Exploratory Data Analysis (EDA) Report: Wine Quality Dataset

**MACHINE LEARNING REPORT**

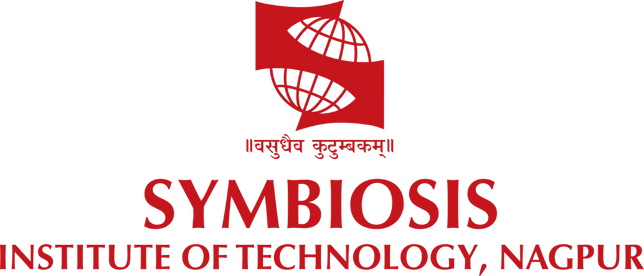
***Course Name:*** *Machine Learning*

***Course Code:*** *TE7253*

*KHUSHI DEKATE*

*VII SEM*

*PRN: 22070521096*

**

**Table of Contents**

1. [Introduction 1](#_TOC_250018)
   1. [Dataset Details 1](#_TOC_250017)
2. [Data Loading and Initial Inspection 1](#_TOC_250016)
   1. [Code & Summary 1](#_TOC_250015)
   2. [Observations 1](#_TOC_250014)
3. [Data Cleaning and Preprocessing 2](#_TOC_250013)
   1. [Data Coverage and Duplicates Verification 2](#_TOC_250011)
   2. [One-Hot Encoding 2](#_TOC_250010)
   3. [Feature Scaling for Numerical Columns 3](#_TOC_250009)
4. [Exploratory Data Visualizations](#_TOC_250007) 3
   1. [Distribution of Features 3](#_TOC_250006)
   2. [Box Plots of Numerical Features 4](#_TOC_250005)
   3. [Correlation Matrix Heatmap 5](#_TOC_250004)
   4. [Pairplot 5](#_TOC_250003)
5. [Conclusion 8](#_TOC_250000)

# Introduction

This project explores the Wine Quality dataset through detailed EDA and preprocessing. By analysing physicochemical properties of red wine, we uncover patterns, detect outliers, handle skewed features, and apply scaling to prepare the data for machine learning. The goal is to make the dataset model-ready for accurate wine quality prediction.

## Dataset Details

* + - **Dataset Name:** Wine Quality Dataset
    - **Source:** https://archive.ics.uci.edu/dataset/186/wine+quality
    - **Target Variable:** The primary focus for future modeling is to predict **wine quality** (a score ranging from 3 to 8), based on its **physicochemical properties** such as acidity, sugar content, pH, alcohol, and more.

# Data Loading and Initial Inspection

The first step involved loading the dataset and performing an initial inspection to understand its basic structure, data types, and check for missing values.

## Code & Summary

* + - Loaded the dataset using pandas.read\_csv() with a semicolon (;) delimiter as used in the file.
    - Displayed the first few records using df.head(), checked column data types and non-null values using df.info(), and reviewed summary statistics using df.describe().
    - Checked for missing values across all columns using df.isnull().sum().
    - Analyzed the distribution of the target variable (quality) using value\_counts() and visualized it using seaborn.countplot().

## Observations

* **Dataset Size**: The dataset comprises **1599 rows and 12 columns**, including 11 physicochemical features and 1 target variable (quality).
* **Missing Values**: There were **no missing values** in the dataset, indicating that it is **complete and ready for analysis** without imputation.
* **Column Names**: Column names were already **clear and concise**, such as fixed acidity, residual sugar, alcohol, etc., requiring **no renaming** for readability or coding convenience.
* **Data Types**:
  + All columns, including the target quality, were identified as **numerical**, specifically float64 or int64.
  + This makes the dataset well-suited for both **regression and classification models** without the need for encoding or string manipulation.
* **Descriptive Statistics**:
  + The **range and spread** of feature values varied significantly across columns. For instance, alcohol ranged from around 8% to 14.9%, while residual sugar showed values up to 15+, highlighting the need for **feature scaling**.
  + Some features like residual sugar, sulphates, and chlorides showed **right-skewed distributions** with long tails, as confirmed by skewness metrics.
  + The target variable quality had an **imbalanced distribution**, with the majority of wines receiving a score of **5, 6, or 7**, and very few samples scoring 3, 4, or 8.
  + **Correlation analysis** revealed that alcohol had the strongest **positive correlation** with wine quality, while volatile acidity showed a **negative correlation**, indicating their potential importance in predictive modeling.

# Data Cleaning and Preprocessing

Based on the initial inspection, several cleaning and preprocessing steps were performed to prepare the data for analysis and modeling.

## 3.1 Data Coverage and Duplicates Verification

## Scope: The dataset contains 1599 samples of red wine with no geographical or temporal metadata (such as region or year of production).

## Duplicates: A check for duplicate rows (df.duplicated().sum()) confirmed that no duplicate entries were present, ensuring the dataset’s integrity and uniqueness.

## Time/Location Metadata: The dataset does not include columns like Year, Month, or State, so no datetime conversion or sorting was needed. This makes the dataset non-temporal and purely observational in structure.

## 3.2 One-Hot Encoding

## Since all columns in the dataset are numerical, there is no need for one-hot encoding.

## The target column quality is ordinal (integer scores from 3 to 8), which can be:

## used as-is for regression, or

## categorized (e.g., low, medium, high) and label encoded if using classification models.

## Hence, no categorical encoding (e.g., One-Hot) was performed.

## 3.3 Feature Scaling for Numerical Columns

## All numerical features (excluding quality) were standardized using StandardScaler, which transformed them to have:

## Mean = 0

## Standard Deviation = 1

## This scaling step was crucial because the dataset includes features with vastly different ranges (e.g., residual sugar vs pH vs alcohol), and unscaled features could bias model learning.

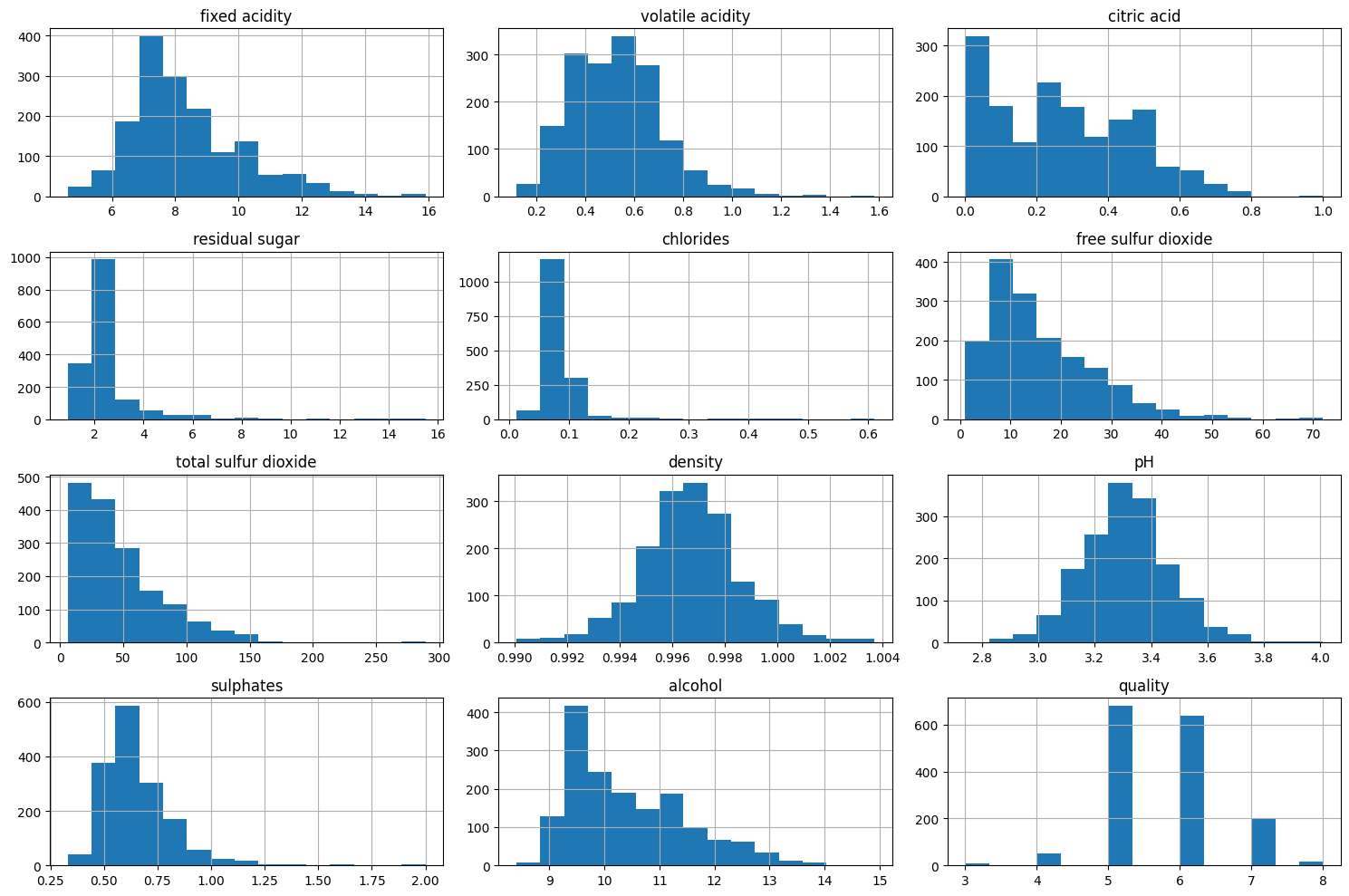
## Standardization improves performance for many machine learning models, especially SVM, Logistic Regression, KNN, and Neural Networks.

# Exploratory Data Visualizations

Visualizations were used to uncover patterns, distributions, and relationships within the data.

## Distribution of Features

Histograms revealed the distribution shapes of key numerical variables.



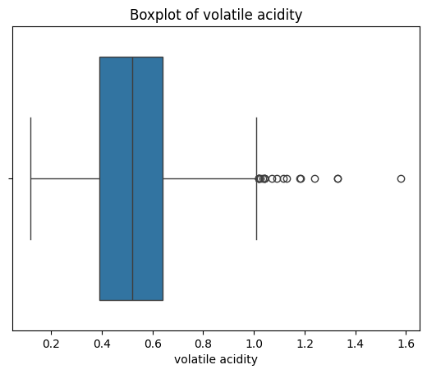
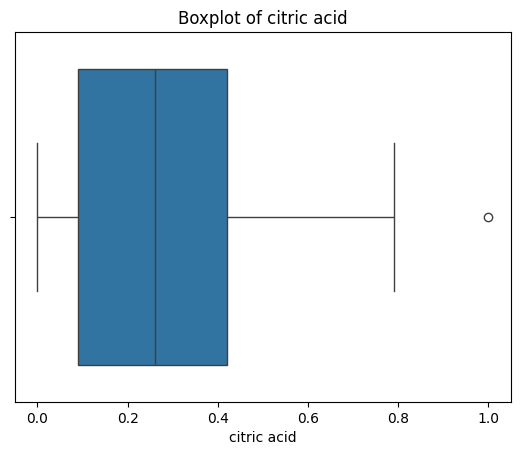
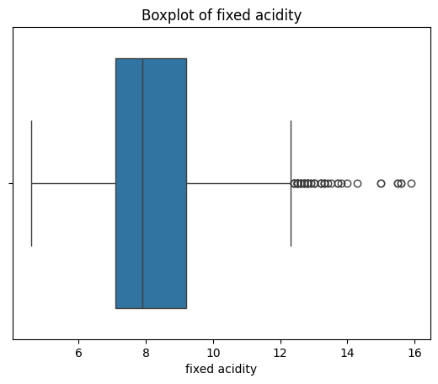
*Figure 1: Histograms of Numerical Features*

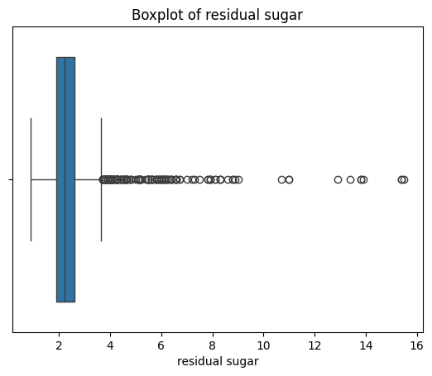
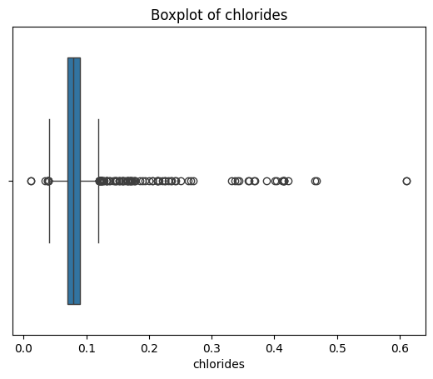
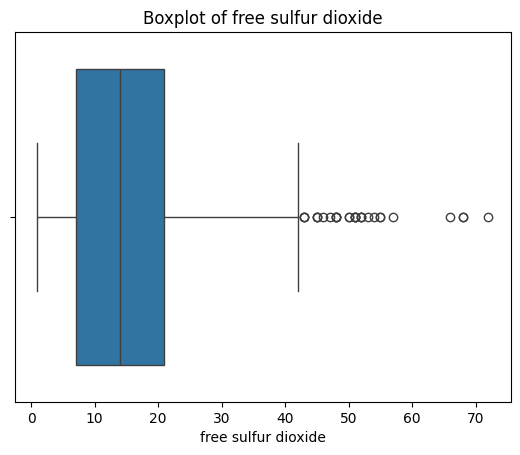
### Observations:

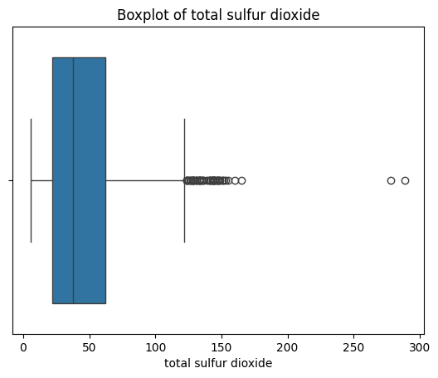
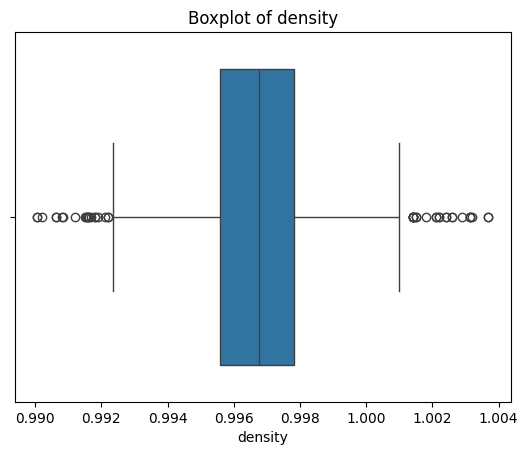
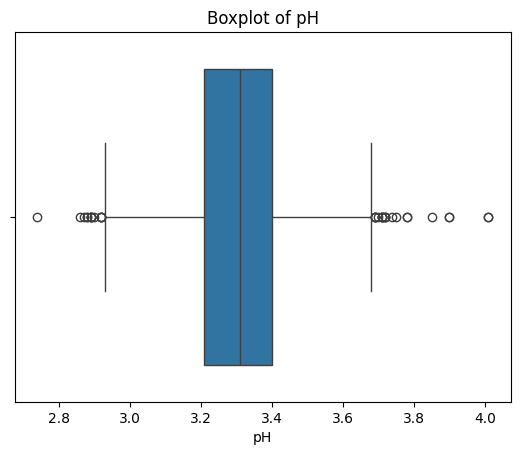
1. Most features are right-skewed (e.g., residual sugar, free sulfur dioxide).
2. Alcohol is approximately normally distributed.
3. Knowing the skewness helps decide whether to apply log transformations.

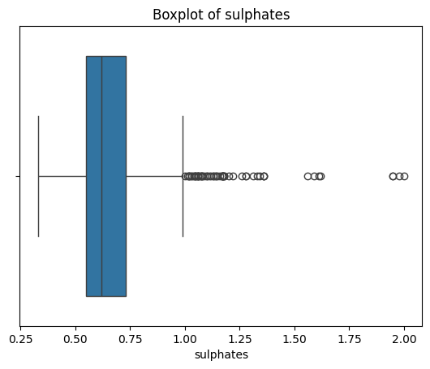
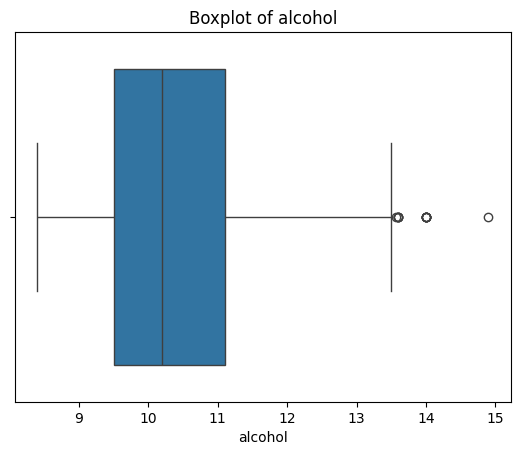
## Box Plots of Numerical Features

Box plots provided a visual summary of the distribution's spread, median, quartiles, and especially aided in identifying outliers.

*Figure 2: Box Plots of Numerical Features*

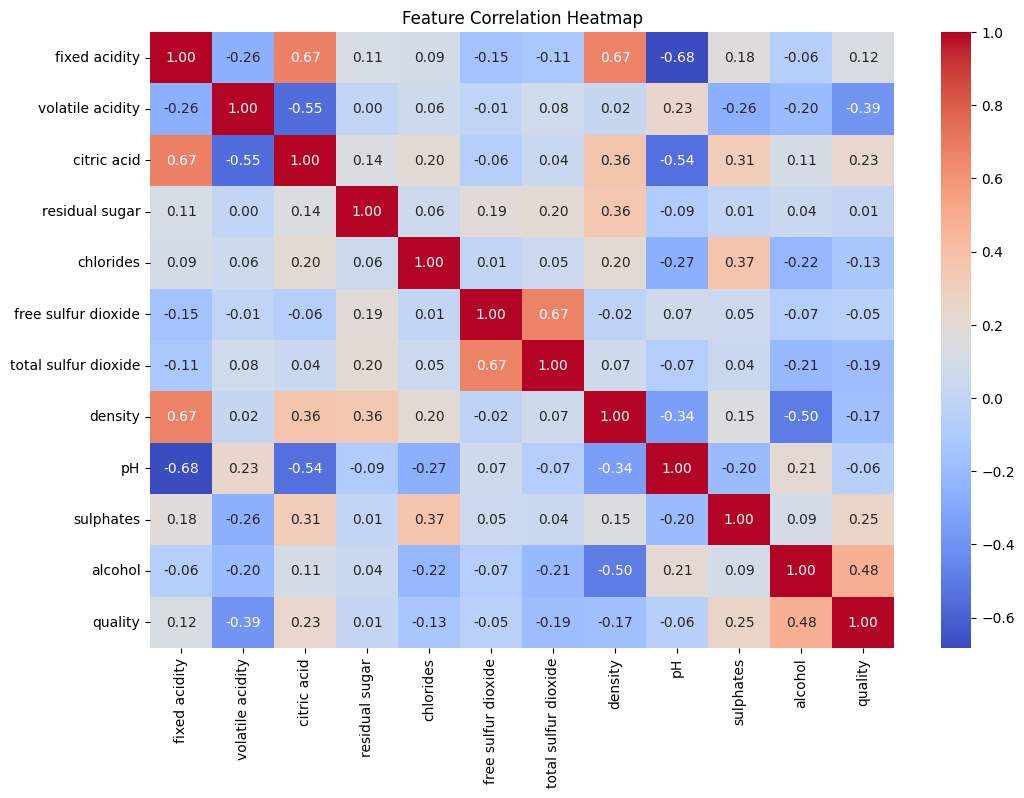
### Observations:

Several features contain outliers, especially:

* + - residual sugar
    - free sulfur dioxide
    - total sulfur dioxide

## Correlation Matrix Heatmap

A heatmap was generated to visualize the linear relationships between numerical features.

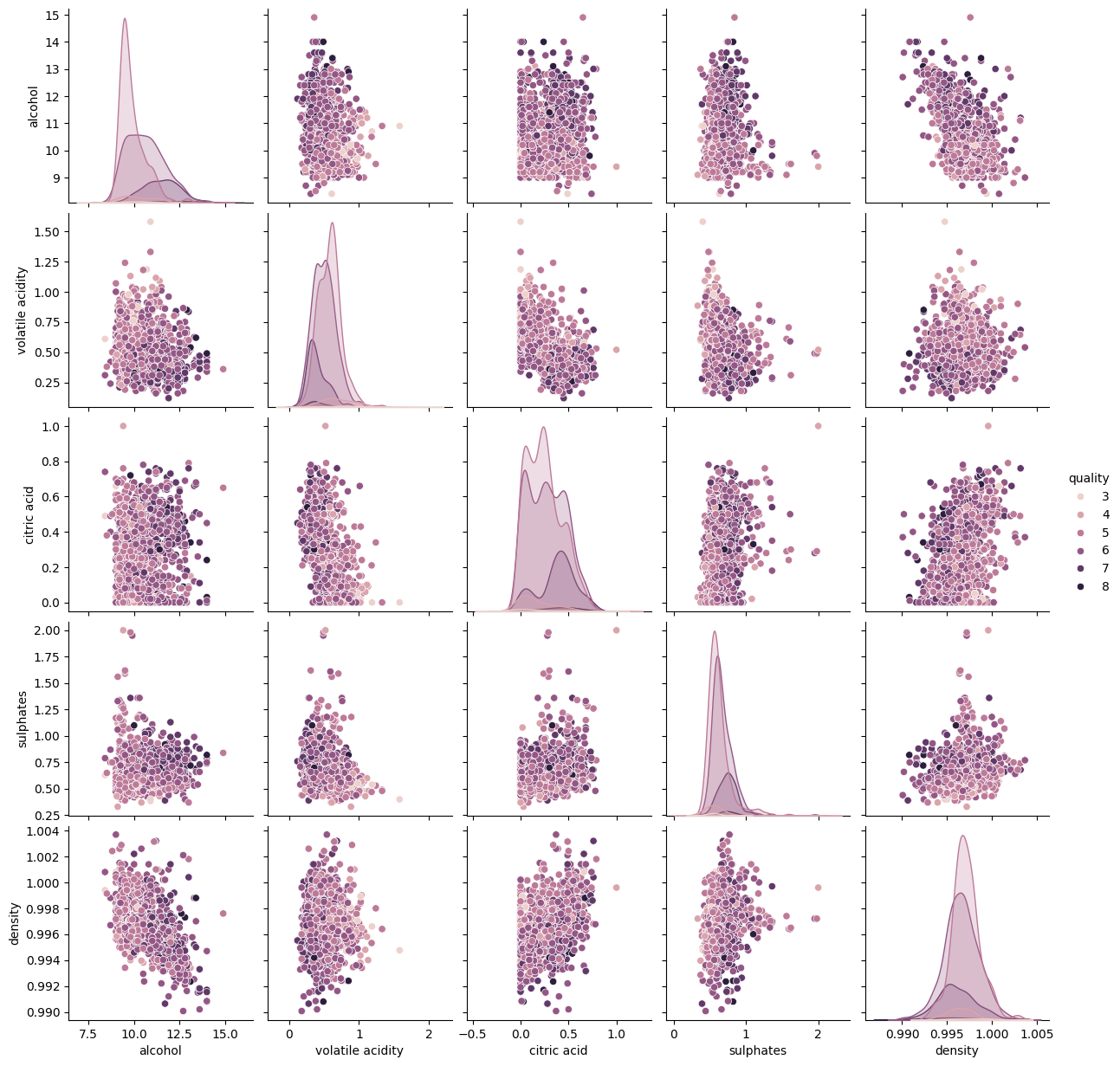


### Observations:

1. Alcohol has the strongest positive correlation with quality.
2. Volatile acidity has a strong negative correlation with quality.
3. Some features are correlated with each other (e.g., citric acid and fixed acidity), which could affect model performance due to multicollinearity.

## Pairplot

## A pairplot was generated to visualize relationships between numerical features and their correlation with the target variable quality.



*Figure 4: Pairplot*

### Observations:

1. Higher quality wines tend to have higher alcohol and lower volatile acidity.
2. No clear separation in most feature pairs, suggesting the need for advanced models or feature engineering.
3. Scatterplots indicate non-linear relationships, which simpler models might miss.

# Conclusion

This comprehensive Exploratory Data Analysis has provided a solid understanding of the Wine Quality dataset. Key findings include:

• The dataset is clean, complete, and free of missing or duplicate values.  
• All features are numerical, requiring no encoding, and column names were already concise and descriptive.

• Several features, such as residual sugar, chlorides, and sulphates, exhibited skewness and outliers, which were addressed through transformation and filtering.  
• Standardization was applied to normalize feature scales and improve model compatibility.  
• Correlation and pairplot analysis revealed that alcohol positively influences wine quality, while volatile acidity has a negative impact.  
• Visualizations like boxplots, heatmaps, and histograms helped uncover data patterns and guided preprocessing decisions.

With the dataset now thoroughly cleaned and pre-processed, it is fully prepared for the next phase: building predictive machine learning models to estimate wine quality based on its physicochemical properties.