Wine Quality Analysis Project

End-to-end data science case study exploring physicochemical drivers of wine quality across +5,000 sample records of red and white wine. We identify chemical correlations, segment wines by profile, and set benchmarks for decision-making on wine quality production.

Project Objectives

Analyze

Uncover chemical features that impact expert and consumer quality ratings.

Segment

Cluster wines by chemical profile and relate segments to quality scores.

Benchmark

Define actionable chemical thresholds winemakers can target.

Key Datasets

Datasets from the UCI Machine Learning Repository containing 12 physicochemical features and expert quality ratings for thousands of red and white wines.

- Red Wine: winequality-red.csv
- White Wine: winequality-white.csv

Dataset Overview

Dataset Features

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality

For practical purposes, the original **quality** variable was renamed to **consumer_score**. Although this score represents the median evaluation of expert sommeliers, it is treated in this study as a proxy for aggregated consumer preference, following the assumption that expert quality assessments are aligned with general consumer perception.

Core Techniques & Workflow

Data Cleaning & EDA

pandas, numpy, seaborn, matplotlib — distributions, missing values, outliers.

Correlation & Insights

Feature relationships analyzed in chemistry context.

Dimensionality Reduction

PCA to summarize chemistry into interpretable axes.

Clustering

k-Means, silhouette, elbow method to segment wines.

Data Cleaning

Initial examination of the wine datasets revealed a high level of data integrity, with key observations regarding missing values and duplicate entries.

No Missing Values (NaN)

Both red and white wine datasets were complete, with no NaN values detected across any chemical features or quality scores.

Duplicate Entries Identified

While no missing data was found, a significant number of duplicate rows were present: **240 in red wine** and **937 in white wine** datasets.

- These duplicates are not indicative of data entry errors but rather represent instances of wines sharing identical chemical profiles and consumer quality ratings.
- Given that these profiles are valid and contribute to the overall statistical representation of wine characteristics, they were intentionally retained in the datasets.

Exploratory Data Analysis (EDA)

Quality Score Distribution

Red Wine

- Count of 1599 testing wine records (rows).
- ~83% score 5–6 (acceptable to good).
- Only ~1% reach very high scores (7–9).

White Wine

- Count of 4898 testing records (rows).
- ~75% score 5–6 (acceptable to good).
- 4% reach 7–9 (slightly higher than red).

Correlation With Consumer Score Analysis (Spearman)

Spearman correlations reveal monotonic relationships between chemistry and consumer scores — useful for ordinal score interpretation.

Red Wine — Key Correlates

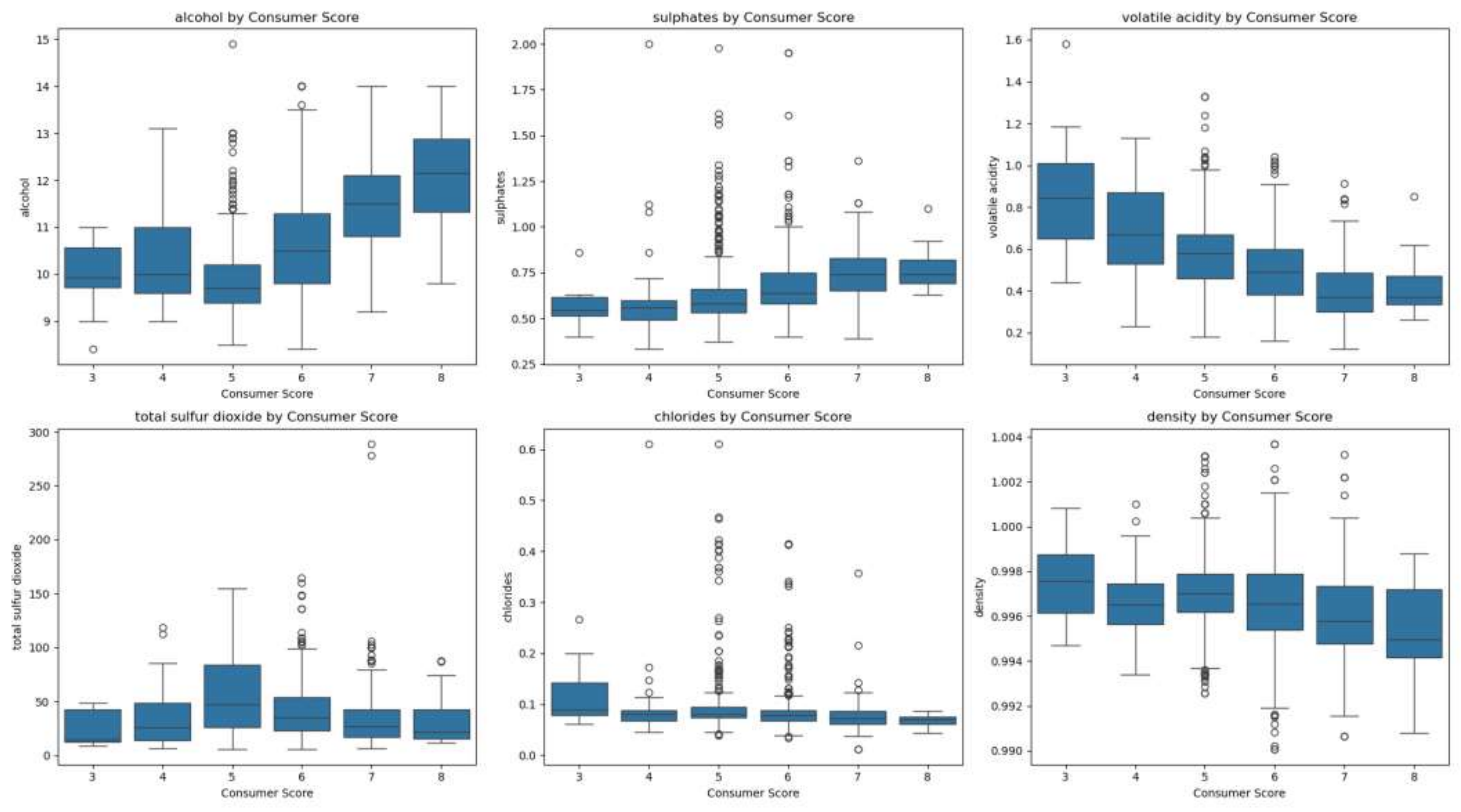
Alcohol & sulphates: positive with consumer score. Volatile acidity, chlorides, total SO_2: negative.

White Wine — Key Correlates

Alcohol & pH: positive. Density, chlorides, total SO_2: strong negatives.

Interpretation ties chemistry to perception: alcohol often signals ripeness and body; volatile acidity and chlorides associate with lower perceived quality.

Distributions of Highly Correlated Variables with Consumer Scores



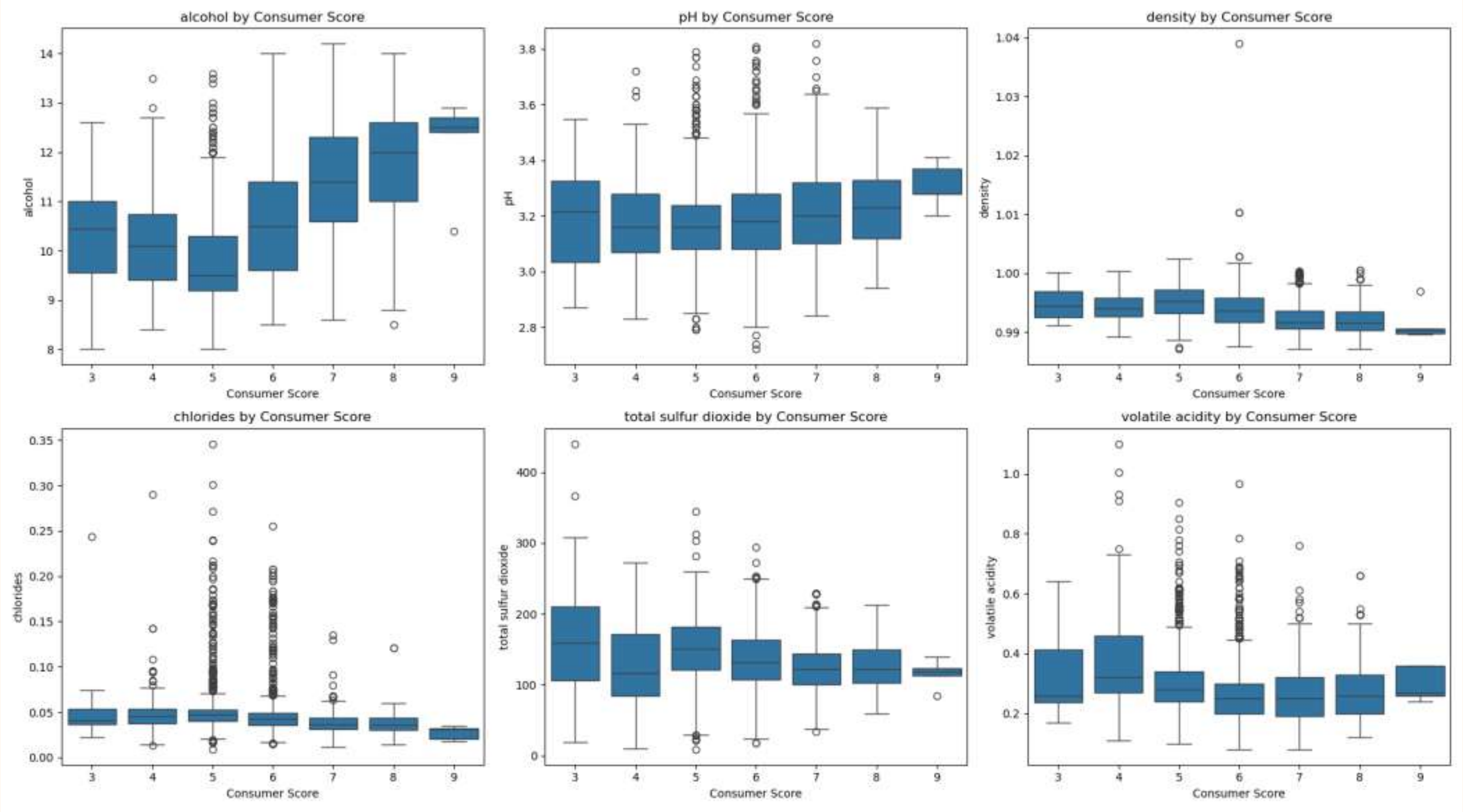
•

Boxplots Red:

Alcohol and sulphates increase with score; volatile acidity, chlorides, density decrease as quality rises.

Distributions of Highly Correlated Variables with Consumer Scores

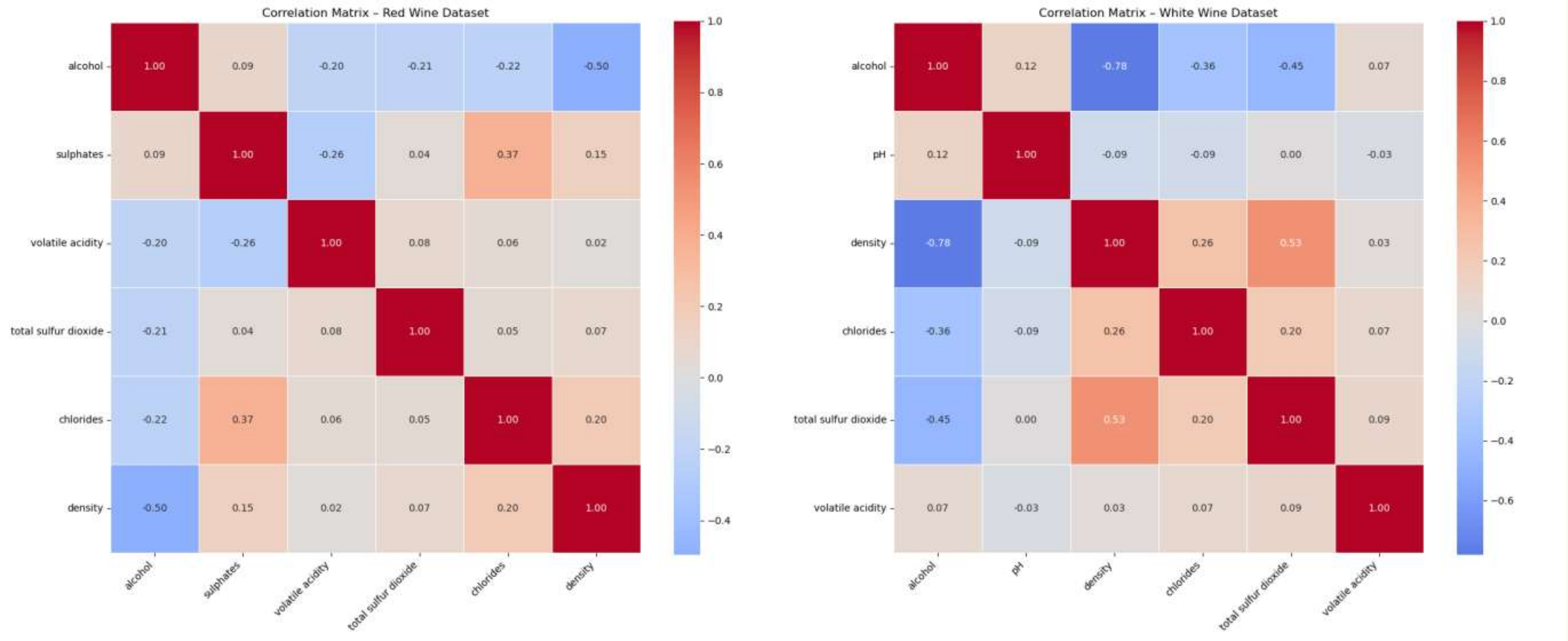
•



Boxplots White:

Alcohol and pH rise with higher scores; density and chlorides peak in low-quality wines.

Correlation Between Chemical Variables



Red:

Alcohol-negative relationship with most variables, while density keep positive trend with minerals and ions.

White:

Even negative relationship with alcohol, while density keep positive trend with minerals and ions.

Correlation Analysis & Chemical Summary Insights

Variable	Typical Correlation	Explanation (Red & White)
Alcohol	Strong Negative (density)	Ethanol lowers density; proxy for ripeness
Sulphates	Slight Positive (chlorides/density)	Added for stabilization; tracks with mineral ions
Volatile Acidity	Negative (alcohol/sulphates)	Indicates microbial/fermentation stress
Total SO_2	Weak	Reflects winemaker additions more than chemistry
Chlorides	Negative	Reflects soil/mineral/underripeness
pH (White only)	Weak/Complex	Controlled via acidification in white wines
Density	Strong Negative (alcohol)	Alcohol and dissolved solids dominate density in whites

Top Associated Features with Quality score by Wine Type

Red Wine

- High Alcohol
- High Sulphates
- Low Volatile Acidity
- Low Chlorides
- High Citric Acid

White Wine

- High Alcohol
- High pH
- Low Density
- Low Chlorides
- Low Total Sulfur Dioxide

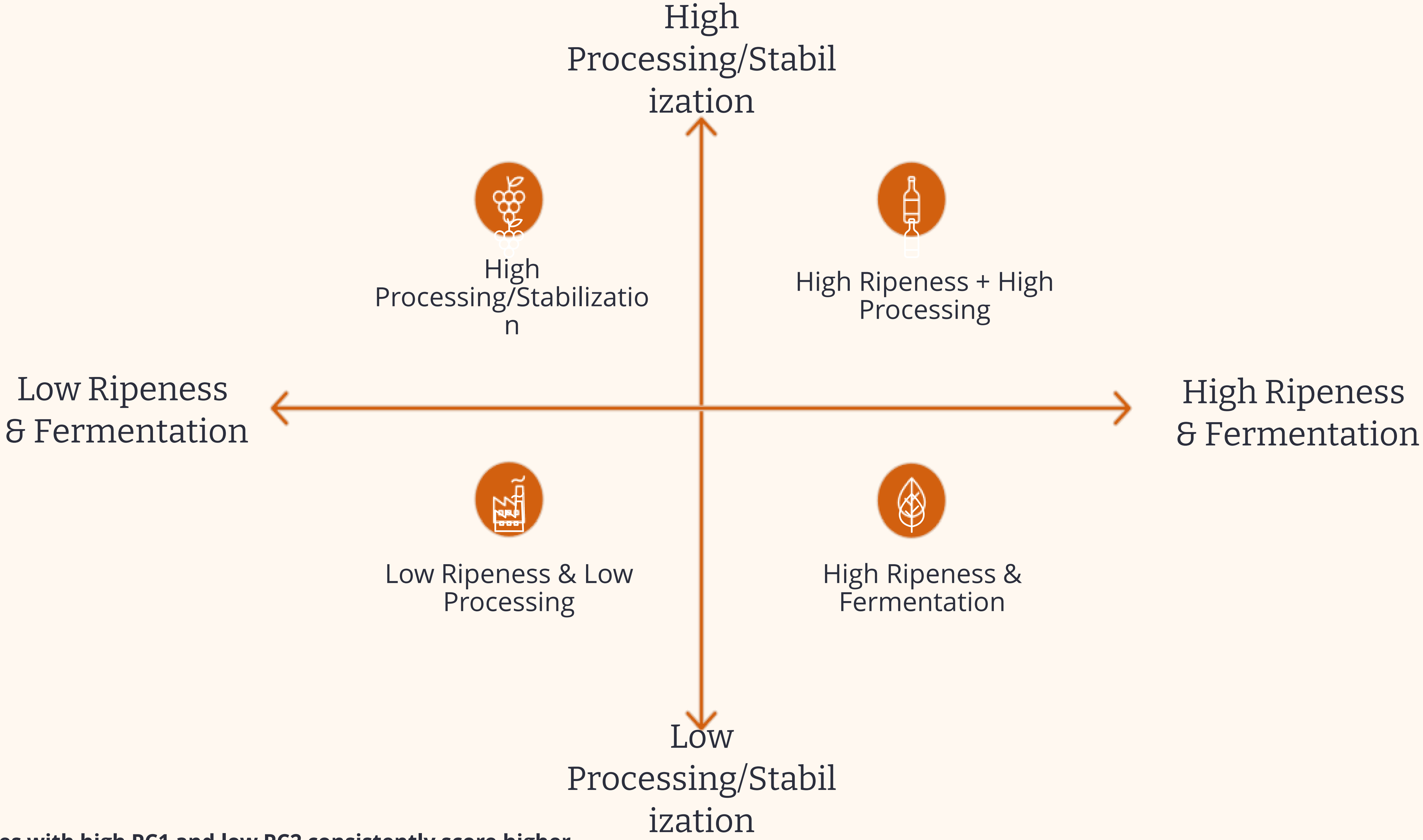
Dimensionality Reduction - PCA Interpretation

PC1 — Ripeness & Fermentation

High alcohol, low density, high citric acid & sulphates, low volatile acidity → higher quality.

PC2 — Processing/Stabilization

High sulfur dioxide, more residual sugar, higher density → processing/stability axis.



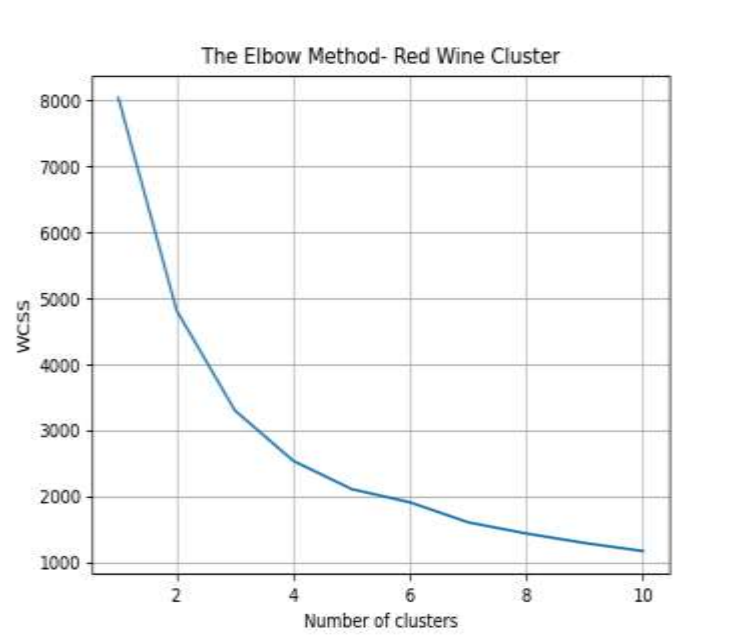
Wines with high PC1 and low PC2 consistently score higher.

K-Means Clustering Methodology

To determine the optimal number of clusters for both the red and white wine datasets, unsupervised K-Means clustering analysis was conducted. This involved employing two widely recognized methods:
- Elbow method.
- Silhouette analysis.

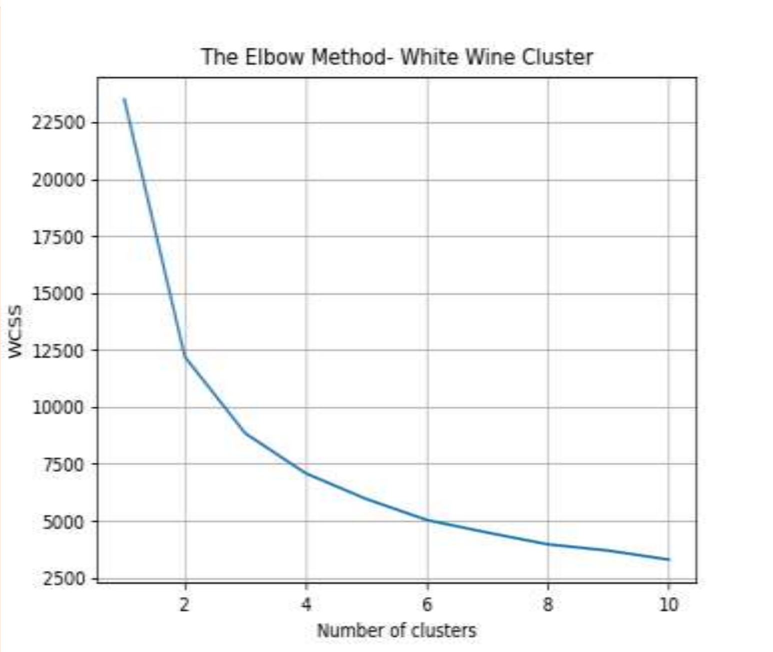
Elbow Method

Identifies the "elbow point" in the WCSS curve where adding more clusters provides diminishing returns, indicating the optimal number of clusters.



Silhouette Analysis

Measures cluster cohesion and separation, with scores ranging from -1 to 1. Higher values indicate better-defined and well-separated clusters.



K-Means Clustering Results

Based on the Elbow method and silhouette analysis, 3 clusters were identified as optimal for both datasets.

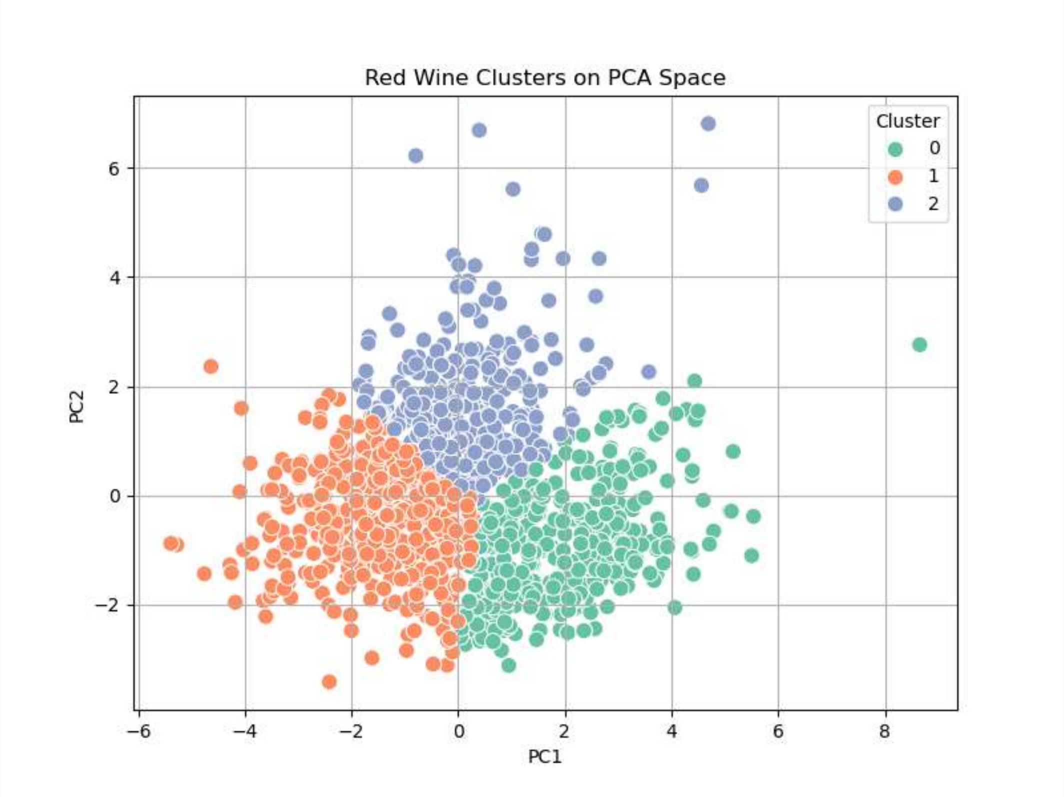
Red Wine Clusters

3 clusters with Silhouette Score of 0.377.

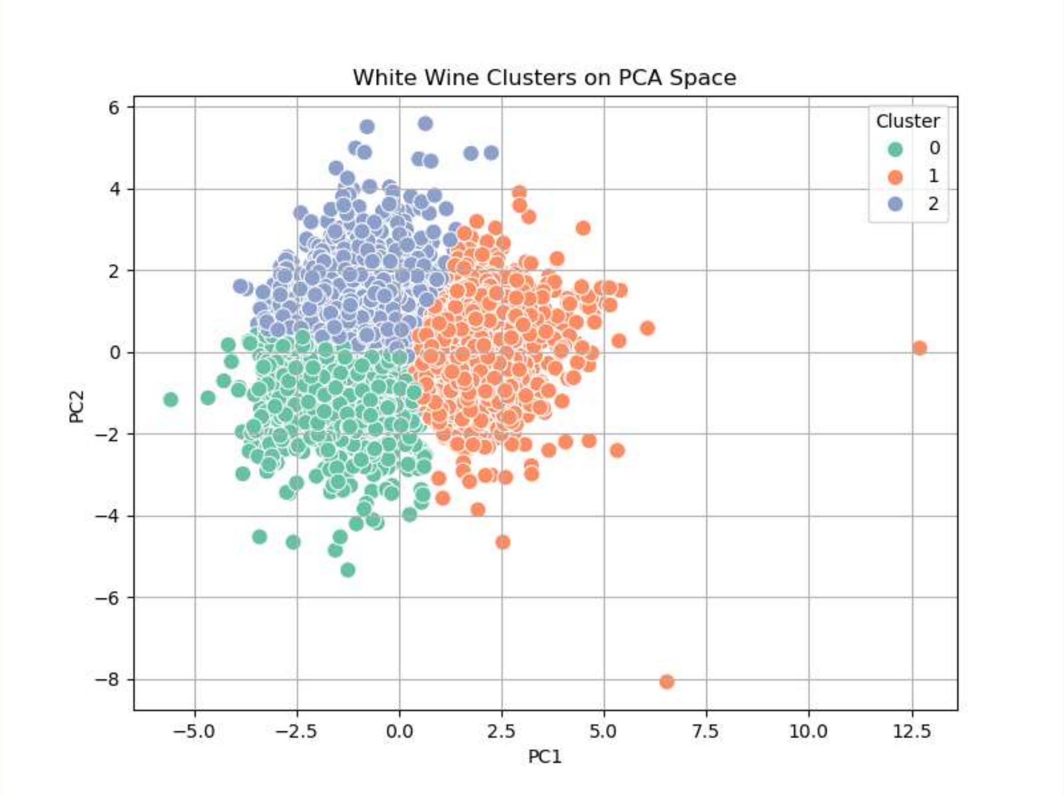
White Wine Clusters

3 clusters with Silhouette Score of 0.378.

Red Wine Clusters Plot



White Wine Clusters Plot



Cluster Benchmarks (Top Clusters)

Red Cluster #0

Highest mean score 5.95, and highest distribution of quality wines (score above 7) with 118 register samples. Benchmarks: Alcohol > 10.6%, Sulphates > 0.75 g/L, VA < 0.41 g/L, Chlorides < 0.10 g/L.

White Cluster #0

Highest mean score 6.17, and highest distribution of quality wines (score above 7) with 169 register samples. Benchmarks: Alcohol > 11.2%, Sulphates > 0.51 g/L, VA < 0.28 g/L, Chlorides < 0.04 g/L.

Actionable Chemical Targets

Practical thresholds derived from clusters 0 to guide winemaking decisions:

Variable	Red Benchmark	White Benchmark
Alcohol (%)	> 10.6	> 11.2
Sulphates (g/L)	> 0.75	> 0.51
Citric Acid (g/L)	> 0.48	> 0.28
Volatile Acidity (g/L)	< 0.41	< 0.28
Chlorides (g/L)	< 0.10	< 0.04
pH	> 3.18	> 3.31
Density	< 0.9978	< 0.9920
Total SO_2 (mg/L)	< 30	< 123
Residual Sugar (g/L)	< 2.6	< 3.4

Thank You

Thank you for your attention to this analysis. Your engagement and interest are greatly appreciated!

For a more detailed exploration of this analysis, please visit my GitHub repository: github.com/RenatoMateo

Contact Information

Name

Renato Silva

+971 58 506 5918.

rmsilvap@bu.edu

LinkedIn

linkedin.com/in/renato-silva-portilla/