DATA ANALYTICS AND MACHINE LEARNING FOR FINANCE

YASIN GUNER

**Table of Contents**

[1. Introduction 3](#_Toc184471592)

[2. Dataset Overview 3](#_Toc184471593)

[3. Exploratory Data Analysis (EDA) 4](#_Toc184471594)

[4. Conclusion 13](#_Toc184471595)

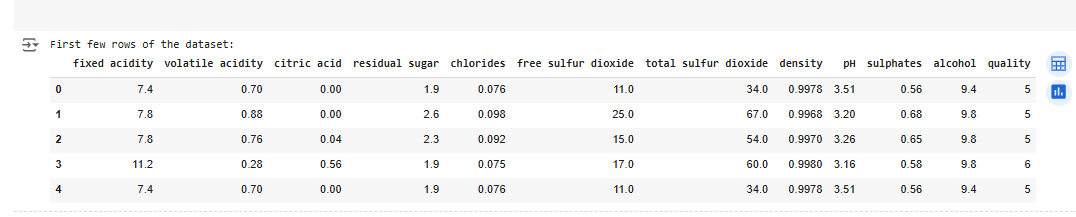
[5. Reflection 14](#_Toc184471596)

[References 15](#_Toc184471597)

# 1. Introduction

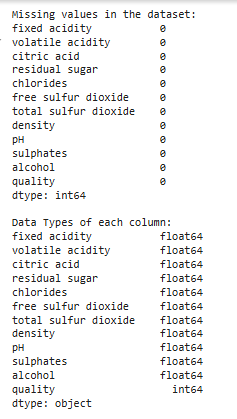
This report presents an ***“Exploratory Data Analysis (EDA)”*** on the "Wine Quality" dataset, which is used to examine the “relationship between various physicochemical properties of red wine and its sensory quality. The dataset contains 12 variables, including both continuous features, such as alcohol content, pH, and residual sugar, and a categorical target variable representing wine quality on a scale from 0 to 10.

# 2. Dataset Overview



**Figure 1: First few rows of the dataset**

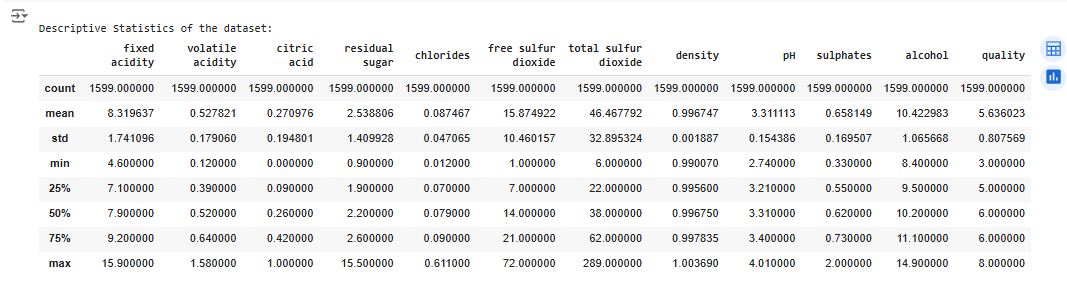
The "Wine Quality" dataset contains information about red variants of the Portuguese "Vinho Verde" wine, focusing on physicochemical properties and sensory quality ratings. The dataset consists of 12 variables, 11 of which represent the wine's chemical composition, and a target variable, "quality," which indicates the sensory rating given to the wine on a scale from 0 to 10.

****

**Figure 2: Dataset information and check the missing values**

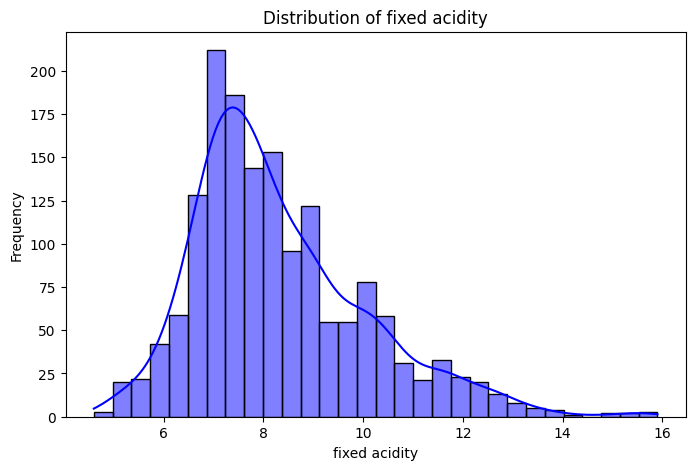
This figure shows the dataset information and checks for missing values. The study shows there are no missing values available in the dataset. These features are as follows: “fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol”.

# 3. Exploratory Data Analysis (EDA)



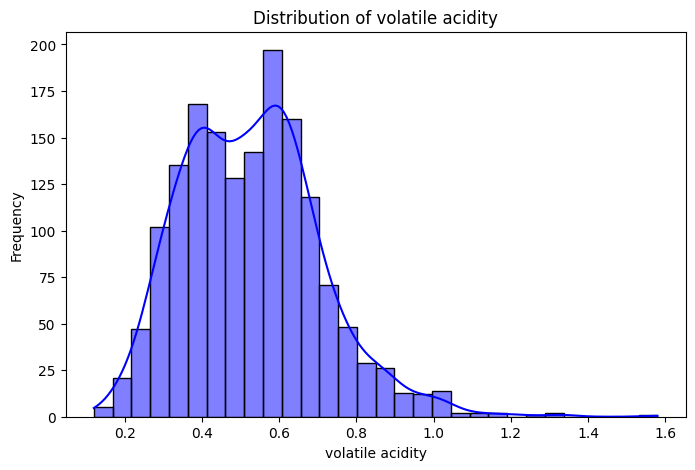
**Figure 3: Description of the dataset**

This figure shows information about the data set, with examples of the first rows and the general layout of the data. Rows describe 12 variables, 11 of which are the continuous input variables corresponding with the physicochemical characteristics of wines, and one output variable, “quality”, refers to the wine taste quality sensing.

****

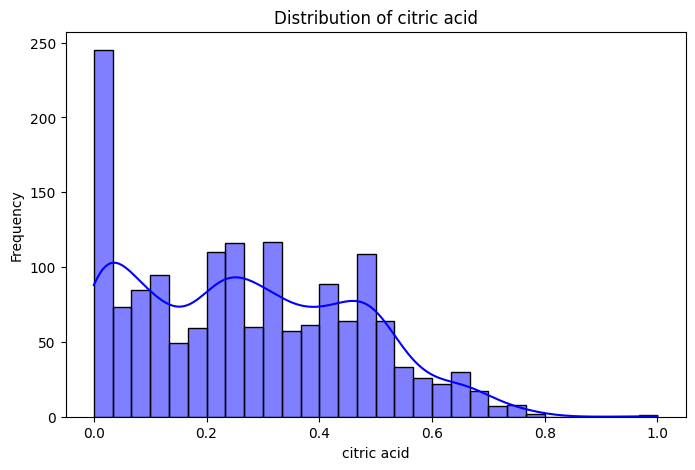
**Figure 4: Distribution of fixed acidity**

This figure exhibits the histogram of the fixed acidity feature, illustrating the measure of how the values of this feature are dispersed across the data set.

****

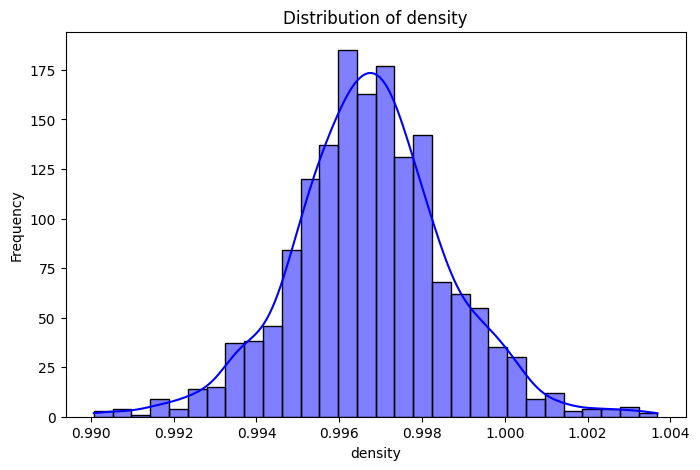
**Figure 5: Distribution of volatile acidity**

The volatile acidity distribution of pre-fermentation, alcoholic fermentation, and total volatile acidity is depicted in the above figure (Kumar *et al.* 2022). Many wines have low levels of volatile acidity, though it rises more significantly at the lower range of the graph.

****

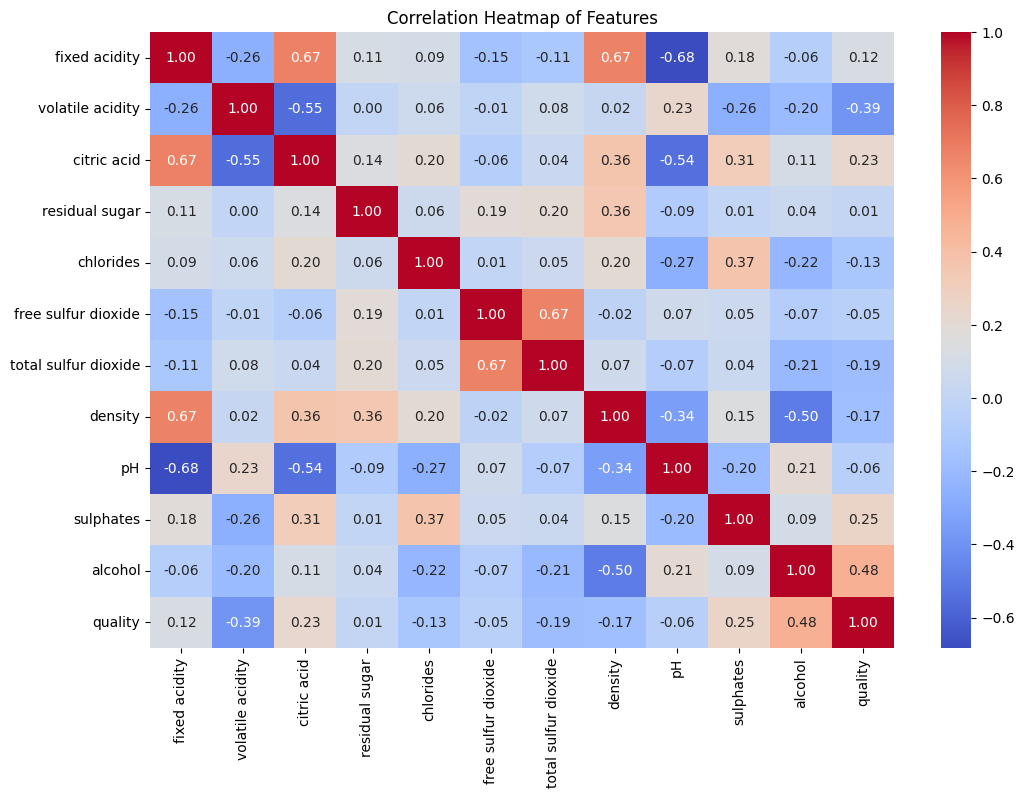
**Figure 6: Distribution of citric acidity**

The distribution of citric acid levels in the dataset is presented in the above figure. The citric acid histogram presents a clear view of the wine's citric acid content where the majority of the wines have low to moderate content of citric acid and most of the wines have no detectable citric acid.

****

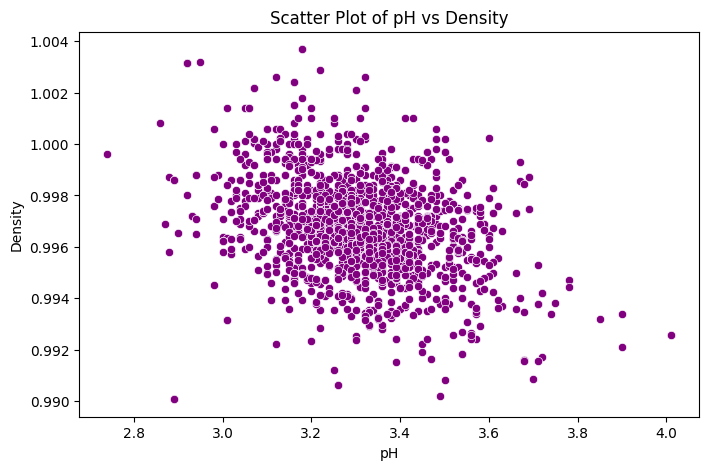
**Figure 8: Distribution of density**

The data represents the amount of density of the wines available in the dataset. This are expected based on the overall distribution depicted in the plot since the density values are shown to be confined at the lower end of the scale thus implying that most wines in the dataset are less denser.

****

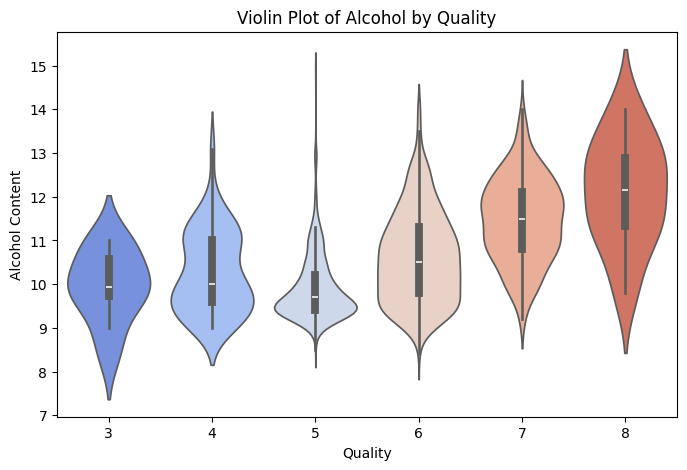
**Figure 9: Correlation Heatmap of Features**

This figure shows a correlation heatmap of the features in the dataset. It is a visual representation of how each feature correlates with one another (Nosratabadi *et al.* 2020). The heatmap helps identify strong positive or negative relationships between variables, such as the correlation between alcohol content and quality, offering valuable insights for understanding the interdependencies among physicochemical properties.

****

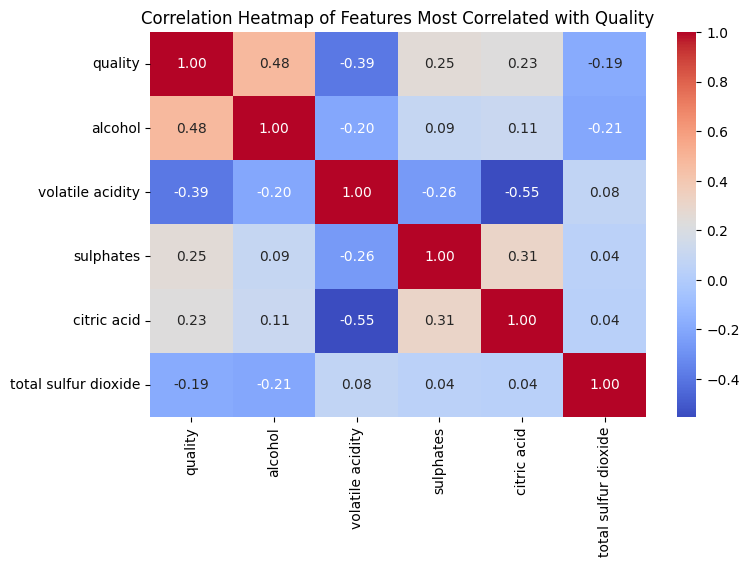
**Figure 10: Scatter Plot of pH vs Density**

A scatter plot of the pH and density of the solution is shown. The plot shows a slight negative relationship and that wines with smaller pH have slightly higher density. This way, possible connections between acidity and density can be traced, and what they might change in the wine, regarding its taste and smell.

****

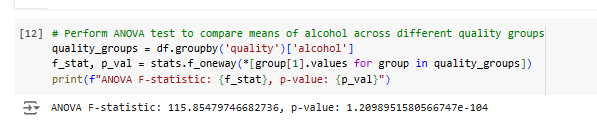
**Figure 11: Violin Plot of Alcohol by Quality**

This figure shows the distributions of the alcohol content of wines according to their quality ratings (Huang *et al.* 2020). This must be seen from an analysis of the plot chart where wines with higher alcohol content exhibit higher quality scores with a median alcohol content rising with the quality ratings.

****

**Figure 12: Correlation Heatmap of Features Most Correlated with Quality**

A correlation heatmap that addresses the features with a high correlation to quality. This means that while there are positive relationships between quality and certain physicochemical variables including alcohol and fixed acidity, others like volatile acidity indicate a negative relationship with quality. This figure is very important to determine which of the factors may have a positive or negative impact on wine quality.



**Figure 13: perform the ANOVA test**

This figure displays the results of the ANOVA test, showing a significant difference in alcohol content across quality groups.

# 4. Conclusion

This red wine dataset analysis proves elucidating on the correlation between the physicochemical properties and wine quality. That is why alcohol and, particularly, fixed acidity are positively related to quality”, while volatile acidity has a negative impact. These patterns have grouped the distribution plots and correlation heatmaps as the most relevant visualization. From this point of view, the study showed that some physicochemical properties are important to describe and predict wine quality, which creates a background for further researches.

# 5. Reflection

In this analysis, I have learned how several physicochemical properties affect wine quality. This form of data visualization assists me in recognizing important distributions and the positive correlation between alcohol content and quality levels. Ad hoc style functions such as Seaborn’s ‘heatmap’ and ‘boxplot’ enhance my analysis and interpretational skills. Furthermore, using statistical tests such, analysis of variance or ANOVA make me more confident in arriving at the data. This exercise strengthens my EDA skills as it further makes me remind how visualization and statistics play significant roles in identifying subsets within data sets.

# References

Huang, J., Chai, J. and Cho, S., 2020. Deep learning in finance and banking: A literature review and classification. Frontiers of Business Research in China, 14(1), p.13.

Kumar, S., Sharma, D., Rao, S., Lim, W.M. and Mangla, S.K., 2022. Past, present, and future of sustainable finance: insights from big data analytics through machine learning of scholarly research. Annals of Operations Research, pp.1-44.

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S.S., Reuter, U., Gama, J. and Gandomi, A.H., 2020. Data science in economics: a comprehensive review of advanced machine learning and deep learning methods. Mathematics, 8(10), p.1799.