Blue Berry Winery

Wine Quality Analytics

Source: https://blog.liebherr.com/appliances/us/vinho-verde-the-ideal-summer-wine-from-portugal/

Anna Szczepara

August 2024

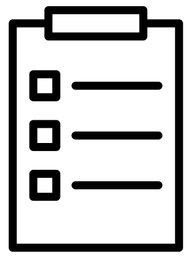
Agenda

- Introduction
- What did we learn from the datasets?
- Machine Learning (ML) Models
- Price prediction

Introduction

Questions asked

- **How does the composition of wines relate to their quality?**
- Are the chemical **compositions of red and white wines comparable**?
- Does the **chemical composition of wine influence its perceived quality**?
- What **additional insights can be drawn from the datasets?**



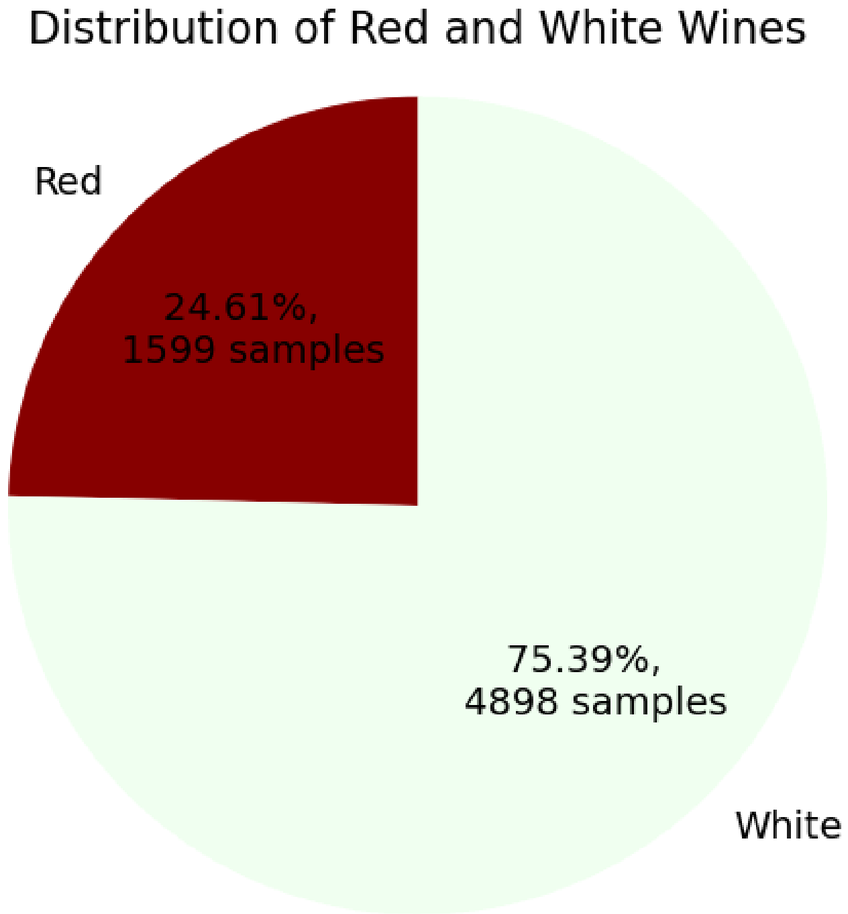
Data description & methodology

Red and white data set description:

- 6497 samples (red + white) wine samples in total,
- White wines dominate the sample size,
- Points (0-10) were used to label ‘low’: <5, ‘medium’: [5-7], ‘high’: >7 quality,
- White wines have higher average residual sugar and sulfur (both free and total),
- Density and pH are similar across both wine types.
- On average, **red wines have a lower quality rating than white wines,**
- Sulphates tend to be higher in red wines compared to white wines.

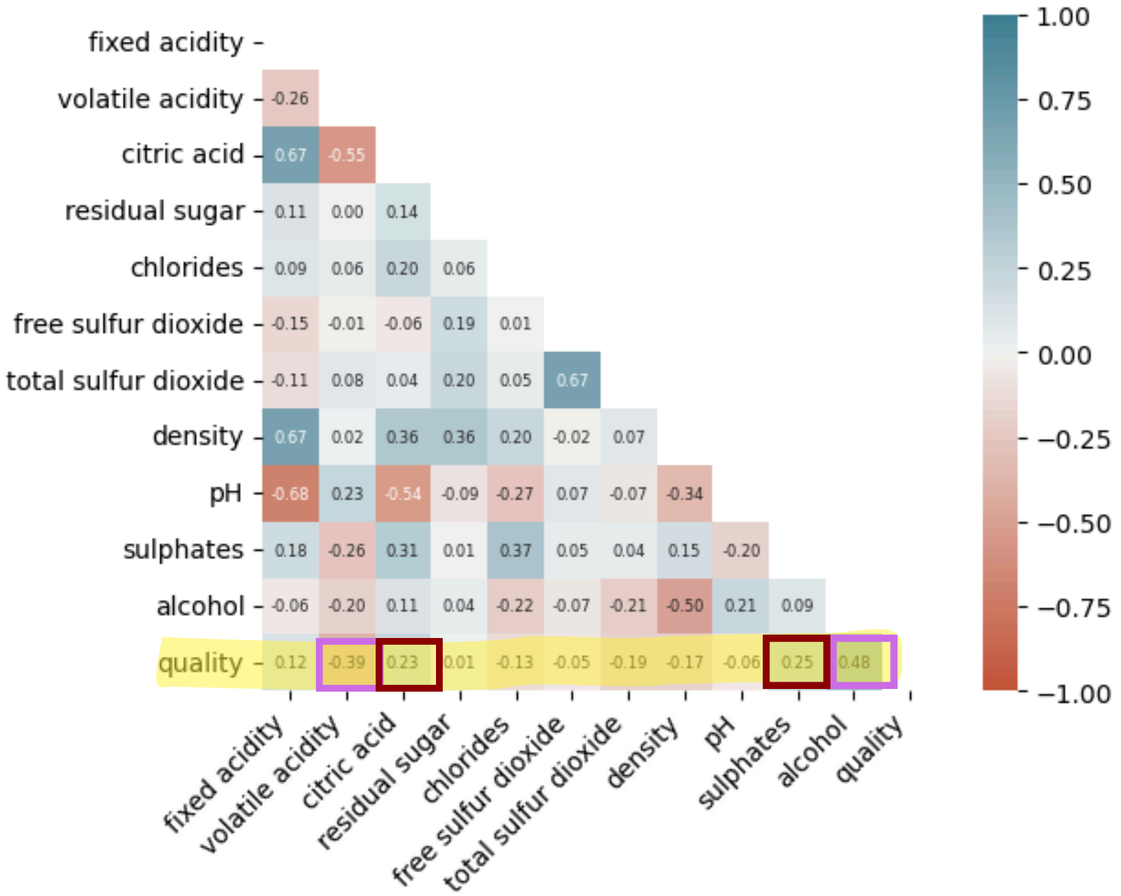
Methodology used in the process of **descriptive analysis:**

- **Data standardisation** - process of rescaling features to have a mean of zero and a standard deviation of one, thereby **transforming them to a common scale without distorting differences in the range of values,**
- **ANOVA (Analysis of Variance)** is a statistical method used to **compare the means of three or more groups to determine if there are any statistically significant differences between them.**
- **Combining red and white data for some parts analysis.**

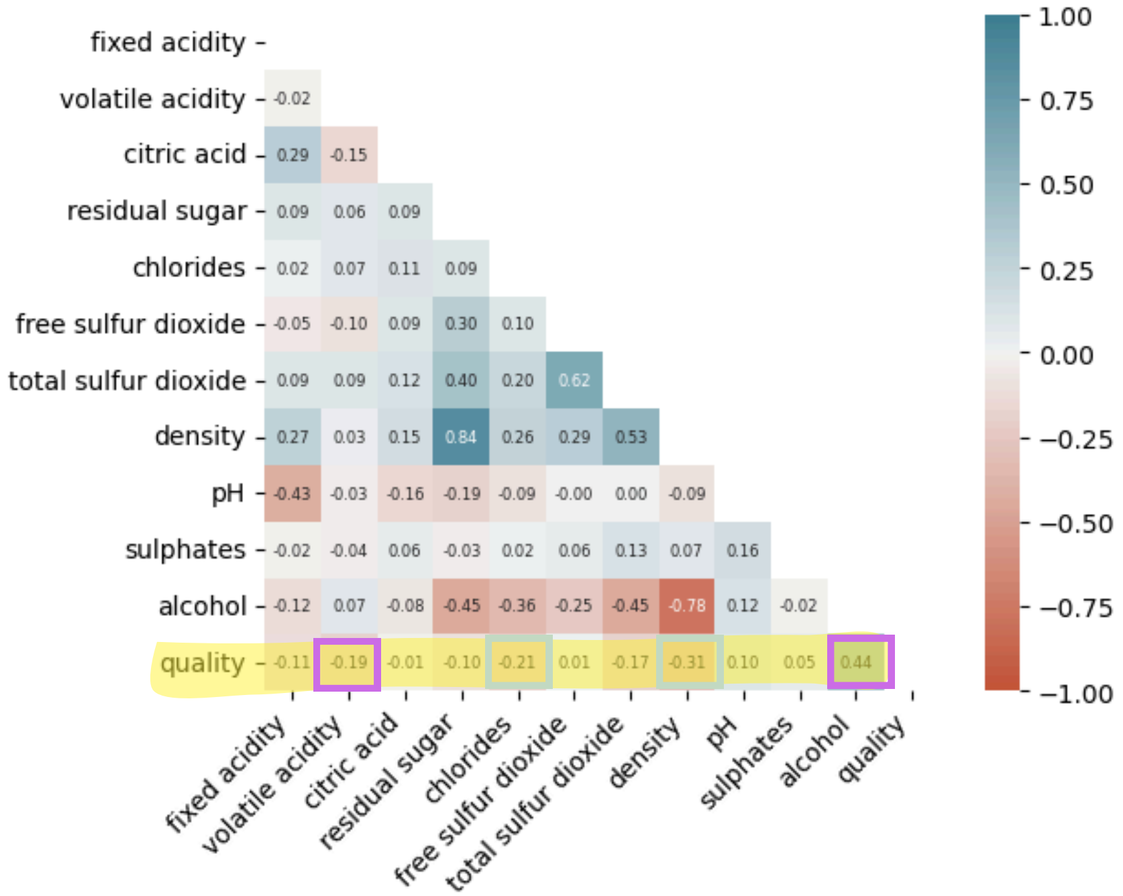


What did we learn from the datasets?

Which features are important
Correlation matrix



Correlation matrix **red** wine



Correlation matrix white wine

Which features are important
ANOVA* tests & external resources

Important features coming from **ANOVA testing** with different quality ratings **confirmed** most of the correlation matrix dependencies:

- **Red** wine (Alcohol, Fixed Acidity (FA), Volatile Acidity (VA), Citric Acid (CA), Sulphates)
- **White** wine (Alcohol, Fixed Acidity (FA), Volatile Acidity (VA), Sulphates, Chlorides)

External research:

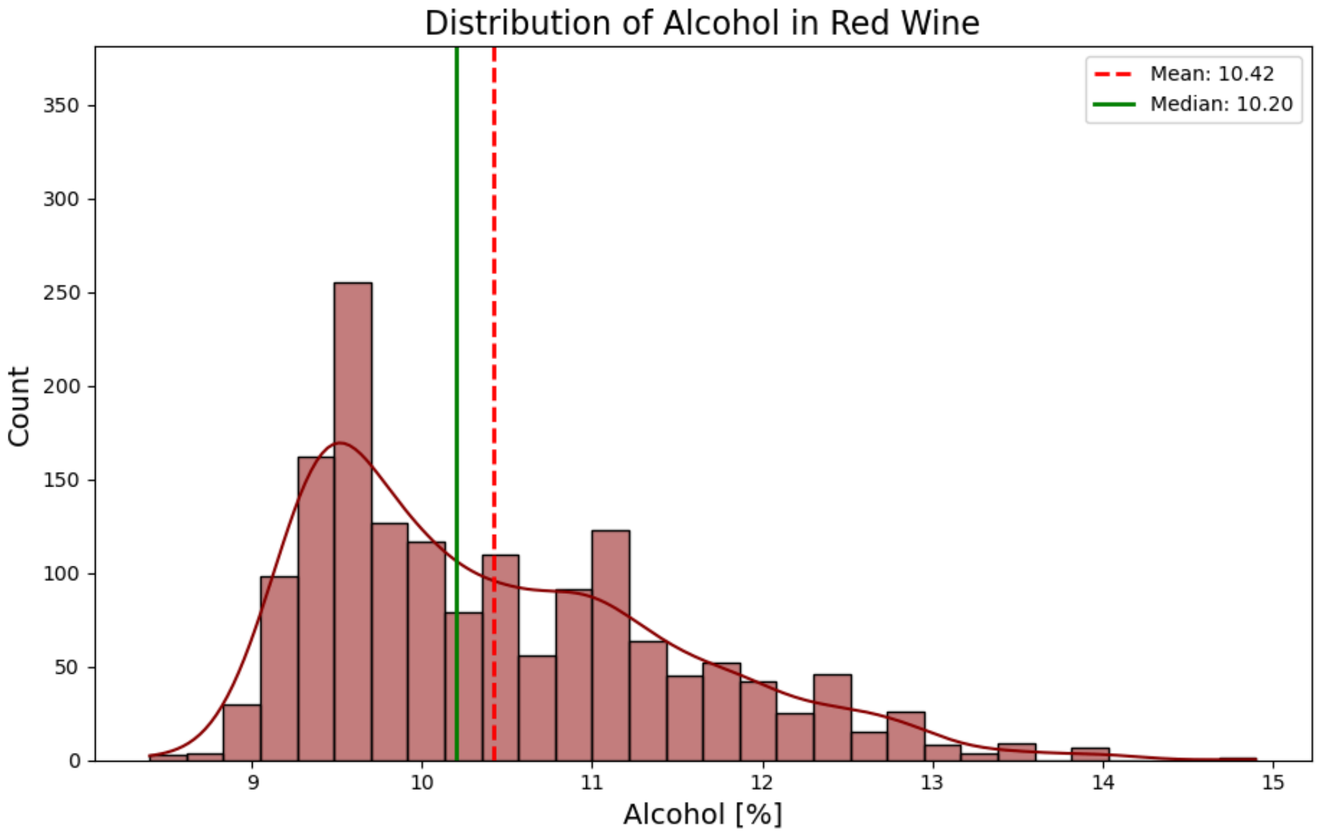
"Winemakers are usually most concerned with acetic acid, which accounts for more than 93% of steam distillable acids in wine."

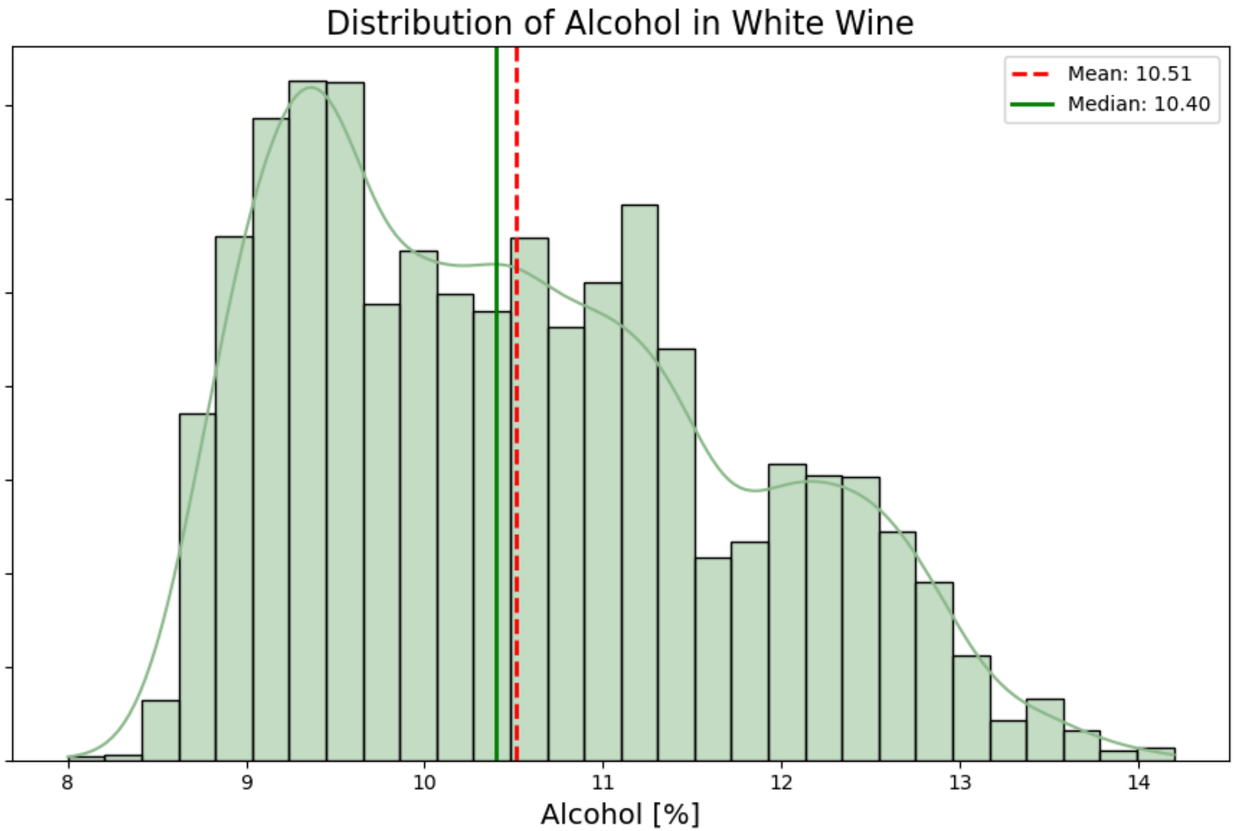
"Analysis of **volatile acidity (VA)** was probably the wine industry's **first measure of wine quality** and is **routinely used as an indicator of wine spoilage**."

https://www.awri.com.au/wp-content/uploads/2018/03/s1982.pdf

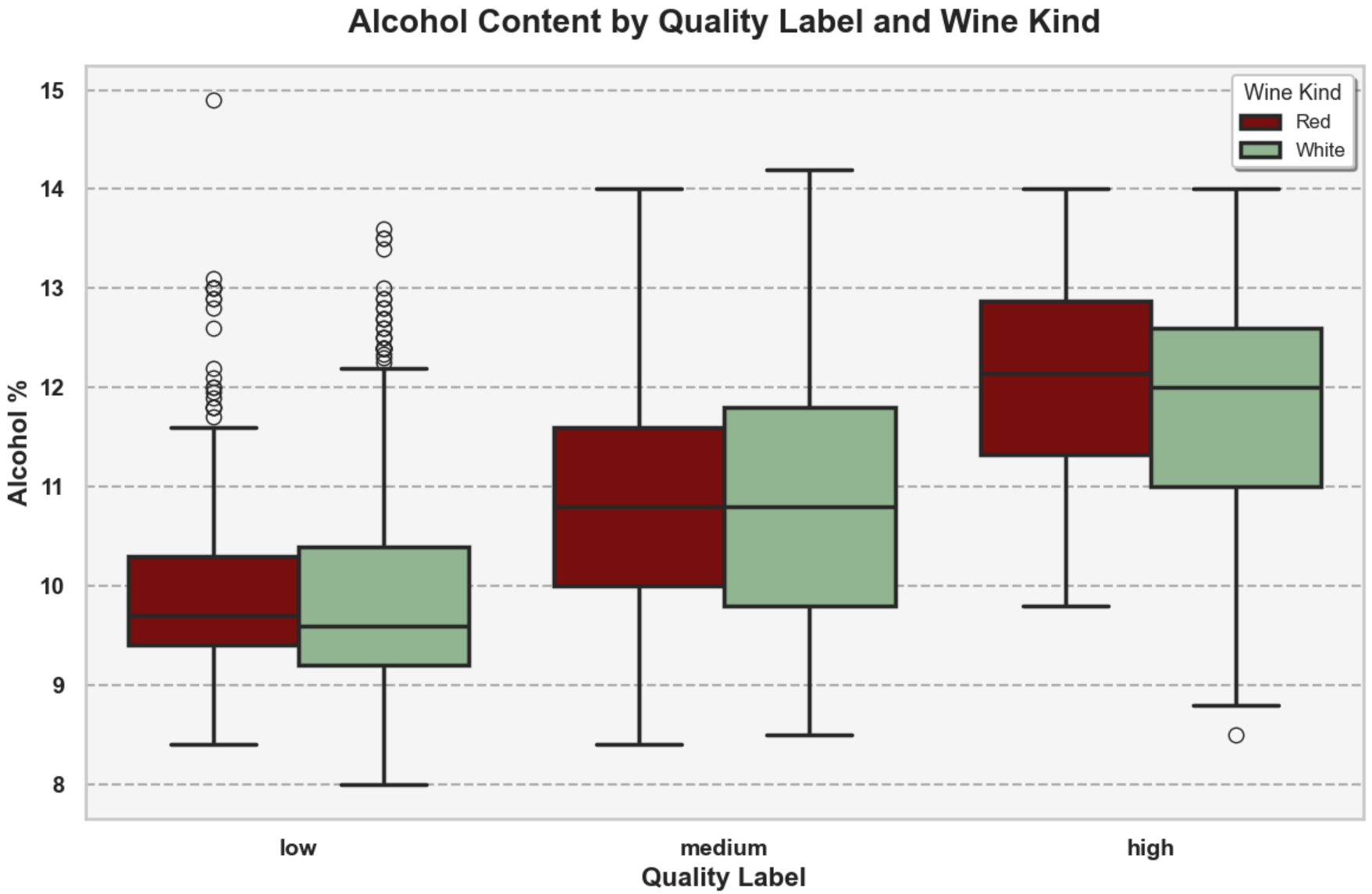
*ANOVA (Analysis of Variance) is a statistical method used to compare **the means of three or more groups** to determine if there are any statistically **significant differences between them**. It tests the null hypothesis that all group means are equal, by analyzing the variance within and between groups.

Exploratory Data Analysis

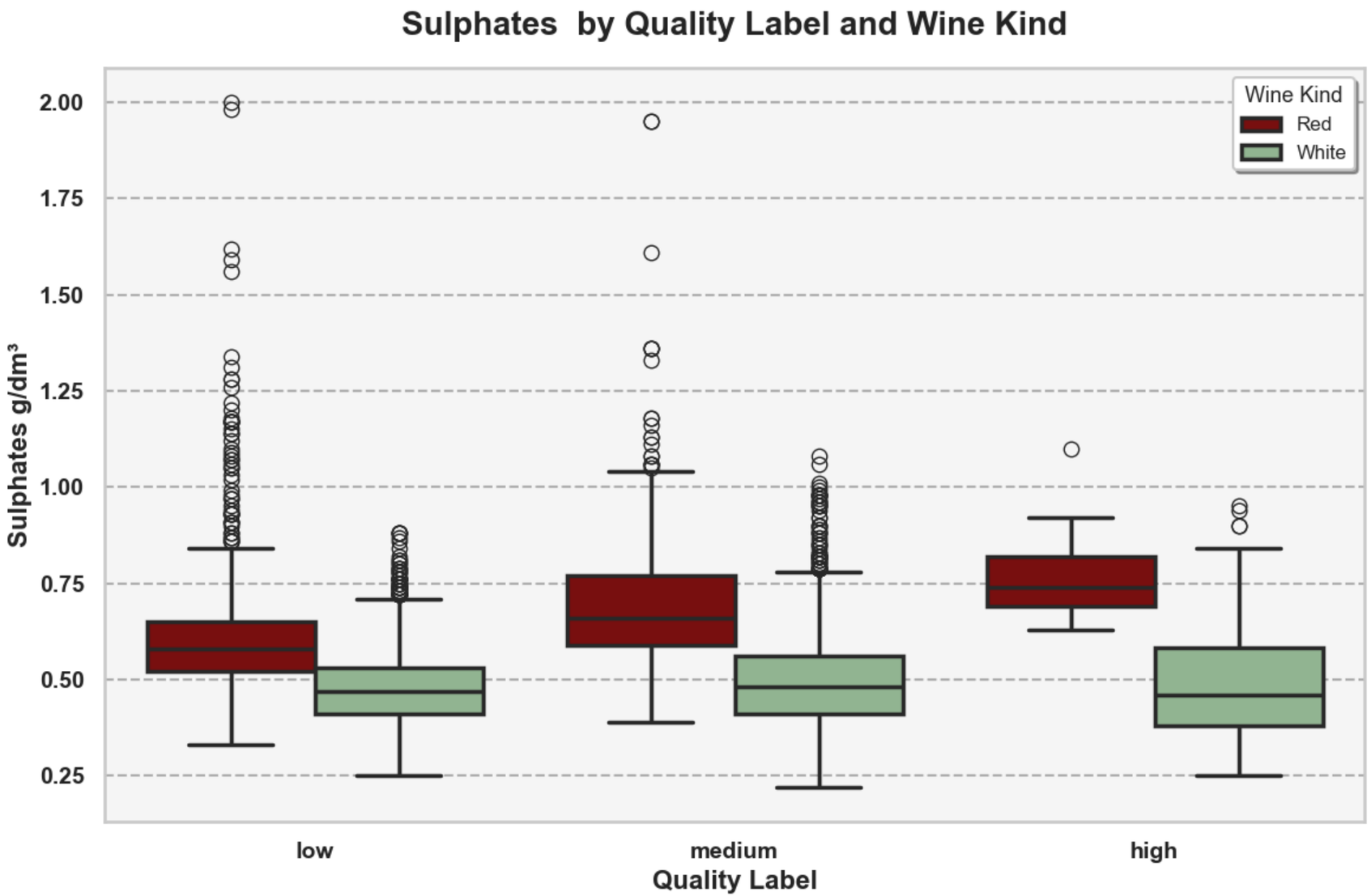




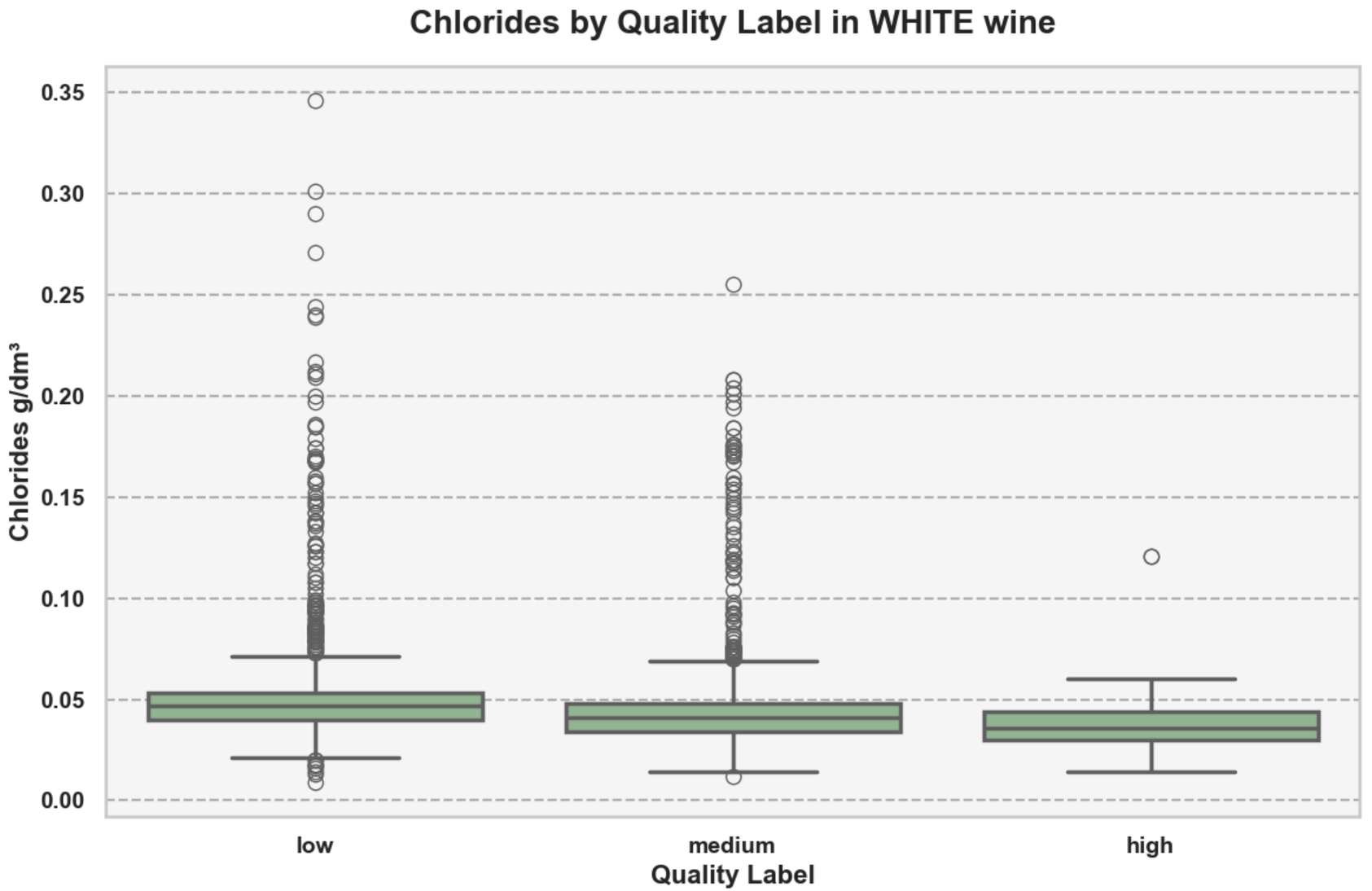
Exploratory Data Analysis



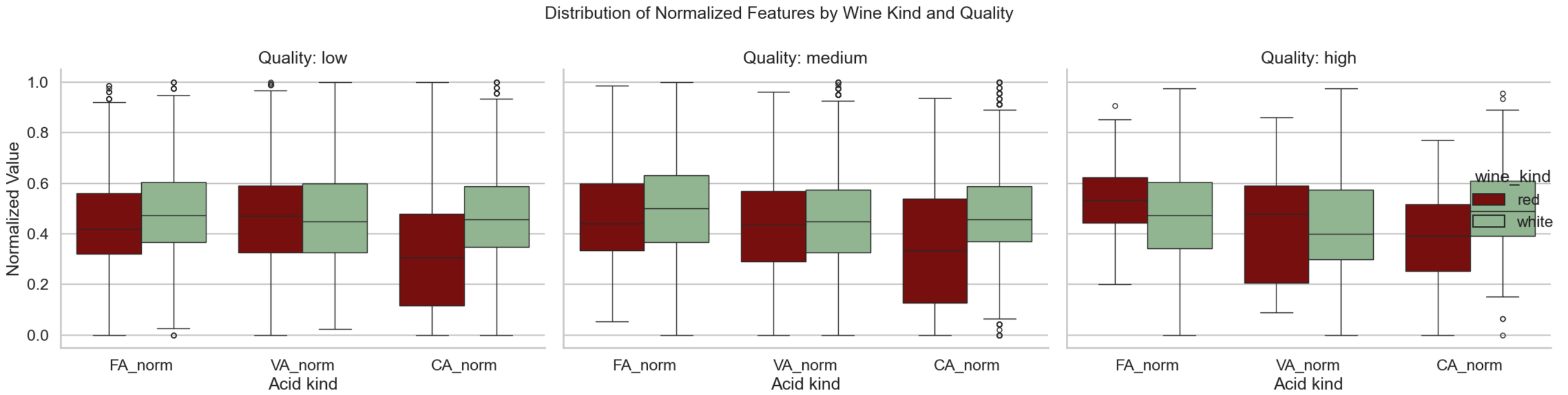
Exploratory Data Analysis



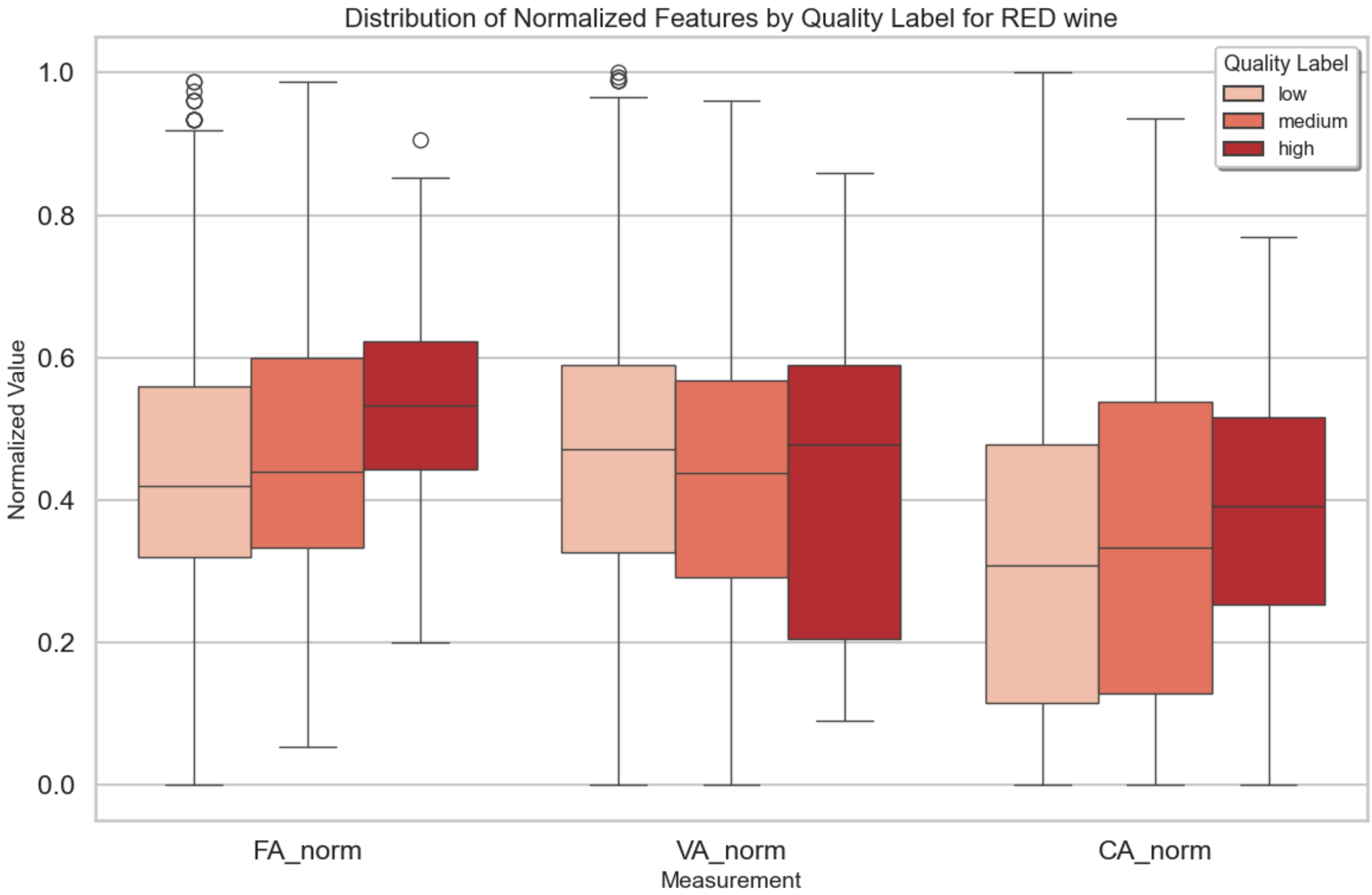
Exploratory Data Analysis



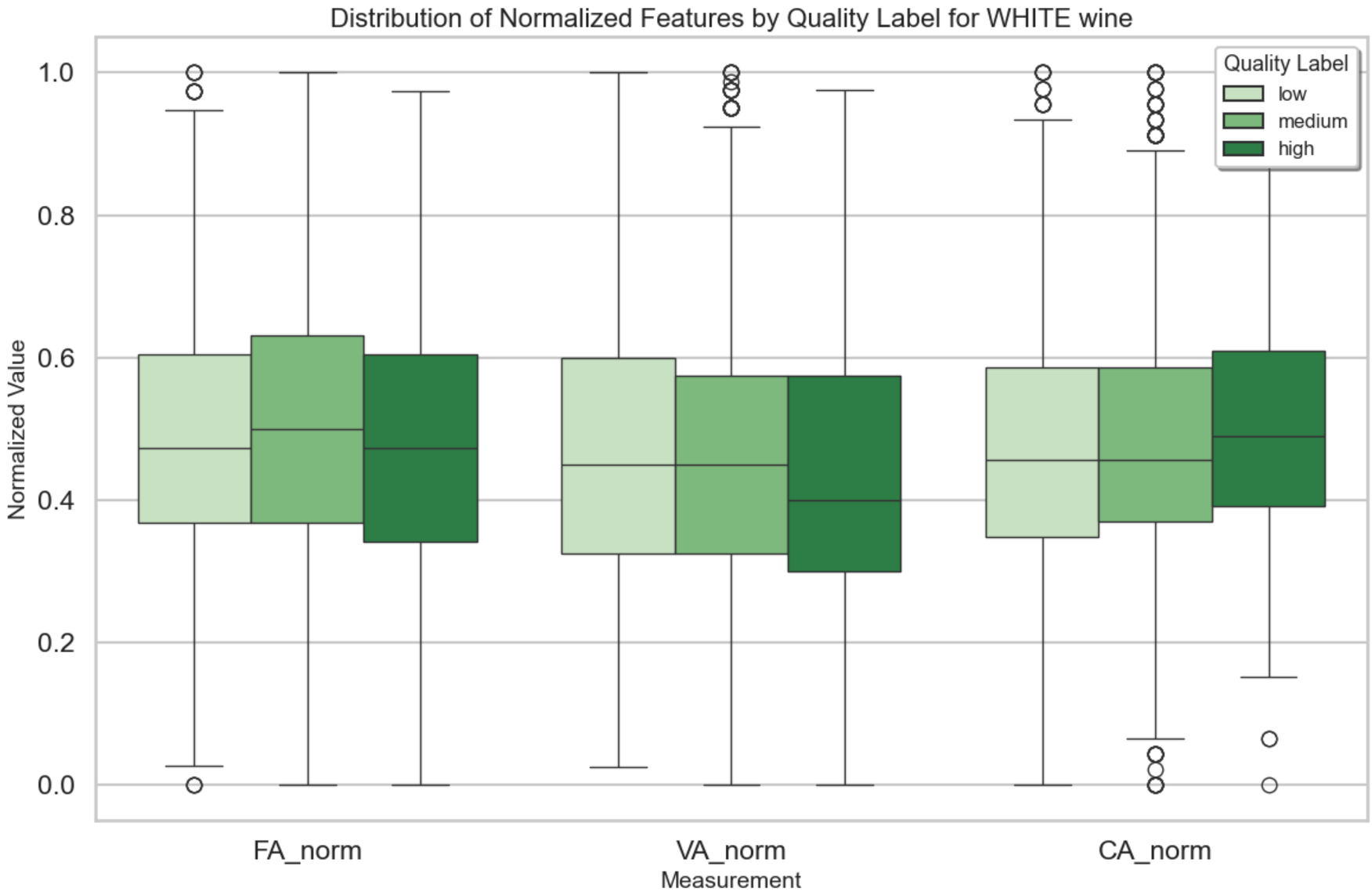
Exploratory Data Analysis



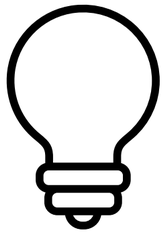
Exploratory Data Analysis



Exploratory Data Analysis



Conclusions Exploratory Data Analysis



- **Alcohol** content is a big factor for quality. **Alcohol % rises with quality.**
- **Sulphates** content play important role in **red wine quality**. Like with alcohol, the **higher amount of sulphates, the better red wine quality** we get.
- **Acids content** is also important for quality of wine.
  - **Red wine: Fixed acidity and citric acid** amount rises with the quality. **Volatile acidity** should decrease with the wine better quality.
  - **White wine:** Generally acidity is less influential on the wine quality. However, we see **similar behaviour as in red wines acidity.**

The results demonstrate that, for **both red and white wines,** characteristics such as **alcohol, fixed acidity, volatile acidity, and sulphates are crucial in determining perceived quality. In general we** confirm **the importance of these chemical variables** identified in the quality analysis.

Machine Learning (ML) models

Used Machine Learning (ML) methods

ML methods used in additional data set analysing **to predict prices according to quality:**

- Logistic Regression: A linear model used for binary or multi-class classification that predicts probabilities using a logistic function.
- Random Forest: An ensemble learning method that builds multiple decision trees and combines their outputs to improve prediction accuracy and reduce overfitting.
- K-Nearest Neighbours (KNN): A non-parametric method that classifies data points based on the majority class of their nearest neighbours in feature space.
- Naive Bayes (NB): A probabilistic classifier that applies Bayes' theorem with the assumption of independence between features.

Validation and **parameters' hypertunig methods:**

- **K-fold cross-validation** is a resampling method used to evaluate a machine learning model's performance. The data is split into **k** equally sized folds, where the model is trained on **k-1** folds and tested on the remaining fold. This process is repeated *k* times, with each fold serving as the test set once, and the results are averaged to provide a more robust estimate of model performance.
- **Confusion matrix, classification report, accuracy**
- **Grid Search CV -** thorough hyperparameters tuning method; it exhaustively searches through all possible combinations of hyperparameters within the specified grid. This ensures that the model finds the best hyperparameter set, making it more likely to achieve optimal performance.

Quality prediction - ML Models' results

	Red wine	White wine
Logistic Regression	0.7219	0.7143
Random Forest	**0.7875**	**0.8244**
KNN	0.7011	0.7173
NB	0.7115	0.6758

Machine learning models, **particularly Random Forest,** were effective in **predicting wine quality,** standing out as the best-performing method. It provides a **solid foundation for future quality predictions** and **decisions** related to improving wine production.

Price prediction

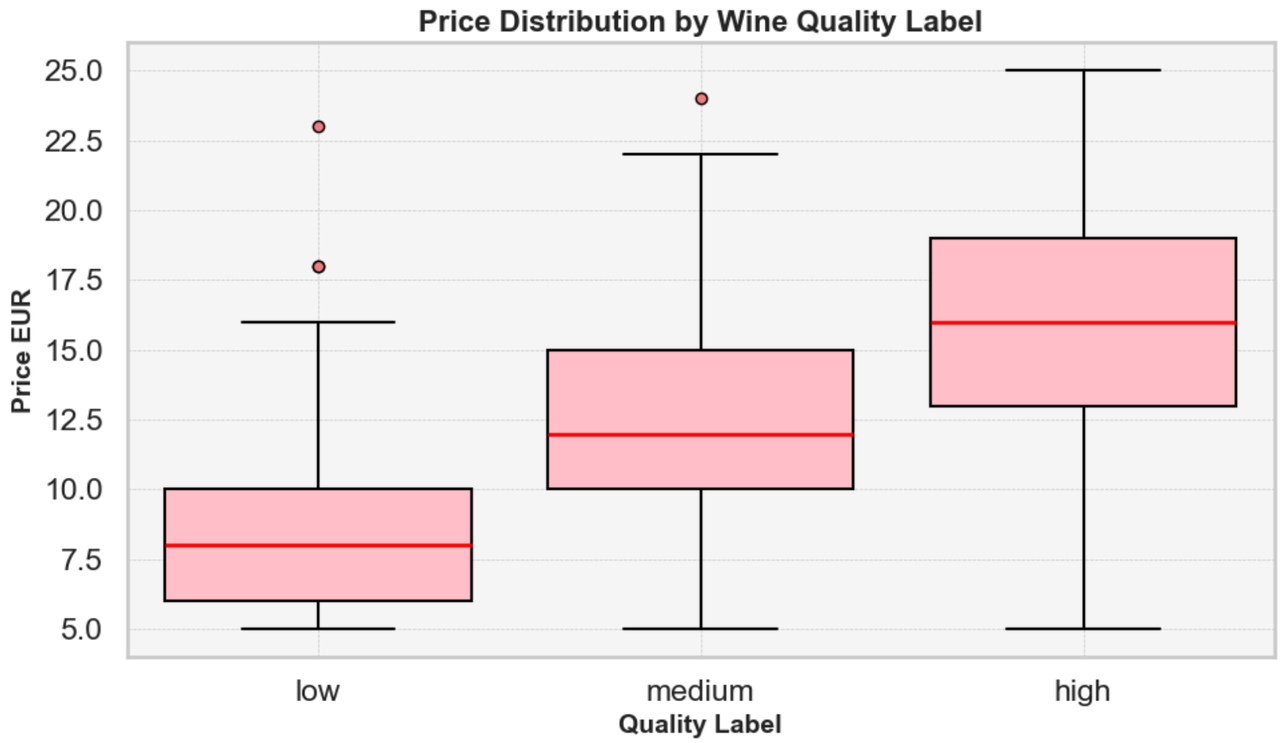
Additional data set describtion

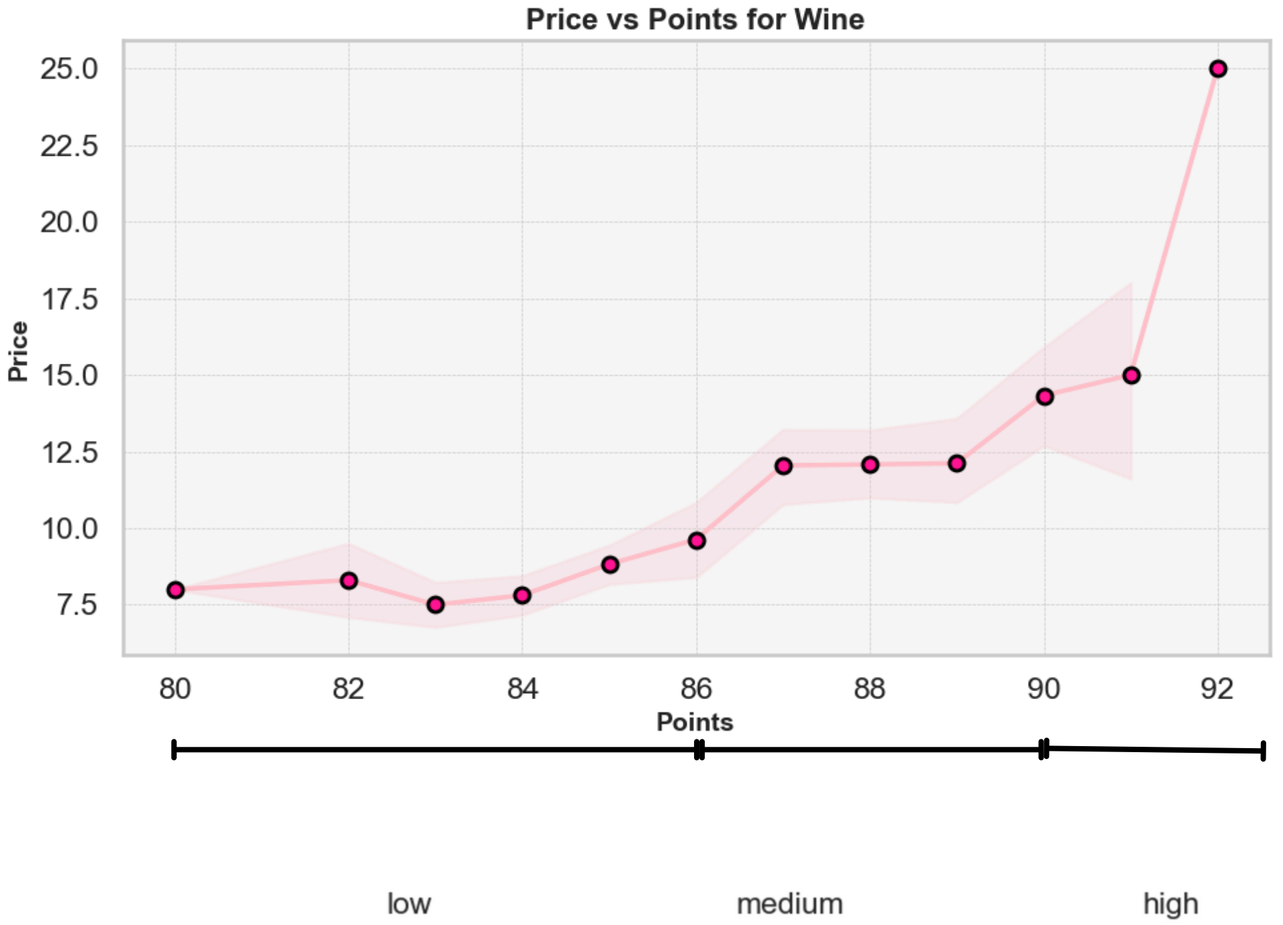
Provided additional data set :

- 150930 wines samples from all around the world,
- 10 features (points (80-100), country, price, province, winery, region (1, 2), description, designation, variety)
- no information about wine kind (red or white)
- the wine rating scale (points) has a different scale than in the ‘original’ data set

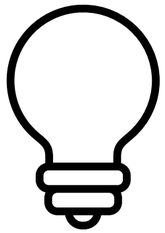
- **Main focus:** wines from **Vinho Verde** Province (396 wines samples)
- wines have been labeled as ‘low’ (<86 points), ‘medium’ ([86; 90 points]), ‘high’(>90)

Pricing in Vinho Verde region





Pricing prompts



Here are some conclusions after analysing the additional data set:

1. **Price Increases with Quality:** There is a clear positive correlation between wine quality and price. Wines with **higher quality labels** (medium and high) **have significantly higher prices** compared to those labeled as low quality.
2. **Price Range Variation**: Higher quality wines not only have higher median prices but also exhibit a broader price range. This suggests that **higher-quality wines are more varied in price**, possibly due to differences in factors like production methods or brand reputation.
3. **Outliers**: In the low and medium-quality categories, there are a few outliers with prices much higher than the typical range, indicating that some wines in these categories may be priced at a premium despite their lower quality rating.
4. **Price Consistency**: Low-quality wines have a narrower price range, indicating more consistent pricing. This could be due to less variability in production costs or consumer perception of value in this category.

Overall, higher quality wines are generally more expensive and have greater price variability.

THANK YOU