WINE QUALTITY PREDICTION

UMBC DATA SCIENCE MASTER'S
CAPSTONE, SPRING 2024

BY PRIYANK SAI PANNEM

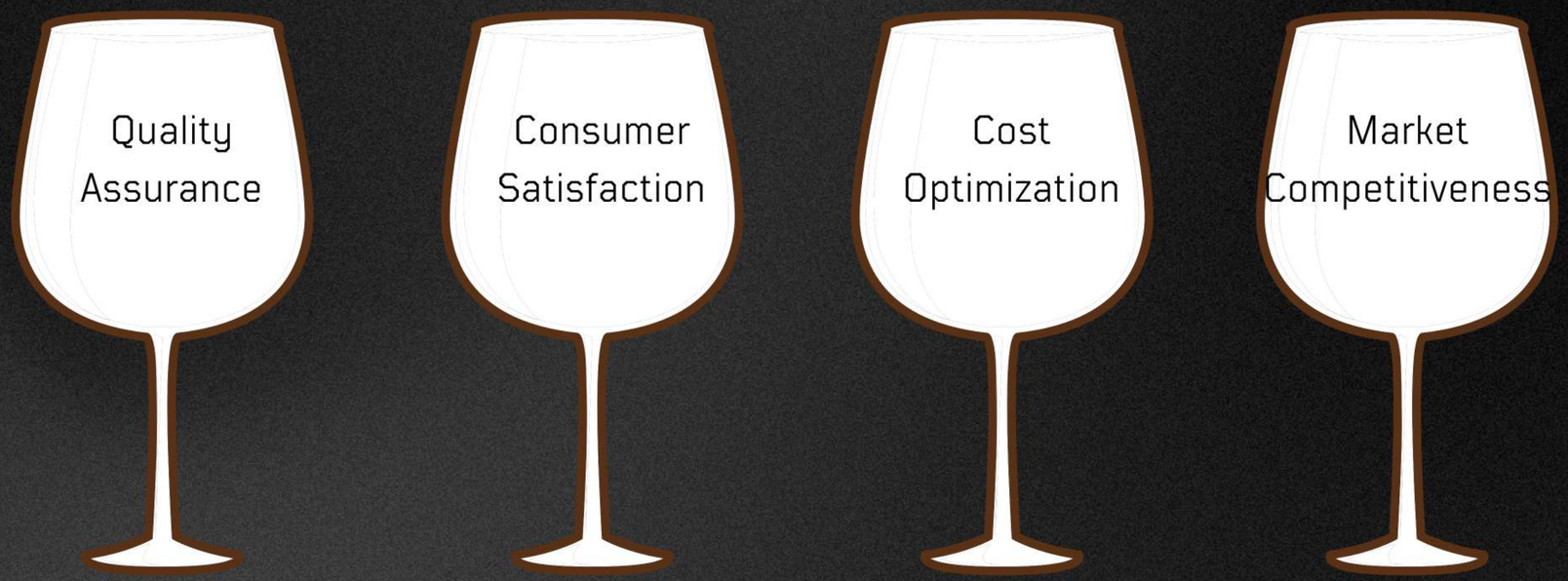
CONTENTS

INTRODUCTION

In the field of viticulture and oenology, wine quality is a critical factor in determining market value and customer happiness. Winemakers are continuously looking for ways to improve and preserve the quality of their goods. To this purpose, predictive modeling may be used to forecast wine quality based on physicochemical and sensory characteristics. The ability to accurately assess wine quality is crucial for both producers and consumers, as it influences manufacturing, pricing, and marketing strategies. This journey utilizes a dataset containing several physicochemical attributes of wine, such as acidity, sugar content, alcohol level, and others, which are believed to be indicative of its overall quality where various statistical and machine learning techniques to establish a predictive model.

OBJECTIVE

- The objective is to build a machine learning model capable of accurately predicting the quality of wine based on its attributes. This model can assist winemakers in assessing the quality of their products and making informed decisions regarding production and refinement processes.
- The other goals that are achieved upon implementing predictive modeling are



DATA OVERVIEW

Data Set Discription

The data contains 12 wine features or ingredients based upon which the quality or the types of wine can be predicted.

Data Size and Shape

The size of the data is 2.43MB with 32486 rows, 14 columns

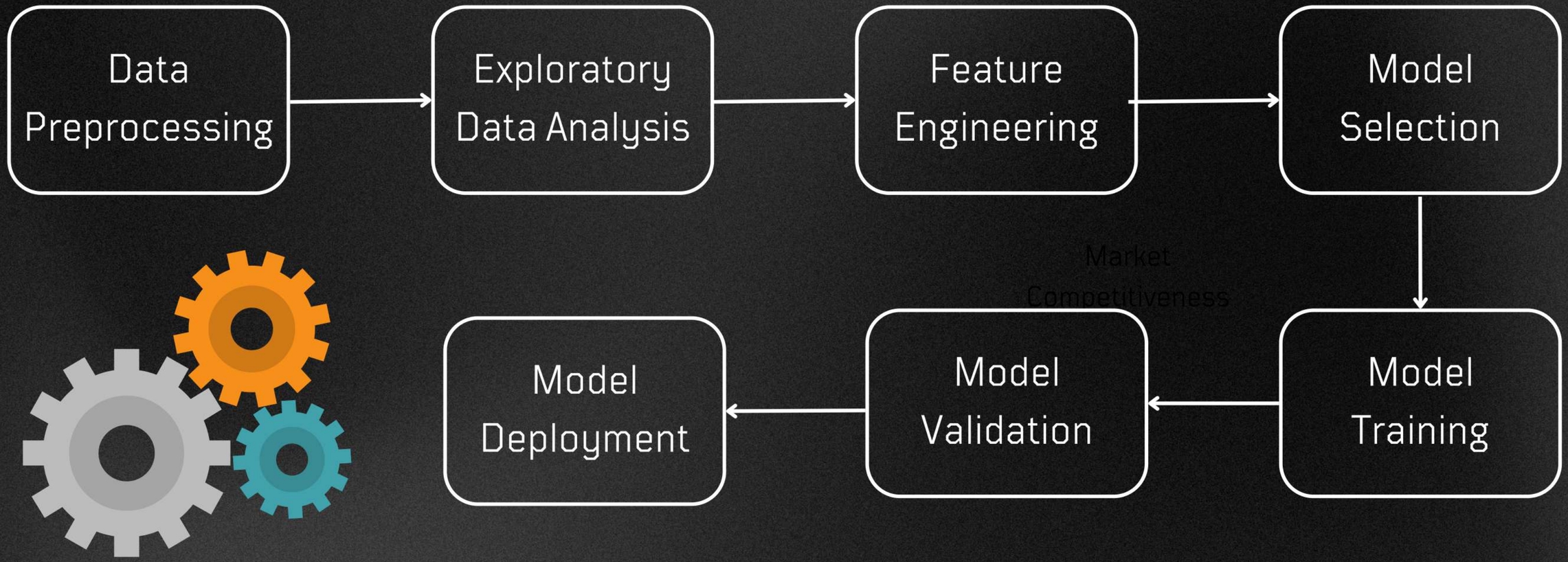
Representation of Each row in data set

Each row in the dataset represents various chemical properties of wine sample such as chlorides, pH, density, etc.

Target Variable

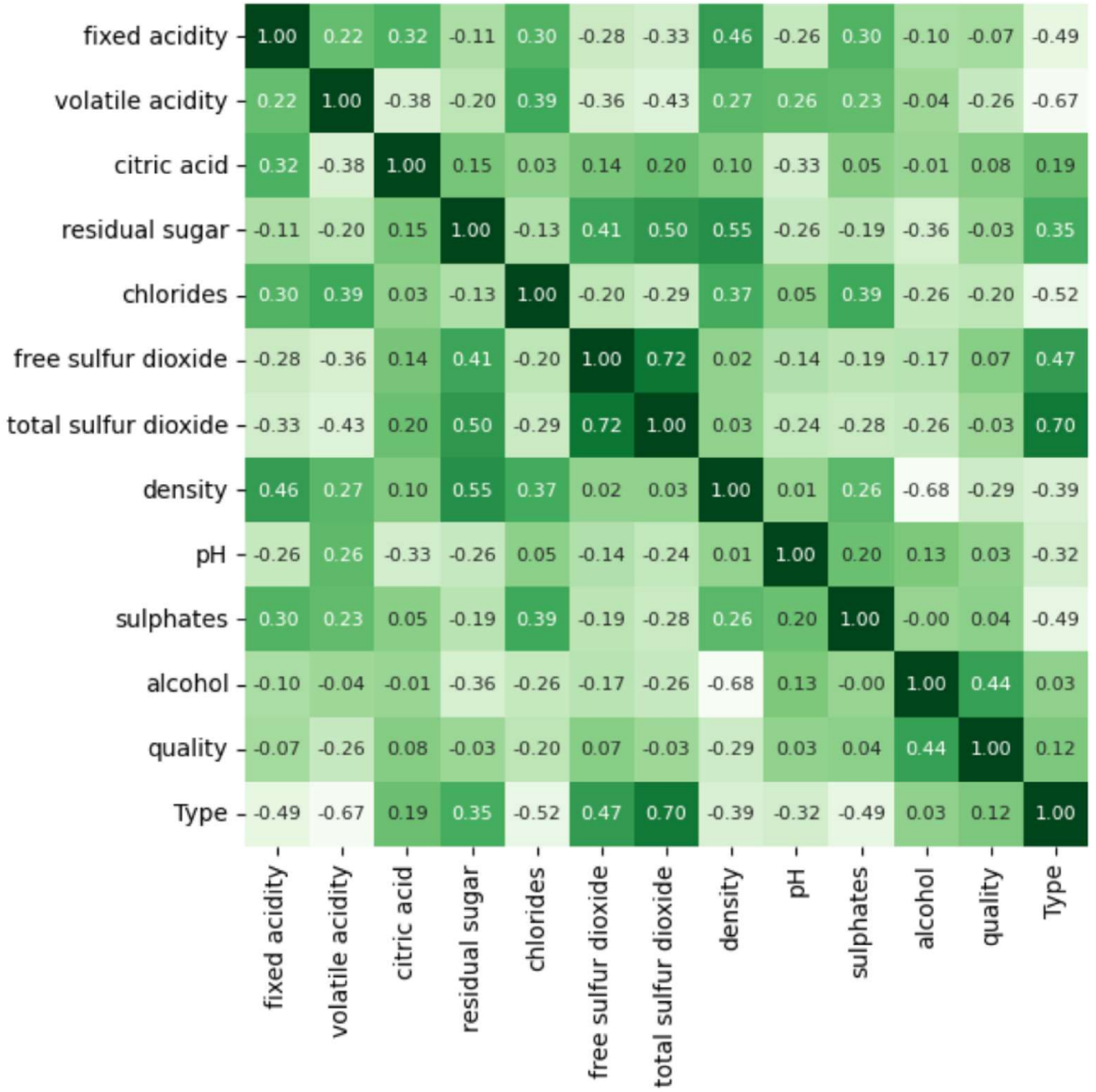
Quality of wine

METHODOLOGY

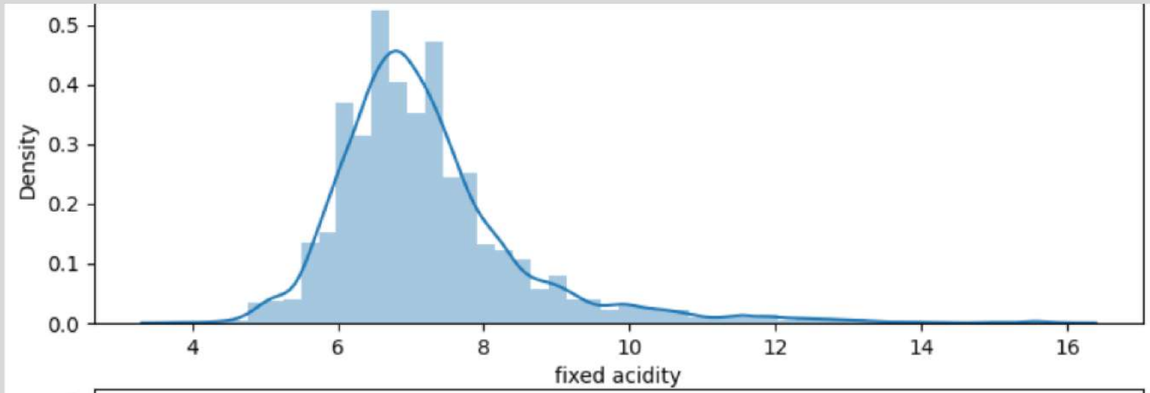


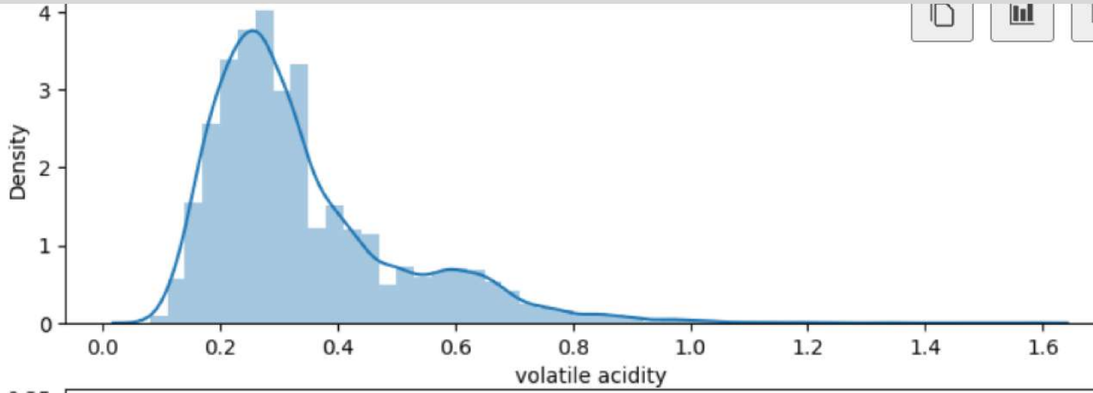
CORRELATION MATRIX

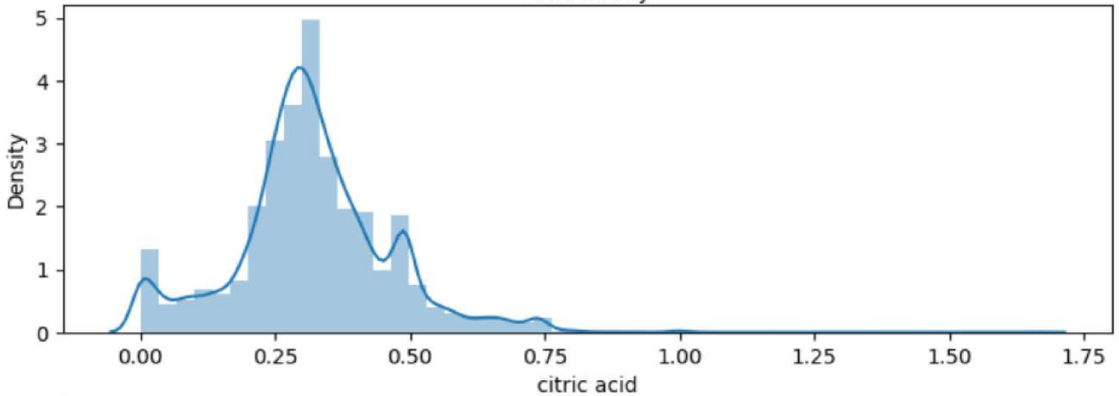
This shows there is no sign of multicollinearity, with this we can move forward without eliminating any columns.

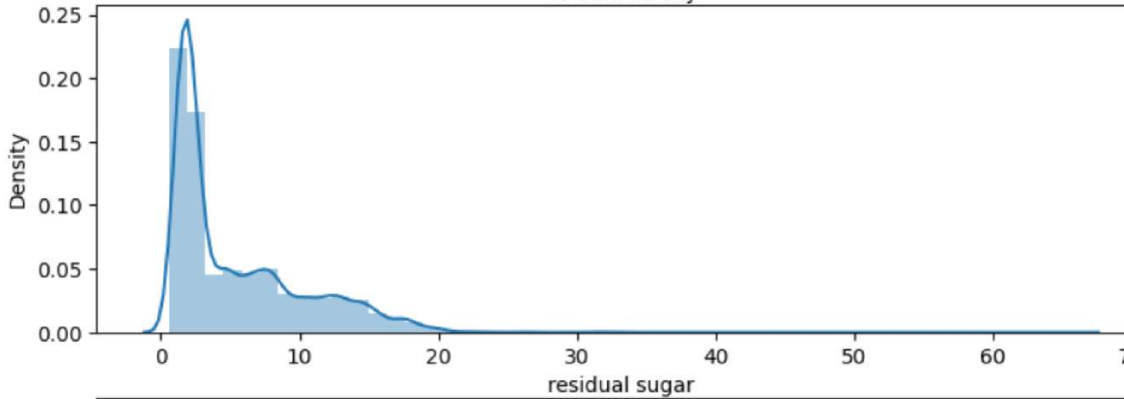


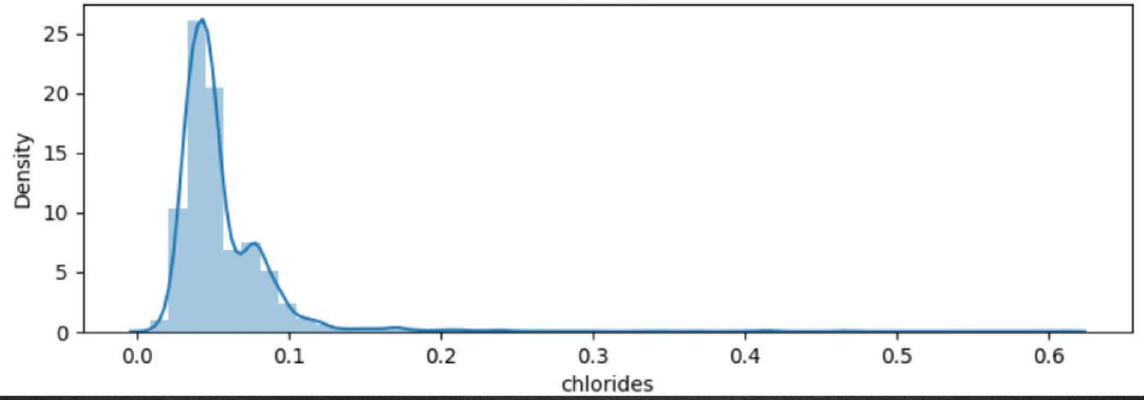
DISTRIBUTION PLOTS

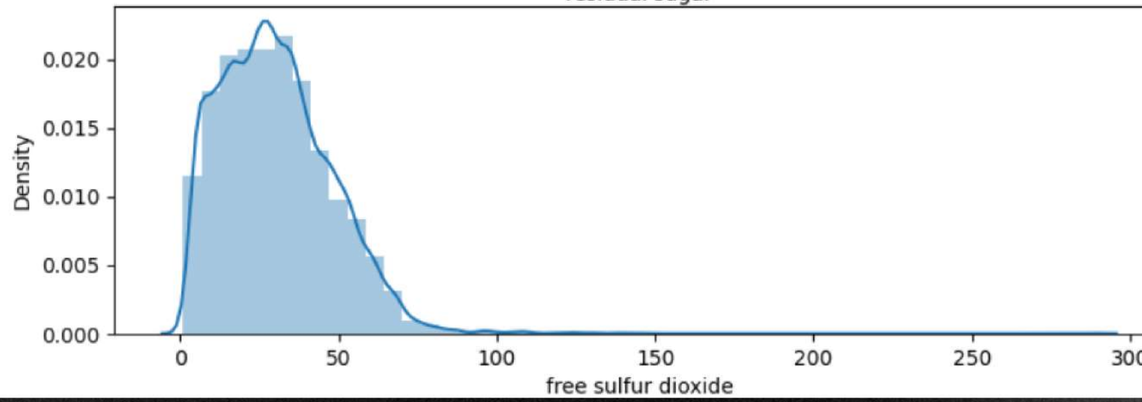






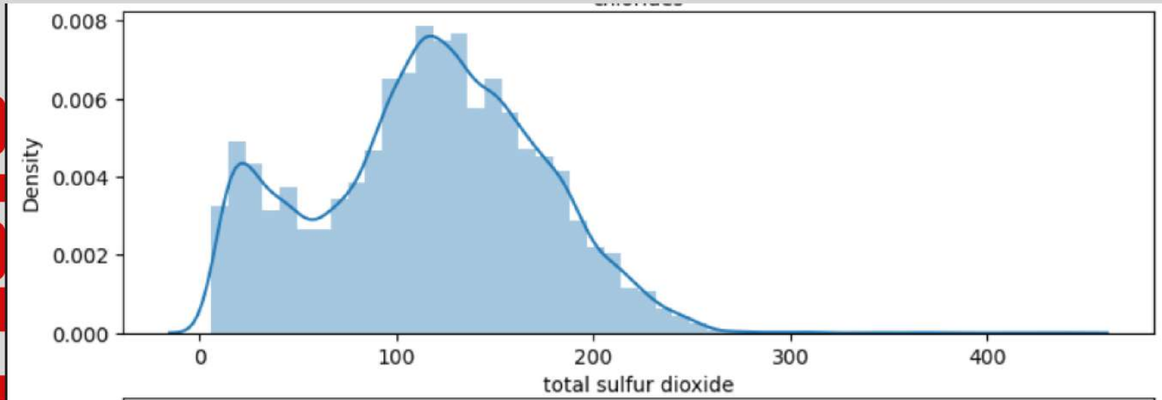


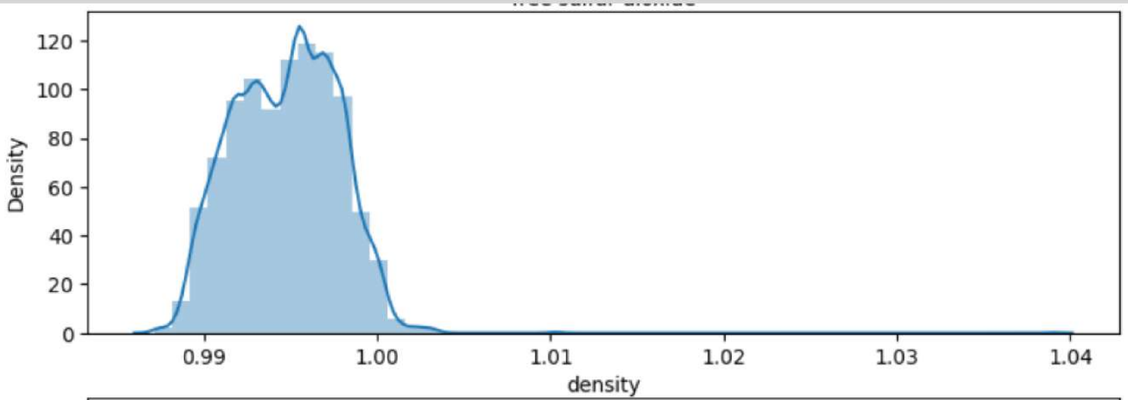


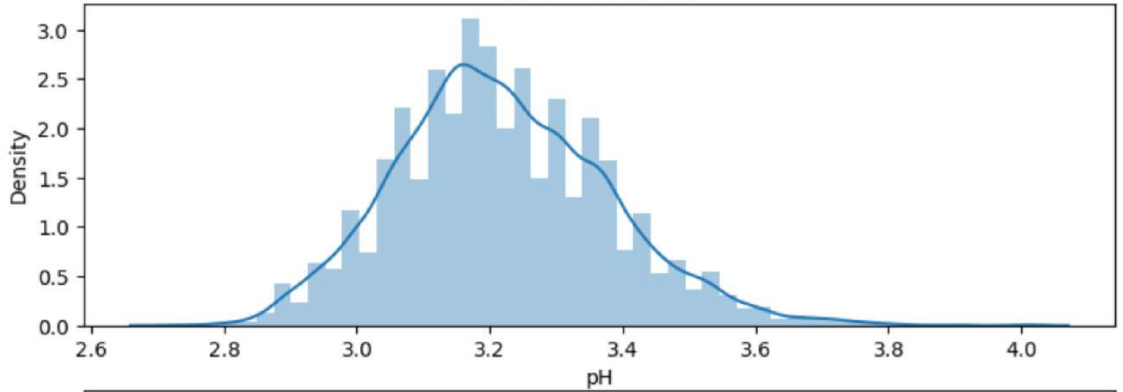


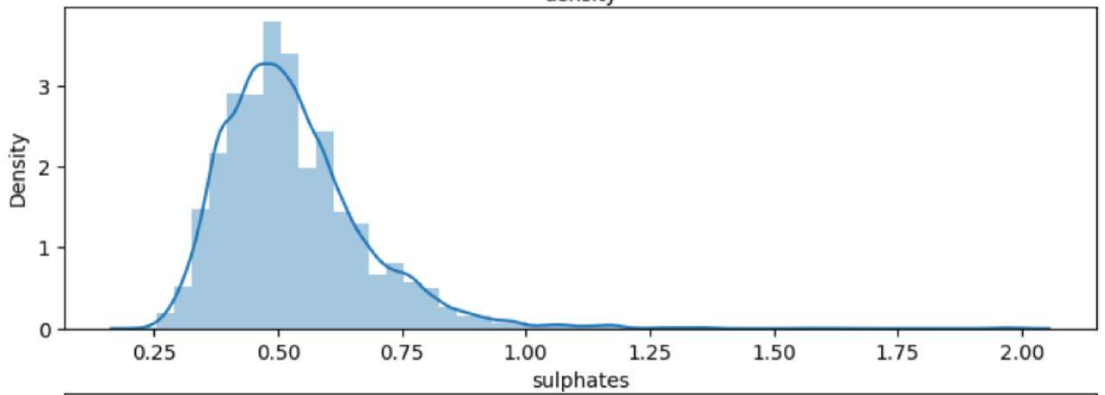
Most wines in the dataset exhibit fixed acidity between 6-8, volatile acidity around 0.3, two common citric acid levels near 0 and 0.5, low residual sugar with some outliers,

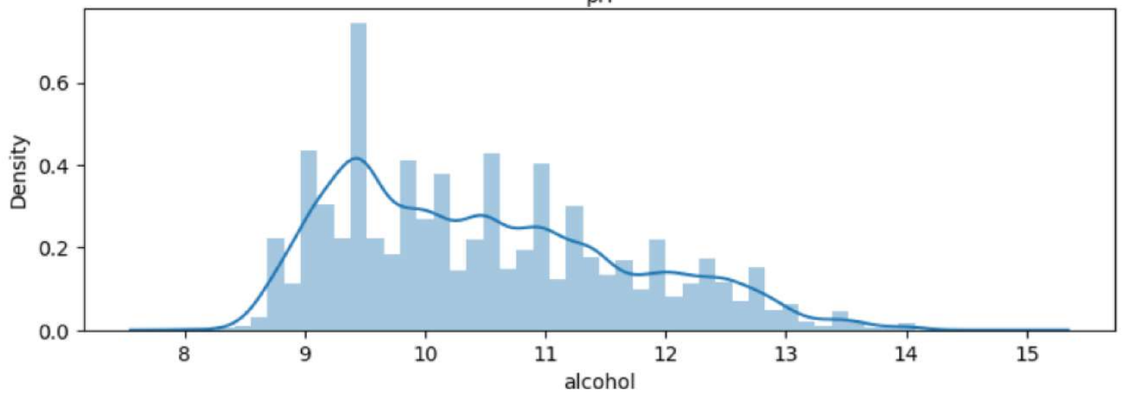
DISTRIBUTION PLOTS

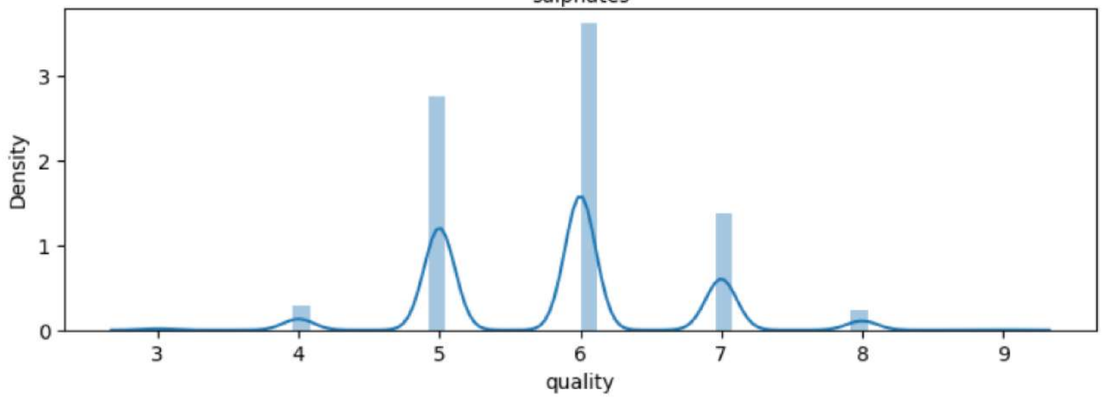






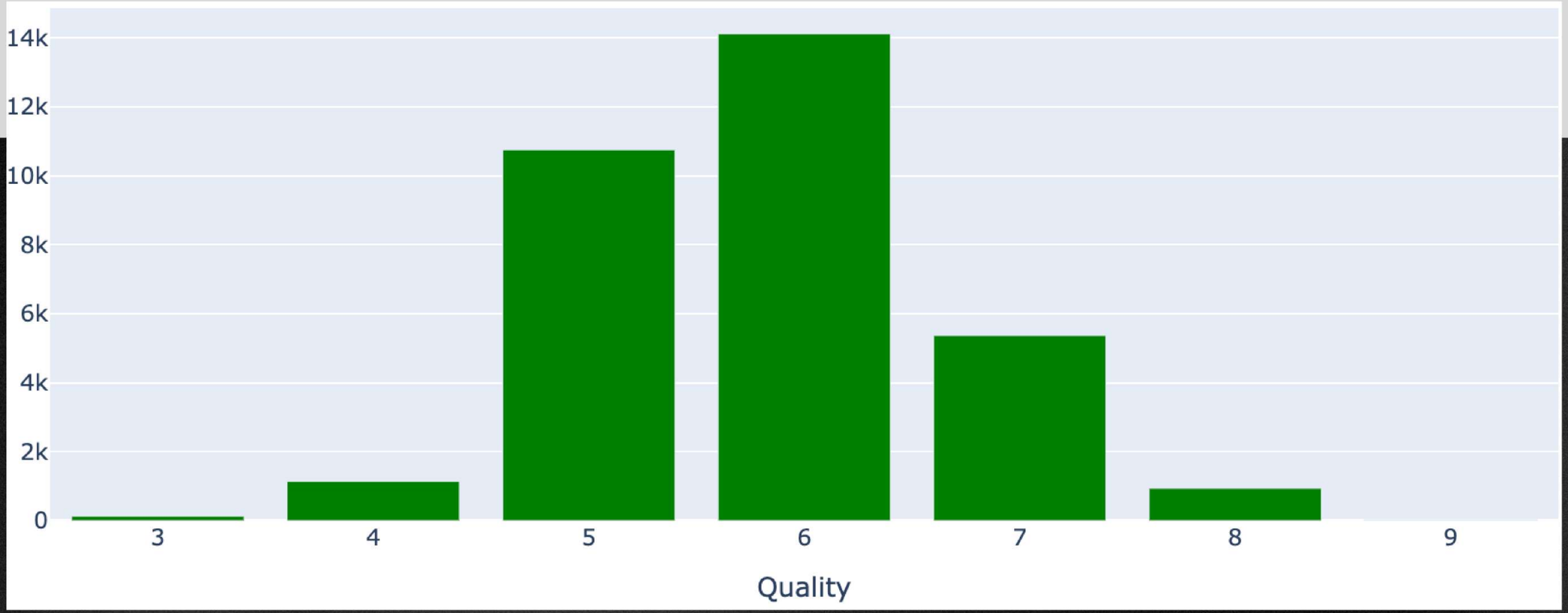






Most wines exhibit total sulfur dioxide around 150 mg/L, density between 0.99-1.0, pH around 3.2, sulphates around 0.5, alcohol content peaking at 9.5% with a wide range, and quality ratings commonly at 5, 6, and 7, indicating varied winemaking practices

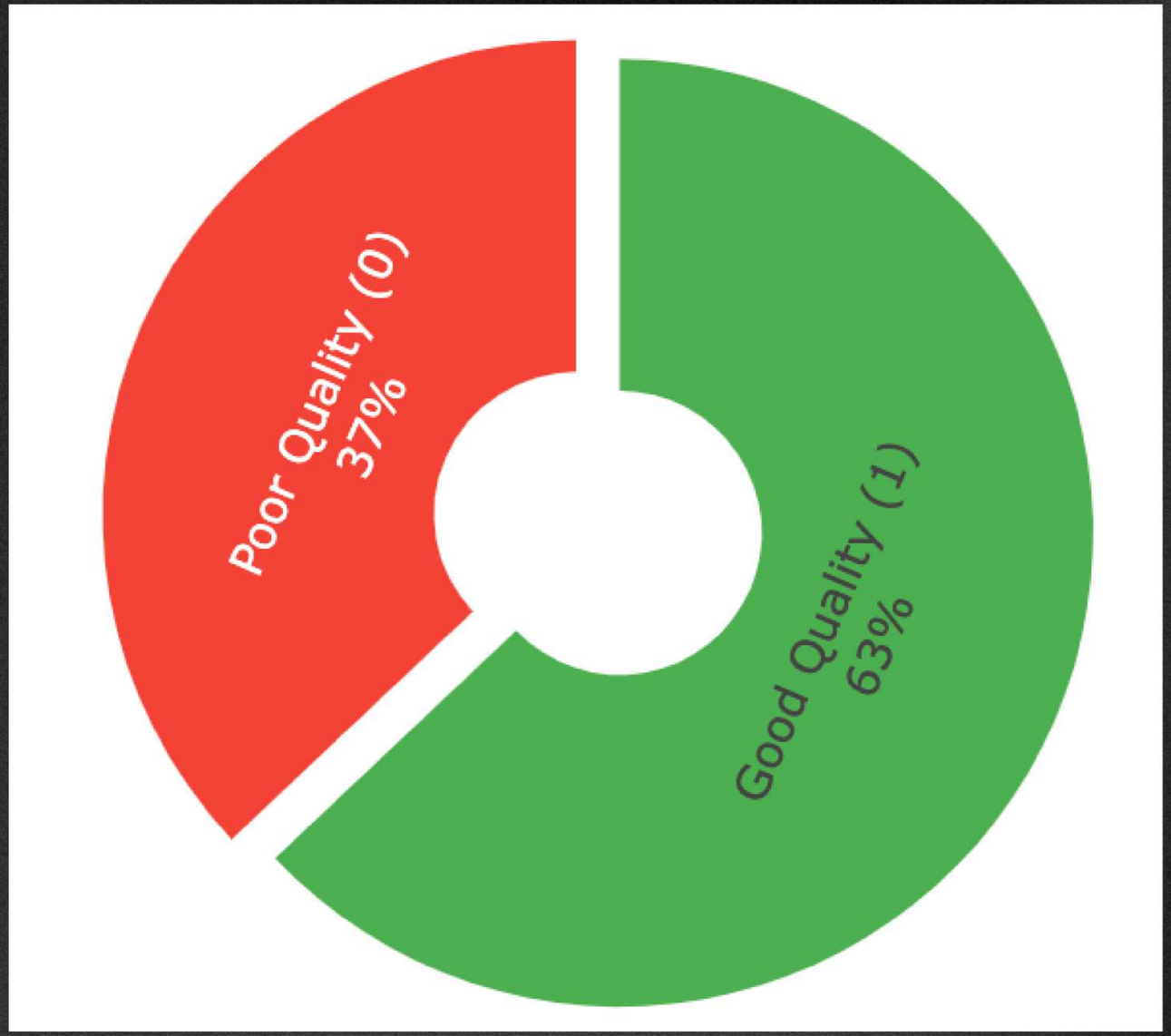
TARGET VARIABLE DISTRIBUTION



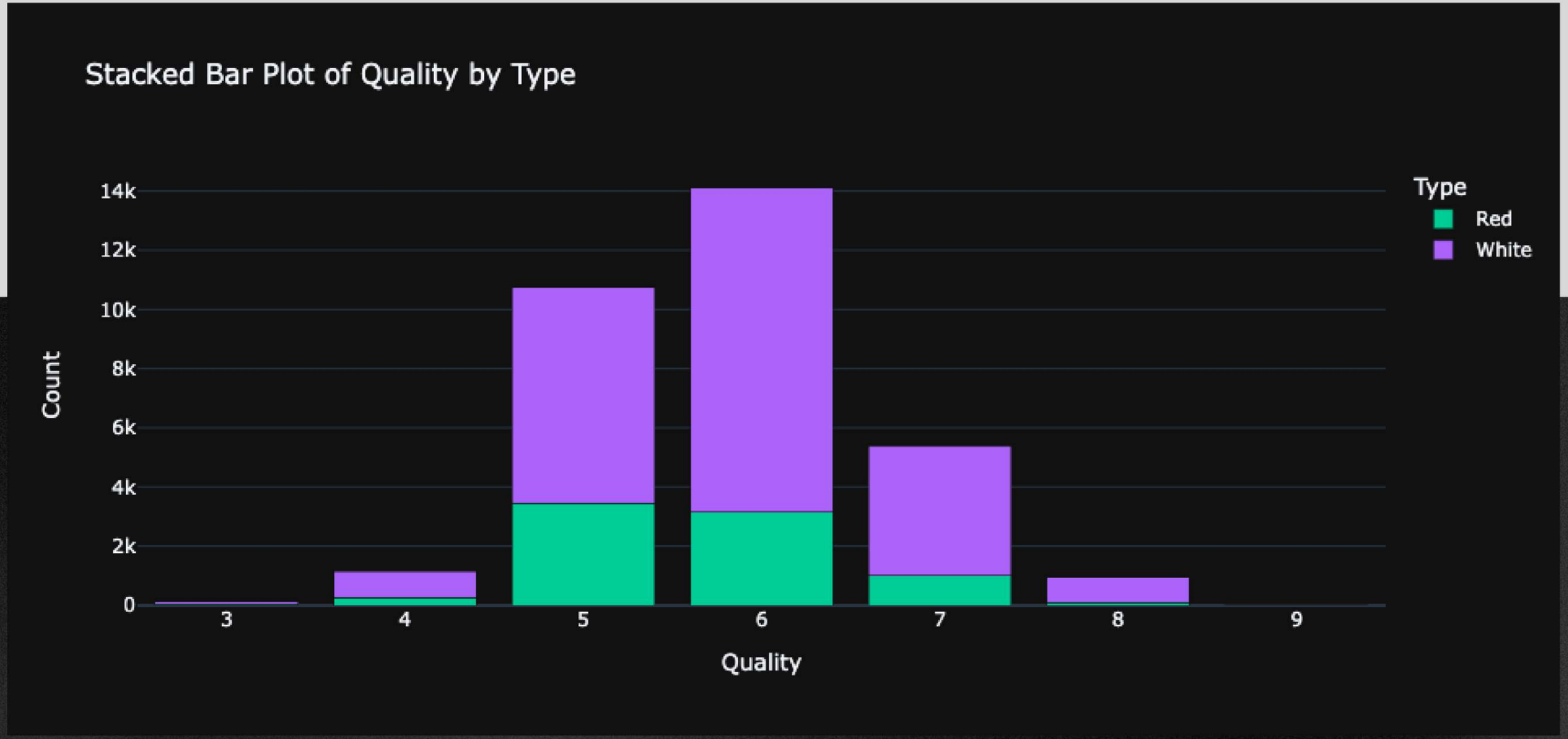
Imbalanced data can pose challenges for machine learning models, especially for classification tasks, as they might become biased towards the majority class.

PERCENTAGE OF BEST QUALITY CATEGORIES

The dataset has 63% "good quality" and 37% "poor quality" wines, with the slight imbalance likely having minimal impact on analysis given the manageability by machine learning algorithms.

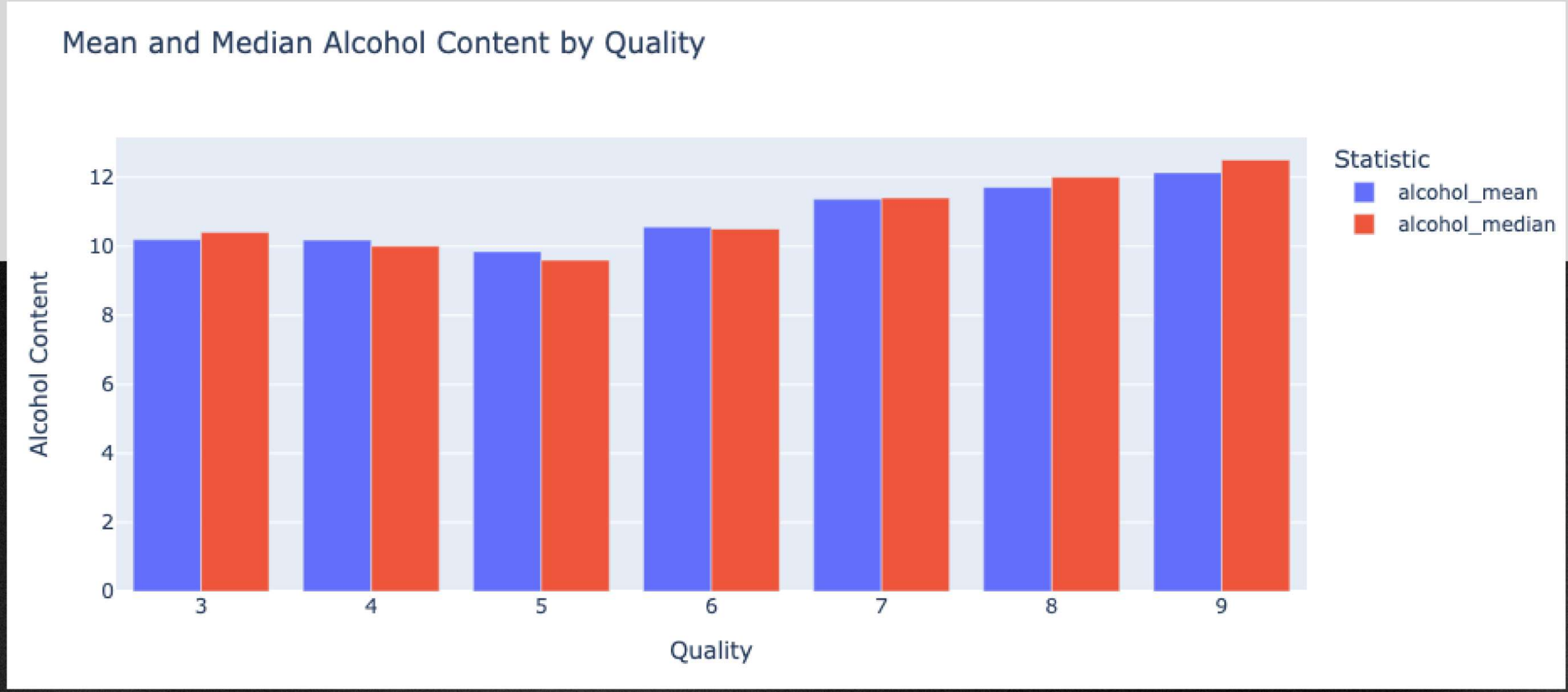


BAR PLOT OF QUALITY BY TYPE



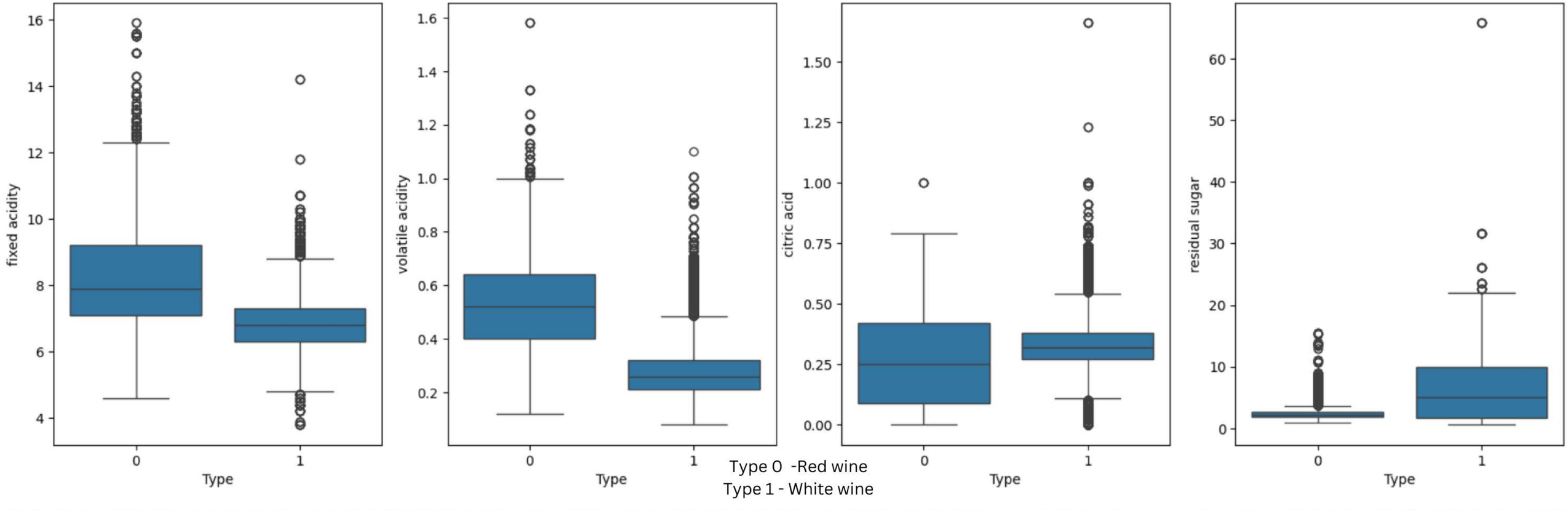
White wines, which dominate the dataset, are mostly rated 5 and 6 but also appear frequently at lower quality ratings, whereas fewer red wines tend to achieve higher ratings, particularly at 7 and 8, with overall quality ratings of 5, 6, and 7 covering most wines.

MEAN & MEDIAN ALCOHOL CONTENT



The increasing trend in alcohol content from lower to higher quality ratings indicates that alcohol content could be a significant factor in determining the quality of wine.

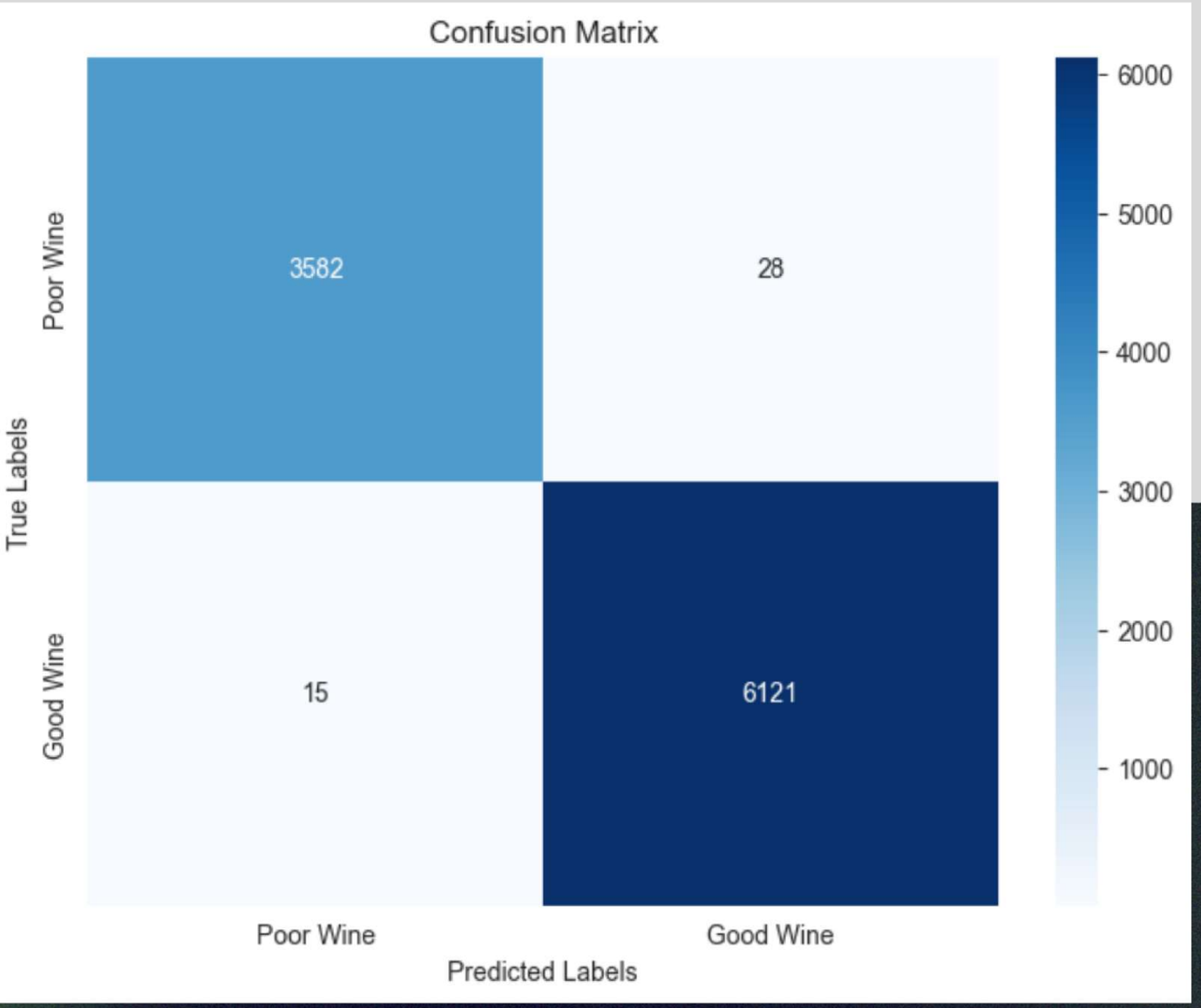
BOX PLOTS



White wine typically has higher volatile acidity, residual sugar, chlorides, content than red wine, which generally has higher fixed acidity and citric acid, while features like density and quality do not differ significantly, with white wine exhibiting greater variability in many features.

MODEL SELECTION & TRAINING

- Split the dataset into training and testing set with 70% and 30%.
- Created a pipeline with **StandardScaler lined up with different ml algorithms**
- **RandomForest classifier outperformed over other classification models**
- Used **GridSearchCV** with a 5-fold cross-validation (**KFold**) to find the optimal hyper parameters, ensuring robustness and avoiding overfitting.
- Model performance is evaluated with f1 score.



PERFORMANCE METRICS

- F1 Score : 100%
- Precision : 100%
- Recall : 100%
- Random Forest achieved perfect performance, XGBoost performed slightly less accurately but still well, while Logistic Regression showed the lowest accuracy among the three classifiers.

STREAMLIT WEB APP

Wine Quality Classifier Web App

Fixed Acidity: 5

Volatile Acidity: 6

Citric Acid: 3

Residual Sugar: 7

Chlorides: 4

Free Sulfur Dioxide: 8.9

Total Sulfur Dioxide: 34

Density: 8

pH: 3.2

Sulphates: 5.6

Alcohol Concentration: 6

Choose wine type (Red or White): White

Predict

The quality of wine is Poor

CONCLUSION

The random forest model demonstrated promising results in predicting wine quality, achieving a notable accuracy on the test set. It effectively differentiated between poor and good quality wines, suggesting that specific features significantly influence wine quality. This model could serve as a useful tool for winemakers and industry professionals looking to assess and enhance their products.

THANK YOU!