Red Wine Quality Data Analysis

Sandra ROCHE

January 14, 2024

1 Introduction

Some wines are higher-quality than others due to several factors from climate to viticulture to winemaking, a myriad of factors make some wines exceptional and others run-of-the-mill.

Indeed, the terroir of wine has a clear-cut influence on its quality. Climate and weather help determine how quickly wine grapes grow, how much flavor and juiciness they have, and how well those grapes can be turned into wine. Moerover, to carry out photosynthesis, grape vines must be exposed to temperatures between 60 and 70 degrees Fahrenheit.

In addition to what the land and sky provide, the ways in which a producer manipulates the vines will also influence the quality of the resultant wine. Among other things, the winemaking process is equally important in determining the final quality of the wine. Wineries follow four main steps when producing their wines, maceration, fermentation, extraction and aging, and they must ensure consistency to get the most from their grapes. Inputs such as sulfur dioxide and processing enzymes, as well as decisions with oak barrel aging and oxygen management, all contribute to the quality of wine – from the exceptional to the insipid.

With high-quality wines, flavors may appear on the palate one after the other, giving you time to savor each one before the next appears. The five components – acidity, tannins, sugar/sweetness, alcohol and fruit – need to be balanced. For wines that need several years of aging to reach maturity, this gives them the time they need to reach optimal balance.

Based on the studied dataset which components influence the wine quality? Which components are liked? Moreover, is-it possible to predict the win quality?

2 Methodology

2.1 Dataset

The data set is provided by the Kaggle website https://www.kaggle.com/datasets/yasserh/wine-quality-dataset which contains a lot of data sets on multiple thematic.

The dataset is related to red variants of the Portuguese ”Vinho Verde” wine, located in the northwest of Portugal. It describes the amount of various chemicals parameters present in wine and a quality score. The dataset contains 1599 lines corresponding to different specific wine.

The study provides detailed statistics for each type of wine, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free and total sulfur dioxide, density, pH, sulfates, and alcohol content.

2.2 Variables

Variables description is summarized in the Table1. These physicochemical features are essential to understand the composition of wine and how each of them can influence the perceived quality of it.

Variables	Explanation
fixed acidity	presence of tartaric acid, essential for the wine's stability and flavor
volatile acidity	amount of acetic acid, high levels can lead to an unpleasant vinegar-like taste
citric acid	one of the main acids present, contribute to a refreshing flavor
residual sugar	amount of sugar remaining, more residual sugar are sweeter win
chlorides	amount of salt present
free sulfur dioxide	refers to free form of SO2, prevent microbial growth and wine oxidation
total sulfur dioxide	combined amount of free and bound forms of SO2
density	concentration of compounds
pH	describe acidity or basicity, scale ranging from 0 to 14
sulphates	salts or acids that contain sulfur, act as an antimicrobial agent
alcohol	volume percentage of alcohol present
quality	determined through sensory evaluation, score between 0 and 10

Table 1: Variables explanation

2.3 Method

First, dataset is imported from Kaggle. At the beginning, data are analyzed from a quality point of view including search of missing values, duplicate data and typography issue. Then, some variables distributions are observed.

Next, a deep analysis is performed to explore the dataset and try to find which features are linked to the win quality.

Finally, limitations and discussion are proposed at the end of the report.

3 Results

The dataset contains none missing values and none duplicate data. Moreover, no typography issue has been found in variables.

First is explored the quality of win corresponding to a score between 0 and 10 : 0 represents a bad quality and 10 a very high quality. As show in the Figure 1, 4% of win have a quality score equal to 3 or 4, 82% have a quality score equal to 5 or 6 and 14% equal to 7 or 8. Thus, majority of Portuguese win from the analyzed dataset are in the average (5.6).

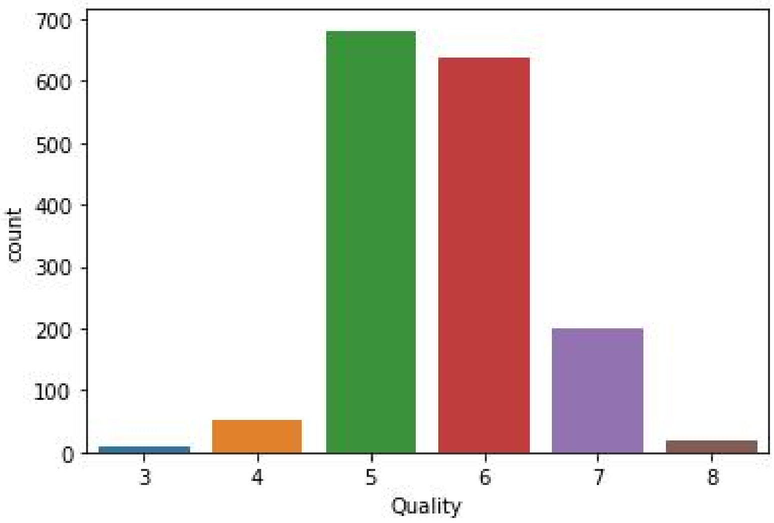

Figure 1: Data distribition for the win quality score

Next, linear relation between features and wine quality is analyzed. As show in the Figure 2, Wine quality is positively related by its alcohol content and negatively related by its volatile acidity. Sulfates and citric acid have also a moderate positive relation with the win quality.

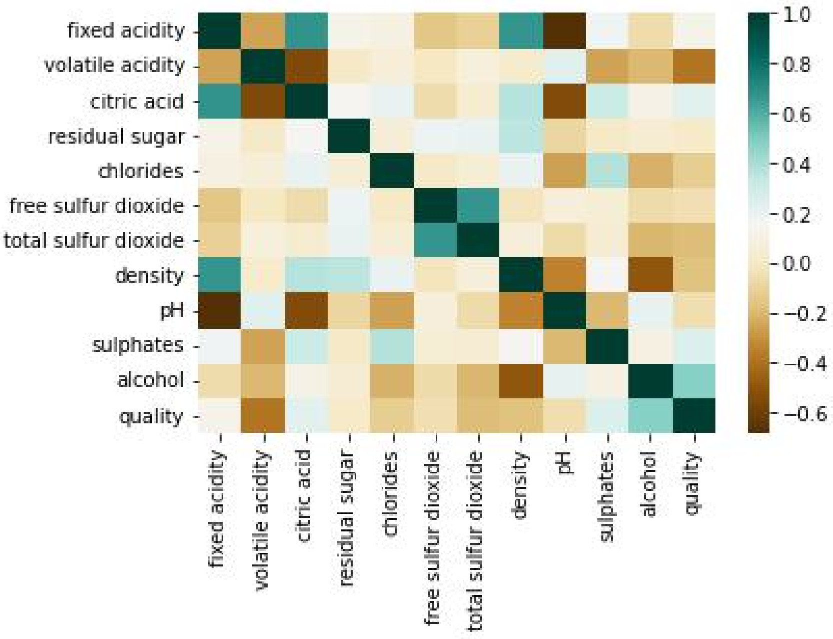

Figure 2: Linear correlation between all features from the dataset

From the Figure 2, deeper analysis show also positive correlation between features like density and fixed acidity, and citric acid and fixed acidity. Results are illustrated in the Figure 3.

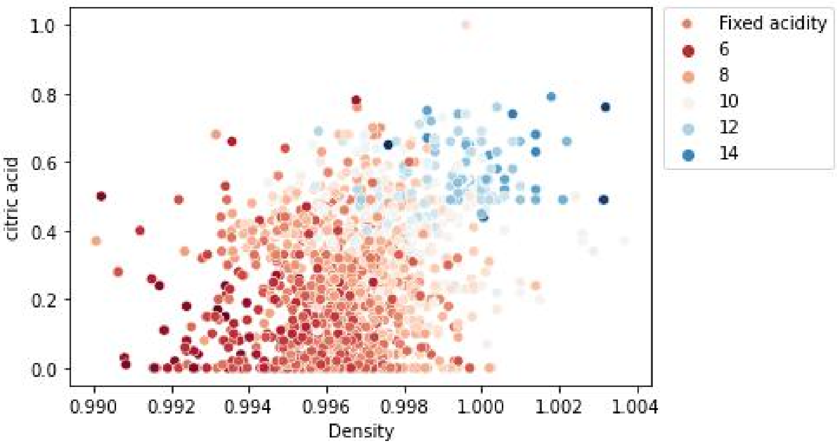

Figure 3: Density VS citric acid per fixed acidity

A random forest classifier model is applied to predict wine quality based on all features including in the dataset. Statistical accuracy refers to the degree to which the results of a statistical analysis are close to the true values. In this case, accuracy value obtained is 0.67 meaning that the model is not high confident in the result (close to 1).

4 Discussion

The dataset includes majority of wine with quality score of 5-6 (82%) which made it harder to identify each factor's different influence on a "high" or "low" quality of the wine. Based on the studied dataset from the Portuguese "Vinho Verde" wine, quality prediction seems to be difficult to be obtained. The wine quality depends on a lot of components that should be finely balanced and make the difference between a bad and a high quality wine. Some of these components presented in the introduction are not included in the dataset: climate, terroir, winemaking.

In fact, wine quality does not necessarily need moderation in each component – indeed, some red wines have higher acidity while others have a higher alcohol content. What makes the difference is that the other components balance things out. Another indicator of wine quality comes from typicity, or how much the wine looks and tastes the way it should. A final indicators of red wine quality are the intensity and finish. High-quality wines will express intense flavors and a lingering finish, with flavors lasting after you have swallowed the wine. Flavors that disappear immediately can indicate that your wine is of moderate quality at best. The better the wine, the longer the flavor finish will last on your palate.