Wine Multi-Dataset Analysis Report

Analysis Date: June 12, 2026 at 15:30

Comprehensive analysis of 42,219 wine samples
across 4 datasets

Datasets Analyzed:

1. **Wine Quality Dataset (General)**
   1,143 samples with 13 chemical features

2. **Red & White Wine Quality Combined**
   6,497 samples (1,599 red, 4,898 white wines)

3. **Red Wine Quality (Cortez et al. 2009)**
   1,599 samples from Portuguese Vinho Verde region

4. **Global Wine Ratings Dataset**
   32,980 wines with ratings, prices, regions, and varieties

Analysis performed by:
Jordi Tarroch Mejón

Table of Contents

1. Executive Summary

This report presents a comprehensive analysis of wine quality and characteristics across four major datasets, totaling 42,219 wine samples. The analysis incorporates chemical composition data, quality ratings, and regional information to provide insights into wine styles, aging potential, and taste profiles.

Key Findings:

1. Quality Predictors (Machine Learning Analysis):
- **Alcohol content** is the strongest single predictor (~13-15% of model importance)
- **Volatile acidity** (vinegar) is the strongest negative correlate (-0.39 with quality)
- **Sulphates** and **citric acid** show modest positive correlations with quality
- A Random Forest predicts the exact 0-10 score with ~68-71% held-out accuracy (5-fold CV ~51-59%), versus a ~43% majority-class baseline -- useful but far from perfect, so chemistry explains only part of perceived quality

2. Wine Style Clusters (Unsupervised Learning):
- Analysis discovered **4 natural wine style groups** based on chemistry
- **Cluster 1** (Full-bodied, 11.2% alcohol) shows highest quality (6.09/10)
- **Cluster 3** (Medium alcohol, 10.2%) shows lowest quality (5.39/10)
- Cluster quality means span only 5.4-6.1/10, so chemistry-based groupings relate to expert ratings only weakly

3. Body & Elegance:
- **Light-bodied** wines (8-11% alcohol): Delicate, floral, elegant
- **Medium-bodied** (11-13%): Balanced, food-friendly, most common
- **Full-bodied** (13%+ alcohol): Rich, powerful, oak aging candidates
- **Oak aging potential**: 13%+ alcohol + 0.8+ g/L sulphates (vanilla/toast notes)

4. Oxidation Risk & Preservation:
- **High risk** (<15 mg/L free SO_2): Develops rusty/metallic notes
- **Protected zone** (20-50 mg/L): Preserves fresh fruit character
- Proper SO_2 levels critical for preventing oxidation faults

5. Red vs White Wine Differences:
- Red wines show higher sulphates (spicy character)
- White wines have higher acidity and SO_2 (preservation)
- Distinct quality distributions: whites more variable, reds more consistent

6. Geographic Excellence:
- Top regions identified from 32,980 global wine ratings
- Quality varies by region and grape variety
- Note: these are curated critic scores ranging only 85-99 (mean 91), i.e. a selection-biased sample of already well-regarded wines -- not a representative cross-section of all wine, so regional rankings should be read as relative standing among reviewed wines, not absolute quality

2. Dataset Overview

Combined Summary

Shape: (6497, 13)
Total Samples: 6,497
Features: 13
Wine Types:
- white: 4,898
- red: 1,599

Quality Distribution:
Quality 3: 30
Quality 4: 216

Ratings Summary

Shape: (32980, 5)
Total Samples: 32,980
Features: 5
Rating Statistics:
Mean: 91.19
Range: 85.0 - 99.0

Red Only Summary

Shape: (1599, 12)
Total Samples: 1,599
Features: 12
Quality Distribution:
Quality 3: 10
Quality 4: 53
Quality 5: 681
Quality 6: 638
Quality 7: 199
Quality 8: 18

Wine Quality Summary

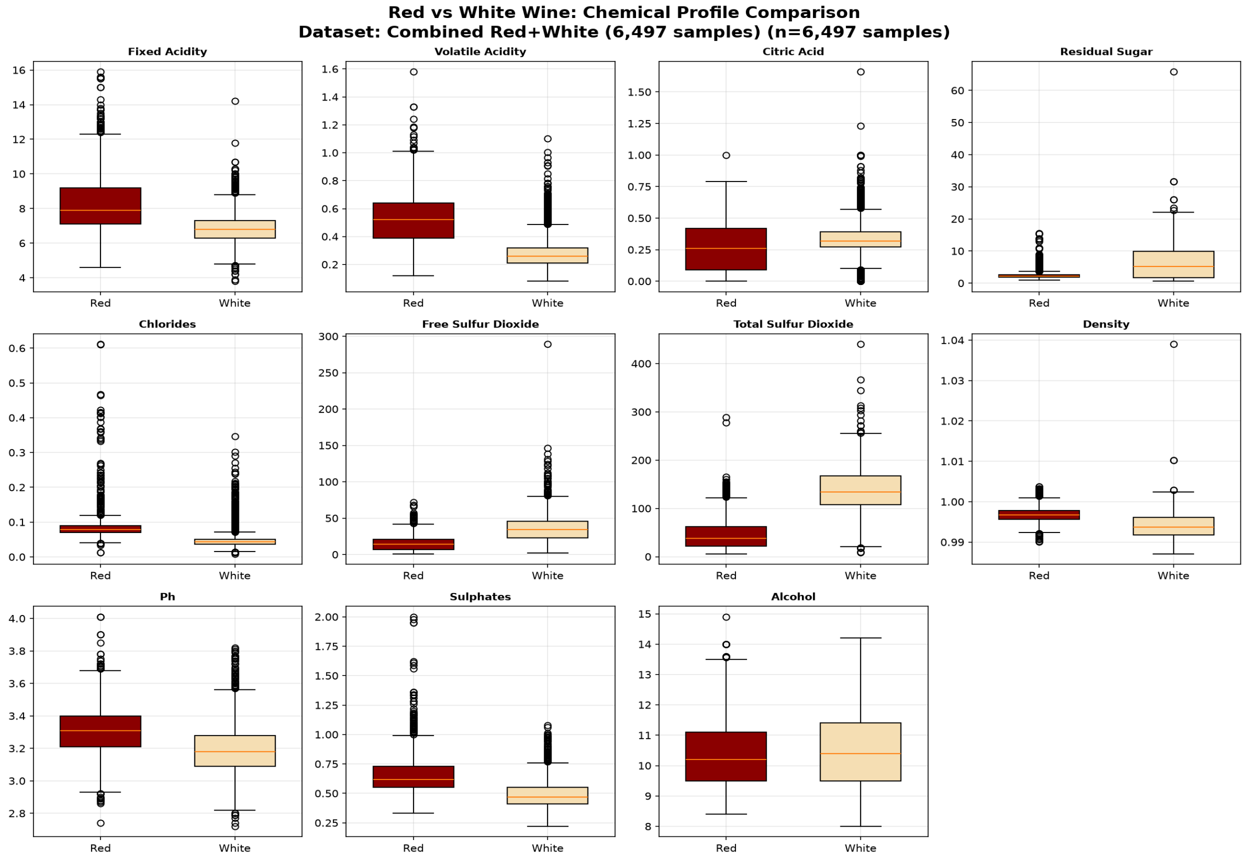
Shape: (1143, 13)
Total Samples: 1,143
Features: 13
Quality Distribution:
Quality 3: 6
Quality 4: 33
Quality 5: 483
Quality 6: 462
Quality 7: 143
Quality 8: 16

3. Wine Style Analysis

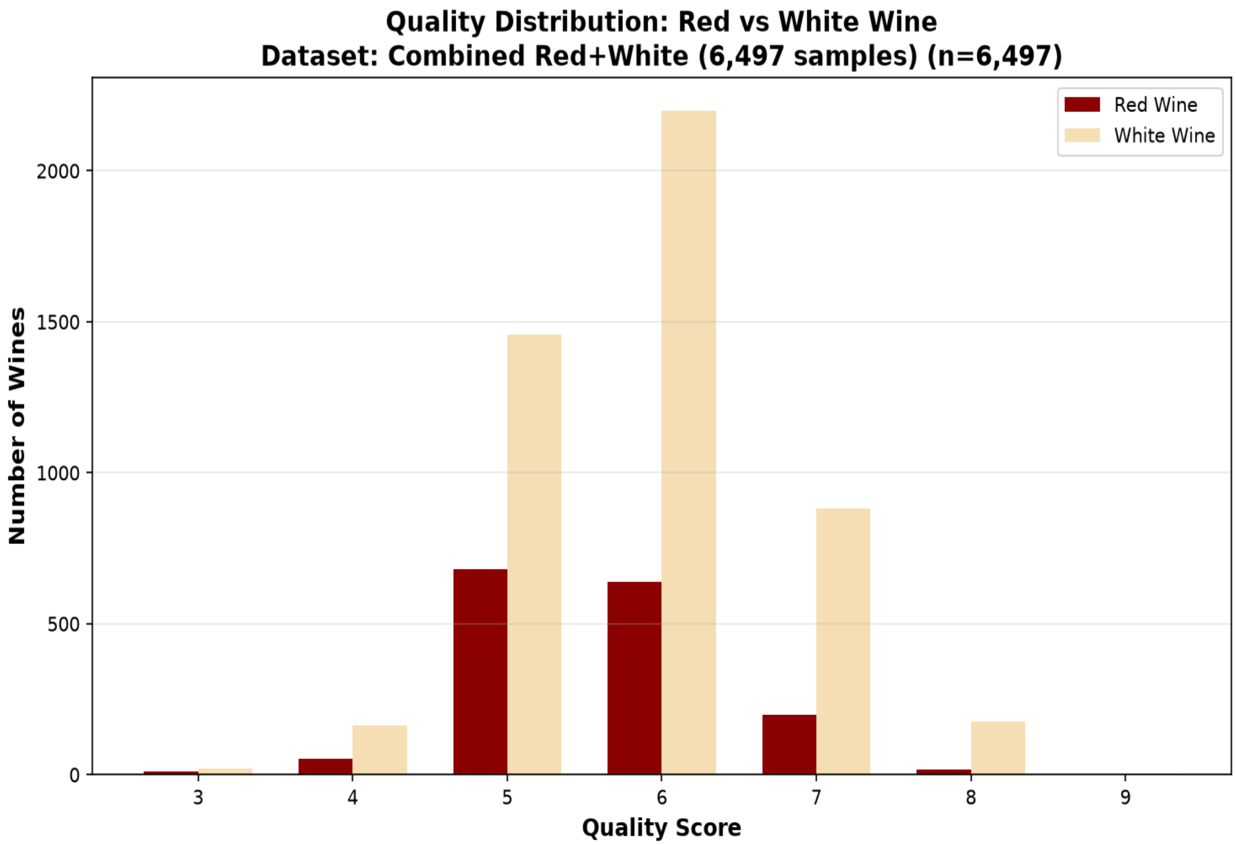
Total wines analyzed: 1,599
Style Distribution:
elegant: 1,529 (95.6%)
light-bodied: 1,132 (70.8%)
oxidation_risk: 846 (52.9%)
crisp: 741 (46.3%)
medium-bodied: 438 (27.4%)
soft: 58 (3.6%)
full-bodied: 29 (1.8%)
oak_candidate: 8 (0.5%)

Wine Style Definitions:
- **Light-bodied:** Alcohol < 11%, floral and elegant characteristics
- **Medium-bodied:** Alcohol 11-13%, balanced structure
- **Full-bodied:** Alcohol 13%+, rich and powerful
- **Oak candidates:** High alcohol + high sulphates (vanilla/toast notes)
- **Elegant:** 11-12.5% alcohol with good acidity balance
- **Oxidation risk:** Free SO2 < 15 mg/L (rusty/metallic notes)

4. Red vs White Wine Comparison



Chemical Profile Comparison



Quality Distribution Comparison

Understanding Oak Aging Potential:

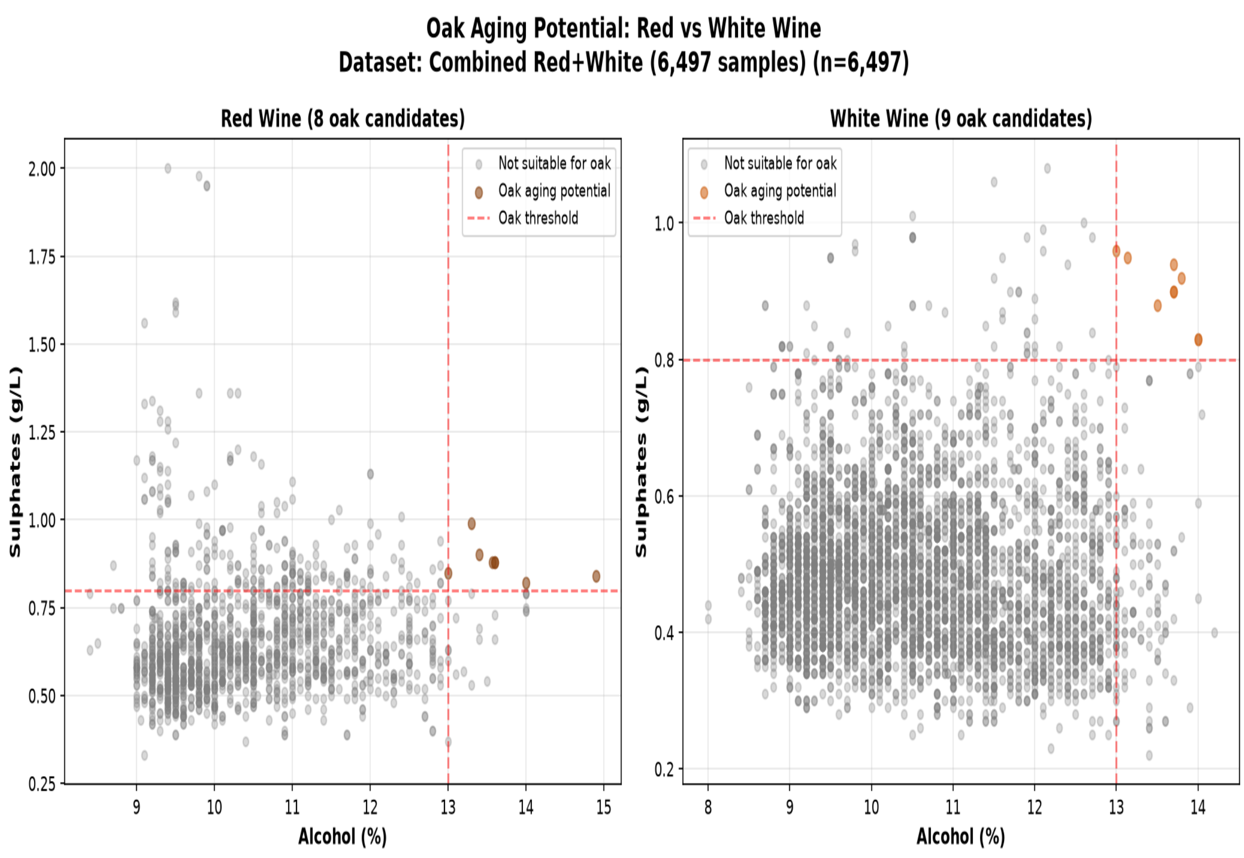
These scatter plots identify wines suitable for oak barrel aging based on two key criteria:

1. Alcohol Content (≥13%): Higher alcohol wines have the structure to benefit from oak aging. Full-bodied wines can absorb oak flavors without being overwhelmed.

2. Sulphates (≥0.8 g/L): Higher sulphates contribute spicy complexity that complements oak character.

Oak Barrel Aging Adds:
- **Vanilla** notes from oak vanillin (especially American oak)
- **Toast/Caramel** flavors from barrel charring
- **Coconut** character from oak lactones
- **Spice complexity** (clove, cinnamon) from oak tannins
- Smoother mouthfeel through micro-oxygenation

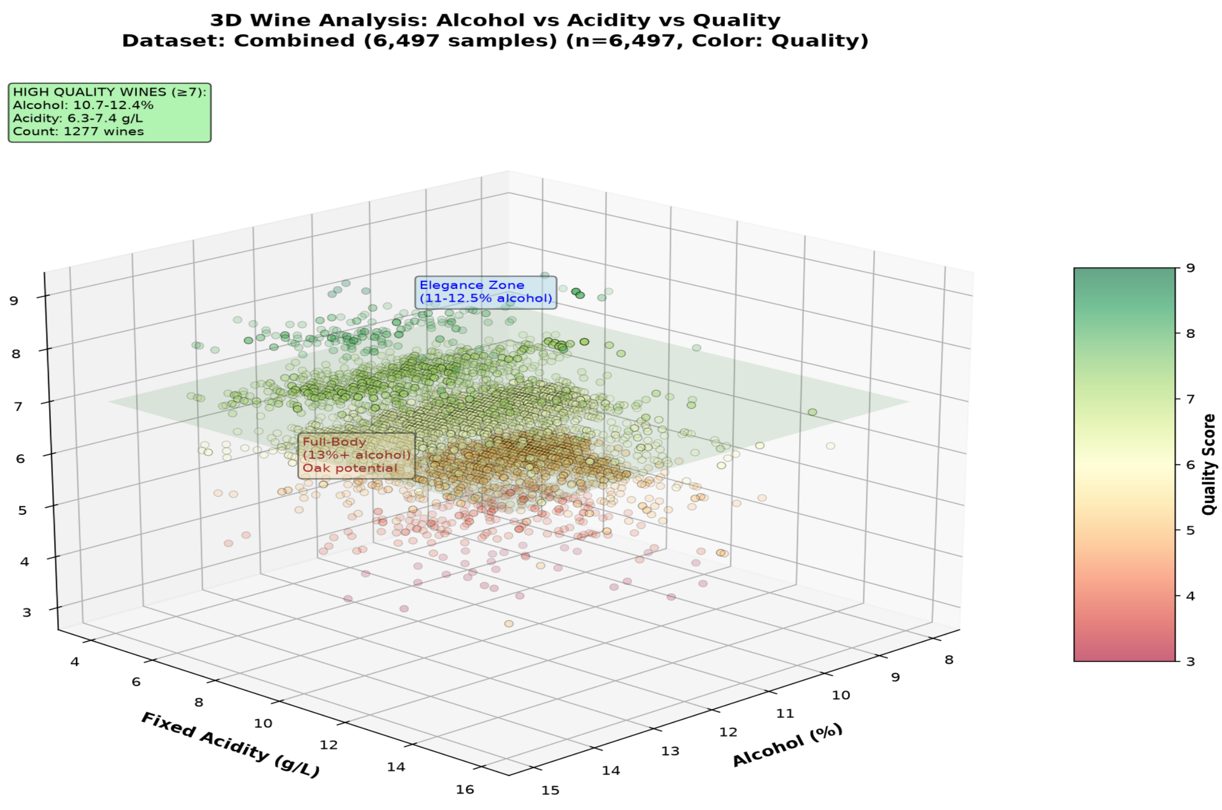
Chart Interpretation:
- **Upper-right quadrant** (red zone): Wines meeting both criteria = Oak aging candidates
- **Gray points**: Wines better suited for stainless steel or minimal oak (preserve fresh fruit)
- Red wines show more oak candidates than whites due to higher structure and sulphates



Oak Aging Potential (Vanilla/Toast Notes)

5. Chemical Profile Analysis (3D)

Three-dimensional visualizations reveal relationships between key chemical parameters and wine quality. These plots show how multiple factors interact to influence overall wine quality ratings.



Alcohol vs Acidity vs Quality

5b. Multi-Dimensional Chemical Profiles

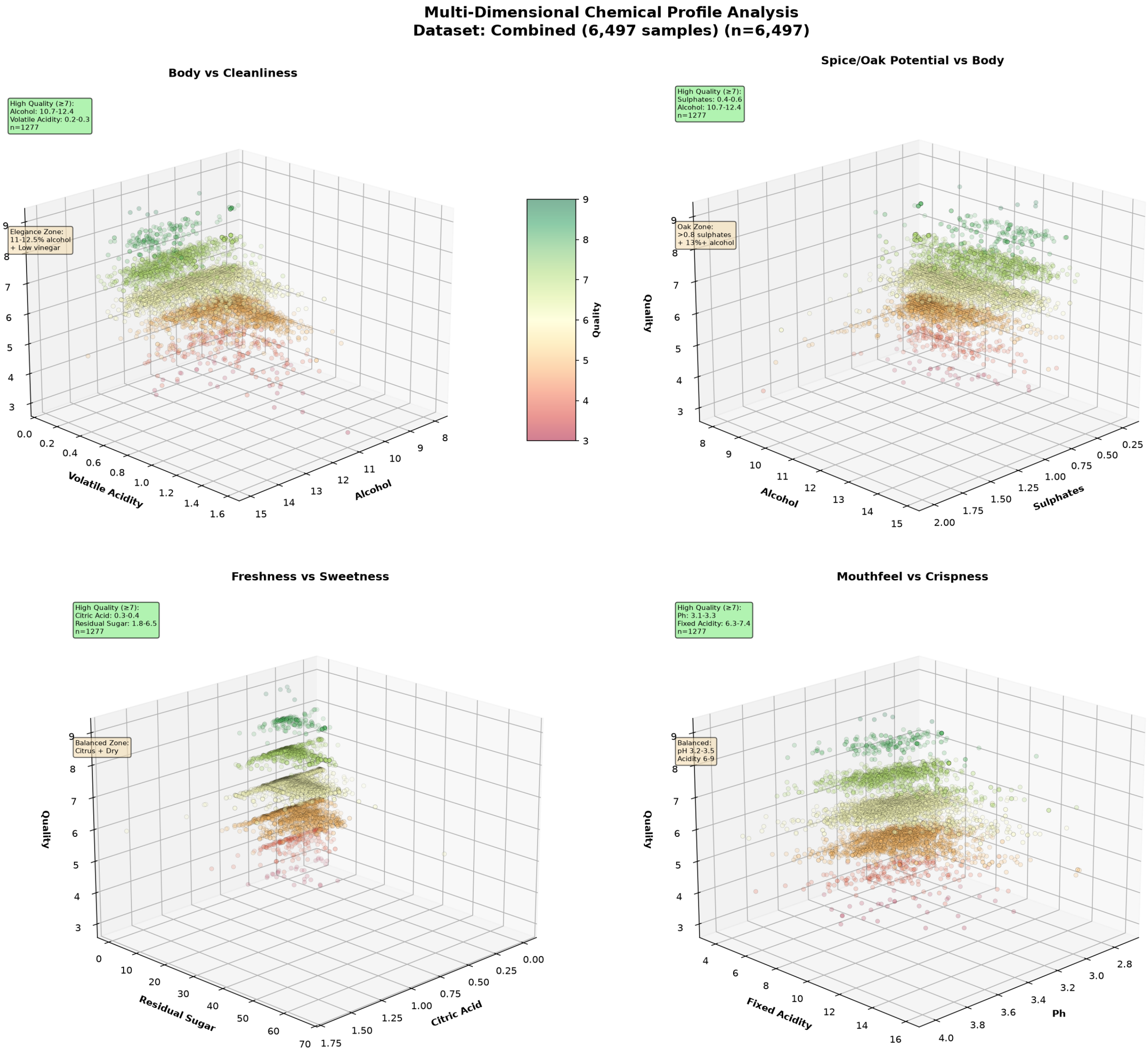
Beyond alcohol and acidity, wine quality depends on multiple chemical interactions. These four views explore different dimension combinations:

Top-Left: Body vs Cleanliness
Alcohol content (body) versus volatile acidity (vinegar). Best wines combine good body with low vinegar character.

Top-Right: Spice/Oak Potential vs Body
Sulphates (spice/oak) versus alcohol (body). High values in both indicate wines suitable for oak barrel aging (vanilla/toast notes).

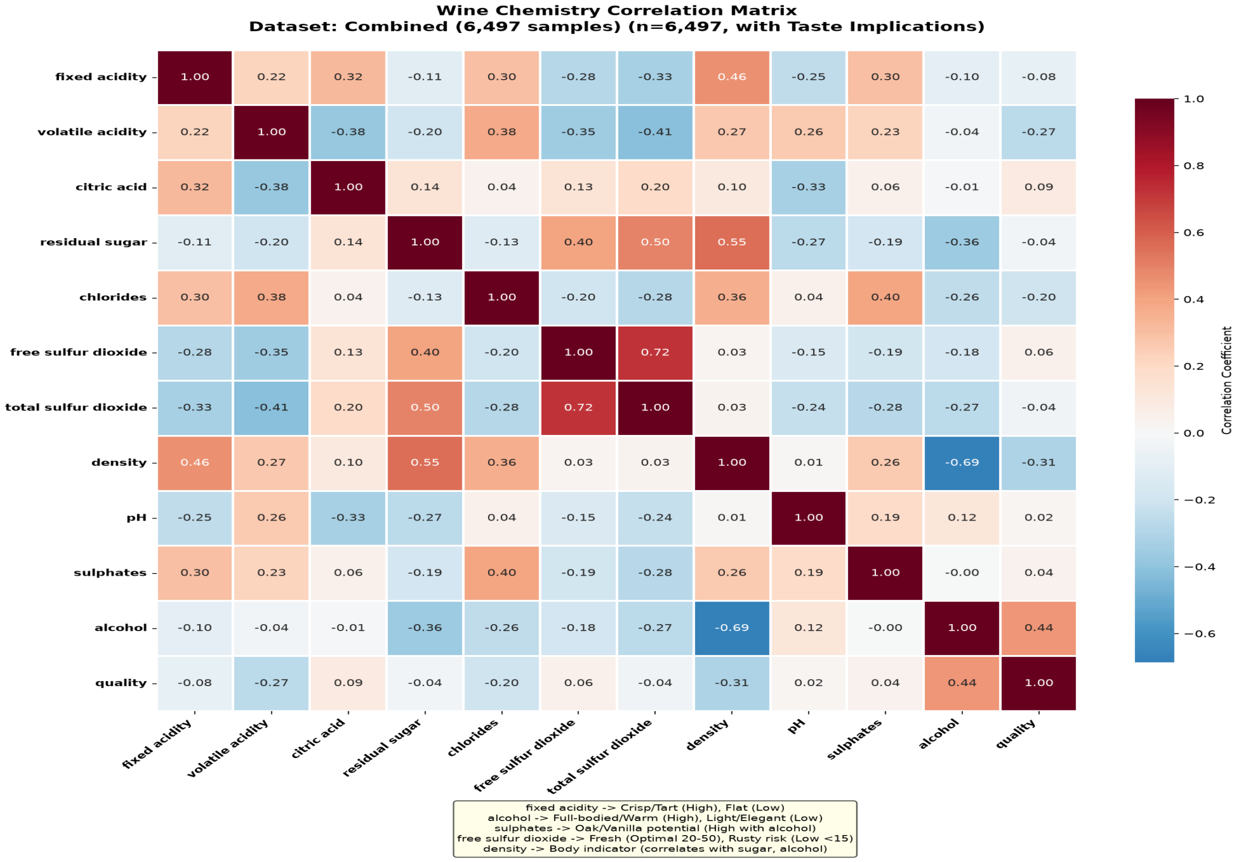
Bottom-Left: Freshness vs Sweetness
Citric acid (freshness) versus residual sugar (sweetness). Shows the balance between crisp citrus character and fruit sweetness.

Bottom-Right: Mouthfeel vs Crispness
pH (mouthfeel) versus fixed acidity (crispness). Balanced wines show moderate values in both dimensions.



Multi-Dimensional Chemical Profile Analysis

6. Correlation Analysis

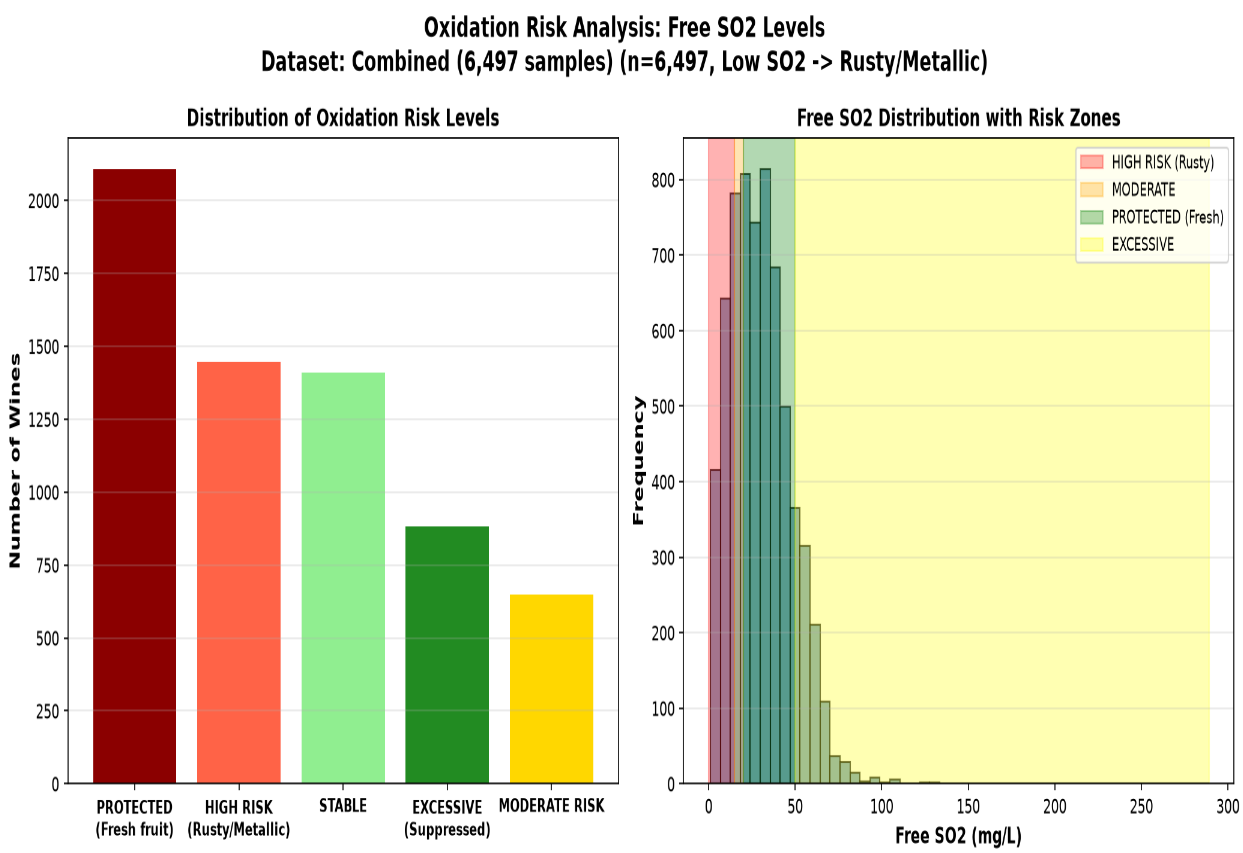


Chemical Correlations with Taste Descriptors

7. Oxidation Risk Assessment

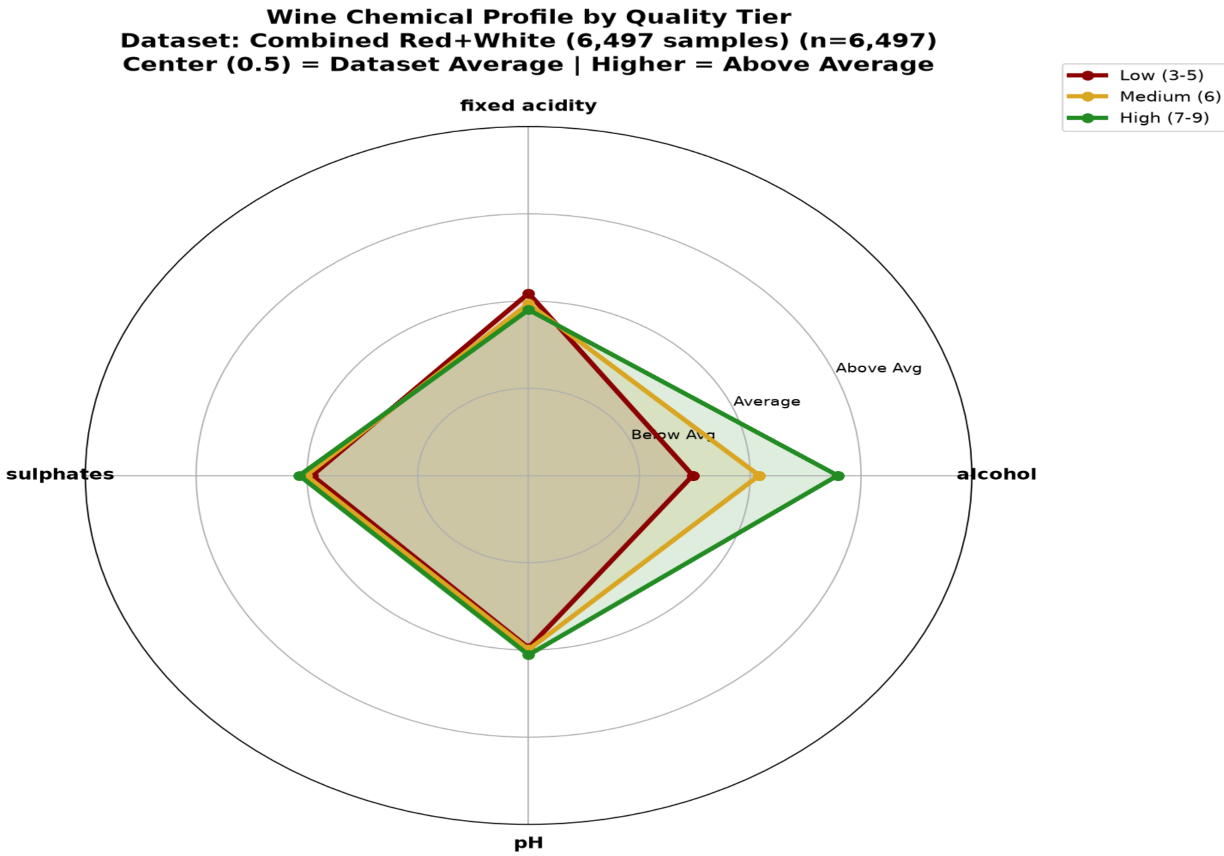
Free sulfur dioxide (SO_2) acts as a preservative in wine. Low levels increase oxidation risk, leading to rusty or metallic notes. Optimal levels preserve fresh fruit characteristics.

Risk Categories:
- HIGH RISK (<15 mg/L): Rusty/metallic notes likely
- MODERATE RISK (15-20 mg/L): Monitor storage conditions
- PROTECTED (20-50 mg/L): Fresh fruit preserved
- EXCESSIVE (>50 mg/L): May suppress aromatic compounds



Free SO2 Distribution and Risk Zones

8. Quality Analysis by Tier

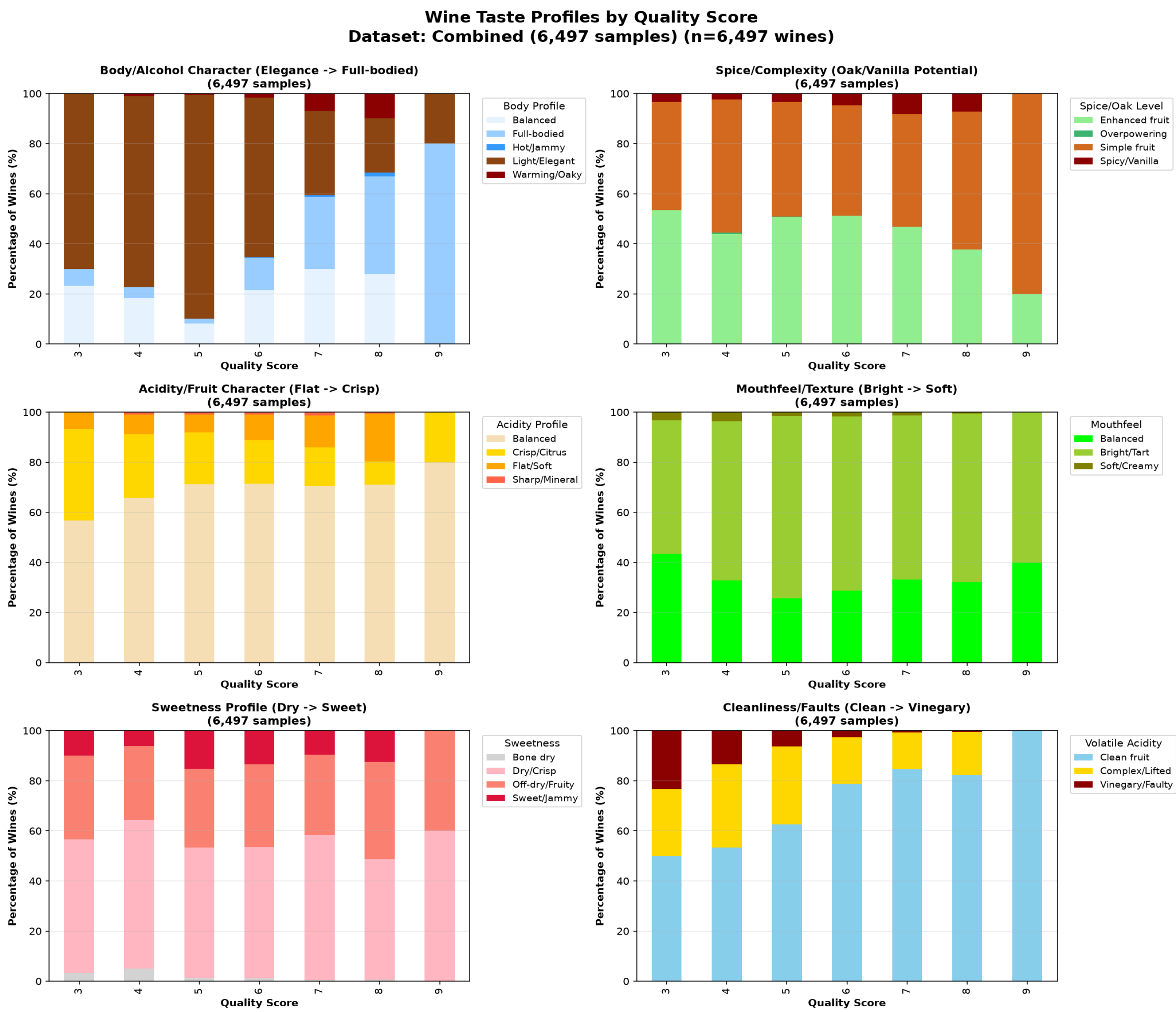


Chemical Profile Radar Chart by Quality Tier

9. Taste Descriptor Profiles

These charts show how chemical parameters translate into taste experiences. Each quality level shows the distribution of wines across different taste profiles:

- **Body:** Light/Elegant to Full-bodied (alcohol content)
- **Spice/Oak:** Simple fruit to Spicy/Vanilla notes (sulphates)
- **Acidity:** Flat to Crisp/Citrus (fixed acidity)
- **Mouthfeel:** Bright to Soft/Creamy (pH)
- **Sweetness:** Bone dry to Sweet (residual sugar)
- **Cleanliness:** Clean to Vinegary (volatile acidity)



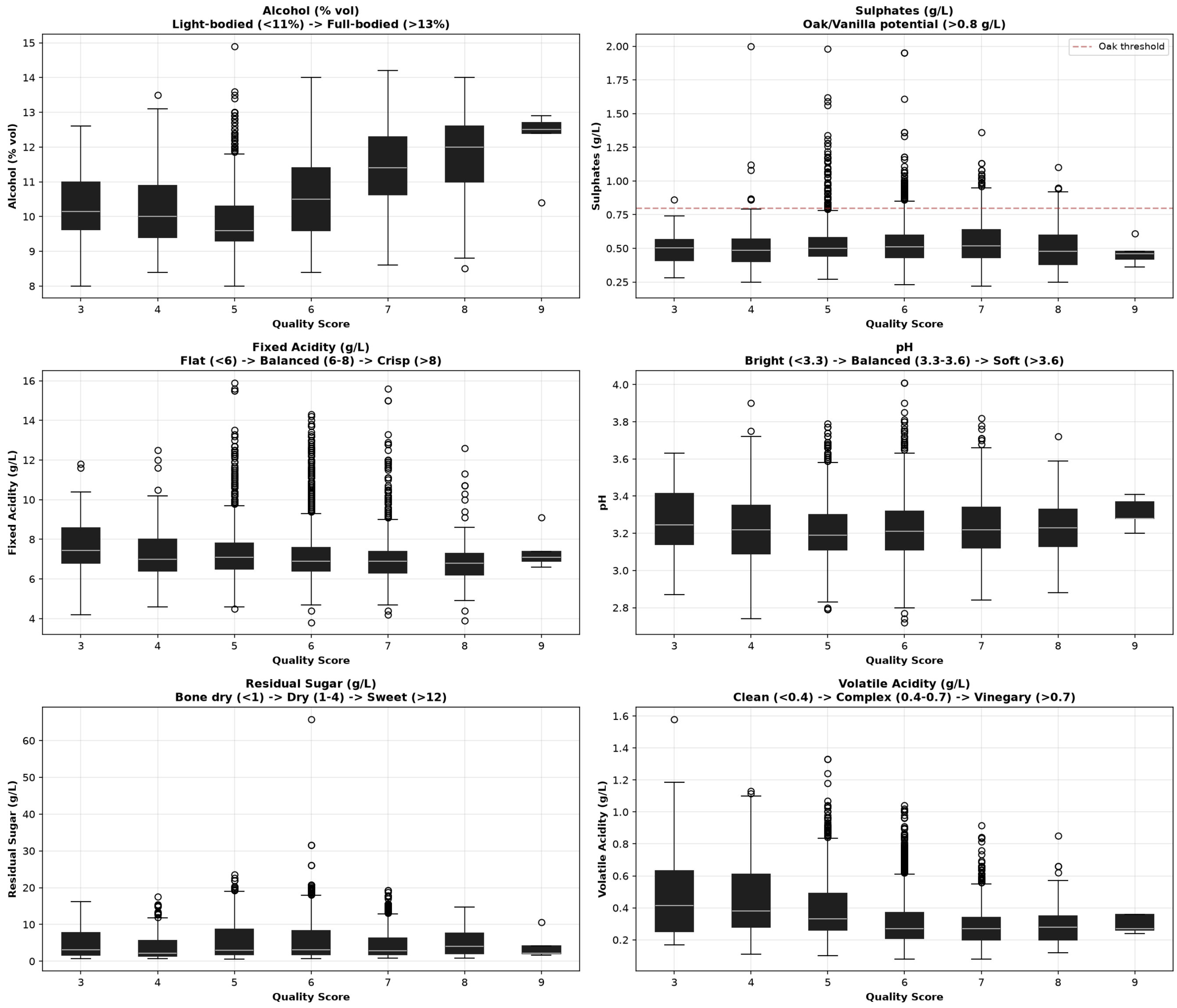
Taste Descriptor Distributions by Quality Score

10. Raw Chemical Values by Quality

These boxplots show the actual chemical measurements (not taste-mapped) for each quality level. Each boxplot displays the distribution of values with:

- **Box:** 25th to 75th percentile (middle 50% of values)
- **Line in box:** Median value
- **Whiskers:** Range of typical values
- **Circles:** Outliers (unusual values)

These raw values show the actual chemical composition differences between quality levels, with taste descriptor thresholds noted in subtitles.



Raw Chemical Value Distributions by Quality Score

11. Feature Importance Analysis

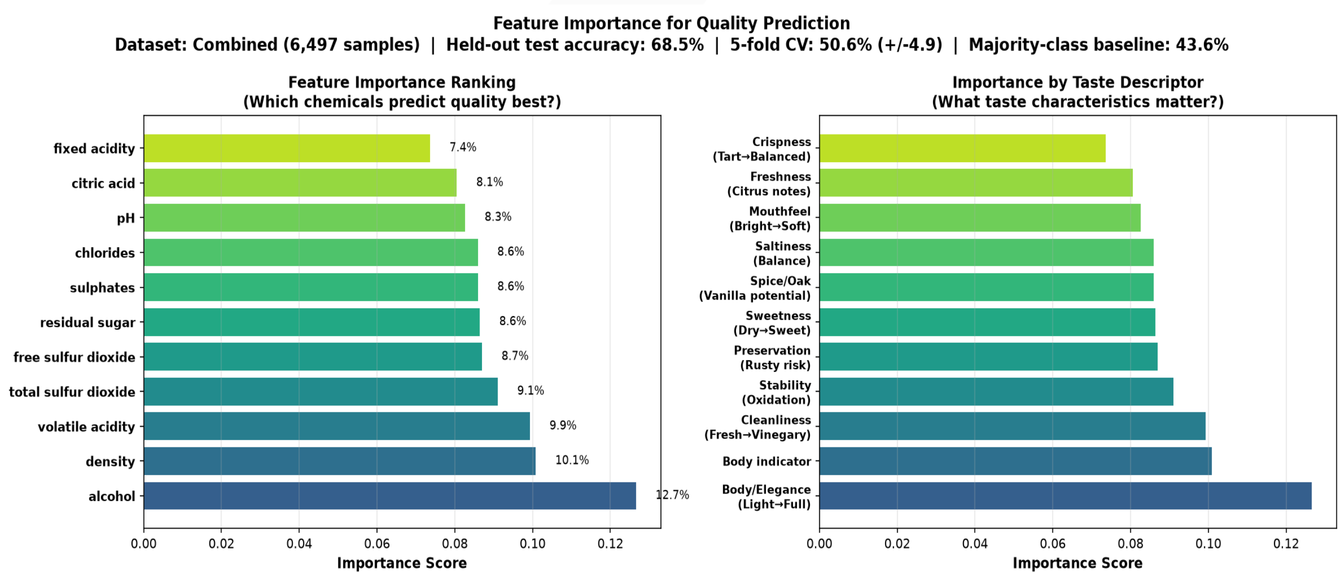
Machine Learning analysis using Random Forest to identify which chemical features most strongly predict wine quality. This reveals which taste characteristics matter most for quality assessment:

- **High importance:** Strong predictor of quality
- **Low importance:** Less relevant to quality prediction

The right panel translates chemical features into taste descriptors, showing what taste characteristics expert tasters value most.

How good is the model, really? Accuracy is measured on a held-out test set the model never saw during training (scoring on training rows would report a meaningless ~100%). See the figure subtitle for held-out, cross-validated and baseline accuracy on this dataset.

QUALITY-PREDICTION MODEL PERFORMANCE (HONEST)
Dataset: Combined (6,497 samples)
Model: RandomForestClassifier(n_estimators=300)
Features used: 11
Held-out test accuracy (20% split): 68.5%
5-fold cross-validation accuracy: 50.6% (+/- 4.9)
Majority-class baseline: 43.6%
Note: accuracy is measured ONLY on rows the model never saw during
training. Scoring on the training rows would report ~100%, which
reflects memorisation, not predictive skill.
Top feature importances:
alcohol 12.7%
density 10.1%
volatile acidity 9.9%
total sulfur dioxide 9.1%
free sulfur dioxide 8.7%
residual sugar 8.6%



Feature Importance for Quality Prediction

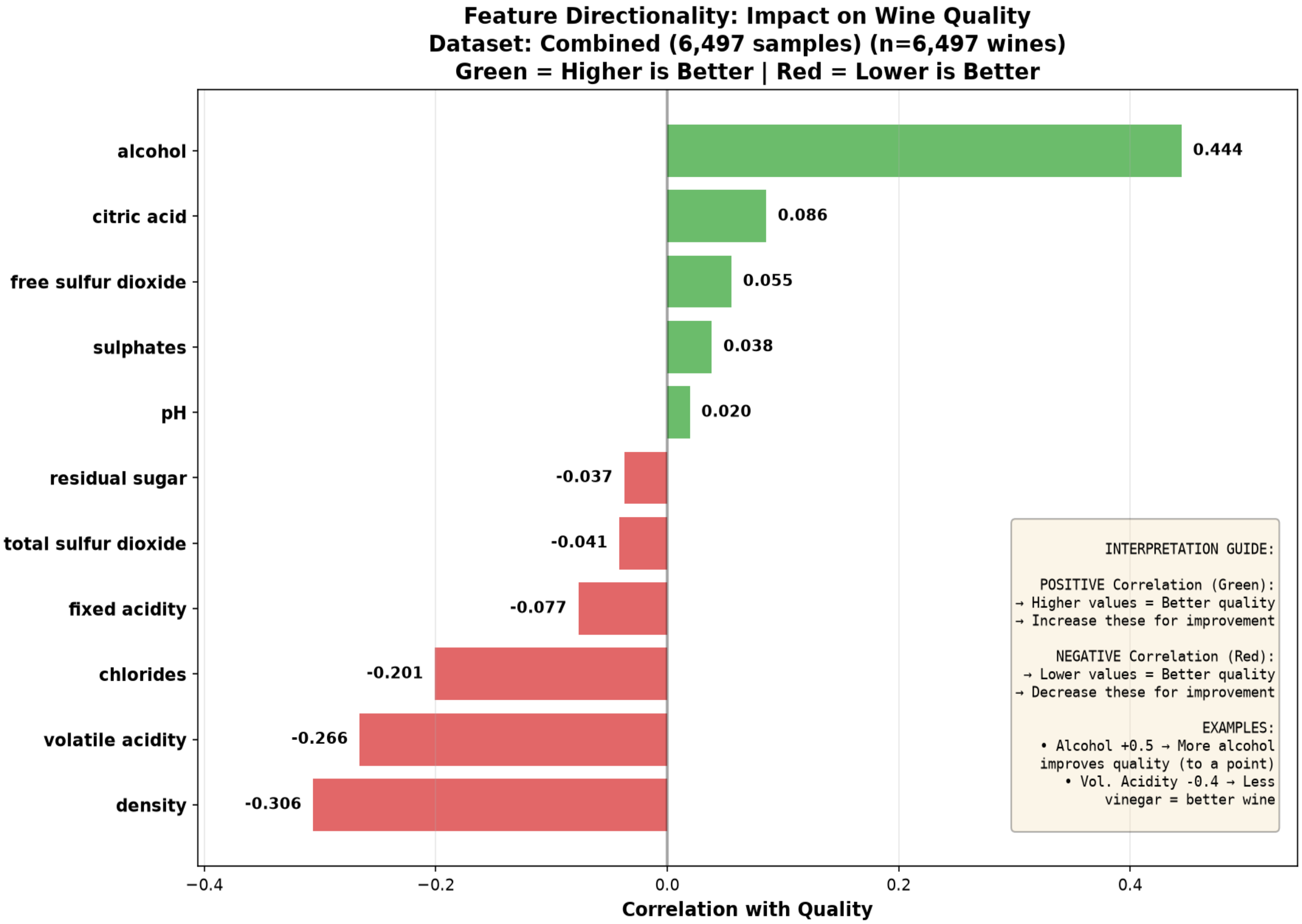
11b. Feature Directionality Analysis

This analysis shows the **correlation direction** between chemical features and quality. It answers a critical question: "Should this feature be higher or lower for better wine?"

GREEN bars (Positive correlation):
- Higher values tend to coincide with higher quality scores
- Example: If alcohol shows +0.5, wines with more alcohol tend to score higher

RED bars (Negative correlation):
- Higher values tend to coincide with lower quality scores
- Example: If volatile acidity shows -0.4, wines with less vinegar score higher

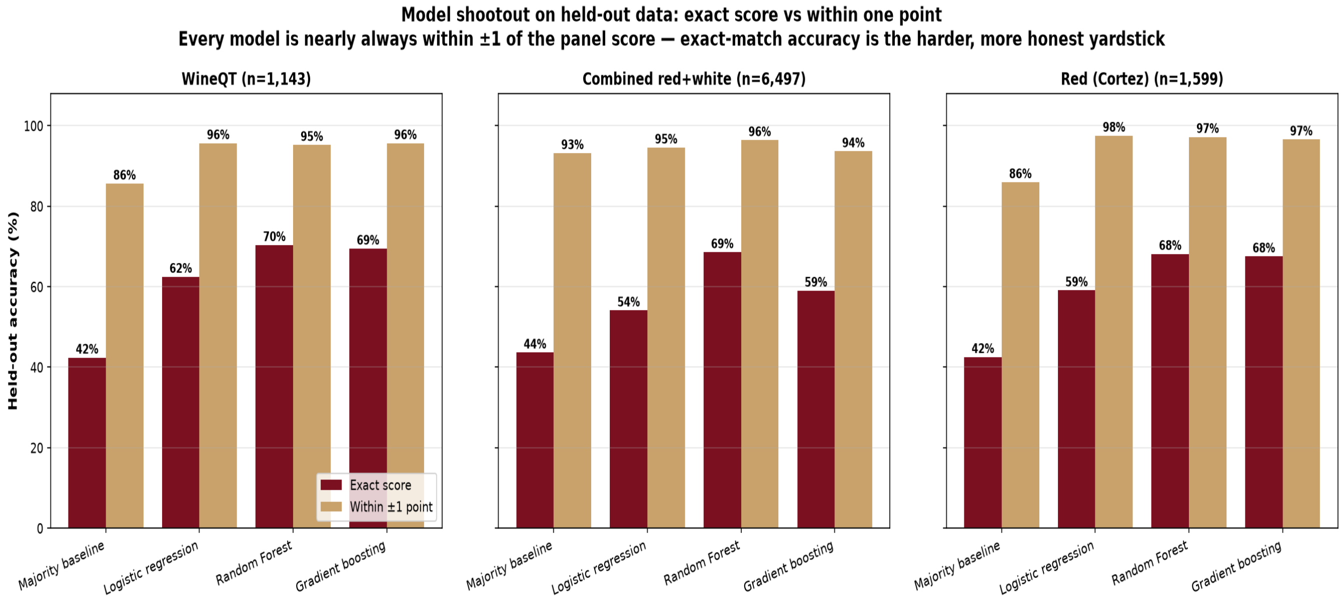
Important caveat (correlation is not causation):
These are observational correlations in finished wines, not controlled experiments. A winemaker cannot simply "add alcohol" to raise a score -- alcohol is itself a downstream marker of ripe fruit, longer fermentation and style. Read these directions as descriptive associations, not production levers.



Feature Direction: Which Features to Increase vs Decrease

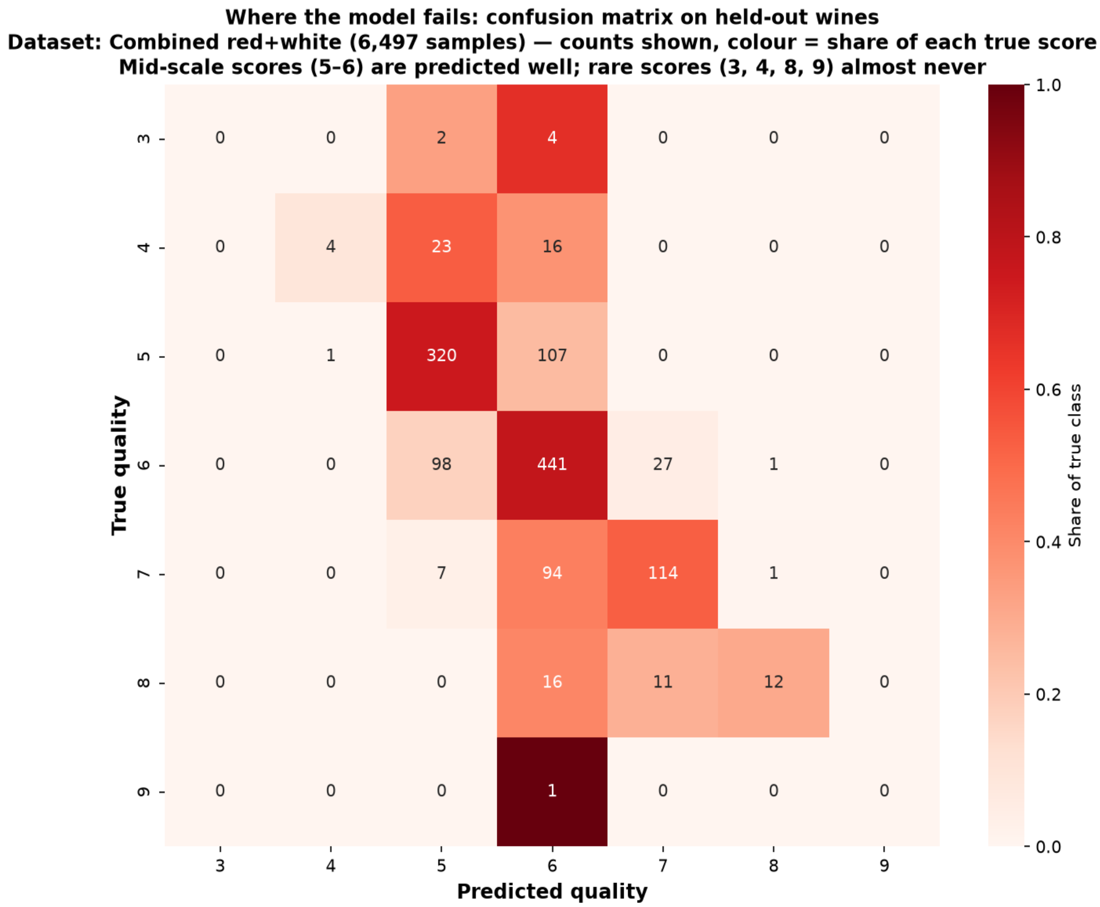
11c. Model Benchmark

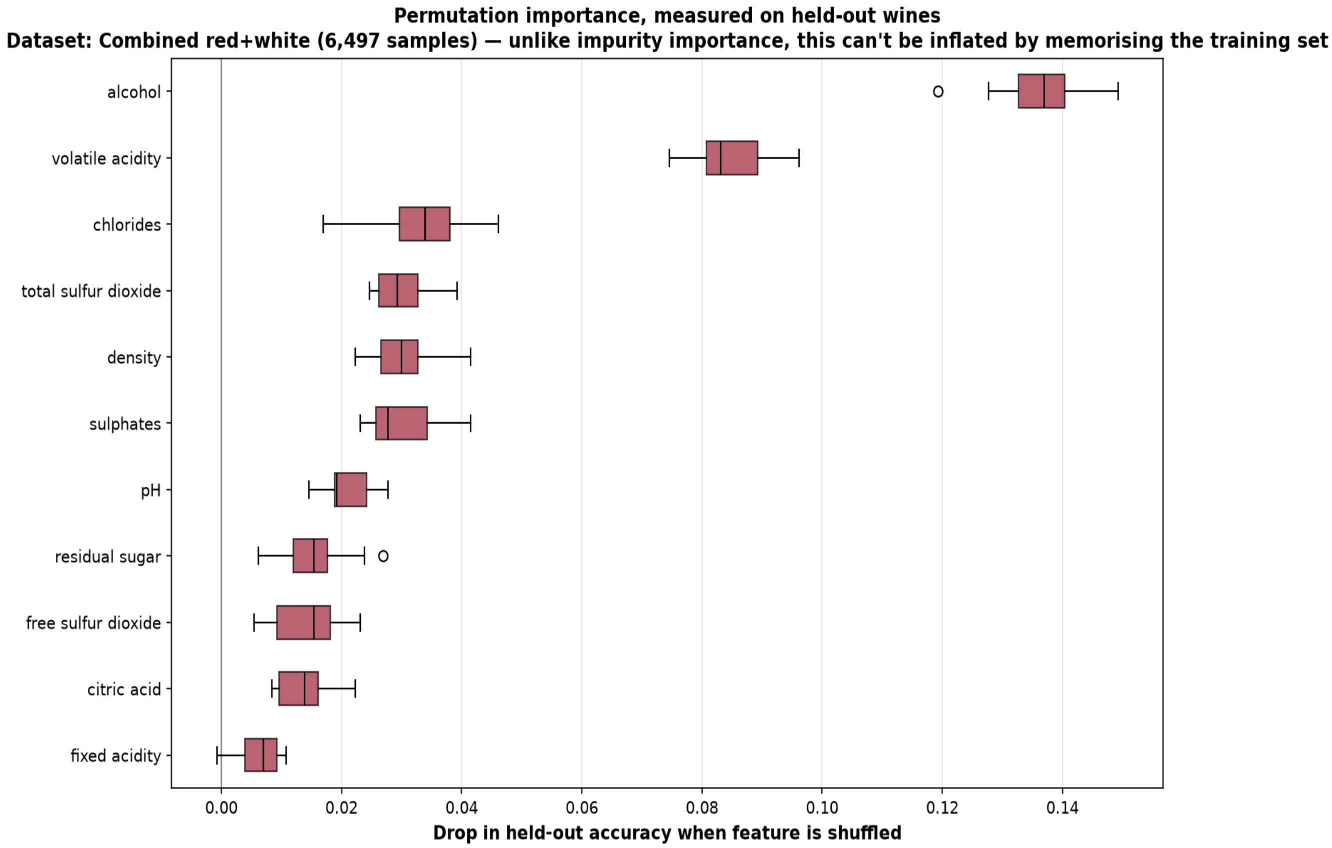
Four models are compared on each quality dataset, all scored on a held-out 20% split: a majority-class baseline, logistic regression, Random Forest and gradient boosting. Two accuracies are shown: **exact score** (the model names the precise 0-10 panel score) and **within one point** (prediction lands within +/-1 of the panel). Because quality is an ordinal scale judged by humans whose own repeat-tasting agreement is imperfect, within-one-point accuracy is arguably the fairer yardstick -- and there every model exceeds 90%.



11d. Error Analysis

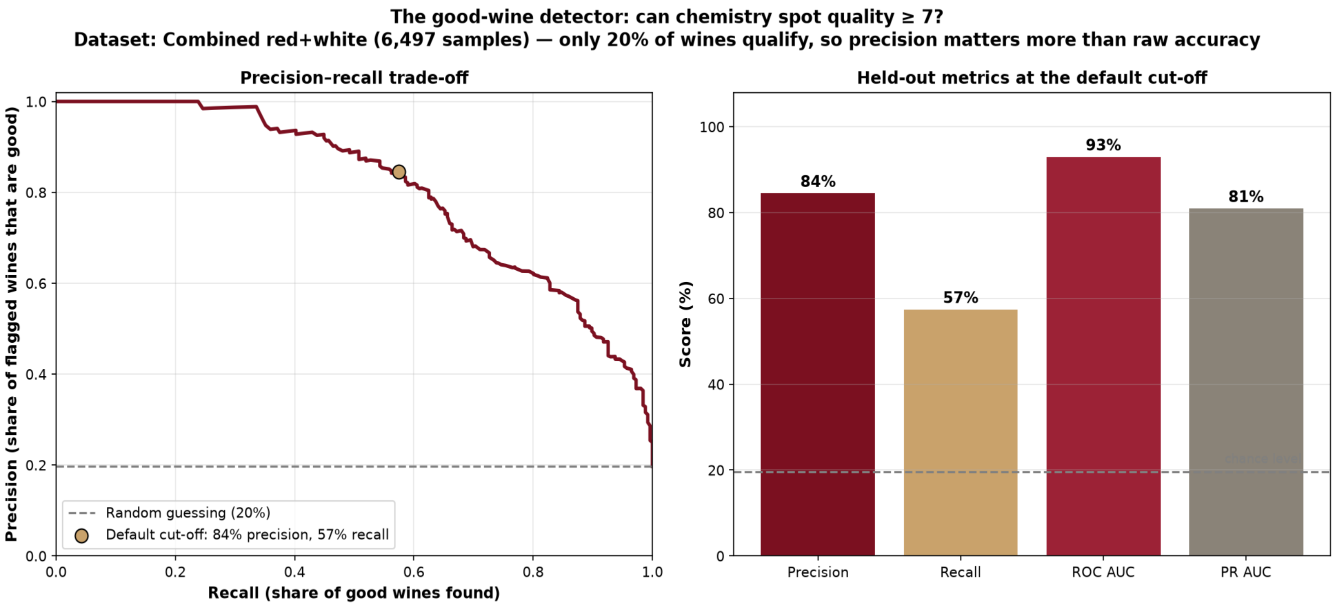
The confusion matrix shows exactly where predictions fail: the model is reliable for the common scores (5-6), drifts by one point for 7s, and essentially never predicts the rare scores (3, 4, 8, 9) -- the class-imbalance cost hidden inside any headline accuracy figure. Permutation importance re-ranks the features by how much held-out accuracy drops when each is shuffled; unlike impurity-based importance it cannot be inflated by memorising the training data.




Permutation importance, measured on held-out wines
Dataset: Combined red+white (6,497 samples) — unlike impurity importance, this can't be inflated by memorising the training set
alcohol
volatile acidity
chlorides
total sulfur dioxide
density
sulphates
pH
residual sugar
free sulfur dioxide
citric acid
fixed acidity
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Drop in held-out accuracy when feature is shuffled

11e. Good-Wine Detector

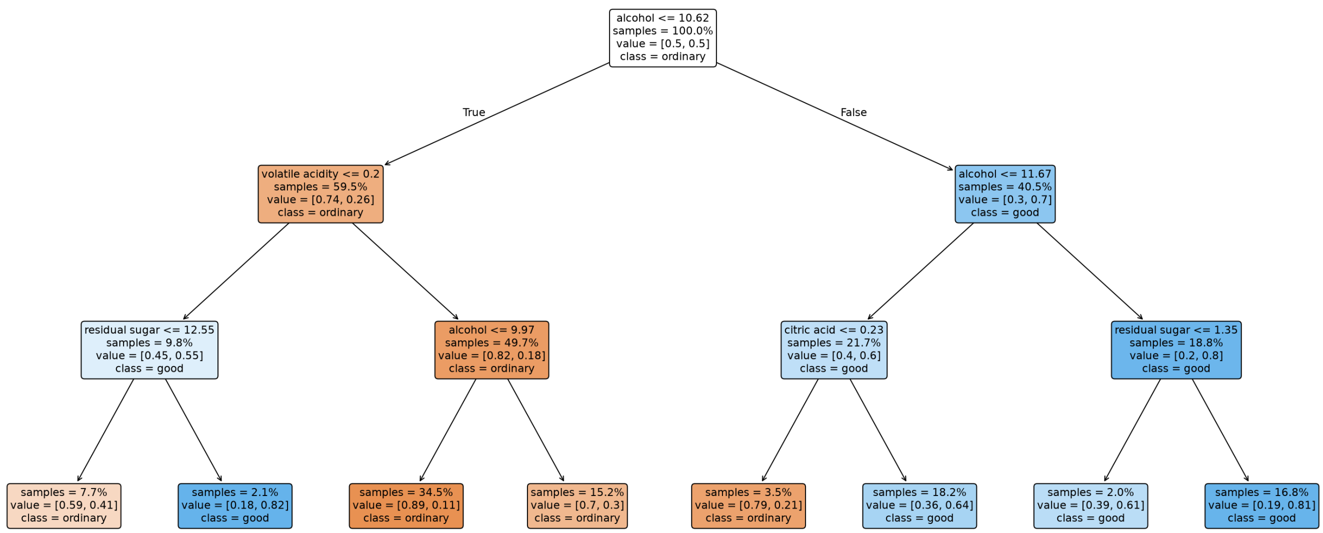
Reframing the task as the question a buyer actually asks -- "is this wine good (quality 7 or above)?" -- turns it into a binary problem on an imbalanced sample. The precision-recall curve shows the available trade-off; the bars report precision, recall and AUC scores on held-out wines against the chance level.

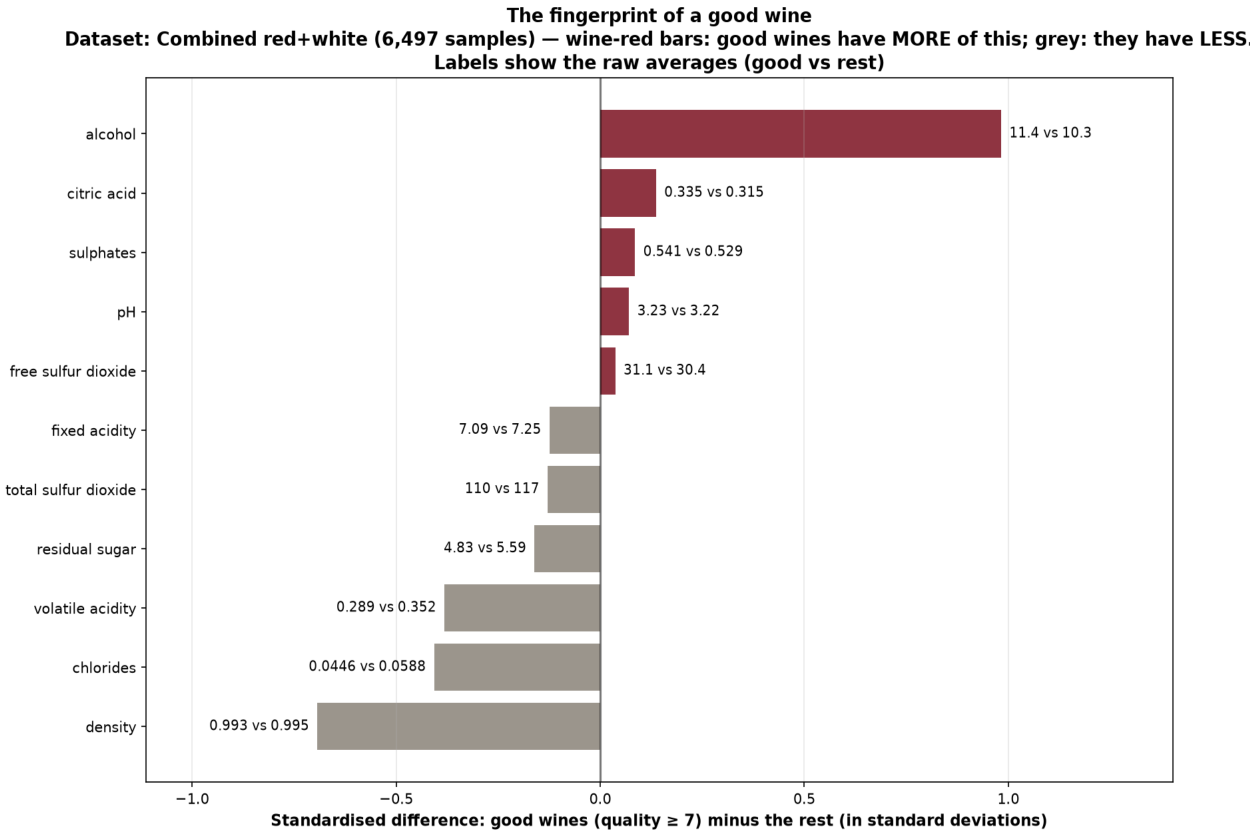


11e-2. What Makes a Good Wine

Two deliberately simple diagrams summarise the whole analysis. The **decision tree** asks at most three questions per wine (read top to bottom; the left branch means the condition is true) and already recovers most good wines on held-out data. The **fingerprint chart** shows how good wines (quality 7+) differ chemically from the rest, in standard deviations: more alcohol, less volatile acidity, lower density and chlorides.

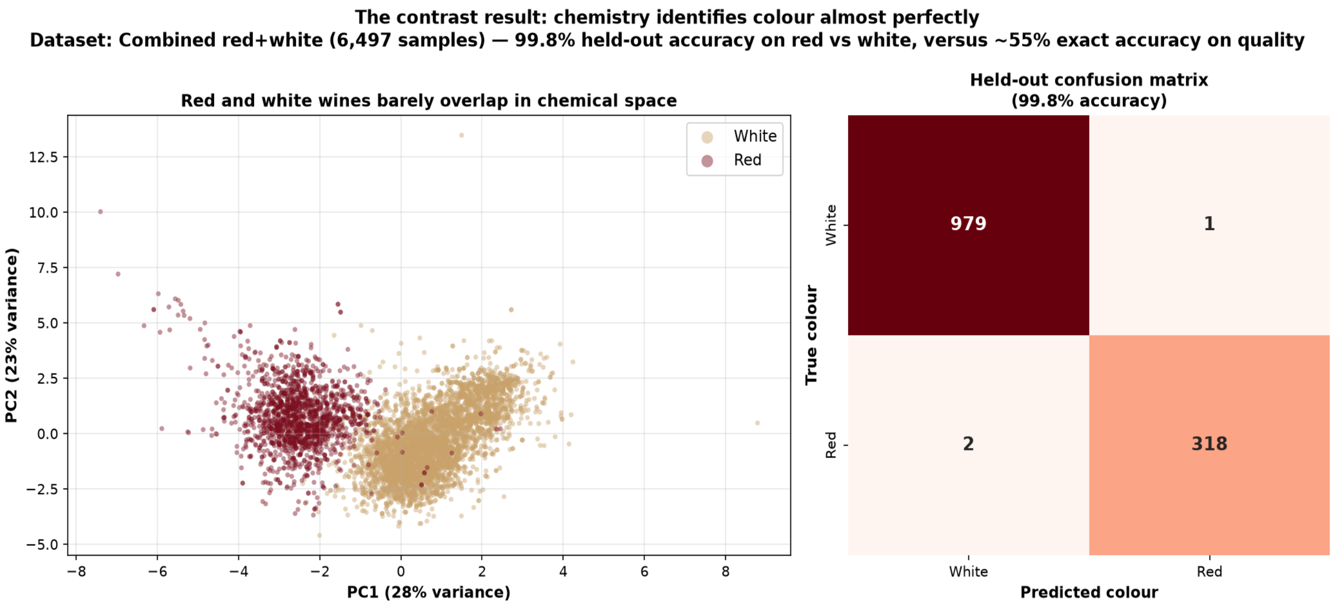
What makes a good wine (quality ≥ 7)? A three-question decision tree
Dataset: Combined red+white (6,497 samples) — read top to bottom: left branch = condition true, right = false. Blue = mostly good wines, orange = mostly ordinary.
This tiny tree alone finds 79% of good wines on held-out data (71% balanced accuracy)




The fingerprint of a good wine
Dataset: Combined red+white (6,497 samples) — wine-red bars: good wines have MORE of this; grey: they have LESS.
Labels show the raw averages (good vs rest)
alcohol
citric acid
sulphates
pH
free sulfur dioxide
fixed acidity
total sulfur dioxide
residual sugar
volatile acidity
chlorides
density
11.4 vs 10.3
0.335 vs 0.315
0.541 vs 0.529
3.23 vs 3.22
31.1 vs 30.4
7.09 vs 7.25
110 vs 117
4.83 vs 5.59
0.289 vs 0.352
0.0446 vs 0.0588
0.993 vs 0.995
−1.0
−0.5
0.0
0.5
1.0
Standardised difference: good wines (quality ≥ 7) minus the rest (in standard deviations)

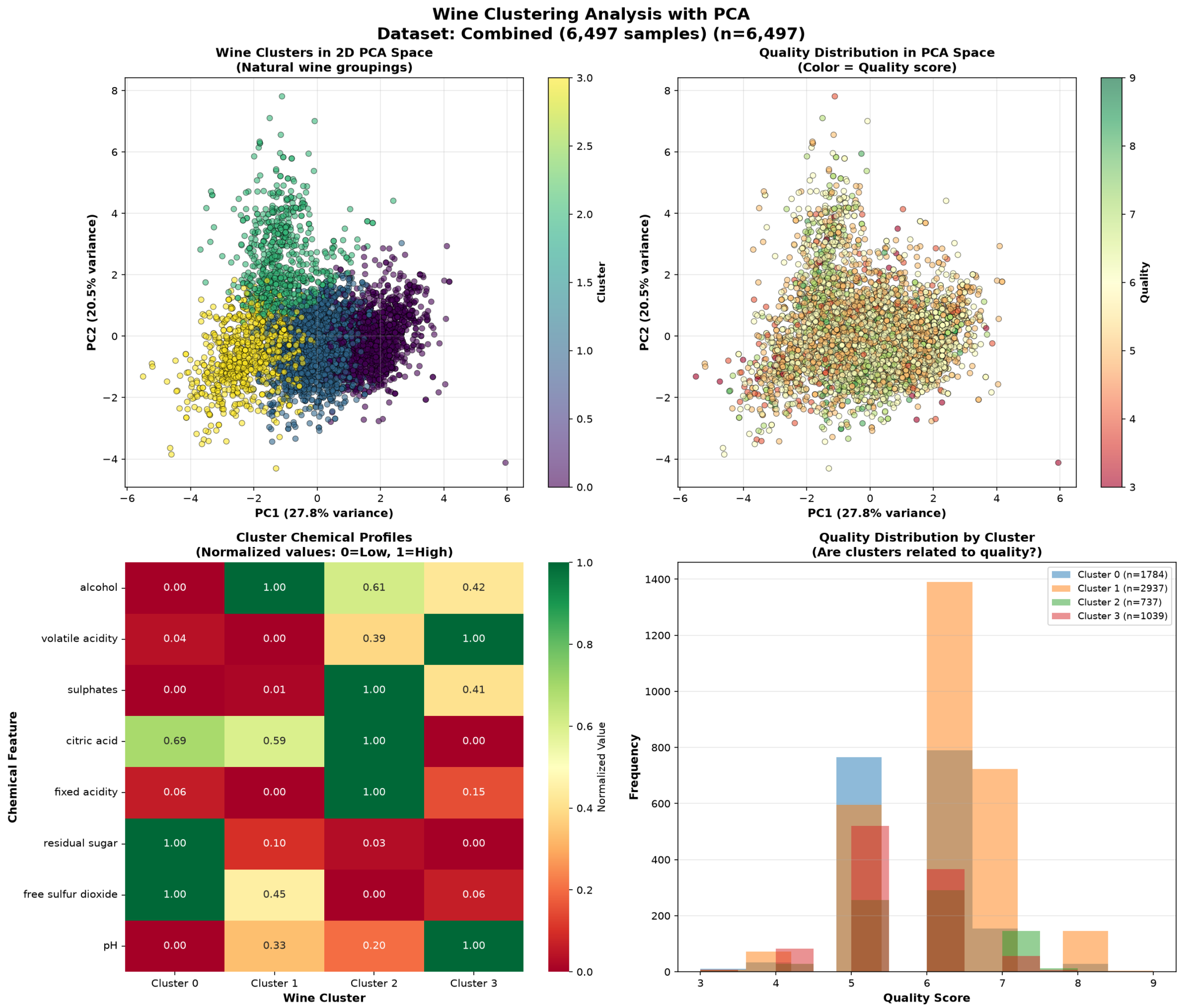
11f. The Contrast Result: Colour from Chemistry

The same eleven chemical measurements that only weakly predict quality identify a wine's colour almost perfectly. This is the cleanest evidence that the quality model's modest accuracy is a property of the target, not the chemistry: colour is written into the chemistry; quality largely is not.



12. Wine Clustering & PCA

Principal Component Analysis (PCA) reduces the 8+ chemical dimensions to 2D, revealing natural wine groupings. K-means clustering identifies 4 distinct wine style clusters based on chemical composition.



Wine Style Clusters Discovered via Machine Learning

Detailed Interpretation:

Dataset: Combined (6,497 samples)
Total wines analyzed: 6,497
Number of clusters identified: 4
PCA variance explained: PC1=27.8%, PC2=20.5%

WHAT DOES THIS MEAN?

1. PRINCIPAL COMPONENT ANALYSIS (PCA)
   - Reduces 8+ chemical dimensions to 2D for visualization
   - PC1 and PC2 capture 48.4% of chemical variation
   - Wines close together = similar chemical composition
   - Wines far apart = different chemical profiles
2. CLUSTER ANALYSIS (K-MEANS)
   - Algorithm found 4 natural wine groups based on chemistry
   - Each cluster represents a distinct 'wine style'

CLUSTER CHARACTERISTICS:

CLUSTER 0: "LIGHT-BODIED" WINES
Size: 1,784 wines (27.5% of dataset)
Average Quality: 5.64/10
Quality Range: 3 - 9
KEY CHARACTERISTICS:
• LOW alcohol (9.5% vs dataset avg 10.5%)
→ Light-bodied, delicate, refreshing, elegant
• LOWER volatile acidity (0.28 vs 0.34 g/L)
→ Clean, pure fruit expression, fresh
• HIGHER sugar (11.4 vs 5.4 g/L)
→ Off-dry, fruity sweetness, ripe fruit character
WINE STYLE SUMMARY: Delicate

CLUSTER 1: "FULL-BODIED" WINES
Size: 2,937 wines (45.2% of dataset)
Average Quality: 6.09/10
Quality Range: 3 - 9
KEY CHARACTERISTICS:
• HIGH alcohol (11.2% vs dataset avg 10.5%)
→ Full-bodied, rich, warming, powerful structure
• LOWER volatile acidity (0.27 vs 0.34 g/L)
→ Clean, pure fruit expression, fresh
WINE STYLE SUMMARY: Powerful

CLUSTER 2: "SPICY & CRISP, MEDIUM-BODIED" WINES
Size: 737 wines (11.3% of dataset)
Average Quality: 5.78/10
Quality Range: 3 - 8
KEY CHARACTERISTICS:
• MEDIUM alcohol (10.5% vs dataset avg 10.5%)
→ Balanced body, versatile, food-friendly
• HIGHER volatile acidity (0.40 vs 0.34 g/L)
→ More complex, some vinegar character, potentially faulty
• HIGH sulphates (0.73 vs 0.53 g/L)
→ Spicy complexity (black pepper), oak aging potential
• HIGH acidity (9.7 vs 7.2 g/L)
→ Crisp, tart, green apple/citrus notes, bright
• HIGH citric acid (0.46 vs 0.32 g/L)
→ Fresh citrus notes, lemon/lime character, zippy
WINE STYLE SUMMARY: Balanced, Oak-Friendly, Tart, Complex Acidity

CLUSTER 3: "MEDIUM-BODIED" WINES
Size: 1,039 wines (16.0% of dataset)
Average Quality: 5.39/10
Quality Range: 3 - 8
KEY CHARACTERISTICS:

• MEDIUM alcohol (10.2% vs dataset avg 10.5%)
→ Balanced body, versatile, food-friendly
• HIGHER volatile acidity (0.61 vs 0.34 g/L)
→ More complex, some vinegar character, potentially faulty
WINE STYLE SUMMARY: Balanced, Complex Acidity
WHAT DIFFERENTIATES THE CLUSTERS?
The 4 clusters are distinguished primarily by:
1. ALCOHOL CONTENT (Body):
  - Cluster with LOWEST alcohol (0): 9.5%
  - Cluster with HIGHEST alcohol (1): 11.2%
  - This creates Light → Medium → Full body spectrum
2. SULPHATES (Spice/Oak Potential):
  - Cluster 2 shows highest sulphates (0.73 g/L)
  - These wines have more spicy character and oak aging potential
3. ACIDITY (Crispness):
  - Cluster 2 has highest acidity (9.7 g/L)
  - Creates spectrum from Soft/Mellow → Crisp/Tart
4. VOLATILE ACIDITY (Cleanliness):
  - Cluster 1 shows cleanest profile (0.27 g/L)
  - Cluster 3 shows more complexity (0.61 g/L)

KEY INSIGHTS:

  - HIGHEST quality cluster: Cluster 1 (avg quality 6.09)
  - LOWEST quality cluster: Cluster 3 (avg quality 5.39)

CONCLUSION:

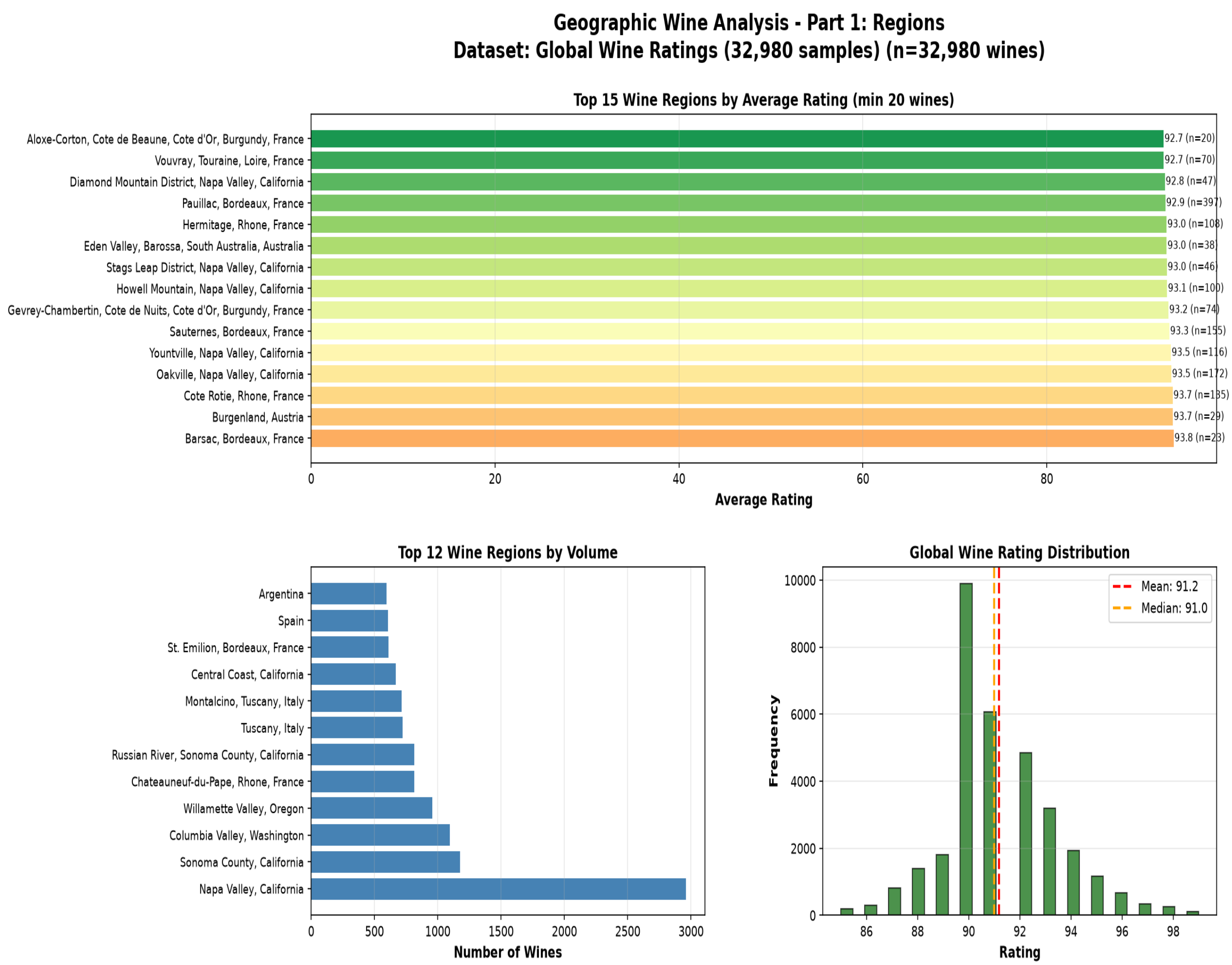
The clustering reveals that wines naturally group into 4 distinct styles based on their chemical composition. Quality ratings show some correlation with these natural groupings, suggesting certain chemical profiles are preferred by wine experts.

13. Geographic Wine Analysis

Analysis of 32,980 wine ratings from around the world, revealing regional quality patterns and production volumes:

- **Top regions by rating:** Highest average scores (minimum 20 wines)
- **Volume leaders:** Most prolific wine-producing regions
- **Rating distribution:** spread of critic scores (note: range is only 85-99, a curated high-end sample, not a true population bell curve)

This data represents professional wine critic scores from major wine regions worldwide, providing insights into relative standing among reviewed wines.



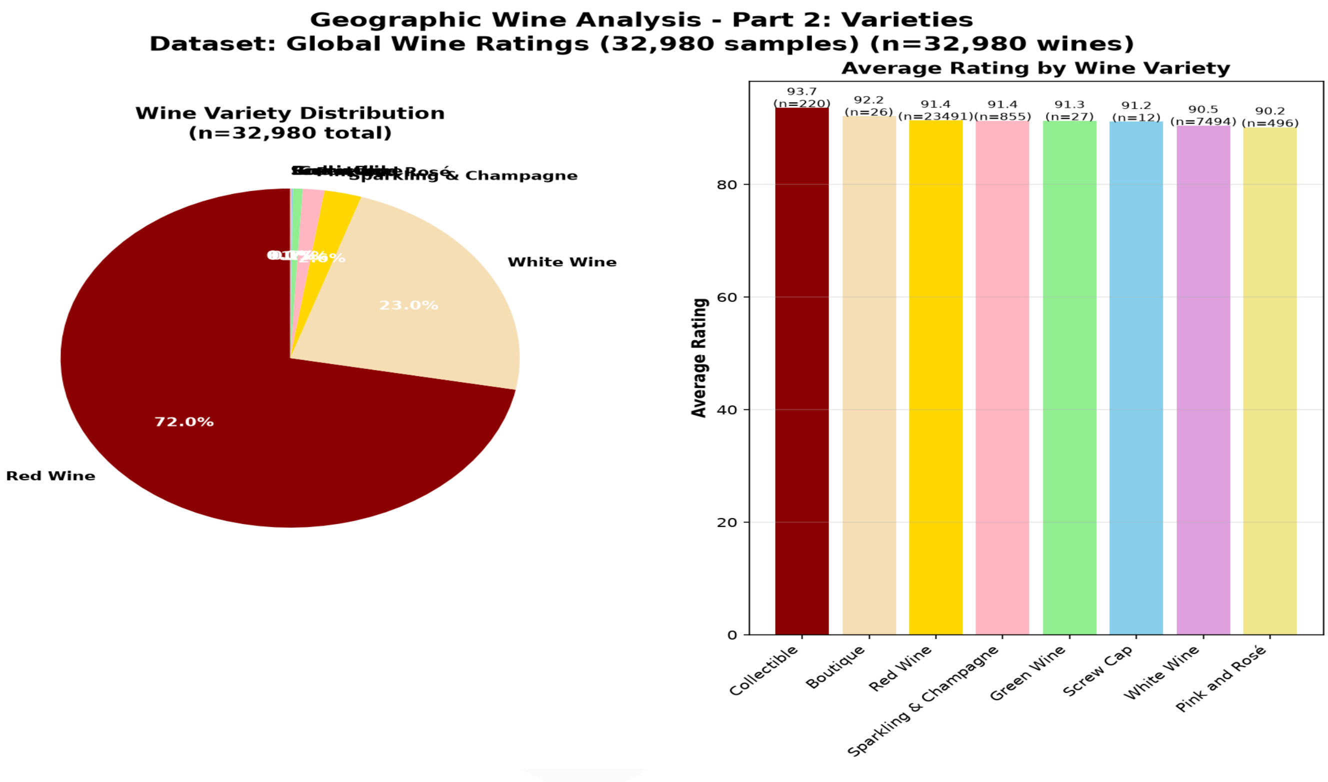
Regional Wine Analysis - Top Performers & Production Leaders

13. Geographic Wine Analysis (continued)

Global wine variety distribution and quality patterns across grape types:

- **Variety breakdown:** Distribution of wine types worldwide
- **Quality by variety:** Average ratings for major grape varieties

This reveals which grape varieties dominate global production and which consistently achieve the highest quality ratings from wine critics.

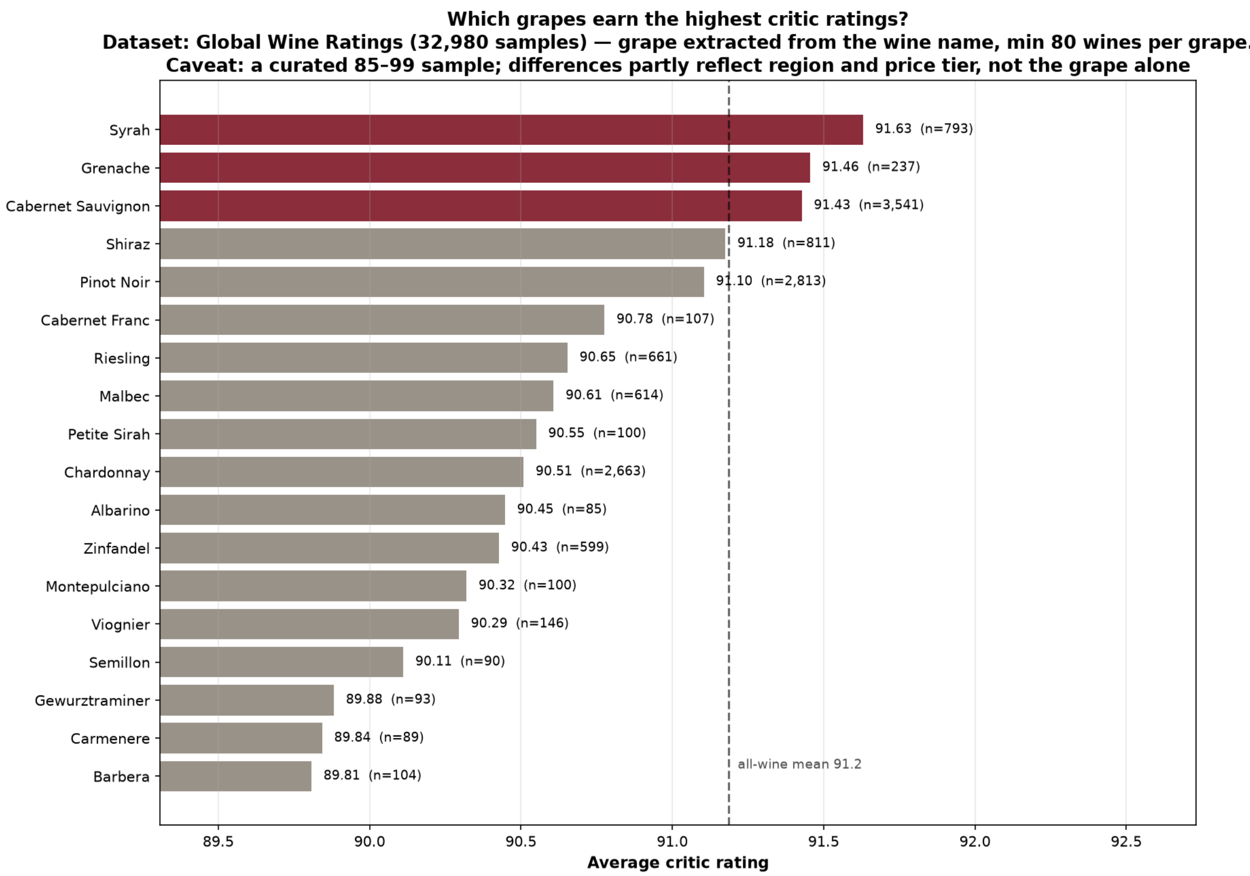


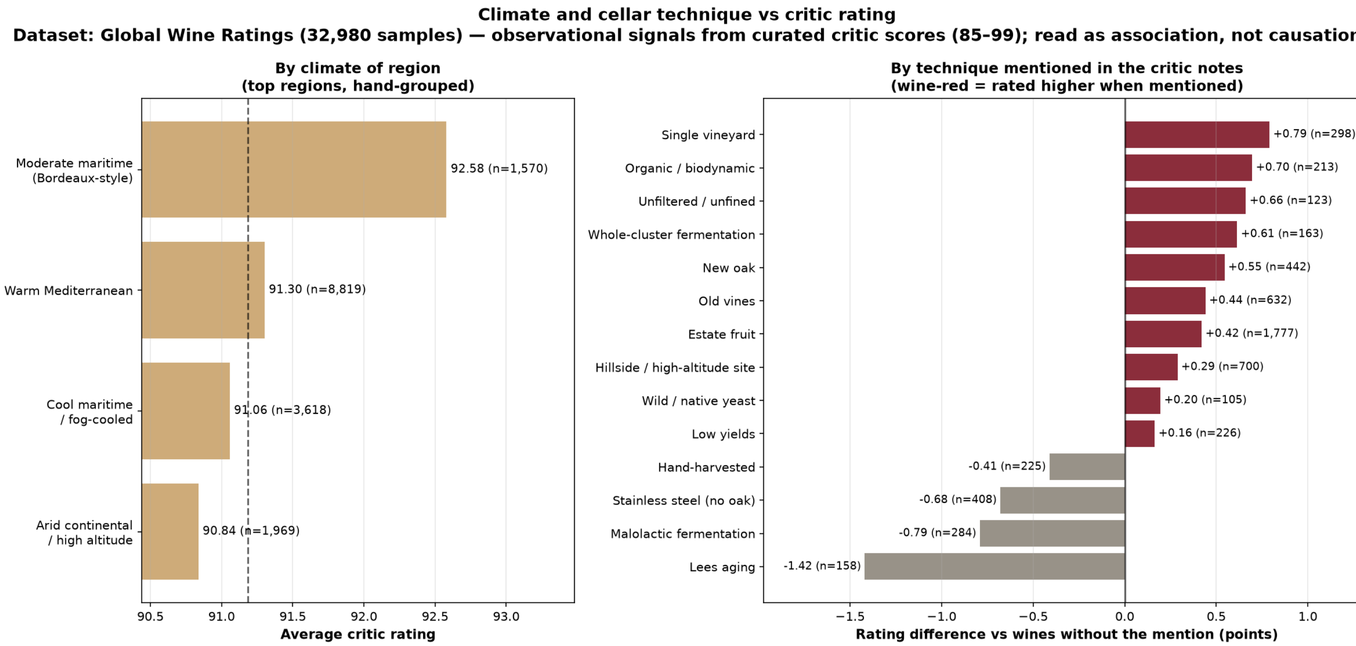
Wine Variety Distribution & Quality Analysis

14. Grapes, Climates & Techniques of the Best Wines

Mining the 32,980-wine ratings dataset for what the highest-rated wines have in common. Grape variety is extracted from the wine name; the top regions are hand-grouped into four coarse climate buckets; and winemaking techniques are matched as keywords in the critic notes (new oak, old vines, single vineyard, whole cluster...).

All of this is observational: the sample is a curated 85-99 critic-score range, the notes are written by marketers, and grape averages partly reflect region and price tier. The deltas are small (under a point) -- read them as consistent associations among already-good wines, never as recipes.





From chemistry to vineyard -- the specifics. Good wines (quality 7+, n=6,497) average **pH 3.23** (middle 50%: 3.12-3.34) and **11.4% alcohol** (10.7-12.4%), implying ~205-230 g/L sugar at harvest. Reaching that fingerprint maps onto measurable viticulture:

Body / alcohol (+0.98 SD): growing-season mean 16-19°C, ~1,400-2,000 growing degree days (warm, Winkler II-III), free-draining poor soils (gravel, schist, galestro), yields capped at ~35-50 hL/ha, fruit picked at 22-25°Brix after a long dry hang.
Low volatile acidity (0.29 vs 0.35 g/L, well under the 0.7 fault line): intact disease-free fruit, prompt cool processing, ferment reds at 26-30°C / whites at 12-18°C, and hold molecular SO2 ≥0.5 mg/L to suppress Acetobacter and Brett.
Fresh acidity / balanced pH: a large 12-18°C diurnal swing from altitude (400-1,500 m) or maritime cooling preserves malic and tartaric acid; pick on acid retention, limit malolactic in crisp whites.
Free SO2 protected zone (31 mg/L free, 110 total): hold free SO2 25-45 mg/L through elevage, top barrels weekly, minimise racking.
Dessert-wine route (sugar >12.6 g/L, volatile acidity ≤0.20): late harvest or noble rot (Botrytis), grapes to 30-40+°Brix, berry-by-berry selection, a cool ferment stopped with sugar intact.

Climate and process figures are standard viticultural references, not quantities measured in this study; pH, alcohol, SO2 and sugar values are this dataset's good-wine averages.