

# Exploratory Data Analysis of Red Wine Quality

## UCI Machine Learning Repository Dataset

Kevin Bell  
Angelie Reyes-Sosa  
Dani Lopez

Fall 2025

# Agenda

- Motivation and dataset overview
- Three guiding questions and hypotheses
- Evidence from descriptive statistics and visuals
- Key anomalies and synthesis of findings
- Future research plan and closing takeaways

# Why This Dataset?

- **Scope:** 1599.00 Portuguese *Vinho Verde* red wines with 11.00 physicochemical features plus quality score.
- **Relevance:** Widely cited benchmark that still leaves room for actionable winemaking insights.
- **Course fit:** Easily satisfies requirements on feature count and observation size.

# Feature Definitions

Feature	Description
Fixed acidity	Non-volatile acids ( $\text{g/dm}^3$ ).
Volatile acidity	Acetic acid ( $\text{g/dm}^3$ ); vinegar aromas.
Citric acid	Adds freshness and structure ( $\text{g/dm}^3$ ).
Residual sugar	Sugar remaining post-fermentation ( $\text{g/dm}^3$ ).
Chlorides	Salt content ( $\text{g/dm}^3$ ).
Free sulfur dioxide	Protects against oxidation ( $\text{mg/dm}^3$ ).
Total sulfur dioxide	Combined bound and free $\text{SO}_2$ ( $\text{mg/dm}^3$ ).
Density	Proxy for sugar/alcohol balance ( $\text{g/cm}^3$ ).
pH	Acidity (unitless).
Sulfates	Potassium sulfate ( $\text{g/dm}^3$ ); antimicrobial.
Alcohol	Ethanol percentage by volume.
Quality	Median sensory rating (3–8 scale).

# Guiding Questions and Analytical Assumptions

- ① Which chemistry attributes best distinguish higher-quality wines?
- ② How do acidity profiles interact with sulfur management across quality tiers?
- ③ Are there latent subgroups that signal distinct wine styles?

## Assumptions and biases

- Treat quality score as approximately continuous for correlation work.
- Lab measurements assumed unbiased; focus is chemistry-centric.

# Working Hypotheses

- Alcohol and sulfates will correlate positively with quality.
- Volatile acidity will correlate negatively with quality.
- Density and residual sugar play limited roles because most wines are dry.

# Summary Statistics Highlights

	Alcohol	Vol. acidity	Citric acid	Sulfates	Density	pH
Mean	10.42	0.53	0.27	0.66	1.00	3.31
Std. dev.	1.07	0.18	0.20	0.17	0.00	0.15

- Alcohol and sulfates show the greatest relative spread → leverage for quality differentiation.
- Volatile acidity variability flags riskier sensory outcomes.

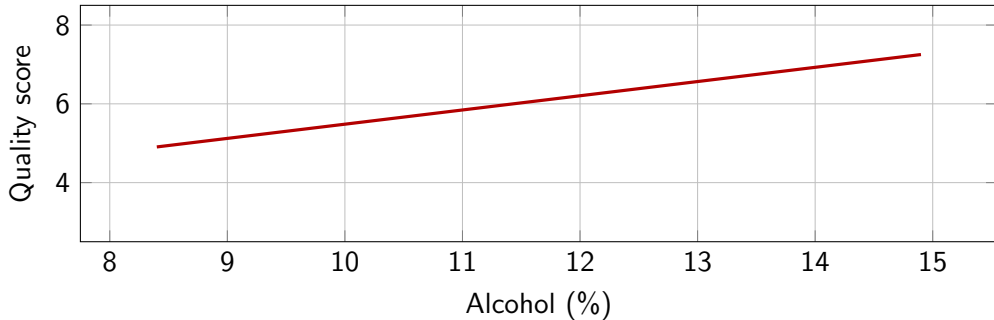
# Means by Quality Tier

Quality	Alcohol	Vol. acidity	Citric acid	Sulfates	Total SO <sub>2</sub>
3	9.96	0.89	0.17	0.57	24.90
4	10.27	0.69	0.17	0.60	36.25
5	9.90	0.58	0.24	0.62	56.51
6	10.63	0.50	0.27	0.68	40.87
7	11.47	0.40	0.38	0.74	35.02
8	12.09	0.42	0.39	0.77	33.44

- Alcohol climbs steadily with quality while volatile acidity drops.
- Total SO<sub>2</sub> peaks at mid-tier quality, pointing to a sweet spot.

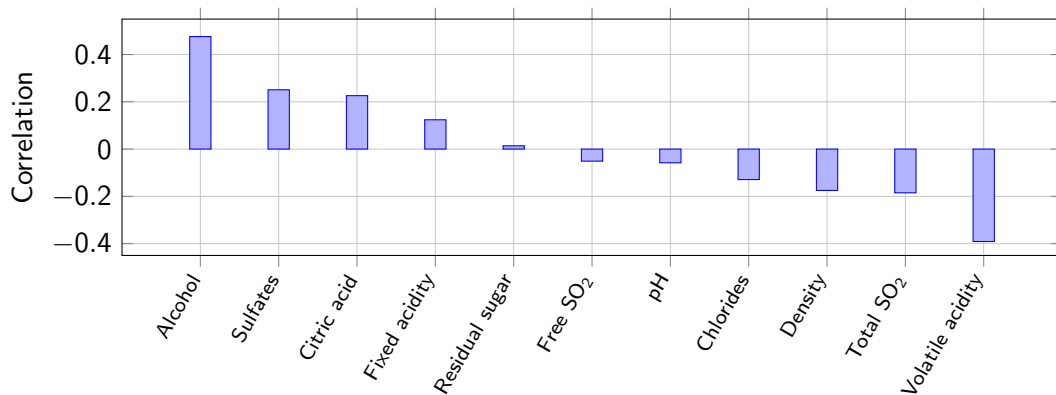


# Alcohol vs. Quality



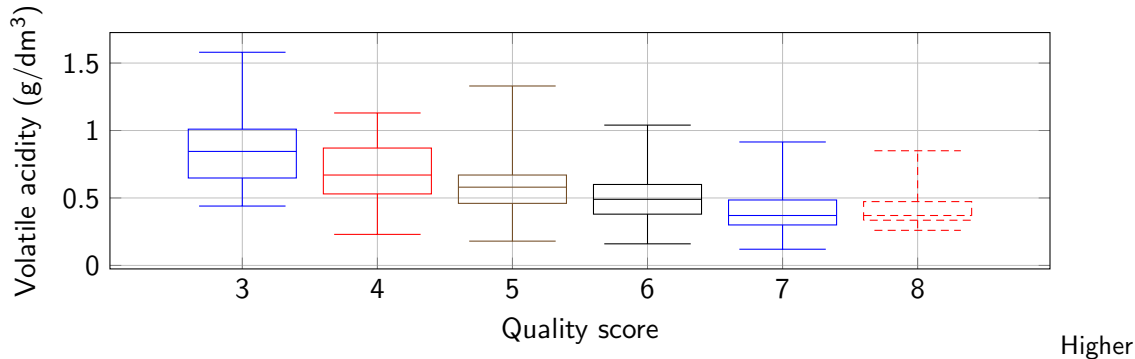
Positive slope reinforces alcohol as a top quality discriminator.

# Feature Correlations with Quality



Alcohol dominates; volatile acidity is the strongest negative signal.

# Volatile Acidity Tightens at Higher Quality



scores cluster at lower, tighter volatility levels.

- Sulfate-heavy outliers: extreme additions sometimes coincide with lower quality despite average upward trend.
- High-chloride, low-quality samples merit lab re-checks and potential process audits.

- High-scoring wines: higher alcohol, moderate sulfates, elevated citric acid, lower volatile acidity.
- Confirms hypotheses on alcohol (positive) and volatile acidity (negative).
- Two apparent segments: mainstream quality (5–6) vs. premium cluster (7–8).

# Future Research and Predictive Plan

- 1 Build predictive models (regularized regression, gradient boosting, tree ensembles) with key interactions.
- 2 Explore segmentation via Gaussian mixtures or density-based clustering.
- 3 Simulate chemistry adjustments with causal inference tools (e.g., propensity score weighting).

Additional questions: missing vintage/producer effects? blending strategies to mitigate deficiencies?

- Met objectives: characterized data, answered core questions, charted next steps.
- Key takeaways: alcohol and balanced sulfates aid quality; volatile acidity detracts; actionable subgroups emerge.
- Thank you! Report and repository contain full reproducible analysis.

Questions?

Team: Kevin Bell, Angelie Reyes-Sosa, Dani Lopez