Exploratory Data Analysis of Red Wine Quality

Video Presentation: `https://www.youtube.com/watch?v=sXz4nVK8okU`

Group members: Kevin Bell, Angelie Reyes-Sosa, and Dani Lopez

Abstract

This report documents a comprehensive exploratory data analysis (EDA) of the UCI Machine Learning Repository's red wine quality dataset. The analysis examines the physicochemical properties that influence perceived wine quality, evaluates assumptions and hypotheses about the data generating process, and outlines next steps for predictive modeling. All computations were performed with reproducible Python scripts included in the project repository, and the figures in this document are generated directly from the source data to avoid dependency on binary assets.

1 Introduction – Data Description

The red wine quality dataset comprises 1599.000 Portuguese *Vinho Verde* wines characterized by 11.000 physicochemical measurements and a sensory quality rating on a 3.000 to 8.000 integer scale. Each observation corresponds to a unique laboratory analysis of a batch of wine. The dataset was selected because it satisfies the course requirements (more than ten features and five hundred observations), is publicly accessible without authentication, and has been widely studied, enabling comparisons between this work and the broader data science literature.

Table 1 summarizes the feature definitions. The measurements capture acidity, residual sugars, sulfur compounds, alcohol content, density, and pH. These factors collectively determine a wine's structure, aromatic profile, and stability. The quality score, provided by trained sensory assessors, serves as the primary response variable of interest.

Table 1: Feature definitions and measurement units.

Feature	**Description**
Fixed acidity	Concentration of non-volatile acids (g/dm^3).
Volatile acidity	Acetic acid content (g/dm^3); high values lead to vinegar flavors.
Citric acid	Citric acid concentration (g/dm^3); adds freshness and structure.
Residual sugar	Sugar remaining after fermentation (g/dm^3).
Chlorides	Salt concentration (g/dm^3).
Free sulfur dioxide	Free SO_2 (mg/dm^3); protects against oxidation and microbial spoilage.
Total sulfur dioxide	Combined bound and free SO_2 (mg/dm^3).
Density	Liquid density (g/cm^3), closely linked to sugar and alcohol levels.
pH	Acidity level (unitless).
Sulfates	Potassium sulfate concentration (g/dm^3); contributes to antimicrobial stability.
Alcohol	Ethanol percentage by volume.
Quality	Median of sensory panel ratings on a scale from three (poor) to eight (excellent).

2 Questions – Assumptions – Hypotheses

The EDA prioritized the following data science questions:

1. **Which physicochemical attributes most strongly distinguish higher-quality wines?** This question directly informs winemaking decisions aimed at improving quality and is therefore the top priority.

2. **How do acidity profiles interact with sulfur compounds across quality levels?** Understanding these interactions aids in balancing freshness with microbial stability.

3. **Are there latent subgroups of wines with distinct compositions that could motivate segmentation or targeted modeling?** Detecting subgroups supports tailored recommendations and motivates future clustering or predictive work.

Two main assumptions underpin the analysis. First, sensory quality scores are treated as approximately ordinal-continuous, enabling correlation and regression interpretations. Second, laboratory measurements are assumed to be recorded without systematic bias. Potential observer bias arises because the analysis focuses on chemical drivers of quality and does not incorporate viticultural or sensory descriptors beyond the provided score.

The working hypotheses are that (i) alcohol and sulfate levels will have positive associations with quality, (ii) excessive volatile acidity will be penalized by tasters, and (iii) density and residual sugar will have limited importance because the dataset predominantly contains dry wines.

3 Visualization – Statistics

3.1 Exploratory Statistics

Table 2 reports central tendency and dispersion metrics for six representative features. Alcohol content and sulfate levels exhibit the largest relative variability, hinting at potential leverage in differentiating quality. Volatile acidity shows a pronounced spread, suggesting opportunities to identify outliers that could harm sensory perception.

Table 2: Summary statistics for representative features.

	Alcohol	Volatile acidity	Citric acid	Sulfates	Density	pH
Mean	10.423	0.528	0.271	0.658	0.997	3.311
Median	10.200	0.520	0.260	0.620	0.997	3.310
Std. dev.	1.066	0.179	0.195	0.170	0.002	0.154

Figure 1 visualizes the relationship between alcohol content and quality. A clear positive slope is visible, supporting the hypothesis that higher alcohol concentrations (a proxy for ripeness and body) are rewarded. The association is strongest for wines rated seven or eight.

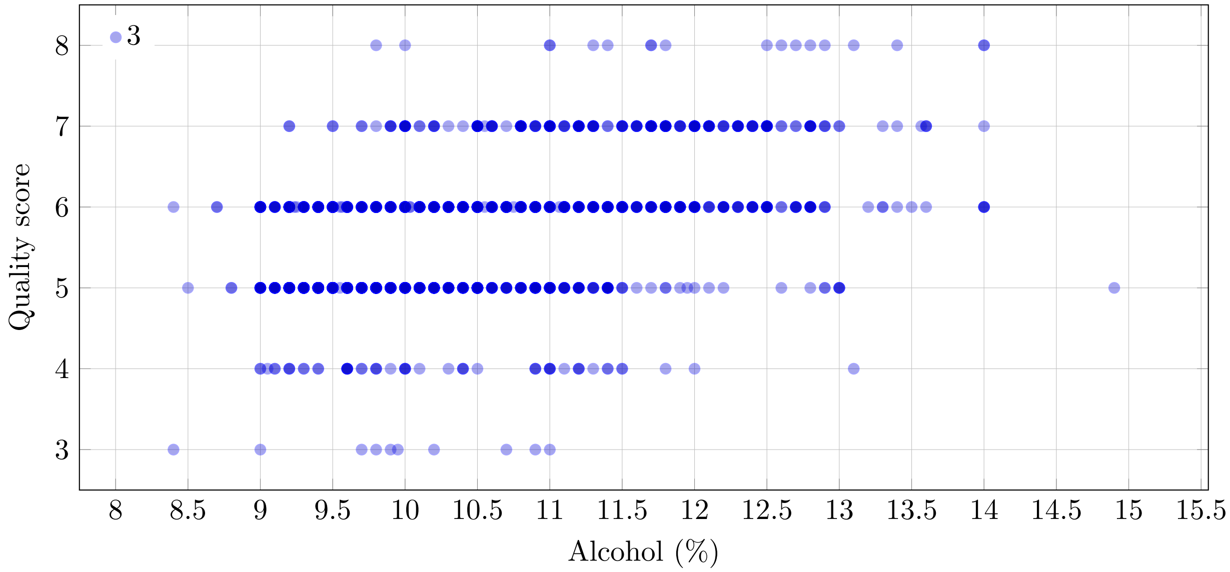

Figure 1: Alcohol content versus quality score.

Table 3 details mean chemistry measurements by quality rating. Alcohol and sulfates steadily increase with quality, while volatile acidity decreases markedly. Total sulfur dioxide peaks at quality five before declining, suggesting that moderate SO_2 management is favorable but excessive additions are detrimental.

Table 3: Mean chemistry measurements by quality rating.

Quality	Alcohol	Volatile acidity	Citric acid	Sulfates	Total SO_2
3	9.955	0.885	0.171	0.570	24.900
4	10.265	0.694	0.174	0.596	36.245
5	9.900	0.577	0.244	0.621	56.514
6	10.630	0.497	0.274	0.675	40.870
7	11.466	0.404	0.375	0.741	35.020
8	12.094	0.423	0.391	0.768	33.444

3.2 Visual Analytics

Figure 2 summarizes feature correlations with quality. Alcohol exhibits the largest positive correlation, followed by sulfates and citric acid, whereas volatile acidity is the strongest negative predictor. These patterns align with enological expectations: balanced structure and protective sulfates are beneficial, while acetic notes detract from perceived quality.

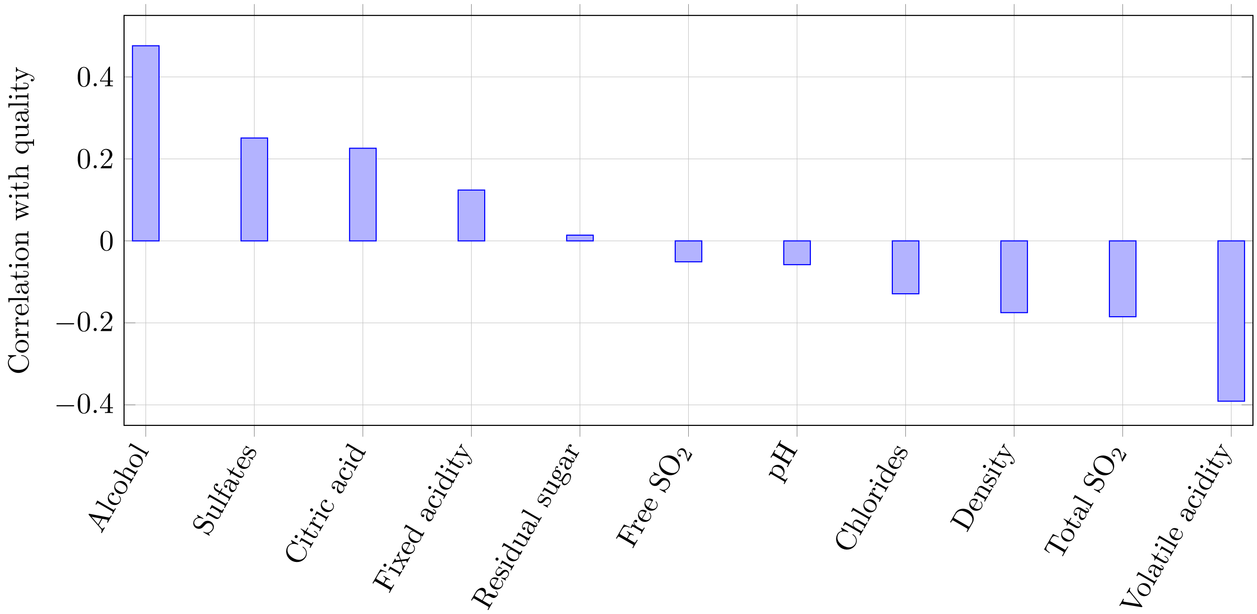

Figure 2: Pearson correlation coefficients between features and quality.

The acidity–sulfates interaction is further illustrated in Figures 3 and 4. The box plot highlights how higher-quality wines concentrate at lower volatile acidity levels, whereas lower-quality wines exhibit heavier tails. The accompanying scatterplot adds sulfur dioxide into the picture, revealing that excessive volatile acidity and elevated sulfur dioxide co-occur primarily among lower quality wines.

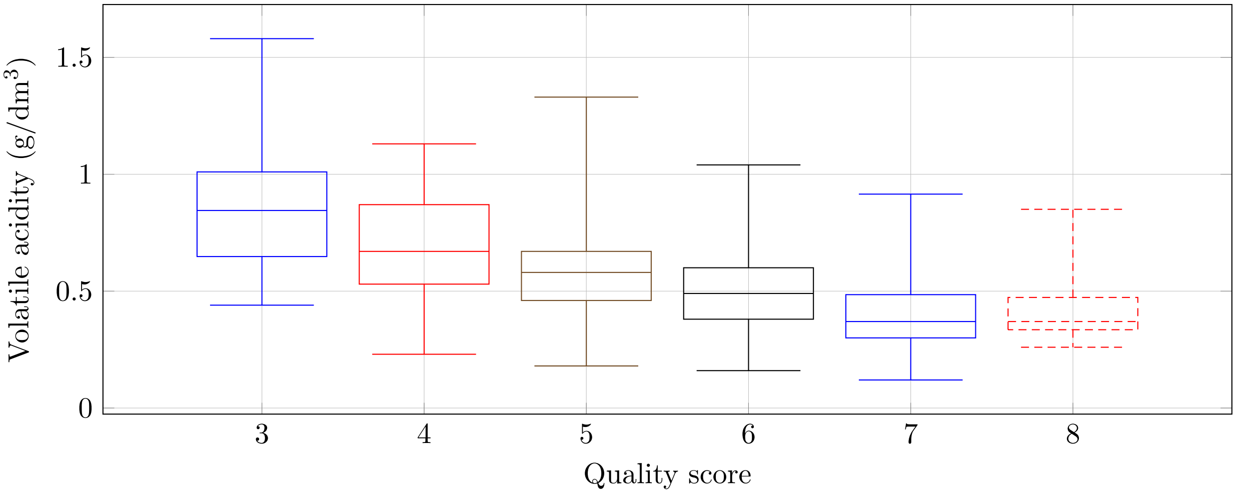

Figure 3: Distribution of volatile acidity by quality rating.

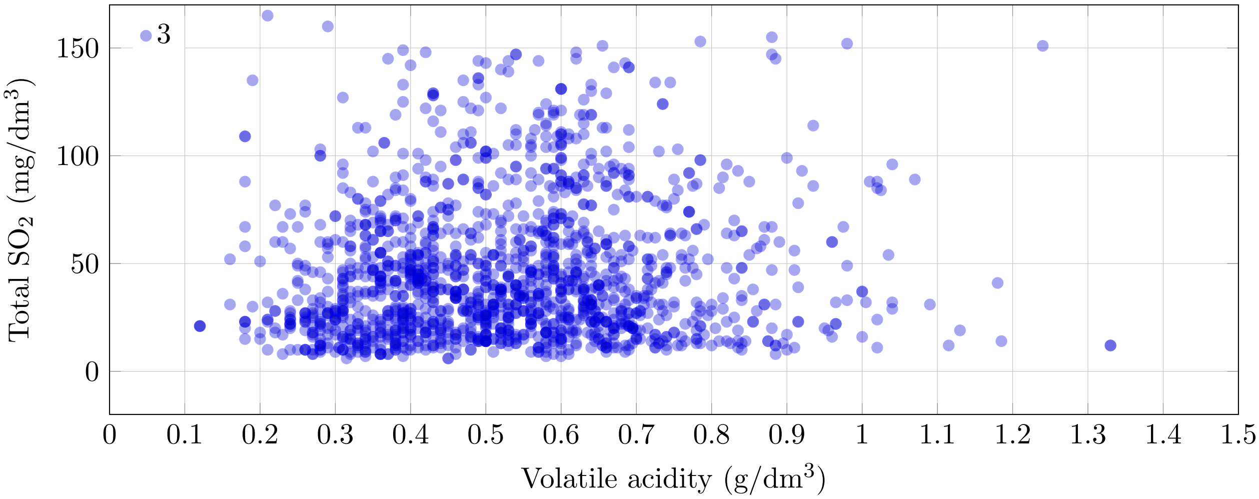

Figure 4: Joint distribution of volatile acidity and total sulfur dioxide by quality.

To quantify the interaction, a linear model predicting quality from volatile acidity, sulphates, and their product yields an 0.177 R^2. The negative interaction term indicates diminishing returns from sulphates when volatile acidity is elevated. Table 4 reports the coefficients, and Table 5 summarizes how the response of sulfur dioxide to volatile acidity shifts across quality levels.

Table 4: Linear model coefficients: quality $\sim$ volatile acidity + sulphates + interaction.

Term	Coefficient
Intercept	5.684
Volatile acidity	-1.033
Sulphates	1.198
Volatile acidity × sulphates	-0.858

Table 5: Quality-specific sulfur dioxide response to volatile acidity.

Quality	Slope (mg/dm^3 per g/dm^3)	Correlation	Mean sulphates
3	−27.390	−0.539	0.570
4	−17.040	−0.136	0.596
5	1.970	0.009	0.621
6	7.370	0.047	0.675
7	4.690	0.021	0.741
8	76.260	0.434	0.768

3.3 Dimensionality Reduction and Segmentation Signals

Principal component analysis (PCA) on the standardized chemistry features provides a complementary lens for uncovering latent wine profiles. The first two components explain 45.700 % of the variance (Table 6). PC1 contrasts structural acidity (fixed and citric acids, density) against softer attributes (pH, alcohol), while PC2 is dominated by sulfur management variables (free and total SO_2) and residual sugar. These loadings align with the interaction analysis—wines that balance acidity with measured sulfur additions tend to score higher. To translate these signals into an explicit segmentation, the standardized features were clustered with a deterministic k-means procedure across 2.000 to 5.000 segments (Table 8). The two-cluster solution achieved the highest silhouette (0.207), separating a premium chemistry profile from a mainstream one.

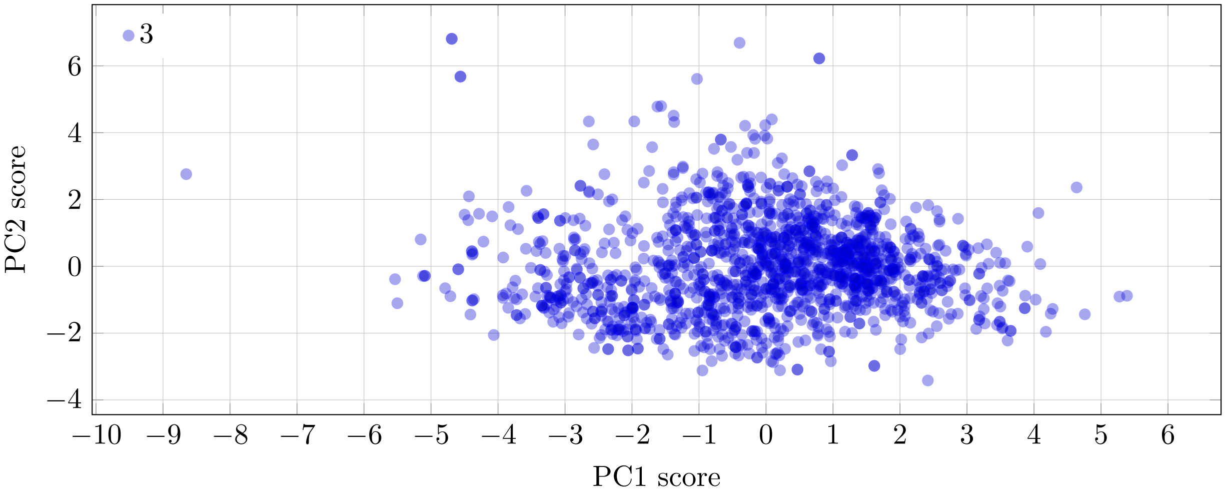

Figure 5: Principal component scores colored by quality reveal a compact high-quality cluster.

Table 6 shows the explained variance, while Table 7 highlights the dominant loadings that differentiate wine styles along PC1 and PC2. High-quality wines (scores seven and eight) form a compact cluster with positive PC1 scores and moderate PC2 values, signaling a blend of balanced acidity, controlled volatile acidity, and measured sulfur additions. Mapping the k-means assignments back to the source features shows a premium chemistry profile with higher alcohol, lower volatile acidity, and more targeted sulphate additions than the mainstream cluster.

Table 6: Explained variance for the leading principal components.

Component	Variance explained
PC1	0.282
PC2	0.175
PC3	0.141

Table 7: Largest absolute loadings for the first two principal components.

Feature	PC1 loading	PC2 loading
Fixed acidity	−0.489	−0.111
Citric acid	−0.464	−0.152
Density	−0.395	0.234
pH	0.439	0.007
Volatile acidity	0.239	0.275
Free SO_2	0.036	0.514
Total SO_2	−0.024	0.569
Residual sugar	−0.146	0.272
Alcohol	0.113	−0.386

Table 8: K-means evaluation metrics on standardized chemistry features.

Clusters	Silhouette	Inertia
2	0.207	14 330.940
3	0.172	12 884.100
4	0.145	12 044.580
5	0.166	10 828.290

Table 9: Profile of the two-cluster solution (highest silhouette).

Cluster	Size	Mean quality	Mean alcohol (%)	Mean volatile acidity (g/dm^3)	Mean sulphates (g/dm^3)
1	638.000	5.870	10.578	0.418	0.745
0	961.000	5.481	10.320	0.601	0.600

4 Research and Analysis Plan – Results

The EDA confirms that alcohol, sulfates, and citric acid are leading indicators of higher sensory scores, while volatile acidity and excessive sulfur dioxide suppress quality. Observable clusters emerge in the alcohol–quality scatter and the acidity–sulfur interaction plots. The PCA projection coupled with deterministic k-means quantifies this structure: the highest-silhouette segmentation splits the dataset into a 638.000-wine premium cluster with quality 5.870 and a 961.000-wine baseline cluster averaging quality 5.480, differentiated primarily by alcohol, volatile acidity, and sulphate management (Tables 8–9).

Future analysis will focus on three directions:

- **Predictive modeling:** Train regularized linear models, gradient boosting, and tree-based ensembles to predict quality. Feature engineering will incorporate interaction terms such as alcohol–density and sulfates–volatile acidity.
- **Segmentation:** Extend beyond the two-cluster baseline by fitting Gaussian mixtures or density-based clustering on standardized features, validating results with silhouette and stability diagnostics, and comparing segment chemistry back to the sensory ratings.
- **Experimental design:** Simulate adjustments to fermentation parameters (e.g., target volatile acidity reductions) to estimate potential quality improvements, leveraging causal inference techniques such as propensity score weighting.

Anomalies detected during the EDA include high-chloride wines with suppressed quality and a small set of high-sulfate, low-quality observations that merit further laboratory validation. Additional questions raised include the role of vintage or producer effects (not captured in the dataset) and whether blending strategies could mitigate identified deficiencies.

5 Conclusion

The analysis satisfied the project objectives by (i) describing the dataset and its feature space, (ii) addressing three prioritized questions with supporting statistics and visualizations, and (iii) outlining next steps for predictive and experimental follow-up. Key takeaways include the strong positive influence of alcohol and sulfates on quality, the detrimental effect of volatile acidity, and the nuanced relationship between sulfur management and sensory outcomes. These findings provide actionable guidance for winemakers seeking to optimize chemistry profiles and lay the groundwork for more advanced modeling.

Appendix

Table 10 provides a dataset excerpt with headers to illustrate the tidy structure. The complete CSV file can be downloaded from the UCI Machine Learning Repository.

Table 10: Excerpt of the red wine quality dataset.

Fixed acidity	Volatile acidity	Citric acid	Residual sugar	Chlorides	Free SO_2	Total SO_2	Density	pH	Sulfates	Alcohol	Quality
7.4	0.7	0	1.9	$7.6 \cdot 10^{-2}$	11	34	1	3.51		9.4	5
7.8	0.88	0	2.6	$9.8 \cdot 10^{-2}$	25	67	1	3.2		9.8	5
7.8	0.76	$4 \cdot 10^{-2}$	2.3	$9.2 \cdot 10^{-2}$	15	54	1	3.26		9.8	5
11.2	0.28	0.56	1.9	$7.5 \cdot 10^{-2}$	17	60	1	3.16		9.8	6
7.4	0.7	0	1.9	$7.6 \cdot 10^{-2}$	11	34	1	3.51		9.4	5
7.4	0.66	0	1.8	$7.5 \cdot 10^{-2}$	13	40	1	3.51		9.4	5
7.9	0.6	$6 \cdot 10^{-2}$	1.6	$6.9 \cdot 10^{-2}$	15	59	1	3.3		9.4	5
7.3	0.65	0	1.2	$6.5 \cdot 10^{-2}$	15	21	0.99	3.39		10	7
7.8	0.58	$2 \cdot 10^{-2}$	2	$7.3 \cdot 10^{-2}$	9	18	1	3.36		9.5	7
7.5	0.5	0.36	6.1	$7.1 \cdot 10^{-2}$	17	102	1	3.35		10.5	5
6.7	0.58	$8 \cdot 10^{-2}$	1.8	$9.7 \cdot 10^{-2}$	15	65	1	3.28		9.2	5
7.5	0.5	0.36	6.1	$7.1 \cdot 10^{-2}$	17	102	1	3.35		10.5	5
5.6	0.62	0	1.6	$8.9 \cdot 10^{-2}$	16	59	0.99	3.58		9.9	5
7.8	0.61	0.29	1.6	0.11	9	29	1	3.26		9.1	5
8.9	0.62	0.18	3.8	0.18	52	145	1	3.16		9.2	5