Wine Exploratory Analysis - STA 135

Lejia Xu
918195599
aijxu@ucdavis.edu

September 11 2025

1 Introduction

This study uses a balanced subset of the Vinho Verde wine dataset with 600 samples. Each sample includes 11 physicochemical variables, a binary wine type, and a quality score (5–7), providing a basis to examine how chemistry differentiates wines by type and quality.

Two questions are addressed: (1) whether red and white wines can be accurately distinguished by chemical attributes, and (2) whether high-quality wines differ systematically from ordinary wines and can be reliably predicted from continuous variables. The first reflects clear separations from winemaking processes, while the second tests whether subtle sensory and compositional differences are captured by chemistry.

To answer these, exploratory data analysis (EDA) was conducted to summarize distributions, compare groups, and assess multivariate normality. Principal component analysis (PCA) was applied to extract major axes of variation, and linear discriminant analysis (LDA) and logistic regression were used as interpretable classifiers.

Findings show that red and white wines are highly separable, with LDA and logistic regression achieving near-perfect accuracy. Quality classification is more difficult: high- and low-quality wines overlap considerably, though alcohol, density, and volatile acidity remain predictive. The remainder of this report presents the EDA, PCA, classification analyses, and conclusions.

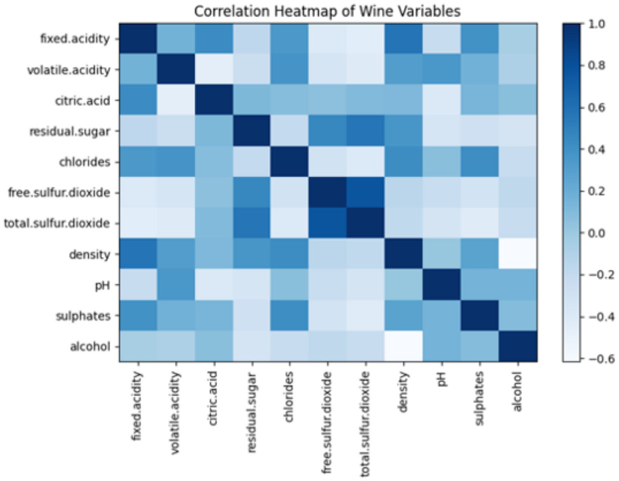
2 Exploratory Data Analysis

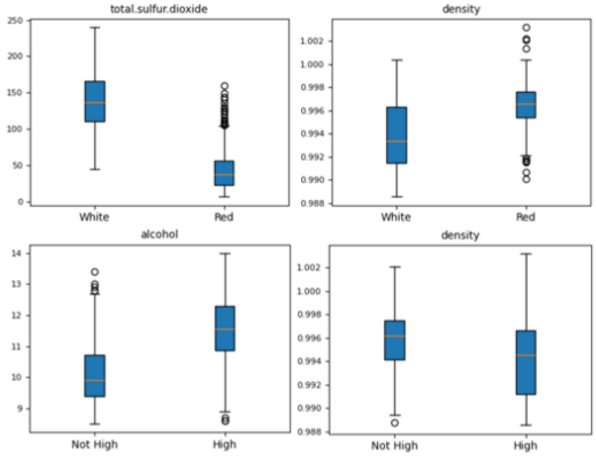
The dataset contains 600 wines, evenly split between 300 red and 300 white. Each is described by 11 continuous physicochemical variables (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) measured on different scales. The dataset also includes wine type and a quality score (5–7), from which a binary high vs. ordinary quality label was derived. No missing values were found.

The heatmap of correlations shows strong links such as free vs. total sulfur dioxide ($r \approx 0.77$) and alcohol vs. density ($r \approx -0.62$), while some pairs are nearly uncorrelated (e.g., pH vs. density, $r \approx 0.02$). This indicates redundancy in some variables and independence in others, supporting PCA for dimensionality reduction.

Boxplots illustrate groupwise distributions (median, quartiles, outliers). With many variables, only alcohol, total sulfur dioxide, and density are shown, while the rest are assessed using descriptive statistics and Mann–Whitney U tests.

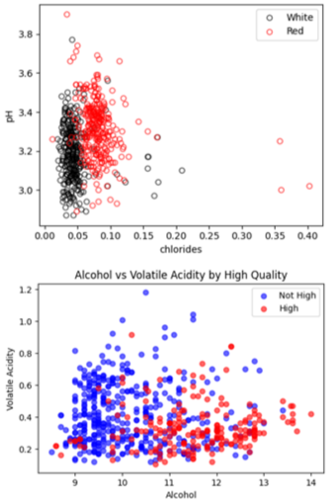
In red–white comparisons, red wines show higher volatile acidity, fixed acidity, density, and chlorides, while white wines are higher in sulfur dioxide and citric acid (all $p < 0.001$). For quality, high-quality wines have higher alcohol, citric acid, and sulphates, but lower density and volatile acidity; pH, free sulfur dioxide, and residual sugar are not significant. Type differences are driven by winemaking and are clear-cut, while quality differences are subtle, with alcohol and acidity balance as key markers. These statistical tests align with boxplot trends: tests confirm significance and direction, while plots provide intuition. Yet univariate analysis cannot capture joint patterns, motivating multivariate methods such as LDA and logistic regression.

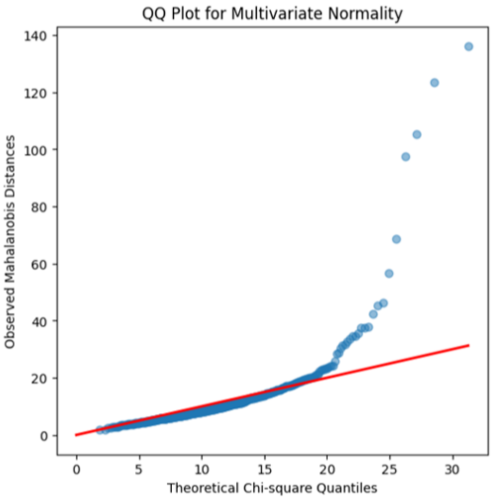


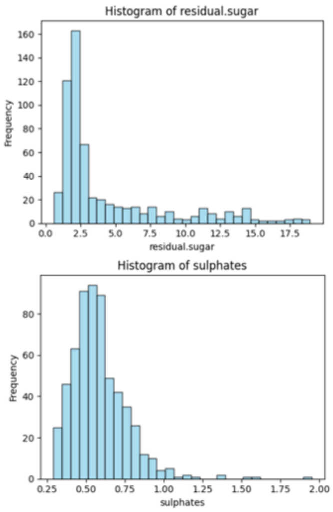


Effect size analysis shows that in red–white classification, total sulfur dioxide ($d \approx -2.78$), free sulfur dioxide ($d \approx -1.50$), and volatile acidity ($d \approx 1.50$) are strongest, with chlorides, sulphates, fixed acidity, residual sugar, and density also notable ($|d| \approx 1$). Alcohol ($d \approx 0.06$) and citric acid ($d \approx -0.30$) contribute little. For quality, effect sizes are weaker, but alcohol ($d \approx 1.32$) dominates, with high-quality wines showing higher alcohol, lower density ($d \approx -0.58$) and volatile acidity ($d \approx -0.44$), and slightly higher citric acid ($d \approx 0.45$) and sulphates ($d \approx 0.34$). Thus, red–white classification depends on strong contrasts, while quality rests on subtler signals.

Scatterplots further illustrate pairwise relationships. For type, chlorides vs. pH separates red and white fairly well, though with some overlap. For quality, alcohol vs. volatile acidity shows high-quality wines clustering in the high-alcohol, low-volatility region, but with substantial overlap. While weak, these trends support applying LDA to extract signals.





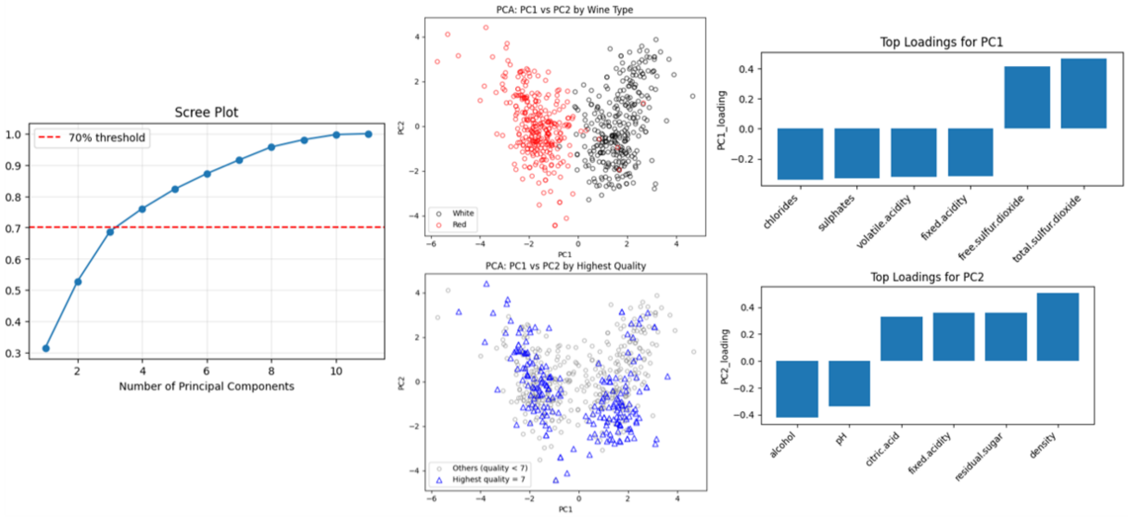


Multivariate normality was checked via a QQ plot of Mahalanobis distances and histograms. The QQ plot showed deviations, especially in the tails, indicating heavy-tailed distributions and outliers. Residual sugar was strongly right-skewed, sulphates skewed with a few extremes near 2 g/dm^3, and free sulfur dioxide long-tailed above 100 mg/L.

Thus, strict multivariate normality is not met, but PCA and LDA are robust to moderate violations, particularly after z-score standardization. All continuous variables were standardized to mean 0 and variance 1, ensuring comparability across different scales (e.g., g/dm^3 for sugar, mg/L for SO_2, % for alcohol, unitless for pH). Standardization also reduces the influence of extreme but valid observations and ensures PCA loadings and LDA covariance estimates are not dominated by variables with larger raw magnitudes.

2.1 Principal Component Analysis

The scree plot shows that PC1 alone explains about 31.6% of the total variance, PC1–PC2 together explain 52.9%, and PC1–PC3 capture 68.8%. An "elbow" appears between the third and fourth components, after which subsequent components contribute only minimal additional variance. This indicates that the first three principal components are sufficient to capture most of the dataset's structure.



The loadings of the first two principal components reveal that PC1 is dominated by total sulfur dioxide and free sulfur dioxide, with substantial contributions from chlorides, sulphates, volatile acidity, and fixed acidity. This axis can be interpreted as a "sulfur and acidity" dimension. PC2, on the other hand, contrasts density (positive) with alcohol (negative), supported by contributions from residual sugar, fixed acidity, pH, and citric acid. It can be interpreted as an "alcohol–density and sweetness" dimension.

In the scatter plot of PC1 versus PC2, the separation between red and white wines is most evident: red wines cluster in the negative region of PC1, while white wines are distributed in the positive region, suggesting that wine type differences are the primary source of chemical variation in the data. By contrast, in the quality classification, high-quality and ordinary wines exhibit substantial overlap in this space. High-quality wines do not form a distinct cluster but show some concentration in the region of higher alcohol and lower density. This indicates that quality-related differences are relatively weak in the principal component space.

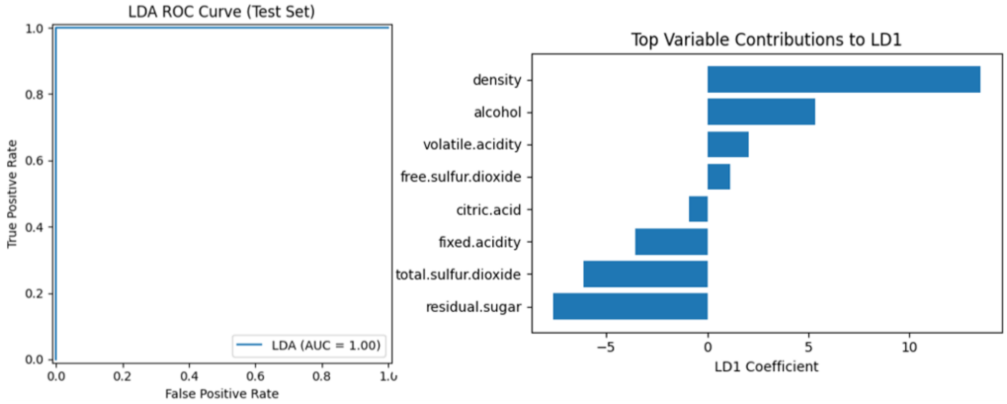
Overall, the first three components capture nearly 70% of the total variance while maintaining good chemical interpretability. The clear separation of red and white wines provides a foundation for subsequent supervised classification, whereas the fuzziness of quality-based grouping suggests the need for more complex multivariate models to extract meaningful signals.

3 Methodology

3.1 Linear Discriminant Analysis

3.1.1 LDA for Red–White Classification

In the red–white classification, LDA achieved near-perfect results. With a 70/30 stratified split, test accuracy was 99.4% and AUC 1.000, with precision and recall both ≈ 1. Cross-validation confirmed robustness (mean accuracy 99.2% $\pm 0.5\%$, AUC 0.996 ± 0.003), and the confusion matrix showed only a few errors. The **ROC curve** lay almost against the upper-left corner, further confirming the model's strong discriminative power. These outcomes reflect strong chemical contrasts that LDA consolidated into a single discriminant axis.



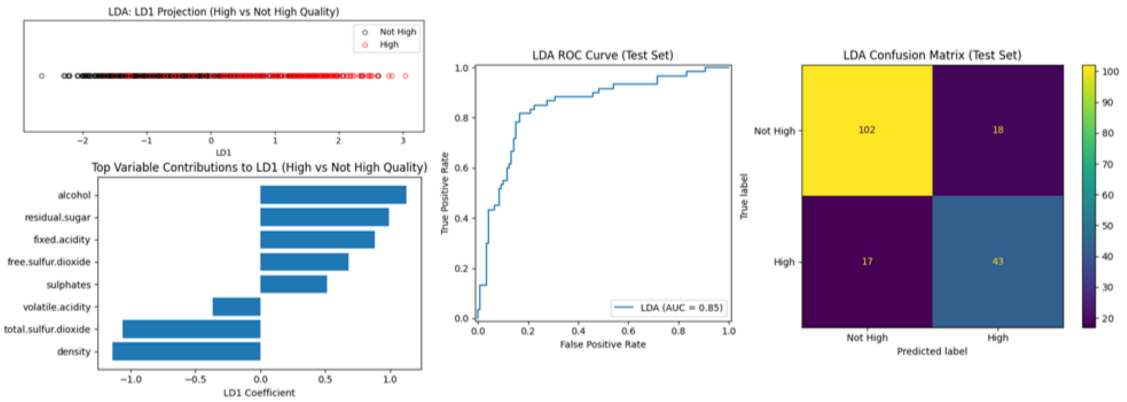
From the **variable contribution plot**, density was the strongest positive driver toward red wines, while residual sugar and total sulfur dioxide were the strongest negatives, marking white wines. Alcohol, volatile acidity, sulphates, and fixed acidity contributed moderately, free sulfur dioxide negatively, and citric acid and pH minimally.

Overall, LDA combined clear chemical differences into one interpretable axis, yielding nearly flawless classification that is both statistically robust and chemically intuitive.

3.1.2 LDA for Quality Classification

For quality classification, **LD1 projection** showed much weaker separation: high and low quality wines largely overlapped, with only a slight clustering of high-quality wines in regions of higher alcohol and lower density. With the same 70/30 stratified split, performance dropped sharply: test accuracy was 72%–75% with AUC 0.77–0.80; precision and recall were limited. Cross-validation confirmed modest results, with mean accuracy about 73% and higher variability.

The **ROC curve** indicated better-than-random classification but far weaker than red–white results. The **confusion matrix** showed frequent errors, especially ordinary wines misclassified as high-quality. The **variable contribution plot** identified alcohol as the dominant positive driver, density and volatile acidity as strong negatives, citric acid and sulphates as moderate positives, and free sulfur dioxide, residual sugar, and pH as negligible. These align with effect size analysis, confirming alcohol and acidity balance as the key signals.

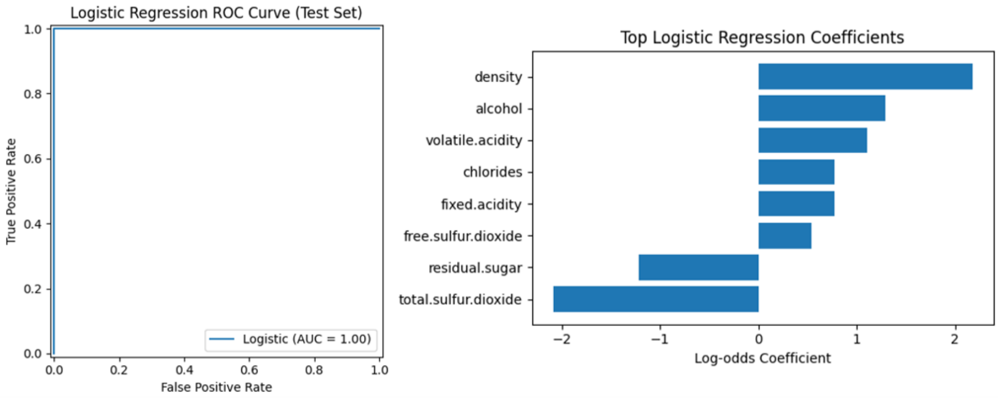


Overall, LDA consolidated weak differences into a linear axis, clarifying the role of alcohol and acidity balance. Yet its performance lagged far behind the red–white case, showing that linear discriminants alone cannot achieve reliable quality prediction, and more advanced multivariate or nonlinear models are needed.

3.2 Logistic Regression

3.2.1 Logistic Regression for Red–White Classification

Logistic regression was applied with a 70/30 stratified split and L2 regularization. Model performance was nearly perfect: test accuracy reached 100%, with both precision and recall equal to 1 and an AUC of 1.000. Cross-validation results were consistent, yielding an average accuracy of 99.2% ($\pm 0.5\%$) and an AUC of 0.993 (± 0.004). These metrics are almost identical to those of LDA, indicating that the chemical differences between red and white wines are so pronounced that linear methods alone can achieve near-perfect separation. The **ROC curve** lies almost against the upper-left corner, confirming the model's strong discriminative power, while the confusion matrix shows only a handful of errors, further demonstrating its stability.



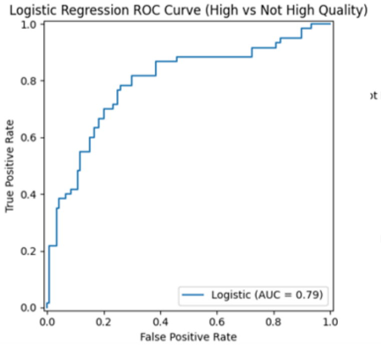
According to the **Top Logistic Regression Coefficients plot** and regression results, density ($\beta = 2.18$) was the strongest positive predictor, with alcohol ($\beta = 1.29$) and volatile acidity ($\beta = 1.10$) also contributing substantially, all pushing samples toward the red class. In contrast, total sulfur dioxide ($\beta = -2.09$) and residual sugar ($\beta = -1.22$) were the strongest negative predictors, reflecting higher values in white wines. Other variables such as chlorides ($\beta = 0.77$), fixed acidity ($\beta = 0.77$), free sulfur dioxide ($\beta = 0.54$), and pH ($\beta = 0.42$) added smaller positive contributions, while sulphates ($\beta = 0.36$) and citric acid ($\beta = -0.35$) played relatively minor roles.

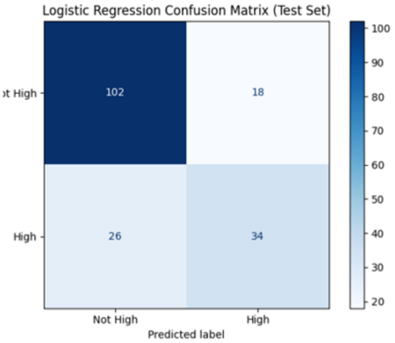
Overall, logistic regression results are highly consistent with those from LDA and the exploratory analysis: red wines are characterized by higher density, alcohol, and volatile acidity, whereas white wines stand out for higher sulfur dioxide and residual sugar. Logistic regression not only achieved near-perfect predictive performance but also provided a probabilistic and interpretable model fully consistent with chemical intuition.

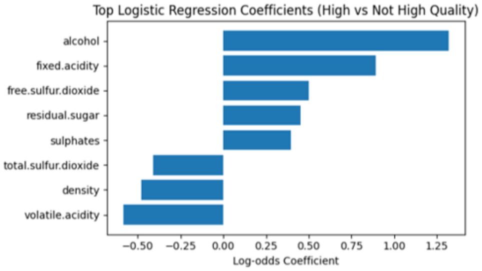
3.2.2 Logistic Regression for Quality classification

In the quality classification task, logistic regression was applied with a 70/30 stratified split and L2 regularization. Model performance was moderate: test accuracy reached 75.6% with an AUC of 0.795. Precision was 0.80 for ordinary wines but only 0.65 for high-quality wines, with recall dropping to 0.57. Cross-validation showed consistent results, with an average accuracy of 78.5% ($\pm 2.3\%$) and an AUC of 0.849 (± 0.035).

As shown in the figures, the **ROC curve** indicates performance clearly above random, but far weaker than the near-perfect red–white results, reflecting limited discriminative signal. The **confusion matrix** further reveals frequent misclassifications, especially 26 high-quality wines predicted as ordinary, leading to a lower recall for the positive class.







According to the **Top Logistic Regression Coefficients plot** and regression results, alcohol ($\beta = 1.32$) is the strongest positive predictor, followed by fixed acidity ($\beta = 0.89$) and free sulfur dioxide ($\beta = 0.50$). Residual sugar ($\beta = 0.45$), sulphates ($\beta = 0.39$), and pH ($\beta = 0.38$) also contribute positively, though less strongly. In contrast, volatile acidity ($\beta = -0.58$), density ($\beta = -0.48$), total sulfur dioxide ($\beta = -0.41$), and chlorides ($\beta = -0.18$) are negatively associated with quality, while citric acid shows almost no effect.

Overall, logistic regression confirms the findings from effect size analysis and LDA: high-quality wines are associated with higher alcohol and better acidity balance. However, due to the substantial overlap between quality groups, linear models have limited discriminative power. While logistic regression provides probabilistic and interpretable results, further improvements in predictive performance will require more complex multivariate or nonlinear approaches.

4 Discussion

This study analyzed the chemical differences between red and white wines as well as between different quality levels, and evaluated the performance of PCA, LDA, and logistic regression.

The separation of red and white wines was highly pronounced: PCA already showed clear distinction, while LDA and logistic regression achieved near-perfect classification. Red wines exhibited higher density, alcohol, and volatile acidity, whereas white wines contained more sulfur dioxide and residual sugar. These contrasts were consistently confirmed by statistical tests, effect sizes, and model coefficients.

Quality classification proved more challenging: although alcohol and acidity balance were key features, high- and low-quality wines showed substantial overlap. Model accuracy was only 72%–78% with AUC around 0.77–0.85. The ROC curve and confusion matrix indicated low recall for high-quality wines, suggesting that physicochemical indicators alone cannot fully capture sensory quality.

In terms of methods, LDA and logistic regression were highly consistent in both performance and interpretation, jointly highlighting the importance of alcohol, density, and acidity. PCA, while not directly used for classification, provided a reasonable basis for dimensionality reduction.

Limitations remain: all models applied were linear and thus unable to capture nonlinear relationships; furthermore, the quality label was simplified into a binary classification, which may overlook finer distinctions.

Overall, this study confirmed the strong chemical differences underlying red–white classification, while also revealing the complexity of quality prediction. Future research should integrate additional chemical and sensory attributes and explore nonlinear or ensemble models to enhance both the accuracy and interpretability of wine quality prediction.

5 Conclusion

This study analyzed 600 wine samples to compare chemical differences in red–white and quality classifications, using PCA, LDA, and logistic regression. The results show that:

- Red and white wines differ strongly in chemical features, allowing linear methods to achieve near-perfect separation.
- Quality classification provides weaker signals; although alcohol and acidity balance are key predictors, overall performance remains limited.

In summary, physicochemical indicators are sufficient to explain type differences but insufficient to fully capture wine quality. Future research should expand the sample size, explore nonlinear or ensemble methods to further improve prediction and interpretation of wine quality.

6 Appendix

Source code available at: GitHub Repository (wine.ipynb)