

Hypothesis Testing with Bootstrap & Confidence Intervals in R

Discover how to leverage bootstrap methods and confidence intervals in R for robust statistical inference with wine quality data.

J by Jemael Nzihou





Topic Overview

Hypothesis Testing

Learn statistical methods to test assumptions about population parameters using sample data.

Bootstrap Methods

Master resampling techniques that create repeated samples with replacement from original dataset.

Confidence Intervals

Understand how to estimate parameter ranges with specified probability of containing true value.



Key Objectives



Perform Hypothesis Testing

Apply bootstrap methods to test statistical hypotheses without parametric assumptions.



Apply Resampling Techniques

Use data-driven simulation to estimate population parameters with greater precision.



Compute Confidence Intervals

Generate and interpret statistical intervals for robust inferential analysis.

Dataset Overview

Portuguese Vinho Verde

The dataset contains physicochemical properties and quality ratings from the UCI Machine Learning Repository.

Each wine sample has multiple attributes plus a quality score between 0-10.

Key Features

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- pH
- Sulphates
- Alcohol content
- Quality score

R Programming Environment



Powerful Syntax

Elegant coding style with specialized functions for statistical analysis.



Visualization Tools

Create professional graphics with ggplot2 and other specialized packages.



Data Manipulation

Efficiently wrangle datasets using tidyverse libraries.



Statistical Libraries

Access comprehensive collection of statistical methods and tests.



Key Insight



Bootstrap Resampling

Generate thousands of resamples with replacement from original data.



Estimate Mean Differences

Calculate alcohol content differences between high and low quality wines.



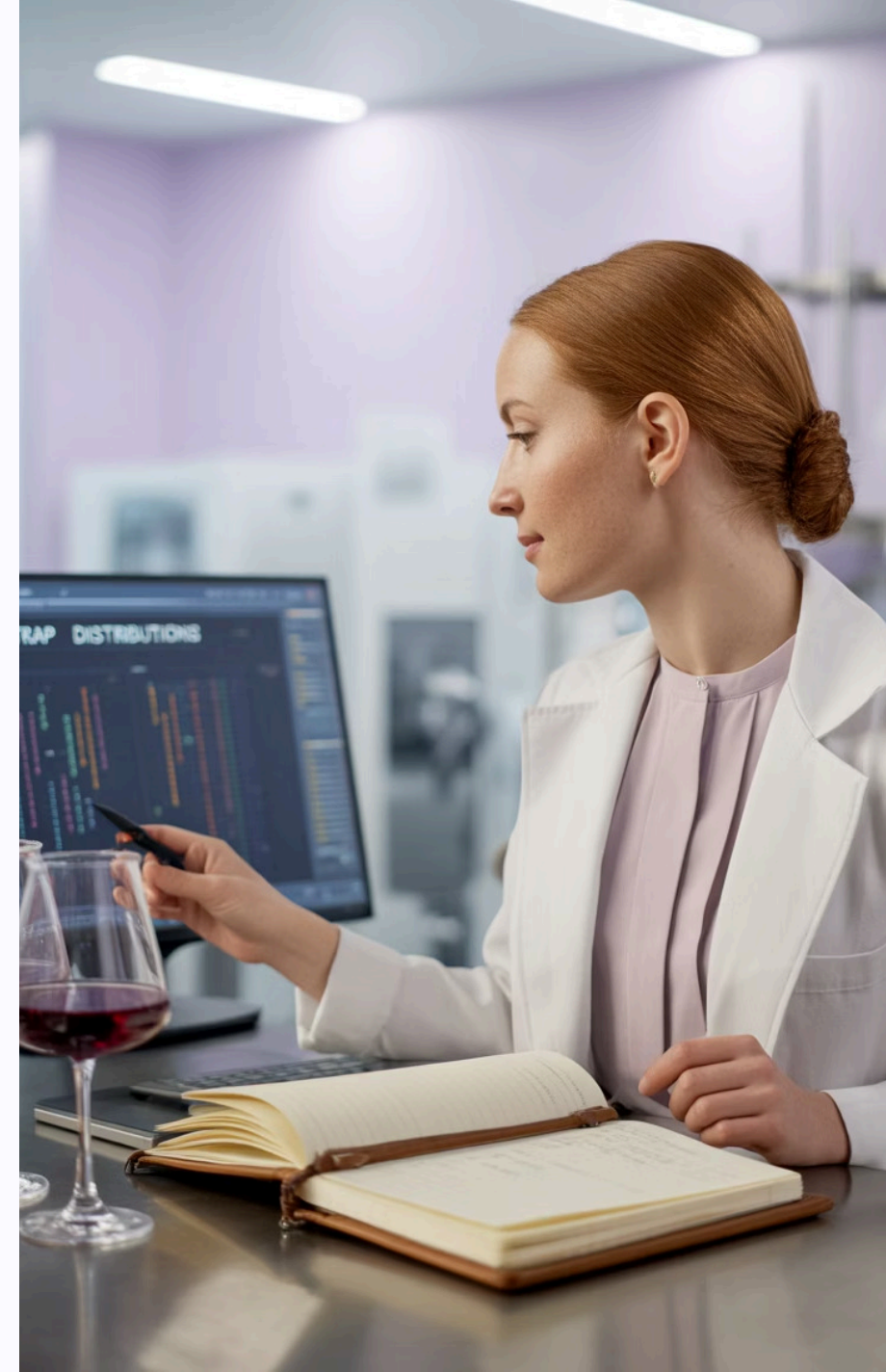
Compute Confidence Intervals

Determine 95% confidence interval for the difference in means.

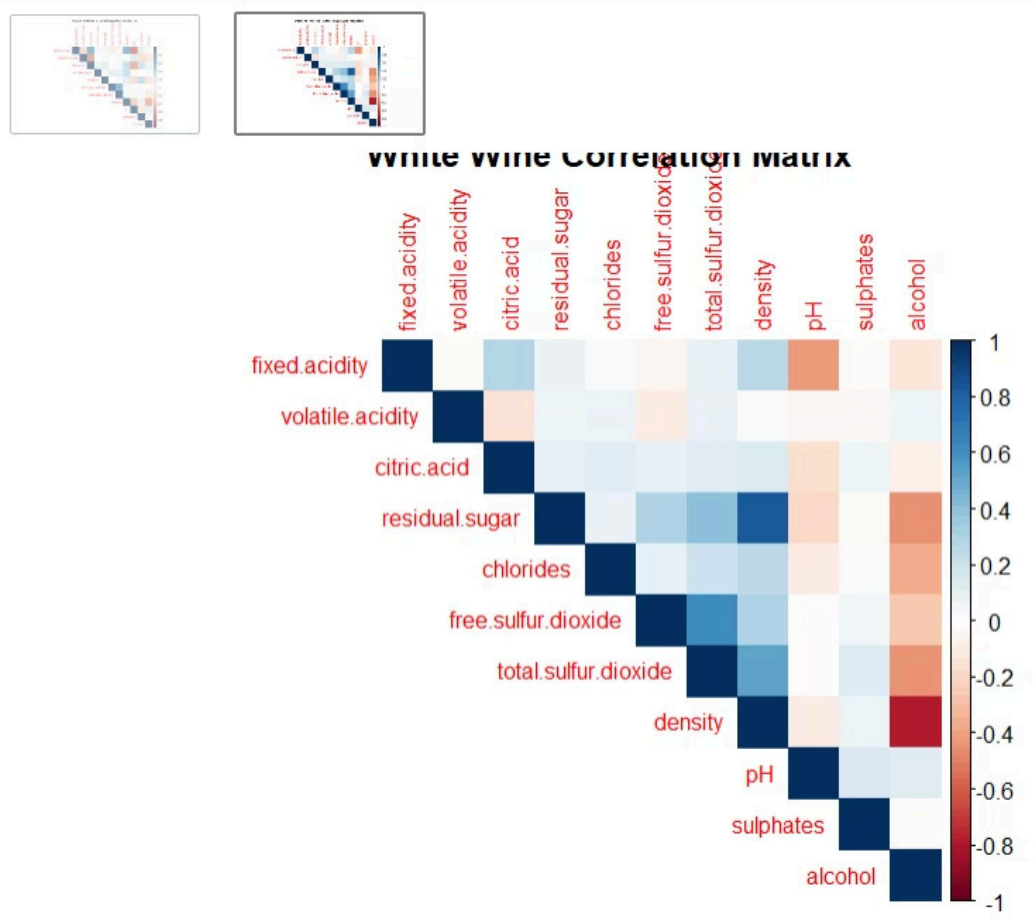


Cross-Validate Results

Compare bootstrap results with traditional two-sample t-test.

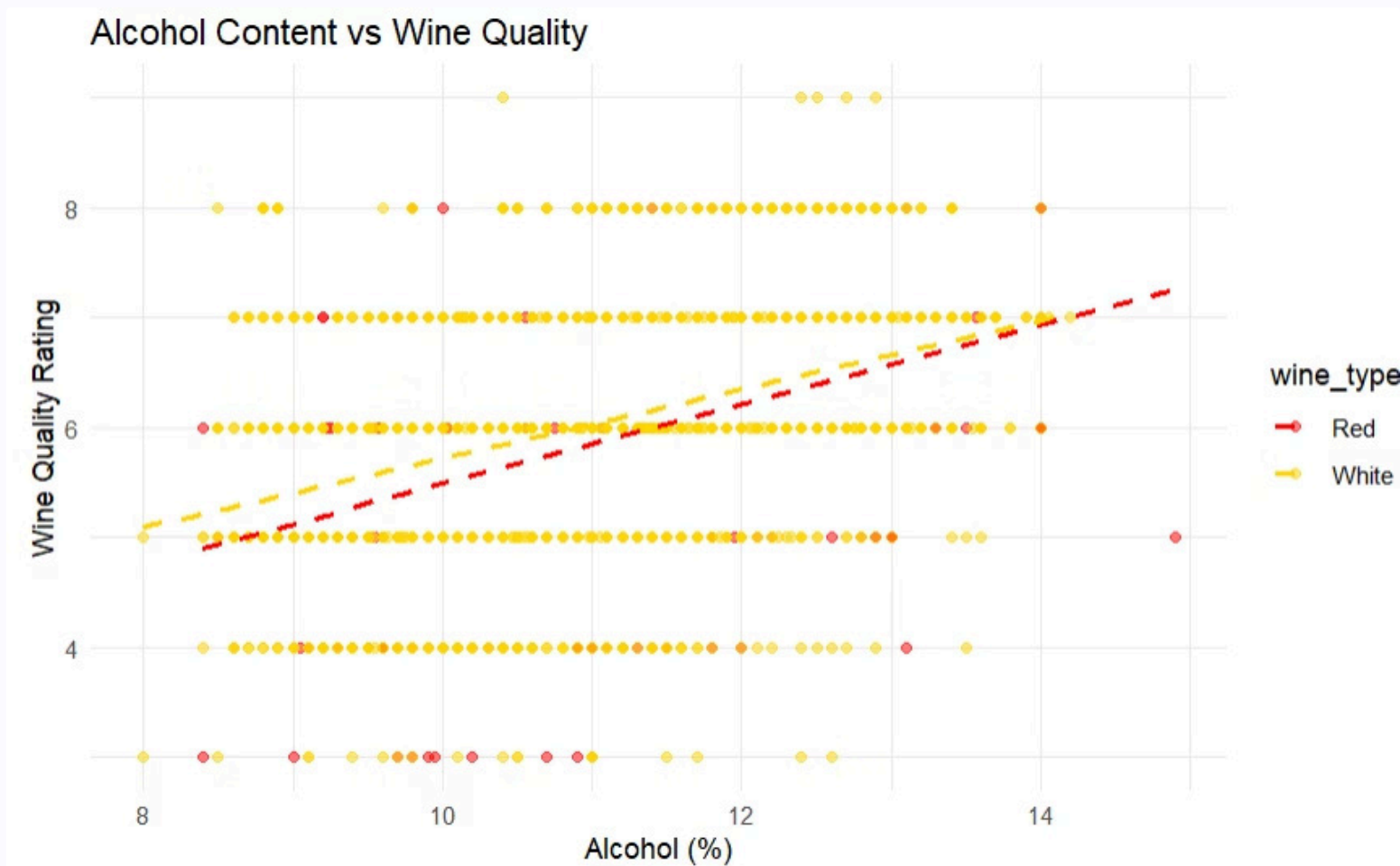


correlation matrix heatmap for the White Wine Quality dataset, specifically visualizing the relationships among physicochemical variables



- **Alcohol vs. Density**
 - ▼ **Strong negative correlation (~ -0.6 to -0.7)**
 - As **alcohol content increases**, **density decreases**.
 - Alcohol is less dense than water, so this is scientifically consistent.
 - **Interpretation:** Alcohol is a good quality indicator; higher alcohol tends to be associated with higher quality wine.
- **Alcohol vs. Residual Sugar & Total Sulfur Dioxide**
 - ▼ **Weak to moderate negative correlations**
 - Wines with higher alcohol tend to have **lower residual sugar** and **lower sulfur dioxide**.
 - **Decision Impact:** Excess residual sugar or sulfur dioxide might reduce consumer satisfaction or quality perception.

Scatter plot shows the relationship between alcohol content and wine quality rating for both red and white wines, including regression trend lines.



Comparison: Red vs White Wine Trends

- The **slopes of the trend lines** for red and white wines are **comparable**, suggesting a consistent relationship between alcohol and quality across wine types.
- White wine has a broader range in both alcohol percentage and quality scores, especially in the 11%–13% alcohol range.
- **Implication:** Despite differences in base composition, alcohol content influences quality similarly in both red and white wines—this allows for **cross-product optimization strategies**.

A Welch Two Sample t-test was used to compare the mean alcohol content between two wine quality groups: High and Low.

Welch Two Sample t-test

data: alcohol by quality_group

t = 17.45, df = 283.78, p-value < 2.2e-16

alternative hypothesis: true difference in means between group High and group Low is not equal to 0

95 percent confidence interval:


1.124097 1.409927

sample estimates:

mean in group High	mean in group Low
11.51805	10.25104

- The **extremely small p-value** (< 0.000001) suggests that the difference in alcohol content between high- and low-quality wines is **statistically significant**.
- **Conclusion:** Reject the null hypothesis. There is **strong evidence** that alcohol content differs between high- and low-quality wines.

Effect Size: Mean Difference

- Mean (High-quality group) = 11.52%
- Mean (Low-quality group) = 10.25%
-  The **difference in means = 1.27%**, which is **practically meaningful** in the context of winemaking.

A ~1.3% higher alcohol content is associated with significantly better wine quality perception.

Confidence Interval (95%)

- CI: [1.1241, 1.4099]
- This interval does **not contain 0**, confirming statistical significance.
- It estimates that wines rated as high quality have **1.12–1.41% more alcohol** than lower-quality ones.



- The **black dots** represent the **mean difference in alcohol content** between high- and low-quality wines.
- The **red error bars** span the **95% confidence interval** of that mean difference.
- Both red and white wine types show **positive mean differences**, confirming that high-quality wines tend to have **more alcohol** than low-quality ones.

Red Wine: Mean difference ~ **1.28%**

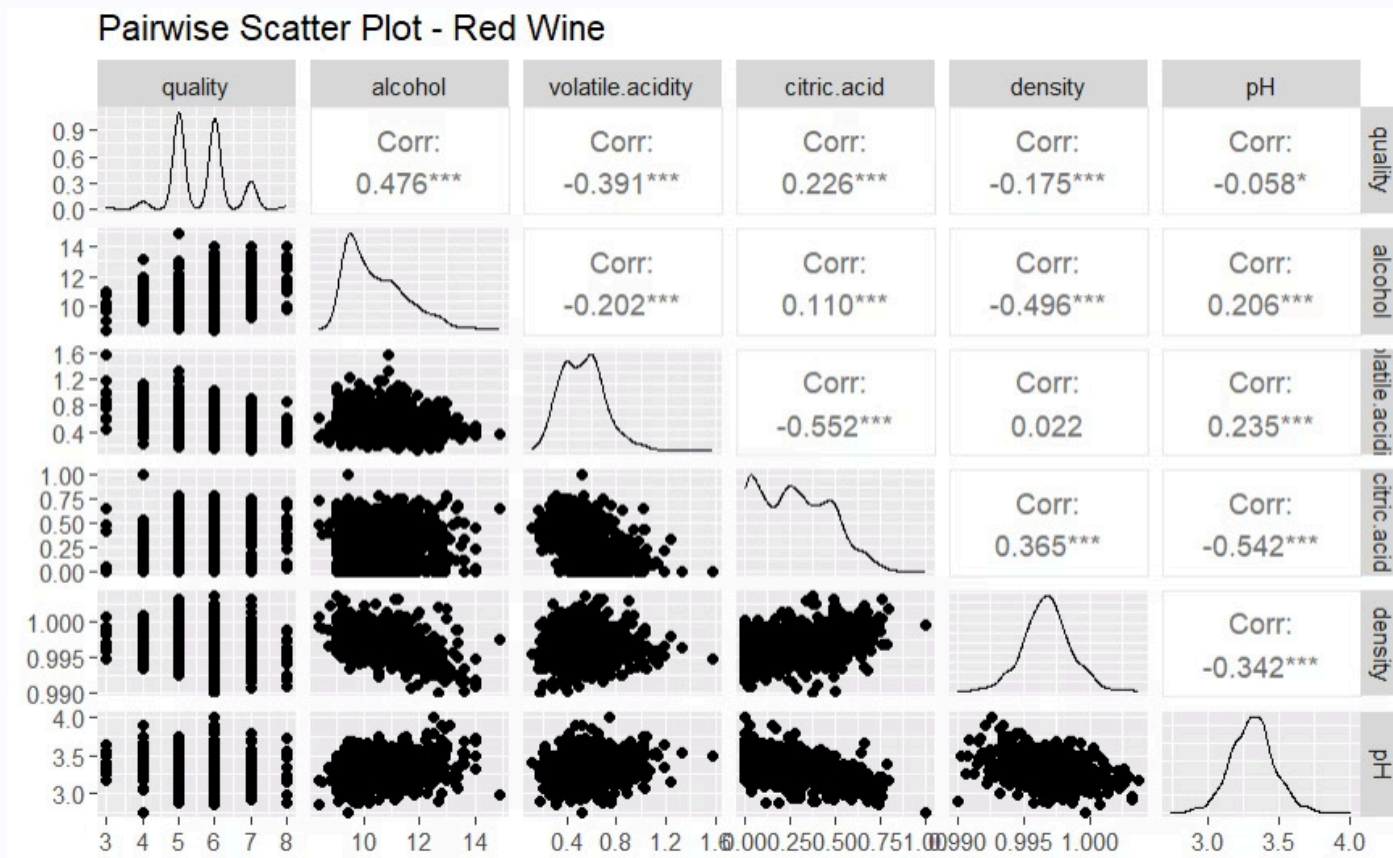
White Wine: Mean difference ~ **1.21%**

Non-Overlapping with Zero

- The entire confidence intervals for both wine types are **above 1%** and **do not include 0**, reinforcing:
 - **Statistical significance**
 - **Practical relevance**

This affirms the finding from the Welch t-test: the alcohol content **contributes meaningfully to wine quality**.

The pairwise scatter plot matrix for Red Wine provides rich exploratory data analysis (EDA)



- ✓ These pairwise scatter plots highlight trends and relationships across features.
- 💡 Look for patterns where wine quality is higher (e.g., high alcohol, low acidity).

Quality vs Alcohol

- Correlation = 0.476* (moderate positive)
- Alcohol content is **the strongest positive predictor** of red wine quality.
- This reinforces earlier findings: **higher alcohol = better quality ratings.**

Quality vs Volatile Acidity

- Correlation = -0.391* (moderate negative)
- Higher volatile acidity is associated with **lower quality.**
- Volatile acidity relates to spoilage or vinegary taste; reducing it can improve quality.

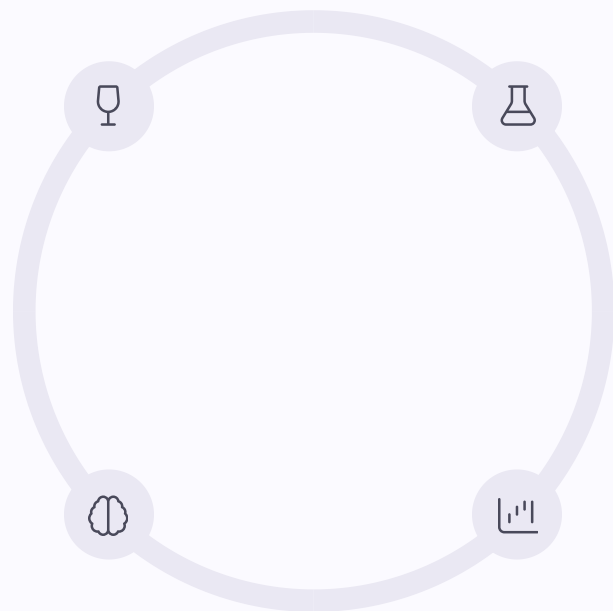
Why This Matters

Wine Quality Control

Enhance production decisions through statistical validation.

Data Interpretation

Better understand uncertainty and variability in results.



Food Science

Apply techniques to product testing and development.

Consumer Analytics

Improve sensory studies with robust statistical methods.

R Code Implementation



Load Libraries & Data

Import boot, dplyr packages and wine quality dataset



Group by Quality

Categorize wines as "High" or "Low" quality



Define Function

Create function to compute mean alcohol difference



Run Bootstrap

Apply with 2000 resamples and calculate confidence intervals

Discussion Points

Parametric vs. Bootstrap

When should we prefer bootstrap over traditional parametric tests?

Real-world Applications

How have you applied these methods in your own work?



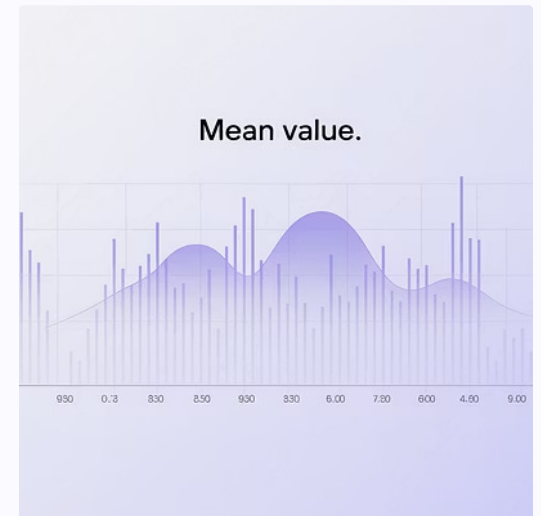
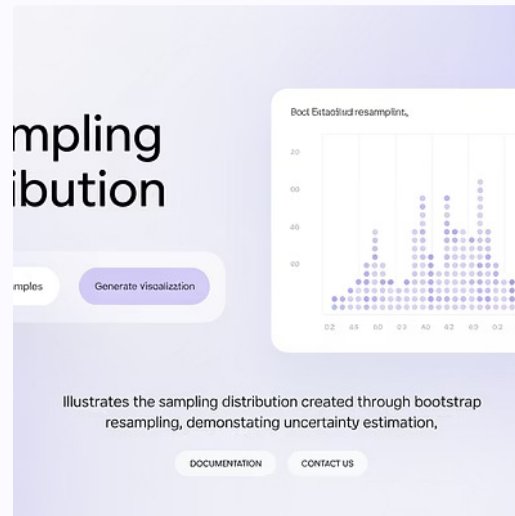
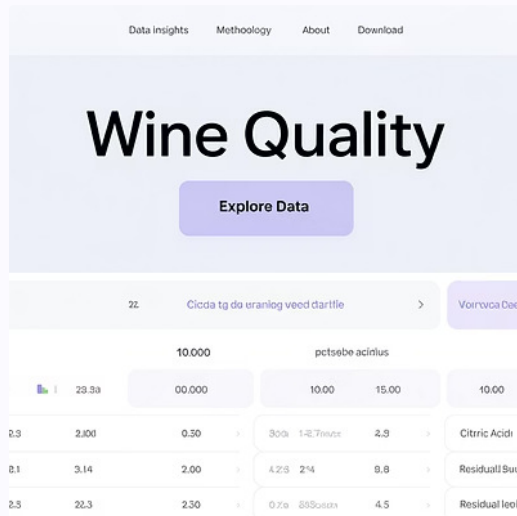
Confidence Level Selection

How does confidence level choice impact decision making?

Computational Considerations

What are the trade-offs between simplicity and computational intensity?

Helpful Resources



Explore these resources to deepen your understanding of bootstrap methods and their applications in R programming.

Connect & Learn More

3+

Statistical Methods

Bootstrap, parametric tests, and confidence intervals for robust analysis.

8+

Wine Features

Key physicochemical properties that influence quality ratings.

2000

Resamples

Recommended minimum bootstrap iterations for reliable results.

95%

Confidence Level

Standard statistical threshold for meaningful inference.

