OGTIP
DATA
SCIENCE
INTERNSHIP

UNCORKED INSIGHTS
WINE QUALITY ANALYSIS

JOSEPH ROBINSON

07 AUG 24

CONTENT

OVERVIEW

Objective: Understand the factors influencing red and white wine quality using statistical and machine learning methods.

Data: Two datasets (red and white wine) with 11 physicochemical attributes and a quality rating (0-10).

Methodology: Descriptive statistics, t-tests, confidence intervals, and logistic regression.

ASSUMPTIONS & DISCLAIMERS

Data Set Limitations:

- **Wine Origin:**
  - The wine samples used in this study only consist of Portuguese “Vinho Verde“ variants
  - This limits the significance of population inferences, since all samples come from the same region

- **Class Distribution:**
  - Class distribution is imbalanced
    - There are more “normal” wines than high/low quality

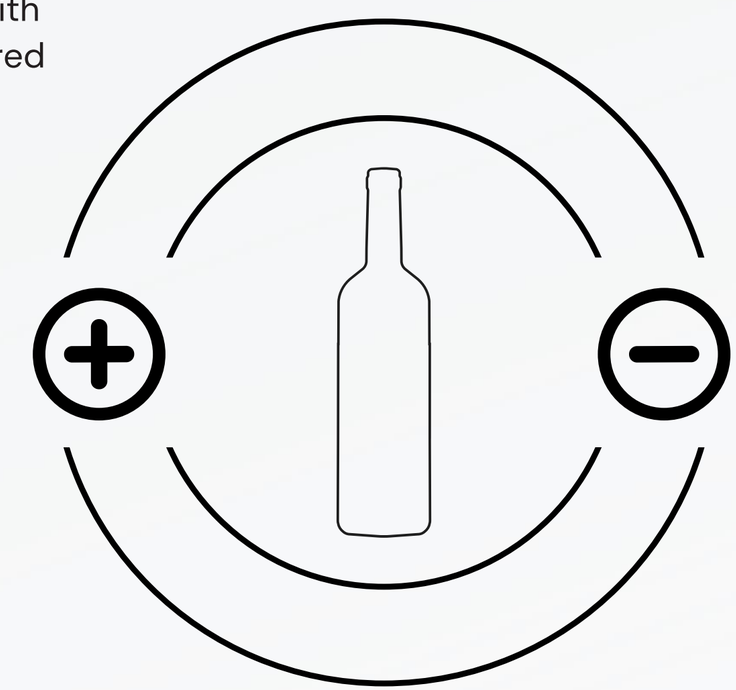
- **Limited variables:**
  - There is no data for grape type, wine type, wine prices, or sales
  - This reduces the scope and impact of findings, as they relate to the business task

KEY INSIGHTS

Common Wine Quality Correlates

Positive Correlates:

- Higher levels of **sulphates** and **alcohol** generally correlate with **higher quality** wine for both red and white varieties.



Negative Correlates:

- Higher levels of **volatile acidity** and **chlorides** are associated with **lower quality** wine for both red and white varieties.

RED WINE INSIGHTS

Statistical Inferences

Correlates:

- **Positive:** Sulphates, Alcohol
- **Negative:** Volatile Activity, Chlorides
- **Negligible:** Citric Acid, Fixed Acidity, Residual Sugar, Total Sulfur Dioxide, pH

Confidence Intervals (95%):

- **Alcohol content:** [10.78 to 10.92]
- **Citric Acid:** [0.28 to 0.31]
- **Sulphates:** [0.68 to 0.70]
- **Volatile Activity:** [0.46 to 0.48]
- **Density:** [0.9963 to 0.9966]
- **Chlorides:** [0.80 to 0.85]

Logistic Regression Analysis

Model Overview:

- Retained **moderate** explanatory power (25%) with pseudo-R-Squared value of 0.2504.
- Model has **strong** predictive power due to the extremely low LLR p-value (1.988e-112).

High Quality Red Wine factors based on Odds Ratio:

- **Strong Positive Impact:**
  - Sulphates: Higher levels increase quality odds by 1443%.
  - Alcohol: Higher content increases quality odds by 148%.
- **Strong Negative Impact:**
  - Volatile Acidity and Chlorides: Decrease quality odds significantly.
- **Weak Negative Impact:**
  - Free Sulfur Dioxide: Slightly positive impact.
- **Negligible Impact:**
  - Citric Acid, Fixed Acidity, Residual Sugar, Total Sulfur Dioxide, pH.

WHITE WINE INSIGHTS

Statistical Inferences

Correlates:

- **Positive**: Sulphates, Alcohol, Free Sulfur Dioxide
- **Negative:** Volatile Activity, Density, Chlorides
- **Negligible:** Citric Acid, Fixed Acidity, Residual Sugar, Total Sulfur Dioxide, pH

Confidence Intervals (95%):

- **Alcohol content:** [11.34 to 11.49]
- **Free Sulfur Dioxide:** [33.71 to 35.38]
- **Sulphates:** [0.492 to 0.508]
- **Volatile Activity:** [0.25 to 0.27]
- **Density:** [0.9922 to 0.9925]
- **Chlorides:** [0.0374 to 0.0388]

Logistic Regression Analysis

Model Overview:

- Retained **mild** explanatory power (18%) with pseudo-R-Squared value of 0.1804.
- Model has **strong** predictive power due to the extremely low LLR p-value (7.632e-192).

Factors impacting white wine quality based on odds ratios:

- **Moderate Positive Impact:**
  - Alcohol (86% increase), Sulphates (29% increase).
- **Weak Positive Impact:**
  - pH, Free Sulfur Dioxide.
- **Strong Negative Impact:**
  - Volatile Acidity (92% decrease), Chlorides (93% decrease).
- **Moderate Negative Impact:**
  - Citric Acid (59% decrease).
- **Weak Negative Impact:**
  - Fixed Acidity, Residual Sugar, Total Sulfur Dioxide.

KEY TAKEAWAYS

1. At first blush, both wines seem to have similar general results.

2. Upon testing, Red Wine revealed more compelling and straightforward insights.

3. White wine was found to be more nuanced in terms of how its variables relate to “quality”.

4. Multinomial Logistic Regression would offer a comprehensive analysis of quality factors for both wines.

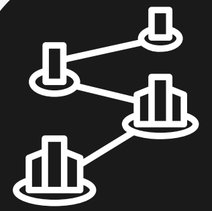
RECOMMENDATIONS

ENHANCE MODEL PERFORMANCE

Optimize through feature engineering and selection.

DEEPEN UNDERSTANDING

Investigate anomalies in the density variable.



EXPAND MODEL COMPARISON

Compare logistic regression with other classification algorithms (e.g., multinomial, random forest, support vector machines).

UNCOVER RELATIONSHIPS

Explore potential interactions between variables.

NEXT STEPS

INCORPORATE ADDITIONAL FEATURES

Include grape type, wine type, price, and sales data in future analysis.

LEVERAGE TIME SERIES DATA

Utilize time series data (dates, months, quarters) for trend analysis.

OGTIP
DATA
SCIENCE
INTERNSHIP

THANK YOU