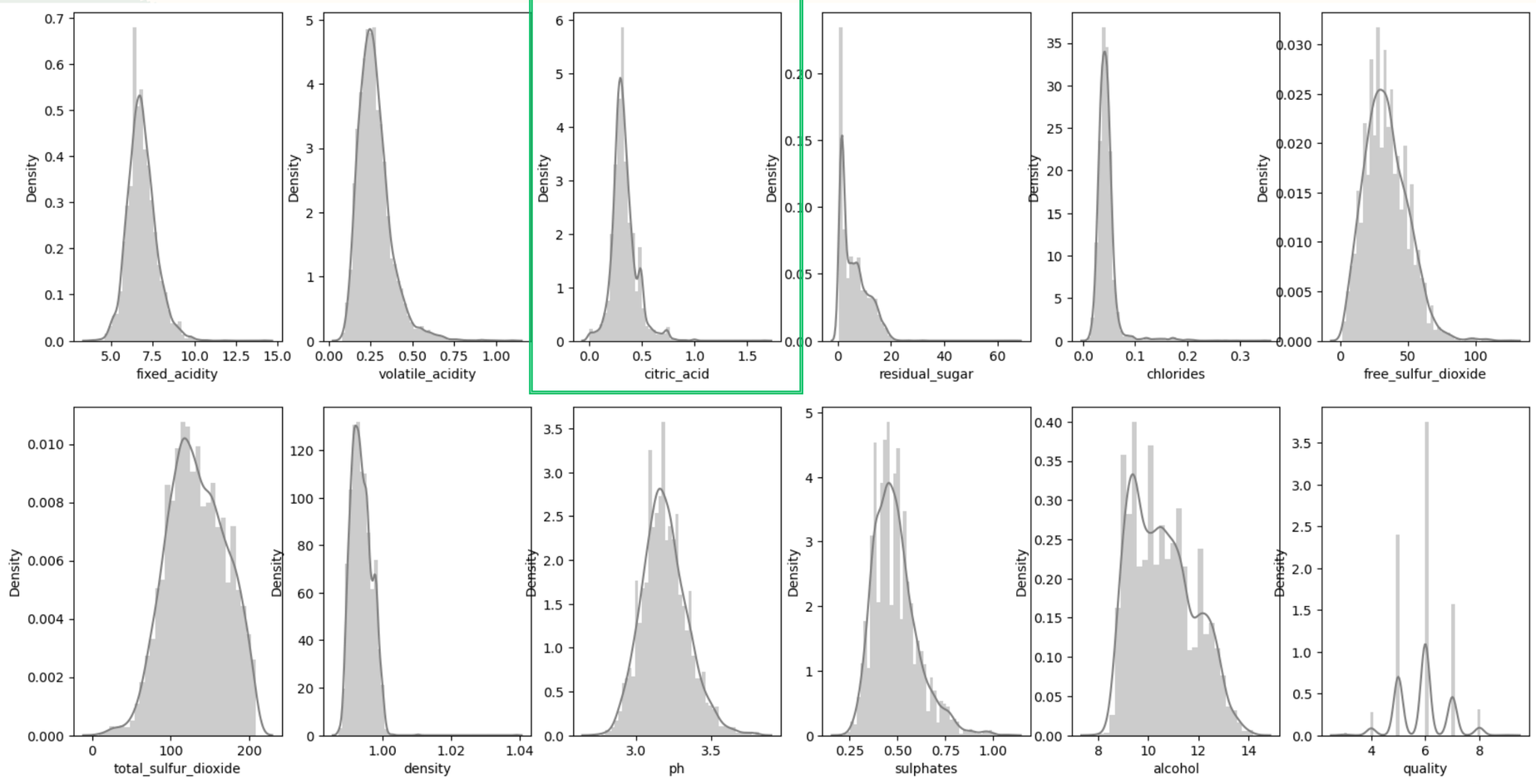
Data Distribution: White wine



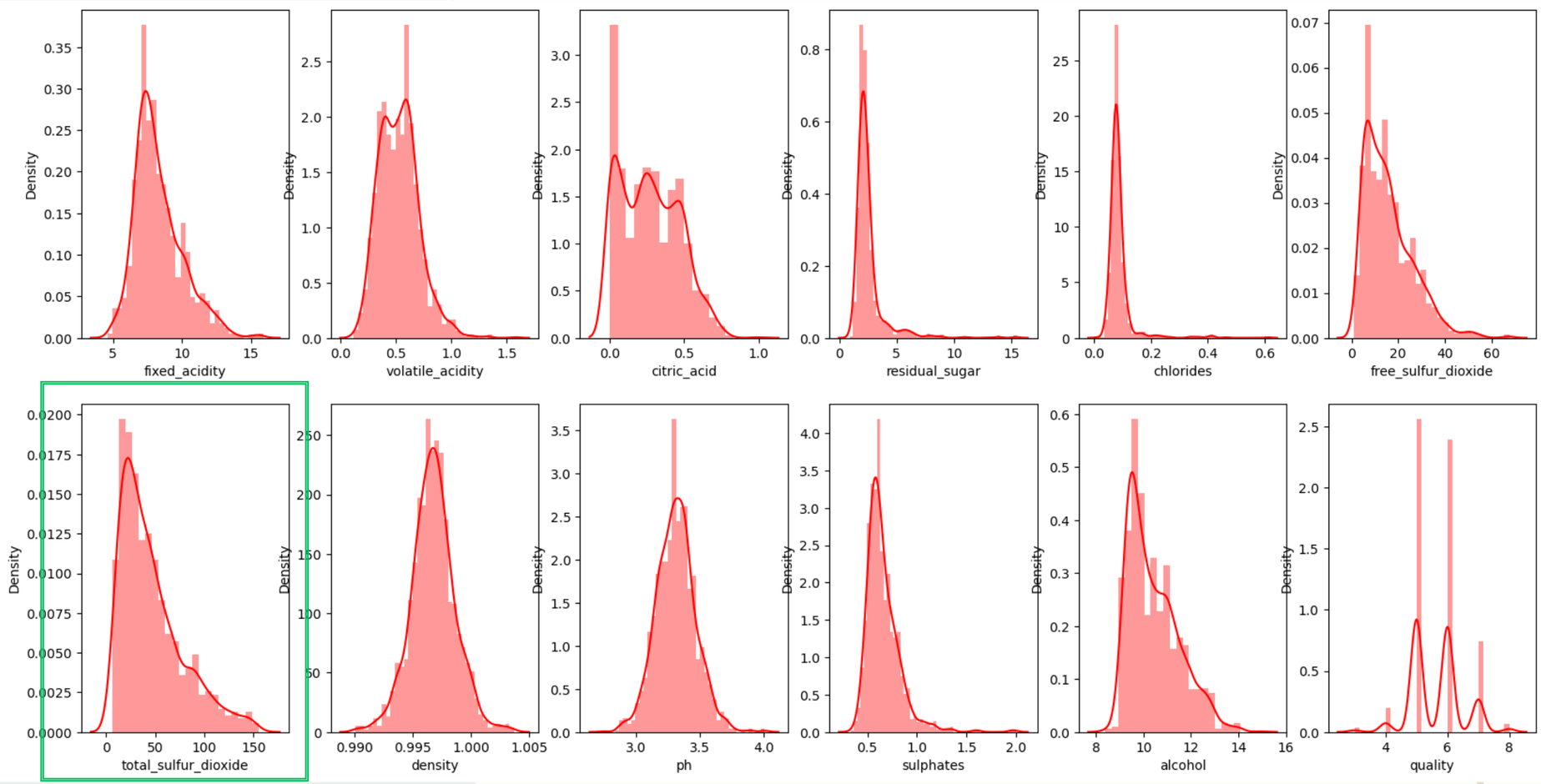
Data Distribution: Red wine



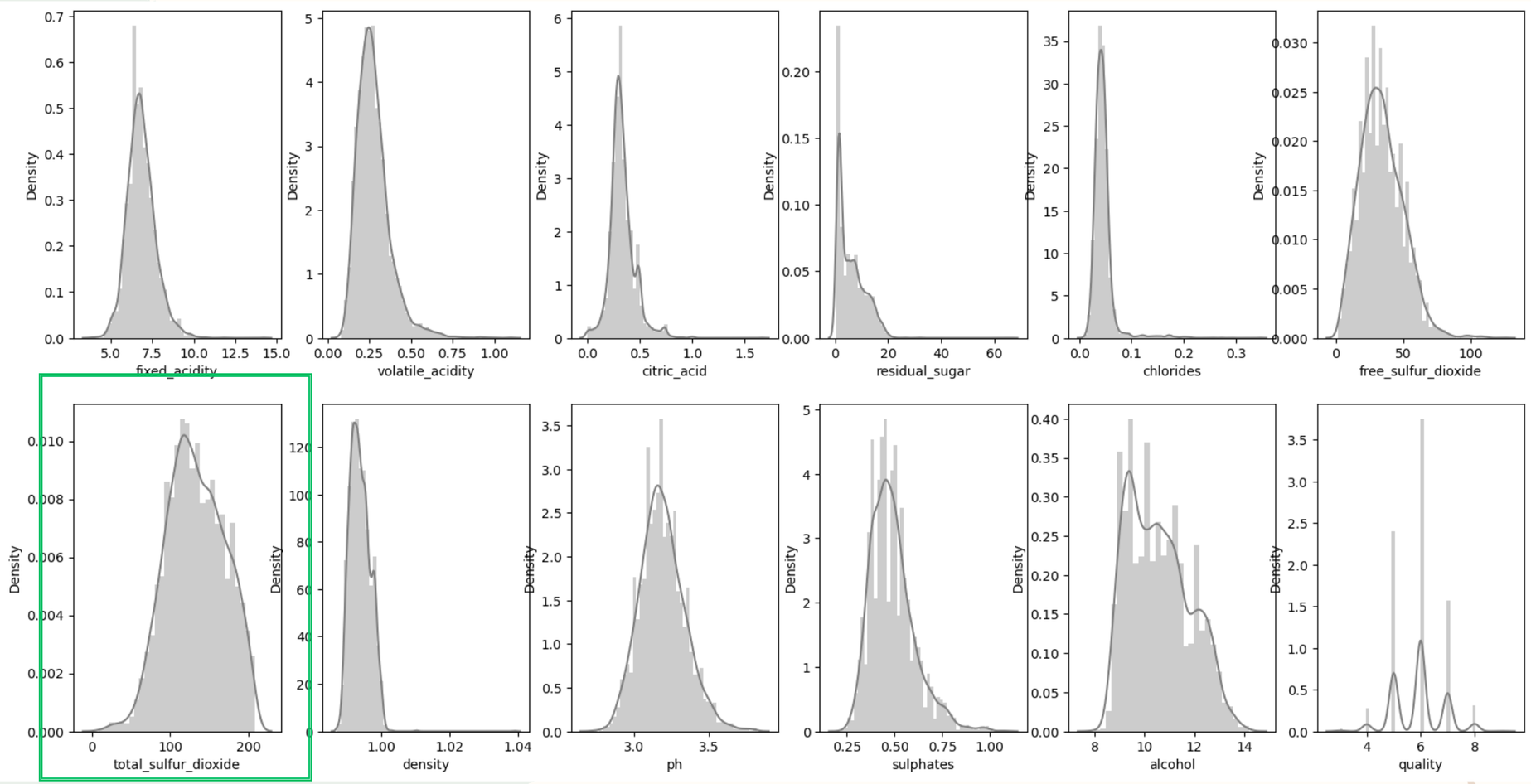
Data Distribution: White wine



Data Distribution: Red wine



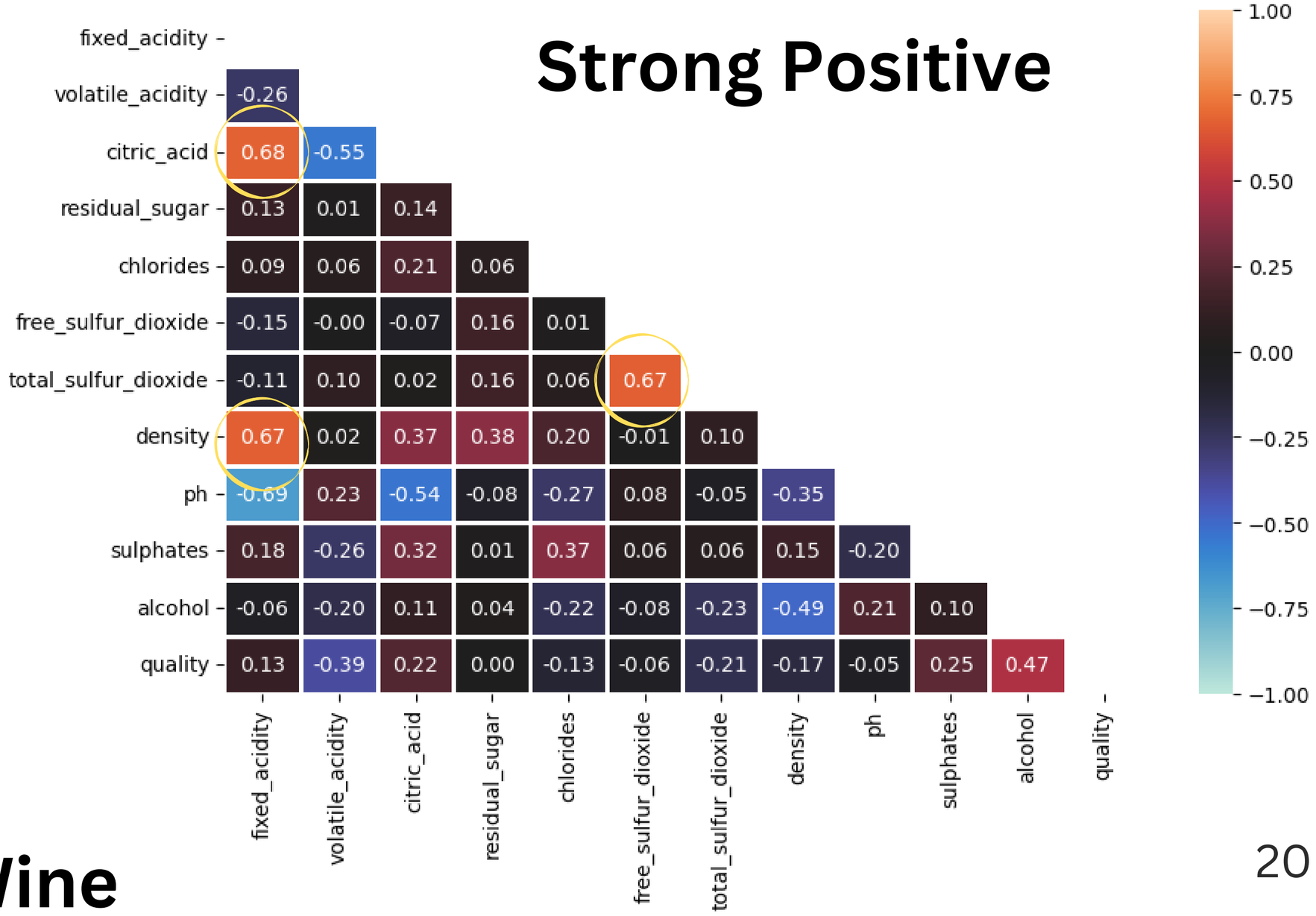
Data Distribution: White wine



From the graph, it appears that the distributions of some variables differ between red and white wines. These differences suggest that there may be some underlying differences in the production or composition of red and white wines that affect these variables differently.

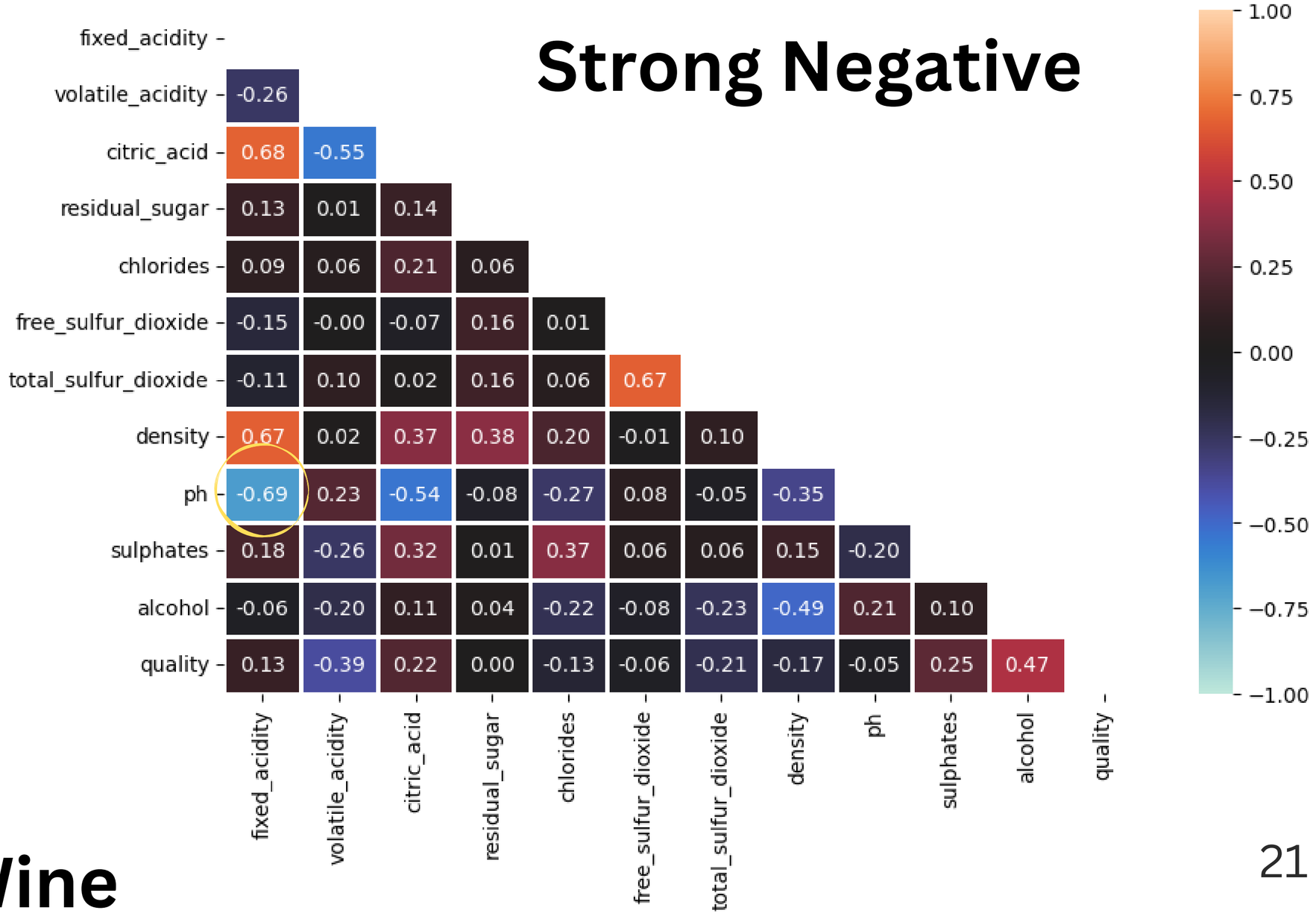
Let's have a look at some Correlations between Physicochemical Variables

Strong Positive



Red Wine

Strong Negative



Red Wine

Red Wine

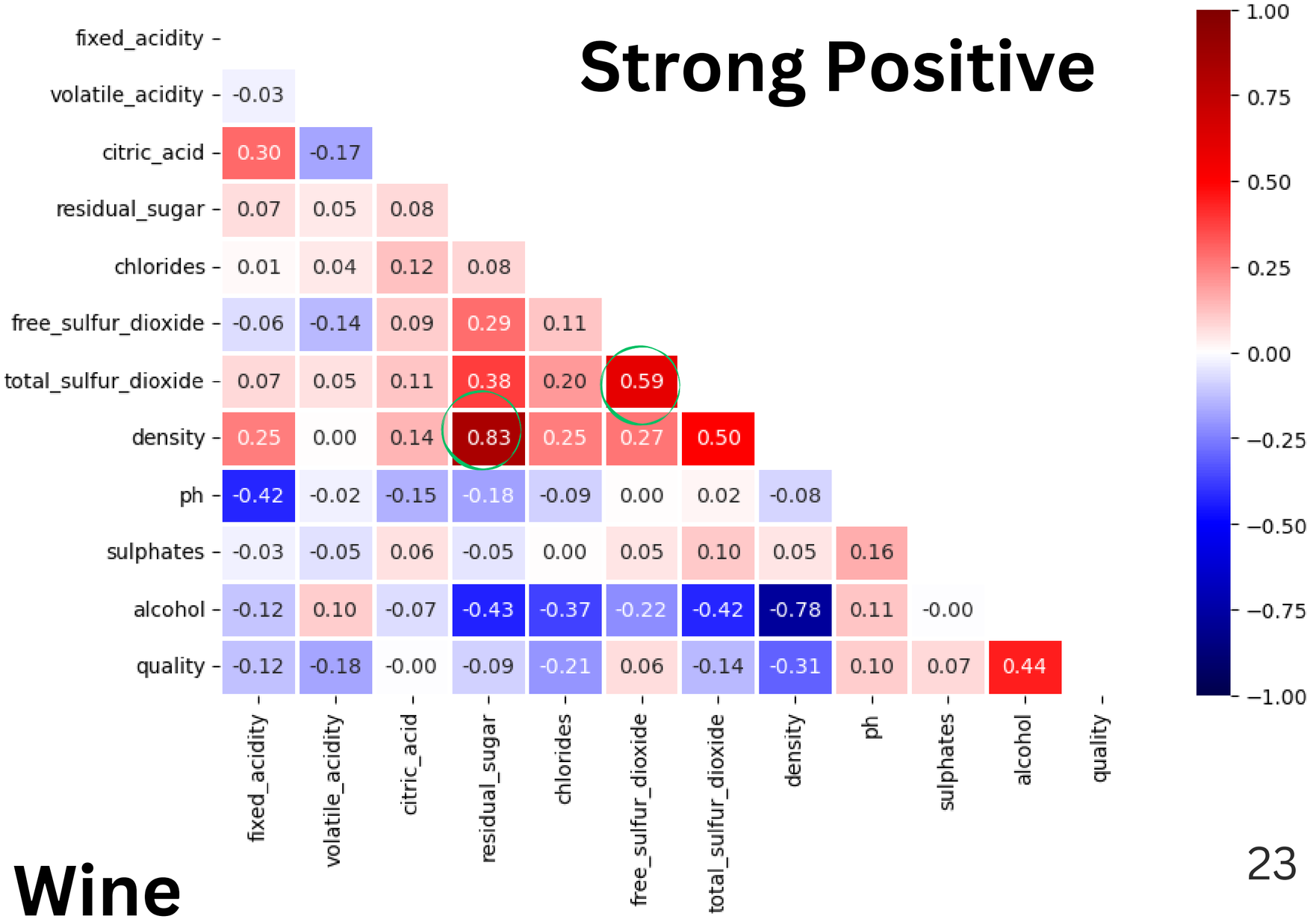
Positive Correlation

1. fixed acidity and density
2. fixed acidity and citric acid
3. total sulfur dioxide and free sulfur dioxide

Negative Correlation

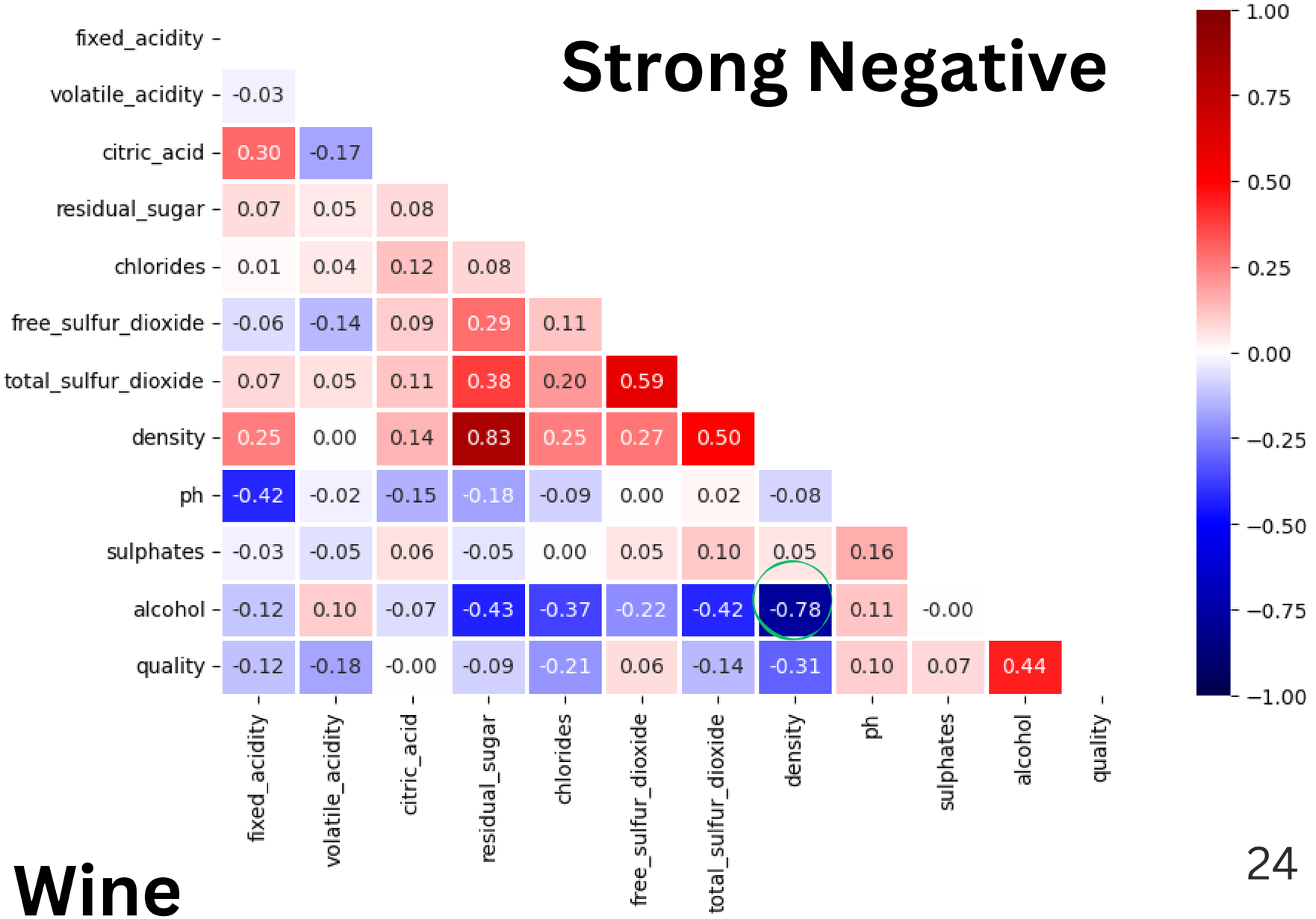
1. fixed acidity and ph

Strong Positive



White Wine

Strong Negative



White Wine

White Wine

Positive Correlations

1. residual sugar and density
2. total sulfur dioxide and free sulfur dioxide

Negative Correlations

1. density and alcohol

Mean table: Red Wine

	Low Quality Wine	Medium Quality Wine	High Quality Wine
	mean	mean	mean
alcohol	9.93	10.82	12.09
volatile_acidity	0.59	0.48	0.42
ph	3.31	3.31	3.27
fixed_acidity	8.14	8.48	8.57
citric_acid	0.24	0.30	0.39
residual_sugar	2.54	2.51	2.58
chlorides	0.09	0.08	0.07
free_sulfur_dioxide	16.57	15.16	13.28
total_sulfur_dioxide	54.65	38.60	33.44
density	1.00	1.00	1.00
sulphates	0.62	0.69	0.77

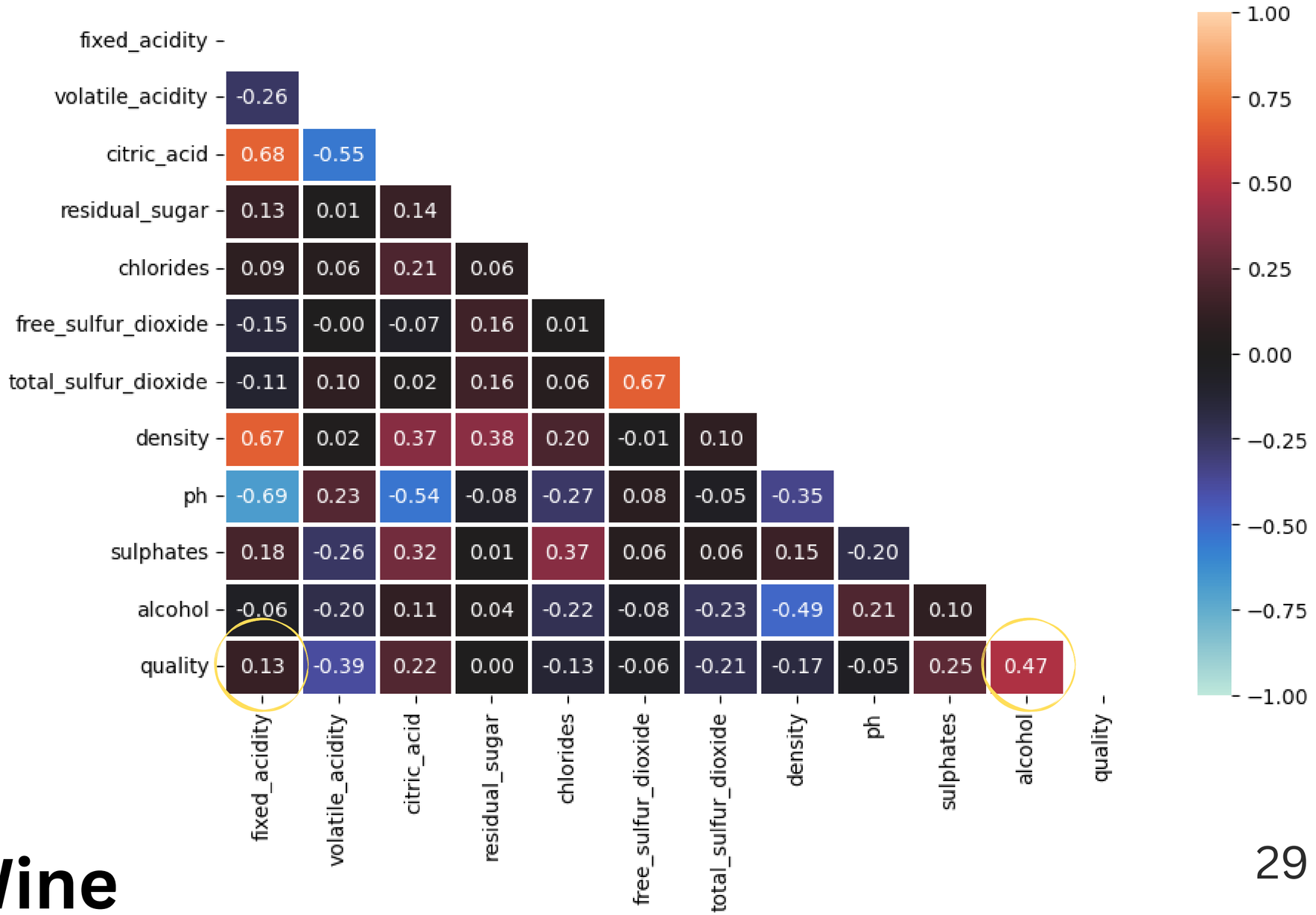
Mean table: Red Wine

	Low Quality Wine	Medium Quality Wine	High Quality Wine
	mean	mean	mean
alcohol	9.93	10.82	12.09
volatile_acidity	0.59	0.48	0.42
ph	3.31	3.31	3.27
fixed_acidity	8.14	8.48	8.57
citric_acid	0.24	0.30	0.39
residual_sugar	2.54	2.51	2.58
chlorides	0.09	0.08	0.07
free_sulfur_dioxide	16.57	15.16	13.28
total_sulfur_dioxide	54.65	38.60	33.44
density	1.00	1.00	1.00
sulphates	0.62	0.69	0.77

Mean table: Red Wine

	Low Quality Wine	Medium Quality Wine	High Quality Wine
	mean	mean	mean
alcohol	9.93	10.82	12.09
volatile_acidity	0.59	0.48	0.42
ph	3.31	3.31	3.27
fixed_acidity	8.14	8.48	8.57
citric_acid	0.24	0.30	0.39
residual_sugar	2.54	2.51	2.58
chlorides	0.09	0.08	0.07
free_sulfur_dioxide	16.57	15.16	13.28
total_sulfur_dioxide	54.65	38.60	33.44
density	1.00	1.00	1.00
sulphates	0.62	0.69	0.77

Red Wine


fixed_acidity
volatile_acidity -0.26
citric_acid 0.68 -0.55
residual_sugar 0.13 0.01 0.14
chlorides 0.09 0.06 0.21 0.06
free_sulfur_dioxide -0.15 -0.00 -0.07 0.16 0.01
total_sulfur_dioxide -0.11 0.10 0.02 0.16 0.06 0.67
density 0.67 0.02 0.37 0.38 0.20 -0.01 0.10
ph -0.69 0.23 -0.54 -0.08 -0.27 0.08 -0.05 -0.35
sulphates 0.18 -0.26 0.32 0.01 0.37 0.06 0.06 0.15 -0.20
alcohol -0.06 -0.20 0.11 0.04 -0.22 -0.08 -0.23 -0.49 0.21 0.10
quality 0.13 -0.39 0.22 0.00 -0.13 -0.06 -0.21 -0.17 -0.05 0.25 0.47
fixed_acidity
volatile_acidity
citric_acid
residual_sugar
chlorides
free_sulfur_dioxide
total_sulfur_dioxide
density
ph
sulphates
alcohol
quality
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

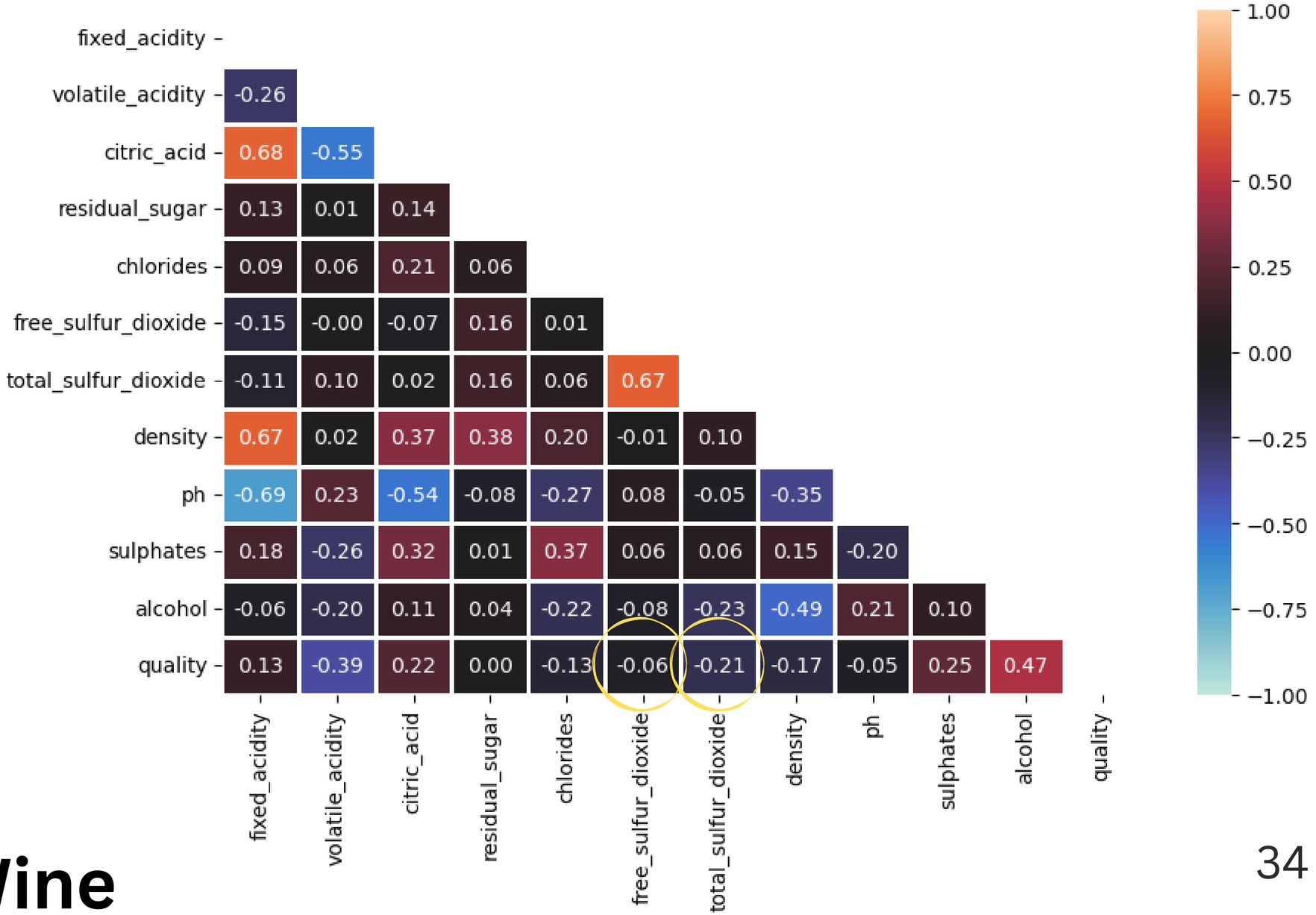
Mean table: Red Wine

	Low Quality Wine	Medium Quality Wine	High Quality Wine
	mean	mean	mean
alcohol	9.93	10.82	12.09
volatile_acidity	0.59	0.48	0.42
ph	3.31	3.31	3.27
fixed_acidity	8.14	8.48	8.57
citric_acid	0.24	0.30	0.39
residual_sugar	2.54	2.51	2.58
chlorides	0.09	0.08	0.07
free_sulfur_dioxide	16.57	15.16	13.28
total_sulfur_dioxide	54.65	38.60	33.44
density	1.00	1.00	1.00
sulphates	0.62	0.69	0.77

Mean table: Red Wine

	Low Quality Wine	Medium Quality Wine	High Quality Wine
	mean	mean	mean
alcohol	9.93	10.82	12.09
volatile_acidity	0.59	0.48	0.42
ph	3.31	3.31	3.27
fixed_acidity	8.14	8.48	8.57
citric_acid	0.24	0.30	0.39
residual_sugar	2.54	2.51	2.58
chlorides	0.09	0.08	0.07
free_sulfur_dioxide	16.57	15.16	13.28
total_sulfur_dioxide	54.65	38.60	33.44
density	1.00	1.00	1.00
sulphates	0.62	0.69	0.77

Red Wine


fixed_acidity
volatile_acidity -0.26
citric_acid 0.68 -0.55
residual_sugar 0.13 0.01 0.14
chlorides 0.09 0.06 0.21 0.06
free_sulfur_dioxide -0.15 -0.00 -0.07 0.16 0.01
total_sulfur_dioxide -0.11 0.10 0.02 0.16 0.06 0.67
density 0.67 0.02 0.37 0.38 0.20 -0.01 0.10
ph -0.69 0.23 -0.54 -0.08 -0.27 0.08 -0.05 -0.35
sulphates 0.18 -0.26 0.32 0.01 0.37 0.06 0.06 0.15 -0.20
alcohol -0.06 -0.20 0.11 0.04 -0.22 -0.08 -0.23 -0.49 0.21 0.10
quality 0.13 -0.39 0.22 0.00 -0.13 -0.06 -0.21 -0.17 -0.05 0.25 0.47
fixed_acidity
volatile_acidity
citric_acid
residual_sugar
chlorides
free_sulfur_dioxide
total_sulfur_dioxide
density
ph
sulphates
alcohol
quality
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

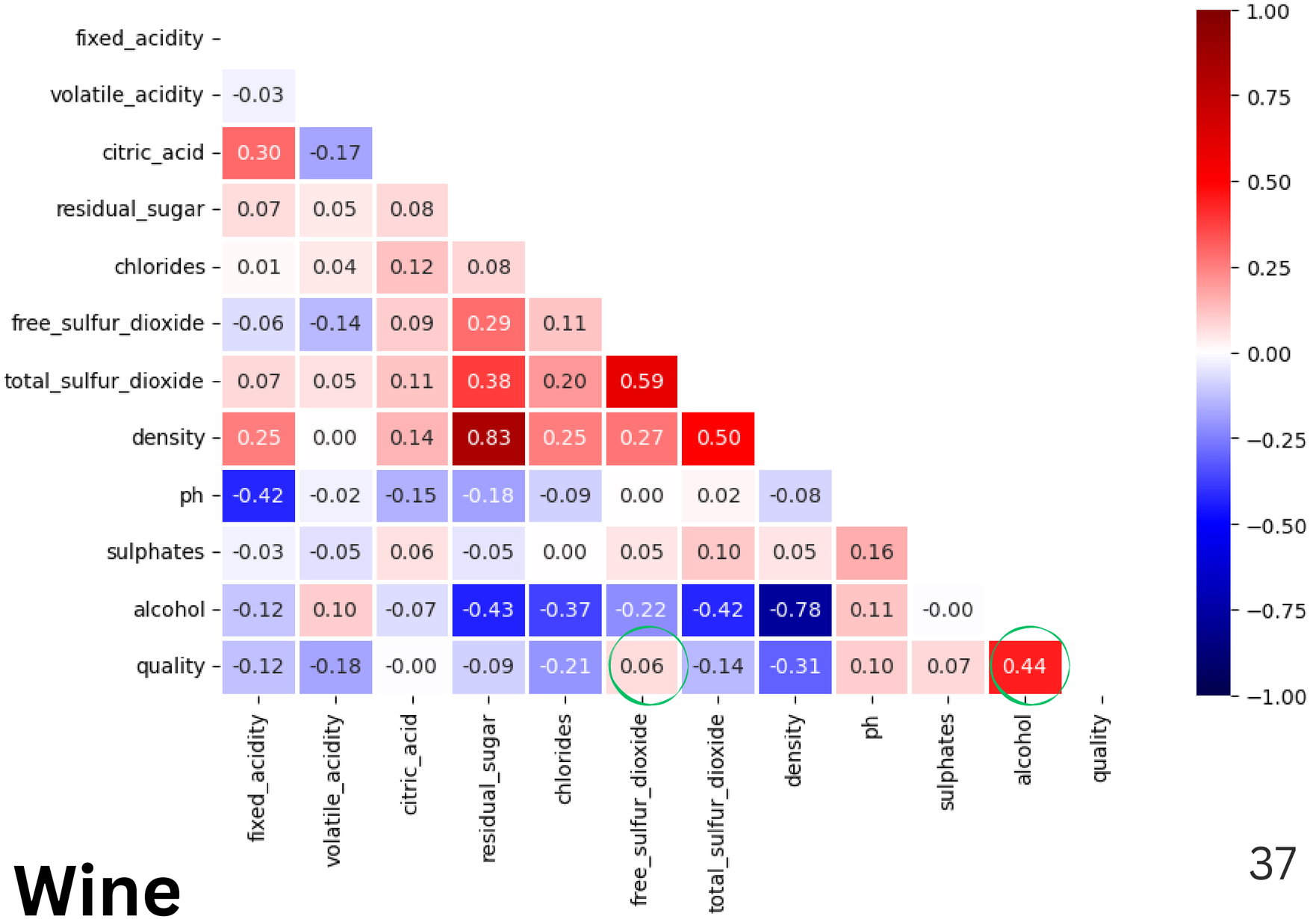
Mean table: White Wine

	Low Quality Wine	Medium Quality Wine	High Quality Wine
	mean	mean	mean
alcohol	9.90	10.85	11.75
volatile_acidity	0.31	0.26	0.28
ph	3.17	3.20	3.22
fixed_acidity	6.96	6.80	6.65
citric_acid	0.33	0.33	0.33
residual_sugar	6.73	5.91	5.33
chlorides	0.05	0.04	0.04
free_sulfur_dioxide	33.17	34.50	36.06
total_sulfur_dioxide	140.72	129.36	122.90
density	0.99	0.99	0.99
sulphates	0.48	0.49	0.49

Mean table: White Wine

	Low Quality Wine	Medium Quality Wine	High Quality Wine
	mean	mean	mean
alcohol	9.90	10.85	11.75
volatile_acidity	0.31	0.26	0.28
ph	3.17	3.20	3.22
fixed_acidity	6.96	6.80	6.65
citric_acid	0.33	0.33	0.33
residual_sugar	6.73	5.91	5.33
chlorides	0.05	0.04	0.04
free_sulfur_dioxide	33.17	34.50	36.06
total_sulfur_dioxide	140.72	129.36	122.90
density	0.99	0.99	0.99
sulphates	0.48	0.49	0.49

White Wine


fixed_acidity
volatile_acidity -0.03
citric_acid 0.30 -0.17
residual_sugar 0.07 0.05 0.08
chlorides 0.01 0.04 0.12 0.08
free_sulfur_dioxide -0.06 -0.14 0.09 0.29 0.11
total_sulfur_dioxide 0.07 0.05 0.11 0.38 0.20 0.59
density 0.25 0.00 0.14 0.83 0.25 0.27 0.50
ph -0.42 -0.02 -0.15 -0.18 -0.09 0.00 0.02 -0.08
sulphates -0.03 -0.05 0.06 -0.05 0.00 0.05 0.10 0.05 0.16
alcohol -0.12 0.10 -0.07 -0.43 -0.37 -0.22 -0.42 -0.78 0.11 -0.00
quality -0.12 -0.18 -0.00 -0.09 -0.21 0.06 -0.14 -0.31 0.10 0.07 0.44
fixed_acidity
volatile_acidity
citric_acid
residual_sugar
chlorides
free_sulfur_dioxide
total_sulfur_dioxide
density
ph
sulphates
alcohol
quality
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

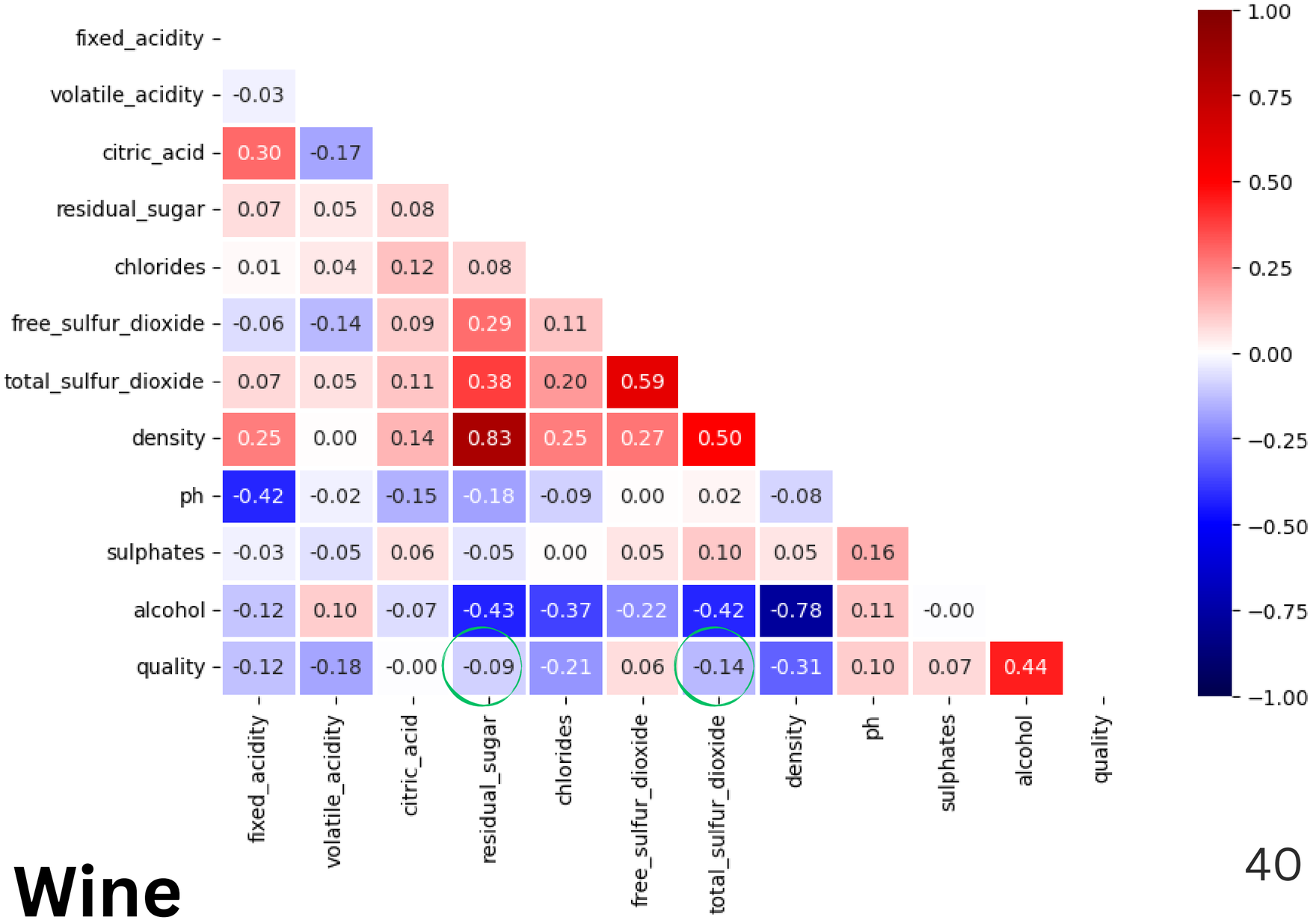
Mean table: White Wine

	Low Quality Wine	Medium Quality Wine	High Quality Wine
	mean	mean	mean
alcohol	9.90	10.85	11.75
volatile_acidity	0.31	0.26	0.28
ph	3.17	3.20	3.22
fixed_acidity	6.96	6.80	6.65
citric_acid	0.33	0.33	0.33
residual_sugar	6.73	5.91	5.33
chlorides	0.05	0.04	0.04
free_sulfur_dioxide	33.17	34.50	36.06
total_sulfur_dioxide	140.72	129.36	122.90
density	0.99	0.99	0.99
sulphates	0.48	0.49	0.49

Mean table: White Wine

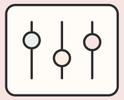
	Low Quality Wine	Medium Quality Wine	High Quality Wine
	mean	mean	mean
alcohol	9.90	10.85	11.75
volatile_acidity	0.31	0.26	0.28
ph	3.17	3.20	3.22
fixed_acidity	6.96	6.80	6.65
citric_acid	0.33	0.33	0.33
residual_sugar	6.73	5.91	5.33
chlorides	0.05	0.04	0.04
free_sulfur_dioxide	33.17	34.50	36.06
total_sulfur_dioxide	140.72	129.36	122.90
density	0.99	0.99	0.99
sulphates	0.48	0.49	0.49

White Wine


fixed_acidity
volatile_acidity -0.03
citric_acid 0.30 -0.17
residual_sugar 0.07 0.05 0.08
chlorides 0.01 0.04 0.12 0.08
free_sulfur_dioxide -0.06 -0.14 0.09 0.29 0.11
total_sulfur_dioxide 0.07 0.05 0.11 0.38 0.20 0.59
density 0.25 0.00 0.14 0.83 0.25 0.27 0.50
ph -0.42 -0.02 -0.15 -0.18 -0.09 0.00 0.02 -0.08
sulphates -0.03 -0.05 0.06 -0.05 0.00 0.05 0.10 0.05 0.16
alcohol -0.12 0.10 -0.07 -0.43 -0.37 -0.22 -0.42 -0.78 0.11 -0.00
quality -0.12 -0.18 -0.00 -0.09 -0.21 0.06 -0.14 -0.31 0.10 0.07 0.44
fixed_acidity
volatile_acidity
citric_acid
residual_sugar
chlorides
free_sulfur_dioxide
total_sulfur_dioxide
density
ph
sulphates
alcohol
quality
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

So from here, we can observe that even if these variables are not having a strong correlation with each other still they can affect the quality.

Conclusion



Data Distribution

The distribution of data is different in the two datasets

Correlation

Even if two variables have no correlation, they still the can affect quality

Impact on Quality

Different variables affecting quality in white and red wine datasets

Next Steps

- Normalize Outliers
- Work separately with two datasets
- Different machine learning Algorithms

References

1 https://www.winesofportugal.com/en/portuguese-wines/wine-styles/

2 https://winefolly.com/tips/red-wine-vs-white-wine-the-real-differences/

3 https://rpubs.com/nimit/Report

Thank you!