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
- Descriptive statistics, visuals, and interaction modeling
- Dimensionality reduction, clustering, and anomalies
- 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 (g/dm^3).
Volatile acidity	Acetic acid (g/dm^3); vinegar aromas.
Citric acid	Adds freshness and structure (g/dm^3).
Residual sugar	Sugar remaining post-fermentation (g/dm^3).
Chlorides	Salt content (g/dm^3).
Free sulfur dioxide	Protects against oxidation (mg/dm^3).
Total sulfur dioxide	Combined bound and free SO_2 (mg/dm^3).
Density	Proxy for sugar/alcohol balance (g/cm^3).
pH	Acidity (unitless).
Sulfates	Potassium sulfate (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

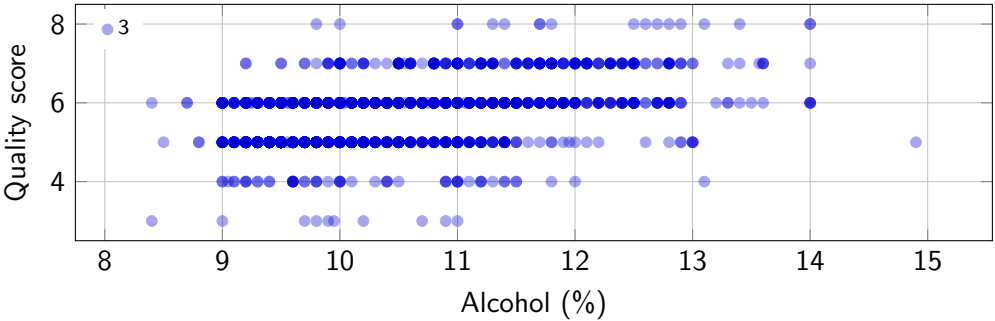
- Alcohol and sulfates show the greatest relative spread $\rightarrow$ leverage for quality differentiation.
- Volatile acidity variability flags riskier sensory outcomes.

Means by Quality Tier

Quality	Alcohol	Vol. acidity	Citric acid	Sulfates	Total SO_2
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

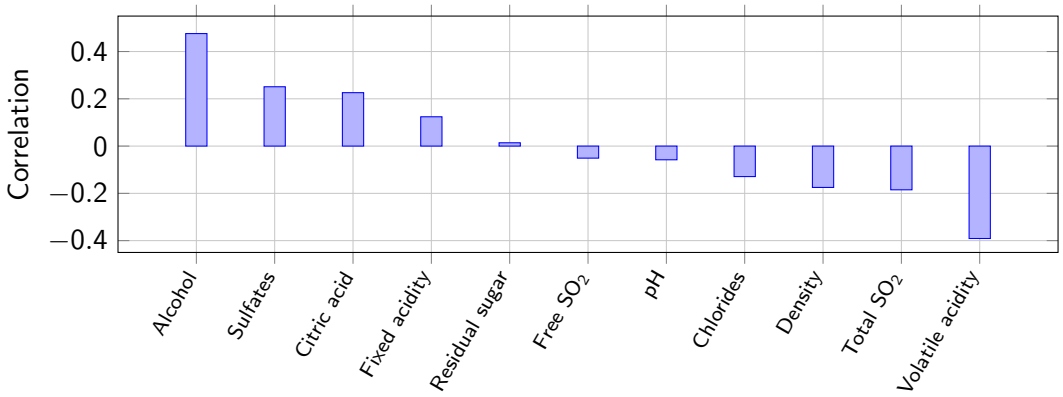
- Alcohol climbs steadily with quality while volatile acidity drops.
- Total SO_2 peaks at mid-tier quality, pointing to a sweet spot.

Alcohol vs. Quality



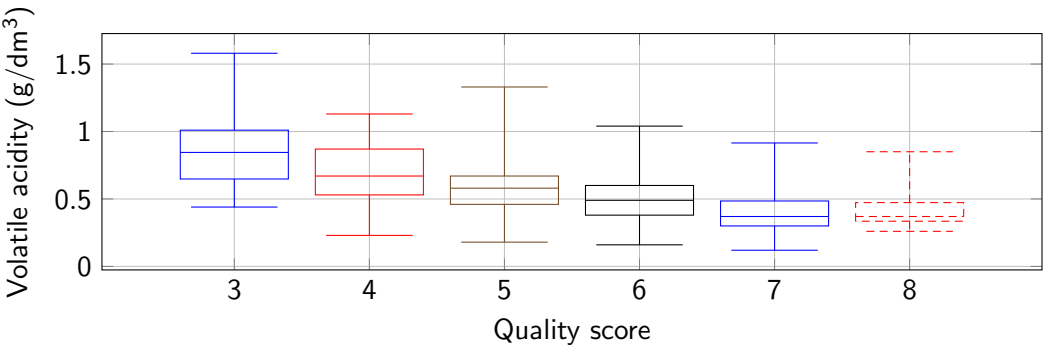
Positive slope reinforces alcohol as a top quality discriminator.

Feature Correlations with Quality



Alcohol dominates; volatile acidity is the strongest negative signal.

Volatile Acidity Tightens at Higher Quality



Higher scores cluster at lower, tighter volatility levels.

Notable Anomalies

- 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.

Interaction Model Insights

- Linear model: $\text{quality} = 5.68 - 1.03\,(\text{vol. acidity}) + 1.20\,(\text{sulfates}) - 0.86\,(\text{vol. acidity} \times \text{sulfates})$.
- Model explains 17.70% of variance; negative interaction captures diminishing sulfate returns.
- High sulfates help only when volatile acidity is under control; guides winemaking trade-offs.

PCA Signals Latent Styles

- First two principal components capture 45.70% of variance (PC1 28.20%, PC2 17.50%).
- PC1 contrasts structural acidity and density against pH and alcohol; PC2 emphasizes sulfur management and residual sugar.
- High-quality wines cluster at higher PC1 scores with controlled volatility, confirming balanced chemistry.

Chemistry Clusters

Cluster	Size	Mean quality	Mean alcohol (%)	Mean vol. acidity	Mean sulfates
Premium	638.00	5.87	10.58	0.42	0.75
Baseline	961.00	5.48	10.32	0.60	0.60

Highest silhouette (0.207) separates a premium profile with higher alcohol and managed volatility from the mainstream cohort.

Synthesis

- High-scoring wines: higher alcohol, moderate sulfates, elevated citric acid, lower volatile acidity.
- Interaction model (R^2=0.177) shows sulfates lose impact when volatile acidity spikes.
- PCA + k-means isolate a premium cluster (638.00 wines, mean quality 5.87) distinct from a baseline group (961.00 wines, 5.48).

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?

Conclusion

- 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.

Thank you!
Team: Kevin Bell, Angelie Reyes-Sosa, Dani Lopez