COSC 5010-03 Practical Machine Learning Fall 2023 Exploratory Data Analysis Report

Michael Elgin

November 13, 2023

1 Introduction

Exploratory data analysis is an important consideration before creating machine-learning models. It can affect the choice of model, the choice of features to prune, or whether samples should be modified or deleted. This report examines the charactersistic of the wine quality dataset[1]. The dataset explored is created by concatenating both the red and white wine datasets together. Later, the class distributions for each of red and white are individually examined.

2 Analysis and its Results

2.1 Dataset Basics

The dataset contains 11 features, all of which are real numbers. The quality (target) is an integer. There are a total of 6497 samples. There are no missing values. Approximately 75% of the samples are white, and 25% are red.

2.2 Feature Correlations

Feature correlations are an important facet of a dataset to explore. The first reason is that there are certain types of models, such as linear models, that expect features not to be correlated. If this is not the case, a linear model may end up assigning unpredictable and inaccurate weights to features. For example, if a linear model predicting circumference of circles is trained on radius as one feature *and* diameter as the second, it may assign an arbitrary weight less than 2π to the radius and likewise an arbitrary weight less than π to the diameter such that the weighted sum always arrives properly at the circumference. But of course in reality this is nonsense; as we know the weights should be *exactly* 2π and π respectively. The issue is correlation, in this case the logical extreme of perfect correlation. Knowing the second value provides no new information and merely creates weights that don't

[1] https://archive.ics.uci.edu/dataset/186/wine+quality

reflect how the feature truly influences the target in reality. The solution is to prune highly correlated features so only one is left.

The second reason is that there is a model-agnostic interpretability method known as partial dependence plots, which are made by creating artificial datasets by setting a feature to be a specific value for all samples, and then finding the model's average prediction for all those samples. Varying the value shows how model prediction changes accordingly (on average). The reason why correlated features are an issue here is that if a plotted feature is correlated with another (unplotted) feature, in reality it won't change by itself. Thus showing how one feature affects model output as it changes independently is misleading because it will never change independently from it's correlated feature. One may observe a change to the feature in the plot and think that the model will also change according to the plot, but it will not because that change in the original feature also involves a hidden change to the correlated feature which will affect model output some way; a way that the partial dependence plot doesn't account for.

Feature correlations are easily visualized by heatmaps. A common technique is to set positively correlated features to one color, negatively correlated features to the opposite color, and no correlations to white. This allows one to easily spot the most correlated features. The correlations for the wine quality dataset are shown in figure 1.

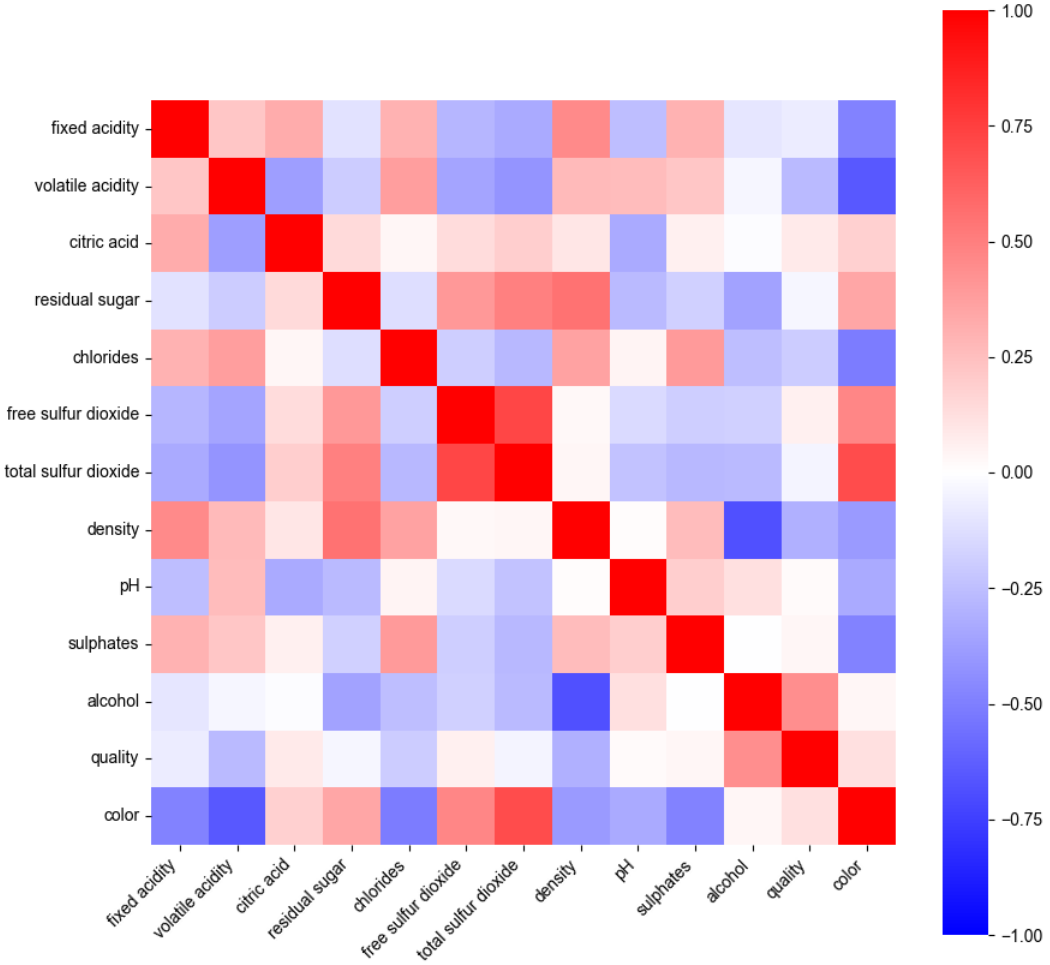


Figure 1: Feature and target correlations

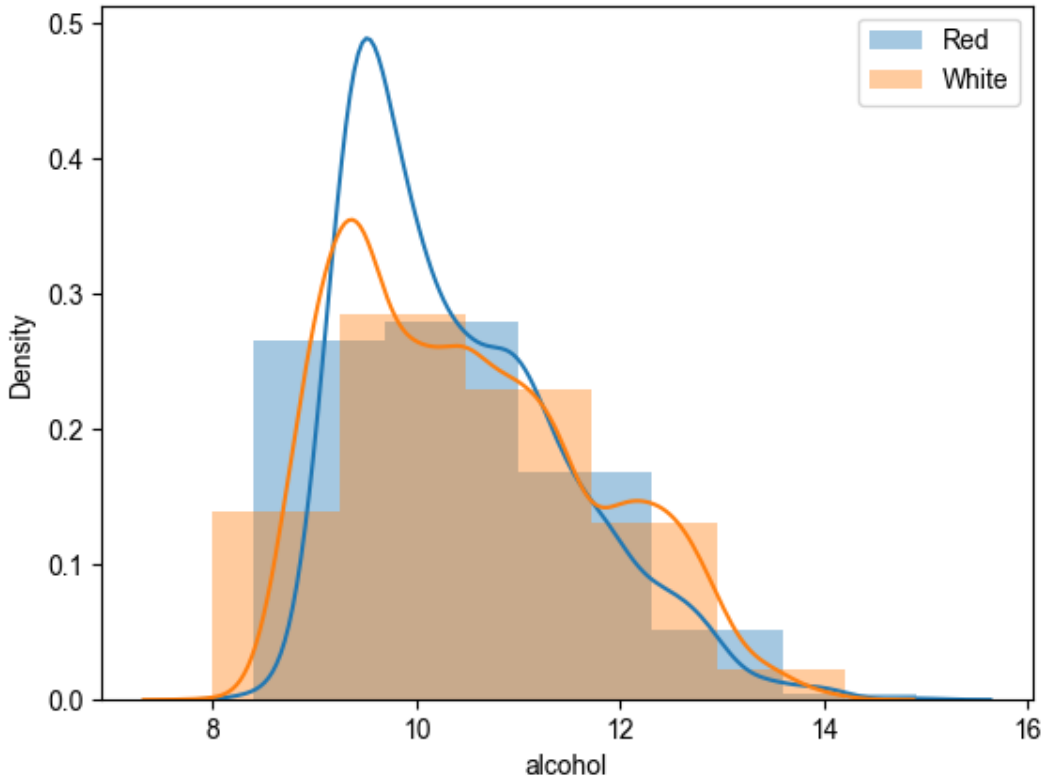
From here, one can spot free sulphur dioxide and total sulfur dioxide as highly correlated features. This is no surprise, for obvious reasons. At least one of the features should be pruned, most likely total sulfur dioxide. Likewise one can also spot the negative correlation between alcohol and density. This makes intuitive sense, since increasing the amount of alcohol (and thus decreasing water) is always going to change the density in the same direction.

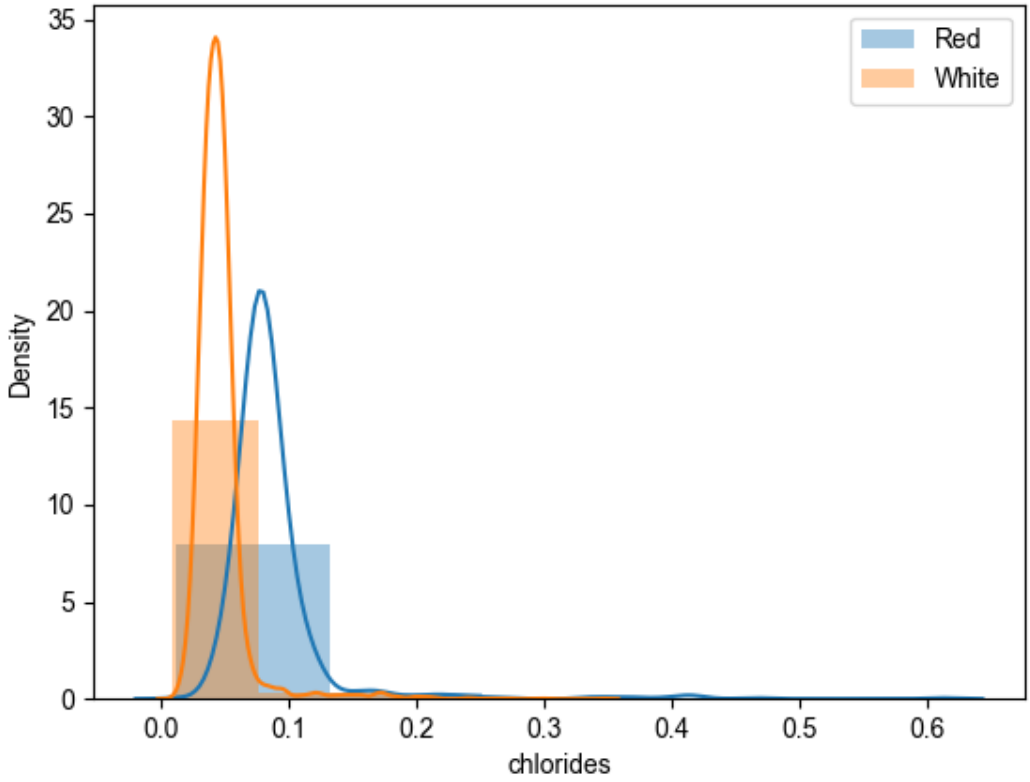
A third reason to know the correlations is to see which features are correlated with the target, as this provides direct insight into which features are influential at changing the target and thus tell something to a model about what target to predict.

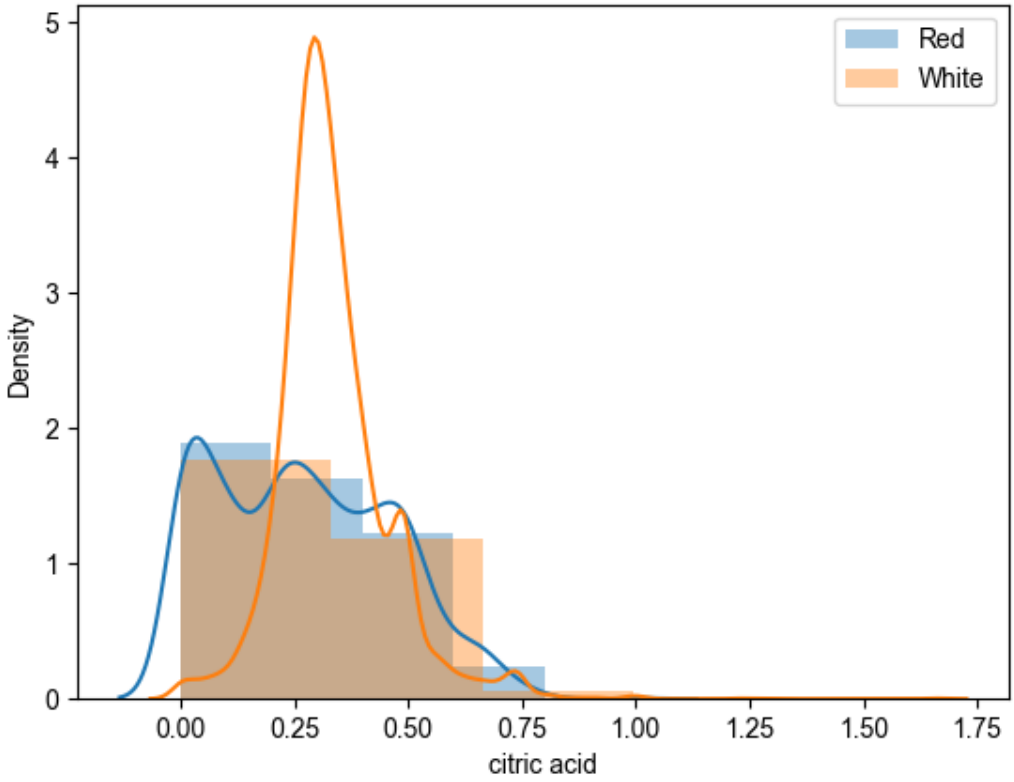
2.3 Class Density Distributions

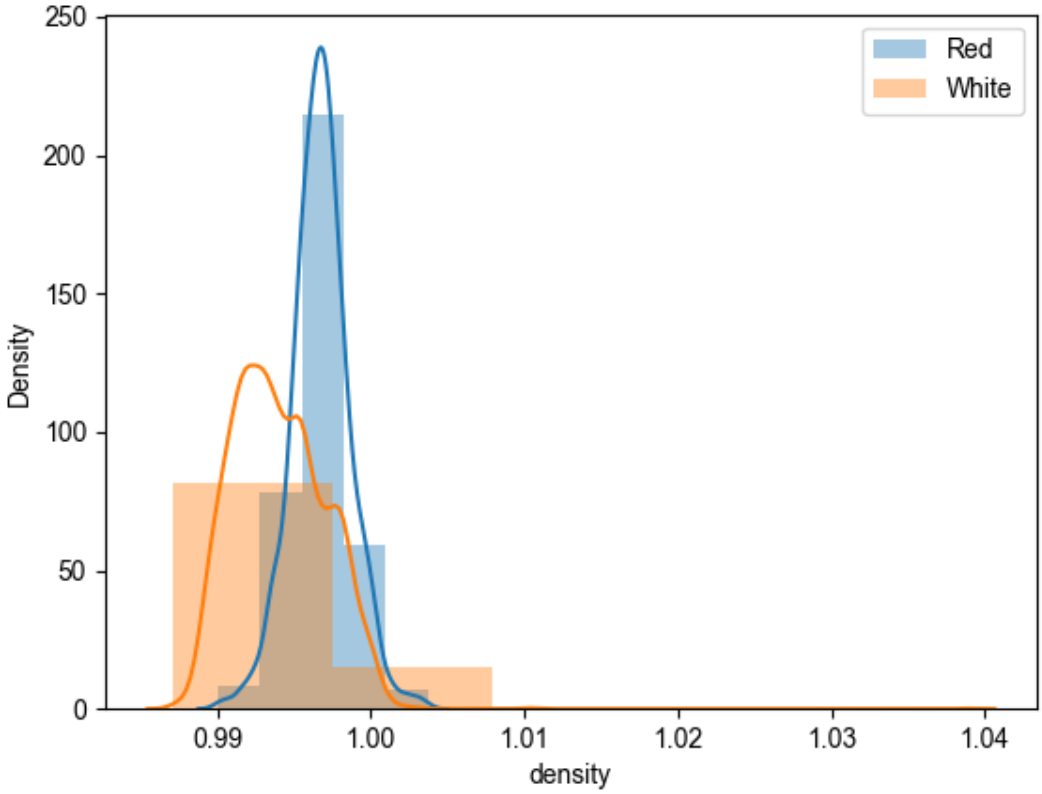
2.3.1 Concatenated Dataset

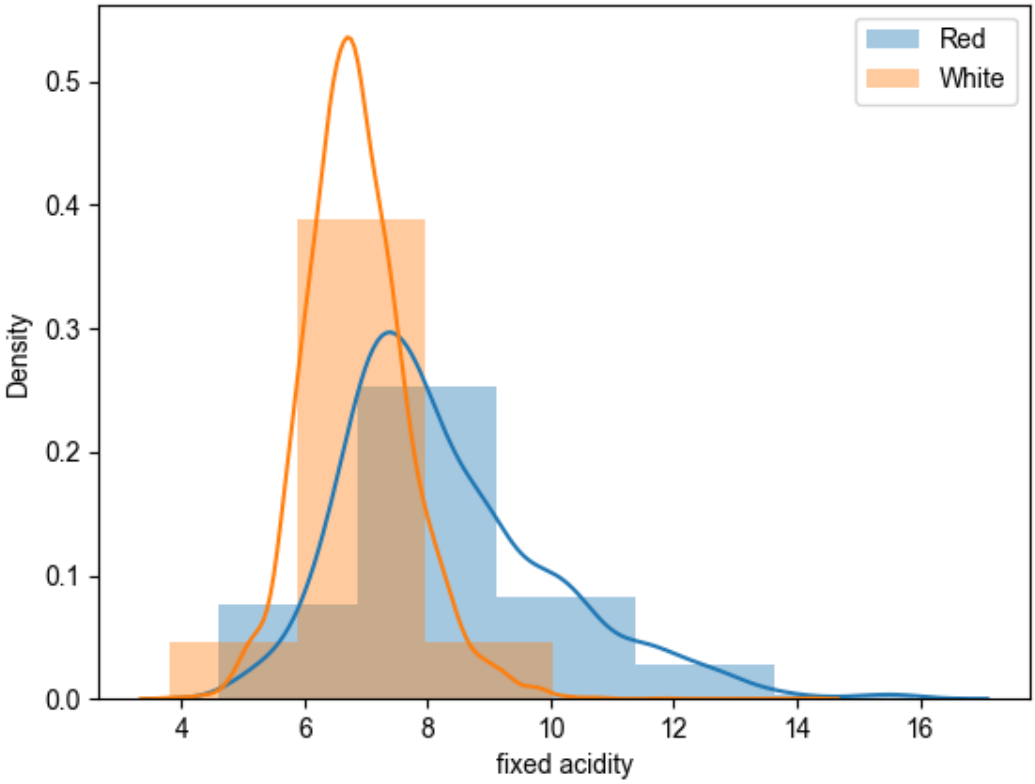
An important part of exploratory data analysis involving continuous features and a classification target is to plot the density functions of each class along each feature in the same graph, to see how much separation a feature may give between the classes. This visually shows the differences in features between the classes.

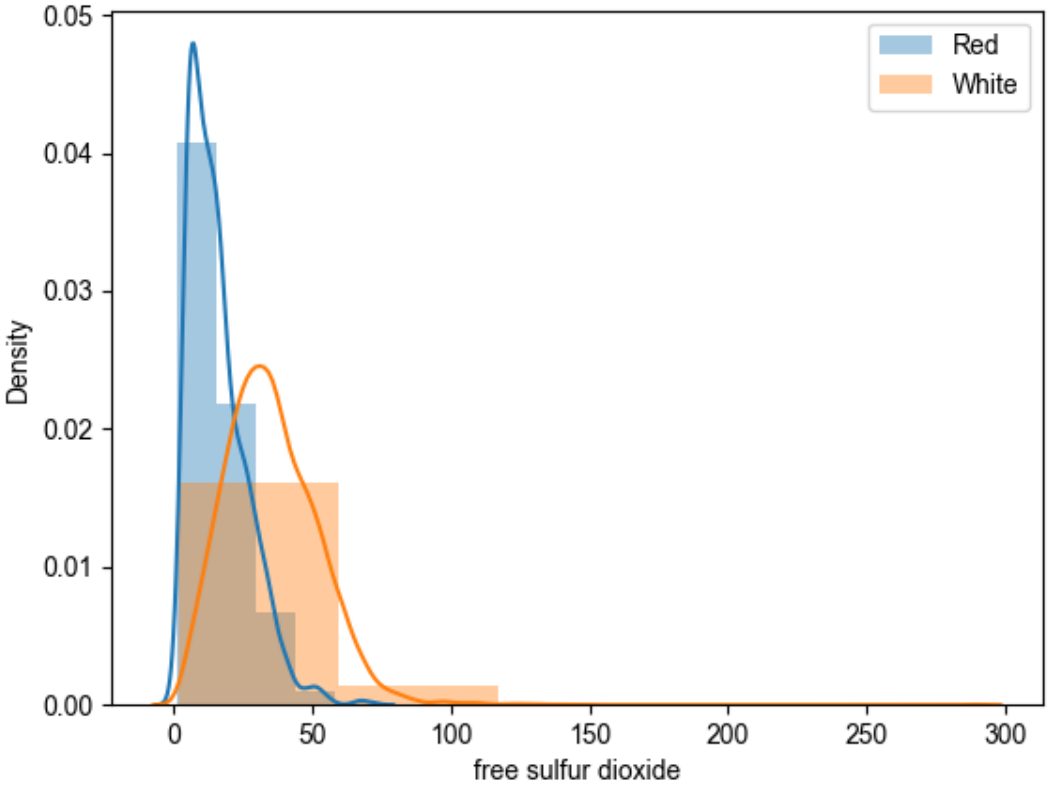


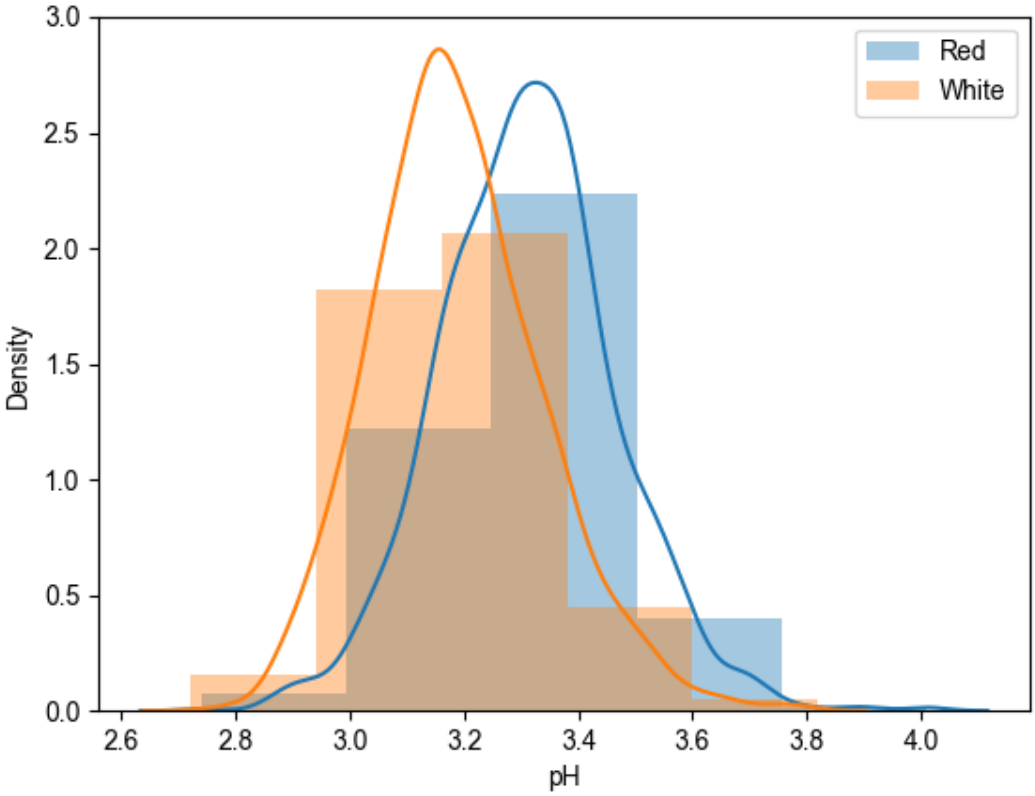

Red
White
Density
chlorides
0
5
10
15
20
25
30
35
0.0
0.1
0.2
0.3
0.4
0.5
0.6

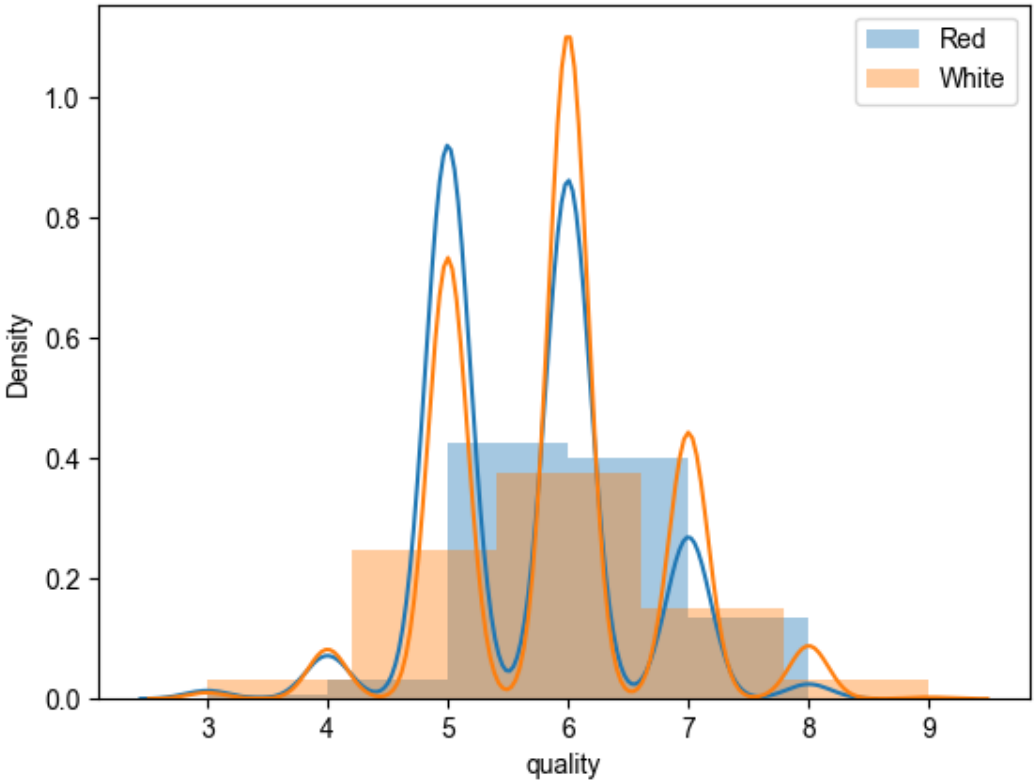

Red
White
Density
citric acid
0
1
2
3
4
5
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

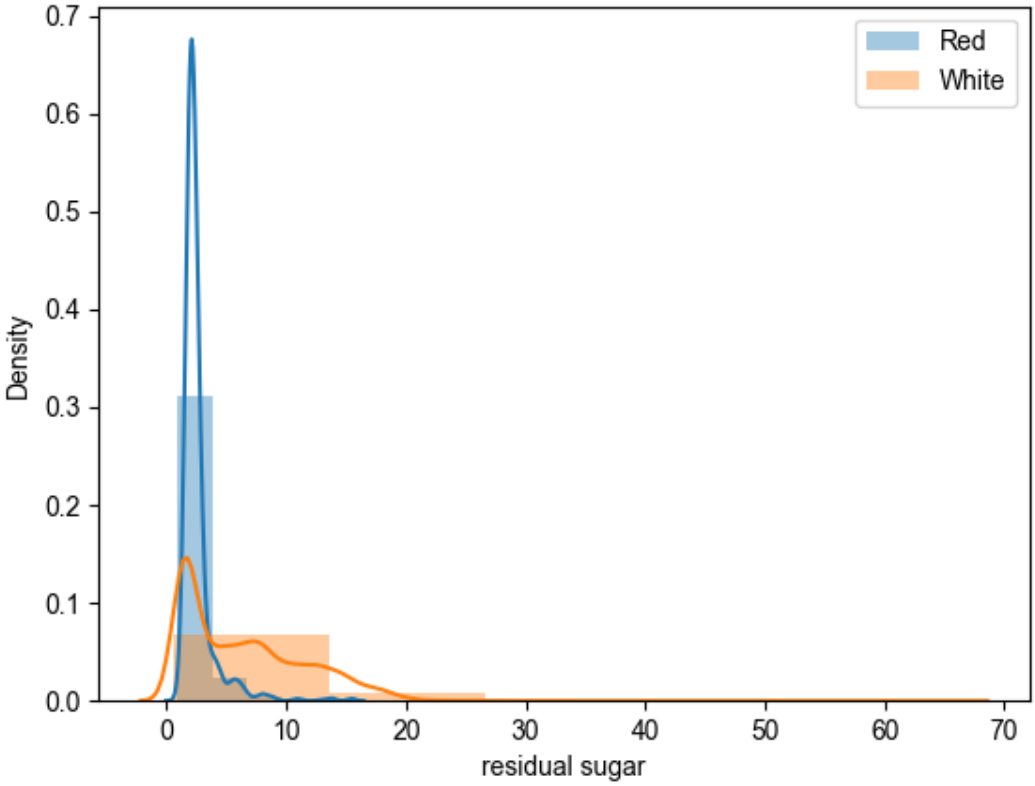

Red
White
Density
density
250
200
150
100
50
0
0.99
1.00
1.01
1.02
1.03
1.04

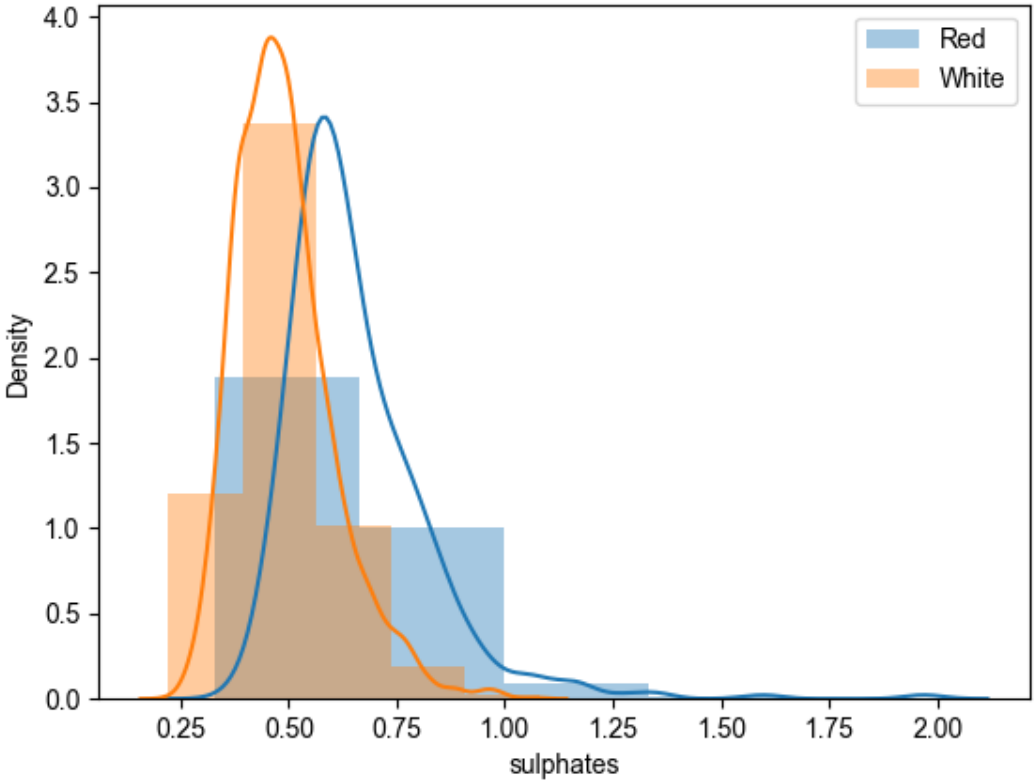

Red
White
Density
0.0
0.1
0.2
0.3
0.4
0.5
4
6
8
10
12
14
16
fixed acidity

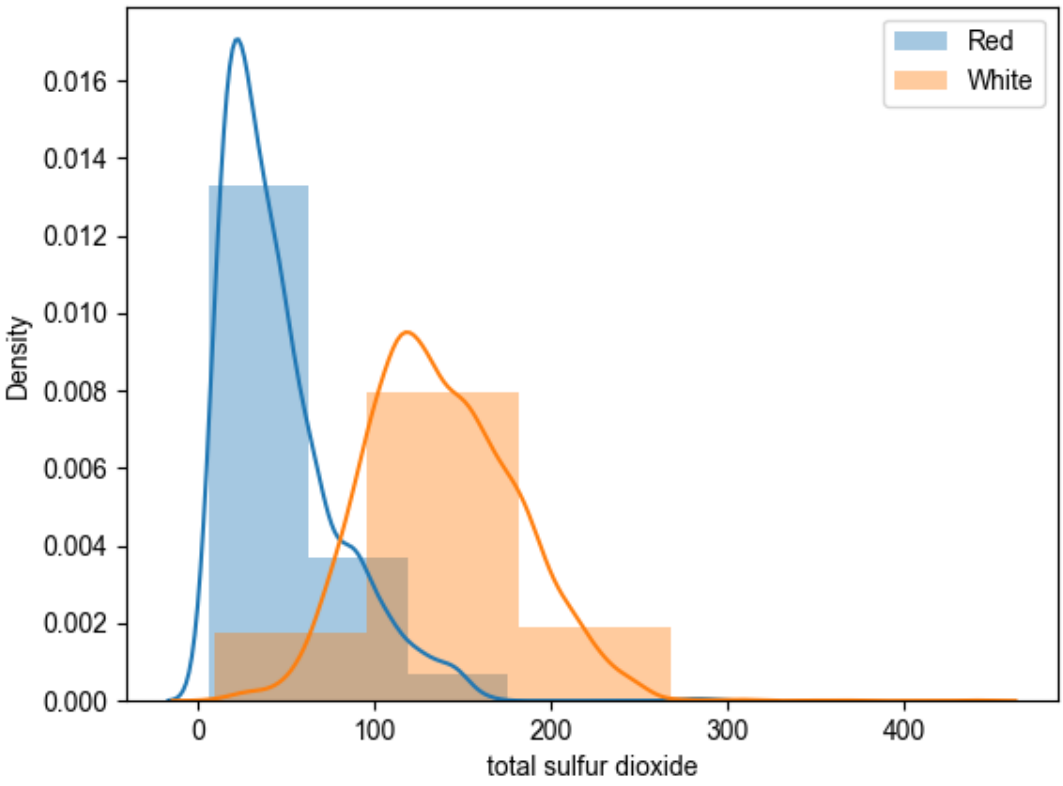

Red
White
Density
0.05
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
250
300
free sulfur dioxide

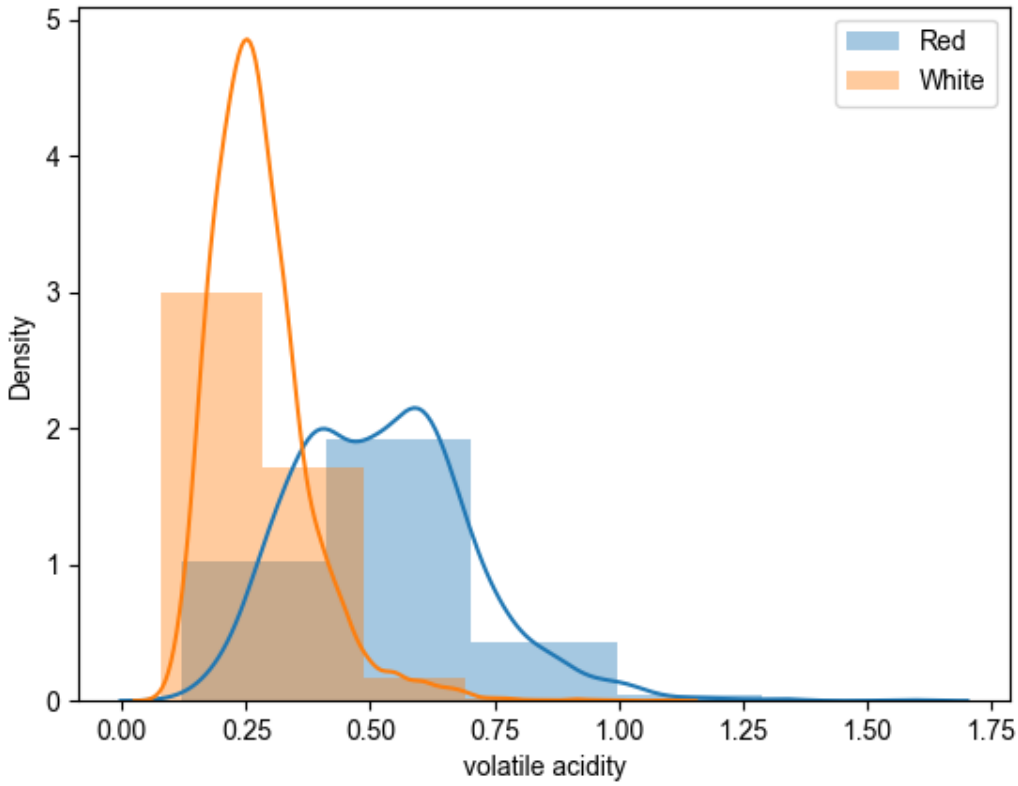

Red
White
Density
pH
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0


Red
White
Density
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
quality


Red
White
Density
residual sugar
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
50
60
70


Red
White
Density
sulphates
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

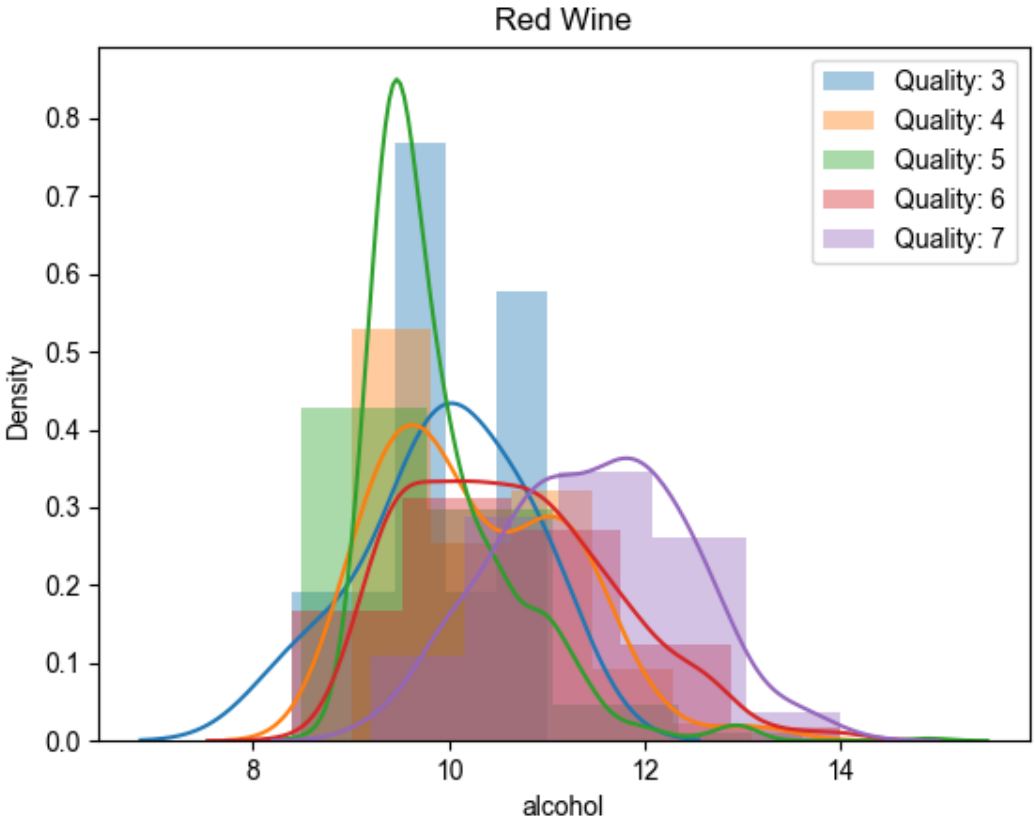

Red
White
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0
100
200
300
400
total sulfur dioxide

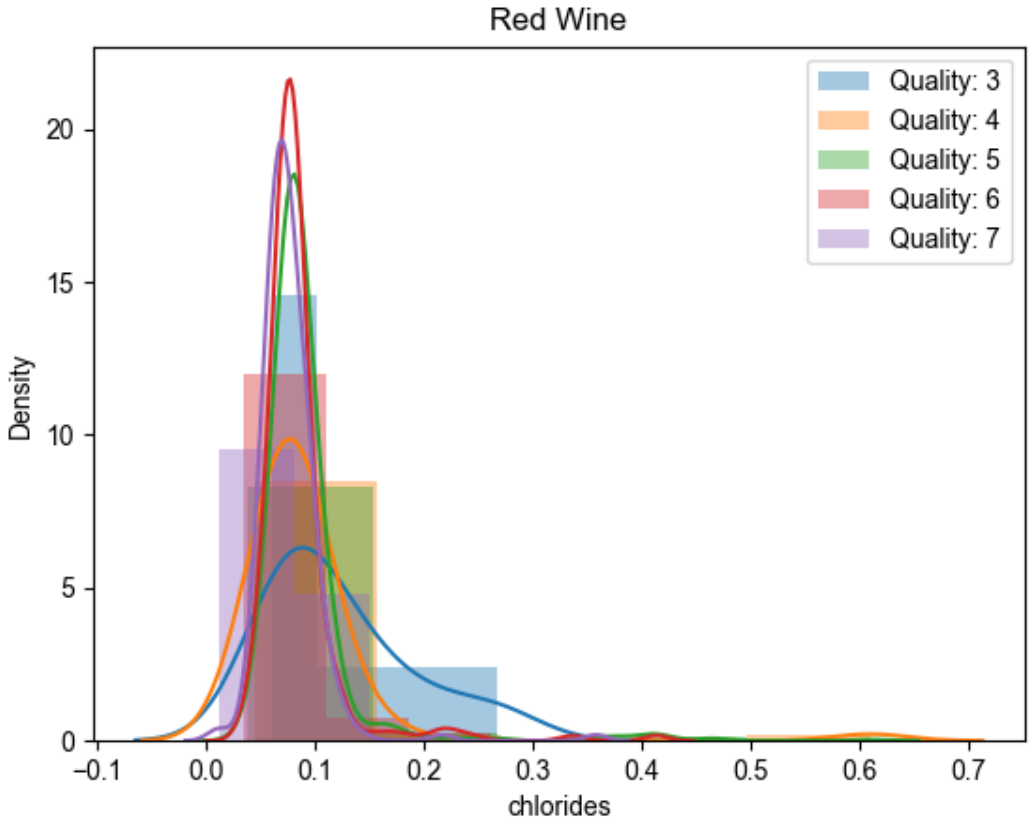


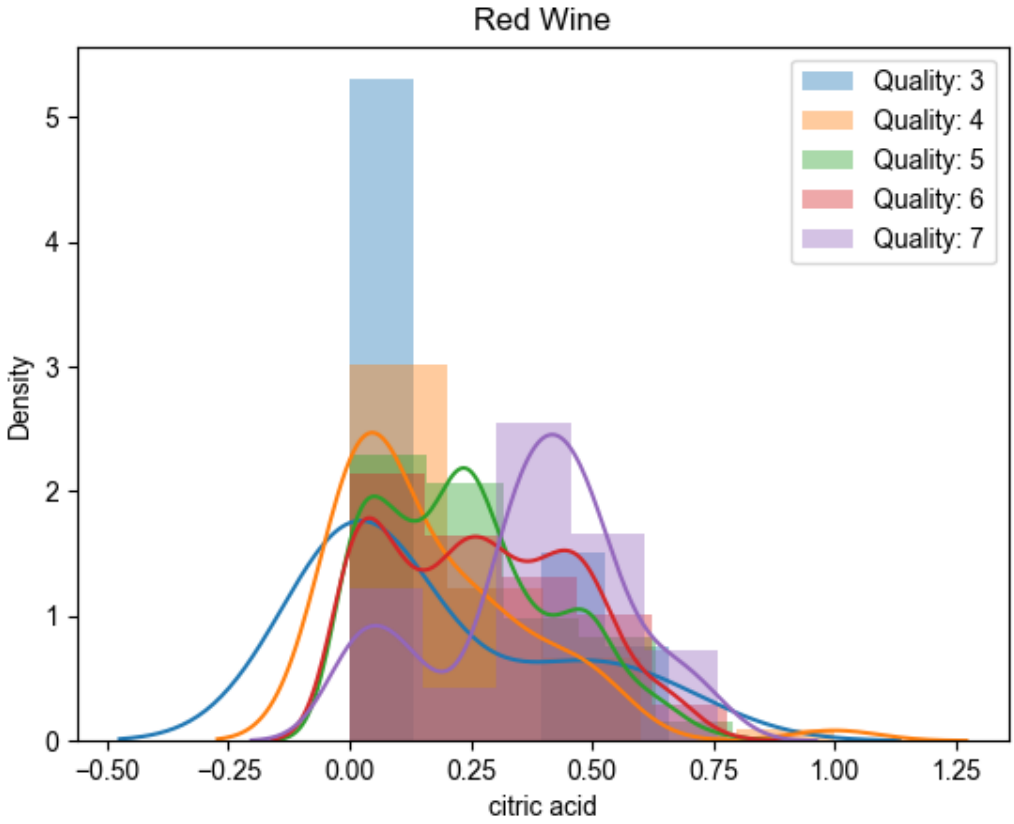
There is not complete separation given by any feature. However there is still some healthy separation in chlorides, volatile acidity, free sulfur dioxide, total sulfur dioxide, and residual sugar. These will be features that are likely to be of most use in classification.

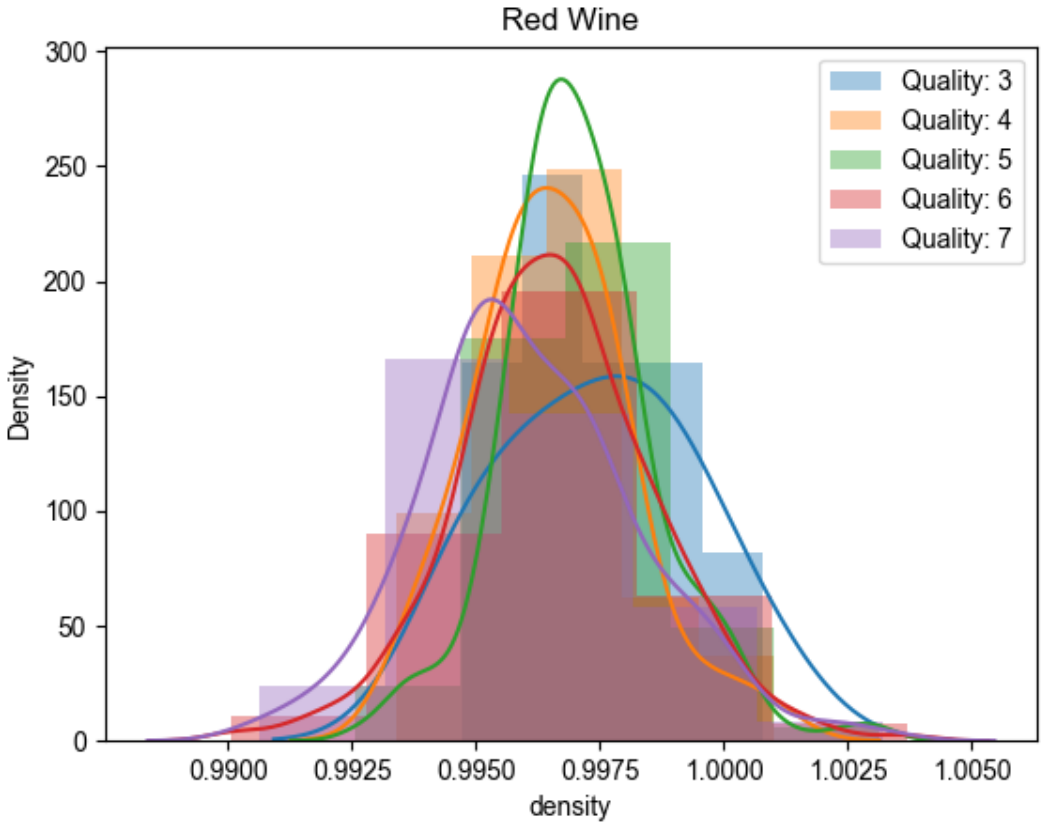
2.3.2 Red and White Datasets

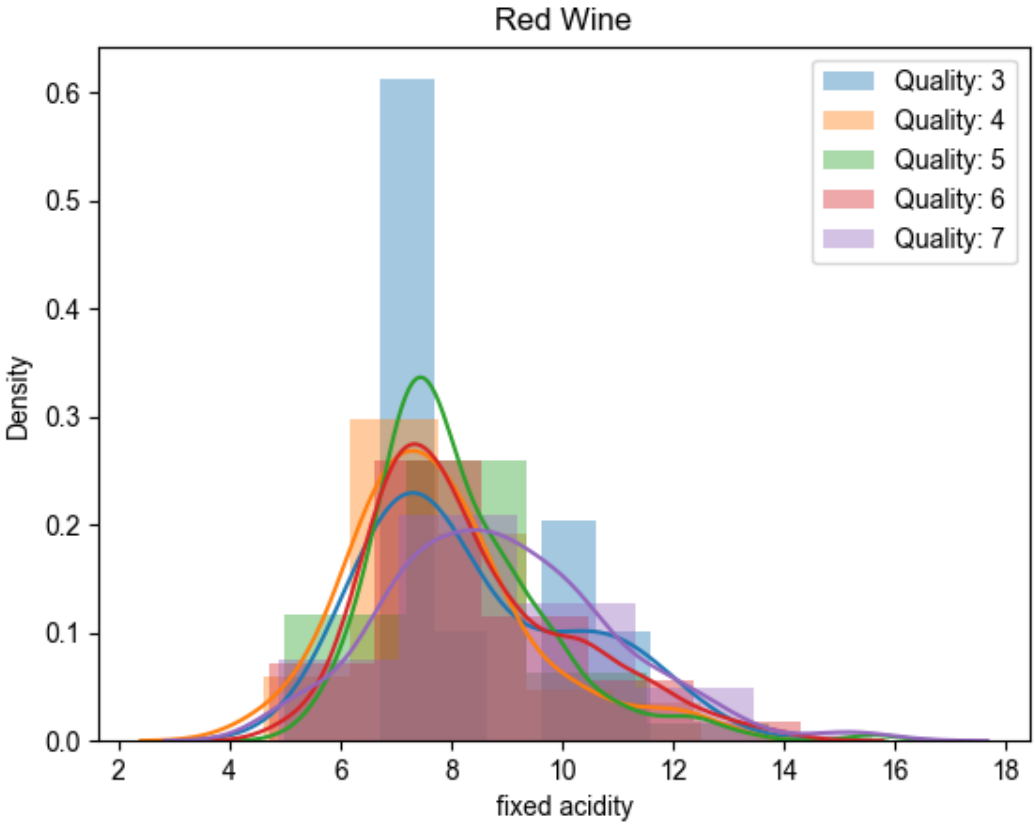
This section plots density functions for each color individually, to see which features may give good separation between quality levels. Red is examined first.

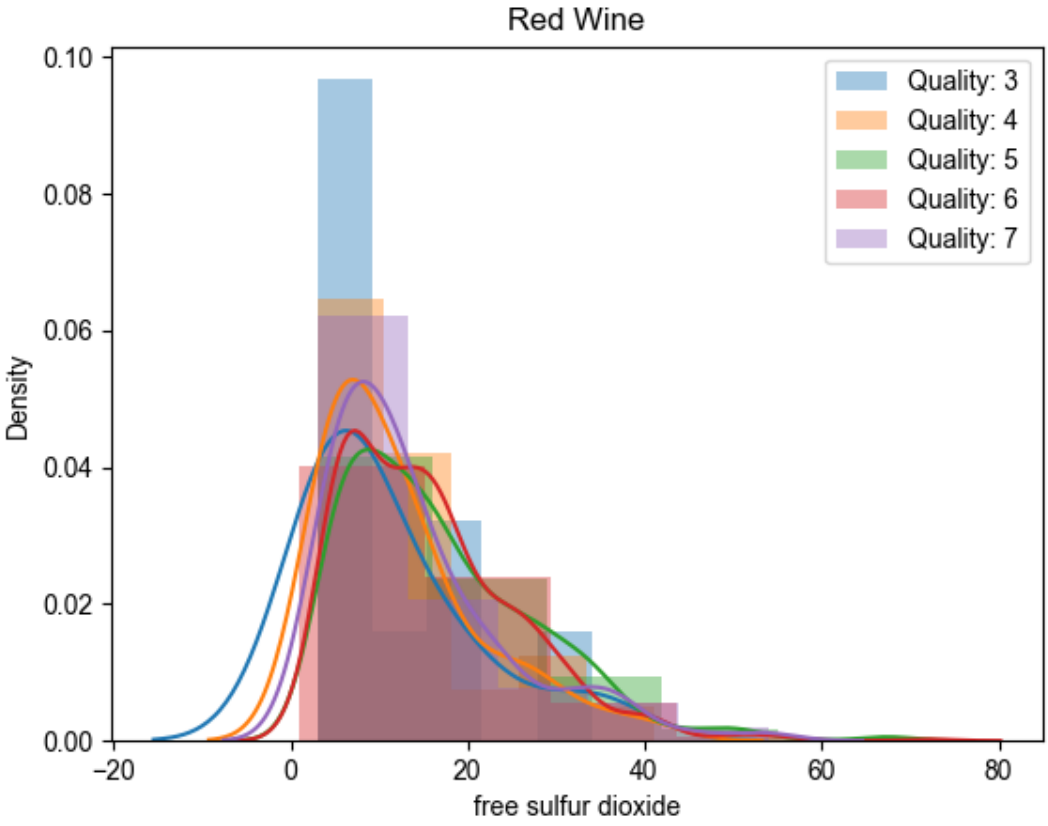

Red Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
8
10
12
14
alcohol

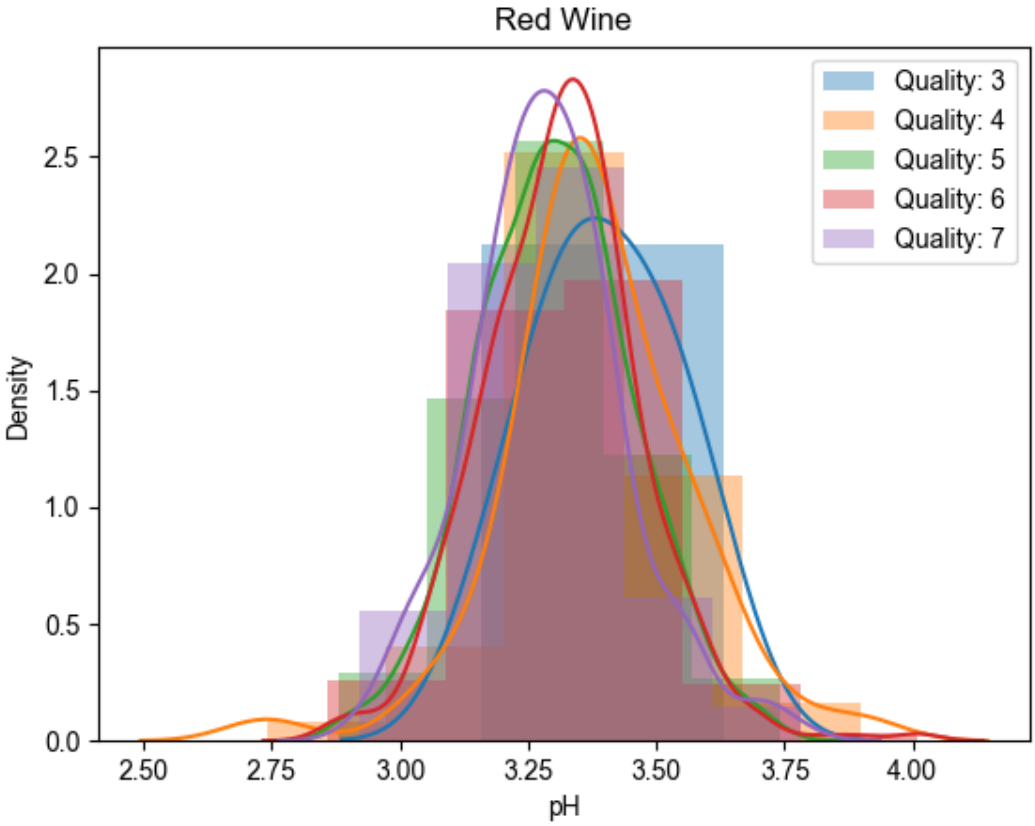

Red Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Density
chlorides
0
5
10
15
20
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

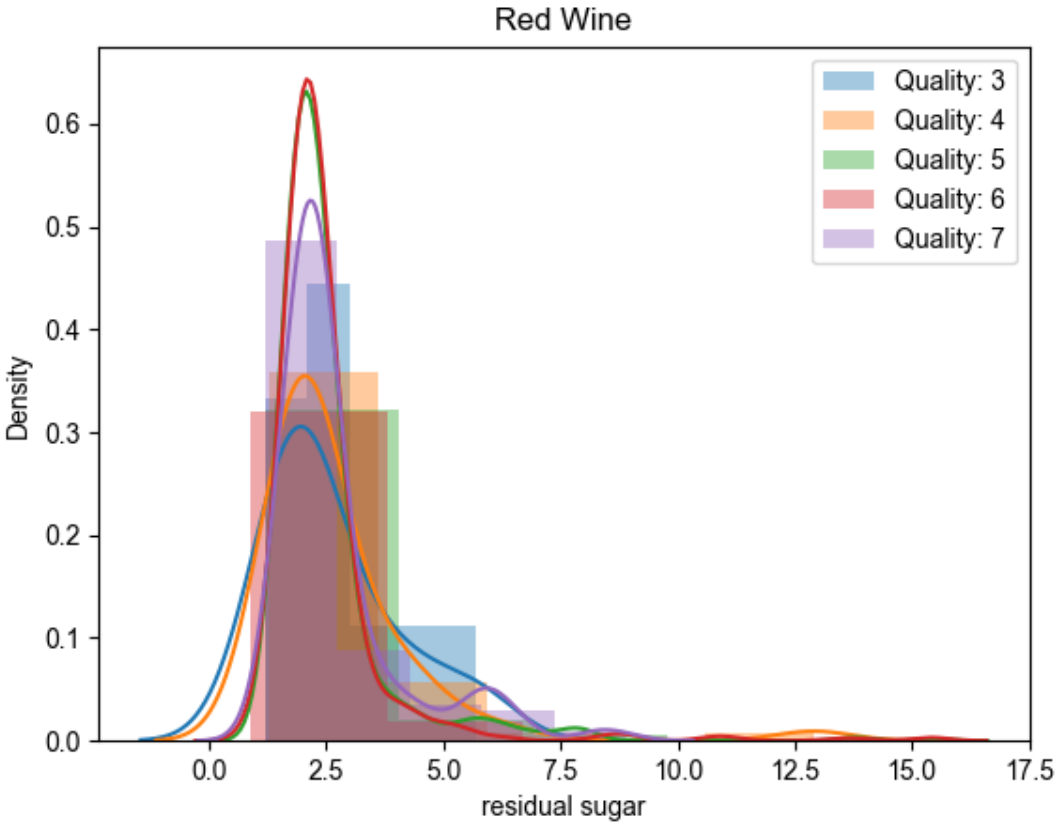

Red Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Density
0
1
2
3
4
5
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
citric acid

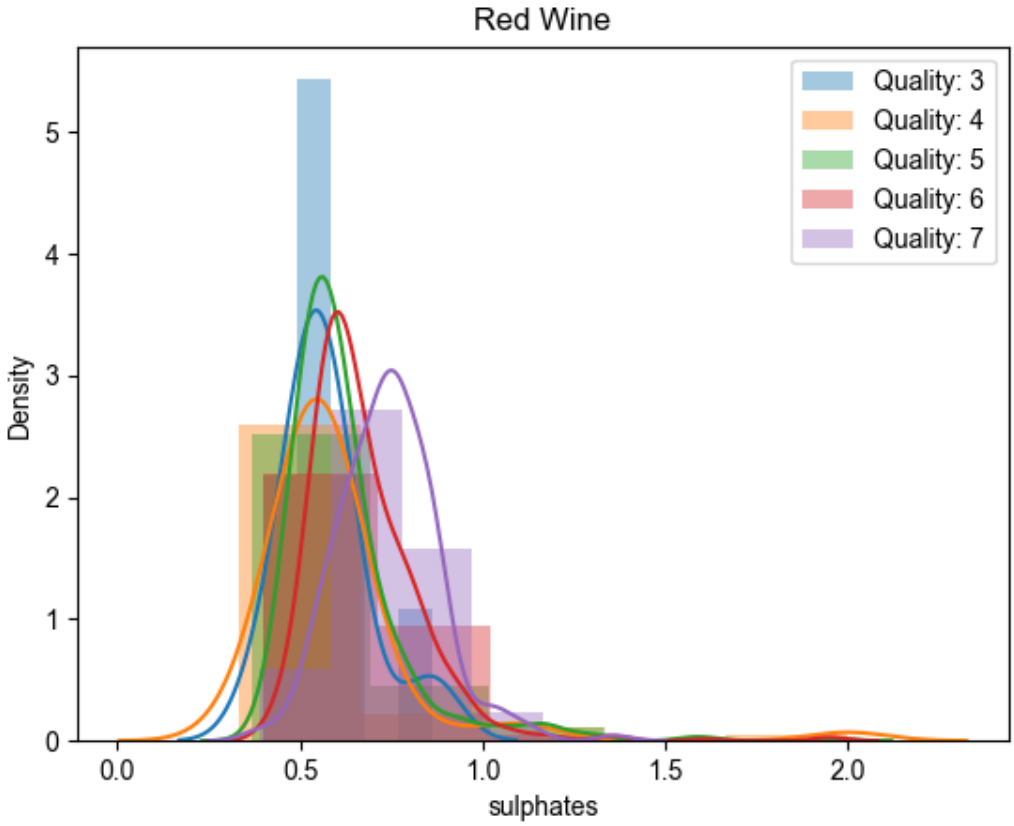

Red Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Density
density
300
250
200
150
100
50
0
0.9900
0.9925
0.9950
0.9975
1.0000
1.0025
1.0050

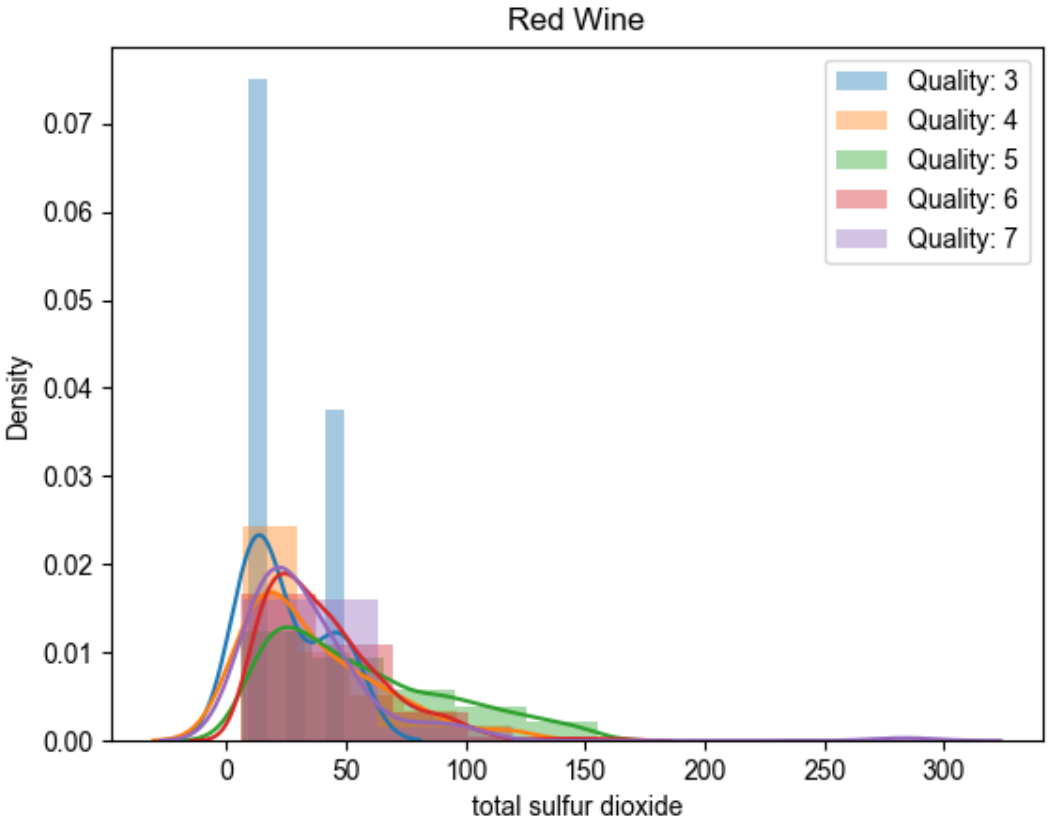

Red Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
2
4
6
8
10
12
14
16
18
fixed acidity

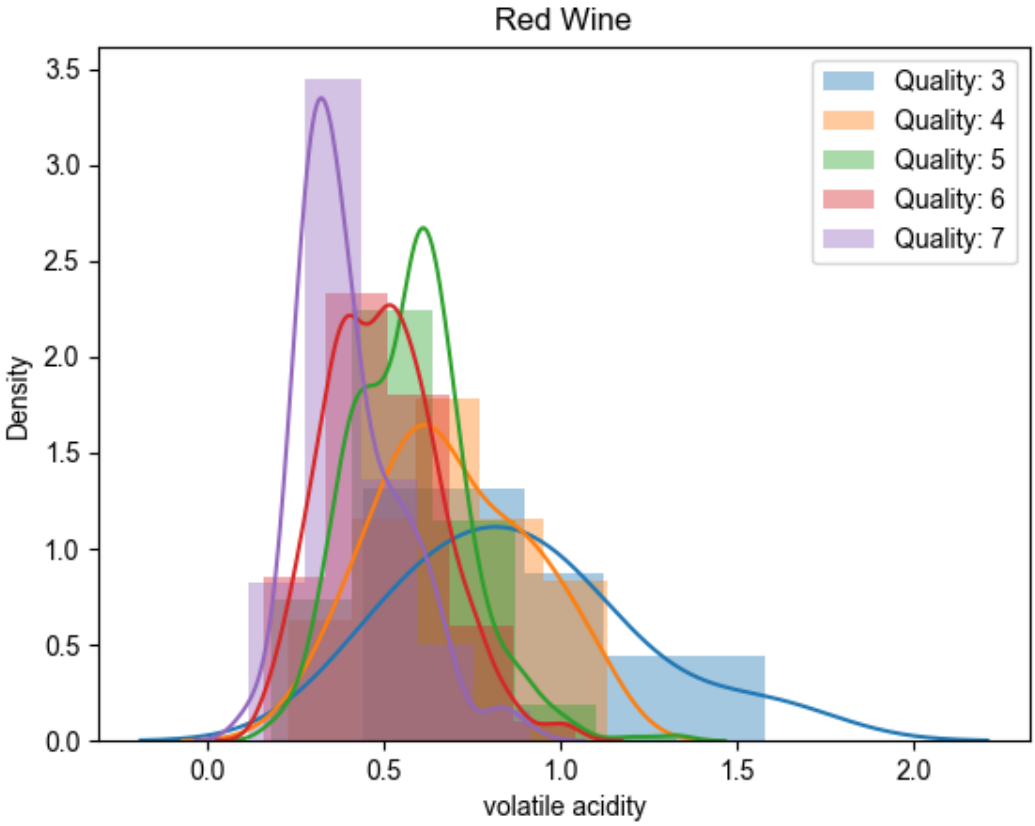

Red Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Density
0.10
0.08
0.06
0.04
0.02
0.00
−20
0
20
40
60
80
free sulfur dioxide


Red Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Density
0.0
0.5
1.0
1.5
2.0
2.5
2.50
2.75
3.00
3.25
3.50
3.75
4.00
pH


Red Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
residual sugar

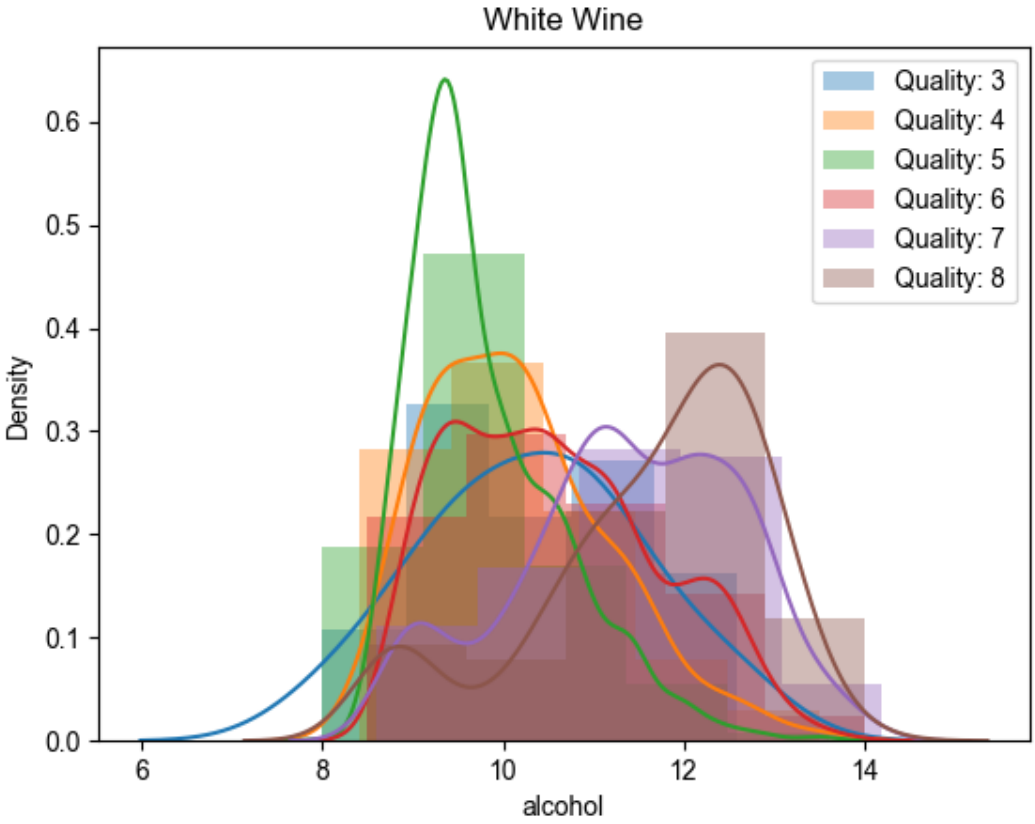

Red Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Density
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
sulphates

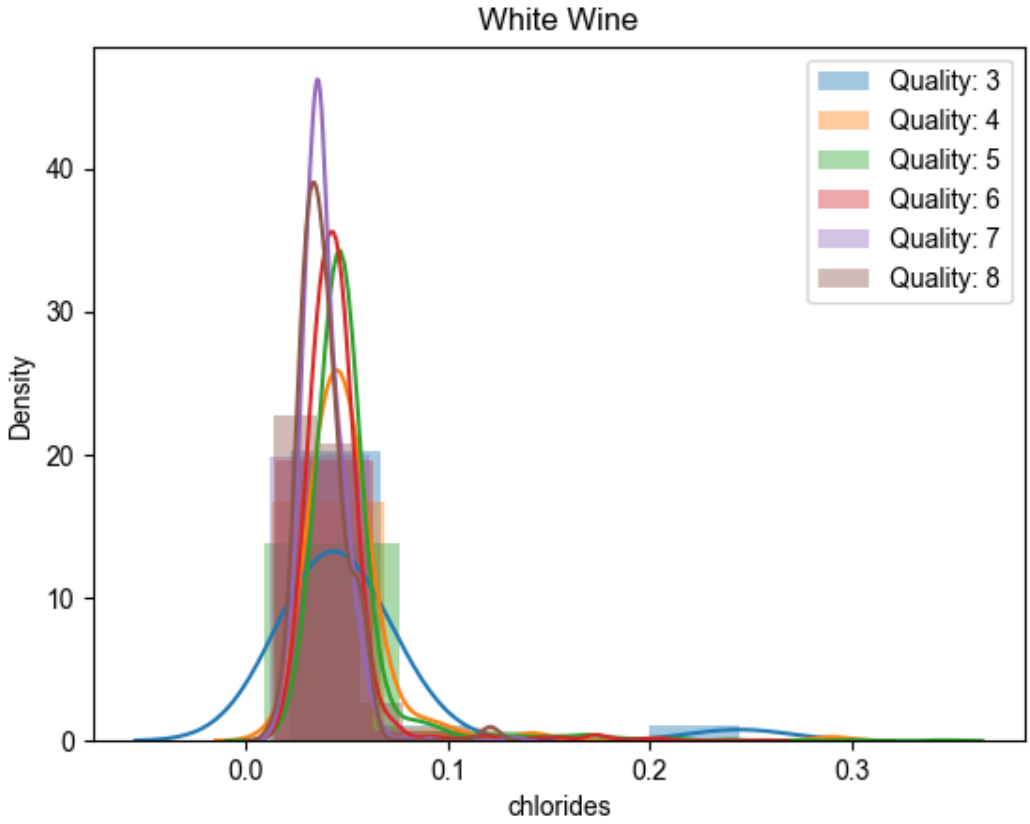

Red Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
50
100
150
200
250
300
total sulfur dioxide

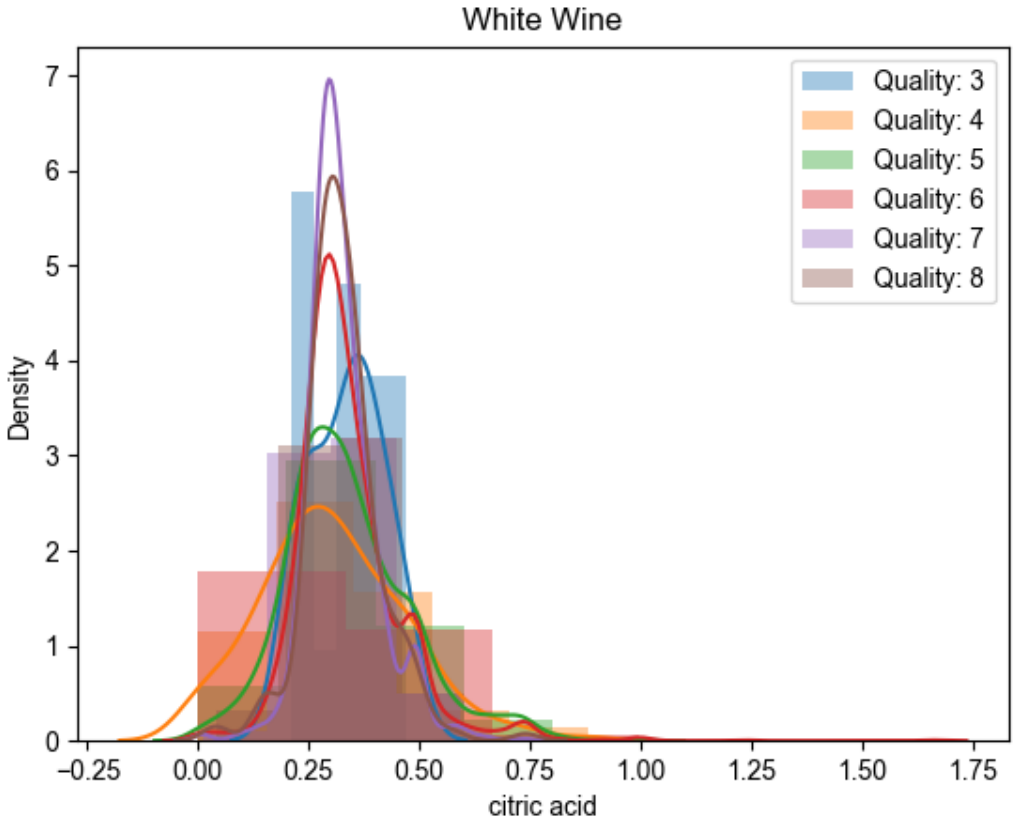


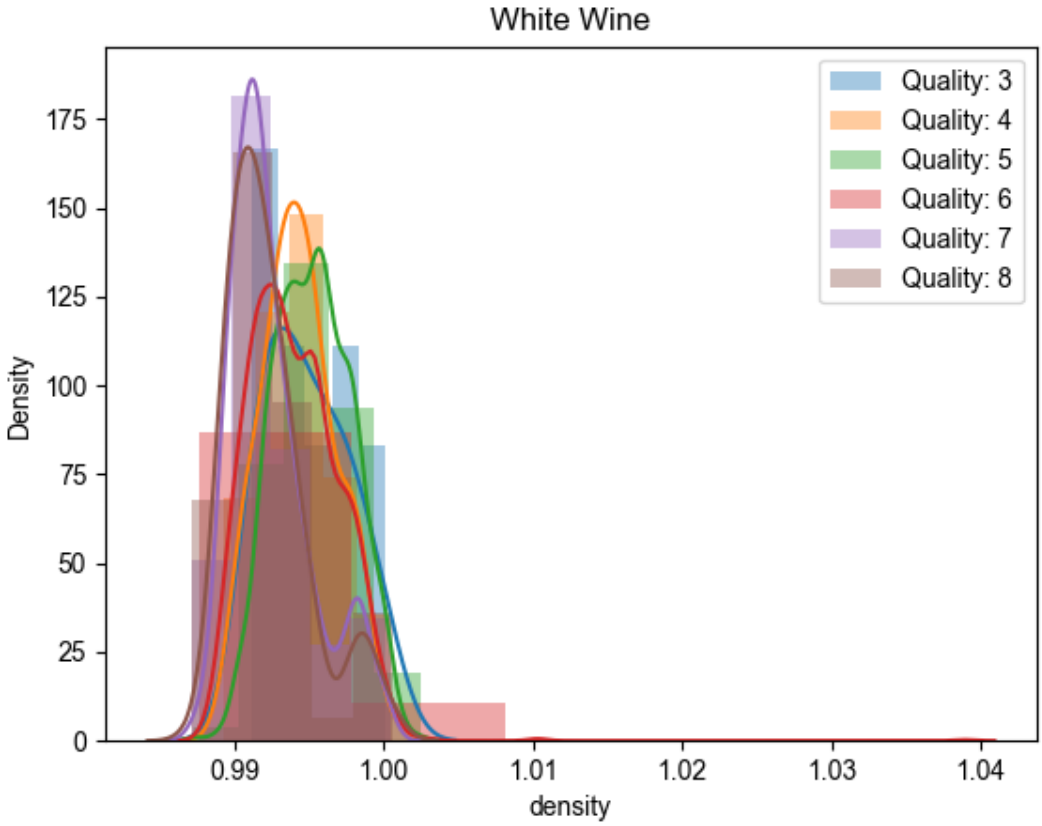
What is most prominent is the difference between low quality wines (blue) and high quality wines (purple). Low quality wines tend to especially have low alcohol, citric acid, free/total sulfur dioxide, and high volatile acidity. High quality wines tend to have the opposite.

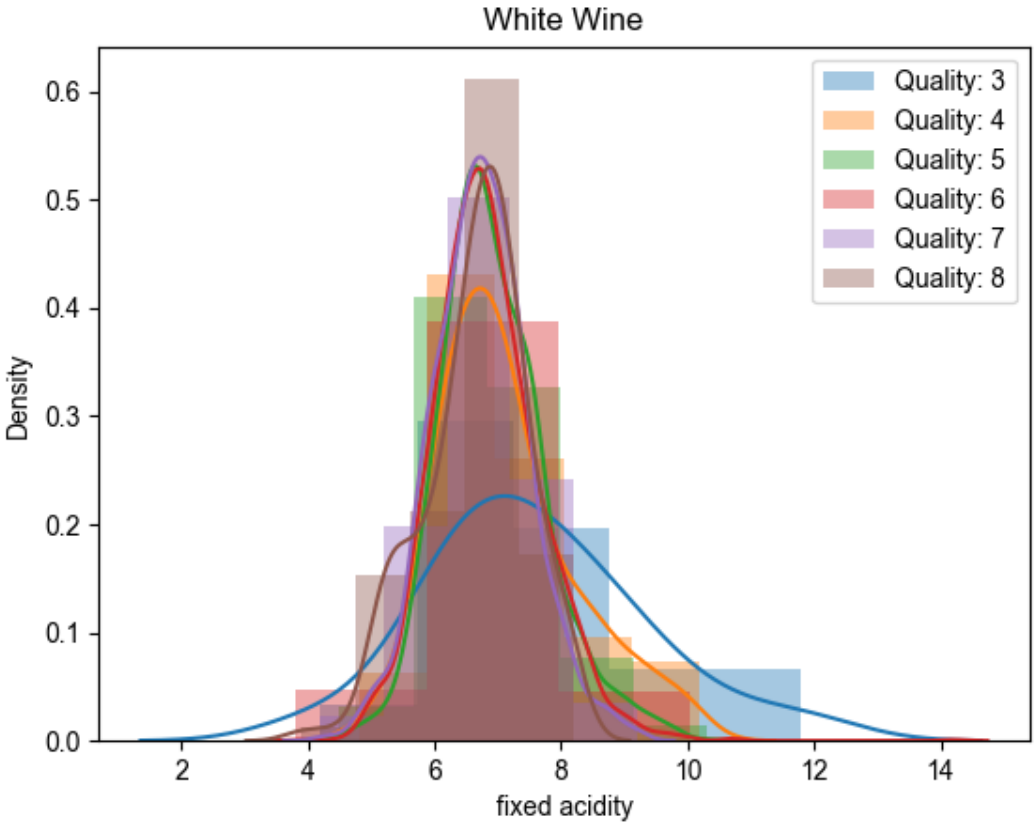
White wine is examined next.

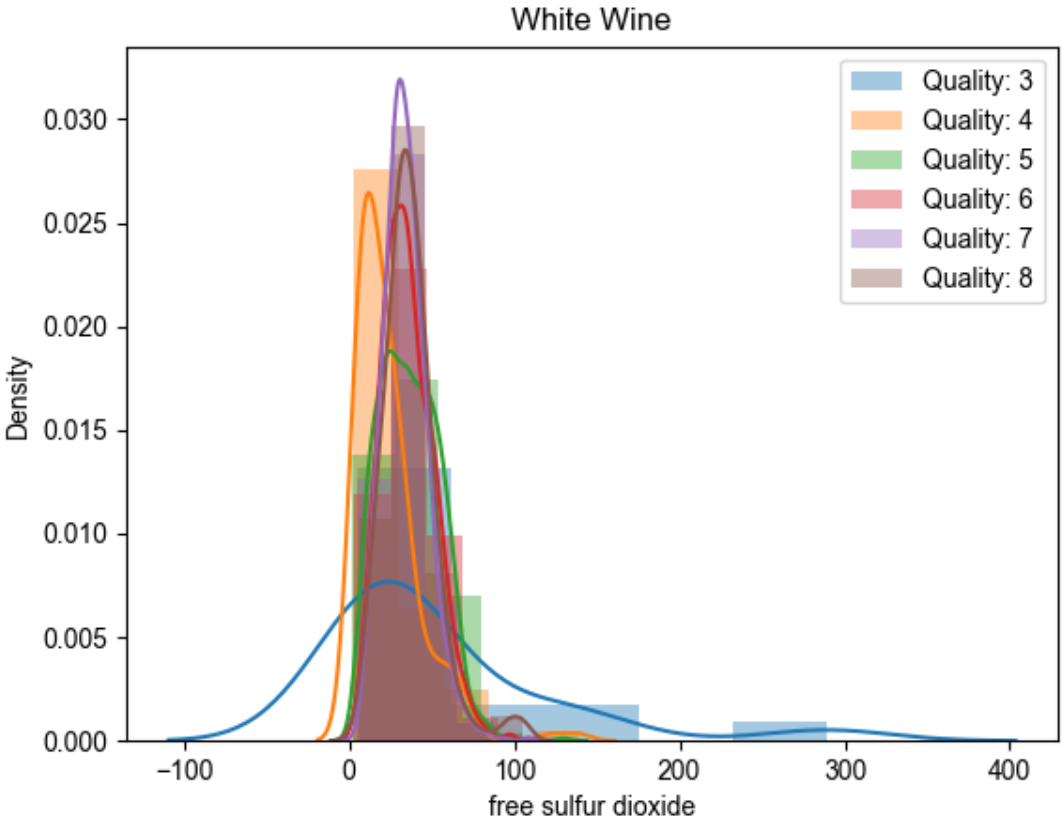

White Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Quality: 8
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
6
8
10
12
14
alcohol

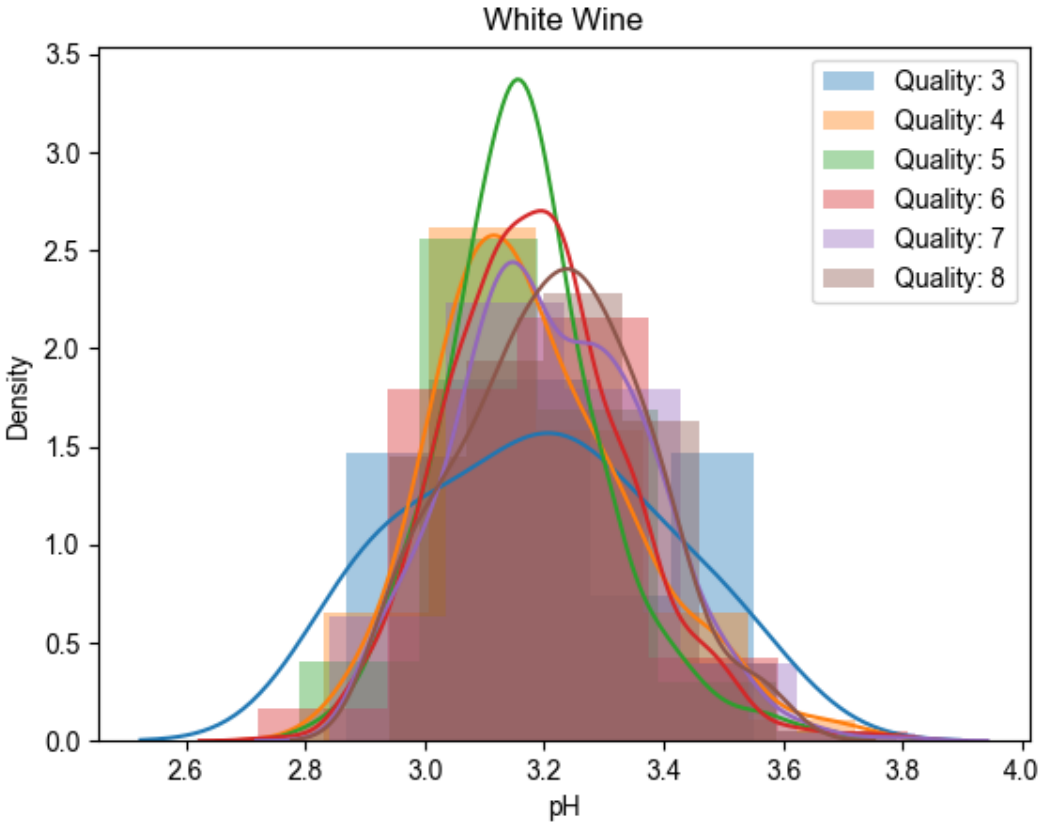

White Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Quality: 8
Density
0
10
20
30
40
0.0
0.1
0.2
0.3
chlorides

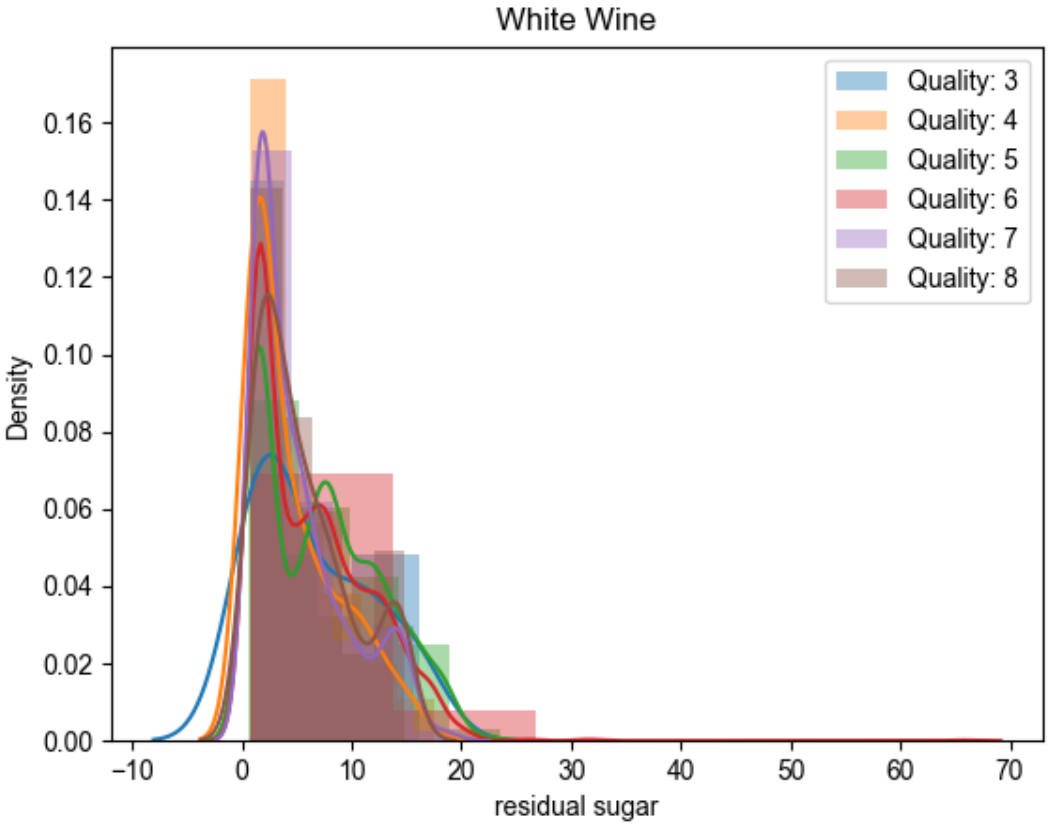

White Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Quality: 8
Density
citric acid
0
1
2
3
4
5
6
7
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

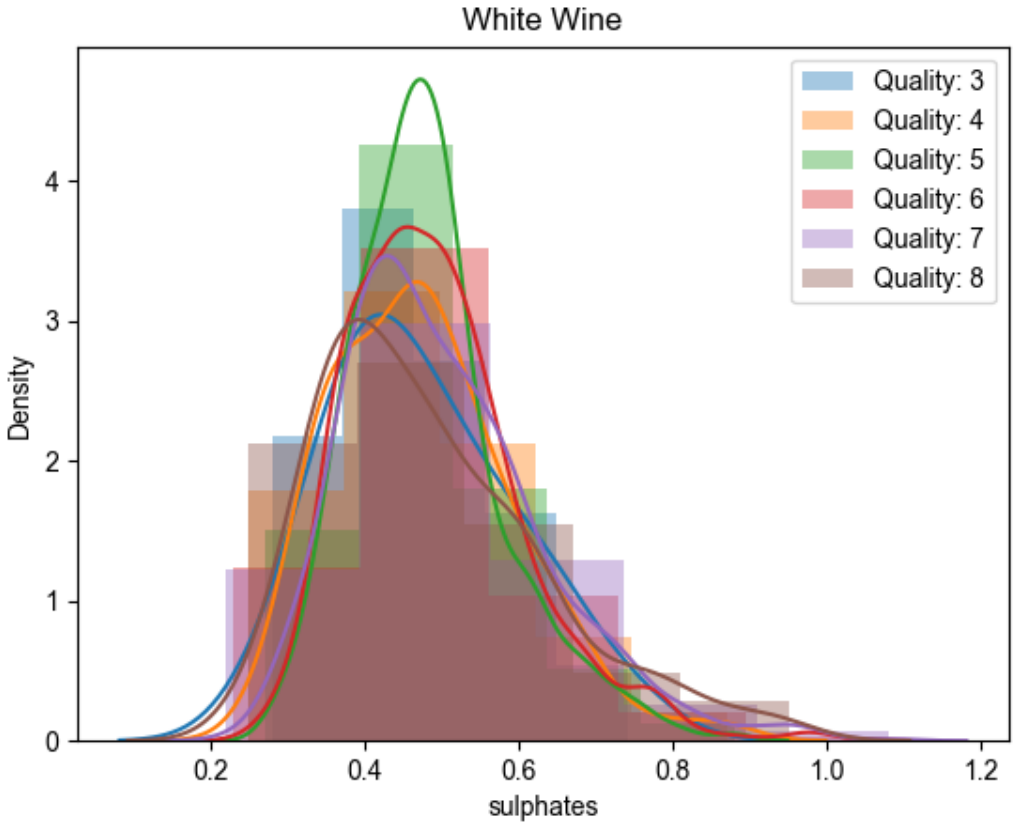

White Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Quality: 8
Density
0
25
50
75
100
125
150
175
0.99
1.00
1.01
1.02
1.03
1.04
density

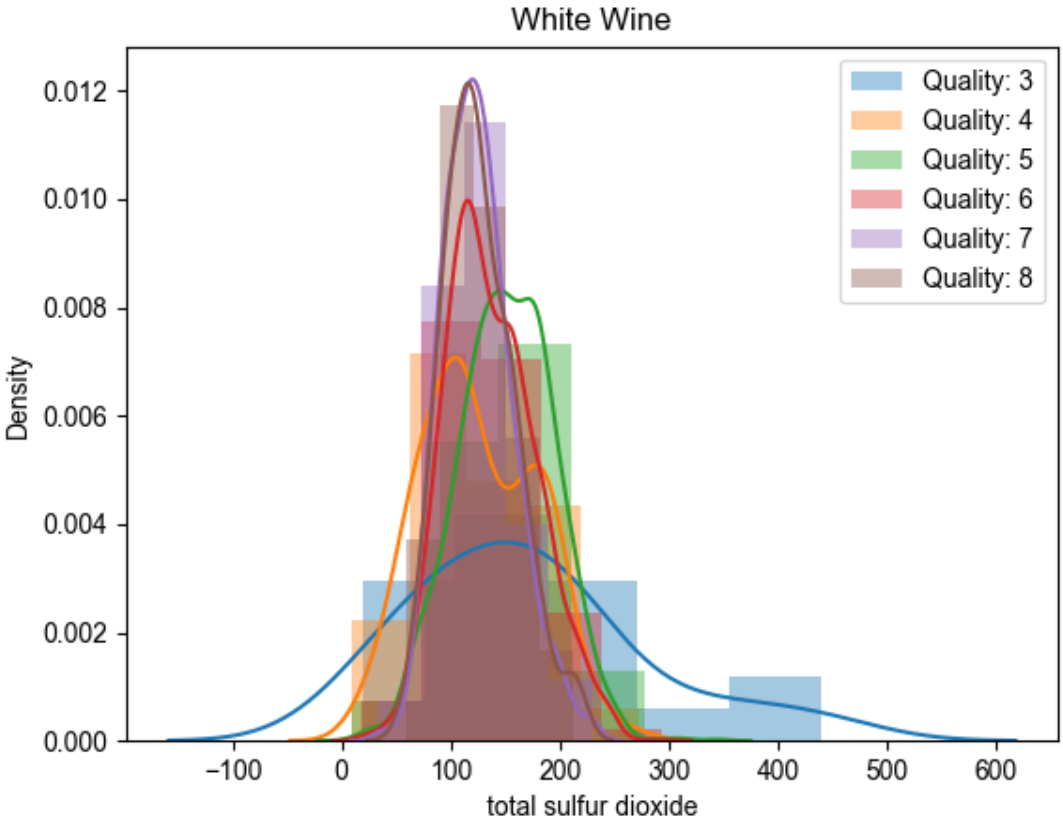

White Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Quality: 8
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
2
4
6
8
10
12
14
fixed acidity

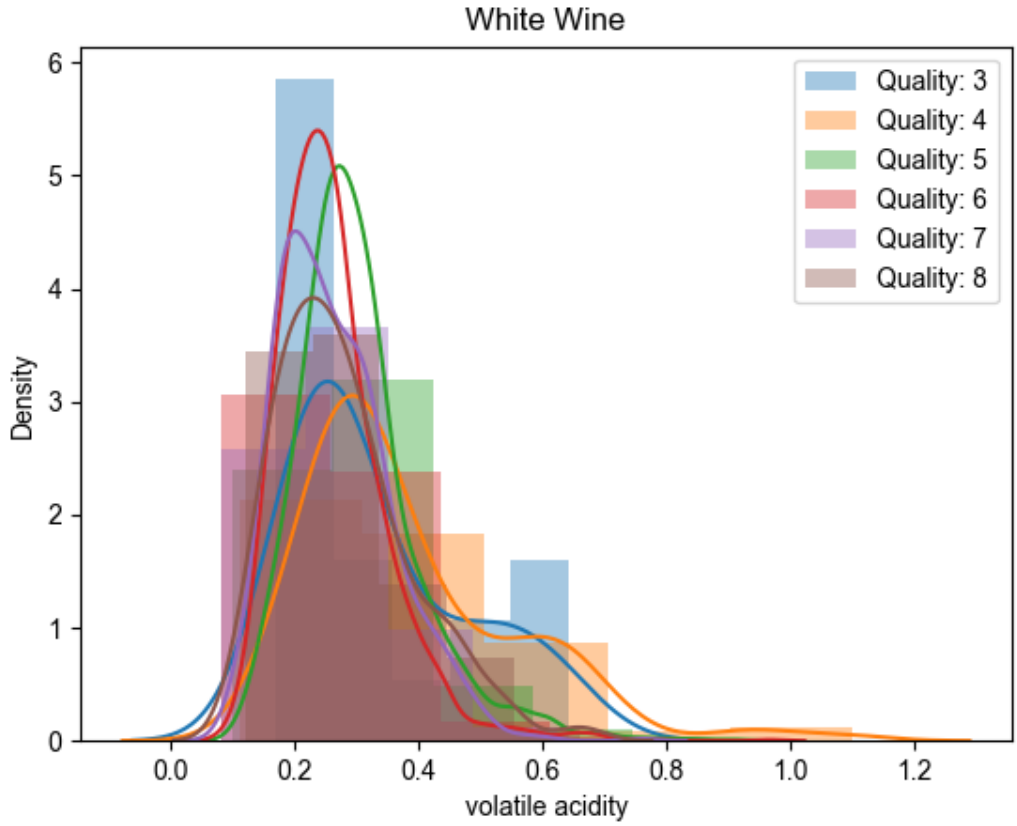

White Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Quality: 8
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
−100
0
100
200
300
400
free sulfur dioxide


White Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Quality: 8
Density
pH
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0


White Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Quality: 8
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−10
0
10
20
30
40
50
60
70
residual sugar


White Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Quality: 8
Density
0
1
2
3
4
0.2
0.4
0.6
0.8
1.0
1.2
sulphates


White Wine
Quality: 3
Quality: 4
Quality: 5
Quality: 6
Quality: 7
Quality: 8
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−100
0
100
200
300
400
500
600
total sulfur dioxide



For white, low quality wines tend to have high fixed acidity, high volatile acidity, high free/total sulfur dioxide.

3 Code

The associated code is in EDA.ipynb