OGTIP
DATA
SCIENCE
INTERNSHIP

UNCORKED INSIGHTS
WINE QUALITY ANALYSIS

JOSEPH ROBINSON

28 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).

Analytical Approach: Descriptive and inferential statistics, t-tests, confidence intervals, and linear regression analysis.

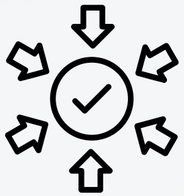
EXECUTIVE SUMMARY

Project: Linear Regression Analysis of Red and White Wine Quality

Objective: Develop predictive models to estimate wine quality based on chemical composition.

Key Findings

- Factors influencing wine quality: volatile activity, chlorides, density, pH, sulphates, and alcohol.
- Variations in importance of these factors between red and white wines.
- Defined optimal ranges for high-quality wines.



Business Implications

- Optimize wine making processes.
- Develop targeted marketing campaigns.
- Establish quality standards for procurement.
- Implement quality-based pricing strategies.

Recommendations

- Expand research to include different regions and factors.
- Refine models and address limitations.
- Collect more comprehensive data.

DATA

Source:

- Cortez, A. Cerdeira, F. Almeida, T. Matos & J. Reis. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.

Details:

- Two data sets (Red and White wine) of Portuguese “Vinho Verde” variants
  - Merged both data sets
  - Created **type** variable
  - Only physicochemical (inputs) and sensory (output) variables
  - **Number of Instances:** red wine - **1599**; white wine - **4898**
- **6497** rows and **13** columns (after merge)
- All numerical data except created **type** feature

Key Metric:

- Target variable **quality** rated **(0 to 10)**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixedacidity          6497 non-null   float64
 1   volatileacidity       6497 non-null   float64
 2   citricacid            6497 non-null   float64
 3   residualsugar         6497 non-null   float64
 4   chlorides             6497 non-null   float64
 5   freesulfurdioxide     6497 non-null   float64
 6   totalsulfurdioxide    6497 non-null   float64
 7   density               6497 non-null   float64
 8   pH                    6497 non-null   float64
 9   sulphates             6497 non-null   float64
 10  alcohol               6497 non-null   float64
 11  quality               6497 non-null   int64
 12  type                  6497 non-null   object
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
```

ANALYTICAL APPROACH

Methodology:

- **Descriptive Statistics:** Data exploration, Summary Statistics
- **Inferential Statistics:** T-tests, Confidence Intervals
- **Linear Regression Analysis**: Model development, Model Interpretation, and Residual Analysis

Tools:

- **Jupyter Notebook:** Interactive environment for data analysis
- **Python Libraries:** NumPy, Pandas, Scikit-learn, Statsmodels, Matplotlib, Seaborn

Process:

1. **Data Acquisition:** Collect and clean data.
2. **EDA**: Explore and analyze data.
3. **Feature Engineering:** Create or transform features.
4. **Split Data:** Divide dataset into training and testing sets.
5. **Model Development:** Build linear regression models.
6. **Model Evaluation:** Assess model performance using metrics.
7. **Residual Analysis:** Check for model assumptions.
8. **Interpretation and Insights:** Analyze results and draw conclusions.
9. **Recommendations:** Provide recommendations based on findings.

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

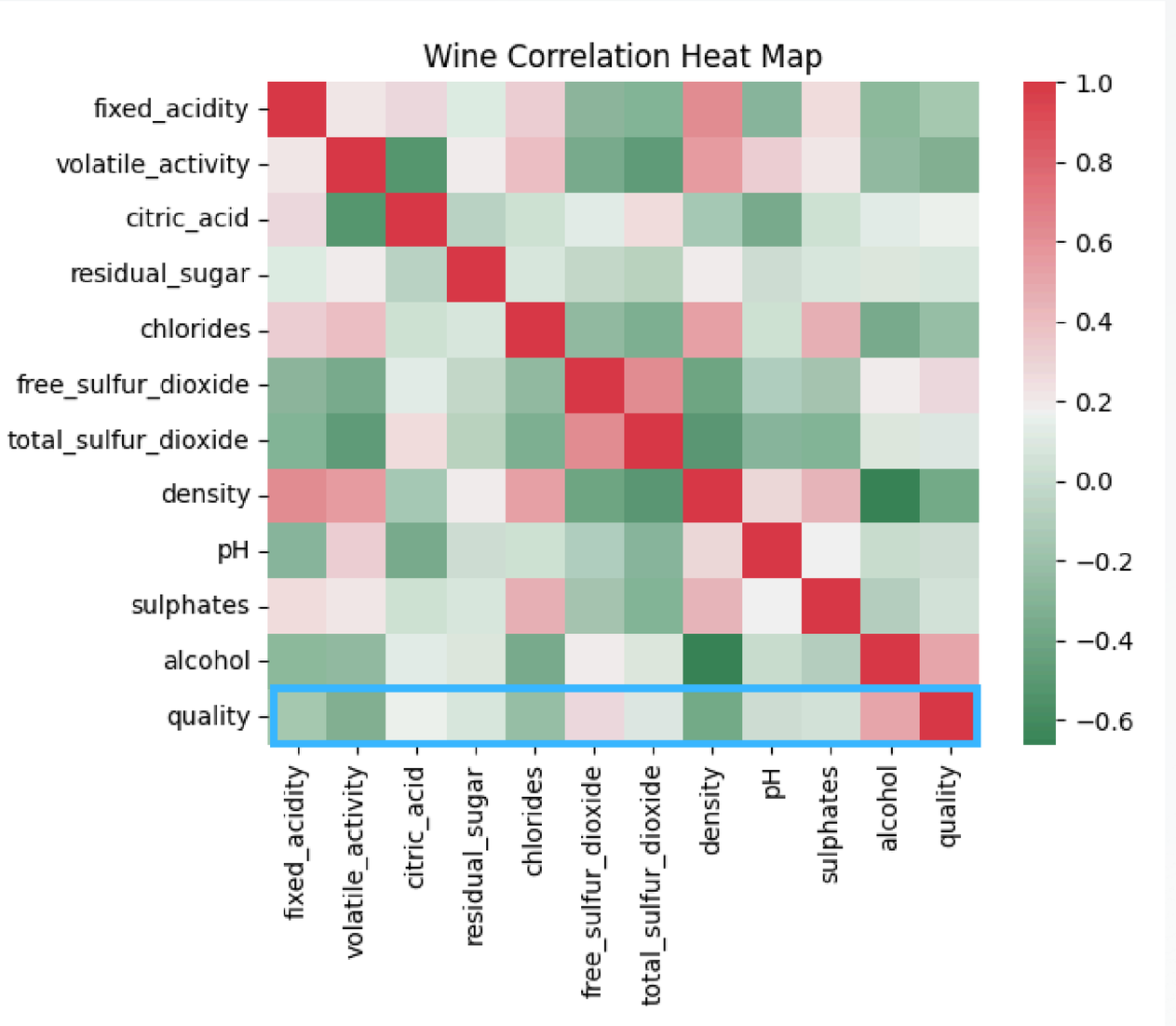
- **Limited variables:**
  - There is no data for grape type, wine type, wine prices, or sales
  - Target variable “Quality” is based on sensory input (subjective)
    - Quality grading criterion is not provided
    - Sommelier credentials not provided
  - This reduces the scope and impact of findings, as they relate to the business task

COMMON INSIGHTS

Wine Quality Correlates

```
Wine Quality Correlations Sorted:

quality                1.000000
alcohol                0.506969
free_sulfur_dioxide    0.268535
citric_acid            0.157989
total_sulfur_dioxide   0.098717
residual_sugar         0.072992
sulphates              0.056488
pH                     0.021752
fixed_acidity         -0.160528
chlorides             -0.219614
volatile_activity     -0.318096
density               -0.376687
Name: quality, dtype: float64
```



RED WINE INSIGHTS

Statistical Inferences & Linear Regression Analysis

Model Overview:

- Retained **moderate** explanatory power (**39.7%**) with R-Squared value of 0.397.
- Model has **strong** predictive power due to the **extremely low** Prob (F-Statistic) of 1.33e-133.

Key Predictors:

- Volatile activity, chlorides, density, pH, sulphates, alcohol, and acidity ratio significantly influence red wine quality.

Relationships:

- **Higher** sulphates & alcohol, **lower** volatile activity, chlorides, density, pH, and acidity ratio correlate with higher quality.

Confidence Intervals (95%):

- **Positive**: Alcohol [10.35-10.47], Sulphates [0.64-0.66]

- **Negative**: Volatile Activity [0.52-0.54], Chlorides [0.9962-0.9964], Fixed Acidity [44.82-48.23]

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.397
Model:	OLS	**Adj. R-squared:**	0.393
Method:	Least Squares	**F-statistic:**	93.48
Date:	Tue, 27 Aug 2024	**Prob (F-statistic):**	1.10e-133
Time:	14:41:34	**Log-Likelihood:**	-1075.7
No. Observations:	1286	**AIC:**	2171.
Df Residuals:	1276	**BIC:**	2223.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.0236	15.214	-0.527	0.598	-37.871	21.824
volatile_activity	-0.6770	0.119	-5.671	0.000	-0.911	-0.443
residual_sugar	0.0022	0.036	0.062	0.951	-0.068	0.072
chlorides	-1.3161	0.396	-3.326	0.001	-2.092	-0.540
density	11.5187	15.160	0.760	0.448	-18.222	41.260
pH	-0.3045	0.122	-2.493	0.013	-0.544	-0.065
sulphates	0.7906	0.110	7.202	0.000	0.575	1.006
alcohol	0.3238	0.024	13.664	0.000	0.277	0.370
acidity_ratio	-0.7579	1.060	-0.715	0.475	-2.837	1.322
total_sulfur	-0.0051	0.001	-5.697	0.000	-0.007	-0.003

WHITE WINE INSIGHTS

Statistical Inferences & Linear Regression Analysis

Model Overview:

- Retained **moderate** explanatory power (**28.4%**) with R-Squared value of 0.284.
- Model has **strong** predictive power due to the **extremely low** Prob (F-Statistic) of 6.14e-118.

Key Predictors:

- Alcohol, volatile activity, chlorides, density, pH significantly influence white wine quality.

Relationships:

- **Higher** alcohol, **lower** volatile activity, chlorides, density, pH correlate with higher quality.

Confidence Intervals (95%):

- **Positive**: Alcohol [11.34-11.46]
- **Negative**: Density [0.9912-0.9913], Chlorides [0.039-0.042], Fixed Acidity [6.66-6.75]

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.284
Model:	OLS	**Adj. R-squared:**	0.281
Method:	Least Squares	**F-statistic:**	84.49
Date:	Tue, 27 Aug 2024	**Prob (F-statistic):**	6.14e-118
Time:	14:51:26	**Log-Likelihood:**	-1796.2
No. Observations:	1711	**AIC:**	3610.
Df Residuals:	1702	**BIC:**	3659.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8871	0.564	1.573	0.116	-0.219	1.994
volatile_activity	-1.0353	0.189	-5.486	0.000	-1.405	-0.665
citric_acid	0.0237	0.170	0.139	0.889	-0.310	0.358
residual_sugar	0.0992	0.019	5.214	0.000	0.062	0.137
chlorides	-1.1079	0.906	-1.223	0.222	-2.885	0.669
pH	0.6902	0.125	5.520	0.000	0.445	0.935
alcohol	0.3201	0.017	19.233	0.000	0.287	0.353
combined_fixed_density	-0.1359	0.045	-2.996	0.003	-0.225	-0.047
combined_sulfur	0.0006	0.001	0.947	0.344	-0.001	0.002

KEY TAKEAWAYS

1. There is a **significant** statistical difference between Red and White wines based on how their chemical attributes relate to **quality** score.

2. Upon testing, Red Wine aligned more with the Linear Regression Model, revealing more compelling and straightforward insights.

3. White wine was found to be more nuanced in terms of how its variables relate to **quality** score, underlining the need for further testing.

4. Further addressing multicollinearity issues, combining features, and exploring different models would be key to finding new insights.

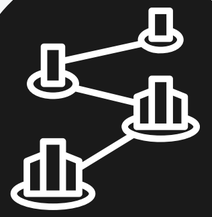
RECOMMENDATIONS

ENHANCE MODEL PERFORMANCE

Optimize through feature engineering and selection.

DEEPEN UNDERSTANDING

Investigate multicollinear, non-linear, and heterodastic relationships.



EXPAND MODEL COMPARISON

Compare linear regression with other regression models (i.e., ridge regression, lasso regression, elastic net).

UNCOVER RELATIONSHIPS

Explore potential interactions between variables and how they relate to quality.

NEXT STEPS

INCORPORATE ADDITIONAL FEATURES

Include grape type/origin, 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