Wine Quality Dataset Analysis (Red & White Wines) – Project Summary

Tools Used

- Software: Jupyter Notebook
- Python Packages:
    - `pandas` (data manipulation)
    - `numpy` (numerical operations)
    - `matplotlib`, `seaborn` (visualization)
    - `IPython.display` (notebook display)

Key Questions and Code-Based Answers

1. What are the summary statistics (mean, median, std, min, max) for each physicochemical property in the wine dataset?

- The notebook computes these statistics for all columns (e.g., fixed acidity, pH, alcohol, etc.) using `describe()` and `agg()`.
- Example: For red wine, alcohol has mean ≈ 10.43%, median = 10.2%, std ≈ 1.08, min = 8.4%, max = 14.9%.
- For white wine, alcohol mean ≈ 10.59%, median = 10.4%, std ≈ 1.22, min = 8.0%, max = 14.2%.

2. Which chemical properties show the greatest and least variation among the wines?

- The code sorts columns by standard deviation.
- Greatest variation:
    - `total_sulfur_dioxide` (std ≈ 56.77 for reds, 43.13 for whites)
    - `free_sulfur_dioxide` and `residual_sugar` are also highly variable, especially in whites.
- Least variation:
    - `density` (std ≈ 0.0019 for reds, 0.0029 for whites)
    - `chlorides` is also low-variation.

3. Are there notable differences in chemical properties between red and white wines?

- The notebook compares means and medians by wine type:

- Red wines: Higher fixed acidity (mean ≈ 8.31), volatile acidity (mean ≈ 0.53), and sulphates (mean ≈ 0.66).
- White wines: Higher residual sugar (mean ≈ 6.84), free sulfur dioxide (mean ≈ 34.89), and total sulfur dioxide (mean ≈ 137.19).
- pH: Slightly higher in reds (mean ≈ 3.31) than whites (mean ≈ 3.20).

4. What are the most common values (modes) for key properties like pH and alcohol?

- The code uses `mode()` for both types:
    - pH mode: 3.3 for reds, 3.1 for whites
    - Alcohol mode: 9.5% for reds, 9.5% for whites (with 10.5% also common in whites)

5. How is wine quality distributed in the dataset, and does it differ by wine type?

- The notebook plots and counts quality scores:
    - Most common score: 6 for both reds and whites
    - Red wine: Most samples are rated 5 or 6
    - White wine: Slightly more samples rated 6 or 7; wider spread at higher scores

Recommendations (Based on Analysis)

- Quality Control: Focus on managing sulfur dioxide and sugar levels, especially for white wines, due to their high variability.
- Product Consistency: Monitor density and chlorides, as they are stable and can serve as quality baselines.
- Marketing: Highlight wines with common, preferred characteristics (e.g., pH ≈ 3.3, alcohol ≈ 10.5%) and quality scores of 6 or higher.

Future Work

- Predictive Modeling: Use machine learning to predict wine quality from chemical properties.
- Feature Importance: Quantify which properties most influence quality.
- Sensory Data Integration: Combine chemical and sensory/tasting data for richer insights.
- Trend Analysis: If vintage/year data is available, analyze changes over time.